Quantitative Economics 9 (2018), 441–482 1759-7331/20180441

Simultaneous selection of optimal bandwidths for the sharp regression discontinuity estimator

Yoichi Arai
School of Social Sciences, Waseda University

Hidehiko Ichimura
Department of Economics, University of Tokyo

A new bandwidth selection method that uses different bandwidths for the local linear regression estimators on the left and the right of the cut-off point is proposed for the sharp regression discontinuity design estimator of the average treatment effect at the cut-off point. The asymptotic mean squared error of the estimator using the proposed bandwidth selection method is shown to be smaller than other bandwidth selection methods proposed in the literature. The approach that the bandwidth selection method is based on is also applied to an estimator that exploits the sharp regression kink design. Reliable confidence intervals compatible with both of the proposed bandwidth selection methods are also proposed as in the work of Calonico, Cattaneo, and Titiunik (2014a). An extensive simulation study shows that the proposed method's performances for the samples sizes 500 and 2000 closely match the theoretical predictions. Our simulation study also shows that the common practice of halving and doubling an optimal bandwidth for sensitivity check can be unreliable.

Keywords. Bandwidth selection, local linear regression, regression discontinuity design, regression kink design, confidence interval.

JEL classification. C13, C14, C21.

1. Introduction

The regression discontinuity (RD) is a quasi-experimental design to evaluate causal effects introduced by Thistlewaite and Campbell (1960) and developed by Hahn, Todd,

Yoichi Arai: yarai@waseda.jp
Hidehiko Ichimura: ichimura@e.u-tokyo.ac.jp

Earlier versions of this paper were titled "Optimal bandwidth selection for differences of nonparametric estimators with an application to the sharp regression discontinuity design" and presented at Academia Sinica, the Japanese Economic Association Spring Meeting, LSE, the North American Winter Meeting of the Econometric Society, UC Berkeley, UCL, and Yale. Valuable comments were received from seminar participants. We are especially grateful to Yoshihiko Nishiyama, Jack Porter, Jim Powell, and three anonymous reviewers for many helpful comments. We also thank Jens Ludwig and Douglas Miller for making the data used in Ludwig and Miller (2007) publicly available. Yoko Sakai and Masaru Nagashima provided expert research assistance. This research was supported by Grants-in-Aid for Scientific Research 22243020, 23330070, and 15H05692 from the Japan Society for the Promotion of Science.

https://doi.org/10.3982/QE590

and Van der Klaauw (2001). A large number of empirical studies are carried out using the RD design in various areas of economics. See Imbens and Lemieux (2008), Van der Klaauw (2008), Lee and Lemieux (2010), and DiNardo and Lee (2011) for an overview and lists of empirical researches. The RD approach has been extended in various directions. For example, Card, Lee, Pei, and Weber (2015) and Dong and Lewbel (2015) consider the regression kink design and its related model, and Frandsen, Frörich, and Melly (2012) and Chiang and Sasaki (2016) consider the quantile treatment effect in the context of the RD and RK designs, respectively.

We first consider the sharp RD design in which whether a value of the assignment variable exceeds a known cut-off point or not determines the treatment status. A parameter of interest is the average treatment effect at the cut-off point. The average treatment effect is given by the difference between the two conditional mean functions at the cut-off point. This implies that estimating the treatment effect amounts to estimating two functions at the boundary point. One of the most frequently used estimation methods is the local linear regression (LLR) because of its superior performance at the boundary. See Fan (1992, 1993) and Porter (2003).

A particular nonparametric estimator, in general, is undefined unless the smoothing parameter selection method is specified, and it is well recognized that choosing an appropriate smoothing parameter is a key implementation issue (Ichimura and Todd (2007)). In the RD setting, it amounts to choosing a bandwidth for each of the LLR estimators at two sides of the cut-off point. Therefore, in the context of RD design, using two bandwidths for estimating two functions is a natural approach. In fact, DesJardins and McCall (2008) propose to use the plug-in method for each side of the cut-off point. However, each of the two bandwidths are chosen optimally to estimate the conditional mean functions considered separately. But the target function is the difference of the two conditional mean functions, which corresponds to the average treatment effect at the cut-off point.

Figure 1 illustrates the situation motivated by Ludwig and Miller (2007) where the cut-off value is depicted by a dotted vertical line. The solid lines depict two conditional mean functions to estimate. And the dashed curve denotes the estimated density of the assignment variable. One can see that the curvatures of the conditional mean functions for the treated and the untreated in the vicinity of the cut-off point differ significantly. This is not an exceptional case but arises naturally in many empirical studies. For example, sharp contrasts in curvatures are observed in Figures 1 and 2 in Ludwig and Miller (2007), Figures IV and V in Card, Mas, and Rothstein (2008), Figures 12 and 14 in DesJardins and McCall (2008), Figures 3 and 5 in Lee (2008), and Figures 1 and 2 in Hinnerich and Pettersson-Lidbom (2014), among others.

The most widely used bandwidth selection methods in the RD context are the method by Imbens and Kalyanaraman (2012) (hereafter IK) and its modification by Calonico, Cattaneo, and Titiunik (2014a) (hereafter CCT).[1] These methods choose the

[1] Calonico, Cattaneo, and Titiunik (2014a) is the first to propose robust confidence intervals for estimators exploiting the sharp design, the fuzzy RD design, or the RK design.

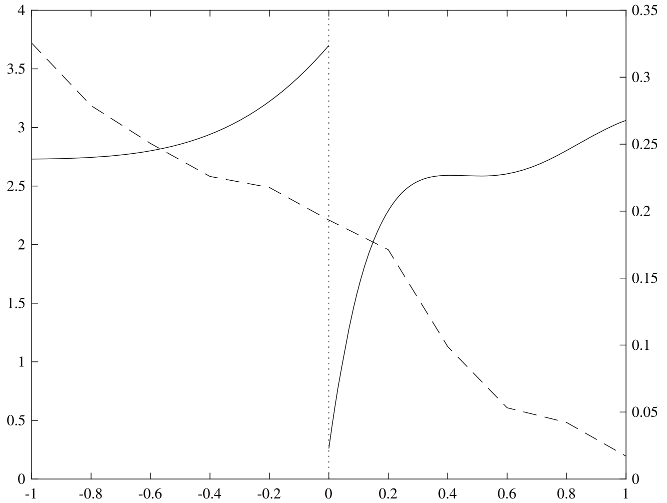


FIGURE 1. Conditional mean functions of outcomes based on Ludwig and Miller (2007). The solid line on the left of the cut-off point, zero, depicts the conditional mean function of the potential outcome for untreated conditional on the assignment variable. Similarly, the solid line on the right of the cut-off point draws the corresponding function for treated. The dashed line depicts the kernel density estimate of the assignment variable based on the rule-of-thumb bandwidth. The left and right vertical axes denote the solid line and the dashed line, respectively.

optimal bandwidth for the difference of the two conditional mean functions, but imposing that the bandwidths are the same for two sides of the cut-off point.[2,3]

If we use the same bandwidth for both sides of the cut-off point, which is relatively large, in the situation described by Figure 1, it will incur a large bias on the right of the cut-off point, for estimating the conditional mean function. On the other hand, using a single bandwidth, which is relatively small, will lead to a smaller bias on the right while it will generate a large variance on the same side as the sample size is about half that on the left of the cut-off point. While a distinct-bandwidth approach would choose a modest bandwidth on the right and a large bandwidth on the left of the cut-off point, a single-bandwidth approach such as IK or CCT tend to choose a bandwidth in between, leading to the larger bias on the right and the larger variance on the left of the cut-off point.

We develop an optimal bandwidth selection method that uses two distinct bandwidths for two sides of the cut-off point taking into account that the target parameter is the difference of two conditional mean functions.

[2] Ludwig and Miller (2005, 2007) used the cross validation method to select a single bandwidth. However, they did not consider an objective function that corresponds to the asymptotic mean square error of the difference of the two conditional mean functions at the cut-off point.

[3] A single bandwidth approach is familiar to empirical researchers in the applications of matching methods (Abadie and Imbens (2011)) since the supports of covariates for treated and untreated individuals overlap and we wish to construct two comparable groups. This reasoning does not apply to the RD estimator since values of the assignment variable never overlap due to the structure of the RD design.

Our theoretical results and simulation results show that the proposed bandwidth selection method produces more accurate point estimates relative to the existing bandwidths especially when the curvatures or the sample sizes on each side of the cut-off point differ significantly. Even when the curvatures and the sample sizes are very similar on each side, the performance of the proposed bandwidth selection method performs reasonably well especially when the sample size is large. Hence there would be almost no loss but a gain for employing the proposed method.

In addition, we show the importance of using an optimal bandwidth through an extensive simulation work and an empirical application. A very popular practice in dealing with the bandwidth selection is to report results of a sensitivity analysis using different bandwidths. Often, after reporting results using an "optimal bandwidth," results using half the bandwidth and those using double the bandwidth are reported. Through a simulation study, we show that this approach is unreliable. That is, we show that this approach results in a large loss in efficiency in terms of mean squared error. An optimal bandwidth typically minimizes an approximation to the mean squared error of an estimator and halving or doubling the bandwidths could lead to a large deviation from the optimal point. Hence the practice should not be appropriate as the robustness check. Our suggestion is to use different "optimal" bandwidths to conduct sensitivity analysis. The proposed method complements the existing methods in the sense that it offers an additional option for the robustness check which is rooted in the different principle from the existing ones.

We also consider a bandwidth selection method for the case of the sharp regression kink (RK) design, a term coined by Nielsen, Sørensen, and Taber (2010) and extensively developed by Card et al. (2015) (hereafter CLPW). We show that this case has a similar structure with the case of the RD design and apply the proposed approach to the sharp RK design. Following the point estimation, an important next step is the statistical inference. As emphasized by CCT, the confidence interval without bias correction using the conventional standard error is not asymptotically valid when it is combined with the optimal bandwidths. Although the conventional confidence interval with bias correction is asymptotically valid, CCT also show via simulation that the resulting coverage probability tends to be much lower than the nominal one. To overcome these issues, they constructed a novel method to construct a reliable confidence interval for the estimators that use a single bandwidth. We show that the approach by CCT can be extended to accommodate the estimators that use the proposed bandwidth selection methods that exploit the sharp RD or the sharp RK designs by constructing the CCT-type robust confidence intervals.

In Section 2 of this paper, we propose a new bandwidth selection method suitable for the RD context. Our asymptotic analysis shows that the proposed method dominates the currently available methods in terms of the asymptotic mean square error (AMSE). In Section 3, we extend the approach to the regression kink (RK) design and consider the CCT-type robust confidence interval. In Section 4, we report results from an extensive simulation work and show that the asymptotic advantages theoretically derived in Section 2 realize in finite sample sizes relevant for empirical works. We also report an empirical illustration in Section 5. Section 6 concludes the paper. In the Appendix, we include the proofs of the main theorems. All the omitted proofs and detailed implementation procedures are provided in the Supplemental Material (available in a supplementary file on the journal website, http://qeconomics.org/supp/590/supplement.pdf).[4]

2. Bandwidth selection of the sharp regression discontinuity estimators

For observation i, we denote potential outcomes with and without treatment by $Y_i(1)$ and $Y_i(0)$, respectively. Let D_i be a binary variable which takes the value 0 or 1 indicating the treatment status. The observed outcome, Y_i, can be written as $Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0)$. In the sharp RD setting, we consider, the treatment status is determined solely by the assignment variable, denoted by X_i: $D_i = \mathbb{I}\{X_i \geq c\}$ where c is a known cut-off point and $\mathbb{I}\{A\}$ takes value 1 if A holds and takes value 0 if A does not hold. Throughout the paper, we assume that $(Y_1, X_1), \dots, (Y_n, X_n)$ are independent and identically distributed observations and X_i has the Lebesgue density f.

Define $m_1(x) = E(Y_i(1)|X_i = x) = E(Y_i|X_i = x)$ for $x \geq c$ and $m_0(x) = E(Y_i(0)|X_i = x) = E(Y_i|X_i = x)$ for $x < c$. Suppose that the limits $\lim_{x\to c+} m_1(x)$ and $\lim_{x\to c-} m_0(x)$ exist where $x \to c+$ and $x \to c-$ mean taking the limits from the right and left, respectively. Denote $\lim_{x\to c+} m_1(x)$ and $\lim_{x\to c-} m_0(x)$ by $m_1(c)$ and $m_0(c)$, respectively. Then the average treatment effect at the cut-off point is given by $\tau_{\mathrm{SRD}}(c) = m_1(c) - m_0(c)$ and $\tau_{\mathrm{SRD}}(c)$ is the parameter of interest in the sharp RD design.

Estimation of $\tau_{\mathrm{SRD}}(c)$ requires to estimate two functions, $m_1(c)$ and $m_0(c)$. The nonparametric estimators that we consider are LLR estimators proposed by Stone (1977) and investigated by Fan (1992). For estimating these limits, the LLR is particularly attractive because it exhibits the automatic boundary adaptive property (Fan (1992, 1993), Hahn, Todd, and Van der Klaauw (2001), and Porter (2003)). The LLR estimator for $m_1(c)$ is given by $\hat{\alpha}_{h_1}(c)$, where

$$\bigl(\hat{\alpha}_{h_1}(c), \hat{\beta}_{h_1}(c)\bigr) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} \bigl\{Y_i - \alpha - \beta(X_i - c)\bigr\}^2 K\left(\frac{X_i - c}{h_1}\right)\mathbb{I}\{X_i \geq c\},$$

where $K(\cdot)$ is a kernel function and h_1 is a bandwidth. A standard choice of the kernel function for the RD estimators is the triangular kernel given by $K(u) = (1 - |u|) \times \mathbb{I}\{|u| < 1\}$ because of its MSE and minimax optimality (Cheng, Fan, and Marron (1997)). The LLR estimator for $m_0(c)$, $\hat{\alpha}_{h_0}(c)$, can be obtained in the same manner, except replacing $\mathbb{I}\{X_i \geq c\}$ with $\mathbb{I}\{X_i \leq c\}$. Denote $\hat{\alpha}_{h_1}(c)$ and $\hat{\alpha}_{h_0}(c)$ by $\hat{m}_1(c)$ and $\hat{m}_0(c)$, respectively. Then $\tau_{\mathrm{SRD}}(c)$ is estimated by $\hat{\tau}_{\mathrm{SRD}}(c) \equiv \hat{m}_1(c) - \hat{m}_0(c)$.

Before we start to discuss issues on difficulties of choosing two bandwidths simultaneously and to propose our optimal bandwidth selection method, we provide a quick exposition of our proposed bandwidths for the sharp RD design. A standard approach to choose a bandwidth for the average treatment effect at the cut-off point, $\hat{\tau}_{\mathrm{SRD}}(c)$, is to

[4] Matlab and Stata codes to implement the proposed methods are available as a supplementary file on the journal website, http://qeconomics.org/supp/590/code_and_data.zip, or at one of the authors' webpage, http://www.f.waseda.jp/yarai/.

minimize the AMSE of $\hat{\tau}_{\mathrm{SRD}}(c)$ given by[5]

$$\mathrm{AMSE}_n(h) = \left\{ \frac{b_1}{2} \left[m_1^{(2)}(c) h_1^2 - m_0^{(2)}(c) h_0^2 \right] \right\}^2 + \frac{v}{nf(c)} \left\{ \frac{\sigma_1^2(c)}{h_1} + \frac{\sigma_0^2(c)}{h_0} \right\},$$

where b_1 and v are constants solely determined by a kernel function, $m_j^{(s)}(c)$ is the sth derivative of $m_j(x)$ at c and $f(c)$ is the density of the assignment variable at c. For example, IK and CCT minimize this object with regularization under the assumption of $h_1 = h_0$. In contrast to the standard approach, the proposed bandwidths minimize the following modified version of the AMSE given by

$$\mathrm{MMSE}_n(h) = \left\{ \frac{b_1}{2} \left[m_1^{(2)}(c) h_1^2 - m_0^{(2)}(c) h_0^2 \right] \right\}^2 + \left\{ b_{2,1}(c) h_1^3 - b_{2,0}(c) h_0^3 \right\}^2 + \frac{v}{nf(x)} \left\{ \frac{\sigma_1^2(x)}{h_1} + \frac{\sigma_0^2(x)}{h_0} \right\},$$

where $b_{2,j}(c)$ is the object which depends on the constant determined by the kernel function, $f(c)$, and the second and third derivatives of $m_j(c)$ and $f(c)$ for $j = 0, 1$. Section 2.1 explains why the standard approach does not work through to choose two bandwidths simultaneously. We discuss, in Section 2.2, how the proposed bandwidths overcome the difficulties and show that they possess theoretically desirable properties relative to the existing methods.

2.1 *The AMSE for the regression discontinuity estimators*

In this paper, we propose a simultaneous selection method for two distinct bandwidths, h_1 and h_0, based on an AMSE. The use of the AMSE as an objective function is the standard approach in the literature.[6] In the standard case, the first-order AMSE formula gives a trade-off of using a narrower bandwidth versus a wider bandwidth. For example, the formula shows that when a narrower bandwidth is used, the bias term is smaller but the variance term is larger. However, because the target parameter is the difference of two conditional mean functions, this trade-off in the first-order AMSE formula can break down in the RD setting. We first describe this issue to motivate the objective function we ultimately use.

The conditional mean squared error (MSE) of the RD estimators of the average treatment effect given the assignment variable, X, is defined by

$$\mathrm{MSE}_n(h) = E\left[\left\{ \left[\hat{m}_1(c) - \hat{m}_0(c) \right] - \left[m_1(c) - m_0(c) \right] \right\}^2 | X \right],$$

where $X = (X_1, X_2, \dots, X_n)'$.[7]

[5] All of the assumptions and details are provided in the following sections. Here, we simply assume that all objects such as derivatives and variances exist.

[6] As IK emphasize, the bandwidth selection problem in the context of the RD setting is how to choose local bandwidths rather than global bandwidths. Thus, bandwidth selection based on either the asymptotic mean "integrated" squared errors or the cross-validation criterion can never be optimal.

[7] Throughout the paper, we use "h" without a subscript to denote a combination of h_1 and h_0, for example, $\mathrm{MSE}_n(h_1, h_0)$ is written as $\mathrm{MSE}_n(h)$. We assume that X is such that this conditional MSE is well-defined.

To describe the trade-off, we examine the first-order AMSE under the standard set of assumptions described below.

Assumption 1. *$K(\cdot):\mathbb{R}\to\mathbb{R}$ is a bounded and symmetric second-order kernel function that is continuous with compact support, that is, K satisfies the following: $K(u)\geq 0$ for any $u\in\mathbb{R}$, $\int_{-\infty}^{\infty}K(u)\,du=1$, $\int_{-\infty}^{\infty}uK(u)\,du=0$, and $\int_{-\infty}^{\infty}u^2K(u)\,du>0$.*

Also let $\mu_s=\int_0^{\infty}u^sK(u)\,du$ and $\nu_s=\int_0^{\infty}u^sK^2(u)\,du$ for the nonnegative integer s.

Assumption 2. *The positive sequence of bandwidths is such that $h_j\to 0$ and $nh_j\to\infty$ as $n\to\infty$ for $j=0,1$.*

Assumptions 1 and 2 are standard assumptions on the kernel functions and the bandwidths in the literature of regression function estimation as well as the RD design.

Let $\mathcal{D}$ be an open set in $\mathbb{R}$, k be a nonnegative integer, $\mathcal{C}_k$ be the family of k times continuously differentiable functions on $\mathcal{D}$, and $g^{(k)}(\cdot)$ be the kth derivative of $g(\cdot)\in\mathcal{C}_k$. Let $\mathcal{G}_k(\mathcal{D})$ be the collection of functions g such that $g\in\mathcal{C}_k$ and

$$\left|g^{(k)}(x)-g^{(k)}(y)\right|\leq M_k|x-y|^{\alpha},\quad x,y\in\mathcal{D}$$

for some positive M_k and some α such that $0<\alpha\leq 1$.

Let $\sigma_1^2(x)$ and $\sigma_0^2(x)$ denote the conditional variances of Y_1 and Y_0 given $X_i=x$, respectively, and let

$$\sigma_1^2(c)=\lim_{x\to c+}\sigma_1^2(x),\qquad \sigma_0^2(c)=\lim_{x\to c-}\sigma_0^2(x),\qquad m_1^{(s)}(c)=\lim_{x\to c+}m_1^{(s)}(x),$$

and

$$m_0^{(s)}(c)=\lim_{x\to c-}m_0^{(s)}(x).$$

The following assumptions are also standard regularity conditions on the underlying Lebesgue density of the assignment variable, conditional variance functions, and the conditional mean functions of the outcome variables, respectively.

Assumption 3. *The Lebesgue density of X_i, denoted f, is an element of $\mathcal{G}_1(\mathcal{D})$ where $\mathcal{D}$ is an open neighborhood of c and is bounded above and strictly positive on $\mathcal{D}$.*

Assumption 4. *Let δ be some positive constant. The conditional variance function σ_1^2 is a element of $\mathcal{G}_0(\mathcal{D}_1)$, where $\mathcal{D}_1$ is a one-sided open neighborhood of c, $(c,c+\delta)$, and $\sigma_1^2(c)$ exists and are bounded above and strictly positive. Analogous conditions hold for σ_0^2 on $\mathcal{D}_0$, where $\mathcal{D}_0$ is a one-sided open neighborhood of c, $(c-\delta,c)$.*

Assumption 5. *Let δ be some positive constant and κ be some positive integer. The conditional mean function m_1 is a element of $\mathcal{G}_\kappa(\mathcal{D}_1)$, and $m_1^{(s)}(c)$, for $s=1,\dots,\kappa$, exist and are bounded. Analogous conditions hold for m_0 on $\mathcal{D}_0$.*

Under Assumptions 1–4 and 5 with $\kappa = 3$, we can obtain the following result analogous to the results by Fan and Gijbels (1992):[8]

$$\begin{aligned}\mathrm{MSE}_n(h) = {} & \left\{\frac{b_1}{2}\left[m_1^{(2)}(c)h_1^2 - m_0^{(2)}(c)h_0^2\right]\right\}^2 + \frac{v}{nf(c)}\left\{\frac{\sigma_1^2(c)}{h_1} + \frac{\sigma_0^2(c)}{h_0}\right\} \\ & + o\left(h_1^4 + h_1^2 h_0^2 + h_0^4 + \frac{1}{nh_1} + \frac{1}{nh_0}\right),\end{aligned} \qquad (1)$$

where

$$b_1 = \frac{\mu_2^2 - \mu_1\mu_3}{\mu_0\mu_2 - \mu_1^2} \quad \text{and} \quad v = \frac{\mu_2^2\nu_0 - 2\mu_1\mu_2\nu_1 + \mu_1^2\nu_2}{\left(\mu_0\mu_2 - \mu_1^2\right)^2}.$$

This suggests that we choose the bandwidths to minimize the following AMSE:

$$\mathrm{AMSE}_n(h) = \left\{\frac{b_1}{2}\left[m_1^{(2)}(c)h_1^2 - m_0^{(2)}(c)h_0^2\right]\right\}^2 + \frac{v}{nf(c)}\left\{\frac{\sigma_1^2(c)}{h_1} + \frac{\sigma_0^2(c)}{h_0}\right\}. \qquad (2)$$

However, this procedure can fail. To see why, consider the case where $m_1^{(2)}(c)m_0^{(2)}(c) > 0$. Now choose $h_0 = [m_1^{(2)}(c)/m_0^{(2)}(c)]^{1/2}h_1$ to remove the first-order bias component completely from the AMSE. Then the first-order AMSE consists only of the variance term:

$$\mathrm{AMSE}_n(h) = \frac{v}{nh_1 f(c)}\left\{\sigma_1^2(c) + \sigma_0^2(c)\left[\frac{m_0^{(2)}(c)}{m_1^{(2)}(c)}\right]^{1/2}\right\}.$$

This implies that the AMSE can be made arbitrarily small by choosing a sufficiently large h_1. When the target parameter is the difference of the conditional mean functions and $m_1^{(2)}(c)m_0^{(2)}(c) > 0$, the first-order bias can be reduced not only by choosing a smaller bandwidth but also by changing a ratio of two bandwidths.

One reason for this problem is that the AMSE given in (2) does not account for higher-order terms. If we account for the higher-order terms for the bias component, setting the bias term in (2) to zero does not eliminate the whole bias component, and thus choosing large values for bandwidths may be punished. However, in what follows, we show that simply incorporating the second-order bias term into the AMSE does not resolve the problem because the bias term can be reduced to a smaller order inclusive of higher-order bias terms when $m_1^{(2)}(c)m_0^{(2)}(c) > 0$. After demonstrating this, we propose a new objective function that avoids this problem.

In order to discuss these, we first show in the next lemma, the second-order expansion of the MSE by generalizing the higher-order approximation of Fan, Gijbels, Hu, and Huang (1996).[9]

[8] The conditions on the first derivative of f and the third derivatives of m_1 and m_0, described in Assumptions 3 and 5, are not necessary to obtain the result (1). They are stated for later use.

[9] Fan et al. (1996) show the higher-order approximation of the MSE for interior points of the support of X. Lemma 1 presents the analogous result for a boundary point. A proof of Lemma 1 is provided in the Supplementary Material. The expression presented in Lemma 1 is the higher-order approximation rather than a better approximation to the bias since the next-order term of the variance is $O(1/n)$.

Lemma 1. *Suppose Assumptions 1–4 and 5 with $\kappa = 4$ hold. Then it follows that*

$$\mathrm{MSE}_n(h) = \left\{\frac{b_1}{2}\big[m_1^{(2)}(c)h_1^2 - m_0^{(2)}(c)h_0^2\big] + \big[b_{2,1}(c)h_1^3 - b_{2,0}(c)h_0^3\big] + o\big(h_1^3 + h_0^3\big)\right\}^2 + \frac{v}{nf(c)}\left\{\frac{\sigma_1^2(c)}{h_1} + \frac{\sigma_0^2(c)}{h_0}\right\} + o\left(\frac{1}{nh_1} + \frac{1}{nh_0}\right),$$

where, for $j = 0, 1$,

$$b_{2,j}(c) = (-1)^{j+1}\left\{\xi_1\left[\frac{m_j^{(2)}(c)}{2}\frac{f^{(1)}(c)}{f(c)} + \frac{m_j^{(3)}(c)}{6}\right] - \xi_2\frac{m_j^{(2)}(c)}{2}\frac{f^{(1)}(c)}{f(c)}\right\},$$

$$\xi_1 = \frac{\mu_2\mu_3 - \mu_1\mu_4}{\mu_0\mu_2 - \mu_1^2}, \quad \text{and} \quad \xi_2 = \frac{\big(\mu_2^2 - \mu_1\mu_3\big)(\mu_0\mu_3 - \mu_1\mu_2)}{\big(\mu_0\mu_2 - \mu_1^2\big)^2}.$$

Given the expression of Lemma 1, one might be tempted to proceed with an AMSE including the second-order bias term:

$$\mathrm{AMSE}_{2n} \equiv \left\{\frac{b_1}{2}\big[m_1^{(2)}(c)h_1^2 - m_0^{(2)}(c)h_0^2\big] + \big[b_{2,1}(c)h_1^3 - b_{2,0}(c)h_0^3\big]\right\}^2 + \frac{v}{nf(c)}\left\{\frac{\sigma_1^2(c)}{h_1} + \frac{\sigma_0^2(c)}{h_0}\right\}. \tag{3}$$

As indicated above, when $m_1^{(2)}(c)m_0^{(2)}(c) > 0$, we show that the bias component, inclusive of higher-order terms, can be made to the order $O(h_1^\ell)$, for an arbitrarily large integer $\ell > 0$ by appropriately choosing h_0. This implies that the minimization problem is not well-defined. Therefore, incorporating higher-order terms in itself does not resolve the problem we discussed when $m_1^{(2)}(c)m_0^{(2)}(c) > 0$.

To gain insight, consider choosing $h_0 = C(h_1, k)^{1/2}h_1$ for any given h_1 where $C(h_1, k) = C_0 + C_1h_1 + C_2h_1^2 + C_3h_1^3 + \cdots + C_kh_1^k$ for some constants $C_0, C_1, \ldots, C_k$ and positive k.[10] We first consider the case of $k = 1$ with $C(h_1, 1) = C_0 + C_1h_1$, where $C_0 = m_1^{(2)}(c)/m_0^{(2)}(c)$. In this case, the sum of the first- and second-order bias terms is

$$\frac{b_1}{2}\big[m_1^{(2)}(c) - C(h_1, 1)m_0^{(2)}(c)\big]h_1^2 + \big[b_{2,1}(c) - C(h_1, 1)^{3/2}b_{2,0}(c)\big]h_1^3 = \left\{-\frac{b_1}{2}C_1m_0^{(2)}(c) + b_{2,1}(c) - C_0^{3/2}b_{2,0}(c)\right\}h_1^3 + O\big(h_1^4\big).$$

By choosing $C_1 = 2[b_{2,1}(c) - C_0^{3/2}b_{2,0}(c)]/[b_1m_0^{(2)}(c)]$, one can make the order of bias $O(h_1^4)$. Next, consider $C(h_1, 2) = C_0 + C_1h_1 + C_2h_1^2$, where C_0 and C_1 are as determined

[10] Given that bandwidths are necessarily positive, we must have $C_0 > 0$, although we allow $C_1, C_2, \ldots, C_k$ to be negative. For sufficiently large n and for any k, we always have $C(h_1, k) > 0$ given $C_0 > 0$ and we assume this without loss of generality.

above. In this case,

$$\begin{aligned}&\frac{b_1}{2}\big[m_1^{(2)}(c)-C(h_1,2)m_0^{(2)}(c)\big]h_1^2+\big[b_{2,1}(c)-C(h_1,2)^{3/2}b_{2,0}(c)\big]h_1^3\\&\quad=-\big\{b_1C_2m_0^{(2)}(c)+3C_0^{1/2}C_1b_{2,0}(c)\big\}h_1^4/2+O\big(h_1^5\big).\end{aligned}$$

Hence, by choosing $C_2=-3C_0^{1/2}C_1b_{2,0}(c)/[b_1m_0^{(2)}(c)]$, one can make the order of the bias term $O(h_1^5)$. Similar arguments can be formulated generally: the discussion above is summarized in the following lemma.

Lemma 2. *Suppose that the conditions stated in Lemma 1 hold. Also suppose $m_1^{(2)}(c)m_0^{(2)}(c)>0$. Then there exist a combination of h_1 and h_0 such that the* AMSE$_{2n}$ *defined in* (3) *becomes*

$$\frac{v}{nh_1f(c)}\left\{\sigma_1^2(c)+\sigma_0^2(c)\left[\frac{m_1^{(2)}(c)}{m_0^{(2)}(c)}\right]^{1/2}\right\}+O\big(h_1^{2(k+3)}\big)$$

for an arbitrary integer $k>0$.

Lemma 2 says, given that our target is to minimize the AMSE$_{2n}$ with respect to two bandwidths, we can make the objective function converge to zero arbitrarily close to the rate $1/n$ by a proper choice of bandwidths when $m_1^{(2)}(c)m_0^{(2)}(c)>0$. This implies nonexistence of the optimal solution.[11] Lemma 2 can be generalized to the case of the AMSE with higher-order bias terms. Lemma A.1 in Appendix A shows, given that our target is to minimize the AMSE with up to $(K-1)$th-order bias terms for any $K>2$, there exists a combination of h_1 and h_0 such that the AMSE can be made converge to 0 at the rate arbitrarily close to $1/n$.

2.2 *Simultaneous selection of bandwidths*

Note the dichotomous nature of the problem with the AMSE objective function or its higher-order version AMSE$_{2n}$. When $m_1^{(2)}(c)m_0^{(2)}(c)>0$, the trade-off breaks down even with AMSE$_{2n}$ as Lemma 2 shows.

We define a new objective function which shows that the bandwidths that minimize it adapt to both situations without knowing the sign of $m_1^{(2)}(c)m_0^{(2)}(c)$. The new objective

[11] In the present approach, we consider choosing the bandwidths for the LLR estimator. In the literature of regression function estimation, it is common to employ local polynomial regression (LPR) of second order when the conditional mean function is three times continuously differentiable because it is known to reduce bias (see, e.g., Fan (1992)). However, we have two reasons for confining our attention to the LLR. First, as shown later, we can achieve the same bias reduction with the LLR when the sign of the product of the second derivatives is positive. When the sign is negative, the existence of the third derivatives becomes unnecessary. Second, even when we use a higher-order LPR, we end up with an analogous problem. For example, the first-order bias term is removed by using the second-order LPR, but when the signs of $b_{2,1}(c)$ and $b_{2,0}(c)$ are the same, the second-order bias term can be eliminated by using an appropriate choice of bandwidths.

function is a modified version of the AMSE with the second-order bias term (MMSE):

$$\mathrm{MMSE}_n(h) = \left\{\frac{b_1}{2}\left[m_1^{(2)}(c)h_1^2 - m_0^{(2)}(c)h_0^2\right]\right\}^2 + \left\{b_{2,1}(c)h_1^3 - b_{2,0}(c)h_0^3\right\}^2 + \frac{v}{nf(x)}\left\{\frac{\sigma_1^2(x)}{h_1} + \frac{\sigma_0^2(x)}{h_0}\right\}.$$

A notable characteristic of the MMSE is that the bias component is represented by the sum of the squared first- and the second-order bias terms. Intuitively, when $m_1^{(2)}(c)m_0^{(2)}(c) < 0$, the first-order bias term dominates the second order and the standard trade-off of the first-order bias term and the variance term emerges. When $m_1^{(2)}(c)m_0^{(2)}(c) > 0$, the first-order bias term can be made small, but the second-order term and the variance term provide the appropriate trade-off.

The bandwidth selection method discussed above is infeasible because the MMSE contains unknown quantities. We propose a feasible bandwidth selection method based on the MMSE by replacing the unknown objects in the MMSE by their nonparametric estimates. Consider the following plug-in version of the MMSE denoted by MMSE^p:

$$\mathrm{MMSE}_n^p(h) = \left\{\frac{b_1}{2}\left[\hat{m}_1^{(2)}(c)h_1^2 - \hat{m}_0^{(2)}(c)h_0^2\right]\right\}^2 + \left\{\hat{b}_{2,1}(c)h_1^3 - \hat{b}_{2,0}(c)h_0^3\right\}^2 + \frac{v}{n\hat{f}(c)}\left\{\frac{\hat{\sigma}_1^2(c)}{h_1} + \frac{\hat{\sigma}_0^2(c)}{h_0}\right\}, \tag{4}$$

where $\hat{m}_j^{(2)}(c)$, $\hat{b}_{2,j}(c)$, $\hat{\sigma}_j^2(c)$, and $\hat{f}(c)$ are consistent estimators of $m_j^{(2)}(c)$, $b_{2,j}(c)$, $\sigma_j^2(c)$, and $f(x)$ for $j = 0, 1$, respectively.[12] Let $(\hat{h}_1, \hat{h}_0)$ be a combination of bandwidths that minimizes the MMSE^p given in (4) and $\hat{h}$ denote $(\hat{h}_1, \hat{h}_0)$.[13] The next theorem characterizes the asymptotic properties of $\hat{h}$. The theorem demonstrates that the bandwidth selection methods adapt to the underlying sign of $m_1^{(2)}(c)m_0^{(2)}(c)$ automatically.

Let the bandwidths, (h_1^*, h_0^*), be the unique minimizer of

$$\mathrm{AMSE}_{1n}(h) = \left\{\frac{b_1}{2}\left[m_1^{(2)}(c)h_1^2 - m_0^{(2)}(c)h_0^2\right]\right\}^2 + \frac{v}{nf(c)}\left\{\frac{\sigma_1^2(c)}{h_1} + \frac{\sigma_0^2(c)}{h_0}\right\}, \tag{5}$$

when $m_1^{(2)}(c)m_0^{(2)}(c) < 0$, and when $m_1^{(2)}(c)m_0^{(2)}(c) > 0$ let (h_1^*, h_0^*) be the unique minimizer of

$$\mathrm{AMSE}_{2n}(h) = \left\{b_{2,1}(c)h_1^3 - b_{2,0}(c)h_0^3\right\}^2 + \frac{v}{nf(c)}\left\{\frac{\sigma_1^2(c)}{h_1} + \frac{\sigma_0^2(c)}{h_0}\right\} \tag{6}$$

subject to the restriction $m_1^{(2)}(c)h_1^2 - m_0^{(2)}(c)h_0^2 = 0$.[14]

[12] The construction of MMSE$_n^p$ requires pilot estimates for $m_j^{(2)}(c)$, $b_{2,j}(c)$, $f(c)$, and $\sigma_j^2(c)$ for $j = 0, 1$. A detailed procedure about how to obtain the pilot estimates is given in the Supplemental Material.

[13] It is also possible to construct another version of the MMSE^p based on the finite sample approximations discussed by Fan and Gijbels (1996, Section 4.3). We do not pursue this direction because it is computationally intensive for a large sample and an unreported simulation produced almost the same result as that based on the MMSE^p given in (4).

[14] Uniqueness of (h_1^*, h_0^*) in each case is verified in Arai and Ichimura (2013).

The bandwidths (h_0^*, h_1^*) defined in the theorem coincides with the bandwidths defined above without a need to know the sign of $m_1^{(2)}(c)m_0^{(2)}(c)$.

Theorem 1. *Suppose that the conditions stated in Lemma 1 hold. Assume further that $b_{2,1}(c) - \{m_1^{(2)}(c)/m_0^{(2)}(c)\}^{3/2}b_{2,0}(c) \neq 0$. Let $\hat{m}_j^{(2)}(c)$, $\hat{b}_{2,j}(c)$, $\hat{f}(c)$, and $\hat{\sigma}_j^2(c)$ satisfy $\hat{m}_j^{(2)}(c) \xrightarrow{p} m_j^{(2)}(c)$, $\hat{b}_{2,j}(c) \xrightarrow{p} b_{2,j}(c)$, $\hat{f}(c) \xrightarrow{p} f(c)$, and $\hat{\sigma}_j^2(c) \xrightarrow{p} \sigma_j^2(c)$ for $j = 0, 1$, respectively. Then the following hold*:

$$\frac{\hat{h}_1}{h_1^*} \xrightarrow{p} 1, \qquad \frac{\hat{h}_0}{h_0^*} \xrightarrow{p} 1, \quad \textit{and} \quad \frac{\mathrm{MMSE}_n^p(\hat{h})}{\mathrm{MSE}_n(h^*)} \xrightarrow{p} 1,$$

where $h_1^* = \theta^* n^{-1/5}$ and $h_0^* = \lambda^* h_1^*$ with

$$\theta^* = \left\{ \frac{v\sigma_1^2(c)}{b_1^2 f(c) m_1^{(2)}(c)\big[m_1^{(2)}(c) - \lambda^{*2} m_0^{(2)}(c)\big]} \right\}^{1/5} \quad \textit{and} \quad \lambda^* = \left\{ -\frac{\sigma_0^2(c) m_1^{(2)}(c)}{\sigma_1^2(c) m_0^{(2)}(c)} \right\}^{1/3}, \tag{7}$$

when $m_1^{(2)}(c)m_0^{(2)}(c) < 0$, and $h_1^* = \theta^* n^{-1/7}$ and $h_0^* = \lambda^* h_1^*$ with

$$\theta^* = \left\{ \frac{v\big[\sigma_1^2(c) + \sigma_0^2(c)/\lambda^*\big]}{6f(c)\big[b_{2,1}(c) - \lambda^{*3} b_{2,0}(c)\big]^2} \right\}^{1/7} \quad \textit{and} \quad \lambda^* = \left\{ \frac{m_1^{(2)}(c)}{m_0^{(2)}(c)} \right\}^{1/2}, \tag{8}$$

when $m_1^{(2)}(c)m_0^{(2)}(c) > 0$.

Theorem 1 also shows that the minimized value of the plug-in version of the MMSE is asymptotically the same as the MSE based on (h_1^*, h_0^*).

As discussed already, Theorem 1 shows that the single objective function adapts to two distinct cases where the optimal bandwidths converge to zero at the different rates, $n^{-1/5}$ and $n^{-1/7}$. To see how this happens, consider when $m_1^{(2)}(c)m_0^{(2)}(c) < 0$. In this case, the square of the first-order bias term serves as the leading penalty and that of the second-order bias term becomes the second-order penalty, which does not affect the asymptotic behavior of the bandwidths. When $m_1^{(2)}(c)m_0^{(2)}(c) > 0$, the square of the second-order bias term works as the penalty and that of the first-order bias term becomes the linear restriction in equation (6) asymptotically.[15]

We next discuss the advantages of the bandwidths, $\hat{h}$, through the asymptotically equivalent bandwidths, (h_1^*, h_0^*). In particular, we show that the bandwidths dominate the existing approaches in the AMSE, irrespective of the values of the second derivatives. To see this, first note that when the product of the second derivatives is positive, the AMSE based on (h_1^*, h_0^*) is of order $n^{-6/7}$ whereas the AMSE based on the optimal bandwidths chosen for each of the regression function separately (we refer to these bandwidths, Independent Bandwidths (IND)) is of order $n^{-4/5}$.[16] The same order holds for

[15] In this case, the first-order bias term can be considered as the regularization term in the sense that it provides additional information on the bandwidths.

[16] The independent selection chooses the bandwidths on the left and the right of the cut-off optimally for each function without paying attention to the relationship between the two functions. The IND bandwidths

the single bandwidth approach such as IK bandwidth unless the two second derivatives are exactly the same. Thus, when the product of the second derivatives is positive, the bandwidths, (h_1^*, h_0^*), are more efficient than either the IK bandwidth or the IND bandwidths in the sense that the AMSE has a faster rate of convergence.

The only exception to this observation is when the second derivatives are the same. In this case, the IK bandwidth is

$$h_{\mathrm{IK}} = \theta_{\mathrm{IK}} n^{-1/7},$$

where

$$\theta_{\mathrm{IK}} = C_{\mathrm{IK}} \left(\frac{\sigma_1^2(c) + \sigma_0^2(c)}{\left[\sigma_1^2(c)\right]^{2/7} \left\{ p_1 \left[m_1^{(3)}(c) \right]^2 \right\}^{5/7} + \left[\sigma_0^2(c)\right]^{2/7} \left\{ p_0 \left[m_0^{(3)}(c) \right]^2 \right\}^{5/7}} \right)^{1/5},$$

$C_{\mathrm{IK}} = [v/(2160 \cdot 3.56^5 \cdot b_1^2 [f(c)]^{5/7})]^{1/5}$, $p_1 = \int_c^\infty f(x)\,dx$, and $p_0 = \int_{-\infty}^c f(x)\,dx$.[17] Although this bandwidth is of the same order with (h_1^*, h_0^*), it is not determined by minimizing the AMSE. In fact, the ratio of the AMSE up to the second-order bias term based on (h_1^*, h_0^*) to that of the IK bandwidth converges to

$$\frac{1}{\frac{1}{7}\gamma^6 + \frac{6}{7}\frac{1}{\gamma}},$$

where $\gamma = \theta_{\mathrm{IK}}/\theta_S$ and θ_S equals θ^* in equation (4) for $\lambda^* = m_1^{(2)}(c)/m_0^{(2)}(c) = 1$.[18] It is easy to show that the ratio is strictly less than one and equals one if and only if $\gamma = 1$. Since the θ_S depends on the second derivatives but θ_{IK} does not, the ratio can be much larger or smaller than 1, and hence the ratio can be arbitrarily close to 0.

When the sign of the product of the second derivatives is negative, the rates of convergence of the AMSEs corresponding to different bandwidth selection rules are the same. By construction, (h_1^*, h_0^*) lead to the lowest AMSE. Hence the issue would be how large the difference in the AMSE could be under what kind of circumstances. The AMSEs

based on the AMSE criterion are given by

$$\check{h}_1 = \left\{ \frac{v\sigma_1^2(c)}{b_1^2 f(c) \left[m_1^{(2)}(c) \right]^2} \right\}^{1/5} n^{-1/5} \quad \text{and} \quad \check{h}_0 = \left\{ \frac{v\sigma_0^2(c)}{b_1^2 f(c) \left[m_0^{(2)}(c) \right]^2} \right\}^{1/5} n^{-1/5}.$$

[17] The derivation of θ_{IK} is provided in the Supplementary Material.

[18] To see why the ratios of the AMSEs converges to the specified limit, note that the ratio of the AMSEs is

$$\frac{\left[b_{2,1}(c) - b_{2,0}(c)\right]^2 \theta_S^6 + \dfrac{v\left[\sigma_1^2(c) + \sigma_0^2(c)\right]}{\theta_S f(c)}}{\left[b_{2,1}(c) - b_{2,0}(c)\right]^2 \theta_{\mathrm{IK}}^6 + \dfrac{v\left[\sigma_1^2(c) + \sigma_0^2(c)\right]}{\theta_{\mathrm{IK}} f(c)}}.$$

Since the first-order condition implies $v[\sigma_1^2(c) + \sigma_0^2(c)]/[\theta_S f(c)] = 6[b_{2,1}(c) - b_{2,0}(c)]^2 \theta_S^6$, substituting this expression and some simple calculations yield the result.

corresponding to (h_1^*, h_0^*), IK bandwidth, and IND bandwidths are, respectively,

$$\mathrm{AMSE}(h^*) = \frac{5}{4} n^{-4/5} C_K \bigl[m_0^{(2)}(c)\bigr]^{2/5} \bigl[\sigma_0^2(c)\bigr]^{4/5} \bigl[\bigl(\gamma_1 \gamma_2^2\bigr)^{1/3} + 1\bigr]^{6/5},$$

$$\mathrm{AMSE}(h_{\mathrm{IK}}) = \frac{5}{4} n^{-4/5} C_K \bigl[m_0^{(2)}(c)\bigr]^{2/5} \bigl[\sigma_0^2(c)\bigr]^{4/5} (\gamma_1 + 1)^{2/5} (\gamma_2 + 1)^{4/5},$$

and

$$\mathrm{AMSE}(h_{\mathrm{IND}}) = \frac{5}{4} n^{-4/5} C_K \bigl[m_0^{(2)}(c)\bigr]^{2/5} \bigl[\sigma_0^2(c)\bigr]^{4/5} \bigl(\bigl(\gamma_1 \gamma_2^2\bigr)^{1/5} + 1\bigr)^2 \bigl(\bigl(\gamma_1 \gamma_2^2\bigr)^{2/5} + 1\bigr),$$

where $\gamma_1 = -m_1^{(2)}(c)/m_0^{(2)}(c)$, $\gamma_2 = \sigma_1^2(c)/\sigma_0^2(c)$, and $C_K = [b_1 v_2/f(c)^2]^{2/5}$.

Clearly, the AMSE based on (h_1^*, h_0^*) relative to that based on the IK depends only on γ_1 and γ_2. It is straightforward to show that the maximum of the ratio is 1 and attained if and only if $\gamma_1 = \gamma_2$. Efficiency as a function of γ_1 given γ_2 and that as a function of γ_2 given γ_1 are plotted in Figure 2(a) and Figure 2(b), and the contour of the ratio is depicted in Figure 2(c). We note that while the region on which the ratio is close to 1 is large, the ratio is less than 0.8 whenever γ_1 and γ_2 are rather different.

The bandwidths (h_1^*, h_0^*) have the advantage over the IND bandwidths, too. Again, clearly the AMSE of (h_1^*, h_0^*) relative to that of the IND only depends on γ_1 and γ_2. The ratio attains its minimum when $\gamma_1 \gamma_2^2 = 1$ and the minimum value is $2^{6/5}/(12/5) \doteq 0.957$. It is an interesting finding that when the sign of the second derivatives differs, there is less than 5% gain in efficiency by (h_1^*, h_0^*) over IND.

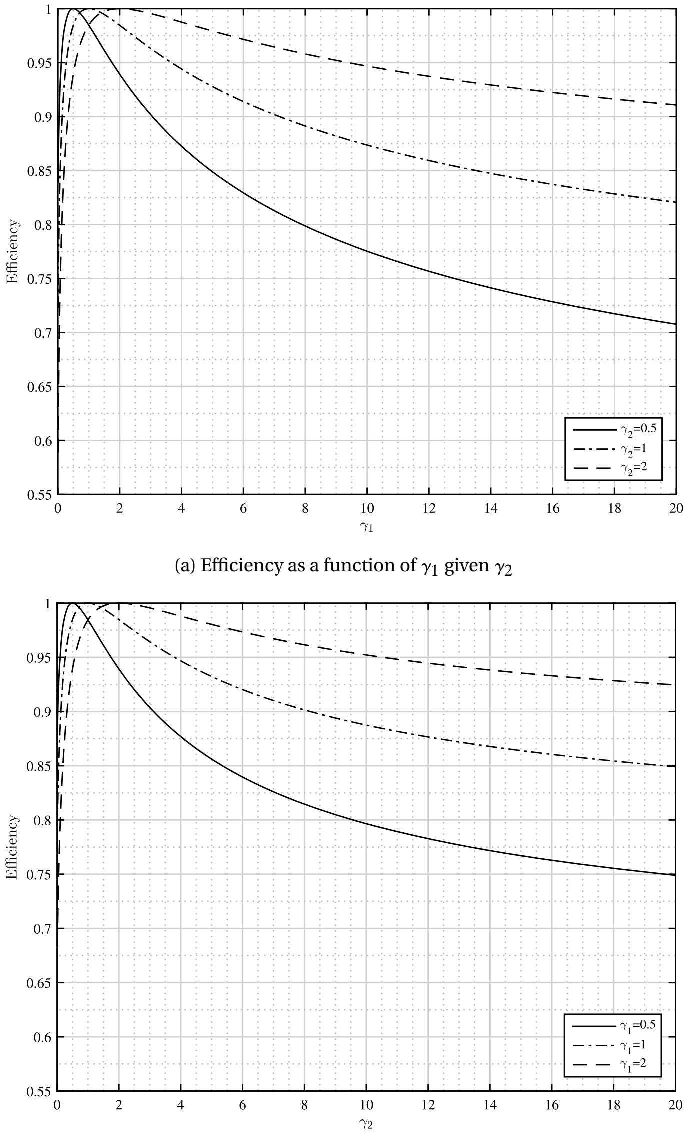
In summary, the bandwidths (h_1^*, h_0^*) improve the rate of convergence of AMSE when the sign of the product of the second derivatives is positive. When the sign is negative, it is more efficient than either the IK bandwidth or the IND bandwidths although the gain over IND is less than 5%.

Another important issue would be the robustness of $\hat{h}$ with respect to the sign of the product of the second derivatives since the behavior of (h_1^*, h_0^*) changes discontinuously at $m_1^{(2)}(c) m_0^{(2)}(c) = 0$. The discontinuous behavior seems unlikely to produce a well-behaved estimator and this was confirmed by our unreported simulation experiments. In fact, the robustness is the reason why we consider the bandwidths based on the MMSE rather than the plug-in version of (h_1^*, h_0^*). We can show that the minimizer of MMSE_n^p, $\hat{h}$, is unique around $m_1^{(2)}(c) m_0^{(2)}(c) = 0$ regardless of the sign of the second derivatives and that $\hat{h}$ changes continuously with respect to the second derivatives.[19] This is how $\hat{h}$ performs stably around $m_1^{(2)}(c) m_0^{(2)}(c) = 0$.

3. Extension

In this section, we extend the framework of the simultaneous bandwidth selection to the problem of the sharp regression kink (RK) design developed by CLPW. We then propose the confidence interval of the RD and RK design estimators based on the proposed bandwidths in the spirit of CCT.

[19] The property of $\hat{h}$ is discussed in the Supplemental Material.



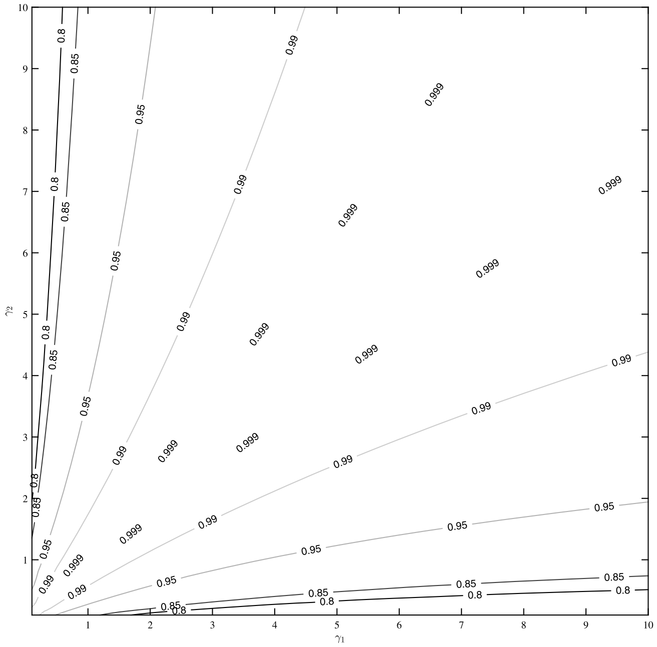
(a) Efficiency as a function of γ_1 given γ_2

(b) Efficiency as a function of γ_2 given γ_1

FIGURE 2. The ratio of the AMSEs, $\text{AMSE}(h^*)/\text{AMSE}(h_{\text{IK}})$, as a function of γ_1 and γ_2. (a) The ratio of the AMSEs as a function of γ_1 given γ_2, (b) the ratio of the AMSEs as a function of γ_2 given γ_1, and (c) the contour of the ratio of the AMSEs, as a function of γ_1 and γ_2.

3.1 *Sharp regression kink design*

The sharp RK design is a class of models where a continuous treatment variable is a known kinked function of an assignment variable. The sharp RK design is discussed extensively by CLPW and applied to investigate the effect of unemployment benefits on unemployment durations. CLPW employed several bandwidths including the optimal



(c) Contour as a function of γ_1 and γ_2

FIGURE 2. Continued.

bandwidth proposed by CCT as well as a rule-of-thumb bandwidth proposed by Fan and Gijbels (1996). CCT extends the approach for the RD design by IK to the case of the RK design. Since the bandwidth by CCT is based on a single bandwidth approach, it would be meaningful to propose to choose distinct bandwidths as in the sharp RD design.

The sharp RK design has a similar structure to the sharp RD design and the treatment-on-treated parameter, denoted τ_{SRK} up to a known multiplicative constant, is given using the notation introduced in the previous section by

$$\tau_{\text{SRK}}(c) = m_1^{(1)}(c) - m_0^{(1)}(c).$$

A standard estimation method is to use the LPR of second order. Denote the LPR estimators by $\hat{m}_1^{(1)}(c)$ and $\hat{m}_0^{(1)}(c)$. Then $\tau_{\text{SRK}}(c)$ is estimated by $\hat{\tau}_{\text{SRK}}(c) = \hat{m}_1^{(1)}(c) - \hat{m}_0^{(1)}(c)$.

We follow the approach taken for the sharp RD design to choose bandwidths for the sharp RK design. The next lemma provides the MSE expansion up to a second-order bias term, which plays a key role to propose the optimal bandwidths.

LEMMA 3. *Suppose Assumptions 1–4 and 5 with $\kappa = 5$ hold. Then it follows that*

$$\begin{aligned}\text{MSE}_n(h) = {} & \bigl\{d_1\bigl[m_1^{(3)}(c)h_1^2 - m_0^{(3)}(c)h_0^2\bigr] + \bigl[d_{2,1}(c)h_1^3 - d_{2,0}(c)h_0^3\bigr] + o\bigl(h_1^3 + h_0^3\bigr)\bigr\}^2 \\ & + \frac{w}{nf(c)}\left\{\frac{\sigma_1^2(c)}{h_1^3} + \frac{\sigma_0^2(c)}{h_0^3}\right\} + o\left(\frac{1}{nh_1^3} + \frac{1}{nh_0^3}\right),\end{aligned}$$

where, for $j = 0, 1$,

$$d_1 = e_1' S_{0,2}^{-1} c_{3,2}/3!, \qquad w = e_1' S_{0,2}^{-1} S_{0,2}^{*} S_{0,2}^{-1} e_1,$$

$$\begin{aligned} d_{2,j}(c) = (-1)^{j+1} \Biggl\{ & \biggl[\frac{m_j^{(3)}(c)}{3!} \frac{f^{(1)}(c)}{f(c)} + \frac{m_j^{(4)}(c)}{4!} \biggr] e_1' S_{0,2}^{-1} c_{4,2} \\ & - \frac{m_j^{(3)}(c)}{2} \frac{f^{(1)}(c)}{f(c)} e_1' S_{0,2}^{-1} S_{1,2} S_{0,2}^{-1} c_{3,2} \Biggr\}, \end{aligned}$$

e_2 *is the unit vector* $(0, 1, 0)'$, $S_{k,p}$ *and* $S_{k,p}^{*}$ *are* $(p+1) \times (p+1)$ *matrices of which* (i, j) *element is given by* $\mu_{k+i+j-2}$ *and* $\mu_{k+i+j-1}$, *respectively, and* $c_{k,p}$ *is a* $(p+1)$*-dimensional column vector of which* j*th element is given by* μ_{k+j-1}.

It is evident that the MSE of the sharp RK estimator possesses the same structure as that of the sharp RD estimator. This fact suggests the bandwidths, which minimize the following MMSE:

$$\begin{aligned} \mathrm{MMSE}_n^p(h) = & \bigl\{ d_1 \bigl[\hat{m}_1^{(3)}(c) h_1^2 - \hat{m}_0^{(3)}(c) h_0^2 \bigr] \bigr\}^2 + \bigl\{ \bigl[\hat{d}_{2,1}(c) h_1^3 - \hat{d}_{2,0}(c) h_0^3 \bigr] \bigr\}^2 \\ & + \frac{w}{n \hat{f}(c)} \biggl\{ \frac{\hat{\sigma}_1^2(c)}{h_1^3} + \frac{\hat{\sigma}_0^2(c)}{h_0^3} \biggr\}, \end{aligned} \tag{9}$$

where, for $j = 0, 1$, $\hat{m}_j^{(3)}(c)$, $\hat{d}_{2,j}(c)$, $\hat{\sigma}_j^2(c)$, and $\hat{f}(c)$ are the consistent estimators of $m_j^{(3)}(c)$, $d_{2,j}(c)$, $\sigma_1^2(c)$, and $f(c)$, respectively.[20] Let $(\hat{h}_1, \hat{h}_0)$ be a pair of the bandwidths that minimizes the MMSE^p. The next theorem describes their asymptotic properties.

Theorem 2. *Suppose that the conditions stated in Lemma 3 hold. Assume further that* $d_{2,1}(c) - \{m_1^{(3)}(c)/m_0^{(3)}(c)\}^{3/2} d_{2,0}(c) \neq 0$. *Then the following hold*:

$$\frac{\hat{h}_1}{h_1^*} \xrightarrow{p} 1, \qquad \frac{\hat{h}_0}{h_0^*} \xrightarrow{p} 1, \quad \textit{and} \quad \frac{\mathrm{MMSE}_n^p(\hat{h})}{\mathrm{MSE}_n(h^*)} \xrightarrow{p} 1,$$

where $h_1^* = \theta^* n^{-1/7}$ *and* $h_0^* = \lambda^* h_1^*$ *with*

$$\theta^* = \biggl\{ \frac{3 w \sigma_1^2(c)}{4 d_1^2 f(c) m_1^{(3)}(c) \bigl[m_1^{(3)}(c) - \lambda^{*2} m_0^{(3)}(c) \bigr]} \biggr\}^{1/7} \quad \textit{and} \quad \lambda^* = \biggl\{ - \frac{\sigma_0^2(c) m_1^{(3)}(c)}{\sigma_1^2(c) m_0^{(3)}(c)} \biggr\}^{1/5}, \tag{10}$$

when $m_1^{(3)}(c) m_0^{(3)}(c) < 0$, *and* $h_1^* = \theta^* n^{-1/9}$ *and* $h_0^* = \lambda^* h_1^*$ *with*

$$\theta^* = \biggl\{ \frac{w \bigl[\sigma_1^2(c) + \sigma_0^2(c)/\lambda^{*3} \bigr]}{2 f(c) \bigl[d_{2,1}(c) - \lambda^{*3} d_{2,0}(c) \bigr]^2} \biggr\}^{1/9} \quad \textit{and} \quad \lambda^* = \biggl\{ \frac{m_1^{(3)}(c)}{m_0^{(3)}(c)} \biggr\}^{1/2}, \tag{11}$$

when $m_1^{(3)}(c) m_0^{(3)}(c) > 0$.

[20] A detailed procedure to obtain these pilot estimates are provided in the Supplemental Material.

The proof of Theorem 2 is analogous to that of Theorem 1 and it is provided in the Supplemental Material.

Theorem 2 shows that the bandwidths that minimize the MMSE is asymptotically equivalent to (h_1^*, h_0^*), and analogously to Theorem 1, the bandwidths adapt automatically to the sign of $m_1^{(3)}(c)m_0^{(3)}(c)$. That is, the bandwidths, (h_1^*, h_0^*), are the unique minimizer of

$$\mathrm{AMSE}_{1n}(h) = \bigl\{d_1\bigl[m_1^{(3)}(c)h_1^2 - m_0^{(3)}(c)h_0^2\bigr]\bigr\}^2 + \frac{w}{nf(c)}\left\{\frac{\sigma_1^2(c)}{h_1^3} + \frac{\sigma_0^2(c)}{h_0^3}\right\},$$

when $m_1^{(3)}(c)m_0^{(3)}(c) < 0$, and when $m_1^{(3)}(c)m_0^{(3)}(c) > 0$, (h_1^*, h_0^*) are the unique minimizer of

$$\mathrm{AMSE}_{2n}(h) = \bigl\{d_{2,1}(c)h_1^3 - d_{2,0}(c)h_0^3\bigr\}^2 + \frac{w}{nf(c)}\left\{\frac{\sigma_1^2(c)}{h_1^3} + \frac{\sigma_0^2(c)}{h_0^3}\right\},$$

subject to the restriction $m_1^{(3)}(c)h_1^2 - m_0^{(3)}(c)h_0^2 = 0$.

One notable difference between the sharp RK and RD designs is that the bandwidths are now of order $n^{-1/7}$ when $m_1^{(3)}(c)m_0^{(3)}(c) < 0$ and $n^{-1/9}$ when $m_1^{(3)}(c)m_0^{(3)}(c) > 0$. The advantages of the bandwidths, (h_1^*, h_0^*), over the existing bandwidths are parallel to the case of the RD design.

3.2 *Confidence intervals*

The main purpose of this paper is to choose bandwidths optimally for point estimation. In practice, estimation of confidence intervals is also important. For the construction of confidence intervals for the RD and RK estimators, as CCT discuss, the conventional confidence intervals are not valid when the optimal bandwidths are employed and that the coverage probabilities of the conventional confidence intervals tend to be much lower than the nominal one. Then CCT propose novel confidence interval estimators based on the bias-corrected RD and RK estimators with the robust standard errors, which account for variability in the bias estimators. See Calonico, Cattaneo, and Farrell (forthcoming) for more discussions on theoretical superiority of the confidence interval proposed by CCT. Here, we propose confidence interval estimators in the spirit of CCT. Through simulations, we demonstrate their usefulness in the next section.

The robust confidence interval estimators proposed by CCT require a modification for the RD and RK estimators based on our bandwidths since the conditions on the bandwidths imposed by Theorems 1 and 2 of CCT are violated by our bandwidths when the sign of the product of the relevant order (second or third) derivatives is positive. However, we show that the approach of CCT can be extended to cover the case by employing the bias correction suitable for the estimators based on our bandwidths. While the bias correction based on the first-order bias term is sufficient for the estimators by CCT, our bias-corrected estimators take the second-order bias term into consideration in addition to the first-order bias term. Hence our confidence interval estimators are based on the following bias-corrected estimators. Let $h = (h_1, h_0)$ be the bandwidth used to estimate the relevant treatment effect. Also let $h_k = (h_{k,1}, h_{k,0})$, for $k = 2, 3, 4$ be the pilot bandwidth to estimate the kth derivatives, where $h_{k,1}$ and $h_{k,0}$ are the bandwidths on the right and left of the cut-off point, respectively:[21]

$$\hat{\tau}^{\text{bc}}_{\text{SRD}}(h, h_2, h_3) = \hat{\tau}_{\text{SRD}}(c) - \hat{B}_{\text{SRD},1}(h, h_2) - \hat{B}_{\text{SRD},2}(h, h_2, h_3),$$
$$\hat{\tau}^{\text{bc}}_{\text{SRK}}(h, h_3, h_4) = \hat{\tau}_{\text{SRK}}(c) - \hat{B}_{\text{SRK},1}(h, h_3) - \hat{B}_{\text{SRK},2}(h, h_3, h_4),$$

where $\hat{\tau}_{\text{SRD}}(c)$ and $\hat{\tau}_{\text{SRK}}(c)$ are the SRD and SRK estimators discussed above, $\hat{B}_{\text{SRD},1}(h, h_2)$ and $\hat{B}_{\text{SRK},1}(h, h_3)$ are the estimators of the first-order bias terms, $\hat{B}_{\text{SRD},2}(h, h_2, h_3)$ and $\hat{B}_{\text{SRK},2}(h, h_3, h_4)$ are the estimators of the second-order bias terms for the SRD and SRK, respectively.[22] When the sign of the second derivatives of the conditional mean functions, $m_1^{(2)}(c)$ and $m_0^{(2)}(c)$ are distinct in the case of the SRD, the estimator of the first-order bias term plays a main role as in CCT. However, when the sign of the second derivatives is the same, the estimator of the first-order bias term converges to zero and then the estimator of the second-order bias term plays the main role. The case for the SRK is analogous.

The next theorem shows asymptotic normality of the bias-corrected estimator and it indicates how to construct the robust confidence interval. The estimators of the second-order bias terms require the pilot estimates, $\hat{f}(c)$ and $\hat{f}^{(1)}(c)$. We estimate them with the bandwidths of optimal orders under the following assumption.[23]

Assumption 6. *The Lebesgue density of X_i, denoted f, is an element of $\mathcal{G}_3(\mathcal{D})$ where $\mathcal{D}$ is an open neighborhood of c and is bounded above and strictly positive on $\mathcal{D}$.*

Theorem 3. *Suppose that the conditions stated in Assumptions 1–4 and 6 hold*:

(i) (*Sharp RD Design*) *In addition, suppose that Assumption 5 with $\kappa = 4$ hold. Assume that, for the bandwidths on the right of the cut-off point, h_1, $h_{2,1}$, and $h_{3,1}$, $n \min\{h_1^5, h_{2,1}^5, h_{3,1}^7/h_j^2\} \times \max\{h_1^4, h_{2,1}^4, h_1^2 h_{3,1}^2\} \to 0$ and $n \min\{h_1, h_{2,1}, h_{3,1}\} \to \infty$ and assume that the analogous conditions hold for the bandwidths on the left. Then*

$$\frac{\hat{\tau}^{\text{bc}}_{\text{SRD}}(h, h_2, h_3) - \tau_{\text{SRD}}(c)}{\sqrt{V^{\text{bc}}_{\text{SRD}}(h, h_2, h_3)}} \xrightarrow{d} N(0, 1),$$

where the exact form of $V^{\text{bc}}_{\text{SRD}}(h, h_2, h_3)$ is given in Appendix C.

[21] As provided in the Supplemental Material, we use the bandwidths, $h_{k,1}$ and $h_{k,0}$, of the same order.

[22] The explicit forms of the bias-correction terms are provided in Appendix C. Our bias correction is slightly different from the one by CCT. While CCT use the finite sample approximation for the components related to the constants that depend solely on the kernel function, we use the constants directly to be consistent for the construction of the MMSE.

[23] The pilot bandwidths to estimate $f(c)$ and $f^{(1)}(c)$ are of order $n^{-1/5}$ and $n^{-1/7}$, respectively. The explicit expressions are provided in the Supplemental Material. We use these bandwidths because they are natural choices and simplify the proof. It is possible to generalize the results for a wide range of pilot bandwidths at the cost of more notation in the proof.

(ii) (*Sharp RK Design*) *In addition, suppose that Assumption* 5 *with* $\kappa = 5$ *holds. Assume that, for the bandwidths on the right of the cut-off point,* h_1, $h_{3,1}$, *and* $h_{4,1}$, $n \min\{h_1^7, h_{3,1}^7, h_{4,1}^9/h_1^2\} \times \max\{h_1^4, h_{3,1}^4, h_1^2 h_{4,1}^2\} \to 0$ *and* $n \min\{h_1, h_{2,1}, h_{3,1}\} \to \infty$ *and assume that the analogous conditions hold for the bandwidths on the left. Then*

$$\frac{\hat{\tau}_{\mathrm{SRK}}^{\mathrm{bc}}(h, h_3, h_4) - \tau_{\mathrm{SRK}}(c)}{\sqrt{V_{\mathrm{SRK}}^{\mathrm{bc}}(h, h_3, h_4)}} \xrightarrow{d} N(0,1),$$

where the exact form of $V_{\mathrm{SRK}}^{\mathrm{bc}}(h, h_3, h_4)$ *is given in Appendix* C.

The proof of Theorem 3 is provided in the Supplemental Material. This theorem is an application of the CCT approach. The robust variances given in Theorem 3 consist of three components. The first component is the conventional variance and the second is the one due to the variability of the bias-correction term for the first-order bias. The sum of the first and the second components is the robust variance in the context of CCT. The third component shows up in the present case because of the variability related to the bias-correction term for the second-order bias.

Theorem 3 suggests the following $100(1-\alpha)$-percent confidence intervals for the sharp RD design

$$I_{\mathrm{SRD}}^{\mathrm{rbc}}(h, h_2, h_3) = \left[\hat{\tau}_{\mathrm{SRD}}^{\mathrm{bc}}(h, h_2, h_3) \pm z_{\alpha/2}\sqrt{\hat{V}_{\mathrm{SRD}}^{\mathrm{bc}}(h, h_2, h_3)}\right], \tag{12}$$

where $z_{\alpha/2}$ is $\alpha/2$ percentile of the standard normal distribution. The same implications hold for the sharp RK design and the $100(1-\alpha)$-percent confidence intervals are given by

$$I_{\mathrm{SRK}}^{\mathrm{rbc}}(h, h_3, h_4) = \left[\hat{\tau}_{\mathrm{SRK}}^{\mathrm{bc}}(h, h_3, h_4) \pm z_{\alpha/2}\sqrt{\hat{V}_{\mathrm{SRK}}^{\mathrm{bc}}(h, h_3, h_4)}\right].$$

As in the results given in CCT, Theorem 3 are flexible in terms of the choice of the bandwidths. For implementation, following CCT, we use the bandwidths of the same order for h, h_2, and h_3 (h, h_3, and h_4) so that all components of the robust variance are of the same order for the sharp RD design (sharp RK design). Again, following the current practice of CCT, we specifically set $h = h_2 = h_3$ for the sharp RD and $h = h_3 = h_4$ for the sharp RK designs.

4. Simulation

To investigate the finite sample performance of the proposed method, we conducted simulation experiments. Our simulation experiments demonstrate that the theoretical advantages of the proposed bandwidths have over the existing bandwidth selection rules, such as the IK, IND, and CCT bandwidths, realize in the sample sizes relevant for empirical studies in general, and especially so for the simulation designs taken directly from empirical studies.

4.1 *Simulation designs*

We consider four designs. Designs 1–3 are the ones used for simulation experiments in the present context by IK and CCT. Designs 1 and 2 are motivated by the empirical studies of Lee (2008) and Ludwig and Miller (2007), respectively. Design 4 mimics the situation considered by Ludwig and Miller (2007) where they investigate the effect of Head Start assistance on Head Start spending in 1968 on child mortality. This design corresponds to Panel A of Figure II in Ludwig and Miller (2007, p. 176).[24]

The designs are depicted in Figure 3. For the first two designs, the sign of the product of the second derivatives is negative so that the AMSE convergence rates for all bandwidth selection rules are the same. For the next two designs, the sign is positive. For these two cases, (h_1^*, h_0^*) has the slower convergence rate compared to IND. Design 3, examined by IK, however, has the same second derivatives on the right and on the left of the cut of point, so that the convergence rate of the AMSE for IK is the same with that for (h_1^*, h_0^*).

For each design, the assignment variable X_i is given by $2Z_i - 1$ where Z_i have a Beta distribution with parameters $\alpha = 2$ and $\beta = 4$. We consider a normally distributed additive error term with mean zero and standard deviation 0.1295. The specification for the assignment variable and the additive error are exactly the same as that considered by IK. We use data sets of 500 and 2000 observations and the results are drawn from 10,000 replications.

4.2 *Results*

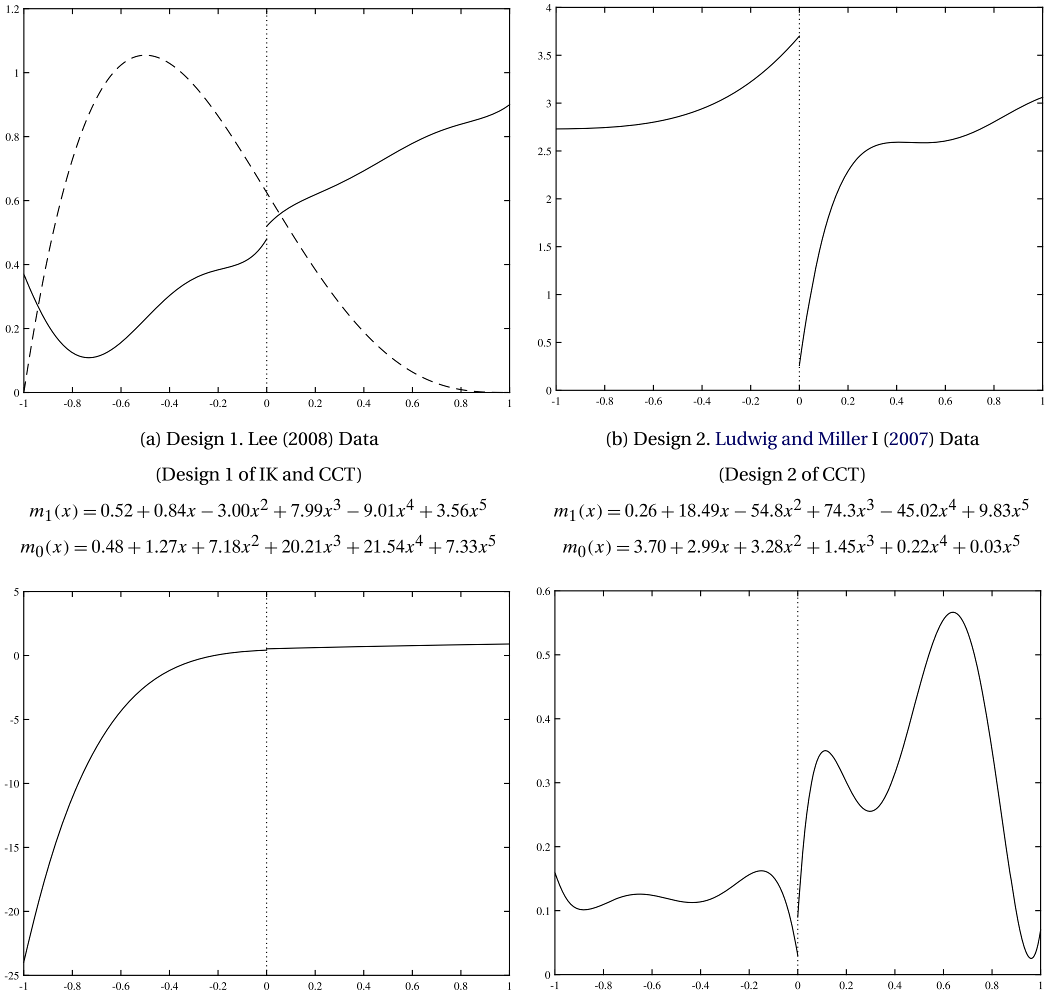
Table 1 presents the simulation results regarding point estimation. The first column shows the sample size and the second column explains the design. The third column reports the method used to obtain the bandwidth(s). MMSE refers to the proposed bandwidth selection rule based on $\text{MMSE}_n^p(h)$ in equation (4). IND is the independent bandwidth. IK is the bandwidth denoted by $\hat{h}_{\text{opt}}$ in Table 2 of IK.[25] CCT is the bandwidth proposed in Section 2.6 of Calonico, Cattaneo, and Titiunik (2014b).[26] The CCT bandwidth is the refined version of the IK bandwidth, which uses the nearest neighbor-type variance and the general regularization term which depends on the variance of the LPR estimator of the second derivatives.

The fourth and fifth columns report the mean (labeled "Mean") and standard deviation (labeled "SD") of the bandwidths for MMSE, IND, IK, and CCT. For MMSE and IND, these columns report the bandwidth obtained for the right side of the cut-off point. The

[24] We followed IK and CCT to obtain the functional form. We fit the fifth-order global polynomial with different coefficients for the right and the left of the cut-off point after rescaling.

[25] Algorithms provided by Imbens and Kalyanaraman (2009) and IK differ slightly for computing the variances and the regularization terms. See Section 4.2 of Imbens and Kalyanaraman (2009) and Section 4.2 of IK for more details. Given that they provide a Stata code for the former and that it is used in many empirical researches, we show the result for the former. Our unreported simulation finds that two algorithms perform very similarly except Design 2 where the former performs significantly better than the latter.

[26] Our Matlab code is constructed based on the R code developed by CCT. See Calonico, Cattaneo, and Titiunik (2015) for details.



(a) Design 1. Lee (2008) Data

(Design 1 of IK and CCT)

$m_1(x) = 0.52 + 0.84x - 3.00x^2 + 7.99x^3 - 9.01x^4 + 3.56x^5$

$m_0(x) = 0.48 + 1.27x + 7.18x^2 + 20.21x^3 + 21.54x^4 + 7.33x^5$

(b) Design 2. Ludwig and Miller I (2007) Data

(Design 2 of CCT)

$m_1(x) = 0.26 + 18.49x - 54.8x^2 + 74.3x^3 - 45.02x^4 + 9.83x^5$

$m_0(x) = 3.70 + 2.99x + 3.28x^2 + 1.45x^3 + 0.22x^4 + 0.03x^5$

(c) Design 3. Constant Additive Treatment Effect

(Design 3 of IK)

$m_1(x) = 0.52 + 0.84x - 3.0x^2 + 7.99x^3 - 9.01x^4 + 3.56x^5$

$m_0(x) = 0.42 + 0.84x - 3.0x^2 + 7.99x^3 - 9.01x^4 + 3.56x^5$

(d) Design 4. Ludwig and Miller II (2007, Figure II. B) Data

$m_1(x) = 0.09 + 5.76x - 42.56x^2 + 120.90x^3 - 139.71x^4 + 55.59x^5$

$m_0(z) = 0.03 - 2.26x - 13.14x^2 - 30.89x^3 - 31.98x^4 - 12.1x^5$

FIGURE 3. Simulation Design. The dashed line in the panel for Design 1 denotes the density of the forcing variable. The supports for $m_1(x)$ and $m_0(x)$ are $x \geq 0$ and $x < 0$, respectively.

sixth and seventh columns report the corresponding ones on the left sides for MMSE and IND. The eighth and ninth columns report the bias (Bias) and the root mean squared error (RMSE) for the sharp RD estimate. Bias and RMSE are 5% trimmed versions since unconditional finite sample variance of local linear estimators is infinite (see Seifert and Gasser (1996)). The tenth column reports efficiency relative to the most efficient bandwidth selection rule based on the RMSE. The eleventh and twelfth columns report RMSE and efficiency based on the true objective functions for the respective bandwidth selec-

Table 1. Bias and RMSE for $\hat{\tau}_{\text{SRD}}(c)$.

			$\hat{h}_1$		$\hat{h}_0$		$\hat{\tau}_{\text{SRD}}$				
n	Design	Method	Mean	SD	Mean	SD	Bias	RMSE	Eff	RMSE*	Eff*
500	1	MMSE	0.333	0.165	0.380	0.158	0.027	0.051	0.959	0.062	1
		IND	0.767	0.574	0.590	0.402	0.039	0.049	1	0.063	0.981
		IK	0.432	0.115			0.038	0.051	0.970	0.063	0.986
		CCT	0.204	0.043			0.019	0.053	0.928	0.063	0.978
	2	MMSE	0.074	0.005	0.187	0.033	0.038	0.076	1	0.081	1
		IND	0.145	0.012	0.279	0.019	0.115	0.126	0.603	0.083	0.979
		IK	0.177	0.010			0.138	0.148	0.513	0.088	0.913
		CCT	0.097	0.011			0.047	0.085	0.886	0.088	0.913
	3	MMSE	0.309	0.159	0.205	0.043	−0.022	0.053	0.942	0.046	1
		IND	0.354	0.283	0.180	0.062	−0.007	0.050	1	0.047	0.988
		IK	0.199	0.029			−0.014	0.051	0.988	0.046	0.998
		CCT	0.154	0.014			−0.007	0.054	0.929	0.051	0.903
	4	MMSE	0.259	0.091	0.701	0.203	0.009	0.054	1	0.039	1
		IND	0.612	0.535	1.218	0.974	0.058	0.065	0.833	0.072	0.530
		IK	0.337	0.073			0.074	0.083	0.654	0.077	0.494
		CCT	0.139	0.025			0.027	0.066	0.823	0.079	0.496
2000	1	MMSE	0.322	0.193	0.264	0.125	0.021	0.033	0.973	0.035	1
		IND	0.730	0.604	0.360	0.120	0.041	0.044	0.723	0.036	0.979
		IK	0.359	0.083			0.036	0.041	0.780	0.036	0.987
		CCT	0.181	0.040			0.016	0.032	1	0.036	0.984
	2	MMSE	0.055	0.002	0.137	0.010	0.021	0.042	1	0.046	1
		IND	0.109	0.004	0.200	0.009	0.066	0.072	0.589	0.047	0.979
		IK	0.121	0.004			0.069	0.076	0.558	0.051	0.913
		CCT	0.070	0.006			0.025	0.048	0.883	0.051	0.913
	3	MMSE	0.299	0.155	0.166	0.026	−0.009	0.028	0.989	0.026	1
		IND	0.288	0.217	0.148	0.067	−0.003	0.028	0.989	0.027	0.950
		IK	0.160	0.024			−0.007	0.028	1	0.026	0.999
		CCT	0.130	0.009			−0.004	0.029	0.965	0.028	0.903
	4	MMSE	0.257	0.082	0.600	0.188	0.014	0.034	1	0.022	1
		IND	0.528	0.442	0.994	0.837	0.054	0.057	0.604	0.041	0.520
		IK	0.274	0.420			0.066	0.070	0.487	0.044	0.485
		CCT	0.106	0.017			0.019	0.039	0.872	0.045	0.488

Note: n is the sample size. "Eff" stands for the efficiency based on RMSE relative to the method which performs best. RMSE* and Eff* are based on the infeasible bandwidths which depend on the true values of parameters.

tion rules. These can be considered as the theoretical predictions based on asymptotic analysis.[27]

We now discuss the simulation results. The sign of the product of the second derivatives is negative for Designs 1 and 2. Thus, the AMSEs for all the bandwidth selection rules converge in the same rate, $n^{-4/5}$, where n is the sample size. For Design 1, theoretical efficiency is not so different across different bandwidth selection rules. Reflecting

[27] A detailed procedure to obtain RMSE* is provided in the Supplemental Material.

this, the simulation results show similar performances, for sample size 500, across different bandwidth selection rules. As the sample size increases, however, the performance of MMSE and CCT improve relative to IND and IK. Note that the relative efficiency of the MMSE is higher than the asymptotic prediction for sample size 2000. This is attained by the finite sample performance of MMSE, in terms of RMSE, realizing close to the theoretical prediction. We conjecture that the same reasoning holds for CCT.

For Design 2, the magnitude of the ratio of the second derivatives is larger for this design compared with Design 1, so that the RMSE is larger relative to Design 1 for the same sample size. For Design 2, we observe the same tendency as that of Design 1 with greater difference. Relative performance of IK is worse for this design compared to the performance in Design 1 reflecting the theoretical relative efficiency loss of IK for this design. We note that CCT improve over IK significantly because of the refinement such as the nearest neighbor variance and regularization term based on the LPR. The efficiency gain of MMSE is about 10% relative to CCT and about 42% relative to IK. The observations made for $n = 500$ also hold when $n = 2000$.

Next, we turn to Designs 3 and 4, in which the sign of the product of the second derivatives is positive. In general, these cases should show the advantage of MMSE over IND, as the AMSE for it converges with rate $n^{-6/7}$ whereas IND's AMSE converges with rate $n^{-4/5}$. For Design 4, the same rate advantage holds for MMSE over IK and CCT. However, the second derivatives are the same for Design 3, which corresponds to the exceptional case as discussed in Section 2.2.

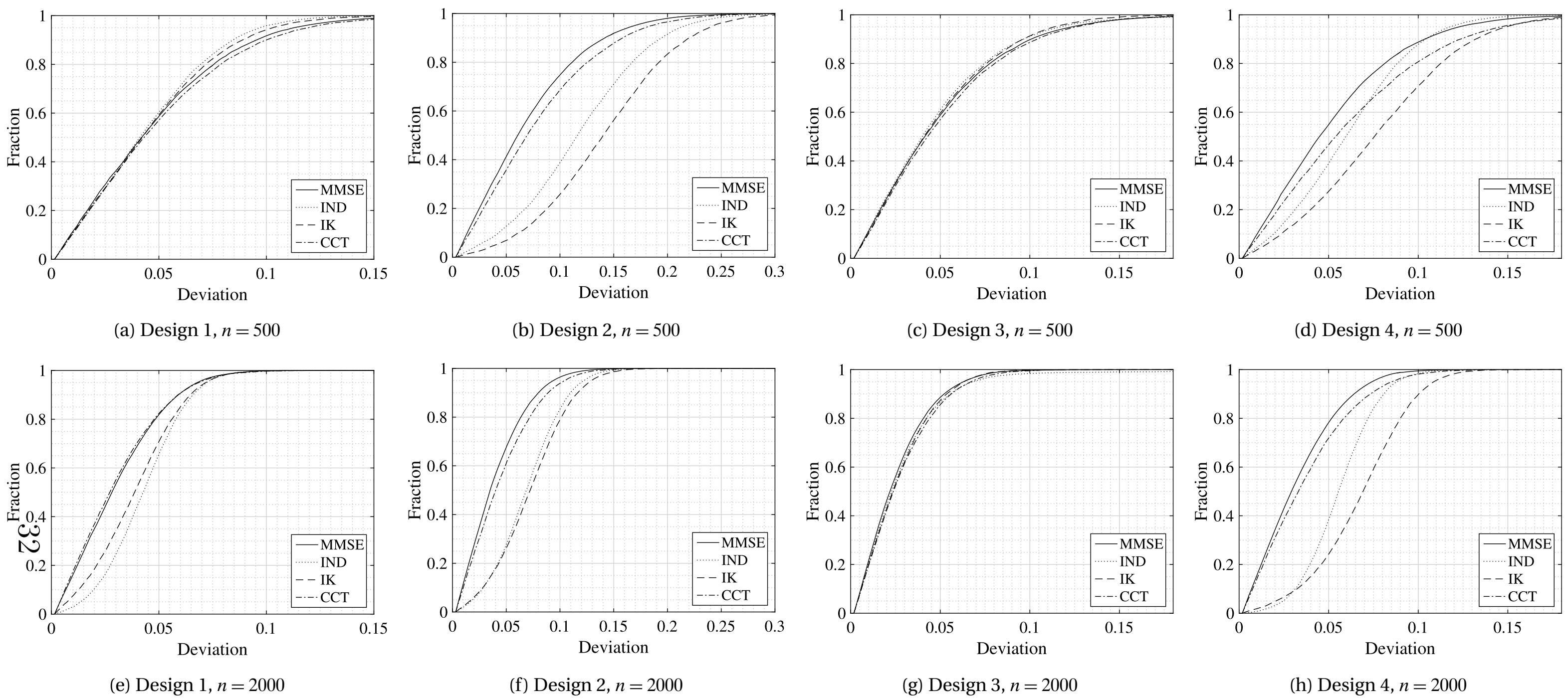
For Design 3, while there are some variations when $n = 500$, the performances of all bandwidth selection rules match the asymptotic theoretical predictions when $n = 2000$.

Design 4 is the design in which the theoretical prediction of the performance of the MMSE clearly dominates other bandwidth selection rules. And the simulation results demonstrate this. IND, IK, and CCT bandwidths tend to lead to larger biases. We emphasize here that the main advantage of using the proposed bandwidth selection rule is to take advantage of situations like Design 4 without incurring much penalty in other cases.

In summary, the simulation demonstrates (i) the performance of MMSE is close to the theoretical prediction while IND and IK suffer from the finite sample approximation, (ii) CCT improves IK significantly, especially when the magnitude of the ratio of the second derivatives is large, and (iii) MMSE dominates others when the magnitude of the ratio of the second derivatives is large.

Note that the comparison based on the RMSE can understate the difference between different bandwidth selection rules. This happens because large bias and very small variance can lead to reasonable size of the RMSE but this implies that RD estimators are concentrated on the biased value. Figure 4 shows the simulated CDF of $|\hat{\tau} - \tau|$ for different bandwidth selection rules for 10,000 simulations. Figure 4 visualizes the results presented in Table 1 and confirms the observations made for Table 1.

Table 2 shows the simulation results analogous to those in Table 1 but based on the bandwidths which are half and double the optimal bandwidths. This experiment conveys an implication of a robustness check, which are commonly implemented in empirical research. We focus on the results of MMSE and CCT since they exhibit a clear



(a) Design 1, $n = 500$

(b) Design 2, $n = 500$

(c) Design 3, $n = 500$

(d) Design 4, $n = 500$

(e) Design 1, $n = 2000$

(f) Design 2, $n = 2000$

(g) Design 3, $n = 2000$

(h) Design 4, $n = 2000$

FIGURE 4. Simulated CDF of $|\hat{\tau} - \tau|$ for different bandwidth selection rules for 10,000 simulations.

Table 2. Bias and RMSE for $\hat{\tau}_{\text{SRD}}(c)$.

			$\hat{h}/2$			$\hat{2}h$		
n	Design	Method	Bias	RMSE	Eff	Bias	RMSE	Eff
500	1	MMSE	0.015	0.058	0.846	0.028	0.040	1.219
		IND	0.033	0.050	0.988	0.029	0.037	1.314
		IK	0.023	0.050	0.986	0.027	0.037	1.332
		CCT	0.007	0.066	0.747	0.035	0.049	1.008
	2	MMSE	0.012	0.096	0.790	0.134	0.145	0.521
		IND	0.033	0.075	1.011	0.360	0.365	0.207
		IK	0.041	0.081	0.933	0.396	0.401	0.189
		CCT	0.016	0.100	0.756	0.155	0.168	0.450
	3	MMSE	−0.004	0.062	0.799	−0.148	0.171	0.292
		IND	0.001	0.064	0.782	−0.091	0.118	0.423
		IK	−0.002	0.066	0.757	−0.131	0.149	0.334
		CCT	−0.001	0.075	0.668	−0.058	0.071	0.700
	4	MMSE	−0.019	0.064	0.841	0.050	0.060	0.894
		IND	0.025	0.051	1.059	0.053	0.060	0.908
		IK	0.042	0.065	0.832	0.063	0.070	0.770
		CCT	0.011	0.082	0.660	0.062	0.076	0.713
2000	1	MMSE	0.009	0.034	0.944	0.031	0.035	0.911
		IND	0.023	0.031	1.016	0.028	0.031	1.019
		IK	0.019	0.031	1.026	0.028	0.031	1.029
		CCT	0.006	0.036	0.896	0.035	0.039	0.814
	2	MMSE	0.006	0.052	0.813	0.078	0.083	0.508
		IND	0.018	0.042	1.000	0.223	0.224	0.188
		IK	0.019	0.046	0.919	0.223	0.225	0.187
		CCT	0.006	0.056	0.752	0.088	0.095	0.455
	3	MMSE	−0.001	0.033	0.856	−0.072	0.078	0.359
		IND	0.001	0.035	0.793	−0.046	0.065	0.428
		IK	−0.001	0.036	0.773	−0.066	0.076	0.370
		CCT	0.000	0.040	0.707	−0.035	0.041	0.688
	4	MMSE	−0.011	0.033	1.043	0.053	0.056	0.612
		IND	0.020	0.041	0.837	0.054	0.056	0.610
		IK	0.032	0.043	0.796	0.068	0.070	0.484
		CCT	0.006	0.046	0.738	0.050	0.057	0.600

Note: n is the sample size. “Eff” stands for the efficiency based on RMSE relative to the method which performs best in Table 1.

pattern although IND and IK perform more or less similarly especially when $n = 2000$. For all designs and sample sizes, the average bias decreases for $\hat{h}/2$ at the cost of more variation, leading to increase in RMSE. For all designs and sample sizes except Design 1 with $n = 500$, the larger bandwidths, $2\hat{h}$, increase bias, and consequently the efficiencies in terms of RMSE deteriorate. This is more evident for the case of $n = 2000$ than $n = 500$. For the case of $n = 2000$, the efficiency losses are at least 9% and can be as large as 64%. These results imply that care must be taken in interpreting results of a commonly implemented robustness check. Unless underlying functional forms of the estimand are close

Table 3. Bias and RMSE for the conditional mean functions, $n = 500$.

		$\hat{m}_1(c)$		$\hat{m}_0(c)$	
Design	Method	Bias	RMSE	Bias	RMSE
1	MMSE	0.005	0.038	−0.021	0.037
	IK	0.011	0.032	−0.027	0.037
2	MMSE	0.028	0.071	−0.010	0.040
	IK	0.128	0.137	−0.010	0.039
3	MMSE	0.004	0.039	0.027	0.049
	IK	0.007	0.039	0.021	0.045
4	MMSE	0.106	0.120	0.098	0.101
	IK	0.139	0.145	0.066	0.074

to be linear, it is very natural to observe that the resulting point estimates are rather different from those based on the optimal bandwidths due to the increased variability for $\hat{h}/2$ and bias for $2\hat{h}$, and hence interpreting them is not straightforward at all. It would be more sensible to use various optimal bandwidths for a robustness check. In this respect, MMSE complements the existing optimal bandwidths nicely since MMSE is based on a different principle from IK and CCT.

Next, we show that the proposed method also estimates not only the treatment effects but also each conditional mean function at the cut-off point reasonably well. The discussion provided in the previous section might have made an impression that the proposed method produces larger bias in estimating the conditional mean functions when the sign of the products of the second derivatives is positive while keeping the bias of the "difference" of the conditional mean functions small because removing the first-order bias term could incur larger bandwidths. This could be true if the second-order bias term does not work well as a penalty. Table 3 reports the bias and the RMSE for the conditional mean functions, $m_1(c)$ and $m_0(c)$, at the cut-off point. There is no evidence that the proposed method estimates the RD parameter with larger bias of the estimates for the conditional mean functions.

Finally, we show the simulation results concerning confidence interval. Table 4 shows empirical coverage (EC) and average interval length (AL) for various methods. "Nonvalid" stands for the $100(1-\alpha)\%$ conventional confidence intervals given by

$$I_{\mathrm{SRD}}(h) = \left[\hat{\tau}_{\mathrm{SRD}}(h) \pm z_{\alpha/2}\sqrt{\hat{V}_{\mathrm{SRD}}(h)}\right],$$

where $\hat{\tau}_{\mathrm{SRD}}(h)$ is the bias-uncorrected sharp RD estimate and $\hat{V}_{\mathrm{SRD}}(h)$ is the commonly used conventional variance using a bandwidth h.

"Conventional" stands for the $100(1-\alpha)\%$ conventional confidence interval, commonly considered in the nonparametric literature, given by

$$I_{\mathrm{SRD}}(h) = \left[\hat{\tau}^{\mathrm{bc}}_{\mathrm{SRD}}(h, h_2, h_3) \pm z_{\alpha/2}\sqrt{\hat{V}_{\mathrm{SRD}}(h)}\right],$$

Table 4. Empirical coverage and average length for 95% confidence interval.

			Nonvalid		Conventional		US	Robust		
n	Design		MMSE	CCT	MMSE	CCT	MMSE	MMSE	CCT	CCF
500	1	EC	0.861	0.902	0.743	0.812	0.937	0.970	0.934	0.931
		AL	0.175	0.202	0.175	0.202	0.271	0.367	0.300	0.354
	2	EC	0.898	0.876	0.660	0.810	0.913	0.958	0.936	0.935
		AL	0.287	0.319	0.287	0.319	0.511	0.635	0.508	0.629
	3	EC	0.922	0.929	0.734	0.806	0.947	0.970	0.929	0.922
		AL	0.204	0.230	0.204	0.230	0.318	0.411	0.344	0.409
	4	EC	0.837	0.878	0.765	0.790	0.898	0.956	0.923	0.921
		AL	0.183	0.246	0.183	0.246	0.256	0.340	0.370	0.443
2000	1	EC	0.768	0.844	0.700	0.793	0.923	0.960	0.935	0.941
		AL	0.095	0.107	0.095	0.107	0.164	0.195	0.157	0.189
	2	EC	0.910	0.879	0.678	0.802	0.941	0.954	0.935	0.931
		AL	0.161	0.170	0.161	0.170	0.297	0.321	0.251	0.306
	3	EC	0.927	0.940	0.753	0.814	0.947	0.963	0.943	0.939
		AL	0.110	0.124	0.110	0.124	0.182	0.207	0.182	0.221
	4	EC	0.815	0.867	0.776	0.796	0.919	0.951	0.933	0.933
		AL	0.094	0.138	0.094	0.138	0.140	0.163	0.203	0.247

Note: n is the sample size. EC and AL stand for empirical coverage (%) and average interval length, respectively. "Nonvalid" and "Conventional" are constructed with the conventional standard error without and with bias correction, respectively, "US" stands for undersmoothing, "Robust" stands for the CCT-type robust confidence interval. "MMSE," "CCT," and "CCF" stand for bandwidths used to construct the confidence intervals.

where $\hat{\tau}^{\text{bc}}_{\text{SRD}}(h, h_2, h_3)$ is the bias-uncorrected sharp RD estimate. For both "Nonvalid" and "Conventional," MMSE uses the proposed bandwidths, $\hat{h}$, and CCT employs the bandwidth, $\hat{h}_{\text{CCT}}$ proposed by Section 2.6 of Calonico, Cattaneo, and Titiunik (2014b).

For MMSE, the confidence interval based on the undersmoothing idea denoted by "US" is also presented. "US" is constructed as

$$I^{\text{US}}_{\text{SRD}}(h) = \left[\hat{\tau}_{\text{SRD}}(h) \pm z_{\alpha/2}\sqrt{\hat{V}_{\text{SRD}}(h)}\right],$$

where h is the undersmoothed bandwidth, $\hat{h} = n^{-1/6}\hat{h}_{\text{MMSE}}$. This multiplying factor $n^{-1/6}$ is ad hoc but turned out to produce reasonable coverage probability. Though the "Nonvalid" confidence interval is asymptotically invalid because the nonnegligible bias term is ignored, the "US" confidence interval is asymptotically valid.

"Robust" signifies the CCT-type robust confidence interval. The robust confidence intervals by CCT and CCF are given by

$$I^{\text{bc}}_{\text{SRD,CCT}}(h, h_2) = \left[\hat{\tau}^{\text{bc}}_{\text{SRD,CCT}}(h, h_2) \pm z_{\alpha/2}\sqrt{\hat{V}^{\text{bc}}_{\text{SRD,CCT}}(h, h_2)}\right],$$

where the bias-corrected sharp RD estimate $\hat{\tau}^{\text{bc}}_{\text{SRD,CCT}}(h, h_2)$ and the robust variance $\hat{V}^{\text{bc}}_{\text{SRD,CCT}}(h, h_2)$ are given in Theorem 1 of CCT. CCT is computed by using $(\hat{h}_{\text{CCT}}, \hat{h}_{\text{CCT}})$

for (h, h_2) and CCF uses $(n^{-1/20}\hat{h}_{\text{CCT}}, n^{-1/20}\hat{h}_{\text{CCT}})$ as discussed in Calonico, Cattaneo, and Farrell (forthcoming).

For MMSE, the robust confidence interval is given by (12). The bandwidth we use for (h, h_2, h_3) is $(\hat{h}_{\text{MMSE}}, \hat{h}_{\text{MMSE}}, \hat{h}_{\text{MMSE}})$ when the estimated signs of the second derivatives are distinct as in CCT and $n^{-1/25} \times (\hat{h}_{\text{MMSE}}, \hat{h}_{\text{MMSE}}, \hat{h}_{\text{MMSE}})$ otherwise. This modification is due to the fact that the bandwidths are of order $n^{-1/7}$ when the signs are the same. When the bandwidths are of order $n^{-1/7}$, the order of the asymptotically negligible component in bias estimation is close to that of the sharp RD estimate and it affects the performance of the confidence interval for Design 4.

Table 4 shows that the performance of the "Nonvalid" and "Conventional" confidence intervals is unstable and tend to be lower than the nominal coverage probability (95%). We observe that the CCT-type robust confidence intervals improve empirical coverage. The CCT-type robust confidence intervals proposed in the previous section perform reasonably well relative to those by CCT and Calonico, Cattaneo, and Farrell (forthcoming). While the empirical coverages of CCT and CCF are below the nominal coverage probability, those of MMSE tend to be conservative.

5. Empirical illustration

We illustrate how the proposed method in this paper can contribute to empirical research. In doing so, we revisit the problem considered by Ludwig and Miller (2007). They investigated the effect of Head Start (hereafter HS) on health and schooling. HS is the federal government's program aimed to provide preschool, health, and other social services to poor children, age three to five and their families. They note that the federal government assisted HS proposals of the 300 poorest counties based on the county's 1960 poverty rate and find that the county's 1960 poverty rate can become the assignment variable where the cut-off value is given by 59.1984.[28] They assess the effect of HS assistance on numerous measures such as HS participation, HS spending, other social spending, health, mortality, and education. The effect of HS on child mortality is extensively reexamined by Cattaneo, Titiunik, and Vazquez-Bare (2017) using a local randomization framework as well as a nonparametric framework.

Here, we revisit the study on the effect of HS assistance on HS spending and mortality provided in Tables II and III of Ludwig and Miller (2007). The outcome variables considered in Tables II and III include HS spending per child in 1968 and 1972, and the mortality rate for the causes of death that could be affected by the Head Start health services to all and black children age five to nine. The 1972 HS spending per child and the mortality rate for all children generated the simulation Designs 2 and 4 in the previous section, respectively. In obtaining the RD estimates, they employ the LLR using a triangular kernel function as proposed by Porter (2003). For bandwidths, they use three different bandwidths, 9, 18, and 36 in a somewhat ad-hoc manner rather than relying on some bandwidths' selection methods. This implies that the bandwidths and the number

[28] Since the poverty rate is based on the county level information, the sampling framework does not exactly correspond to the one considered in the paper. However, in this illustration we follow the estimation framework used by Ludwig and Miller (2007), which fits into our framework.

of observations with nonzero weight used for estimation are independent of outcome variables.

Columns 3 to 5 in Table 5 reproduce the results presented in Tables II and III of Ludwig and Miller (2007) for comparison. The point estimates for 1968 HS spending per child range from 114.711 to 137.251. Perhaps we may say that they are not very sensitive to the choice of bandwidth in this case. However, the point estimates for 1972 HS spending per child range from 88.959 to 182.396. What is more troubling would be the fact that they produce mixed results in statistical significance. For 1968 HS spending per child, the point estimate with the bandwidth of 36 produce the result, which is statistically significant at the 5% level while the estimates with bandwidths of 9 and 18 are not statistically significant even at the 10% level. The results for 1972 HS spending per child are similar in the sense that the estimates based on the bandwidths of 9 and 36 are statistically significant at the 10% level while the estimate based on the bandwidth of 18 is not at the same level. We also note that the 1% and 5% statistical significance denoted by $*$ and $**$, respectively, for the LM are the ones reported in LM. They are based on the percentile-T bootstrap and they may not be asymptotically valid as pointed by CCT.

The results on the mortality rate for all children age five to nine exhibit statistical significance though the point estimates range from -1.895 to -1.114 depending on which bandwidth to employ. The point estimate for the mortality rate for black children age five to nine with bandwidth 18 is -2.719, which is statistically significant at the 5% level while the point estimates with bandwidths 9 and 36 are -2.275 and -1.589, respectively, which are not statistically significant even at 10% level. It would be meaningful to see what sophisticated bandwidth selection methods can offer under situations where the results based on ad-hoc approaches cannot be interpreted easily.

Columns 6 to 8 in Table 5 present the result based on the bandwidth selection methods based on MMSE, IK and CCT. The p-value and confidence interval are constructed by the CCT-type robust approach in contrast to that used by LM. For 1968 and 1972 HS spending per child, the point estimates differ substantially but they are all statistically insignificant. For the mortality rate for all children age five to nine, MMSE and IK methods produce similar point estimates while the point estimate by CCT is slightly bigger. They all agree on the statistical significance. For black children, the point estimates differ but they are all statistically insignificant. To summarize, we found large but statistically insignificant point estimates for HS spending and statistically significant estimates for mortality rates for all children but not for black children. What is comforting is that MMSE, IK, and CCT agree on statistical significance although they produce the different point estimates. The results presented in Table 5 alone do not imply any superiority of the proposed method over the existing methods because we never know true causal relationships. However, the results based on the proposed method should provide a meaningful perspective given the simulation experiments demonstrated in the previous section especially when the curvatures of the conditional mean functions look rather different.

Finally, we also present, in Table 6, the estimation results based on $\hat{h}/2$ and $2\hat{h}$ as presented in various empirical research as a robustness check. We can observe considerable variation in the point estimates as well as statistical significance, which is expected by the simulation results.

Table 5. RD estimates of the effect of head start assistance.

Variable			LM		MMSE	IK	CCT
1968 HS spending	Bandwidth	9	18	36	⟨11.261, 11.488⟩	19.013	6.561
	No. of obs.	{310, 217}	{674, 287}	{1877, 300}	{385, 238}	{727, 290}	{222, 178}
	RD estimate	137.251	114.711	134.491**	139.333	108.128	137.035
	Robust p-value				0.689	0.219	0.381
	Robust 95% CI				[−518.597, 784.742]	[−75.224, 328.065]	[−112.445, 293.898]
	US 95% CI				[−194.963, 417.464]		
1972 HS spending	Bandwidth	9	18	36	⟨27.006, 18.224⟩	20.9235	6.917
	No. of obs.	[217, 310]	[287, 674]	[300, 1877]	[754, 251]	[824, 294]	[238, 185]
	RD estimate	182.119*	88.959	130.153*	106.338	89.102	118.593
	Robust p-value				0.512	0.272	0.568
	Robust 95% CI				[−337.215, 575.674]	[−89.353, 316.962]	[−170.124, 310.274]
	US 95% CI				[−293.484, 383.256]		
Child mortality (All)	Bandwidth	9	18	36	⟨16.028, 6.346⟩	7.074	5.225
	No. of obs.	[217, 310]	[287, 674]	[300, 1877]	[587, 170]	[243, 184]	[177, 147]
	RD estimate	−1.895**	−1.198*	−1.114**	−2.285	−2.359	−3.017
	Robust p-value				0.030	0.007	0.008
	Robust 95% CI				[−6.108, −0.306]	[−6.322, −0.981]	[−6.390, −0.946]
	US 95% CI				[−6.043, −0.302]		
Child mortality (Black)	Bandwidth	9	18	36	⟨34.865, 22.372⟩	9.832	7.402
	No. of obs.	[217, 310]	[287, 674]	[300, 1877]	[936, 252]	[312, 209]	[243, 178]
	RD estimate	−2.275	−2.719**	−1.589	−2.751	−1.394	−0.741
	Robust p-value				0.748	0.735	0.230
	Robust 95% CI				[−6.132, 4.404]	[−5.803, 4.093]	[−8.268, 1.986]
	US 95% CI				[−7.354, 4.733]		

Note: The results for LM is reproduced based on Tables II and III of Ludwig and Miller (2007) for reference. The 1% and 5% statistical significance denoted by $*$ and $**$, respectively, for the LM are the ones reported in LM. They are based on the percentile-T bootstrap and they may not be asymptotically valid. The bandwidths on the left and right of the cut-off points are presented in angle brackets for the MMSE. The numbers of observations with nonzero weight on the left and right of the cut-off are shown in curly brackets. "Robust p-value" is obtained by the CCT-type robust t value based on the bias correction and robust standard error as in Cattaneo, Titiunik, and Vazquez-Bare (2017). For IK and CCT, the "Robust 95% CI" is computed as in CCT. For MMSE, "Robust 95% CI" and "US 95% CI" are the CCT-type robust and undersmoothed CIs constructed by the method described in Section 3.2, respectively.

Table 6. RD estimates of the effect of head start assistance.

Variable		$\hat{h}/2$			$2\hat{h}$		
		MMSE	IK	CCT	MMSE	IK	CCT
1968 HS spending	Bandwidth	⟨5.630, 5.744⟩	9.5063	3.2804	⟨22.522, 22.975⟩	38.025	13.121
	No. of obs.	[186, 162]	[331, 226]	[100, 96]	[909, 298]	[2031, 300]	[463, 250]
	RD estimate	125.672	137.918	114.071	117.623	122.619	124.494
	SE	357.771	102.96	100.781	364.667	102.055	100.193
	Robust p-value	0.917	0.249	0.984	0.780	0.403	0.139
	Robust 95% CI	[−663.811, 738.652]	[−83.119, 320.485]	[−195.528, 199.532]	[−612.920, 816.573]	[−114.617, 285.437]	[−48.159, 344.598]
	US 95% CI	[−299.201, 323.115]			[−159.142, 438.088]		
1972 HS spending	Bandwidth	⟨13.503, 9.112⟩	10.462	3.458	⟨54.012, 36.448⟩	41.847	13.833
	No. of obs.	[342, 180]	[360, 230]	[107, 101]	[2123, 299]	[2321, 300]	[500, 258]
	RD estimate	92.860	150.714	134.262	108.645	119.644	119.936
	SE	274.927	107.076	148.10	285.382	101.826	103.991
	Robust p-value	0.848	0.432	0.096	0.857	0.492	0.144
	Robust 95% CI	[−486.087, 591.627]	[−125.777, 293.960]	[−43.879, 536.674]	[−508.023, 610.673]	[−129.678, 269.481]	[−52.001, 355.644]
	US 95% CI	[−130.223, 468.597]			[−157.145, 369.694]		
Child mortality (All)	Bandwidth	⟨8.014, 3.173⟩	3.537	2.612	⟨32.056, 12.692⟩	14.147	10.449
	No. of obs.	[280, 88]	[108, 103]	[80, 72]	[1561, 244]	[509, 260]	[357, 228]
	RD estimate	−3.12	−3.435	−2.908	−1.71	−1.888	−2.03
	SE	1.876	1.918	2.431	1.383	1.226	1.338
	Robust p-value	0.095	0.201	0.389	0.033	0.048	0.03
	Robust 95% CI	[−6.806, 0.548]	[−6.209, 1.308]	[−6.857, 2.671]	[−5.654, −0.233]	[−4.827, −0.021]	[−5.524, −0.279]
	US 95% CI	[−6.591, 2.012]			[−5.127, −0.989]		
Child mortality (Black)	Bandwidth	⟨17.432, 11.186⟩	4.916	3.701	⟨69.729, 44.744⟩	19.663	14.804
	No. of obs.	[402, 192]	[153, 126]	[102, 97]	[2372, 267]	[663, 263]	[481, 246]
	RD estimate	−2.029	−2.141	−3.188	−1.421	−2.407	−2.103
	SE	4.248	3.270	3.993	5.178	1.514	2.065
	Robust p-value	0.832	0.349	0.312	0.598	0.295	0.500
	Robust 95% CI	[−9.227, 7.424]	[−10.543, 2.276]	[−11.8611, 3.790]	[−12.880, 7.418]	[−4.554, 1.381]	[−5.441, 2.653]
	US 95% CI	[−5.957, 1.674]			[−6.629, 1.326]		

Note: $\hat{h}/2$ and $2\hat{h}$ stand for the bandwidths which are half and twice as large as those in Table 5. The numbers of observations with nonzero weight on the left and right of the cut-off are shown in curly brackets. "Robust p-value" is obtained by the CCT-type robust t value based on the bias correction and robust standard error as in Cattaneo, Titiunik, and Vazquez-Bare (2017). For IK and CCT, the "Robust 95% CI" is computed as in CCT. For MMSE, "Robust 95% CI" and "US 95% CI" are the CCT-type robust and undersmoothed CIs constructed by the method described in Section 3.2, respectively.

6. Conclusion

In this paper, we proposed a new bandwidth selection method for the RD estimators. A main feature of the proposed method is that we choose two bandwidths simultaneously. When we allow two bandwidths to be distinct, we showed that the minimization problem of the AMSE exhibits dichotomous characteristics depending on the sign of the product of the second derivatives of the underlying functions, but we also showed that the proposed method automatically adapted to the underlying conditions.

We provided a discussion on the validity of the simultaneous choice of the bandwidths and their theoretical advantages and illustrated through simulations that the proposed bandwidths produce results comparable to the theoretical results in the sample sizes relevant for empirical works. A simulation study based on designs motivated by existing empirical literatures exhibits nonnegligible gain of the proposed method over existing methods. We also demonstrated how the proposed method can be implemented via an empirical example. In addition, we demonstrate that the common robustness check should not be appropriate and propose the robustness check based on several optimal bandwidths.

We extended the proposed approach for the sharp RD design in two ways. First, we proposed the bandwidth selection rule for the sharp RK design following the idea developed for the sharp RD design. Second, we proposed the robust confidence intervals for the sharp RD and RK designs following CCT. We show that the general approach proposed by CCT can be implemented for the LLR estimator based on the proposed bandwidths.

The bandwidths selection rule for the fuzzy RD design is not investigated in the paper. The main obstacle to extend the idea developed in the paper is that we need to choose four bandwidths simultaneously for the fuzzy RD design. While we discuss the difficulty of the bandwidth selection problem for the sharp RD design when we allow two distinct bandwidths, the difficulty is amplified significantly when we allow four distinct bandwidths for the fuzzy RD design.[29] The extension is nontrivial, and hence left as future research.

Appendix A: Generalization of Lemma 2

Lemma A.1. *Suppose that the bias component is expanded up to $(K-1)$th-order for any integer $K>2$ and it has the following form*:

$$\left(\alpha_{1,2}h_1^2-\alpha_{0,2}h_0^2\right)+\left(\alpha_{1,3}h_1^3-\alpha_{0,3}h_0^3\right)+\cdots+\left(\alpha_{1,K}h_1^K-\alpha_{0,K}h_0^K\right).$$

Also suppose $\alpha_{1,2}\alpha_{0,2}>0$. Then there exists a combination of h_1 and h_0 such that the AMSE including up to the $(K-1)$th-order bias term becomes

$$\frac{v}{nh_1f(c)}\left\{\sigma_1^2(c)+\sigma_0^2(c)\left[\frac{\alpha_{1,2}}{\alpha_{0,2}}\right]^{1/2}\right\}+O\left(h_1^{2(\ell+K)}\right)$$

for an arbitrary nonnegative integer ℓ.

[29] This difficulty is partially solved by reducing the selection of four bandwidths to that of two bandwidths in Arai and Ichimura (2016). The development is based on the present paper.

Proof. It suffices to show that the bias component can be made arbitrarily small. Let $h_0^2 = C(h_1, k)h_1^2$ where $C(h_1, k) = C_0 + C_1 h + \cdots + C_k h^k$ for $k > K - 2$. Let $\mathcal{C}_j = \{C_0, C_1, \ldots, C_j\}$ and we write $f(C_0, C_2, \ldots, C_j)$ as $f(\mathcal{C}_j)$ for arbitrary function f which has $C_1, C_2, \ldots, C_j$ as arguments. Observe that, for $j > 2$, $C(h_1, k)^{j/2}$ can be written as

$$C(h_1, k)^{j/2} = \sum_{s=0}^{\infty} \phi_{j,s}(\mathcal{C}_s)h_1^s$$

for some $\phi_{j,s}$. For even j, $\phi_{j,s} \equiv 0$ when $s > jk/2$. We choose C_0 such that $\alpha_{1,2} - \alpha_{0,2}C_0 = 0$ holds. This choice of C_0 removes the first-order bias term. Next, we select C_1 given C_0 by

$$-C_1\alpha_{0,2} + \bigl(\alpha_{1,3} - \phi_{3,0}(\mathcal{C}_0)\alpha_{0,3}\bigr) = 0.$$

This choice of C_1 combined with C_0 described above removes the second-order bias term too. In general, given $(C_0, C_1, \ldots, C_{j-1})$, we can choose C_j for $j \le K - 2$ by setting

$$-C_j\alpha_{0,2} - \phi_{3,j-1}(\mathcal{C}_{j-1})\alpha_{0,3} - \cdots - \phi_{j+1,1}(\mathcal{C}_1)\alpha_{0,j+1} + \bigl(\alpha_{j+2,1} - \phi_{j+2,0}(\mathcal{C}_0)\alpha_{j+2,0}\bigr) = 0.$$

For j satisfying $K - 2 < j \le k$, we can choose C_j successively by

$$-C_j\alpha_{0,2} - \phi_{3,j-1}(\mathcal{C}_{j-1})\alpha_{0,3} - \cdots - \phi_{j+1,1}(\mathcal{C}_1)\alpha_{0,j+1} - \phi_{K,j-K+2}(\mathcal{C}_{j-K+2})\alpha_{K,0} = 0.$$

The choice $(C_0, C_1, \ldots, C_k)$ makes the order of bias $O(h_1^{k+3})$. This completes the proof. □

Appendix B: Proofs of Theorem 1

Recall that the objective function is

$$\mathrm{MMSE}_n^p(h) = \left\{\frac{b_1}{2}\bigl[\hat{m}_1^{(2)}(c)h_1^2 - \hat{m}_0^{(2)}(c)h_0^2\bigr]\right\}^2 + \bigl[\hat{b}_{2,1}(c)h_1^3 - \hat{b}_{2,0}(c)h_0^3\bigr]^2 + \frac{v}{n\hat{f}(c)}\left\{\frac{\hat{\sigma}_1^2(c)}{h_1} + \frac{\hat{\sigma}_0^2(c)}{h_0}\right\}.$$

To begin with, we show that $\hat{h}_1$ and $\hat{h}_0$ satisfy Assumption 2. If we choose a sequence of h_1 and h_0 to satisfy Assumption 2, then $\mathrm{MMSE}_n^p(h)$ converges to 0. Assume to the contrary that either one or both of $\hat{h}_1$ and $\hat{h}_0$ do not satisfy Assumption 2. Since $m_0^{(2)}(c)^3 b_{2,1}(c)^2 \neq m_1^{(2)}(c)^3 b_{2,0}(c)^2$ by assumption, $\hat{m}_0^{(2)}(c)^3\hat{b}_{2,1}(c)^2 \neq \hat{m}_1^{(2)}(c)^3\hat{b}_{2,0}(c)^2$ with probability approaching 1. Without loss of generality, we assume this as well. Then at least one of the first-order bias terms, the second-order bias term, and the variance term of $\mathrm{MMSE}_n^p(\hat{h})$ does not converge to zero in probability. Then $\mathrm{MMSE}_n^p(\hat{h}) > \mathrm{MMSE}_n^p(h)$ holds for some n. This contradicts the definition of $\hat{h}$. Hence $\hat{h}$ satisfies Assumption 2.

We first consider the case in which $m_1^{(2)}(c)m_0^{(2)}(c) < 0$. In this case, with probability approaching 1, $\hat{m}_1^{(2)}(c)\hat{m}_0^{(2)}(c) < 0$, so that we assume this without loss of generality. When this holds, note that the leading terms are the first term and the last term of $\mathrm{MMSE}_n^p(\hat{h})$ since $\hat{h}_1$ and $\hat{h}_0$ satisfy Assumption 2. Define the plug-in version of $\mathrm{AMSE}_{1n}(h)$ provided in (5) by

$$\mathrm{AMSE}_{1n}^p(h) = \left\{ \frac{b_1}{2} \left[\hat{m}_1^{(2)}(c) h_1^2 - \hat{m}_0^{(2)}(c) h_0^2 \right] \right\}^2 + \frac{v}{n \hat{f}(c)} \left\{ \frac{\hat{\sigma}_1^2(c)}{h_1} + \frac{\hat{\sigma}_0^2(c)}{h_0} \right\}.$$

A calculation yields $\tilde{h}_1 = \tilde{C}_1 n^{-1/5}$ and $\tilde{h}_0 = \tilde{C}_0 n^{-1/5}$ where

$$\tilde{C}_1 = \left\{ \frac{v \hat{\sigma}_1^2(c)}{b_1^2 \hat{f}(c) \hat{m}_1^{(2)}(c) \left[\hat{m}_1^{(2)}(c) - \hat{\lambda}_1^2 \hat{m}_0^{(2)}(c) \right]} \right\}^{1/5}, \quad \hat{\lambda}_1 = \left\{ - \frac{\hat{\sigma}_0^2(c) \hat{m}_1^{(2)}(c)}{\hat{\sigma}_1^2(c) \hat{m}_0^{(2)}(c)} \right\}^{1/3},$$

and $\tilde{C}_0 = \tilde{C}_1 \hat{\lambda}_1$. With this choice, $\mathrm{AMSE}_{1n}^p(\tilde{h})$, and hence $\mathrm{MMSE}_n^p(\tilde{h})$ converges at the rate of $n^{-4/5}$. Note that if $\hat{h}_1$ or $\hat{h}_0$ converges at the rate slower than $n^{-1/5}$, then the bias term converges at the rate slower than $n^{-4/5}$. If $\hat{h}_1$ or $\hat{h}_0$ converges at the rate faster than $n^{-1/5}$, then the variance term converges at the rate slower than $n^{-4/5}$. Thus, the minimizer of $\mathrm{MMSE}_n^p(h)$, $\hat{h}_1$, and $\hat{h}_0$ converges to 0 at rate $n^{-1/5}$.

Thus, we can write $\hat{h}_1 = \hat{C}_1 n^{-1/5} + o_p(n^{-1/5})$ and $\hat{h}_0 = \hat{C}_0 n^{-1/5} + o_p(n^{-1/5})$ for some $O_P(1)$ sequences $\hat{C}_1$ and $\hat{C}_0$ that are bounded away from 0 and ∞ as $n \to \infty$. Using this expression,

$$\begin{aligned} \mathrm{MMSE}_n^p(\hat{h}) = {} & n^{-4/5} \left\{ \frac{b_1}{2} \left[\hat{m}_1^{(2)}(c) \hat{C}_1^2 - \hat{m}_0^{(2)}(c) \hat{C}_0^2 \right] \right\}^2 \\ & + \frac{v}{n^{4/5} \hat{f}(c)} \left\{ \frac{\hat{\sigma}_1^2(c)}{\hat{C}_1} + \frac{\hat{\sigma}_0^2(c)}{\hat{C}_0} \right\} + o_p(n^{-4/5}). \end{aligned}$$

Note that

$$\begin{aligned} \mathrm{MMSE}_n^p(\tilde{h}) = {} & n^{-4/5} \left\{ \frac{b_1}{2} \left[\hat{m}_1^{(2)}(c) \tilde{C}_1^2 - \hat{m}_0^{(2)}(c) \tilde{C}_0^2 \right] \right\}^2 \\ & + \frac{v}{n^{4/5} \hat{f}(c)} \left\{ \frac{\hat{\sigma}_1^2(c)}{\tilde{C}_1} + \frac{\hat{\sigma}_0^2(c)}{\tilde{C}_0} \right\} + O_P(n^{-6/5}). \end{aligned}$$

Since $\hat{h}$ is the optimizer, $\mathrm{MMSE}_n^p(\hat{h}) / \mathrm{MMSE}_n^p(\tilde{h}) \le 1$. Thus,

$$\frac{\left\{ \frac{b_1}{2} \left[\hat{m}_1^{(2)}(c) \hat{C}_1^2 - \hat{m}_0^{(2)}(c) \hat{C}_0^2 \right] \right\}^2 + \frac{v}{\hat{f}(c)} \left\{ \frac{\hat{\sigma}_1^2(c)}{\hat{C}_1} + \frac{\hat{\sigma}_0^2(c)}{\hat{C}_0} \right\} + o_p(1)}{\left\{ \frac{b_1}{2} \left[\hat{m}_1^{(2)}(c) \tilde{C}_1^2 - \hat{m}_0^{(2)}(c) \tilde{C}_0^2 \right] \right\}^2 + \frac{v}{\hat{f}(c)} \left\{ \frac{\hat{\sigma}_1^2(c)}{\tilde{C}_1} + \frac{\hat{\sigma}_0^2(c)}{\tilde{C}_0} \right\} + O_P(n^{-2/5})} \le 1.$$

Note that the denominator converges to

$$\left\{ \frac{b_1}{2} \left[m_1^{(2)}(c) C_1^{*2} - m_0^{(2)}(c) C_0^{*2} \right] \right\}^2 + \frac{v}{f(c)} \left\{ \frac{\sigma_1^2(c)}{C_1^*} + \frac{\sigma_0^2(c)}{C_0^*} \right\},$$

where C_1^* and C_0^* are the unique optimizers of

$$\left\{\frac{b_1}{2}\left[m_1^{(2)}(c)C_1^2 - m_0^{(2)}(c)C_0^2\right]\right\}^2 + \frac{v}{f(c)}\left\{\frac{\sigma_1^2(c)}{C_1} + \frac{\sigma_0^2(c)}{C_0}\right\},$$

with respect to C_1 and C_0. This implies that $\hat{C}_1$ and $\hat{C}_0$ also converge to the same respective limit C_1^* and C_0^* because the inequality will be violated otherwise.

Next, we consider the case with $m_1^{(2)}(c)m_0^{(2)}(c) > 0$. In this case, with probability approaching 1, $\hat{m}_1^{(2)}(c)\hat{m}_0^{(2)}(c) > 0$, so that we assume this without loss of generality.

When these conditions hold, define $h_0 = \hat{\lambda}_2 h_1$ where $\hat{\lambda}_2 = \{\hat{m}_1^{(2)}(c)/\hat{m}_0^{(2)}(c)\}^{1/2}$. This sets the first-order bias term of $\text{MMSE}_n^p(h)$ equal to 0. Define the plug-in version of $\text{AMSE}_{2n}(h)$ in (6) by

$$\text{AMSE}_{2n}^p(h) = \left\{\hat{b}_{2,1}(c)h_1^3 - \hat{b}_{2,0}(c)h_0^3\right\}^2 + \frac{v}{n\hat{f}(c)}\left\{\frac{\hat{\sigma}_1^2(c)}{h_1} + \frac{\hat{\sigma}_0^2(c)}{h_0}\right\}.$$

Choosing h_1 to minimize $\text{AMSE}_{2n}^p(h)$, we define $\tilde{h}_1 = \tilde{C}_1 n^{-1/7}$ and $\tilde{h}_0 = \tilde{C}_0 n^{-1/7}$ where

$$\hat{\theta}_2 = \left\{\frac{v\left[\hat{\sigma}_1^2(c) + \hat{\sigma}_0^2(c)/\hat{\lambda}_2\right]}{6\hat{f}(c)\left[\hat{b}_{2,1}(c) - \hat{\lambda}_2^3\hat{b}_{2,0}(c)\right]^2}\right\}^{1/7} \quad \text{and} \quad \tilde{C}_0 = \tilde{C}_1\hat{\lambda}_2. \tag{13}$$

Then $\text{MMSE}_n^p(\tilde{h})$ can be written as

$$\text{MMSE}_n^p(\tilde{h}) = n^{-6/7}\left\{\hat{b}_{2,1}(c)\tilde{C}_1^3 - \hat{b}_{2,0}(c)\tilde{C}_0^3\right\}^2 + n^{-6/7}\frac{v}{\hat{f}(c)}\left\{\frac{\hat{\sigma}_1^2(c)}{\tilde{C}_1} + \frac{\hat{\sigma}_0^2(c)}{\tilde{C}_0}\right\}.$$

In order to match this rate of convergence, both $\hat{h}_1$ and $\hat{h}_0$ need to converge at the rate slower than or equal to $n^{-1/7}$ because the variance term needs to converge at the rate $n^{-6/7}$ or faster. In order for the first-order bias term to match this rate,

$$\hat{m}_1^{(2)}(c)\hat{h}_1^2 - \hat{m}_0^{(2)}(c)\hat{h}_0^2 \equiv B_{1n} = n^{-3/7}b_{1n},$$

where $b_{1n} = O_P(1)$ so that under the assumption that $m_0^{(2)}(c) \neq 0$, with probability approaching 1, $\hat{m}_0^{(2)}(c)$ is bounded away from 0 so that assuming this without loss of generality, we have $\hat{h}_0^2 = \hat{\lambda}_2^2\hat{h}_1^2 - B_{1n}/\hat{m}_0^{(2)}(c)$. Substituting this expression to the second term and the third term of MMSE_n^p, we have

$$\begin{aligned}\text{MMSE}_n^p(\hat{h}) = {} & \left\{\frac{b_1}{2}B_{1n}\right\}^2 + \left\{\hat{b}_{2,1}(c)\hat{h}_1^3 - \hat{b}_{2,0}(c)\left[\hat{\lambda}_2^2\hat{h}_1^2 - B_{1n}/\hat{m}_0^{(2)}(c)\right]^{3/2}\right\}^2 \\ & + \frac{v}{n\hat{f}(c)}\left\{\frac{\hat{\sigma}_1^2(c)}{\hat{h}_1} + \frac{\hat{\sigma}_0^2(c)}{\left[\hat{\lambda}_2^2\hat{h}_1^2 - B_{1n}/\hat{m}_0^{(2)}(c)\right]^{1/2}}\right\}.\end{aligned}$$

Suppose $\hat{h}_1$ is of order slower than $n^{-1/7}$. Then because $\hat{m}_0^{(2)}(c)^3\hat{b}_{2,1}(c)^2 \neq \hat{m}_1^{(2)}(c)^3\hat{b}_{2,0}(c)^2$ and this holds even in the limit, the second-order bias term is of order slower than $n^{-6/7}$. If $\hat{h}_1$ converges to 0 faster than $n^{-1/7}$, then the variance term converges at the rate slower than $n^{-6/7}$. Therefore, we can write $\hat{h}_1 = \hat{C}_1 n^{-1/7} + o_p(n^{-1/7})$ for some $O_P(1)$ sequence $\hat{C}_1$ that is bounded away from 0 and ∞ as $n \to \infty$ and as before $\hat{h}_0^2 = \hat{\lambda}_2^2 \hat{h}_1^2 - B_{1n}/\hat{m}_0^{(2)}(c)$. Using this expression, we can write

$$\begin{aligned}\mathrm{MMSE}_n^p(\hat{h}) = n^{-6/7}\left\{\frac{b_1}{2} b_{1n}\right\}^2 \\ &+ n^{-6/7}\left\{\hat{b}_{2,1}(c)\hat{C}_1^3 + o_p(1) - \hat{b}_{2,0}(c)\left[\hat{\lambda}_2^2\hat{C}_1^2 + o_p(1) - n^{-1/7} b_{1n}/\hat{m}_0^{(2)}(c)\right]^{3/2}\right\}^2 \\ &+ n^{-6/7}\frac{v}{\hat{f}(c)}\left\{\frac{\hat{\sigma}_1^2(c)}{\hat{C}_1 + o_p(1)} + \frac{\hat{\sigma}_0^2(c)}{\left[\hat{\lambda}_2^2\hat{C}_1^2 + o_p(1) - n^{-1/7} b_{1n}/\hat{m}_0^{(2)}(c)\right]^{1/2}}\right\}.\end{aligned}$$

Thus, b_{1n} converges in probability to 0. Otherwise, the first-order bias term remains and that contradicts the definition of $\hat{h}_1$.

Since $\hat{h}$ is the optimizer, $\mathrm{MMSE}_n^p(\hat{h})/\,\mathrm{MMSE}_n^p(\tilde{h}) \le 1$. Thus,

$$\frac{o_p(1) + \left\{\hat{b}_{2,1}(c)\hat{C}_1^3 - \hat{b}_{2,0}(c)\left[\hat{\lambda}_2^2\hat{C}_1^2 + o_p(1)\right]^{3/2}\right\}^2 + \dfrac{v}{\hat{f}(c)}\left\{\dfrac{\hat{\sigma}_1^2(c)}{\hat{C}_1 + o_p(1)} + \dfrac{\hat{\sigma}_0^2(c)}{\left[\hat{\lambda}_2^2\hat{C}_1^2 + o_p(1)\right]^{1/2}}\right\}}{\left\{\hat{b}_{2,1}(c)\tilde{C}_1^3 - \hat{b}_{2,0}(c)\tilde{C}_0^3\right\}^2 + \dfrac{v}{\hat{f}(c)}\left\{\dfrac{\hat{\sigma}_1^2(c)}{\tilde{C}_1} + \dfrac{\hat{\sigma}_0^2(c)}{\tilde{C}_0}\right\}} \le 1.$$

If $\hat{C}_1 - \tilde{C}_1$ does not converge to 0 in probability, then the ratio is not less than 1 at some point. Hence $\hat{C}_1 - \tilde{C}_1 = o_p(1)$. Therefore, $\hat{h}_0/\tilde{h}_0 \xrightarrow{p} 1$ as well.

The results shown above also imply that $\mathrm{MMSE}_n^p(\hat{h})/\,\mathrm{MSE}_n(h^*) \xrightarrow{p} 1$ in both cases. □

Appendix C: Detailed description of bias correction and Theorem 3

This section describes the bias correction terms and Theorem 3 discussed in Section 3.2. Let $\hat{\beta}_{1,p}(h_1)$ and $\hat{\beta}_{0,p}(h_0)$ be the LPR estimators of order p ($p \in \mathbb{N}$) with bandwidth h_1 and h_0 for the right and left of the cut-off point, respectively. Then the LPR estimator on the right of the cut-off point can be expressed as $\hat{\beta}_{1,p}(h_1) = (X_p(c)' W_{1,h_1}(c) X_p(c))^{-1} X_p(c)' W_{1,h_1}(c) Y$, where $X_p(c)$ is an $n \times (p+1)$ matrix whose ith row is given by $(1, X_i - c, \dots, (X_i - c)^p)$, $Y = (Y_1, \dots, Y_n)'$, and $W_{1,h}(c) = \mathrm{diag}\{K((X_i - c)/h)\mathbb{I}\{X_i \ge c\}/h\}$. The LPR estimator on the left of the cut-off point can be written analogously with $W_{0,h}(c) = \mathrm{diag}\{K((X_i - c)/h)\mathbb{I}\{X_i < c\}/h\}$.

Note that the LPR estimator of $m_j^{(\nu)}(c)$ based on the LPR of order p can be written as $\nu! e_\nu \hat{\beta}_{j,p}(h_j)$, for $j = 0, 1$, where e_ν is the conformable unit vector having one in the $(\nu+1)$th entry and zero in the other entry. Define $\hat{\tau}_{\nu,p}(h) = \hat{\tau}_{1,\nu,p}(h_1) - \hat{\tau}_{0,\nu,p}(h_0)$ where $\hat{\tau}_{j,\nu,p}(h_j) = \nu! e_\nu' \hat{\beta}_{j,p}(h_j)$. It follows that $\hat{\tau}_{\mathrm{SRD}}(c) = \hat{\tau}_{0,1}(h)$ and $\hat{\tau}_{\mathrm{SRK}}(c) = \hat{\tau}_{1,2}(h)$. Suppose that we estimate the $(p+1)$th and $(p+2)$th derivatives by the LPR of order q $(= p+2)$. The bias-corrected estimator of $\hat{\tau}_{\nu,p}(h)$, denoted $\hat{\tau}_{\nu,p,q}^{\mathrm{bc}}(h, h_{p+1}, h_{p+2})$ is defined by

$$\hat{\tau}_{\nu,p,q}^{\mathrm{bc}}(h, h_{p+1}, h_{p+2}) = \hat{\tau}_{1,\nu,p,q}^{\mathrm{bc}}(h_1, h_{p+1,1}, h_{p+2,1}) - \hat{\tau}_{0,\nu,p,q}^{\mathrm{bc}}(h_0, h_{p+1,0}, h_{p+2,0}),$$

where

$$\begin{aligned}\hat{\tau}^{\mathrm{bc}}_{j,\nu,p,q}(h_j, h_{p+1,j}, h_{p+2,j}) &= \hat{\tau}_{j,\nu,p}(h_j) \\ &\quad - h_j^{p+1-\nu}\mathcal{B}_{j,\nu,p,q}(h_{p+1,j}) - h_j^{p+2-\nu}\mathcal{C}_{j,\nu,p,q}(h_{p+1,j}, h_{p+2,j}), \\ \mathcal{B}_{j,\nu,p,q}(h_{r,j}) &= e'_{p+1}\hat{\beta}_{j,q}(h_{p+1,j})\vartheta_{j,\nu,p,p+1}, \\ \mathcal{C}_{j,\nu,p,q}(h_{p+1,j}, h_{p+2,j}) &= e'_{p+1}\hat{\beta}_{j,q}(h_{p+1,j})\frac{\hat{f}^{(1)}(c)}{\hat{f}(c)}\varphi_{j,\nu,p} + e'_{p+2}\hat{\beta}_{j,q}(h_{p+2,j})\vartheta_{j,\nu,p,p+2}, \\ \vartheta_{j,\nu,p,r} &= \nu! e'_{\nu} S^{-1}_{j,0,p} c_{j,r,p}, \\ \varphi_{j,\nu,p} &= \nu! e'_{\nu} S^{-1}_{j,0,p}\bigl(c_{j,p+2,p} - S_{j,1,p}S^{-1}_{j,0,p}c_{j,p+1,p}\bigr),\end{aligned}$$

with $S_{j,k,p} = (\mu_{j,k+\ell_1+\ell_2})_{0\le\ell_1,\ell_2\le p}$, $c_{j,k,p} = (\mu_{j,k+\ell})_{0\le\ell\le p}$, $\mu_{1,s} = \int_0^\infty u^s K(s)\,ds$, $\mu_{0,s} = \int_{-\infty}^0 u^s\,ds$ for $j = 0, 1$. This implies that

$$\begin{aligned}\hat{B}_{\mathrm{SRD},1}(h, h_2) &= h_1^2\mathcal{B}_{1,0,1,3}(h_{2,1}) - h_0^2\mathcal{B}_{0,0,1,3}(h_{2,0}), \\ \hat{B}_{\mathrm{SRD},2}(h, h_2, h_3) &= h_1^3\mathcal{C}_{1,0,1,3}(h_{2,1}, h_{3,1}) - h_0^3\mathcal{C}_{0,0,1,3}(h_{2,0}, h_{3,0}), \\ \hat{B}_{\mathrm{SRK},1}(h, h_3) &= h_1^2\mathcal{B}_{1,1,2,4}(h_{3,1}) - h_0^2\mathcal{B}_{0,1,2,4}(h_{3,0}),\end{aligned}$$

and

$$\hat{B}_{\mathrm{SRK},2}(h, h_3, h_4) = h_1^3\mathcal{C}_{1,1,2,4}(h_{3,1}, h_{4,1}) - h_0^3\mathcal{C}_{0,1,2,4}(h_{3,0}, h_{4,0}).$$

Let the conditional variance of the bias-corrected estimators, $\hat{\tau}^{\mathrm{bc}}_{\nu,p,q}(h, h_{p+1}, h_{p+2})$, be $V^{\mathrm{bc}}_{\nu,p,q}(h, h_{p+1}, h_{p+2})$. Then it follows that

$$V^{\mathrm{bc}}_{\nu,p,q}(h, h_{p+1}, h_{p+2}) = V^{\mathrm{bc}}_{1,\nu,p,q}(h, h_{p+1}, h_{p+2}) + V^{\mathrm{bc}}_{0,\nu,p,q}(h, h_{p+1}, h_{p+2,j}),$$

where, for $j = 0, 1$,

$$\begin{aligned}V^{\mathrm{bc}}_{j,\nu,p,q}(h_j, h_{p+1,j}, h_{p+2,j}) &= V^{(0)}_{j,\nu,p}(h_j) + V^{(1)}_{j,\nu,p,q}(h_j, h_{p+1,j}) + V^{(2)}_{j,\nu,p,q}(h_j, h_{p+1}, h_{p+2,j}) \\ &\quad - 2C^{(0,1)}_{j,\nu,p,q}(h_j, h_{p+1,j}) - 2C^{(0,2)}_{j,\nu,p,q}(h_j, h_{p+1}, h_{p+2,j}) \\ &\quad + 2C^{(1,2)}_{j,\nu,p,q}(h_j, h_{p+1,j}, h_{p+2}), \\ V^{(0)}_{j,\nu,p}(h_j) &= \nu!^2 e'_{\nu}(h_j)S^{-1}_{j,0,p}(h_j)T_{j,p,p}(h_j, h_j)S^{-1}_{j,0,p}(h_j)e_{\nu}, \\ V^{(1)}_{j,\nu,p,q}(h_j, h_{p+1,j}) &= h_j^{2(p+1-\nu)}\vartheta^2_{j,\nu,p,p+1}e'_{p+1}S^{-1}_{j,0,q}(h_{p+1,j}) \\ &\quad \times T_{j,q,q}(h_{p+1,j}, h_{p+1,j})S^{-1}_{j,0,q}(h_{p+1,j})e_{p+1}, \\ V^{(2)}_{j,\nu,p,q}(h_j, h_{p+1}, h_{p+2,j}) &= h_j^{2(p+2-\nu)} \\ &\quad \times \bigl\{\bigl(\hat{f}^{(1)}(c)/\hat{f}(c)\bigr)^2 e'_{p+1}S^{-1}_{j,0,q}(h_{p+1,j}) \\ &\quad \times T_{j,q,q}(h_{p+1,j}, h_{p+1,j})S^{-1}_{j,0,q}(h_{p+1,j})e_{p+1}\end{aligned}$$

$$
\begin{aligned}
&+\vartheta^2_{j,\nu,p,p+2}(h_j)e'_{p+2}S^{-1}_{j,0,q}(h_{p+2,j})\\
&\times T_{j,q,q}(h_{p+2,j},h_{p+2,j})S^{-1}_{j,0,q}(h_{p+2,j})e_{p+2}\\
&+2\big(\hat{f}^{(1)}(c)/\hat{f}(c)\big)\varphi_{j,\nu,p}\vartheta_{j,\nu,p,p+2}e'_{p+1}S^{-1}_{j,0,q}(h_{p+1,j})\\
&\times T_{j,q,q}(h_{p+1,j},h_{p+2,j})S^{-1}_{j,0,q}(h_{p+2,j})e_{p+2}\big\},
\end{aligned}
$$

$$
\begin{aligned}
C^{(0,1)}_{j,\nu,p,q}(h_j,h_{p+1,j}) &= h_j^{p+1-\nu}\nu!\vartheta_{j,\nu,p,r}(h_j)e'_{\nu}S^{-1}_{j,0,p}(h_j)\\
&\quad\times T_{j,p,q}(h_j,h_{p+1,j})S^{-1}_{j,0,q}(h_{p+1,j})e_{p+1},\\
C^{(0,2)}_{j,\nu,p,q}(h_j,h_{p+1,j},h_{p+2,j}) &= h_j^{p+2-\nu}\nu!\\
&\quad\times\big\{\big(\hat{f}^{(1)}(c)/\hat{f}(c)\big)\varphi_{j,\nu,p}e'_{\nu}S^{-1}_{j,0,p}(h_j)\\
&\quad\times T_{j,p,q}(h_j,h_{p+1,j})S^{-1}_{j,0,q}(h_{p+1,j})e_{p+1}\\
&\quad+\vartheta_{j,\nu,p,p+2}e'_{\nu}S^{-1}_{j,0,p}(h_j)\\
&\quad\times T_{j,p,q}(h_j,h_{p+2,j})S^{-1}_{j,0,q}(h_{p+2,j})e_{p+2}\big\},\\
C^{(1,2)}_{j,\nu,p,q}(h_j,h_{p+1,j},h_{p+2,j}) &= h_j^{2p+3-2\nu}\\
&\quad\times\big\{\big(\hat{f}^{(1)}(c)/\hat{f}(c)\big)\varphi_{j,\nu,p}e'_{p+1}S^{-1}_{j,0,q}(h_{p+1,j})\\
&\quad\times T_{j,q,q}(h_{p+1,j},h_{p+1,j})S^{-1}_{j,0,q}(h_{p+1,j})e_{p+1}\\
&\quad+\vartheta_{j,\nu,p,p+2}e'_{p+1}S^{-1}_{j,0,q}(h_{p+1,j})\\
&\quad\times T_{j,q,q}(h_{p+1,j},h_{p+2,j})S^{-1}_{j,0,q}(h_{p+2,j})e_{p+2}\big\},
\end{aligned}
$$

$S_{j,k,p}(h_j)=(s_{j,k+\ell_1+\ell_2}(h_j))_{0\le\ell_1,\ell_2\le p}$, $s_{j,k}(h)=\sum_{i=1}^n K_{j,h}(X_i-c)(X_i-c)^k$, $T_{j,k,\ell}(b_1,b_2)=\hat{\sigma}^2_j(c)X_k(c)'W_{j,b_1}(c)W_{j,b_2}(c)X_\ell(c)$, and $\hat{\sigma}^2_j$ is the consistent estimator of $\sigma^2_j(c)$ and its explicit form is provided in Section C of the Supplemental Material. Using the notation introduced in the Appendix, $V^{\mathrm{bc}}_{\mathrm{SRD}}(h,h_2,h_3)$ and $V^{\mathrm{bc}}_{\mathrm{SRK}}(h,h_3,h_4)$ can be expressed as $V^{\mathrm{bc}}_{0,1,3}(h,h_2,h_3)$ and $V^{\mathrm{bc}}_{0,2,4}(h,h_3,h_4)$, respectively.

References

Abadie, A. and G. W. Imbens (2011), “Bias-corrected matching estimators for average treatment effects.” *Journal of Business & Economic Statistics*, 29, 1–11. [443]

Arai, Y. and H. Ichimura (2013), “Supplement to ‘Optimal bandwidth selection for differences of nonparametric estimators with an application to the sharp regression discontinuity design’.” Report. [451]

——— (2016), “Optimal bandwidth selection for the fuzzy regression discontinuity estimator.” *Economics Letters*, 141, 103–106. [473]

Calonico, S., M. D. Cattaneo, and M. H. Farrell (forthcoming), "On the effect of bias estimation on coverage accuracy in nonparametric inference." *Journal of the American Statistical Association*. [458, 469]

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014a), "Robust nonparametric bias-corrected inference in the regression discontinuity design." *Econometrica*, 82, 2295–2326. [441, 442]

——— (2014b), "Supplement to 'Robust nonparametric confidence intervals for regression-discontinuity design'." *Econometrica Supplemental Material*, 82, http://dx.doi.org/10.3982/ECTA11757. [461, 468]

——— (2015), "rdrobust: An R package for robust inference in regression discontinuity design." *R Journal*, 7, 38–51. [461]

Card, D., D. S. Lee, Z. Pei, and A. Weber (2015), "Inference on causal effects in a generalized regression kink design." *Econometrica*, 83, 2453–2483. [442, 444]

Card, D., A. Mas, and J. Rothstein (2008), "Tipping and the dynamics of segregation." *Quarterly Journal of Economics*, 123, 177–218. [442]

Cattaneo, M. D., R. Titiunik, and G. Vazquez-Bare (2017), "Comparing inference approaches in RD designs: A reexamination of the effect of Head Start on Child Martality." *Journal of Policy Analysis and Management*, 36, 643–681. [469, 471, 472]

Cheng, M. Y., J. Fan, and J. S. Marron (1997), "On automatic boundary corrections." *The Annals of Statistics*, 25, 1691–1708. [445]

Chiang, H. D. and Y. Sasaki (2016), "Causal inference by quantile regression kink designs." Johns Hopkins University. [442]

DesJardins, S. L. and B. P. McCall (2008), "The impact of the Gates Millennium scholars program on the retention, college finance- and work-related choices, and future educational aspirations of low-income minority students." Report. [442]

DiNardo, J. and D. S. Lee (2011), "Program evaluation and research designs." In *Handbook of Labor Economics*, Vol. 4A (O. Ashenfelter and D. Card, eds.), 463–536, Elsevier, Amsterdam. [442]

Dong, Y. and A. Lewbel (2015), "Identifying the effect of changing the policy threshold in regression discontinuity models." *Review of Economics and Statistics*, 97, 1081–1092. [442]

Fan, J. (1992), "Design-adaptive nonparametric regression." *Journal of the American Statistical Association*, 87, 998–1004. [442, 445, 450]

——— (1993), "Local linear regression smoothers and their minimax efficiencies." *The Annals of Statistics*, 21, 196–216. [442, 445]

Fan, J. and I. Gijbels (1992), "Variable bandwidth and local linear regression smoothers." *The Annals of Statistics*, 20, 2008–2036. [448]

——— (1996), *Local Polynomial Modeling and Its Applications*. Chapman & Hall, Boca Raton, FL. [451, 456]

Fan, J., I. Gijbels, T.-C. Hu, and L.-S. Huang (1996), "A study of variable bandwidth selection for local polynomial regression." *Statistica Sinica*, 6, 113–127. [448]

Frandsen, B. R., M. Frörich, and B. Melly (2012), "Quantile treatment effects in the regression discontinuity design." *Journal of Econometrics*, 168, 382–395. [442]

Hahn, J., P. Todd, and W. Van der Klaauw (2001), "Identification and estimation of treatment effects with a regression-discontinuity design." *Econometrica*, 69, 201–209. [441, 442, 445]

Hinnerich, B. T. and P. Pettersson-Lidbom (2014), "Democracy, redistribution, and political participation: Evidence from Sweden 1919–1938." *Econometrica*, 82, 961–993. [442]

Ichimura, H. and P. E. Todd (2007), "Implementing nonparametric and semiparametric estimators." In *Handbook of Econometrics*, Vol. 6 (J. J. Heckman and E. E. Leamer, eds.), Chapter 74, 5369–5468, Elsevier, Amsterdam. [442]

Imbens, G. W. and K. Kalyanaraman (2009), "Optimal bandwidth choice for the regression discontinuity estimator." IZA Discussion Paper 3995. [461]

——— (2012), "Optimal bandwidth choice for the regression discontinuity estimator." *Review of Economic Studies*, 79, 933–959. [442]

Imbens, G. W. and T. Lemieux (2008), "Regression discontinuity designs: A guide to practice." *Journal of Econometrics*, 142, 615–635. [442]

Lee, D. S. (2008), "Randomized experiments from non-random selection in U.S. house elections." *Journal of Econometrics*, 142, 675–697. [442, 461]

Lee, D. S. and T. Lemieux (2010), "Regression discontinuity designs in economics." *Journal of Economic Literature*, 48, 281–355. [442]

Ludwig, J. and D. L. Miller (2005), "Does Head Start improve children's life changes? Evidence from a regression discontinuity design." NBER Working Paper 11702. [443]

——— (2007), "Does Head Start improve children's life changes? Evidence from a regression discontinuity design." *Quarterly Journal of Economics*, 122, 159–208. [441, 442, 443, 461, 462, 469, 470, 471]

Nielsen, H. S., T. Sørensen, and C. Taber (2010), "Estimating the effect of student aid on college enrollment: Evidence from a government grant policy reform." *American Economic Journal: Economic Policy*, 2, 185–215. [444]

Porter, J. (2003), "Estimation in the regression discontinuity model." Report. [442, 445, 469]

Seifert, B. and T. Gasser (1996), "Finite-sample variance of local polynomials: Analysis and solutions." *Journal of the American Statistical Association*, 91, 267–275. [462]

Stone, C. J. (1977), "Consistent nonparametric regression." *The Annals of Statistics*, 5, 595–645. [445]

Thistlewaite, D. and D. Campbell (1960), "Regression-discontinuity analysis: An alternative to the ex post facto experiment." *Journal of Educational Psychology*, 51, 309–317. [441]

Van der Klaauw, W. (2008), "Regression-discontinuity analysis: A survey of recent developments in economics." *Labour*, 22, 219–245. [442]

Co-editor Rosa L. Matzkin handled this manuscript.

Manuscript received 30 June, 2015; final version accepted 29 April, 2017; available online 21 June, 2017.