Kitagawa, Toru

Working Paper

A bootstrap test for instrument validity in heterogeneous treatment effect models

cemmap working paper, No. CWP53/13

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

Suggested Citation: Kitagawa, Toru (2013) : A bootstrap test for instrument validity in heterogeneous treatment effect models, cemmap working paper, No. CWP53/13, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2013.5313

This Version is available at:
https://hdl.handle.net/10419/97387

cemmap
centre for microdata methods and practice

A bootstrap test for instrument validity in heterogeneous treatment effect models

Toru Kitagawa

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP53/13

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

A Bootstrap Test for Instrument Validity in Heterogeneous Treatment Effect Models

Toru Kitagawa*
CeMMAP and *Department of Economics, UCL*

This draft: October, 2013

Abstract

This paper develops a specification test for the instrument validity conditions in the heterogeneous treatment effect model with a binary treatment and a discrete instrument. A necessary testable implication for the joint restriction of instrument exogeneity and instrument monotonicity is given by nonnegativity of point-identifiable complier's outcome densities. Our specification test infers this testable implication using a Kolmogorov-Smirnov type test statistic. We provide a bootstrap algorithm to implement the proposed test and show its asymptotic validity. The proposed test procedure can apply to both discrete and continuous outcome cases.

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

*Email: t.kitagawa@ucl.ac.uk. I am deeply grateful to Josh Angrist, Guido Imbens and Frank Kleibergen for helpful comments. I also thank Mario Fiorini, Stefan Hoderlein, Martin Huber, Giovanni Mellace, and Katrine Stevens for beneficial discussions. This version substantially revises the earlier draft circulated in 2008 under the same title. All remaining errors are mine. 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 is gratefully acknowledged.

1 Introduction

The instrumental variable method is a common tool to extract identifying information of causal effects when selection to treatment is present. As shown in Imbens and Angrist (1994), Angrist, Imbens and Rubin (1996), Imbens and Rubin (1997), and Heckman and Vytracil (1999, 2001), an instrument variable Z satisfying two key conditions enables us to identify the average treatment effects for those whose participation decision to treatment is affected by the instrument (local average treatment effect), which is referred to as LATE, hereafter. The two key conditions are, namely, (i) *random treatment assignment (RTA):* an instrument is assigned independently of any unobserved heterogeneities affecting one's potential outcomes and treatment selection response, and ii) *monotonic selection response to instrument (MSR):* every individual in the population has weakly monotonic treatment selection response to the instrument.

When we analyze experimental data with treatment incompliance, we often use the initial treatment assignment as an instrument. In this case, the instrumental validity assumptions would be credible if the initial treatment assignment is strictly randomized and possible incompliance is only one-sided, i.e., subjects who are allowed to switch their treatment status are only those who are initially assigned to the treatment group. See, for example, Abadie, Angrist, and Imbens (2002) and Kling, Liebman, and Katz (2007) for experimental data with one-sided incompliance. In contrast, if the incompliance is allowed for every subject irrespective of their initial treatment assignment status, validity of the instrument becomes less credible, since the sampling design cannot guarantee MSR. Examples of data subject to the two-sided incompliance include the well-known draft lottery data of Angrist (1991) and applications of the fuzzy regression discontinuity design (Campbell (1969), Hahn, Todd, and Van der Klaauw (2001)), where eligibility for a treatment based on one's attribute is used as an instrument. In case of multi-valued treatment status, Angrist and Imbens (1995) propose a specification test for MSR by inferring the stochastic dominance of the distribution functions of the treatment status conditional on the instrument; see also Barua and Lang (2009) and Fiorini, Stevens, Taylor, and Edwards (2013) for applications of the Angrist and Imbens test. In the binary treatment case, on the other hand, the Angrist and Imbens test cannot be applied.

In case where observational data are used for LATE analysis, strict randomization of Z is no longer guaranteed, so, not only violation of MSR, but also violation of RTA becomes a threat for instrument validity. Although credibility of LATE estimate critically hinges on validity of the employed instrument, there has been no test procedure proposed for empirically diagnosing instrument validity for the case with a binary treatment. As a result, causal inference studies using an instrument assumes its validity based solely on some background knowledge or indirect evidence

outside of data, whose credibility often remains controversial in many empirical contexts.

The main contribution of this paper is to develop a specification test for the instrument validity (the joint restriction of RTA and MSR) in the binary treatment case. Our specification test builds on the testable implication for instrument validity obtained by Balke and Pearl (1997) and by Heckman and Vytlacil (2005, Proposition A.5). These testable implications can be equivalently interpreted as the nonnegativity conditions for the density functions of complier's potential outcomes, which can be identified under RTA and MSR as shown in Imbens and Rubin (1997). Imbens and Rubin (1997) noted that the estimates of the complier's outcome densities can be negative on some region in the outcome support. Our test procedure focuses on this phenomenon as a clue to refute the instrumental validity. That is, if the complier's treated outcome or control outcome density is estimated to be negative over some regions in the outcome support, we interpret it as a counter-evidence for the joint restriction of RTA and MSR, since probability density function cannot be negative by definition. We demonstrate that the refuting rule based on the negativity of complier's outcome densities is most powerful for screening out invalid instruments, in the sense that any other feature of data distribution does not contribute to screening out more violations of MSR or RTA.

We propose a variance-weighted Kolmogorov-Smirnov type test statistic to measure how serious the nonnegativity of the compliers outcome density is violated in data. The asymptotic distribution of the proposed test statistic is analytically less tractable. We therefore develop a resampling algorithm to obtain asymptotically valid critical values, and demonstrate that the test procedure attains asymptotically correct size uniformly over a large class of data generating processes. As argued in Romano (1988), bootstrap is widely applicable and easy to implement to obtain the critical values for general Kolmogorov-Smirnov type test statistic, and it has been instrumental in the context of stochastic dominance testing; see, e.g., Abadie (2002), Barret and Donald (2003), Horváth, Kokoszka, and Zitikis (2006), and Linton, Maasoumi, and Whang (2005).

It is important to note that the joint restriction of MSR and RTA is a refutable but non-verifiable hypothesis. That is, rejecting the null hypothesis of nonnegativity of the complier's outcome densities enables us to reject validity of instrument, but accepting the null does not confirm that the instrument is valid. Such limitation on learnability of instrument validity condition is common in other contexts, such as the classical over-identification test in the generalized method of moments, and the test of MSR in the multi-valued treatment case proposed by Angrist and Imbens (1995). See Breusch (1986) for general discussion on hypothesis tests for refutable but non-verifiable assumptions.

This paper concerns the exogeneity of instrument defined in terms of statistical independence.

Given MSR, identification of LATE in fact can be attained under a slightly weaker set of assumptions, where 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 (2011) show that this weaker LATE identifying condition has a testable implication given by the finite number of moment inequalities. Our test builds on the distributional restrictions implied from statistical independence, and our test screens out a larger class of data generating processes than the test of Huber and Mellace (2011). Also, the set of alternatives to which our test is consistent is invariant to any monotonic transformation of the outcome variables, whereas this invariance property does not hold with the mean independence type restrictions considered in the Huber and Mellace's approach.

The rest of the paper is organized as follows. In Section 2, we analyze the testable implication of the instrumental validity in the heterogeneous treatment effect model with a binary treatment. Section 3 proposes a hypothesis test for the testable implication obtained in Section 2 and provide an algorithm of the bootstrap procedure for a binary instrument case. Section 4 extends the analysis to cases with a multi-valued instrument. Monte Carlo simulations and two empirical applications are provided in Section 5. Proofs are provided in the appendices.

2 Model

We consider a model with a binary treatment, where observed treatment status is denoted by D; $D = 1$ when one receives the treatment while $D = 0$ if she does not. Let Y_1 be potential outcomes with treatment, and Y_0 be a potential outcome without the treatment, whose support is denoted by $\mathcal{Y} \subset \mathbb{R}$. The observed outcome Y satisfies $Y = Y_1 D + Y_0(1-D)$. We denote an instrumental variable by Z, which is assumed to be binary. We discuss an extension to a multi-valued Z in Section 4. Following Angrist and Imbens (1994), we introduce D_1 as the potential selection response that one would take given $Z = 1$. Similarly, we define D_0 for $Z = 0$. Associated with the potential selection indicators, we define the individual type T that indicates individual selection response to the instrument Z.

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

The following two assumptions guarantee point-identification of the local average treatment effects for compliers, and, simultaneously, the marginal distributions of the counterfactual outcomes for compliers (see Imbens and Angrist (1994) and Imbens and Rubin (1997)) and the quantile treatment effects for compliers (Abadie, Angrist, and Imbens (2002)).

Assumption IV

1. *Random Treatment Assignment (RTA):* Z is jointly independent of (Y_1, Y_0, D_1, D_0).
2. *Monotonic Selection Response to Instrument (MSR): Without loss of generality, assume* $\Pr(D = 1|Z = 1) \geq \Pr(D = 1|Z = 0)$. The potential participation indicators satisfy $D_1 \geq D_0$ *with probability one.*

Note that the above assumptions are defined in terms of the potential variables. RTA incorporates the instrument exogeneity and the instrument exclusion restrictions in the form of joint statistical independence, and it can be interpreted that the instrument is assigned randomly with respect to any of the individual unobserved heterogeneities affecting treatment selection responses and/or potential outcomes. MSR means that there are no defiers in the population $\Pr(T = d) = 0$. Since we never observe all the potential variables of the same individual, we cannot directly examine these assumptions from data, and necessary and sufficient testable implications for these assumptions are not available.

In order to present a necessary condition for the instrumental validity, we introduce the following notations. Let P and Q be the conditional probability distributions of $(Y, D) \in \mathcal{Y} \times \{1, 0\}$ given $Z = 1$ and $Z = 0$ respectively, which can be consistently estimated from data We view the data generating processes to have the two-sample structure in terms of the assigned value of Z. For Borel set $B \subset \mathcal{Y}$ and $d = 1, 0$, define

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

We now present the refutability result of the instrumental validity. A proof is provided in Appendix A.

Proposition 2.1 *If a population distribution of* (Y_1, Y_0, D_1, D_0, Z) *satisfies RTA and MSR, then, the distribution of observables* P *and* Q *satisfies the following inequalities for every Borel set* B *in* $\mathcal{Y}$*,*

$$\begin{aligned} &P(B, 1) \geq Q(B, 1), \\ &P(B, 0) \leq Q(B, 0). \end{aligned} \tag{2.1}$$

Conversely, if the data generating process P and Q satisfies these inequalities for all Borel set B in $\mathcal{Y}$, and $P(\cdot, d)$ and $Q(\cdot, d)$ are absolutely continuous with respect to a common dominating measure on $\mathcal{Y}$ for each $d = 0, 1$, then there exists a joint probability law of (Y_1, Y_0, D_1, D_0, Z) that is compatible with the data generating process P and Q, RTA, and MSR.

Balke and Pearl (1997) obtain the testable implication (2.1) for the case of binary Y and binary Z, and Heckman and Vytlacil (2005) obtain a more general form of the testable implication, in which Y and Z can be continuous. The converse statement of the proposition clarifies that the refuting rule based on Proposition 1 is most powerful in screening out violations of the instrument validity, and no other features of data distribution can further contributes to detecting invalid instrument. Note that Proposition 2.1 does not give an if and only if statement for instrument validity, so knowing that the data distribution satisfies (2.1) cannot guarantee that the instrument is valid. In this sense, the instrument validity condition of RTA and MSR is refutable but non-verifiable.

Let $p(y, d)$ and $q(y, d)$ be the probability density function of P and Q on $\mathcal{Y} \times \{d\}$, defined by

$$\begin{aligned} P(A, d) &= \int_A p\,(y, d)\, d\mu, \\ Q(A, d) &= \int_A q(y, d) d\mu, \end{aligned}$$

where μ is a dominating measure on $\mathcal{Y}$. In terms of these density functions, the inequalities of (2.1) can be equivalently written as

$$\begin{aligned} &p(y, 1) \geq q(y, 1), \quad \mu\text{-a.e.}, \\ &p(y, 0) \leq q(y, 0), \quad \mu\text{-a.e.} \end{aligned}$$

Figures 1 and 2 provide visual illustration of Proposition 1. There, the left-hand side figures correspond to Y_1's distribution and the right-hand side figures corresponds to Y_0's distribution. The solid lines represent $p(y, d)$ and $q(y, d)$, which are identifiable by data. The dotted line in each figure represents the marginal probability density of the potential outcomes, i.e., $f_{Y_1}(y)$ is the marginal density of Y_1 and $f_{Y_0}(y)$ is the marginal density of Y_0, which are not identified by data. Note that integrations of $p(y, d)$ and $q(y, d)$ are equal to the probability of $D = d$ conditional on Z, so the areas of $p(y, d)$ and $q(y, d)$ are smaller than those of $f_{Y_d}(\cdot)$ and $f_{Y_0}(\cdot)$, respectively. Furthermore, under RTA, both $p(y, d)$ and $q(y, d)$ must lie below the potential outcome densities $f_{Y_d}(\cdot)$. If RTA and MSR hold in the population, Proposition 2.1 implies that the two identifiable density functions $p(y, d)$ and $q(y, d)$ must be nested as shown in Figure 1. For the treated outcome densities, $p(y, 1)$ must nest $q(y, 1)$ and for the control outcome densities, $q(y, 0)$ must nest $p(y, 0)$. These nesting

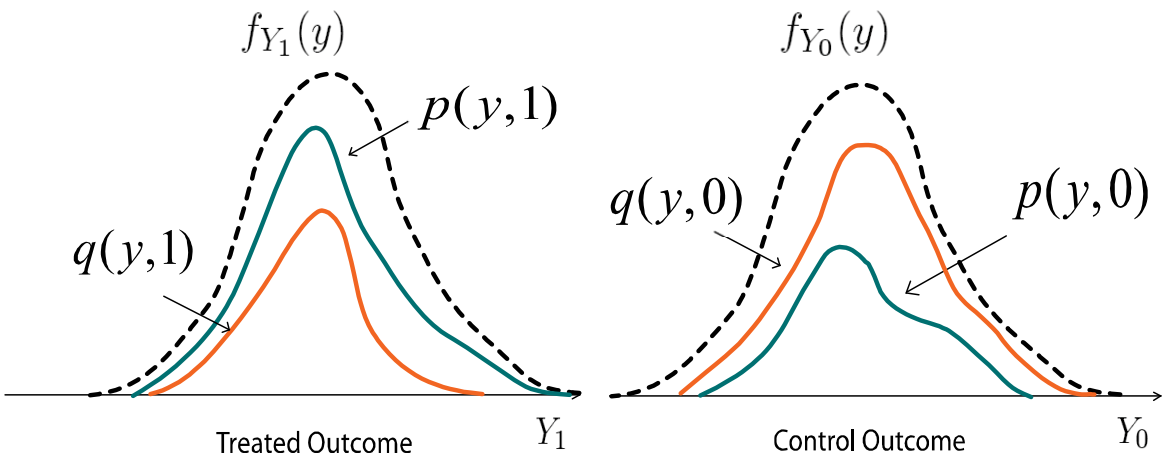

Figure 1: *When we observe that the observable densities* $p(y, D_{obs} = 1)$ and $q(y, D_{obs} = d)$ *are nested as in this figure, the instrumental validity is not refuted.*

structures of the subdensities are equivalent to nonnegativity of the complier's potential outcome densities since, under RTA and MSR, Imbens and Rubin (1997) show

$$\begin{aligned} p(y,1) - q(y,1) &= f_{Y_1|T}(y|T=c) \times \Pr(T=c) \quad \text{and} \\ q(y,0) - p(y,0) &= f_{Y_0|T}(y|T=c) \times \Pr(T=c). \end{aligned}$$

If we observe the densities like Figure 2, we can refute at least one of the instrumental validity conditions since some of the inequalities (2.1) are violated on some subsets of the outcome support. These subsets are labeled as V_1 and V_2 in Figure 2.

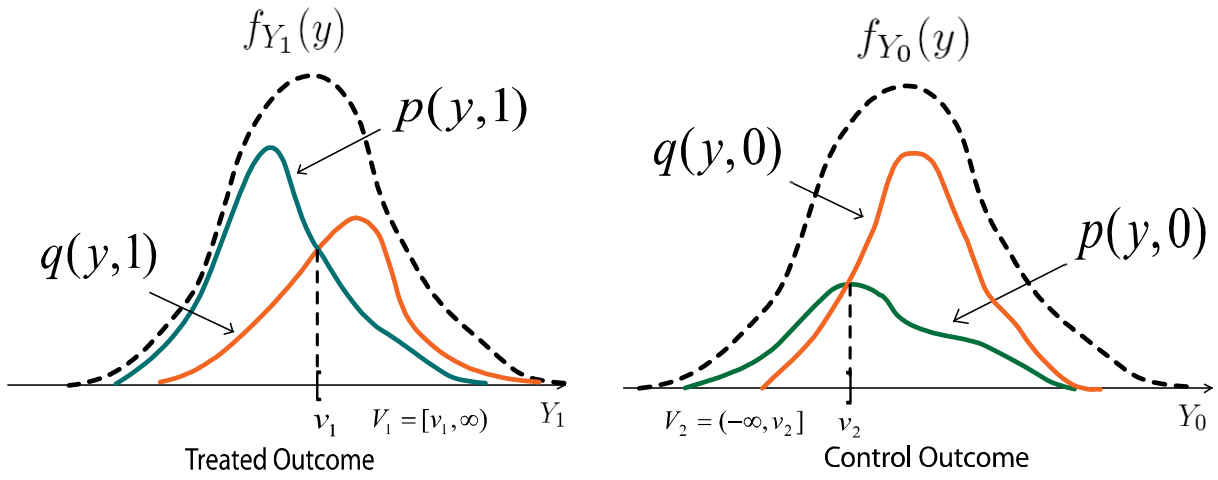

Figure 2: *When we observe the above configuration of the densities, we can refute the instrumental validity since at the subset $V_1=[v_1,\infty)$, the first inequality in Proposition 1 is violated. The right-hand side picture shows that the second inequality in Proposition 1 is violated at $V_2=(-\infty,v_2]$.*

3 Test Procedure

P and Q are point-identified by the sampling process, and therefore we can examine inequalities (2.1) by inferring whether estimators for P and Q satisfy them or not. Assume a sample consist of N i.i.d observations of (Y,D,Z). We divide the sample into two subsamples based on the value of Z. Let m be the sample size with $Z_i=1$ and n be the sample size with $Z_i=0$. We make our asymptotic analysis conditional on a sequence of instrument values, $\{Z_1,Z_2\dots\}$ with $\hat{\lambda}=m/N\to\lambda$ as $N\to\infty$, where λ is bounded away from zero and one. Let (Y_i^1,D_i^1), $i=1,\dots,m$ be the observations with $Z=1$ and (Y_j^0,D_j^0), $j=1,\dots,n$ be those with $Z=0$. Consider estimating P and Q by the empirical distributions,

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

To test the null hypothesis given by inequalities (2.1), we consider a variance-weighted Kolmogorov-Smirnov type statistic,

$$T_N=\left(\frac{mn}{N}\right)^{1/2}\max\left\{\begin{array}{c}\sup_{V\in\mathbb{V}_1}\left\{\sigma^{-1}_{P_m,Q_n}(V,1)\left(Q_n(V,1)-P_m(V,1)\right)\right\},\\ \sup_{V\in\mathbb{V}_0}\left\{\sigma^{-1}_{P_m,Q_n}(V,1)\left(P_m(V,0)-Q_n(V,0)\right)\right\}\end{array}\right\}, \tag{3.1}$$

where $\mathbb{V}_1$ and $\mathbb{V}_0$ are collection of subsets in $\mathcal{Y}$, and $\sigma^2_{P_m,Q_n}(V,d)$ is a consistent estimator for the asymptotic variance of $\left(\frac{mn}{N}\right)^{1/2}(P_m(V,d)-Q_n(V,d))$,

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

If the sample counterpart of the first inequality of (2.1) is violated for some subset V, then, the first supremum in the max operator is positive. Similarly, when the sample counterpart of the second inequality of (2.1) is violated for some subset V, then the second term in the max operator becomes positive. The weighting term $\sigma^{-1}_{P_m,Q_n}(V,d)$ adjusts the sample variations of the difference of the empirical probabilities across different (V,d). Thus, the proposed test statistics quantifies a variance-adjusted maximal violation of the inequalities (2.1) over prespecified class of subsets.

As far as asymptotically valid test size is concerned, we can also consider using a non-weighted Kolmogorov-Smirnov statistic,

$$T_{N,nw} = \left(\frac{mn}{N}\right)^{1/2} \max\left\{ \begin{array}{c} \sup_{V\in\mathbb{V}_1}\{Q_n(V,1)-P_m(V,1)\}, \\ \sup_{V\in\mathbb{V}_0}\{P_m(V,0)-Q_n(V,0)\} \end{array} \right\}. \tag{3.2}$$

As illustrated in the Monte Calro studies of Section 5, the null distribution of the non-weighted statistics $T_{N,nw}$ can be better approximated by bootstrap in small sample situations. In cases where the sample size is moderately large, on the other hand, T_N appears to have a better finite sample power than $T_{N,nw}$ for a wide class of alternatives, since T_N can better capture violations of the inequalities (2.1) in the tail parts of the P and Q than $T_{N,nw}$ does. Our informal recommendation is, therefore, to use T_N when the two samples have moderate sample sizes, e.g., $m \geq 500$ and $n \geq 500$, and to use $T_{N,nw}$ instead if either or both of the samples have small sample sizes.

Although the testable implication of Proposition 2.1 states that the inequalities hold for every subset in $\mathcal{Y}$, we cannot take $\mathbb{V}_1$ and $\mathbb{V}_0$ as rich as the Borel σ-algebra unless Y is discrete. In order for the above test statistic to have a nontrivial asymptotic distribution, a specified $\mathbb{V}_1$ and $\mathbb{V}_0$ has to guarantee the uniform convergence property of the empirical processes of P_m and Q_n. As a class of subsets which meets this requirement, we consider a Vapnik-Červonenkis class (VC-class) of subsets satisfying Assumption 1(a) below. See e.g., Dudley (1999) and van der Vaart and Wellner (1996) for the definition and examples of a VC-class of subsets.

We will employ two specific constructions of a VC-class in our Monte Carlo studies and empirical applications in the following sections. The first specification is a *half-unbounded-interval class* $\mathbb{V}_{half}$, which is simply a collection of half-unbounded intervals,

$$\mathbb{V}_{half} = \{(-\infty, y] : y \in \mathcal{Y}^*\} \cup \{[y,\infty) : y \in \mathcal{Y}^*\}, \quad \mathcal{Y}^* \subset \mathcal{Y}, \tag{3.3}$$

The second specification is a union of $\mathbb{V}_{half}$ and a class of connected intervals, which we referred to as an *interval class*

$$\mathbb{V}_{int} = \mathbb{V}_{half} \cup \left[\bigcup_{h \in [\underline{h}, \bar{h}]} \mathbb{V}_{bin}(h) \right],$$

where $\mathbb{V}_{bin}(h)$ is a collection of connected intervals with width $h > 0$,

$$\mathbb{V}_{bin}(h) = \{[y, y+h] : y \in \mathcal{Y}^*\}.$$

An advantage of considering this richer class over $\mathbb{V}_{half}$ is that the test statistics can asymptotically screen out a larger class of data generating processes, while a potential drawback is that, given fixed sample size, the quality of the asymptotic approximation may deteriorate as the VC-class becomes richer. In Section 5, we provide a further discussion and a practical recommendation on a convenient choice of the VC-class based on our Monte Carlo findings.

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. In order to determine H in a data-driven way, we focus on the following representation of H,

$$H = (1-\lambda)P + \lambda Q, \tag{3.4}$$

and aim to estimate the quantiles of the null distribution of the test statistic T_N or $T_{N,nw}$ as if the data are generated from $P = Q = H$.

As discussed in Romano (1988), the resampling method is an attractive approach to estimate asymptotically valid critical values for the Kolmogorov-Smirnov type test statistic since its asymptotic distribution generally does not have an analytically tractable distribution function. Given that our focus is on approximating the sampling distribution of the test statistic under $P = Q = H$, we draw bootstrap samples from the empirical analogue of (3.4), $H_N = (1-\hat{\lambda})P_m + \hat{\lambda}Q_n$. Note that this specification of H_N is different from the pooled empirical measure, $\hat{\lambda}P_m + (1-\hat{\lambda})Q_n$, which the standard resampling-based Kolmogorov-Smirnov uses to generate the bootstrap samples. The reason that we focus on H_N rather than the pooled empirical measure is that, for our non-standard form of the one-sided null hypothesis, we can guarantee that the test with the bootstrap samples being drawn from $H_N = (1-\hat{\lambda})P_m + \hat{\lambda}Q_n$ achieve asymptotically uniformly correct test size with a general construction of VC-classes, whereas we do not have a proof for the uniform validity of the test if the bootstrap samples are drawn from the pooled empirical measure.

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

Algorithm 3.1:

1. S*ample* (Y_i^*, D_i^*), $i = 1, \ldots, m$ *randomly with replacement from* $H_N = (1 - \hat{\lambda})P_m + \hat{\lambda}Q_n$ *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^*.[1]

2. Let $H_N^* = (1 - \hat{\lambda})P_m^* + \hat{\lambda}Q_n^*$ and compute $\sigma^2_{H_N^*}(V, d) = \max\{H_N^*(V, d)(1 - H_N^*(V, d)), \xi\}$ for every $V \in \mathbb{V}_1$ or $\mathbb{V}_0$, and $d \in \{1, 0\}$, where $\xi > 0$ is some positive constant smaller than κ defined in Condition RG (iv) below.

3. *Calculate a bootstrap realization of test statistic*

$$T_N^* = \left(\frac{mn}{N}\right)^{1/2} \max \left\{ \begin{array}{c} \sup_{V \in \mathbb{V}_1} \left\{ \sigma^{-1}_{H_N^*}(V, 1) \left(Q_n^*(V, 1) - P_m^*(V, 1)\right) \right\}, \\ \sup_{V \in \mathbb{V}_0} \left\{ \sigma^{-1}_{H_N^*}(V, 0) \left(P_m^*(V, 0) - Q_n^*(V, 0)\right) \right\} \end{array} \right\}.$$

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

5. *Reject the null hypothesis if* $T_N > c_{N,1-\alpha}$.

When the non-weighted test statistics $T_{N,nw}$ is used, Step 2 of Algorithm 3.1 is not needed, while the rest of the steps is unchanged except that the bootstrap statistic is computed by

$$T_{N,nw}^* = \left(\frac{mn}{N}\right)^{1/2} \max \left\{ \begin{array}{c} \sup_{V \in \mathbb{V}_1} \left\{ Q_n^*(V, 1) - P_m^*(V, 1) \right\}, \\ \sup_{V \in \mathbb{V}_0} \left\{ P_m^*(V, 0) - Q_n^*(V, 0) \right\} \end{array} \right\}.$$

To formally claim that the test procedure of Algorithm 3.1 is asymptotically valid uniformly over a certain class of data generating processes, we introduce the following notations. Let $\mathcal{P}$ be a

[1] In terms of the point mass measure, H_N is written as

$$H_N = \sum_{i=1}^{m} \frac{n}{Nm} \delta_{Y_i^1, D_i^1} + \sum_{j=1}^{n} \frac{m}{Nn} \delta_{Y_j^0, D_j^0}.$$

Hence, resampling from H_N is done by sampling with replacement the observations in the original sample with the corresponding probability weight, i.e, probability $\frac{n}{Nm}$ for the observations with $Z_i = 1$ and probability $\frac{m}{Nn}$ for the observations with $Z_i = 0$.

set of probability measures defined on the Borel σ-algebra of $\mathcal{Y} \times \{0,1\}$, to which P and Q belong. Denote by $\mathcal{H}_0$ the set of data generating processes satisfying the null,

$$\mathcal{H}_0 = \left\{(P,Q) \in \mathcal{P}^2 : \text{ inequalities (2.1) hold.}\right\}.$$

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

Condition RG:

(a) $\mathbb{V}_1$ and $\mathbb{V}_0$ are VC-classes of subsets in $\mathcal{Y}$.

(b) Probability measures in $\mathcal{P}$ have a common dominating measure μ for $\mathcal{Y}$-coordinate, and the density functions $p(y,d) \equiv \frac{dP(\cdot,d)}{d\mu}$ are bounded, 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 $P \in \mathcal{P}$.

(c) $\mathcal{P}$ is uniformly tight, i.e., for arbitrary $\epsilon > 0$, there exists a compact set $K \subset \mathcal{Y} \times \{0,1\}$ such that

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

(d) Given the VC-classes $\mathbb{V}_1$ and $\mathbb{V}_0$ and $\lambda = \lim_{N\to\infty} \frac{m}{N}$,

$$\sigma^2_{P,Q}(V,d) \equiv (1-\lambda)P(V,d)(1-P(V,d)) + \lambda Q(V,d)(1-Q(V,d))$$

is bounded away from zero uniformly over (V,d) and $(P,Q) \in \mathcal{H}_0$, i.e., there exists $\kappa > 0$ such that

$$\inf_{(P,Q)\in\mathcal{H}_0, V\in\mathbb{V}_1} \sigma^2_{P,Q}(V,d) \geq \kappa \text{ and } \inf_{(P,Q)\in\mathcal{H}_0, V\in\mathbb{V}_0} \sigma^2_{P,Q}(V,0) \geq \kappa.$$

For discrete Y case, what is relevant among these conditions is only Condition RG (d), requiring that every point in the support of Y occurs with positive probability in terms of either P or Q. For continuous Y case, Condition RG (b) imposes mild conditions on the density functions of P and Q. Condition RG (d) requires that, at every $V \in \mathbb{V}_d$ and every null data generating process, $P(V,d)$ or $Q(V,d)$ is bounded away from zero. In practical term, this condition imposes that we should specify $\mathbb{V}_1$ and $\mathbb{V}_0$ in such way that the the probabilities on each subset is well supported by data generating processes.

The asymptotic validity of the proposed test is stated in the next proposition.

Proposition 3.1 *Suppose Condition RG. Let $\alpha \in (0,1)$.*

(i) The test procedure of Algorithm 3.1 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) The test procedure of Algorithm 3.1 modified for the non-weighted test statistic $T_{N,nw}$ has asymptotically uniformly correct size for null hypothesis $\mathcal{H}_0$. This claim does not rely on Condition RG (d).

(iii) Suppose Condition RG. If, for a fixed alternative, there exist some $V \in \mathbb{V}_1$ or $V \in \mathbb{V}_0$ at which the corresponding inequalities of (2.1) are violated, the tests based on T_N and $T_{N,nw}$ are consistent, i.e., the rejection probabilities converge to one as $N \to \infty$.

Proof. A proof for claim (i) and (iii) are given in Appendix B. A proof for claim (ii) is omitted, since the proof of claim (i) covers it. ■

This proposition establishes asymptotic uniform validity of the proposed test procedure over $\mathcal{P}$ characterized by Condition RG (b)-(d). The third claim of the proposition is on the asymptotic power of the test, and it emphasizes that the set of alternatives that can be consistently rejected hinges on how $\mathbb{V}_1$ and $\mathbb{V}_0$ are specified. Accordingly, in practical terms, a choice of $\mathbb{V}_1$ and $\mathbb{V}_0$ can reflect the degree of importance on what alternatives should be detected or user's opinion on what type of alternatives is more likely to be true. For instance, consider the case where RTA is guaranteed by a sampling design, while MSR is not so that it is the hypothesis of concern. Suppose that there exists a behavioral model that says, if the defiers are present, the ratio of the conditional type probabilities given Y_1,

$$\frac{\Pr(T = d|Y_1 = y)}{\Pr(T = c|Y_1 = y)},$$

is weakly monotonically decreasing in y. If the set of alternatives are restricted to those, then specifying $\mathbb{V}_1$ to $\{(-\infty, y] : y \in \mathcal{Y}\}$ suffices to screen out every alternative violating the first inequality of (2.1) at some subset.[2]

This example illustrates that a *qualitative* assumption imposed on the distribution of potential outcomes and selection types can justify a parsimonious specification of $\mathbb{V}_1$ and/or $\mathbb{V}_0$. With help of such assumption, our test would become particularly attractive since the procedure does not require any smoothing parameters even though the model and the hypothesis to be tested is fully nonparametric.

[2] Consider a model where RTA holds, but defiers can exist. Suppose that the ratio conditional type probabilities given (Y_1, Y_0) satisfies

$$\frac{\Pr(T = d|Y_1 = y, Y_0 = y')}{\Pr(T = c|Y_1 = y, Y_0 = y')} \leq 1.$$

Then, we can show no alternatives violate the inequalities of (2.1), implying our test procedure has no power detecting this type of violation of MSR. However, Chaisemartin (2013) shows that, even under this type of alternatives, the Wald estimator still estimates an average causal effects for a well-defined subpopulation of "comvivors".

In case no such assumption is available, our recommendation based on the following Monte Carlo studies is to try $\mathbb{V}_{int}$ with various choice of smallest binwidth $\underline{h}$.

4 Extension and Discussion

4.1 Multi-valued instrument

The test procedure proposed above can be extended 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, the following assumptions guarantees 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*

1. *RTA*:* Z is jointly independent of $(Y_1, Y_0, D_{z_1}, \ldots, D_{z_K})$.
2. *MSR*:* *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 $P(B, 1|z_k) = \Pr(Y \in B, D = 1|Z = z_k)$, $k = 1, \ldots, K$. The testable implication of the binary instrument case is now generalized to the following set of inequalities; under RTA* and MSR* of Assumption 2,

$$\begin{aligned} &P(B,1|z_1) \leq P(B,1|z_2) \leq \cdots \leq P(B,1|z_K) \quad \text{and} \\ &P(B,0|z_1) \geq P(B,0|z_2) \geq \cdots \geq P(B,0|z_K) \end{aligned} \tag{4.1}$$

holds with every measurable subset B in $\mathcal{Y}$. Using the test statistic of the previous section to measure the violation of the functional inequalities across the neighboring value of Z, we can develop a statistic that jointly tests the inequalities of (4.1),

$$T_N = \max\left\{T_{N,1}, \ldots, T_{N,K-1}\right\}, \tag{4.2}$$

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

$$\begin{aligned}
T_{N,k} &= \left(\frac{m_k m_{k+1}}{m_k + m_{k+1}}\right)^{1/2} \max\left\{ \begin{array}{c} \sup_{V\in\mathbb{V}_{1,k}} \left\{\sigma_k^{-1}(V,1)\left(P_{m_k}(V,1|z_k) - P_{m_{k+1}}(V,1|z_{k+1})\right)\right\}, \\ \sup_{V\in\mathbb{V}_{0,k}} \left\{\sigma_k^{-1}(V,0)\left(P_{m_{k+1}}(V,0|z_{k+1}) - P_{m_k}(V,0|z_k)\right)\right\} \end{array} \right\}, \\
\sigma_k^2(V,d) &= \left(\frac{m_k}{m_k + m_{k+1}}\right) P_{m_{k+1}}(V,d|z_{k+1})\left(1 - P_{m_{k+1}}(V,d|z_{k+1})\right) \\
&\quad + \left(\frac{m_{k+1}}{m_k + m_{k+1}}\right) P_{m_k}(V,d|z_k)\left(1 - P_{m_k}(V,d|z_k)\right),
\end{aligned}$$

$m_k = \#\{i : Z_i = z_k\}$, P_{m_k} is the empirical probability measure on $\mathcal{Y}\times\{1,0\}$ of the subsample with $Z = z_k$, and $\mathbb{V}_{1,k}$ and $\mathbb{V}_{0,k}$ are VC-class of subsets. Critical values can be obtained by applying a resampling algorithm of the previous section to each $T_{N,k}$ simultaneously.

Algorithm 4.1:

1. Let $H_{N,k}(\cdot) = \left(\frac{m_k}{m_k+m_{k+1}}\right) P_{m_{k+1}}(\cdot|z_{k+1}) + \left(\frac{m_{k+1}}{m_k+m_{k+1}}\right) P_{m_k}(\cdot|z_k)$ be a weighted average of the two empirical measures of 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 $k = 1, \ldots, (K-1)$,

$$\begin{aligned}
H^*_{N,k} &= \left(\frac{m_k}{m_k + m_{k+1}}\right) P^*_{m_{k+1}} + \left(\frac{m_{k+1}}{m_k + m_{k+1}}\right) P^*_{m_k}, \\
\sigma_k^{*2}(V,d) &= H^*_{N,k}(V,d)(1 - H^*_{N,k}(V,d)), \\
T^*_{N,k} &= \left(\frac{m_k m_{k+1}}{m_k + m_{k+1}}\right)^{1/2} \\
&\quad \times \max\left\{ \begin{array}{c} \sup_{V\in\mathbb{V}_{1,k}} \left\{\sigma_k^{*-1}(V,1)\left(P^*_{m_k}(V,1|z_k) - P^*_{m_{k+1}}(V,1|z_{k+1})\right)\right\}, \\ \sup_{V\in\mathbb{V}_{0,k}} \left\{\sigma_k^{*-1}(V,0)\left(P^*_{m_{k+1}}(V,0|z_{k+1}) - P^*_{m_k}(V,0|z_k)\right)\right\} \end{array} \right\}
\end{aligned}$$

   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 and get the empirical distribution of* T^*_N, and *obtain the bootstrapped critical value* $c_{N,1-\alpha}$ *from its empirical* $(1-\alpha)$*-th quantile .*

4. *Reject the null hypothesis if* $T_N > c_{N,1-\alpha}$.

4.2 Discussion: Improving Finite Sample Power and Alternative Approaches

The bootstrap procedure considered in this paper draws a critical value from a boundary null hypothesis $P = Q$. As is known in the moment inequality literature, obtaining critical values from a least favorable null may sacrifice a finite sample power for some alternatives. The finite sample power can be improved if critical values are obtained from the null distribution of the supremum statistic over estimated contact sets, $\{V \in \mathbb{V}_1 : P(V,1) = Q(V,1)\}$ and $\{V \in \mathbb{V}_0 : P(V,0) = Q(V,0)\}$; see Linton, Song, and Whang (2010)). By a similar idea, inference procedures for moment inequalities involving generalized moment selection (Andrews and Soares (2010), Andrews and Shi (2013)) can apply to the current context and can improve the finite sample power performance of our test. In these approaches, estimation of the contact sets (the set of binding inequalities) requires the user to specify a value of slackness parameters. To our knowledge, a recommended choice for such tuning parameters is not known for the functional inequality case (cf. Andrews and Barwick (2012)), and the test size can be sensitive to its choice. This paper therefore does not pursue these approaches, and leave potential power improvement based on a reliable estimater for the contact set for future research.

The asymptotic analysis in this paper assumes complexity of the VC-classes does not grow with sample size. It results in limiting the set of alternatives that can be rejected consistently. One way to enable the test to consistently screen out all the alternatives with the interval VC-class $\mathbb{V}_{int}$ is to let $\underline{h}$ decrease to zero at a certain rate. Recently, Armstrong and Chan (2012) derives an asymptotic distribution of the Kolmgorov-Smirnov statistic of such form, and demonstrated its desirable asymptotic property. An alternative approach to screening out all the alternatives of the functional inequalities (2.1) is to focus on a one-sided L^1-type quantity, such as $\int \max\{q(y,1) - p(y,1), 0\}\, dy + \int \max\{p(y,0) - q(y,0), 0\}\, dy$, and to form a statistic by plugging in the kernel density estimators of $p(y,d)$ and $q(y,d)$ with bandwidth shrinking to zero as $N \to \infty$. Anderson, Linton, and Whang (2011) and Lee, Song, and Whang (2011) develop a test for the functional inequalities based on such one-sided L^1-type statistic. Applicability of these approaches to the current instrument test, and comparisons of test performance between ours and these approaches are worth examining in future work.

5 Monte Carlo Studies and Empirical Applications

5.1 Small sample performance

This section examines the finite sample performance of the bootstrap test by Monte Carlo. We consider a data generating process on a boundary of $\mathcal{H}_0$, so that the type I error of the test equals

Table 1: Monte Carlo Test Size

Monte Carlo iterations 3000, Bootstrap iterations 500.

	Specification of $\mathbb{V}_1$ and $\mathbb{V}_0$																	
	$\mathbb{V}_{half}$						$\mathbb{V}_{int}$ with $h \in \{.3, .5, .7\}$						$\mathbb{V}_{int}$ with $h \in \{.1, .3, .5\}$					
Test Statistic	$T_{N,nw}$			T_N			$T_{N,nw}$			T_N			$T_{N,nw}$			T_N		
Nominal size	.10	.05	.01	.10	.05	.01	.10	.05	.01	.10	.05	.01	.10	.05	.01	.10	.05	.01
(m,n):(100,100)	.12	.06	.01	.13	.07	.02	.11	.06	.01	.12	.06	.01	.11	.06	.01	.12	.06	.02
(100,500)	.12	.06	.01	.15	.08	.02	.12	.06	.02	.15	.08	.02	.10	.06	.01	.14	.08	.02
(500,500)	.11	.06	.01	.11	.06	.01	.10	.05	.01	.11	.06	.01	.11	.06	.01	.12	.06	.02

to a nominal size asymptotically.

$$p(y, D = 1) = q(y, D = 1) = 0.5 \times \mathcal{N}(1, 1),$$
$$p(y, D = 0) = q(y, D = 0) = 0.5 \times \mathcal{N}(0, 1).$$

We consider two specifications of $\mathbb{V}$. One is the half-unbounded interval class $\mathbb{V}_{half}$ and the other is a connected interval class $\mathbb{V}_{int}$ defined in Section 3. When $\mathbb{V}_d = \mathbb{V}_{half}$ is used, we set $\mathcal{Y}^*$ at the 2.5% and 97.5% sample quantile range of the $\{D_i = d\}$ sample. We also look at When $\mathbb{V}_d = \mathbb{V}_{int}$ is used, we include $\bigcup_{h \in \{.3,.5,.7\}} \mathbb{V}_{bin}(h)$ in addition to $\mathbb{V}_{half}$, where $\mathcal{Y}^*$ in the construction of $\mathbb{V}_{bin}(h)$ is equally distanced 128 grid points between the 2.5% and 97.5% sample quantiles of the $\{D_i = d\}$ sample. We set the trimming constant for the sample variance estimate at $\xi = 10^{-4}$.

Table 1 shows the simulated test size. With balanced sample sizes, the test has good size performance even with the sample sizes as small as $(m, n) = (100, 100)$. The unbalanced sample case, $(m, n) = (100, 500)$, shows a slight size distortion for the variance weighted statistic. The test size is not sensitive to whether $\mathbb{V}_{half}$ or $\mathbb{V}_{int}$ is used. Furthermore, Table 3 also shows that making binwidths finer in the construction of $\mathbb{V}_{int}$ does not distort the test size.

In order to see finite sample power of our test procedure, we simulate the empirical rejection rate of the bootstrap test against a fixed alternative. The data generating process is specified as

$$p(y, D = 1) = 0.55 \times N(1, 1.44), \qquad q(y, D = 1) = 0.45 \times N(0.2, 1)$$
$$p(y, D = 0) = 0.45 \times N(0, 1), \qquad q(y, D = 0) = 0.55 \times N(0, 1).$$

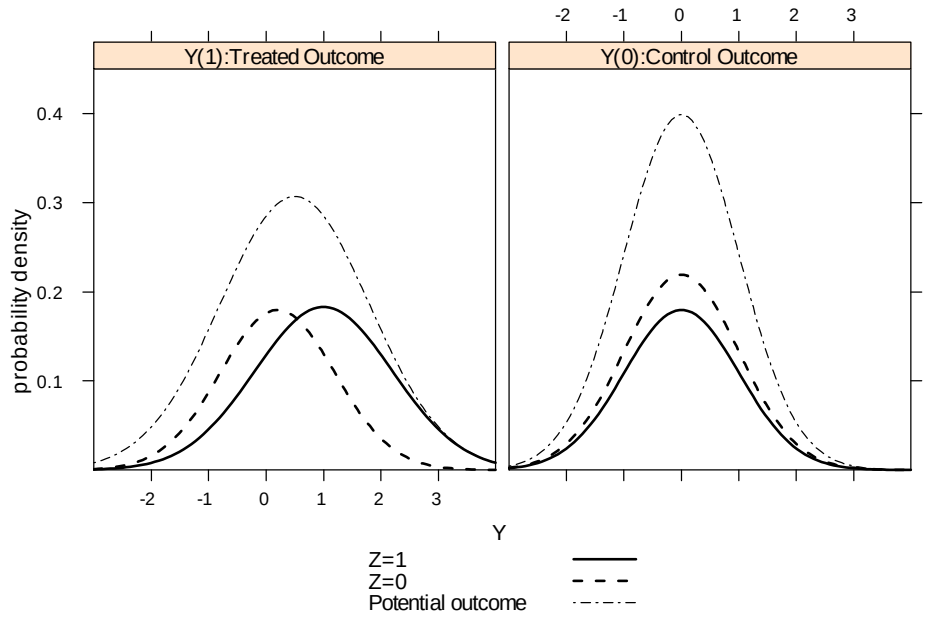

Figure 3: **Monte Carlo for Test Power: Specification of Densities.** *The instrument validity is refuted since for the treated outcomes the two observable densities intersect. In each panel, the density function that covers the other two is a probability density function of the potential outcomes that integrates to one.*

Table 2: Rejection Probabilities Against the Fixed Alternative

Monte Carlo iterations 3000, Bootstrap iterations 500.

	Specification of $\mathbb{V}_1$ and $\mathbb{V}_0$											
	$\mathbb{V}_{half}$				$\mathbb{V}_{int}$ with $h \in \{.3, .5, .7\}$				$\mathbb{V}_{int}$ with $h \in \{.1, .3, .5\}$			
Test Statistic	$T_{N,nw}$		T_N		$T_{N,nw}$		T_N		$T_{N,nw}$		T_N	
Nominal Test Size	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01
(m,n): (100,100)	.20	.06	.28	.13	.20	.06	.30	.14	.20	.05	.28	.13
(100,500)	.33	.12	.37	.17	.33	.11	.39	.16	.33	.12	.40	.18
(500,500)	.84	.61	.90	.77	.87	.63	.97	.86	.84	.60	.95	.80

Figure 4 presents the densities of the specified data generating process. The treated outcome densities since $p(y, D = 1)$ intersects with $q(y, D = 1)$ show counter-evidence to the instrument validity since they intersect. Table 2 presents the simulated rejection probabilities. The first two specifications of $\mathbb{V}_1$ and $\mathbb{V}_0$ in Table 2 are same as the ones employed in the Monte Carlo studies of Table 1. In the third specification of $\mathbb{V}_1$ and $\mathbb{V}_0$, we specify the set of binwidths to include a finer one, $h \in \{0.15, 0.3, 0.45\}$, to see a sensitivity of the rejection probabilities when the class of subsets include finer bins. First, the weighted version of the test statistics has higher rejection probabilities than the non-weighted one at every sample size.[3] Second, the rejection probabilities do not deteriorate as we enrich the class of subsets from $\mathbb{V}_{half}$ to $\mathbb{V}_{int}$. Furthermore, a comparison between $\mathbb{V}_{int}$ with $\underline{h} = 0.3$ and $\mathbb{V}_{int}$ with $\underline{h} = 0.1$ shows that the test power is not sensitive to a choice of the smallest binwidths in the construction of $\mathbb{V}_{int}$. Although this paper does not formally demonstrate to what extent these Monte Carlo findings can be generalized to other specifications of an alternative, we recommend $\mathbb{V}_{int}$ as a convenient choice for $\mathbb{V}_1$ and $\mathbb{V}_0$ in terms of robustness of both size and power properties with respect to a choice of smallest binwidth $\underline{h}$, and (ii) the ability in screening out a large class of alternatives than $\mathbb{V}_{half}$.

5.2 Empirical Applications

We illustrate a use of our test using the following two data sets. The first dataset is the draft lottery data during Vietnam era used in Angrist (1991). The second dataset is from Card (1993) on returns to schooling using geographical proximity to college as an instrument.

[3] In the small sample and the unbalanced sample cases, the power gain of the weighted statistic can be driven by its slight upward size distortion as seen in Table 1.

5.2.1 Draft Lottery Data

The draft lottery data consist of a sample of 10,101 white men, born in 1950-1953 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 or not. Since the enrollment for the military service possibly involves self-selection based on one's future earning, the veteran status is not considered to be randomly assigned. In order to solve this endogeneity issue, Angrist (1991) constructs the binary indicator of the draft eligibility, which is randomly assigned based on one's birthdate through the draft lotteries. A justification of the instrumental validity here is that the instrument is generated being independent of any individual characteristics. Hence, it is reasonable to argue that the instrument satisfies RTA. On the other hand, the validity of MSR is less credible since the existence of defiers are not eliminated by the sampling design, i.e., in the sample there are observations who participate to the military service even though they are not initially drafted.

As a exploratory tool for summarizing the shapes of $p(y,d)$ and $q(y,d)$, Figure 4 plots the kernel density estimates for the observed outcome distribution multiplied by the selection probability. We observe that, except for slight violations of the inequalities (2.1) at the tail parts of $Y(0)$'s densities, the kernel densities overall exhibit the nested structures. Table 3 shows the result of our test. We specify classes of subsets to be interval classes, $\mathbb{V}_1 = \mathbb{V}_0 = \mathbb{V}_{int}$, where the binwidths for $\mathbb{V}_{bin}(h)$ are 20 grid points between 0.1 and 2.0. The p-values of the bootstrap test are one for both $T_{N,nw}$ and T_N, so we do not refute the instrumental validity from the data.

5.2.2 Returns to Education: Proximity to College Data

The Card data is based on National Longitudinal Survey of Young Men (NLSYM) began in 1966 with age 14-24 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. Observations of years of education and wage level are based on the follow-ups' educational attainment and wage level responded in the interview in 1976.

The idea of using proximity to college as an instrument is stated as follows. Presence of a nearby college reduces a cost of college education by allowing students to live at home, while one's inherited ability is presumably independent of his birthplace. Compliers in this context can be considered to be those who grew up in relatively low-income families and who were not able to go to college without living with their parents. We make the educational level as a binary treatment

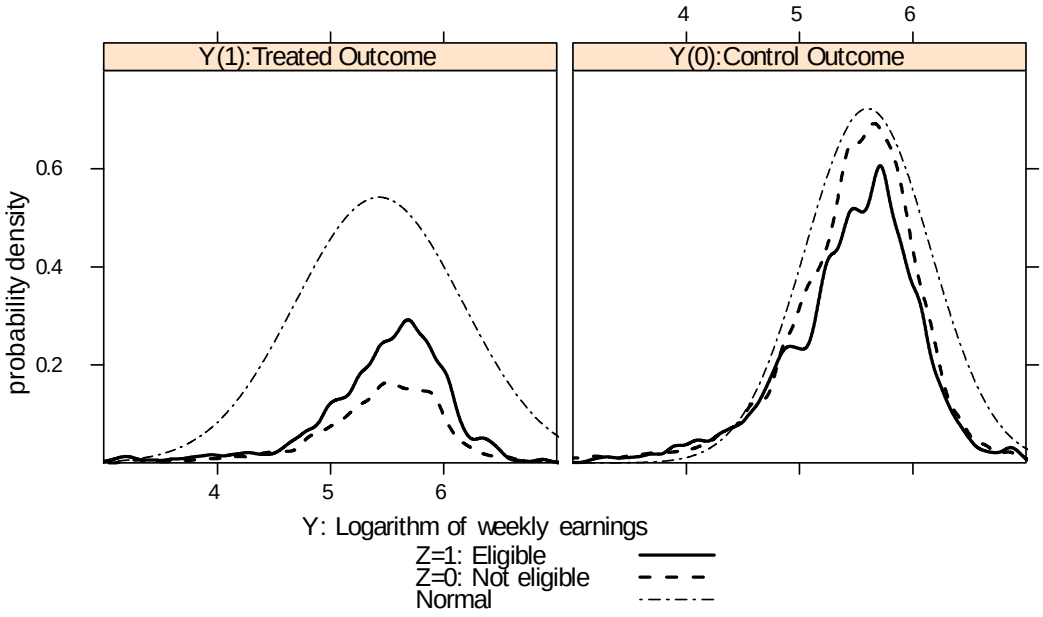

Figure 4: **Kernel Density Estimates for the Draft Lottery Data.** The *Gaussian kernel with bandwidth 0.06 is used. In each panel, we draw a normal density to illustrate the scale of the subdensities.*

Table 3: Test Results of the Empirical Applications

Bootstrap iterations 500

Specifications of $\mathbb{V}_1$ and $\mathbb{V}_0$	$\mathbb{V}_1$=$\mathbb{V}_0$=$\mathbb{V}_{int}\cup\left[\bigcup_{h\in\{.1,.2,...,2.0\}}\mathbb{V}_{bin}(h)\right]$					
	Draft lottery data		Proximity to college data			
	Full sample		Full sample		Restricted sample	
sample size (m,n)	(2780,7321)		(2053,957)		(1047,144)	
$\Pr(D=1\|Z=1)$, $\Pr(D=1\|Z=0)$	0.31, 0.19		0.29, 0.22		0.35,0.24	
Test Statistics	$T_{N,nw}$	T_N	$T_{N,nw}$	T_N	$T_{N,nw}$	T_N
Bootstrap test, p-value	1.00	1.00	0.00	0.00	0.00	0.00

which indicates one's education years to be greater or equal to 16 years, meaning that the treatment can be roughly 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 any demographic covariates. This simplification raises a concern for the violation of RTA. For instance, one's region of residence, or whether they were born in the standard metropolitan area or rural area may affect one's wage levels and the proximity to colleges if the urban areas are more likely to have colleges and has higher wage level compared with the rural areas. This kind of confounder may contaminate the validity of RTA. In fact, Card (1993) emphasizes an importance of controlling for regions, residence in the urban area, race, job experience, and parent's education in order to make use of the college proximity as an instrument.

Figure 5 presents the kernel density estimates for observed outcome densities. In contrast to Figure 4, we observe that the kernel density estimates in Figure 5 clearly intersect, especially, for those of control outcome. Our test procedure yields zero p-value and this provides an empirical evidence that, without controlling for any covariates, college proximity is not a valid instrument.

Consider next how the test result changes once we control for some covariates. Controlling discrete covariates can be done by simply making the whole analysis conditional on the specified value of the covariates. We consider restricting the sample to be white workers (black dummy is zero), not living in south states in 1966 (south66 dummy is zero), and living in a metropolitan area in 1966 (SMSA66 dummy is one). That is, we are controlling for race, whether or not one grew up in southern states, and whether or not one grew up in urban area. The size of the restricted

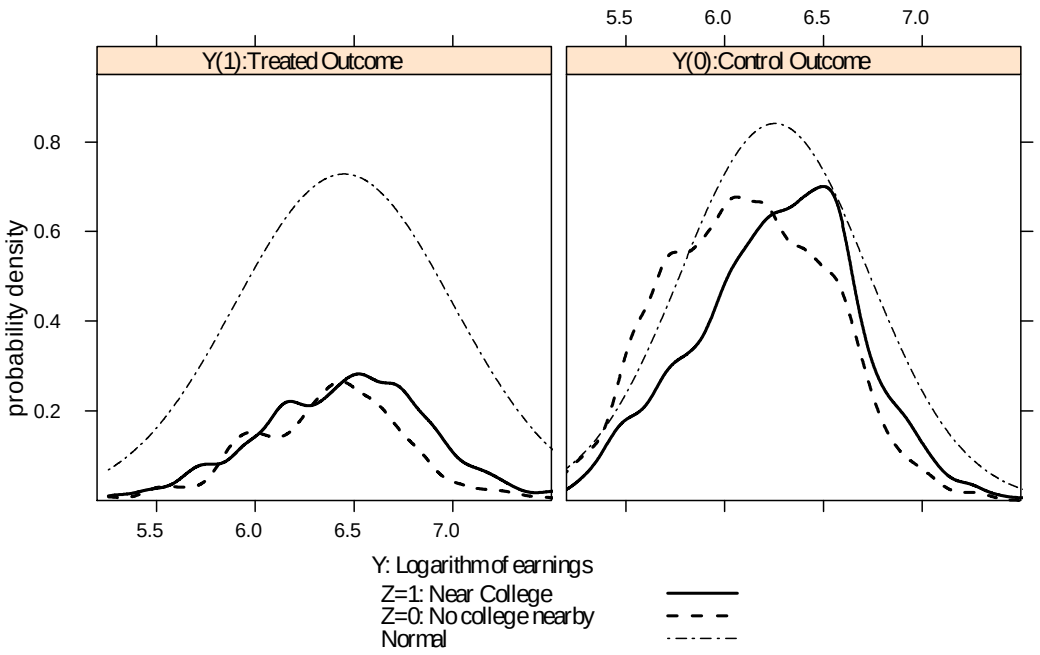

Figure 5: **Kernel Density Estimates for the Proximity to College Data (No covariates controlled)**. *The Gaussian kernel with bandwidth 0.07 is used. In each panel, we draw a normal density to illustrate the scale of the estimated subdensities.*

sample is 1191 ($m = 1047$, $n = 144$). Figure 6 indicates that the kernel density estimates present less evident violation of instrument validity compared with Figure 5. The p-value of our test turns out to be zero. Hence, the instrument validity is refuted even with these covariates conditioned. Note that the test results presented here do not invalidate the estimation results of Card (1993), because he treats education years are treated as a multi-valued treatment and uses a richer set of covariates including continuous ones.

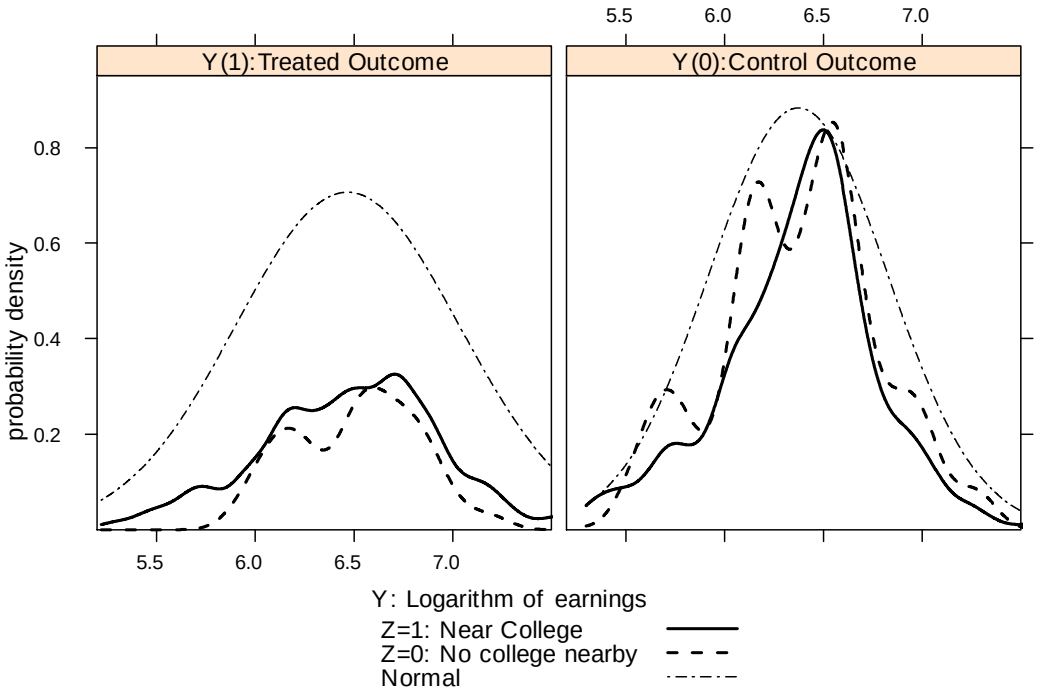

Figure 6: **Kernel Density Estimates for the Proximity to College Data (white workers, not living in south states, and living in a metropolitan area)**. *The Gaussian kernel with bandwidth 0.1 is used. In each panel, we draw a normal density to illustrate the scale of the estimated subdensities.*

6 Concluding Remarks

In this paper, we develop a test procedure to empirically check the conditions of the instrumental validity of Imbens and Angrist (1994). The test statistic measures negativity of the complier's outcome densities by the supremum statistic, and the bootstrap algorithm is developed for obtaining the asymptotically valid critical values. Regarding a choice for classes of subsets $\mathbb{V}_1$ and $\mathbb{V}_0$, the interval classes can be attractive specifications in terms of robustness of test performance and the asymptotic screening power.

In every empirical study where a discrete instrument is used to infer causal effects of a binary treatment, we recommend to report the p-values of our test, with acknowledging that the instrument validity is a refutable but non-verifiable assumption.

A Appendix A: Proof of Proposition 2.1

Proof of Proposition 1. Denote the population distribution of the types by $\pi_t \equiv \Pr(T = t)$, $t \in \{c, n, a, df\}$. Under RTA, $P(B, 1)$ for any Borel set $B \subset Y$ is expressed as,

$$\begin{aligned} P(B,1) &= \Pr(Y_1 \in B, D_1 = 1 | Z = 1) \\ &= \Pr(Y_1 \in B, D_1 = 1) \\ &= \Pr\left(Y_1 \in B, T \in \{a, c\}\right) \\ &= \Pr(Y_1 \in B | T = a)\pi_a + \Pr(Y_1 \in B | T = c)\pi_c. \end{aligned} \tag{A.1}$$

The first line follows because the event $\{Y \in B, D = 1 | Z = 1\}$ is identical to $\{Y_1 \in B, D_1 = 1 | Z = 1\}$. The second equality follows by RTA, and the third equality follows by the definition of selection types.

The similar operation to $Q(B, 1)$ yields

$$Q(B,1) = \Pr(Y_1 \in B | T = a)\pi_a + \Pr(Y_1 \in B | T = df)\pi_{df}. \tag{A.2}$$

Under MSR, $\pi_{df} = 0$, so the difference between (A.1) and (A.2) is

$$P(B,1) - Q(B,1) = \Pr(Y_1 \in B | T = c)\pi_c \geq 0.$$

This proves the first inequality of the proposition. The second inequality of the proposition is proven in a similar way.

For a proof of converse statement, let P and Q satisfying the inequalities (2.1) be given. Let $p(y, d)$ and $q(y, d)$ be the densities of $P(\cdot, d)$ and $Q(\cdot, d)$ with respect to a common dominating

measure μ on Y. In what follows, we show that there exists a joint distribution of (Y_1, Y_0, T, Z) that is compatible with P and Q and satisfies RTA and MSR. 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. Consider nonnegative functions $h_{Y_d,t}(y)$, $d = 1, 0$, $t \in \{c, n, a, df\}$,

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

where $\gamma_{Y_1}(y)$ and $\gamma_{Y_0}(y)$ are arbitrary nonnegative functions supported on Y that satisfy $\int_{\mathcal{Y}} \gamma_{Y_1}(y)d\mu = P(Y,0)$ and $\int_{\mathcal{Y}} \gamma_{Y_0}(y)d\mu = Q(Y,1)$. We construct a probability law of (Y_1, Y_0, T) given Z on the product σ-algebra as

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

Note that this is a probability measure on the product sigma-algebra of $\mathcal{Y}^2 \times \{c, a, n, df\}$, 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 RTA and MSR by the construction, and it induces the given data generating process. i.e., $\Pr(Y \in B, D = d|Z = z)$ implied by the proposed probability distribution of $(Y_1, Y_0, T|Z)$ coincides with the given P and Q. This completes the proof. ■

B Appendix B: Proof of Proposition 2

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 $\mathcal{X} \equiv \mathcal{Y} \times \{0,1\}$ generated by the VC-classes $\mathbb{V}_1$ and $\mathbb{V}_0$.

$$\mathcal{F} = \left\{1_{\{V,1\}}(Y,D) : V \in \mathbb{V}_1\right\} \cup \left\{1_{\{V,0\}}(Y,D) : V \in \mathbb{V}_0\right\},$$

where $1_{\{V,d\}}(Y,D)$ is an indicator function for event $\{Y \in V, D = d\}$. The Borel σ-algebra of $\mathcal{X}$ is denoted by $\mathcal{B}(\mathcal{X})$. Note that, given $\mathbb{V}_1$ and $\mathbb{V}_0$ are VC-classes, $\mathcal{F}$ is a VC-class of functions. 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, $P(f) \equiv P(V,d)$ and $P_m(f) \equiv P_m(V,d)$ for $f = 1\{V,d\} \in \mathcal{F}$. We denote the empirical process 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 $\rho_P(f,f') = P(|f - f'|)$ be a seminorm on $\mathcal{F}$ defined with respect to probability measure $P \in \mathcal{P}$. 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 samples probability measures that drift with the sample sizes, 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 $\sigma_{P_m,Q_n}^2(\cdot,\cdot)$ be its sample analogue,

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

Note that, with the current notation, $\sigma_{P_m,Q_m}^2(V,d)$ defined in Section 3 of the main text is equivalent to $\sigma_{P_m,Q_n}^2(f,f)$, for $f = 1_{\{V,d\}}$. 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 that, for every $\epsilon > 0$, $\lim_{N\to\infty} \Pr_{P^{[m]},Q^{[n]}}(|W_N - c| > \epsilon) = 0$. 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 proof of Proposition 2.

Lemma B.1 *Suppose Condition RG (a). 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. The assumption that $\mathcal{F}$ is a class of indicator functions for a VC-class of subsets guarantees application of the Glivenko-Cantelli theorem uniform in $\mathcal{P}$ (Theorem 2.8.1 of van der Vaart and Wellner (1996)). Hence, this lemma follows as its corollary. ■

Lemma B.2 *Suppose Condition RG (b) and (c). 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,*

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

where $\mathcal{B}(\mathcal{X})$ *is the Borel* σ*-algebra on* $\mathcal{X}$*.*

Proof. Suppose $\lim_{m\to\infty} \sup_{B \in \mathcal{B}(\mathcal{X})} \left| \left(P^{[m]} - P_0\right)(B) \right| = 0$ is false, that is, there exists $\xi > 0$ and a sequence $\{B_m \in \mathcal{B}(\mathcal{X}) : m = 1, 2, \dots\}$ such that $\limsup_{m\to\infty} \left| \left(P^{[m]} - P_0\right)(B_m) \right| > \xi$ holds. By uniform tightness of Condition RG (c), there exist a compact set $K \subset \mathcal{B}(\mathcal{X})$ such that

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

holds. We metricize $\mathcal{B}(\mathcal{X})$ by 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 (b) and δ_d is the mass measure on $\{0, 1\}$. Since $\{B_m \cap K : m = 1, 2 \dots, \}$ is a sequence in a compact subset of $\mathcal{B}(\mathcal{X})$, there exists a subsequence b_m, such that$\{B_{b_m} \cap K\}$ converges to $B^* \in \mathcal{B}(\mathcal{X})$ in terms of metric $d_{\mathcal{B}(\mathcal{X})}(\cdot, \cdot)$ and

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

holds. Under the bounded density assumption of Condition RG (b), it holds that, for a finite constant M of Condition RG (b),

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

Hence, (B.1) implies

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

By Condition RG (b), P_0 as a weak limit of $\left\{P^{[m]} : m = 1, 2, \dots\right\}$ is absolutely continuous in $\mu \times \delta_d$, so, for $\bar{B}^*$ the closure of B^*, $P_0\left(\bar{B}^*\right) = P_0(B^*)$ holds. 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), so $\lim_{m\to\infty} \sup_{B \in \mathcal{B}(\mathcal{X})} \left| \left(P^{[m]} - P_0\right)(B) \right| = 0$ holds. ■

Lemma B.3 *Suppose Condition RG (a)-(c). 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* $m \to \infty$.

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

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

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

$$\begin{aligned}
&(i) \qquad \sup_{f,g \in \mathcal{F}} \left| \sigma^2_{P_m^{[m]}, Q_n^{[m]}}(f,g) - \sigma^2_{P^{[m]}, Q^{[m]}}(f,g) \right| \underset{P^{[m]}, Q^{[n]}}{\longrightarrow} 0, \\
&(ii) \qquad \sup_{f,g \in \mathcal{F}} \left| \sigma_{P_m^{[m]}, Q_n^{[m]}}(f,g) - \sigma_{P^{[m]}, Q^{[m]}}(f,g) \right| \underset{P^{[m]}, Q^{[n]}}{\longrightarrow} 0, \quad \text{as } N \to \infty
\end{aligned}$$

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.3. Furthermore, since class of indicator functions $\{fg : f, g \in \mathcal{F}\}$ is also a VC-class, $\sup_{f,g \in \mathcal{F}} \left| \left(P_m^{[m]} - P^{[m]} \right)(fg) \right| \underset{P^{[m]}}{\longrightarrow} 0$ also holds by Lemma B.3. 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 first conclusion (i) follows. (ii) is an immediate corollary of (i). ■

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 2.8.2 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 $\left[R(|f - f'|^2)\right]^{1/2}$,[4] and (ii) there exists $P^* \in \mathcal{P}$ such that $\lim_{m\to\infty} \sup_{f,g\in\mathcal{F}} \left\{\left|\rho_{P^{[m]}}(f, g) - \rho_{P^*}(f, g)\right|\right\} = 0$, 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, e.g., 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

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

Hence, to validate Condition (ii) with $P^* = P_0$, it suffices to have $\lim_{m\to\infty} \sup_{B\in\mathcal{B}(\mathcal{X})} \left|\left(P^{[m]} - P_0\right)(B)\right| = 0$, 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 two samples, which converges weakly to* (P_0, Q_0)*, as* $N \to \infty$*. Then, stochastic processes indexed by VC-class of indicator functions* $\mathcal{F}$*,*

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

converges weakly to mean zero Gaussian processes $v_0(\cdot) = \frac{(1-\lambda)^{1/2} G_{P_0}(\cdot) - \lambda^{1/2} G_{Q_0}(\cdot)}{\sigma_{P_0,Q_0}(\cdot,\cdot)}$*, 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 of seminorms such as, $\rho_{P^{[m]}+Q^{[n]}} = P^{[m]}\left(|f - f'|\right) + Q^{[m]}\left(|f - f'|\right)$, 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 (i) and Lemma B.5, and (iii) follows as a corollary

[4] 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 Lemma B.2. To verify (ii), consider, for $f, g \in \mathcal{F}$ with $\rho_{P^{[m]}+Q^{[n]}}(f,g) < \delta$,

$$\begin{aligned} |v_N(f) - v_N(g)| &\leq \left| \frac{1}{\sigma_{P_m^{[m]},Q_n^{[m]}}(f,f)} - \frac{1}{\sigma_{P_m^{[m]},Q_n^{[m]}}(g,g)} \right| \left| (1-\lambda)^{1/2} G_{m,P^{[m]}}(g) - \lambda^{1/2} G_{n,Q^{[n]}}(g) \right| \qquad \text{(B.7)} \\ &\quad + \frac{(1-\lambda)^{1/2} \left| G_{m,P^{[m]}}(f) - G_{m,P^{[m]}}(g) \right| + \lambda^{1/2} \left| G_{n,Q^{[n]}}(f) - G_{n,Q^{[n]}}(g) \right|}{\sigma_{P_m^{[m]},Q_n^{[m]}}(g,g)} + o\left(\left| \hat{\lambda} - \lambda \right| \right). \end{aligned}$$

Note that

$$\begin{aligned} \left| \frac{1}{\sigma_{P_m^{[m]},Q_n^{[m]}}(f,f)} - \frac{1}{\sigma_{P_m^{[m]},Q_n^{[m]}}(g,g)} \right| &= \left| \frac{1}{\sigma_{P^{[m]},Q^{[n]}}(f,f)} - \frac{1}{\sigma_{P^{[m]},Q^{[n]}}(g,g)} \right| + o_{P^{[m]},Q^{[n]}}(1) \\ &= \frac{\left| \sigma_{P^{[m]},Q^{[n]}}(f,f) - \sigma_{P^{[m]},Q^{[n]}}(g,g) \right|}{\sigma_{P^{[m]},Q^{[n]}}(f,f)\sigma_{P^{[m]},Q^{[n]}}(g,g)} + o_{P^{[m]},Q^{[n]}}(1) \\ &\leq \frac{1}{\kappa} \left| \sigma_{P^{[m]},Q^{[n]}}(f,f) - \sigma_{P^{[m]},Q^{[n]}}(g,g) \right| + o_{P^{[m]},Q^{[n]}}(1), \qquad \text{(B.8)} \end{aligned}$$

where the first equality follows from Lemma B.4 (i), and the third inequality follows by Condition RG (d). Noting the following inequalities

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

we have

$$\left| \sigma_{P^{[m]},Q^{[m]}}(f,f) - \sigma_{P^{[m]},Q^{[m]}}(g,g) \right| \leq \frac{\rho_{P^{[m]}+Q^{[n]}}(f,g)}{2\kappa^{1/2}}. \qquad \text{(B.9)}$$

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

$$\begin{aligned} \left| \frac{1}{\sigma_{P_m^{[m]},Q_n^{[m]}}(f,f)} - \frac{1}{\sigma_{P_m^{[m]},Q_n^{[m]}}(g,g)} \right| &\leq \frac{1}{2\kappa^{3/2}} \rho_{P^{[m]}+Q^{[n]}}(f,g) + o_{P^{[m]},Q^{[m]}}(1) \qquad \text{(B.10)} \\ &\leq \frac{\delta}{2\kappa^{3/2}} + o_{P^{[m]},Q^{[m]}}(1). \end{aligned}$$

Hence,

$$\begin{aligned} \sup_{\rho_{P^{[m]}+Q^{[n]}}(f,g)<\delta} |v_N(f) - v_N(g)| &\leq \frac{\delta}{2\kappa^{3/2}} \left| (1-\lambda)^{1/2} G_{m,P^{[m]}}(g) - \lambda^{1/2} G_{n,Q^{[n]}}(g) \right| \qquad \text{(B.11)} \\ &\quad + \left(\frac{1-\lambda}{\kappa} \right)^{1/2} \sup_{\rho_{P^{[m]}+Q^{[n]}}(f,g)<\delta} \left| G_{m,P^{[m]}}(f) - G_{m,P^{[m]}}(g) \right| \\ &\quad + \left(\frac{\lambda}{\kappa} \right)^{1/2} \sup_{\rho_{P^{[m]}+Q^{[n]}}(f,g)<\delta} \left| G_{n,Q^{[n]}}(f) - G_{n,Q^{[n]}}(g) \right| + o_{P^{[m]},Q^{[m]}}(1). \end{aligned}$$

By noting $\rho_{P^{[m]}}(f,g) \leq \rho_{P^{[m]}+Q^{[n]}}(f,g)$ 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 this equality follows since the uniform convergence of $G_{m,P^{[m]}}(f)$ as established by Lemma B.5 implies

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

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). ■

B.3 Proof of Proposition 3.1

Let $\mathcal{F}_1 = \left\{1_{\{V,1\}}(Y,D) : V \in \mathbb{V}_1\right\}$ and $\mathcal{F}_0 = \left\{1_{\{V,0\}}(Y,D) : V \in \mathbb{V}_0\right\}$ be a subclass of $\mathcal{F}$, corresponding to the VC-class of subsets in $\mathcal{Y}$ with $d=1$ and $d=0$, respectively. We want to show

$$\lim\sup_{N\to\infty} \sup_{(P,Q)\in\mathcal{H}_0} \Pr\left(T_N > c_{N,1-\alpha}\right) \leq \alpha, \tag{B.12}$$

where

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

Consider a sequence $\left(P^{[m(N)]}, Q^{[n(N)]}\right) \in \mathcal{H}_0$ at which $\Pr_{P^{[m(N)]},Q^{[n(N)]}}(T_N > c_{N,1-\alpha})$ 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(c)), 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 $N \to \infty$. 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 $\lim\sup_{N\to\infty} \Pr_{P^{[m]},Q^{[n]}}(T_N > c_{N,1-\alpha}) \leq \alpha$ holds.

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

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

where

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

By the almost sure representation theorem (see, e.g., Theorem 9.4 of Pollard (1990)), weak convergence of $\left(v_N(\cdot), P_m^{[m]}(\cdot), Q_n^{[n]}(\cdot), \hat{\sigma}^2_{P^{[m]}, Q^{[n]}}(\cdot,\cdot)\right)$ 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)

$$\sup_{f \in \mathcal{F}} |\tilde{v}_N(f) - \tilde{v}_0(f)| \to 0, \tag{B.13}$$

$$\sup_{f \in \mathcal{F}} \left|\tilde{P}_m^{[m]}(f) - P_0(f)\right| \to 0, \tag{B.14}$$

$$\sup_{f \in \mathcal{F}} \left|\tilde{Q}_n^{[n]}(f) - Q_0(f)\right| \to 0, \text{ and} \tag{B.15}$$

$$\sup_{f,g \in \mathcal{F}} \left|\tilde{\sigma}^2_{P_m^{[m]}, Q_m^{[n]}}(f,g) - \sigma^2_{P_0,Q_0}(f,g)\right| \to 0, \text{ as } N \to \infty \ \mathbb{P}\text{-a.s.} \tag{B.16}$$

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 $\tilde{h}_N(f) = \sqrt{\frac{mn}{N}} \frac{P^{[m]}(f) - Q^{[n]}(f)}{\tilde{\sigma}^2_{P_m^{[m]}, Q_n^{[n]}}(f,f)}$. 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 B.7 shown below, $\tilde{c}_{N,1-\alpha} \to c_{1-\alpha}$ as $N \to \infty$, $\mathbb{P}$-a.s. holds, where $c_{1-\alpha}$ is the $(1-\alpha)$-th quantile of statistic

$$T_H \equiv \max \left\{ \begin{array}{c} \sup_{f \in \mathcal{F}_1} \{-G_{H_0}(f)/\sigma_{P_0,Q_0}(f,f)\} \\ \sup_{f \in \mathcal{F}_0} \{G_{H_0}(f)/\sigma_{P_0,Q_0}(f,f)\} \end{array} \right\}, \tag{B.17}$$

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

Since $\Pr_{P^{[m]},Q^{[n]}}(T_N > c_{N,1-\alpha}) = \mathbb{P}\left(\tilde{T}_N > \tilde{c}_{N,1-\alpha}\right)$ 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 $(\Omega, B(\Omega), \mathbb{P})$, 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

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

Hence, in what follows, we aim to find 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 $\mathcal{F}_1$,

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

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

$$\begin{aligned}\sup_{f\in\mathcal{F}_1}\left\{-\tilde{v}_N(f) - \tilde{h}_N(f)\right\} &= \max\left\{\begin{array}{c}\sup_{f\in\mathcal{F}_{1,\eta_N}}\left\{-\tilde{v}_N(f) - \tilde{h}_N(f)\right\} \\ \sup_{f\in\mathcal{F}_1\setminus\mathcal{F}_{1,\eta_N}}\left\{-\tilde{v}_N(f) - \tilde{h}_N(f)\right\}\end{array}\right\} \\ &\leq \max\left\{\begin{array}{c}\sup_{f\in\mathcal{F}_{1,\eta_N}}\{-\tilde{v}_N(f)\} \\ \sup_{f\in\mathcal{F}_1\setminus\mathcal{F}_{1,\eta_N}}\left\{-\tilde{v}_N(f) - \tilde{h}_N(f)\right\}\end{array}\right\} \\ &\leq \max\left\{\begin{array}{c}\sup_{f\in\bigcup_{N'\geq N}\mathcal{F}_{1,\eta_{N'}}}\{-\tilde{v}_N(f)\} \\ \sup_{f\in\mathcal{F}_1\setminus\mathcal{F}_{1,\eta_N}}\{-\tilde{v}_N(f)\} - \eta_N\end{array}\right\}, \quad \text{for every } N, \qquad \text{(B.18)}\end{aligned}$$

where the second line follows since $\tilde{h}_N(f) \geq 0$ for all $f \in \mathcal{F}_1$ under the assumption that $\left(P^{[m]}, Q^{[n]}\right) \in \mathcal{H}_0$, the third line follows because $\tilde{h}_N(f) \geq \eta_N$ 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

$$\sup_{f\in\mathcal{F}_1\setminus\mathcal{F}_{1,\eta_N}}\{-\tilde{v}_N(f)\} - \eta_N \to -\infty, \text{ as } N \to \infty, \mathbb{P}\text{-a.s.} \qquad \text{(B.19)}$$

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

$$\sup_{f\in\bigcup_{N'\geq N}\mathcal{F}_{1,\eta_{N'}}}\{-\tilde{v}_N(f)\} \to \sup_{f\in\mathcal{F}_{1,\infty}}\{-\tilde{v}_0(f)\}, \text{ as } N \to \infty, \mathbb{P}\text{-a.s.}, \qquad \text{(B.20)}$$

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

$$\liminf_{N\to\infty} \left\{ \sqrt{\hat{\lambda}(1-\hat{\lambda})} \frac{P^{[m]}(f) - Q^{[n]}(f)}{\tilde{\sigma}^2_{P_m^{[m]},Q_n^{[n]}}(f,f)} \right\} = 0. \tag{B.21}$$

Since $\tilde{\sigma}^2_{P_m^{[m]},Q_n^{[n]}}(f,f)$ converges to a constant bounded away from zero by (B.16) and Condition RG (c), and $P^{[m]}(f) - Q^{[n]}(f)$ converges to $P_0(f) - Q_0(f)$ by Lemma B.2, any f satisfying (B.21) belongs to $\mathcal{F}_1^*$. Hence, we have

$$\sup_{f\in\mathcal{F}_{1,\infty}} \{-\tilde{v}_0(f)\} \leq \sup_{f\in\mathcal{F}_1^*} \{-\tilde{v}_0(f)\} \quad \mathbb{P}\text{-a.s.} \tag{B.22}$$

By combining (B.18), (B.19), (B.20), and (B.22), we obtain

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

In a similar manner, it can be shown that

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

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, continuity of the cdf of $\tilde{T}^*$ at $c_{1-\alpha}$ follows by the absolute continuity theorem for the supremum of Gaussian processes (Tsirelson (1975)). To establish the second requirement of (B), note that statistic T_H defined in (B.17) can be written as

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

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 (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 $\mathcal{F}$, $\mathcal{F}^* = \mathcal{F}_1^* \cup \mathcal{F}_0^*$,

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

Note that $Var(u(f)) = Var(u_H(f)) = 1$ holds at every $f \in \mathcal{F}^*$. As for the covariance kernels of $u(\cdot)$ and $u_H(\cdot)$, we have, for $f, g \in \mathcal{F}_1^*$ or $f, g \in \mathcal{F}_0^*$,

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

where the second equality follows since $P_0(\cdot) = Q_0(\cdot)$ on $f \in \mathcal{F}^*$, and the third equality follows since $H_0(f) = P_0(f) = Q_0(f)$ and $\sigma_{H_0}(f,f) = \sigma_{P_0,Q_0}(f,f)$ holds for all $f \in \mathcal{F}^*$. Also, for $f \in \mathcal{F}_1^*$ and $g \in \mathcal{F}_0^*$,

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

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, $P\left(\tilde{T}^* > c_{1-\alpha}\right) \leq \Pr(T_H > c_{1-\alpha}) = \alpha$. This completes the proof of Proposition 3.1 (i).

A proof of (ii) proceeds in a similar way, but slightly simpler than the proof of (i). We omit a proof of (ii) for brevity.

To prove claim (iii), assume that the first inequality of (2.1) is violated at some $f^* \in \mathcal{F}_1$, i.e., the true data generating process satisfies $P(f^*) < Q(f^*)$. Then, we have

$$\begin{aligned} T_N &= \max\left\{ \begin{array}{l} \sup_{f\in\mathcal{F}_1}\left\{\sigma_{P_m,Q_n}^{-1}(f,f)\left(\hat{\lambda}^{1/2}Q_n(f) - (1-\hat{\lambda})^{1/2}P_m(f)\right)\right\} \\ \sup_{f\in\mathcal{F}_0}\left\{\sigma_{P_m,Q_n}^{-1}(f,f)\left((1-\hat{\lambda})^{1/2}P_m(f) - \hat{\lambda}^{1/2}Q_n(f)\right)\right\} \end{array} \right\} \\ &\geq \sigma_{P_m,Q_n}^{-1}(f^*,f^*)\left(\hat{\lambda}^{1/2}G_{n,Q}(f^*) - (1-\hat{\lambda})^{1/2}G_{m,P}(f^*)\right) + \sqrt{\frac{mn}{N}}\frac{Q(f^*) - P(f^*)}{\sigma_{P_m,Q_n}^{-1}(f^*,f^*)}, \end{aligned} \tag{B.23}$$

where the second term of (B.23) 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.

B.4 Lemma on Convergence of the Bootstrap Critical Values

The proof of Proposition 3.1 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 $(\Omega, B(\Omega), \mathbb{P})$ and the random objects with "tilde" referred to in what follows are the ones defined in the proof of Proposition 3.1 (i) by the almost sure representation theorem.

Lemma B.7 *Suppose Condition RG. Let $\tilde{c}_{N,1-\alpha}$ be the bootstrap critical value of Algorithm 3.1 constructed from $\tilde{H}_N^{[N]} = (1-\hat{\lambda})\tilde{P}_m^{[m]} + \hat{\lambda}\tilde{Q}_n^{[n]}$, 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*

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

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

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

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

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 $\left\{\tilde{H}_N^{[N]} : N = 1, 2, \dots\right\}$. We shall show that $G^*_{m,H_N^{[N]}}(\cdot)$ converges weakly to H_0-brownian bridge processes $G_{H_0}(\cdot)$ for $\mathbb{P}$-almost every sequence $\left\{\tilde{H}_N^{[N]} : N = 1, 2, \dots\right\}$. Let $\left(X_i^1 \in \mathcal{X} : i = 1, \dots, m\right)$ be the points of support of $\tilde{P}_m^{[m]}$ and $\left(X_j^0 \in \mathcal{X} : j = 1, \dots, n\right)$ be the points of support of $\tilde{Q}_n^{[n]}$. Using the multinomial random variables and the point mass measure δ_X, bootstrap empirical process $G^*_{m,\tilde{H}_N^{[N]}}$ can be written as

$$G^*_{m,\tilde{H}_N^{[N]}} = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \left[M_{N,i}^1 - (1-\hat{\lambda})\right] \delta_{X_i^1} + \left(\frac{1-\lambda}{\lambda}\right)^{1/2} \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left[M_{N,j}^0 - \frac{m}{n}\hat{\lambda}\right] \delta_{X_j^0},$$

where $\left(M_{N,1}^1, \dots, M_{N,m}^1, M_{N,1}^0, \dots, M_{N,n}^0\right)$ are multinomial random variables following

$$\left(M_{N,1}^1, \dots, M_{N,m}^1, M_{N,1}^0, \dots, M_{N,n}^0\right) \sim \mathcal{MN}(m, \frac{1-\hat{\lambda}}{m}, \dots, \frac{1-\hat{\lambda}}{m}, \frac{\hat{\lambda}}{n}, \dots, \frac{\hat{\lambda}}{n}),$$

and independent of $\left(\tilde{P}_m^{[m]}, \tilde{Q}_n^{[n]}\right)$. Following Theorem 3.6.1 of van der Vaart and Wellner (1996), we can replace the multinomial random variables with the independent Poisson variables, and obtain

the following approximations

$$\begin{aligned} G^*_{m,\tilde{H}_N^{[N]}}(\cdot) &= \frac{1}{\sqrt{m}}\sum_{i=1}^{m}\left[\tilde{M}^1_{N,i}-(1-\lambda)\right]\left(\delta_{X_i^1}-P^{[m]}\right) \\ &\quad+\left(\frac{1-\lambda}{\lambda}\right)^{1/2}\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left[\tilde{M}^0_{N,j}-\frac{\lambda^2}{1-\lambda}\right]\left[\delta_{X_j^0}-Q^{[n]}\right] \\ &\quad+O\left(\sup_{f\in\mathcal{F}}\left|\tilde{P}_m^{[m]}-P^{[m]}\right|\right)+O\left(\sup_{f\in\mathcal{F}}\left|\tilde{Q}_n^{[n]}-Q^{[n]}\right|\right), \end{aligned}$$

where $\left(\tilde{M}^1_{N,i}: i=1,\ldots,m\right)$ are iid Poisson random variables with mean $(1-\lambda)$ and $\left(\tilde{M}^0_{N,j}: j=1,\ldots,n\right)$ are iid Poisson random variables with mean $\frac{\lambda^2}{1-\lambda}$, and $O\left(\sup_{f\in\mathcal{F}}\left|\tilde{P}_m^{[m]}-P^{[m]}\right|\right)+O\left(\sup_{f\in\mathcal{F}}\left|\tilde{Q}_n^{[n]}-Q^{[n]}\right|\right)=o(1)$, $\mathbb{P}$-a.s. by Lemma B.2 and the convergence properties of $\left(\tilde{P}_m^{[m]},\tilde{Q}_n^{[n]}\right)$ as given in (B.14) and (B.15). By Lemma B.4, we have

$$\begin{aligned} &\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\left(\delta_{X_i^1}-P^{[m]}\right)(\cdot) \quad\rightsquigarrow\quad G_{P_0}(\cdot), \\ &\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left(\delta_{X_j^0}-Q^{[n]}\right)(\cdot) \quad\rightsquigarrow\quad G_{Q_0}(\cdot),\ G_{P_0}(\cdot) \text{ and } G_{Q_0}(\cdot) \text{ are independent.} \end{aligned}$$

Accordingly, an application of the multiplier central limit theorem (Theorem 2.9.2 and Corollary 2.9.4 in van der Vaart and Wellner (1996)) with $Var(\tilde{M}^1_{N,i})=1-\lambda$ and $Var(\tilde{M}^0_{N,j})=\frac{\lambda^2}{1-\lambda}$ yields

$$G^*_{m,\tilde{H}_n^{[N]}}(\cdot)\rightsquigarrow(1-\lambda)^{1/2}G_{P_0}(\cdot)+\lambda^{1/2}G_{Q_0}(\cdot),\ \mathbb{P}\text{-a.s. sequences of }\left\{\tilde{H}_N^{[N]}\right\}. \tag{B.24}$$

Since the covariance kernel of $(1-\lambda)^{1/2}G_{P_0}(\cdot)+\lambda^{1/2}G_{Q_0}(\cdot)$ coincides with that of H_0-brownian bridges, we conclude $G^*_{m,\tilde{H}_N^{[N]}}(\cdot)\rightsquigarrow G_{H_0}(\cdot)$, $\mathbb{P}$-a.s. sequences of $\left\{\tilde{H}_N^{[N]}\right\}$. By the same argument, it holds $G^{*\prime}_{n,\tilde{H}_N^{[N]}}(\cdot)\rightsquigarrow G'_{H_0}(\cdot)$, $\mathbb{P}$-a.s. sequences of $\left\{\tilde{H}_N^{[N]}\right\}$, where $G'_{H_0}(\cdot)$ are H_0-brownian bridges independent of $G_{H_0}(\cdot)$.

Regarding the bootstrap covariance kernel, we have convergence of $\sup_{f,g\in\mathcal{F}}\left|\sigma^2_{H_N^*}(f,g)-\sigma^2_{H_0}(f,g)\right|$ 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

$$\sup_{f,g\in\mathcal{F}}\left|\sigma^2_{H_N^*}(f,g)-\sigma^2_{H_0}(f,g)\right|\le\sup_{f,g\in\mathcal{F}}\left|\sigma^2_{H_N^*}(f,g)-\sigma^2_{\tilde{H}_N^{[N]}}(f,g)\right|+\sup_{f,g\in\mathcal{F}}\left|\sigma^2_{\tilde{H}_N^{[N]}}(f,g)-\sigma^2_{H_0}(f,g)\right|, \tag{B.25}$$

where the first term in the right hand side follows by the Glivenko-Cantelli theorem for the triangular arrays as stated in Lemma B.1, and the convergence of the second term follows from (B.14) and (B.15).

By putting together (B.24) and (B.25), and repeating the proof of the asymptotic uniform equicontinuity

as given in (B.11) above, we obtain

$$\begin{aligned} v_N^*(\cdot) &\rightsquigarrow \frac{(1-\lambda)^{1/2} G_{H_0}(\cdot) - \hat{\lambda}^{1/2} G'_{H_0}(\cdot)}{\sigma_{H_0}(\cdot,\cdot)} \\ &\sim \frac{G_{H_0}(\cdot)}{\sigma_{H_0}(\cdot,\cdot)}, \text{ as } N \to \infty, \mathbb{P}\text{-almost every sequence of } \left\{\tilde{H}_N^{[N]}\right\}. \end{aligned}$$

where the second line follows since the covariance kernel of the mean zero Gaussian processes $(1-\lambda)^{1/2} G_{H_0}(\cdot) - \hat{\lambda}^{1/2} G'_{H_0}(\cdot)$ is identical to that of $G_{H_0}(\cdot)$. The bootstrapped test statistics T_N^* is a continuous functional of $v_N^*(\cdot)$, so the continuous mapping theorem leads to

$$T_N^* \rightsquigarrow T_H = \max \left\{ \begin{array}{c} \sup_{f \in \mathcal{F}_1} \{-G_{H_0}(f)/\sigma_{H_0}(f,f)\} \\ \sup_{f \in \mathcal{F}_0} \{G_{H_0}(f)/\sigma_{H_0}(f,f)\} \end{array} \right\} \text{ as } N \to \infty, \mathbb{P}\text{-almost every sequence of } \left\{\tilde{H}_N^{[N]}\right\}.$$

Since T_H has continuous cdf, the bootstrap critical values $\tilde{c}_{N,1-\alpha}$ converges to $c_{1-\alpha}$, $\mathbb{P}$-a.s. ■

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., 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.

[3] Anderson, G., O. Linton, and Y. Whang (2012): "Nonparametric Estimation and Inference about the Overlap of Two Distributions," Journal of Econometrics, 171, 1-23.

[4] Andrews, D.W.K. and P. Jia Barwick, "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] Balke, A. and J. Pearl (1997): "Bounds on Treatment Effects from Studies with Imperfect Compliance," Journal of the American Statistical Association, 92, 1171-1176.

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

[13] 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.

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

[15] Campbell, D. T. (1969): "Reforms as Experiments," American Psychologist, 24 (4), 409-429.

[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] Chaisemartin, C. (2013): "Late with Defiers," unpublished manuscript, University of Warwick.

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

[19] 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.

[20] Hahn, J., Todd, P. E., and Van der Klaauw, W. (2001), "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," Econometrica 69 (1), 201-209.

[21] Heckman, J. J. and E. Vytlacil (1999): "Local Instrumental Variables and Latent Variables Models for Identifying and Bounding Treatment Effects," Proceedings of the National Academy of Sciences, 96, 4730-4734.

[22] Heckman, J. J. and E. Vytlacil (2001): "Local Instrumental Variables," in C. Hsiao, K. Morimune, and J. Powell editors, Nonlinear Statistical Model: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya, 1-46. Cambridge University Press, Cambridge UK.

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

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

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

[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] Kling, J. R., J. B. Liebman, and L. F. Katz (2007): "Experimental Analysis of Neighborhood Effects," Econometrica, 75, 83-119.

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

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

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

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

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

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

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