Journal of Econometrics 189 (2015) 132–147

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Regression discontinuity designs with unknown discontinuity points: Testing and estimation☆

Jack Porter[a], Ping Yu[b,*]

[a] Department of Economics, University of Wisconsin, 6448 Social Science Building, 1180 Observatory Drive, Madison, WI 53706-1393, USA

[b] School of Economics and Finance, The University of Hong Kong, Pokfulam Road, Hong Kong

ARTICLE INFO

Article history:
Received 19 March 2014
Received in revised form
23 November 2014
Accepted 11 June 2015
Available online 20 July 2015

JEL classification:
C12
C13
C14
C21

Keywords:
Regression discontinuity design
Unknown discontinuity point
Specification testing
Nonparametric structural change
Wild bootstrap
Degenerate *U*-statistic
Difference kernel estimator
Cross validation

ABSTRACT

The regression discontinuity design has become a common framework among applied economists for measuring treatment effects. A key restriction of the existing literature is the assumption that the discontinuity point is known, which does not always hold in practice. This paper extends the applicability of the regression discontinuity design by allowing for an unknown discontinuity point. First, we construct a unified test statistic to check whether there are selection or treatment effects. Our tests are shown to be consistent, and local powers are derived. Also, a bootstrap method is proposed to obtain critical values. Second, we estimate the treatment effect by first estimating the nuisance discontinuity point. It is shown that estimating the discontinuity point does not affect the efficiency of the treatment effect estimator. Simulation studies illustrate the usefulness of our procedures in finite samples.

1. Introduction

Since its invention by Thistlethwaite and Campbell (1960), the regression discontinuity design (RDD) has attracted much attention among econometricians; see Imbens and Lemieux (2008), van der Klaauw (2008) and Lee and Lemieux (2010) for excellent reviews on up-to-date theoretical developments and applications and Yu (2013) for a summary of treatment effects estimators in RDDs. In RDDs, an observable covariate is used to completely determine the treatment status, and is called the *forcing* (*running* or *assignment*) variable. When the value of the covariate for an individual is above a threshold or discontinuity point, the individual will be treated; otherwise, the individual will be put in the control group. Usually, the discontinuity point is set by the policy maker and is publicly known. However, such information is not always available in practice. Sometimes, the discontinuity point is only known to the policy maker but is unknown to the public (including econometricians) due to ethical reasons or privacy. In the classical application of RDDs in the effect of the scholarship offers on student enrollment decisions by van der Klaauw (2002), the forcing variable is an underlying index of various individual characteristics. To avoid manipulation by individuals or competition from other schools, the discontinuity point may not be disclosed. Another example with an unknown discontinuity point is Card et al. (2008) who analyze the tipping effect in the dynamic of segregation. Specifically, when the minority share in a neighborhood exceeds a "tipping point", all the whites leave. Such a tipping point depends on the strength of white distaste for minority neighbors and is generally unknown.

To date, all existing literature on RDDs assumes that the discontinuity point is known, especially to econometricians. This paper studies the testing and estimation problem when the discontinu-

☆ Thanks go to the co-editor Jianqing Fan, one associate editor, two referees and the seminar participants at Auckland, Northwestern, and ESAM 2011 for helpful comments. The second author wants to thank FRDF at the University of Auckland for financial supports.

* Corresponding author.
E-mail addresses: jrporter@ssc.wisc.edu (J. Porter), pingyu@hku.hk (P. Yu).

http://dx.doi.org/10.1016/j.jeconom.2015.06.002
0304-4076/

ity point is unknown. In testing, we try to check whether there are selection or treatment effects in our experiment. The first test is to check whether there is selection among individuals. This test is new in the literature. The second test is to check the presence of treatment effects. This test is very close to the nonparametric structural change test, and there have been at least three tests designed for this purpose in the statistical literature. Our test is inspired by the nonparametric specification testing literature started from Bierens (1982), and is novel in our context. We solve both testing problems by a unified test statistic, but adapt it to different problems by varying a smoothing parameter. The test statistic is constructed under the null, so is similar to the score test in spirit. In estimation, our main interest lies in the treatment effects evaluation, but a primary input, the discontinuity point, is unknown. So we estimate the discontinuity point first by an estimator called the *difference kernel estimator* (DKE). We show its superconsistency and find its asymptotic distribution, and then estimate the treatment effect as if the discontinuity point were known. It is surprising that estimation of the discontinuity point does not affect the efficiency of the treatment effect estimator asymptotically. The model we consider has many applications beyond RDDs; see Müller (1992), Wu and Chu (1993a), Wang (1995), Müller and Stadtmüller (1999), and the reference therein for applications in statistics.

This paper is organized as follows. In Section 2, we set up our framework and specify some regularity assumptions. Especially, we clarify what selection means and what it implies to observations. Section 3 presents our specification test statistic and develops its asymptotic distributions in different testing problems. Furthermore, a bootstrap method is suggested to obtain critical values which may have better finite sample performances. Alternative tests of nonparametric structural change in the statistical literature are also reviewed and compared with our test. Section 4 considers the estimation problem. We provide estimators of the discontinuity point and the treatment effect, and develop their asymptotic distributions. In Section 5, we extend the results in Sections 3 and 4 to other settings and solve an important practical issue, the bandwidth selection, in both specification testing and estimation. Section 6 includes some simulation results and Section 7 concludes. To save space, we put some intuitions and all technical proofs in the online supplementary materials (see Appendix A).

Throughout this paper, we concentrate on the sharp design and discuss the fuzzy design only briefly in Sections 5.1 and 5.2. Such an arrangement allows us to focus on the main idea of this paper. In the sharp design, we assume that only the response variable and the forcing variable are observable, while the treatment status is not. Otherwise, the testing and estimation problem will degenerate to the case with a known discontinuity point; see Section 2 of Yu (2012) and Section 2.2 of Yu and Zhao (2013) for a detailed discussion on this point. We further concentrate on the sharp design with at most one discontinuity point; generalization to finite and unknown number of discontinuity points is only discussed briefly in Section 5.3.

A word on notation: the letter C is used as a generic positive constant, which need not be the same in each occurrence. WLOG means "without loss of generality". DGP means "data generating process". LLS means the "local linear smoother" popularized by Fan (1992, 1993) and Fan and Gijbels (1996). The symbol $\approx$ means that the higher-order terms are omitted or a constant term is omitted (depending on the context). w.p.a.1 means "with probability approaching one". For a nonnegative real s, $[s]$ is its integer part. For any two random variables x and y, $f(x)$ means the density of x and $f(y|x)$ means the conditional density of y given x. The letter π_0 represents the true discontinuity point and π represents a generic discontinuity point in the parameter space $\Pi = [\underline{\pi}, \overline{\pi}]$, where $\underline{\pi}$ and $\overline{\pi}$ are constants and $\underline{\pi} < \pi_0 < \overline{\pi}$. For a function $g(x)$, $g(x)$ has a cusp at $x = \pi$ means that $g(x)$ is continuous at π but $g'(\pi+) \neq g'(\pi-)$, that is, the left and right derivatives at π are not the same. "Discontinuity" and "jump" are used exchangeably.

2. Framework and assumptions

We first put RDDs in the usual treatment framework and discuss a key "selection" assumption. Such a framework can be treated as the structural form of RDDs. We then impose some smoothness assumptions on the usual reduced-form formulation of RDDs. Such assumptions are necessary for the development of the testing and estimation procedures in this paper. Finally, we sketch the basic ideas of our specification testing and estimation.

2.1. Selection and treatment effects

Following Lee (2008), suppose the response $y = y(x, U)$, where x is the one-dimensional forcing variable which is observable, and U is the unobservable component such as students' ability in the scholarship example of van der Klaauw (2002). We assume that there can be any correlation between U and x.[1] Also, $y(x, U)$ satisfies the following smoothness assumptions.

Assumption Y. (a) If there is no treatment or there is treatment but are no treatment effects, $y(x, U) = \underline{y}(x, U)$ is continuous in (x, U) and is **continuously** differentiable in x for each U.
(b) If there are treatment effects, $y(x, U) = \underline{y}(x, U) + \alpha(U)1(x \geq \pi)$ with $\alpha(U)$ being continuous.

Under Assumption Y, when there are no treatment effects, the response y is a **smooth** function of x after controlling for all specific characters (except x) of an individual. In the special case where $\underline{y}$ takes the additively separable form, $\underline{y} = g(x) + U$, $g(\cdot)$ is assumed to be continuously differentiable. This is understandable because it is hard to imagine y changes dramatically when x changes from $x-\Delta$ to $x+\Delta$ for a small Δ given that human beings usually behave smoothly. Even if there are treatment effects, $y(x, U)$ only changes its size at $x = \pi$, and the slopes at the left and right sides of π remain the same.

Under Assumption Y, using notations of the conventional average treatment effects literature such as Heckman and Vytlacil (2007a,b), we can express the responses of the control and treated group as follows:

$$Y_0 = \mu_0(x, U_0), \qquad Y_1 = \mu_1(x, U_1) \quad \text{and} \quad D = 1(x \geq \pi),$$

where $\mu_0(x, U_0) = \underline{y}(x, U)$ with $U_0 = U$ and $\mu_1(x, U_1) = \underline{y}(x, U) + \alpha(U)$ with $U_1 = U$. The main difference of RDDs from the conventional average treatment effects framework is that the treatment status is determined by a single **observable** x, so the treatment status can be sharply observed (or there is a discontinuity in the propensity score at π, or the unconfoundedness condition is trivially satisfied). Such an advantage is not free. The usual overlap assumption is violated because given any x, we can observe either Y_0 or Y_1 but not both. We must rely on the continuity of $\mu_0(x, U_0)$ in the left neighborhood of π to predict its behavior in the right neighborhood of π, and similarly for $\mu_1(x, U_1)$. As a result, we only use the local information around π to identify the treatment effects, which makes the treatment effects estimator achieve only a

[1] From the Skorohod representation, y can be expressed as $Q(U|x)$, where $Q(\cdot|x)$ is the conditional quantile function of y given x, and $U|x \sim U(0, 1)$. Note here that U and x can have any correlation; see Section 2.1 of Yu (2014b) for more discussions on this point. Only in quantile regression, we assume U and x are independent.

nonparametric convergence rate rather than the usual $\sqrt{n}$ rate of average treatment effects estimators. Usually, we express the relationship between y and x in a reduced form,

$$y = m(x) + \varepsilon = m_\pi(x) + \alpha_\pi d_\pi + \varepsilon$$
$$\text{with } E[\varepsilon|x, d_\pi] = 0, \quad d_\pi = D, \tag{1}$$

where

$$y = Y_0(1-D) + Y_1 D,$$
$$m(x) = \mu_0(x)(1-D) + \mu_1(x)D = E[y|x]$$

with

$$\mu_0(x) = E[\underline{y}(x,U)|x], \qquad \mu_1(x) = E[\underline{y}(x,U) + \alpha(U)|x],$$
$$\varepsilon = \varepsilon_0(1-D) + \varepsilon_1 D$$

with

$$\varepsilon_0 = \underline{y}(x,U) - \mu_0(x), \qquad \varepsilon_1 = \underline{y}(x,U) + \alpha(U) - \mu_1(x),$$
$$\alpha_\pi = \mu_1(\pi+) - \mu_0(\pi-) = E[y|x=\pi+] - E[y|x=\pi-] \equiv m_+(\pi) - m_-(\pi),$$
$$m_\pi(x) = m(x) - \alpha_\pi d_\pi \text{ is continuous.}$$

Note that U is different from ε (or ε_0) in our setup since $\varepsilon = y - E[y|x]$, which is not equal to U even if we assume an additively separable form of $\underline{y}$ (unless $E[U|x] = 0$ and $\alpha(U)$ does not depend on U). Also, ε_0 and ε_1 generally follow different distributions. Especially, $\sigma^2(x) = E\left[\varepsilon^2|x\right]$ is different in the left and right neighborhoods of π.

The special structure of RDDs allows us to express $E[y|x] = m_\pi(x) + \alpha_\pi d_\pi$. But when there is treatment, no matter there are treatment effects or not, individuals can respond to the treatment actively although they cannot control x precisely; see Lee (2008) and Lee and Lemieux (2010) for discussions on the role of imprecise control of x by individuals in RDDs. Such attempts to control x generate a cusp at π in $f(x|U)$ and consequently in $m_\pi(x)$; see the figures in Lee (2008) for some intuitive impressions. Since $f(u|x) = \frac{f(x|U=u)f_U(u)}{\int f(x|U=u)f_U(u)du}$, $f(u|x)$ will also have a cusp at $x = \pi$. In Appendix A of the supplementary materials (see Appendix A), we provide a concrete example to clarify this point. Therefore, we impose the following smoothness assumptions on $f(U|x)$, and call this assumption the *selection* assumption and explain the reason for this name at the end of this subsection.

Assumption S. The marginal density of U, $f_U(u)$, is continuously differentiable. For each U, $f(U|x)$ is continuous for all x.

(a) When there is no selection, $f(U|x)$ is **continuously** differentiable in x for each U.
(b) When there is selection, $f(U|x)$ is **continuously** differentiable in x for each U except at $x = \pi$.

Assumption S(b) implies that $f(U|x)$ is a nontrivial function of x; that is, U is not independent of x. Also, under Assumption S(b), it is possible for $f(U|x)$ to have a cusp at $x = \pi$.

Under Assumptions Y(a) and S(a), $m(x) = E[y|x] = \int y(x,U) f(U|x)dU$ is continuously differentiable in x, so there is no cusp in $m(x)$. But different combinations of Y(b) and S(b) imply different behaviors of $m(x)$. First, suppose Y(a) and S(b) hold; then $m(x)$ is obviously continuous, but it is not smooth at $x = \pi$. Note that

$$\begin{aligned} m'_+(\pi) &\equiv m'(\pi+) \\ &= \int \frac{\partial y(\pi+,U)}{\partial x} f(U|x=\pi+)dU \\ &\quad + \int y(\pi+,U)\frac{\partial f(U|x=\pi+)}{\partial x}dU, \\ m'_-(\pi) &\equiv m'(\pi-) \\ &= \int \frac{\partial y(\pi-,U)}{\partial x} f(U|x=\pi-)dU \\ &\quad + \int y(\pi-,U)\frac{\partial f(U|x=\pi-)}{\partial x}dU. \end{aligned}$$

Table 1
The behavior of $m(\cdot)$ and $m'(\cdot)$ at π with and without selection or treatment effect.

	No selection	Selection
No treatment effect	$m_+(\pi) = m_-(\pi)$ $m'_+(\pi) = m'_-(\pi)$	$m_+(\pi) = m_-(\pi)$ $m'_+(\pi) \neq m'_-(\pi)$
Treatment effect	$m_+(\pi) \neq m_-(\pi)$ $m'_+(\pi) = m'_-(\pi)$	$m_+(\pi) \neq m_-(\pi)$ $m'_+(\pi) \neq m'_-(\pi)$

Given that $\frac{\partial y(\pi+,U)}{\partial x} = \frac{\partial y(\pi-,U)}{\partial x}$ and $f(U|x)$ is continuous at $x = \pi$,

$$\begin{aligned} &m'_+(\pi) - m'_-(\pi) \\ &= \int \left(y(\pi+,U)\frac{\partial f(U|x=\pi+)}{\partial x} - y(\pi-,U)\frac{\partial f(U|x=\pi-)}{\partial x} \right) dU \\ &= \int y(\pi,U)\left(\frac{\partial f(U|x=\pi+)}{\partial x} - \frac{\partial f(U|x=\pi-)}{\partial x} \right) dU, \end{aligned}$$

which can be but generally is not zero. Of course, if U is independent of x (that is, there is no selection), then $\frac{\partial f(U|x=\pi+)}{\partial x} = \frac{\partial f(U|x=\pi-)}{\partial x} = 0$, and $m'_+(\pi) = m'_-(\pi)$. Second, suppose Y(b) and S(a) hold; then

$$m_+(\pi) - m_-(\pi) = \int \alpha(U) f(U|x=\pi)dU,$$

which can be but generally is not zero. For example, if $\alpha(U)$ has the same sign for any U, then $m_+(\pi) - m_-(\pi)$ cannot be zero. But it is easy to show that $m'_+(\pi) = m'_-(\pi)$. Finally, suppose Y(b) and S(b) hold; then neither $m_+(\pi) - m_-(\pi)$ nor $m'_+(\pi) - m'_-(\pi)$ is necessarily zero. The analysis above is summarized in Table 1 and intuitively shown in Fig. 1.

In a word, $m'_+(\pi) \neq m'_-(\pi)$ is sufficient but not necessary for the existence of selection. Similarly, $m_+(\pi) \neq m_-(\pi)$ is sufficient but not necessary for the existence of treatment effects; when some extra assumptions (such as $\alpha(U)$ has the same sign for any U) hold, it is also necessary. Notice that $f(x) = \int f(x|U=u)f_U(u)du$, so similar arguments show that $f'_+(\pi) \neq f'_-(\pi)$ is sufficient but not necessary for the existence of selection, and $f_+(\pi) \neq f_-(\pi)$ is sufficient but not necessary for the existence of precise control, where $f_\pm(\pi)$ and $f'_\pm(\pi)$ are similarly defined as $m_\pm(\pi)$ and $m'_\pm(\pi)$. If manipulation is monotonic, $f_+(\pi) \neq f_-(\pi)$ is also necessary for the existence of precise control as argued in McCrary (2008).

Finally, we briefly explain why we call Assumption S the selection assumption. In the usual treatment effect literature, *selection* means $Cov(D, U_0|x) \neq 0$; that is, control and treatment groups include individuals with different characteristics under the control treatment. In RDDs, selection is trivially avoided because given x, D is fixed at either 1 or 0. However, if x is not observed, then $Cov(D, U_0) = Cov(1(x \geq \pi), U) \neq 0$ as discussed above. So Assumption S essentially corresponds to the selection assumption in the average treatment effects literature; that is, individuals are trying to benefit from managing their treatment status. Such an argument also reveals the very fact that all specialties of RDDs are from the observability of the forcing variable x which completely determines the treatment status.

2.2. Regularity assumptions

For either specification testing or estimation, only the structures of $m(x)$ on $\Pi_\epsilon \equiv \left(\underline{\pi} - \epsilon, \overline{\pi} + \epsilon\right)$ for some $\epsilon > 0$ are relevant, so WLOG, we assume the support of x to be Π_ϵ. The conditional distribution of ε given x is assumed to satisfy the following conditions.

Assumption E. (a) $f(\varepsilon|x)$ is continuous in (ε, x) for $x \in \Pi_\epsilon^+ \equiv [\pi, \overline{\pi} + \epsilon)$ and $x \in \Pi_\epsilon^- \equiv \left(\underline{\pi} - \epsilon, \pi\right)$, respectively.

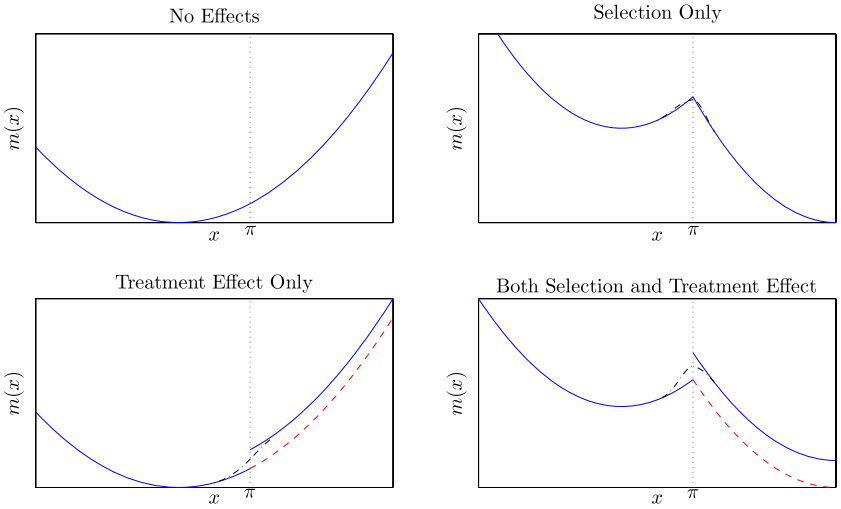

Fig. 1. $m(\cdot)$, $m_\pi(\cdot)$ and $\overline{m}(\cdot)$ in models with and without selection or treatment effect: solid line for $m(\cdot)$, dashed line for $m_\pi(\cdot)$, and dash–dot line for $\overline{m}(\cdot)$.

(b) $E\left[|\varepsilon|^4 \middle| x\right]$ is uniformly bounded on Π_ϵ.
(c) $\sigma^2(x)$ is Lipschitz for $x \in \Pi_\epsilon^+$ and $x \in \Pi_\epsilon^-$.

(a) implies that there is no point mass in the distribution of ε, which excludes the binary y case. This assumption will be used to uniquely detect the location of π. Also, (a) (combined with (b)) implies that both $\sigma^2(x)$ and $E\left[|\varepsilon|^4 \middle| x\right]$ are finite for $x \in \Pi_\epsilon$ and continuous for $x \neq \pi$. Such finiteness and continuity restrictions guarantee that the possible discontinuity at π in $\sigma^2(x)$ and $E\left[|\varepsilon|^4 \middle| x\right]$ is of the first kind. (b) is used in specification testing; in estimation, we need only $E\left[|\varepsilon|^{2+\zeta} \middle| x\right]$ bounded for some $\zeta > 0$. (c) is a standard smoothness assumption on $\sigma^2(x)$.

We must impose some smoothness assumptions on $m(x)$ to continue our analysis. For $s \geq 1$, let the Hölder class $\mathcal{C}_\pi^+(L, s)$ be the set of maps $m(\cdot)$ from Π_ϵ^+ to $\mathbb{R}$ with

$$\mathcal{C}_\pi^+(L,s) = \left\{ m(\cdot) : \left| m(x+\Delta) - \sum_{j=0}^{[s]} \frac{m^{(j)}(x)}{j!}\Delta^j \right| \leq L|\Delta|^s \right.$$
$$\left. \text{for all } x, x+\Delta \in \Pi_\epsilon^+ \text{ and } \sup_{j=0,\ldots,[s]} \sup_{x\in\Pi_\epsilon^+} \left|m^{(j)}(x)\right| \leq L \right\},$$

where $m^{(0)}(x) = m(x)$, $m^{(j)}(x)$ is the jth-order derivative when $x \neq \pi$, and is the jth-order right hand derivative at π. Similarly, we can define $\mathcal{C}_\pi^-(L, s)$ with Π_ϵ^+ replaced by Π_ϵ^-, and define $\mathcal{C}(L, s)$ with Π_ϵ^+ replaced by Π_ϵ. Further define

$$\mathcal{C}_\pi(L,s) \equiv \mathcal{C}_\pi^+(L,s) \cap \mathcal{C}_\pi^-(L,s),$$
$$\mathcal{M}_\Pi(L,s) \equiv \bigcup_{\pi\in\Pi} \mathcal{C}_\pi(L,s), \qquad \mathcal{C}_\Pi(L,s) \equiv \mathcal{M}_\Pi(L,s) \cap \mathcal{C}_0, \tag{2}$$

where $\mathcal{C}_0$ is the set of continuous functions on Π_ϵ. Note that $\mathcal{C}_\pi(L, s)$ is the Hölder class of functions on Π_ϵ^+ and Π_ϵ^- which may not be continuous at π, $\mathcal{M}_\Pi(L, s)$ is the collection of all such Hölder class of functions with a possible discontinuity at $\pi \in \Pi$, and $\mathcal{C}_\Pi(L, s)$ includes the continuous (although may not be smooth at $\pi \in \Pi$) function among $\mathcal{M}_\Pi(L, s)$. We impose the following assumptions on $m(\cdot)$.

Assumption M. $m(x) \in \mathcal{M}_\Pi(L, s), m_\pi(x) \in \mathcal{C}_\Pi(L, s), s \geq 1$.

Note that $\mathcal{C}_\Pi(L, s)$ includes functions with a cusp at π. Such functions imply the presence of selection. Without selection, $m_\pi(\cdot)$ is smooth. Fig. 1 shows these two cases. Assumption M states that although there can be a cusp in $m_\pi(\cdot)$, the cusp cannot be too sharp; see Wang (1995) for the case with a sharp cusp. Such an assumption is consistent with the analysis in the last subsection: $m'(\pi+) - m'(\pi-) = \int y(\pi, U)\left(\frac{\partial f(U|x=\pi+)}{\partial x} - \frac{\partial f(U|x=\pi-)}{\partial x}\right) dU$ should not be very large if $\frac{\partial f(U|x=\pi+)}{\partial x} - \frac{\partial f(U|x=\pi-)}{\partial x}$ is not large for each U. We next impose some smoothness conditions on $f(x)$:

Assumption F. $f(x) \in \mathcal{C}_0$. $0 < \underline{f} \leq f(x) \leq \overline{f} < \infty$ for $x \in \Pi_\epsilon$.

(a) $f(x) \in \mathcal{C}(\overline{f}, \lambda), \lambda \geq 1$.
(b) $f(x) \in \mathcal{C}_\Pi(\overline{f}, \lambda), \lambda \geq 1$.

Assumption F corresponds to Assumption S. F(a) corresponds to the no selection case S(a), and F(b) corresponds to the selection case S(b); see Section 5.1 for tests on these two Assumptions F(a) and F(b). Both $\mathcal{C}(\overline{f}, \lambda)$ and $\mathcal{C}_\Pi(\overline{f}, \lambda)$ include $\mathcal{C}_0$, so x cannot be precisely controlled.

2.3. Sketch of the paper

We start from two kinds of specification tests. First, we test whether there is selection. Specifically, the hypotheses are

$$H_0^{(1)} : m(\cdot) \in \mathcal{C}(L,s), \quad s \geq 1; \tag{3}$$
$$H_1^{(1)} : m(\cdot) \in \mathcal{C}_\Pi(L,s) \setminus \mathcal{C}(L,s), \quad s \geq 1;$$

where $\mathcal{C}_\Pi(L, s)$ and $\mathcal{C}(L, s)$ are defined around Eq. (2), $H_0^{(1)}$ corresponds to no effects (the upper-left panel of Fig. 1), and $H_1^{(1)}$ corresponds to selection only (the upper-right panel of Fig. 1). Second, we test whether there are treatment effects (regardless of whether selection is present). Specifically, the hypotheses are

$$H_0^{(2)} : m(\cdot) \in \mathcal{C}_\Pi(L,s), \quad s \geq 1; \tag{4}$$
$$H_1^{(2)} : m(\cdot) \in \mathcal{M}_\Pi(L,s) \setminus \mathcal{C}_\Pi(L,s), \quad s \geq 1;$$

where $\mathcal{C}_\Pi(L, s)$ and $\mathcal{M}_\Pi(L, s)$ are defined in (2), $H_0^{(2)}$ corresponds to two cases: no effects and selection only (the upper two panels of Fig. 1), and $H_1^{(2)}$ corresponds to the other two cases: treatment effect only and both selection and treatment effect (the lower two panels of Fig. 1). Both tests are relevant in practice, and can be sequentially conducted: first test (4); if $H_0^{(2)}$ cannot be rejected, then test (3). Denote the class of probability measures under $H_0^{(\ell)}$ as $\mathcal{H}_0^{(\ell)}$ and under $H_1^{(\ell)}$ as $\mathcal{H}_1^{(\ell)}$, $\ell = 1, 2$. Both $\mathcal{H}_0^{(\ell)}$ and $\mathcal{H}_1^{(\ell)}$, $\ell = 1, 2$, are characterized by $m(\cdot)$. In what follows, we acknowledge the dependence of the distribution of y given x upon the regression

function by denoting probabilities and expectations as P_m and E_m, respectively.

We develop a unified test statistic for both cases. We first find the best approximation of $m(x)$, $\overline{m}(\cdot)$, in $\mathcal{C}(L,s)$ in both cases,

$$\begin{aligned}\overline{m}(\cdot) &= \arg\inf_{\widetilde{m}\in\mathcal{C}(L,s)} E\left[(m(x)-\widetilde{m}(x))^2 1_x^{\Pi}\right]\\ &= \arg\inf_{\widetilde{m}\in\mathcal{C}(L,s)} E\left[(y-\widetilde{m}(x))^2 1_x^{\Pi}\right],\end{aligned}$$

where $1_x^{\Pi} = 1(x\in\Pi)$, and the second equality is from the fact that

$$\begin{aligned}E\left[(y-\widetilde{m}(x))^2 1_x^{\Pi}\right] &= E\left[(\varepsilon+m(x)-\widetilde{m}(x))^2 1_x^{\Pi}\right]\\ &= E\left[\sigma^2(x)1_x^{\Pi}\right]+E\left[(m(x)-\widetilde{m}(x))^2 1_x^{\Pi}\right],\end{aligned}$$

and $E\left[\sigma^2(x)1_x^{\Pi}\right]$ does not depend on $\widetilde{m}(\cdot)$. In short words, $\overline{m}(\cdot)$ is the closest $\widetilde{m}\in\mathcal{C}(L,s)$ to $m(\cdot)$ (or the observable y) under the L^2-metric. It can be estimated by some standard nonparametric techniques. We then define a measure of the distance between $\overline{m}(\cdot)$ and $m(\cdot)$ to detect the deviation from the null. Our test is inspired by Fan and Li (1996) and Zheng (1996) who consider different testing problems. The test statistic is constructed under the null, so is similar to the score test in spirit. The difference of our test statistic in testing (3) and (4) is to choose different smoothing parameters. This is intuitively illustrated in Fig. 1. In testing (3), we oversmooth $m(x)$, and then the bias in such smoothing will generate power. In contrast, in testing (4), we undersmooth $m(x)$. Now, the bias from the cusp will disappear asymptotically, while the jump in $m(\cdot)$ at π still generates power.

Of course, our tests need satisfy some basic properties. Abuse the notations a little bit here: H_0 represents either $H_0^{(1)}$ or $H_0^{(2)}$, and H_1 represents either $H_1^{(1)}$ or $H_1^{(2)}$. For a randomized test t_n, its *size* (or *level*) is defined as

$$\alpha(t_n) = \sup_{m(\cdot)\in H_0} P_m(t_n=1).$$

t_n is *consistent* if its type II error, $P_m(t_n=0)$, converges to zero for any $m(\cdot)\in H_1$ under the Neyman–Pearson restriction $\alpha(t_n)\le\alpha+o(1)$ for some $\alpha\in(0,1)$. We show that our tests are consistent in both cases and we also derive their local powers.

If the specification test rejects $H_0^{(2)}$, we would believe that there is a discontinuity in $m(x)$. The ensued question is how to estimate the discontinuity location π and size α_π. In the RDD literature, the main interest lies in α_π instead of π although an estimator of π is a primary input to estimate α_π. Suppose we have an estimator $\widehat{\pi}$ in hand; then we can use the local polynomial estimator (LPE) as in Porter (2003) to estimate α_π. Denote the estimator as $\widehat{\alpha}(\widehat{\pi})$ with

$$\widehat{\alpha}(\pi) = \widehat{m}_+(\pi)-\widehat{m}_-(\pi),$$

where $\widehat{m}_+(\pi)$ is the LPE determined by the minimizer $\widehat{a}$ in the following minimization problem:

$$\min_{a,b_1,\ldots,b_p}\frac{1}{n}\sum_{j=1}^{n}k_h(x_j-\pi)d_j(\pi)\left[y_j-a-b_1(x_j-\pi)-\cdots-b_p(x_j-\pi)^p\right]^2, \tag{5}$$

where p is the suitable order of local polynomials defined in Porter (2003) and is usually set to be 1 as in Hahn et al. (2001), $d_j(\pi)=1(x_j\ge\pi)$, $k_h(\cdot)=\frac{1}{h}k\left(\frac{\cdot}{h}\right)$, $k(\cdot)$ is a kernel function, and h is the bandwidth. Similarly, we define $\widehat{m}_-(\pi)$ but using the data in the left neighborhood of π. Actually, $\widehat{\alpha}(\widehat{\pi})$ is a corollary of our estimation of π because π is estimated by maximizing $\widehat{\alpha}^2(\pi)$. We show that $\widehat{\pi}$ is n-consistent, so it will not affect the asymptotic distribution of $\widehat{\alpha}(\widehat{\pi})$. As a result, our estimator of α_π achieves the optimal rate of convergence developed in Porter (2003).

3. Consistent specification testing

This section presents consistent specification testing of (3) and (4). It begins with consistent tests and the corresponding asymptotic distributions, followed by a bootstrap method to obtain critical values, and concludes with a discussion about three alternative tests.

3.1. Tests construction and asymptotics

First rewrite (1) as

$$y=\overline{m}(x)+e, \tag{6}$$

where $E[e|x]=0$ under $H_0^{(1)}$ for almost all $x\in\Pi$ and $E[e|x]\neq0$ under $H_1^{(1)}$ for a set of $x\in\Pi$ with positive Lebesgue measure. Observing that $E\left[eE[e|x]1_x^{\Pi}\right]=E\left[E[e|x]^2 1_x^{\Pi}\right]=E\left[(m(x)-\overline{m}(x))^2 1_x^{\Pi}\right]\ge0$ and the equality holds if and only if $H_0^{(1)}$ is true, we can construct a consistent test for (3) based on $E\left[eE[e|x]1_x^{\Pi}\right]$. $E[e|x]$ can be estimated by a kernel estimator. To avoid the random denominator problem in kernel estimation, we choose to estimate a density weighted version of $E\left[eE[e|x]1_x^{\Pi}\right]$ given by $E\left[eE[e|x]f(x)1_x^{\Pi}\right]$. If e_i and $E[e_i|x_i]f(x_i)$ were available, we could estimate $E\left[eE[e|x]f(x)1_x^{\Pi}\right]$ by its sample analogue $n^{-1}\sum_{i=1}^n e_iE[e_i|x_i]f_i1_i^{\Pi}$, where $1_i^{\Pi}=1(x_i\in\Pi)$, and $f_i=f(x_i)$. To get a feasible test statistic, we need to estimate e_i by the corresponding residual from (6) and $E[e_i|x_i]f_i$ by an appropriate kernel estimator. Specifically, we estimate e_i by $\widehat{e}_i=y_i-\widehat{y}_i$, where $\widehat{y}_i$ is a kernel estimator of $\overline{m}(x_i)$ defined as

$$\widehat{y}_i=\frac{1}{n-1}\sum_{j\neq i}y_jL_{b,ij}\Big/\widehat{f}_i \tag{7}$$

with $\widehat{f}_i$ being the corresponding kernel estimator of f_i given by

$$\widehat{f}_i=\frac{1}{n-1}\sum_{j\neq i}L_{b,ij},$$

$L_{b,ij}=l_b(x_i-x_j)=\frac{1}{b}l\left(\frac{x_i-x_j}{b}\right)$, b is the bandwidth, and $l(\cdot)$ is some kernel function. Excluding (x_i,y_i) in estimating $\overline{m}(x_i)$ is to convert a V-statistic to a U-statistic. We estimate $E[e_i|x_i]f_i$ by $\frac{1}{n-1}\sum_{j\neq i}\widehat{e}_jK_{h,ij}1_j^{\Pi}$, where $K_{h,ij}=k_h(x_i-x_j)=\frac{1}{h}k\left(\frac{x_i-x_j}{b}\right)$ is similarly defined as $L_{b,ij}$. Eventually, our test statistic is based on

$$I_n=\frac{nh^{1/2}}{n(n-1)}\sum_i\sum_{j\neq i}1_i^{\Pi}1_j^{\Pi}K_{h,ij}\widehat{e}_i\widehat{e}_j,$$

where

$$\widehat{e}_i=y_i-\widehat{y}_i=(m_i-\widehat{m}_i)+(\varepsilon_i-\widehat{\varepsilon}_i), \tag{8}$$

$m_i=m(x_i)$, and $\widehat{m}_i$ and $\widehat{\varepsilon}_i$ are defined in the same way as $\widehat{y}_i$ in (7) with y_j replaced by $m(x_j)$ and ε_j, respectively. Under $H_0^{(1)}$, $\widehat{e}_i$ is a good estimate of ε_i, while under $H_1^{(1)}$, $\widehat{e}_i$ includes a bias term in the neighborhood of π, which generates power. The indicator function 1_i^{Π} in I_n can improve the power properties by shrinking the asymptotic variance of I_n.

Fan and Li (1996) further eliminate the random denominators in $\widehat{e}_i$ and $\widehat{e}_j$ in I_n by replacing $\widehat{e}_i$ and $\widehat{e}_j$ with $\widehat{e}_i\widehat{f}_i$ and $\widehat{e}_j\widehat{f}_j$, but the form of I_n we use is more convenient for practitioners since it allows $\widehat{e}_i$'s to be estimated by other nonparametric methods. For example, although the LPE is asymptotically equivalent to a higher-order kernel smoother in (7), its various forms, especially the LLS because of its minimax efficiency shown in Fan (1993), can improve the finite-sample performance of I_n and are more popular than (7).

However, it is not easy to define the random denominator for the LLS if Fan and Li's formulation is used. Of course, our form of I_n is not new; Gu et al. (2007) consider a similar test statistic in testing whether there are omitted variables.

To use I_n testing (4), we must guarantee that the estimator of $E\left[eE\left[e|x\right]f(x)1_x^{\Pi}\right]$ is approximately zero under $H_0^{(2)}$. Note that under $H_0^{(2)}$, $E\left[eE\left[e|x\right]f(x)1_x^{\Pi}\right] = E\left[(m_\pi(x)-\overline{m}(x))^2 f(x)1_x^{\Pi}\right]$, so we must use a smaller b to make the bias in approximating $m_\pi(x)$ by $\overline{m}(x)$ disappear; see the upper-right panel of Fig. 1 for some intuition. Assumption B imposes such kind of restrictions on b (and h) which are suitable in different cases.

Assumption B. $b \to 0, nh \to \infty, h^{1/2}/b \to 0$. For λ and s defined in Assumptions M and F,

(a) $nh^{1/2}b^{2\eta} \to 0, nh^{1/2}b^3 \to \infty$, where $\eta = \min(\lambda+1, s) > 1.5$.
(b) $nh^{1/2}\left(b^{2\eta}+b^3\right) \to 0, nh^{1/2}b \to \infty$, where $\eta = \min(\lambda+1, s) \geq 1$.

Assumption B implies that $\max\{h, b\} \to 0$ and $n\min\{h, b\} \to \infty$, which ensure that the kernel estimators involved are consistent. In B(a), $nh^{1/2}b^{2\eta}$ is the bias under $H_0^{(1)}$, and the bias under $H_1^{(1)}$ is $nh^{1/2}\left(b^{2\eta}+b^3\right)$. The first term $b^{2\eta}$ is the bias from the area out of a b neighborhood of π, and the second term b^3 is the bias from the b neighborhood of π. Because $\eta > 1.5$, the second term dominates the first term. The assumption $nh^{1/2}b^{2\eta} \to 0$ ensures that I_n is centered correctly at zero under $H_0^{(1)}$. Corresponding arguments can be applied to B(b) for testing (4). We indeed require $m(\cdot)$ to be a little smoother under $H_0^{(1)}$ than $H_0^{(2)}$ to control the bias, but it seems irrelevant in practice. The assumption $h^{1/2}/b \to 0$ implies that h is smaller than b.[2] This condition not only ensures the bias disappear under $H_0^{(1)}$ and $H_0^{(2)}$, but generates power under $H_1^{(1)}$ and $H_1^{(2)}$. Because $\widehat{e}_j$ is estimated using the data in a b neighborhood of x_j, and the distance between x_j and x_i is restricted to be no greater than $O(h)$ by $K_{h,ij}$, we can treat x_i and x_j as the same point without changing the asymptotic results. In consequence, $\widehat{e}_i\,\widehat{e}_j$ is like a squared term which generates power. Take $b \sim n^{-c}$ and $h = n^{-(2c+\delta)}$ for an arbitrarily small positive δ; then in B(a), $c < 1/4$ and in B(b), $\frac{1}{4} \leq c < \frac{1}{2}$ when $s \geq 1.5$. This matches the intuition above that a larger bandwidth is required in testing (3) than in testing (4).

Next, we impose some standard constraints on the kernel functions $l(\cdot)$ and $k(\cdot)$.

Assumption K. Both $l(\cdot)$ and $k(\cdot)$ are bounded, symmetric, Lipschitz function, zero outside a bounded set $[-M, M]$, $\int u^i l(u)du = \delta_{i0}, i = 0, \ldots, [s]+[\lambda]$, where δ_{i0} is Kronecker's delta, and $k(\cdot)$ is non-negative with $k(0) > 0$.

From Assumption K, $l(\cdot)$ can be a higher-order kernel to reduce the bias in estimating $\overline{m}(\cdot)$, but $k(\cdot)$ is a usual second-order kernel. Such an assumption can simplify our proof. To further simplify our discussion, we assume $M = 1$ throughout this paper. Now, we state the asymptotic distribution of I_n under the null and the local power of the test based on I_n in different cases.

Theorem 1. *Under Assumptions B(a), E, K and M,*

(i) *if F(a) holds, then*

$$I_n \xrightarrow{d} N(0, \Sigma)$$

uniformly over $\mathcal{H}_0^{(1)}$, where

$$\Sigma = 2E\left[1_x^{\Pi} f(x)\sigma^4(x)\right]\int k^2(u)du$$

can be consistently estimated by

$$v_n^2 = \frac{2h}{n(n-1)}\sum_i\sum_{j\neq i} 1_i^{\Pi}1_j^{\Pi}K_{h,ij}^2\widehat{e}_i^2\widehat{e}_j^2.$$

As a result, the test based on the studentized test statistic $T_n = I_n/v_n$

$$t_n = 1(T_n > z_\alpha),$$

has the significance level α, where z_α is the $1-\alpha$ quantile of the standard normal distribution.

(ii) *if F(b) holds and $\frac{m'_+(\pi)-m'_-(\pi)}{n^{-1/2}h^{-1/4}b^{-3/2}} \to \delta$, then*

$$I_n \xrightarrow{d} N\left(\kappa f^2(\pi)\delta^2, \Sigma\right) \quad \text{and}$$

$$T_n \xrightarrow{d} N\left(\kappa f^2(\pi)\delta^2/\sqrt{\Sigma}, 1\right)$$

where $\kappa = 2\int_0^1\left(v\int_v^1 l(u)\,du - \int_v^1 ul(u)\,du\right)^2 dv$. This implies that the test t_n is consistent; that is, $P_m(T_n > z_\alpha) \to 1$ for any fixed $m(\cdot) \in \mathcal{C}_{\Pi}(L, s)\setminus\mathcal{C}(L, s)$. Furthermore, z_α can be replaced by any nonstochastic constant $C_n = o\left(nh^{1/2}b^3\right)$, and the result still holds.

Theorem 2. *Under Assumptions B(b), F(b), E, K and M,*

(i) $I_n \xrightarrow{d} N(0, \Sigma)$
uniformly over $\mathcal{H}_0^{(2)}$, where Σ is the same as in Theorem 1, and can be consistently estimated by v_n^2. As a result, the test $t_n = 1(T_n > z_\alpha)$ has the significance level α, where T_n and z_α are defined in Theorem 1(i).

(ii) *if $\frac{\alpha_\pi}{n^{-1/2}h^{-1/4}b^{-1/2}} \to \delta$, then*

$$I_n \xrightarrow{d} N\left(\kappa f^2(\pi)\delta^2, \Sigma\right) \quad \text{and}$$

$$T_n \xrightarrow{d} N\left(\kappa f^2(\pi)\delta^2/\sqrt{\Sigma}, 1\right)$$

where $\kappa = 2\int_0^1\left(\int_v^1 l(u)du\right)^2 dv$. This implies that the test t_n is consistent; that is, $P_m(T_n > z_\alpha) \to 1$ for any fixed $m(\cdot)$ such that $\alpha_\pi \neq 0$. Furthermore, z_α can be replaced by any nonstochastic constant $C_n = o\left(nh^{1/2}b\right)$, and the result still holds.

The asymptotic distributions of I_n under the null in Theorems 1 and 2 are the same, but under different assumptions. To provide a nondegenerate asymptotic distribution of I_n, we use the central limit theorem for the second order degenerate U-statistics. The appearance of degenerate U-statistics is standard in model specification tests when a nonparametric component is present; see, e.g., Fan and Li (1996) and Zheng (1996).[3] Note that the local power in Theorems 1 and 2 depends on $m(\cdot)$ only through $m'_\pm(\pi)$ or $m_\pm(\pi)$ rather than the whole $m(x)$ on Π, which is the key difference between the tests here and those in the nonparametric specification testing literature (see, e.g., Theorem 3 of Zheng (1996) and Theorem 3.1 of Fan and Li (2000)). Intuitively, the power of I_n is only from the local deviation (from H_0) of $m(x)$ around π, while the power of the usual specification tests can be from global deviations of $m(x)$. Also, although I_n is designed for the case where the number of cusps (or jumps) is at most one, it is easy to see that when the number of cusps (or jumps) is greater than one but finite, the specification tests based on I_n remain valid.

[2] In the classical specification tests such as Zheng (1996), $\widehat{e}_i$ is obtained under the null parametric specification, which corresponds to $b = C$, a fixed number. Since $h \to 0, h/b = h/C \to 0$ and h is smaller than b.

[3] Because Fan and Li (1996) also eliminate the denominator of $\widehat{e}_i$, they need the fourth order U-statistics theory to derive their asymptotic distribution.

3.2. Bootstrapping critical values

From the proofs of Theorems 1 and 2, the convergence rate of T_n to standard normal is very slow.[4] As argued in the literature of specification testing (see, e.g., Härdle and Mammen (1993), Li and Wang (1998), Stute et al. (1998), Delgado and Manteiga (2001), and Gu et al. (2007)), a better way to approximate the finite-sample distribution is to use the wild bootstrap of Wu (1986) and Liu (1988).[5] Specifically, the following procedure is used in testing both (3) and (4).

Algorithm WB. Step 1: For $i = 1, \ldots, n$, generate the two-point wild bootstrap residual $\varepsilon_i^* = \widehat{e}_i\left(1-\sqrt{5}\right)/2$ with probability $\left(1+\sqrt{5}\right)/\left(2\sqrt{5}\right)$, and $\varepsilon_i^* = \widehat{e}_i\left(1+\sqrt{5}\right)/2$ with probability $\left(\sqrt{5}-1\right)/\left(2\sqrt{5}\right)$; then $E^*\left[\varepsilon_i^*\right]=0$, $E^*\left[\varepsilon_i^{*2}\right]=\widehat{e}_i^2$ and $E^*\left[\varepsilon_i^{*3}\right]=\widehat{e}_i^3$, where $E^*[\cdot]=E[\cdot|\mathcal{F}_n]$ and $\mathcal{F}_n=\{(x_i,y_i)\}_{i=1}^n$.

Step 2: Generate the bootstrap resample $\left\{y_i^*, x_i\right\}_{i=1}^n$ by

$$y_i^* = \widehat{y}_i + \varepsilon_i^*.$$

Then obtain the bootstrap residuals $\widehat{e}_i^* = y_i^* - \widehat{y}_i^*$, where $\widehat{y}_i^*$ is similarly defined as $\widehat{y}_i$ with the only difference being that y_j in (7) is replaced by y_j^*.

Step 3: Use $\left\{\widehat{e}_i^*, x_i\right\}_{i=1}^n$ to compute the test statistic

$$I_n^* = \frac{nh^{1/2}}{n(n-1)}\sum_i\sum_{j\neq i} 1_i^{\Pi}1_j^{\Pi}K_{h,ij}\widehat{e}_i^*\,\widehat{e}_j^*$$

and the estimated asymptotic variance

$$v_n^{*2} = \frac{2h}{n(n-1)}\sum_i\sum_{j\neq i} 1_i^{\Pi}1_j^{\Pi}K_{h,ij}^2\widehat{e}_i^{*2}\widehat{e}_j^{*2}.$$

Then the studentized bootstrap statistic is $T_n^* = I_n^*/v_n^*$. Here, the same b and h are used as in I_n and v_n^2 in Theorems 1 and 2.

Step 4: Repeat Step 1 through 3 B times, and use the empirical distribution of $\left\{T_{nk}^*\right\}_{k=1}^B$ to approximate the null distribution of T_n. We reject H_0 if $T_n > T_{n(\alpha B)}^*$ or $p \equiv B^{-1}\sum_{k=1}^B 1(T_{nk}^* \geq T_n) \leq \alpha$ at the significance level α, where $T_{n(\alpha B)}^*$ is the upper α-percentile of $\left\{T_{nk}^*\right\}_{k=1}^n$.

In Algorithm WB, the only difference in testing (3) and (4) is to use different b and h as defined in Assumptions B(a) and B(b), respectively. In Step 1, a more popular way to simulate ε_i^* is based on $\widehat{e}_i$'s centralized counterpart $\widetilde{\widehat{e}}_i = \widehat{e}_i - \bar{\widehat{e}}$ instead of $\widehat{e}_i$ itself, where $\bar{\widehat{e}} = \sum_{i=1}^n \widehat{e}_i 1_i^{\Pi_b}/\sum_{i=1}^n 1_i^{\Pi_b}$, $\Pi_b = \left(\underline{\pi}-b, \overline{\pi}+b\right)$; see, e.g., Gijbels and Goderniaux (2004) and Su and Xiao (2008). Such a formulation can ensure $\sum_{i=1}^n \widetilde{\widehat{e}}_i 1_i^{\Pi_b}/\sum_{i=1}^n 1_i^{\Pi_b} = 0$, which will not affect the asymptotic results, but may affect the finite-sample performance of Algorithm WB especially under the alternative.

The bootstrap sample is generated by imposing the null hypothesis. Therefore, the bootstrap statistic T_n^* will mimic the null distribution of T_n even when the null hypothesis is false. When the null is false, $\widehat{e}_i$ is not a consistent estimate of ε_i in the neighborhood of π. Nevertheless, the following theorem shows that the above bootstrap procedure is valid. This is because our studentized test statistic T_n is invariant to the variance of ε. In other words, the wild bootstrap procedure may not work well if the test statistic I_n instead of T_n is used.

Theorem 3. *Suppose Assumptions E, F(b), K and M hold. In testing (3), under Assumption B(a), and in testing (4), under Assumption B(b),*

$$\sup_{z\in\mathbb{R}}\left|P\left(T_n^* \leq z|\mathcal{F}_n\right) - \Phi(z)\right| = o_p(1),$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable.

The above theorem only proves the first-order validity of the wild bootstrap procedure. The higher-order theory can follow the line of Li and Wang (1998) and Fan and Linton (2003). Also, the validity of the wild bootstrap using the bandwidth based on cross-validation (CV) can be justified by extending the technique developed in Hsiao et al. (2007). A formal development of these results is beyond the scope of this paper.

3.3. Alternative specification tests

There are no tests designed for testing (3) in the statistical literature, while at least three alternative tests exist for testing (4). All of them assume a fixed, equally spaced design on [0, 1] and ε_i being i.i.d. sampled.[6] Nevertheless, these tests provide some insights for testing (4).

A straightforward test is based on $\int_{\underline{\pi}}^{\overline{\pi}}\left(\widehat{m}_+(\pi)-\widehat{m}_-(\pi)\right)^2 d\pi$, where $\widehat{m}_\pm(\pi)$ are the local constant estimators of $m_\pm(\pi)$. To avoid the random denominator problem as in I_n, we can use

$$\widetilde{I}_n = nh^{1/2}\int_{\underline{\pi}}^{\overline{\pi}}\left(\widehat{m}_+(\pi)\widehat{f}_+(\pi)-\widehat{m}_-(\pi)\widehat{f}_-(\pi)\right)^2 d\pi,$$

where $\widehat{f}_+(\pi) = n^{-1}\sum_{j=1}^n k_h\left(x_j-\pi\right)d_j(\pi)$, $\widehat{f}_-(\pi) = n^{-1}\sum_{j=1}^n k_h\left(x_j-\pi\right)\left(1-d_j(\pi)\right)$. Wu and Lai (1998) show that

$$\widetilde{I}_n - \frac{\sigma^2}{\sqrt{h}}K_1 \xrightarrow{d} 2\sigma^2 N(0,K_2),$$

where $\sigma^2 = E\left[\varepsilon_i^2\right]$, and K_1 and K_2 are constants only related to $k(\cdot)$. From the form of $\widetilde{I}_n$ and I_n, we can see their key difference: I_n is based on $m(\cdot)$ under the null (e.g., $\overline{m}(\cdot)$ is close to $m(\cdot)$ under the null, and π does not appear at all), while $\widetilde{I}_n$ is based on $m(\cdot)$ under the alternative (e.g., it integrates jumping sizes at different possible jumping locations).[7] The key problem of $\widetilde{I}_n$ is that it does not have a zero mean because it takes a quadratic form. Also, numerical integration is required, which is not attractive in practice. In our setup, σ^2 will change to a functional of $f(\cdot)$ and $\sigma^2(\cdot)$; see, e.g., Härdle and Mammen (1993). Estimation of such nuisance parameters introduces extra troubles to practitioners; nevertheless, see Gao et al. (2008) for such a development.

[4] In the classical specification tests, the bias in the test statistic of Zheng (1996) is $O(h^{1/2})$, which is $O(n^{-1/10})$ when $h = n^{-1/5}$, as shown in Li and Wang (1998). From the discussion after Assumption B, the bias is $nh^{1/2}b^{2\eta}$ under $H_0^{(1)}$ and $nh^{1/2}b^3$ under $H_0^{(2)}$.

[5] The naive bootstrap is not valid as argued in Härdle and Mammen (1993). This is because the regression function under the bootstrap distribution is not the conditional expectation of the observation: $E^*\left[y_i^*|x_i^*\right] = y_i^*$ which is typically different from the fitted value of (7) at x_i^*.

[6] Throughout this paper, we mean this setup when we discuss the statistical literature if no further specification. Note that Π is a subinterval of [0, 1] now.

[7] Note that I_n is like the score test, $\widetilde{I}_n$ is like the Wald test, and the test of Härdle and Mammen (1993) is like the likelihood ratio test in comparison with the three classical asymptotically equivalent tests.

Both I_n and $\widetilde{I}_n$ measure the L^2 distance between $m(\cdot)$ and $\overline{m}(\cdot)$, so they are closely related; see Fan and Li (2000) for a clear argument for this point in the classical specification testing. Such test statistics correspond to the average-statistic of Andrews and Ploberger (1994) in the parametric structural change environment. Another popular metric used in the structural change literature is the sup-norm; e.g., the sup-statistic in Andrews (1993) uses this metric. In our case, a natural test statistic under the sup-norm is $\sup_{\pi\in\Pi}|\widehat{\alpha}(\pi)|$. To avoid the random denominator problem as in I_n and $\widetilde{I}_n$, we use the following test statistic

$$\begin{aligned} S_n &= \sup_{\pi\in\Pi}\sqrt{nh}\left|\widehat{m}_+(\pi)\widehat{f}_+(\pi)-\widehat{m}_-(\pi)\widehat{f}_-(\pi)\right| \\ &= \sup_{\pi\in\Pi}\left|\frac{\sqrt{nh}}{n}\sum_{j=1}^{n}k_h\left(x_j-\pi\right)y_jd_j(\pi)\right. \\ &\quad\left.-\frac{\sqrt{nh}}{n}\sum_{j=1}^{n}k_h\left(x_j-\pi\right)y_j\left(1-d_j(\pi)\right)\right| \\ &\equiv \sup_{\pi\in\Pi}|S_{n\pi}|. \end{aligned}$$

Wu and Chu (1993a) show that the asymptotic distribution of S_n is related to the type-I extreme value distribution. Specifically,

$$P(S_n < a_n + b_n x) \to \exp(-2\exp(-x)),$$

where

$$a_n = \sigma K_3\left(C_\Pi + \frac{\log\left(3\sqrt{4\pi}\right)}{C_\Pi}\right), \qquad b_n = \frac{\sigma K_3}{C_\Pi},$$

K_3 is a constant only related to $k(\cdot)$, and $C_\Pi = \sqrt{C_1 + C_2\log n}$ with C_1 related to the length of Π and C_2 related to the width of h.

We provide some intuition for this result here. First, by the central limit theory, the average form $S_{n\pi}$ in S_n will follow a normal distribution asymptotically. Second, because $S_{n\pi}$ only uses the data in a h neighborhood of π, $S_{n\pi_1}$ and $S_{n\pi_2}$ are asymptotically independent when the distance between π_1 and π_2 is greater than $2h$. Third, since x_i is evenly sampled, and ε_i is homoskedastic, the normal distributions that $S_{n\pi}$ will follow are the same. In summary, S_n is asymptotically equivalent to the supremum of infinite i.i.d. normal random variables. By the extreme value theory (see, e.g., Example 21.16 of van der Vaart (1998)), for n i.i.d. standard normal random variables X_i, $i = 1, \ldots, n$,

$$\begin{aligned} &P\left(\max_i X_i < \sqrt{2\log n} - \frac{1}{2}\frac{\log\log n + \log 4\pi}{\sqrt{2\log n}} + \frac{1}{\sqrt{2\log n}}x\right) \\ &\quad\to \exp(-\exp(-x)). \end{aligned}$$

It is easy to see the similarity of the asymptotic distributions of S_n and $\max_i X_i$. The "2" in $\exp(-2\exp(-x))$ is because S_n is the supremum of the absolute value of $S_{n\pi}$ instead of $S_{n\pi}$ itself. On the contrary, $\widetilde{I}_n$ takes an average form of $S_{n\pi}$, so its asymptotic distribution is normal. The simulation studies in Wu and Chu (1993a) show that the type I error using S_n to test H_0 is much higher than the nominal level, which is partially due to the low convergence rate of S_n to its asymptotic distribution. It is unknown whether the wild bootstrap is valid for S_n.

It is noteworthy that we need normalize S_n by a location constant a_n to get a nondegenerate asymptotic distribution. When there is no normalizing constant a_n, the asymptotic distribution of S_n will degenerate. To appreciate this result, consider $\max_i X_i$ again. It is shown in Gnedenko (1943) that for any $\varepsilon > 0$,

$$P\left(\left|\frac{\max_i X_i}{\sqrt{2\log n}} - 1\right| < \varepsilon\right) \to 1.$$

In this case, $\max_i X_i$ is called *relatively stable* by Gnedenko (1943).

Although the arguments above are elegant, they cannot be applied in our case. The key point here is that x_i is not evenly sampled and ε_i may be heteroskedastic, but we need $S_{n\pi}$ to be homoskedastic to apply the extreme value theory. Let us check the effect of either of the two conditions on S_n. If ε_i is homoskedastic and x_i is not evenly sampled, because h is the same for all $S_{n\pi}$, the asymptotic variance of $S_{n\pi}$ will be smaller when $f(\pi)$ is larger since more data are used in $S_{n\pi}$, and vice versa. To make the asymptotic variance invariant to π, we must normalize $S_{n\pi}$ by an estimator of $\sqrt{f(\pi)}$ or use variable bandwidths as in Fan and Gijbels (1992). Second, if ε_i is heteroskedastic and x_i is evenly sampled, similarly, the asymptotic variance of $S_{n\pi}$ will depend on π. We need divide $S_{n\pi}$ by an estimator of $\sigma(\pi)$ to make it homoskedastic; see Eubank and Speckman (1993) for discussions on this issue in confidence band construction of nonparametric regression. In summary, we need estimators of $f(\cdot)$ and $\sigma^2(\cdot)$ to make the extreme value theory apply. When $\varepsilon_i = \sigma(x_i)\epsilon_i$ with ϵ_i i.i.d. sampled, Hamrouni (1999) normalizes $S_{n\pi}$ by $K_4\sqrt{f(\pi)/\sigma^2(\pi)}$ to get the asymptotic distribution $\exp(-2\exp(-x))$, where K_4 is a constant only related to $k(\cdot)$. To our knowledge, there is no literature to consider normalizing $S_{n\pi}$ by an estimator of $K_4\sqrt{f(\pi)/\sigma^2(\pi)}$. Even there were, such complicated procedures are not attractive to practitioners.

The third test is proposed by Müller and Stadtmüller (1999). This test is based on sums of squared differences of the data, formed with various span sizes:

$$Z_k = \sum_{j=1}^{n-L}\frac{\left(y_{j+k}-y_j\right)^2}{n-L}, \quad 1\le k\le L,$$

where $L = L(n) \ge 1$ is a sequence of integers depending on n. Müller and Stadtmüller (1999) show that the statistics Z_k can be interpreted as dependent variables within the following two-parameter asymptotic linear model[8]

$$Z_k = 2\sigma^2 + \frac{k}{n-L}\alpha_\pi^2 + \eta_k, \quad 1\le k\le L. \tag{9}$$

Now, α_π^2 can be estimated by least squares and the asymptotic distribution of the estimator is as follows:

$$\sqrt{L}\left(\widehat{\alpha}_\pi^2 - \alpha_\pi^2\right) \xrightarrow{d} N\left(0, \frac{12}{5}\left(E\left[\varepsilon_i^4\right] - \sigma^4\right)\right).$$

A usual t test can be conducted to test whether $\alpha_\pi = 0$. Müller and Stadtmüller (1999) derive this test under the fixed design with i.i.d. errors, and it is unknown how to adjust this procedure to the case with random design and heterogeneous ε_i's.

4. Efficient estimation of the treatment effect

If the specification tests reject $H_0^{(2)}$ that $\alpha_\pi = 0$, we need to estimate α_π with an unknown π. Usually, we can use a two-step procedure: first estimate π, and then estimate α_π as if π were known.

4.1. Estimation of the discontinuity point

Because of the weighted average nature of kernel smoothers, $\widehat{\alpha}(\pi)$ would be near to zero when there was no jump at a

[8] This approximation is straightforward when $k = 1$. Müller and Stadtmüller (1999) allow an unknown number of jump points, and in this case, α_π^2 in (9) should change to the sum of squared jumping sizes.

given point. Otherwise, the difference would be near the jump magnitude. So a natural estimator of π is as follows:

$$\widehat{\pi} = \arg\max_{\pi \in \Pi} \widehat{\alpha}^2(\pi). \tag{10}$$

Such an estimator is called the *Difference Kernel Estimator* (DKE) in Qiu et al. (1991). In practice, we can specify $\Pi = [x_{(0.15n)}, x_{(0.85n)}]$ with $x_{(i)}$ being the ith order statistic of $\{x_i\}_{i=1}^n$; also, we need only check whether $\pi = x_i$, $x_i \in \Pi$, maximizes $\widehat{\alpha}^2(\pi)$ as in the parametric threshold regression scenario. Yu (2012) suggests to check the middle points of contiguous x_i's to improve the efficiency of $\widehat{\pi}$. We expect either way will generate similar estimates of α_π which is of main interest in RDDs. Before stating the asymptotic distribution of $\widehat{\pi}$, we impose the following standard restrictions on the bandwidth h.

Assumption H. $h \to 0$, $nh \to \infty$, and $\frac{\sqrt{nh}}{\ln n} \to \infty$.

Note that the this bandwidth h may be different from the h in Assumption B of specification testing.

The following theorem provides the asymptotic distribution of $\widehat{\pi}$.

Theorem 4. *Suppose Assumptions E, F(b), H and M hold, and $k(\cdot)$ satisfies the same conditions as specified in Assumption K; then $\widehat{\pi} - \pi_0 = O_p\left(n^{-1}\right)$, and*

$$n\left(\widehat{\pi} - \pi_0\right) \xrightarrow{d} \arg\max_{v \in \mathbb{R}} D(v),$$

where

$$D(v) = \begin{cases} \sum_{i=1}^{N_-(|v|)} \left(-\alpha_{\pi_0}^2 + 2\alpha_{\pi_0}\varepsilon_i^-\right), & \text{if } v < 0; \\ \sum_{i=1}^{N_+(v)} \left(-\alpha_{\pi_0}^2 - 2\alpha_{\pi_0}\varepsilon_i^+\right), & \text{if } v \geq 0; \end{cases}$$

$N_-(\cdot)$ and $N_+(\cdot)$ are Poisson processes with intensity $f(\pi_0)$, ε_i^- has the density $f(\varepsilon | x = \pi_0-)$, ε_i^+ has the density $f(\varepsilon | x = \pi_0+)$, and $\left\{\left\{\varepsilon_i^+\right\}_{i\geq 1}, \left\{\varepsilon_i^-\right\}_{i\geq 1}, N_-(\cdot), N_+(\cdot)\right\}$ are independent of each other. Furthermore, $D(v)$ is caglad with $D(0) = 0$ almost surely, and the asymptotic distribution of $\widehat{\pi}$ is the same as that in the case when α_{π_0} is known.

Note that there is an interval $(M_-, M_+]$ maximizing $D(v)$, so we must determine which point on the maximizing interval is taken as $\arg\max_{v \in \mathbb{R}} D(v)$. If we check $x_i \in \Pi$ when maximizing $\widehat{\alpha}^2(\pi)$ in (10), then $\arg\max_{v \in \mathbb{R}} D(v) = M_+$; if we check middle points, then $\arg\max_{v \in \mathbb{R}} D(v) = \frac{M_+ + M_-}{2}$. Also, when ε is independent of x, ε_i^- and ε_i^+ have the same distribution as ε_i. The explicit form of the asymptotic distribution of $\widehat{\pi}$ can be found in Appendix D of Yu (2012).

It is surprising that $\widehat{\pi}$ is n-consistent although $m(\cdot)$ takes a nonparametric form. In Appendix A of the supplementary materials, we provide some intuitions on this result. We also illustrate why the asymptotic distribution of $\widehat{\pi}$ is the same as in a parametric model and is the same as the least squares estimator. The result in Theorem 4 hinges on the assumption that $k(0) > 0$; that is, we indeed use some information in the neighborhood of π to estimate $m_\pm(\pi)$. When $k(0) = 0$ (and $k'(0) > 0$), the convergence rate is $\sqrt{nh}/h = n/\sqrt{nh}$ slower than n; see Müller (1992), Wu and Chu (1993a), and Delgado and Hidalgo (2000) for the details. Loader (1996) puts forward a similar estimator as ours but he only considers the case with the fixed design and the error term following the standard normal. Gréoire and Hamrouni (2002) essentially use the objective function $|\widehat{\alpha}(\pi)|$, but their asymptotic distribution of $\widehat{\pi}$ is questionable. This is because $\arg\max_{\pi \in \Pi} \widehat{\alpha}^2(\pi) = \arg\max_{\pi \in \Pi} |\widehat{\alpha}(\pi)|^\beta$ for any $\beta \geq 1$, and their asymptotic distribution should be the same as that in Theorem 4. As to the inferences of π, see Section 6 of Yu (2014a) for a summary in the parametric model. Since we are interested in α_π rather than π in RDDs, inference on π is not very important, and we only need a precise estimator of π such that its impact on the estimation of α_π is small.

When there is only one discontinuity point, Korostelev (1987) shows that n^{-1} is the optimal minimax rate to estimate π in the ideal Gaussian white noise model. Actually, we can adapt the proof idea of Yu (2008) to show that π can even be adaptively estimated. When the number of discontinuities is greater than 1, Spokoiny (1998) shows that the optimal minimax rate changes to $n^{-1}\log n$.

4.2. Estimation of the treatment effect

Given $\widehat{\pi}$, α_π is estimated by

$$\widehat{\alpha}_\pi = \widehat{\alpha}(\widehat{\pi}),$$

so $\widehat{\alpha}_\pi$ is a natural by-product of the estimation of π. The asymptotic distribution of $\widehat{\alpha}_\pi$ is as follows.

Theorem 5. *Under the assumptions in Theorem 4,*

$$\sqrt{nh}\left(\widehat{\alpha}(\widehat{\pi}) - \widehat{\alpha}(\pi_0)\right) = o_p(1).$$

Furthermore, $\widehat{\alpha}(\widehat{\pi})$ is asymptotically independent of $\widehat{\pi}$.

Theorem 5 claims that the estimation of π will not affect the efficiency of $\widehat{\alpha}_\pi$. In other words, π can be treated as known in evaluating the treatment effects in RDDs even if it is not. Combining this result with that in Theorem 4, we can treat α_π as if were known when estimating π, and treat π as if were known when estimating α_π. From Yu (2008), π is a "middle" boundary of x, and the information used in estimating the regular parameter α_π and the nonregular parameter π does not affect each other. Given this result, the asymptotic distribution of $\widehat{\alpha}(\widehat{\pi})$ is the same as that given in Theorem 3 of Porter (2003). Because the model with π unknown is more difficult than that with π known, while our estimator has the same efficiency as in the latter case, it achieves the optimal rate of convergence in the minimax sense as derived in Section 4 of Porter (2003).

When there are multiple, say q, discontinuity points, $\pi_1, \ldots, \pi_q$ in ascending order (WLOG, suppose $\alpha_{\pi_1} > \alpha_{\pi_2} > \cdots > \alpha_{\pi_{q-1}} > \alpha_{\pi_q}$), a sequential procedure can be used to detect them. Specifically, we first estimate π_1 by maximizing $\widehat{\alpha}^2(\pi)$ on $\Pi_1 = \Pi$, and then sequentially estimate π_j by maximizing $\widehat{\alpha}^2(\pi)$ on $\Pi_j = \Pi_{j-1} \setminus \bigcup_{k=1}^{j-1} [\widehat{\pi}_k - 2h, \widehat{\pi}_k + 2h]$, for $j = 2, \ldots, q$. Given $\widehat{\pi}_j$'s, α_{π_j} can be estimated straightforward; see Bertanha (2014) and references therein for estimating all kinds of treatment effects when $\pi_1, \ldots, \pi_q$ are known. Before stating the asymptotic distributions of $\widehat{\pi}_j$ and $\widehat{\alpha}_{\pi_j}$, we must assume π_j's are separated apart. Rigorously,

Assumption A. $\min_{j=2,\ldots,q} \left(\pi_j - \pi_{j-1}\right) > \epsilon > 0$.

The following corollary provides the joint asymptotic distribution of $\left\{\widehat{\pi}_j, \widehat{\alpha}_{\pi_j}\right\}_{j=1}^q$.

Corollary 1. *Suppose Assumptions A, E, F(b), H and M hold, and $k(\cdot)$ satisfies the same conditions as specified in Assumption K,*

(i) *$\widehat{\pi}_j - \pi_{j0} = O_p\left(n^{-1}\right)$, $j = 1, \ldots, q$, and $n\left(\widehat{\pi}_j - \pi_{j0}\right)$ has the same asymptotic distribution as $n\left(\widehat{\pi} - \pi_0\right)$ in Theorem 3 except that π_0 is replaced by π_{j0} in $D(v)$.*

(ii) $\sqrt{nh}\left(\widehat{\alpha}_{\pi_j}-\widehat{\alpha}_{\pi_{j0}}\right)=o_p(1)$ *for all* $1\le j\le q$.
(iii) *All* $\widehat{\pi}_j$*'s and* $\widehat{\alpha}_{\pi_j}$*'s are asymptotically independent of each other.*

This corollary claims that each $\left(\pi_j,\alpha_{\pi_j}\right)$ can be estimated independent of the others and as if the others were known. This is because each $\widehat{\pi}_j$ is n-consistent, and w.p.a.1. π_{j0} will stay in the h neighborhood of $\widehat{\pi}_j$. Given that we exclude a $2h$ neighborhood of $\widehat{\pi}_j$, we essentially use an independent data set to estimate $\left(\pi_{j+1},\alpha_{\pi_{j+1}}\right)$, and consequently, $\left(\widehat{\pi}_{j+1},\widehat{\alpha}_{\pi_{j+1}}\right)$ is asymptotically independent of $\left(\widehat{\pi}_j,\widehat{\alpha}_{\pi_j}\right)$ and is not affected by $\left(\widehat{\pi}_j,\widehat{\alpha}_{\pi_j}\right)$.

There is also some literature discussing the L^p-consistency of the whole $m(\cdot)$ function, where $p\ge 1$. Müller (1992) uses the boundary kernel developed in Gasser and Müller (1979) to deal with the estimation of $m(\cdot)$ in the neighborhood of the discontinuity point, and derives the L^p convergence rate of his estimator. Oudshoorn (1998) uses the L^2 norm, and allows multiple discontinuity points (with unknown numbers and locations) in the ideal white noise model. He concentrates on the asymptotically minimax estimation of $m(\cdot)$ based on series, and shows that the minimax rate is the same as that in Stone (1982) where no jumps exist.

5. Extensions and a practical issue

In this section, we discuss three extensions of the results in Sections 3 and 4. Also, we deal with an important practical issue, the bandwidth selection, in both specification testing and estimation.

5.1. Two auxiliary tests

In this subsection, we will provide two auxiliary tests. The first one is to check Assumption F. There are two parallel tests as in Section 2.3. The first test is to check whether $f(\cdot)$ is smooth at π; that is, we want to know whether there is imprecise control of x. The second test is to check whether there is a jump in $f(\cdot)$ at π. McCrary (2008) interprets the second test as an indicator of whether there is precise manipulation of x. He assumes π is known and uses the local linear density estimator of Cheng (1994) and Cheng et al. (1997) to test this hypothesis in the spirit of the first or second alternative test (which are equivalent when π is known) in Section 3.3. Our test statistic T_n can be extended to apply in both tests even if π is unknown. The only adjustment is to redefine x_i and y_i. For this purpose, we first divide Π_ϵ into $N=\left[\frac{\overline{\pi}+\epsilon-(\underline{\pi}-\epsilon)}{h}\right]+1$ subintervals, and denote the middle-points of these intervals as $\{\pi_l\}_{l=1}^N$. Then define $y_l=\left(n\widetilde{h}\right)^{-1}\sum_{i=1}^n 1\left(\pi_l-\frac{\widetilde{h}}{2}\le x_i<\pi_l+\frac{\widetilde{h}}{2}\right)$. Now, substitute (x_i,y_i) in I_n by (π_l,y_l), and make sure that $\widetilde{h}=o(h)$ and $h=o(b)$; then the results in Theorems 1 and 2 apply. As argued in Section 3.2 of McCrary (2008), his procedure is robust to the choice of h, and a similar argument can be applied in our case. As to the choice of h and b, see Section 5.4. As noted in McCrary (2008), "a running variable with a continuous density is neither necessary nor sufficient for identification except under **auxiliary** assumptions". Actually, Lee (2008) claims that the treatment effect can be identified even if there were manipulation as long as the treatment status cannot be precisely controlled by individuals. A testable corollary of his framework is that the conditional mean of any pre-determined variable given x is continuous at π. Of course, our testing procedure can also be used to test such a hypothesis.

The second test is to check whether there are enough compliers in the fuzzy design. To describe this test, first recall the identification structure in the fuzzy design. From Hahn et al. (2001), under the local unconfoundedness condition, the treatment effect Δ_π can be identified, and

$$\Delta_\pi\equiv E\left[Y_1-Y_0|x=\pi\right]=\frac{\alpha_\pi}{\beta_\pi},\tag{11}$$

where $\beta_\pi=p_+(\pi)-p_-(\pi)$, and $p_\pm(\pi)=E\left[D|x=\pi\pm\right]$ are propensity scores in the right and left neighborhoods of π with D being the treatment status. Under the monotonicity assumption as used in Imbens and Angrist (1994), the denominator of Δ_π, β_π, can be interpreted as the probability to be a complier. To check whether $\Delta_\pi=0$ by testing whether $\alpha_\pi=0$, we must make sure $\beta_\pi\ne 0$; that is, there are enough compliers. Our test statistic T_n can be applied directly except that D now plays the role of y. Note also that because $|\beta_\pi|<1$, the test of $\Delta_\pi=0$ based on testing $\alpha_\pi=0$ as in Theorem 2 is conservative.

5.2. Estimation in the fuzzy design

We first discuss estimation of π. In the fuzzy design, the available data are $\{y_i,x_i,D_i\}_{i=1}^n$. The extra data D_i can provide extra information about π. To be specific, our estimator of π is

$$\widetilde{\pi}=\arg\max_{\pi\in\Pi}\left[\widehat{\alpha}^2(\pi)+\widehat{\beta}^2(\pi)\right],$$

where $\widehat{\beta}(\pi)=\widehat{p}_+(\pi)-\widehat{p}_-(\pi)$ is the estimated jump of the propensity score at π, and $\widehat{p}_\pm(\pi)$ are nonparametric estimators of $p_\pm(\pi)$. The asymptotic distribution of $\widetilde{\pi}$ is the same as that of $\widehat{\pi}$ in Theorem 3 except that the $D(v)$ process is redefined as

$$D(v)=\begin{cases}\sum_{i=1}^{N_-(|v|)}\left(-\alpha_{\pi_0}^2+2\alpha_{\pi_0}\varepsilon_i^- -\beta_{\pi_0}^2+2\beta_{\pi_0}\eta_i^-\right), & \text{if } v<0;\\ \sum_{i=1}^{N_+(v)}\left(-\alpha_{\pi_0}^2-2\alpha_{\pi_0}\varepsilon_i^+ -\beta_{\pi_0}^2-2\beta_{\pi_0}\eta_i^+\right), & \text{if } v\ge 0;\end{cases}$$

where $\eta_i=D_i-E\left[D_i|x=x_i\right]$, and $\eta_i^\pm$ are similarly defined as $\varepsilon_i^\pm$.[9] Because $E\left[-\alpha_{\pi_0}^2\pm 2\alpha_{\pi_0}\varepsilon_i^\mp-\beta_{\pi_0}^2\pm 2\beta_{\pi_0}\eta_i^\mp\right]=-\alpha_{\pi_0}^2-\beta_{\pi_0}^2<-\alpha_{\pi_0}^2=E\left[-\alpha_{\pi_0}^2\pm 2\alpha_{\pi_0}\varepsilon_i^\mp\right]$, we expect that $\widetilde{\pi}$ is more efficient than $\widehat{\pi}$.

Given an estimator of π, the treatment effect Δ_π defined in (11) can be estimated as in Section 3.6 of Porter (2003) or in Section 3 of Yu (2013). Of course, due to the superconsistency of $\widetilde{\pi}$, the asymptotic distribution of the Δ_π estimator is the same as that in the case when π_0 is known. For this asymptotic distribution, we refer to Proposition 1 of Porter (2003) and Theorems 3 and 4 of Yu (2013).

5.3. The number of discontinuity points is unknown

Usually, the number of discontinuity points q is known in RDDs. When q is unknown, there are three classes of procedures to get information about q. To describe these procedures, we maintain Assumption A.

First, we can conduct tests about q sequentially; Wu and Chu (1993a) belongs to this class. For this method, we use the notations in Section 4.2. Specifically, we first test $H_0:q=0$ versus $H_1:q>0$ using some test statistic T_n on Π_1. If H_0 is rejected, then estimate π_1 using the procedure in Section 4.2. Generally, test $H_0:q=j$ versus $H_1:q>j$ using T_n on Π_{j+1}. Keep on this procedure until H_0 cannot be rejected. Remark 4 of Wu and Chu (1993a) uses the sup-statistic S_n as T_n. Of course, our test statistic T_n can also be used. Because of the type-I error in testing, this method tends to overestimate q. Of course, we can let the significance level in each test shrink to zero to make the tests pick q consistently.

[9] Note that $\eta_i^\pm$ are discretely distributed. Nevertheless, the jumps in $D(v)$ are still continuously distributed as $\varepsilon_i^\pm$ are continuously distributed, so $\arg\max_v D(v)$ is uniquely defined. Of course, if π is estimated based on $\arg\max_{\pi\in\Pi}\widehat{\beta}^2(\pi)$, the maximizer of $D(v)$ may not be uniquely defined; see Yu (2010) for conditions that guarantee the uniqueness of $\arg\max_v D(v)$.

Second, we can estimate q; Yin (1988) and Qiu (1994) belong to this class. We briefly describe the procedure of Qiu (1994) here.[10] Because the kernel estimator uses the data in a $2h$ range, we divide Π into $N = \left[\frac{\overline{\pi}-\underline{\pi}}{2h}\right] + 1$ (closed) subintervals. Denote the right endpoints of these intervals as π_l, where $\pi_l = \underline{\pi} + \left(\overline{\pi} - \underline{\pi}\right) l/N$, $l = 1, \ldots, N$. Calculate $\widehat{\alpha}(\pi)$ at all π_l's, and find the π_l's such that $|\widehat{\alpha}(\pi_l)| > B_n$, where $B_n = O\left(\beta_n \sqrt{\frac{\log n}{nh}}\right)$ with $\lim_{n\to\infty} \beta_n = \infty$ such as $\beta_n = \log\log n$. Such a B_n is understandable because when there is no jump, $\sup_{\pi\in\Pi} |\widehat{\alpha}(\pi)| = O_p\left(\sqrt{\frac{\log n}{nh}}\right)$ from Section 3.3. Denote such π_l's as $\left\{\pi_{l_1}, \ldots, \pi_{l_r}, \ldots, \pi_{l_R}\right\}$. In the neighborhood of a discontinuity point, there may be more than one π_{l_r}'s such that $\left|\widehat{\alpha}(\pi_{l_r})\right| > B_n$, so the contiguous π_{l_r}'s are indicating the same discontinuity point. Combining the contiguous subintervals whose left endpoint is π_{l_r}, we get intervals $\mathcal{I}_1, \ldots, \mathcal{I}_k, \ldots \mathcal{I}_{\widehat{q}}$. $\widehat{q}$ is the estimator of q, and the middle point of each $\mathcal{I}_k$ or the maximizer of $|\widehat{\alpha}(\pi)|$ on each $\mathcal{I}_k$, $k = 1, \ldots, \widehat{q}$, is the estimator of the location of jumps. Qiu (1994) shows the a.s. consistency and convergence rate of $\widehat{q}$.

Third, we can select q to obtain an optimal fitting. This approach is proposed in Braun and Müller (1998) and used in Müller and Stadtmüller (1999) and Gijbels and Goderniaux (2004), but no theoretical justification is provided yet. Specifically, we can choose q and h together when minimizing the objective function (12) of cross-validation below, where $\widehat{m}_{-i}(x_i)$ is a function of both q and h.

5.4. Bandwidth selection in specification testing and estimation

Most statistical and econometric literature concentrates on the bandwidth selection in estimation instead of specification testing (in classical specification tests, Hsiao et al. (2007) is the only exception). We will first provide an algorithm for the bandwidth selection in estimation, and then adjust the algorithm to select bandwidth in specification testing. Throughout our discussion, we assume there is at most one cusp or jump under the null and alternative.

For statistical literature on bandwidth selection when $m(\cdot)$ is discontinuous, see, e.g., Remark 2 of Wu and Chu (1993a,b), and Section 5 of Müller (1992).[11] It seems that only Wu and Chu (1993b) give a rigorous analysis. There is also some econometric literature on bandwidth selection in the estimation of α_π when π is *known*. Ludwig and Miller (2005) use the conventional CV approach to select the bandwidth. To avoid the boundary problem, they use the one-side kernel in estimating $m(\cdot)$ on both sides of π. Such a scheme induces some efficiency loss in data usage; see Ludwig and Miller (2005) for other disadvantages of this boundary CV method. DesJardins and McCall (2008) and Imbens and Kalyanaraman (2012) use the plug-in method, but as argued in Loader (1999), "plug-in approaches are tuned by arbitrary specification of pilot estimators and prone to oversmoothing when presented with difficult smoothing problems". See Imbens and Kalyanaraman (2012) for detailed descriptions on these econometric methods.

Cross-validation remains the dominating approach in bandwidth selection for practitioners. Our approach is based on and extends the CV method in Wu and Chu (1993b). Different from the CV method of Ludwig and Miller (2005), Wu and Chu (1993b) eliminate the boundary effect by projecting the data near the discontinuity point. Hall and Wehrly (1991) propose a similar procedure but reflect the data near the boundary; their idea is recently used in testing for smooth structural changes by Chen and Hong (2012). Both Hall and Wehrly (1991) and Wu and Chu (1993b) use the kernel smoother (i.e., the local constant smoother). Given that the LLS is dominating in practice, we adapt their procedures to our case by using the LLS. Because the procedure based on projection or reflection is not popular in the literature of RDDs, we provide a detailed description here.

Algorithm CV. Step 1: Fix $h \in H_n \equiv \left[An^{-1+\rho}, Bn^{-\rho}\right]$ for arbitrary small ρ and constants A and B.

Step 2: Estimate π using (10) with the bandwidth ch, where c is determined by

$$c = \frac{C\left(k_+^*\right)}{C\left(k\right)},$$

where $C(K) = \left\{\frac{\int K^2(u)du}{\left(\int u^2 K(u)du\right)^2}\right\}^{1/5}$ for a kernel function $K(u)$,

$$k_+^*(u) = \frac{\left(\int_0^1 t^2 k(t)dt\right) k(u) - \left(\int_0^1 tk(t)dt\right) uk(u)}{\int_0^1 k(t)dt \int_0^1 t^2 k(t)dt - \left(\int_0^1 tk(t)dt\right)^2}$$

is the equivalent kernel of $k_+(u) \equiv k(u)1(0 \le u \le 1)$ in the local linear regression.[12] To avoid the difficulty in deciding that x_i should be used in estimating $m_+(x_i)$ or $m_-(x_i)$ and improve efficiency of $\widehat{\pi}$, check the middle points of contiguous x_i's in Π in maximizing $\widehat{\alpha}^2(\pi)$.[13]

Step 3: Given $\widehat{\pi}$, $\widehat{m}_+(\widehat{\pi})$ and $\widehat{m}_-(\widehat{\pi})$ in step 2, generate pseudo-data $(\widetilde{x}_{i+}, \widetilde{y}_{i+}) \equiv (2\widehat{\pi} - x_{i-}, 2\widehat{m}_-(\widehat{\pi}) - y_{i-})$, $i = 1, \ldots, L_h$, in the right h neighborhood and $(\widetilde{x}_{i-}, \widetilde{y}_{i-}) \equiv (2\widehat{\pi} - x_{i+}, 2\widehat{m}_+(\widehat{\pi}) - y_{i+})$, $i = 1, \ldots, R_h$, in the left h neighborhood of $\widehat{\pi}$, where $\{x_{i-}, y_{i-}\}_{i=1}^{L_h}$ are the data points such that $\widehat{\pi} - x_{i-} \le h$, and $\{x_{i+}, y_{i+}\}_{i=1}^{R_h}$ are the data points such that $x_{i+} - \widehat{\pi} \le h$.

Step 4: For $x_i \in \left[\underline{\pi}, \widehat{\pi}\right]$, use $\{x_{i-}, y_{i-}\}_{i=1}^{L} \bigcup \{(\widetilde{x}_{i+}, \widetilde{y}_{i+})\}_{i=1}^{L_h}$ with $x_{i-} \in \left[\underline{\pi} - h, \widehat{\pi}\right]$ and for $x_i \in [\widehat{\pi}, \overline{\pi}]$, use $\{x_{i+}, y_{i+}\}_{i=1}^{R} \bigcup \{(\widetilde{x}_{i-}, \widetilde{y}_{i-})\}_{i=1}^{R_h}$ with $x_{i+} \in [\widehat{\pi}, \overline{\pi} + h]$ to get the "leave-one-out" version of $\widehat{m}(x_i)$. Here, the bandwidth of the LLS is h, and the resulting estimates are denoted as $\widehat{m}_{-i}(x_i)$.

Step 5: Get the objective function of CV:

$$CV(h) = \sum_{i=1}^{n} \left(\widehat{m}_{-i}(x_i) - y_i\right)^2 1_i^{\Pi} \tag{12}$$

for each h, and minimize $CV(h)$ for $h \in H_n$ to find the CV bandwidth $\widehat{h}_{CV}$.

[10] Oudshoorn (1998) uses a similar procedure based on wavelet in the ideal white noise model. Yin (1988) first uses a sequential procedure as in Section 4.2 to estimate the location and size of jumps, and then estimates the number of jumps based on the truncated cumulative magnitude of jumps plus a penalty term, just as in the AIC and BIC selection of OLS regressors. Yin (1988)'s procedure checks every design point to find the jump location, while Qiu (1994) only checks multiples of the bandwidth.

[11] See also Gijbels and Goderniaux (2004) for the bandwidth selection in a two-step estimation scheme proposed in Gijbels et al. (1999). They suggest to use different bandwidths in estimating π and α_π, which is now a common sense in the literature, but their bandwidth selection in estimating π critically depends on the assumption that ε is independent of x.

[12] The constant c is determined as follows. It is well known that the optimal bandwidth minimizing the MSE is $n^{-1/5}C(K)C\left(m, f, \sigma^2\right)$ for both the interior point and the boundary point, where $C\left(m, f, \sigma^2\right)$ is a constant related to $m(\cdot)$, $f(\cdot)$ and $\sigma^2(\cdot)$ but not to $K(\cdot)$, so the optimal bandwidths for these two kinds of points are different only by a constant ratio c if $m(\cdot)$, $f(\cdot)$ and $\sigma^2(\cdot)$ are stable on the interested area.

[13] Also, in Step 3 below, we do not need to consider the reflection of (x_i, y_i) itself.

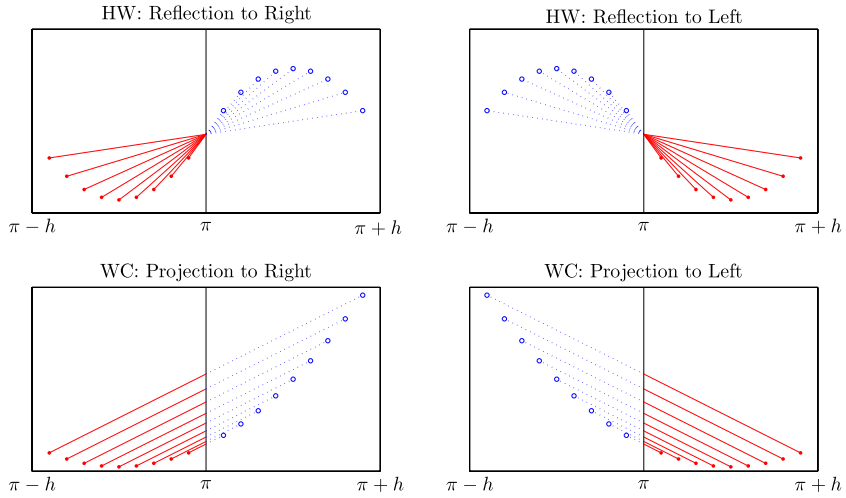

Fig. 2. Comparison of reflection in Hall and Wehrly (1991) and projection in Wu and Chu (1993b): HW for Hall and Wehrly (1991) and WC for Wu and Chu (1993b), red dot for real data, and blue circle for reflected or projected data. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

This algorithm is designed for the case where π is unknown. When π is known, it can be simplified in an obvious way. Note that $\widehat{\pi}$ and $\widehat{\alpha}_\pi$ are by-products of Algorithm CV; also note that $c \cdot \widehat{h}_{CV}$ rather than $\widehat{h}_{CV}$ is used in estimating α_π. In Step 2, we use a delicate bandwidth in estimating π to achieve an optimal fitting in (12). But ch may not be a good choice in estimating π. For example, suppose $k(u) = 3\left(1-u^2\right)1\left(|u| \leq 1\right)/4$, the Epanechnikov kernel, it can be shown that $c = 1.554$ which is greater than one. This is reasonable because we need more data in estimating $m(\cdot)$ at the boundary to improve the efficiency. But our purpose is to precisely estimate π rather than precisely estimate $m(\cdot)$ around π. Because π is identified by the jump of $m(\cdot)$, a small bias in estimating $m(\cdot)$ is more important than a small variance. So we suggest to use a smaller bandwidth than h in estimating π. This is also a common sense in simulation studies in the literature.

Now, we describe how to determine H_n and minimize $CV(h)$. Suppose a vector $\left(x_{(1)}, \ldots, x_{(\underline{n})}\right)$ includes the sorted x_i's in Π_ϵ, where Π_ϵ can be specified as $[x_{(0.1n)}, x_{(0.9n)}]$ in practice. Define the lower bound of H_n as $\underline{h} \equiv \max\left\{x_{(i+1)} - x_{(i)} : i = 1, \ldots, \underline{n}-1\right\}$. This $\underline{h}$ guarantees that there are at least two data points in a $\underline{h}$ neighborhood of any $x \in \Pi$, so the multicollinearity problem in the LLS can be avoided. The upper bound $\overline{h}$ of H_n can be set as $\frac{1}{2}\left(\overline{\pi} - \underline{\pi}\right)$, which roughly corresponds to the parametric estimation of $m(\cdot)$ when π is in the middle of Π. When minimizing $CV(h)$, we need only search over a discretized subset D_h of H_n; e.g., $D_h = \left\{\underline{h} + i \cdot \text{step} : i = 0, 1, \ldots, \left[\frac{\overline{h}-\underline{h}}{\text{step}}\right]\right\}$, where step $= \frac{1}{2}\min\left\{x_{(i+1)} - x_{(i)} : i = 1, \ldots, \underline{n}-1\right\}$. The step size in D_h ensures that at most one more data point is covered by the $\underline{h}+(i+1)\cdot$ step ball than the $\underline{h}+i\cdot$step ball centered at any $x \in \Pi$, so essentially all possible LLS estimates are considered for a fixed data set. If such a step size is too small, we may set step $= \frac{1}{2}\frac{\overline{\pi}-\underline{\pi}}{\underline{n}}$. In this case, $CV(h)$ is evaluated at roughly $\underline{n}$ points. Such a specification of H_n and D_h is not rigid. In practice, we must make sure that the minimizer is not obtained at the boundary points, $\underline{h}$ and $\overline{h}$, of H_n, and hope that $CV(h)$ is relatively flat near its minimizer so that the estimation of α_π is relatively robust to the bandwidth selection.[14]

Wu and Chu (1993b) use a little different approach in Step 2 because their estimation procedure assumes $k(0) = 0$ as mentioned in Section 4.1. This results in a smaller bandwidth in estimating π. A smaller bandwidth can make the uncertainty of $\widehat{\pi}$ in evaluating $CV(h)$ disappear. Also, they generate a different pseudo-data set in Step 3. In their case, $(\widetilde{x}_{i+}, \widetilde{y}_{i+}) = \left(2\widehat{\pi} - x_{i-}, y_{i-} + 2\widehat{m}'_-(\widehat{\pi})(\widehat{\pi} - x_{i-})\right)$, $i = 1, \ldots, L_h$, and $(\widetilde{x}_{i-}, \widetilde{y}_{i-}) \equiv \left(2\widehat{\pi} - x_{i+}, y_{i-} + 2\widehat{m}'_+(\widehat{\pi})(x_{i+} - \widehat{\pi})\right)$, $i = 1, \ldots, R_h$. Our procedure in Step 3 is based on Hall and Wehrly (1991). Comparison of these two procedures is illustrated in Fig. 2. From Fig. 2, when reflecting the data to the right (left) neighborhood of $\widehat{\pi}$, Hall and Wehrly (1991) reflect the ray of light through the fixed point $(\widehat{\pi}, \widehat{m}_-(\widehat{\pi}))$ $((\widehat{\pi}, \widehat{m}_+(\widehat{\pi})))$, while Wu and Chu (1993b) project the ray of light parallelly with the slope $\widehat{m}'_-(\widehat{\pi})$ $(\widehat{m}'_+(\widehat{\pi}))$. Intuitively, both methods maintain the linear component of $m(\cdot)$ in the neighborhood of $\widehat{\pi}$, so the bias term in estimating $m(\cdot)$ would be $O\left(h^2\right)$ which is of the same order at the interior point. Because we only reflect the data in a h neighborhood of $\widehat{\pi}$, the resulting contribution to the MISE in $CV(h)$ is $O\left(\left(h^2\right)^2 h\right) = O\left(h^5\right)$. This contribution is negligible because the bias component of the MISE from other area in $CV(h)$ is $O\left(h^4\right)$. Without reflection or projection, this contribution would be $O\left(h^2 h\right) = O\left(h^3\right)$ and is not negligible. We rate Hall and Wehrly (1991)'s method over Wu and Chu (1993b)'s because levels $\widehat{m}_\pm(\widehat{\pi})$ can be estimated more precisely than slopes $\widehat{m}'_\pm(\widehat{\pi})$.

We next describe our bandwidth selection of b (and h) in specification testing. The key difference between specification testing and estimation is that there may or may not be a cusp or discontinuity in the data, so we do not need Steps 2 and 3 in Algorithm CV. In Step 4, we use the original data to estimate $\widehat{m}_{-i}(x_i)$. Such an algorithm is to choose the best smooth fit of the original data, so matches the idea in the construction of I_n. If the DGP is from $H_0^{(1)}$, then balancing the integrated bias squared $(O(b^4))$ and variance $\left(O\left(\frac{1}{nb}\right)\right)$, we get $\widehat{b}_{CV} = O\left(n^{-1/5}\right)$, the most popular rate of optimal bandwidth. If the DGP is from $H_1^{(1)}$, then balancing the integrated bias squared $(O(b^3))$ and variance $\left(O\left(\frac{1}{nb}\right)\right)$, we get $\widehat{b}_{CV} = O\left(n^{-1/4}\right)$. At last, if the DGP is from $H_1^{(2)}$, then balancing the integrated bias squared $(O(b))$ and variance $\left(O\left(\frac{1}{nb}\right)\right)$, we get $\widehat{b}_{CV} = O\left(n^{-1/2}\right)$. So the CV procedure is adaptive to both the null DGP and the alternative DGP: when $m(\cdot)$ gets rougher, our bandwidth gets smaller, which is exactly what we want in Assumption B. Given $\widehat{b}_{CV}$, $\widehat{h}$ can be set as $\widehat{b}_{CV}^{2.1}$ to guarantee that $h^{1/2}/b \to 0$.

As mentioned above, we recommend to use a small bandwidth in estimating π when $\alpha_\pi \neq 0$. The bandwidth selection method

[14] See the figures in Lee and Lemieux (2010) for some intuitions on this fact. This is also why the CV bandwidth converges to the optimal value in a very slow rate $n^{-1/10}$ as shown in Wu and Chu (1993b).

in specification testing can serve this purpose. In this case, the bandwidth is $O\left(n^{-1/2}\right)$, much smaller than the optimal bandwidth (which is $O(n^{-1/5})$) in estimating α_π. Actually, this choice is also suggested in Wu and Chu (1993b).

We do not suggest to use the bandwidths above blindly. Instead, our procedure of bandwidth selection better serves as a method of strengthening and rigorizing the intuition in a practical problem. Before conducting any analysis in this paper, we suggest to plot local moving averages for a range of bandwidths (centering at the CV bandwidth or the rule of thumb bandwidth of Silverman (1986)) or globally polynomial fitting for different orders to get some senses about where π is and whether there are selection or treatment effects. Even in a rigorous analysis, we suggest to check a range of bandwidths to explore the sensitivity of our estimators and testing procedures to the bandwidth selection. For example, in specification testing, we can draw a graph of decision (0 for acceptance and 1 for rejection) against bandwidth and check how the decision changes when the bandwidth is smaller or larger than $\widehat{b}_{CV}$.

6. Monte Carlo results

The goal of this Monte Carlo study is to check how our tests and estimators depend on the bandwidth selection. We will use the four setups of Fig. 1 in our simulations.

$$DGP1: y = x^2 1(-2 \le x \le 3) + \varepsilon,$$
$$DGP2: y = x^2 1(-2 \le x < 1) + \left[(x-3)^2 - 3\right] 1(1 \le x \le 3) + \varepsilon,$$
$$DGP3: y = x^2 1(-2 \le x < 1) + \left(x^2 + 1\right) 1(1 \le x \le 3) + \varepsilon,$$
$$DGP4: y = x^2 1(-2 \le x < 1) + \left[(x-3)^2 - 2\right] 1(1 \le x \le 3) + \varepsilon.$$

In all setups, ε's are i.i.d. sampled and follow $N(0, 0.2^2)$, and x is uniformly distributed on $[-2, 3]$. The parameter space $\Pi = [0.5, 1.5]$, and the number of discontinuity points is known as 1. In specification testing, we will study size under the null and also power under the alternative. The rejection probability under $DGP1$ is the size for testing both (3) and (4). The rejection probability under $DGP2$ is the power for testing (3) and the size for testing (4). The rejection probabilities under $DGP3$ and $DGP4$ are the power for testing (4). The significance level is set at 5%. In estimation, we will consider only $DGP3$ and $DGP4$ and study the bias and variance properties of the estimators of π and α_π.

Because we will check the performance of our procedures for each bandwidth in a reasonable range, the computation burden is tremendous. To save simulation time, we let $n = 500$ which guarantees that about 100 data points fall in Π. The number of replications is set as 100. In the bootstrap method of Section 3.2, B is set as 199. Even in this small simulation study, many of our theoretical results are confirmed. Throughout the simulations, all estimators are based on the LLS with the kernel function $k(u) = 3\left(1 - u^2\right) 1(|u| \le 1)/4$.

6.1. Specification testing

In this simulation study, h is set as $b^{2.1}$. The construction of the test statistic and the bootstrap method of obtaining critical values can be found in Sections 3.1 and 3.2, respectively. The simulation results are summarized in Fig. 3. A few conclusions from Fig. 3 are as follows. (i) For the same bandwidth, when $m(\cdot)$ gets rougher, the probability of rejection gets higher. This matches the construction of T_n because the bias will increase when $m(\cdot)$ gets rougher. Similarly, given a DGP, the larger the bandwidth, the higher the probability of rejection. (ii) In both tests, the bootstrap distribution has a better finite-sample approximation to the distribution of T_n than the asymptotic distribution under the null. In testing (3), the bootstrap method has an almost correct size for a wide range of bandwidths, while the asymptotic method is either over-sized or under-sized so that it is hard to pick the right bandwidth. In consideration of the power properties, it seems that [0.25, 0.35] (the bandwidths between the two magenta vertical lines) is a suitable range for b. In testing (4), similarly, the size property of the asymptotic method is not satisfying. For the bootstrap method, [0.1, 0.15] (the bandwidth below the cyan vertical line) seems to be a suitable range of b because under the null which covers both $DGP1$ and $DGP2$, the sizes match the target 5% for b in this range. These ranges of bandwidths also justify Assumption B: a smaller bandwidth should be used in testing (4) than in testing (3). At last, the bootstrap method has a better power in testing (4) than the asymptotic method for $b \in [0.1, 0.15]$. The powers under $DGP3$ and $DGP4$ are similar. This is because the power is mostly generated by the jump instead of the cusp in testing (4). The lesson here is that the bootstrap method is indeed preferable than the asymptotic method in both the robustness to the bandwidth selection and size and power properties of the specification tests.

6.2. Estimation

In this simulation study, besides checking how our estimators of π and α_π depend on the bandwidth selection, we also illustrate the efficiency loss in estimating α_π with an unknown π. Our simulation results are summarized in Fig. 4 and Fig. 5 for $DGP3$ and $DGP4$, respectively. The optimal h's in estimating π in the two figures are based on the RMSE risk of $\widehat{\pi}$.

Some common structures in these two figures are as follows. (i) The risk of $\widehat{\pi}$ is much smaller than $\widehat{\alpha}_\pi$, which confirms the superconsistency of $\widehat{\pi}$. (ii) When the bias of $\widehat{\pi}$ is small, the risk of $\widehat{\alpha}_\pi$ is close to the case with π known, which confirms the result that $\widehat{\pi}$ will not affect the efficiency of $\widehat{\alpha}_\pi$. Especially, when the optimal bandwidth in estimating π is used, the risk of $\widehat{\alpha}_\pi$ is almost not affected by the estimation of π. But when the bias of $\widehat{\pi}$ is large, there are indeed considerable efficiency losses in $\widehat{\alpha}_\pi$ compared with the case where π is known. (iii) The bandwidth suitable for $\widehat{\pi}$ should be much smaller than that for $\widehat{\alpha}_\pi$. Due to the smoothness of $m(\cdot)$ in our specification, the RMSE risk of $\widehat{\alpha}_\pi$ is decreasing even at $h = 1$.

There are indeed some differences between Figs. 4 and 5. Under $DGP3$, the slope and curvature of $m(\cdot)$ in the left and right neighborhoods of π are the same, so even if we use the same bandwidth to estimate π and α_π, there is not much efficiency loss in estimating α_π. While under $DGP4$, where the structures of $m(\cdot)$ are very different in the left and right neighborhoods of π, choosing a different bandwidth in estimating α_π is very important to get satisfactory performance of $\widehat{\alpha}_\pi$. The lesson here is that when we suspect there is selection in the data, bandwidth selection should be very careful in estimating π and α_π.

7. Conclusion and future research

This paper studies the specification testing and estimation in regression discontinuity designs with unknown discontinuity point. In the specification testing, we construct a unified test statistic in testing the presence of both the selection effect and the treatment effect; the only difference is to use different bandwidths. Also, a bootstrap procedure is suggested to obtain critical values in finite samples. In estimation, we first estimate the unknown discontinuity point, and then estimate the treatment effect as if the discontinuity point were known. It is shown that the estimation of the discontinuity point will not affect the efficiency of the treatment effect estimator.

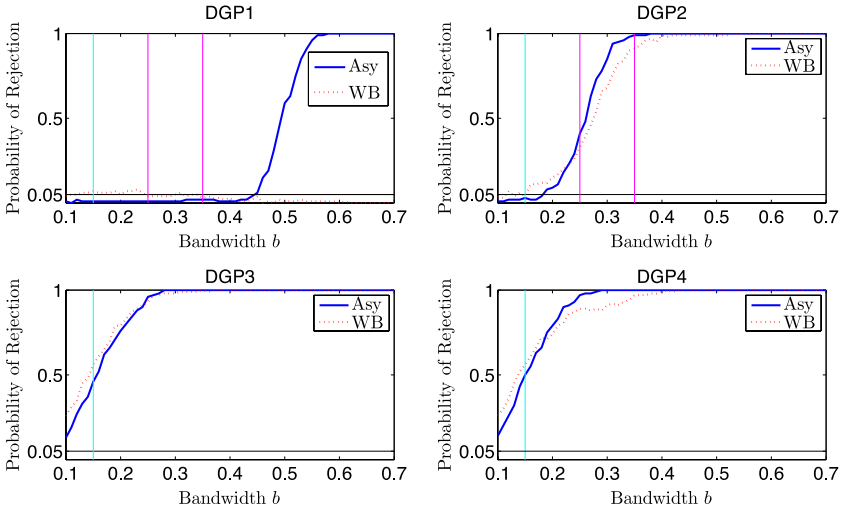

Fig. 3. Probability of rejection based on T_n under different DGPs.

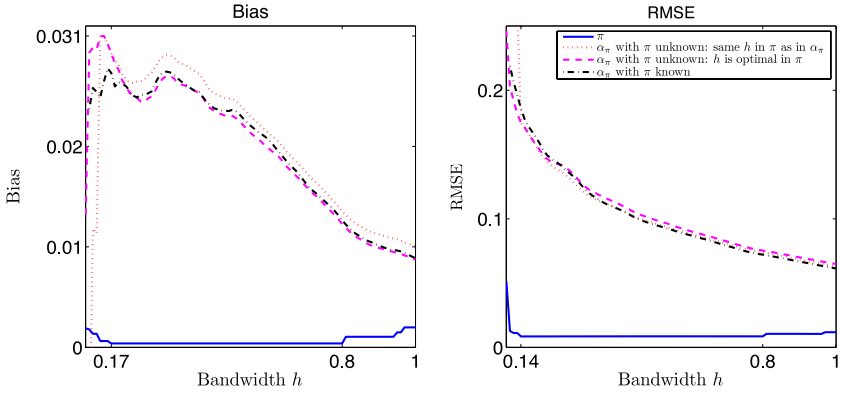

Fig. 4. Comparison of estimators of π and α_π in bias and RMSE in *DGP*3.

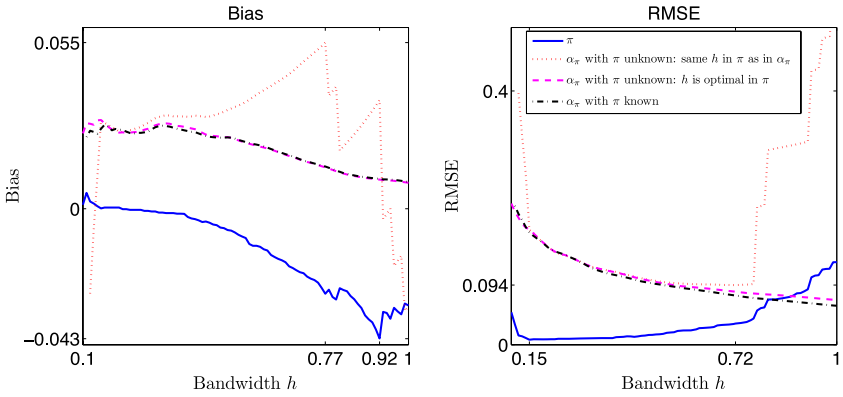

Fig. 5. Comparison of estimators of π and α_π in bias and RMSE in *DGP*4.

There are some interesting problems unsolved in this paper, some of which have already been mentioned in the main text. First, the Edgeworth expansion of the wild bootstrap test statistic is attractable for a theoretical justification of our testing procedure. Second, a key assumption of this paper is that the smoothness index s is known beforehand. If s is unknown, we can extend Horowitz and Spokoiny (2001) in specification testing and Sun (2005) in estimation to adapt to the unknown s. Third, consistency is only a basic requirement for a test; a more challenging problem is to find the minimax optimal test in our testing environment. Relevant literature in the classical specification testing includes Ingster (1993) and Guerre and Lavergne (2002). Fourth, as mentioned in the introduction of Oudshoorn (1998), the limit experiment of the model we considered should be the ideal Gaussian white noise model with discontinuities. A rigorous development of this result is theoretically intriguing. The counterpart in the nonparametric regression without discontinuities can be found in Brown and Low (1996). Fifth, we test whether there is selection after excluding the possibility of nonzero treatment effects. If we want to test the presence of selection regardless of the presence of treatment effects, we must exclude the influence of treatment effects (if they are present) on the test statistic. Construction of such a test is challenging. Sixth, our asymptotic arguments in both testing and estimation are based on a fixed sequence of bandwidths; a rigorous analysis on the asymptotic behaviors based on a data-driven bandwidth such as the CV bandwidth in Section 5.4 is preferable.

Seventh, more simulation studies should be conducted to provide more practical suggestions on the bandwidth selection in both specification testing and estimation. Finally, our tests and estimation concentrate on the case where no covariates exist. When y is affected by other covariates and the covariate space is partitioned by a discontinuity frontier, the procedures in this paper can be applied with some technical complication; related discussions on this topic can be found in Yu (2014c) and Yu and Phillips (2014).

Appendix A. Supplementary material

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jeconom.2015.06.002.

References

Andrews, D.W.K., 1993. Tests for parameter instability and structural change with unknown change point. Econometrica 61, 821–856.

Andrews, D.W.K., Ploberger, W., 1994. Optimal tests when a nuisance parameter is present only under an alternative. Econometrica 62, 1383–1414.

Bertanha, M., 2014. Regression Discontinuity Design with Many Thresholds. Mimeo, Stanford.

Bierens, H.J., 1982. Consistent model specification tests. J. Econometrics 20, 105–134.

Braun, J.V., Müller, H.G., 1998. Statistical methods for DNA sequence segmentation. Statist. Sci. 13, 142–162.

Brown, L.D., Low, M.G., 1996. Asymptotic equivalence of nonparametric regression and white noise. Ann. Statist. 24, 2384–2398.

Card, D., Mas, A., Rothstein, J., 2008. Tipping and the dynamics of segregation. Q. J. Econ. 123, 177–218.

Chen, B., Hong, Y., 2012. Testing for smooth structural changes in time series models via nonparametric regression. Econometrica 80, 1157–1183.

Cheng, M.Y., 1994. On Boundary Effects of Smooth Curve Estimators (dissertation), Unpublished Manuscript Series # 2319. Institute for Statistics, University of North Carolina.

Cheng, M.Y., Fan, J., Marron, J.S., 1997. On automatic boundary corrections. Annals of Statistics 25, 1691–1708.

Delgado, M.A., Hidalgo, J., 2000. Nonparametric inference on structural breaks. Journal of Econometrics 96, 113–144.

Delgado, M.A., Manteiga, W.G., 2001. Significance testing in nonparametric regression based on the bootstrap. Annals of Statistics 29, 1469–1507.

DesJardins, S.L., McCall, B.P., 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, Department of Economics, University of Michigan.

Eubank, R.L., Speckman, P.L., 1993. Confidence bands in nonparametric regression. Annals of Statistics 88, 1287–1301.

Fan, J., 1992. Design-adaptive nonparametric regression. J. Amer. Statist. Assoc. 87, 998–1004.

Fan, J., 1993. Local linear regression smoothers and their minimax efficiency. Annals of Statistics 21, 196–216.

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

Fan, J., Gijbels, I., 1996. Local Polynomial Modelling and its Applications. Chapman & Hall, London.

Fan, Y., Li, Q., 1996. Consistent model specification tests: omitted variables and semiparametric functional forms. Econometrica 64, 865–890.

Fan, Y., Li, Q., 2000. Consistent model specification tests: Kernel-based tests versus Bierens' ICM tests. Econometric Theory 16, 1016–1041.

Fan, Y., Linton, O., 2003. Some higher-order theory for a consistent non-parametric model specification test. J. Statist. Plann. Inference 109, 125–154.

Gao, J., Gijbels, I., Van Bellegem, S., 2008. Nonparametric simultaneous testing for structural breaks. Journal of Econometrics 143, 123–142.

Gasser, T., Müller, H.G., 1979. Kernel estimation of regression functions. In: Smoothing Techniques for Curve Estimation. In: Lecture Notes in Mathematics, vol. 757. Springer-Verlag, New York, pp. 23–68.

Gijbels, I., Goderniaux, A., 2004. Bandwidth selection for changepoint estimation in nonparametric regression. Technometrics 46, 76–86.

Gijbels, I., Hall, P., Kneip, A., 1999. On the estimation of jump points in smooth curves. Ann. Inst. Statist. Math. 51, 231–251.

Gnedenko, B.V., 1943. Sur la distribution limité du terme d' une série aléatoire. Ann. of Math. 44, 423–453. Translated and reprinted in S. Kotz & N.L. Johnson, eds., 1992, Breakthroughs in statistics, vol I, Berlin: Springer-Verlag, pp. 195–225.

Gréoire, G., Hamrouni, Z., 2002. Change point estimation by local linear smoothing. J. Multivariate Anal. 83, 56–83.

Gu, J., Li, D., Liu, D., 2007. Bootstrap non-parametric significance test. Nonparametric Statistics 19, 215–230.

Guerre, E., Lavergne, P., 2002. Optimal minimax rates for nonparametric specification testing in regression models. Econometric Theory 18, 1139–1171.

Hahn, J., Todd, P., Van der Klaauw, W., 2001. Identification and estimation of treatment effects with a regression-discontinuity design. Econometrica 69, 201–209.

Hall, P., Wehrly, T.E., 1991. A geometrical method for removing edge effects from kernel-type nonparametric regression estimators. J. Amer. Statist. Assoc. 86, 665–672.

Hamrouni, Z., 1999, Inference statistique par lissage linéaire local pour une fonction de régression présentant des discontinuité, Unpublished Thesis, Laboratoire LMC/IMAG, Université Joseph Fourier.

Härdle, W., Mammen, E., 1993. Comparing nonparametric versus parametric regression fits. Annals of Statistics 21, 1926–1947.

Heckman, J.J., Vytlacil, E.J., 2007a. Econometric evaluation of social programs, Part I: causal models, structural models and econometric policy evaluation, In: Heckman, J., Leamer, E. (Eds.), Handbook of Econometrics, vol. 6B. Elsevier Science B.V, pp. 4779–4874. Ch. 70.

Heckman, J.J., Vytlacil, E.J., 2007b. Econometric evaluation of social programs, Part II: using the marginal treatment effect to organize alternative economic estimators to evaluate social programs and to forecast their effects in new environments, In: Heckman, J., Leamer, E. (Eds.), Handbook of Econometrics, vol. 6B. Elsevier Science B.V, pp. 4875–5143. Ch. 71.

Horowitz, J.L., Spokoiny, V.G., 2001. An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative. Econometrica 69, 4875–5143.

Hsiao, C., Li, Q., Racine, J.S., 2007. A consistent model specification test with mixed discrete and continuous data. Journal of Econometrics 140, 802–826.

Imbens, G.W., Angrist, J.D., 1994. Identification and estimation of local average treatment effects. Econometrica 62, 467–475.

Imbens, G.W., Lemieux, T., 2008. Regression discontinuity designs: a guide to practice. Journal of Econometrics 142, 615–635.

Imbens, G.W., Kalyanaraman, K., 2012. Optimal bandwidth choice for the regression discontinuity estimator. Rev. Econom. Stud. 79, 933–959.

Ingster, Y.I., 1993. Asymptotically minimax hypothesis testing for nonparametric alternatives, I, II, and III. Math. Methods Statist. 2, 85–114, 171–189, and 249–268.

Korostelev, A., 1987. On minimax estimation of a discontinuous signal. Theory Probab. Appl. 32, 727–730.

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

Lee, D.S., Lemieux, T., 2010. Regression discontinuity designs in economics. Journal of Economic Literature 48, 281–355.

Li, Q., Wang, S., 1998. A simple consistent bootstrap test for a parametric regression function. Journal of Econometrics 87, 145–165.

Liu, R.Y., 1988. Bootstrap procedures under some non i.i.d. models. Annals of Statistics 16, 1696–1708.

Loader, C., 1996. Change point estimation using nonparametric regression. Annals of Statistics 24, 1667–1678.

Loader, C., 1999. Bandwidth selection: classical or plug-in? Annals of Statistics 27, 415–438.

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

McCrary, J., 2008. Testing for manipulation of the running variable in the regression discontinuity design. Journal of Econometrics 142, 698–714.

Müller, H.G., 1992. Change-points in nonparametric regression analysis. Annals of Statistics 20, 737–761.

Müller, H.G., Stadtmüller, U., 1999. Discontinuous versus smooth regression. Annals of Statistics 27, 299–337.

Oudshoorn, C.G.M., 1998. Asymptotically minimax estimation of a function with jumps. Bernoulli 4, 15–33.

Porter, J., 2003. Estimation in the Regression Discontinuity Model. Mimeo, Department of Economics, University of Wisconsin at Madison.

Qiu, P., 1994. Estimation of the number of jumps of the jump regression functions. Comm. Statist.-Theory Methods 23, 2141–2155.

Qiu, P., Asano, C., Li, X., 1991. Estimation of jump regression function. Bull. Inform. Cybernet. 24, 197–212.

Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman and Hall, New York.

Spokoiny, V.G., 1998. Estimation of a function with discontinuities via local polynomial fit with an adaptive window choice. Annals of Statistics 26, 1356–1378.

Stone, C.J., 1982. Optimal global rates of convergence for nonparametric regression. Annals of Statistics 10, 1040–1053.

Stute, W., Manteiga, W.G., Quindimil, M.P., 1998. Bootstrap approximation in model checks for regression. J. Amer. Statist. Assoc. 93, 141–149.

Su, L., Xiao, Z., 2008. Testing structural change in time-series nonparametric regression models. Stat. Interface 347–366.

Sun, Y., 2005. Adaptive Estimation of the Regression Discontinuity Model. Mimeo, Department of Economics, University of California at San Diego.

Thistlethwaite, D., Campbell, D., 1960. Regression-discontinuity analysis: an alternative to the ex-post facto experiment. J. Educ. Psychol. 51, 309–317.

van der Klaauw, W., 2002. Estimating the effect of financial aid offers on college enrollment: a regression-discontinuity approach. Internat. Econom. Rev. 43, 1249–1287.

van der Klaauw, W., 2008. Regression-discontinuity analysis: a survey of recent development in economics. Labour 22, 219–245.

van der Vaart, A.W., 1998. Asymptotic Statistics. Cambridge University Press, New York.

Wang, Y., 1995. Jump and sharp cusp detection by wavelets. Biometrika 82, 385–397.

Wu, C.F.J., 1986. Jackknife, bootstrap and other resampling methods in regression analysis. Annals of Statistics 14, 1261–1295.
Wu, C.K., Chu, J.S., 1993a. Kernel-type estimators of jump points and values of a regression function. Annals of Statistics 21, 1545–1566.
Wu, C.K., Chu, J.S., 1993b. Nonparametric function estimation and bandwidth selection for discontinuous regression functions. Statistica Sinica 3, 557–576.
Wu, C.K., Lai, Y.T., 1998. Note on changing-point testing for nonparametric regression function. Tamsui Oxf. J. Math. Sci. 14, 1–10.
Yin, Y.Q., 1988. Detection of the number, locations and magnitudes of jumps. Stoch. Models 4, 445–455.
Yu, P., 2008. Adaptive estimation of the threshold point in threshold regression, forthcoming. Journal of Econometrics.
Yu, P., 2010. Integrated Quantile Threshold Regression and Distributional Threshold Effects. Mimeo, HKU.
Yu, P., 2012. Likelihood estimation and inference in threshold regression. Journal of Econometrics 167, 274–294.
Yu, P., 2013. Understanding estimators of treatment effects in regression discontinuity designs, forthcoming. Econometric Rev..
Yu, P., 2014a. The bootstrap in threshold regression. Econometric Theory 30, 676–714.
Yu, P., 2014b. Marginal Quantile Treatment Effect. Mimeo, HKU.
Yu, P., 2014c. Threshold Regression With A Threshold Boundary. Mimeo, HKU.
Yu, P., Phillips, P.C.B., 2014. Threshold Regression with Endogeneity. Mimeo, HKU.
Yu, P., Zhao, Y., 2013. Asymptotics for threshold regression under general conditions. Econom. J. 16, 430–462.
Zheng, J.X., 1996. A consistent test of functional form via nonparametric estimation techniques. Journal of Econometrics 75, 263–289.