

Approximate Permutation Tests and Induced Order Statistics in the Regression Discontinuity Design *

Ivan A. Canay
Department of Economics
Northwestern University
iacanay@northwestern.edu

Vishal Kamat
Department of Economics
Northwestern University
v.kamat@u.northwestern.edu

October 12, 2017

Abstract

In the regression discontinuity design (RDD), it is common practice to assess the credibility of the design by testing whether the means of baseline covariates do not change at the cutoff (or threshold) of the running variable. This practice is partly motivated by the *stronger* implication derived by Lee (2008), who showed that under certain conditions the distribution of baseline covariates in the RDD must be continuous at the cutoff. We propose a permutation test based on the so-called induced ordered statistics for the null hypothesis of continuity of the distribution of baseline covariates at the cutoff; and introduce a novel asymptotic framework to analyze its properties. The asymptotic framework is intended to approximate a *small* sample phenomenon: even though the total number n of observations may be large, the number of *effective* observations local to the cutoff is often small. Thus, while traditional asymptotics in RDD require a growing number of observations local to the cutoff as $n \rightarrow \infty$, our framework keeps the number q of observations local to the cutoff *fixed* as $n \rightarrow \infty$. The new test is easy to implement, asymptotically valid under weak conditions, exhibits finite sample validity under stronger conditions than those needed for its asymptotic validity, and has favorable power properties relative to tests based on means. In a simulation study, we find that the new test controls size remarkably well across designs. We then use our test to evaluate the plausibility of the design in Lee (2008), a well-known application of the RDD to study incumbency advantage.

KEYWORDS: Regression discontinuity design, permutation tests, randomization tests, induced ordered statistics, rank tests.

JEL classification codes: C12, C14.

*We thank the Co-Editor and four anonymous referees for helpful comments. We also thank Azeem Shaikh, Alex Torgovitsky, Magne Mogstad, Matt Notowidigdo, Matias Cattaneo, and Max Tabord-Meehan for valuable suggestions. Finally, we thank Mauricio Olivares-Gonzalez and Ignacio Sarmiento-Barbieri for developing the R package. The research of the first author was supported by the National Science Foundation Grant SES-1530534. First version: CeMMAP working paper CWP27/15.

1 Introduction

The regression discontinuity design (RDD) has been extensively used in recent years to retrieve causal treatment effects - see [Lee and Lemieux \(2010\)](#) and [Imbens and Lemieux \(2008\)](#) for exhaustive surveys. The design is distinguished by its unique treatment assignment rule. Here individuals receive treatment when an observed covariate (commonly referred to as the running variable) crosses a known cutoff or threshold, and receive control otherwise. [Hahn et al. \(2001\)](#) illustrates that such an assignment rule allows nonparametric identification of the average treatment effect (ATE) at the cutoff, provided that potential outcomes have continuous conditional expectations at the cutoff. The credibility of this identification strategy along with the abundance of such discontinuous rules in practice have made the RDD increasingly popular in empirical applications.

The continuity assumption that is necessary for nonparametric identification of the ATE at the cutoff is fundamentally untestable. Empirical studies however assess the plausibility of their RDD by exploiting two testable implications of a stronger identification assumption proposed by [Lee \(2008\)](#). We can describe the two implications as follows: (i) individuals have imprecise control over the running variable, which translates into the density of the running variable being continuous at the cutoff; and (ii) the treatment is locally randomized at the cutoff, which translates into the distribution of all observed baseline covariates being continuous at the cutoff. The second prediction is particularly intuitive and, quite importantly, analogous to the type of restrictions researchers often inspect or test in a fully randomized controlled experiment. The practice of judging the reliability of RDD applications by assessing either of the two above stated implications (commonly referred to as manipulation, falsification or placebo tests) is ubiquitous in the empirical literature.¹ However, in regards to the second testable implication, researchers often verify continuity of the means of baseline covariates at the cutoff, which is a weaker requirement than Lee’s implication.

This paper proposes a novel permutation test for the null hypothesis on the second testable implication, i.e., the distribution of baseline covariates is continuous at the cutoff.² The new test has a number of attractive properties. First, our test controls the limiting null rejection probability under fairly mild conditions, and delivers finite sample validity under stronger, but yet plausible, conditions. Second, our test is more powerful against some alternatives than those aimed at testing the continuity of the *means* of baseline covariates at the cutoff, which appears to be a dominant practice in the empirical literature. Third, our test is arguably simple to implement as it only involves computing order statistics and empirical cdfs with a fixed number of observations

¹Table 5 surveys RDD empirical papers in four leading applied economic journals during the period 2011-2015, see Appendix E for further details. Out of 62 papers, 43 of them include some form of manipulation, falsification or placebo test. In fact, the most popular practice involves evaluating the continuity of the *means* of baseline covariates at the cutoff (42 papers).

²It is important to emphasize that the null hypothesis we test in this paper is neither necessary nor sufficient for identification of the ATE at the cutoff. See Section 2 for a discussion on this.

closest to the cutoff. This contrasts with the few existing alternatives that require local polynomial estimation and undersmoothed bandwidth choices. Finally, we have developed companion **Stata** and **R** packages to facilitate the adoption of our test.³

The construction of our test is based on the simple intuition that observations close to the cutoff are *approximately* (but not exactly) identically distributed to either side of it when the null hypothesis holds. This allows us to permute these observations to construct an *approximately* valid test. In other words, the formal justification for the validity of our test is asymptotic in nature and recognizes that traditional arguments advocating the use of permutation tests are not necessarily valid under the null hypothesis of interest; see Section 3.2 for a discussion on this distinction. The novel asymptotic framework we propose aims at capturing a *small* sample problem: the number of observations close to the cutoff is often small even if the total sample size is large. More precisely, our asymptotic framework is one in which the number of observations q that the test statistic contains from either side of the cutoff is *fixed* as the total sample size n goes to infinity. Formally, we exploit the recent asymptotic framework developed by [Canay, Romano and Shaikh \(2017\)](#) for randomization tests, although we introduce novel modifications to accommodate the RDD setting. Further, in an important intermediate stage, we use induced order statistics, see [Bhattacharya \(1974\)](#) and (8), to frame our problem and develop some insightful results of independent interest in Theorem 4.1.

An important contribution of this paper is to show that permutation tests can be justified in RDD settings through a novel asymptotic framework that aims at embedding a small sample problem. The asymptotic results are what primarily separates this paper from others in the RDD literature that have advocated for the use of permutation tests (see, e.g., [Cattaneo et al., 2015](#); [Sales and Hansen, 2015](#); [Ganong and Jäger, 2015](#)). In particular, all previous papers have noticed that permutation tests become appropriate for testing null hypotheses under which there is a *neighborhood* around the cutoff where the RDD can be viewed as a randomized experiment. This, however, deviates from traditional RDD arguments that require such local randomization to hold only *at the cutoff*. Therefore, as explained further in Section 3.2, this paper is the first to develop and to provide formal results that justify the use of permutation tests asymptotically for these latter null hypotheses. Another contribution of this paper is to exploit the testable implication derived by [Lee \(2008\)](#), which is precisely a statement on the distribution of baseline covariates, and note that permutation tests arise as natural candidates to consider. Previous papers have focused attention on hypotheses about distributional treatment effects, which deviates from the predominant interest in ATEs, and do not directly address the testing problem we consider in this paper.

The remainder of the paper is organized as follows. Section 2 introduces the notation and discusses the hypothesis of interest. Section 3 introduces our permutation test based on a fixed

³The **Stata** package `rdperm` and the **R** package `RATest` can be downloaded from <http://sites.northwestern.edu/iac879/software/>.

number of observations closest to the cutoff, discusses all aspects related to its implementation in practice, and compares it with permutation tests previously proposed in the RDD setting. Section 4 contains all formal results, including the description of the asymptotic framework, the assumptions, and the main theorems. Section 5 studies the finite sample properties of our test via Monte Carlo simulations. Finally, Section 6 implements our test to reevaluate the validity of the design in Lee (2008), a familiar application of the RDD to study incumbency advantage. All proofs are in the Appendix.

2 Testable implications of local randomization

Let $Y \in \mathbf{R}$ denote the (observed) outcome of interest for an individual or unit in the population, $A \in \{0, 1\}$ denote an indicator for whether the unit is treated or not, and $W \in \mathcal{W}$ denote observed baseline covariates. Further denote by $Y(1)$ the potential outcome of the unit if treated and by $Y(0)$ the potential outcome if not treated. As usual, the (observed) outcome and potential outcomes are related to treatment assignment by the relationship

$$Y = Y(1)A + Y(0)(1 - A) . \tag{1}$$

The treatment assignment in the (sharp) Regression Discontinuity Design (RDD) follows a discontinuous rule,

$$A = I\{Z \geq \bar{z}\} ,$$

where $Z \in \mathcal{Z}$ is an observed scalar random variable known as the running variable and \bar{z} is the known threshold or cutoff value. For convenience, we normalize $\bar{z} = 0$. This treatment assignment rule allows us to identify the average treatment effect (ATE) at the cutoff; i.e.,

$$E[Y(1) - Y(0)|Z = 0] .$$

In particular, Hahn et al. (2001) establish that identification of the ATE at the cutoff relies on the discontinuous treatment assignment rule and the assumption that

$$E[Y(1)|Z = z] \quad \text{and} \quad E[Y(0)|Z = z] \quad \text{are both continuous in } z \text{ at } z = 0 . \tag{2}$$

Reliability of the RDD thus depends on whether the mean outcome for units marginally below the cutoff identifies the true counterfactual for those marginally above the cutoff.

Despite the continuity assumption appearing weak, Lee (2008) states two practical limitations for empirical researchers. First, it is difficult to determine whether the assumption is plausible as it is not a description of a treatment-assigning process. Second, the assumption is fundamentally untestable. Motivated by these limitations, Lee (2008, Condition 2b) considers an alternative (and arguably stronger) sufficient condition for identification. The new condition is intuitive and leads

to clean testable implications that are easy to assess in an applied setting. In RDD empirical studies, these implications are often presented (with different levels of formality) as falsification, manipulation, or placebo tests (see Table 5 for a survey).

In order to describe Lee’s alternative condition, let U be a scalar random variable capturing the unobserved type or heterogeneity of a unit in the population. Assume there exist measurable functions $m_0(\cdot)$, $m_1(\cdot)$, and $m_w(\cdot)$, such that

$$Y(1) = m_1(U), \quad Y(0) = m_0(U), \quad \text{and} \quad W = m_w(U) .$$

Condition 2b in Lee (2008) can be stated in our notation as follows.

Assumption 2.1. *The cdf of Z conditional on U , $F(z|u)$, is such that $0 < F(0|u) < 1$, and is continuously differentiable in z at $z = 0$ for each u in the support of U . The marginal density of Z , $f(z)$, satisfies $f(0) > 0$.*

This assumption has a clear behavioral interpretation - see Lee (2008) and Lee and Lemieux (2010) for a lengthy discussion of this assumption and its implications. It allows units to have control over the running variable, as the distribution of Z may depend on U in flexible ways. Yet, the condition $0 < F(0|u) < 1$ and the continuity of the conditional density ensure that such control may not be fully *precise* - i.e., it rules out *deterministic* sorting around the cutoff. For example, if for some u' we had $\Pr\{Z < 0|u'\} = 0$, then units with $U = u'$ would be all on one side of the cutoff and deterministic sorting would be possible - see Lee and Lemieux (2010) for concrete examples.

Lee (2008, Proposition 2) shows that Assumption 2.1 implies the continuity condition in (2) is sufficient for identification of the ATE at the cutoff, and further implies that

$$H(w|z) \equiv \Pr\{W \leq w|Z = z\} \text{ is continuous in } z \text{ at } z = 0 \text{ for all } w \in \mathcal{W} . \quad (3)$$

In other words, the behavioral assumption that units do not precisely control Z around the cutoff implies that the treatment assignment is locally randomized *at the cutoff*, which means that the (conditional) distribution of baseline covariates should not change discontinuously at the cutoff.

In this paper we propose a test for this null hypothesis of continuity in the distribution of the baseline covariates W at the cutoff $Z = 0$, i.e. (3). To better describe our test, it is convenient to define two auxiliary distributions that capture the local behavior of W to either side of the cutoff. To this end, define

$$H^-(w|0) = \lim_{z \uparrow 0} H(w|z) \quad \text{and} \quad H^+(w|0) = \lim_{z \downarrow 0} H(w|z) . \quad (4)$$

Using this notation, the continuity condition in (3) is equivalent to the requirement that $H(w|z)$ is right continuous at $z = 0$ and that

$$H^-(w|0) = H^+(w|0) \text{ for all } w \in \mathcal{W} . \quad (5)$$

The advantage of the representation in (5) is that it facilitates the comparison between two sample testing problems and the one we consider here. It also facilitates the comparison between our approach and alternative ones advocating the use of permutation tests on the grounds of favorable finite sample properties, see Section 3.2.

Remark 2.1. In randomized controlled experiments where the treatment assignment is exogenous by design, the empirical analysis usually begins with an assessment of the comparability of treated and control groups in baseline covariates, see Bruhn and McKenzie (2008). This practice partly responds to the concern that, if covariates differ across the two groups, the effect of the treatment may be confounded with the effect of the covariates - casting doubts on the validity of the experiment. The local randomization nature in RDD leads to the analogous (local) implication in (5).

■

Remark 2.2. Assumption 2.1 requires continuity of the conditional density of Z given U at $z = 0$, which implies continuity of the marginal density of Z , $f(z)$, at $z = 0$. McCrary (2008) exploits this testable implication and proposes a test for the null hypothesis of continuity of $f(z)$ at the cutoff. Our test exploits a different implication of Assumption 2.1 and therefore should be viewed as a complement, rather than a substitute, to the density test proposed by McCrary (2008). ■

Remark 2.3. Gerard et al. (2016) study the consequences of discontinuities in the density of Z at the cutoff. In particular, the authors consider a situation in which manipulation occurs only in one direction for a subset of the population (i.e., there exists a subset of participants such that $Z \geq 0$ a.s.) and use the magnitude of the discontinuity of $f(z)$ at $z = 0$ to identify the proportion of always-assigned units among all units close to the cutoff. Using this setup, Gerard et al. (2016) show that treatment effects in RDD are not point identified but that the model still implies informative bounds (i.e., treatment effects are partially identified). ■

A common practice in applied research is to test the hypothesis

$$E[W|Z = z] \text{ is continuous in } z \text{ at } z = 0, \tag{6}$$

which is an implication of the null in (3). Table 5 in Appendix E shows that out of 62 papers published in leading journals during the period 2011-2015, 42 of them include a formal (or informal via some form of graphical inspection) test for the null in (6). However, if the fundamental hypothesis of interest is the implication derived by Lee (2008), testing the hypothesis in (6) has important limitations. First, tests designed for (6) have low power against certain distributions violating (3). Indeed, these tests may incorrectly lead the researcher to believe that baseline covariates are “continuous” at the cutoff, when some features of the distribution of W (other than the mean) may be discontinuous. Second, tests designed for (6) may exhibit poor size control in cases where usual smoothness conditions required for local polynomial estimation do not hold. Section 5 illustrates both of these points.

Before moving to describe the test we propose in this paper, we emphasize two aspects about Assumption 2.1 and the testable implication in (3). First, Assumption 2.1 is sufficient but not necessary for identification of the ATE at the cutoff. Second, the continuity condition in (3) is neither necessary nor sufficient for identification of the ATE at the cutoff. Assessing whether (3) holds or not is simply a sensible way to argue in favor or against the credibility of the design.

3 A permutation test based on induced ordered statistics

Let P be the distribution of (Y, W, Z) and $X^{(n)} = \{(Y_i, W_i, Z_i) : 1 \leq i \leq n\}$ be a random sample of n i.i.d. observations from P . Let q be a small (relative to n) positive integer. The test we propose is based on $2q$ values of $\{W_i : 1 \leq i \leq n\}$, such that q of these are associated with the q closest values of $\{Z_i : 1 \leq i \leq n\}$ to the right of the cutoff $\bar{z} = 0$, and the remaining q are associated with the q closest values of $\{Z_i : 1 \leq i \leq n\}$ to the left of the cutoff $\bar{z} = 0$. To be precise, denote by

$$Z_{n,(1)} \leq Z_{n,(2)} \leq \dots \leq Z_{n,(n)} \quad (7)$$

the order statistics of the sample $\{Z_i : 1 \leq i \leq n\}$ and by

$$W_{n,[1]}, W_{n,[2]}, \dots, W_{n,[n]} \quad (8)$$

the corresponding values of the sample $\{W_i : 1 \leq i \leq n\}$, i.e., $W_{n,[j]} = W_k$ if $Z_{n,(j)} = Z_k$ for $k = 1, \dots, n$. The random variables in (8) are called *induced order statistics* or *concomitants* of order statistics, see [David and Galambos \(1974\)](#); [Bhattacharya \(1974\)](#).

In order to construct our test statistic, we first take the q closest values in (7) to the right of the cutoff and the q closest values in (7) to the left of the cutoff. We denote these ordered values by

$$Z_{n,(q)}^- \leq \dots \leq Z_{n,(1)}^- < 0 \text{ and } 0 \leq Z_{n,(1)}^+ \leq \dots \leq Z_{n,(q)}^+ , \quad (9)$$

respectively, and the corresponding induced values in (8) by

$$W_{n,[q]}^-, \dots, W_{n,[1]}^- \text{ and } W_{n,[1]}^+, \dots, W_{n,[q]}^+ . \quad (10)$$

Note that while the values in (9) are ordered, those in (10) are not necessarily ordered.

The random variables $(W_{n,[1]}^-, \dots, W_{n,[q]}^-)$ are viewed as an independent sample of W conditional on Z being “close” to zero from the left, while the random variables $(W_{n,[1]}^+, \dots, W_{n,[q]}^+)$ are viewed as an independent sample of W conditional on Z being “close” to zero from the right. We therefore use each of these two samples to compute empirical cdfs as follows,

$$\hat{H}_n^-(w) = \frac{1}{q} \sum_{j=1}^q I\{W_{n,[j]}^- \leq w\} \text{ and } \hat{H}_n^+(w) = \frac{1}{q} \sum_{j=1}^q I\{W_{n,[j]}^+ \leq w\} .$$

Finally, letting

$$S_n = (S_{n,1}, \dots, S_{n,2q}) = (W_{n,[1]}^-, \dots, W_{n,[q]}^-, W_{n,[1]}^+, \dots, W_{n,[q]}^+) , \quad (11)$$

denote the pooled sample of induced order statistics, we can define our test statistic as

$$T(S_n) = \frac{1}{2q} \sum_{j=1}^{2q} (\hat{H}_n^-(S_{n,j}) - \hat{H}_n^+(S_{n,j}))^2 . \quad (12)$$

The statistic $T(S_n)$ in (12) is a Cramér Von Mises test statistic, see [Hajek et al. \(1999, p. 101\)](#).

We propose to compute the critical values of our test by a permutation test as follows. Let \mathbf{G} denote the set of all permutations $\pi = (\pi(1), \dots, \pi(2q))$ of $\{1, \dots, 2q\}$. We refer to \mathbf{G} as the group of permutations (in this context, “group” is understood as a mathematical group). Let

$$S_n^\pi = (S_{n,\pi(1)}, \dots, S_{n,\pi(2q)}) ,$$

be the permuted values of S_n in (11) according to π . Let $M = |\mathbf{G}|$ be the cardinality of \mathbf{G} and denote by

$$T^{(1)}(S_n) \leq T^{(2)}(S_n) \leq \dots \leq T^{(M)}(S_n)$$

the ordered values of $\{T(S_n^\pi) : \pi \in \mathbf{G}\}$. For $\alpha \in (0, 1)$, let $k = \lceil M(1 - \alpha) \rceil$ and define

$$\begin{aligned} M^+(S_n) &= |\{1 \leq j \leq M : T^{(j)}(S_n) > T^{(k)}(S_n)\}| \\ M^0(S_n) &= |\{1 \leq j \leq M : T^{(j)}(S_n) = T^{(k)}(S_n)\}| , \end{aligned} \quad (13)$$

where $\lceil x \rceil$ is the smallest integer greater than or equal to x . The test we propose is given by

$$\phi(S_n) = \begin{cases} 1 & T(S_n) > T^{(k)}(S_n) \\ a(S_n) & T(S_n) = T^{(k)}(S_n) \\ 0 & T(S_n) < T^{(k)}(S_n) \end{cases} , \quad (14)$$

where

$$a(S_n) = \frac{M\alpha - M^+(S_n)}{M^0(S_n)} .$$

Remark 3.1. The test in (14) is possibly randomized. The non-randomized version of the test that rejects when $T(S_n) > T^{(k)}(S_n)$ is also asymptotically level α by [Theorem 4.2](#). In our simulations, the randomized and non-randomized versions perform similarly when M is not too small. ■

Remark 3.2. When M is too large the researcher may use a stochastic approximation to $\phi(S_n)$ without affecting the properties of our test. More formally, let

$$\hat{\mathbf{G}} = \{\pi_1, \dots, \pi_B\} ,$$

where $\pi_1 = (1, \dots, 2q)$ is the identity permutation and π_2, \dots, π_B are i.i.d. Uniform(\mathbf{G}). [Theorem 4.2](#) in [Section 4](#) remains true if, in the construction of $\phi(S_n)$, \mathbf{G} is replaced by $\hat{\mathbf{G}}$. ■

Remark 3.3. Our results are not restricted to the Cramér Von Mises test statistic in (12) and apply to other rank statistics satisfying our assumptions in [Section 4](#), e.g., the Kolmogorov-Smirnov statistics. We restrict our discussion to the statistic in (12) for simplicity of exposition. ■

3.1 Implementing the new test

In this section we discuss the practical considerations involved in the implementation of our test, highlighting how we addressed these considerations in the companion **Stata** and **R** packages.

The only tuning parameter of our test is the number q of observations closest to the cutoff. The asymptotic framework in Section 4 is one where q is fixed as $n \rightarrow \infty$, so this number should be small relative to the sample size. In this paper we use the following rule of thumb,

$$q_{\text{rot}} = \left\lceil f(0)\sigma_Z\sqrt{1-\rho^2}\frac{n^{0.9}}{\log n} \right\rceil, \quad (15)$$

where $f(0)$ is the density of Z at zero, ρ is the correlation between W and Z , and σ_Z^2 is the variance of Z . The motivation for this rule of thumb is as follows. First, the rate $\frac{n^{0.9}}{\log n}$ arises from the proof of Theorem 4.1, which suggests that q may increase with n as long as $n - q \rightarrow \infty$ and $q \log n / (n - q) \rightarrow 0$. Second, the constant arises by considering the special case where (W, Z) are bivariate normal. In such a case, it follows that

$$\left. \frac{\partial \Pr\{W \leq w | Z = z\}}{\partial z} \right|_{z=0} \propto \frac{-1}{\sigma_Z\sqrt{1-\rho^2}} \text{ at } w = E[W|z = 0].$$

Intuitively, one would like q to adapt to the slope of this conditional cdf. When the derivative is close to zero, a large q would be desired as in this case $H(w|0)$ and $H(w|z)$ should be similar for small values of $|z|$. When the derivative is high, a small value of q is desired as in this case $H(w|z)$ could be different than $H(w|0)$ even for small values of $|z|$. Our rule of thumb is thus inversely proportional to this derivative to capture this intuition. Finally, we scale the entire expression by the density of Z at the cutoff, $f(0)$, which accounts for the potential number of observations around the cutoff and makes q_{rot} scale invariant when (W, Z) are bivariate normal. All these quantities can be estimated to deliver a feasible \hat{q}_{rot} .⁴

Given q , the implementation of our test proceeds in the following six steps.

Step 1. Compute the order statistics of $\{Z_i : 1 \leq i \leq n\}$ at either side of the cutoff as in (9).

Step 2. Compute the associated values of $\{W_i : 1 \leq i \leq n\}$ as in (10).

Step 3. Compute the test statistic in (12) using the observations from Step 2.

Step 4. Generate random permutations $\hat{\mathbf{G}} = \{\pi_1, \dots, \pi_B\}$ as in Remark 3.2 for a given B .

Step 5. Evaluate the test statistic in (12) for each permuted sample: $T(S_n^{\pi_\ell})$ for $\ell \in \{1, \dots, B\}$.

⁴We have also considered the alternative rule of thumb $q_{\text{rot}} = \left\lceil f(0)\sigma_Z\sqrt{10(1-\rho^2)}\frac{n^{3/4}}{\log n} \right\rceil$ in the simulations of Section 5 and found similar results to those reported there. This alternative rule of thumb grows at a slower rate but has a larger constant in front of the rate.

Step 6. Compute the p -value of the test as follows,

$$p_{\text{value}} = \frac{1}{B} \sum_{\ell=1}^B I\{T(S_n^{\pi_\ell}) \geq T(S_n)\} . \quad (16)$$

Note that p_{value} is the p -value associated with the non-randomized version of the test, see Remark 3.1. The default values in the **Stata** package, and the values we use in the simulations in Section 5, are $B = 999$ and $q = \hat{q}_{\text{rot}}$, as described in Appendix D.

Remark 3.4. The recommended choice of q in (15) is simply a sensible rule of thumb and is not an optimal rule in any formal sense. Given our asymptotic framework where q is fixed as n goes to infinity, it is difficult, and out of the scope of this paper, to derive optimal rules for choosing q . ■

Remark 3.5. The number of observations q on either side of the cutoff need not be symmetric. All our results go through with two fixed values, q_l and q_r , to the left and right of the cutoff respectively. However, we restrict our attention to the case where q is the same on both sides as it simplifies deriving a rule of thumb for q and makes the overall exposition cleaner. ■

3.2 Relation to other permutation tests in the literature

Permutation tests have been previously discussed in the RDD literature for doing inference on distributional treatment effects. In particular, Cattaneo et al. (2015, CFT) provide conditions in a randomization inference context under which the RDD can be interpreted as a local randomized controlled experiment (RCE) and develop exact finite-sample inference procedures based on such an interpretation. Ganong and Jäger (2015) and Sales and Hansen (2015) build on the same framework and consider related tests for the kink design and projected outcomes, respectively.

The most important distinction with our paper is that permutation tests have been previously advocated on the grounds of finite sample validity. Such a justification requires, essentially, a different type of null hypothesis than the one we consider. In particular, suppose it was the case that for some $b > 0$, $H(w|z) = \Pr\{W \leq w|Z = z\}$ was constant in z for all $z \in [-b, b]$ and $w \in \mathcal{W}$. In other words, suppose the treatment assignment is locally randomized in a *neighborhood* of zero as opposed to “at zero”. The null hypothesis in this case could be written as

$$H(w|z \in [-b, 0)) = H(w|z \in [0, b]) \text{ for all } w \in \mathcal{W} . \quad (17)$$

Under the null hypothesis in (17), a permutation test applied to the sample with observations $\{(W_i, Z_i) : -b \leq Z_i < 0, 1 \leq i \leq n\}$ and $\{(W_i, Z_i) : 0 \leq Z_i \leq b, 1 \leq i \leq n\}$, leads to a test that is valid in finite samples (i.e., its finite sample size does not exceed the nominal level). The proof of this result follows from standard arguments (see Lehmann and Romano, 2005, Theorem 15.2.1). For these arguments to go through, the null hypothesis must be the one in (17) for a

known b . Indeed, CFT clearly state that the key assumption for the validity of their approach is the existence of a *neighborhood* around the cutoff where a randomization-type condition holds. In our notation, this is captured by (17).

Contrary to those arguments, our paper shows that permutation tests can be used for the null hypothesis in (5), which only requires local randomization at zero, and shows that the justification for using permutation tests may be asymptotic in nature (see Remark 4.1 for a technical discussion). The asymptotics are non-standard as they intend to explicitly capture a situation where the number of effective observations (q in our notation) is small relative to the total sample size (n in our notation). This is possible in our context due to the recent asymptotic framework developed by Canay et al. (2017) for randomization tests, although we introduce novel modifications to make it work in the RDD setting - see Section 4.2. Therefore, even though the test we propose in this paper may be “mechanically” equivalent to the one in CFT, the formal arguments that justify their applicability are markedly different (see also the recent paper by Sekhon and Titiunik (2016) for a discussion on local randomization at the cutoff versus in a neighborhood). Importantly, while our test can be viewed as a test for (3), which is the actual implication in Lee (2008, Proposition 2), the test in CFT is a test for (17), which does not follow from Assumption 2.1.

Remark 3.6. The motivation behind the finite sample analysis in Cattaneo et al. (2015) is that only a few observations might be available close enough to the cutoff where a local randomization-type condition holds, and hence standard large-sample procedures may not be appropriate. They go on to say that “...small sample sizes are a common phenomenon in the analysis of RD designs...”, referring to the fact that the number of effective observations typically used for inference (those local to the cutoff) are typically small even if the total number of observations, n , is large. Therefore, the motivation behind their finite sample analysis is precisely the motivation behind our asymptotic framework where, as $n \rightarrow \infty$, the effective number of observations q that enter our test are taken to be fixed. By embedding this finite sample situation into our asymptotic framework, we can construct tests for the hypothesis in (3) as opposed to the one in (17). ■

Remark 3.7. In Remark 2.1 we made a parallel between our testing problem and the standard practice in RCEs of comparing the treated and control groups in baseline covariates. However, the testable implication in RCEs is a global statement about the conditional distribution of W given $A = 1$ and $A = 0$. With large sample sizes, there exists a variety of asymptotically valid tests that are available to test $\Pr\{W \leq w|A = 1\} = \Pr\{W \leq w|A = 0\}$, and permutation tests are one of the many methods that may be used. On the contrary, in RDD the testable implication is “local” in nature, which means that few observations are actually useful for testing the hypothesis in (5). Finite sample issues, and permutation tests in particular, thus become relevant. ■

Another difference between the aforementioned papers and our paper is that their goal is to conduct inference on the (distributional) treatment effect and not on the hypothesis of continuity

of covariates at the cutoff. Indeed, they essentially consider (sharp) hypotheses of the form

$$Y_i(1) = Y_i(0) + \tau_i \text{ for all } i \text{ such that } Z_i \in [-b, b]$$

(for $\tau_i = 0 \forall i$ in the case of no-treatment effect), which deviates from the usual interest on average treatment effects (Ganong and Jäger, 2015, is about the kink design but similar considerations apply). On the contrary, the testable implication in Lee (2008, Proposition 2) is *precisely* a statement about conditional distribution functions (i.e. (3)), so our test is designed by construction for the hypothesis of interest.

Remark 3.8. Sales and Hansen (2015), building on CFT, also use small-sample justifications in favor of permutation tests. However, they additionally exploit the assumption that the researcher can correctly specify a model for variables of interest (outcomes in their paper and covariates in our setting) as a function of the running variable Z . Our results do not require such modeling assumptions and deliver a test for the hypothesis in (3) as opposed to (17). ■

Remark 3.9. Shen and Zhang (2016) also investigate distributional treatment effects in the RDD. In particular, they are interested in testing $\Pr\{Y(0) \leq y|Z = 0\} = \Pr\{Y(1) \leq y|Z = 0\}$, and propose a Kolmogorov-Smirnov-type test statistic based on local linear estimators of distributional treatment effects. Their asymptotic framework is standard and requires $nh \rightarrow \infty$ (where h is a bandwidth), which implies that the effective number of observations at the cutoff increases as the sample size increases. Although not mentioned in their paper, their test could be used to test the hypothesis in (3) whenever W is continuously distributed. We therefore compare the performance our test to the one in Shen and Zhang (2016) in Sections 5 and 6. ■

Remark 3.10. Our test can be used (replacing W with Y) to perform distributional inference on the outcome variable as in CFT and Shen and Zhang (2016). We do not focus on this case here. ■

4 Asymptotic framework and formal results

In this section we derive the asymptotic properties of the test in (14) using an asymptotic framework where q is fixed and $n \rightarrow \infty$. We proceed in two parts. We first derive a result on the asymptotic properties of induced order statistics in (10) that provides an important milestone in proving the asymptotic validity of our test. We then use this intermediate result to prove our main theorem.

4.1 A result on induced order statistics

Consider the order statistics in (7) and the induced order statistics in (8). As in the previous section, denote the q closest values in (7) to the right and left of the cutoff by

$$Z_{n,(q)}^- \leq \dots \leq Z_{n,(1)}^- < 0 \text{ and } 0 \leq Z_{n,(1)}^+ \leq \dots \leq Z_{n,(q)}^+,$$

respectively, and the corresponding induced values in (8) by

$$W_{n,[q]}^-, \dots, W_{n,[1]}^- \text{ and } W_{n,[1]}^+, \dots, W_{n,[q]}^+ .$$

To prove the main result in this section we make the following assumption.

Assumption 4.1. *For any $\epsilon > 0$, Z satisfies $\Pr\{Z \in [-\epsilon, 0)\} > 0$ and $\Pr\{Z \in [0, \epsilon]\} > 0$.*

Assumption 4.1 requires that the distribution of Z is locally dense to the left of zero, and either locally dense to the right of zero or has a mass point at zero, i.e. $\Pr\{Z = 0\} > 0$. Importantly, Z could be continuous with a density $f(z)$ discontinuous at zero; or could have mass points anywhere in the support except in a neighborhood to the left of zero.

Theorem 4.1. *Let Assumptions 4.1 and (3) hold. Then,*

$$\Pr \left\{ \prod_{j=1}^q \{W_{n,[j]}^- \leq w_j^-\} \prod_{j=1}^q \{W_{n,[j]}^+ \leq w_j^+\} \right\} = \prod_{j=1}^q H^-(w_j^-|0) \cdot \prod_{j=1}^q H^+(w_j^+|0) + o(1) ,$$

as $n \rightarrow \infty$, for any $(w_1^-, \dots, w_q^-, w_1^+, \dots, w_q^+) \in \mathbf{R}^{2q}$.

Theorem 4.1 states that the joint distribution of the induced order statistics are asymptotically independent, with the first q random variables each having limit distribution $H^-(w|0)$ and the remaining q random variables each having limit distribution $H^+(w|0)$. The proof relies on the fact that the induced order statistics $S_n = (W_{n,[q]}^-, \dots, W_{n,[1]}^-, W_{n,[1]}^+, \dots, W_{n,[q]}^+)$ are conditionally independent given (Z_1, \dots, Z_n) , with conditional cdfs

$$H(w|Z_{n,(q)}^-), \dots, H(w|Z_{n,(1)}^-), H(w|Z_{n,(1)}^+), \dots, H(w|Z_{n,(q)}^+) .$$

The result then follows by showing that $Z_{n,(j)}^- = o_p(1)$ and $Z_{n,(j)}^+ = o_p(1)$ for all $j \in \{1, \dots, q\}$, and invoking standard properties of weak convergence.

Theorem 4.1 plays a fundamental role in the proof of Theorem 4.2 in the next section. It is the intermediate step that guarantees that, under the null hypothesis in (3), we have

$$S_n \xrightarrow{d} S = (S_1, \dots, S_{2q}) , \tag{18}$$

where (S_1, \dots, S_{2q}) are i.i.d. with cdf $H(w|0)$. This implies that $S^\pi \stackrel{d}{=} S$ for all permutations $\pi \in \mathbf{G}$, which means that the limit random variable S is indeed invariant to permutations.

Remark 4.1. Under the null hypothesis in (3) it is not necessarily true that the distribution of S_n is invariant to permutations. That is, $S_n^\pi \not\stackrel{d}{=} S_n$. Invariance of S_n to permutations is exactly the condition required for a permutation test to be valid in finite samples, see [Lehmann and Romano \(2005\)](#). The lack of invariance in finite samples lies behind the fact that the random variables in S_n are not draws from $H^-(w|0)$ and $H^+(w|0)$, but rather from $H(w|Z_{n,(j)}^-)$ and $H(w|Z_{n,(j)}^+)$, $j \in$

$\{1, \dots, q\}$. Under the null hypothesis in (3), the latter two distributions are not necessarily the same and therefore permuting the elements of S_n may not keep the joint distribution unaffected. However, under the continuity implied by the null hypothesis, it follows that a sample from $H(w|Z_{n,(j)}^-)$ exhibits a similar behavior to a sample from $H^-(w|0)$, at least for n sufficiently large. This is the value of Theorem 4.1 to prove the results in the following section. ■

In addition to Assumptions 4.1, we also require that the random variable W is either continuous or discrete to prove the main result of the next section. Below we use $\text{supp}(\cdot)$ to denote the support of a random variable.

Assumption 4.2. *The scalar random variable W is continuously distributed conditional on $Z = 0$.*

Assumption 4.3. *The scalar random variable W is discretely distributed with $|\mathcal{W}| = m \in \mathbf{N}$ points of support and such that $\text{supp}(W|Z = z) \subseteq \mathcal{W}$ for all $z \in \mathcal{Z}$.*

We note that Theorem 4.1 does not require either Assumption 4.2 or Assumption 4.3. We however use each of these assumptions as a primitive condition of Assumptions 4.4 and 4.5 below, which are the high-level assumptions we use to prove the asymptotic validity of the permutation test in (14) for the scalar case. For ease of exposition, we present the extension to the case where W is a vector of possibly continuous and discrete random variables in Appendix C.

Remark 4.2. Our assumptions are considerably weaker than those used by Shen and Zhang (2016) to do inference on distributional treatment effects. In particular, while Assumption 4.1 allows Z to be discrete everywhere except in a local neighborhood to the left of zero, Shen and Zhang (2016, Assumption 3.1) require the density of Z to be bounded away from zero and twice continuously differentiable with bounded derivatives. Similar considerations apply to their conditions on $H(w|z)$. In addition, the test proposed by Shen and Zhang (2016) does not immediately apply to the case where W is discrete, as it requires an alternative implementation based on the bootstrap. On the contrary, our test applies indistinctly to continuous and discrete variables. ■

4.2 Asymptotic validity under approximate invariance

We now present our theory of permutation tests under approximate invariance. By approximate invariance we mean that only S is assumed to be invariant to $\pi \in \mathbf{G}$, while S_n may not be invariant - see Remark 4.1. The insight of approximating randomization tests when the conditions required for finite sample validity do not hold in finite samples, but are satisfied in the limit, was first developed by Canay et al. (2017) in a context where the group of transformations \mathbf{G} was essentially sign-changes. Here we exploit this asymptotic framework but with two important modifications. First, our arguments illustrate a concrete case in which the framework in Canay et al. (2017) can be used for the group \mathbf{G} of permutations as opposed to the group \mathbf{G} of sign-changes. The result in Theorem

4.1 provides a fundamental milestone in this direction. Second, we adjust the arguments in Canay et al. (2017) to accommodate rank test statistics, which happen to be discontinuous and do not satisfy the so-called no-ties condition in Canay et al. (2017). We do this by exploiting the specific structure of rank test statistics, together with the requirement that the limit random variable S is either continuously or discretely distributed. We formalize our requirements for the continuous case in the following assumption, where we denote the set of distributions $P \in \mathbf{P}$ satisfying the null in (3) as

$$\mathbf{P}_0 = \{P \in \mathbf{P} : \text{condition (3) holds}\} .$$

Assumption 4.4. *If $P \in \mathbf{P}_0$, then*

(i) $S_n = S_n(X^{(n)}) \xrightarrow{d} S$ under P .

(ii) $S^\pi \stackrel{d}{=} S$ for all $\pi \in \mathbf{G}$.

(iii) S is a continuous random variable taking values in $\mathcal{S} \subseteq \mathbf{R}^{2q}$.

(iv) $T : \mathcal{S} \rightarrow \mathbf{R}$ is invariant to rank preserving transformations, i.e., it only depends on the order of the elements in (S_1, \dots, S_{2q}) .

Assumption 4.4 states the high-level conditions that we use to show the asymptotic validity of the permutation test we propose in (14) and formally stated in Theorem 4.2 below. The assumption is also written in a way that facilitates the comparison with the conditions in Canay et al. (2017). In our setting, Assumption 4.4 follows from Assumptions 4.1-4.2, which may be easier to interpret and impose clear restrictions on the primitives of the model. To see this, note that Theorem 4.1, and the statement in (18) in particular, imply that Assumptions 4.4.(i)-(ii) follow from Assumption 4.1. In turn, Assumption 4.4.(iii) follows directly from Assumption 4.2. Finally, Assumption 4.4.(iv) holds for several rank test statistics and for the test statistic in (12) in particular.

To see the last point more clearly, it is convenient to write the test statistic in (12) using an alternative representation. Let

$$R_{n,i} = \sum_{j=1}^{2q} I\{S_{n,j} \leq S_{n,i}\} , \tag{19}$$

be the rank of $S_{n,i}$ in the pooled vector S_n in (11). Let $R_{n,1}^* < R_{n,2}^* < \dots < R_{n,q}^*$ denote the increasingly ordered ranks $R_{n,1}, \dots, R_{n,q}$ corresponding to the first sample (i.e., first q values) and $R_{n,q+1}^* < \dots < R_{n,2q}^*$ denote the increasingly ordered ranks $R_{n,q+1}, \dots, R_{n,2q}$ corresponding to the second sample (i.e., remaining q values). Letting

$$T^*(S_n) = \frac{1}{q} \sum_{i=1}^q (R_{n,i}^* - i)^2 + \frac{1}{q} \sum_{j=1}^q (R_{n,q+j}^* - j)^2 \tag{20}$$

it follows that

$$T(S_n) = \frac{1}{q} T^*(S_n) - \frac{4q^2 - 1}{12q} ,$$

see Hajek et al. (1999, p. 102). The expression in (20) immediately shows two properties of the statistic $T(s)$. First, $T(s)$ is not a continuous function of s as the ranks make discrete changes with s . Second, $T(s) = T(s')$ whenever s and s' share the same ranks (our Assumption 4.4(iv)), which immediately follows from the definition of $T^*(s)$. This property is what makes rank test statistics violate the no-ties condition in Canay et al. (2017).

We next formalize our requirements for the discrete case in the following assumption.

Assumption 4.5. *If $P \in \mathbf{P}_0$, then*

$$(i) S_n = S_n(X^{(n)}) \xrightarrow{d} S \text{ under } P.$$

$$(ii) S^\pi \stackrel{d}{=} S \text{ for all } \pi \in \mathbf{G}.$$

(iii) S_n are discrete random variables taking values in $\mathcal{S}_n \subseteq \mathcal{S} \equiv \otimes_{j=1}^{2q} \mathcal{S}_1$, where $\mathcal{S}_1 = \bigcup_{k=1}^m \{a_k\}$ is a collection of m distinct singletons.

Parts (i) and (ii) of Assumption 4.5 coincide with parts (i) and (ii) of Assumption 4.4 and, accordingly, follow from Assumption 4.1. Assumption 4.5.(iii) accommodates a case not allowed by Assumption 4.4.(iii), which required S to be continuous. This is important as many covariates are discrete in empirical applications, including the one in Section 6. Note that here we require the random variable S_n to be discrete, which in turn implies that S is discrete too. However, Assumption 4.5 does not impose any requirement on the test statistic $T : \mathcal{S} \rightarrow \mathbf{R}$.

We now formalize our main result in Theorem 4.2, which shows that the permutation test defined in (14) leads to a test that is asymptotically level α whenever either Assumption 4.4 or Assumption 4.5 hold. In addition, the same theorem also shows that Assumptions 4.1-4.3 are sufficient primitive conditions for the asymptotic validity of our test.

Theorem 4.2. *Suppose that either Assumption 4.4 or Assumption 4.5 holds and let $\alpha \in (0, 1)$. Then, $\phi(S_n)$ defined in (14) satisfies*

$$E_P[\phi(S_n)] \rightarrow \alpha \tag{21}$$

as $n \rightarrow \infty$ whenever $P \in \mathbf{P}_0$. Moreover, if $T : \mathcal{S} \rightarrow \mathbf{R}$ is the test statistic in (12) and Assumptions 4.1-4.2 hold, then Assumption 4.4 also holds and (21) follows. Additionally, if instead Assumptions 4.1 and 4.3 hold, then Assumption 4.5 also holds and (21) follows.

Theorem 4.2 shows the validity of the test in (14) when the scalar random variable W is either discrete or continuous. However, the test statistic in (12) and the test construction in (14) immediately apply to the case where W is a vector consisting of a combination of discrete and continuously distributed random variables. In Appendix C we show the validity of the test in (14) for the vector case, which is a result we use in the empirical application of Section 6. Also note that Theorem 4.2 implies that the proposed test is asymptotically similar, i.e., has limiting rejection probability equal to α if $P \in \mathbf{P}_0$.

Remark 4.3. If the distribution P is such that (17) holds and q is such that $-b \leq Z_{n,(q)}^- < Z_{n,(q)}^+ \leq b$, then $\phi(S_n)$ defined in (14) satisfies

$$E_P[\phi_n(S_n)] = \alpha \text{ for all } n .$$

Since (17) implies (3), it follows that our test exhibits finite sample validity for some of the distributions in \mathbf{P}_0 . ■

Remark 4.4. As in Canay et al. (2017), our asymptotic framework is such that the number of permutations in \mathbf{G} , $|\mathbf{G}| = 2q!$, is fixed as $n \rightarrow \infty$. An alternative asymptotic approximation would be one requiring that $|\mathbf{G}| \rightarrow \infty$ as $n \rightarrow \infty$ - see, for example, Hoeffding (1952), Romano (1989), Romano (1990), and more recently, Chung and Romano (2013) and Bugni, Canay and Shaikh (2017). This would require an asymptotically “large” number of observations local to the cutoff and would therefore be less attractive for the problem we consider here. From the technical point of view, these two approximations involve quite different formal arguments. ■

5 Monte Carlo Simulations

In this section, we examine the finite-sample performance of several different tests of (3), including the one introduced in Section 3, with a simulation study. The data for the study is simulated as follows. The scalar baseline covariate is given by

$$W_i = \begin{cases} m(Z_i) + U_{0,i} & \text{if } Z_i < 0 \\ m(Z_i) + U_{1,i} & \text{if } Z_i \geq 0 \end{cases}, \quad (22)$$

where the distribution of $(U_{0,i}, U_{1,i})$ and the functional form of $m(z)$ varies across specifications. In the baseline specification, we set $U_{0,i} = U_{1,i} = U_i$, where U_i is i.i.d. $N(0, 0.15^2)$, and use the same function $m(z)$ as in Shen and Zhang (2016), i.e.

$$m(z) = 0.61 - 0.02z + 0.06z^2 + 0.17z^3 .$$

The distribution of Z_i also varies across the following specifications.

Model 1: $Z_i \sim 2\text{Beta}(2, 4) - 1$ where $\text{Beta}(a, b)$ is the Beta distribution with parameters (a, b) .

Model 2: As in Model 1, but $Z_i \sim \frac{1}{2}(2\text{Beta}(2, 8) - 1) + \frac{1}{2}(1 - 2\text{Beta}(2, 8))$.

Model 3: As in Model 1, but values of Z_i with $Z_i \geq 0$ are scaled by $\frac{1}{4}$.

Model 4: As in Model 1, but Z_i is discretely distributed uniformly on the support

$$\{-1, -0.95, -0.90, \dots, -0.15, -0.10, -\frac{3}{\sqrt{n}}, 0, 0.05, 0.10, 0.15, \dots, 0.90, 0.95, 1\} .$$

Model 5: As in Model 1, but

$$m(z) = \begin{cases} 1.6 + z & \text{if } z < -0.1 \\ 1.5 - 0.4(z + 0.1) & \text{if } z \geq -0.1 \end{cases} .$$

Model 6: As in Model 5, but $Z_i \sim \frac{1}{2} (2\text{Beta}(2, 8) - 1) + \frac{1}{2} (1 - 2\text{Beta}(2, 8))$.

Model 7: as in Model 1, but

$$m(z) = \Phi \left(\frac{-0.85z}{1 - 0.85^2} \right) ,$$

where $\Phi(\cdot)$ denotes the cdf of a standard normal random variable.

The baseline specification in Model 1 has two features: (i) Z_i is continuously distributed with a large number of observations around the cutoff; and (ii) the functional form of $m(z)$ is well behaved - differentiable and relatively flat around the cutoff, see Figures 1a and 1b. The other specifications deviate from the baseline as follows. Models 2 to 4 violate (i) in three different ways, see Figures 1c and 1e. Model 5 violates (ii) by introducing a kink close to the cutoff, see Figure 1d. Model 6 combines Model 2 and 5 to violate both (i) and (ii). Finally, Model 7 is a difficult case (see Kamat, 2017, for a formal treatment of why this case is expected to introduce size distortions in finite samples) where the conditional mean of W exhibits a high first-order derivative at the threshold, see Figure 1f. These variations from the baseline model are partly motivated by the empirical application in Almond et al. (2010), where the running variable may be viewed as discrete as in Model 4, having heaps as in Figure 1c, or exhibiting discontinuities as in Figure 1e.

We consider sample sizes $n \in \{1000, 2500, 5000\}$, a nominal level of $\alpha = 5\%$, and perform 10,000 Monte Carlo repetitions. Models 1 to 7 satisfy the null hypothesis in (3). We additionally consider the same models but with $U_{0,i} \stackrel{d}{\neq} U_{1,i}$ to examine power under the alternative.

Model P1-P7: Same as Models 1-7, but $U_{1,i} \sim \frac{1}{2}N(0.2, 0.15^2) + \frac{1}{2}N(-0.2, 0.15^2)$.

We report results for the following tests.

RaPer and **Per**: the permutation test we propose in this paper in its two versions. The randomized version (RaPer) in (14) and the non-randomized version (Per) that rejects when p_{value} in (16) is below α , see Remark 3.1. We include the randomized version only in the results on size to illustrate the differences between the randomized and non-randomized versions of the test. For power results, we simply report Per, which is the version of the test that practitioners will most likely use. The tuning parameter q is set to

$$q \in \{10, 25, 50, q_{\text{rot}}, \hat{q}_{\text{rot}}\} ,$$

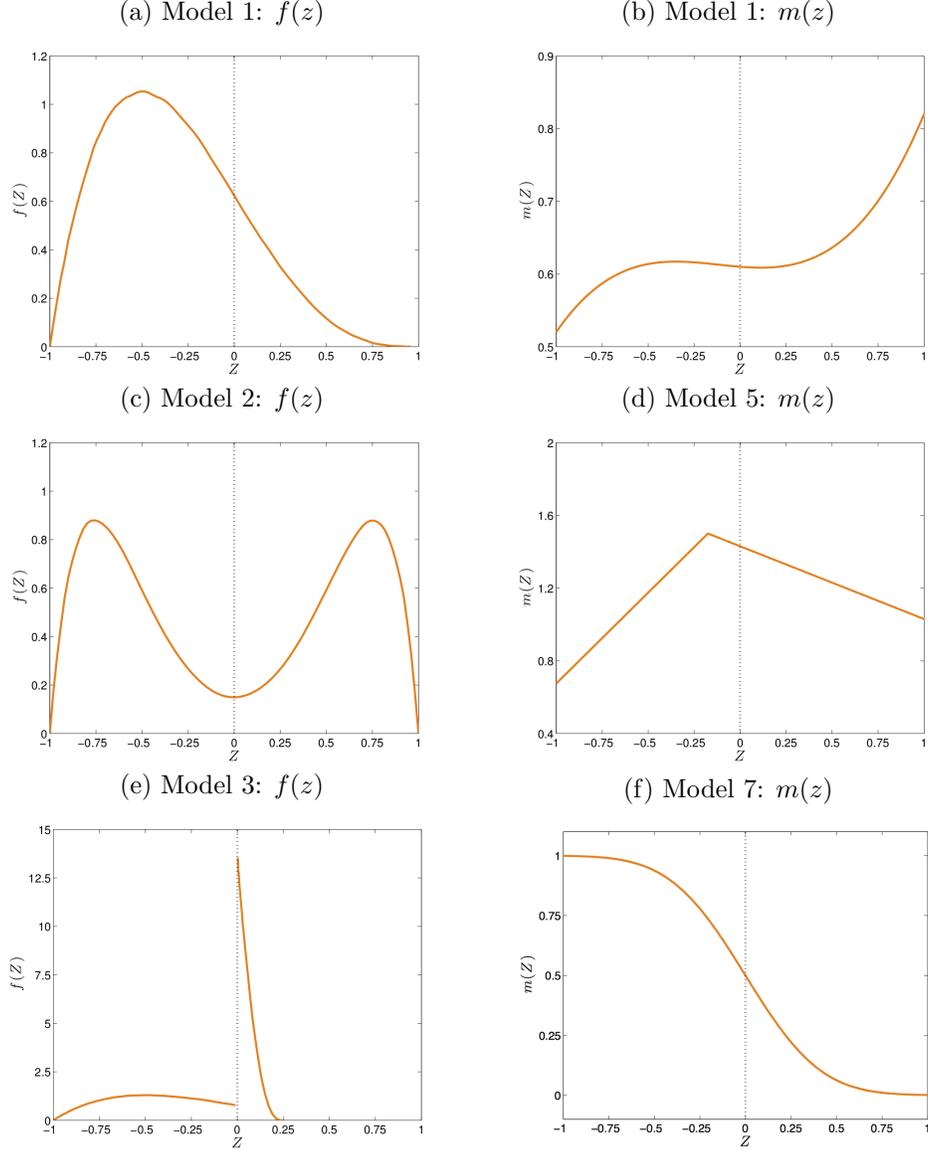


Figure 1: Density of Z (left column) and function $m(z)$ (right column) used in the Monte Carlo model specifications.

where q_{rot} is the rule of thumb in (3.4) and \hat{q}_{rot} is a feasible q_{rot} with all unknown quantities non-parametrically estimated - see Appendix D for details. We set $B = 999$ for the random number of permutations, see Remark 3.2.

SZ: the test proposed by Shen and Zhang (2016) for the null hypothesis of no distributional treatment effect at the cutoff. When used for the null in (3) at $\alpha = 5\%$, this test rejects when

$$A \left(\frac{n}{2} \tilde{f}_n \right)^2 \sup_w \left| \tilde{H}_n^-(w) - \tilde{H}_n^+(w) \right| ,$$

exceeds 1.3581. Here A is a known constant based on the implemented kernel, \tilde{f}_n is a non-

Model	n	RaPer					Per					SZ			CCT
		q					q					c_h			
		10	25	50	q_{rot}	\hat{q}_{rot}	10	25	50	q_{rot}	\hat{q}_{rot}	4.0	4.5	5.0	
1	1000	5.18	4.83	4.92	4.85	4.89	5.05	4.82	4.92	4.79	4.87	3.86	4.29	5.03	5.54
	2500	4.67	5.08	4.86	4.81	4.76	4.57	5.06	4.85	4.80	4.75	4.10	4.70	5.44	4.90
	5000	5.34	5.23	4.75	4.57	4.53	5.24	5.21	4.75	4.56	4.53	4.02	4.54	5.08	4.28
2	1000	5.17	5.31	4.98	5.06	5.10	5.04	5.30	4.97	4.94	4.99	3.89	5.49	7.34	6.36
	2500	5.15	5.02	5.01	5.09	4.85	5.01	5.02	5.00	5.03	4.77	4.37	6.13	8.65	5.39
	5000	5.02	5.35	4.92	5.18	5.35	4.93	5.34	4.92	5.16	5.34	4.55	6.23	9.03	4.87
3	1000	5.17	4.86	4.90	4.86	4.77	5.05	4.84	4.90	4.80	4.77	13.54	13.84	13.97	7.80
	2500	4.67	5.06	4.82	4.78	4.74	4.58	5.05	4.82	4.77	4.74	12.60	12.85	13.31	5.90
	5000	5.35	5.23	4.74	4.60	4.64	5.25	5.21	4.73	4.59	4.64	13.53	13.73	14.06	5.40
4	1000	4.84	4.63	4.69	4.93	5.02	4.75	4.62	4.69	4.82	5.01	26.80	18.21	15.00	3.94
	2500	5.09	5.06	5.00	4.92	4.96	5.00	5.05	5.00	4.91	4.96	16.19	11.73	10.52	4.50
	5000	4.59	5.01	4.76	4.98	4.80	4.53	4.98	4.76	4.97	4.80	7.66	7.18	7.75	5.30
5	1000	5.37	6.18	17.29	5.43	5.49	5.27	6.16	17.27	5.32	5.38	4.93	8.23	13.00	5.71
	2500	4.66	5.34	6.71	5.02	5.11	4.54	5.34	6.71	4.99	5.08	6.73	14.36	25.22	5.05
	5000	5.38	5.34	5.32	5.19	5.05	5.30	5.32	5.32	5.19	5.05	8.64	21.29	35.04	3.97
6	1000	6.77	18.20	16.50	6.81	6.85	6.62	18.15	16.50	6.61	6.74	7.30	13.91	21.02	10.19
	2500	5.67	10.00	33.07	5.65	5.62	5.58	9.98	33.05	5.53	5.60	12.02	26.10	39.26	12.17
	5000	5.03	6.48	14.40	5.91	6.43	4.89	6.46	14.38	5.91	6.42	16.94	40.55	56.54	13.55
7	1000	6.03	19.74	85.07	5.99	5.98	5.94	19.70	85.07	5.88	5.86	5.10	7.05	9.86	5.69
	2500	4.83	7.22	24.08	5.88	5.68	4.71	7.22	24.05	5.84	5.64	5.02	6.97	10.14	5.06
	5000	5.33	6.08	9.25	6.35	6.34	5.26	6.05	9.24	6.35	6.34	4.83	6.36	9.08	4.26

Table 1: Rejection probabilities (in %) under the null hypothesis. 10,000 replications.

parametric estimate of the density of Z_i at $Z_i = 0$, and $\tilde{H}_n^-(w)$ and $\tilde{H}_n^+(w)$ are local linear estimates of the cdfs in (4). The kernel is set to a triangular kernel. Shen and Zhang (2016) propose using the following (undersmoothed) rule of thumb bandwidth for the nonparametric estimates,

$$h_n = h_n^{CCT} n^{1/5-1/c_h}, \quad (23)$$

where h_n^{CCT} is a sequential bandwidth based on Calonico et al. (2014), and c_h is an under-smoothing parameter - see Appendix D for details. We follow Shen and Zhang (2016) and report results for $c_h \in \{4.0, 4.5, 5.0\}$, where $c_h = 4.5$ is their recommended choice.

CCT: the test proposed by Calonico et al. (2014) for the null hypothesis of no average treatment effect at the cutoff. When used for the null in (3) at $\alpha = 5\%$, this test rejects when

$$\frac{|\hat{\mu}_n^{-,bc} - \hat{\mu}_n^{+,bc}|}{\hat{V}_n^{bc}},$$

Model	n	Per					SZ			SZ (Size Adj.)			CCT
		q					c_h			c_h			
		10	25	50	q_{rot}	\hat{q}_{rot}	4.0	4.5	5.0	4.0	4.5	5.0	
P1	1000	8.23	19.20	52.62	12.77	12.04	13.80	18.71	23.75	17.36	20.98	23.59	5.93
	2500	8.46	21.17	53.76	30.39	30.15	40.95	55.67	67.06	45.58	57.09	65.13	4.72
	5000	8.43	20.07	53.05	60.70	60.53	81.70	92.50	96.90	86.19	93.68	96.77	4.97
P2	1000	8.73	20.17	53.10	8.80	8.69	7.17	11.28	17.19	8.90	10.33	11.88	7.40
	2500	8.38	19.22	52.69	10.41	11.24	17.18	29.61	44.97	19.18	26.11	30.92	5.47
	5000	8.24	20.45	53.74	18.57	21.00	42.75	65.16	81.28	44.86	59.10	69.75	5.19
P3	1000	8.23	19.20	52.59	12.56	20.89	16.08	19.38	22.23	4.48	5.52	6.58	6.47
	2500	8.44	21.17	53.84	30.25	59.68	34.92	43.00	51.26	13.30	17.40	22.73	5.29
	5000	8.43	20.05	52.96	60.81	92.56	66.50	78.38	86.36	33.33	46.69	57.67	4.95
P4	1000	8.16	20.58	53.92	8.70	15.85	43.75	38.91	41.33	5.33	7.95	12.04	4.59
	2500	8.41	20.08	52.85	16.12	41.59	61.45	71.04	80.12	15.88	36.91	57.09	4.87
	5000	8.52	20.50	53.36	33.62	78.25	91.78	97.52	99.05	84.72	94.94	97.96	4.83
P5	1000	8.40	20.43	56.84	9.47	9.43	15.42	22.01	29.66	15.54	15.05	14.27	5.84
	2500	8.46	21.18	53.86	19.83	20.39	39.59	53.19	65.70	32.20	26.11	20.93	4.79
	5000	8.55	20.25	52.99	40.69	41.58	70.74	83.28	90.95	55.88	37.09	23.39	4.81
P6	1000	9.24	25.94	46.50	9.24	9.16	11.36	20.53	31.43	8.14	9.02	10.62	9.37
	2500	8.68	21.89	62.01	10.42	10.89	20.26	34.00	46.84	10.40	10.70	12.46	10.02
	5000	8.16	21.57	56.69	17.85	19.14	30.25	44.32	57.73	13.02	12.03	13.65	12.72
P7	1000	8.89	26.94	81.03	8.81	9.01	16.58	24.87	32.98	16.27	18.04	18.17	5.92
	2500	8.48	21.77	58.89	16.06	16.02	48.57	66.94	80.30	48.40	57.06	58.06	4.78
	5000	8.53	20.33	54.33	31.83	31.11	85.00	95.62	98.64	85.56	93.38	95.72	4.85

Table 2: Rejection probabilities (in %) under the alternative hypothesis. 10,000 replications.

exceeds 1.96. Here $\hat{\mu}_n^{-,bc}$ and $\hat{\mu}_n^{+,bc}$ are bias corrected local linear estimates of the conditional means of W_i to the left and right of $Z_i = 0$, and \hat{V}_n^{bc} is a novel standard error formula that accounts for the variance of the estimated bias. The kernel is set to a triangular kernel. We implement their test using their proposed bandwidth - see Appendix D for details.

Table 1 reports rejection probabilities under the null hypothesis for all models and all tests considered. Across all cases, the permutation test controls size remarkably well. In particular, the feasible rule of thumb \hat{q}_{rot} in (3.4) delivers rejection rates between 4.53% and 6.74%. On the other hand, SZ returns rejection rates between 4.29% and 40.55% for their recommended choice of $c_h = 4.5$. Except in the baseline Model 1 where SZ performs similarly to Per, in all other models Per clearly dominates SZ in terms of size control. Finally, CCT controls size very well in all models except Model 6, where the lack of smoothness affects the local polynomial estimators and returns

Model	n	Per		SZ			CCT
		q		c_h			
		q_{rot}	\hat{q}_{rot}	4.0	4.5	5.0	
1	1000	17.00	16.59	90.76	109.95	128.11	137.48
	2500	33.00	32.93	219.93	273.20	324.91	349.21
	5000	56.00	56.08	427.10	540.87	653.45	699.11
2	1000	10.00	10.00	49.60	67.43	88.09	98.84
	2500	14.00	14.93	120.33	170.24	230.87	255.08
	5000	23.00	24.52	230.78	335.18	466.67	500.58
3	1000	17.00	25.91	62.50	73.20	82.61	86.50
	2500	33.00	54.23	147.40	176.64	202.83	213.31
	5000	56.00	95.59	283.75	346.32	402.97	423.25
4	1000	11.00	19.91	116.78	141.49	165.08	179.68
	2500	21.00	40.58	260.21	324.02	385.11	415.71
	5000	36.00	69.88	494.24	624.02	755.82	801.74
5	1000	12.00	11.89	94.85	114.90	133.89	124.83
	2500	23.00	23.48	222.03	275.94	328.25	281.12
	5000	40.00	39.89	401.72	508.99	614.76	487.69
6	1000	10.00	10.00	51.20	69.97	91.80	89.92
	2500	12.00	13.60	115.61	163.21	221.12	199.77
	5000	21.00	22.30	208.29	300.11	414.91	324.12
7	1000	10.00	10.05	94.81	114.86	133.90	132.01
	2500	18.00	18.42	203.43	252.76	300.72	319.24
	5000	31.00	31.22	347.02	439.56	531.06	604.18

Table 3: Average number of observations (to one side) used in the tests reported in Table 1.

rejection rates between 10.19% and 13.55%. Table 3 reports the average number of observations⁵ used by each of the tests and illustrates how both SZ and CCT consistently use a larger number of observations around the cutoff than Per.

Two final lessons arise from Table 1. First, the differences between RaPer and Per are negligible, even when $q = 10$. Second, Per is usually less sensitive to the choice of q than SZ is to the choice of c_h . The notable exceptions are Model 6, where both tests appear to be equally sensitive; and Model 7, where Per is more sensitive for $n = 1,000$ and $n = 2,500$. Recall that Model 7 is a particularly difficult case in RDD (see Kamat, 2017), but even in this case Per controls size well for n sufficiently large or q sufficiently small. Most importantly, the rejection probabilities under the null hypothesis are very close to the nominal level for our suggested rule of thumb \hat{q}_{rot} .

Table 2 reports rejection probabilities under the alternative hypothesis for all models and all

⁵In the case of SZ and CCT, we compute the average of the number of observations to the left and right of the cutoff, and then take an average across simulations. In the case of Per, we simply average q across simulations.

tests considered. Since SZ may severely over-reject under the null hypothesis, we report both raw and size-adjusted rejection rates. For the recommended values of tuning parameters, the size adjusted power of SZ is consistently above the one of Per in Models P1, P2, and P7. In Models P3-P6, Per delivers higher power than SZ in 9 out of the 12 cases considered; while in the remaining three cases (P4 with $n = 5,000$ and P5 with $n \in \{1,000; 2,500\}$), SZ delivers higher power. This is remarkable as Table 3 shows that Per uses considerably fewer observations than SZ does.⁶ The power of CCT, as expected, does not exceed the rejection probabilities under the null hypothesis.

6 Empirical application

In this section we reevaluate the validity of the design in Lee (2008). Lee studies the benefits of incumbency on electoral outcomes using a discontinuity constructed with the insight that the party with the majority wins. Specifically, the running variable Z is the difference in vote shares between Democrats and Republicans in time t . The assignment rule then takes a cutoff value of zero that determines the treatment of incumbency to the Democratic candidate, which is used to study their election outcomes in time $t + 1$. The data set contains six covariates that contain electoral information on the Democrat runner and the opposition in time $t - 1$ and t . Out of the six variables, one is continuous (Democrat vote share $t - 1$) and the remaining are discrete. The total number of observations is 6,559 with 2,740 below the cutoff. The dataset is publicly available at <http://economics.mit.edu/faculty/angrist/data1/mhe>.

Lee assessed the credibility of the design in this application by inspecting discontinuities in means of the baseline covariates. His test is based on local linear regressions with observations in different margins around the cutoff. The estimates and graphical illustrations of the conditional means are used to conclude that there are no discontinuities at the cutoff in the baseline covariates. Here, we frame the validity of the design in terms of the hypothesis in (3) and use the newly developed permutation test as described in Section 3.1, using \hat{q}_{rot} as our default choice for the number of observations q .⁷ Our test allows for continuous or discrete covariates, and so it does not require special adjustments to accommodate discrete covariates; cf. Remark 4.2. In addition, our test allows the researcher to test for the hypothesis of continuity of individual covariates, in which case W includes a single covariate; as well as continuity of the entire vector of covariates, in which case W includes all six covariates. Finally, we also report the results of test CCT, as described in Section 5, for the continuity of means at the cutoff.

Table 4 reports the p -values for continuity of each of the six covariates individually, as well as

⁶We computed the equivalent of Table 3 for the results in Table 2 and obtained very similar numbers, so we only report Table 3 to save space.

⁷We also computed our test using $0.8\hat{q}_{\text{rot}}$, $1.2\hat{q}_{\text{rot}}$, and the alternative rule of thumb discussed in footnote 4, and found similar results.

Variable	Per	CCT	SZ
Democrat vote share $t - 1$	4.60	83.74	31.21
Democrat win $t - 1$	1.20	7.74	–
Democrat political experience t	0.30	21.43	–
Opposition political experience t	3.60	83.14	–
Democrat electoral experience t	13.31	25.50	–
Opposition electoral experience t	4.20	92.79	–
Joint Test - CvM statistic	16.42		
Joint Test - Max statistic	1.70		

Table 4: Test results with p -value (in %) for covariates in Lee (2008)

the joint test for the continuity of the six dimensional vector of covariates; see Appendix C for details. Our results show that the null hypothesis of continuity of the conditional distributions of the covariates at the cutoff is rejected for most of the covariates at a 5% significance level, in contrast to the results reported by Lee (2008) and the results of the CCT test in Table 4. The differences between our test and tests based on conditional means can be illustrated graphically. Figure 2(a)-(b) displays the histogram and empirical cdf (based on \hat{q}_{rot} observations on each side) of the continuous covariate *Democrat vote share* $t - 1$. The histogram exhibits a longer right tail for observations to the right of the threshold (in orange), and significantly more mass at shares below 50% for observations to the left of the threshold (in blue). The empirical CDFs are similar up until the 40th quantile, approximately, and then are markedly different. Our test formally shows that the observed differences are statistically significant. On the contrary, the conditional means from the left and from the right appear to be similar around the cutoff and so tests for the null hypothesis in (6) fail to reject the null in (3); see Figure 2c. A similar intuition applies to the rest of the covariates. Finally, we note that \hat{q}_{rot} in the implementation of our test ranges from 80 to 115, depending on the covariate, while the average number of effective observations (i.e. the average of observations to the left and right of the cutoff) used by CCT ranges from 880 to 1113. This is consistent with one asymptotic framework assuming few effective observations around the cutoff and another assuming a large and growing number of observations around the cutoff.

Table 4 also reports the test by Shen and Zhang (2016), as described in Section 5, for the only continuously distributed covariate of this application. This tests fails to reject the null hypothesis with a p -value of 31.21%. The rest of the covariates in this empirical application are discrete and so the results in Shen and Zhang (2016) do not immediately apply; see Remark 4.2.

The standard practice in applied work appears to be to test the hypothesis of continuity individually for each covariate. This is informative as it can provide information as to which covariate may or may not be problematic. However, testing many individual hypotheses may lead to spurious rejections (due to a multiple testing problem). In addition, the statement in (3) is a statement

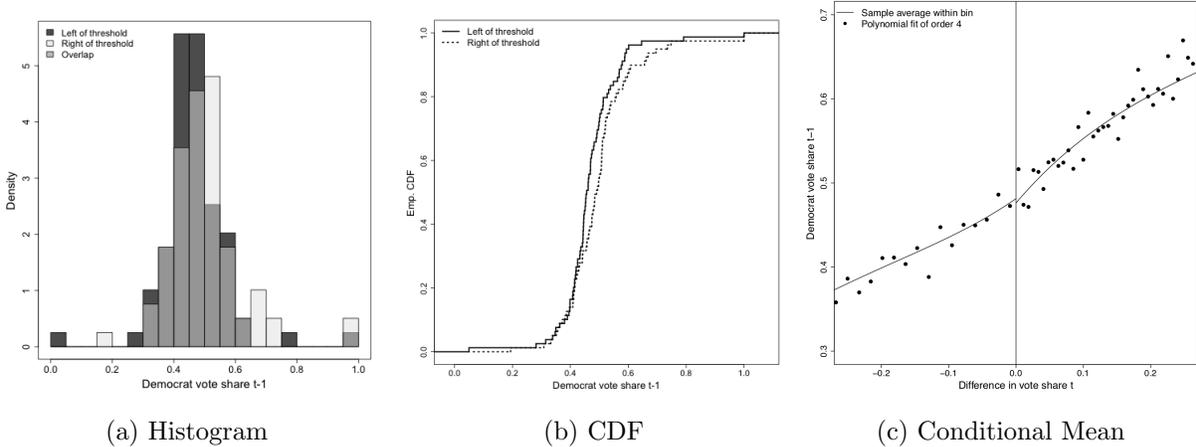


Figure 2: Histogram, CDF, and conditional means for *Democrat vote share $t - 1$*

about the vector W that includes all baseline covariates in the design. We therefore report in Table 4, in addition to each individual test, the results for the joint test that uses all six covariates in the construction of the test statistic - as explained in detail in Section C. Table 4 shows that the results for the joint test depend on the choice of test statistic used in its construction. If one uses the Cramér Von Mises test statistic in (12), the null hypothesis in (3) is not rejected, with a p -value of 17.62%. If one instead uses the max-type test statistic introduced in Appendix C, see (C-34), the null hypothesis in (3) is rejected, with a p -value of 1.70%. In unreported simulations we found that the max-type test statistic appears to have significantly higher power than the Cramér Von Mises test statistic in the multivariate case, which is consistent with the results of this particular application. It is worth noting that in the case of scalar covariates, these two test statistics are numerically identical. We therefore recommend the Max test statistic in (C-34) for the multivariate case, which is the default option in the companion `rdperm` Stata package.

7 Concluding remarks

In this paper we propose an asymptotically valid permutation test for the hypothesis of continuity of the distribution of baseline covariates at the cutoff in the regression discontinuity design (RDD). The asymptotic framework for our test is based on the simple intuition that observations close to the cutoff are *approximately* identically distributed on either side of it when the null hypothesis holds. This allows us to permute these observations to conduct an approximately valid test. Formally, we exploit the framework, with novel additions, from Canay et al. (2017), which first developed the

insight of approximating randomization tests in this manner. Our results also represent a novel application of induced order statistics to frame our problem, and we present a result on induced order statistics that may be of independent interest.

A final aspect we would like to highlight of our test is its simplicity. The test only requires computing two empirical cdfs for the induced order statistic, and does not involve kernels, local polynomials, bias correction, or bandwidth choices. Importantly, we have developed the `rdperm` **Stata** package and the `RATtest` **R** package that allow for effortless implementation of the test we propose in this paper.

A Proof of Theorem 4.1

First, note that the joint distribution of the induced order statistics $W_{n,[q]}^-, \dots, W_{n,[1]}^-, W_{n,[1]}^+, \dots, W_{n,[q]}^+$ are conditionally independent given (Z_1, \dots, Z_n) , with conditional cdfs

$$H(w|Z_{n,(q)}^-), \dots, H(w|Z_{n,(1)}^-), H(w|Z_{n,(1)}^+), \dots, H(w|Z_{n,(q)}^+).$$

A proof of this result can be found in [Bhattacharya \(1974, Lemma 1\)](#). Now let $\mathcal{A} = \sigma(Z_1, \dots, Z_n)$ be the sigma algebra generated by (Z_1, \dots, Z_n) . It follows that

$$\begin{aligned} \Pr \left\{ \bigcap_{j=1}^q \{W_{n,[j]}^- \leq w_j^-\} \bigcap_{j=1}^q \{W_{n,[j]}^+ \leq w_j^+\} \right\} &= E \left[\Pr \left\{ \bigcap_{j=1}^q \{W_{n,[j]}^- \leq w_j^-\} \bigcap_{j=1}^q \{W_{n,[j]}^+ \leq w_j^+\} \mid \mathcal{A} \right\} \right] \\ &= E \left[\prod_{j=1}^q H(w_j^- | Z_{n,(j)}^-) \cdot \prod_{j=1}^q H(w_j^+ | Z_{n,(j)}^+) \right]. \end{aligned}$$

The first equality follow from the law of iterated expectations and the last equality follows from the conditional independence of the induced order statistics.

Let $f_{n,(q^-, \dots, q^+)}(z_{q^-}, \dots, z_{q^+})$ denote the joint density of

$$Z_{n,(q)}^- \leq \dots \leq Z_{n,(1)}^- < 0 \leq Z_{n,(1)}^+ \leq \dots \leq Z_{n,(q)}^+,$$

so that we can write the last term in the previous display as

$$\int_0^\infty \int_0^{z_{q^+}} \dots \int_0^{z_{(q-1)^-}} \prod_{j=1}^q H(w_j^- | z_{j^-}) \cdot \prod_{j=1}^q H(w_j^+ | z_{j^+}) f_{n,(q^-, \dots, q^+)}(z_{q^-}, \dots, z_{q^+}) dz_{q^-}, \dots, dz_{q^+}.$$

By (3), the integrand term

$$\prod_{j=1}^q H(w_j^- | z_{j^-}) \cdot \prod_{j=1}^q H(w_j^+ | z_{j^+})$$

is a bounded continuous function of $(z_{q^-}, \dots, z_{1^-}, z_{1^+}, \dots, z_{q^+})$ at $(0, 0, \dots, 0)$. Suppose that the order statistics $Z_{n,(j)}^-$ and $Z_{n,(q)}^+$, for $j \in \{1, \dots, q\}$, converge in distribution to a degenerate distribution with mass at $(0, 0, \dots, 0)$. It would then follow from the definition of weak convergence, the asymptotic uniform integrability of the integrand term above, and [van der Vaart \(1998, Theorem 2.20\)](#) that

$$\lim_{n \rightarrow \infty} E \left[\prod_{j=1}^q H(w_j^- | z_{j^-}) \cdot \prod_{j=1}^q H(w_j^+ | z_{j^+}) \right] = E \left[\prod_{j=1}^q H^-(w_j^- | 0) \cdot \prod_{j=1}^q H^+(w_j^+ | 0) \right].$$

Hence, it is sufficient to prove that for any given $j \in \{1, \dots, q\}$, $Z_{n,(j)}^- = o_p(1)$ and $Z_{n,(q)}^+ = o_p(1)$. We prove $Z_{n,(q)}^+ = o_p(1)$ by complete induction, and omit the other proof as the result follows from similar arguments.

Take $j = 1$ and let $\epsilon > 0$. By Assumption 4.1, it follows that

$$F^+(\epsilon) = \Pr\{Z_i \in [0, \epsilon]\} > 0.$$

Next, note that

$$\begin{aligned} F_{n,(1)}^+(\epsilon) &\equiv \Pr\{Z_{n,(1)}^+ \leq \epsilon\} = \Pr\{\text{at least 1 of the } Z_i \text{ is such that } Z_i \in [0, \epsilon]\} \\ &= \sum_{i=1}^n \binom{n}{i} [F^+(\epsilon)]^i [1 - F^+(\epsilon)]^{n-i} \\ &= \sum_{i=0}^n \binom{n}{i} [F^+(\epsilon)]^i [1 - F^+(\epsilon)]^{n-i} - [1 - F^+(\epsilon)]^n \\ &= 1 - [1 - F^+(\epsilon)]^n. \end{aligned} \tag{A-24}$$

Since $F^+(\epsilon) > 0$ for any $\epsilon > 0$, it follows that $\Pr\{Z_{n,(1)}^+ > \epsilon\} = [1 - F^+(\epsilon)]^n \rightarrow 0$ as $n \rightarrow \infty$ and $Z_{n,(1)}^+ = o_p(1)$. Now let $F_{n,(j)}^+(\epsilon)$ denote the cdf of $Z_{n,(j)}^+$, which is given by

$$\begin{aligned} F_{n,(j)}^+(\epsilon) &= \Pr\{Z_{n,(j)}^+ \leq \epsilon\} \\ &= \Pr\{\text{at least } j \text{ of the } Z_i \text{ are such that } Z_i \in [0, \epsilon]\} \\ &= \sum_{i=j}^n \binom{n}{i} [F^+(\epsilon)]^i [1 - F^+(\epsilon)]^{n-i} \\ &= F_{n,(j+1)}^+(\epsilon) + \binom{n}{j} [F^+(\epsilon)]^j [1 - F^+(\epsilon)]^{n-j}, \end{aligned}$$

so that we can write

$$1 - F_{n,(j+1)}^+(\epsilon) = 1 - F_{n,(j)}^+(\epsilon) - \binom{n}{j} [F^+(\epsilon)]^j [1 - F^+(\epsilon)]^{n-j} \text{ for } j \in \{1, \dots, q-1\}. \quad (\text{A-25})$$

It follows from (A-24) that $1 - F_{n,(1)}^+(\epsilon) \rightarrow 0$ for any $\epsilon > 0$ as $n \rightarrow \infty$. In order to complete the proof we assume that $1 - F_{n,(j)}^+(\epsilon) \rightarrow 0$ for $j \in \{1, \dots, q-1\}$ and show that this implies that $1 - F_{n,(j+1)}^+(\epsilon) \rightarrow 0$. By (A-25) this is equivalent to showing that

$$\binom{n}{j} [F^+(\epsilon)]^j [1 - F^+(\epsilon)]^{n-j} \rightarrow 0.$$

To this end, note that

$$\binom{n}{j} [F^+(\epsilon)]^j [1 - F^+(\epsilon)]^{n-j} \leq n^j [1 - F^+(\epsilon)]^{n-j} = \left[e^{\frac{j \log n}{n-j}} [1 - F^+(\epsilon)] \right]^{n-j} \rightarrow 0,$$

where the convergence follows after noticing that there exists $N \in \mathbf{R}$ such that $e^{\frac{j \log n}{n-j}} [1 - F^+(\epsilon)] < 1$ for all $n > N$ and any $j \in \{1, \dots, q-1\}$. The result follows. ■

B Proof of Theorem 4.2

Part 1.

Continuous case: Let $P_n = \otimes_{i=1}^n P$ with $P \in \mathbf{P}_0$ be given. By Assumption 4.4(i) and the Almost Sure Representation Theorem (see van der Vaart, 1998, Theorem 2.19), there exists \tilde{S}_n , \tilde{S} , and $U \sim U(0, 1)$, defined on a common probability space $(\Omega, \mathcal{A}, \tilde{P})$, such that

$$\tilde{S}_n \rightarrow \tilde{S} \text{ w.p.1,}$$

$\tilde{S}_n \stackrel{d}{=} S_n$, $\tilde{S} \stackrel{d}{=} S$, and $U \perp (\tilde{S}_n, \tilde{S})$. Consider the permutation test based on \tilde{S}_n , this is,

$$\tilde{\phi}(\tilde{S}_n, U) \equiv \begin{cases} 1 & T(\tilde{S}_n) > T^{(k)}(\tilde{S}_n) \text{ or } T(\tilde{S}_n) = T^{(k)}(\tilde{S}_n) \text{ and } U < a(\tilde{S}_n) \\ 0 & T(\tilde{S}_n) < T^{(k)}(\tilde{S}_n) \end{cases}.$$

Denote the randomization test based on \tilde{S} by $\tilde{\phi}(\tilde{S}, U)$, where the same uniform variable U is used in $\tilde{\phi}(\tilde{S}_n, U)$ and $\tilde{\phi}(\tilde{S}, U)$.

Since $\tilde{S}_n \stackrel{d}{=} S_n$, it follows immediately that $E_{P_n}[\phi(S_n)] = E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)]$. In addition, since $\tilde{S} \stackrel{d}{=} S$, Assumption 4.4(ii) implies that $E_{\tilde{P}}[\tilde{\phi}(\tilde{S}, U)] = \alpha$ by the usual arguments behind randomization tests, see [Lehmann and Romano \(2005, Chapter 15\)](#). It therefore suffices to show

$$E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)] \rightarrow E_{\tilde{P}}[\tilde{\phi}(\tilde{S}, U)] . \quad (\text{B-26})$$

In order to show (B-26), let E_n be the event where the ordered values of $\{\tilde{S}_j : 1 \leq j \leq 2q\}$ and $\{\tilde{S}_{n,j} : 1 \leq j \leq 2q\}$ correspond to the same permutation π of $\{1, \dots, 2q\}$, i.e., if $S_{\pi(j)} = S_{(k)}$ then $S_{n,\pi(j)} = S_{n,(k)}$ for $1 \leq j \leq 2q$ and $1 \leq k \leq 2q$. We first claim that $I\{E_n\} \rightarrow 1$ w.p.1. To see this, note that Assumption 4.4(iii) and $\tilde{S} \stackrel{d}{=} S$ imply that

$$\tilde{S}_{(1)}(\omega) < \tilde{S}_{(2)}(\omega) < \dots < \tilde{S}_{(2q)}(\omega) \quad (\text{B-27})$$

for all ω in a set with probability one under \tilde{P} . Moreover, since $\tilde{S}_n \rightarrow \tilde{S}$ w.p.1, there exists a set Ω^* with $\tilde{P}\{\Omega^*\} = 1$ such that both (B-27) and $\tilde{S}_n(\omega) \rightarrow \tilde{S}(\omega)$ hold for all $\omega \in \Omega^*$. For all ω in this set, let $\pi(1, \omega), \dots, \pi(2q, \omega)$ be the permutation that delivers the order statistics in (B-27). It follows that for any $\omega \in \Omega^*$ and any $j \in \{1, \dots, 2q-1\}$, if $\tilde{S}_{\pi(j, \omega)}(\omega) < \tilde{S}_{\pi(j+1, \omega)}(\omega)$ then

$$\tilde{S}_{n,\pi(j, \omega)}(\omega) < \tilde{S}_{n,\pi(j+1, \omega)}(\omega) \text{ for } n \text{ sufficiently large} . \quad (\text{B-28})$$

We can therefore conclude that

$$I\{E_n\} \rightarrow 1 \text{ w.p.1} ,$$

which proves the first claim.

We now prove (B-26) in two steps. First, we note that

$$E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)I\{E_n\}] = E_{\tilde{P}}[\tilde{\phi}(\tilde{S}, U)I\{E_n\}] . \quad (\text{B-29})$$

This is true because, on the event E_n , the rank statistics in (19) of the vectors \tilde{S}_n^π and \tilde{S}^π coincide for all $\pi \in \mathbf{G}$, and by Assumption 4.4(iv), the test statistic $T(S)$ only depends on the order of the observations, leading to $\tilde{\phi}(\tilde{S}_n, U) = \tilde{\phi}(\tilde{S}, U)$ on E_n . Second, since $I\{E_n\} \rightarrow 1$ w.p.1 it follows that $\tilde{\phi}(\tilde{S}, U)I\{E_n\} \rightarrow \tilde{\phi}(\tilde{S}, U)$ w.p.1 and $\tilde{\phi}(\tilde{S}_n, U)I\{E_n^c\} \rightarrow 0$ w.p.1. We can therefore use (B-29) and invoke the dominated convergence theorem to conclude that,

$$\begin{aligned} E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)] &= E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)I\{E_n\}] + E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)I\{E_n^c\}] \\ &= E_{\tilde{P}}[\tilde{\phi}(\tilde{S}, U)I\{E_n\}] + E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)I\{E_n^c\}] \\ &\rightarrow E_{\tilde{P}}[\tilde{\phi}(\tilde{S}, U)] . \end{aligned}$$

This completes the proof of the first part of the statement of the theorem for the continuous case.

Discrete case: The proof for the discrete setting is similar to the continuous one with few intuitive differences. We reproduce it here for completeness.

Let $P_n = \otimes_{i=1}^n P$ with $P \in \mathbf{P}_0$ be given. By Assumption 4.5(i) and the Almost Sure Representation Theorem (see [van der Vaart, 1998, Theorem 2.19](#)), there exists \tilde{S}_n, \tilde{S} , and $U \sim U(0, 1)$, defined on a common probability space $(\Omega, \mathcal{A}, \tilde{P})$, such that

$$\tilde{S}_n \rightarrow \tilde{S} \text{ w.p.1} ,$$

$\tilde{S}_n \stackrel{d}{=} S_n$, $\tilde{S} \stackrel{d}{=} S$, and $U \perp (\tilde{S}_n, \tilde{S})$. Consider the permutation test based on \tilde{S}_n , this is,

$$\tilde{\phi}(\tilde{S}_n, U) \equiv \begin{cases} 1 & T(\tilde{S}_n) > T^{(k)}(\tilde{S}_n) \text{ or } T(\tilde{S}_n) = T^{(k)}(\tilde{S}_n) \text{ and } U < a(\tilde{S}_n) \\ 0 & T(\tilde{S}_n) < T^{(k)}(\tilde{S}_n) \end{cases} .$$

Denote the randomization test based on \tilde{S} by $\tilde{\phi}(\tilde{S}, U)$, where the same uniform variable U is used in $\tilde{\phi}(\tilde{S}_n, U)$ and $\tilde{\phi}(\tilde{S}, U)$.

Since $\tilde{S}_n \stackrel{d}{=} S_n$, it follows immediately that $E_{P_n}[\phi(S_n)] = E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)]$. In addition, since $\tilde{S} \stackrel{d}{=} S$, Assumption 4.5(ii) implies that $E_{\tilde{P}}[\tilde{\phi}(\tilde{S}, U)] = \alpha$ by the usual arguments behind randomization tests, see [Lehmann and Romano \(2005, Chapter 15\)](#). It therefore suffices to show

$$E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)] \rightarrow E_{\tilde{P}}[\tilde{\phi}(\tilde{S}, U)] . \quad (\text{B-30})$$

In order to show (B-30), let E_n be the event where $\tilde{S}_n = \tilde{S}$. We first claim that $I\{E_n\} \rightarrow 1$ w.p.1. To see this, note that by Assumption 4.5(iii), the discrete random variable \tilde{S}_n takes values in $\mathcal{S}_n \subseteq \mathcal{S} \equiv \otimes_{j=1}^{2q} \mathcal{S}_1$ for all $n \geq 1$. The set \mathcal{S} is closed by virtue of being a finite collection of singletons, and by the Portmanteau Lemma (see [van der Vaart, 1998, Lemma 2.2](#)) it follows that

$$1 = \limsup_{n \rightarrow \infty} \tilde{P}\{\tilde{S}_n \in \mathcal{S}_n\} \leq \limsup_{n \rightarrow \infty} \tilde{P}\{\tilde{S}_n \in \mathcal{S}\} \leq \tilde{P}\{\tilde{S} \in \mathcal{S}\} , \quad (\text{B-31})$$

meaning that $\text{supp}(\tilde{S}) \subseteq \mathcal{S}$. Moreover, since $\tilde{S}_n \rightarrow \tilde{S}$ w.p.1, there exists a set Ω^* with $\tilde{P}\{\Omega^*\} = 1$ such that $\tilde{S}_n(\omega) \rightarrow \tilde{S}(\omega)$ holds for all $\omega \in \Omega^*$. It follows that for any $\omega \in \Omega^*$ and any $j \in \{1, \dots, 2q\}$,

$$\tilde{S}_{n,j}(\omega) = \tilde{S}_j(\omega) \text{ for } n \text{ sufficiently large} , \quad (\text{B-32})$$

which follows from the fact that both \tilde{S} and \tilde{S}_n are discrete random variables taking values in (possibly a subset of) the finite collection of points in $\mathcal{S} \equiv \otimes_{j=1}^{2q} \mathcal{S}_1 = \otimes_{j=1}^{2q} \bigcup_{k=1}^m \{a_k\}$. We conclude that

$$I\{E_n\} \rightarrow 1 \text{ w.p.1} ,$$

which proves the first claim.

We now prove (B-30) in two steps. First, we note that

$$E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)I\{E_n\}] = E_{\tilde{P}}[\tilde{\phi}(\tilde{S}, U)I\{E_n\}] . \quad (\text{B-33})$$

This is true because, on the event E_n , \tilde{S}_n^π and \tilde{S}^π coincide for all $\pi \in \mathbf{G}$, leading to $\tilde{\phi}(\tilde{S}_n, U) = \tilde{\phi}(\tilde{S}, U)$ on E_n . Second, since $I\{E_n\} \rightarrow 1$ w.p.1 it follows that $\tilde{\phi}(\tilde{S}, U)I\{E_n\} \rightarrow \tilde{\phi}(\tilde{S}, U)$ w.p.1 and $\tilde{\phi}(\tilde{S}_n, U)I\{E_n^c\} \rightarrow 0$ w.p.1. We can therefore use (B-33) and invoke the dominated convergence theorem to conclude that,

$$\begin{aligned} E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)] &= E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)I\{E_n\}] + E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)I\{E_n^c\}] \\ &= E_{\tilde{P}}[\tilde{\phi}(\tilde{S}, U)I\{E_n\}] + E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)I\{E_n^c\}] \\ &\rightarrow E_{\tilde{P}}[\tilde{\phi}(\tilde{S}, U)] . \end{aligned}$$

This completes the proof for the discrete case and the first part of the statement of the theorem.

Part 2.

Let $P_n = \otimes_{i=1}^n P$ with $P \in \mathbf{P}_0$ be given and note that by Theorem 4.1 it follows that

$$S_n = (S_{n,1}, \dots, S_{n,2q}) = (W_{n,[1]}^-, \dots, W_{n,[q]}^-, W_{n,[1]}^+, \dots, W_{n,[q]}^+) \xrightarrow{d} (S_1, \dots, S_{2q}),$$

where (S_1, \dots, S_{2q}) are i.i.d. with cdf $H(w|0)$. The conditions in Assumption 4.4.(i)-(ii) (or Assumption 4.5.(i)-(ii)) immediately follow as $(S_1, \dots, S_{2q}) \stackrel{d}{=} (S_{\pi(1)}, \dots, S_{\pi(2q)})$ for any $\pi \in \mathbf{G}$. Assumption 4.4.(iii) follows the fact that (S_1, \dots, S_{2q}) are i.i.d. with cdf $H(w|0)$, where $H(w|0)$ is the cdf of a continuous random variable by Assumption 4.2.

We are left to prove that the test statistic in (12) satisfies Assumption 4.4.(iv). To show this, note that $T(S)$ as in (12) admits the alternative representation

$$T(S) = \frac{1}{q} T^*(S) - \frac{4q^2 - 1}{12q},$$

where

$$T^*(S) = \frac{1}{q} \sum_{i=1}^q (R_i^* - i)^2 + \frac{1}{q} \sum_{j=1}^q (R_{q+j}^* - j)^2,$$

$R_1^* < R_2^* < \dots < R_q^*$ denote the increasingly ordered ranks R_1, \dots, R_q of the first q variables in S , and $R_{q+1}^* < \dots < R_{2q}^*$ are the increasingly ordered ranks R_{q+1}, \dots, R_{2q} of the last q values in S . It follows immediately that this test statistic satisfies Assumption 4.4.(iv). This completes the proof of the second part of the statement of the theorem. ■

C The multidimensional case

In this appendix we discuss the case where W is a K -dimensional vector. The test statistic in (12) and the test construction in (14) immediately apply to this case where W is a vector consisting of a combination of discrete and continuously distributed random variables. However, in the multidimensional case we also consider an alternative test statistic that may exhibit better power when W includes several components and some are continuous and some are discontinuous at the threshold. We call this test statistic the max test statistic and define it as follows,

$$T_{\max}(S_n) = \max_{c \in \hat{\mathbf{C}}} T(c' S_n), \quad (\text{C-34})$$

where $T(\cdot)$ is the test statistic in (12),

$$c' S_n = (c' S_{n,1}, \dots, c' S_{n,2q}) = (c' W_{n,[1]}^-, \dots, c' W_{n,[q]}^-, c' W_{n,[1]}^+, \dots, c' W_{n,[q]}^+), \quad (\text{C-35})$$

and $\hat{\mathbf{C}}$ is a collection of elements from the unit sphere $\mathbf{C} \equiv \{c \in \mathbf{R}^K : \|c\| = 1\}$. The intuition behind this test statistic arises from observing that the null hypothesis in (3) is equivalent to the same statement applied to any univariate projection $c'W$ of W , i.e.,

$$\Pr\{c'W \leq w | Z = z\} \text{ is continuous in } z \text{ at } z = 0 \text{ for all } w \in \mathbf{R} \text{ and all } c \in \mathbf{C}. \quad (\text{C-36})$$

In the empirical application of Section 6 we choose $\hat{\mathbf{C}}$ to include $100 - K$ i.i.d. draws from $\text{Uniform}(\mathbf{C})$ together with the K canonical elements (i.e., vectors c with zeros in all coordinates except for one). We also set \hat{q}_{rot} to be the minimum value across the rule of thumb across each individual covariate, i.e., $\hat{q}_{\text{rot}} = \min\{\hat{q}_{\text{rot},1}, \dots, \hat{q}_{\text{rot},K}\}$.

Given a test statistic, here we show that the permutation test for this setting is also asymptotically valid. We first state the primitive conditions required to prove this.

Assumption C.1. *For any $\epsilon > 0$, Z satisfies $\Pr\{Z \in (-\epsilon, 0)\} > 0$ and $\Pr\{Z \in [0, \epsilon]\} > 0$.*

Assumption C.2. *The random vector W takes values in \mathbf{R}^{d_w} and has components W_k , for $k \in \{1, \dots, d_w\}$, that satisfy either Assumption 4.2 or Assumption 4.3 with $|W_k| = m_k \in \mathbf{N}$ points of support.*

Assumption C.1 is the same as Assumption 4.1, which is required for Theorem 4.1 to hold. Moreover, Assumption C.2 essentially requires that each component of the vector W satisfies one of the two assumptions we used for the scalar case.

We formalize the high level assumptions required for the validity of the permutation test for the vector case in the following assumption.

Assumption C.3. *If $P \in \mathbf{P}_0$, then*

- (i) $S_n = S_n(X^{(n)}) \xrightarrow{d} S$ under P_n .
- (ii) $S^\pi \stackrel{d}{=} S$ for all $\pi \in \mathbf{G}$.
- (iii) $S_n = (S_{n,1}, \dots, S_{n,2q})$ is such that each $S_{n,j}$, $j \in \{1, \dots, 2q\}$, takes values in \mathbf{R}^{d_w} and has single components $S_{n,j,k}$, $k \in \{1, \dots, d_w\}$, that are either continuously distributed taking values in \mathbf{R} or discretely distributed taking values $S_{n,k} \subseteq S_1 = \bigcup_{\ell=1}^m \{a_\ell\}$ with $a_\ell \in \mathbf{R}$ distinct. In addition, for each component $S_{n,j,k}$, $k \in \{1, \dots, d_w\}$, that is continuously distributed, the corresponding component in $S = (S_1, \dots, S_{2q})$, $S_{j,k}$, is also continuously distributed.
- (iv) $T : \mathcal{S} \rightarrow \mathbf{R}$ is invariant to rank with respect to each continuous component, i.e., it only depends on the order of the elements of each continuous component.

We now formalize our result for the vector case in Theorem C.1, which shows that the permutation test defined in (14) leads to a test that is asymptotically level α whenever Assumption C.3 holds. In addition, the same theorem also shows that Assumption C.1-C.2 are sufficient primitive conditions for the asymptotic validity of our test.

Theorem C.1. *Suppose that Assumption C.3 holds and let $\alpha \in (0, 1)$. Then, $\phi(S_n)$ defined in (14) satisfies*

$$E_P[\phi(S_n)] \rightarrow \alpha \tag{C-37}$$

as $n \rightarrow \infty$ whenever $P \in \mathbf{P}_0$. Moreover, if $T : \mathcal{S} \rightarrow \mathbf{R}$ is the Cramér Von Mises test statistic in (12) and Assumptions C.1-C.2 hold, then Assumption C.3 also holds and (C-37) follows.

C.1 Proof of Theorem C.1

Part 1.

Let $P_n = \otimes_{i=1}^n P$ with $P \in \mathbf{P}_0$ be given. By Assumption C.3(i) and the Almost Sure Representation Theorem (see van der Vaart, 1998, Theorem 2.19), there exists \tilde{S}_n , \tilde{S} , and $U \sim U(0,1)$, defined on a common probability space $(\Omega, \mathcal{A}, \tilde{P})$, such that

$$\tilde{S}_n \rightarrow \tilde{S} \text{ w.p.1 ,}$$

$\tilde{S}_n \stackrel{d}{=} S_n$, $\tilde{S} \stackrel{d}{=} S$, and $U \perp (\tilde{S}_n, \tilde{S})$. Consider the permutation test based on \tilde{S}_n , this is,

$$\tilde{\phi}(\tilde{S}_n, U) \equiv \begin{cases} 1 & T(\tilde{S}_n) > T^{(k)}(\tilde{S}_n) \text{ or } T(\tilde{S}_n) = T^{(k)}(\tilde{S}_n) \text{ and } U < a(\tilde{S}_n) \\ 0 & T(\tilde{S}_n) < T^{(k)}(\tilde{S}_n) \end{cases} .$$

Denote the randomization test based on \tilde{S} by $\tilde{\phi}(\tilde{S}, U)$, where the same uniform variable U is used in $\tilde{\phi}(\tilde{S}_n, U)$ and $\tilde{\phi}(\tilde{S}, U)$.

Since $\tilde{S}_n \stackrel{d}{=} S_n$, it follows immediately that $E_{P_n}[\phi(S_n)] = E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)]$. In addition, since $\tilde{S} \stackrel{d}{=} S$, Assumption C.3(ii) implies that $E_{\tilde{P}}[\tilde{\phi}(\tilde{S}, U)] = \alpha$ by the usual arguments behind randomization tests, see Lehmann and Romano (2005, Chapter 15). It therefore suffices to show

$$E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)] \rightarrow E_{\tilde{P}}[\tilde{\phi}(\tilde{S}, U)] . \quad (\text{C-38})$$

Before we show (C-38), we introduce the additional notation to easily refer to the different components of the vectors S_j and $S_{n,j}$ for $j \in \{1, \dots, 2q\}$. Let the first K^c elements of S_j and $S_{n,j}$ denote the continuous components, where each component is denoted by $S_{j,k}^c$ and $S_{n,j,k}^c$ for $1 \leq k \leq K^c$. Let the remaining subvector S_j^d and $S_{n,j}^d$ of dimension $K^d = K - K^c$ denote the discrete component of S_j and $S_{n,j}$. Arguing as in the proof of Theorem 4.2, it follows that S_j^d has support in (a possible subset of) the same finite collection of points that $S_{n,j}^d$ may take values. For simplicity here and wlog, denote by (s_1^*, \dots, s_L^*) the common points of support of S_j^d and $S_{n,j}^d$. Using this notation, we can partition S_j and $S_{n,j}$ as (S_j^c, S_j^d) and $(S_{n,j}^c, S_{n,j}^d)$, respectively. A similar partition applies to \tilde{S}_j and $\tilde{S}_{n,j}$.

In order to show (C-38), let E_n be the event where the following holds. First, the ordered values of each continuous component $\{\tilde{S}_{j,k}^c : 1 \leq j \leq 2q\}$ and $\{\tilde{S}_{n,j,k}^c : 1 \leq j \leq 2q\}$ correspond to the same permutation π_k of $\{1, \dots, 2q\}$ for $1 \leq k \leq K^c$, i.e., if $S_{k,\pi_k(j)}^c = S_{k,(l)}^c$ then $S_{n,k,\pi_k(j)}^c = S_{n,k,(l)}^c$ for $1 \leq j, l \leq 2q$ and $1 \leq k \leq K^c$. Second, the discrete subvectors $\{\tilde{S}_j^d : 1 \leq j \leq 2q\}$ and $\{\tilde{S}_{n,j}^d : 1 \leq j \leq 2q\}$ coincide, i.e., $\tilde{S}_j^d = \tilde{S}_{n,j}^d$ for $1 \leq j \leq 2q$.

We first claim that $I\{E_n\} \rightarrow 1$ w.p.1. To see this, note that Assumption C.3(iii) and $\tilde{S} \stackrel{d}{=} S$ imply that for all ω in a set with probability one under \tilde{P} we have for each continuous component k of S that

$$\tilde{S}_{k,(1)}^c(\omega) < \tilde{S}_{k,(2)}^c(\omega) < \dots < \tilde{S}_{k,(2q)}^c(\omega) , \quad (\text{C-39})$$

and for the discrete subvector of \tilde{S} that

$$\tilde{S}_j^d(\omega) = s_l^* , \quad (\text{C-40})$$

for $1 \leq j \leq 2q$ and some $1 \leq l \leq L$. Moreover, since $\tilde{S}_n \rightarrow \tilde{S}$ w.p.1, there exists a set Ω^* with $\tilde{P}\{\Omega^*\} = 1$ such that (C-39), (C-40) and $\tilde{S}_n(\omega) \rightarrow \tilde{S}(\omega)$ hold for all $\omega \in \Omega^*$. For all ω in this set, let $\pi_k(1, \omega), \dots, \pi_k(2q, \omega)$

be the permutation that delivers the order statistics in (C-39) for the k^{th} continuous component. It follows that for any $\omega \in \Omega^*$ and any $j \in \{1, \dots, 2q-1\}$, if for any continuous component k we have $\tilde{S}_{k, \pi_k(j, \omega)}^c(\omega) < \tilde{S}_{k, \pi_k(j+1, \omega)}^c(\omega)$ then

$$\tilde{S}_{n, k, \pi_k(j, \omega)}^c < \tilde{S}_{n, k, \pi_k(j+1, \omega)}^c \text{ for } n \text{ sufficiently large } , \quad (\text{C-41})$$

and moreover, if for the discrete subvector we have $\tilde{S}_j^d(\omega) = s_l^*$ then

$$\tilde{S}_{n, j}^d(\omega) = s_l^* \text{ for } n \text{ sufficiently large } , \quad (\text{C-42})$$

which follows from the fact that both $\{\tilde{S}_j^d : 1 \leq j \leq 2q\}$ and $\{\tilde{S}_{n, j}^d : 1 \leq j \leq 2q\}$ are discretely distributed with common support points (s_1^*, \dots, s_L^*) . We can therefore conclude that

$$I\{E_n\} \rightarrow 1 \text{ w.p.1 } ,$$

which proves the first claim.

We now prove (C-38) in two steps. First, we note that

$$E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)I\{E_n\}] = E_{\tilde{P}}[\tilde{\phi}(\tilde{S}, U)I\{E_n\}] . \quad (\text{C-43})$$

This is true because, on the event E_n , the following two hold. First, for each continuous component the rank statistics in (19) of the vectors $\tilde{S}_{n, k}^{c, \pi}$ and $\tilde{S}_k^{c, \pi}$ coincide for $1 \leq k \leq K^c$ and for all $\pi \in \mathbf{G}$. Then we have by Assumption C.3(iv) that the test statistic $T(S)$ only depends on the order of the elements of each continuous component. Second, the discrete subvectors $\tilde{S}_n^{d, \pi}$ and $\tilde{S}^{d, \pi}$ coincide for all $\pi \in \mathbf{G}$. These two properties in turn result in, on the event E_n , $T(\tilde{S}_n^\pi)$ equaling $T(\tilde{S}^\pi)$ for all $\pi \in \mathbf{G}$, which leads to $\tilde{\phi}(\tilde{S}_n, U) = \tilde{\phi}(\tilde{S}, U)$ on E_n .

Then for the second step in proving (C-38), since $I\{E_n\} \rightarrow 1$ w.p.1 it follows that $\tilde{\phi}(\tilde{S}, U)I\{E_n\} \rightarrow \tilde{\phi}(\tilde{S}, U)$ w.p.1 and $\tilde{\phi}(\tilde{S}_n, U)I\{E_n^c\} \rightarrow 0$ w.p.1. We can therefore use (C-43) and invoke the dominated convergence theorem to conclude that,

$$\begin{aligned} E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)] &= E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)I\{E_n\}] + E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)I\{E_n^c\}] \\ &= E_{\tilde{P}}[\tilde{\phi}(\tilde{S}, U)I\{E_n\}] + E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)I\{E_n^c\}] \\ &\rightarrow E_{\tilde{P}}[\tilde{\phi}(\tilde{S}, U)] . \end{aligned}$$

This completes the proof of the first part of the statement of the theorem.

Part 2.

Let $P_n = \otimes_{i=1}^n P$ with $P \in \mathbf{P}_0$ be given and note that by Theorem 4.1 it follows that

$$\begin{aligned} S_n &= (S_{n,1}, \dots, S_{n,2q}) = (W_{n,[1]}^-, \dots, W_{n,[q]}^-, W_{n,[1]}^+, \dots, W_{n,[q]}^+) \\ &\xrightarrow{d} (S_1, \dots, S_{2q}) , \end{aligned}$$

where (S_1, \dots, S_{2q}) are i.i.d. with cdf $H(w|0)$. The conditions in Assumption C.3.(i)-(ii) immediately follow as $(S_1, \dots, S_{2q}) \stackrel{d}{=} (S_{\pi(1)}, \dots, S_{\pi(2q)})$ for any $\pi \in \mathbf{G}$. Assumption C.3.(iii) also follows immediately by Assumption C.2. Finally, to show Assumption C.3.(iv) we first demonstrate that the test statistic in (12)

admits an alternate representation. By Assumption C.2, let without loss of generality the first K^c components be continuous and the rest be discrete. Denote by S_i^d the discrete subvector of S_i and by

$$R_{i,k} = \sum_{j=1}^{2q} I\{S_{j,k}^c \leq S_{i,k}^c\},$$

the rank of the k^{th} continuous component of S_i for $1 \leq i \leq 2q$ and $1 \leq k \leq K^c$. Finally, the test statistic can be rewritten in the following alternate representation

$$T(S) = \frac{1}{2q} \sum_{j=1}^{2q} \left(\frac{1}{q} \sum_{i=1}^q \left[I\{S_i^d \leq S_j^d\} \prod_{k=1}^{K^c} 1\{R_{i,k} \leq R_{j,k}\} \right] - \frac{1}{q} \sum_{i=q+1}^{2q} \left[I\{S_i^d \leq S_j^d\} \prod_{k=1}^{K^c} \{R_{i,k} \leq R_{j,k}\} \right] \right)^2.$$

The above representation follows from first rewriting

$$I\{S_i \leq S_j\} = I\{S_i^d \leq S_j^d\} \prod_{k=1}^{K^c} I\{S_{i,k}^c \leq S_{j,k}^c\},$$

and then noticing that for $1 \leq k \leq K^c$

$$I\{S_{i,k}^c \leq S_{j,k}^c\} = I\{R_{i,k} \leq R_{j,k}\}.$$

This representation illustrates that for the continuous components the test statistic only depends on their individual orderings. It then follows immediately that this test statistic satisfies Assumption C.3.(iv). This completes the proof of the second part of the statement of the theorem. ■

D Additional details on the simulations

In this appendix we document some computational details on the simulations of Section 5. The `Matlab` codes to replicate all our results are available online and include a discussion on the details mentioned here.

Details on \hat{q}_{rot} . The feasible rule of thumb for q is computed (in our simulations and as an option in the companion `Stata` and `R` packages) as follows:

$$\hat{q}_{\text{rot}} = \left\lceil \max \left\{ \min \left\{ \hat{f}_n(0) \hat{\sigma}_{Z,n} (1 - \hat{\rho}_n^2)^{1/2} \frac{n^{0.9}}{\log n}, q_{UB} \right\}, q_{LB} \right\} \right\rceil,$$

where q_{LB} and q_{UB} are a lower and upper bounds, respectively. We set $q_{LB} = 10$, as less than 10 observations leads to tests where the randomized and non-randomized versions of the permutation test differ. We then set $q_{UB} = \frac{n^{0.9}}{\log n}$, as $\frac{n}{\log n}$ is the rate that violates the conditions we require for q in the proof of Theorem 4.1. The estimator $\hat{f}_n(0)$ of $f(0)$ is a kernel estimator with a triangular kernel and a bandwidth h computed using Silverman's rule of thumb. The estimators $\hat{\rho}_n$ and $\hat{\sigma}_{Z,n}^2$ are the sample correlation between W_i and Z_i and sample variance of Z_i . For additional details on the `R` implementation, see [Oliveres-González and Sarmiento-Barbieri \(2017\)](#).

Details on SZ bandwidth. Shen and Zhang (2016) propose the rule of thumb bandwidth in (23), where h_n^{CCT} is a two step bandwidth estimate based on Calonico et al. (2014). In the first step, a pilot bandwidth is selected using CCT for estimating the average treatment effect at the cutoff. Note that this is the same bandwidth used in the CCT test. Then, in the second step, CCT is used again with the dependent variables as $I\{W_i \leq \tilde{w}\}$, where \tilde{w} corresponds to the minimum amongst the values that attain the maximum estimated distributional treatment effect. In our simulations, however, this results in no variation in the dependent variable in some models, which leads to the termination of the program. In such cases when there is no variation, for example Model 6, we first take \tilde{w} to be the estimated median value of W_i using the whole sample of data. If this additionally fails, we take h_n^{CCT} to be the pilot bandwidth. Shen and Zhang (2016) additionally propose an alternative rule of thumb based on the bandwidth proposed by Imbens and Kalyanaraman (2012), and find similar results. We hence do not include results of this alternative choice in our comparisons, but the results are available upon request.

We observe that these bandwidths have practical difficulties in some model, where one faces matrix inversion issues. In Model 4, the program terminates for either bandwidths, which we believe is due to the discreteness of the running variable. To deal with this, we impose the bandwidth to take a minimum value of 0.125 for the SZ test and a minimum value of 0.175 for the CCT test. Note that the average bandwidth in Model 4 (across simulations) is well over this lower bound across all undersmoothing parameters and sample sizes. In Model 5 and Model 6, we observe similar matrix inversion issues for the bandwidths based on IK, but the program does not terminate. In this case we hence do not make any adjustments.

E Surveyed papers on RDD

Table 5 displays the list of papers we surveyed in leading journals that use regression discontinuity designs. We specifically note whether these papers test for any of the two implications we mention in the introduction, namely, validating the continuity of the density of the running variable and validating the continuity of the means of the baseline covariates.

We briefly describe the criteria used to prepare our list. The journals selected were the *American Economic Review* (AER), the *American Economic Journal: Applied Economics* (AEJ:AppEcon), the *Quarterly Journal of Economics* (QJE), and the *Review of Economics and Statistics* (ReStat), and the years used were from the beginning of 2011 to the end of 2015. All papers in each volumes were surveyed with the exception of the May volume for AER. We first categorized papers using regression discontinuity methods by searching the main text for the keywords ‘regression discontinuity’. We then individually inspected the papers along with their appendices for whether they validated their design, and, if so, by either checking the continuity of the density of the running variable or the continuity of the means of the baseline covariates, or both. We allowed for both formal test results as well as informal graphical evidence.

We find that out of the 62 papers that use regression discontinuity methods, 35 validate by checking the continuity of the density, 42 validate by checking continuity of the baseline covariates, 34 validate using both tests, and 19 do not include any form of manipulation or placebo test.

Authors (Year)	Journal	Density Test	Mean Test	Authors (Year)	Journal	Density Test	Mean Test
Schmieder et al. (2016)	AER	✓	✓	Miller et al. (2013)	AEJ:AppEcon	✓	✓
Feldman et al. (2016)	AER	✓	✓	Litschig and Morrison (2013)	AEJ:AppEcon	✓	✓
Jayaraman et al. (2016)	AER	×	×	Dobbie and Skiba (2013)	AEJ:AppEcon	✓	✓
Dell (2015)	AER	✓	✓	Kazianga et al. (2013)	AEJ:AppEcon	✓	✓
Hansen (2015)	AER	✓	✓	Magruder (2012)	AEJ:AppEcon	×	×
Anderson (2014)	AER	×	×	Dustmann and Schmberg (2012)	AEJ:AppEcon	×	×
Martin et al. (2014)	AER	×	×	Clots-Figueras (2012)	AEJ:AppEcon	✓	✓
Dahl et al. (2014)	AER	✓	✓	Manacorda et al. (2011)	AEJ:AppEcon	✓	✓
Shigeoka (2014)	AER	×	✓	Chetty et al. (2014)	QJE	✓	✓
Crost et al. (2014)	AER	×	✓	Michalopoulos and Papaioannou (2014)	QJE	×	✓
Kostol and Mogstad. (2014)	AER	✓	✓	Fredriksson et al. (2013)	QJE	✓	✓
Clark and Royer (2013)	AER	×	✓	Schmieder et al. (2012)	QJE	✓	✓
Brollo et al. (2013)	AER	✓	✓	Lee and Mas (2012)	QJE	×	×
Bharadwaj et al. (2013)	AER	✓	✓	Saez et al. (2012)	QJE	×	×
Pop-Eleches and Urquiola (2013)	AER	✓	✓	Barreca et al. (2011)	QJE	×	×
Lacetera et al. (2012)	AER	✓	×	Almond et al. (2011)	QJE	✓	✓
Dufo et al. (2012)	AER	×	×	Malamud and Pop-Eleches (2011)	QJE	✓	✓
Gopinath et al. (2011)	AER	✓	✓	Fulford (2015)	ReStat	×	✓
Auffhammer and Kellogg (2011)	AER	×	×	Snider and Williams (2015)	ReStat	×	×
Dufo et al. (2011)	AER	×	×	Doleac and Sanders (2015)	ReStat	×	×
Ferraz and Finan (2011)	AER	×	×	Coşar et al. (2015)	ReStat	×	×
McCrary and Royer (2011)	AER	×	✓	Avery and Brevoort (2015)	ReStat	×	×
Beland (2015)	AEJ:AppEcon	✓	✓	Carpenter and Dobkin (2015)	ReStat	×	✓
Buser (2015)	AEJ:AppEcon	✓	✓	Black et al. (2014)	ReStat	✓	✓
Fack and Grenet (2015)	AEJ:AppEcon	✓	✓	Anderson et al. (2014)	ReStat	×	×
Cohodes and Goodman (2014)	AEJ:AppEcon	✓	✓	Alix-Garcia et al. (2013)	ReStat	×	✓
Haggag and Paci (2014)	AEJ:AppEcon	✓	✓	Albouy (2013)	ReStat	×	×
Dobbie and Fryer (2014)	AEJ:AppEcon	✓	✓	Garibaldi et al. (2012)	ReStat	✓	✓
Sekhri (2014)	AEJ:AppEcon	✓	✓	Manacorda (2012)	ReStat	✓	✓
Schumann (2014)	AEJ:AppEcon	✓	✓	Martorell and McFarlin (2011)	ReStat	✓	✓
Lucas and Mbiti (2014)	AEJ:AppEcon	✓	✓	Grosjean and Senik (2011)	ReStat	×	×

Table 5: Papers using manipulation/placebo tests from 2011 – 2015.

References

- ALMOND, D., DOYLE JR, J. J., KOWALSKI, A. and WILLIAMS, H. (2010). Estimating marginal returns to medical care: Evidence from at-risk newborns. *The Quarterly Journal of Economics*, **125** 591–634.
- BHATTACHARYA, P. (1974). Convergence of sample paths of normalized sums of induced order statistics. *The Annals of Statistics* 1034–1039.
- BRUHN, M. and MCKENZIE, D. (2008). In pursuit of balance: Randomization in practice in development field experiments. *World Bank Policy Research Working Paper*, **4752**.
- BUGNI, F. A., CANAY, I. A. and SHAIKH, A. M. (2017). Inference under covariate-adaptive randomization. *Journal of the American Statistical Association*, doi:10.1080/01621459.2017.1375934.
- CALONICO, S., CATTANEO, M. D. and TITIUNIK, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, **82**.
- CANAY, I. A., ROMANO, J. P. and SHAIKH, A. M. (2017). Randomization tests under an approximate symmetry assumption. *Econometrica*, **85** 1013–1030.

- CATTANEO, M. D., FRANDBEN, B. R. and TITIUNIK, R. (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the us senate. *Journal of Causal Inference*, **3** 1 – 24.
- CHUNG, E. and ROMANO, J. P. (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics*, **41** 484–507.
- DAVID, H. and GALAMBOS, J. (1974). The asymptotic theory of concomitants of order statistics. *Journal of Applied Probability* 762–770.
- GANONG, P. and JÄGER, S. (2015). A permutation test for the regression kink design. Tech. rep., Working paper.
- GERARD, F., ROKKANEN, M. and ROTHE, C. (2016). Bounds on treatment effects in regression discontinuity designs with a manipulated running variable, with an application to unemployment insurance in brazil. Working paper.
- HAHN, J., TODD, P. and KLAUW, W. V. D. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, **69** pp. 201–209. URL <http://www.jstor.org/stable/2692190>.
- HAJEK, J., SIDAK, Z. and SEN, P. K. (1999). *Theory of rank tests*. 2nd ed. Academic press.
- HOEFFDING, W. (1952). The large-sample power of tests based on permutations of observations. *The Annals of Mathematical Statistics*, **23** pp. 169–192. URL <http://www.jstor.org/stable/2236445>.
- IMBENS, G. and KALYANARAMAN, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, **79** 933–959.
- IMBENS, G. W. and LEMIEUX, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of econometrics*, **142** 615–635.
- KAMAT, V. (2017). On nonparametric inference in the regression discontinuity design. *Econometric Theory*, doi:10.1017/S0266466617000196 1–10.
- LEE, D. S. (2008). Randomized experiments from non-random selection in u.s. house elections. *Journal of Econometrics*, **142** 675 – 697. The regression discontinuity design: Theory and applications, URL <http://www.sciencedirect.com/science/article/pii/S0304407607001121>.
- LEE, D. S. and LEMIEUX, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, **48** 281–355.
- LEHMANN, E. and ROMANO, J. P. (2005). *Testing Statistical Hypotheses*. 3rd ed. Springer, New York.

- MCCRARY, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, **142** 698 – 714. The regression discontinuity design: Theory and applications, URL <http://www.sciencedirect.com/science/article/pii/S0304407607001133>.
- OLIVARES-GONZÁLEZ, M. and SARMIENTO-BARBIERI, I. (2017). RATest: An R package for randomization tests with an application to testing the continuity of the baseline covariates in RDD using approximate permutation tests.
- ROMANO, J. P. (1989). Bootstrap and randomization tests of some nonparametric hypotheses. *The Annals of Statistics*, **17** 141–159. URL <http://dx.doi.org/10.1214/aos/1176347007>.
- ROMANO, J. P. (1990). On the behavior of randomization tests without a group invariance assumption. *Journal of the American Statistical Association*, **85** pp. 686–692. URL <http://www.jstor.org/stable/2290003>.
- SALES, A. and HANSEN, B. B. (2015). Limitless regression discontinuity. *arXiv preprint arXiv:1403.5478*.
- SEKHON, J. S. and TITIUNIK, R. (2016). On interpreting the regression discontinuity design as a local experiment. Manuscript.
- SHEN, S. and ZHANG, X. (2016). Distributional tests for regression discontinuity: Theory and empirical examples. *Review of Economics and Statistics*. Forthcoming.
- VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.