The Review of Economic Studies Advance Access published November 24, 2011

Optimal Bandwidth Choice for the Regression Discontinuity Estimator*

Guido Imbens[†]
and Karthik Kalyanaraman[‡]

First Draft: June 2008
This Draft: September 2011

Abstract

We investigate the choice of the bandwidth for the regression discontinuity estimator. We focus on estimation by local linear regression, which was shown to have attractive properties (Porter, 2003). We derive the asymptotically optimal bandwidth under squared error loss. This optimal bandwidth depends on unknown functionals of the distribution of the data and we propose simple and consistent estimators for these functionals to obtain a fully data-driven bandwidth algorithm. We show that this bandwidth estimator is optimal according to the criterion of Li (1987), although it is not unique in the sense that alternative consistent estimators for the unknown functionals would lead to bandwidth estimators with the same optimality properties. We illustrate the proposed bandwidth, and the sensitivity to the choices made in our algorithm, by applying the methods to a data set previously analyzed by Lee (2008), as well as by conducting a small simulation study.

JEL Classification: C13, C14, C21
Keywords: *Optimal Bandwidth Selection, Local Linear Regression, Regression Discontinuity Designs, Cross-Validation*

[*] Financial support for this research was generously provided through NSF grants 0452590, 0820361, and 1024841. We are grateful to David Lee for making his data available, and to Joshua Angrist, Tom Cook, Tom Lemieux, Doug Miller, Fernando Yu, Fanyin Zheng, a co-editor and three referees for comments.

[†] Department of Economics, Harvard University, 1805 Cambridge Street, Cambridge, MA 02138, and NBER. Electronic correspondence: imbens@harvard.edu, http://www.economics.harvard.edu/faculty/imbens/imbens.html.

[‡] University College London, Gower Street, London. Electronic correspondence: kkalyanar@gmail.com.

1 Introduction

Regression discontinuity (RD) designs for evaluating causal effects of interventions, where assignment to the intervention is (partly) determined by the value of an observed covariate exceeding a threshold, were introduced by Thistlewaite and Campbell (1960). See Cook (2008) for a historical perspective. A recent surge of applications in economics includes studies of the impact of financial aid offers on college acceptance (Van Der Klaauw, 2002), school quality on housing values (Black, 1999), class size on student achievement (Angrist and Lavy, 1999), air quality on health outcomes (Chay and Greenstone, 2005), incumbency on re-election (Lee, 2008), and many others. Recent important theoretical work has dealt with identification issues (Hahn, Todd, and Van Der Klaauw, 2001, HTV from hereon), optimal estimation (Porter, 2003), tests for validity of the design (McCrary, 2008), quantile effects (Frandsen, 2008; Frölich and Melly, 2008), and the inclusion of covariates (Frölich, 2007). General surveys include Imbens and Lemieux (2008), Van Der Klaauw (2008), and Lee and Lemieux (2010).

In RD settings analyses typically focus on the average effect of the treatment for units with values of the forcing variable close to the threshold, using local linear, or global polynomial series estimators. Fan and Gijbels (1992) and Porter (2003) show that local linear estimators are rate optimal and have attractive bias properties. A key decision in implementing local methods is the choice of bandwidth. In current practice researchers use a variety of *ad hoc* approaches for bandwidth choice, such as standard plug-in and cross-validation methods from the general nonparametric regression literature (e.g., Härdle 1992, Fan and Gijbels, 1992, Wand and Jones, 1994). These are typically based on objective functions which take into account the performance of the estimator of the regression function over the entire support, and do not yield optimal bandwidths for the problem at hand. There are few papers in the literature that use bandwidths which focus specifically on the RD setting (Ludwig and Miller, 2007; DesJardins and McCall, 2008; see discussion later in the paper), and none with optimality properties. In this paper we build on this literature by (i) deriving the asymptotically optimal bandwidth under squared error loss, taking account of the special features of the RD setting, and (ii) providing a fully data-dependent method for choosing the bandwidth that is asymptotically optimal in the sense of Li (1987).[1] Although optimal in large samples, the proposed algorithm involves initial bandwidth choices and is not unique. We analyze the sensitivity of the results to these choices. We illustrate our proposed algorithm using a data set previously analyzed by

[1] Matlab and Stata software for implementing this bandwidth rule is available on the website http://www.economics.harvard.edu/faculty/imbens/imbens.html.

Lee (2008), and compare our procedure to global methods and other local methods based on other error criteria. Simulations indicate that our proposed algorithm works well in realistic settings.

2 Basic model

In the basic RD setting researchers are interested in the causal effect of a binary treatment. In the setting we consider we have a sample of N units, drawn randomly from a large population. For unit i, for $i = 1, \ldots, N$, using Rubin's (1974) potential outcome notation, the variable $Y_i(1)$ denotes the potential outcome for unit i given treatment, and $Y_i(0)$ denotes the potential outcome without treatment. For unit i we observe the treatment received, W_i, equal to one if unit i was exposed to the treatment and 0 otherwise, and the outcome corresponding to the treatment received:

$$Y_i = Y_i(W_i) = \begin{cases} Y_i(0) & \text{if} \quad W_i = 0, \\ Y_i(1) & \text{if} \quad W_i = 1. \end{cases}$$

We also observe for each unit a scalar covariate, called the forcing variable, denoted by X_i. In Section 5 we discuss the case with additional covariates. Define

$$m(x) = \mathbb{E}[Y_i|X_i = x],$$

to be the conditional expectation of the outcome given the forcing variable. The idea behind the Sharp Regression Discontinuity (SRD) design is that the treatment W_i is determined solely by the value of the forcing variable X_i being on either side of a fixed and known threshold c, or:

$$W_i = \mathbf{1}_{X_i \geq c}.$$

In Section 5 we extend the SRD setup to the case with additional covariates and to the Fuzzy Regression Discontinuity (FRD) design, where the probability of receiving the treatment jumps discontinuously at the threshold for the forcing variable, but not necessarily from zero to one.

In the SRD design the focus is on average effect of the treatment for units with covariate values equal to the threshold:

$$\tau_{\text{SRD}} = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = c].$$

Now suppose that the conditional distribution functions $F_{Y(0)|X}(y|x)$ and $F_{Y(1)|X}(y|x)$ are continuous in x for all y, and that the conditional first moments $\mathbb{E}[Y_i(1)|X_i = x]$ and $\mathbb{E}[Y_i(0)|X_i = x]$

exist, and are continuous at $x = c$. Then

$$\tau_{\text{SRD}} = \mu_+ - \mu_-, \qquad \text{where} \quad \mu_+ = \lim_{x \downarrow c} m(x), \qquad \text{and} \quad \mu_- = \lim_{x \uparrow c} m(x).$$

Thus, the estimand is the difference of two regression functions evaluated at boundary points.

We focus on estimating τ_{SRD} by separate local linear regressions on both sides of the threshold. We view local nonparametric methods as attractive in this setting compared to methods based on global approximations to the regression function (e.g., higher order polynomials applied to the full data set) because local methods build in robustness by ensuring that observations with values for the forcing variable far away from the threshold do not affect the point estimates. Furthermore, in the RD setting local linear regression estimators are preferred to the standard Nadaraya-Watson kernel estimator, because local linear methods have attractive bias properties in estimating regression functions at the boundary (Fan and Gijbels, 1992), and enjoy rate optimality (Porter, 2003).

To be explicit, we estimate the regression function $m(\cdot)$ at x as

$$\hat{m}_h(x) = \begin{cases} \hat{\alpha}_-(x) & \text{if } x < c, \\ \hat{\alpha}_+(x) & \text{if } x \geq c. \end{cases} \tag{2.1}$$

where,

$$(\hat{\alpha}_-(x), \hat{\beta}_-(x)) = \arg\min_{\alpha,\beta} \sum_{i=1}^{N} \mathbf{1}_{X_i < x} \cdot (Y_i - \alpha - \beta(X_i - x))^2 \cdot K\left(\frac{X_i - x}{h}\right),$$

and

$$(\hat{\alpha}_+(x), \hat{\beta}_+(x)) = \arg\min_{\alpha,\beta} \sum_{i=1}^{N} \mathbf{1}_{X_i > x} \cdot (Y_i - \alpha - \beta(X_i - x))^2 \cdot K\left(\frac{X_i - x}{h}\right),$$

Then we can write the estimator for τ_{SRD} as the difference in two regression estimators,

$$\hat{\tau}_{\text{SRD}} = \hat{\mu}_+ - \hat{\mu}_-,$$

where the two regression estimators are

$$\hat{\mu}_- = \lim_{x \uparrow c} \hat{m}_h(x) = \hat{\alpha}_-(c) \qquad \text{and} \quad \hat{\mu}_+ = \lim_{x \downarrow c} \hat{m}_h(x) = \hat{\alpha}_+(c).$$

The focus in this paper is on the optimal choice for the bandwidth h.

3 Error Criterion and Infeasible Optimal Bandwidth Choice

In this section we discuss the objective function, and derive the optimal bandwidth h_{opt} under that criterion.

3.1 Error Criteria

The primary question studied in this paper concerns the optimal choice of the bandwidth h. In the current empirical literature researchers often choose the bandwidth by either cross-validation or *ad hoc* methods. See Härdle (1992), Fan and Gijbels (1992), and Wand and Jones (1994) for textbook discussions of cross-validation and related methods, and Ludwig and Miller (2007) for a specific implementation in the RD settings. Conventional cross-validation yields a bandwidth that is optimal for fitting a curve over the entire support of the data. Typically it leads to a bandwidth choice that minimizes an approximation to the mean integrated squared error criterion (MISE),

$$\text{MISE}(h) = \mathbb{E}\left[\int_x \left(\hat{m}_h(x) - m(x)\right)^2 f(x)dx\right],$$

where $f(x)$ is the density of the forcing variable. This criterion is not directly relevant for the problem at hand: we wish to choose a bandwidth that is optimal for estimating τ_{SRD}. This estimand has two special features that are not captured in the MISE criterion. First, τ_{SRD} depends on $m(x)$ only through two values, and specifically their difference. Second, both these values are boundary values.

Our proposed criterion is based on the expectation of the asymptotic expansion, around $h = 0$, of the squared error $(\hat{\tau}_{\text{SRD}} - \tau_{\text{SRD}})^2$. First, define the mean squared error:

$$\text{MSE}(h) = \mathbb{E}\left[\left(\hat{\tau}_{\text{SRD}} - \tau_{\text{SRD}}\right)^2\right] = \mathbb{E}\left[\left((\hat{\mu}_+ - \mu_+) - (\hat{\mu}_- - \mu_-)\right)^2\right], \tag{3.2}$$

and let h^* be the optimal bandwidth that minimizes this criterion:

$$h^* = \arg\min_h \text{MSE}(h). \tag{3.3}$$

This criterion is difficult to work with directly. The problem is that in many cases even as the sample sizes become infinite, the optimal bandwidth h^* will not converge to zero. This is because biases in the parts of the regression function away from the threshold may be offsetting.[2] In such cases the optimal bandwidth h^* is very sensitive to the actual distribution and regression function. Moreover it does not seem appropriate to base estimation on global criteria when identification is local. We therefore follow the standard bandwidth choice literature in statistics

[2] To be explicit, consider a simple example where we are interested in estimating a regression function $g(x)$ at a single point, say $g(0)$. Suppose the covariate X has a uniform distribution on $[0, 1]$. Suppose the regression function is $g(x) = (x - 1/4)^2 - 1/16$. With a uniform kernel the estimator for $g(0)$ is, for a bandwidth h, equal to $\sum_{i:X_i<h} X_i / \sum_{i:X_i<h} 1$. As a function of the bandwidth h the bias is equal to $h^2/3 - h/4$, conditional on $\sum_{i:X_i<h} 1$. Thus, the bias is zero at $h = 3/4$, and if we minimize the expected squared error, the optimal bandwidth will converge to $3/4$ as the sample size gets large.

by focusing on the bandwidth that minimizes a first order approximation to MSE(h), what we call the asymptotic mean squared error or AMSE(h).

A second comment concerns our focus on a single bandwidth. Because the estimand, $\tau_{\rm SRD}$, is a function of the regression function at two points, an alternative would be to allow for a different bandwidth for these two points, h_- for estimating μ_-, and h_+ for estimating μ_+ and focus on an objective function that is an approximation to

$$\text{MSE}(h_-, h_+) = \mathbb{E}\left[\Big((\hat{\mu}_+(h_+) - \mu_+) - (\hat{\mu}_-(h_-) - \mu_-)\Big)^2\right], \tag{3.4}$$

instead of focusing on an approximation to MSE(h). Doing so would raises an important issue. We focus on minimizing mean squared error, equal to variance plus bias squared. Suppose that for both estimators the biases, $\mathbb{E}[\hat{\mu}_-(h_-)] - \mu_-$ and $\mathbb{E}[\hat{\mu}_+(h_+)] - \mu_+$ are strictly increasing (or both strictly decreasing) functions of the bandwidth. Then there is a function $h_+(h_-)$ such that the bias of the RD estimate, i.e. the difference between the above biases cancel out: $(\mathbb{E}[\hat{\mu}_-(h_-)] - \mu_-) - (\mathbb{E}[\hat{\mu}_+(h_+(h_-)) - \mu_+]) = 0$. Hence we can minimize the mean-squared-error by letting h_- get large (the variance is generally a decreasing function of the bandwidth), and choosing $h_+ = h_+(h_-)$. Even if this does not hold exactly, the point is that a problem may arise that even for large bandwidths the difference in bias may be close to zero. In practice it is unlikely that one can effectively exploit the cancellation of biases for large bandwidths. This would make it difficult to construct practical bandwidth algorithms. Therefore, in order to avoid this problem, we focus in this discussion on a single bandwidth choice.

3.2 An Asymptotic Expansion of the Expected Error

The next step is to derive an asymptotic expansion of MSE(h) given (3.2) and formally define the asymptotic approximation AMSE(h). First we state the key assumptions. Not all of these will be used immediately, but for convenience we state them all here.

Assumption 3.1: (Y_i, X_i), *for* $i = 1, \ldots, N$, *are independent and identically distributed.*

Assumption 3.2: *The marginal distribution of the forcing variable* X_i, *denoted* $f(\cdot)$, *is continuous and bounded away from zero at the threshold* c.

Assumption 3.3: *The conditional mean* $m(x) = \mathbb{E}[Y_i|X_i = x]$ *has at least three continuous derivatives in an open neigborhood of* $X = c$. *The right and left limits of the* k^{th} *derivative of* $m(x)$ *at the threshold* c *are denoted by* $m_+^{(k)}(c)$ *and* $m_-^{(k)}(c)$.

Assumption 3.4: *The kernel $K(\cdot)$ is nonnegative, bounded, differs from zero on a compact interval $[0, a]$, and is continuous on $(0, a)$.*

Assumption 3.5: *The conditional variance function $\sigma^2(x) = \mathrm{Var}(Y_i|X_i = x)$ is bounded in an open neigborhood of $X = c$, and right and left continuous at c. The right and left limit at the threshold are denoted by $\sigma^2_+(c)$ and $\sigma^2_-(c)$ respectively. $\sigma^2_+(c) > 0$ and $\sigma^2_-(c) > 0$*

Assumption 3.6: *The second derivatives from the right and the left differ at the threshold: $m^{(2)}_+(c) \neq m^{(2)}_-(c)$.*

Now define the Asymptotic Mean Squared Error (AMSE) as a function of the bandwidth h:

$$\mathrm{AMSE}(h) = C_1 \cdot h^4 \cdot \left(m^{(2)}_+(c) - m^{(2)}_-(c)\right)^2 + \frac{C_2}{N \cdot h} \cdot \left(\frac{\sigma^2_+(c)}{f(c)} + \frac{\sigma^2_-(c)}{f(c)}\right). \tag{3.5}$$

The constants C_1 and C_2 in this approximation are functions of the kernel:

$$C_1 = \frac{1}{4}\left(\frac{\nu_2^2 - \nu_1\nu_3}{\nu_2\nu_0 - \nu_1^2}\right)^2, \qquad \text{and} \quad C_2 = \frac{\nu_2^2\pi_0 - 2\nu_1\nu_2\pi_1 + \nu_1^2\pi_2}{(\nu_2\nu_0 - \nu_1^2)^2}, \tag{3.6}$$

where

$$\nu_j = \int_0^\infty u^j K(u) du, \qquad \text{and} \quad \pi_j = \int_0^\infty u^j K^2(u) du.$$

The first term in (3.5) corresponds to the square of the bias, and the second term corresponds to the variance. The expression for AMSE(h) clarifies the role that Assumption 3.6 will play. The leading term in the expansion of the bias is of order h^4 if the left and right limits of the second derivative differ. If these two limits are equal, the bias converges to zero faster, allowing for estimation of τ_{SRD} at a faster rate of convergence. It is difficult to exploit the improved convergence rate that would result from this in practice, because it would be difficult to establish sufficiently fast that two second derivatives are indeed equal, and therefopre we focus on optimality results given Assumption 3.6. Note however, that even if the second derivatives are identical, our proposed estimator for τ_{SRD} will be consistent.

An alternative approach would be to focus on a bandwidth choice that is optimal if the second derivatives from the left and right are identical. It is possible to construct such a bandwidth choice, and still maintain consistency of the resulting estimator for τ_{SRD} irrespective of the difference in second derivatives. However, such an bandwidth choice would generally not be optimal if the difference in second derivatives is nonzero. Thus there is a choice between a bandwidth choice that is optimal under $m^{(2)}_+(c) \neq m^{(2)}_-(c)$ and a bandwidth choice that is optimal under $m^{(2)}_+(c) = m^{(2)}_-(c)$. In the current paper we choose to focus on the first case.

Lemma 3.1: (Mean Squared Error Approximation and Optimal Bandwidth)
(i) Suppose Assumptions 3.1-3.5 hold. Then

$$\text{MSE}(h) = \text{AMSE}(h) + o_p\left(h^4 + \frac{1}{N \cdot h}\right).$$

(ii) Suppose that also Assumption 3.6 holds. Then

$$h_{\text{opt}} = \arg\min_h \text{AMSE}(h) = C_K \cdot \left(\frac{\sigma_+^2(c) + \sigma_-^2(c)}{f(c) \cdot \left(m_+^{(2)}(c) - m_-^{(2)}(c)\right)^2}\right)^{1/5} \cdot N^{-1/5}, \tag{3.7}$$

where $C_K = (C_2/(4 \cdot C_1))^{1/5}$, indexed by the kernel $K(\cdot)$.

For the edge kernel, with $K(u) = \mathbf{1}_{|u| \leq 1}(1 - |u|)$, shown by Cheng, Fan and Marron (1997) to have optimality properties for boundary estimation problems, the constant is $C_{K,\text{edge}} \approx 3.4375$. For another commonly used kernel, the uniform kernel with $K(u) = \mathbf{1}_{|u| \leq 1/2}$, the constant is approximately $C_{K,\text{uniform}} \approx 5.40$.

4 Feasible Optimal Bandwidth Choice

In this section we develop an estimator for the bandwidth and discuss its asymptotic properties. The proposed bandwith estimator is fully data-driven and based on substituting consistent estimators for the various components of the optimal bandwidth given in (3.7). It involves a number of choices for initial smoothing parameters in order to estimate these components. As is typically the case with plug-in estimators, these choices are not unique and can be replaced by others without affecting the asymptotic optimality of the procedure.

4.1 A Simple Plug-in Bandwidth

A natural choice for the estimator for the optimal bandwidth estimator is to replace the six unknown quantities in the expression for the optimal bandwidth h_{opt} in (3.7) by consistent estimators, leading to

$$\tilde{h}_{\text{opt}} = C_K \cdot \left(\frac{\hat{\sigma}_-^2(c) + \hat{\sigma}_+^2(c)}{\hat{f}(c) \cdot \left(\hat{m}_+^{(2)}(c) - \hat{m}_-^{(2)}(c)\right)^2}\right)^{1/5} \cdot N^{-1/5}. \tag{4.8}$$

One potential concern here however is that the first order bias may be extremely small or vanish in finite samples. This could happen for instance in the constant additive treatment effects case where $m_{+}^{(2)}(x) = m_{-}^{(2)}(x)$ for any x. In this case the bandwidth that minimizes first order mean squares is infinite (the denominator term is zero in (4.8))[3].

More generally, even if the true value of the bias term is not zero, the precision with which we estimate the second derivatives $m_{+}^{(2)}(c)$ and $m_{-}^{(2)}(c)$ is likely to be low. Thus the estimated optimal bandwidth $\tilde{h}_{\text{opt}}$ will occasionally be very large, even when the data are consistent with a substantial degree of curvature. Thus estimates of the bandwidth will be very imprecise and will have a large variance across repeated datasets. Moreover such a bandwidth may lead to estimators for τ_{SRD} with poor properties, because the true finite sample bias depends on global properties of the regression function which are not captured by the asymptotic approximation used to calculate the bandwidth.[4]

4.1.1 Regularization

Motivated by the above concern that due to the error in the estimation of the true curvature, the error in the estimation of its squared reciprocal could potentially be large, leading to very large and ill-performing bandwidths, we modify the bandwidth estimator using ideas from the regularization literature.[5] A simple calculation establishes that the bias in the plug-in estimator for the reciprocal of the squared difference in second derivatives is

$$\mathbb{E}\left[\frac{1}{\left(\hat{m}_{+}^{(2)}(c) - \hat{m}_{-}^{(2)}(c)\right)^2} - \frac{1}{\left(m_{+}^{(2)}(c) - m_{-}^{(2)}(c)\right)^2}\right] = \left(\frac{3 \cdot \left(\mathbb{V}\left(\hat{m}_{+}^{(2)}(c)\right) + \mathbb{V}\left(\hat{m}_{-}^{(2)}(c)\right)\right)}{\left(m_{+}^{(2)}(c) - m_{-}^{(2)}(c)\right)^4}\right) + o(N^{-2\alpha}).$$

This implies that, for $r = 3 \cdot \left(\mathbb{V}\left(\hat{m}_{-}^{(2)}(c)\right) + \mathbb{V}\left(\hat{m}_{+}^{(2)}(c)\right)\right)$, the bias in the modified estimator for the reciprocal of the squared difference in second derivatives is of lower order:

$$\mathbb{E}\left[\frac{1}{\left(\hat{m}_{+}^{(2)}(c) - \hat{m}_{-}^{(2)}(c)\right)^2 + r} - \frac{1}{\left(m_{+}^{(2)}(c) - m_{-}^{(2)}(c)\right)^2}\right] = o(N^{-2\alpha}).$$

[3] This problem is not unique to our specific estimator. In the general case of estimating a regression at an interior point, this occurs when the second derivative at that point is zero

[4] As an aside, the same formal argument applies to the estimator of the density. If the estimated density is close to zero, the bandwidth estimator might become unstable. However, in practice that is rarely a concern: if the true density is so close to zero that one cannot estimate the density accurately at the threshold, it is unlikely that any estimates of the discontinuity will be precise enough to be of interest. We therefore focus on the complications arising from the difference in second derivatives being estimated to be close to zero.

[5] Kalyanaraman(2008) has developed some theory about regularization in bandwidth selection in the different context of estimated smooth regression functionals.

This in turn motivates the modified bandwidth estimator

$$\hat{h}_{\text{opt}} = C_K \cdot \left(\frac{\hat{\sigma}_-^2(c) + \hat{\sigma}_+^2(c)}{\hat{f}(c)\left(\left(\hat{m}_+^{(2)}(c) - \hat{m}_-^{(2)}(c) \right)^2 + \hat{r}_- + \hat{r}_+ \right)} \right)^{1/5} \cdot N^{-1/5}, \tag{4.9}$$

where

$$\hat{r}_- = 3 \cdot \hat{\mathbb{V}}\left(\hat{m}_-^{(2)}(c) \right), \qquad \text{and} \;\; \hat{r}_+ = 3 \cdot \hat{\mathbb{V}}\left(\hat{m}_+^{(2)}(c) \right).$$

Note that this bandwidth will not become infinite even in the cases when the difference in curvatures at the threshold is zero.

4.1.2 Implementing the Regularization

Consider first a simplification of the regularization term $r = r_- + r_+$, where r_- and r_+ are three times the variance of the estimated curvatures on the left and the right respectively. To be explicit, we estimate the second derivative $m_+^{(2)}(c)$ by fitting a quadratic function to the observations with $X_i \in [c, c+h]$. The initial bandwidth h here will be different from the bandwidth $\hat{h}_{\text{opt}}$ used in the estimation of τ_{SRD}, and its choice will be discussed in Section 4.2. Let $N_{h,+}$ be the number of units with covariate values in this interval. We assume homoskedasticity with error variance $\sigma^2(c)$ in this interval. Let

$$\hat{\mu}_{j,h,+} = \frac{1}{N_{h,+}} \sum_{c \le X_i \le c+h} (X_i - \overline{X})^j, \qquad \text{where} \;\; \overline{X} = \frac{1}{N_{h,+}} \sum_{c \le X_i \le c+h} X_i,$$

be the j-th (centered) moment of the X_i in the interval $[c, c+h]$. We can derive the following explicit formula for three times the conditional variance of the curvature on the left, denoted by r_+, in terms of these moments:

$$r_+ = \frac{12}{N_{h,+}} \cdot \left(\frac{\sigma_+^2(c)}{\hat{\mu}_{4,h,+} - (\hat{\mu}_{2,h,+})^2 - (\hat{\mu}_{3,h,+})^2 / \hat{\mu}_{2,h,+}} \right)$$

However, because fourth moments are difficult to estimate precisely, we approximate this expression exploiting the fact that for small h, the distribution of the forcing variable can be approximated by a uniform distribution on $[c, c+h]$, so that $\mu_{2,h,+} \approx h^2/12$, $\mu_{3,h,+} \approx 0$, and $\mu_{4,h,+} \approx h^4/60$. After substituting $\hat{\sigma}_-^2(c)$ for $\sigma_-^2(c)$ and $\hat{\sigma}_+^2(c)$ for $\sigma_+^2(c)$ this leads to

$$\hat{r}_+ = \frac{2160 \cdot \hat{\sigma}_+^2(c)}{N_{h,+} \cdot h^4}, \qquad \text{and similarly} \;\; \hat{r}_- = \frac{2160 \cdot \hat{\sigma}_-^2(c)}{N_{h,-} \cdot h^4}.$$

The proposed bandwidth is now obtained by adding the regularization term $\hat{r} = \hat{r}_- + \hat{r}_+$ to the squared difference-in-curvature term in the bias term of MSE expansion:

$$\hat{h}_{\text{opt}} = C_K \cdot \left(\frac{\hat{\sigma}_-^2(c) + \hat{\sigma}_+^2(c)}{\hat{f}(c) \left(\left(\hat{m}_+^{(2)}(c) - \hat{m}_-^{(2)}(c) \right)^2 + \hat{r}_- + \hat{r}_+ \right)} \right)^{1/5} \cdot N^{-1/5}. \tag{4.10}$$

To operationalize this proposed bandwidth, we need specific estimators $\hat{f}(c)$, $\hat{\sigma}_-^2(c)$, $\hat{\sigma}_+^2(c)$, $\hat{m}_-^{(2)}(c)$, and $\hat{m}_+^{(2)}(c)$. In the next section we discuss a specific way of doing so, leading to a completely data-driven bandwidth choice. This bandwidth estimator will be shown to have certain optimality properties. It should be noted though that our proposed bandwidth estimator is not unique in having these optimality properties. Any combination of consistent estimators for $f(c)$, $\sigma_-^2(c)$, $\sigma_+^2(c)$, $m_-^{(2)}(c)$, and $m_+^{(2)}(c)$ substituted into expression (4.10), with or without the regularity terms, will have the same optimality properties. Within this class, our proposed estimator is relatively simple, but the more important point is that it is a specific estimator, in the same spirit as the Silverman rule-of-thumb bandwidth for nonparametric density estimation: it gives a convenient starting point and benchmark for doing a sensitivity analyses regarding bandwidth choice.

In addition we will address the sensitivity of our bandwidth estimator to the choices made in our algorithm in a simulation study. In general we find the bandwidth selection algorithm to be relatively robust to these choices. This is not surprising given that the presence of the power 1/5 in the expression for the optimal bandwidth: for example, doubling the estimate for both $\sigma_-^2(c)$ and $\sigma_+^2(c)$ only increases the estimated bandwidth by a factor $2^{1/5} \approx 1.18$.

4.2 An Algorithm for bandwidth selection

The reference bandwidth $\hat{h}_{\text{opt}}$ is a function of estimates for $f(c)$, $\sigma_-^2(c)$, $\sigma_+^2(c)$, $m_-^{(2)}(c)$, and $m_+^{(2)}(c)$ and the kernel $K(\cdot)$. Here we give a specific algorithm for implementation. In practice we recommend using the theoretically optimal edge kernel, where $K(u) = 1_{|u|\leq 1} \cdot (1 - |u|)$, which also has consistently superior performance in simulations, although the algorithm is easily modified for other kernels by changing the kernel-specific constant C_K. To calculate the bandwidth we also need estimators for the density at the threshold, $f(c)$, the conditional variances at the threshold, $\sigma_-^2(c)$ and $\sigma_+^2(c)$, and the limits of the second derivatives at the threshold from the right and the left, $m_-^{(2)}(c)$, $m_+^{(2)}(c)$. (The other components of (4.10), $\hat{r}_-$ and $\hat{r}_+$ are functions of these four components.) The first three functionals are calculated in step 1, the last two in step 2. Step 3 puts these together with the appropriate kernel constant

C_K to produce the reference bandwidth $\hat{h}_{\text{opt}}$.

We make the following choices in this algorithm. First, an initial bandwidth h_1 and a kernel to estimate the density $f_X(c)$ and the conditional outcome variances $\sigma^2_-(c)$ and $\sigma^2_+(c)$. Second, a pair of bandwidths $h_{2,-}$ and $h_{2,+}$ for estimating the second derivatives $m^{(2)}_-(c)$, $m^{(2)}_+(c)$. We choose these two bandwidths $h_{2,-}$ and $h_{2,+}$ optimally given the third derivative, which in turn we estimate globally. The choices for h_1, the initial kernel and the estimator for the third derivative do not affect the asymptotic optimality properties of the bandwidth estimator, but they do affect the finite sample properties.

Step 1: Estimation of density $f(c)$ and conditional variances $\sigma^2_-(c)$ and $\sigma^2_+(c)$

First calculate the sample variance of the forcing variable, $S^2_X = \sum\left(X_i - \overline{X}\right)^2/(N-1)$. We now use the Silverman rule to get a pilot bandwidth for calculating the density and variance at c. The standard Silverman rule of $h = 1.06 \cdot S_X \cdot N^{-1/5}$ is based on a normal kernel and a normal reference density. We modify this for the uniform kernel on $[-1,1]$ and the normal reference density, and calculate the pilot bandwidth h_1 as:

$$h_1 = 1.84 \cdot S_X \cdot N^{-1/5}.$$

We assess the sensitivity of the choice of a uniform kernel in the final simulations. We choose the uniform kernel because we are interested in a simple estimate of density, i.e. proportion of observations near the threshold (which is a kernel density estimate with a uniform kernel). Using alternative kernels would not affect the optimality properties in Theorem 4.1.

Calculate the number of units on either side of the threshold, and the average outcomes on either side as

$$N_{h_1,-} = \sum_{i=1}^{N} \mathbf{1}_{c-h_1 \le X_i < c}, \qquad N_{h_1,+} = \sum_{i=1}^{N} \mathbf{1}_{c \le X_i \le c+h_1},$$

$$\overline{Y}_{h_1,-} = \frac{1}{N_{h_1,-}} \sum_{i:c-h_1 \le X_i < c} Y_i, \qquad \text{and} \quad \overline{Y}_{h_1,+} = \frac{1}{N_{h_1,+}} \sum_{i:c \le X_i \le c+h_1} Y_i.$$

Now estimate the density of X_i at c as

$$\hat{f}(c) = \frac{N_{h_1,-} + N_{h_1,+}}{2 \cdot N \cdot h_1}, \tag{4.11}$$

and estimate the limit of the conditional variances of Y_i given $X_i = x$, at $x = c$, from the left and the right, as

$$\hat{\sigma}^2_-(c) = \frac{1}{N_{h_1,-} - 1} \sum_{i:c-h_1 \le X_i < c} \left(Y_i - \overline{Y}_{h_1,-}\right)^2, \tag{4.12}$$

and

$$\hat{\sigma}^2_+(c) = \frac{1}{N_{h_1,+} - 1} \sum_{i:c \le X_i \le c+h_1} \left(Y_i - \overline{Y}_{h_1,+}\right)^2 . \tag{4.13}$$

The main property we will need for these estimators is that they are consistent for the density and the conditional variance respectively. They need not be efficient for the optimality properties in Theorem 4.1. Because the bandwidth goes to zero at rate $N^{-1/5}$, Assumptions 3.2 and 3.5 are sufficient for consistency of these estimators.

Step 2: Estimation of second derivatives $\hat{m}^{(2)}_+(c)$ and $\hat{m}^{(2)}_-(c)$

First we need pilot bandwidths $h_{2,-}$ and $h_{2,+}$. We base this on a simple, not necessarily consistent, estimator of the third derivative of $m(\cdot)$ at c. Fit a third order polynomial to the data, including an indicator for $X_i \ge 0$. Thus, estimate the regression function

$$Y_i = \gamma_0 + \gamma_1 \cdot 1_{X_i \ge c} + \gamma_2 \cdot (X_i - c) + \gamma_3 \cdot (X_i - c)^2 + \gamma_4 \cdot (X_i - c)^3 + \varepsilon_i, \tag{4.14}$$

and estimate $m^{(3)}(c)$ as $\hat{m}^{(3)}(c) = 6 \cdot \hat{\gamma}_4$. This will be our estimate of the third derivative of the regression function. Note that $\hat{m}^{(3)}(c)$ is in general not a consistent estimate of $m^{(3)}(c)$ but will converge to some constant at a parametric rate. However we do not need a consistent estimate of the third derivative at c here to obtain a consistent estimator for the second derivative. Calculate $h_{2,+}$, using the $\hat{\sigma}^2_-(c)$, $\hat{\sigma}^2_+(c)$ and $\hat{f}(c)$ from Step 1, as

$$h_{2,+} = 3.56 \cdot \left(\frac{\hat{\sigma}^2_+(c)}{\hat{f}(c) \cdot \left(\hat{m}^{(3)}(c)\right)^2} \right)^{1/7} \cdot N_+^{-1/7}, \tag{4.15}$$

and

$$h_{2,-} = 3.56 \cdot \left(\frac{\hat{\sigma}^2_-(c)}{\hat{f}(c) \cdot \left(\hat{m}^{(3)}(c)\right)^2} \right)^{1/7} \cdot N_-^{-1/7},$$

where N_- and N_+ are the number of of observations to the left and right of the threshold, respectively. These bandwidths, $h_{2,-}$ and $h_{2,+}$, are estimates of the optimal bandwidth for calculation of the second derivative at a boundary point using a local quadratic and a uniform kernel. See the Appendix for details. Again alternative consistent estimators for these second derivatives would also lead to optimality for the corresponding bandwidth estimator $\hat{h}_{\text{opt}}$.

Given the pilot bandwidth $h_{2,+}$, we estimate the curvature $m^{(2)}_+(c)$ by a local quadratic fit. To be precise, temporarily discard the observations other than the $N_{2,+}$ observations with $c \le X_i \le c + h_{2,+}$. Label the new data $\hat{\mathbf{Y}}_+ = (Y_1, \ldots, Y_{N_{2,+}})$ and $\hat{\mathbf{X}}_+ = (X_1, \ldots, X_{N_{2,+}})$

each of length $N_{2,+}$. Fit a quadratic to the new data. I.e. let $\mathbf{T} = [\iota \quad \mathbf{T}_1 \quad \mathbf{T}_2]$ where ι is a column vector of ones, and $\mathbf{T}_j' = \left((X_1 - c)^j, \ldots, (X_{N_{2,+}} - c)^j\right)$, for $j = 1, 2$. Estimate the regression coefficiens $\hat{\lambda} = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\hat{\mathbf{Y}}$. Calculate the curvature as $\hat{m}_+^{(2)}(c) = 2 \cdot \hat{\lambda}_3$. This is a consistent estimate of $m_+^{(2)}(c)$. To estimate $m_-^{(2)}(c)$ follow the same procedure using the data with $c - h_{2,-} \leq X_i < c$.

Step 3: Calculation of Regularization Terms $\hat{r}_-$ and $\hat{r}_+$, and Calculation of $\hat{h}_{\text{opt}}$

Given the previous steps, the regularization terms are calculated as

$$\hat{r}_+ = \frac{2160 \cdot \hat{\sigma}_+^2(c)}{N_{2,+} \cdot h_{2,+}^4}, \qquad \text{and} \quad \hat{r}_- = \frac{2160 \cdot \hat{\sigma}_-^2(c)}{N_{2,-} \cdot h_{2,-}^4}. \tag{4.16}$$

We now have all the pieces to calculate the proposed bandwidth:

$$\hat{h}_{\text{opt}} = C_K \cdot \left(\frac{\hat{\sigma}_-^2(c) + \hat{\sigma}_+^2(c)}{\hat{f}(c) \cdot \left(\left(\hat{m}_+^{(2)}(c) - \hat{m}_-^{(2)}(c) \right)^2 + (\hat{r}_+ + \hat{r}_-) \right)} \right)^{1/5} \cdot N^{-1/5}. \tag{4.17}$$

where C_K is, as in Lemma 3.1, a constant that depends on the kernel used. For the edge kernel, with $K(u) = (1 - |u|) \cdot \mathbf{1}_{|u| \leq 1}$, the constant is $C_K \approx 3.4375$.

Given the bandwidth $\hat{h}_{\text{opt}}$, we estimate τ_{SRD} as

$$\hat{\tau}_{\text{SRD}} = \lim_{x \downarrow c} \hat{m}_{\hat{h}_{\text{opt}}}(x) - \lim_{x \uparrow c} \hat{m}_{\hat{h}_{\text{opt}}}(x),$$

where $\hat{m}_h(x)$ is the local linear regression estimator defined in (2.1).

4.3 Properties of Algorithm

For the bandwidth choice based on this algorithm we establish some asymptotic properties. First, the resulting RD estimator $\hat{\tau}_{\text{SRD}}$ is consistent at the best rate for nonparametric regression functions at a point (Stone, 1982). Second, the estimated constant term in the reference bandwidth converges to the best constant. Third, we also have a Li (1987) type optimality result for the mean squared error and consistency at the optimal rate for the RD estimate. The optimality result implies that asymptotically the procedure with the estimated bandwidth $\hat{h}_{\text{opt}}$ performs as well as the infeasible procedure with the optimal bandwidth h_{opt}.

Theorem 4.1: (Properties of $\hat{h}_{\text{opt}}$)

Suppose Assumptions 3.1-3.5 hold. Then:

(i) (consistency) if Assumption 3.6 holds, then

$$\hat{\tau}_{\text{SRD}} - \tau_{\text{SRD}} = O_p\left(N^{-2/5}\right). \tag{4.18}$$

(ii) (consistency) if Assumption 3.6 does not hols, then

$$\hat{\tau}_{\mathrm{SRD}} - \tau_{\mathrm{SRD}} = O_p\left(N^{-3/7}\right). \tag{4.19}$$

(iii) (convergence of bandwidth)

$$\frac{\hat{h}_{\mathrm{opt}} - h_{\mathrm{opt}}}{h_{\mathrm{opt}}} = o_p(1), \tag{4.20}$$

and (iv) (Li optimality):

$$\frac{\mathrm{MSE}(\hat{h}_{\mathrm{opt}}) - \mathrm{MSE}(h_{\mathrm{opt}})}{\mathrm{MSE}(h_{\mathrm{opt}})} = o_p(1). \tag{4.21}$$

Note that when Assumption 3.6 holds, the convergence rate ($N^{-2/5}$) for $\hat{\tau}_{\mathrm{SRD}}$ is slower than when Assumption 3.6 does not hold (namely $N^{-3/7}$). This is because failure of Assumption (3.6) implies that the second derivatives from the left and right are equal, implying in turn that the leading term of the bias vanishes, which, as one might expect, would improve convergence. In fact the new convergence rate is the optimal one under the new setting as well.[6] In other words our proposed optimal bandwidth automatically adapts to the vanishing of the difference in second derivatives. In other words though we have a single bandwidth irrespective of whether the curvatures on either side of the threshold are equal or not, the procedure has an 'oracle' property of adapting to the unknown optimal rate.

4.4 DesJardins-McCall Bandwidth Selection

DesJardins and McCall (2008) use an alternative method for choosing the bandwidth. They focus separately on the limits of the regression function to the left and the right, rather than on the difference in the limits. This implies a focus on minimizing an objective criterion based on the sum of the squared differences between $\hat{\mu}_-$ and μ_-, and between $\hat{\mu}_+$ and μ_+:

$$\mathbb{E}\left[(\hat{\mu}_+ - \mu_+)^2 + (\hat{\mu}_- - \mu_-)^2\right],$$

instead of our criterion, which focuses on the squared difference between $(\hat{\mu}_+ - \hat{\mu}_-)$ and $(\mu_+ - \mu_-)$,

$$\mathbb{E}\left[((\hat{\mu}_+ - \mu_+) - (\hat{\mu}_- - \mu_-))^2\right] = \mathbb{E}\left[\left(\hat{\tau}_{\mathrm{SRD}} - \tau_{\mathrm{SRD}}\right)^2\right].$$

[6] I.e. when Assumption 3.6 fails, the new leading term in bias squared is now $O(h^6)$. Because the variance remains of order $O((Nh)^{-1})$, the optimal rate for the bandwidth, based on balancing the bias-squared and the variance, becomes $N^{-1/7}$. This leads to the optimal rate for $\hat{\tau}_{\mathrm{SRD}} - \tau_{\mathrm{SRD}}$ becoming $N^{-3/7}$, although the constant is no longer optimal.

The single optimal bandwidth based on the DesJardins and McCall criterion is

$$h_{DM} = C_K \cdot \left(\frac{\sigma_+^2(c) + \sigma_-^2(c)}{f(c) \cdot \left(m_+^{(2)}(c)^2 + m_-^{(2)}(c)^2 \right)} \right)^{1/5} \cdot N^{-1/5}.$$

This will in large samples lead to a smaller bandwidth than our proposed bandwidth choice if the second derivatives are of the same sign. DesJardins and McCall actually use different bandwidths on the left and the right, and also use a Epanechnikov kernel instead of the optimal edge kernel. In the simulations and bandwidth comparisons below we use the better performing edge kernel to facilitate the comparison with our proposed bandwidth $\hat{h}_{\text{opt}}$.

4.5 Ludwig-Miller Crossvalidation

In this section we briefly describe the cross-validation method used by Ludwig and Miller (2005, LM from hereon), which we compare to our proposed bandwidth in the application and simulations. See also Lee and Lemieux (2010). The LM bandwidth is the only cross-validation bandwidth selection procedure in the literature that is specifically aimed at the regression discontinuity setting. Let N_- and N_+ be the number of observations with $X_i < c$ and $X_i \geq c$ respectively. For $\delta \in (0, 1)$, let $\theta_-(\delta)$ and $\theta_+(\delta)$ be the δ-th quantile of the X_i among the subsample of observations with $X_i < c$ and $X_i \geq c$ respectively, so that

$$\theta_-(\delta) = \arg\min_a \left\{ a \,\middle|\, \left(\sum_{i=1}^{N} 1_{X_i \leq a} \right) \geq \delta \cdot N_- \right\},$$

and

$$\theta_+(\delta) = \arg\min_a \left\{ a \,\middle|\, \left(\sum_{i=1}^{N} 1_{c \leq X_i \leq a} \right) \geq \delta \cdot N_+ \right\}.$$

Now the LM cross-validation criterion we use is of the form:

$$CV_\delta(h) = \sum_{i=1}^{N} 1_{\theta_-(1-\delta) \leq X_i \leq \theta_+(\delta)} \cdot \left(Y_i - \hat{m}_h(X_i) \right)^2.$$

(In fact, LM use a slightly different criterion function, where they sum up over all observations within a distance h_0 from the threshold.) The estimator for the regression function here is $\hat{m}_h(x)$ defined in equation (2.1). A key feature of $\hat{m}_h(x)$ is that for values of $x < c$, it only uses observations with $X_i < x$ to estimate $m(x)$, and for values of $x \geq c$, it only uses observations with $X_i > x$ to estimate $m(x)$, so that $\hat{m}_h(X_i)$ does not depend on Y_i, as is necessary for cross-validation. By using a value for δ close to zero, we only use observations close to the threshold to evaluate the cross-validation criterion. Apart from the choice on needs to make of δ, the concern is that by using too small value of δ, we may not get a precisely estimated cross-validation bandwidth. In a minor modification of the LM proposal we use the edge kernel instead of the Epanechnikov kernel they suggest. In our calculations we use $\delta = 0.5$.

Any fixed value for δ is unlikely to lead to an optimal bandwidth in general, as it is implicitly based on a criterion function that is appropriate for fitting the entire regression function between the $(1-\delta)$-quantile for the observations on the left and the δ-quantile for observations on the right. Moreover, the criterion focuses implicitly on minimizing a criterion more akin to $\mathbb{E}\left[(\hat{\mu}_+ - \mu_+)^2 + (\hat{\mu}_- - \mu_-)^2\right]$, (with the errors in estimating μ_- and μ_+ squared before adding them up), rather than rather than $\text{MSE}(h) = \mathbb{E}[((\hat{\mu}_+ - \mu_+) - (\hat{\mu}_- - \mu_-))^2]$ where the error in the difference $\mu_+ - \mu_-$ is squared. As a result even letting $\delta \to 0$ with the sample size in the cross-validation procecure will not result in an optimal bandwidth.

5 Extensions

In this section we discuss two extensions. First we consider the fuzzy regression discontinuity design, and second we allow for the presence of covariates.

5.1 The Fuzzy Regression Design

In the Fuzzy Regression Discontinuity Design (FRD) the treatment W_i is not a deterministic function of the forcing variable. Instead the probability $\Pr(W_i = 1|X_i = x)$ changes discontinuously at the threshold c. The focus is on the ratio

$$\tau_{\text{FRD}} = \frac{\lim_{x\downarrow c} \mathbb{E}\left[Y_i|X_i = x\right] - \lim_{x\uparrow c} \mathbb{E}\left[Y_i|X_i = x\right]}{\lim_{x\downarrow c} \mathbb{E}\left[W_i|X_i = x\right] - \lim_{x\uparrow c} \mathbb{E}\left[W_i|X_i = x\right]}.$$

In an important theoretical paper Hahn, Todd and VanderKlaauw (2001) discuss identification in this setting, and show that in settings with heterogenous effects the estimand has an interpretation as a local average treatment effects (Imbens and Angrist, 1994). In the FRD case we need to estimate two regression functions, each at two boundary points: the expected outcome given the forcing variable $\mathbb{E}[Y_i|X_i = x]$ to the right and left of the threshold c and the expected value of the treatment variable given the forcing variable $\mathbb{E}[W_i|X_i = x]$, again both to the right and left of c. Again we focus on a single bandwidth, now the bandwidth that minimizes the mean squared error to this ratio. Define

$$\tau_Y = \lim_{x\downarrow c} \mathbb{E}\left[Y_i|X_i = x\right] - \lim_{x\uparrow c} \mathbb{E}\left[Y_i|X_i = x\right], \quad \text{and } \tau_W = \lim_{x\downarrow c} \mathbb{E}\left[W_i|X_i = x\right] - \lim_{x\uparrow c} \mathbb{E}\left[W_i|X_i = x\right],$$

with $\hat{\tau}_Y$ and $\hat{\tau}_W$ denoting the corresponding estimators, so that $\tau_{\text{FRD}} = \tau_Y/\tau_W$, and $\hat{\tau}_{\text{FRD}} = \hat{\tau}_Y/\hat{\tau}_W$. In large samples we can approximate the difference $\hat{\tau}_{\text{FRD}} - \tau_{\text{FRD}}$ by

$$\hat{\tau}_{\text{FRD}} - \tau_{\text{FRD}} = \frac{1}{\tau_W}(\hat{\tau}_Y - \tau_Y)) - \frac{\tau_Y}{\tau_W^2}(\hat{\tau}_W - \tau_W)) + o_p((\hat{\tau}_Y - \tau_Y) + (\hat{\tau}_W - \tau_W)).$$

This is the basis for the asymptotic approximation to the MSE around $h = 0$:

$$\text{AMSE}_{\text{FRD}}(h) = C_1 h^4 \left(\frac{1}{\tau_W} \left(m_{Y,+}^{(2)}(c) - m_{Y,-}^{(2)}(c) \right) - \frac{\tau_Y}{\tau_W^2} \left(m_{W,+}^{(2)}(c) - m_{W,-}^{(2)}(c) \right) \right)^2 \tag{5.22}$$

$$+ \frac{C_2}{Nhf(c)} \left(\frac{1}{\tau_W^2} \left(\sigma_{Y,+}^2(c) + \sigma_{Y,-}^2(c) \right) + \frac{\tau_Y^2}{\tau_W^4} \left(\sigma_{W,+}^2(c) + \sigma_{W,-}^2(c) \right) - \frac{2\tau_Y}{\tau_W^3} \left(\sigma_{YW,+}(c) + \sigma_{YW,-}(c) \right) \right).$$

In this expression the constants C_1 and C_2 are the same as before in Equation (3.6). The second derivatives of the regression functions, $m_{Y,-}^{(2)}(c)$, $m_{Y,+}^{(2)}(c)$, $m_{W,-}^{(2)}(c)$, and $m_{W,+}^{(2)}(c)$, are now defined separately for the treatment W and the outcome Y. In addition the conditional variances are indexed by the treatment and outcome. Finally the AMSE also depends on the right and left limit of the covariance of W and Y conditional on the forcing variable, at the threshold, denoted by $\sigma_{YW,+}(c)$ and $\sigma_{YW,-}(c)$ respectively.

The bandwidth that minimizes the AMSE in the fuzzy design is

$$h_{\text{opt,FRD}} = C_K \cdot N^{-1/5} \tag{5.23}$$

$$\times \left(\frac{\left(\sigma_{Y,+}^2(c) + \sigma_{Y,-}^2(c) \right) + \tau_{\text{FRD}}^2 \left(\sigma_{W,+}^2(c) + \sigma_{W,-}^2(c) \right) - 2\tau_{\text{FRD}} \left(\sigma_{YW,+}(c) + \sigma_{YW,-}(c) \right)}{f(c) \cdot \left(\left(m_{Y,+}^{(2)}(c) - m_{Y,-}^{(2)}(c) \right) - \tau_{\text{FRD}} \left(m_{W,+}^{(2)}(c) - m_{W,-}^{(2)}(c) \right) \right)^2} \right)^{1/5}.$$

The analogue of the bandwidth proposed for the sharp regression discontinuity is

$$\hat{h}_{\text{opt,FRD}} = C_K \cdot N^{-1/5}$$

$$\times \left(\frac{\left(\hat{\sigma}_{Y,+}^2(c) + \hat{\sigma}_{Y,-}^2(c) \right) + \hat{\tau}_{\text{FRD}}^2 \left(\hat{\sigma}_{W,+}^2(c) + \hat{\sigma}_{W,-}^2(c) \right) - 2\hat{\tau}_{\text{FRD}} \left(\hat{\sigma}_{YW,+}(c) + \hat{\sigma}_{YW,-}(c) \right)}{\hat{f}(c) \cdot \left(\left(\left(\hat{m}_{Y,+}^{(2)}(c) - \hat{m}_{Y,-}^{(2)}(c) \right) - \hat{\tau}_{\text{FRD}} \left(\hat{m}_{W,+}^{(2)}(c) - \hat{m}_{W,-}^{(2)}(c) \right) \right)^2 + \hat{r}_{Y,+} + \hat{r}_{Y,-} + \hat{\tau}_{\text{FRD}} \left(\hat{r}_{W,+} + \hat{r}_{W,-} \right) \right)} \right)^{1/5}.$$

We can implement this as follows. First, using the algorithm described for the sharp RD case separately for the treatment indicator and the outcome, calculate $\hat{\tau}_{\text{FRD}}$, $\hat{f}(c)$, $\hat{\sigma}_{Y,+}^2$, $\hat{\sigma}_{Y,-}^2$, $\hat{\sigma}_{W,+}^2$, $\hat{\sigma}_{W,-}^2$, $\hat{m}_{Y,+}^{(2)}(c)$, $\hat{m}_{Y,-}^{(2)}(c)$, $\hat{m}_{W,+}^{(2)}(c)$, $\hat{m}_{W,-}^{(2)}(c)$, $\hat{r}_{Y,+}$, $\hat{r}_{Y,-}$, $\hat{r}_{W,+}$, and $\hat{r}_{W,-}$. Second, using the initial Silverman bandwidth use the deviations from the means to estimate the conditional covariances $\hat{\sigma}_{YW,+}(c)$ and $\hat{\sigma}_{YW,-}(c)$. Then substitute everything into the expression for the bandwidth. By the same argument as for the sharp RD case the resulting bandwidth has the asymptotic no-regret property.

5.2 Additional covariates

Typically the presence of additional covariates does not affect the regression discontinuity analyses very much. In most cases the distribution of the additional covariates does not exhibit any discontinuity around the threshold for the forcing variable, and as a result those covariates are approximately independent of the treatment indicator for samples constructed to be close to the threshold. In that case the covariates only affect the precision of the estimator, and one can modify the previous analysis using the conditional variance of Y_i given all covariates at the threshold, $\sigma_{-}^{2}(c|x)$ and $\sigma_{+}^{2}(c|x)$ instead of the variances $\sigma_{-}^{2}(c)$ and $\sigma_{+}^{2}(c)$ that condition only on the forcing variable. In practice this modification does not affect the optimal bandwidth much unless the additional covariates have great explanatory power (recall that the variance enters to the power 1/5), and the basic algorithm is likely to perform adequately even in the presence of covariates. For example, if the conditional variances are half the size of the unconditional ones, using the basic algorithm with unconditional variances will mean that the bandwidth will be off only by a factor $(1 - 1/2^{1/5})$, or approximately 0.17.

6 An Illustration and Some Simulations

6.1 Data

To illustrate the implementation of these methods we use a data set previously analyzed by Lee (2008) in a recent influential application of regression discontinuity designs. Lee studies the incumbency advantage in elections. His identification strategy is based on the discontinuity generated by the rule that the party with a majority vote share wins. The forcing variable X_i is the difference in vote share between the Democratic and Republican parties in one election, with the threshold $c = 0$. The outcome variable Y_i is vote share at the second election. There are 6,558 observations (districts) in this data set, 3,818 with $X_i > 0$, and 2,740 with $X_i < 0$. The average difference in voting percentages at the last election for the Democrats was 0.13, with a standard deviation of 0.46. Figure 1 plots the density of the forcing variable, in bins with width 0.05. Figure 2 plots the average value of the outcome variable, in 40 bins with width 0.05, against the forcing variable. The discontinuity is clearly visible in the raw data, lending credibility to any positive estimate of the incumbency effect. The vertical line indicate the optimal bandwidth calculated below.

6.2 IK algorithm on Lee Data

In this section we implement our proposed bandwidth on the Lee dataset. For expositional reasons we gave all the intermediate steps.

Step 1: Estimation of density $f(0)$ and conditional variance $\sigma^2(0)$

We start with the modified Silverman bandwidth,

$$h_1 = 1.84 \cdot S_X \cdot N^{-1/5} = 1.84 \cdot 0.4553 \cdot 6558^{-1/5} = 0.1445.$$

There are $N_{h_1,-} = 836$ units with values for X_i in the interval $[-h_1, 0)$, with an average outcome of $\overline{Y}_{h_1,-} = 0.4219$ and a sample variance of $S^2_{Y,h_1,-} = 0.1047^2$, and $N_{h_1,+} = 862$ units with values for X_i in the interval $[0, h_1]$, with an average outcome of $\overline{Y}_{h_1,+} = 0.5643$ and a sample variance of $S^2_{Y,h_1,+} = 0.1202^2$. This leads to

$$\hat{f}(0) = \frac{N_{h_1,-} + N_{h_1,+}}{2 \cdot N \cdot h_1} = \frac{836 + 862}{2 \cdot 6558 \cdot 0.1445} = 0.8962,$$

and

$$\hat{\sigma}^2_-(0) = S^2_{Y,h_1,-} = 0.1047^2 \qquad \text{and} \quad \hat{\sigma}^2_+(0) = S^2_{Y,h_1,+} = 0.1202^2.$$

Step 2: Estimation of second derivatives $\hat{m}^{(2)}_+(0)$ and $\hat{m}^{(2)}_-(0)$

To estimate the curvature at the threshold, we first need to choose bandwidths $h_{2,+}$ and $h_{2,-}$. We choose these bandwidths based on an estimate of $\hat{m}^{(3)}(0)$, obtained by fitting a global cubic with a jump at the threshold:

$$Y_i = \gamma_0 + \gamma_1 \cdot 1_{X_i \geq c} + \gamma_2 \cdot (X_i - c) + \gamma_3 \cdot (X_i - c)^2 + \gamma_4 \cdot (X_i - c)^3 + \varepsilon_i,$$

The least squares estimate for γ_4 is $\hat{\gamma}_4 = -0.1686$, and thus the third derivative at the threshold is estimated as $\hat{m}^{(3)}(0) = 6 \cdot \hat{\gamma}_4 = -1.0119$. This leads to the two bandwidths

$$h_{2,+} = 3.56 \cdot \left(\frac{\hat{\sigma}^2_+(0)}{\hat{f}(0) \cdot \left(\hat{m}^{(3)}(0)\right)^2} \right)^{1/7} \cdot N_+^{-1/7} = 0.6057, \qquad \text{and} \quad h_{2,-} = 0.6105.$$

The two pilot bandwidths are used to fit two quadratics. The quadratic to the right of 0 is fitted on $[0, 0.6057]$, yielding $\hat{m}^{(2)}_+(0) = 0.0455$ and the quadratic to the left is fitted on [-0.6105, 0] yielding $\hat{m}^{(2)}_-(0) = -0.8471$.

Step 3: Calculation of Regularization Terms $\hat{r}_-$ and $\hat{r}_+$, and Calculation of $\hat{h}_{\text{opt}}$

Next, the regularization terms are calculated. We obtain

$$\hat{r}_+ = \frac{2160 \cdot \hat{\sigma}^2_+(0)}{N_{2,+} h^4_{2,+}} = \frac{2160 \cdot 0.1202^2}{2814 \cdot 0.6057^4} = 0.0825 \qquad \text{and} \quad \hat{r}_- = \frac{2160 \cdot \hat{\sigma}^2_-(0)}{N_{2,-} h^4_{2,-}} = 0.0675.$$

Now we have all the ingredients to calculate the optimal bandwidth under different kernels and the corresponding RD estimates. Using the edge kernel with $C_K = 3.4375$, we obtain

$$\hat{h}_{\text{opt}} = C_K \left(\frac{\hat{\sigma}^2_-(0) + \hat{\sigma}^2_+(0)}{\hat{f}(0) \cdot \left(\left(\hat{m}^{(2)}_+(0) - \hat{m}^{(2)}_-(0) \right)^2 + (\hat{r}_+ + \hat{r}_-) \right)} \right)^{1/5} N^{-1/5} = 0.2939.$$

6.3 Thirteen Estimates for the Lee Data

Here we calculate thirteen estimates of the ultimate object of interest, the size of the discontinuity in $m(x)$ at zero. The first eight are based on local linear regression, and the last five on global polynomial regressions. The first is based on our proposed bandwidth. The second drops the regularization terms. The third uses a normal kernel and the corresponding Silverman bandwidth for estimating the density function at the threshold ($h_1 = 1.06 \cdot S_x \cdot N^{-1/5}$). The fourth estimates separate cubic regressions on the left and the right of the threshold to derive the bandwidth for estimating the second derivatives. The fifth estimates the conditional variance at the threshold assuming its left and right limit are identical. The sixth uses a uniform kernel on $[-1/2, 1/2]$ instead of the optimal edge kernel. The seventh bandwidth is based on the DesJardin-McCall criterion, where we modify the procedure to use the edge kernel instead of the Epanechikov kernel that DesJardin-McCall use. The eighth bandwidth is based on the Ludwig-Miller cross-validation criterion. The last five estimates for τ_{SRD} are based on global linear, quadratic, cubic, quartic, and quintic regressions. The point estimates and robust standard errors are presented in Table 1. To investigate the overall sensitivity of the point estimates to the bandwidth choice, Figure 3 plots the RD estimates $\hat{\tau}_{\text{SRD}}(h)$, and the associated 95% confidence intervals, as a function of the bandwidth, for h between 0 and 1. The solid vertical line indicates the optimal bandwidth ($\hat{h}_{\text{opt}} = 0.2939$).

6.4 A Small Simulation Study

Next we conduct a small Monte Carlo study to assess the properties of the proposed bandwidth selection rule in practice. We consider four designs, the first based on the Lee data, the second on a very simple low order polynomial, and the third and fourth on a case of constant average treatment effect.

In the first design, based on the Lee data, we use a Beta distribution for the forcing variable. Let Z have a beta distribution with parameters $\alpha = 2$ and $\beta = 4$, then the forcing variable is $X = 2 \cdot Z - 1$. The regression function is a 5-th order polynomial, with separate coefficients for $X_i < 0$ and $X_i > 0$, with the coefficients estimated on the Lee data (after discarding observations with past vote share differences greater than 0.99 and less than -0.99), leading to

$$m_{\text{Lee}}(x) = \begin{cases} 0.48 + 1.27x + 7.18x^2 + 20.21x^3 + 21.54x^4 + 7.33x^5 & \text{if } x < 0, \\ 0.52 + 0.84x - 3.00x^2 + 7.99x^3 - 9.01x^4 + 3.56x^5 & \text{if } x \geq 0. \end{cases}$$

The error variance is $\sigma_\varepsilon^2 = 0.1295^2$. We use data sets of size 500 (smaller than the Lee data set with 6,558 observations, but more in line with common sample sizes).

In the second design we use the same distribution for the forcing variable as in the first design. We again have 500 observations per sample, and the true regression function is quadratic both to the left and to the right of the threshold, but with different coefficients:

$$m_{\text{quad}}(x) = \begin{cases} 3x^2 & \text{if } x < 0, \\ 4x^2 & \text{if } x \geq 0, \end{cases}$$

implying the data generating process is close to the point where the bandwidth h_{opt} is fairly large (because the left and right limit of the second derivative are 6 and 8 respectively), and one may expect some effect from the regularization. The error variance is the same as in the first design, $\sigma_\varepsilon^2 = 0.1295^2$.

Under the third design we have a constant additive treatment effect (CATE), and consequently the second derivatives on both sides of the threshold are equal. Here one might expect the DesJardins-McCall bandwidth to work particularly well, because it assumes equality of the second derivatives. We base the design on the Lee data, where we use the following regressions, where note that the regression for the treated group (right of the threshold) is an additive shift (of 0.1, approximately the discontinuity in the original sample) of the treatment effect regression for the control (left of threshold). In other words we test a scenario where the treatment effect is constant across values of the forcing variable.

$$m_{\text{CATE}(1)}(x) = 0.42 + 0.1 \cdot \mathbf{1}_{x \geq 0} + 0.84x - 3.00x^2 + 7.99x^3 - 9.01x^4 + 3.56x^5$$

Our fourth design is a modification of the above. We look at the the constant additive treatment effect case where the curvature at the threshold is zero on both sides (for instance, in locally linear regression functions). To do this, we simply use $m_{\text{CATE}(1)}(x)$, but set the coefficients on the squared term to zero:

$$m_{\text{CATE}(2)}(x) = 0.42 + 0.1 \cdot \mathbf{1}_{x \geq 0} + 0.84x + 7.99x^3 - 9.01x^4 + 3.56x^5.$$

The other parameters for the data generating process are set as in the simulations based on the Lee data. In the last two cases one might expect substantial effects from regularization because the infeasible optimal bandwidth in both cases is infinite. Moreover in the last case, even methods that are based on separately estimating left and right end points will need regularization.

In Tables 2 and 3 we report results for the same estimators as we reported in Table 1 for the real data. We include one additional bandwidth choice, namely the infeasible optimal bandwidth h_{opt}, which can be derived because we know the data generating process. In Table 2 and Table 3 we present for both designs in each case the mean (Mean) and standard deviation (Std) of the bandwidth choices, and the bias (Bias) and the root-mean-squared-error (RMSE) of the estimator for τ.

First consider the design motivated by the Lee data. All feasible bandwidth selection methods combined with local linear estimation perform fairly similarly under this design as far as $\hat{\tau}_{\text{SRD}}$ is concerned, and close to the infeasible h_{opt}. There is considerably more variation in the performance of the global polynomial estimators. The quadratic estimator performs very well, but adding a third order term increases both bias and RMSE. The quintic approximation does very well in terms of bias, not surprising given the regression that generated the data was a fifth order polynomial, but has a higher RMSE than the local methods.

In the second design the regularization matters, and the bandwidth choices based on different criterion functions perform worse than the proposed bandwidth in terms of RMSE, increasing it by about 35%. The global quadratic estimator obviously performs well here because it corresponds to the data generating process, but it is interesting that the local linear estimator based on $\hat{h}_{\text{opt}}$ has a RMSE very similar to that for the global quadratic estimator.

In the third and forth designs, as expected, regularization matters even more. Again the bandwidth choices based on different criterion functions perform worse. In particular in the case where the regression function has no curvature at the threshold, methods based on estimating end points separately perform poorly (RMSE nearly the size of the RD estimate itself). This is partly explained by the fact that in this case these bandwidth choices would benefit from regularization as well. Note that across all four simulations, the standard deviation of the estimated bandwidth with regularization is lower than that of the bandwidth without regularization, sometimes by a factor 10. This is because regularization has the added benefit of reducing the instability of the estimated bandwidth.

7 Conclusion

In this paper we propose a fully data-driven, asymptotically optimal bandwidth choice for regression discontinuity settings. Although this choice has asymptotic optimality properties, it still relies on somewhat arbitrary initial bandwidth choices. Rather than relying on a single bandwidth, we therefore encourage researchers to use this bandwidth choice as a reference point for assessing sensitivity to bandwidth choice in regression discontinuity settings. The bandwidth selection procedures commonly used in this literature are typically based on different objectives, for example on global measures, not tailored to the specific features of the regression discontinuity setting. We compare our proposed bandwidth selection procedure to these and find that our proposed method works well in realistic settings, including one motivated by data previously analyzed by Lee (2008).

Appendix

To obtain the MSE expansions for the RD estimand, we first obtain the bias and variance estimates from estimating a regression function at a boundary point. Fan and Gijbels (1992) derive a version of Lemma A.1 under different assumptions (such as thin tailed rather than compact kernels) and hence their proof is less transparent and not easily generalizable to multiple dimensions and derivatives. The proof we outline is based on Ruppert and Wand (1994) but since they only cursorily indicate the approach for a boundary point in multiple dimensions, we provide a simple proof for our case.

Lemma A.1: (MSE for Estimation of a Regression Function at the Boundary)
Suppose (i) we have N pairs (Y_i, X_i), independent and identically distributed, with $X_i \geq 0$, (ii), $m(x) = \mathbb{E}[Y_i|X_i = x]$ is three times continuously differentiable, (iii), the density of X_i, $f(x)$, is continuously differentiable at $x = 0$, with $f(0) > 0$, (iv), the conditional variance $\sigma^2(x) = \mathrm{Var}(Y_i|X_i = x) > 0$ is bounded, and continuous at $x = 0$, (v), we have a kernel $K : \mathbb{R}^+ \mapsto \mathbb{R}$, with $K(u) = 0$ for $u \geq \overline{u}$, and $\int_0^{\overline{u}} K(u)du = 1$, and define $K_h(u) = K(u/h)/h$. Define $\mu = m(0)$, and

$$(\hat{\mu}_h, \hat{\beta}_h) = \arg\min_{\mu,\beta} \sum_{i=1}^{N} (Y_i - \mu - \beta \cdot X_i)^2 \cdot K_h(X_i).$$

Then:

$$\mathbb{E}\left[\hat{\mu}\middle| X_1, \ldots, X_N\right] - \mu = C_1^{1/2} m^{(2)}(0) h^2 + o_p\left(h^2\right), \tag{A.1}$$

$$\mathbb{V}(\hat{\mu}| X_1, \ldots, X_N) = C_2 \frac{\sigma^2(0)}{f(c)Nh} + o_p\left(\frac{1}{Nh}\right), \tag{A.2}$$

and

$$\mathbb{E}\left[(\hat{\mu} - \mu)^2 \middle| X_1, \ldots, X_N\right] = C_1 \left(m^{(2)}(0)\right)^2 h^4 + C_2 \frac{\sigma^2(0)}{f(0)Nh} + o_p\left(h^4 + \frac{1}{Nh}\right), \tag{A.3}$$

where the kernel-specific constants C_1 and C_2 are those given in Lemma 3.1.

Before proving Lemma A.1, we state and prove two preliminary results.

Lemma A.2: *Define $F_j = \frac{1}{N}\sum_{i=1}^N K_h(X_i)X_i^j$. Under the assumptions in Lemma A.1, (i), for nonnegative integer j,*

$$F_j = h^j f(0)\nu_j + o_p(h^j) \equiv h^j(F_j^* + o_p(1)),$$

with $\nu_j = \int_0^\infty t^j K(t)dt$ and $F_j^* \equiv f(0)\nu_j$, and (ii), If $j \geq 1$, $F_j = o_p(h^{j-1})$.

Proof: F_j is the average of independent and identically distributed random variables, so

$$F_j = \mathbb{E}[F_j] + O_p\left(\mathrm{Var}(F_j)^{1/2}\right).$$

The mean of F_j is, using a change of variables from z to $x = z/h$,

$$\begin{aligned}\mathbb{E}[F_j] &= \int_0^\infty \frac{1}{h} K\left(\frac{z}{h}\right) z^j f(z)dz = h^j \int_0^\infty K(x)\, x^j f(hx)dx \\ &= h^j \int_0^\infty K(x)\, x^j f(0)dx + h^{j+1} \int_0^\infty K(x)\, x^{j+1} \frac{f(hx) - f(0)}{hx} dx = h^j f(0)\nu_j + O\left(h^{j+1}\right).\end{aligned}$$

The variance of F_j can be bounded by

$$\frac{1}{N}\mathbb{E}\left[(K_h(X_i))^2 X_i^{2j}\right] = \frac{1}{Nh^2}\mathbb{E}\left[\left(K\left(\frac{X_i}{h}\right)\right)^2 \cdot X_i^{2j}\right] = \frac{1}{Nh^2}\int_0^\infty \left(K\left(\frac{z}{h}\right)\right)^2 \cdot z^{2j} f(z)dz.$$

By a change of variables from z to $x = z/h$, this is equal to

$$\frac{h^{2j-1}}{N}\int_0^\infty (K(x))^2 \cdot x^{2j} f(hx)dx = O\left(\frac{h^{2j-1}}{N}\right) = o\left(\left(\frac{h^j}{hN^{1/2}}\right)^2\right) = o\left(\left(h^j\right)^2\right).$$

Hence

$$F_j = \mathbb{E}[F_j] + o_p\left(h^j\right) = h^j f(0)\nu_j + o_p\left(h^j\right) = h^j \cdot \left(f(0)\nu_j + o_p(1)\right).$$

□

Lemma A.3: *Let* $G_j = \frac{1}{N}\sum_{i=1}^{N} K_h^2(X_i) X_i^j \sigma^2(X_i)$. *Under the assumptions from Lemma A.1,*

$$G_j = h^{j-1}\sigma^2(0) f(0)\pi_j \left(1 + o_p\left(1\right)\right), \qquad \text{with} \quad \pi_j = \int_0^\infty t^j K^2(t) dt.$$

Proof: This claim is proved in a manner exactly like Lemma A.1, here using in addition the continuity of the conditional variance function. □

Proof of Lemma A.1: Define $R = [\iota \quad X]$, where ι is a N-dimensional column of ones, define the diagonal weight matrix W with (i,i)th element equal to $K_h(X_i)$, and define $e_1 = (1\ 0)'$. Then

$$\hat{m}(0) = \hat{\mu} = e_1'(R'WR)^{-1}R'WY.$$

The conditional bias is $B = \mathbb{E}[\hat{m}(0)|X_1, \ldots, X_N] - m(0)$. Note that $\mathbb{E}(\hat{m}(0)|X) = e_1'(R'WR)^{-1}R'WM$ where $M = (m(X_1), \ldots, m(X_N))'$. Let $m^{(k)}(x)$ denote the kth derivative of $m(x)$ with respect to x. Using Assumption (ii) in Lemma A.1, a Taylor expansion of $m(X_i)$ yields:

$$m(X_i) = m(0) + m^{(1)}(0)X_i + \frac{1}{2}m^{(2)}(0)X_i^2 + T_i,$$

where

$$|T_i| \leq \sup_x \left| m^{(3)}(x) \right| \cdot \left| X_i^3 \right|.$$

Thus we can write the vector M as

$$M = R \begin{pmatrix} m(0) \\ m^{(1)}(0) \end{pmatrix} + S + T.$$

where the vector S has ith element equal to $S_i = m^{(2)}(0)X_i^2/2$, and the vector T has typical element T_i. Therefore the bias can be written as

$$B = e_1'(R'WR)^{-1}R'WM - m(0) = e_1'(R'WR)^{-1}R'W(S+T).$$

Using Lemma A.2 we have

$$\begin{aligned}
\left(\frac{1}{N}R'WR\right)^{-1} &= \begin{pmatrix} F_0 & F_1 \\ F_1 & F_2 \end{pmatrix}^{-1} = \frac{1}{F_0F_2 - F_1^2}\begin{pmatrix} F_2 & -F_1 \\ -F_1 & F_0 \end{pmatrix} \\
&= \begin{pmatrix} \frac{F_2^*}{F_0^*F_2^* - (F_1^*)^2} + o_p\left(1\right) & -\frac{1}{h}\left(\frac{F_1^*}{F_0^*F_2^* - (F_1^*)^2} + o_p\left(1\right)\right) \\ -\frac{1}{h}\left(\frac{F_1^*}{F_0^*F_2^* - (F_1^*)^2} + o_p\left((1\right)\right) & \frac{1}{h^2}\left(\frac{F_0^*}{F_0^*F_2^* - (F_1^*)^2} + o_p\left(1\right)\right) \end{pmatrix} \\
&= \begin{pmatrix} \frac{\nu_2}{(\nu_0\nu_2 - \nu_1^2)f(0)} + o_p\left(1\right) & -\frac{\nu_1}{(\nu_0\nu_2 - \nu_1^2)f(0)h} + o_p\left(\frac{1}{h}\right) \\ -\frac{\nu_1}{(\nu_0\nu_2 - \nu_1^2)f(0)h} + o_p\left(\frac{1}{h}\right) & O_p\left(\frac{1}{h^2}\right) \end{pmatrix} \\
&= \begin{pmatrix} O_p\left(1\right) & O_p\left(\frac{1}{h}\right) \\ O_p\left(\frac{1}{h}\right) & O_p\left(\frac{1}{h^2}\right) \end{pmatrix}.
\end{aligned}$$

Next

$$\left|\frac{1}{N}R'WT\right| = \sup_x m^{(3)}(x) \cdot \begin{pmatrix} F_3 \\ F_4 \end{pmatrix} \leq \begin{pmatrix} o_p\left(h^2\right) \\ o_p\left(h^3\right) \end{pmatrix}.$$

Thus

$$e_1'(R'WR)^{-1}R'WT = O_p(1) \cdot o_p(h^2) + O_p\left(\frac{1}{h}\right) \cdot o_p(h^3) = o_p\left(h^2\right),$$

implying

$$B = e_1'(R'WR)^{-1}R'WS + o_p(h^2).$$

Similarly:

$$\frac{1}{N}(R'WS) = \frac{1}{2}m^{(2)}(0)\begin{pmatrix} \frac{1}{N}\sum_{i=1}^N K_h(X_i)X_i^2 \\ \frac{1}{N}\sum_{i=1}^N K_h(X_i)X_i^3 \end{pmatrix} = \frac{1}{2}m^{(2)}(0)f(0)\begin{pmatrix} \nu_2 h^2 + o_p(h^2) \\ \nu_3 h^3 + o_p(h^3) \end{pmatrix}.$$

Therefore:

$$B = e_1'(R'WR)^{-1}R'WS + o_p(h^2) = \frac{1}{2}m^{(2)}(0)\left(\frac{\nu_2^2 - \nu_3\nu_1}{\nu_0\nu_2 - \nu_1^2}\right)h^2 + o_p(h^2).$$

This finishes the proof for the first part of the result in Lemma A.1, equation (A.1).
Next, we consider the expression for the conditional variance in (A.2).

$$V = \mathbb{V}(\hat{m}(0)|X_1,\ldots,X_N) = e_1'(R'WR)^{-1}R'W\Sigma WR(R'WR)^{-1}e_1,$$

where Σ is the diagonal matrix with (i,i)th element equal to $\sigma^2(X_i)$.
Consider the middle term

$$\frac{1}{N}R'W\Sigma WR = \begin{pmatrix} \frac{1}{N}\sum_i K_h^2(X_i)\sigma^2(X_i) & \frac{1}{N}\sum_i K_h^2(X_i)X_i\sigma^2(X_i) \\ \frac{1}{N}\sum_i K_h^2(X_i)X_i\sigma^2(X_i) & \frac{1}{n}\sum_i K_h^2(X_i)X_i^2\sigma^2(X_i) \end{pmatrix} = \begin{pmatrix} G_0 & G_1 \\ G_1 & G_2 \end{pmatrix}.$$

Thus we have:

$$\begin{aligned} NV &= \frac{1}{(F_0F_2 - F_1^2)^2}e_1'\begin{pmatrix} F_2 & -F_1 \\ -F_1 & F_0 \end{pmatrix}\begin{pmatrix} G_0 & G_1 \\ G_1 & G_2 \end{pmatrix}\begin{pmatrix} F_2 & -F_1 \\ -F_1 & F_0 \end{pmatrix}e_1 \\ &= \frac{F_2^2G_0 - 2F_1F_2G_1 + F_1^2G_2}{(F_0F_2 - F_1^2)^2} \end{aligned}$$

Applying lemmas A.2 and A.3 this leads to

$$V = \frac{\sigma^2(0)}{f(0)Nh}\cdot\left(\frac{\nu_2^2\pi_0 - 2\nu_1\nu_2\pi_1 + \nu_1^2\pi_2}{(\nu_0\nu_2 - \nu_1^2)^2}\right) + o_p\left(\frac{1}{Nh}\right).$$

This finishes the proof for the statement in (A.2). The final result in (A.3) follows directly from the first two results. □

Proof of Lemma 3.1: Applying Lemma A.1 to the N_+ units with $X_i \geq c$, implies that

$$\mathbb{E}\left[\hat{\mu}_+ - \mu_+\right| X_1,\ldots,X_N] = C_1^{1/2}m_+^{(2)}(c)h^2 + o_p\left(h^2\right),$$

and

$$\mathbb{V}\left(\hat{\mu}_+ - \mu_+\right| X_1,\ldots,X_N) = C_2\frac{\sigma_+^2(c)}{f_{X|X\geq c}(c)N_+h} + o_p\left(\frac{1}{N_+h}\right).$$

Because $N_+/N = \text{pr}(X_i \geq c) + O_p(1/N)$, and $f_{X|X\geq c}(x) = f(x)/\Pr(X_i \geq c)$, it follows that

$$\mathbb{V}\left(\hat{\mu}_+ - \mu_+\right| X_1,\ldots,X_N) = C_2\frac{\sigma_+^2(c)}{f(c)Nh} + o_p\left(\frac{1}{Nh}\right).$$

Conditional on $X_1,\ldots,X_N$ the covariance between $\hat{\mu}_+$ and $\hat{\mu}_-$ is zero, and thus, combining the results from applying Lemma A.1 also to the units with $X_i < c$, we find

$$\begin{aligned} &\mathbb{E}\left[(\hat{\tau}_{\text{SRD}} - \tau_{\text{SRD}})^2\right| X_1,\ldots,X_N] = \mathbb{E}\left[(\hat{\mu}_+ - \hat{\mu}_- - (\mu_+ - \mu_-))^2\right| X_1,\ldots,X_N] \\ &\quad = \mathbb{E}\left[(\hat{\mu}_+ - \mu_+)^2\right| X_1,\ldots,X_N] + \mathbb{E}\left[(\hat{\mu}_- - \mu_-)^2\right| X_1,\ldots,X_N] \\ &\qquad -2\cdot\mathbb{E}\left[\hat{\mu}_+ - \mu_+\right| X_1,\ldots,X_N]\cdot\mathbb{E}\left[\hat{\mu}_- - \mu_-\right| X_1,\ldots,X_N] \\ &\quad = C_1\cdot h^4\cdot\left(m_+^{(2)}(c) - m_-^{(2)}(c)\right)^2 + \frac{C_2}{N\cdot h}\cdot\left(\frac{\sigma_+^2(c)}{f(c)} + \frac{\sigma_-^2(c)}{f(c)}\right). + o_p\left(h^4 + \frac{1}{N\cdot h}\right), \end{aligned}$$

proving the first result in Lemma 3.1.
For the second part of Lemma 3.1, solve

$$h_{\text{opt}} = \arg\min_h\left(C_1h^4\left(m_+^{(2)}(c) - m_-^{(2)}(c)\right)^2 + C_2\left(\frac{\sigma_+^2(c)}{f(c)Nh} + \frac{\sigma_-^2(c)}{f(c)Nh}\right)\right),$$

which leads to

$$h_{\text{opt}} = \left(\frac{C_2}{4C_1}\right)^{1/5}\left(\frac{\frac{\sigma_+^2(c)}{f(c)} + \frac{\sigma_-^2(c)}{f(c)}}{\left(m_+^{(2)}(c) - m_-^{(2)}(c)\right)^2}\right)^{1/5}N^{-1/5}.$$

□

Motivation for the Bandwidth Choice in Equation (4.15) in Step 2 of bandwidth algorithm
Fan and Gijbels (1996 Theorem 3.2) give an asymptotic approximation to the MSE for an estimator of the ν-th derivative of a regression function at a boundary point, using a p-th order local polynomial (using the notation in Fan and Gijbels). Specializing this to our case, with the boundary point c, a uniform one-sided kernel $K(t) = 1_{0\le t\le 1}$, and interest in the 2-nd derivative using a local quadratic approximation ($\nu = p = 2$, their MSE formula simplifies to

$$\text{MSE} = \left(\frac{1}{9}K_1^2\left(m_+^{(3)}(c)\right)^2 h^2 + 4K_2\frac{1}{Nh^5}\frac{\sigma_+^2(c)}{f(c)}\right)(1+o_p(1))$$

Here

$$K_1 = \int t^3 K^*(t)dt \qquad K_2 = \int (K^*(t))^2 dt,$$

where

$$K^*(t) = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}' \begin{pmatrix} \mu_0 & \mu_1 & \mu_2 \\ \mu_1 & \mu_2 & \mu_3 \\ \mu_2 & \mu_3 & \mu_4 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ t \\ t^2 \end{pmatrix} \cdot K(t), \qquad \text{with } \ \mu_k = \int q^k K(q)dq = \frac{1}{(k+1)},$$

so that

$$K^*(t) = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}' \begin{pmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ t \\ t^2 \end{pmatrix} \cdot K(t) = (30 - 180t + 180t^2)\cdot \mathbf{1}_{[0,1]},$$

and therefore, $K_1 = 1.5$ and $K_2 = 180$. Thus

$$\text{MSE} = \left(\frac{1}{4}\left(m_+^{(3)}(c)\right)^2 h^2 + 720\frac{1}{Nh^5}\frac{\sigma_+^2(c)}{f_+(c)}\right)(1+o_p(1)).$$

Minimizing this over h leads to

$$h_{2,+} = 7200^{1/7}\cdot\left(\frac{\sigma_+^2(c)}{f(c)\left(m_+^{(3)}(c)\right)^2}\right)^{1/7} N_+^{-1/7} \approx 3.56\cdot\left(\frac{\sigma_+^2(c)}{f(c)\left(m_+^{(3)}(c)\right)^2}\right)^{1/7} N_+^{-1/7}.$$

□

Proof of Theorem 4.1: Before directly proving the three claims in the theorem, we make some preliminary observations. Write

$$h_{\text{opt}} = C_{\text{opt}}\cdot N^{-1/5}, \qquad \text{with } \ C_{\text{opt}} = C_K\cdot\left(\frac{\sigma_-^2(c)+\sigma_+^2(c)}{f(c)\cdot\left(\left(m_+^{(2)}(c) - m_-^{(2)}(c)\right)^2\right)}\right)^{1/5},$$

and

$$\hat{h}_{\text{opt}} = \hat{C}_{\text{opt}}\cdot N^{-1/5}, \qquad \text{with } \ \hat{C}_{\text{opt}} = C_K\cdot\left(\frac{\hat{\sigma}_-^2(c)+\hat{\sigma}_+^2(c)}{\hat{f}(c)\cdot\left(\left(\hat{m}_+^{(2)}(c) - \hat{m}_-^{(2)}(c)\right)^2 + \hat{r}_+ + \hat{r}_-\right)}\right)^{1/5}.$$

First we show that the various estimates of the functionals in $\hat{C}_{\text{opt}}$, $\hat{\sigma}_-^2(c)$, $\hat{\sigma}_+^2(c)$, $\hat{f}(c)$, $\hat{m}_+^{(2)}(c)$ and $\hat{m}_-^{(2)}(c)$ converge to their counterparts in C_{opt}, $\sigma_-^2(c)$, $\sigma_+^2(c)$, $f(c)$, $m_+^{(2)}(c)$ and $m_-^{(2)}(c)$. Consider $\hat{f}(c)$. This is a histogram estimate of density at c, with bandwidth $h = CN^{-1/5}$. Hence $\hat{f}(c)$ is consistent for $f(c)$ if $f_-(c) = f_+(c) = f(c)$, if the left and righthand limit are equal, and for $(f_-(c) + f_+(c))/2$ if they are different.
Next, consider $\hat{\sigma}_-^2(c)$ (and $\hat{\sigma}_+^2(c)$). Because it is based on a bandwidth $h = C\cdot N^{-1/5}$ that converges to zero, it is consistent for $\sigma_-^2(c)$ if $\sigma_-^2(c) = \sigma_+^2(c) = \sigma^2(c)$.

Third, consider $\hat{m}_+^{(2)}(c)$. This is a local quadratic estimate using a one sided uniform kernel. From Fan and Gijbels (1996), Theorem 3.2, it follows that to guarantee consistency of $\hat{m}_+^{(2)}(c)$ for $m_+^{(2)}(c)$ we need both

$$h_{2,+} = o_p(1) \qquad \text{and} \quad \left(N h_{2,+}^5\right)^{-1} = o_p(1). \tag{A.4}$$

Let m_3 be the probability limit of $\hat{m}^{(3)}(c)$. This probability limit need not be equal to $m^{(3)}(c)$, but it will exist under the assumptions in Theorem 4.1. As long as this probability limit differs from zero, then $h_{2,+} = O_p(N^{-1/7})$, so that the two conditions in (A.4) are satisfied and $\hat{m}_+^{(2)}(c)$ is consistent for $m_+^{(2)}(c)$.

Fourth, consider $\hat{r}_+ = 2160\hat{\sigma}_+^2(c)/(N_{2,+}h_{2,+}^4)$. The numerator converges to $2160\sigma_+^2(c)$. The denominator is $N_{2,+}\cdot h_{2,+}^4 = C\cdot(N\cdot h_{2,+})\cdot N^{-4/7}(1+o_p(1)) = C\cdot N^{2/7}(1+o_p(1))$, so that the ratio is $C\cdot N^{-2/7}(1+o_p(1)) = o_p(1)$. A similar result holds for $\hat{r}_-$.

Now we turn to the statements in Theorem 4.1. We will prove (iii), then (iv), and then (i) and (ii). First consider (iii). If $m_+^{(2)}(c) - m_-^{(2)}(c)$ differs from zero, then C_{opt} is finite. Moreover, in that case $(\hat{m}_+^{(2)}(c) - \hat{m}_-^{(2)}(c))^2 + \hat{r}_+ + \hat{r}_-$ converges to $(\hat{m}_+^{(2)}(c) - \hat{m}_-^{(2)}(c))^2$, and $\hat{C}_{\text{opt}}$ converges to C_{opt}. These two implications in turn lead to the result that $(\hat{h}_{\text{opt}} - h_{\text{opt}})/h_{\text{opt}} = (\hat{C}_{\text{opt}} - C_{\text{opt}})/C_{\text{opt}} = o_p(1)$, finishing the proof for (iii).

Next, we prove (iv). Because $h_{\text{opt}} = C_{\text{opt}}\cdot N^{-1/5}$, it follows that

$$\text{MSE}\left(h_{\text{opt}}\right) = \text{AMSE}\left(h_{\text{opt}}\right) + o_p\left(h_{\text{opt}}^4 + \frac{1}{N\cdot h_{\text{opt}}}\right) = \text{AMSE}\left(h_{\text{opt}}\right) + o_p\left(N^{-4/5}\right).$$

Because $\hat{h}_{\text{opt}} = (\hat{C}_{\text{opt}}/C_{\text{opt}})\cdot C_{\text{opt}}N^{-1/5}$, and $\hat{C}_{\text{opt}}/C_{\text{opt}} \to 1$ it follows that

$$\text{MSE}\left(\hat{h}_{\text{opt}}\right) = \text{AMSE}\left(\hat{h}_{\text{opt}}\right) + o_p\left(N^{-4/5}\right).$$

Therefore

$$N^{4/5}\cdot\left(\text{MSE}\left(\hat{h}_{\text{opt}}\right) - \text{MSE}\left(h_{\text{opt}}\right)\right) = N^{4/5}\cdot\left(\text{AMSE}\left(\hat{h}_{\text{opt}}\right) - \text{AMSE}\left(h_{\text{opt}}\right)\right) + o_p(1),$$

and

$$\begin{aligned}
\frac{\text{MSE}\left(\hat{h}_{\text{opt}}\right) - \text{MSE}\left(h_{\text{opt}}\right)}{\text{MSE}\left(h_{\text{opt}}\right)} &= \frac{N^{4/5}\cdot\left(\text{MSE}\left(\hat{h}_{\text{opt}}\right) - \text{MSE}\left(h_{\text{opt}}\right)\right)}{N^{4/5}\cdot\text{MSE}\left(h_{\text{opt}}\right)} \\
&= \frac{N^{4/5}\cdot\left(\text{AMSE}\left(\hat{h}_{\text{opt}}\right) - \text{AMSE}\left(h_{\text{opt}}\right)\right) + o_p(1)}{N^{4/5}\cdot\text{AMSE}\left(h_{\text{opt}}\right) + o_p(1)}. \\
&= \frac{N^{4/5}\cdot\left(\text{AMSE}\left(\hat{h}_{\text{opt}}\right) - \text{AMSE}\left(h_{\text{opt}}\right)\right)}{N^{4/5}\cdot\text{AMSE}\left(h_{\text{opt}}\right)} + o_p(1).
\end{aligned}$$

Because $N^{4/5}\cdot\text{AMSE}(h_{\text{opt}})$ converges to a nonzero constant, all that is left to prove in order to establish (iv) is that

$$N^{4/5}\cdot\left(\text{AMSE}\left(\hat{h}_{\text{opt}}\right) - \text{AMSE}\left(h_{\text{opt}}\right)\right) = o_p(1). \tag{A.5}$$

Substituting in, we have

$$\begin{aligned}
&N^{4/5}\cdot\left(\text{AMSE}\left(\hat{h}_{\text{opt}}\right) - \text{AMSE}\left(h_{\text{opt}}\right)\right) \\
&= C_1\cdot\left(m_+^{(2)}(c) - m_-^{(2)}(c)\right)^2\cdot\left((N^{1/5}h_{\text{opt}})^4 - N^{1/5}\hat{h}_{\text{opt}})^4\right) + \left(\frac{C_2}{N^{1/5}\cdot h_{\text{opt}}} - \frac{C_2}{N^{1/5}\cdot\hat{h}_{\text{opt}}}\right)\cdot\left(\frac{\sigma_+^2(c)}{f(c)} + \frac{\sigma_-^2(c)}{f(c)}\right) \\
&= o_p(1),
\end{aligned}$$

because $N^{1/5}h_{\text{opt}} - N^{1/5}\hat{h}_{\text{opt}} = C_{\text{opt}} - \hat{C}_{\text{opt}} = o_p(1)$, so that A.5 holds, and therefore (iv) holds.

Now we turn to (i). If Assumption 3.6 holds, $\hat{h}_{\text{opt}} = \hat{C}_{\text{opt}}N^{-1/5}$, with $\hat{C}_{\text{opt}} \to C_{\text{opt}}$, a nonzero constant. Then Lemma 3.1 implies that $\text{MSE}(\hat{h}_{\text{opt}})$ is $O_p(\hat{h}_{\text{opt}}^4 + N^{-1}\hat{h}_{\text{opt}}^{-1}) = O_p(N^{-4/5})$ so that $\hat{\tau}_{\text{SRD}} - \tau_{\text{SRD}} = O_p(N^{-2/5})$.

Next consider (ii). If Assumption 3.6 does not hold and $m_+^{(2)}(c) - m_+^{(2)}(c) = 0$. Because $h_{2,+} = CN^{-1/7}$, it follows that $r_+ = CN_{2,+}^{-1}h^{-4} = CN^{-2/7}(1 + o_p(1))$ (with each time different constants C), it follows that $\hat{h}_{\text{opt}} = C(N^{2/7})^{1/5}N^{-1/5} = CN^{-1/7}$, so that the $\text{MSE}(h) = CN^{-6/7} + \tilde{C}N^{-6/7} = CN^{-6/7}$ (note that the leading bias term is now $O(h^3)$ so that the square of the bias is $O(h^6) = O(N^{-6/7})$) and thus $\hat{\tau}_{\text{SRD}} - \tau_{\text{SRD}} = O_p(N^{-3/7})$, and thus the result holds. □

References

Cook, T., (2008), ""Waiting for Life to arrive": A history of the regression-discontinuity design in Psychology, Statistics and Economics," *Journal of Econometrics*, 142, 636-654.

Cheng, M.-Y., Fan, J. and Marron, J.S., (1997), "On Automatic Boundary Corrections," *The Annals of Statistics*, 25, 1691-1708.

DesJardins, S., and B. 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," Unpublished Manuscript.

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

Fan, J. and Gijbels, I., (1996), *Local polynomial modeling and its implications*, Monographs on Statistics and Applied Probability 66, Chapman and Hall/CRC, Boca Raton, FL.

Frandsen, B., (2008), "A Nonparametric Estimator for Local Quantile Treatment Effects in the Regression Discontinuity Design," Unpublished Working Paper, Dept of Economics, MIT.

Frölich, M., (2007), "Regression Discontinuity Design with Covariates," IZA Discussion Paper 3024 Bonn.

Frölich, M., and B. Melly, (2008), "Quantile Treatment Effects in the Regression Discontinuity Design," IZA Discussion Paper 3638, Bonn.

Hahn, J., Todd, P., and Van Der Klaauw, W., (2001), "Regression discontinuity," *Econometrica*, 69(1), 201-209.

Härdle, W., (1992), *Applied Nonparametric Regression*, Cambridge University Press, Cambridge, UK.

Imbens, G., and J. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects,"*Econometrica*, Vol. 61, No. 2, 467-476.

Imbens, G., and Lemieux, T., (2008), "Regression discontinuity designs," *Journal of Econometrics*, 142, 615-635.

Kalyanaraman, K., (2008), "Bandwidth selection for linear functionals of the regression function," Working Paper, Harvard University Department of Economics, June, 2008.

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

Lee, D., and T. Lemieux, (2010), "Regression Discontinuity Designs in Economics," *Journal of Economic Literature*, 48(2): 281-355.

Li, K.-C., (1987), "Asymptotic Optimality for C_p, C_L, Cross-validation and Generalized Cross-validation: Discrete Index Set,"*The Annals of Statistics*, vol. 15(3), 958-975.

Ludwig, J. and Miller, D., (2005), "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design," NBER Working Paper 11702.

Ludwig, J. and Miller, D., (2007), "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design," *Quarterly Journal of Economics*.

McCrary, J., (2008), "Manipulation of the running variable in the regression discontinuity design: A density test," *Journal of Econometrics*, 142, 698-714.

Porter, J., (2003), "Estimation in the Regression Discontinuity Model," Working Paper, Harvard University Department of Economics, draft date September 25, 2003.

Powell, J., and T. Stoker, (1996), "Optimal Bandwidth Choice for Density Weighted Averages," *Journal of Econometrics*, Vol 75, 291-316

Rubin, D. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies," *Journal of Educational Psychology*, 66, 688-701.

Ruppert, D. and Wand, M.P., (1994), "Multivariate locally weighted least squares regression," *The Annals of Statistics*, 22, 1346-1370

Shadish, W., T. Campbell and D. Cook, (2002), *Experimental and Quasi-experimental Designs for Generalized Causal Inference*, Houghton and Mifflin, Boston.

Stone, C., (1982), "Optimal global rates of convergence for nonparametric regression," *The Annals of Statistics*, 10, 1040-1053

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

Van Der Klaauw, W., (2008), "Regression–Discontinuity Analysis: A Survey of Recent Developments in Economics," *Labour*, 22(2): 219-245.

Wand, M. and M. Jones, (1994), *Kernel Smoothing*, Chapman and Hall.

Table 1: RD estimates and bandwidths for Lee Data

Procedure	h	$\hat{\tau}_{\text{SRD}}$	(s.e.)
$\hat{h}_{\text{opt}}$	0.2939	0.0799	0.0083
no regularization	0.3042	0.0802	0.0082
$f(c)$ estimated using normal kernel	0.2938	0.0799	0.0083
third order polynomial separate on left and right	0.2546	0.0774	0.0089
homoskedastic variance	0.2940	0.0799	0.0083
uniform kernel	0.4617	0.0806	0.0087
DesJardin-McCall	0.3105	0.0804	0.0081
Ludwig-Miller cross-validation ($\delta = 0.5$)	0.9750	0.0788	0.0056
Linear	global	0.1182	0.0056
Quadratic	global	0.0519	0.0071
Cubic	global	0.1115	0.093
Quartic	global	0.0766	0.0113
Quintic	global	0.0433	0.0132

Table 2: Simulations, 5,000 Replications

	$\hat{h}$		$\hat{\tau}_{\text{SRD}}$	
	Mean	Std	Bias	RMSE
Lee Design				
h_{opt} (infeasible)	0.166		0.017	0.060
$\hat{h}_{\text{opt}}$	0.480	0.058	0.040	0.054
no regularization	0.757	0.680	0.037	0.051
$f(c)$ estimated using normal kernel	0.480	0.058	0.040	0.054
third order polynomial separate on left and right	0.336	0.037	0.038	0.056
homoskedastic variance	0.478	0.058	0.041	0.054
uniform kernel	0.377	0.046	0.034	0.056
DesJardins-McCall	0.556	0.134	0.037	0.051
Ludwig-Miller cross-validation ($\delta = 0.5$)	0.423	0.115	0.037	0.054
Linear	global		0.048	0.055
Quadratic	global		-0.019	0.043
Cubic	global		0.087	0.100
Quartic	global		0.028	0.068
Quintic	global		0.001	0.074
Quadratic Design				
h_{opt} (infeasible)	0.418		0.003	0.037
$\hat{h}_{\text{opt}}$	0.422	0.070	0.006	0.036
no regularization	0.473	0.268	0.015	0.045
$f(c)$ estimated using normal kernel	0.422	0.070	0.006	0.036
third order polynomial separate on left and right	0.372	0.060	0.003	0.040
homoskedastic variance	0.421	0.070	0.006	0.036
uniform kernel	0.332	0.055	-0.041	0.067
DesJardins-McCall	0.223	0.010	-0.002	0.049
Ludwig-Miller cross-validation ($\delta = 0.5$)	0.220	0.023	-0.002	0.050
Linear	global		0.245	0.251
Quadratic	global		-0.000	0.037
Cubic	global		-0.000	0.048
Quartic	global		-0.000	0.060
Quintic	global		-0.000	0.073

Table 3: Simulations, 5,000 Replications

	$\hat{h}$		$\hat{\tau}_{\text{SRD}}$	
	Mean	Std	Bias	RMSE
CATE(1), non-zero curvature				
h_{opt} (infeasible)	∞		-3.758	3.767
$\hat{h}_{\text{opt}}$	0.174	0.016	-0.008	0.058
no regularization	0.257	0.206	-0.067	0.303
$f(c)$ estimated using normal kernel	0.174	0.016	-0.008	0.058
third order polynomial separate on left and right	0.164	0.013	-0.007	0.059
homoskedastic variance	0.175	0.016	-0.009	0.058
uniform kernel	0.137	0.013	0.003	0.069
DesJardins-McCall	0.206	0.045	-0.015	0.065
Ludwig-Miller cross-validation ($\delta = 0.5$)	0.113	0.013	-0.003	0.073
Linear	global		-3.758	3.767
Quadratic	global		1.367	1.373
Cubic	global		-0.207	0.214
Quartic	global		0.015	0.062
Quintic	global		-0.000	0.074
CATE(2), zero curvature				
h_{opt} (infeasible)	∞		-3.453	3.462
$\hat{h}_{\text{opt}}$	0.173	0.016	-0.007	0.057
no regularization	0.252	0.184	-0.055	0.260
$f(c)$ estimated using normal kernel	0.173	0.016	-0.007	0.057
third order polynomial separate on left and right	0.163	0.013	-0.006	0.058
homoskedastic variance	0.172	0.016	-0.007	0.057
uniform kernel	0.135	0.013	-0.003	0.068
DesJardins-McCall	0.239	0.073	-0.026	0.095
Ludwig-Miller cross-validation ($\delta = 0.5$)	0.120	0.011	-0.004	0.069
Linear	global		-3.453	3.462
Quadratic	global		1.365	1.371
Cubic	global		-0.209	0.216
Quartic	global		0.015	0.061
Quintic	global		-0.000	0.073