Kitagawa, Toru

Working Paper

A test for instrument validity

cemmap working paper, No. CWP34/14

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Kitagawa, Toru (2014) : A test for instrument validity, cemmap working paper, No. CWP34/14, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2014.3414

This Version is available at:
https://hdl.handle.net/10419/111380

cemmap
centre for microdata methods and practice

A Test for Instrument Validity

Toru Kitagawa

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP34/14

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

A Test for Instrument Validity*

Toru Kitagawa†

CeMMAP and *Department of Economics, UCL*

This draft: August, 2014

Abstract

This paper develops a specification test for instrument validity in the heterogeneous treatment effect model with a binary treatment and a discrete instrument. The strongest testable implication for instrument validity is given by the condition for non-negativity of point-identifiable complier's outcome densities. Our specification test infers this testable implication using a variance-weighted Kolmogorov-Smirnov test statistic. Implementation of the proposed test does not require smoothing parameters, even though the testable implications involve non-parametric densities. The test can be applied to both discrete and continuous outcome cases, and an extension of the test to settings with conditioning covariates is provided.

Keywords: Treatment Effects, Instrumental Variable, Specification Test, Bootstrap.
JEL Classification: C12, C15, C21.

*This paper is a revised version of a chapter of my Ph.D thesis submitted to Brown University in 2009. This paper replaces the previous versions titled as "A Bootstrap Test for Instrument Validity in the Heterogeneous Treatment Effect Models."

†Email: t.kitagawa@ucl.ac.uk. I would like to thank Frank Kleibergen for guidance and continuous encouragement. I would also like to thank Josh Angrist, Tim Armstrong, Clément de Chaisemartin, Le-Yu Chen, Mario Fiorini, James Heckman, Stefan Hoderlein, Yu-Chin Hsu, Martin Huber, Hide Ichimura, Guido Imbens, Giovanni Mellace, and Katrine Stevens, and the seminar participants at Academia Sinica, GRIPS, Penn State, SFU, and UCL, for helpful comments and beneficial discussions. I also thank a co-editor and three anonymous referees for valuable suggestions that significantly improved the paper. Financial support from the ESRC through the ESRC Center for Microdata Methods and Practice (CEMMAP) (grant number RES-589-28-0001) and the Merit Dissertation Fellowship from the Graduate School of Economics in Brown University are gratefully acknowledged.

1 Introduction

Consider a heterogeneous causal effect model of Angrist and Imbens (1994) with a binary treatment and a binary instrument. We denote an observed outcome by $Y \in \mathcal{Y} \subset \mathbb{R}$, an observed treatment status by $D \in \{1,0\}$; $D = 1$ when one receives the treatment while $D = 0$ when one does not, and a binary non-degenerate instrument by $Z \in \{1,0\}$. Let $\{Y_{dz} \in \mathcal{Y} : d \in \{1,0\}, z \in \{1,0\}\}$ be the potential outcomes that would have been observed if the treatment status were set at $D = d$ and the assigned instrument were set at $Z = z$. Furthermore, $\{D_z : z \in \{1,0\}\}$ are the potential treatment responses that would have been observed if $Z = 1$ and $Z = 0$, respectively. The seminal works of Imbens and Angrist (1994) and Angrist, Imbens and Rubin (1996) showed that, given $\Pr(D = 1|Z = 1) > \Pr(D = 1|Z = 0)$, the instrument variable Z that satisfies the three conditions involving the potential variables is able to identify the average treatment effects for those whose selection to treatment is affected by the instrument (local average treatment effect, LATE hereafter). The three key conditions, of which the joint validity is hereafter referred to as *IV-validity,* are[1]

Assumption: IV-validity for binary Z

1. *Instrument Exclusion*: $Y_{d1} = Y_{d0}$ for $d = 1, 0$, with probability one.
2. *Random Assignment:* Z is jointly independent of $(Y_{11}, Y_{10}, Y_{01}, Y_{00}, D_1, D_0)$.
3. *Instrument Monotonicity (No-defier):* The potential participation indicators satisfy $D_1 \geq D_0$ *with probability one.*

Despite the fact that the credibility of LATE analysis relies on the validity of the employed instrument, no test procedure has been proposed to empirically diagnose IV-validity. As a result, causal inference studies have assumed IV-validity based solely on some background knowledge or out-of-sample evidence, and, accordingly, its credibility often remains controversial in many empirical contexts.

The main contribution of this paper is to develop a specification test for IV-validity in the LATE model. Our specification test builds on the testable implication obtained by Balke

[1]Note that the null hypothesis of IV-validity tested in this paper does not include the *instrument relevance* assumption, $\Pr(D = 1|Z = 1) > \Pr(D = 1|Z = 0)$. The instrument relevance assumption can be assessed by inferring the coefficient in the first-stage regression of D onto Z.

and Pearl (1997) and Heckman and Vytlacil (2005, Proposition A.5). Let P and Q be the conditional probability distributions of $(Y, D) \in \mathcal{Y} \times \{1, 0\}$ given $Z = 1$ and $Z = 0$, i.e.,

$$\begin{aligned} P(B, d) &= \Pr(Y \in B, D = d | Z = 1), \\ Q(B, d) &= \Pr(Y \in B, D = d | Z = 0), \end{aligned}$$

for Borel set B in $\mathcal{Y}$ and $d = 1, 0$. Imbens and Rubin (1997) showed that, under IV-validity,

$$\begin{aligned} P(B, 1) - Q(B, 1) &= \Pr(Y_1 \in B, D_1 > D_0) \text{ and} \\ Q(B, 0) - P(B, 0) &= \Pr(Y_0 \in B, D_1 > D_0) \end{aligned}$$

hold for every B in $\mathcal{Y}$. Since the quantities in the right-hand sides are nonnegative by the definition of probabilities, we obtain the testable implication of Balke and Pearl (1997) and Heckman and Vytlacil (2005);

$$\begin{aligned} P(B, 1) - Q(B, 1) &\geq 0, \\ Q(B, 0) - P(B, 0) &\geq 0, \end{aligned} \tag{1.1}$$

for every Borel set B in $\mathcal{Y}$.[2] Figures 1 and 2 provide visual illustration of these testable implications for a continuous Y case. The solid lines, $p(y, d)$ and $q(y, d)$, each show the probability density of $P(\cdot, d)$ and $Q(\cdot, d)$ that is identifiable by data. If the instrument is valid, $p(y, 1)$ must nest $q(y, 1)$ for treatment outcome, and $q(y, 0)$ must nest $p(y, 0)$ for control outcome, as plotted in Figure 1.

In contrast, if we observed the densities as plotted in Figure 2, we could refute at least one of the IV-validity assumptions since some of the inequalities (1.1) are violated at some subsets in the support of Y, e.g., those labeled as V_1 and V_2 in Figure 2.

To see how densities of $p(y, d)$ and $q(y, d)$ look like in real data, Figure 3 plots kernel density estimates of $p(y, d)$ and $q(y, d)$ for the draft lottery data of Angrist (1991), where the logarithm of one's post-war weekly earning is Y, the veteran status is D, and the draft eligibility determined by the lottery is Z. The estimated densities overall exhibit the nesting relationships similar to those illustrated in Figure 1; therefore, no strong evidence against IV-validity appears to be available. Contrasting density plots are shown in Figure 4, where

[2] As is clear from the derivation, the testable implication can be equivalently interpreted as the nonnegativity conditions for the complier's potential outcome distributions, $\Pr(Y_d \in B | D_1 > D_0) \geq 0$, which are identifiable under IV-validity. Imbens and Rubin (1997) noted that, depending on data, the estimates of the complier's outcome densities can be negative over some region in the outcome support.

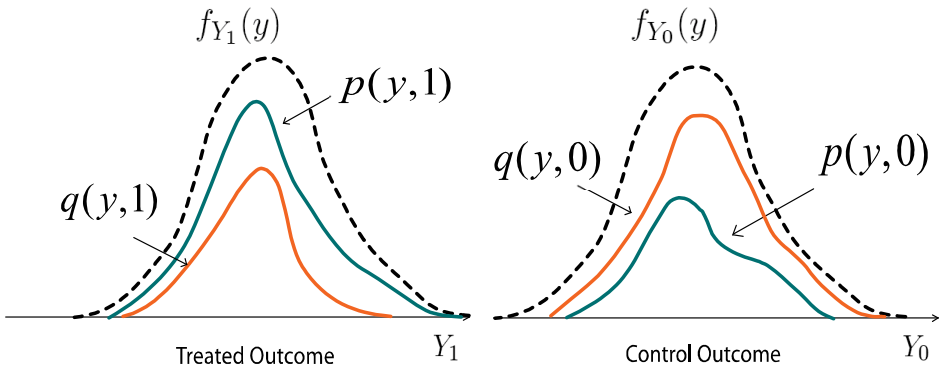

Figure 1: *When the identifiable densities* $p(y, D = d)$ *and* $q(y, D = d)$ *are nested as in this figure, IV-validity cannot be refuted. The dotted lines show the marginal probability densities of the potential outcomes, i.e.,* $f_{Y_d}(y)$ *is the marginal probability density of* $Y_d \equiv Y_{d1} = Y_{d0}$*, which is not identifiable. Under the instrument exclusion and random assignment, both* $p(y, d)$ *and* $q(y, d)$ *must lie below the potential outcome densities* $f_{Y_d}(\cdot)$*.*

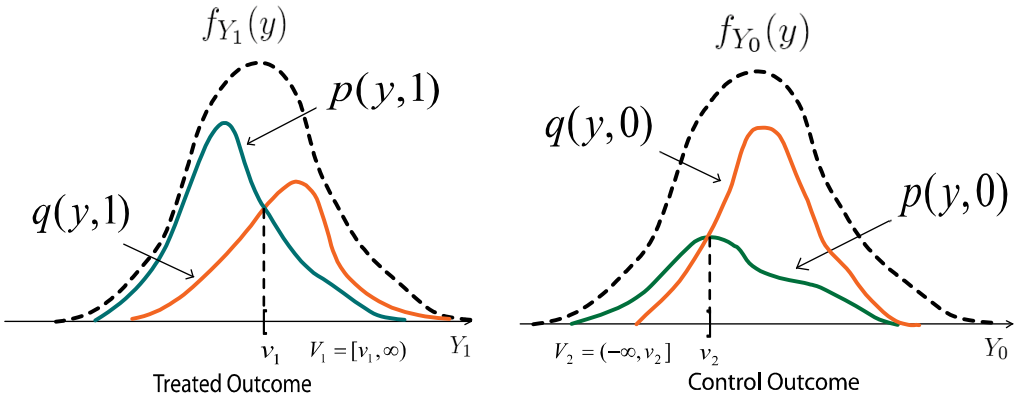

Figure 2: *When we observe* $p(y, d)$ *intersects with* $q(y, d)$ *for at least one of* $d = 1, 0$*, we can refute IV-validity.*

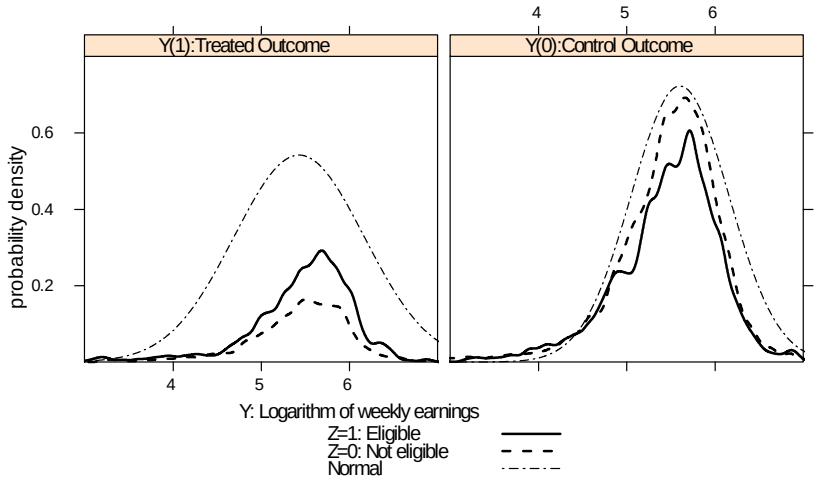

Figure 3: **Kernel Density Estimates for the Draft Lottery Data.** The *Gaussian kernel with bandwidth 0.06 is used. In each panel, we draw a normal density to indicate the scale of the subdensities.*

the data are from Card (1993), Y is one's weekly earning, D indicates whether one graduated from a four-year college, Z indicates whether a four-year college is located in the area of one's residence. No conditioning covariates are controlled for when drawing the densities. Here, we observe that the density estimates intersect, especially for the control outcome. This is an in-sample visual evidence against IV-validity. These eyeball-based assessments are indeed intuitive and useful, but they fail (i) to take into account sampling uncertainty and (ii) to quantify the strength of evidence for or against IV-validity without relying on a specific and often ad hoc choice of smoothing parameters. A hypothesis test procedure proposed in this paper solves these important practical issues.

The above derivation of (1.1) shows only that inequalities (1.1) are *necessary* implications of IV-validity, so it is natural to ask (i) whether the testable implications of (1.1) can be further strengthened and (ii) whether there exist some P and Q for which (1.1) becomes

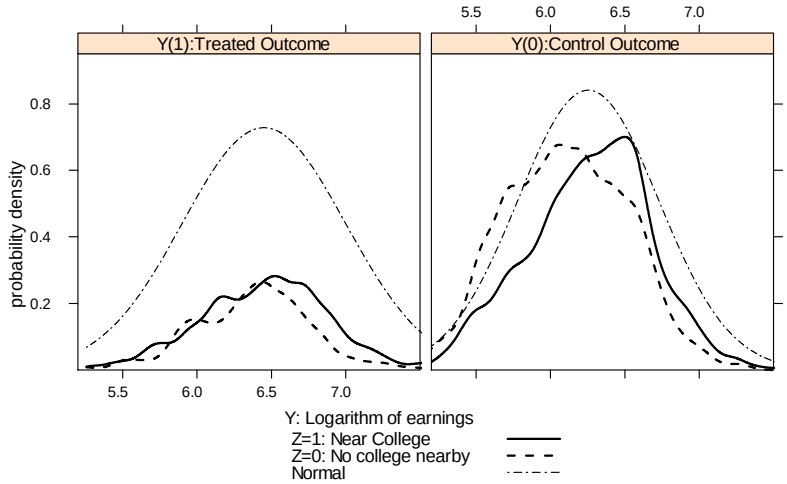

Figure 4: **Kernel Density Estimates, the Proximity to College Data**. *The Gaussian kernel with bandwidth 0.08 is used.*

a necessary and sufficient condition for IV-validity. The next proposition shows that the answers to these questions are negative (see Appendix A of the supplementary material for a proof).

Proposition 1.1 *Assume that $P(\cdot, d)$ and $Q(\cdot, d)$ have a common dominating measure μ on $\mathcal{Y}$. (i) If distributions of observables, P and Q, satisfy inequalities (1.1), then there exists a joint probability law of $(Y_{11}, Y_{10}, Y_{01}, Y_{00}, D_1, D_0, Z)$ that satisfies IV-validity and induces the P and Q.*

(ii) For any P and Q satisfying inequalities (1.1), we can construct a joint probability law of $(Y_{11}, Y_{10}, Y_{01}, Y_{00}, D_1, D_0, Z)$ that violates IV-validity.

The first claim of the proposition shows an optimality of the testable implication, in the sense that any other feature of the data distribution cannot contribute to screening out invalid instruments further than the testable implication of (1.1). The second claim shows that accepting the null hypothesis of (1.1) never allows us to confirm IV-validity regardless of the sample size.[3] In this precise sense, the IV-validity is a *refutable* but *non-verifiable*

[3]de Chaisemartin (2014) shows that the Wald estimand can identify an average causal effect of a well-

hypothesis. Such limitation in terms of learnability of instrument validity is known in other contexts, such as the classical over-identification test in the linear instrumental variable method with homogeneous effect[4] and the test of instrument monotonicity in the multi-valued treatment case proposed in Angrist and Imbens (1995).[5] See Breusch (1986) for a general discussion on hypothesis testing of refutable but non-verifiable assumptions.

Our test uses a variance-weighted Kolmogorov-Smirnov test statistic (KS-statistic, hereafter) to measure the magnitude of violations of inequalities (1.1) in the data. We provide a resampling algorithm to obtain critical values and demonstrate that the test procedure attains asymptotically correct size uniformly over a large class of data generating processes, and consistently rejects all the data generating processes violating (1.1). A similar variance weighted KS-statistic has been considered in the literature of conditional moment inequalities, as in Andrews and Shi (2013), Armstrong (2014), Armstrong and Chan (2013), and Chetverikov (2012). As shown in Romano (1988), bootstrap is widely applicable and easy to implement to obtain the critical values for general KS-statistic, and it has been instrumental in the context of stochastic dominance testing; see, e.g., Abadie (2002), Barrett and Donald (2003), Donald and Hsu (forthcoming), Horváth, Kokoszka, and Zitikis (2006), and Linton, Maasoumi, and Whang (2005).

Our test concerns the exogeneity of instrument defined in terms of statistical independence, and it can be applied to the context, in which objects of interest are distributional features of complier's potential outcome distribution, e.g., the quantile treatment effects for compliers (Abadie, Angrist, and Imbens (2002)). On the other hand, if solely the mean

defined subpopulation under the "Compliers-defiers" assumption, which is weaker than the instrument monotonicity assumption of Imbens and Angrist (1994). He shows that, even when instrument monotonicity is replaced by the Complier-definers condition, the testable implication of (1.1) is a necessary, but is not an if-and-only-if testable implication.

[4]If the instrument is multi-valued, we can naively perform the classical over-identification test by treating the multi-valued instrument as a collection of binary instruments. However, as discussed in Imbens (2014) and Lee (2014), the over-identification test should not be used if causal effects are considered to be heterogeneous, since heterogeneity of causal effects can lead to misspecified over-identifying restrictions, even when LATE IV-validity is true.

[5]In case of multi-valued treatment status, Angrist and Imbens (1995) propose a specification test to assess instrument monotonicity by inferring the stochastic dominance of the distribution functions of the treatment status conditional on the instrument; see Barua and Lang (2009) and Fiorini, Stevens, Taylor, and Edwards (2013) for applications of the Angrist and Imbens test. In the binary treatment case, however, Angrist and Imbens test cannot be applied.

effect is concerned, identification of LATE can in fact be attained under a slightly weaker set of assumptions, such that the instrument is statistically independent of the selection types while the potential outcomes are only mean independent of Z conditional on each selection type. Huber and Mellace (2013) show that this weaker LATE identifying condition has a testable implication given by a finite number of moment inequalities. Since our test builds on the distributional restrictions implied from statistical independence, it screens out a larger class of data generating processes compared to the test of Huber and Mellace. In addition, the set of detectable alternatives and the p-value of our test are invariant to any monotonic transformation of the outcome variables, whereas this invariance property does not hold for the Huber and Mellace's test. Mourifié and Wan (2014) recently consider testing the same instrument validity condition as ours by applying the inference technique developed by Chernozhukov, Lee, and Rosen (2013). Our test differs from theirs at least in the following three aspects. First, validity of their test assumes continuity of the density of Y, which the LATE identification and its consistent estimation do not require, whereas our test does not require such continuity assumption. Second, in implementing their test, one has to specify smoothing parameters, whereas our test is based solely on the empirical distributions so its implementation is free from a choice of smoothing parameters. Third, our test has nontrivial power against a class of nonparametric local alternatives shrinking to null at $N^{-1/2}$-rate, while, as shown in Chernozhukov et al (2013), the approach of Mourifié and Wan cannot attain nontrivial power against any nonparametric $N^{-1/2}$-local alternatives.

The rest of the paper is organized as follows. Section 2 presents implementation of our test when D and Z are binary and shows its asymptotic validity. Section 3 extends the analysis to settings with a multi-valued instrument and with conditioning covariates. Two empirical applications are provided in Section 4. The online supplementary material provides proofs and results of Monte Carlo experiments.

2 Test

2.1 Test Statistics and Implementation

Let a sample be given by N observations of $(Y, D, Z) \in \mathcal{Y} \times \{1, 0\}^2$. We divide the sample into two subsamples based on the value of Z, and we consider the sampling process as being conditional on a sequence of instrument values. Let (Y_i^1, D_i^1), $i = 1, \ldots, m$ be observations

with $Z = 1$ and (Y_j^0, D_j^0), $j = 1, \ldots, n$ be those with $Z = 0$, and assume that the observations of (Y_i^1, D_i^1) and (Y_j^0, D_j^0) are drawn independently and identically from P and Q, respectively. We assume a deterministic sequence $\hat{\lambda} = m/N \to \lambda$ as $N \to \infty$, where λ is bounded away from zero and one.[6] We denote the empirical distributions of P and Q by

$$\begin{aligned} P_m(B, d) &\equiv \frac{1}{m} \sum_{i=1}^{m} I\{Y_i^1 \in B, D_i^1 = d\}, \\ Q_n(B, d) &\equiv \frac{1}{n} \sum_{j=1}^{n} I\{Y_j^0 \in B, D_j^0 = d\}. \end{aligned}$$

To test the null hypothesis given by inequalities (1.1), we consider a variance-weighted KS-statistic,

$$T_N = \left(\frac{mn}{N}\right)^{1/2} \max \left\{ \begin{array}{c} \sup_{-\infty \le y \le y' \le \infty} \left\{ \frac{Q_n([y,y'],1) - P_m([y,y'],1)}{\xi \vee \sigma_{P_m,Q_n}([y,y'],1)} \right\}, \\ \sup_{-\infty \le y \le y' \le \infty} \left\{ \frac{P_m([y,y'],0) - Q_n([y,y'],0)}{\xi \vee \sigma_{P_m,Q_n}([y,y'],0)} \right\} \end{array} \right\}, \tag{2.1}$$

where ξ is a positive constant specified by the user and

$$\sigma^2_{P_m,Q_n}([y, y'], d) = (1-\hat{\lambda}) P_m([y, y'], d)(1 - P_m([y, y'], d)) + \hat{\lambda} Q_n([y, y'], d)(1 - Q_n([y, y'], d)).$$

If the sample counterpart of the first (second) inequality of (1.1) is violated at some interval, then, the first (second) supremum in the max operator becomes positive. For each interval $[y, y']$, $\sigma^2_{P_m,Q_n}([y, y'], d)$ is a consistent estimator of the asymptotic variance of $\left(\frac{mn}{N}\right)^{1/2} (P_m([y, y'], d) - Q_n([y, y'], d))$. Thus, the proposed test statistics quantifies a variance-adjusted maximal violation of the inequalities (1.1) over a class of connected intervals including unbounded ones. The exact suprema can be computed by evaluating the maximand at the finite number of intervals, because, to compute the first (second) supremum in the statistic, it suffices to evaluate the differences of the empirical distribution functions at every interval, the boundaries of which are given by a pair of Y values observed in the subsample of $\{D = 1, Z = 0\}$ ($\{D = 0, Z = 1\}$). The suprema are searched over a smaller class of subsets than the class of Borel sets for which the population inequalities (1.1) are examined. Nevertheless, this reduction of the class of sets does not cause any loss of information, in the sense that any data generating processes that violate (1.1) for at least one Borel set can be screened out asymptotically (Theorem 2.1 (ii) below). Note that the

[6] If one wants to perform the test without conditioning on observations of Z, instruments need be resampled as well in the bootstrap algorithm given below.

proposed test statistic and asymptotic validity of the test are not restricted to a continuous Y case. The same statistic can be used for any ordered discrete Y or a mixture of discrete and continuous Y.[7]

The user-specified trimming constant ξ plays a role in ensuring that the inverse weighting terms are sufficiently away from zero. Note that when $\xi \geq 1/2$, the proposed test statistic is identical up to a constant to the non-weighted KS-statistic,

$$T_{N,nw} = \left(\frac{mn}{N}\right)^{1/2} \max \left\{ \begin{array}{c} \sup_{-\infty \leq y \leq y' \leq \infty} \left\{ Q_n([y,y'],1) - P_m([y,y'],1) \right\}, \\ \sup_{-\infty \leq y \leq y' \leq \infty} \left\{ P_m([y,y'],0) - Q_n([y,y'],0) \right\} \end{array} \right\}. \tag{2.2}$$

Hence, variance-weighting is effective only when ξ is smaller than $1/2$. The Monte Carlo studies presented in Appendix D of the supplementary material show that the test size is insensitive to a choice of ξ even in small sample situations. The finite sample power of the test, on the other hand, can be sensitive to a choice of ξ depending on a specification of alternative. Specifically, when violations of the testable implications occur at the tail parts of P and Q, our Monte Carlo experiments suggest that smaller ξ yields a higher power. In contrast, if violations occur at an interval where P and Q have high probabilities, a larger ξ tends to show a slightly higher power. Although a formal discussion regarding an optimal choice of ξ is out of scope of this paper, our informal recommendation is to specify $\xi = 0.05 \sim 0.1$ in order to avoid a big power loss when violations are occurring at the tail parts of P and Q. Alternatively, reporting the test results with several choices of ξ is also recommended in order to showcase the range of p-values across different choices of ξ.

To obtain asymptotically valid critical values for the test, we focus on a data generating processes on the boundary of the one-sided null hypothesis, such that P and Q are identical to some probability measure H. Specifically, we set H at the pooled probability measure (unconditional distribution of (Y, D)),

$$H = \lambda P + (1 - \lambda) Q, \tag{2.3}$$

and aim to estimate the quantiles of the null distribution of the statistic as if the data were generated from $P = Q = H$.[8]

We now summarize a bootstrap algorithm for obtaining critical values for T_N.

[7] A similar test statistic can be defined also for unordered discrete Y and multi-dimensional Y. In case of unordered discrete Y, the supremum can be defined over every support point of Y, and in case of multi-dimensional Y, the supremum can be defined over a class of rectangles in the support of Y.

[8] The finite sample power may be improved if critical values are obtained from the null distribution of

Algorithm 2.1:

1. *Sample* (Y_i^*, D_i^*), $i = 1, \ldots, m$ *randomly with replacement from* $H_N = \hat{\lambda} P_m + \left(1 - \hat{\lambda}\right) Q_n$ *and construct empirical distribution* P_m^*. *Similarly, sample* (Y_j^*, D_j^*), $j = 1, \ldots, n$ *randomly with replacement from* H_N *and construct empirical distribution* Q_n^*.

2. *Calculate a bootstrap realization of test statistic*[9]

$$T_N^* = \left(\frac{mn}{N}\right)^{1/2} \max \left\{ \begin{array}{c} \sup_{-\infty \le y \le y' \le \infty} \left\{ \frac{Q_n^*([y,y'],1) - P_m^*([y,y'],1)}{\xi \vee \sigma_{P_m^*, Q_n^*}([y,y'],1)} \right\}, \\ \sup_{-\infty \le y \le y' \le \infty} \left\{ \frac{P_m^*([y,y'],0) - Q_n^*([y,y'],0)}{\xi \vee \sigma_{P_m^*, Q_n^*}([y,y'],0)} \right\} \end{array} \right\},$$

   where $\sigma^2_{P_m^*, Q_n^*}([y,y'],d) = (1-\hat{\lambda}) P_m^*([y,y'],d)(1 - P_m^*([y,y'],d)) + \hat{\lambda} Q_n^*([y,y'],d)(1 - Q_n^*([y,y'],d))$.

3. *Iterate Step 1 - 3 many times and get the empirical distribution of* T_N^*. *For a chosen nominal level* $\alpha \in (0, 1/2)$, *we obtain a bootstrap critical value* $c_{N,1-\alpha}$ *from its empirical* $(1-\alpha)$*-th quantile .*

4. *Reject the null hypothesis (1.1) if* $T_N > c_{N,1-\alpha}$. *The bootstrap p-value is obtained according to the proportion of bootstrap repetitions such that* T_N^* exceeds T_N.

2.2 Asymptotic Uniform Size Control and Consistency

This section formally claims that the test procedure of Algorithm 2.1 is asymptotically valid uniformly over a certain class of data generating processes. Let $\mathcal{P}$ be a set of probability measures defined on the Borel σ-algebra of $\mathcal{Y} \times \{0,1\}$, and the set of data generating processes satisfying (1.1) is denoted by

$$\mathcal{H}_0 = \left\{ (P,Q) \in \mathcal{P}^2 : \text{ inequalities (1.1) hold.} \right\}.$$

---

the supremum statistic over a pre-estimated set of y where $p(y,d) = q(y,d)$ (contact set). See Lee, Song, and Whang (2011), Linton, Song, and Whang (2010)), Donald and Hsu (forthcoming), and the literatures on generalized moment selection including Andrews and Barwick (2012), Andrews and Shi (2013), Andrews and Soares (2010), among others. Estimation of the contact set relies on a user-specified tuning parameter, and the test size can be affected by its choice.

[9] Since P_m^* and Q_n^* are drawn from the common pooled empirical distribution, recentering of the bootstrap empirical measures with respect to the original P_m and Q_n are not needed.

The uniform validity of our test procedure is based on the following two weak regularity conditions.

Condition-RG:
(a) Probability measures in $\mathcal{P}$ are nondegenerate and have a common dominating measure μ for the $\mathcal{Y}$-coordinate, where μ is the Lebesgue measure, a point mass measure with finite support points, or their mixture. The density functions $p(y,d) \equiv \frac{dP(\cdot,d)}{d\mu}$ are bounded uniformly over $\mathcal{P}$, i.e., there exists $M < \infty$ such that $p(y,d) \leq M$ holds at μ-almost every $y \in \mathcal{Y}$ and $d = 0, 1$ for all $P \in \mathcal{P}$.
(b) $\mathcal{P}$ is uniformly tight, i.e., for arbitrary $\epsilon > 0$, there exists a compact set $K \subset \mathcal{Y} \times \{0,1\}$ such that

$$\sup_{P \in \mathcal{P}} \{P(K^c)\} < \epsilon.$$

The asymptotic validity of the proposed test is stated in the next proposition (see Appendix B of the supplementary material for a proof).

Theorem 2.1 *Suppose Condition-RG. Let $\alpha \in (0, 1/2)$. (i) The test procedure of Algorithm 2.1 has asymptotically uniformly correct size for null hypothesis $\mathcal{H}_0$,*

$$\limsup_{N \to \infty} \sup_{(P,Q) \in \mathcal{H}_0} \Pr\left(T_N > c_{N,1-\alpha}\right) \leq \alpha.$$

(ii) For a fixed data generating process that violates inequalities (1.1) for some Borel set B, the test based on T_N is consistent, i.e., the rejection probability converges to one as $N \to \infty$.

This theorem establishes asymptotic uniform validity of the proposed test procedure over $\mathcal{P}$. The second claim of the proposition concerns the power of the test, and it shows that any fixed alternative within $\mathcal{P}$ can be consistently rejected.

2.3 Power against $N^{-1/2}$-local Alternatives

In this section, we show that the proposed test has nontrivial power against a class of non-parametric $N^{-1/2}$-local alternatives. Let $\left\{\left(P^{[N]}, Q^{[N]}\right) \in \mathcal{P}^2 : N = 1, 2, \dots\right\}$ denote a sequence of probability measures on $\mathcal{Y} \times \{1, 0\}$ shrinking to $(P_0, Q_0) \in \mathcal{H}_0$. The next assumption defines a class of local alternatives, against which we derive power of our test.

Assumption-LA:

A sequence of true alternatives $\left\{\left(P^{[N]}, Q^{[N]}\right) \in \mathcal{P}^2 : N = 1, 2, \dots\right\}$ is represented by

$$\begin{aligned} P^{[N]} &= P_0 + N^{-1/2}\beta_1^{[N]} \text{ and} \\ Q^{[N]} &= Q_0 + N^{-1/2}\beta_0^{[N]}, \end{aligned} \tag{2.4}$$

where $(P_0, Q_0) \in \mathcal{P}^2$ is a pair of probability measures on $\mathcal{Y}\times\{1,0\}$ and $\left\{\left(\beta_1^{[N]}, \beta_0^{[N]}\right) : N = 1, 2, \dots\right\}$ is a sequence of bounded signed measures on $\mathcal{Y} \times \{1, 0\}$.

(a) $(P_0, Q_0) \in \mathcal{H}_0$ and $P_0\left([y, y'], d\right) = Q_0\left([y, y'], d\right) > 0$ for some $-\infty \leq y \leq y' \leq \infty$, and $d \in \{1, 0\}$.

(b) For all N, $-N^{1/2}P_0 \leq \beta_1^{[N]} < \infty$ and $-N^{1/2}Q_0 \leq \beta_0^{[N]} < \infty$ hold and $\beta_1^{[N]}(\mathcal{Y}, d) = \beta_0^{[N]}(\mathcal{Y}, d) = 0$ for $d = 1, 0$.

(c) $\beta_1^{[N]} - \beta_0^{[N]}$ converges in terms of the sup metric over Borel sets to a bounded signed measure $\Delta\beta$ as $N \to \infty$.

(d) For some $([y, y'], d)$ satisfying (a), $\Delta\beta\left([y, y'], 1\right) < 0$ and/or $\Delta\beta\left([y, y'], 0\right) > 0$ hold.

Assumption-LA (a) says that $(P_0, Q_0) \in \mathcal{H}_0$, to which $\left(P^{[N]}, Q^{[N]}\right)$ converges, has a nonempty contact set with a positive measure in terms of $P_0 = Q_0$. Assumption-LA (b) ensures $\left(P^{[N]}, Q^{[N]}\right) \in \mathcal{P}^2$ and $\Pr(D = 1|Z = 1) \geq \Pr(D = 1|Z = 0)$ for all N, and $P^{[N]}$ and $Q^{[N]}$ are in an $N^{-1/2}$-neighborhood of P_0 and Q_0 in terms of the total variation distance. Assumption-LA (c) implies that $\sqrt{N}\left(P^{[N]} - Q^{[N]}\right)\left([y, y'], d\right) \to \Delta\beta\left([y, y'], d\right)$ at every $[y, y']$ contained in the contact set of P_0 and Q_0. Accordingly, combined with Assumption-LA (d), $\left(P^{[N]}, Q^{[N]}\right)$ violates the IV-validity testable implication at some $[y, y']$ contained in the contact set for all large N.

The next theorem provides a lower bound of the power of our test along $N^{-1/2}$-local alternatives satisfying Assumption-LA.

Theorem 2.2 *Assume Condition-RG and* $\left\{\left(P^{[N]}, Q^{[N]}\right) \in \mathcal{P}^2 : N = 1, 2, \dots\right\}$ *satisfies Assumption-LA. Then,*

$$\lim_{N\to\infty} \Pr_{P^{[N]}, Q^{[N]}}\left(T_N > c_{N,1-\alpha}\right) \geq 1 - \Phi(t),$$

where $\Phi(\cdot)$ *is the cumulative distribution function of the standard normal distribution,*

$$t = \left(\frac{\sigma^2_{P_0,Q_0}\left([y, y'], 1\right)}{\xi^2} \wedge 1\right)^{-1}\left[c_{1-\alpha} - \left[\lambda(1-\lambda)\right]^{1/2}\frac{\left|\Delta\beta\left([y, y'], d\right)\right|}{\xi \vee \sigma_{P_0,Q_0}\left([y, y'], 1\right)}\right],$$

$c_{1-\alpha}$ is the limit of the bootstrap critical value of Algorithm 2.1 that is bounded and depends only on $(\alpha, \xi, \lambda, P_0, Q_0)$, and $([y, y'], d)$ is as defined in Assumption-LA (a) and (d).

Note that the provided lower bound of the power is increasing in $|\Delta\beta([y, y'], d)|$ and it approaches one as a deviation from the null $|\Delta\beta([y, y'], d)|$ gets larger. Hence, we conclude that, for some $N^{-1/2}$-local alternatives satisfying Assumption-LA, the power is greater than the size of the test for every $\alpha \in (0, 1/2)$.

3 Extensions

3.1 A Multi-valued Discrete Instrument

The test procedure proposed above can be extended straightforwardly to a case with a multi-valued discrete instrument, $Z \in \{z_1, z_2, \ldots, z_K\}$. Let $p(z_k) = \Pr(D = 1|Z = z_k)$, and assume knowledge of the ordering of $p(z_k)$, so that without loss of generality we assume $p(z_1) \leq \cdots \leq p(z_K)$. With the multi-valued instrument, we denote the potential outcomes indexed by treatment and instrument status by $\{Y_{dz} : d = 0, 1, z = z_1, \ldots, z_K\}$, and potential selection responses by $\{D_z : z = z_1, \ldots, z_K\}$. The following assumptions guarantee that the linear two-stage least squares estimator can be interpreted as a weighted averages of the compliers average treatment effects (Imbens and Angrist (1994)).

Assumption: IV-validity for Multi-valued Discrete Z

1. *Instrument Exclusion*: $Y_{dz_1} = Y_{dz_2} = \cdots = Y_{dz_K}$ for $d = 1, 0$, with probability one.
2. *Random Assignment:* Z is jointly independent of $(Y_{1z_1}, \ldots, Y_{1z_K}, Y_{0z_1}, \ldots, Y_{0z_K})$ *and* $(D_{z_1}, \ldots, D_{z_K})$.
3. *Instrument Monotonicity: Given* $p(z_1) \leq \cdots \leq p(z_K)$, the potential selection indicators satisfy $D_{z_{k+1}} \geq D_{z_k}$ *with probability one for every* $k = 1, \ldots, (K-1)$.

Let $P(B, d|z_k) = \Pr(Y \in B, D = d|Z = z_k)$, $k = 1, \ldots, K$, and $P_{m_k}(B, d|z_k)$ be its empirical distribution based on the subsample of $Z_i = z_k$ with size m_k. The testable implication of the binary instrument case is now generalized to the following set of inequalities,

$$\begin{aligned} &P(B, 1|z_1) \leq P(B, 1|z_2) \leq \cdots \leq P(B, 1|z_K) \quad \text{and} \\ &P(B, 0|z_1) \geq P(B, 0|z_2) \geq \cdots \geq P(B, 0|z_K) \end{aligned} \tag{3.1}$$

for all Borel set B in $\mathcal{Y}$. Using the test statistic for the binary Z case to measure the violation of the inequalities across the neighboring values of Z, we can develop a statistic that jointly tests the inequalities of (3.1),

$$T_N = \max\left\{T_{N,1}, \ldots, T_{N,K-1}\right\}, \tag{3.2}$$

where, for $k = 1, \ldots, (K-1)$,

$$\begin{aligned}
T_{N,k} &= \left(\frac{m_k m_{k+1}}{m_k + m_{k+1}}\right)^{1/2} \max\left\{ \begin{array}{l} \sup_{-\infty \le y \le y' \le \infty} \left\{ \frac{P_{m_k}([y,y'],1|z_k) - P_{m_{k+1}}([y,y'],1|z_{k+1})}{\xi_k \vee \sigma_k([y,y'],1)} \right\}, \\ \sup_{-\infty \le y \le y' \le \infty} \left\{ \frac{P_{m_{k+1}}([y,y'],0|z_{k+1}) - P_{m_k}([y,y'],0|z_k)}{\xi_k \vee \sigma_k([y,y'],0)} \right\} \end{array} \right\}, \\
\sigma_k^2([y,y'],d) &= \left(\frac{m_{k+1}}{m_k + m_{k+1}}\right) P_{m_k}([y,y'],d|z_k)\left(1 - P_{m_k}([y,y'],d|z_k)\right) \\
&\quad + \left(\frac{m_k}{m_k + m_{k+1}}\right) P_{m_{k+1}}([y,y'],d|z_k)(1 - P_{m_{k+1}}([y,y'],d|z_k)),
\end{aligned}$$

and $\xi_1, \ldots, \xi_{K-1}$ are positive constants. Critical values can be obtained by applying a resampling algorithm of the previous section to each $T_{N,k}$ simultaneously.

Algorithm 3.1:

1. Let $H_{N,k}(\cdot) = \left(\frac{m_{k+1}}{m_k + m_{k+1}}\right) P_{m_{k+1}}(\cdot|z_{k+1}) + \left(\frac{m_k}{m_k + m_{k+1}}\right) P_{m_k}(\cdot|z_k)$ be the pooled empirical measures that pool the sample of $Z_i = z_{k+1}$ and that of $Z_i = z_k$. *Sample* (Y_i^*, D_i^*), $i = 1, \ldots, m_{k+1}$ *randomly with replacement from* $H_{N,k}$ *and construct the bootstrap empirical distribution* $P^*_{m_{k+1}}(\cdot|z_{k+1})$. *Similarly, sample* (Y_j^*, D_j^*), $j = 1, \ldots, m_k$ *randomly with replacement from* $H_{N,k}$ *and construct the bootstrap empirical distribution* $P^*_{m_k}(\cdot|z_k)$.

2. Apply step 1 for every $k = 1, \ldots, (K-1)$, and obtain $(K-1)$ pairs of the resampled empirical measures, $(P^*_{m_1}, P^*_{m_2})$, $(P^*_{m_2}, P^*_{m_3}), \ldots, (P^*_{m_{K-1}}, P^*_{m_K})$. Define, for $k = 1, \ldots, (K-1)$,

$$\begin{aligned}
\sigma_k^{*2}([y,y'],d) &= \left(\frac{m_{k+1}}{m_k + m_{k+1}}\right) P^*_{m_k}([y,y'],d|z_k)\left(1 - P^*_{m_k}([y,y'],d|z_k)\right) \\
&\quad + \left(\frac{m_k}{m_k + m_{k+1}}\right) P^*_{m_{k+1}}([y,y'],d|z_{k+1})(1 - P^*_{m_{k+1}}([y,y'],d|z_{k+1})), \\
T^*_{N,k} &= \left(\frac{m_k m_{k+1}}{m_k + m_{k+1}}\right)^{1/2} \\
&\quad \times \max\left\{ \begin{array}{l} \sup_{-\infty \le y \le y' \le \infty} \left\{ \frac{P^*_{m_k}([y,y'],1|z_k) - P^*_{m_{k+1}}([y,y'],1|z_{k+1})}{\xi_k \vee \sigma^*_k([y,y'],1)} \right\}, \\ \sup_{-\infty \le y \le y' \le \infty} \left\{ \frac{P^*_{m_{k+1}}([y,y'],0|z_{k+1}) - P^*_{m_k}([y,y'],0|z_k)}{\xi_k \vee \sigma^*_k([y,y'],0)} \right\} \end{array} \right\},
\end{aligned}$$

where ξ_k, $k = 1, \ldots, (K-1)$, are positive constants. *The bootstrap statistic* T_N^* is computed accordingly by $T_N^* = \max\left\{T_{N,1}^*, \ldots, T_{N,K-1}^*\right\}$.

3. *Iterate Step 1 -3 many times, get the empirical distribution of* T_N^*, and *obtain a bootstrap critical value* $c_{N,1-\alpha}$ *from its empirical* $(1-\alpha)$*-th quantile .*

4. *Reject the null hypothesis (3.1) if* $T_N > c_{N,1-\alpha}$. *The bootstrap p-value is obtained by the proportion of* T_N^*'s *greater than* T_N.

3.2 Conditioning Covariates

Empirical studies commonly use observable conditioning covariates in the context of instrumental variable methods. One of the major motivations for introducing them is to control for potential confounders that invalidate the random assignment assumption. This section briefly discusses how to extend IV-validity test proposed above to the settings with conditioning covariates, $X \in \mathbb{X} \subset \mathbb{R}^{d_x}$, used for this purpose.

IV-validity to be tested in this case consists of the joint restriction of instrument exclusion, instrument monotonicity, and the conditional version of the instrument random assignment assumption, $(Y_{11}, Y_{10}, Y_{01}, Y_{00}, D_1, D_0) \perp Z|X$. These three assumptions combined with the first stage rank condition, $\Pr(D = 1|Z = 1, X) \neq \Pr(D = 1|Z = 0, X)$ for some X, guarantees that the linear two stage least squares with a function of (Z, X) used as an instrument (e.g. $\Pr(D = 1|Z, X)$) estimates a certain weighted average of the complier's conditional causal effects $E(Y_1 - Y_0|T = c, X)$ (Heckman and Vytlacil (2005)). Moreover, under the same set of assumptions, the semiparametric IV estimator developed by Abadie (2003) consistently estimates the unconditional complier's causal effect $E(Y_1 - Y_0|T = c)$.

A testable implication with the largest screening power in the sense similar to Proposition 1.1 is given by the conditional version of the inequalities (1.1), i.e., for every Borel set $B \subset \mathcal{Y}$ and $X \in \mathbb{X}$,

$$\begin{aligned} &\Pr(Y \in B, D = 1|Z = 1, X) - \Pr(Y \in B, D = 1|Z = 0, X) \geq 0, \\ &\Pr(Y \in B, D = 0|Z = 0, X) - \Pr(Y \in B, D = 0|Z = 1, X) \geq 0. \end{aligned} \tag{3.3}$$

As shown in Lemma B.8 in the supplementary material, the use of Theorem 3.1 of Abadie (2003) and the instrument function argument for conditional moment inequalities as given in Andrews and Shi (2013) and Khan and Tamer (2009) enable us to reduce (3.3) to the

following unconditional moment inequalities without loss of any information,[10]

$$\begin{aligned} E\left[\kappa_1\left(D,Z,X\right)g(Y,X)\right] &\geq 0, \\ E\left[\kappa_0\left(D,Z,X\right)g(Y,X)\right] &\geq 0, \quad \text{for all } g(\cdot,\cdot)\in\mathcal{G} \end{aligned} \tag{3.4}$$

where

$$\begin{aligned} \kappa_1\left(D,Z,X\right) &= D\frac{Z-\Pr(Z=1|X)}{\Pr(Z=0|X)\Pr(Z=1|X)}, \\ \kappa_0\left(D,Z,X\right) &= (1-D)\frac{(1-Z)-\Pr(Z=0|X)}{\Pr(Z=0|X)\Pr(Z=1|X)}, \end{aligned}$$

and $\mathcal{G}$ is the class of indicator functions for boxes in $\mathcal{Y}\times\mathcal{X}$,

$$\mathcal{G}=\left\{ \begin{array}{c} 1\left\{(Y,X)\in C\right\} : C=[y,y']\times[x_1,x_1']\times\cdots\times\left[x_{d_x},x_{d_x}'\right],\ -\infty\leq y\leq y'\leq\infty\ , \\ -\infty\leq x_l\leq x_l'\leq\infty,\quad l=1,\ldots,d_x. \end{array} \right\}. \tag{3.5}$$

Accordingly, a variance-weighted KS statistic to infer (3.4) can be proposed as

$$T_N=\sqrt{N}\max\left\{ \begin{array}{c} \sup_{g\in\mathcal{G}}\frac{-E_N[\hat{\kappa}_1(D,Z,X)g(Y,X)]}{\xi\vee\hat{\sigma}_1(g)}, \\ \sup_{g\in\mathcal{G}}\frac{-E_N[\hat{\kappa}_0(D,Z,X)g(Y,X)]}{\xi\vee\hat{\sigma}_0(g)} \end{array} \right\},$$

where $\hat{\kappa}_d$ is an estimate of κ_d with estimated $\Pr(Z=1|X)$ plugged in, $E_N\left(\cdot\right)$ is the sample average, and $\hat{\sigma}_d^2\left(g\right)$ is the sample variance of $\hat{\kappa}_d\left(D_i,Z_i,X_i\right)g(Y_i,X_i)$. Ignoring estimation uncertainty of $\hat{\kappa}_d$, the critical values can be obtained by bootstrapping the supremum statistic of the recentered moments,

$$T_N^*=\sqrt{N}\max\left\{ \begin{array}{c} \sup_{g\in\mathcal{G}}\frac{-\left[E_N^*[\hat{\kappa}_1(D,Z,X)g(Y,X)]-E_N[\hat{\kappa}_1(D,Z,X)g(Y,X)]\right]}{\xi\vee\hat{\sigma}_1^*(g)}, \\ \sup_{g\in\mathcal{G}}\frac{-\left[E_N^*[\hat{\kappa}_0(D,Z,X)g(Y,X)]-E_N[\hat{\kappa}_0(D,Z,X)g(Y,X)]\right]}{\xi\vee\hat{\sigma}_0^*(g)} \end{array} \right\},$$

where $E_N^*\left(\cdot\right)$ is the sample average based on a bootstrap sample that is obtained by resampling (Y,D,Z,X) from the original sample, and $\hat{\sigma}_d^{*2}\left(g\right)$ is the variance estimate based on the bootstrap sample.

[10] If the random assignment assumption is strengthened to $(Y_{11},Y_{10},Y_{01},Y_{00},D_1,D_0,X)\perp Z$, then it can be shown that the moment conditions of (3.4) are reduced to

$$\begin{aligned} &\Pr((Y,X)\in C,D=1|Z=1)-\Pr((Y,X)\in C,D=1|Z=0)\geq 0, \\ &\Pr((Y,X)\in C,D=0|Z=0)-\Pr((Y,X)\in C,D=0|Z=1)\geq 0, \end{aligned}$$

for any box C in $\mathcal{Y}\times\mathcal{X}$. As a result, the test procedure for no-covariate case can be extended straightforwardly to this case.

Regarding implementations of this procedure in practice, a couple of issues need to be addressed. First, in the presence of many covariates, computation of the statistic involves an optimization over a large class of indicator functions. This raises a computational difficulty in implementing the test. Second, the proposed approach requires estimation of $\Pr(Z = 1|X)$, which the linear two stage least squares estimation do not require. Hence, if a parametric estimation for $\Pr(Z = 1|X)$ is used to implement the test, the validity of the test relies on the assumption of a correct functional form for $\Pr(Z = 1|X)$, and rejection can be driven by its misspecification rather than the violation of IV-validity.

4 Empirical Applications

We illustrate a use of our test using the two data sets mentioned in Introduction.

4.1 Draft Lottery Data

The draft lottery data of Angrist (1991) consist of a sample of 10,101 white men, born between 1950 and 1953 extracted from March Current Population Surveys of 1979 and 1981-1985. The outcome variable is measured in terms of the logarithm of weekly earnings imputed by the annual labor earnings divided by weeks worked. The treatment is whether one has a Vietnam veteran status. In order to cope with the selection to treatment status, Angrist (1991) constructed the binary indicator of the draft eligibility, which is randomly assigned based on one's birthdate through the draft lotteries. It is reasonable to believe that the constructed instrument is independent of any individual characteristics, and defiers do not exist in the population even though the sampling design does not exclude the possibility of having them. A less credible assumption in the current context would be instrument exclusion. For instance, the draft lottery can directly affect control outcomes for some never-takers if those who were drafted change their career choice, school years, or migration choice for the purpose of escaping from the military service.

Table 1 shows the result of our test. We present the bootstrap p-values of our test for several different specifications of the trimming constant. All of them are one, so we do not reject validity of the draft lottery instrument from the data.

4.2 Returns to Education: Proximity to College Data

The Card data is based on National Longitudinal Survey of Young Men (NLSYM) that began in 1966 with 14-24 years old men and continued with follow-up surveys through 1981. Based on the respondents' county of residence at 1966, the Card data provides the presence of a 4-year college in the local labor market. The observations of years of education and wages were based on the follow-ups' educational attainment and wages reported in the interview in 1976.

Proximity to college was used as an instrument, because the presence of a nearby college reduces the cost of college education by allowing students to live at their home, while one's unobservable ability is presumably independent of student's residence during their teenage years. Compliers in this context can be considered as those who grew up in relatively low-income families and who were not able to go to college without living with their parents. Being different from the original Card's study, we treat the educational level as a binary treatment, with years of education greater than or equal to 16 years, that is, the treatment can be considered as a four year college degree.

We specify the measure of outcome to be the logarithm of weekly earnings. In the first specification, we do not control for any demographic covariates. This raises a concern regarding the violation of random assignment assumption. For instance, one's region of residence, or whether they were born in the standard metropolitan area or rural area. may well be dependent on one's wage levels and the proximity to colleges if the urban areas are more likely to have colleges and higher wage levels compared to the rural areas.

Our test procedure yields zero p-values for each choice of trimming constant. This provides an empirical evidence that, without controlling for any covariates, college proximity is not a valid instrument.

Table 1: Test Results of the Empirical Applications

Bootstrap iterations 500

data	Draft lottery data			Proximity to college data					
	No Covariate			No Covariate			With Covariates*		
sample size (m,n)	(2780,7321)			(2053,957)			(2053,957)		
$\Pr(D=1\|Z=1)$, $\Pr(D=1\|Z=0)$	0.31, 0.19			0.29, 0.22					
trimming constant ξ	0.07	0.3	1	0.07	0.3	1	0.07	0.3	1
Bootstrap test, p-value	1.00	1.00	1.00	0.00	0.00	0.00	0.89	0.71	0.91

* five dummy variables indicating (1) residence in a standard metropolitan area (SMSA) in 1976, (2) residence in SMSA in 1966, (3) race is black or not, (4) residence in southern states in 1976, and (5) residence in southern states in 1966.

The original study of Card (1993) indeed emphasized the importance of controlling for regions, residence in the urban area, race, job experience, and parent's education, and he included them in his specification of the two stage least square estimation. In our second specification, we control for the covariates listed at the bottom of Table 1, which are all binary variables. We estimate $\Pr(Z=1|X)$ using a linear probability regression with these five dummy variables. The class of indicator functions $\mathcal{G}$ we use is

$$\mathcal{G} = \left\{ \begin{array}{c} 1\{(Y,X)\in C\} : C = [y_q, y_{q'}] \times \{x_1\} \times \cdots \times \{x_{d_x}\}, \\ y_q \text{ is the empirical } q\text{-th quantile of } Y, \\ q,\, q' \in \{0, 0.05, \ldots, 0.95, 1\},\, q < q' \\ x_l \in \{0,1\}, \quad l = 1, \ldots, d_x. \end{array} \right\}$$

With these covariates, the p-values turn out to be large. We therefore conclude that we do not reject validity of the college proximity instrument when these covariates are controlled for.

References

[1] Abadie, A. (2002): "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models," Journal of the American Statistical Association, 97, 284-292.

[2] Abadie, A. (2003): "Semiparametric Instrumental Variable Estimation of Treatment Response Models," Journal of Econometrics, 113, 231 - 263.

[3] Abadie, A., J. D. Angrist, and G. W. Imbens. (2002): "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," Econometrica, 70, 91-117.

[4] Andrews, D.W.K. and P. Jia Barwick (2012): "Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure," Econometrica, 80, 2805-2826.

[5] Andrews, D.W.K. and X. Shi (2013): "Inference Based on Conditional Moment Inequalities," Econometrica, 81, 609-666.

[6] Andrews, D.W.K. and G. Soares (2010): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," Econometrica, 78, 119-157.

[7] Angrist, J. D. (1991): "The Draft Lottery and Voluntary Enlistment in the Vietnam Era," Journal of the American Statistical Association, 86, 584-595

[8] Angrist, J.D., G.W. Imbens (1995): "Two-stage Least Squares Estimation of Average Causal Effects Using Instrumental Variables," Journal of the American Statistical Association, 91, 444-455.

[9] Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996): "Identification of Causal Effects Using Instrumental Variables," Journal of the American Statistical Association, 91, 444-455.

[10] Armstrong, T.B. and H.P. Chan (2013): "Multiscale Adaptive Inference on Conditional Moment Inequalities," Cowles Foundation Discussion Paper, No. 1885. Yale University.

[11] Armstrong, T.B. (2014): "Weighted KS Statistics for Inference on Conditional Moment inequalities," Journal of Econometrics, 181, 92-116.

[12] Balke, A. and J. Pearl (1997): "Bounds on Treatment Effects from Studies with Imperfect Compliance," Journal of the American Statistical Association, 92, 1171-1176.

[13] Barrett, G.F. and S.G. Donald (2003): "Consistent Tests for Stochastic Dominance," Econometrica 71, 71-104.

[14] Barua, R. and K. Lang (2009): "School Entry, Educational Attainment and Quarter of Birth: A Cautionary Tale of LATE." NBER Working Paper 15236, National Bureau of Economic Research.

[15] Breusch, T.S. (1986): "Hypothesis Testing in Unidentified Models," Review of Economic Studies, 53, 4, 635-651.

[16] Card, D. (1993): "Using Geographical Variation in College Proximity to Estimate the Returns to Schooling", National Bureau of Economic Research Working Paper No. 4, 483.

[17] Chernozhukov, V., S. Lee, and A.M. Rosen (2013): "Intersection Bounds: Estimation and Inference," Econometrica, 81, 667–737.

[18] Chetverikov, D. (2012): "Adaptive Test of Conditional Moment Inequalities," unpublished manuscript, UCLA.

[19] de Chaisemartin (2014): "Tolerating defiance: local average treatment effects without monotonicity." unpublished manuscript, University of Warwick.

[20] Donald, S. and Y.-C. Hsu (forthcoming): "Improving the Power of Tests of Stochastic Dominance," forthcoming in Econometric Reviews.

[21] Fiorini, M., K. Stevens, M. Taylor, and B. Edwards (2013): "Monotonically Hopeless? Monotonicity in IV and Fuzzy RD Designs," unpublished manuscript, University of Technology Sydney, University of Sydney, and Australian Institute of Family Studies.

[22] Heckman, J. J. and E. Vytlacil (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," Econometrica 73, 669-738.

[23] Horváth, L., P. Kokoszka, and R. Zitikis (2006): "Testing for Stochastic Dominance Using the Weighted McFadden-type statistic," Journal of Econometrics, 133, 191-205.

[24] Huber, M. and G. Mellace (2013) "Testing Instrument Validity for LATE Identification based on Inequality Moment Constraints," unpublished manuscript, University of Sankt Gallen.

[25] Imbens, G.W. (forthcoming) "Instrumental Variables: An Econometrician's Perspective," forthcoming in Statistical Science.

[26] Imbens, G. W. and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," Econometrica, 62, 467-475.

[27] Imbens, G. W. and D. B. Rubin (1997): "Estimating Outcome Distributions for Compliers in Instrumental Variable Models," Review of Economic Studies, 64, 555-574.

[28] Khan, S. and E. Tamer (2009): "Inference on Endogenously Censored Regression Models Using Conditional Moment Inequalities," Journal of Econometrics, 152, 104 - 119.

[29] Lee, SJ. (2014): "On Variance Estimation for 2SLS When Instruments Identify Different LATEs," unpublished manuscript, University of New South Wales.

[30] Lee, S., K. Song, and Y. Whang (2011): "Testing Functional Inequalities," cemmap working paper 12/11, University College London.

[31] Linton, O., E. Maasoumi, and Y. Whang (2005): "Consistent Testing for Stochastic Dominance under General Sampling Schemes," Review of Economic Studies, 72, 735-765.

[32] Linton, O., K. Song, and Y. Whang (2010): "An Improved Bootstrap Test of Stochastic Dominance," Journal of Econometrics, 154, 186-202.

[33] Mourifié, I. and Y. Wan (2014): "Testing LATE Assumptions," unpublished manuscript, University of Toronto.

[34] Romano, J. P. (1988): "A Bootstrap Revival of Some Nonparametric Distance Tests." Journal of American Statistical Association, 83, 698-708.

Supplimentary Material for "A Test for Instrument Validity"

Toru Kitagawa*

CeMMAP and *Department of Economics, University College London*

Aug, 2014

A Proof of Proposition 1.1

In addition to the notations introduced in the main text, we introduce the individual type indicator T,

$$T = c\text{: } \textit{complier} \quad \text{if } D_1 = 1, D_0 = 0$$
$$T = n\text{: } \textit{never-taker} \quad \text{if } D_1 = 0, D_0 = 0$$
$$T = a\text{: } \textit{always-taker} \quad \text{if } D_1 = 1, D_0 = 1$$
$$T = df\text{: } \textit{defier} \quad \text{if } D_1 = 0, D_0 = 1.$$

When instrument exclusion is imposed, we suppress the z subscript in the potential outcome notation, and define $Y_1 \equiv Y_{11} = Y_{10}$ and $Y_0 \equiv Y_{01} = Y_{00}$ as a pair of the potential outcomes indexed solely by $D = 1$ and 0. Note that the joint restriction of instrument exclusion and random assignment is equivalent to $(Y_1, Y_0, T) \perp Z$.

A.1 Proof of Proposition 1.1

(i) Let P and Q satisfying the inequalities (1.1) be given and assume instrument exclusion. Our goal is to show that there exists a joint distribution of (Y_1, Y_0, T, Z) that is consistent

*Email: t.kitagawa@ucl.ac.uk. Financial support from the ESRC through the ESRC Center for Microdata Methods and Practice (CEMMAP) (grant number RES-589-28-0001) and the Merit Dissertation Fellowship from the Graduate School of Economics in Brown University are gratefully acknowledged.

with the given P and Q, and satisfies $(Y_1, Y_0, T) \perp Z$ and instrument monotonicity. Since the marginal distribution of Z is not important in the following argument, we focus on constructing the conditional distribution of (Y_1, Y_0, T) given Z. Let $p(\cdot, d) = \frac{dP(\cdot,d)}{d\mu}$ and $q(y, d) = \frac{dQ(\cdot,d)}{d\mu}$. Define nonnegative functions,

$$\begin{aligned}
h_{Y_1,c}(y) &\equiv p(y,1) - q(y,1), \\
h_{Y_0,c}(y) &\equiv q(y,0) - p(y,0), \\
h_{Y_1,a}(y) &= q(y,1), \\
h_{Y_0,n}(y) &= p(y,0) \\
h_{Y_1,df}(y) &= 0, \\
h_{Y_0,df}(y) &= 0,
\end{aligned}$$

and $h_{Y_0,a}(y)$ and $h_{Y_1,n}(y)$ are arbitrary nonnegative functions supported on $\mathcal{Y}$ and satisfy $\int_{\mathcal{Y}} h_{Y_0,a}(y)d\mu = \Pr(D = 1|Z = 1)$ and $\int_{\mathcal{Y}} h_{Y_1,n}(y)d\mu = \Pr(D = 1|Z = 0)$. These nonnegative functions, $h_{Y_d,t}(y)$, $d \in \{1, 0\}$, $t \in \{c, n, a, df\}$, are introduced for the purpose of imputing a probability density of $\frac{\partial}{\partial \mu} \Pr(Y_d \in \cdot, T = t)$ that match the data distribution P and Q. Consider the following probability law of (Y_1, Y_0, T) given Z defined on the product σ-algebra of $\mathcal{Y} \times \mathcal{Y} \times \{c, n, a, df\}$,

$$\begin{aligned}
&\Pr(Y_1 \in B_1, Y_0 \in B_0, T = c|Z = 1) = \Pr(Y_1 \in B_1, Y_0 \in B_0, T = c|Z = 0) \\
\equiv\ &\begin{cases} \frac{\int_{B_1} h_{Y_1,c}(y)d\mu}{\int_{\mathcal{Y}} h_{Y_1,c}(y)d\mu} \times \frac{\int_{B_0} h_{Y_0,c}(y)d\mu}{\int_{\mathcal{Y}} h_{Y_0,c}(y)d\mu} \times [P(\mathcal{Y},1) - Q(\mathcal{Y},1)] & \text{if } [P(\mathcal{Y},1) - Q(\mathcal{Y},1)] > 0, \\ 0 & \text{if } [P(\mathcal{Y},1) - Q(\mathcal{Y},1)] = 0, \end{cases} \\
&\Pr(Y_1 \in B_1, Y_0 \in B_0, T = n|Z = 1) = \Pr(Y_1 \in B_1, Y_0 \in B_0, T = n|Z = 0) \\
\equiv\ &\begin{cases} \frac{\int_{B_1} h_{Y_1,n}(y)d\mu}{\int_{\mathcal{Y}} h_{Y_1,n}(y)d\mu} \times \frac{\int_{B_0} h_{Y_0,n}(y)d\mu}{\int_{\mathcal{Y}} h_{Y_0,n}(y)d\mu} \times P(\mathcal{Y},0) & \text{if } P(\mathcal{Y},0) > 0, \\ 0 & \text{if } P(\mathcal{Y},0) = 0, \end{cases} \\
&\Pr(Y_1 \in B_1, Y_0 \in B_0, T = a|Z = 1) = \Pr(Y_1 \in B_1, Y_0 \in B_0, T = a|Z = 0) \\
\equiv\ &\begin{cases} \frac{\int_{B_1} h_{Y_1,a}(y)d\mu}{\int_{\mathcal{Y}} h_{Y_1,a}(y)d\mu} \times \frac{\int_{B_0} h_{Y_0,a}(y)d\mu}{\int_{\mathcal{Y}} h_{Y_0,a}(y)d\mu} \times Q(\mathcal{Y},1) & \text{if } Q(\mathcal{Y},1) > 0, \\ 0 & \text{if } Q(\mathcal{Y},1) = 0, \end{cases} \\
&\Pr(Y_1 \in B_1, Y_0 \in B_0, T = df|Z = 1) = \Pr(Y_1 \in B_1, Y_0 \in B_0, T = df|Z = 0) \\
\equiv\ &0,
\end{aligned}$$

where $P(\mathcal{Y}, d) = \Pr(D = d|Z = 1)$ and $Q(\mathcal{Y}, d) = \Pr(D = d|Z = 0)$. Note that this is a probability measure on the product sigma-algebra of $\mathcal{Y} \times \mathcal{Y} \times \{c, a, n, df\}$, since it is

nonnegative, additive, and sums up to one,

$$\sum_{t\in\{c,n,a,df\}} \Pr(Y_1 \in \mathcal{Y}, Y_0 \in \mathcal{Y}, T = t|Z = z) = 1, \quad z = 1, 0.$$

The proposed probability distribution of $(Y_1, Y_0, T|Z)$ clearly satisfies the joint independence and instrument monotonicity by the construction, and it induces the given data generating process. i.e., the proposed probability distribution of $(Y_1, Y_0, T|Z)$ satisfies

$$\begin{aligned}
P(B,1) &= \Pr(Y_1 \in B, Y_0 \in \mathcal{Y}, T = a|Z = 1) + \Pr(Y_1 \in B, Y_0 \in \mathcal{Y}, T = c|Z = 1), \quad \text{(A.1)}\\
Q(B,1) &= \Pr(Y_1 \in B, Y_0 \in \mathcal{Y}, T = a|Z = 0) + \Pr(Y_1 \in B, Y_0 \in \mathcal{Y}, T = df|Z = 0),\\
P(B,0) &= \Pr(Y_1 \in \mathcal{Y}, Y_0 \in B, T = n|Z = 1) + \Pr(Y_1 \in \mathcal{Y}, Y_0 \in B, T = df|Z = 1),\\
Q(B,0) &= \Pr(Y_1 \in \mathcal{Y}, Y_0 \in B, T = n|Z = 0) + \Pr(Y_1 \in \mathcal{Y}, Y_0 \in B, T = c|Z = 0).
\end{aligned}$$

This completes the proof of the first claim.

(ii) Let arbitrary P and Q satisfying inequalities (1.1) be given. We maintain instrument exclusion, so, in what follows, we construct a probability law of (Y_1, Y_0, T) given Z that is consistent to the P and Q, but violates $(Y_1, Y_0, T) \perp Z$. Consider the following probability distribution of (Y_1, Y_0, T) given Z,

$$\begin{aligned}
&\Pr(Y_1 \in B_1, Y_0 \in B_0, T = c|Z = 1) = 0,\\
&\Pr(Y_1 \in B_1, Y_0 \in B_0, T = c|Z = 0) = \begin{cases} \frac{Q(B_1,0)Q(B_0,0)}{Q(\mathcal{Y},0)} & \text{if } Q(\mathcal{Y},0) > 0,\\ 0 & \text{if } Q(\mathcal{Y},0) = 0, \end{cases}\\
&\Pr(Y_1 \in B_1, Y_0 \in B_0, T = n|Z = 1) = \begin{cases} \frac{P(B_1,0)P(B_0,0)}{P(\mathcal{Y},0)} & \text{if } P(\mathcal{Y},0) > 0,\\ 0 & \text{if } P(\mathcal{Y},0) = 0, \end{cases}\\
&\Pr(Y_1 \in B_1, Y_0 \in B_0, T = n|Z = 0) = 0,\\
&\Pr(Y_1 \in B_1, Y_0 \in B_0, T = a|Z = 1) = \begin{cases} \frac{P(B_1,1)P(B_0,1)}{P(\mathcal{Y},1)} & \text{if } P(\mathcal{Y},1) > 0,\\ 0 & \text{if } P(\mathcal{Y},1) = 0, \end{cases}\\
&\Pr(Y_1 \in B_1, Y_0 \in B_0, T = a|Z = 0) = 0,\\
&\Pr(Y_1 \in B_1, Y_0 \in B_0, T = df|Z = 1) = 0,\\
&\Pr(Y_1 \in B_1, Y_0 \in B_0, T = df|Z = 0) = \begin{cases} \frac{Q(B_1,1)Q(B_0,1)}{Q(\mathcal{Y},1)} & \text{if } Q(\mathcal{Y},1) > 0,\\ 0 & \text{if } Q(\mathcal{Y},1) = 0. \end{cases}
\end{aligned}$$

Note that, in this construction, Z and T are dependent, i.e., $Z = 1$ is assigned to only never takers and always takers, and $Z = 0$ is assigned to only compliers and defiers, so it violates

$T \perp Z$ (and the no-defier condition as well if $Q(Y,1) > 0$). Furthermore, the proposed distribution of $(Y_1, Y_0, T|Z)$ satisfies (A.1), so it is consistent with the P and Q. Since the proposed construction is feasible for any P and Q, we conclude that for any P and Q that meet the testable implications, there exists a distribution of (Y_1, Y_0, T, Z) that violates IV-validity.

B Appendix B: Proof of Theorem 2.1

B.1 Notations

In addition to the notations introduced in the main text, we introduce the following notations that are used throughout this appendix. Let $\mathcal{F}$ be a set of indicator functions defined on $\mathcal{X} \equiv \mathcal{Y} \times \{0,1\}$,

$$\mathcal{F} = \left\{1_{\{[y,y'],1\}}(Y,D) : -\infty \leq y \leq y' \leq \infty\right\} \cup \left\{1_{\{[y,y'],0\}}(Y,D) : -\infty \leq y \leq y' \leq \infty\right\},$$

where $1_{\{B,d\}}(Y,D)$ is the indicator function for event $\{Y \in B, D = d\}$. The Borel σ-algebra of $\mathcal{X}$ is denoted by $\mathcal{B}(\mathcal{X})$. Note that $\mathcal{F}$ is a VC-class of functions since a class of connected intervals is a VC-class of subsets. We denote a generic element of $\mathcal{F}$ by f. For generic $P \in \mathcal{P}$, let P_m be an empirical probability measure constructed by a size m iid sample from P. we define short-hand notations, $P(f) \equiv P([y,y'],d)$ and $P_m(f) \equiv P_m([y,y'],d)$. Denote empirical processes indexed by $\mathcal{F}$ by

$$G_{m,P}(\cdot) = \sqrt{m}\,(P_m - P)(\cdot).$$

For a probability measure P on $\mathcal{X}$, we denote the mean zero P-brownian bridge processes indexed by $\mathcal{F}$ by $G_P(\cdot)$. Let $\rho_\omega(f,f') = [\omega((f-f')^2)]^{1/2}$ be a seminorm on $\mathcal{F}$ defined in terms of the L_2-metric with respect to a finite measure ω on $\mathcal{X}$. Given a deterministic sequence of the sizes of two samples, $\{(m(N), n(N)) : N = 1, 2, \dots\}$, let $\left\{(P^{[m(N)]}, Q^{[n(N)]}) \in \mathcal{P}^2 : N = 1, 2, \dots\right\}$ be a sequence of the two sample probability measures that drift with the sample sizes $(m(N), n(N))$, where superscripts with brackets index a sequence. We often omit the arguments of $(m(N), n(N))$ unless any confusion arises.

Let $\sigma_P^2(\cdot,\cdot) : \mathcal{F}^2 \to \mathbb{R}$ denote the covariance kernel of P-brownian bridges, $\sigma_P^2(f,g) = P(fg) - P(f)P(g)$. We denote by $\sigma_{P,Q}^2(f,g) : \mathcal{F}^2 \to \mathbb{R}$ the covariance kernel of the independent two-sample brownian bridge processes $(1-\lambda)^{1/2} G_P(\cdot) - \lambda^{1/2} G_Q(\cdot)$,

$$\sigma_{P,Q}^2(f,g) = (1-\lambda)\sigma_P^2(f,g) + \lambda\sigma_Q^2(f,g),$$

and $\sigma^2_{P_m,Q_n}(\cdot,\cdot)$ be its sample analogue,

$$\sigma^2_{P_m,Q_n}(f,g) = (1-\hat{\lambda})\left[P_m(fg) - P_m(f)P_m(g)\right] + \hat{\lambda}\left[Q_n(fg) - Q_n(f)Q_n(g)\right].$$

Note that, with the current notation, $\sigma^2_{P_m,Q_m}\left([y,y'],d\right)$ defined in the main text is equivalent to $\sigma^2_{P_m,Q_n}(f,f)$, for $f = 1_{\{[y,y'],d\}}$. For a sequence of random variables $\{W_N : N = 1,2,\dots\}$ whose probability law is governed by a sequence of two sample probability measures $(P^{[m(N)]}, Q^{[n(N)]})$, $W_N \underset{P^{[m]},Q^{[n]}}{\longrightarrow} c$ denotes convergence in probability in the sense of, for every $\epsilon > 0$,

$$\lim_{N\to\infty} \Pr_{P^{[m]},Q^{[n]}}\left(|W_N - c| > \epsilon\right) = 0.$$

In particular, if $W_N \underset{P^{[m]},Q^{[n]}}{\longrightarrow} 0$, we notate as $W_N = o_{P^{[m]},Q^{[n]}}(1)$.

B.2 Auxiliary Lemmas

We first present a set of lemmas to be used in the proofs of Theorems 2.1 and 2.2.

Lemma B.1 *Let $\left\{P^{[m]} \in \mathcal{P} : m = 1,2\dots\right\}$ be a sequence of probability measures on $\mathcal{X}$. Then,*

$$\sup_{f\in\mathcal{F}} \left|\left(P_m^{[m]} - P^{[m]}\right)(f)\right| \underset{P^{[m]}}{\longrightarrow} 0.$$

Proof. $\mathcal{F}$ is the class of indicator functions corresponding to the interval VC-class of subsets, so an application of the Glivenko-Cantelli theorem uniform in $\mathcal{P}$ (Theorem 2.8.1 of van der Vaart and Wellner (1996)) yields the claim. ■

Lemma B.2 *Suppose Condition-RG. Let $\left\{P^{[m]} \in \mathcal{P} : m = 1,2\dots\right\}$ be a sequence of data generating processes on $\mathcal{X}$ that weakly converges to $P_0 \in \mathcal{P}$ as $m \to \infty$. Then,*

$$\sup_{B\in\mathcal{B}(\mathcal{X})} \left|\left(P^{[m]} - P_0\right)(B)\right| \to 0 \text{ as } m \to \infty.$$

Proof. We first consider the case of μ being the Lebesgue measure. Suppose the conclusion is false, that is, there exists $\xi > 0$ and a sequence $\{B_m \in \mathcal{B}(\mathcal{X}) : m = 1,2,\dots\}$ such that $\limsup_{m\to\infty} \left|\left(P^{[m]} - P_0\right)(B_m)\right| > \xi$. By uniform tightness of Condition-RG (b), there exists a compact set $K \in \mathcal{B}(\mathcal{X})$ such that

$$\lim_{m\to\infty} \sup \left|\left(P^{[m]} - P_0\right)(B_m \cap K)\right| > \xi/2$$

holds. Let $\{b_m\}$ be a subsequence of $\{m\}$ such that $\left|\left(P^{[b_m]} - P_0\right)\left(B_{b_m} \cap K\right)\right| > \xi/2$ holds for all $b_m \geq b_m^*$. We metricize $\mathcal{B}(\mathcal{X})$ by the L_1-metric, $d_{\mathcal{B}(\mathcal{X})}(B, B') = (\mu \times \delta_d)(B \triangle B')$, where μ is the measure defined in Condition-RG (a) and δ_d is the mass measure on $d \in \{0, 1\}$. Since $\{B_{b_m} \cap K : m = 1, 2 \ldots, \}$ is a sequence in a compact subset of $\mathcal{B}(\mathcal{X})$, there exists a subsequence c_{b_m} of b_m, such that$\left\{B_{c_{b_m}} \cap K\right\}$ converges to $B^* \in \mathcal{B}(\mathcal{X})$ in terms of metric $d_{\mathcal{B}(\mathcal{X})}(\cdot, \cdot)$, and

$$\left|\left(P^{[c_{b_m}]} - P_0\right)\left(B_{c_{b_m}} \cap K\right)\right| > \xi/2 \tag{B.1}$$

holds by the construction of $\{b_m\}$ for all $c_{b_m} \geq c_{b_m^*}$. Under the bounded density assumption of Condition-RG (a), it holds that

$$\begin{aligned} & \left|\left(P^{[c_{b_m}]} - P_0\right)\left(B_{c_{b_m}} \cap K\right) - \left(P^{[c_{b_m}]} - P_0\right)(B^*)\right| \\ \leq \;& 2M d_{\mathcal{B}(\mathcal{X})}(B_{c_{b_m}} \cap K, B^*) \to 0, \text{ as } m \to \infty. \end{aligned}$$

Hence, (B.1) implies

$$\limsup_{m \to \infty} \left|\left(P^{[c_{b_m}]} - P_0\right)(B^*)\right| > \xi/2. \tag{B.2}$$

Since μ is the Lebesgue measure and, by Condition-RG (a), P_0 as a weak limit of $\left\{P^{[m]} : m = 1, 2, \ldots\right\}$ is absolutely continuous in $\mu \times \delta_d$, we have $P_0(\delta B^*) = 0$ where δB^* is the boundary of B^*. Accordingly, by applying the Portmanteau theorem (see, e.g., Theorem 1.3.4 of van der Vaart and Wellner (1996)), we obtain $\lim_{m\to\infty}\left|\left(P^{[m]} - P_0\right)(B^*)\right| = 0$. This contradicts (B.2). Hence, $\lim_{m\to\infty}\sup_{B\in\mathcal{B}(\mathcal{X})}\left|\left(P^{[m]} - P_0\right)(B)\right| = 0$ holds.

When μ is a discrete mass measure with finite support points, then the weak convergence of $P^{[m]}$ to P_0 is equivalent to the point wise convergence of the probability mass functions, and the $\sup_{B\in\mathcal{B}(\mathcal{X})}\left|\left(P^{[m]} - P_0\right)(B)\right|$ is equivalent to the supremum over power sets of the finite support points. Hence, the claim follows.

For the case of μ being a mixture of the Lebesgue and a discrete mass measure with finite support points, the claim holds as an immediate corollary of each of the two cases already shown. ■

Lemma B.3 *Suppose Condition-RG. Let $\left\{P^{[m]} \in \mathcal{P} : m = 1, 2 \ldots\right\}$ be a sequence of data generating processes on $\mathcal{X}$ that weakly converges to $P_0 \in \mathcal{P}$ as $m \to \infty$.*

$$\sup_{f\in\mathcal{F}}\left|\left(P_m^{[m]} - P_0\right)(f)\right| \underset{P^{[m]}}{\longrightarrow} 0.$$

Proof. This lemma is a corollary of Lemma B.1 and B.2. ■

Lemma B.4 *Suppose Condition-RG. Let* $\left\{(P^{[m(N)]}, Q^{[n(N)]}) \in \mathcal{P}^2 : N = 1, 2, \dots\right\}$ *be a sequence of two-sample probability measures with sample size* $(m, n) = (m(N), n(N)) \to (\infty, \infty)$ *as* $N \to \infty$. *We have*

$$\sup_{f,g\in\mathcal{F}} \left|\sigma^2_{P_m^{[m]},Q_n^{[m]}}(f,g) - \sigma^2_{P^{[m]},Q^{[m]}}(f,g)\right| \underset{P^{[m]},Q^{[n]}}{\longrightarrow} 0.$$

Proof. Consider

$$\begin{aligned} & \left|\sigma^2_{P_m^{[m]},Q_n^{[m]}}(f,g) - \sigma^2_{P^{[m]},Q^{[m]}}(f,g)\right| \qquad \text{(B.3)} \\ \leq\ & (1-\lambda)\left|P_m^{[m]}(fg) - P_m^{[m]}(f)P_m^{[m]}(g) - P^{[m]}(fg) + P^{[m]}(f)P^{[m]}(g)\right| \\ & +\lambda\left|Q_n^{[n]}(fg) - Q_n^{[n]}(f)Q_n^{[n]}(g) - Q^{[n]}(fg) + Q^{[n]}(f)Q^{[n]}(g)\right| + o(1), \end{aligned}$$

where $o(1)$ is the approximation error of order $\left|\hat{\lambda} - \lambda\right|$. Regarding the first term in the right-hand side of this inequality, the following inequalities hold,

$$\begin{aligned} & (1-\lambda)\left|P_m^{[m]}(fg) - P_m^{[m]}(f)P_m^{[m]}(g) - P^{[m]}(fg) + P^{[m]}(f)P^{[m]}(g)\right| \\ \leq\ & \left|\left(P_m^{[m]} - P^{[m]}\right)(fg)\right| + \left|P_m^{[m]}(f)P_m^{[m]}(g) - P^{[m]}(f)P^{[m]}(g)\right| \\ \leq\ & \left|\left(P_m^{[m]} - P^{[m]}\right)(fg)\right| + \left|\left(P_m^{[m]} - P^{[m]}\right)(f)P_m^{[m]}(g)\right| + \left|\left(P_m^{[m]} - P^{[m]}\right)(g)P^{[m]}(f)\right| \\ \leq\ & \left|\left(P_m^{[m]} - P^{[m]}\right)(fg)\right| + \left|\left(P_m^{[m]} - P^{[m]}\right)(f)\right| + \left|\left(P_m^{[m]} - P^{[m]}\right)(g)\right|. \qquad \text{(B.4)} \end{aligned}$$

The second and the third term of (B.4) is $o_{P^{[m]}}(1)$ uniformly in $\mathcal{F}$ by Lemma B.1. Furthermore, since class of indicator functions $\{fg : f, g \in \mathcal{F}\}$ is also a VC-class,

$$\sup_{f,g\in\mathcal{F}} \left|\left(P_m^{[m]} - P^{[m]}\right)(fg)\right| \underset{P^{[m]}}{\longrightarrow} 0$$

holds also by Lemma B.1. This proves the first term in the right-hand side of (B.3) converges to zero uniformly in $f, g \in \mathcal{F}$. So is the case for the second term of (B.3) by the same argument. Hence, the conclusion follows. ■

Lemma B.5 *Suppose Condition-RG. Let* $\left\{P^{[m]} \in \mathcal{P} : m = 1, 2, \dots\right\}$ *be a sequence of probability measures, which converges weakly to* $P_0 \in \mathcal{P}$. *Then, the empirical processes* $G_{m,P^{[m]}}(\cdot)$ *on index set* $\mathcal{F}$ *converge weakly to* P_0*-brownian bridges* $G_{P_0}(\cdot)$.

Proof. To prove this lemma, we apply a combination of Theorem 2.8.3 and Lemma 2.8.8 of van der Vaart and Wellner (1996) restricted to a class of indicator functions. It claims that, given $\mathcal{F}$ be a class of measurable indicator functions and a sequence of probability measure $\left\{P^{[m]} : m = 1, 2, \dots\right\}$ in $\mathcal{P}$, if (i) $\int_0^1 \sup_R \sqrt{\log N\left(\epsilon, \mathcal{F}, L_2\left(R\right)\right)} d\epsilon < \infty$, where R ranges over all finitely discrete probability measures and $N\left(\epsilon, \mathcal{F}, L_2\left(R\right)\right)$ is the covering number of $\mathcal{F}$ with radius ϵ in terms of $L_2\left(R\right)$-metric $\left[R(|f - f'|^2)\right]^{1/2}$,[1] and (ii) there exists $P^* \in \mathcal{P}$ such that $\lim_{m\to\infty} \sup_{f,g\in\mathcal{F}} \left\{|\rho_{P^{[m]}}(f, g) - \rho_{P^*}(f, g)|\right\} = 0$, then $G_{m,P^{[m]}}\left(\cdot\right)$ weakly converges to P^*-brownian bridge process $G_{P^*}\left(\cdot\right)$. Condition (i) is known to hold if $\mathcal{F}$ is a VC-class (see Theorem 2.6.4 of van der Vaart and Wellner (1996)).

Therefore, what remains to show is Condition (ii). By the construction of seminorm $\rho_P(f, g)$, we have

$$\sup_{f,g\in\mathcal{F}} \left|\rho^2_{P^{[m]}}(f, g) - \rho^2_{P_0}(f, g)\right| \leq \sup_{B\in\mathcal{B}(\mathcal{X})} \left|\left(P^{[m]} - P_0\right)\left(B\right)\right|.$$

Hence, to validate Condition (ii) with $P^* = P_0$, it suffices to have $\lim_{m\to\infty} \sup_{B\in\mathcal{B}(\mathcal{X})} \left|\left(P^{[m]} - P_0\right)\left(B\right)\right| = 0$, which follows from Lemma B.2. ■

Lemma B.6 *Suppose Condition-RG. Let* $\left\{(P^{[m(N)]}, Q^{[n(N)]}) \in \mathcal{P}^2 : N = 1, 2, \dots\right\}$ *be a sequence of probability measures of the independent two samples, which converges weakly to* (P_0, Q_0)*, as* $N \to \infty$*. Then, stochastic processes indexed by VC-class of indicator functions* $\mathcal{F}$*,*

$$v_N(\cdot) = \frac{\left(1 - \hat{\lambda}\right)^{1/2} G_{m,P^{[m]}}\left(\cdot\right) - \hat{\lambda}^{1/2} G_{n,Q^{[n]}}(\cdot)}{\xi \vee \sigma_{P_m^{[m]},Q_n^{[n]}}(\cdot, \cdot)}, \quad \xi > 0, \tag{B.5}$$

converges weakly to mean zero Gaussian processes $v_0(\cdot) = \frac{(1-\lambda)^{1/2} G_{P_0}(\cdot) - \lambda^{1/2} G_{Q_0}(\cdot)}{\xi \vee \sigma_{P_0,Q_0}(\cdot,\cdot)}$*, where* $G_{P_0}\left(\cdot\right)$ *and* $G_{Q_0}(\cdot)$ *are independent brownian bridge processes.*

Proof. VC-class $\mathcal{F}$ is totally bounded with seminorm ρ_P for any finite measure P. Hence, following Section 2.8.3 of van der Vaart and Wellner (1996), what we want to show for the weak convergence of $v_N(\cdot)$ are that (i) finite dimensional marginal, $(v_N(f_1), \dots, v_N(f_K))$, converges to that of $v_0(\cdot)$, (ii) $v_N(\cdot)$ is asymptotically uniformly equicontinuous along a sequence

[1] The covering number $N\left(\epsilon, \mathcal{F}, L_2\left(R\right)\right)$ is defined as the minimal number of balls of radious ϵ needed to cover $\mathcal{F}$.

of seminorms such as $L_2(P^{[m]}+Q^{[n]})$ norm, $\rho_{P^{[m]}+Q^{[n]}}(f,g)=\left[\left(P^{[m]}+Q^{[n]}\right)\left((f-g)^2\right)\right]^{1/2}$, i.e., for arbitrary $\epsilon>0$,

$$\lim_{\delta\searrow 0}\limsup_{N\to\infty} P^*_{P^{[m]},Q^{[n]}}\left(\sup_{\rho_{P^{[m]}+Q^{[n]}}(f,g)<\delta}|v_N(f)-v_N(g)|>\epsilon\right)=0, \tag{B.6}$$

where $P^*_{P^{[m]},Q^{[n]}}$ is the outer probability, and (iii) $\sup_{f,g\in\mathcal{F}}\left|\rho_{P^{[m]}+Q^{[n]}}(f,g)-\rho_{P_0+Q_0}(f,g)\right|\to 0$ as $N\to\infty$. Note that (i) is implied by Lemma B.4 and Lemma B.5, and (iii) follows as a corollary of Lemma B.2, since

$$\begin{aligned}\sup_{f,g\in\mathcal{F}}\left|\rho^2_{P^{[m]}+Q^{[n]}}(f,g)-\rho^2_{P_0+Q_0}(f,g)\right| &\le \sup_{B\in\mathcal{B}(\mathcal{X})}\left|\left(P^{[m]}-P_0\right)(B)\right|+\sup_{B\in\mathcal{B}(\mathcal{X})}\left|\left(Q^{[n]}-Q_0\right)(B)\right|\\ &\to 0 \quad \text{as } N\to\infty.\end{aligned}$$

To verify (ii), consider, for $f,g\in\mathcal{F}$ with $\rho_{P^{[m]}+Q^{[n]}}(f,g)<\delta$,

$$\begin{aligned}&|v_N(f)-v_N(g)| \\ \le\ & \left|\frac{1}{\xi\vee\sigma_{P^{[m]}_m,Q^{[n]}_n}(f,f)}-\frac{1}{\xi\vee\sigma_{P^{[m]}_m,Q^{[n]}_n}(g,g)}\right|\left|(1-\lambda)^{1/2}G_{m,P^{[m]}}(g)-\lambda^{1/2}G_{n,Q^{[n]}}(g)\right| \\ &+\frac{(1-\lambda)^{1/2}\left|G_{m,P^{[m]}}(f)-G_{m,P^{[m]}}(g)\right|+\lambda^{1/2}\left|G_{n,Q^{[n]}}(f)-G_{n,Q^{[n]}}(g)\right|}{\xi\vee\sigma_{P^{[m]}_m,Q^{[n]}_n}(g,g)} \\ &+o\left(\left|\hat\lambda-\lambda\right|\right).\end{aligned} \tag{B.7}$$

Note that

$$\begin{aligned}&\left|\frac{1}{\xi\vee\sigma_{P^{[m]}_m,Q^{[n]}_n}(f,f)}-\frac{1}{\xi\vee\sigma_{P^{[m]}_m,Q^{[n]}_n}(g,g)}\right| \\ =\ &\left|\frac{1}{\xi\vee\sigma_{P^{[m]},Q^{[n]}}(f,f)}-\frac{1}{\xi\vee\sigma_{P^{[m]},Q^{[n]}}(g,g)}\right|+o_{P^{[m]},Q^{[n]}}(1) \\ \le\ &\frac{1}{\xi^2}\left|\xi\vee\sigma_{P^{[m]},Q^{[n]}}(f,f)-\xi\vee\sigma_{P^{[m]},Q^{[n]}}(g,g)\right|+o_{P^{[m]},Q^{[n]}}(1) \\ \le\ &\frac{1}{\xi^2}\left|\sigma_{P^{[m]},Q^{[n]}}(f,f)-\sigma_{P^{[m]},Q^{[n]}}(g,g)\right|+o_{P^{[m]},Q^{[n]}}(1),\end{aligned} \tag{B.8}$$

where the first line follows from Lemma B.4. By noting the following inequalities,

$$\begin{aligned}&\left|\sigma_{P^{[m]},Q^{[m]}}(f,f)-\sigma_{P^{[m]},Q^{[m]}}(g,g)\right|^2 \\ \le\ &\left|\sigma_{P^{[m]}_m,Q^{[n]}_n}(f,f)-\sigma_{P^{[m]},Q^{[m]}}(g,g)\right|\left|\sigma_{P^{[m]}_m,Q^{[n]}_n}(f,f)+\sigma_{P^{[m]},Q^{[m]}}(g,g)\right| \\ =\ &\left|\sigma^2_{P^{[m]},Q^{[m]}}(f,f)-\sigma^2_{P^{[m]},Q^{[m]}}(g,g)\right|\end{aligned}$$

and

$$\begin{aligned}\left|\sigma^2_{P^{[m]},Q^{[m]}}(f,f)-\sigma^2_{P^{[m]},Q^{[m]}}(g,g)\right| &\leq \left|(1-\lambda)\left(P^{[m]}(f)-P^{[m]}(g)\right)\left(1-P^{[m]}(f)-P^{[m]}(g)\right)\right| \\ &\quad+\left|\lambda\left(Q^{[n]}(f)-Q^{[n]}(g)\right)\left(1-Q^{[n]}(f)-Q^{[n]}(g)\right)\right| \\ &\leq \left|(1-\lambda)\left(P^{[m]}(f)-P^{[m]}(g)\right)\right|+\left|\lambda\left(Q^{[n]}(f)-Q^{[n]}(g)\right)\right| \\ &\leq (1-\lambda)\rho^2_{P^{[m]}}(f,g)+\lambda\rho^2_{Q^{[n]}}(f,g) \\ &\leq \rho^2_{P^{[m]}+Q^{[n]}}(f,g),\end{aligned}$$

we have

$$\left|\sigma_{P^{[m]},Q^{[m]}}(f,f)-\sigma_{P^{[m]},Q^{[m]}}(g,g)\right|\leq\rho_{P^{[m]}+Q^{[n]}}(f,g). \tag{B.9}$$

Combining (B.8) and (B.9) then leads to

$$\left|\frac{1}{\xi\vee\sigma_{P^{[m]}_m,Q^{[n]}_n}(f,f)}-\frac{1}{\xi\vee\sigma_{P^{[m]}_m,Q^{[n]}_n}(g,g)}\right|\leq\frac{\rho_{P^{[m]}+Q^{[n]}}(f,g)}{\xi^2}+o_{P^{[m]},Q^{[m]}}(1) \tag{B.10}$$

Hence, (B.7) and (B.10) yield

$$\begin{aligned}\sup_{\rho_{P^{[m]}+Q^{[n]}}(f,g)<\delta}|v_N(f)-v_N(g)| &\leq \frac{\delta}{\xi^2}\left|(1-\lambda)^{1/2}G_{m,P^{[m]}}(g)-\lambda^{1/2}G_{n,Q^{[n]}}(g)\right| \\ &\quad+\frac{(1-\lambda)^{1/2}}{\xi}\sup_{\rho_{P^{[m]}+Q^{[n]}}(f,g)<\delta}\left|G_{m,P^{[m]}}(f)-G_{m,P^{[m]}}(g)\right| \\ &\quad+\frac{\lambda^{1/2}}{\xi}\sup_{\rho_{P^{[m]}+Q^{[n]}}(f,g)<\delta}\left|G_{n,Q^{[n]}}(f)-G_{n,Q^{[n]}}(g)\right|+o_{P^{[m]},Q^{[m]}}(1).\end{aligned} \tag{B.11}$$

Since $\rho_{P^{[m]}}(f,g)\leq\rho_{P^{[m]}+Q^{[n]}}(f,g)$ for every $f,g\in\mathcal{F}$, we have

$$\begin{aligned}\sup_{\rho_{P^{[m]}+Q^{[n]}}(f,g)<\delta}\left|G_{m,P^{[m]}}(f)-G_{m,P^{[m]}}(g)\right| &\leq \sup_{\rho_{P^{[m]}}(f,g)<\delta}\left|G_{m,P^{[m]}}(f)-G_{m,P^{[m]}}(g)\right| \\ &= o^*_{P^{[m]}}(\delta),\end{aligned}$$

where $o^*_{P^{[m]}}(\delta)$ denotes the convergence to zero in outer probability along $\left\{P^{[m]}\right\}$ as $\delta\searrow 0$, and the equality follows since the uniform convergence of $G_{m,P^{[m]}}(f)$ as established by Lemma B.5 implies

$$\lim_{\delta\searrow 0}\limsup_{m\to\infty}P^*_{P^{[m]}}\left(\sup_{\rho_{P^{[m]}}(f,g)<\delta}\left|G_{m,P^{[m]}}(f)-G_{m,P^{[m]}}(g)\right|>\epsilon\right)=0.$$

Similarly, we obtain $\sup_{\rho_{P^{[m]}+Q^{[n]}}(f,g)<\delta}\left|G_{n,Q^{[n]}}(f)-G_{n,Q^{[n]}}(g)\right|=o^*_{Q^{[n]}}(\delta)$.

Since $\left|(1-\lambda)^{1/2} G_{m,P^{[m]}}(g) - \lambda^{1/2} G_{n,Q^{[n]}}(g)\right|$ converges weakly to the tight Gaussian processes, (B.11) is written as

$$\begin{aligned}\sup_{\rho_{P^{[m]}+Q^{[n]}}(f,g)<\delta} |v_N(f) - v_N(g)| &= \delta O_{P^{[m]},Q^{[n]}}(1) + o^*_{P^{[m]},Q^{[m]}}(\delta) + o_{P^{[m]},Q^{[m]}}(1) \\ &= o^*_{P^{[m]},Q^{[m]}}(\delta)\end{aligned}$$

where $O_{P^{[m]},Q^{[n]}}(1)$ stands for that $\lim_{N\to\infty} \Pr_{P^{[m]},Q^{[n]}}(|W_N| > a_N) = 0$ for every diverging sequence $a_N \to \infty$. This establishes the asymptotic uniform equicontinuity (B.6). ■

The next lemma states that the null hypothesis of our test defined by inequalities (1.1) for every Borel set B can be reduced without loss of information to the hypothesis that inequalities (1.1) hold for all connected intervals. This lemma is a direct corollary of Lemma C1 in Andrews and Shi (2013).

Lemma B.7 *$P(B,1) - Q(B,1) \geq 0$ and $Q(B,0) - P(B,0) \geq 0$ hold for every Borel set B if and only if $P(V,1) - Q(V,1) \geq 0$ and $Q(V,0) - P(V,0) \geq 0$ hold for all $V \in \mathcal{V} \equiv \{[y,y'] : -\infty \leq y \leq y' \leq \infty\}$.*

Proof. The only-if statement is obvious. To prove the if statement, we apply Lemma C1 of Andrews and Shi (2013). By viewing $\mathcal{V}$ as $\mathcal{R}$ and $P(\cdot,1) - Q(\cdot,1)$ as $\mu(\cdot)$ in the notation of Lemma C1 of Andrews and Shi (2013), it follows that $P(B,1) - Q(B,1) \geq 0$ for all B in the Borel σ-algebra generated by $\mathcal{V}$. Since the Borel σ-algebra generated by $\mathcal{V}$ coincides with $\mathcal{B}(\mathcal{Y})$, $P(V,1) - Q(V,1) \geq 0$ for every $V \in \mathcal{V}$ implies $P(B,1) - Q(B,1) \geq 0$ for every $B \in \mathcal{B}(\mathcal{Y})$. The same results hold for the other inequalities $Q(\cdot,0) - P(\cdot,0) \geq 0$. ■

The next lemma shows that the version of testable implications with conditioning covariates as given in (3.3) can be reduced without any loss of information to the unconditional moment inequalities of (3.4).

Lemma B.8 *Assume that $\Pr(Z=1|X)$ is bounded away from zero and one, X-a.s. Then,*

$$\begin{aligned}&\Pr(Y \in B, D = 1|Z = 1, X) - \Pr(Y \in B, D = 1|Z = 0, X) \geq 0, \\ &\Pr(Y \in B, D = 0|Z = 0, X) - \Pr(Y \in B, D = 0|Z = 1, X) \geq 0.\end{aligned} \tag{B.12}$$

hold for all $B \in \mathcal{B}(\mathcal{Y})$, X-a.s. if and only if

$$\begin{aligned} E\left[\kappa_1\left(D,Z,X\right)g(Y,X)\right] &\geq 0, \\ E\left[\kappa_0\left(D,Z,X\right)g(Y,X)\right] &\geq 0, \quad \textit{for all } g(\cdot,\cdot) \in \mathcal{G}, \end{aligned}$$

where κ_1, κ_0, and $\mathcal{G}$ are as defined in Section 3.2 of the main text.

Proof. By applying Theorem 3.1 of Abadie (2003) with conditioning of X, the first inequalities of (B.12) can be equivalently written as

$$E\left[1\left\{Y \in B\right\}\kappa_1(D,Z,X)|X\right] \geq 0, \; X\text{-a.s.} \tag{B.13}$$

Hence, the only-if statement immediately follows.

To show the if statement, we again invoke Lemma C1 in Andrews and Shi (2013). Let us read $\mathcal{R}$ and $\mu\left(\cdot\right)$ of their notation as

$$\mathcal{V} \equiv \left\{ \begin{array}{c} [y,y'] \times [x_1,x_1'] \times \cdots \times \left[x_{d_x},x_{d_x}'\right] : -\infty \leq y \leq y' \leq \infty, \\ -\infty \leq x_l \leq x_l' \leq \infty, \quad l = 1,\ldots,d_x \end{array} \right\},$$

and $\mu\left(\cdot\right) = E\left[\kappa_1\left(D,Z,X\right)1\left\{(Y,X) \in \cdot\right\}\right]$, respectively. By the assumption that $\Pr(Z = 1|X)$ is bounded away from zero and one, κ_1 is bounded X-a.s. Hence, the thus-defined $\mu\left(\cdot\right)$ satisfies the boundedness condition to apply Lemma C1 in Andrews and Shi (2013). Moreover, $\mathcal{V}$ meets the condition for a semiring. Hence, $\mu\left(V\right) = E\left[\kappa_1\left(D,Z,X\right)1\left\{(Y,X) \in V\right\}\right] \geq 0$ for all $V \in \mathcal{V}$ implies $\mu\left(C\right) = E\left[\kappa_1\left(D,Z,X\right)1\left\{(Y,X) \in C\right\}\right] \geq 0$ for all C in the Borel σ-algebra generated by $\mathcal{V}$. Since the Borel σ-algebra generated by $\mathcal{V}$ coincides with $\mathcal{B}(\mathcal{Y}\times\mathbb{X})$, and any product set $B \times V_x$, $B \in \mathcal{B}(\mathcal{Y})$ and $V_x \in \mathcal{B}(\mathbb{X})$, belongs to $\mathcal{B}(\mathcal{Y}\times\mathbb{X})$, it implies $E\left[1\left\{Y \in B\right\}\kappa_1(D,Z,X)1\left\{X \in V_x\right\}\right] \geq 0$ for all $B \in \mathcal{B}(\mathcal{Y})$ and $V_x \in \mathcal{B}(\mathbb{X})$. Hence, (B.13) follows. A similar line of reasoning yields the equivalence of the second inequalities of (B.12) to $E\left[\kappa_0\left(D,Z,X\right)g(Y,X)\right] \geq 0$ for all $g(\cdot,\cdot) \in \mathcal{G}$. ■

B.3 Proof of Theorem 2.1

Let $\mathcal{F}_1 = \left\{1_{\{[y,y'],1\}}(Y,D) : -\infty \leq y \leq y' \leq \infty\right\}$ and $\mathcal{F}_0 = \left\{1_{\{[y,y'],0\}}(Y,D) : -\infty \leq y \leq y' \leq \infty\right\}$. We want to show

$$\limsup_{N\to\infty} \sup_{(P,Q)\in\mathcal{H}_0} \Pr\left(T_N > c_{N,1-\alpha}\right) \leq \alpha, \tag{B.14}$$

where

$$T_N = \max\left\{ \begin{array}{c} \sup_{f\in\mathcal{F}_1}\left\{\frac{\hat{\lambda}^{1/2}Q_n(f)-(1-\hat{\lambda})^{1/2}P_m(f)}{\xi\vee\sigma_{P_m,Q_n}(f,f)}\right\} \\ \sup_{f\in\mathcal{F}_0}\left\{\frac{\left((1-\hat{\lambda})^{1/2}P_m(f)-\hat{\lambda}^{1/2}Q_n(f)\right)}{\xi\vee\sigma_{P_m,Q_n}(f,f)}\right\} \end{array} \right\}.$$

Consider a sequence $\left(P^{[m(N)]}, Q^{[n(N)]}\right) \in \mathcal{H}_0$ at which $\Pr_{P^{[m(N)]},Q^{[n(N)]}}\left(T_N > c_{N,1-\alpha}\right)$ differs from its supremum over $\mathcal{H}_0$ by $\epsilon_N > 0$ or less with $\epsilon_N \to 0$ as $N \to \infty$. Since $\left(P^{[m(N)]}, Q^{[n(N)]}\right) \in \mathcal{P}^2$ are sequences in the uniformly tight class of probability measures (Condition-RG (b)), there exists a_N subsequence of N such that $\left(P^{[m(a_N)]}, Q^{[n(a_N)]}\right)$ converges weakly to $(P_0, Q_0) \in \mathcal{P}^2$ as $N \to \infty$. Note that (P_0, Q_0) lies in $\mathcal{H}_0$ since $\left(P^{[m(N)]}, Q^{[n(N)]}\right) \in \mathcal{H}_0$ for all N and by Lemma B.2. With abuse of notations, we read a_N as N and $(m(a_N), n(a_N))$ as (m, n) with $m+n = N$. Along such sequence, we aim to show $\limsup_{N\to\infty} \Pr_{P^{[m]},Q^{[n]}}\left(T_N > c_{N,1-\alpha}\right) \leq \alpha$ holds.

Using the notation of the weighted empirical processes introduced in Lemma B.6, we can write the test statistic as

$$T_N = \max\left\{ \begin{array}{c} \sup_{f\in\mathcal{F}_1}\left\{-v_N(f) - h_N(f)\right\} \\ \sup_{f\in\mathcal{F}_0}\left\{v_N(f) + h_N(f)\right\} \end{array} \right\},$$

where

$$h_N(f) = \sqrt{\frac{mn}{N}}\frac{P^{[m]}(f) - Q^{[n]}(f)}{\xi \vee \sigma_{P_m^{[m]},Q_n^{[n]}}(f,f)}, \;\; d = 1, 0.$$

By the almost sure representation theorem (see, e.g., Theorem 9.4 of Pollard (1990)), weak convergence of $\left(v_N(\cdot), P_m^{[m]}(\cdot), Q_n^{[n]}(\cdot), \sigma^2_{P_m^{[m]},Q_n^{[n]}}(\cdot,\cdot)\right)$ to $\left(v_0(\cdot), P_0(\cdot), Q_0(\cdot), \sigma^2_{P_0,Q_0}(\cdot,\cdot)\right)$, as established in Lemma B.3, B.4, and B.6, implies existence of a probability space $(\Omega, \mathcal{B}(\Omega), \mathbb{P})$ and random objects $\tilde{v}_0(\cdot)$, $\tilde{v}_N(\cdot)$, $\tilde{P}_m^{[m]}(\cdot)$, $\tilde{Q}_n^{[n]}(\cdot)$, and $\tilde{\sigma}^2_{P_m^{[m]},Q_m^{[n]}}(\cdot,\cdot)$ defined on it, such that (i) $\tilde{v}_0(\cdot)$ has the same probability law as $v_0(\cdot)$ (ii) $\left(\tilde{v}_N(\cdot), \tilde{P}_m^{[m]}(\cdot), \tilde{Q}_n^{[n]}(\cdot), \tilde{\sigma}^2_{P_m^{[m]},Q_n^{[n]}}(\cdot,\cdot)\right)$ has the same probability law as $\left(v_N(\cdot), P_m^{[m]}(\cdot), Q_n^{[n]}(\cdot), \sigma^2_{P_m^{[m]},Q_n^{[n]}}(\cdot,\cdot)\right)$ for all N, and (iii)

$$\sup_{f\in\mathcal{F}}\left|\tilde{v}_N(f) - \tilde{v}_0(f)\right| \;\to\; 0, \tag{B.15}$$

$$\sup_{f\in\mathcal{F}}\left|\tilde{P}_m^{[m]}(f) - P_0(f)\right| \;\to\; 0, \tag{B.16}$$

$$\sup_{f\in\mathcal{F}}\left|\tilde{Q}_n^{[n]}(f) - Q_0(f)\right| \;\to\; 0, \text{ and} \tag{B.17}$$

$$\sup_{f,g\in\mathcal{F}}\left|\tilde{\sigma}^2_{P_m^{[m]},Q_m^{[n]}}(f,g) - \sigma^2_{P_0,Q_0}(f,g)\right| \;\to\; 0, \text{ as } N\to\infty, \mathbb{P}\text{-a.s.} \tag{B.18}$$

Let $\tilde{T}_N$ be the analogue of T_N defined on probability space $(\Omega, \mathcal{B}(\Omega), \mathbb{P})$,

$$\tilde{T}_N = \max \left\{ \begin{array}{c} \sup_{f \in \mathcal{F}_1} \left\{ -\tilde{v}_N(f) - \tilde{h}_N(f) \right\} \\ \sup_{f \in \mathcal{F}_0} \left\{ \tilde{v}_N(f) + \tilde{h}_N(f) \right\} \end{array} \right\},$$

where $\tilde{h}_N(f) = \sqrt{\frac{mn}{N}} \frac{P^{[m]}(f) - Q^{[n]}(f)}{\xi \vee \tilde{\sigma}^2_{P_m^{[m]}, Q_n^{[n]}}(f,f)}$. Let $\tilde{c}_{N,1-\alpha}$ be the bootstrap critical values, which we view as a random object defined on the same probability space as $\left(\tilde{v}_N, \tilde{P}_m^{[m]}, \tilde{Q}_n^{[n]}, \tilde{\sigma}^2_{P_m^{[m]}, Q_n^{[n]}} \right)$ are defined. Note that the probability law of $\tilde{c}_{N,1-\alpha}$ under $\mathbb{P}$ is identical to the probability law of bootstrap critical value $c_{N,1-\alpha}$ under $\left(P^{[m]}, Q^{[n]} \right)$ for every N, because the distributions of $\tilde{c}_{N,1-\alpha}$ and $c_{N,1-\alpha}$ are determined by the distributions of $\left(\tilde{P}_m^{[m]}, \tilde{Q}_n^{[n]} \right)$ and $\left(P_m^{[m]}, Q_n^{[n]} \right)$, respectively, and $\left(\tilde{P}_m^{[m]}, \tilde{Q}_n^{[n]} \right) \sim \left(P_m^{[m]}, Q_n^{[n]} \right)$ for every N, as claimed by the almost sure representation theorem.

By the Lemma C.1 shown below, $\tilde{c}_{N,1-\alpha} \to c_{1-\alpha}$ as $N \to \infty$, $\mathbb{P}$-a.s., where $c_{1-\alpha}$ is the $(1-\alpha)$-th quantile of statistic

$$T_H \equiv \max \left\{ \begin{array}{c} \sup_{f \in \mathcal{F}_1} \left\{ -G_{H_0}(f) / \left(\xi \vee \sigma_{H_0}(f,f) \right) \right\} \\ \sup_{f \in \mathcal{F}_0} \left\{ G_{H_0}(f) / \left(\xi \vee \sigma_{H_0}(f,f) \right) \right\} \end{array} \right\}, \tag{B.19}$$

where $H_0 = \lambda P_0 + (1-\lambda) Q_0$.

Since $\Pr_{P^{[m]}, Q^{[n]}} \left(T_N > c_{N,1-\alpha} \right) = P\left(\tilde{T}_N > \tilde{c}_{N,1-\alpha} \right)$ for all N and $\tilde{c}_{N,1-\alpha} \to c_{1-\alpha}$ as $N \to \infty$, $\mathbb{P}$-a.s, if there exists a random variable $\tilde{T}^*$ defined on $(\Omega, \mathcal{B}(\Omega), \mathbb{P})$, such that

(A) : $\limsup_{N \to \infty} \tilde{T}_N \leq \tilde{T}^*$, $\mathbb{P}$-a.s., and

(B) : The cdf of $\tilde{T}^*$ is continuous at $c_{1-\alpha}$ and $\mathbb{P}\left(\tilde{T}^* > c_{1-\alpha} \right) \leq \alpha$,

then, the claim of the proposition follows from

$$\begin{aligned} \limsup_{N \to \infty} \Pr_{P^{[m]}, Q^{[n]}} \left(T_N > c_{N,1-\alpha} \right) &= \limsup_{N \to \infty} \mathbb{P}\left(\tilde{T}_N > \tilde{c}_{N,1-\alpha} \right) \\ &\leq \mathbb{P}\left(\tilde{T}^* > c_{1-\alpha} \right) \\ &\leq \alpha, \end{aligned}$$

where the second line follows from Fatou's lemma. Hence, in what follows, we aim to find a random variable $\tilde{T}^*$ that satisfies (A) and (B).

Let η_N be a deterministic sequence that satisfies $\eta_N \to \infty$ and $\eta_N/\sqrt{N} \to 0$. Fix $\omega \in \Omega$ and define a sequence of subclass of $\mathcal{F}_1$,

$$\begin{aligned}\mathcal{F}_{1,\eta_N} &= \left\{ f \in \mathcal{F}_1 : \tilde{h}_N(f) \leq \eta_N \right\} \\ &= \left\{ f \in \mathcal{F}_1 : \sqrt{\hat{\lambda}(1-\hat{\lambda})} \frac{P^{[m]}(f) - Q^{[n]}(f)}{\xi \vee \tilde{\sigma}^2_{P_m^{[m]}, Q_n^{[n]}}(f,f)} \leq \frac{\eta_N}{\sqrt{N}} \right\}.\end{aligned}$$

The first term in the maximum operator of $\tilde{T}_N$ satisfies

$$\begin{aligned}\sup_{f \in \mathcal{F}_1} \left\{ -\tilde{v}_N(f) - \tilde{h}_{1,N}(f) \right\} &= \max \left\{ \begin{array}{c} \sup_{f \in \mathcal{F}_{1,\eta_N}} \left\{ -\tilde{v}_N(f) - \tilde{h}_N(f) \right\} \\ \sup_{f \in \mathcal{F}_1 \setminus \mathcal{F}_{1,\eta_N}} \left\{ -\tilde{v}_N(f) - \tilde{h}_N(f) \right\} \end{array} \right\} \\ &\leq \max \left\{ \begin{array}{c} \sup_{f \in \mathcal{F}_{1,\eta_N}} \left\{ -\tilde{v}_N(f) \right\} \\ \sup_{f \in \mathcal{F}_1 \setminus \mathcal{F}_{1,\eta_N}} \left\{ -\tilde{v}_N(f) - \tilde{h}_N(f) \right\} \end{array} \right\} \\ &\leq \max \left\{ \begin{array}{c} \sup_{f \in \bigcup_{N' \geq N} \mathcal{F}_{1,\eta_{N'}}} \left\{ -\tilde{v}_N(f) \right\} \\ \sup_{f \in \mathcal{F}_1 \setminus \mathcal{F}_{1,\eta_N}} \left\{ -\tilde{v}_N(f) \right\} - \eta_N \end{array} \right\}, \end{aligned} \tag{B.20}$$

for every N,where the second line follows since $\tilde{h}_{1,N}(f) \geq 0$ for all $f \in \mathcal{F}_1$ under the assumption that $\left(P^{[m]}, Q^{[n]}\right) \in H_0$, the third line follows because $\tilde{h}_N(f) > \eta_N$ for all $f \in \mathcal{F}_1 \setminus \mathcal{F}_{1,\eta_N}$. Since $\tilde{v}_N(\cdot)$ is $\mathbb{P}$-a.s. bounded and $\eta_N \to \infty$, it holds

$$\sup_{f \in \mathcal{F}_1 \setminus \mathcal{F}_{1,\eta_N}} \left\{ -\tilde{v}_N(f) \right\} - \eta_N \to -\infty, \text{ as } N \to \infty, \mathbb{P}\text{-a.s.} \tag{B.21}$$

On the other hand, since $\tilde{v}_N(\cdot)$ $\mathbb{P}$-a.s converges to $\tilde{v}_0(\cdot)$ uniformly in $\mathcal{F}$, we have

$$\sup_{f \in \bigcup_{N' \geq N} \mathcal{F}_{1,\eta_{N'}}} \left\{ -\tilde{v}_N(f) \right\} \to \sup_{f \in \mathcal{F}_{1,\infty}} \left\{ -\tilde{v}_0(f) \right\}, \text{ as } N \to \infty, \ \mathbb{P}\text{-a.s.}, \tag{B.22}$$

where $\mathcal{F}_{1,\infty} = \lim_{N \to \infty} \bigcup_{N' \geq N} \mathcal{F}_{1,\eta_{N'}}$. Let $\mathcal{F}_1^* = \{ f \in \mathcal{F}_1 : P_0(f) = Q_0(f) \}$. By the construction of $\mathcal{F}_{1,\eta_N}$, every $f \in \mathcal{F}_{1,\infty}$ satisfies

$$\liminf_{N \to \infty} \left\{ \sqrt{\hat{\lambda}(1-\hat{\lambda})} \frac{P^{[m]}(f) - Q^{[n]}(f)}{\xi \vee \tilde{\sigma}^2_{P_m^{[m]}, Q_n^{[n]}}(f,f)} \right\} = 0. \tag{B.23}$$

Since $P^{[m]}(f) - Q^{[n]}(f)$ converges to $P_0(f) - Q_0(f)$ by Lemma B.2, any f satisfying (B.23) belongs to $\mathcal{F}_1^*$. Hence, we have

$$\sup_{f \in \mathcal{F}_{1,\infty}} \left\{ -\tilde{v}_0(f) \right\} \leq \sup_{f \in \mathcal{F}_1^*} \left\{ -\tilde{v}_0(f) \right\} \quad \mathbb{P}\text{-a.s.} \tag{B.24}$$

By combining (B.20), (B.21), (B.22), and (B.24), we obtain

$$\limsup_{N\to\infty} \sup_{f\in\mathcal{F}_1} \left\{-\tilde{v}_N(f) - \tilde{h}_N(f)\right\} \leq \sup_{f\in\mathcal{F}_1^*} \left\{-\tilde{v}_0(f)\right\}, \ \mathbb{P}\text{-a.s.}$$

In a similar manner, it can be shown that

$$\limsup_{N\to\infty} \sup_{f\in\mathcal{F}_0} \left\{\tilde{v}_N(f) + \tilde{h}_N(f)\right\} \leq \sup_{f\in\mathcal{F}_0^*} \left\{\tilde{v}_0(f)\right\}, \ \mathbb{P}\text{-a.s.},$$

where $\mathcal{F}_0^* = \{f \in \mathcal{F}_0 : P_0(f) = Q_0(f)\}$. Hence, $\tilde{T}^*$ defined by

$$\tilde{T}^* = \max\left\{ \begin{array}{c} \sup_{f\in\mathcal{F}_1^*} \{-\tilde{v}_0(f)\} \\ \sup_{f\in\mathcal{F}_0^*} \{\tilde{v}_0(f)\} \end{array} \right\}$$

satisfies condition (A).

Next, we show that the thus-defined $\tilde{T}^*$ satisfies (B). First, we show that $\tilde{T}^*$ is stochastically dominated by T_H. Note that statistic T_H defined in (B.19) can be written as

$$\begin{aligned} T_H &= \max\left\{T_H^*, \sup_{f\in\mathcal{F}_1\setminus\mathcal{F}_1^*}\left\{-\frac{G_{H_0}(f)}{\xi\vee\sigma_{H_0}(f,f)}\right\}, \sup_{f\in\mathcal{F}_0\setminus\mathcal{F}_0^*}\left\{\frac{G_{H_0}(f)}{\xi\vee\sigma_{H_0}(f,f)}\right\}\right\}, \\ \text{where } T_H^* &= \max\left\{ \begin{array}{c} \sup_{f\in\mathcal{F}_1^*} \{-G_{H_0}(f)/(\xi\vee\sigma_{H_0}(f,f))\}, \\ \sup_{f\in\mathcal{F}_0^*} \{G_{H_0}(f)/(\xi\vee\sigma_{H_0}(f,f))\} \end{array} \right\}. \end{aligned}$$

If the distribution of T_H^* is identical to $\tilde{T}^*$, then the distribution of T_H stochastically dominates $\tilde{T}^*$ so that we can ascertain the second part of (B). Hence, in what follows we show that T_H^* and $\tilde{T}^*$ follow the same probability law. Define stochastic processes defined on subdomain of $\mathcal{F}$, $\mathcal{F}^* = \mathcal{F}_1^* \cup \mathcal{F}_0^*$,

$$\begin{aligned} u(f) &= -v_0(f)1\{f\in\mathcal{F}_1^*\} + v_0(f)1\{f\in\mathcal{F}_0^*\}, \\ u_H(f) &= -\frac{G_{H_0}(f)}{\xi\vee\sigma_{H_0}(f,f)}1\{f\in\mathcal{F}_1^*\} + \frac{G_{H_0}(f)}{\xi\vee\sigma_{H_0}(f,f)}1\{f\in\mathcal{F}_0^*\}. \end{aligned}$$

Note first that, for $f \in \mathcal{F}^*$, $P_0(f) = Q_0(f) = H_0(f)$ implies that

$$\sigma^2_{P_0,Q_0}(f,f) = P_0(f)(1-P_0(f)) = \sigma^2_{H_0}(f,f).$$

Hence, $Var(u(f)) = Var(u_H(f))$ holds for every $f \in \mathcal{F}^*$. To also show equivalence of the covariance kernels of $u(\cdot)$ and $u_H(\cdot)$, consider, for $f, g \in \mathcal{F}^*$,

$$\begin{aligned} Cov(u(f),u(g)) &= \frac{(1-\lambda)[P_0(fg) - P_0(f)P_0(g)] + \lambda[Q_0(fg) - Q_0(f)Q_0(g)]}{(\xi\vee\sigma_{P_0,Q_0}(f,f))(\xi\vee\sigma_{P_0,Q_0}(g,g))} \\ &= \frac{[(1-\lambda)P_0 + \lambda Q_0](fg) - H_0(f)H_0(g)}{(\xi\vee\sigma_{H_0}(f,f))(\xi\vee\sigma_{H_0}(g,g))}. \end{aligned}$$

If $f \in \mathcal{F}_1^*$ and $g \in \mathcal{F}_0^*$, $P_0(fg) = Q_0(fg) = H_0(fg) = 0$. If $f, g \in \mathcal{F}_1^*$, then $(P_0, Q_0) \in \mathcal{H}_0$ implies $0 \geq (P_0 - Q_0)(fg) \geq (P_0 - Q_0)(f) = 0$, so $P_0(fg) = Q_0(fg) = H_0(fg)$. Similarly, if $f, g \in \mathcal{F}_0^*$, $(P_0, Q_0) \in \mathcal{H}_0$ implies $0 \leq (P_0 - Q_0)(fg) \leq (P_0 - Q_0)(f) = 0$, so $P_0(fg) = Q_0(fg) = H_0(fg)$ holds as well. Thus, we obtain

$$\begin{aligned} Cov(u(f), u(g)) &= \frac{H_0(fg) - H_0(f)H_0(g)}{(\xi \vee \sigma_{H_0}(f,f))(\xi \vee \sigma_{H_0}(g,g))} \\ &= Cov(u_H(f), u_H(g)) \end{aligned}$$

for every $f, g \in \mathcal{F}^*$. Equivalence of the covariance kernels imply equivalence of the probability laws of the mean zero Gaussian processes, so we conclude $T_H^* \sim \tilde{T}^*$. Hence, $P\left(\tilde{T}^* > c_{1-\alpha}\right) \leq \Pr(T_H > c_{1-\alpha}) = \alpha$.

To check the first requirement of (B), we show continuity of the cdf of $\tilde{T}^*$ at $c_{1-\alpha}$ by applying the absolute continuity theorem for the supremum of Gaussian processes (Tsirelson (1975)), which says the supremum of Gaussian processes has a continuous cdf except at the left limit of its support. By the definition of $u_H(\cdot)$, T_H can be equivalently written as $T_H = \sup_{f \in \mathcal{F}} \{u_H(f)\}$. Note first that the support of T_H contains 0 since $\mathcal{F}$ contains an indicator function for a singleton set in $\mathcal{X}$ at which $u_{H_0}(f) = 0$ holds with probability one. Following the symmetry argument of the mean zero Gaussian process, which we borrowed from the proof of Proposition 2.2 in Abadie (2002), we have

$$\Pr(T_H \leq 0) = \Pr\left((\nexists f \in \mathcal{F}, u_H(f) > 0)\right) = \Pr\left((\nexists f \in \mathcal{F}, u_H(f) < 0)\right).$$

By Condition-RG (a), $u_H(\cdot)$ is not a degenerate process, so

$$\Pr\left((\nexists f \in \mathcal{F}, u_H(f) < 0) \cap (\nexists f \in \mathcal{F}, u_H(f) < 0)\right) = 0.$$

Hence,

$$\begin{aligned} 1 &\geq \Pr\left((\nexists f \in \mathcal{F}, u_H(f) < 0) \cup (\nexists f \in \mathcal{F}, u_H(f) < 0)\right) \\ &= 2\Pr(T_H \leq 0), \end{aligned}$$

implying that the probability mass that T_H can have at the left limit of its support is less than or equal to $1/2$. As a result, $c_{1-\alpha}$ for $\alpha \in (0, 1/2)$ lies in the region where the cdf of T_H is continuous. Since $\tilde{T}^*$ is also a supremum of mean zero Gaussian process and, as already shown, it is stochastically dominated by T_H, the cdf of $\tilde{T}^*$ is also continuous at $c_{1-\alpha}$. This completes the proof of Theorem 2.1 (i).

To prove claim (ii), assume that the first inequality of (1.1) is violated for some Borel set $B \subset \mathcal{Y}$. By lemma B.7, there exists some $f^* \in \mathcal{F}_1$ such that $0 \leq P(f^*) < Q(f^*)$ holds. Then, we have

$$\begin{aligned} T_N &= \max \left\{ \begin{array}{c} \sup_{f \in \mathcal{F}_1} \left\{ \frac{\hat{\lambda}^{1/2} Q_n(f) - (1-\hat{\lambda})^{1/2} P_m(f)}{\xi \vee \sigma_{P_m, Q_n}(f,f)} \right\} \\ \sup_{f \in \mathcal{F}_0} \left\{ \frac{(1-\hat{\lambda})^{1/2} P_m(f) - \hat{\lambda}^{1/2} Q_n(f)}{\xi \vee \sigma_{P_m, Q_n}(f,f)} \right\} \end{array} \right\} \\ &\geq \frac{\left(\hat{\lambda}^{1/2} G_{n,Q}(f^*) - (1-\hat{\lambda})^{1/2} G_{m,P}(f^*) \right)}{\xi \vee \sigma_{P_m, Q_n}(f^*, f^*)} + \sqrt{\frac{mn}{N}} \frac{Q(f^*) - P(f^*)}{\xi \vee \sigma_{P_m, Q_n}(f^*, f^*)}, \end{aligned} \tag{B.25}$$

where the second term of (B.25) diverges to positive infinity, while the first term is stochastically bounded asymptotically. Since the bootstrap critical values $c_{N,1-\alpha}$ converges to $c_{1-\alpha} < \infty$ irrespective of the null holds true or not, the rejection probability converges to one.

C Convergence of the Bootstrap Critical Values and a Proof of Theorem 2.2

C.1 A Lemma on Convergence of the Bootstrap Critical Values

The proof of Theorem 2.1 given in the previous section assumes $\mathbb{P}$-almost sure convergence of the bootstrap critical value $\tilde{c}_{N,1-\alpha}$ to $c_{1-\alpha}$. This convergence claim is proven by the next lemma. The probability space $(\Omega, \mathcal{B}(\Omega), \mathbb{P})$ and the random objects with "tilde" used in the following proof are the ones defined in the proof of Theorem 2.1 (i) by the almost sure representation theorem.

Lemma C.1 *Suppose Condition-RG. Let $\tilde{c}_{N,1-\alpha}$ be the bootstrap critical value of Algorithm 2.1 constructed from $\tilde{H}_N^{[N]} = \hat{\lambda} \tilde{P}_m^{[m]} + (1-\hat{\lambda}) \tilde{Q}_n^{[n]}$, which is viewed as a sequence of random variables $\{\tilde{c}_{N,1-\alpha} : N = 1, 2, \dots\}$ defined on probability space $(\Omega, \mathcal{B}(\Omega), \mathbb{P})$. It holds that $\tilde{c}_{N,1-\alpha}$ converges to $c_{1-\alpha}$ as $N \to \infty$, $\mathbb{P}$-a.s, where $c_{1-\alpha}$ is the $(1-\alpha)$-th quantile of statistic*

$$T_H = \max \left\{ \begin{array}{c} \sup_{f \in \mathcal{F}_1} \left\{ -G_{H_0}(f) / \left(\xi \vee \sigma_{H_0}(f,f) \right) \right\} \\ \sup_{f \in \mathcal{F}_0} \left\{ G_{H_0}(f) / \left(\xi \vee \sigma_{H_0}(f,f) \right) \right\} \end{array} \right\},$$

where $H_0 = \lambda P_0 + (1-\lambda) Q_0$.

Proof. Let sequence$\left\{ \tilde{H}_N^{[N]} : N = 1, 2, \dots \right\}$ be given, and let P_m^* and Q_n^* be the bootstrap empirical probability measures with size m and size n, respectively, drawn iid from $\tilde{H}_N^{[N]}$. Define bootstrap weighted empirical processes indexed by $f \in \mathcal{F}$ as

$$\begin{aligned} v_N^*(\cdot) &= \sqrt{\frac{mn}{N}} \frac{P_m^*(\cdot) - Q_n^*(\cdot)}{\xi \vee \sigma_{P_m^*, Q_n^*}(\cdot, \cdot)} \\ &= \frac{(1-\hat{\lambda})^{1/2} G^*_{m, \tilde{H}_N^{[N]}}(\cdot) - \hat{\lambda}^{1/2} G^{*\prime}_{n, \tilde{H}_N^{[N]}}(\cdot)(f)}{\xi \vee \sigma_{P_m^*, Q_n^*}(\cdot, \cdot)}, \end{aligned}$$

where $G^*_{m, \tilde{H}_N^{[N]}}(\cdot) = \sqrt{m}\left(P_m^* - \tilde{H}_N^{[N]}\right)(\cdot)$ and $G^{*\prime}_{n, \tilde{H}_N^{[N]}}(\cdot) = \sqrt{n}\left(Q_n^* - \tilde{H}_N^{[N]}\right)(\cdot)$ are two independent bootstrap empirical processes given $\left\{ \tilde{H}_N^{[N]} : N = 1, 2, \dots \right\}$. Let $(X_1, \dots, X_N)$ be the N support points of $\tilde{H}_N^{[N]}$, and let δ_X be the point-mass measure at X. To apply the uniform central limit theorem with exchangeable multipliers (Theorem 3.6.13 of van der Vaart and Wellner (1996)), we introduce multinomial random vector $(M_{m,1}, \dots, M_{m,N})$ that is independent of $(X_1, \dots, X_N)$ and has parameters $\left(m, \frac{1}{N}, \dots, \frac{1}{N}\right)$. We express $G^*_{m, \tilde{H}_N^{[N]}}(\cdot)$ as

$$\begin{aligned} G^*_{m, \tilde{H}_N^{[N]}}(\cdot) &= \frac{1}{\sqrt{m}} \sum_{i=1}^{N} \left(M_{m,i} - \frac{m}{N} \right) \delta_{X_i}(\cdot) \\ &= \frac{1}{\sqrt{m}} \sum_{i=1}^{N} \left(M_{m,i} - \frac{m}{N} \right) \left(\delta_{X_i} - H^{[N]} \right)(\cdot) \\ &= \hat{\lambda}^{-1/2} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \xi_{m,i} \left(\delta_{X_i} - H^{[N]} \right)(\cdot), \end{aligned}$$

where $\xi_{m,i} = M_{m,i} - \frac{m}{N}$, $i = 1, \dots, N$. Note that $\left(\xi_{m,1}, \dots, \xi_{m,N}\right)$ are exchangeable random variables by construction and $E\left(\frac{1}{N}\sum_{i=1}^{N} \xi_i^2\right) = \frac{m}{N}\left(1 - \frac{1}{N}\right) \to \lambda$, as $N \to \infty$. On the other hand, since $H^{[N]}$ converges weakly to H_0, an application of Lemma B.5 yields $\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(\delta_{X_i} - H^{[N]} \right)(\cdot) \rightsquigarrow G_{H_0}(\cdot)$. Hence, the uniform central limit theorem with exchangeable multipliers (Theorem 3.6.13 of van der Vaart and Wellner (1996)) leads to $G^*_{m, \tilde{H}_N^{[N]}}(\cdot) \rightsquigarrow G_{H_0}(\cdot)$ for $\mathbb{P}$-almost every sequence $\left\{ \tilde{H}_N^{[N]} : N = 1, 2, \dots \right\}$. By the same reasoning, we have $G^{*\prime}_{n, \tilde{H}_N^{[N]}}(\cdot) \rightsquigarrow G'_{H_0}(\cdot)$ for $\mathbb{P}$-almost every sequence $\left\{ \tilde{H}_N^{[N]} : N = 1, 2, \dots \right\}$, where $G'_{H_0}(\cdot)$ is an H_0-brownian bridge process independent of $G_{H_0}(\cdot)$.

Hence, the numerator of $v_N^*(\cdot)$ converges weakly to $(1-\lambda)^{1/2}G_{H_0}(\cdot) - \lambda^{1/2}G'_{H_0}(\cdot)$, $\mathbb{P}$-a.s. sequences of $\left\{\tilde{H}_N^{[N]}\right\}$. Note that the covariance kernel of $(1-\lambda)^{1/2}G_{H_0}(\cdot) - \lambda^{1/2}G'_{H_0}(\cdot)$ coincides with that of H_0-brownian bridge, so we conclude that

$$(1-\hat{\lambda})^{1/2}G^*_{m,\tilde{H}_N^{[N]}}(\cdot) - \hat{\lambda}^{1/2}G^{*\prime}_{n,\tilde{H}_N^{[N]}}(\cdot)(f) \rightsquigarrow G_{H_0}(\cdot), \ \mathbb{P}\text{-a.s. sequences of } \left\{\tilde{H}_N^{[N]}\right\}. \quad \text{(C.1)}$$

Regarding the bootstrap covariance kernel, we have convergence of $\sup_{f\in\mathcal{F}}\left|\sigma_{P_m^*,Q_n^*}(f,f) - \sigma_{H_0}(f,f)\right|$ to zero (in probability in terms of the probability law of bootstrap resampling given $\tilde{H}_N^{[N]}$) for $\mathbb{P}$-a.s. sequences of $\left\{\tilde{H}_N^{[N]}\right\}$, since

$$\sup_{f\in\mathcal{F}}\left|\sigma^2_{P_m^*,Q_n^*}(f,f) - \sigma^2_{H_0}(f,f)\right| \leq \sup_{f\in\mathcal{F}}\left|\sigma^2_{P_m^*,Q_n^*}(f,f) - \sigma^2_{\tilde{H}_N^{[N]}}(f,f)\right| + \sup_{f\in\mathcal{F}}\left|\sigma^2_{\tilde{H}_N^{[N]}}(f,f) - \sigma^2_{H_0}(f,f)\right|, \quad \text{(C.2)}$$

where the first term in the right hand side converges to zero (in probability in terms of the probability law of bootstrap resampling) by applying the Glivenko-Cantelli theorem for the triangular arrays as given in Lemma B.1, and the convergence to zero $\mathbb{P}$-a.s. for the second term follows from the almost sure representation theorem, (B.16) and (B.17).

By putting together (C.1) and (C.2), and repeating the proof of the asymptotic uniform equicontinuity as given in (B.11) above, we obtain

$$\begin{aligned} v_N^*(\cdot) &\rightsquigarrow \frac{(1-\lambda)^{1/2}G_{H_0}(\cdot) - \hat{\lambda}^{1/2}G'_{H_0}(\cdot)}{\xi \vee \sigma_{H_0}(\cdot,\cdot)} \\ &\sim \frac{G_{H_0}(\cdot)}{\xi \vee \sigma_{H_0}(\cdot,\cdot)}, \quad \text{as } N \to \infty, \end{aligned}$$

for $\mathbb{P}$-almost every sequence of $\left\{\tilde{H}_N^{[N]}\right\}$. The bootstrap test statistics T_N^* is a continuous functional of $v_N^*(\cdot)$, so the continuous mapping theorem leads to

$$T_N^* \rightsquigarrow T_H = \max\left\{\begin{array}{c} \sup_{f\in\mathcal{F}_1}\{-G_{H_0}(f)/(\xi \vee \sigma_{H_0}(f,f))\} \\ \sup_{f\in\mathcal{F}_0}\{G_{H_0}(f)/(\xi \vee \sigma_{H_0}(f,f))\} \end{array}\right\} \quad \text{as } N \to \infty,$$

for $\mathbb{P}$-almost every sequence of $\left\{\tilde{H}_N^{[N]}\right\}$. We already showed in the proof of Theorem 2.1 (i) that the cdf of T_H is continuous at $c_{1-\alpha}$ for $\alpha \in (0, 1/2)$. Hence, the bootstrap critical values $\tilde{c}_{N,1-\alpha}$ converges to $c_{1-\alpha}$, $\mathbb{P}$-a.s. ■

C.2 Proof of Theorem 2.2

Proof. By Assumption-LA(c) and the Portmanteau theorem, $\left(P^{[N]}, Q^{[N]} \in \mathcal{P}^2 : N = 1, 2 \dots\right)$ converges weakly to $(P_0, Q_0) \in \mathcal{H}_0$. We can therefore apply all the lemmas established in Appendix B and C.1, and, as done in the proof of Theorem 2.1 (i), we can define via the almost sure representation theorem a probability space $(\Omega, \mathcal{B}(\Omega), \mathbb{P})$ and random objects with "tilde", that copy the ones defined in a sequence of probability spaces in terms of $\left(P^{[N]}, Q^{[N]} : N = 1, 2, \dots\right)$. By Lemma C.1, the bootstrap critical values $\tilde{c}_{N,1-\alpha}$ converges to $c_{1-\alpha}$ the $(1-\alpha)$-th quantile of T_H, $\mathbb{P}$-a.s., which depends only on $(\alpha, \xi, \lambda, P_0, Q_0)$. Suppose that $([y, y'], d = 1)$ satisfies Assumption-LA (a) and (d). Let $\tilde{v}_N(\cdot) = \frac{\left(1-\hat{\lambda}\right)^{1/2} \tilde{G}_{m,P^{[N]}}(\cdot) - \hat{\lambda}^{1/2} \tilde{G}_{n,Q^{[N]}}(\cdot)}{\xi \vee \tilde{\sigma}_{P_m^{[N]}, Q_n^{[N]}}(\cdot,\cdot)}$ be the weighted empirical process defined on $(\Omega, \mathcal{B}(\Omega), \mathbb{P})$, where $\tilde{G}_{m,P^{[N]}}(\cdot) = \sqrt{m}\left(\tilde{P}_m^{[N]} - P^{[N]}\right)(\cdot)$ and $\tilde{G}_{m,Q^{[N]}}(\cdot) = \sqrt{m}\left(\tilde{Q}_n^{[N]} - Q^{[N]}\right)(\cdot)$. Note the probability law of the test statistic is that of

$$\tilde{T}_N = \max \left\{ \begin{array}{c} \sup_{f \in \mathcal{F}_1} \left\{ -\tilde{v}_N(f) - \tilde{h}_N(f) \right\} \\ \sup_{f \in \mathcal{F}_0} \left\{ \tilde{v}_N(f) + \tilde{h}_N(f) \right\} \end{array} \right\},$$

induced by $\mathbb{P}$, where

$$\tilde{h}_N(f) = \sqrt{\frac{mn}{N}} \frac{P^{[N]}(f) - Q^{[N]}(f)}{\xi \vee \tilde{\sigma}_{P_m^{[N]}, Q_n^{[N]}}(f, f)}.$$

Since $\tilde{T}_N$ is bounded from below by

$$-\tilde{v}_N([y, y'], 1) - \tilde{h}_N([y, y'], 1),$$

the rejection probability is also bounded from below by

$$\mathbb{P}(-\tilde{v}_N([y, y'], 1) - \tilde{h}_N([y, y'], 1) \geq \tilde{c}_{N,1-\alpha}).$$

By Assumption-LA (c), and by applying Lemmas B.4 and B.6, $\tilde{v}_N([y, y'], 1) - \tilde{h}_N([y, y'], 1)$ converges $\mathbb{P}$-a.s. to

$$-\tilde{v}_0([y, y'], 1) - \frac{[\lambda(1-\lambda)]^{1/2} \Delta\beta([y, y'], 1)}{\xi \vee \sigma_{P_0,Q_0}([y, y'], 1)},$$

which follows Gaussian with mean $-\frac{[\lambda(1-\lambda)]^{1/2}\Delta\beta([y,y'],1)}{\xi\vee\sigma_{P_0,Q_0}([y,y'],1)}$ and variance $\min\left\{\frac{\sigma^2_{P_0,Q_0}([y,y'],1)}{\xi^2},1\right\}$. Hence, we obtain

$$\begin{aligned}
& \mathbb{P}(-\tilde{v}_N([y,y'],1)-\tilde{h}_N([y,y'],1)\geq \tilde{c}_{N,1-\alpha}) \\
\to\ & \mathbb{P}(-\tilde{v}_0([y,y'],1)-\frac{[\lambda(1-\lambda)]^{1/2}\,\Delta\beta([y,y'],1)}{\xi\vee\sigma_{P_0,Q_0}([y,y'],1)}\geq c_{1-\alpha}) \\
=\ & 1-\Phi\left(\left(\frac{\sigma^2_{P_0,Q_0}([y,y'],1)}{\xi^2}\wedge 1\right)^{-1}\left(c_{1-\alpha}-\frac{[\lambda(1-\lambda)]^{1/2}\,|\Delta\beta([y,y'],1)|}{\xi\vee\sigma_{P_0,Q_0}([y,y'],1)}\right)\right).
\end{aligned}$$

In case $([y,y'],d=0)$ satisfies Assumption-LA (i) and (iv), a similar argument yields the same lower bound. ■

D Monte Carlo Studies

This section examines the finite sample performance of the test by Monte Carlo. In assessing finite sample type I errors of the test, we consider a data generating process on a boundary of $\mathcal{H}_0$, so that the theoretical type I error of the test equals to a nominal size asymptotically.

$$\begin{aligned}
p(y,D=1)&=q(y,D=1)=0.5\times\mathcal{N}(1,1),\\
p(y,D=0)&=q(y,D=0)=0.5\times\mathcal{N}(0,1),
\end{aligned}$$

where $N(\mu,\sigma^2)$ is the probability density of a normal random variable with mean μ and σ^2.

In computing the first (second) supremum of the test statistic, the boundaries points of intervals are chosen by every pair of Y-values observed in the subsample of $\{D=1,Z=0\}$ ($\{D=0,Z=1\}$). In order to assess how the test performance depends on a choice of trimming constant, we run simulations for each of the following four specification of the trimming constant,

$$\begin{aligned}
\xi_1 &= \sqrt{0.005\,(1-0.005)}\approx 0.07,\\
\xi_2 &= \sqrt{0.05(1-0.05)}\approx 0.22,\\
\xi_3 &= \sqrt{0.1(1-0.1)}=0.3,\\
\xi_4 &= 1.
\end{aligned}$$

Note that ξ_k, $k=1,2,3$, has the form of $\sqrt{\pi_k(1-\pi_k)}$, and π_k can be interpreted as that, if both $P_m([y,y'],d)$ and $Q_n([y,y'],d)$ are less than π_k, we weigh the difference of the empirical distribution by the inverse of ξ instead of the inverse of its standard deviation

Table 2: Monte Carlo Test Size

Monte Carlo iterations 1000, Bootstrap iterations 300.

Trimming constant	$\xi_1 \approx 0.07$			$\xi_2 \approx 0.21$			$\xi_3 = 0.3$			$\xi_4 = 1$		
Nominal size	.10	.05	.01	.10	.05	.01	.10	.05	.01	.10	.05	.01
(m,n):(100,100)	.13	.07	.01	.13	.07	.01	.14	.06	.01	.13	.06	.01
(100,500)	.11	.06	.01	.10	.06	.01	.11	.05	.01	.10	.05	.01
(500,500)	.13	.06	.02	.12	.07	.02	.11	.06	.02	.12	.05	.01
(100,1000)	.12	.06	.02	.12	.06	.01	.13	.06	.02	.12	.06	.02
(1000,1000)	.14	.07	.02	.13	.08	.02	.13	.06	.02	.12	.06	.01

Note: The statistic is equivalent to the non-weighted KS-statistics when $\xi_4 = 1$.

estimate. Accordingly, as π_k becomes larger, we put relatively less weight on the differences of the empirical probabilities for thinner probability events. The fourth choice of ξ, $\xi_4 = 1$, makes the test statistic identical to the non-weighted KS-statistic.

Table 2 shows the simulated test size. The rejection probabilities are slightly upward biased relative to the nominal sizes, while they are overall showing good size performance even in the cases with the sample sizes being as small as $(m, n) = (100, 100)$ and being unbalanced as much as $(m, n) = (100, 1000)$. It is also worth noting that these test sizes are not sensitive to a choice of trimming constant.

In order to see finite sample power performance of our test, we simulate the rejection probabilities of the bootstrap test against four different specifications of fixed alternatives. These four data generating processes share

$$\begin{aligned} \Pr(Z &= 1) = \frac{1}{2}, \quad \Pr(D = 1|Z = 1) = 0.55, \quad \Pr(D = 1|Z = 0) = 0.45 \\ p(y, D &= 1) = 0.55 \times \mathcal{N}(0, 1), \\ p(y, D &= 0) = 0.45 \times \mathcal{N}(0, 1), \qquad\qquad q(y, D = 0) = 0.55 \times \mathcal{N}(0, 1), \end{aligned}$$

while they differ in terms of specifications of the treated outcome distribution conditional

on $Z = 0$,

$$
\begin{aligned}
&\text{DGP 1:} && q(y, D=1) = 0.45 \times \mathcal{N}(-0.7, 1), \\
&\text{DGP 2:} && q(y, D=1) = 0.45 \times \mathcal{N}(0, 1.675^2), \\
&\text{DGP 3:} && q(y, D=1) = 0.45 \times \mathcal{N}(0, 0.515^2), \\
&\text{DGP 4:} && q(y, D=1) = 0.45 \times \sum_{l=1}^{5} w_l \mathcal{N}(\mu_l, 0.125^2), \\
& && (w_1, \ldots, w_5) = (0.15, 0.2, 0.3, 0.2, 0.15)\,, \\
& && (\mu_1, \ldots, \mu_5) = (-1, -0.5, 0, 0.5, 1)\,.
\end{aligned}
$$

In all these specifications, violations of the testable implication occur only for the treatment outcome densities. As plotted in Figure 6, the ways that the densities $p(y,1)$ and $q(y,1)$ intersect differ across the DGPs. In DGP 1, $p(y,1)$ and $q(y,1)$ is differentiated horizontally, and they intersect only once. In DGP 2, the violations occurs at the tail parts of $p(y,1)$ and $q(y,1)$, whereas, in DGP 3, the violation occurs around the modes of $p(y,1)$ and $q(y,1)$. In DGP 4, $q(y,1)$ is specified to be oscillating sharply around $p(y,1)$ and they intersect many times. In all these specifications, $p(y,1)$ and $q(y,1)$ are designed to be equally distant in terms of the one-sided total variation distance, i.e., $\int_{-\infty}^{\infty} \max\{(q(y,1) - p(y,1)), 0\}\, dy \approx 0.092$ for all the DGPs.

Table 3 shows the simulated rejection probabilities, based on which several remarks follow. First, we observe that the rejection probabilities vary depending on the DGPs and the choices of trimming constant. When the violations occur for the tail parts of the densities (DGP2), smaller ξ yields a significantly higher power. In contrast, if violations occur on a fatter part of the densities (DGPs 1, 3 and 4), middle-range ξ's and $\xi = 1$ tend to exhibit a slightly higher power than the smallest choice of ξ. This suggests that, if a likely violation of the testable implications is expected at the tail parts of the distributions, it is important to use a variance weighted statistic with a sufficiently small ξ such as $\xi = 0.07$. Given these simulation findings that a power loss by choosing $\xi = 0.07$ instead of the medium size ξ or $\xi = 1$ is not so severe in the other cases, we can argue that, in case there is no prior knowledge available about a likely alternative, one default choice of ξ is as small as 0.07. At the same time, it is also worth reporting the test results with several other choices of $\xi \in (0, 0.5]$. Second, the rows of unbalanced sample sizes indicate that the magnitude of

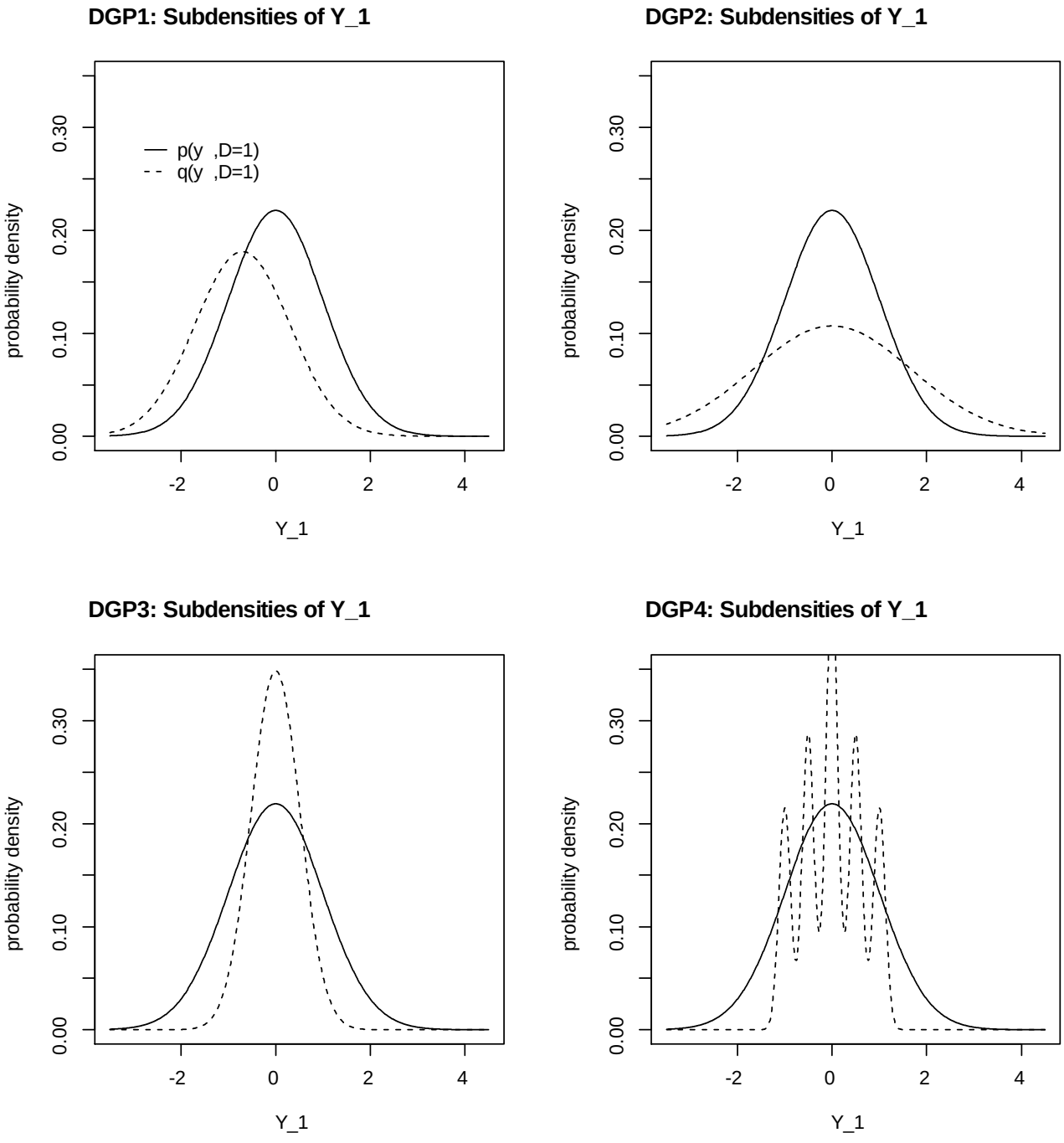

Figure 1: **Specification of Densities in Monte Carlo Experiments of Test Power**

the rejection probabilities tend to depend on a smaller sample size of (m, n), rather than the total sample size N, so a lack of power should be acknowledged when one of the sample size is small. Third, for the magnitudes of violations considered in these simulations, the rejection probabilities are sufficiently close to one (for some smaller choices ξ only for DGP 2) if the sample sizes are as large as $(m, n) = (1000, 1000)$.

References

[1] Abadie, A. (2002): "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models," Journal of the American Statistical Association, 97, 284-292.

[2] Abadie, A. (2003): "Semiparametric Instrumental Variable Estimation of Treatment Response Models," Journal of Econometrics, 113, 231 - 263.

[3] Andrews, D.W.K. and X. Shi (2013): "Inference Based on Conditional Moment Inequalities," Econometrica, 81, 609-666.

[4] Dudley, R. M. (1999): Uniform Central Limit Theorem. Cambridge University Press.

[5] Pollard, D. (1990): Empirical Processes: Theory and Applications, NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 2.

[6] Romano, J. P. (1988): "A Bootstrap Revival of Some Nonparametric Distance Tests." Journal of American Statistical Association, 83, 698-708.

[7] Tsirelson, V.S. (1975): "The density of the maximum of a Gaussian Process," Theory of Probability and Its Applications, 20, 817-856.

[8] van der Vaart, A. W., and J. A. Wellner (1996): Weak Convergence and Empirical Processes: With Applications to Statistics, New York: Springer.

Table 3: Rejection Probabilities against Fixed Alternatives

Monte Carlo iterations 1000, Bootstrap iterations 300.

	Trimming constant	$\xi_1 \approx 0.07$			$\xi_2 = 0.22$			$\xi_3 = 0.3$			$\xi_4 = 1$		
	Nominal size	.10	.05	.01	.10	.05	.01	.10	.05	.01	.10	.05	.01
DGP1	(m,n):(100,100)	.31	.22	.10	.31	.21	.10	.30	.21	.10	.23	.15	.05
	(100,500)	.42	.31	.14	.56	.43	.22	.57	.44	.21	.38	.24	.07
	(500,500)	.93	.88	.77	.95	.91	.78	.96	.92	.79	.89	.80	.52
	(100,1000)	.38	.28	.12	.58	.46	.24	.59	.46	.23	.39	.26	.09
	(1000,1000)	.99	.98	.94	1.00	.99	.97	.99	.98	.94	.99	.98	.93
DGP2	(m,n):(100,100)	.16	.09	.02	.15	.09	.02	.08	.04	.00	.01	.00	.00
	(100,500)	.35	.23	.07	.17	.10	.02	.07	.02	.00	.01	.00	.00
	(500,500)	.95	.91	.73	.86	.77	.53	.56	.40	.16	.10	.03	.01
	(100,1000)	.40	.26	.08	.20	.09	.02	.06	.03	.00	.01	.00	.00
	(1000,1000)	1.00	1.00	1.00	1.00	.99	.97	.96	.90	.67	.52	.27	.05
DGP3	(m,n):(100,100)	.30	.20	.09	.30	.20	.09	.33	.22	.09	.34	.22	.09
	(100,500)	.32	.21	.08	.51	.38	.19	.59	.46	.23	.55	.40	.15
	(500,500)	.77	.69	.54	.83	.76	.57	.87	.79	.62	.89	.82	.61
	(100,1000)	.30	.18	.06	.53	.40	.18	.61	.47	.25	.54	.41	.18
	(1000,1000)	.98	.96	.89	.99	.98	.93	1.00	.99	.96	1.00	.99	.95
DGP4	(m,n):(100,100)	.12	.07	.02	.11	.07	.02	.09	.05	.02	.09	.05	.01
	(100,500)	.23	.13	.04	.20	.11	.03	.15	.09	.02	.12	.05	.01
	(500,500)	.46	.33	.17	.45	.33	.16	.33	.22	.10	.23	.13	.03
	(100,1000)	.26	.15	.05	.22	.13	.05	.15	.10	.03	.11	.06	.01
	(1000,1000)	.78	.67	.48	.80	.69	.50	.51	.38	.19	.45	.30	.11