CIRJE-F-889

Optimal Bandwidth Selection for Differences of Nonparametric Estimators with an Application to the Sharp Regression Discontinuity Design

Yoichi Arai
National Graduate Institute for Policy Studies (GRIPS)

Hidehiko Ichimura
University of Tokyo

June 2013

Optimal Bandwidth Selection for Differences of Nonparametric Estimators with an Application to the Sharp Regression Discontinuity Design*

Yoichi Arai† and Hidehiko Ichimura‡

Abstract

We consider the problem of choosing two bandwidths simultaneously for estimating the difference of two functions at given points. When the asymptotic approximation of the mean squared error (AMSE) criterion is used, we show that minimization problem is not well-defined when the sign of the product of the second derivatives of the underlying functions at the estimated points is positive. To address this problem, we theoretically define and construct estimators of the asymptotically first-order optimal (AFO) bandwidths which are well-defined regardless of the sign. They are based on objective functions which incorporate a second-order bias term. Our approach is general enough to cover estimation problems related to densities and regression functions at interior and boundary points. We provide a detailed treatment of the sharp regression discontinuity design.

Key words: Bandwidth selection, kernel density estimation, local linear regression, regression discontinuity design

*Earlier versions of this paper were presented at the Japanese Economic Association Spring Meeting, the North American Winter Meeting of the Econometric Society, LSE, UC Berkeley and Yale. Valuable comments were received from seminar participants. We are especially grateful to Yoshihiko Nishiyama, Jack Porter and Jim Powell for many helpful comments. This research was supported by Grants-in-Aid for Scientific Research No. 22243020 and No. 23330070 from the Japan Society for the Promotion of Science.

†National Graduate Institute for Policy Studies (GRIPS), 7-22-1 Roppongi, Minato-ku, Tokyo 106-8677, Japan; yarai@grips.ac.jp

‡Department of Economics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan; ichimura@e.u-tokyo.ac.jp

1 Introduction

Given a particular nonparametric estimator, it is well recognized that choosing an appropriate smoothing parameter is a key implementation issue about which various methods have been proposed. Among myriad developments in nonparametric estimation methods, those in program evaluation highlight the need to estimate the difference of two functions at particular points rather than an unknown function itself. Examples include applications of the average treatment effect (ATE), the local average treatment effect (LATE), and the regression discontinuity design (RDD).

The standard approach in empirical researches is to estimate two functions by kernel-type nonparametric estimators. Two bandwidths are required to estimate two functions and are selected independently by using the plug-in or the cross-validation method proposed to estimate a single function. For example, Ludwig and Miller (2005, 2007) and DesJardins and McCall (2008) used the cross-validation and the plug-in method, respectively, in the context of the sharp RDD. One notable exception is the bandwidth selection procedure proposed by Imbens and Kalyanaraman (2012) (hereafter IK) developed for the RDD estimator to choose the same bandwidth to estimate two functions on both sides of a discontinuity point. The bandwidth proposed by IK is obtained by minimizing the asymptotic approximation of the mean squared error (AMSE) with regularization.

In this paper, we propose to choose two bandwidths simultaneously to estimate the difference of two functions based on minimizing a version of the AMSE. Empirical studies using the RDD estimators by DesJardins and McCall (2008), Lee (2008) and Ludwig and Miller (2005, 2007) among others reveal that the curvatures on the right- and left-side of the threshold often differ. Since we should allow this possibility in general, it is natural to choose two bandwidths simultaneously for both sides of the threshold. Although a simultaneous choice of two bandwidths seems natural, it has not yet been considered in either the econometrics or the statistics literature. We show that this natural approach leads to a nonstandard problem. To illustrate the main issue of the problem, we consider estimating the difference of densities evaluated at two

distinct points by kernel density estimator with a second-order kernel function because density estimation problems are the simplest, but have all the essential features that we explore.

We show that when the sign of the product of the second derivatives of the density functions at two distinct points is negative, the bandwidths that minimize the AMSE are well-defined. But when the sign of the product is positive, the trade-off between bias and variance, which is a key aspect of optimal bandwidth selection, breaks down, and the AMSE can be made arbitrarily small without increasing the bias component. This happens because there exists a specific ratio of bandwidths that can remove the bias term completely, and we can make the variance arbitrarily small by choosing large values of the bandwidths keeping the ratio constant.

To address this problem, we theoretically define asymptotically first-order optimal (AFO) bandwidths based on objective functions which incorporates a second-order bias term. The AFO bandwidths are defined as the minimizer of the standard AMSE when the sign is negative while they are the minimizer of the AMSE with a second-order bias term subject to the restriction that the first-order bias term is equal to zero when the sign is positive. We construct an estimator which is shown to be asymptotically equivalent to the AFO bandwidths.

We investigate the problems of nonparametric estimation of the difference of regression functions at interior and boundary points. The nonparametric regression estimators we consider are LLR estimators proposed by Stone (1977) and investigated by Fan (1992). An important application of the boundary cases is the sharp RDD. We show that the essential features of the problems are exactly the same as those for the estimation problem of the difference of densities and the results are generalized to cover these cases.

We conducted a simulation study to investigate the finite sample properties of the proposed method. We concentrated on the case of the sharp RDD, which is most empirically relevant. Our experiment showed that the proposed method performs well for all six designs considered in the paper and particularly well for designs in which there exists a large difference in the absolute magnitudes of the second derivatives.

More specifically, the proposed bandwidths are more stable and perform better than existing bandwidths in terms of the root mean squared error.

The remainder of the paper is organized as follows. In Section 2, all the essential features of our approach are presented through the estimation problem of the difference of densities at given points. We generalize the proposed method to the estimation problem of regression functions at interior and boundary points with emphasis on the sharp RDD in Section 3. In Section 4, we demonstrate the finite sample behavior of our approach via a simulation study. Section 5 concludes. Omitted discussions, an algorithm to implement the proposed method for the sharp RDD and all proofs for main results are provided in the supplemental material (Arai and Ichimura, 2013).

2 Nonparametric Estimation of Differences of Densities

2.1 The AMSE for Differences of Kernel Density Estimators

We consider estimating a difference of a density function at two given points, i.e., $f(x_1) - f(x_2)$, for $x_1 \neq x_2$, where f is a Lebesgue density.[1] Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from a univariate distribution with the Lebesgue density f. Then, $f(x_1) - f(x_2)$ is estimated by $\hat{f}_{h_1}(x_1) - \hat{f}_{h_2}(x_2)$, where $\hat{f}_{h_j}(x_j)$ is the kernel density estimator of f given by $\hat{f}_{h_j}(x_j) = \sum_{i=1}^{n} K\left((x_j - X_i)/h_j\right)/(nh_j)$, where K is a kernel function, and h_j is a bandwidth used to estimate the density f at x_j for $j = 1, 2$. For simplicity we use the same kernel function K to estimate both $\hat{f}_{h_1}(x_1)$ and $\hat{f}_{h_2}(x_2)$.

In this paper, we propose a simultaneous selection method of two distinct bandwidths based on an approximate MSE in a broad sense. In the standard context of kernel density estimation, numerous methods have been proposed to choose

[1] Throughout this section, we consider the difference of kernel density estimators for a "single" density at two distinct points. A straightforward generalization shows that we can apply the discussions in this section to bandwidth choices for the difference of kernel density estimators of two distinct densities, f and g, at two points, x and y; i.e., $f(x) - g(y)$ based on the two random samples $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_n\}$.

a bandwidth. One of the most popular and frequently used methods is to choose a bandwidth based on the AMSE.[2] The MSE for the difference of the two density estimators is defined by

$$MSE_n(h) = E\left\{\left[\left(\hat{f}_{h_1}(x_1) - \hat{f}_{h_2}(x_2)\right) - \left(f(x_1) - f(x_2)\right)\right]^2\right\},$$

where the expectation is taken using f as the density for the observations.[3] A standard approach is to obtain the AMSE, ignoring higher-order terms, and to choose the bandwidths that minimize that. To do so, we make the following assumptions. (The integral sign $\int$ refers to an integral over the range $(-\infty, \infty)$ unless stated otherwise.)

ASSUMPTION 1 *$K(\cdot) : \mathbb{R} \to \mathbb{R}$ is a symmetric second-order kernel function that is continuous with compact support; i.e., K satisfies the following: $\int K(u)du = 1$, $\int uK(u)du = 0$, and $\int u^2K(u)du \neq 0$.*

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

$$\left|f^{(k)}(x) - f^{(k)}(y)\right| \leq M_k \left|x - y\right|^\alpha, \quad \varepsilon < f(z) < M, \quad x, y, z \in \mathcal{D},$$

for some positive M_k, ε and M such that $0 < \varepsilon < M < \infty$ and some α such that $0 < \alpha \leq 1$.

ASSUMPTION 2 *The density f is an element of $\mathcal{F}_2(\mathcal{D}_j)$ where $\mathcal{D}_j$ is an open neighborhood of x_j for $j = 1, 2$.*

ASSUMPTION 3 *The positive sequence of bandwidths is such that $h_j \to 0$ and $nh_j \to \infty$ as $n \to \infty$ for $j = 1, 2$.*

[2] As IK emphasize, the bandwidth selection problem in the context of the RDD as well as the other problems considered in this paper are 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.

[3] Throughout the paper, we use "h" without a subscript to denote a combination of h_1 and h_2; e.g., $MSE_n(h_1, h_2)$ is written as $MSE_n(h)$.

Assumptions 1, 2 and 3 are standard in the literature of kernel density estimation. Under Assumptions 1, 2 and 3, standard calculation yields

$$MSE_n(h) = \left\{ \frac{\mu_2}{2} \left[f^{(2)}(x_1) h_1^2 - f^{(2)}(x_2) h_2^2 \right] \right\}^2 + \frac{\nu_0}{n} \left\{ \frac{f(x_1)}{h_1} + \frac{f(x_2)}{h_2} \right\} + o\left(h_1^4 + h_1^2 h_2^2 + h_2^4 + \frac{1}{nh_1} + \frac{1}{nh_2} \right),$$

where $\mu_j = \int u^j K(u) du$ and $\nu_j = \int u^j K^2(u) du$ (see, e.g., Prakasa Rao, 1983, Section 2.1). This suggests that we choose the bandwidths to minimize the following AMSE:

$$AMSE_n(h) = \left\{ \frac{\mu_2}{2} \left[f^{(2)}(x_1) h_1^2 - f^{(2)}(x_2) h_2^2 \right] \right\}^2 + \frac{\nu_0}{n} \left\{ \frac{f(x_1)}{h_1} + \frac{f(x_2)}{h_2} \right\}. \tag{1}$$

However, this procedure may fail. To see why, let h_1, $h_2 \in H$, where $H = (0, \infty)$, and consider the case in which $f^{(2)}(x_1) f^{(2)}(x_2) > 0$. Now choose $h_2 = [f^{(2)}(x_1)/f^{(2)}(x_2)]^{1/2} h_1$. Then, we have

$$AMSE_n(h) = \frac{\nu_0}{nh_1} \left\{ f(x_1) + f(x_2) \left[\frac{f^{(2)}(x_2)}{f^{(2)}(x_1)} \right]^{1/2} \right\}.$$

This implies that the bias component can be removed completely from the AMSE by choosing a specific ratio of bandwidths and the AMSE can be made arbitrarily small by choosing a sufficiently large h_1.

One reason for this nonstandard behavior is that the AMSE given in (1) does not account for higher-order terms. If non-removable higher-order terms for the bias component are present, they should punish the act of choosing large values for bandwidths. In what follows, we incorporate a second-order bias term into the AMSE assuming densities are smooth.

ASSUMPTION 4 *The density f is an element of $\mathcal{F}_4(\mathcal{D}_j)$ where $\mathcal{D}_j$ is an open neighborhood of x_j for $j = 1, 2$.*

In the literature of kernel density estimation, it is common to employ higher-order kernel functions when the density is four times differentiable because it is known to reduce bias (see, e.g., Silverman, 1986, Section 3.6). However, we have several

reasons for confining our attention to the second-order kernel functions. First, as shown later, we can achieve the same bias reduction without employing higher-order kernel functions when the sign of the product of the second derivatives is positive. When the sign is negative, Assumption 4 is unnecessary. Second, even when we use a higher-order kernel functions, we end up with an analogous problem. For example, the first-order bias term is removed by using higher-order kernel functions, but when the signs of the fourth derivatives are the same, the second-order bias term can be eliminated by using an appropriate choice of bandwidths.

The next lemma shows the asymptotic property of the MSE under the smoothness condition. This straightforward extension of the standard result (see, e.g., Prakasa Rao, 1983, Section 2.1) is presented without proof.

LEMMA 1 *Suppose Assumptions 1, 3 and 4 hold. Then, it follows that*

$$MSE_n(h) = \left\{ \frac{\mu_2}{2} \left[f^{(2)}(x_1) h_1^2 - f^{(2)}(x_2) h_2^2 \right] + \frac{\mu_4}{4!} \left[f^{(4)}(x_1) h_1^4 - f^{(4)}(x_2) h_2^4 \right] + o\left(h_1^4 + h_2^4 \right) \right\}^2 + \frac{\nu_0}{n} \left\{ \frac{f(x_1)}{h_1} + \frac{f(x_2)}{h_2} \right\} + o\left(\frac{1}{nh_1} + \frac{1}{nh_2} \right). \tag{2}$$

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

$$\left\{ \frac{\mu_2}{2} \left[f^{(2)}(x_1) h_1^2 - f^{(2)}(x_2) h_2^2 \right] + \frac{\mu_4}{4!} \left[f^{(4)}(x_1) h_1^4 - f^{(4)}(x_2) h_2^4 \right] \right\}^2 + \frac{\nu_0}{n} \left\{ \frac{f(x_1)}{h_1} + \frac{f(x_2)}{h_2} \right\}. \tag{3}$$

We show that a straightforward minimization of this approximate MSE does not overcome the problem discussed earlier. That is, the minimization problem is not well-defined when $f^{(2)}(x_1) f^{(2)}(x_2) > 0$. In particular, we show that one can make the order of the bias term $O(h_1^{2k})$, with k being an arbitrary positive integer, by choosing $h_2^2 = C(h_1, k) h_1^2$ and $C(h_1, k) = C_0 + C_1 h_1^2 + C_2 h_1^4 + C_3 h_1^6 + \ldots + C_k h_1^{2k}$ for some constants C_0, C_1, ..., C_k when the sign of the product of the second derivatives is positive. Given that bandwidths are necessarily positive, we must have $C_0 > 0$, although we allow C_1, C_2, ..., C_k to be negative.

To gain insight, consider choosing $C(h_1, 1) = C_0 + C_1 h_1^2$, where $C_0 = f^{(2)}(x_1)/f^{(2)}(x_2)$.

In this case, the sum of the first- and second-order bias terms is

$$\frac{\mu_2}{2}\left[f^{(2)}(x_1) - C(h_1,1)f^{(2)}(x_2)\right]h_1^2 + \frac{\mu_4}{4!}\left[f^{(4)}(x_1) - C(h_1,1)^2 f^{(4)}(x_2)\right]h_1^4$$
$$= \left\{-\frac{\mu_2}{2}C_1 f^{(2)}(x_2) + \frac{\mu_4}{4!}\left[f^{(4)}(x_1) - C_0^2 f^{(4)}(x_2)\right]\right\}h_1^4 + O(h_1^6).$$

By choosing $C_1 = \mu_4\left[f^{(4)}(x_1) - C_0^2 f^{(4)}(x_2)\right] / \left[12\mu_2 f^{(2)}(x_2)\right]$, one can make the order of bias $O(h_1^6)$. Next, consider $C(h_1,2) = C_0 + C_1 h_1^2 + C_2 h_1^4$, where C_0 and C_1 are as determined above. In this case,

$$\frac{\mu_2}{2}\left[f^{(2)}(x_1) - C(h_1,2)f^{(2)}(x_2)\right]h_1^2 + \frac{\mu_4}{4!}\left[f^{(4)}(x_1) - C(h_1,2)^2 f^{(4)}(x_2)\right]h_1^4$$
$$= -\left\{\frac{\mu_2}{2}C_2 f^{(2)}(x_2) + \frac{\mu_4}{12}C_0 C_1 f^{(4)}(x_2)\right\}h_1^6 + O(h_1^8).$$

Hence, by choosing $C_2 = -\mu_4 C_0 C_1 f^{(4)}(x_2)/[6\mu_2 f^{(2)}(x_2)]$, one can make the order of bias term $O(h_1^8)$. Similar arguments can be formulated for arbitrary k and the resulting approximate MSE is given by

$$\frac{\nu_0}{nh_1}\left\{f(x_1) + f(x_2)\left[\frac{f^{(2)}(x_2)}{f^{(2)}(x_1)}\right]^{1/2}\right\} + O\left(h_1^{2k}\right).$$

The discussion above is summarized in the following lemma.

LEMMA 2 *Suppose Assumptions 1, 3 and 4 hold. Then there exist a combination of h_1 and h_2 such that the approximate MSE including the second-order bias term defined in (3) becomes*

$$\frac{\nu_0}{nh_1}\left\{f(x_1) + f(x_2)\left[\frac{f^{(2)}(x_2)}{f^{(2)}(x_1)}\right]^{1/2}\right\} + O\left(h_1^{2k}\right).$$

for an arbitrary positive integer k.

This implies that one can make the approximate MSE arbitrarily small by appropriate choices of h_1 and k, leading to non-existence of the optimal solution. It is straightforward to generalize this discussion to the case of the AMSE with higher-order bias terms.

2.2 AFO Bandwidths

We observed that the optimal bandwidths that minimize the AMSE are not well-defined when the sign of the product of the second derivatives is positive. We also discovered that simply introducing higher-order bias terms does not help to avoid disappearance of the trade-off. Hence, we propose a new optimality criterion termed "asymptotic first-order optimality".

First, we discuss the case in which $f^{(2)}(x_1)f^{(2)}(x_2) < 0$. Remember that the standard AMSE is given by equation (1). In this situation, the square of the first-order bias term cannot be removed by any choice of the bandwidths and dominates the second-order bias term asymptotically. This implies that there is a bias-variance trade-off. Hence, it is reasonable to choose the bandwidths that minimize the AMSE given in (1). This case will turn out to be similar to the existing bandwidth selection methods considered by DesJardins and McCall (2008) and Imbens and Kalyanaraman (2012) in the sense that the order of the bandwidths is $n^{-1/5}$, although they differ from the bandwidths considered here by constant multiples, reflecting the simultaneous selection of two bandwidths.

When $f^{(2)}(x_1)f^{(2)}(x_2) > 0$, by choosing $h_2^2 = C_0 h_1^2$ with $C_0 = f^{(2)}(x_1)/f^{(2)}(x_2)$, the bias component with the second-order term becomes

$$\left\{ \frac{\mu_4}{4!} \left[f^{(4)}(x_1) - C_0^2 f^{(4)}(x_2) \right] \right\} h_1^4 + o\left(h_1^4\right).$$

unless $f^{(2)}(x_2)^2 f^{(4)}(x_1) = f^{(2)}(x_1)^2 f^{(4)}(x_2)$. With this bias component, there exists a bias-variance trade-off and the bandwidths can be determined. The above discussion is formalized in the following definition and the resulting bandwidths are termed "AFO bandwidths."

DEFINITION 1 *The AFO bandwidths for the difference of densities* minimize the AMSE defined by

$$AMSE_{1n}(h) = \left\{ \frac{\mu_2}{2} \left[f^{(2)}(x_1)h_1^2 - f^{(2)}(x_2)h_2^2 \right] \right\}^2 + \frac{\nu_0}{n} \left\{ \frac{f(x_1)}{h_1} + \frac{f(x_2)}{h_2} \right\} \qquad (4)$$

when $f^{(2)}(x_1)f^{(2)}(x_2) < 0$, and their explicit expressions are given by $h_1^* = \theta^* n^{-1/5}$ and $h_2^* = \lambda^* h_1^*$, where

$$\theta^* = \left\{ \frac{\nu_0 f(x_1)}{\mu_2^2 f^{(2)}(x_1)\left[f^{(2)}(x_1) - \lambda^{*2} f^{(2)}(x_2)\right]} \right\}^{1/5} \quad \text{and} \quad \lambda^* = \left\{ -\frac{f(x_2)f^{(2)}(x_1)}{f(x_1)f^{(2)}(x_2)} \right\}^{1/3}.$$

When $f^{(2)}(x_1)f^{(2)}(x_2) > 0$, *the AFO bandwidths for the difference of densities* minimizes the AMSE defined by

$$AMSE_{2n}(h) = \left\{ \frac{\mu_4}{4!}\left[f^{(4)}(x_1)h_1^4 - f^{(4)}(x_2)h_2^4\right] \right\}^2 + \frac{\nu_0}{n}\left\{ \frac{f(x_1)}{h_1} + \frac{f(x_2)}{h_2} \right\} \qquad (5)$$

subject to the restriction $f^{(2)}(x_1)h_1^2 - f^{(2)}(x_2)h_2^2 = 0$ under the assumption of $f^{(2)}(x_2)^2 f^{(4)}(x_1) \neq f^{(2)}(x_1)^2 f^{(4)}(x_2)$, and their explicit expressions are given by $h_1^{**} = \theta^{**} n^{-1/9}$ and $h_2^{**} = \lambda^{**} h_1^{**}$, where

$$\theta^{**} = \left\{ \frac{72\nu_0[f(x_1) + f(x_2)/\lambda^{**}]}{\mu_4^2\left[f^{(4)}(x_1) - \lambda^{**4} f^{(4)}(x_2)\right]^2} \right\}^{1/9} \quad \text{and} \quad \lambda^{**} = \left\{ \frac{f^{(2)}(x_1)}{f^{(2)}(x_2)} \right\}^{1/2}.$$

Definition 1 is stated with assuming that the first- and the second-order bias terms do not vanish simultaneously, i.e., $f^{(2)}(x_2)^2 f^{(4)}(x_1) \neq f^{(2)}(x_1)^2 f^{(4)}(x_2)$.[4] This type of assumption is made for the optimal bandwidth selection for the standard kernel density estimation at a point; namely $f^{(2)}(x) \neq 0$.[5]

The proposed bandwidths are called the AFO bandwidths because the $AMSE_{2n}(h)$ is minimized under the restriction that the first-order bias term is removed when the sign is positive. It is worth noting that the order of the optimal bandwidths exhibits

[4]Uniqueness of the AFO bandwidths in each case is verified in Arai and Ichimura (2013).

[5]Definition 1 can be generalized to cover the excluded case in a straightforward manner if we are willing to assume the existence of the sixth derivative of f and if $f^{(4)}(x_2)^3 f^{(6)}(x_1)^2 \neq f^{(4)}(x_1)^3 f^{(6)}(x_2)^2$. This case corresponds to the situation in which the first- and the second-order bias terms can be removed simultaneously by choosing appropriate bandwidths and the third-order bias term works as a penalty for large bandwidths. When f is continuously differentiable an infinite number of times, the excluded case becomes $f^{(2j)}(x_2)^{j+1} f^{(2(j+1))}(x_1)^j = f^{(2j)}(x_1)^{j+1} f^{(2(j+1))}(x_2)^j$ for all integers j. Another excluded case by Definition 1 is when $f^{(2)}(x_1)f^{(2)}(x_2) = 0$. However, it is possible to extend the idea of the AFO bandwidths when both $f^{(2)}(x_1) = 0$ and $f^{(2)}(x_2) = 0$ hold and when the fourth and the sixth derivatives satisfy certain assumptions. This generalization corresponds to that in Definition 1 (i) and (ii) with $f^{(2)}(x_1)$, $f^{(2)}(x_2)$, $f^{(4)}(x_1)$, $f^{(4)}(x_2)$ and other parameters being replaced by $f^{(4)}(x_1)$, $f^{(4)}(x_2)$, $f^{(6)}(x_1)$, $f^{(6)}(x_2)$ and corresponding parameters.

dichotomous behavior depending on the sign of the product of the second derivatives. Let h^* and h^{**} be (h_1^*, h_2^*) and (h_1^{**}, h_2^{**}), respectively. It is easily seen that the orders of $AMSE_{1n}(h^*)$ and $AMSE_{2n}(h^{**})$ are $O_p(n^{-4/5})$ and $O_p(n^{-8/9})$, respectively. This implies that, when the sign is positive, the AFO bandwidths reduce bias without increasing variance and explains why we need not use higher-order kernel functions even when the fourth derivative of $f(\cdot)$ exists.[6]

We provide a discussion on relationships between the AFO bandwidths and other potential bandwidths. First, as we saw, the bandwidths that minimize the AMSE given in equation (1) become rate-optimal under Assumption 2 when the sign is negative but the minimization problem is not well-defined when the sign is positive.

Second, the bandwidths based on a fourth-order kernel function suffer from the same issue. When the sign of the product of the fourth derivatives is negative, the bandwidths are well-defined and become rate-optimal under Assumption 4. But the minimization problem is not well-defined when the sign is positive.

Third, when we minimize the AMSE given in equation (1) under the restriction that two bandwidths are the same, the bandwidth is well-defined irrespective of the sign of the second derivatives under Assumption 2. However, when the sign of the product of the second derivatives is negative, the restriction is unnecessary. When the sign is positive, the restriction works to determine a bandwidth under Assumption 2 although there is no particular reason for imposing the restriction. Under Assumption 4, it is not rate-optimal.

In contrast , when the sign of the product of the second derivatives is negative, the AFO bandwidths are well-defined and become rate-optimal under Assumption 2. When the sign is positive, the AFO bandwidths become rate-optimal under Assumption 4, achieving the same bias reduction as the approach with a fourth order kernel function does.

Next, we show that the asymptotically higher-order optimal bandwidths can be proposed under a sufficient smoothness condition. To be concise, we only discuss the asymptotically second-order optimal (ASO) bandwidths when $f^{(2)}(x_1)f^{(2)}(x_2) > 0$ un-

[6]The advantages of not using higher-order kernel functions also lies in that one need not worry about having negative values for density estimates.

der the assumption that f is six times continuously differentiable in the neighborhood of x_j with $f(x_j) > 0$ for $j = 1, 2$.

Consider choosing $C(h_1, 1) = C_0 + C_1 h_1^2$, where $C_0 = f^{(2)}(x_1)/f^{(2)}(x_2)$. In this case, the bias component is

$$\begin{aligned}
&\frac{\mu_2}{2}\left[f^{(2)}(x_1) - C(h_1,1)f^{(2)}(x_2)\right]h_1^2 + \frac{\mu_4}{4!}\left[f^{(4)}(x_1) - C(h_1,1)^2 f^{(4)}(x_2)\right]h_1^4 \\
&\quad + \frac{\mu_6}{6!}\left[f^{(6)}(x_1) - C(h_1,1)^3 f^{(6)}(x_2)\right]h_1^6 + o\left(h_1^6\right) \\
&= \left\{-\frac{\mu_2}{2}C_1 f^{(2)}(x_2) + \frac{\mu_4}{4!}\left[f^{(4)}(x_1) - C_0^2 f^{(4)}(x_2)\right]\right\}h_1^4 \\
&\quad + \left\{\frac{\mu_6}{6!}\left[f^{(6)}(x_1) - C_0^3 f^{(6)}(x_2)\right] - \frac{\mu_6}{12}C_0 C_1\right\}h_1^6 + o\left(h_1^6\right)
\end{aligned}$$

where the equality follows by the definition of C_0. By choosing

$$C_1 = \mu_4\left[f^{(4)}(x_1) - C_0^2 f^{(4)}(x_2)\right] / \left[12\mu_2 f^{(2)}(x_2)\right],$$

one can make the order of bias component $O(h_1^6)$. The ASO bandwidths h_1^{**} can be determined by minimizing the following AMSE

$$AMSE_{3n}(h) = \left\{\frac{\mu_6}{6!}\left[f^{(6)}(x_1) - C_0^3 f^{(6)}(x_2)\right] - \frac{\mu_6}{12}C_0 C_1\right\}^2 h_1^6 + \frac{\nu_0}{nh_1}\left[f(x_1) + \frac{f(x_2)}{C_0^{1/2}}\right]$$

and h_2^{**} can be obtained by the relationship $h_2^{**2} = \left(C_0 + C_1 h_1^{**2}\right)h_1^{**2}$ when $f^{(2)}(x_2)^2 f^{(4)}(x_1) \neq f^{(2)}(x_1)^2 f^{(4)}(x_2)$, $(\mu_6/6!)\left[f^{(6)}(x_1) - C_0^3 f^{(6)}(x_2)\right] \neq (\mu_6/12)C_0 C_1$ and $C_0 + C_1 h_1^{**2} > 0$. The ASO bandwidths are of order $n^{-1/13}$. A potential drawback of the ASO bandwidths is that they are not well-defined when $C_0 + C_1 h_1^{**2} \leq 0$. Similar arguments can be formulated for arbitrary k with a sufficient smoothness condition. This implies that one can make the bias component arbitrarily small by choosing h_1 and k.

If one believes that the underlying function is very smooth (say, six times continuously differentiable), it would be reasonable to consider the ASO bandwidths. However, we typically avoid imposing strong assumptions on the density because the true smoothness is almost always unknown. In addition, the following discussion shows that implementing the ASO bandwidths require the estimation of the sixth

derivatives, which is very challenging in practice. Thus we concentrate on the AFO bandwidths in this paper.

2.3 Feasible Automatic Bandwidth Choice

The AFO bandwidths are clearly not feasible because they depend on unknown quantities such as $f(\cdot)$, $f^{(2)}(\cdot)$, $f^{(4)}(\cdot)$ and, most importantly, on the sign of the product of the second derivatives.

An obvious plug-in version of the AFO bandwidths can be implemented by estimating the second derivatives, $\hat{f}^{(2)}(x_1)$ and $\hat{f}^{(2)}(x_2)$. Depending on the estimated sign of the product, we can construct the plug-in version of the AFO bandwidths provided in Definition 1. We refer to these as "the direct plug-in AFO bandwidths". They are defined by

$$\hat{h}_1^D = \hat{\theta}_1 n^{-1/5}\mathbb{I}\{\hat{f}^{(2)}(x_1)\hat{f}^{(2)}(x_2) < 0\} + \hat{\theta}_2 n^{-1/9}\mathbb{I}\{\hat{f}^{(2)}(x_1)\hat{f}^{(2)}(x_2) \geq 0\},$$
$$\hat{h}_2^D = \hat{\theta}_1\hat{\lambda}_1 n^{-1/5}\mathbb{I}\{\hat{f}^{(2)}(x_1)\hat{f}^{(2)}(x_2) < 0\} + \hat{\theta}_2\hat{\lambda}_2 n^{-1/9}\mathbb{I}\{\hat{f}^{(2)}(x_1)\hat{f}^{(2)}(x_2) \geq 0\},$$

where $\mathbb{I}$ denotes the indicator function,

$$\hat{\theta}_1 = \left\{ \frac{\nu_0 \hat{f}(x_1)}{\mu_2^2 \hat{f}^{(2)}(x_1)\left[\hat{f}^{(2)}(x_1) - \hat{\lambda}_1^2 \hat{f}^{(2)}(x_2)\right]} \right\}^{1/5}, \quad \hat{\lambda}_1 = \left[-\frac{\hat{f}(x_2)\hat{f}^{(2)}(x_1)}{\hat{f}(x_1)\hat{f}^{(2)}(x_2)} \right]^{1/3}, \tag{6}$$

$$\hat{\theta}_2 = \left\{ \frac{72\nu_0[\hat{f}(x_1) + \hat{f}(x_2)/\hat{\lambda}_2]}{\mu_4^2\left[\hat{f}^{(4)}(x_1) - \hat{\lambda}_2^4 \hat{f}^{(4)}(x_2)\right]^2} \right\}^{1/9}, \quad \text{and} \quad \hat{\lambda}_2 = \left[\frac{\hat{f}^{(2)}(x_1)}{\hat{f}^{(2)}(x_2)} \right]^{1/2}. \tag{7}$$

These bandwidths switch depending on the estimated sign. We can show that the direct plug-in AFO bandwidths are asymptotically as good as the AFO bandwidths in large samples. That is, we can prove that a version of Theorem 1 below also holds for the direct plug-in AFO bandwidths. However, our unreported simulation experiments show a poor performance of the direct plug-in AFO bandwidths under the designs described in Section 4 since they misjudge the rate of the bandwidths whenever the sign is misjudged. Hence we do not pursue the direct plug-in approach

further.

Instead, we propose an alternative procedure for choosing bandwidths that switch between two bandwidths more smoothly. To propose feasible bandwidths, we present a modified version of the AMSE (MMSE) defined by

$$MMSE_n(h) = \left\{ \frac{\mu_2}{2} \left[f^{(2)}(x_1) h_1^2 - f^{(2)}(x_2) h_2^2 \right] \right\}^2 + \left\{ \frac{\mu_4}{4!} \left[f^{(4)}(x_1) h_1^4 - f^{(4)}(x_2) h_2^4 \right] \right\}^2 + \frac{\nu_0}{n} \left\{ \frac{f(x_1)}{h_1} + \frac{f(x_2)}{h_2} \right\}.$$

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. A key characteristic of the MMSE is that its bias component cannot be made arbitrarily small by any choices of bandwidths even when the sign is positive, unless $f^{(2)}(x_2)^2 f^{(4)}(x_1) = f^{(2)}(x_1)^2 f^{(4)}(x_2)$. Thus, either term can penalize large bandwidths regardless of the sign, in which case, the MMSE preserves the bias-variance trade-off. More precisely, when $f^{(2)}(x_1) f^{(2)}(x_2) < 0$, 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. On the other hand, when $f^{(2)}(x_1) f^{(2)}(x_2) > 0$, the square of the second-order bias term works as the penalty and that of the first-order bias term becomes the linear restriction that shows up in the definition of the AFO bandwidths. In fact, the bandwidths that minimize the MMSE are asymptotically equivalent to the AFO bandwidths. This claim can be proved rigorously as a special case of the following theorem.

We propose a feasible bandwidth selection method based on the MMSE. The proposed method for bandwidth selection can be considered as a generalization of the traditional plug-in method (see, e.g., Wand and Jones, 1994, Section 3.6). Let $\hat{f}(\cdot)$, $\hat{f}^{(2)}(\cdot)$ and $\hat{f}^{(4)}(\cdot)$ be some consistent estimators of $f(\cdot)$, $f^{(2)}(\cdot)$ and $f^{(4)}(\cdot)$. Consider the following plug-in version of the MMSE denoted by $\widehat{MMSE}$:

$$\widehat{MMSE}_n(h) = \left\{ \frac{\mu_2}{2} \left[\hat{f}^{(2)}(x_1) h_1^2 - \hat{f}^{(2)}(x_2) h_2^2 \right] \right\}^2 + \left\{ \frac{\mu_4}{4!} \left[\hat{f}^{(4)}(x_1) h_1^4 - \hat{f}^{(4)}(x_2) h_2^4 \right] \right\}^2 + \frac{\nu_0}{n} \left\{ \frac{\hat{f}(x_1)}{h_1} + \frac{\hat{f}(x_2)}{h_2} \right\}. \tag{8}$$

Let $(\hat{h}_1, \hat{h}_2)$ be a combination of bandwidths that minimizes the MMSE and $\hat{h}$ be $(\hat{h}_1, \hat{h}_2)$. In the next theorem, we show that $(\hat{h}_1, \hat{h}_2)$ is asymptotically as good as the AFO bandwidths in the sense of Hall (1983) (see equation (2.2) of Hall, 1983). We remark that constructing the MMSE does not require prior knowledge of the sign. Moreover the next theorem shows that the proposed bandwidths automatically adjust to each situation asymptotically.

THEOREM 1 *Suppose that the conditions stated in Lemma 1 hold. Assume further that, for* $j = 1, 2$, $\hat{f}(x_j)$, $\hat{f}^{(2)}(x_j)$ *and* $\hat{f}^{(4)}(x_j)$ *satisfy* $\hat{f}(x_j) \to f(x_j)$, $\hat{f}^{(2)}(x_j) \to f^{(2)}(x_j)$ *and* $\hat{f}^{(4)}(x_j) \to f^{(4)}(x_j)$ *in probability, respectively. Let* $\hat{h}$ *be a combination of bandwidths that minimizes the MMSE defined in (8). Then, the following hold.*

(i) *When* $f^{(2)}(x_1)f^{(2)}(x_2) < 0$,

$$\frac{\hat{h}_1}{h_1^*} \to 1, \quad \frac{\hat{h}_2}{h_2^*} \to 1, \quad \textit{and} \quad \frac{\widehat{MMSE}_n(\hat{h})}{MSE_n(h^*)} \to 1$$

in probability.

(ii) *When* $f^{(2)}(x_1)f^{(2)}(x_2) > 0$ *and* $f^{(2)}(x_2)^2 f^{(4)}(x_1) \neq f^{(2)}(x_1)^2 f^{(4)}(x_2)$,

$$\frac{\hat{h}_1}{h_1^{**}} \to 1, \quad \frac{\hat{h}_2}{h_2^{**}} \to 1, \quad \textit{and} \quad \frac{\widehat{MMSE}_n(\hat{h})}{MSE_n(h^{**})} \to 1$$

in probability.

The first part of Theorem 1 (i) and (ii) implies that the bandwidths that minimize the MMSE are asymptotically equivalent to the AFO bandwidths regardless of the sign of the product.[7] The second part shows that the minimized value of the plug-in version of the MMSE is asymptotically the same as the MSE evaluated at the AFO bandwidths. These two findings show that the bandwidths that minimize the MMSE possess the desired asymptotic properties. These findings also justify the use of the MMSE as a criterion function.

[7] Observe that the assumptions of Theorem 1 require pilot estimates of $f(x_j)$, $f^{(2)}(x_j)$ and $f^{(4)}(x_j)$ for $j = 1, 2$. We can use the standard kernel density and kernel density derivative estimators. See Wand and Jones, 1994 for a basic treatment of density and density derivative estimation.

3 Nonparametric Estimation for Differences of Regression Functions

In this section, we extend the approach proposed in the previous section to the nonparametric estimation of the difference of regression functions. The nonparametric regression estimators that we consider are LLR estimators proposed by Stone (1977) and investigated by Fan (1992). Let Y_i be a scalar random variable, and let X_i be a scalar variable having common density $f(\cdot)$. Throughout this section, we assume that (X_1, Y_1), ..., (X_n, Y_n) are independent and identically distributed observations. We use $\sigma^2(x)$ to denote the conditional variance of Y_i given $X_i = x$. Suppose we are interested in estimating the difference of the conditional expectation functions at two points x_1 and x_2, i.e., $m(x_1) - m(x_2)$ where $m(x) = E(Y_i|X_i = x)$. The LLR estimator for the conditional mean function at x_1 is the solution for α to the following problem:

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \{Y_i - \alpha - \beta(X_i - x_1)\}^2 K\left(\frac{X_i - x_1}{h_1}\right),$$

where $K(\cdot)$ is a kernel function and h_1 is a bandwidth. The solution to this minimization problem can be expressed as

$$\begin{bmatrix} \hat{\alpha}_{h_1}(x_1) \\ \hat{\beta}_{h_1}(x_1) \end{bmatrix} = (X(x_1)'W(x_1)X(x_1))^{-1} X(x_1)'W(x_1)Y$$

where $X(x_1)$ is an $n \times 2$ matrix whose ith row is given by $(1, X_i - x_1)$, $Y = (Y_1, \ldots, Y_n)'$, $W(x_1) = \text{diag}(K_{h_1}(X_i - x_1))$ and $K_{h_1}(\cdot) = K(\cdot/h_1)/h_1$. The LLR estimator of $m(x_1)$ can also be written as $\hat{\alpha}_{h_1}(x_1) = e_1' (X(x_1)'W(x_1)X(x_1))^{-1} X(x_1)'W(x_1)Y$, where e_1 is a 2×1 vector having one in the first entry and zero in the other entry. $\hat{\alpha}_{h_2}(x_2)$ can be obtained analogously. Denote $\hat{\alpha}_{h_1}(x_1)$ and $\hat{\alpha}_{h_2}(x_2)$ by $\hat{m}_1(x_1)$ and $\hat{m}_2(x_2)$, respectively. Then the estimated difference of the regression functions is $\hat{m}_1(x_1) - \hat{m}_2(x_2)$.

We first consider the case in which both x_1 and x_2 are interior points of the support of f. Then, we consider the case in which they are near the boundary. According to the standard discussion of LLR estimators, the basic characteristics

of bias and variance for interior points are the same as those for boundary points. However, essentially different behaviors arise because we take a second-order bias term into consideration as we have done for density estimation.

3.1 Differences of LLR Estimators at Interior Points

In this subsection, we proceed under the following assumptions.

ASSUMPTION 5 *The conditional variance $\sigma^2(\cdot)$ is an element of $\mathcal{F}_0(\mathcal{D}_j)$ where $\mathcal{D}_j$ is an open neighborhood of x_j for $j = 1, 2$.*

ASSUMPTION 6 *The conditional mean function $m(\cdot)$ is an element of $\mathcal{F}_2(\mathcal{D}_j)$ where $\mathcal{D}_j$ is an open neighborhood of x_j for $j = 1, 2$.*

Let $m^{(j)}(\cdot)$ denote the jth derivative of $m(\cdot)$. Under Assumptions 1, 2, 3, 5 and 6, a straightforward extension of Theorem 1 in Fan (1992) shows

$$\begin{aligned} MSE_n(h) &= E\Big[\left\{ [\hat{m}_1(x_1) - \hat{m}_2(x_2)] - [m(x_1) - m(x_2)] \right\}^2 \Big| X \Big] \\ &= \left\{ \frac{\mu_2}{2} \left[m^{(2)}(x_1) h_1^2 - m^{(2)}(x_2) h_2^2 \right] \right\}^2 + \frac{\nu_0}{n} \left\{ \frac{\sigma^2(x_1)}{h_1 f(x_1)} + \frac{\sigma^2(x_2)}{h_2 f(x_2)} \right\} \\ &\quad + o\left(h_1^4 + h_1^2 h_2^2 + h_2^4 + \frac{1}{n h_1} + \frac{1}{n h_2} \right), \end{aligned}$$

where $X = (X_1, X_2, \ldots, X_n)'$. This implies that we encounter the same problem as before when trying to minimize the AMSE based on this MSE. Hence, as in the case of density estimation, we must consider the MSE with a second-order bias term. A result concerning the higher-order approximation of the MSE is provided by Fan, Gijbels, Hu, and Huang (1996). However, because their result is up to an order that disappears when symmetric kernel functions are used, it is not sufficient for our purpose. Hence, the next lemma, which is analogous to Lemma 1, generalizes the higher-order approximation of Fan, Gijbels, Hu, and Huang (1996). We proceed under the following assumption:

ASSUMPTION 7 *The conditional mean function $m(\cdot)$ is an element of $\mathcal{F}_4(\mathcal{D}_j)$ where $\mathcal{D}_j$ is an open neighborhood of x_j for $j = 1, 2$.*

It is common to use local polynomial regression (LPR) estimators instead of LLR estimators when the conditional mean function is four times continuously differentiable. However, we proceed with the LLR estimators for exactly the same reason that we employ second-order kernel functions rather than higher-order kernel functions for the problem of density estimation.

LEMMA 3 *Suppose Assumptions 1, 2, 3, 5 and 7 hold. Then, it follows that*

$$MSE_n(h) = \left\{ \frac{\mu_2}{2} \left[m^{(2)}(x_1)h_1^2 - m^{(2)}(x_2)h_2^2 \right] + \left[b(x_1)h_1^4 - b(x_2)h_2^4 \right] + o\big(h_1^4 + h_2^4\big) \right\}^2 + \frac{\nu_0}{n} \left\{ \frac{\sigma^2(x_1)}{h_1 f(x_1)} + \frac{\sigma^2(x_2)}{h_2 f(x_2)} \right\} + o\left(\frac{1}{nh_1} + \frac{1}{nh_2} \right),$$

where

$$b(x) = \frac{1}{4} \left\{ \frac{m^{(2)}(x)}{f(x)^2} (\mu_4 - \mu_2) \left[f^{(2)}(x)f(x) - f^{(1)}(x)^2 \right] + \frac{m^{(4)}(x)}{6} \mu_4 \right\}.$$

Based on the MSE provided in Lemma 3, the AFO optimal bandwidths used to estimate the difference of regression functions at two interior points are obtained in the manner described in Definition 1.

DEFINITION 2 *The AFO bandwidths for the difference of regression functions at interior points* minimize the AMSE defined by

$$AMSE_{1n}(h) = \left\{ \frac{\mu_2}{2} \left[m^{(2)}(x_1)h_1^2 - m^{(2)}(x_2)h_2^2 \right] \right\}^2 + \frac{\nu_0}{n} \left\{ \frac{\sigma^2(x_1)}{h_1 f(x_1)} + \frac{\sigma^2(x_2)}{h_2 f(x_2)} \right\}$$

when $m^{(2)}(x_1)m^{(2)}(x_2) < 0$. Their explicit expressions are given by $h_1^* = \theta^* n^{-1/5}$ and $h_2^* = \lambda^* h_1^*$, where

$$\theta^* = \left\{ \frac{\nu_0 \sigma^2(x_1)}{\mu_2^2 f(x_1) m^{(2)}(x_1) \left[m^{(2)}(x_1) - \lambda^{*2} m^{(2)}(x_2) \right]} \right\}^{1/5}, \quad \text{and}$$

$$\lambda^* = \left\{ -\frac{\sigma^2(x_2) f(x_1) m^{(2)}(x_1)}{\sigma^2(x_1) f(x_2) m^{(2)}(x_2)} \right\}^{1/3}.$$

When $m^{(2)}(x_1)m^{(2)}(x_2) > 0$, *the AFO bandwidths for the difference of regression*

functions at interior points minimize the AMSE defined by

$$AMSE_{2n}(h) = \left\{b(x_1)h_1^4 - b(x_2)h_2^4\right\}^2 + \frac{\nu_0}{n}\left\{\frac{\sigma^2(x_1)}{h_1 f(x_1)} + \frac{\sigma^2(x_2)}{h_2 f(x_2)}\right\}$$

subject to the restriction $m^{(2)}(x_1)h_1^2 - m^{(2)}(x_2)h_2^2 = 0$ under the assumption of $m^{(2)}(x_2)^2 b(x_1) \neq m^{(2)}(x_1)^2 b(x_2)$. Their explicit expressions are given by $h_1^{**} = \theta^{**} n^{-1/9}$ and $h_2^{**} = \lambda^{**} h_1^{**}$, where

$$\theta^{**} = \left\{\frac{\nu_0}{8\left[m^{(4)}(x_1) - \lambda^{**4} m^{(4)}(x_2)\right]^2}\left[\frac{\sigma^2(x_1)}{f(x_1)} + \frac{\sigma^2(x_2)}{\lambda^{**} f(x_2)}\right]\right\}^{1/9} \quad \text{and} \quad \lambda^{**} = \left\{\frac{m^{(2)}(x_1)}{m^{(2)}(x_2)}\right\}^{1/2}.$$

The dichotomous behavior of the AFO bandwidths is evident.[8] In this context, the MMSE is defined by

$$MMSE_n(h) = \left\{\frac{\mu_2}{2}\left[m^{(2)}(x_1)h_1^2 - m^{(2)}(x_2)h_2^2\right]\right\}^2 + \left\{b(x_1)h_1^4 - b(x_2)h_2^4\right\}^2 + \frac{\nu_0}{n}\left\{\frac{\sigma^2(x_1)}{h_1 f(x_1)} + \frac{\sigma^2(x_2)}{h_2 f(x_2)}\right\},$$

and its plug-in version is defined by

$$\widehat{MMSE}_n(h) = \left\{\frac{\mu_2}{2}\left[\hat{m}^{(2)}(x_1)h_1^2 - \hat{m}^{(2)}(x_2)h_2^2\right]\right\}^2 + \left\{\hat{b}(x_1)h_1^4 - \hat{b}(x_2)h_2^4\right\}^2 + \frac{\nu_0}{n}\left\{\frac{\hat{\sigma}^2(x_1)}{h_1 \hat{f}(x_1)} + \frac{\hat{\sigma}^2(x_2)}{h_2 \hat{f}(x_2)}\right\}, \tag{9}$$

where $\hat{m}^{(2)}(x_j)$, $\hat{b}(x_j)$, $\hat{\sigma}^2(x_j)$ and $\hat{f}(x_j)$ are consistent estimators of $m^{(2)}(x_j)$, $b(x_j)$, $\sigma^2(x_j)$ and $f(x_j)$ for $j = 1, 2$, respectively. Let $(\hat{h}_1, \hat{h}_2)$ be a combination of bandwidths that minimizes the MMSE given in (9) and $\hat{h}$ denote $(\hat{h}_1, \hat{h}_2)$. The next theorem is presented without proof because it is analogous to Theorem 1.

THEOREM 2 *Suppose that the conditions stated in Lemma 3 hold. Assume further that, for* $j = 1, 2$, $\hat{m}^{(2)}(x_j)$, $\hat{b}(x_j)$, $\hat{f}(x_j)$ *and* $\hat{\sigma}^2(x_j)$ *satisfy* $\hat{m}^{(2)}(x_j) \to m^{(2)}(x_j)$, $\hat{b}(x_j) \to b(x_j)$, $\hat{f}(x_j) \to f(x_j)$ *and* $\hat{\sigma}^2(x_j) \to \sigma^2(x_j)$ *in probability, respectively. Then,*

[8] Uniqueness of the AFO bandwidths for the difference of regression functions at interior points can be verified in the same manner as that of density functions.

the following hold.

(i) *When* $m^{(2)}(x_1)m^{(2)}(x_2) < 0$,

$$\frac{\hat{h}_1}{h_1^*} \to 1, \quad \frac{\hat{h}_2}{h_2^*} \to 1, \quad \textit{and} \quad \frac{\widehat{MMSE}_n(\hat{h})}{MSE_n(h^*)} \to 1$$

in probability.

(ii) *When* $m^{(2)}(x_1)m^{(2)}(x_2) > 0$ *and* $m^{(2)}(x_2)^2 b(x_1) \neq m^{(2)}(x_1)^2 b(x_2)$

$$\frac{\hat{h}_1}{h_1^{**}} \to 1, \quad \frac{\hat{h}_2}{h_2^{**}} \to 1, \quad \textit{and} \quad \frac{\widehat{MMSE}_n(\hat{h})}{MSE_n(h^{**})} \to 1$$

in probability.

Analogous remarks to those made for Theorem 1 apply for Theorem 2.

3.2 Differences of LLR Estimators Near the Boundary

Next, we consider estimating the difference of functions at given points near the boundary by using the difference of local linear estimators of functions. Recall that the results for cases in which the estimand is the difference of a density function or a regression curve at interior points can be generalized to cases where the estimand is the difference of two distinct densities or regression curves. As we make clear later, this also applies to the difference of regression curves near boundary points. However, for boundary cases, there are more cases to consider because a boundary point can be either the left or the right boundary. Here we consider the problem of the sharp RDD because of its empirical relevance. Define $m_1(z) = E(Y_i|X_i = z)$ for $z \geq x$ and $m_2(z) = E(Y_i|X_i = z)$ for $z < x$. Suppose that the limits $\lim_{z \to x+} m_1(z)$ and $\lim_{z \to x-} m_2(z)$ exist where $z \to x+$ and $z \to x-$ mean taking the limits from the right and left, respectively. Denote $\lim_{z \to x+} m_1(z)$ and $\lim_{z \to x-} m_2(z)$ by $m_1(x)$ and $m_2(x)$, respectively. The parameter of interest in the analysis of the sharp RDD is given by $\tau(x) = m_1(x) - m_2(x)$.[9] For estimating these limits, the LLR is particularly attractive

[9] See Hahn, Todd, and Van Der Klaauw (2001).

because it exhibits the automatic boundary adaptive property (Fan and Gijbels, 1992 and Hahn, Todd, and Van Der Klaauw, 2001). The LLR estimator for $m_1(x)$ is given by $\hat{\alpha}_{h,1}(x)$, where

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

where $K(\cdot)$ is a kernel function and h_1 is a bandwidth. The solution can be expressed as

$$\begin{bmatrix} \hat{\alpha}_{h_1,1}(x) \\ \hat{\beta}_{h_1,1}(x) \end{bmatrix} = \left(X(x)'W_1(x)X(x)\right)^{-1} X(x)'W_1(x)Y,$$

where $W_1(x) = \text{diag}(K_{h_1,1}(X_i - x))$ and $K_{h_1,1}(\cdot) = K(\cdot/h_1)1_{\{\cdot \geq 0\}}/h_1$, and $X(x)$ and Y are as defined in the previous subsection. Similarly, the LLR estimator for $m_2(x)$, denoted by $\hat{\alpha}_{h_2,2}(x)$, can be obtained by replacing $W_1(x)$ with $W_2(x)$, where $W_2(x) = \text{diag}(K_{h_2,2}(X_i - x))$ and $K_{h_2,2}(\cdot) = K(\cdot/h_2)1_{\{\cdot<0\}}/h_2$. Denote $\hat{\alpha}_{h_1,1}$ and $\hat{\alpha}_{h_2,2}$ by $\hat{m}_1(x)$ and $\hat{m}_2(x)$, respectively. Then, $\tau(x)$ is estimated by $\hat{\tau}(x) = \hat{m}_1(x) - \hat{m}_2(x)$, and its conditional MSE given X is given by

$$MSE_n(h) = E\Big[\{(\hat{m}_1(x) - \hat{m}_2(x)) - (m_1(x) - m_2(x))\}^2 | X\Big].$$

Define the conditional variance function σ_1^2 and σ_2^2 analogously. Also define $\sigma_1^2(x) = \lim_{z \to x+} \sigma_1^2(z)$, $\sigma_2^2(x) = \lim_{z \to x-} \sigma_2^2(z)$, $m_1^{(2)}(x) = \lim_{z \to x+} m_1^{(2)}(z)$, $m_2^{(2)}(x) = \lim_{z \to x-} m_2^{(2)}(z)$, $m_1^{(3)}(x) = \lim_{z \to x+} m_1^{(3)}(z)$, $m_2^{(3)}(x) = \lim_{z \to x-} m_2^{(3)}(z)$, $\mu_{j,0} = \int_0^\infty u^j K(u)du$ and $\nu_{j,0} = \int_0^\infty u^j K^2(u)du$ for nonnegative integer j. We proceed under the following assumption.

ASSUMPTION 8 *The density f is an element of $\mathcal{F}_1(\mathcal{D})$ where $\mathcal{D}$ is an open neighborhood of x.*

ASSUMPTION 9 *Let δ be some positive constant. The conditional mean function m_1 and the conditional variance function σ_1^2 are elements of $\mathcal{F}_3(\mathcal{D}_1)$ and $\mathcal{F}_0(\mathcal{D}_1)$, respectively, where $\mathcal{D}_1$ is a one-sided open neighborhood of x, $(x, x+\delta)$, and $m_1(x)$, $m_1^{(2)}(x)$, $m_1^{(3)}(x)$ and $\sigma_1^2(x)$ exist and are bounded. Similarly, m_2 and σ_2^2 are elements*

of $\mathcal{F}_3(\mathcal{D}_2)$ and $\mathcal{F}_0(\mathcal{D}_2)$, respectively, where $\mathcal{D}_2$ is a one-sided open neighborhood of x, $(x-\delta, x)$, and $m_2(x)$, $m_2^{(2)}(x)$, $m_2^{(3)}(x)$ and $\sigma_2^2(x)$ exist and are bounded.

Under Assumptions 1, 3, 8 and 9, we can easily generalize the result obtained by Fan and Gijbels (1992) to get

$$MSE_n(h) = \left\{ \frac{b_1}{2} \left[m_1^{(2)}(x) h_1^2 - m_2^{(2)}(x) h_2^2 \right] \right\}^2 + \frac{v}{nf(x)} \left\{ \frac{\sigma_1^2(x)}{h_1} + \frac{\sigma_2^2(x)}{h_2} \right\} + o\left(h_1^4 + h_1^2 h_2^2 + h_2^4 + \frac{1}{nh_1} + \frac{1}{nh_2} \right),$$

where

$$b_1 = \frac{\mu_{2,0}^2 - \mu_{1,0}\mu_{3,0}}{\mu_{0,0}\mu_{2,0} - \mu_{1,0}^2}, \quad \text{and} \quad v = \frac{\mu_{2,0}^2 \nu_{0,0} - 2\mu_{1,0}\mu_{2,0}\nu_{1,0} + \mu_{1,0}^2 \nu_{2,0}}{(\mu_{0,0}\mu_{2,0} - \mu_{1,0}^2)^2}.$$

Again, it is evident that the trade-off between bias and variance can break down when we try to minimize the AMSE based on this MSE. Thus, we need to consider the MSE that includes a second-order bias term. The next lemma presents the MSE with a second-order bias term for the boundary points.

LEMMA 4 *Suppose Assumptions 1, 3, 8 and 9 hold. Then, it follows that*

$$MSE_n(h) = \left\{ \frac{b_1}{2} \left[m_1^{(2)}(x) h_1^2 - m_2^{(2)}(x) h_2^2 \right] + \left[b_{2,1}(x) h_1^3 - b_{2,2}(x) h_2^3 \right] + o\left(h_1^3 + h_2^3 \right) \right\}^2 + \frac{v}{nf(x)} \left\{ \frac{\sigma_1^2(x)}{h_1} + \frac{\sigma_2^2(x)}{h_2} \right\} + o\left(\frac{1}{nh_1} + \frac{1}{nh_2} \right),$$

where

$$b_{2,j}(x) = (-1)^{j+1} \left\{ c_1 \left[\frac{m_j^{(2)}(x)}{2} \frac{f^{(1)}(x)}{f(x)} + \frac{m_j^{(3)}(x)}{6} \right] - c_2 \frac{m_j^{(2)}(x)}{2} \frac{f^{(1)}(x)}{f(x)} \right\}$$

$$c_1 = \frac{\mu_{2,0}\mu_{3,0} - \mu_{1,0}\mu_{4,0}}{\mu_{0,0}\mu_{2,0} - \mu_{1,0}^2}, \quad \textit{and} \quad c_2 = \frac{\left(\mu_{2,0}^2 - \mu_{1,0}\mu_{3,0}\right)\left(\mu_{0,0}\mu_{3,0} - \mu_{1,0}\mu_{2,0}\right)}{(\mu_{0,0}\mu_{2,0} - \mu_{1,0}^2)^2}.$$

The result given above is essentially different from the one at interior points because the second-order bias terms now involve h^3 rather than h^4. This is because the terms that disappear because of the symmetry of the kernel functions remain for

a one-sided kernel. Based on the MSE provided in Lemma 4, the AFO bandwidths for estimating the difference of regression functions at the boundary points can be defined.

DEFINITION 3 *The AFO bandwidths for the difference of regression functions at the boundary points* minimize the AMSE defined by

$$AMSE_{1n}(h) = \left\{ \frac{b_1}{2} \left[m_1^{(2)}(x) h_1^2 - m_2^{(2)}(x) h_2^2 \right] \right\}^2 + \frac{v}{nf(x)} \left\{ \frac{\sigma_1^2(x)}{h_1} + \frac{\sigma_2^2(x)}{h_2} \right\}.$$

when $m_1^{(2)}(x) m_2^{(2)}(x) < 0$. Their explicit expressions are given by $h_1^* = \theta^* n^{-1/5}$ and $h_2^* = \lambda^* h_1^*$, where

$$\theta^* = \left\{ \frac{v \sigma_1^2(x)}{b_1^2 f(x) m_1^{(2)}(x) \left[m_1^{(2)}(x) - \lambda^{*2} m_2^{(2)}(x) \right]} \right\}^{1/5} \quad \text{and} \quad \lambda^* = \left\{ -\frac{\sigma_2^2(x) m_1^{(2)}(x)}{\sigma_1^2(x) m_2^{(2)}(x)} \right\}^{1/3}.$$

When $m_1^{(2)}(x) m_2^{(2)}(x) > 0$, *the AFO bandwidths for the difference of regression functions at the boundary points* minimize the AMSE defined by

$$AMSE_{2n}(h) = \left\{ b_{2,1}(x) h_1^3 - b_{2,2}(x) h_2^3 \right\}^2 + \frac{v}{nf(x)} \left\{ \frac{\sigma_1^2(x)}{h_1} + \frac{\sigma_2^2(x)}{h_2} \right\}$$

subject to the restriction $m_1^{(2)}(x) h_1^2 - m_2^{(2)}(x) h_2^2 = 0$ under the assumption of $m_2^{(2)}(x)^3 b_{2,1}(x)^2 \neq m_1^{(3)}(x)^3 b_{2,2}(x)^2$. Their explicit expressions are given by $h_1^{**} = \theta^{**} n^{-1/7}$ and $h_2^{**} = \lambda^{**} h_1^{**}$, where

$$\theta^{**} = \left\{ \frac{v \left[\sigma_1^2(x) + \sigma_2^2(x)/\lambda^{**} \right]}{6 f(x) \left[b_{2,1}(x) - \lambda^{**3} b_{2,2}(x) \right]^2} \right\}^{1/7} \quad \text{and} \quad \lambda^{**} = \left\{ \frac{m_1^{(2)}(x)}{m_2^{(2)}(x)} \right\}^{1/2}.$$

Again, it is evident that the AFO bandwidths exhibit the dichotomous behavior.[10] However, the most important difference between these bandwidths and those for interior points is that when the sign is positive, the order of the bandwidths is $n^{-1/7}$.

[10] Uniqueness of the AFO bandwidths for the difference of regression functions at the boundary points can be verified in the same manner as that of density functions.

In the present context, the MMSE used to construct feasible automatic bandwidths is defined by

$$MMSE_n(h) = \left\{ \frac{b_1}{2} \left[m_1^{(2)}(x) h_1^2 - m_2^{(2)}(x) h_2^2 \right] \right\}^2 + \left\{ b_{2,1}(x) h_1^3 - b_{2,2}(x) h_2^3 \right\}^2 + \frac{v}{nf(x)} \left\{ \frac{\sigma_1^2(x)}{h_1} + \frac{\sigma_2^2(x)}{h_2} \right\},$$

and its plug-in version is defined by

$$\widehat{MMSE}_n(h) = \left\{ \frac{b_1}{2} \left[\hat{m}_1^{(2)}(x) h_1^2 - \hat{m}_2^{(2)}(x) h_2^2 \right] \right\}^2 + \left\{ \hat{b}_{2,1}(x) h_1^3 - \hat{b}_{2,2}(x) h_2^3 \right\}^2 + \frac{v}{n\hat{f}(x)} \left\{ \frac{\hat{\sigma}_1^2(x)}{h_1} + \frac{\hat{\sigma}_2^2(x)}{h_2} \right\}, \tag{10}$$

where $\hat{m}_j^{(2)}(x)$, $\hat{b}_{2,1}(x)$, $\hat{b}_{2,2}(x)$, $\hat{\sigma}_j^2(x)$ and $\hat{f}(x)$ are consistent estimators of $m_j^{(2)}(x)$, $b_{2,1}(x)$, $b_{2,1}(x)$, $\sigma_j^2(x)$ and $f(x)$ for $j = 1, 2$, respectively. Let $(\hat{h}_1, \hat{h}_2)$ be a combination of bandwidths that minimizes this plug-in version of the MMSE and $\hat{h}$ denote $(\hat{h}_1, \hat{h}_2)$. Then, the next theorem shows that the bandwidths that minimize the MMSE are again asymptotically as good as the AFO bandwidths. The proof of Theorem 3 is similar to that of Theorem 1 and it is provided in Arai and Ichimura (2013).

THEOREM 3 *Suppose that the conditions stated in Lemma 4 hold. Assume further that, for* $j = 1, 2$, $\hat{m}_j^{(2)}(x)$, $\hat{b}_{2,j}(x)$, $\hat{f}(x)$ *and* $\hat{\sigma}_j^2(x)$ *satisfy* $\hat{m}_j^{(2)}(x) \to m_j^{(2)}(x)$, $\hat{b}_{2,j}(x) \to b_{2,j}(x)$, $\hat{f}(x) \to f(x)$ *and* $\hat{\sigma}_j^2(x) \to \sigma_j^2(x)$ *in probability for* $j = 1, 2$, *respectively. Then, the following hold.*

(i) *When* $m_1^{(2)}(x) m_2^{(2)}(x) < 0$,

$$\frac{\hat{h}_1}{h_1^*} \to 1, \quad \frac{\hat{h}_2}{h_2^*} \to 1, \quad \textit{and} \quad \frac{\widehat{MMSE}_n(\hat{h})}{MSE_n(h^*)} \to 1$$

in probability.

(ii) *When* $m_1^{(2)}(x) m_2^{(2)}(x) > 0$ *and* $m_2^{(2)}(x)^3 b_{2,1}(x)^2 \neq m_1^{(2)}(x)^3 b_{2,2}(x)^2$

$$\frac{\hat{h}_1}{h_1^{**}} \to 1, \quad \frac{\hat{h}_2}{h_2^{**}} \to 1, \quad \textit{and} \quad \frac{\widehat{MMSE}_n(\hat{h})}{MSE_n(h^{**})} \to 1$$

in probability.

The remarks made for Theorem 1 essentially apply for Theorem 3. Similar to Theorems 1 and 2, Theorem 3 requires pilot estimates for $m_j^{(2)}(x)$, $b_{2,j}(x)$, $f(x)$ and $\sigma_j^2(x)$. A detailed explanation of how to obtain the pilot estimates is given in Arai and Ichimura (2013).

Fan and Gijbels (1996, Section 4.3) points out that replacing constants depending on a kernel function with finite sample approximations can improve finite sample performance. This leads to the following version of the estimated MMSE:

$$\widehat{MMSE}_n^E(h) = \left\{\tilde{b}_{1,1}(x) - \tilde{b}_{1,2}(x)\right\}^2 + \left\{\tilde{b}_{2,1}(x) - \tilde{b}_{2,2}(x)\right\}^2 + \hat{\sigma}_1^2(x)\tilde{v}_1(x) + \hat{\sigma}_2^2(x)\tilde{v}_2(x), \tag{11}$$

where

$$\begin{aligned}
\tilde{b}_{1,j}(x) &= \frac{\hat{m}_1^{(2)}(x)}{2} e_1' \tilde{S}_{n,0,j}^{-1} \tilde{c}_{n,2,j}, \\
\tilde{b}_{2,j}(x) &= \left\{ \frac{\hat{m}_1^{(2)}(x)}{2} \cdot \frac{\hat{f}^{(1)}(x)}{\hat{f}(x)} + \frac{\hat{m}_j^{(3)}(x)}{3!} \right\} e_1' \tilde{S}_{n,0,j}^{-1} c_{n,3,j} - \frac{\hat{m}_1^{(2)}(x)}{2} \cdot \frac{\hat{f}^{(1)}(x)}{\hat{f}(x)} e_1' \tilde{S}_{n,0,j}^{-1} S_{n,1,j} \tilde{S}_{n,0,j}^{-1} \tilde{c}_{n,2,j}, \\
\tilde{v}_j(x) &= e_1' S_{n,0,j}^{-1} T_{n,0,j} S_{n,0,j}^{-1} e_1, \quad \tilde{S}_{n,0,j} = S_{n,0,j} - \frac{\hat{f}^{(1)}(x)}{\hat{f}(x)} S_{n1,j}, \quad \tilde{c}_{n,2,j} = c_{n,2,j} - \frac{\hat{f}^{(1)}(x)}{\hat{f}(x)} c_{n,3,j}, \\
S_{n,k,j} &= \begin{bmatrix} s_{n,k,j} & s_{n,k+1,j} \\ s_{n,k+1,j} & s_{n,k+2,j} \end{bmatrix}, \quad T_{n,k,j} = \begin{bmatrix} t_{n,k,j} & t_{n,k+1,j} \\ t_{n,k+1,j} & t_{n,k+2,j} \end{bmatrix}, \quad c_{n,k,j} = \begin{bmatrix} s_{n,k,j} \\ s_{n,k+1,j} \end{bmatrix}, \\
s_{n,k,j} &= \sum_{i=1}^n K_{h,j}(X_i - x)(X_i - x)^k, \quad t_{n,k,j} = \sum_{i=1}^n K_{h,j}^2(X_i - x)(X_i - x)^k,
\end{aligned} \tag{12}$$

for $j = 1, 2$. Let $(\hat{h}_1^E, \hat{h}_2^E)$ minimize the MMSE defined by (11), and let $\hat{h}^E$ denote $(\hat{h}_1^E, \hat{h}_2^E)$. Then, the following extension of Theorem 3 holds.

COROLLARY 1 *Suppose that the conditions stated in Lemma 4 hold for each case. Also assume that the second derivative of the density f exists in the neighborhood of x. Then, the results for $\hat{h}_1$, $\hat{h}_2$ and $\widehat{MMSE}_n(\hat{h})$ also hold for $\hat{h}_1^E$, $\hat{h}_2^E$ and $\widehat{MMSE}_n^E(\hat{h}^E)$.*

It is also possible to use the heteroskedasticity-robust variance estimator for the variance component (Eicker, 1967, Huber, 1967 and White, 1980). In this case,

the estimated MMSE is defined by

$$\widehat{MMSE}_n^R(h) = \left\{\tilde{b}_{1,1}(x) - \tilde{b}_{1,2}(x)\right\}^2 + \left\{\tilde{b}_{2,1}(x) - \tilde{b}_{2,2}(x)\right\}^2 + \tilde{\omega}_1(x) + \tilde{\omega}_2(x), \quad (13)$$

where

$$\tilde{\omega}_j(x) = e_1' S_{n,0,j}^{-1} \tilde{T}_{n,0,j} S_{n,0,j}^{-1} e_1, \quad \tilde{T}_{n,k,j} = \begin{bmatrix} \tilde{t}_{n,k,j} & \tilde{t}_{n,k+1,j} \\ \tilde{t}_{n,k+1,j} & \tilde{t}_{n,k+2,j} \end{bmatrix},$$
$$\tilde{t}_{n,k,j} = \sum_{i=1}^{n} \tilde{\epsilon}_i^2 K_{h,j}^2 (X_i - x)(X_i - x)^k, \quad \tilde{\epsilon}_i = Y_i - \tilde{Y}_i,$$

and $\tilde{Y}_i$ are the fitted values from the third-order LPR used to estimate the second derivatives. Let $(\hat{h}_1^R, \hat{h}_2^R)$ minimize the MMSE defined by (13), and let $\hat{h}^R$ denote $(\hat{h}_1^R, \hat{h}_2^R)$. Then, the following extension of Theorem 3 holds. Its proof is not presented because it is standard given the results in Corollary 1.

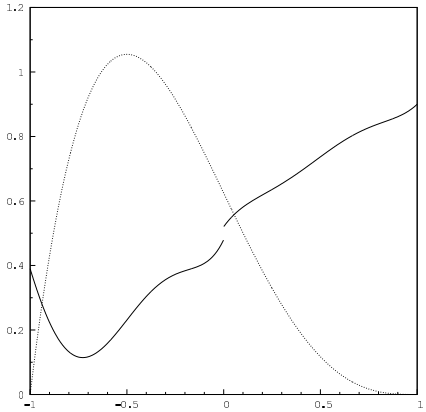
COROLLARY 2 *Suppose that the conditions stated in Lemma 4 hold for each case. Also assume that the second derivative of the density f exists in the neighborhood of x. Then, the results for $\hat{h}_1$, $\hat{h}_2$ and $\widehat{MMSE}_n(\hat{h})$ also hold for $\hat{h}_1^R$, $\hat{h}_2^R$ and $\widehat{MMSE}_n^R(\hat{h}^R)$.*

4 Simulation

To investigate the finite sample performance of the proposed method, we conducted simulation experiments. We focused on the case of the sharp RDD because it is the most empirically relevant case and because there are competing bandwidth selection methods in the literature.

4.1 Simulation Designs

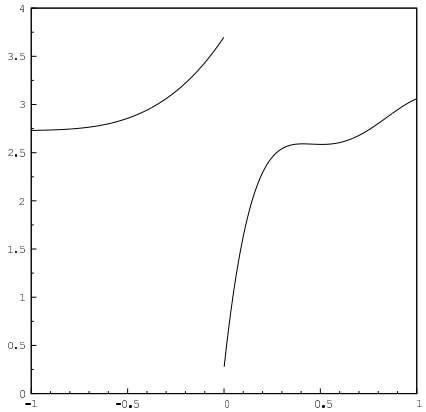
The objective of the RDD application is to estimate $\tau(x)$ defined in Section 3.2. We consider six designs. Five of them are the ones studied by Calonico, Cattaneo, and Titiunik (2012) (hereafter CCT) and IK, and the other is a modification of CCT's Design 3. The designs investigated are given in Figure 1.



1. Lee (2008) Data (Design 1 of IK and CCT)

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

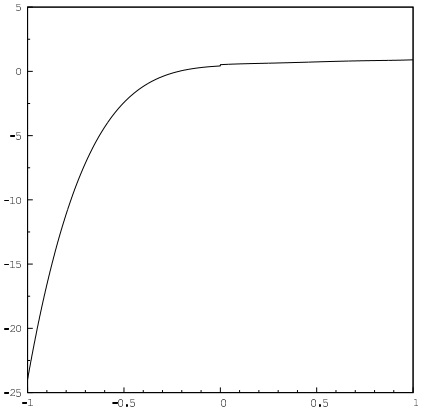
$m_2(z) = 0.48 + 1.27z + 7.18z^2 + 20.21z^3 + 21.54z^4 + 7.33z^5$



2. Ludwign and Miller (2007) Data (Design 2 of CCT)

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

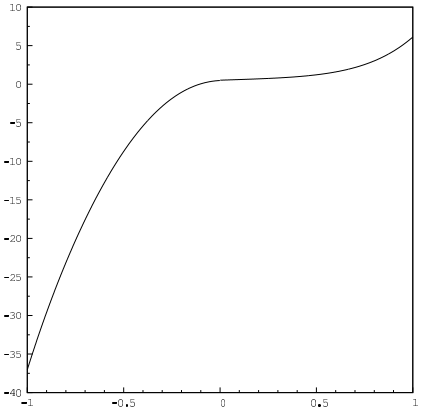
$m_2(z) = 3.70 + 2.99z + 3.28z^2 + 1.45z^3 + 0.22z^4 + 0.03z^5$



3. Constant Additive Treatment Effect (Design 3 of IK)

$m_1(z) = 1.42 + 0.84z - 3.0z^2 + 7.99z^3 - 9.01z^4 + 3.56z^5$

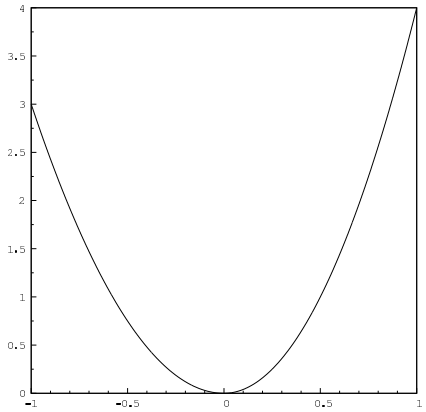
$m_2(z) = 0.42 + 0.84z - 3.0z^2 + 7.99z^3 - 9.01z^4 + 3.56z^5$



4. Modified Version of Design 3 of CCT

$m_1(z) = 0.52 + 0.84z - 0.30z^2 + 2.397z^3 - 0.901z^4 + 3.56z^5$

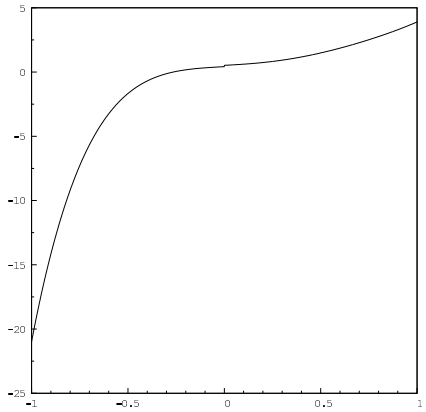
$m_2(z) = 0.48 + 1.27z - 28.72z^2 + 20.21z^3 + 23.694z^4 + 10.995z^5$



5. Quadratic (Design 2 of IK)

$m_1(z) = 4.0z^2$

$m_2(z) = 3.0z^2$



6. Constant Additive Treatment Effect 2 (Design 4 of IK)

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

$m_2(z) = 0.42 + 0.84z + 7.99z^3 - 9.01z^4 + 3.56z^5$

Figure 1. Simulation Design (The dotted line in the panel for Design 1 denotes the density of the forcing variable. The supports for $m_1(z)$ and $m_2(z)$ are $z \geq 0$ and $z < 0$, respectively.)

For the first two designs, the sign of the product of the second derivatives is negative. The ratio of the second derivative on the right to the one on the left in absolute value is moderate for Design 1, whereas it is rather large for Design 2. For the next two designs, the sign is positive. Design 3 has exactly the same second derivative on both sides, and Design 4 has a relatively large ratio of second derivatives. The last two designs are excluded cases of Theorem 3. The sign is positive, but the values of the third derivatives are zero for Design 5. The values of the second derivatives are zero for Design 6.

For each design, we consider a normally distributed additive error term with mean zero and standard deviation 0.1295. We use data sets of 500 observations and the results are drawn from 10,000 replications. The specification for the forcing variable is exactly the same as that considered by IK.[11] A detailed algorithm to implement the proposed method is described in the supplemental material (Arai and Ichimura, 2013).

4.2 Results

The simulation results are presented in Tables 1, 2 and 3. Table 1 reports the results for Designs 1 and 2. The first column explains the design. The second column reports the method used to obtain the bandwidth(s). AFO is the infeasible AFO bandwidths. MMSE-T is also the infeasible bandwidths that minimize the MMSE based on theoretical values. MMSE, MMSE-E and MMSE-R refer to the proposed methods based on $\widehat{MMSE}_n(h)$, $\widehat{MMSE}_n^E(h)$ and $\widehat{MMSE}_n^R(h)$, respectively. IK corresponds to the bandwidth denoted by $\hat{h}_{opt}$ in Table 2 of IK.

The cross-validation bandwidth used by Ludwig and Miller (2005, 2007) is denoted by LM; its implementation is described in Section 4.5 of IK.[12] Note that

[11] In IK the forcing variable is generated by a Beta distribution. More precisely, let Z_i have a Beta distribution with parameters $\alpha = 2$ and $\beta = 4$. Then, the forcing variable X_i is given by $2Z_i - 1$.

[12] MMSE-T, MMSE, MMSE-E, MMSE-R, and LM involve numerical optimization. For MMSE-T, MMSE, MMSE-E and MMSE-R, the minimum of search region is determined by the 3rd nearest neighbor from the discontinuity point on each side of the threshold. For the minimum of search region on each side of the threshold for LM, we first obtain the 3rd nearest neighbor for each observation point X_i in a direction away from the origin. Then the maximum taken for each side of the threshold and the maximum of the two maximums is used for LM. The maximum of search region is one for

the cross-validation bandwidth involves one ad hoc parameter although other methods presented here are fully data-driven.[13] DM is the plug-in bandwidths used by DesJardins and McCall (2008) as explained in Section 4.4 of IK.[14]

The third and fourth columns report the mean (labeled 'Mean') and standard deviation (labeled 'SD') of the bandwidths for IK, LM, and DM. For the others, these columns report the bandwidth obtained for the right sides of the threshold.[15] The fifth and sixth columns report the corresponding bandwidths on the left sides of the threshold. The seventh and eighth columns report the bias (Bias) and the root mean squared error (RMSE) for the sharp RDD estimate, denoted by $\hat{\tau}$.

First, we look at the designs in which the signs of the second derivatives are distinct. The top panel of Table 1, which reports the results for Design 1, demonstrates that all methods perform similarly. DM performs only marginally better. Given similar magnitude for the second derivatives in absolute value, choosing a single bandwidth might be appropriate. The bottom panel of Table 1 reports the results for Design 2, in which there exists a large difference in the magnitudes of the second derivatives. Now MMSE, MMSE-E, MMSE-R perform significantly better than the other methods, followed by LM. IK and DM perform very poorly.

Next, we examine designs in which the sign of the product of the second derivatives is positive. The top panel of Table 2 show that all methods except AFO perform reasonably well for Design 3. The bottom panel of Table 2 reports that MMSE, MMSE-E and MMSE-R work quite well for Design 4, reflecting the advantage of allowing distinct bandwidths. Remember that the second derivatives differ quite substantially.

Next, we look at the designs that do not satisfy the assumptions of Theorem 3. The top panel of Table 3 reports the results for Design 5. All methods perform reasonably well. This may be because Design 5 is such a simple model and that the

all methods. Nine initial values of 0.1, 0.2 ,. . ., and 0.9 are tried for all methods.

[13]See Section 4.5 of IK for the ad hoc parameter δ used in the cross-validation method. δ is set to 0.5 as in IK.

[14]The plug-in method used by DesJardins and McCall (2008) is proposed by Fan and Gijbels (1992, 1995).

[15]No SD concerning AFO or MMSE-T is presented for Designs 1-4. No result of AFO or MMSE-T is presented for Designs 5 and 6 because the AFO and MMSE-T bandwidths are not well-defined.

Bias and RMSE for the Sharp RDD, n=500

		$\hat{h}_1$		$\hat{h}_2$		$\hat{\tau}$	
DGP	Method	Mean	SD	Mean	SD	Bias	RMSE
Design 1	AFO	0.262		0.196		0.024	0.054
	MMSE-T	0.255		0.195		0.024	0.055
	MMSE	0.389	0.191	0.381	0.159	0.033	0.057
	MMSE-E	0.457	0.255	0.396	0.172	0.033	0.056
	MMSE-R	0.434	0.268	0.380	0.186	0.033	0.058
	IK	0.448	0.046			0.041	0.054
	LM	0.424	0.118			0.037	0.054
	DM	0.556	0.135			0.037	0.051
Design 2	AFO	0.091		0.232		0.057	0.087
	MMSE-T	0.091		0.232		0.057	0.087
	MMSE	0.076	0.005	0.187	0.026	0.039	0.085
	MMSE-E	0.077	0.007	0.188	0.033	0.041	0.084
	MMSE-R	0.062	0.027	0.172	0.075	0.041	0.085
	IK	0.249	0.016			0.237	0.245
	LM	0.129	0.013			0.078	0.107
	DM	0.267	0.020			0.264	0.272

Bias and RMSE for the Sharp RDD, n=500

		$\hat{h}_1$		$\hat{h}_2$		$\hat{\tau}$	
DGP	Method	Mean	SD	Mean	SD	Bias	RMSE
Design 3	AFO	0.345		0.345		-0.081	0.091
	MMSE-T	0.345		0.345		-0.081	0.091
	MMSE	0.372	0.213	0.209	0.056	-0.024	0.068
	MMSE-E	0.393	0.227	0.181	0.033	-0.013	0.071
	MMSE-R	0.363	0.241	0.159	0.058	-0.012	0.061
	IK	0.163	0.012			-0.008	0.060
	LM	0.112	0.008			-0.003	0.071
	DM	0.204	0.041			-0.016	0.063
Design 4	AFO	0.896		0.082		0.031	0.071
	MMSE-T	0.741		0.125		-0.007	0.053
	MMSE	0.412	0.185	0.119	0.027	-0.032	0.074
	MMSE-E	0.525	0.261	0.126	0.028	-0.029	0.071
	MMSE-R	0.481	0.268	0.111	0.044	-0.021	0.071
	IK	0.145	0.007			-0.070	0.096
	LM	0.088	0.006			-0.025	0.085
	DM	0.144	0.006			-0.070	0.095

Bias and RMSE for the Sharp RDD, n=500

		$\hat{h}_1$		$\hat{h}_2$		$\hat{\tau}$	
DGP	Method	Mean	SD	Mean	SD	Bias	RMSE
Design 5	MMSE	0.374	0.158	0.414	0.119	0.017	0.058
	MMSE-E	0.375	0.183	0.368	0.093	0.004	0.058
	MMSE-R	0.358	0.192	0.354	0.110	0.005	0.058
	IK	0.410	0.062			0.005	0.036
	LM	0.220	0.022			-0.003	0.051
	DM	0.223	0.010			-0.003	0.049
Design 6	MMSE	0.298	0.084	0.214	0.044	-0.030	0.065
	MMSE-E	0.302	0.088	0.188	0.032	-0.024	0.068
	MMSE-R	0.273	0.112	0.168	0.058	-0.022	0.069
	IK	0.162	0.012			-0.007	0.060
	LM	0.118	0.009			-0.003	0.069
	DM	0.241	0.075			-0.027	0.099

method of bandwidth selection may not matter much. The results for Design 6 are given in the bottom panel of Table 3. All methods except DM perform reasonably well.

In summary, for the designs that satisfy the assumptions of Theorem 3, MMSE, MMSE-E and MMSE-R perform equally well except that MMSE-R works best for Design 3. IK and DM exhibits disappointing performance for some designs. MMSE, MMSE-E, MMSE-R and LM display stable performance for all designs. MMSE-R performs significantly better than LM for Design 2, 3 and 4, and it is outperformed by LM only marginally for Design 1. MMSE-R appears very promising.

5 Conclusion

In this paper, we have proposed a bandwidth selection method for the nonparametric estimation of the difference of two functions at particular points. 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 and that the optimal bandwidths that minimize the AMSE are not well-defined when the sign is positive. We introduced the concept of the AFO bandwidths, which are well-defined regardless of the sign. We proposed a feasible version of these bandwidths

that can be constructed without knowledge of the sign. The feasible bandwidths are asymptotically as good as the AFO bandwidths. Our framework can accommodate estimation problems relating to the differences of densities and differences of functions at interior and boundary points. Our Monte Carlo experiment for the sharp RDD showed that the proposed bandwidth selection method is practically useful.

Generalization of the proposed method is on our research agenda. First, we intend to address the problem of estimating the ratio of the difference of two functions. Special cases of this estimation problem are the LATE and the fuzzy RDD estimator. This is nontrivial problem because one must choose four distinct bandwidths simultaneously . Second, we intend to generalize the proposed method to the ATE estimator. This requires generalizing the results presented in Section 3.1. This is important because it requires analyzing the difference of functions on the whole support of covariates. We plan to address these issues in a separate paper.

Appendix A: Proofs

Proof of Theorem 1: Recall that the objective function is

$$\widehat{MMSE}_n(h) = \left\{\frac{\mu_2}{2}\left[\hat{f}^{(2)}(x_1)h_1^2 - \hat{f}^{(2)}(x_2)h_2^2\right]\right\}^2 + \left\{\frac{\mu_4}{4!}\left[\hat{f}^{(4)}(x_1)h_1^4 - \hat{f}^{(4)}(x_2)h_2^4\right]\right\}^2 + \frac{\nu_0}{n}\left\{\frac{\hat{f}(x_1)}{h_1} + \frac{\hat{f}(x_2)}{h_2}\right\}.$$

To begin with, we show that $\hat{h}_1$ and $\hat{h}_2$ satisfy Assumption 3. Let h_1 and h_2 be sequences that satisfy Assumption 3. Then $\widehat{MMSE}_n(h)$ converges to zero in probability by conditions of Theorem 1. Assume to the contrary that either one or both of $\hat{h}_1$ and $\hat{h}_2$ do not satisfy Assumption 3. Since $f^{(4)}(x_1)[f^{(2)}(x_2)]^2 \neq f^{(4)}(x_2)[f^{(2)}(x_1)]^2$ by assumption, $\hat{f}^{(4)}(x_1)[\hat{f}^{(2)}(x_2)]^2 \neq \hat{f}^{(4)}(x_2)[\hat{f}^{(2)}(x_1)]^2$ with probability approaching 1. Without loss of generality, we assume this as well. Then at least one of the first-order bias term, the second-order bias term and the variance term of $\widehat{MMSE}_n(\hat{h})$ does not converge to zero in probability regardless of the sign of $f^{(2)}(x_1)f^{(2)}(x_2)$. Then $\widehat{MMSE}_n(\hat{h}) > \widehat{MMSE}_n(h)$ holds for some n. This contradicts the definition of $\hat{h}$.

Hence $\hat{h}$ satisfies Assumption 3.

We first consider the case in which $f^{(2)}(x_1)f^{(2)}(x_2) < 0$. In this case, with probability approaching 1, $\hat{f}^{(2)}(x_1)\hat{f}^{(2)}(x_2) < 0$, 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 since $\hat{h}_1$ and $\hat{h}_2$ satisfy Assumption 3. Define the plug-in versions of $AMSE_{1n}(h)$ by

$$\widehat{AMSE}_{1n}(h) = \left\{ \frac{\mu_2}{2} \left[\hat{f}^{(2)}(x_1)h_1^2 - \hat{f}^{(2)}(x_2)h_2^2 \right] \right\}^2 + \frac{\nu_0}{n} \left\{ \frac{\hat{f}(x_1)}{h_1} + \frac{\hat{f}(x_2)}{h_2} \right\}.$$

Denote the minimizer of $\widehat{AMSE}_{1n}(h)$ by $\tilde{h}_1$ and $\tilde{h}_2$. As it is clear from Definition 1, we have $\tilde{h}_1 = \hat{\theta}_1 n^{-1/5} \equiv \tilde{C}_1 n^{-1/5}$ and $\tilde{h}_2 = \hat{\lambda}_1 \tilde{h}_1 \equiv \tilde{C}_2 n^{-1/5}$ where $\hat{\theta}_1$ and $\hat{\lambda}_1$ are defined in (6). With this choice, $\widehat{AMSE}_{1n}(h)$ and hence $\widehat{MMSE}_n(\tilde{h})$ converges at the rate of $n^{-4/5}$. Note that if $\hat{h}_1$ or $\hat{h}_2$ 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 $\hat{h}_2$ converges at the rate faster than $n^{-1/5}$, then the variance term converges at the rate slower than $n^{-4/5}$. These contradict the definition of $\hat{h}$. Thus the minimizer of $\widehat{MMSE}_n(h)$, $\hat{h}_1$ and $\hat{h}_2$ 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 $\hat{h}_2 = \hat{C}_2 n^{-1/5} + o_p(n^{-1/5})$ for some $O_P(1)$ sequences $\hat{C}_1$ and $\hat{C}_2$ that are bounded away from 0 as $n \to \infty$. Using this expression,

$$\widehat{MMSE}_n(\hat{h}) = n^{-4/5} \left\{ \frac{\mu_2}{2} \left[\hat{f}^{(2)}(x_1)\hat{C}_1^2 - \hat{f}^{(2)}(x_2)\hat{C}_2^2 \right] \right\}^2 + \frac{\nu_0}{n^{4/5}} \left\{ \frac{\hat{f}(x_1)}{\hat{C}_1} + \frac{\hat{f}(x_2)}{\hat{C}_2} \right\} + o_p(n^{-4/5}).$$

Note that

$$\widehat{MMSE}_n(\tilde{h}) = n^{-4/5} \left\{ \frac{\mu_2}{2} \left[\hat{f}^{(2)}(x_1)\tilde{C}_1^2 - \hat{f}^{(2)}(x_2)\tilde{C}_2^2 \right] \right\}^2 + \frac{\nu_0}{n^{4/5}} \left\{ \frac{\hat{f}(x_1)}{\tilde{C}_1} + \frac{\hat{f}(x_2)}{\tilde{C}_2} \right\} + O_P(n^{-8/5}).$$

Since $\hat{h}$ is the optimizer, $\widehat{MMSE}_n(\hat{h})/\widehat{MMSE}_n(\tilde{h}) \leq 1$. Thus

$$\frac{\left\{\frac{\mu_2}{2}\left[\hat{f}^{(2)}(x_1)\hat{C}_1^2 - \hat{f}^{(2)}(x_2)\hat{C}_2^2\right]\right\}^2 + \nu_0\left\{\frac{\hat{f}(x_1)}{\hat{C}_1} + \frac{\hat{f}(x_2)}{\hat{C}_2}\right\} + o_p(1)}{\left\{\frac{\mu_2}{2}\left[\hat{f}^{(2)}(x_1)\tilde{C}_1^2 - \hat{f}^{(2)}(x_2)\tilde{C}_2^2\right]\right\}^2 + \nu_0\left\{\frac{\hat{f}(x_1)}{\tilde{C}_1} + \frac{\hat{f}(x_2)}{\tilde{C}_2}\right\} + O_P(n^{-4/5})} \leq 1.$$

Since the denominator converges to

$$\left\{\frac{\mu_2}{2}\left[f^{(2)}(x_1)C_1^{*2} - f^{(2)}(x_2)C_2^{*2}\right]\right\}^2 + \nu_0\left\{\frac{f(x_1)}{C_1^*} + \frac{f(x_2)}{C_2^*}\right\},$$

where C_1^* and C_2^* are optimizers of

$$\left\{\frac{\mu_2}{2}\left[f^{(2)}(x_1)C_1^2 - f^{(2)}(x_2)C_2^2\right]\right\}^2 + \nu_0\left\{\frac{f(x_1)}{C_1} + \frac{f(x_2)}{C_2}\right\}$$

with respect to C_1 and C_2. This implies that $\hat{C}_1$ and $\hat{C}_2$ also converge to the same respective limit C_1^* and C_2^* because the inequality will be violated otherwise.

Next we consider the case in which $f^{(2)}(x_1)f^{(2)}(x_2) > 0$. In this case, with probability approaching 1, $\hat{f}^{(2)}(x_1)\hat{f}^{(2)}(x_2) > 0$, so that we assume this without loss of generality.

When these conditions hold, let $h_2 = \hat{\lambda}_2 h_1$ where $\hat{\lambda}_2$ is defined in (7). This sets the first bias term of $\widehat{MMSE}_n(h)$ equal to zero. Define the plug-in versions of $AMSE_{2n}(h)$ by

$$\widehat{AMSE}_{2n}(h) = \left\{\frac{\mu_4}{4!}\left[\hat{f}^{(4)}(x_1)h_1^4 - \hat{f}^{(4)}(x_2)h_2^4\right]\right\}^2 + \frac{\nu_0}{n}\left\{\frac{\hat{f}(x_1)}{h_1} + \frac{\hat{f}(x_2)}{h_2}\right\}.$$

Choosing h_1 to minimize $\widehat{AMSE}_{2n}(h)$, we define $\tilde{h}_1 = \hat{\theta}_2 n^{-1/9} \equiv \tilde{C}_1 n^{-1/9}$ and $\tilde{h}_2 = \hat{\lambda}_2\tilde{h}_1 \equiv \tilde{C}_2 n^{-1/9}$ where $\hat{\theta}_2$ is defined in (7). Then $\widehat{MMSE}_n(\tilde{h})$ can be written as

$$\widehat{MMSE}_n(\tilde{h}) = n^{-8/9}\left\{\frac{\mu_4}{4!}\left[\hat{f}^{(4)}(x_1)\tilde{C}_1^4 - \hat{f}^{(4)}(x_2)\tilde{C}_2^4\right]\right\}^2 + \nu_0 n^{-8/9}\left\{\frac{\hat{f}(x_1)}{\tilde{C}_1} + \frac{\hat{f}(x_2)}{\tilde{C}_2}\right\}.$$

In order to match this rate of convergence, both $\hat{h}_1$ and $\hat{h}_2$ need to converge at the rate slower than or equal to $n^{-1/9}$ because the variance term needs to converge at

the rate $n^{-8/9}$ or faster. In order for the first-order bias term to match this rate,

$$\hat{f}^{(2)}(x_1)\hat{h}_1^2 - \hat{f}^{(2)}(x_2)\hat{h}_2^2 \equiv B_{1n} = n^{-4/9}b_{1n},$$

where $b_{1n} = O_P(1)$. Under the assumption that $f^{(2)}(x_2) \neq 0$, $\hat{f}^{(2)}(x_2)$ is bounded away from 0, with probability approaching 1. Assuming this without loss of generality, we have $\hat{h}_2^2 = \hat{\lambda}_2^2\hat{h}_1^2 - B_{1n}/\hat{f}^{(2)}(x_2)$. Then, it follows that

$$\widehat{MMSE}_n(\hat{h}) = \left\{\frac{\mu_2}{2}B_{1n}\right\}^2 + \left\{\frac{\mu_4}{4!}\left[\hat{f}^{(4)}(x_1)\hat{h}_1^4 - \hat{f}^{(4)}(x_2)\{\hat{\lambda}_2^2\hat{h}_1^2 - B_{1n}/\hat{f}^{(2)}(x_2)\}^2\right]\right\}^2 + \frac{\nu_0}{n}\left\{\frac{\hat{f}(x_1)}{\hat{h}_1} + \frac{\hat{f}(x_2)}{\{\hat{\lambda}_2^2\hat{h}_1^2 - B_{1n}/\hat{f}^{(2)}(x_2)\}^{1/2}}\right\}.$$

Suppose $\hat{h}_1$ is of order slower than $n^{-1/9}$. Then because $\hat{f}^{(4)}(x_1)[\hat{f}^{(2)}(x_2)]^2 - \hat{f}^{(4)}(x_2)[\hat{f}^{(2)}(x_1)]^2 \neq 0$ and this holds even in the limit, the second-order bias term is of order slower than $n^{-8/9}$. This contradicts the definition of $\hat{h}_1$, implying that $\hat{h}_1$ is of order $n^{-1/9}$. Therefore we can write $\hat{h}_1 = \hat{C}_1 n^{-1/9} + o_p(n^{-1/9})$ for some $O_P(1)$ sequence $\hat{C}_1$ that is bounded away from 0 as $n \to \infty$ and as before $\hat{h}_2^2 = \hat{\lambda}_2^2\hat{h}_1^2 - B_{1n}/\hat{f}^{(2)}(x_2)$. Using this expression, we can write

$$\begin{aligned}\widehat{MMSE}_n(\hat{h}) &= n^{-8/9}\left\{\frac{\mu_2}{2}b_{1n}\right\}^2 \\ &+ n^{-8/9}\left\{\frac{\mu_4}{4!}\left[\hat{f}^{(4)}(x_1)\hat{C}_1^4 + o_p(1) - \hat{f}^{(4)}(x_2)\{\hat{\lambda}_2^2\hat{C}_1^2 + o_p(1) - n^{-2/9}b_{1n}/\hat{f}^{(2)}(x_2)\}^2\right]\right\}^2 \\ &+ \nu_0 n^{-8/9}\left\{\frac{\hat{f}(x_1)}{\hat{C}_1 + o_p(1)} + \frac{\hat{f}(x_2)}{\{\hat{\lambda}_2^4\hat{C}_1^2 + o_p(1) - n^{-2/9}b_{1n}/\hat{f}^{(2)}(x_2)\}^{1/2}}\right\}.\end{aligned}$$

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, $\widehat{MMSE}_n(\hat{h})/\widehat{MMSE}_n(\tilde{h}) \leq 1$. Thus

$$\frac{o_p(1) + \left\{\frac{\mu_4}{4!}\left[\hat{f}^{(4)}(x_1)\hat{C}_1^4 - \hat{f}^{(4)}(x_2)\hat{\lambda}_2^4\hat{C}_1^2 + o_p(1)\}\right]\right\}^2 + \nu_0\left\{\frac{\hat{f}(x_1)}{\hat{C}_1+o_p(1)} + \frac{\hat{f}(x_2)}{\{\hat{\lambda}_2^4\hat{C}_1^2+o_p(1)\}^{1/2}}\right\}}{\left\{\frac{\mu_4}{4!}\left[\hat{f}^{(4)}(x_1)\tilde{C}_1^4 - \hat{f}^{(4)}(x_2)\tilde{C}_2^4\right]\right\}^2 + \nu_0\left\{\frac{\hat{f}(x_1)}{\tilde{C}_1} + \frac{\hat{f}(x_2)}{\tilde{C}_2}\right\}} \leq 1.$$

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 $\hat{h}_2/\tilde{h}_2$ converges in probability to 1 as well.

The result above also show that $\widehat{MMSE}_n(\hat{h})/MSE_n(h^*)$ converges to 1 in probability in both cases. ■

Proof of Lemma 3: A contribution to the MSE from a variance component is standard. See Fan and Gijbels (1996) for the details. Here, we derive the contribution made by the bias component. Denote $\hat{\gamma} = \left(\hat{\alpha}_h(x),\ \hat{\beta}_h(x)\right)'$. The conditional bias is given by

$$\text{Bias}(\hat{\gamma}|X) = (X(x)'W(x)X(x))^{-1}X(x)W(x)(m - X(x)\gamma),$$

where $m = (m(X_1), \ldots, m(X_n))'$ and $\gamma = (m(x), m^{(1)}(x))'$. Let $s_{n,k} = \sum_{i=1}^n K_h(X_i - x)(X_i - x)^k$. We use the following notation:

$$S_{n,k} = \begin{bmatrix} s_{n,k} & s_{n,k+1} \\ s_{n,k+1} & s_{n,k+2} \end{bmatrix}, \quad S_k = \begin{bmatrix} \mu_k & \mu_{k+1} \\ \mu_{k+1} & \mu_{k+2} \end{bmatrix}, \quad c_{n,k} = \begin{bmatrix} s_{n,k} \\ s_{n,k+1} \end{bmatrix}, \quad c_k = \begin{bmatrix} \mu_k \\ \mu_{k+1} \end{bmatrix}. \tag{14}$$

Note that $S_{n,0} = X(x)'W(x)X(x)$. The argument made by Fan, Gijbels, Hu, and Huang (1996) can be generalized to yield

$$s_{n,k} = nh^k \left\{ f(x)\mu_k + hf^{(1)}(x)\mu_{k+1} + \frac{h^2 f^{(2)}(x)}{2}\mu_{k+2} + o_p\left(h^2\right) \right\}. \tag{15}$$

Then, it follows that

$$S_{n,0} = nH \left\{ f(x)S_0 + hf^{(1)}(x)S_1 + \frac{h^2 f^{(2)}(x)}{2}S_2 + o_p\left(h^2\right) \right\} H,$$

where $H = \text{diag}(1, h)$. By using the fact that

$$(A + hB + h^2C)^{-1} = A^{-1} - hA^{-1}BA^{-1} - h^2A^{-1}CA^{-1} + h^2A^{-1}BA^{-1}BA^{-1} + o\left(h^2\right),$$

we obtain

$$S_{n,0}^{-1} = n^{-1}H^{-1}\left\{\frac{1}{f(x)}A_0 - \frac{hf^{(1)}(x)}{f(x)^2}A_1 - \frac{h^2f^{(2)}(x)}{2f(x)^2}A_2 + \frac{h^2f^{(1)}(x)^2}{f(x)^3}A_3 + o_p\left(h^2\right)\right\}H^{-1}, \tag{16}$$

where

$$A_0 = \begin{bmatrix} 1 & 0 \\ 0 & \mu_2^{-1} \end{bmatrix}, \quad A_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} \mu_2 & 0 \\ 0 & \mu_4/\mu_2^2 \end{bmatrix}, \quad A_3 = \begin{bmatrix} \mu_2 & 0 \\ 0 & 1 \end{bmatrix}.$$

This matrix structure is simplified considerably by using a symmetric kernel function.

Next, we consider $X(x)W(x)(m - X(x)\beta)$. A Taylor expansion of $m(\cdot)$ yields

$$X(x)W(x)(m - X(x)\beta) = \frac{m^{(2)}(x)}{2}c_{n,2} + \frac{m^{(3)}(x)}{3!}c_{n,3} + \frac{m^{(4)}(x)}{4!}c_{n,4} + o_p\left(nh^4\right). \tag{17}$$

The definition of $c_{n,j}$ in (14), in conjunction with (15), yields

$$c_{n,k} = nh^kH\left\{f(x)c_k + hf^{(1)}(x)c_{k+1} + \frac{h^2f^{(2)}(x)}{2}c_{k+2} + o_p\left(h^2\right)\right\}.$$

Combining this with (16) and (17) and extracting the first element gives

$$\text{Bias}(\hat{\alpha}_h(x)|X) = \frac{h^2m^{(2)}(x)}{2}\mu_2 + \frac{h^4}{4}\left\{\frac{m^{(2)}(x)}{f(x)^2}(\mu_4 - \mu_2)\left(f^{(2)}(x)f(x) - f^{(1)}(x)^2\right) + \frac{m^{(4)}(x)}{3!}\mu_4\right\}.$$

This expression gives the required result. ■

Proof of Lemma 4: Again, we consider the contribution made by the bias component because that of the variance component is standard. We present the proof only for $\hat{\alpha}_{h,1}(x)$. The proof for $\hat{\alpha}_{h,2}$ is parallel and hence is omitted. Denote $\hat{\gamma}_1 = \left(\hat{\alpha}_{h,1}(x),\ \hat{\beta}_{h,1}(x)\right)'$. The conditional bias is given by

$$\text{Bias}(\hat{\gamma}_1|X) = (X(x)'W_1(x)X(x))^{-1}X(x)W_1(x)(m_1 - X(x)\gamma_1),$$

where $m_1 = (m_1(X_1), \ldots, m_1(X_n))'$ and $\gamma_1 = (m_1(x), m_1^{(1)}(x))'$. Note that $S_{n,0,1} = X(x)'W_1(x)X(x)$. The argument made by Fan, Gijbels, Hu, and Huang (1996) can

be generalized to yield

$$s_{n,k,1} = nh^k \left\{ f(x)\mu_{k,0} + hf^{(1)}(x)\mu_{k+1,0} + o_p(h) \right\}. \tag{18}$$

Then, it follows that

$$S_{n,0,1} = nH \left\{ f(x)S_{0,1} + hf^{(1)}(x)S_{1,1} + o_p(h) \right\} H,$$

where $H = \text{diag}(1, h)$. By using the fact that $(A+hB)^{-1} = A^{-1} - hA^{-1}BA^{-1} + o(h)$, we obtain

$$S_{n,0,1}^{-1} = n^{-1}H^{-1} \left\{ \frac{1}{f(x)} A_{0,1} - \frac{hf^{(1)}(x)}{f(x)^2} A_{1,1} + o_p(h) \right\} H^{-1}, \tag{19}$$

where

$$A_{0,1} = \begin{bmatrix} \mu_{2,0} & -\mu_{1,0} \\ -\mu_{1,0} & \mu_{0,0}^{-1} \end{bmatrix},$$

$$A_{1,1} = \frac{1}{\mu_{0,0}\mu_{2,0} - \mu_{1,0}^2} \begin{bmatrix} -\mu_{1,0}(\mu_{2,0}^2 - \mu_{1,0}\mu_{3,0}) & \mu_{2,0}(\mu_{2,0}^2 - \mu_{1,0}\mu_{3,0}) \\ \mu_{2,0}(\mu_{2,0}^2 - \mu_{1,0}\mu_{3,0}) & \mu_{1,0}^3 - 2\mu_{0,0}\mu_{1,0}\mu_{2,0} + \mu_{0,0}^2\mu_{3,0} \end{bmatrix}.$$

Next, we consider $X(x)W_1(x)(m_1 - X(x)\gamma_1)$. A Taylor expansion of $m_1(\cdot)$ yields

$$X(x)W_1(x)(m_1 - X(x)\gamma_1) = \frac{m_1^{(2)}(x)}{2} c_{n,2,1} + \frac{m_1^{(3)}(x)}{3!} c_{n,3,1} + o_p\left(nh^3\right). \tag{20}$$

The definition of $c_{n,k,j}$ in (12), in conjunction with (18), yields

$$c_{n,k,1} = nh^k H \left\{ f(x)c_{k,1} + hf^{(1)}(x)c_{k+1,1} + o_p(h) \right\}. \tag{21}$$

Combining this with (19) and (20) and extracting the first element gives

$$\text{Bias}(\hat{\alpha}_{h,1}(x)|X) = \frac{h^2 b_1 m_1^{(2)}(x)}{2} + b_{2,1}(x)h_1^3.$$

This expression gives the required result. ■

Proof of Corollary 1: Observe that equations (18) and (21) imply

$$e_1'\tilde{S}_{n,0,j}^{-1}\tilde{c}_{n,2,j} \to b_1, \quad e_1'\tilde{S}_{n,0,j}^{-1}c_{n,3,j} \to (-1)^{j+1}c_1,$$
$$e_1'\tilde{S}_{n,0,j}^{-1}S_{n,1,j}\tilde{S}_{n,0,j}^{-1}\tilde{c}_{n,2,j} \to (-1)^{j+1}c_2 \quad \text{and} \quad e_1'S_{n,0,j}^{-1}T_{n,0,j}S_{n,0,j}^{-1}e_1 \to v$$

in probability uniformly. With these properties, each step of the proof of Theorem 3 is valid even if $\widehat{MMSE}_n(h)$ is replaced by $\widehat{MMSE}^E{}_n(h)$, thus completing the proof of Corollary 1. ■

References

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," mimeo.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2012): "Robust nonparametric bias-corrected inference in the regression discontinuity design," Mimeo.

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," Mimeo.

Eicker, F. (1967): "Limit theorems for regressions with unequal and dependnet errors," in *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics and Probability*, vol. 1, pp. 59–82. University of California Press.

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

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

——— (1995): "Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaption," *Journal of the Royal Statistical Society, Series B*, 57, 371–394.

——— (1996): *Local polynomial modeling and its applications*. Chapman & Hall.

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

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.

Hall, P. (1983): "Large sample optimality of least squares cross-validation in density estimation," *Annals of Statistics*, 11, 1156–1174.

Huber, P. J. (1967): "The behavior of maximum likelihood estimates under nonstandard conditions," in *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics and Probability*, vol. 1, pp. 221–233. University of California Press.

Imbens, G. W., and K. Kalyanaraman (2012): "Optimal bandwidth choice for the regression discontinuity estimator," *Review of Economic Studies*, 79, 933–959.

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

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.

——— (2007): "Does head start improve children's life changes? Evidence from a regression discontinuity design," *Quarterly Journal of Economics*, 122, 159–208.

Prakasa Rao, B. L. S. (1983): *Nonparametric Functional Estimation*. Academic Press, Orland, Florida.

Silverman, B. W. (1986): *Density estimation for statistics and data analyasis*. Chapman & Hall, London.

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

Wand, M. P., and M. C. Jones (1994): *Kernel Smoothing*. Chapman & Hall.

White, H. (1980): "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, 48(4), 817–838.

Supplement to "Optimal Bandwidth Selection for Differences of Nonparametric Estimators with an Application to the Sharp Regression Discontinuity Design"

Yoichi Arai and Hidehiko Ichimura

A Introduction

In this supplemental material, we present omitted discussions, an algorithm to implement the proposed method for the sharp RDD and proofs for the main results.

B Uniqueness of the AFO Bandwidths for the Difference of Densities

In this section, we verify the uniqueness of the AFO bandwidths for the difference of densities.

(i) When $f^{(2)}(x_1)f^{(2)}(x_2) < 0$, the first-order conditions are given by

$$\left.\frac{\partial AMSE_{1n}(h)}{\partial h_1}\right|_{h_1=h_1^*,h_2=h_2^*} = \mu_2^2 f^{(2)}(x_1)h_1^*\left[f^{(2)}(x_1)h_1^{*2} - f^{(2)}(x_2)h_2^{*2}\right] - \frac{\nu_0}{n}\frac{f(x_1)}{h_1^{*2}} = 0,$$
$$\left.\frac{\partial AMSE_{1n}(h)}{\partial h_2}\right|_{h_1=h_1^*,h_2=h_2^*} = -\mu_2^2 f^{(2)}(x_2)h_2^*\left[f^{(2)}(x_1)h_1^{*2} - f^{(2)}(x_2)h_2^{*2}\right] - \frac{\nu_0}{n}\frac{f(x_2)}{h_2^{*2}} = 0.$$

Solving these gives the explicit forms of h_1^* and h_2^*.

To show that h_1^* and h_2^* are global minimizers, it is sufficient to show that $AMSE_{1n}(h)$ is strictly convex with respect to h_1 and h_2. For strict convexity, we

must show that the Hessian matrix is positive definite; i.e. that

$$\frac{\partial^2 AMSE_{1n}(h)}{\partial h_1^2} > 0, \quad \frac{\partial^2 AMSE_{1n}(h)}{\partial h_1^2} \cdot \frac{\partial^2 AMSE_{1n}(h)}{\partial h_2^2} - \left[\frac{\partial^2 AMSE_{1n}(h)}{\partial h_1 \partial h_2}\right]^2 > 0.$$

Given that $f^{(2)}(x_1)$ and $f^{(2)}(x_2)$ have different signs, it follows that

$$\frac{\partial^2 AMSE_{1n}(h)}{\partial h_1^2} = \mu_2^2 f^{(2)}(x_1)\left[f^{(2)}(x_1)h_1^2 - f^{(2)}(x_2)h_2^2\right] + 2\left[\mu_2 f^{(2)}(x_1)h_1\right]^2 + \frac{2\nu_0 f(x_1)}{nh_1^3} > 0,$$

because $f(\cdot)$, μ_2, ν_0, n, h_1 and h_2 are all positive. We can also show that

$$\begin{aligned}
&\frac{\partial^2 AMSE_{1n}(h)}{\partial h_1^2} \cdot \frac{\partial^2 AMSE_{1n}(h)}{\partial h_2^2} - \left[\frac{\partial^2 AMSE_{1n}(h)}{\partial h_1 \partial h_2}\right]^2 \\
&= \left\{\mu_2^2 f^{(2)}(x_1)\left[f^{(2)}(x_1)h_1^2 - f^{(2)}(x_2)h_2^2\right] + 2\left[\mu_2 f^{(2)}(x_1)h_1\right]^2 + \frac{2\nu_0 f(x_1)}{nh_1^3}\right\} \\
&\quad \times \left\{-\mu_2^2 f^{(2)}(x_2)\left[f^{(2)}(x_1)h_1^2 - f^{(2)}(x_2)h_2^2\right] + 2\left[\mu_2 f^{(2)}(x_2)h_2\right]^2 + \frac{2\nu_0 f(x_2)}{nh_2^3}\right\} \\
&\quad - \left[2\mu_2^2 f^{(2)}(x_1)f^{(2)}(x_2)h_1 h_2\right]^2.
\end{aligned}$$

Note that if we ignore the first and third terms in the two brackets of the first term on the right-hand side, what is left coincides with the last term on the right-hand side. However, both the first and third terms are positive as discussed earlier. Thus, the difference of the two terms are positive.

(ii) Next, we consider the case where $f^{(2)}(x_1)f^{(2)}(x_2) > 0$. With the restriction that $f^{(2)}(x_1)h_1^2 - f^{(2)}(x_2)h_2^2 = 0$, $AMSE_{2n}(h)$ can be written as

$$AMSE_{2n}(h) = \left\{\frac{\mu_4}{4!}\left[f^{(4)}(x_1) - \lambda^{**4} f^{(4)}(x_2)\right] h_1^4\right\}^2 + \frac{\nu_0}{nh_1}\left\{f(x_1) + \frac{f(x_2)}{\lambda^{**}}\right\}.$$

The first-order condition becomes

$$\left.\frac{dAMSE_{2n}(h)}{dh_1}\right|_{h_1 = h_1^{**}} = \frac{1}{72}\mu_4^2\left\{f^{(4)}(x_1) - \lambda^{**4} f^{(4)}(x_2)\right\}^2 h_1^{**7} - \frac{\nu_0}{nh_1^{**2}}\left\{f(x_1) + \frac{f(x_2)}{\lambda^{**}}\right\} = 0.$$

Solving this with respect to h_1^{**} yields the expression of Definition 1. To see that

$AMSE_{2n}(h_1)$ has a unique minimum, observe that

$$\frac{d^2 AMSE_{2n}(h)}{dh_1^2} = \frac{7}{56}\mu_4^2 \left\{ f^{(4)}(x_1) - \lambda^{**4} f^{(4)}(x_2) \right\}^2 h_1^6 + \frac{2\nu_0}{h_1^3} \left\{ f(x_1) + \frac{f(x_2)}{\lambda^{**}} \right\}.$$

Both terms on the right-hand side being positive proves strict convexity. ■

C Implementation for the Sharp RDD

In this section, we provide a detailed procedure to implement the proposed method for the sharp RDD. To obtain the proposed bandwidths, we need pilot estimates of the density and its first derivative, the second and third derivatives of the regression functions, and the conditional variances at the discontinuity point. We obtain these pilot estimates in a number of steps. Before we describe them, note that the discontinuity points for all designs are at $x = 0$. When the discontinuity point is at $x = c$ rather than $x = 0$, one proceeds by replacing X_i with $X_i - c$ in the following steps.

C.1 Step 1: Obtain pilot estimates for the density $f(0)$ and its first derivative $f^{(1)}(0)$

We calculate the density of the forcing variable at the discontinuity point $f(0)$, which is estimated by using the kernel density estimator with an Epanechnikov kernel.[1] A pilot bandwidth for kernel density estimation is chosen by using the normal scale rule, given by $\hat{\sigma} \cdot (15\phi(c)/(n\phi^{(2)}(c)^2))^{1/5}$ evaluated at $c = 0$, where $\hat{\sigma}$ is the square root of the sample variance of X_i and $\phi(\cdot)$ is the normal density (see Wand and Jones, 1994 for the normal scale rules). The first derivative of the density is estimated by using the method proposed by Jones (1994). The kernel first derivative density estimator is given by $\sum_{i=1}^{n} L((c - X_i)/h)/(nh^2)$, where L is the kernel function proposed by Jones (1994), $L(u) = -15u(1-u^2)1_{\{|u|<1\}}/4$ and c is the evaluation point (zero in our experiments). Again, a pilot bandwidth is obtained by using the normal scale rule,

[1] IK estimated the density in a simpler manner (see Section 4.2 of IK). We used the kernel density estimator to be consistent with the estimation method used for the first derivative. Our unreported simulation experiments produced similar results for both methods.

given by $\hat{\sigma} \cdot (105\phi(c/\hat{\sigma})/(n\phi^{(3)}(c/\hat{\sigma}))^{1/7}$ evaluated at $c = 0.1$.[2]

C.2 Step 2: Obtain pilot bandwidths for estimating the second and third derivatives $m_j^{(2)}(0)$ and $m_j^{(3)}(0)$ for $j = 1, 2$

We next estimate the second and third derivatives by using the third-order LPR. We obtain pilot bandwidths for the LPR based on the estimated fourth derivatives $m_1^{(4)}(0) = \lim_{x \to 0+} m^{(4)}(x)$ and $m_2^{(4)}(0) = \lim_{x \to 0-} m^{(4)}(x)$. Following IK, we use estimates that are not necessarily consistent by fitting global polynomial regressions. In doing so, we construct a matrix whose ith row is given by $[1 \ X_i \ X_i^2 \ X_i^3 \ X_i^4]$. It turns out that the matrix has an average condition number (the ratio of the largest eigenvalue to the smallest.) of 28.80. This number suggests potential multicollinearity, which typically makes the polynomial regression estimates very unstable. Hence, we use the ridge regression proposed by Hoerl, Kennard, and Baldwin (1975). This is implemented in two steps. First, using observations for which $X_i \geq 0$, we regress Y_i on 1, X_i, X_i^2, X_i^3 and X_i^4 to obtain the standard OLS coefficients $\hat{\gamma}_1$ and the variance estimate $\hat{s}_1^2$. This yields the ridge coefficient proposed by Hoerl, Kennard, and Baldwin (1975): $r_1 = (5\hat{s}_1^2)/(\hat{\gamma}_1'\hat{\gamma}_1)$. Using the data with $X_i < 0$, we repeat the procedure to obtain the ridge coefficient, r_2. Let Y be a vector of Y_i, and let X be the matrix whose ith row is given by $[1 \ X_i \ X_i^2 \ X_i^3 \ X_i^4]$ for observations with $X_i \geq 0$, and let I_k be the $k \times k$ identity matrix. The ridge estimator is given by $\hat{\beta}_{r1} = (X'X + r_1 I_5)^{-1} X'Y$, and $\hat{\beta}_{r2}$ is obtained in the same manner. The estimated fourth derivatives are $\hat{m}_1^{(4)}(0) = 24 \cdot \hat{\beta}_{r1}(5)$ and $\hat{m}_2^{(4)}(0) = 24 \cdot \hat{\beta}_{r2}(5)$, where $\hat{\beta}_{r1}(5)$ and $\hat{\beta}_{r2}(5)$ are the fifth elements of $\hat{\beta}_{r1}$ and $\hat{\beta}_{r2}$, respectively. The estimated conditional variance is $\sigma_{r1}^2 = \sum_{i=1}^{n_1} (Y_i - \hat{Y}_i)^2/(n_1 - 5)$, where $\hat{Y}_i$ denotes the fitted values, n_1 is the number of observations for which $X_i \geq 0$, and the summation is over i with $X_i \geq 0$. σ_{r2}^2 is obtained analogously. The plug-in bandwidths for the third-order LPR used to

[2] The normal scale rules do not work when the evaluation point is zero because the third derivative of the normal density at zero is equal to zero. Hence, we use $c = 0.1$. The following results are robust to the value of c, unless c differs greatly from zero.

estimate the second and third derivatives are calculated by

$$h_{\nu,j} = C_{\nu,3}(K) \left(\frac{\sigma_{rj}^2}{\hat{f}(0) \cdot \hat{m}_j^{(4)}(0) \cdot n_j} \right)^{1/9},$$

where $j = 1, 2$ (see Fan and Gijbels, 1996, Section 3.2.3 for information on plug-in bandwidths and the definition of $C_{\nu,3}$). We use $\nu = 2$ and $\nu = 3$ for estimating the second and third derivatives, respectively.

C.3 Step 3: Estimation of the second and third derivatives $m_j^{(2)}(0)$ and $m_j^{(3)}(0)$ as well as the conditional variances $\hat{\sigma}_j^2(0)$ for $j = 1, 2$

We estimate the second and third derivatives at the threshold by using the third-order LPR with the pilot bandwidths obtained in Step 2. Following IK, we use the uniform kernel, which yields constant values of $C_{2,3} = 5.2088$ and $C_{3,3} = 4.8227$. To estimate $\hat{m}_1^{(2)}(0)$, we construct a vector $Y_a = (Y_1, \ldots, Y_{n_a})'$ and an $n_a \times 4$ matrix, X_a, whose ith row is given by $[1 \; X_i \; X_i^2 \; X_i^3]$ for observations with $0 \leq X_i \leq h_{2,3}$, where n_a is the number of observations with $0 \leq X_i \leq h_{2,3}$. The estimated second derivative is given by $\hat{m}_1^{(2)}(0) = 2 \cdot \hat{\beta}_{2,1}(3)$, where $\hat{\beta}_{2,1}(3)$ is the third element of $\hat{\beta}_{2,1}$ and $\hat{\beta}_{2,1} = ({X_a}'X_a)^{-1}X_aY_a$. We estimate $\hat{m}_2^{(2)}(0)$ in the same manner. Replacing $h_{2,3}$ with $h_{3,3}$ leads to an estimated third derivative of $\hat{m}_1^{(3)}(0) = 6 \cdot \hat{\beta}_{3,1}(4)$, where $\hat{\beta}_{3,1}(4)$ is the fourth element of $\hat{\beta}_{3,1}$, $\hat{\beta}_{3,1} = ({X_b}'X_b)^{-1}X_bY_b$, $Y_b = (Y_1, \ldots, Y_{n_b})'$, X_b is an $n_b \times 4$ matrix whose ith row is given by $[1 \; X_i \; X_i^2 \; X_i^3]$ for observations with $0 \leq X_i \leq h_{3,3}$, and n_b is the number of observations with $0 \leq X_i \leq h_{3,3}$. The conditional variance at the threshold $\sigma_1^2(0)$ is calculated as $\hat{\sigma}_1(0) = \sum_{i=1}^{n_2}(Y_i - \hat{Y}_i)^2/(n-4)$, where $\hat{Y}_i$ denotes the fitted values from the regression used to estimate the second derivative.[3] $\hat{\beta}_{2,2}$ and $\hat{\beta}_{3,2}$ can be obtained analogously.

[3] One can use the fitted values from the regression used to estimate the third derivatives, having replaced n_a with n_b. However, because these values produce simulation results that are almost identical to those produced by the fitted values described in the main text, we present the latter.

C.4 Step 4

The final step is to plug the pilot estimates into the MMSE given by (10) and to use numerical minimization over the compact region to obtain $\hat{h}_1$ and $\hat{h}_2$. Unlike $AMSE_{1n}(h)$ and $AMSE_{2n}(h)$ subject to the restriction given in Definition 3, the MMSE is not necessarily strictly convex, particularly when the sign of the product is positive. In conducting numerical optimization, it is important to try optimization with several initial values, so as to avoid finding only a local minimum. $(\hat{h}_1^E, \hat{h}_2^E)$ and $(\hat{h}_1^R, \hat{h}_2^R)$ can be computed using the MMSE given by (11) and (13), respectively.

D Proof of Theorem 3

Recall that the objective function is:

$$\widehat{MMSE}_n(h) = \left\{ \frac{b_1}{2} \left[\hat{m}_1^{(2)}(x) h_1^2 - \hat{m}_2^{(2)}(x) h_2^2 \right] \right\}^2 + \left[\hat{b}_{2,1}(x) h_1^3 - \hat{b}_{2,2}(x) h_2^3 \right]^2 + \frac{\nu}{n\hat{f}(x)} \left\{ \frac{\hat{\sigma}_1^2(x)}{h_1} + \frac{\hat{\sigma}_2^2(x)}{h_2} \right\}.$$

To begin with, we show that $\hat{h}_1$ and $\hat{h}_2$ satisfy Assumption 3. If we choose a sequence of h_1 and h_2 to satisfy Assumption 3, then $\widehat{MMSE}_n(h)$ converges to 0. Assume to the contrary that either one or both of $\hat{h}_1$ and $\hat{h}_2$ do not satisfy Assumption 3. Since $m_2^{(2)}(x)^3 b_{2,1}(x)^2 \neq m_1^{(2)}(x)^3 b_{2,2}(x)^2$ by assumption, $\hat{m}_2^{(2)}(x)^3 \hat{b}_{2,1}(x)^2 \neq \hat{m}_1^{(2)}(x)^3 \hat{b}_{2,2}(x)^2$ with probability approaching 1. Without loss of generality, we assume this as well. Then at least one of the first-order bias term, the second-order bias term and the variance term of $\widehat{MMSE}_n(\hat{h})$ does not converge to zero in probability. Then $\widehat{MMSE}_n(\hat{h}) > \widehat{MMSE}_n(h)$ holds for some n. This contradicts the definition of $\hat{h}$. Hence $\hat{h}$ satisfies Assumption 3.

We first consider the case in which $m_1^{(2)}(x) m_2^{(2)}(x) < 0$. In this case, with probability approaching 1, $\hat{m}_1^{(2)}(x) \hat{m}_2^{(2)}(x) < 0$, 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 $\widehat{MMSE}_n(\hat{h})$ since $\hat{h}_1$ and $\hat{h}_2$ satisfy Assumption 3. Define the plug-in

version of $AMSE_{1n}(h)$ provided in Definition 3 by

$$\widehat{AMSE}_{1n}(h) = \left\{ \frac{b_1}{2} \left[\hat{m}_1^{(2)}(x) h_1^2 - \hat{m}_2^{(2)}(x) h_2^2 \right] \right\}^2 + \frac{\nu}{n \hat{f}(x)} \left\{ \frac{\hat{\sigma}_1^2(x)}{h_1} + \frac{\hat{\sigma}_2^2(x)}{h_2} \right\}.$$

Let the minimizer of $\widehat{AMSE}_{1n}(h)$ by $\tilde{h}_1$ and $\tilde{h}_2$. Also define

$$\hat{\theta}_1 = \left\{ \frac{v \hat{\sigma}_1^2(x)}{\hat{b}_1^2 \hat{f}(x) \hat{m}_1^{(2)}(x) \left[\hat{m}_1^{(2)}(x) - \hat{\lambda}_1^2 \hat{m}_2^{(2)}(x) \right]} \right\}^{1/5} \quad \text{and} \quad \hat{\lambda}_1 = \left\{ -\frac{\hat{\sigma}_2^2(x) \hat{m}_1^{(2)}(x)}{\hat{\sigma}_1^2(x) \hat{m}_2^{(2)}(x)} \right\}^{1/3}.$$

A calculation yields $\tilde{h}_1 = \hat{\theta}_1 n^{-1/5} \equiv \tilde{C}_1 n^{-1/5}$ and $\tilde{h}_2 = \hat{\theta}_1 \hat{\lambda}_1 n^{-1/5} \equiv \tilde{C}_2 n^{-1/5}$. With this choice, $\widehat{AMSE}_{1n}(\tilde{h})$ and hence $\widehat{MMSE}_n(\tilde{h})$ converges at the rate of $n^{-4/5}$. Note that if $\hat{h}_1$ or $\hat{h}_2$ 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 $\hat{h}_2$ 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 $\widehat{MMSE}_n(h)$, $\hat{h}_1$ and $\hat{h}_2$ 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 $\hat{h}_2 = \hat{C}_2 n^{-1/5} + o_p(n^{-1/5})$ for some $O_P(1)$ sequences $\hat{C}_1$ and $\hat{C}_2$ that are bounded away from 0 and ∞ as $n \to \infty$. Using this expression,

$$\begin{aligned} \widehat{MMSE}_n(\hat{h}) &= n^{-4/5} \left\{ \frac{b_1}{2} \left[\hat{m}_1^{(2)}(x) \hat{C}_1^2 - \hat{m}_2^{(2)}(x) \hat{C}_2^2 \right] \right\}^2 \\ &\quad + \frac{\nu}{n^{4/5} \hat{f}(x)} \left\{ \frac{\hat{\sigma}_1^2(x)}{\hat{C}_1} + \frac{\hat{\sigma}_2^2(x)}{\hat{C}_2} \right\} + o_p(n^{-4/5}). \end{aligned}$$

Note that

$$\begin{aligned} \widehat{MMSE}_n(\tilde{h}) &= n^{-4/5} \left\{ \frac{b_1}{2} \left[\hat{m}_1^{(2)}(x) \tilde{C}_1^2 - \hat{m}_2^{(2)}(x) \tilde{C}_2^2 \right] \right\}^2 \\ &\quad + \frac{\nu}{n^{4/5} \hat{f}(x)} \left\{ \frac{\hat{\sigma}_1^2(x)}{\tilde{C}_1} + \frac{\hat{\sigma}_2^2(x)}{\tilde{C}_2} \right\} + O_P(n^{-8/5}). \end{aligned}$$

Since $\hat{h}$ is the optimizer, $\widehat{MMSE}_n(\hat{h})/\widehat{MMSE}_n(\tilde{h}) \leq 1$. Thus

$$\frac{\left\{\frac{b_1}{2}\left[\hat{m}_1^{(2)}(x)\hat{C}_1^2 - \hat{m}_2^{(2)}(x)\hat{C}_2^2\right]\right\}^2 + \frac{\nu}{\hat{f}(x)}\left\{\frac{\hat{\sigma}_1^2(x)}{\hat{C}_1} + \frac{\hat{\sigma}_2^2(x)}{\hat{C}_2}\right\} + o_p(1)}{\left\{\frac{b_1}{2}\left[\hat{m}_1^{(2)}(x)\tilde{C}_1^2 - \hat{m}_2^{(2)}(x)\tilde{C}_2^2\right]\right\}^2 + \frac{\nu}{\hat{f}(x)}\left\{\frac{\hat{\sigma}_1^2(x)}{\tilde{C}_1} + \frac{\hat{\sigma}_2^2(x)}{\tilde{C}_2}\right\} + O_P(n^{-4/5})} \leq 1.$$

Note that the denominator converges to

$$\left\{\frac{b_1}{2}\left[m_1^{(2)}(x)C_1^{*2} - m_2^{(2)}(x)C_2^{*2}\right]\right\}^2 + \frac{\nu}{f(x)}\left\{\frac{\sigma_1^2(x)}{C_1^*} + \frac{\sigma_2^2(x)}{C_2^*}\right\},$$

where C_1^* and C_2^* are the unique optimizers of

$$\left\{\frac{b_1}{2}\left[m_1^{(2)}(x)C_1^2 - m_2^{(2)}(x)C_2^2\right]\right\}^2 + \frac{\nu}{f(x)}\left\{\frac{\sigma_1^2(x)}{C_1} + \frac{\sigma_2^2(x)}{C_2}\right\},$$

with respect to C_1 and C_2. This implies that $\hat{C}_1$ and $\hat{C}_2$ also converge to the same respective limit C_1^* and C_2^* because the inequality will be violated otherwise.

Next we consider the case with $m_1^{(2)}(x)m_2^{(2)}(x) > 0$. In this case, with probability approaching 1, $\hat{m}_1^{(2)}(x)\hat{m}_2^{(2)}(x) > 0$, so that we assume this without loss of generality.

When these conditions hold, define

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

and let $h_2 = \hat{\lambda}_2 h_1$. This sets the first-order bias term of $\widehat{MMSE}_n(h)$ equal to 0. Define the plug-in version of $AMSE_{2n}(h)$ by

$$\widehat{AMSE}_{2n}(h) = \left\{\hat{b}_{2,1}(x)h_1^3 - \hat{b}_{2,2}(x)h_2^3\right\}^2 + \frac{v}{n\hat{f}(x)}\left\{\frac{\hat{\sigma}_1^2(x)}{h_1} + \frac{\hat{\sigma}_2^2(x)}{h_2}\right\}$$

Choosing h_1 to minimize $\widehat{AMSE}_{2n}(h)$, we define $\tilde{h}_1 = \hat{\theta}_2 n^{-1/7} \equiv \tilde{C}_1 n^{-1/7}$ and $\tilde{h}_2 =$

$\hat{\lambda}_2\tilde{h}_1 \equiv \tilde{C}_2 n^{-1/7}$. Then $\widehat{MMSE}_n(\tilde{h})$ can be written as

$$\widehat{MMSE}_n(\tilde{h}) = n^{-6/7}\left\{\hat{b}_{2,1}(x)\tilde{C}_1^3 - \hat{b}_{2,2}(x)\tilde{C}_2^3\right\}^2 + n^{-6/7}\frac{\nu}{\hat{f}(x)}\left\{\frac{\hat{\sigma}_1^2(x)}{\tilde{C}_1} + \frac{\hat{\sigma}_2^2(x)}{\tilde{C}_2}\right\}.$$

In order to match this rate of convergence, both $\hat{h}_1$ and $\hat{h}_2$ 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,

$$\hat{m}_1^{(2)}(x)\hat{h}_1^2 - \hat{m}_2^{(2)}(x)\hat{h}_2^2 \equiv B_{1n} = n^{-3/7}b_{1n},$$

where $b_{1n} = O_P(1)$ so that under the assumption that $m_2^{(2)}(x) \neq 0$, with probability approaching 1, $\hat{m}_2^{(2)}(x)$ is bounded away from 0 so that assuming this without loss of generality, we have $\hat{h}_2^2 = \hat{\lambda}_2^2\hat{h}_1^2 - B_{1n}/\hat{m}_2^{(2)}(x)$. Substituting this expression to the second term and the third term, we have

$$\widehat{MMSE}_n(\hat{h}) = \left\{\frac{b_1}{2}B_{1n}\right\}^2 + \left\{\hat{b}_{2,1}(x)\hat{h}_1^3 - \hat{b}_{2,2}(x)\{\hat{\lambda}_2^2\hat{h}_1^2 - B_{1n}/\hat{m}_2^{(2)}(x)\}^{3/2}\right\}^2 + \frac{\nu}{n\hat{f}(x)}\left\{\frac{\hat{\sigma}_1^2(x)}{\hat{h}_1} + \frac{\hat{\sigma}_2^2(x)}{\{\hat{\lambda}_2^2\hat{h}_1^2 - B_{1n}/\hat{m}_2^{(2)}(x)\}^{1/2}}\right\}.$$

Suppose $\hat{h}_1$ is of order slower than $n^{-1/7}$. Then because $\hat{m}_2^{(2)}(x)^3\hat{b}_{2,1}(x)^2 \neq \hat{m}_1^{(2)}(x)^3\hat{b}_{2,2}(x)^2$ 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 $\hat{h}_2^2 = \hat{\lambda}_2^2\hat{h}_1^2 - B_{1n}/\hat{m}_2^{(2)}(x)$. Using this expression, we can write

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

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, $\widehat{MMSE}_n(\hat{h})/\widehat{MMSE}_n(\tilde{h}) \leq 1$. Thus

$$\frac{o_p(1) + \left\{\left[\hat{b}_{2,1}(x)\hat{C}_1^3 - \hat{b}_{2,2}(x)\{\hat{\lambda}_2^2\hat{C}_1^2 + o_p(1)\}^{3/2}\right]\right\}^2 + \frac{\nu}{\hat{f}(x)}\left\{\frac{\hat{\sigma}_1^2(x)}{\hat{C}_1+o_p(1)} + \frac{\hat{\sigma}_2^2(x)}{\{\hat{\lambda}_2^2\hat{C}_1^2+o_p(1)\}^{1/2}}\right\}}{\left\{\hat{b}_{2,1}(x)\tilde{C}_1^3 - \hat{b}_{2,2}(x)\tilde{C}_2^3\right\}^2 + \frac{\nu}{\hat{f}(x)}\left\{\frac{\hat{\sigma}_1^2(x)}{\tilde{C}_1} + \frac{\hat{\sigma}_2^2(x)}{\tilde{C}_2}\right\}} \leq 1.$$

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 $\hat{h}_2/\tilde{h}_2$ converges in probability to 1 as well.

The result above also shows that $\widehat{MMSE}_n(\hat{h})/MSE_n(h^*)$ converges to 1 in probability in both cases. ■

References

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

Hoerl, A. E., R. W. Kennard, and K. F. Baldwin (1975): "Ridge regression: some simulations," *Communications in Statistics, Theory and Methods*, 4, 105–123.

Jones, M. C. (1994): "On kernel density derivative estimation," *Communications in Statistics, Theory and Methods*, 23, 2133–2139.

Wand, M. P., and M. C. Jones (1994): *Kernel Smoothing*. Chapman & Hall.