Testing Conditional Rank Similarity With and Without Covariates*

Ping Yu†
University of Hong Kong

First Version: July 2016
This Version: February 2017

Abstract

This paper studies the testing of conditional rank similarity, a key identification assumption for the instrumental variable quantile regression estimator (IV-QRE) of Chernozhukov and Hansen (2005). Different from the unconditional rank similarity test, no covariates are required in the conditional test. The basic idea is that conditional rank similarity implies no "difference (across treatment statuses) in difference (across marginal populations)", so "cross (marginal populations) mismatching" under the two treatment statuses would generate power. Tests are developed in the framework of Heckman and Vytlacil (2005) and three cases are considered: discrete instruments and no covariates, discrete instruments and general covariates, and general instruments and general covariates. In each case, two tests are proposed and extensions are discussed, and the exchangeable bootstrap is suggested to obtain critical values. Simulations corroborate the theoretical results and the tests are applied to the empirical study of Chernozhukov and Hansen (2006) to show the validity of IV-QRE.

KEYWORDS: conditional rank similarity, cross matching, distribution regression, IV-QRE, bootstrap
JEL-CLASSIFICATION: C12, C21

*I want to thank Yuanyuan Wan, Kaspar Wüthrich and seminar participants at 2017ESAM and 2017NASM for helpful comments.

†School of Economics and Finance, The University of Hong Kong, Pokfulam Road, Hong Kong; corresponding author email: pingyu@hku.hk.

1 Introduction

Quantile treatment effects (QTEs), as an alternative of average treatment effects (ATEs), have attracted much attention in recent developments of program evaluation; see Abbring and Heckman (2007) and Yu (2014a) for a summary of relevant literature. Roughly speaking, there are three frameworks to study the QTEs. The first framework assumes unconfoundedness; see, e.g., Firpo (2007) and Donald and Hsu (2014). The other two frameworks assume the existence of endogeneity but in different forms. Chernozhukov and Hansen (2005) (CH hereafter) assume only the selection effect, so can identify the "global" QTE by the so-called instrumental variable quantile regression estimator (IV-QRE). This is not possible when the essential heterogeneity also exists. With the presence of essential heterogeneity and a binary instrumental variable, Abadie et al. (2002) estimate the local quantile treatment effect (LQTE), which is the counterpart of the local average treatment effect (LATE) of Imbens and Angrist (1994) (IA hereafter). When continuous instruments exist, Carneiro and Lee (2009) and Yu (2014a) identify and estimate the marginal quantile treatment effect (MQTE), which is the counterpart of the marginal treatment effect (MTE) of Heckman and Vytlacil (2005) (HV hereafter). For a careful comparison of these two endogenous quantile frameworks, see Yu (2016).

This paper is about testing a key identification assumption of CH. As will be explained in Section 2, to identify the "global" QTE, CH impose a key assumption called (conditional) rank similarity. This assumption excludes the essential heterogeneity in the QTE context. CH's identification scheme is important because the "global" QTE has better external validity than the "local" QTEs such as LQTE and MQTE. However, to apply this scheme, we must guarantee rank similarity holds; otherwise, CH's IV-QRE will converge to some pseudo-true value as explained in Wüthrich (2015b) and Yu (2016). The goal of this paper is to test this key assumption in the framework of HV. As detailed in Section 2, HV's framework accommodates the essential heterogeneity and meanwhile imposes a monotonicity assumption which facilitates the testing procedure.

In Section 3, we overview our testing ideas; specifically, we motivate and develop the population version of our test statistics. We first express the null hypothesis in forms of objects that can be identified in an ideal scenario, and then consider three practical cases where only part of these objects can be identified. The basic idea of our tests is that conditional rank similarity implies no "difference (across treatment statuses) in difference (across marginal populations)", so "cross (marginal populations) mismatching" under the two treatment statuses would generate power. The three practical cases under consideration include (i) discrete instruments and no covariates; (ii) discrete instruments and general covariates; (iii) general instruments and general covariates. In each case, two "oracle" forms of test statistics are proposed and shown to be optimal in the sense that they exhaust the testable implication of the data distribution, and then the corresponding "practical" forms of test statistics are provided. In the following three sections, we consider the finite-sample analogs of the practical test statistics in each case. Specifically, we construct the test statistics, develop their asymptotic distributions under the null and alternative, obtain critical values by the exchangeable bootstrap and prove its validity. We consider also two extensions of our test statistics in each case. We use Case (i) to illustrate our testing idea and provide most details, and treat Case (ii) and (iii) as technical extensions.

Before proceeding, we briefly discuss an independent work by Kim and Park (2016) where the authors try to conduct the same test as in this paper. Their paper uses the framework of LQTE and ours uses the framewok of HV. As explained in Vytlacil (2002) and Yu (2015b), these two frameworks are essentially equivalent. Nevertheless, there are indeed a few key differences between these two papers. First, Kim and Park's test is based on the inverse probability weighting method of Abadie (2003) and is essentially the first test of our two main tests, while our tests are based on the distribution regression suggested by Yu (2014a) and take two main forms and two extensive forms. Second, they consider only Case (ii) especially for the case where the instrument takes only three values, while we consider also Case (i) to show that no covariates

are required to generate power and Case (iii) where instruments can take continuous values. Third, their test involves the counterfactual mapping of Vuong and Xu (2016), while we express such a mapping explicitly in our first test. Consequently, it is easy to derive local powers for our tests while it is hard for theirs (they actually did not derive the local power for their test). These local power expressions make it clear why a binary instrument cannot generate power no matter covariates exist or not. In summary, our paper and theirs are more complements than substitutes.

In Section 7, we discuss the connection and difference of this paper with other related literature. Especially, we show that although Heckman et al. (2010)'s tests of correlated random coefficient models cannot be applied to the problem in this paper, the tests in both papers are overidentification tests. We also show that the unconditional rank preservation (URP) tests in Yu (2015a) are also overidentification tests except using different overidentification information from that used in this paper. There are also some papers considering similar topics. For example, Dong and Shen (2015) test the "unconditional" rank similarity (URS) for *all* "unconditional" compliers (i.e., all compliers regardless of the covariate value) in the framework of HV with a binary instrument (see also Frandsen and Lefgren (2015) for a regression-based implementation); however, CH require only "conditional" rank similarity for *all* "conditional" individuals (i.e., all individuals given a specific covariate value).[1] Consequently, different from Dong and Shen (2015) or Frandsen and Lefgren (2015), we do not require covariates to generate power. Section 8 provides some simulation results, Section 9 applies our tests to the dataset from Angrist and Krueger (1991) which is also used as an illustration of IV-QRE in Chernozhukov and Hansen (2006), and Section 10 concludes. To save space, we relegate some discussions to two supplementary materials S.1 and S.2. S.1 contains the proofs that are not given in the main text, and S.2 contains points that we do not want to expand in the main text.

Some notations are collected here for future reference. The letter d is always used for indicating the two treatment statuses, so is not written out explicitly as "$d = 0, 1$" throughout the paper. Y_d is used for potential outcome and X is used for covariates. $\text{supp}(X)$ for a random variable X denotes the support of the distribution of X. The capital letters such as X denote random variables and the corresponding lower case letters such as x denote the potential values they may take. For a random variable Y and a random vector X, $F_{Y|X}$, $f_{Y|X}$, $Q_{Y|X}$ mean the conditional cumulative distribution function (cdf), the conditional probability density function (pdf) and the conditional quantile function of Y given X. F_d, f_d and Q_d mean the unconditional cdf, pdf and quantile function of Y_d. F_d^{-1} and Q_d are used exchangeably for notational convenience. $\Phi(\cdot)$ and $\phi(\cdot)$ are the cdf and pdf of the standard normal distribution. $\rightsquigarrow$ and $\overset{*}{\rightsquigarrow}$ signify the weak convergence and the weak convergence in probability, respectively.[2] n is the sample size. U denotes the common rank under the null - rank similarity. $C(\mathcal{I})$ is the space of continuous functions on a set $\mathcal{I}$. $\ell^\infty(\mathcal{F})$ as the space of real-valued bounded functions on the index set $\mathcal{F}$ equipped with the supremum norm $\|\cdot\|_{\ell^\infty(\mathcal{F})}$. VW is a short for van der Vaart and Wellner (1996), and CFM for Chernozhukov, Fernández-Val and Melly (2013).

[1] In a comment on Dong and Shen (2015), Edward Vytlacil mentioned a simple example to show that CRS is more interesting than URS: if $Y_1 = \alpha_1 + \beta_1 X + \varepsilon_1$, $Y_1 = \alpha_0 + \beta_0 X + \varepsilon_0$ with $(X, D) \perp (\varepsilon_1, \varepsilon_0)$, $P(X = 1) = p \in (0, 1)$, $P(X = 0) = 1 - p$, and ε_1 and ε_0 are continuously distributed, then CRS holds. Suppose $\varepsilon_1 = \sigma\varepsilon_0$; URS holds if and only if $\beta_1 = \sigma\beta_0$. A similar example can be found in Example 2 of Yu (2015a). Nevertheless, as shown in Dong and Shen (2015), Frandsen and Lefgren (2015) and Yu (2015a), URS is indeed relevant in practice.

[2] Recall from Section 2.9 of VW that $Z_n^* \overset{*}{\rightsquigarrow} Z$ in a separable normed space (D, d) if $\sup_{h \in BL_1(D)} |E^*[h(Z_n^*)] - E[Z]| \xrightarrow{p} 0$, where $BL_1 = \{h : D \to [0, 1] | \, |h(x) - h(y)| \leq \|x - y\| \text{ for all } x, y\}$, and E^* is the expectation with respect to the bootstrap measure conditional on the original data.

2 Frameworks and Hypotheses

We first repeat the framework of CH in our notations; see Chernozhukov and Hansen (2013) for most updated developments on their framework. By the Skorohod representation, the outcome equation is

$$Y_d = Q_d(U_d|X), d = 0, 1, \tag{1}$$

where $Q_d(U_d|X)$ is written as $q\left(d, X, U_d\right)$ in CH, $Q_d(\tau|x)$ is the τth conditional quantile of Y_d given $X = x$, and U_d conditional on $X = x$ is the rank variable of Y_d for the subpopulation $X = x$.[3] U_d represents some unobserved characteristic of Y_d, e.g., ability or proneness, and $U_d|X \sim U(0, 1)$, the uniform distribution on $(0, 1)$. The selection equation is

$$D = \delta(Z, X, V) \tag{2}$$

for some unknown function δ and random vector V, where Z represents the instruments (usually some policy variables) for the choice process. Both X and Z appearing as the arguments of δ does not lose generality since $\delta(Z, X, V)$ may not depend on all elements of X. Because Z does not show up in (1), the exclusion assumption is implicitly assumed. CH further impose the following assumptions on the model:

A1. Potential Outcomes: $Q_d(\tau|x)$ is strictly increasing in τ for $d = 0, 1$ and any $x \in \text{supp}(X)$.
A2. Independence: $(U_1, U_0) \perp Z|X$, where "$\perp$" denotes independence (c.f., Dawid (1979)) and variables to the right of "|" are the conditioning variables.
A3. Rank Similarity (RS): $F_{U_1|X,Z,V}\left(u|x, z, v\right) = F_{U_0|X,Z,V}\left(u|x, z, v\right)$ for $u \in (0, 1), (x, z, v) \in \text{supp}(X, Z, V)$.
A4. Observed Variables: Observed variables consist of $W \equiv (Y, D, X, Z)$, where $Y = DY_1 + (1 - D)Y_0 = Q_D(U_D|X)$.

A2 is the orthogonal condition; note that it does not require $V \perp Z|X$ (see IA for some important examples in which the instrument Z is assigned depending on D). CH also discuss a stronger version of RS, called rank invariance (RI), but for identification of $Q_d(\tau|x)$, RS is enough. Because this version of RS is conditional on X, we label it as conditional rank similarity (CRS). The target of CH is to estimate the QTE,

$$\Delta\left(\tau|x\right) = Q_1(\tau|x) - Q_0(\tau|x),$$

for $\tau \in (0, 1)$ and $x \in \text{supp}(X)$.

As argued in Yu (2016), CH's framework imposes stronger restrictions on the outcome equation and less restrictions on the selection equation compared with HV's framework. For concreteness, we repeat the framework of HV in the quantile context (see Yu (2014a)). The selection equation is

$$D = 1(\mu_D(X, Z) - V \geq 0),$$

where V is a scalar random error in the participation decision, and $1(\cdot)$ is the indicator function. By transforming $\mu_D(X, Z)$ and V by the conditional cdf $F_{V|X,Z}$, we can rewrite

$$D = 1(p(X, Z) - V \geq 0), \tag{3}$$

where we still use V to represent the transformed error to conform to CH's notations as much as possible,

[3] This U_d is different from the U_d in Yu (2015a) where U_d represents the unconditional rank of Y_d. For comparison, $U_d|_{X=x}$ in this paper is the U_d^x in Yu (2015a); in other words, U_d in this paper is the U_d^X in Yu (2015a), where $U_d|_{X=x}$ means U_d is restricted at $X = x$.

$V|X,Z \sim U(0,1)$, and $p(X,Z)$ is the propensity score.[4] Compared with (2), (3) takes the additive latent index form with only one random error while (2) imposes almost no restrictions on D. (3) implies the monotonicity assumption of IA, which is called the uniformity assumption by HV. The outcome equation is

$$Y_d = Q_d\left(U_d|X,V\right) \text{ with } U_d|\left(X,V\right) \sim U(0,1). \tag{4}$$

Compared with (1), Y_d in (4) has two random errors. In other words, the quantile of Y_d depends not only on X but on V; i.e., for each slice of individuals with $V=v$ who have the same propensity of participation, even if their X values are the same, their quantile functions of potential outcomes need not be the same. In term of Y which can be represented as $Q_D\left(U_D|X,V\right) \equiv q\left(D,X,V,U_D\right)$, its quantile depends not only on D but on the error term of D separately. Of course, by applying the Skorohod representation, we can always represent Y_d in the form of (1) although the definitions of U_d in (1) and (4) are different. In other words, although we can identify the quantile function of Y_d for each slice of individuals with $X=x$ and $V=v$, we may be only interested in the quantile function of Y_d for the slice of individuals with $X=x$ (i.e., combining the v slices). As to the relationship between Z and the error terms, HV assume the ignorability condition $(U_1,U_0,V) \perp Z|X$, while CH require only $(U_1,U_0) \perp Z|X$ and Z can be dependent of V.

To ease our testing procedure and study the fine structure of CRS assumption, we maintain the framework of HV throughout the paper. This framework is equivalent to IA's framework in some sense; see Yu (2015b) for a bare comparison of these two frameworks. Although this is a serious restriction, fortunately, we can test the validity of this framework; see Kitagawa (2015) and references therein. In HV's framework, due to $(U_1,U_0,V) \perp Z|X$, the CRS assumption can be restated as

$$H_0 : F_{U_1|X,V}(u|x,v) = F_{U_0|X,V}(u|x,v) \text{ for any } u \in (0,1), v \in (0,1) \text{ and } x \in \text{supp}\left(X\right).$$

This is our null hypothesis. Its opposite statement

$$H_1 : F_{U_1|X,V}(u|x,v) \neq F_{U_0|X,V}(u|x,v) \text{ for some } u \in (0,1), v \in (0,1) \text{ and } x \in \text{supp}\left(X\right)$$

is the alternative hypothesis. Although CH interpret H_0 as the "slippage" distributions of U_1 and U_0 from a common rank U are the same, it should be emphasized that theoretically, U_1 and U_0 can be independent. To see why, suppose (Y_1,Y_0) are fully unconfounded, which is a special case of CH's framework; then $F_{U_d|X,V}(u|x,v) = F_{U_d}(u) = u$ for both U_1 and U_0 even if $U_1 \perp U_0$. Essentially, the CRS assumption is a restriction on the two marginal distributions, while the rank invariance assumption, $U_1|\left(X,V\right) = U_0|\left(X,V\right)$, is a further restriction on the joint distribution.

We conclude this section by summarizing the framework and the corresponding assumptions we maintain in this paper. The outcome and selection equations are

$$\begin{aligned} &Y_d = Q_d(U_d|X), d=0,1, \\ &D = 1(p(X,Z) - V \geq 0), V|X,Z \sim U(0,1). \end{aligned} \tag{5}$$

The maintained assumptions are

Assumption M:

(M1) $\text{supp}(Y_d|X=x,V=v) = \mathcal{S}_{xd}$, the conditional density $f_{Y_d|X,V}(y_d|x,v)$ is continuous in (y_d,v) for any $x \in \text{supp}(X)$, $y_d \in \mathcal{S}_{xd}$, $d=0,1$, and $v \in (0,1)$, where $\mathcal{S}_{x0} = \left[\underline{y}_{x0}, \overline{y}_{x0}\right]$ and $\mathcal{S}_{x1} = \left[\underline{y}_{x1}, \overline{y}_{x1}\right]$ need not be compact and need not be the same.

[4] In (3), we use V rather then U_D as in HV because we have already used U_D for $DU_1 + (1-D)U_0$ as in CH.

(M2) $(U_1, U_0, V) \perp Z|X$.
(M3) $p(X, Z)$ is a nondegenerate random variable conditional on X.
(M4) $1 > P(D = 1|X) > 0$.
(M5) $X_1 = X_0$, where X_d denotes a potential value of X if D is set to d.

(M1) assumes Y_d to be continuously distributed. $\text{supp}(Y_d)$ can be bounded or unbounded. We also allow $\text{supp}(Y_d|X = x, V = v)$ to depend on d and x as in usual applications. We may also allow $\text{supp}(Y_d|X = x, V = v)$ to depend on v and $f_{Y_d|X,V}(y_d|x, v)$ to be discontinuous at some v values, but we impose such regularity conditions to avoid unnecessary complicity in notations. (M2) is the ignorability assumption. As shown in Yu (2016), in the framework of HV, (M3) is equivalent to the monotone likelihood ratio condition of CH and can be used to identify $Q_d(\cdot|\cdot)$ in (5). (M4) is the usual overlap assumption. (M5) is the "no feedback" condition which excludes the effect of D on X so conditioning on X does not mask the effects of D. Note that the framework and Assumption M are understood to be maintained only almost surely. For example, "for any $x \in$supp(X)" in (M1) and H_0, H_1 can be replaced by "for P_X almost sure x", but we will not impose such a qualifier everywhere unless necessary. Also, when covariates do not exist (i.e., $X = 1$), we neglect the subscript x in all notations, e.g., $\mathcal{S}_{xd}$ is written as $\mathcal{S}_d$; similarly, some notations are adjusted correspondingly, e.g., $W = (Y, D, Z)$.

3 An Overview of Testing Ideas

We first state a key testing implication of H_0. To simplify notations, we will depress the conditioning on X unless necessary.

Theorem 1 *In the framework (5), suppose Assumption M holds. Then H_0 is satisfied if and only if*

$$F_{Y_1|X,V}\left(Q_{Y_1|X}(\tau|x)|x, v\right) = F_{Y_0|X,V}\left(Q_{Y_0|X}(\tau|x)|x, v\right) \tag{6}$$

for any $\tau \in (0, 1), v \in (0, 1)$ and $x \in$supp(X).

Proof. Note that

$$F_{Y_1|V}(y_1|v) = P\left(Y_1 \leq y_1|V = v\right) = P\left(U_1 \leq F_1\left(y_1\right)|V = v\right) = F_{U_1|V}(F_1\left(y_1\right)|v). \tag{7}$$

Similarly,

$$F_{Y_0|V}(y_0|v) = F_{U_0|V}(F_0\left(y_0\right)|v).$$

As a result,

$$\begin{aligned} F_{Y_1|V}\left(Q_1(\tau)|v\right) &= F_{U_1|V}(F_1\left(Q_1(\tau)\right)|v) = F_{U_1|V}(\tau|v) \\ &= F_{U_0|V}(\tau|v) = F_{U_0|V}(F_0\left(Q_0(\tau)\right)|v) = F_{Y_0|V}\left(Q_0(\tau)|v\right) \end{aligned}$$

where the middle equality is from H_0. ■

From the proof, $F_{Y_1|V}\left(Q_1(\tau)|v\right) - F_{Y_0|V}\left(Q_0(\tau)|v\right) = F_{U_1|V}(\tau|v) - F_{U_0|V}(\tau|v)$. In other words, we can plot $F_{Y_1|V}\left(Q_1(\tau)|v\right)$ and $F_{Y_0|V}\left(Q_0(\tau)|v\right)$ as a function of τ for each v; the difference between them is the cdf difference of U_1 and U_0 for the subpopulation $V = v$. The intuition for why compound functions $F_{Y_d|V}\left(Q_d(\cdot)|\cdot\right)$ in (6) are used is as follows: from the Skorohod representation, Y_d contains two parts of information - the value information $Q_d\left(\cdot\right)$ and the rank information U_d; as rank similarity is related only to the rank information, we use a compound function to offset the value information such that only the rank

information remains. It is also interesting to observe that H_0 is weaker than unconfoundedness since H_0 states $F_{U_1|V}(u|v) = F_{U_0|V}(u|v)$ for any u and v while unconfoundedess implies further that both are equal to u. In terms of $F_{Y_1|V}\left(Q_1(\tau)|v\right)$ and $F_{Y_0|V}\left(Q_0(\tau)|v\right)$, unconfoundedness implies they are not only equal but equal to τ.

The following corollary states more implications of (6).

Corollary 1 *Under the assumptions of Theorem 1,* $F_{Y_1|X,V}\left(Q_{Y_1|X}(\tau|x)|x,v\right) = F_{Y_0|X,V}\left(Q_{Y_0|X}(\tau|x)|x,v\right)$ *for any* $\tau \in (0,1), v \in (0,1)$ *and* $x \in supp\,(X)$ *is equivalent to that for any* $x \in supp\,(X)$*,*

$$\begin{aligned} Q_{Y_1|X,V}(F_{Y_0|X,V}(y_0|x,v)|x,v) &= Q_{Y_1|X}\left(F_{Y_0|X}(y_0|x)|x\right) \text{ for } y_0 \in \mathcal{S}_{0x}, \\ Q_{Y_0|X,V}(F_{Y_1|X,V}(y_1|x,v)|x,v) &= Q_{Y_0|X}\left(F_{Y_1|X}(y_1|x)|x\right) \text{ for } y_1 \in \mathcal{S}_{1x}, \end{aligned}$$

or

$$F_{Y_1|X,V}(Q_{Y_1|X,V}\left(\tau|x,v\right)|x,v') = F_{Y_0|X,V}(Q_{Y_0|X,V}\left(\tau|x,v\right)|x,v') \text{ for any } \tau \in (0,1), \tag{8}$$

where $v, v' \in (0,1)$*,* $v' \neq v$*,* v' *is arbitrary and* v *can be fixed or arbitrary.*

The first group of results are labelled as *counterfactual-quantiles matching* in Yu (2016), and *counterfactual mapping* in Vuong and Xu (2016). The second group of results show that the matching can be applied not only between the marginal population $V = v$ and the population but across different marginal populations.[5] This group of results also imply $F_1\left(Q_{Y_1|V}(\tau|v)\right) - F_0\left(Q_{Y_0|V}(\tau|v)\right) = \int F_{Y_1|V}\left(Q_{Y_1|V}(\tau|v)|v'\right)dv' - \int F_{Y_0|V}\left(Q_{Y_0|V}(\tau|v)|v'\right)dv' = 0$. Without H_0, although $F_{Y_1|V}\left(Q_{Y_1|V}(\tau|v)|v\right) - F_{Y_0|V}\left(Q_{Y_0|V}(\tau|v)|v\right) = \tau - \tau = 0$, $F_1\left(Q_{Y_1|V}(\tau|v)\right) - F_0\left(Q_{Y_0|V}(\tau|v)\right)$ need not be zero. It is also interesting to note that cross matching with a fixed v is equivalent to cross matching with an arbitrary v; the fixed v is like a "hub" for other v's to get connected through it.

For comparison, consider the unconfounded case again, where $F_{Y_d|V}\left(y_d|v\right) = F_{Y_d|V}\left(y_d|v'\right) = F_d\left(y_d\right)$, so $F_d\left(Q_{Y_d|V}(\tau|v)|v'\right) = F_d\left(Q_d(\tau)\right) = \tau$. In the current case, $F_{Y_d|V}\left(y_d|v\right)$ need not equal $F_d\left(y_d\right)$ so that $F_d\left(Q_{Y_d|V}(\tau|v)|v'\right)$ need not equal τ; however, it still holds that $F_1\left(Q_{Y_1|V}(\tau|v)|v'\right) = F_0\left(Q_{Y_0|V}(\tau|v)|v'\right)$. In other words, unconfoundedness implies no "difference" in $F_d\left(y_d|v\right)$ across v (for any d) while H_0 implies no "difference in difference" in $F_d\left(y_d|v\right)$, where the first "difference" is across d and the second "difference" is across v. This is somewhat like the usual difference-in-difference (DID) method or the changes-in-changes (CIC) method of Athey and Imbens (2006), where the time T plays the role of V. In the former, although $E\left[U_d|T=0\right] \neq E\left[U_d|T=1\right]$, $E\left[U_1|T=1\right] - E\left[U_1|T=0\right] = E\left[U_0|T=1\right] - E\left[U_0|T=0\right]$, where $U_d = Y_d - E[Y_d]$. In the latter, although $F_d\left(y_d|T=0\right) \neq F_d\left(y_d|T=1\right)$, $F_1\left(Q_1\left(\tau|T=0\right)|T=1\right) = F_0\left(Q_0\left(\tau|T=0\right)|T=1\right)$. The de-meaning in the former and the de-scaling in the latter are corresponding operations to nullify the value information in Y_d in the ATE and QTE evaluation, respectively.

In practice, the variation in Z may not be large enough to identify all V marginals. We will consider three practical cases in this paper. Case (i): discrete Z and no X. We use this case to illustrate that unlike the unconditional rank similarity test in Dong and Shen (2015) and Frandsen and Lefgren (2015), no covariates are required to test conditional rank similarity. Case (ii): discrete Z and general X. This is the most practical case since most data sets in applications satisfy this condition. In some applications, both X and Z may include continuous variables, so we also discuss Case (iii): general Z and general X. We will first overview here how our test statistics are constructed for these three cases. In the following three sections, we will provide construction details and asymptotics for each case.

[5] Note that we are conditioning on $X = x$, so the marginal population is actually $V = v, X = x$, and the population is $X = x$.

3.1 Case (i)

Suppose Z can take K values $\{z_1, \cdots, z_K\}$ such that $0 = p_0 < p_1 < p_2 < \cdots < p_K < p_{K+1} = 1$ with $p_k = p(z_k)$.[6] Then the always-takers $\mathcal{A} \equiv \{V \leq p_1\}$, the kth compliers $\mathcal{C}_k \equiv \{p_k < V \leq p_{k+1}\}$, $k = 1, \cdots, K-1$, are individuals who are induced to switch from $D = 0$ to $D = 1$ as Z changes from z_k to z_{k+1}, and the never-takers $\mathcal{N} \equiv \{V > p_K\}$. The proportion of each group of individuals in the population is $p_1, p_{k+1} - p_k$, $k = 1, \cdots, K-1$, and $1 - p_K$, respectively.

We will propose two tests. The first test is based on the following corollary which is the version of (6) in Case (i).

Corollary 2 *In the framework (5), suppose Assumption M holds with $X = 1$ and Z being discrete. If H_0 is satisfied, then*

$$F_{Y_1|\mathcal{C}_k}(Q_1(\tau)) = F_{Y_0|\mathcal{C}_k}(Q_0(\tau)) \tag{9}$$

for any $\tau \in (0,1)$ and $k = 1, \cdots, K-1$.

Proof. Note that

$$F_{Y_d|\mathcal{C}_k}(y_d) = \frac{1}{p_{k+1} - p_k} \int_{p_k}^{p_{k+1}} F_{Y_d|V}(y_d|v)\, dv = \frac{1}{p_{k+1} - p_k} \int_{p_k}^{p_{k+1}} F_{U_d|V}(F_d(y_d)|v) dv.$$

where the last equality is from (7). If $F_{U_1|V}(u|v) = F_{U_0|V}(u|v) = F_{U|V}(u|v)$, $u, v \in (0,1)$, then

$$F_{Y_1|\mathcal{C}_k}(y_1) = F_{U|\mathcal{C}_k}(F_1(y_1)) \text{ and } F_{Y_0|\mathcal{C}_k}(y_0) = F_{U|\mathcal{C}_k}(F_0(y_0)),$$

where

$$F_{U|\mathcal{C}_k}(u) \equiv \frac{1}{p_{k+1} - p_k} \int_{p_k}^{p_{k+1}} F_{U|V}(u|v) dv$$

As a result,

$$F_{Y_1|\mathcal{C}_k}(Q_1(\tau)) = F_{U|\mathcal{C}_k}(F_1(Q_1(\tau))) = F_{U|\mathcal{C}_k}(\tau) = F_{U|\mathcal{C}_k}(F_0(Q_0(\tau))) = F_{Y_0|\mathcal{C}_k}(Q_0(\tau)).$$

■

Different from Theorem 1, (9) does not imply H_0 because we cannot observe all v's. In other words, H_0 is refutable but nonverifiable - if the null is rejected, then we are sure that H_0 is wrong; while if the null is not rejected, H_0 need not be correct. See Breusch (1986) for a general discussion and Kitagawa (2015) for a recent example of this kind of hypothesis. On the other hand, since $\mathcal{C}_k, k = 1, \cdots, K-1$, is the only identifiable set of V marginals, we expect (9) exhausts the testable implication of the data distribution. The following proposition rigorously states this result.

Proposition 1 *Suppose the assumptions of Corollary 2 hold. If $F_{Y_1|\mathcal{C}_k}(Q_1(\tau))$ and $F_{Y_0|\mathcal{C}_k}(Q_0(\tau))$ satisfy (9), then there exists a joint distribution of (Y_1, Y_0, V, Z) that satisfies H_0, generates the joint data distribution of (Y, D, Z), and induces $F_{Y_d|\mathcal{C}_k}(Q_d(\tau))$ for $\tau \in (0,1)$ and $k = 1, \cdots, K-1$.*

This proposition implies that a test based on

$$T_{1o} = \sum_{k=1}^{K-1} w_k \int_0^1 \left[F_{Y_1|\mathcal{C}_k}(Q_1(\tau)) - F_{Y_0|\mathcal{C}_k}(Q_0(\tau))\right]^2 d\mu_k(\tau)$$

[6] If $p_1 = 0$ or $p_K = 1$, some of the following analysis can be simplified.

is optimal in the sense that if $T_{1o} = 0$ then any other feature of the data distribution cannot contribute to invalidate CRS further, where $w_k > 0$, $\sum_{k=1}^{K-1} w_k = 1$, $\mu_k(\tau) > 0$ for almost every $\tau \in (0,1)$, $\int_0^1 d\mu_k(\tau) = 1$ and the subscript o is for "oracle". Note that the choice of w_k and $\mu_k(\cdot)$ does not affect the optimality of T_{1o}, but may improve its power in case that we have prior information on the group of compliers or quantile indices on which $F_{U_1|\mathcal{C}_k}(\tau) \neq F_{U_0|\mathcal{C}_k}(\tau)$. Uniform priors $w_k = 1/k$ and $\mu_k(\tau) = 1$ indicate that no prior information is imposed. In practice, w_k and $\mu_k(\tau)$ can be constructed based on some pilot tests.

From Corollary 2 and Proposition 1, our first test statistic in Case (i) can be based on a truncated version of T_{1o},

$$T_1 = \sum_{k=1}^{K-1} w_k \int_{\mathcal{T}} \left[F_{Y_1|\mathcal{C}_k}\left(Q_1(\tau)\right) - F_{Y_0|\mathcal{C}_k}\left(Q_0(\tau)\right)\right]^2 d\mu_k(\tau),$$

where $\mathcal{T}$ is a compact subset of $[\epsilon, 1-\epsilon]$ for some $\epsilon > 0$. We truncate the quantile index for three reasons. First, when the supports of Y_1 and Y_0 are not bounded, e.g., Y is the weekly wage rate, we can avoid the technical difficulties in estimating extremal quantiles (see, e.g., Chernozhukov (2005) and Chernozhukov and Fernández-Val (2011)). Second, it is commonly believed that at extremal quantiles, the CRS assumption is easier to hold. For example, the extreme rich (poor) complier tends to be extremely rich (poor) after participating a social program. Third, if Y is censored at bottom or top as in, e.g., weekly wage rate, we can avoid the contribution from point masses of censored quantiles to T_1.

Our second test is based on the following corollary which is the version of (8) in Case (i).

Corollary 3 *Suppose the assumptions of Corollary 2 hold. Under H_0,*

$$F_{Y_1|\mathcal{C}_k}\left(Q_{Y_1|\mathcal{C}_1}(\tau)\right) = F_{Y_0|\mathcal{C}_k}\left(Q_{Y_0|\mathcal{C}_1}(\tau)\right) \tag{10}$$

for any $\tau \in (0,1), k \in \{2, \cdots, K-1\}$.

Proof. By the similar logic as in the proof of Corollary 2, $F_{Y_1|\mathcal{C}_k}\left(Q_{Y_1|\mathcal{C}_1}(\tau)\right) = F_{U|\mathcal{C}_k}\left(F^{-1}_{U|\mathcal{C}_1}(\tau)\right) = F_{Y_0|\mathcal{C}_k}\left(Q_{Y_0|\mathcal{C}_1}(\tau)\right)$. ■

From the proof above, $F_{U_d|\mathcal{C}_k}$ need not equal $F_{U_d|\mathcal{C}_1}$ as in the unconfounded case; however, under the null, the difference between $F_{U_1|\mathcal{C}_k}$ and $F_{U_1|\mathcal{C}_1}$ ($F_{U_1|\mathcal{C}_k}\left(F^{-1}_{U_1|\mathcal{C}_1}(\tau)\right)$) and the difference between $F_{U_0|\mathcal{C}_k}$ and $F_{U_0|\mathcal{C}_1}$ ($F_{U_0|\mathcal{C}_k}\left(F^{-1}_{U_0|\mathcal{C}_1}(\tau)\right)$) are similar. In other words, the second test is based on the "similarity between differences". Like (8) in Corollary 1, (10) is equivalent to $F_{Y_1|\mathcal{C}_k}\left(Q_{Y_1|\mathcal{C}_l}(\tau)\right) = F_{Y_0|\mathcal{C}_k}\left(Q_{Y_0|\mathcal{C}_l}(\tau)\right)$ for $k, l \in \{1, \cdots, K-1\}$ and $k \neq l$. For future reference, we label this result as *cross compliers matching* (or simply *cross matching*) and its opposite that $F_{Y_1|\mathcal{C}_k}\left(F^{-1}_{Y_1|\mathcal{C}_1}(\tau)\right) \neq F_{Y_0|\mathcal{C}_k}\left(F^{-1}_{Y_0|\mathcal{C}_1}(\tau)\right)$ for some $k \in \{2, \cdots, K-1\}$ and $\tau \in (0,1)$ as *cross compliers mismatching* (or simply *cross mismatching*).

Like (9), (10) does not imply H_0. It seems also that it does not imply (9) because $Q_d(\cdot)$ cannot be recovered solely from (10). Actually, (9) and (10) include the same testable information as shown in the following proposition.

Proposition 2 *Suppose the assumptions of Corollary 2 hold. If $F_{Y_1|\mathcal{C}_k}\left(Q_{Y_1|\mathcal{C}_1}(\tau)\right)$ and $F_{Y_0|\mathcal{C}_k}\left(Q_{Y_0|\mathcal{C}_1}(\tau)\right)$ satisfy (10), then there exists a joint distribution of (Y_1, Y_0, V, Z) that satisfies H_0, generates the joint data distribution of (Y, D, Z), and induces $F_{Y_1|\mathcal{C}_k}\left(Q_{Y_1|\mathcal{C}_1}(\tau)\right)$ and $F_{Y_0|\mathcal{C}_k}\left(Q_{Y_0|\mathcal{C}_1}(\tau)\right)$ for $\tau \in (0,1)$ and $k = 2, \cdots, K-1$.*

This seemingly surprising result is due to the fact that $Q_d(\cdot)$ in (9) cannot be identified from the data distribution. We in Section 4 identify it by imposing H_0, which loses one more degree of freedom. That is why (9) includes $K-1$ constraints while (10) includes only $K-2$ constraints. A straightforward corollary

of the proof of Proposition 2 is that if $K = 2$, H_0 does not impose any restriction on the data generating process, i.e., no tests can have power.

As in the cae of T_{1o}, a test based on

$$T_{2o} = \sum_{k=2}^{K-1} w_k \int_0^1 \left[F_{Y_1|\mathcal{C}_k}\left(Q_{Y_1|\mathcal{C}_1}(\tau)\right) - F_{Y_0|\mathcal{C}_k}\left(Q_{Y_0|\mathcal{C}_1}(\tau)\right)\right]^2 d\mu_k(\tau)$$

is optimal, where w_k and $\mu_k(\tau)$ are similarly defined as in T_{1o}. Similar to T_1, our second test statistic is based on a truncated version of T_{2o},

$$T_2 = \sum_{k=2}^{K-1} w_k \int_{\mathcal{T}} \left[F_{Y_1|\mathcal{C}_k}\left(Q_{Y_1|\mathcal{C}_1}(\tau)\right) - F_{Y_0|\mathcal{C}_k}\left(Q_{Y_0|\mathcal{C}_1}(\tau)\right)\right]^2 d\mu_k(\tau).$$

For $K-1$ groups of compliers, there are totally $K-2$ cross matchings. Here, we select $\mathcal{C}_1$ as the hub complier; in practice, we can select $\mathcal{C}_l$ with the largest proportion $(p_{l+1} - p_l)$ as the hub complier to improve finite-sample performance.

That T_{1o} and T_{2o} explore the same testable implication of the data distribution does not mean they would have the same power; see Section 4.2 for a detailed analysis. This is just like that both tests have power in any feasible direction of violating H_0 but their power functions may be very different. Also, as suggested after Theorem 1, we can plot $F_{Y_1|\mathcal{C}_k}(Q_1(\tau))$ versus $F_{Y_0|\mathcal{C}_k}(Q_0(\tau))$ in (9) or $F_{Y_1|\mathcal{C}_k}\left(Q_{Y_1|\mathcal{C}_1}(\tau)\right)$ versus $F_{Y_0|\mathcal{C}_k}\left(Q_{Y_0|\mathcal{C}_1}(\tau)\right)$ in (10) as a function of τ to check the violation of H_0.

3.2 Case (ii)

Suppose $\text{supp}(X) = \mathcal{X}$ and for each $X = x$, Z can take K values $\{z_1, \cdots, z_K\}$, $K \geq 2$. For each x, $p(x, Z)$ takes K distinct values $\{p_{x1}, \cdots, p_{xK}\}$, with $p_{xk} = p(x, z_k)$ and $0 = p_{x0} < p_{x1} < p_{x2} < \cdots < p_{xK} < p_{x,K+1} = 1$. We can extend to the case where the support of (X, Z) is not a Cartesian product of two sets, but such a generalization will not add any further insights except complicate notations.[7] Also, we can allow $p(x, z)$ not to be increasing in z, but by rearranging the Z values within each x value we can always make $p(x, z)$ strictly increase in z. We assume $p(x, z_k)$ is distinct for each z_k; if two z_k values generate the same propensity score, they are combined and relabelled as one value. Under these notations, among the individuals with $X = x$, the always-takers $\mathcal{A}_x \equiv \{V \leq p_{x1}\}$, the compliers $\mathcal{C}_{xk} \equiv \{p_{xk} < V \leq p_{x,k+1}\}$, $k = 1, \cdots, K-1$, and never-takers $\mathcal{N}_x \equiv \{V > p_{xK}\}$. As in Case (i), we can show the test based on either

$$T_{1o}^X = \sum_{k=1}^{K-1} \int_{\mathcal{X}} w_k(x) \int_0^1 \left[F_{Y_1|X}^k\left(Q_{Y_1|X}(\tau|x)|x\right) - F_{Y_0|X}^k\left(Q_{Y_0|X}(\tau|x)|x\right)\right]^2 d\mu_k(\tau|x)\, dx$$

or

$$T_{2o}^X = \sum_{k=2}^{K-1} \int_{\mathcal{X}} w_k(x) \int_0^1 \left[F_{Y_1|X}^k\left(Q_{Y_1|X}^1(\tau|x)|x\right) - F_{Y_0|X}^k\left(Q_{Y_0|X}^1(\tau|x)|x\right)\right]^2 d\mu_k(\tau|x)\, dx$$

[7] Think about the binary Z case. If we interpret Z as the theoretical assignment and D as the realized treatment status, then the independence of $\text{supp}(Z)$ on X means that the assignment Z is genuinely randomized for each value of X. This assumption corresponds to the assumption $0 < E[Z|X] < 1$ in IA.

is optimal, where $F^k_{Y_d|X}(\cdot|x)$ is the cdf of Y_d for $\mathcal{C}_{xk}$ given $X = x$.[8] The weight $w_k(x)$, satisfying $w_k(x) > 0$ for P_X almost sure x and $\int_{\mathcal{X}} w_k(x)dx = 1$, can depend on both x and $\mathcal{C}_{xk}$. Similarly, $\mu_k(\tau|x)$, satisfying $\mu_k(\tau|x) > 0$ for almost every $\tau \in (0,1)$ and $\int_0^1 d\mu_k(\tau|x) = 1$ for P_X almost sure x, can depend on both x and $\mathcal{C}_{xk}$. As in the case of T_1 and T_2, our two tests in Case (ii) are based on

$$T_1^X = \sum_{k=1}^{K-1} \int_{\mathcal{X}} w_k(x) \int_{\mathcal{T}} \left[F^k_{Y_1|X}\left(Q_{Y_1|X}(\tau|x)|x\right) - F^k_{Y_0|X}\left(Q_{Y_0|X}(\tau|x)|x\right)\right]^2 d\mu_k(\tau|x)\, dx$$

and

$$T_2^X = \sum_{k=2}^{K-1} \int_{\mathcal{X}} w_k(x) \int_{\mathcal{T}} \left[F^k_{Y_1|X}\left(Q^1_{Y_1|X}(\tau|x)|x\right) - F^k_{Y_0|X}\left(Q^1_{Y_0|X}(\tau|x)|x\right)\right]^2 d\mu_k(\tau|x)\, dx,$$

where $Q^1_{Y_d|X}(\cdot|x)$ is the quantile function of Y_d for $\mathcal{C}_{x1}$ given $X = x$. As to the choice of $w_k(x)$, if X is discrete and can take J values $\{x_1, \cdots, x_J\}$, then $w_k(x) = w_{jk}$; if $P(X = x_j)$ for some x_j is small, we can truncate such x_j and pick only x_j's with enough data points; when the proportion of some $\mathcal{C}_{x_jk}$ within x_j is small, we can further truncate it by setting the corresponding $w_{jk} = 0$. Typically, we can set $w_k(x)dx = dF_X(x)$ independent of k. Similarly, we can set $\mu_k(\tau|x) = 1$ independent of both x and k.

As mentioned in the introduction, Kim and Park's test is related to our T_1^X, where $\left[F^k_{Y_1|X}\left(Q_{Y_1|X}(\tau|x)|x\right) - F^k_{Y_0|X}\left(Q_{Y_0|X}(\tau\right.\right.$ is replaced by $\inf_{y_0 \in \mathcal{Y}_0(x)} \left|F^k_{Y_1|X}(y_1|x) - F^k_{Y_0|X}(y_0|x)\right|$ with $\mathcal{Y}_0(x) = \mathcal{S}_{x0}$ under our assumption M1. Under H_0, the infimum 0 is actually achieved at $y_0 = Q_{Y_0|X}(F_{Y_1|X}(y_1|x)|x)$. In other words, we express their infimum explicitly in our T_1^X.

3.3 Case (iii)

Suppose $\text{supp}(X) = \mathcal{X}$, $\text{supp}(Z|X = x) = \mathcal{Z}$ independent of x and $\text{supp}(p(X,Z)|X = x) = \mathcal{P}_x$ dependent on x. We impose the following assumption on $\mathcal{P}_x$.

Assumption P: $\mathcal{P}_x = \left[\underline{p}_x, \overline{p}_x\right] \subset [0,1]$ for each $x \in \mathcal{X}$. In other words, the conditional density function of $p(X,Z)$ given $X = x$, say $f_{P(X,Z)|X}(p|x)$, is positive for $p \in \left(\underline{p}_x, \overline{p}_x\right)$.

This assumption tries to reflect the reality of the support of $p(X,Z)$. Usually, the support of $p(X,Z)$ is different for different x values and is a subset of $[0,1]$ staying in the middle of $[0,1]$. We can relax this assumption by allowing the support of $p(X,Z)$ to be segments of intervals without difficulty, but we find Assumption P will provide clean results without losing the essence of our problem. For each x, we can identify counterfactual distributions of the marginal population $V = v \in \mathcal{P}_x$.

Following the similar proof idea of Proposition 1, we can show that

$$\int_{\mathcal{X}} \int_{\mathcal{P}_x} w(x,v) \int_0^1 \left[F_{Y_1|X,V}\left(Q_{Y_1|X}(\tau|x)|x,v\right) - F^k_{Y_0|X,V}\left(Q_{Y_0|X}(\tau|x)|x,v\right)\right]^2 d\mu(\tau|x,v)\, dvdx$$

is optimal, where $w(x,v)$ is a weight function on $\mathcal{X}\mathcal{P}_x \equiv \{(x,p_x)|x \in \mathcal{X}, p_x \in \mathcal{P}_x\}$, and $\mu(\tau|x,v)$ is a weight function on $(0,1)$ which may depend on x and v. This suggests a test statistic such as

$$\int_{\mathcal{X}} \int_{\mathcal{P}_x} w(x,v) \int_{\mathcal{T}} \left[F_{Y_1|X,V}\left(Q_{Y_1|X}(\tau|x)|x,v\right) - F^k_{Y_0|X,V}\left(Q_{Y_0|X}(\tau|x)|x,v\right)\right]^2 d\mu(\tau|x,v)\, dvdx, \tag{11}$$

[8] In Kim and Park (2016), $\left[F^k_{Y_1|X}\left(Q_{Y_1|X}(\tau|x)|x\right) - F^k_{Y_0|X}\left(Q_{Y_0|X}(\tau|x)|x\right)\right]^2$ is replaced by $\inf_{y_0 \in \mathcal{Y}_0(x)} \left|F^k_{Y_1|X}(y_1|x) - F^k_{Y_0|X}(y_0|x)\right|$, where $\mathcal{Y}_0(x) = \mathcal{S}_{x0}$ under our assumption M1. Under H_0, the infimum 0 is actually achieved at $y_0 = Q_{Y_0|X}(F_{Y_1|X}(y_1|x)|x)$. In other words, we express the infimum explicitly in our T_{1o}^X.

which can be treated as an extension of T_1^X. However, this form of test is not feasible due to a techinical difficulty in the inference of $Q_{Y_d|X}(\tau|x)$ as explained in Section 6. As an alternative, we suggest two variants of T_2^X. Specifically,

$$T_3^X = \int_{\mathcal{X}} \int_{\mathcal{P}_x: v \neq v_o} w(x,v) \int_{\mathcal{T}} \left[F_{Y_1|X,V}\left(Q_{Y_1|X,V}(\tau|x,v_o)|x,v\right) - F_{Y_0|X,V}\left(Q_{Y_0|X,V}(\tau|x,v_o)|x,v\right)\right]^2 d\mu\left(\tau|x,v\right) dvdx,$$

where v_o is the hub v. Following Proposition 2, we can show if $\mathcal{T}$ is replaced by $(0,1)$, then T_3^X is optimal. We can select v_o as the median of $F_{p(X,Z)|X}(\cdot|x)$ so it may depend on x. When v is close to v_o, $F_{Y_1|X,V}\left(Q_{Y_1|X,V}(\tau|x,v_o)|x,v\right)$ is close to $F_{Y_0|X,V}\left(Q_{Y_0|X,V}(\tau|x,v_o)|x,v\right)$ since both are close to τ. As a result, such v's will include more noise than information to power, and it is better to kick out a neighborhood of v_o in the integration about v. However, it is hard to determine how large a neighborhood should be kicked out or a good kick-out may be data-dependent. To avoid such a difficulty, we suggest another variant of T_2^X:

$$\begin{aligned} T_4^X \quad = \quad & \max\left\{ \int_{\mathcal{X}} \int_{v_o}^{\overline{p}_x} w(x,v) \int_{\mathcal{T}} \left[F_{Y_1|X,V}\left(Q_{Y_1|X,V}(\tau|x,v_1)|x,v\right) - F_{Y_0|X,V}\left(Q_{Y_0|X,V}(\tau|x,v_1)|x,v\right)\right]^2 d\mu\left(\tau|x,v\right) dvdx \right. , \\ & \left. \int_{\mathcal{X}} \int_{\underline{p}_x}^{v_o} w(x,v) \int_{\mathcal{T}} \left[F_{Y_1|X,V}\left(Q_{Y_1|X,V}(\tau|x,v_2)|x,v\right) - F_{Y_0|X,V}\left(Q_{Y_0|X,V}(\tau|x,v_2)|x,v\right)\right]^2 d\mu\left(\tau|x,v\right) dvdx \right\}, \end{aligned}$$

where $v_1 \in (\underline{p}_x, v_o)$ and $v_2 \in (v_o, \overline{p}_x)$. We can select v_1, v_0, v_2 as the first quartile, median and third quartile of $F_{p(X,Z)|X}(\cdot|x)$. Note that in T_4^X, we have two hub v's - v_1 and v_2. The range of integration relative to the first hub is $[v_o, \overline{p}_x]$, which is far from v_1; similarly for the integration relative to v_2. Also, because v_o is included in both ranges of integration, as argued after Corollary 1, all cross-v comparisons are covered. As a result, if $\mathcal{T}$ is replaced by $(0,1)$ in T_4^X, then it is optimal.

We close this section by some additional comments. All the test statistics above take the form of Cramer-von Mises (CM) statistics; an alternative formulation is in the form of Kolmogorov-Smirnov (KS) statistics. The reason for departing from KS statistics is that our simulation experiments suggested that CM-type statistics have somewhat better power properties than those based on the KS-type statistics. In the following sections, we will also discuss some extensions of $T_1, T_2, T_1^X, T_2^X, T_3^X$ and T_4^X; we did not discuss these extensions here to emphasize the basic ideas of our tests.

4 Testing With Discrete Instruments and No Covariates

To construct the finite-sample analog of T_1, we need to estimate $Q_d(\cdot)$ and $F_{Y_d|\mathcal{C}_k}(\cdot)$. The estimation of $Q_d(\cdot)$ is based on the IV-QRE of CH, which is consistent under H_0. Specifically, the moment conditions used by the IV-QRE are

$$E\left[\begin{pmatrix} 1 \\ p(Z) \end{pmatrix} \left(\tau - 1\left(Y \leq Q_0^*(\tau) + D \cdot \Delta^*(\tau)\right)\right) \right] = 0, \tag{12}$$

where $\Delta^*(\tau) = Q_1^*(\tau) - Q_0^*(\tau)$, and we use the superscript $*$ to indicate that $Q_d^*(\tau)$ and $\Delta^*(\tau)$ identified by the moment conditions need not be the true values when H_0 does not hold. As shown in Yu (2016) (see also Wütherich (2015) for the case with Z being the instrument), the moment conditions imply

$$\begin{aligned} F_1^*(y_1) \quad = \quad & p_1 F_{Y_1|\mathcal{A}}(y_1) + \sum_{k=1}^{K-1} (p_{k+1} - p_k) \left\{ F_{Y_1|\mathcal{C}_k}(y_1)\left[1 - F_{p(Z)}(p_k)\right] + F_{Y_0|\mathcal{C}_k}(\widetilde{F}_0^{-1}\widetilde{F}_1(y_1)) F_{p(Z)}(p_k) \right\} \\ & + (1 - p_K) F_{Y_0|\mathcal{N}}(\widetilde{F}_0^{-1}\widetilde{F}_1(y_1)) \end{aligned}$$

and

$$F_0^*(y_0) = p_1 F_{Y_1|\mathcal{A}}(\widetilde{F}_1^{-1}\widetilde{F}_0(y_0)) + \sum_{k=1}^{K-1}(p_{k+1}-p_k)\left\{F_{Y_1|\mathcal{C}_k}(\widetilde{F}_1^{-1}\widetilde{F}_0(y_0))\left[1-F_{p(Z)}(p_k)\right] + F_{Y_0|\mathcal{C}_k}(y_0)F_{p(Z)}(p_k)\right\} + (1-p_K)F_{Y_0|\mathcal{N}}(y_0),$$

where

$$\widetilde{F}_d(y_d) = \sum_{k=1}^{K-1}(p_{k+1}-p_k)h(p_{k+1})F_{Y_d|\mathcal{C}_k}(y_d) \text{ with } h(p_{k+1}) = \frac{Cov(p(Z), 1(p(Z)\geq p_{k+1}))}{Var(p(Z))}$$

is the potential cdf of a mixed compliers with weights $((p_2-p_1)h(p_1),\cdots,(p_K-p_{K-1})h(p_K))$ on $\mathcal{C}_k$, so we can estimate $F_d^*(\cdot)$ by estimating $F_{Y_1|\mathcal{A}}$, $F_{Y_d|\mathcal{C}_k}$ and $F_{Y_0|\mathcal{N}}$ and then plugging in these two formulas. Specifically, from Abadie (2002) and Imbens and Rubin (1997),

$$\begin{aligned}
F_{Y_1|\mathcal{A}}(y_1) &= \frac{E[1(Y\leq y_1)D|Z=z_1]}{P(D=1|Z=z_1)},\\
F_{Y_d|\mathcal{C}_k}(y_d) &= \frac{E[1(Y\leq y_d)\cdot 1(D=d)|Z=z_{k+1}] - E[1(Y\leq y_d)\cdot 1(D=d)|Z=z_k]}{P(D=d|Z=z_{k+1}) - P(D=d|Z=z_k)},\\
F_{Y_0|\mathcal{N}}(y_0) &= \frac{E[1(Y\leq y_0)(1-D)|Z=z_K]}{P(D=0|Z=z_K)}.
\end{aligned}$$

An alternative estimation method is the inverse quantile regression estimator of Chernozhukov and Hansen (2006), but their estimator does not use the information in (3) and usually targets on a constant quantile treatment effect rather than Q_d so will not be used in this paper. As discussed in Wütherich (2015) and Yu (2016), F_d^* involves using the estimables to replace the unestimables, e.g., the true $F_1(y_1) = p_1F_{Y_1|\mathcal{A}}(y_1) + \sum_{k=1}^{K-1}(p_{k+1}-p_k)F_{Y_1|\mathcal{C}_k}(y_1) + (1-p_K)F_{Y_1|\mathcal{N}}(y_1)$, but $F_{Y_1|\mathcal{N}}$ is not estimable so F_1^* uses $F_{Y_0|\mathcal{N}}(\widetilde{F}_0^{-1}\widetilde{F}_1)$ to substitute it and meanwhile uses a weighted average of $F_{Y_1|\mathcal{C}_k}(y_1)$ and $F_{Y_0|\mathcal{N}}(\widetilde{F}_0^{-1}\widetilde{F}_1)$ to substitute $F_{Y_1|\mathcal{C}_k}$. There are also two methods to estimate $F_{Y_d|\mathcal{C}_k}$. The first method is to employ the $F_{Y_d|\mathcal{C}_k}$ above. The second method is the inverse probability weighting method suggested by Abadie (2003). We use the first method to unify the estimation procedure in all three cases.

4.1 Construction of Test Statistics

Our test statistics are constructed as follows.

Step 1: Let

$$\begin{aligned}
\widehat{F}_{Y_1|\mathcal{A}}(y_1) &= \frac{n_1^{-1}\sum_{i=1}^n 1(Y_i\leq y_1)D_i 1(Z_i=z_1)}{n_1^{-1}\sum_{i=1}^n D_i 1(Z_i=z_1)},\\
\widehat{F}_{Y_d|\mathcal{C}_k}(y_d) &= \frac{n_{k+1}^{-1}\sum_{i=1}^n 1(Y_i\leq y_d)1(D_i=d)1(Z_i=z_{k+1}) - n_k^{-1}\sum_{i=1}^n 1(Y_i\leq y_d)1(D_i=d)1(Z_i=z_k)}{n_{k+1}^{-1}\sum_{i=1}^n 1(D_i=d)1(Z_i=z_{k+1}) - n_k^{-1}\sum_{i=1}^n 1(D_i=d)1(Z_i=z_k)},\\
\widehat{F}_{Y_0|\mathcal{N}}(y_0) &= \frac{n_K^{-1}\sum_{i=1}^n 1(Y_i\leq y_0)(1-D_i)1(Z_i=z_K)}{n_K^{-1}\sum_{i=1}^n (1-D_i)1(Z_i=z_K)},
\end{aligned}$$

and

$$\widehat{p}_k = n_k^{-1}\sum_{i=1}^n D_i 1(Z_i=z_k), \widehat{h}(\widehat{p}_{k+1}) = \frac{\sum_{l=k+1}^K n_l\left(\widehat{p}_l-\overline{\widehat{p}}\right)}{\sum_{l=1}^K n_l\left(\widehat{p}_l-\overline{\widehat{p}}\right)^2},$$

where $n_k = \sum_{i=1}^n 1(Z_i = z_k)$, and $\overline{\widehat{p}} = \sum_{l=1}^K \frac{n_l}{n}\widehat{p}_l = n^{-1}\sum_{i=1}^n D_i$.

Step 2: Let

$$\begin{aligned}\widehat{F}_1(y_1) &= \widehat{p}_1\widehat{F}_{Y_1|\mathcal{A}}(y_1) + \sum_{k=1}^{K-1}(\widehat{p}_{k+1}-\widehat{p}_k)\left\{\widehat{F}_{Y_1|\mathcal{C}_k}(y_1)\left[1-\widehat{F}_Z(z_k)\right] + \widehat{F}_{Y_0|\mathcal{C}_k}\left(\widehat{\widehat{F}}_0^{-1}\widehat{\widehat{F}}_1(y_1)\right)\widehat{F}_Z(z_k)\right\} \\ &\quad + (1-\widehat{p}_K)\widehat{F}_{Y_0|\mathcal{N}}\left(\widehat{\widehat{F}}_0^{-1}\widehat{\widehat{F}}_1(y_1)\right)\end{aligned}$$

and

$$\begin{aligned}\widehat{F}_0(y_0) &= \widehat{p}_1\widehat{F}_{Y_1|\mathcal{A}}\left(\widehat{\widehat{F}}_1^{-1}\widehat{\widehat{F}}_0(y_0)\right) + \sum_{k=1}^{K-1}(\widehat{p}_{k+1}-\widehat{p}_k)\left\{\widehat{F}_{Y_1|\mathcal{C}_k}\left(\widehat{\widehat{F}}_1^{-1}\widehat{\widehat{F}}_0(y_0)\right)\left[1-\widehat{F}_Z(z_k)\right] + \widehat{F}_{Y_0|\mathcal{C}_k}(y_0)\widehat{F}_Z(z_k)\right\} \\ &\quad + (1-\widehat{p}_K)\widehat{F}_{Y_0|\mathcal{N}}(y_0),\end{aligned}$$

which are consistent to $F_d^*(y_d)$, where

$$\widehat{F}_Z(z_k) = n^{-1}\sum_{i=1}^n 1(Z \le z_k), \widehat{\widehat{F}}_d(y_d) = \sum_{k=1}^{K-1}(\widehat{p}_{k+1}-\widehat{p}_k)\widehat{h}(\widehat{p}_{k+1})\widehat{F}_{Y_d|\mathcal{C}_k}(y_d).$$

Step 3: Let

$$T_{1n} = \sum_{k=1}^{K-1}\int_{\mathcal{T}}\left[\widehat{F}_{Y_1|\mathcal{C}_k}\left(\widehat{F}_1^{-1}(\tau)\right) - \widehat{F}_{Y_0|\mathcal{C}_k}\left(\widehat{F}_0^{-1}(\tau)\right)\right]^2 d\tau,$$

and

$$T_{2n} = \sum_{k=2}^{K-1}\int_{\mathcal{T}}\left[\widehat{F}_{Y_1|\mathcal{C}_k}\left(\widehat{F}_{Y_1|\mathcal{C}_1}^{-1}(\tau)\right) - \widehat{F}_{Y_0|\mathcal{C}_k}\left(\widehat{F}_{Y_0|\mathcal{C}_1}^{-1}(\tau)\right)\right]^2 d\tau,$$

where we use a uniform prior on $\mathcal{C}_k$ and $\mathcal{T}$ for simplicity. In practice, the integration in T_{1n} and T_{2n} can be replaced by a summation with respect to $\{\tau_i\}_{i=1}^{\kappa_n}$, where $\kappa_n \to \infty$, and $\{\tau_i\}_{i=1}^{\kappa_n}$ gets dense in $\mathcal{T}$ as $\kappa_n \to \infty$.

Even if $F_{Y_d|\mathcal{C}_k}(y_d)$ is a genuine cdf, $\widehat{F}_{Y_d|\mathcal{C}_k}$ need not be, e.g., $\widehat{F}_{Y_d|\mathcal{C}_k}(y_d)$ may be out of $[0,1]$ and nonmonotonic.[9] So we suggest to conduct the monotone rearrangement operator of Chernozhukov et al. (2010) and truncate $\widehat{F}_{Y_d|\mathcal{C}_k}(y_d)$ to $[0,1]$ before further operations on it.[10] Note also that in practical implementation,

$$\begin{aligned}\widehat{\widehat{F}}_1(y_1) &= \sum_{k=1}^{K-1}\widehat{h}(\widehat{p}_{k+1})\left[n_{k+1}^{-1}\sum_{i=1}^n 1(Y_i \le y_1)D1(Z_i = z_{k+1}) - n_k^{-1}\sum_{i=1}^n 1(Y_i \le y_1)D1(Z_i = z_k)\right], \\ \widehat{\widehat{F}}_0(y_0) &= \sum_{k=1}^{K-1}\widehat{h}(\widehat{p}_{k+1})\left[n_k^{-1}\sum_{i=1}^n 1(Y_i \le y_0)(1-D)1(Z_i = z_k) - n_{k+1}^{-1}\sum_{i=1}^n 1(Y_i \le y_0)(1-D)1(Z_i = z_{k+1})\right]\end{aligned}$$

do not involve any denominator; similarly, $\widehat{F}_1(y_1)$ and $\widehat{F}_0(y_0)$ do not involve any denominator which may be too small to affect the finite-sample performance of our tests.

[9] Note that $\widehat{F}_{Y_1|\mathcal{A}}(y_1)$ and $\widehat{F}_{Y_0|\mathcal{N}}(y_0)$ must be monotone and in $[0,1]$.

[10] We can also rearrange $\widehat{\widehat{F}}_d$ and $\widehat{F}_d$ until we need to do it. Usually, $\widehat{\widehat{F}}_d$ and $\widehat{F}_d$ are more likely to be monotone and stay in $[0,1]$ than $\widehat{F}_{Y_d|\mathcal{C}_k}$ because they are averages of a few correlated cdfs, while the two terms in the influence function of $\widehat{F}_{Y_d|\mathcal{C}_k}$ are independent such that the variation in $\widehat{F}_{Y_d|\mathcal{C}_k}$ is quite large. When K is large and there are covariates, we need to pay more attention since each $\mathcal{C}_{xk}$ cell usually contains fewer data points. In the simulation study of Section 8, rearranging $\widehat{F}_{Y_d|\mathcal{C}_k}$ before further operations on it works better.

4.2 Asymptotics for T_{1n} and T_{2n}

The following theorems state the asymptotic distribution of T_{1n} and T_{2n} under H_0. We also consider the local alternative

$$H_1^\delta : F_{U_1|V}(u|v) - F_{U_0|V}(u|v) = \frac{\delta(u|v)}{\sqrt{n}},$$

where $\delta(u|v)$ falls in

$$\mathcal{F} = \left\{ g(\cdot|\cdot) \in C\left([0,1]^2\right) \middle| \, g(0|v) = g(1|v) = 0, \int_0^1 g(u|v)dv = 0, u, v \in [0,1] \right\},$$

where $g(0|v) = g(1|v) = 0$ because $F_{U_1|V}(0|v) = F_{U_0|V}(0|v) = 0$ and $F_{U_1|V}(1|v) = F_{U_0|V}(1|v) = 1$, and $\int_0^1 g(u|v)dv = 0$ because $\int_0^1 F_{U_1|V}(u|v)dv = \int_0^1 F_{U_0|V}(u|v)dv = u$. Because $\int_0^1 g(u|v)dv = 0$, $g(u|v)$ cannot be the same for all $v \in [0,1]$ unless $g(u|v) = 0$. We assume $g(\cdot|\cdot) \in C\left([0,1]^2\right)$; otherwise, there is a point mass shift in u for some v, or there is a sharp change in the shape of $F_{U_d|V}(u|v)$ for two close v values. Actually, since we have only three types of individuals, we can equivalently specify

$$H_1^\delta : F_{U_1|\text{type}}(\tau) - F_{U_0|\text{type}}(\tau) = \frac{\delta_{\text{type}}(\tau)}{\sqrt{n}}$$

where type $\in \{\mathcal{A}, \mathcal{C}_k, k = 1, \cdots, K-1, \mathcal{N}\}$, and $\delta(\tau) \equiv \left(\delta_{\mathcal{A}}(\tau), \{\delta_{\mathcal{C}_k}(\tau)\}_{k=1}^{K-1}, \delta_{\mathcal{N}}(\tau)\right)'$ falls in

$$\mathcal{F}' = \left\{ g(\cdot) \in C\left([0,1]\right)^{K+1} \middle| \, g(0) = g(1) = \mathbf{0} \equiv (0, \cdots, 0)'_{K+1}, \right.$$

$$\left. p_1 g_1(\tau) + \sum\nolimits_{k=1}^{K-1} (p_{k+1} - p_k) g_{k+1}(\tau) + (1 - p_K) g_{K+1}(\tau) = 0, \tau \in [0,1] \right\}.$$

The relationship between $\delta(\tau)$ and $\delta(u|v)$ is $\delta_k(\tau) = \frac{1}{p_{k+1}-p_k} \int_{p_k}^{p_{k+1}} \delta(\tau|v)dv$ where $k = 0, 1, \cdots, K-1, K$ corresponds to type $= \mathcal{A}, \{\mathcal{C}_k\}_{k=1}^{K-1}, \mathcal{N}$, $p_0 = 0$ and $p_{K+1} = 1$. We also impose the following assumption.

Assumption LA: The joint distribution of W implied by the local alternative is contiguous to that implied under H_0.

The requirement for the contiguity of the local alternative to the null is standard in analyzing the local power. A sufficient condition for contiguity in Case (i) is that

$$\sup_{y: f^{(0)}_{Y_d|\text{type}}(y) > 0} \frac{f^{(1)}_{Y_d|\text{type}}(y)}{f^{(0)}_{Y_d|\text{type}}(y)} < \infty,$$

where $f^{(1)}_{Y_d|\text{type}}(y)$ and $f^{(0)}_{Y_d|\text{type}}(y)$ are the densities of Y_d for different types under H_1^δ and H_0 respectively. Intuitively, this would be the case when $f^{(1)}_{Y_d|\text{type}}(y)$ has lighter tails than $f^{(0)}_{Y_d|\text{type}}(y)$.

To ease the statement of the following theorems, define $\{\varphi_i\}_{i=1}^\infty$ as a class of orthonormal and complete basis functions of $L^2(\mathcal{T})$, that is, $\int_{\mathcal{T}} \varphi_i(\tau)\varphi_j(\tau)d\tau = 1(i = j)$ and any function f such that $\int_{\mathcal{T}} f(\tau)^2 d\tau < \infty$ can be approximated arbitrarily well by selecting enough many φ_i's, where $1(\cdot)$ is the indicator function; define the eigenvalues of a real valued positive semi-definite continuous function on $\mathcal{T} \times \mathcal{T}$, say, $\Sigma(\tau_1, \tau_2)$, as $\{\lambda_i\}_{i=1}^\infty$ such that

$$\int_{\mathcal{T}} \Sigma(\tau_1, \tau_2)\varphi_i(\tau_2)\, d\tau = \lambda_i \varphi_i(\tau_1),$$

where $\lambda_i \geq 0$ need not be distinct, and $\sum_{i=1}^\infty \lambda_i < \infty$.

Theorem 2 *Suppose the assumptions of Corollary 2 hold.*

(i) *Under* H_0,

$$nT_{1n} \rightsquigarrow \sum_{k=1}^{K-1}\sum_{i=1}^{\infty}\lambda_{ki}^{(1)}\varepsilon_{ki}^{(1)2},$$

where $\varepsilon_{ki}^{(1)}$ *are iid* $N(0,1)$ *random variables,* $\lambda_{ki}^{(1)}$*'s are eigenvalues of* $\Sigma_k^{(1)}(\tau_1,\tau_2) \equiv E\left[Z_k^{(1)}(\tau_1)Z_k^{(1)}(\tau_2)\right]$ *with the zero-mean Gaussian process* $Z_k^{(1)}(\tau), \tau \in \mathcal{T}$, *defined in the proof, and the covariance between* $\varepsilon_{ki}^{(1)}$ *and* $\varepsilon_{lj}^{(1)}$, $k,l \in \{1,\cdots,K-1\}$ *and* $k \neq l$, $i,j = 1,2,\cdots$, *is determined by the correlation structure of* $Z_k^{(1)}(\tau)$ *and* $Z_l^{(1)}(\tau)$ *and is defined in the proof.*

(ii) *Under* H_1^{δ} *and Assumption LA,*

$$nT_{1n} \rightsquigarrow \sum_{k=1}^{K-1}\sum_{i=1}^{\infty}\left(b_{ki}^{(1)} + \sqrt{\lambda_{ki}^{(1)}}\varepsilon_{ki}^{(1)}\right)^2,$$

where $b_{ki}^{(1)} = \int_{\mathcal{T}} b_k^{(1)}(\tau)\,\varphi_i(\tau)\,d\tau$ *with*

$$b_k^{(1)}(\tau) = \delta_{\mathcal{C}_k}(\tau) - \sum\nolimits_{l=1}^{K-1}(p_{l+1}-p_l)\,h(p_{l+1})\delta_{\mathcal{C}_l}(\tau)$$

being a linear map of $\delta(\tau)$. *Thus, for any* $c > 0$, $P\left(nT_{1n} > c|H_1^{\delta}\right) \geq P\left(nT_{1n} > c|H_0\right)$, *where the equality holds if and only if* $b_{ki}^{(1)} = 0$ *for any* $k = 1,\cdots,K-1$ *and* $i = 1,2,\cdots$.

(iii) *Under the fixed alternative* H_1 *with* $T_1 = p\lim_{n\to\infty} T_{1n} > 0$,

$$\lim_{n\to\infty} P\left(nT_{1n} > c_n\right) = 1$$

for any sequence of random variables $\{c_n : n \geq 1\}$ *with* $c_n = O_p(1)$.

A corollary of Theorem 2 is the asymptotic distribution of $\widehat{F}_d(\cdot)$ and the QTE,

$$\widehat{\Delta}(\tau) \equiv \widehat{F}_1^{-1}(\tau) - \widehat{F}_0^{-1}(\tau).$$

Corollary 4 *Suppose the assumptions of Corollary 2 hold. Under* H_0,

$$\sqrt{n}\begin{pmatrix}\widehat{F}_1(y_1) - F_1(y_1)\\ \widehat{F}_0(y_0) - F_0(y_0)\end{pmatrix} \rightsquigarrow \begin{pmatrix}Z_1(y_1)\\ Z_0(y_0)\end{pmatrix} \text{ in } \ell^{\infty}(\mathcal{Y}_1)\times\ell^{\infty}(\mathcal{Y}_0),$$

and

$$\sqrt{n}\left(\widehat{\Delta}(\tau) - \Delta(\tau)\right) \rightsquigarrow \frac{Z_0\left(F_0^{-1}(\tau)\right)}{f_0\left(F_0^{-1}(\tau)\right)} - \frac{Z_1\left(F_1^{-1}(\tau)\right)}{f_1\left(F_1^{-1}(\tau)\right)} \text{ in } \ell^{\infty}(\mathcal{T}),$$

where $Z_d(y_d)$ *is defined in the proof, and* $\mathcal{Y}_d$ *is a compact interval in* $\mathcal{S}_d$ *and contains an* ϵ*-enlargement of the set* $\left\{F_d^{-1}(\tau)|, \tau \in \mathcal{T}\right\}$.

Theorem 3 *Suppose the assumptions of Corollary 2 hold.*

(i) *Under* H_0,

$$nT_{2n} \rightsquigarrow \sum_{k=2}^{K-1}\sum_{i=1}^{\infty}\lambda_{ki}^{(2)}\varepsilon_{ki}^{(2)2},$$

where $\varepsilon_{ki}^{(2)}$ *are iid* $N(0,1)$ *random variables,* $\lambda_{ki}^{(2)}$*'s are eigenvalues of* $\Sigma_k^{(2)}(\tau_1,\tau_2) \equiv E\left[Z_k^{(2)}(\tau_1)Z_k^{(2)}(\tau_2)\right]$ *with the zero-mean Gaussian process* $Z_k^{(2)}(\tau), \tau \in \mathcal{T}$*, defined in the proof, and the covariance between* $\varepsilon_{ki}^{(2)}$ *and* $\varepsilon_{lj}^{(2)}$*,* $k,l \in \{2,\cdots,K-1\}$ *and* $k \neq l$*,* $i,j = 1,2,\cdots$*, is determined by the correlation structure of* $Z_k^{(2)}(\tau)$ *and* $Z_l^{(2)}(\tau)$ *and is defined in the proof.*

(ii) *Under* H_1^δ *and Assumption LA,*

$$nT_{2n} \rightsquigarrow \sum_{k=2}^{K-1}\sum_{i=1}^{\infty}\left(b_{ki}^{(2)} + \sqrt{\lambda_{ki}^{(2)}}\varepsilon_{ki}^{(2)}\right)^2,$$

where $b_{ki}^{(2)} = \int_{\mathcal{T}} b_k^{(2)}(\tau)\varphi_i(\tau)\,d\tau$ *with*

$$b_k^{(2)}(\tau) = \delta_{\mathcal{C}_k}(F_{U|\mathcal{C}_1}^{-1}(\tau)) - \frac{f_{U|\mathcal{C}_k}\left(F_{U|\mathcal{C}_1}^{-1}(\tau)\right)}{f_{U|\mathcal{C}_1}\left(F_{U|\mathcal{C}_1}^{-1}(\tau)\right)}\delta_{\mathcal{C}_1}(F_{U|\mathcal{C}_1}^{-1}(\tau))$$

being a linear map of $\delta(\tau)$*. Thus, for any* $c > 0$*,* $P\left(nT_{2n} > c|H_1^\delta\right) \geq P\left(nT_{2n} > c|H_0\right)$*, where the equality holds if and only if* $b_{ki}^{(2)} = 0$ *for any* $k = 2,\cdots,K-1$ *and* $i = 1,2,\cdots$*.*

(iii) *Under the fixed alternative* H_1 *with* $T_2 = p\lim_{n\to\infty} T_{2n} > 0$*,*

$$\lim_{n\to\infty} P\left(nT_{2n} > c_n\right) = 1$$

for any sequence of random variables $\{c_n : n \geq 1\}$ *with* $c_n = O_p(1)$*.*

We provide some comments on the asymptotic distributions of T_{1n} and T_{2n} under H_0. From the proof, both T_{1n} and T_{2n} converge weakly to a sum of $\int_{\mathcal{T}} Z_k(\tau)^2\,d\tau$ for some zero-mean Gaussian processes $Z_k(\cdot)$. We express $\int_{\mathcal{T}} Z_k(\tau)^2\,d\tau$ in the two theorems as $\sum_{i=1}^{\infty}\lambda_{ki}\varepsilon_{ki}^2$ for λ_{ki} being the eigenvalues of $Z_k(\cdot)$'s covariance kernel $\Sigma_k(\tau_1,\tau_2)$ and ε_{ki} being iid $N(0,1)$. Such a representation for a single $Z_k(\cdot)$ process was first introduced to the econometric literature by Birens and Ploberger (1997) through Mercer's theorem although such a form of weak limit as $\int_{\mathcal{T}} Z_k(\tau)^2\,d\tau$ appears even in the classical CM statistic. Though such a representation is standard in econometrics now, here we mention some intuition to aid understanding. First note that $Z_k(\cdot)$ is a collection of uncountably many normal random variables; however, the independent "variation" in $Z_k(\cdot)$ is only countable - $\{\varepsilon_{ki}\}_{i=1}^{\infty}$. The key reason is that $Z_k(\cdot)$ is a continuous process on a compact set $\mathcal{T}$, where the continuity is implied by the continuity of $\Sigma_k(\tau_1,\tau_2)$. Since a continuous function on a compact set must be square-integrable, i.e., $Z_k(\cdot) \in L^2(\mathcal{T})$, while $L^2(\mathcal{T})$ is separable, $\int_{\mathcal{T}} Z_k(\cdot)\,d\tau$ can be represented as a sum of countably many normal random variables. As to why $\{\lambda_{ki}\}_{i=1}^{\infty}$ appear, think of the distribution of $\sum_{l=1}^{L} Z_l^2$, where $(Z_1,\cdots,Z_L)$ are jointly normal with positive semi-definite covariance matrix Σ. From some elementary analysis, it can be represented as $\sum_{l=1}^{L}\lambda_i\varepsilon_i^2$, where $\lambda_i \geq 0$, $i = 1,\cdots,L$, are the eigenvalues of Σ and need not be distinct, and ε_i, $i = 1,\cdots,L$, are iid $N(0,1)$; in other words, $\sum_{l=1}^{L} Z_l^2$ follows a mixed chi-square distribution. $\int_{\mathcal{T}} Z_k(\tau)^2\,d\tau$ is a natural extension of $\sum_{l=1}^{L} Z_l^2$; note here that the eigenvalues get smaller and smaller since $\sum_{i=1}^{\infty}\lambda_{ki} < \infty$. When there are multiple Gaussian processes, we need to take into account the correlation structure between different processes, which is expressed through the correlation between ε_{ki} and ε_{lj}, $k \neq l$, $i,j = 1,2,\cdots$, in the theorems.

We next provide some comments on the local power of T_{1n} and T_{2n}. First, although $\widehat{F}_d$ in T_{1n} uses $\widehat{F}_{Y_0|\mathcal{N}}(y_0)$ and $\widehat{F}_{Y_1|\mathcal{A}}(y_1)$, $b_k^{(1)}(\tau) = \delta_{\mathcal{C}_k}(\tau) - \sum_{l=1}^{K-1}(p_{l+1}-p_l)\,h(p_{l+1})\delta_{\mathcal{C}_l}(\tau)$ does not involve $\delta_{\mathcal{A}}(\tau)$ and $\delta_{\mathcal{N}}(\tau)$. Intuitively, since we can only identify $F_{Y_1|\mathcal{A}}$ and $F_{Y_0|\mathcal{N}}$ but not $F_{Y_0|\mathcal{A}}$ and $F_{Y_1|\mathcal{N}}$, there is no

information on the contrast of $F_{U_1|\mathcal{A}}$ vs. $F_{U_0|\mathcal{A}}$ and $F_{U_1|\mathcal{N}}$ vs. $F_{U_0|\mathcal{N}}$. Second, $b_k^{(1)}(\tau)$ comes from two resources - one from the misspecification of F_d as F_d^* under H_1^δ and the other from $F_{Y_1|\mathcal{C}_k}\left(F_1^{-1}(\tau)\right) - F_{Y_0|\mathcal{C}_k}\left(F_0^{-1}(\tau)\right) = F_{U_1|\mathcal{C}_k}(\tau) - F_{U_0|\mathcal{C}_k}(\tau) \neq 0$ under H_1^δ. Third, note that $b_k^{(1)}(\tau)$ is $\delta_{\mathcal{C}_k}(\tau)$ minus a weighted average of $\{\delta_{\mathcal{C}_l}(\tau)\}_{l=1}^{K-1}$ since $(p_{l+1}-p_l)\,h(p_{l+1}) > 0$ and $\sum_{l=1}^{K-1}(p_{l+1}-p_l)\,h(p_{l+1}) = 1$. So unless $\delta_{\mathcal{C}_l}(\tau) = \delta_{\mathcal{C}_k}(\tau)$ for all $k \neq l$, $b_k(\tau)$ cannot be zero for all $k = 1, \cdots, K-1$, and our test will have power. In the function space $\mathcal{F}'$, this event can rarely happen; the larger K is, the rarer this event is. To see why this event is rare, note that

$$p_1\delta_{\mathcal{A}}(\tau) + \sum_{k=1}^{K-1}(p_{k+1}-p_k)\,\delta_{\mathcal{C}_k}(\tau) + (1-p_K)\,\delta_{\mathcal{N}}(\tau) = 0,$$

so $\delta_{\mathcal{C}_k}(\tau) = \delta_{\mathcal{C}}(\tau)$ for all k implies

$$p_1\delta_{\mathcal{A}}(\tau) + (p_K-p_1)\,\delta_{\mathcal{C}}(\tau) + (1-p_K)\,\delta_{\mathcal{N}}(\tau) = 0;$$

i.e., a weighted average of three functions in $C([0,1])$ is luckily to be zero. Fourth, since $\sum_{k=1}^{K-1}(p_{k+1}-p_k)\,h(p_{k+1})b_k^{(1)}(\tau) = 0$, the "net" source of power is $(K-2)$ dimensional. In other words, when $K=2$ (i.e., Z is binary), T_{1n} does not have power as discussed after Proposition 2. Two dimensions are lost because we need to estimate two parameters in (12). So in essence, T_{1n} is an overidentification test; the power comes from cross mismatching; see Section 7 for more discussions on this point. Fifth, the power of T_{2n} also comes from cross mismatching. However, unlike $b_k^{(1)}(\tau)$ which is a function of only $\delta_{\mathcal{C}_k}(\tau)$, $b_k^{(2)}(\tau)$ involves also nuisance densities $f_{U|\mathcal{C}_k}\left(F_{U|\mathcal{C}_1}^{-1}(\tau)\right)$, $k = 1, \cdots, K-1$. Generally, $\left\{b_k^{(2)}(\tau)\right\}_{k=2}^{K-1}$ is also $K-2$ dimensional because $b_k^{(2)}(\tau) = 0$ requires

$$\frac{\delta_{\mathcal{C}_k}(F_{U|\mathcal{C}_1}^{-1}(\tau))}{\delta_{\mathcal{C}_1}(F_{U|\mathcal{C}_1}^{-1}(\tau))} = \frac{f_{U|\mathcal{C}_k}\left(F_{U|\mathcal{C}_1}^{-1}(\tau)\right)}{f_{U|\mathcal{C}_1}\left(F_{U|\mathcal{C}_1}^{-1}(\tau)\right)},$$

which can rarely happen. Finally, the asymptotic distribution under H_0 in (i) is a special case of that under H_1^δ, and the consistency of the test statistics in (iii) is well understood, so we will state only the asymptotic distribution under H_1^δ in sections 5 and 6.

As discussed above, the power dimension of T_{1n} is $K-2$, while T_{1n} is a sum of $K-1$ Gaussian processes, so it would improve power by combining these $K-1$ processes into $K-2$ processes. More specifically, let

$$T'_{1n} = \sum_{k=1}^{K-2}\int_{\mathcal{T}}\left[\widehat{F}_{Y_1|\mathcal{C}}^{(k)}\left(\widehat{F}_1^{-1}(\tau)\right) - \widehat{F}_{Y_0|\mathcal{C}}^{(k)}\left(\widehat{F}_0^{-1}(\tau)\right)\right]^2 d\tau,$$

where

$$\widehat{F}_{Y_d|\mathcal{C}}^{(k)}(y_d) = \sum_{l=1}^{K-1}\omega_{kl}\widehat{F}_{Y_d|\mathcal{C}_l}(y_d), \omega_k \equiv (\omega_{k1}, \cdots, \omega_{k,K-1}) \in \mathbb{S}_{K-2}, k = 1, \cdots, K-2. \tag{13}$$

with $\mathbb{S}_{n-2} \equiv \left\{(x_1, \cdots, x_{n-1}) \,|x_i \geq 0, \sum_{i=1}^{n-1} x_i = 1\right\}$ being a $(n-2)$-simplex. In practice, we can choose the weights ω_k distinct enough from $\widehat{\omega} \equiv \left((\widehat{p}_2-\widehat{p}_1)\,\widehat{h}(\widehat{p}_2), \cdots, (\widehat{p}_K-\widehat{p}_{K-1})\,\widehat{h}(\widehat{p}_K)\right)$ to magnify power. For example, we can choose the weights as the first $K-2$ maximizers of $\max_{\omega_l}\|\widehat{\omega}-\omega_l\|$, where $\|\cdot\|$ is the Euclidean norm, and ω_l is the lth standard base vector in $\mathbb{R}^{K-1}$. In a similar fashion, we can modify T_{2n} as T'_{2n} below to improve its finite-sample performance,

$$T'_{2n} = \sum_{k=1}^{K-2}\int_{\mathcal{T}}\left[\widehat{F}_{Y_1|\mathcal{C}}^{(k)}\left(\widehat{\widehat{F}}_1^{-1}(\tau)\right) - \widehat{F}_{Y_0|\mathcal{C}}^{(k)}\left(\widehat{\widehat{F}}_0^{-1}(\tau)\right)\right]^2 d\tau,$$

where the mixed compliers is used as the hub, and $\widehat{F}^{(k)}_{Y_1|\mathcal{C}}$ is defined in (13). Other mixed compliers with mixing weights different from those of $\widetilde{F}_d$ can also be used. For example, if we use weight $\frac{p_{k+1}-p_k}{p_K-p_1}$ on $\mathcal{C}_k$, then we are replacing $\widehat{\widetilde{F}}_d(y_d)$ by

$$\widehat{\overleftrightarrow{F}}_d(y_d) = \frac{n_K^{-1}\sum_{i=1}^n 1(Y_i \le y_d)1(D_i = d)1(Z_i = z_K) - n_1^{-1}\sum_{i=1}^n 1(Y_i \le y_d)1(D_i = d)1(Z_i = z_1)}{n_K^{-1}\sum_{i=1}^n 1(D_i = d)1(Z_i = z_K) - n_1^{-1}\sum_{i=1}^n 1(D_i = d)1(Z_i = z_1)}.$$

The following corollary states the asymptotic properties of T'_{1n} and T'_{2n} under H_1^δ.

Corollary 5 *Suppose the assumptions of Corollary 2 hold. Under H_1^δ and Assumption LA,*

$$nT'_{1n} \rightsquigarrow \sum_{k=1}^{K-2}\sum_{i=1}^{\infty}\left(b'^{(1)}_{ki} + \sqrt{\lambda'^{(1)}_{ki}}\varepsilon'^{(1)}_{ki}\right)^2,$$

and

$$nT'_{2n} \rightsquigarrow \sum_{k=1}^{K-2}\sum_{i=1}^{\infty}\left(b'^{(2)}_{ki} + \sqrt{\lambda'^{(2)}_{ki}}\varepsilon'^{(2)}_{ki}\right)^2,$$

where $b'^{(1)}_{ki} = \sum_{l=1}^{K-1}\omega_{kl}b^{(1)}_{li}$, $\varepsilon'^{(1)}_{ki} = \sum_{l=1}^{K-1}\omega_{kl}\varepsilon^{(1)}_{li}$, $\lambda'^{(1)}_{ki}$'s are eigenvalues of

$$\Sigma'^{(1)}_k(\tau_1,\tau_2) \equiv E\left[\left(\sum_{l=1}^{K-1}\omega_{kl}Z^{(1)}_l(\tau_1)\right)\left(\sum_{l=1}^{K-1}\omega_{kl}Z^{(1)}_l(\tau_2)\right)\right] = \sum_{l=1}^{K-1}\sum_{m=1}^{K-1}\omega_{kl}\omega_{km}E\left[Z^{(1)}_l(\tau_1)Z^{(1)}_m(\tau_2)\right],$$

$b'^{(2)}_{ki} = \int_{\mathcal{T}} b'^{(2)}_k(\tau)\varphi_i(\tau)\,d\tau$ *with*

$$\begin{aligned}
b'^{(2)}_k(\tau) &= \sum_{l=1}^{K-1}\omega_{kl}\delta_{\mathcal{C}_k}(\widetilde{F}_U^{-1}(\tau)) - \frac{\sum_{l=1}^{K-1}\omega_{kl}f_{U|\mathcal{C}_k}\left(\widetilde{F}_U^{-1}(\tau)\right)}{\widetilde{f}_U\left(\widetilde{F}_U^{-1}(\tau)\right)}\widetilde{\delta}(\widetilde{F}_U^{-1}(\tau)),\\
\widetilde{F}_U(u) &= \sum_{k=1}^{K-1}(p_{k+1}-p_k)h(p_{k+1})F_{U|\mathcal{C}_k}(u), \widetilde{f}_U(u) = \sum_{k=1}^{K-1}(p_{k+1}-p_k)h(p_{k+1})f_{U|\mathcal{C}_k}(u),\\
\widetilde{\delta}(u) &= \sum_{k=1}^{K-1}(p_{k+1}-p_k)h(p_{k+1})\delta_{\mathcal{C}_k}(u),
\end{aligned}$$

and $\lambda'^{(2)}_{ki}$ and $\varepsilon'^{(2)}_{ki}$ are defined in the proof.

Although $b'^{(1)}_{ki}$ and $\varepsilon'^{(1)}_{ki}$, $k = 1,\cdots,K-2$, have straightforward relationship with $b^{(1)}_{ki}$ and $\varepsilon^{(1)}_{ki}$, $k = 1,\cdots,K-1$, it seems hard to express $\lambda'^{(1)}_{ki}$, $b'^{(2)}_{ki}$, $\varepsilon'^{(2)}_{ki}$ and $\lambda'^{(2)}_{ki}$ as functions of $\lambda^{(1)}_{ki}$, $b^{(2)}_{ki}$, $\varepsilon^{(2)}_{ki}$ and $\lambda^{(2)}_{ki}$.

4.3 Bootstrapping Critical Values of T_{1n} and T_{2n}

The eigenvalues $\lambda^{(1)}_{ki}$ and $\lambda^{(2)}_{ki}$ are necessary inputs to determine the critical values of our tests, but they depend on the data-generating process under the null and are hard to estimate.[11] To make our testing procedure more applicable, we suggest to use the bootstrap to obtain the critical values.

Suppose we use the exchangeable bootstrap to conduct the inference; the detailed procedure is as follows. Let $(\omega_1,\cdots,\omega_n)$ be a vector of nonnegative random variables that satisfy the following Assumption EB. For example, $(\omega_1,\cdots,\omega_n)$ is a multinomial vector with dimension n and probabilities $(1/n,\cdots,1/n)$ in the

[11] Nevertheless, Bierens and Ploberger (1997) provide case-independent upper bounds of the asymptotic critical values of the ICM test; Horowitz (2006) and Blundell and Horowitz (2007) consistently estimate the asymptotic critical values in two specification tests. The situation in our case is more complicated since the correlation structure among ε_{ki} and ε_{lj} is also required.

empirical bootstrap. The exchangeable bootstrap uses the components of $(\omega_1, \cdots, \omega_n)$ as random sampling weights in the construction of the bootstrap version of the estimators. More specifically, we use the following three parallel steps as in the construction of T_{1n} and T_{2n}, where the objects with a superscript $*$ in this subsection indicate the samples or estimators based on the bootstrap measure.

Step 1: Let

$$\begin{aligned}
\widehat{F}^*_{Y_1|\mathcal{A}}(y_1) &= \frac{n^{*-1}\sum_{i=1}^n \omega_i 1(Y_i \le y_1) D_i 1(Z_i = z_1)}{n^{*-1}\sum_{i=1}^n \omega_i D_i 1(Z_i = z_1)}, \\
\widehat{F}^*_{Y_d|\mathcal{C}_k}(y_d) &= \frac{n_{k+1}^{*-1}\sum_{i=1}^n \omega_i 1(Y_i \le y_d) 1(D_i = d) 1(Z_i = z_{k+1}) - n_k^{*-1}\sum_{i=1}^n \omega_i 1(Y_i \le y_d) 1(D_i = d) 1(Z_i = z_k)}{n_{k+1}^{*-1}\sum_{i=1}^n \omega_i 1(D_i = d)(Z_i = z_{k+1}) - n_k^{*-1}\sum_{i=1}^n \omega_i 1(D_i = d)(Z_i = z_k)}, \\
\widehat{F}^*_{Y_0|\mathcal{N}}(y_0) &= \frac{n^{*-1}\sum_{i=1}^n \omega_i 1(Y_i \le y_0)(1 - D_i) 1(Z_i = z_K)}{n^{*-1}\sum_{i=1}^n \omega_i (1 - D_i) 1(Z_i = z_K)},
\end{aligned}$$

and

$$\widehat{p}^*_k = n_k^{*-1}\sum_{i=1}^n \omega_i D_i 1(Z_i = z_k), \widehat{h}^*(\widehat{p}^*_{k+1}) = \frac{\sum_{l=k+1}^K n_l^* \left(\widehat{p}^*_l - \overline{\widehat{p}}^*\right)}{\sum_{l=1}^K n_l^* \left(\widehat{p}^*_l - \overline{\widehat{p}}^*\right)^2},$$

where $n^* = \sum_{i=1}^n \omega_i$, $n_k^* = \sum_{i=1}^n \omega_i 1(Z_i = z_k)$, and $\overline{\widehat{p}}^* = \sum_{l=1}^K \frac{n_l^*}{n^*}\widehat{p}^*_l = n^{*-1}\sum_{i=1}^n \omega_i D_i$.

Step 2: Let

$$\begin{aligned}
\widehat{F}^*_1(y_1) &= \widehat{p}^*_1 \widehat{F}^*_{Y_1|\mathcal{A}}(y_1) + \sum_{k=1}^{K-1}\left(\widehat{p}^*_{k+1} - \widehat{p}^*_k\right)\left\{\widehat{F}^*_{Y_1|\mathcal{C}_k}(y_1)\left[1 - \widehat{F}^*_Z(z_k)\right] + \widehat{F}^*_{Y_0|\mathcal{C}_k}\left(\widehat{\widehat{F}}_0^{*-1}\widehat{\widehat{F}}^*_1(y_1)\right)\widehat{F}^*_Z(z_k)\right\} \\
&\quad + (1 - \widehat{p}^*_K)\widehat{F}^*_{Y_0|\mathcal{N}}\left(\widehat{\widehat{F}}_0^{*-1}\widehat{\widehat{F}}^*_1(y_1)\right)
\end{aligned}$$

and

$$\begin{aligned}
\widehat{F}^*_0(y_0) &= \widehat{p}^*_1 \widehat{F}^*_{Y_1|\mathcal{A}}\left(\widehat{\widehat{F}}_1^{*-1}\widehat{\widehat{F}}^*_0(y_0)\right) + \sum_{k=1}^{K-1}\left(\widehat{p}^*_{k+1} - \widehat{p}^*_k\right)\left\{\widehat{F}^*_{Y_1|\mathcal{C}_k}\left(\widehat{\widehat{F}}_1^{*-1}\widehat{\widehat{F}}^*_0(y_0)\right)\left[1 - \widehat{F}^*_Z(z_k)\right] + \widehat{F}^*_{Y_0|\mathcal{C}_k}(y_0)\widehat{F}^*_Z(z_k)\right\} \\
&\quad + (1 - \widehat{p}^*_K)\widehat{F}^*_{Y_0|\mathcal{N}}(y_0),
\end{aligned}$$

which are consistent to $\widehat{F}_d(y_d)$, where

$$\widehat{F}^*_Z(z_k) = n^{*-1}\sum_{i=1}^n \omega_i 1(Z \le z_k), \widehat{\widehat{F}}^*_d(y_d) = \sum_{k=1}^{K-1}\left(\widehat{p}^*_{k+1} - \widehat{p}^*_k\right)\widehat{h}^*(\widehat{p}^*_{k+1})\widehat{F}^*_{Y_d|\mathcal{C}_k}(y_d).$$

Step 3: Let

$$T^*_{1n} = \sum_{k=1}^{K-1}\int_{\mathcal{T}}\left[\left(\widehat{F}^*_{Y_1|\mathcal{C}_k}\left(\widehat{F}_1^{*-1}(\tau)\right) - \widehat{F}_{Y_1|\mathcal{C}_k}\left(\widehat{F}_1^{-1}(\tau)\right)\right) - \left(\widehat{F}^*_{Y_0|\mathcal{C}_k}\left(\widehat{F}_0^{*-1}(\tau)\right) - \widehat{F}_{Y_0|\mathcal{C}_k}\left(\widehat{F}_0^{-1}(\tau)\right)\right)\right]^2 d\tau,$$

and

$$T^*_{2n} = \sum_{k=2}^{K-1}\int_{\mathcal{T}}\left[\left(\widehat{F}^*_{Y_1|\mathcal{C}_k}\left(\widehat{F}_{Y_1|\mathcal{C}_1}^{*-1}(\tau)\right) - \widehat{F}_{Y_1|\mathcal{C}_k}\left(\widehat{F}_{Y_1|\mathcal{C}_1}^{-1}(\tau)\right)\right) - \left(\widehat{F}^*_{Y_0|\mathcal{C}_k}\left(\widehat{F}_{Y_1|\mathcal{C}_1}^{*-1}(\tau)\right) - \widehat{F}_{Y_0|\mathcal{C}_k}\left(\widehat{F}_{Y_1|\mathcal{C}_1}^{-1}(\tau)\right)\right)\right]^2 d\tau.$$

Use the $(1-\alpha)$th quantile of $n^*T_{1n}^*$, $c_{1n}^*(\alpha)$, and the $(1-\alpha)$th quantile of $n^*T_{2n}^*$, $c_{2n}^*(\alpha)$, as the critical values for nT_{1n} and nT_{2n}, respectively, where $\alpha \in (0, 1/2)$ is some prespecified significance level.

In practice, we can simulate T_{1n}^* B times to get $\{T_{1nb}^*\}_{b=1}^B$ for B large enough, and then reject H_0 if $nT_{1n} > \widehat{c}_{1n}^*(\alpha)$, where $\widehat{c}_{1n}^*(\alpha)$ is the $(1-\alpha)$th quantile of $\{n^*T_{1nb}^*\}_{b=1}^B$ which approximates $c_{1n}^*(\alpha)$. Of course, we can also check whether the p-value $B^{-1}\sum_{b=1}^B 1(n^*T_{1nb}^* \geq nT_{1n})$ is less than α to decide whether to reject H_0. A similar simulation scheme can be applied to T_{2n}^*.

We now describe Assumption EB.

Assumption EB: $(\omega_1, \cdots, \omega_n)$ is an exchangeable, nonnegative random vector, which is independent of the data $\{W_i\}_{i=1}^n$ such that for some $\epsilon > 0$,

$$E\left[\omega_1^{2+\epsilon}\right] < \infty, n^{-1}\sum\nolimits_{i=1}^n (\omega_i - \overline{\omega})^2 \xrightarrow{P^*} 1, \overline{\omega} = n^{-1}\sum\nolimits_{i=1}^n \omega_i \xrightarrow{P^*} 1,$$

where $\xrightarrow{P^*}$ signifies the convergence in the probability of bootstrap measure.[12]

By appropriately selecting $(\omega_1, \cdots, \omega_n)$, the exchangeable bootstrap covers many bootstrap schemes (besides the empirical bootstrap) as special cases. For example, the weighted bootstrap corresponds to the case where $\omega_1, \cdots, \omega_n$ are iid nonnegative random variables with $E[\omega_1] = Var(\omega_1) = 1$, e.g., standard exponential. The m out of n bootstrap corresponds to letting $(\omega_1, \cdots, \omega_n)$ be equal to $\sqrt{n/m}$ times multinomial vectors with parameter m and probabilities $(1/n, \cdots, 1/n)$. The subsampling bootstrap corresponds to letting $(\omega_1, \cdots, \omega_n)$ be a row in which the number $m(n-m)^{-1/2}m^{-1/2}$ appears m times and 0 appears $n-m$ times ordered at random, independent of the data. See Section 3.6.2 of VW for more detailed descriptions. Each bootstrap scheme is useful to a specific application. For example, in small samples with categorical covariates, we might want to use the weighted bootstrap to gain good accuracy and robustness to "small cells", whereas in large samples, where computational tractability can be an important consideration, we might prefer subsampling.

Theorem 4 *Suppose the assumptions of Corollary 2 and Assumption EB hold.*

(i) *Under H_0,*

$$\lim_{n\to\infty} P\left(nT_{1n} > c_{1n}^*(\alpha)\right) = \alpha.$$

(ii) *Under H_1^δ and Assumption LA,*

$$\lim_{n\to\infty} P\left(nT_{1n} > c_{1n}^*(\alpha)\right) \geq \alpha.$$

(iii) *Under the fixed alternative H_1 with $T_1 = p\lim_{n\to\infty} T_{1n} > 0$,*

$$\lim_{n\to\infty} P\left(nT_{1n} > c_{1n}^*(\alpha)\right) = 1.$$

(iv) *(i)-(iii) hold also for T_{2n} and $c_{2n}^*(\alpha)$.*

(i) implies that under H_0, $c_{1n}^*(\alpha) \xrightarrow{p} c_1(\alpha)$, where $c_1(\alpha)$ is the $(1-\alpha)$th quantile of the asymptotic distribution of nT_{1n}, and the randomness in the probability convergence includes both the randomness of the original sample and the independent randomness of the bootstrap simulations (this also applies to other statements in Theorem 4). (ii) states that T_{1n} using $c_{1n}^*(\alpha)$ as the critical value is asymptotically locally unbiased. (iii) implies that T_{1n} using $c_{1n}^*(\alpha)$ as the critical value is consistent. This result is a corollary of

[12] This assumption can be relaxed a little bit as in (3.6.8) of VW.

Theorem 2(iii) since $c_{1n}^*(\alpha)$ is bounded in probability under the fixed alternative. Finally, (iv) implies that these comments also apply to T_{2n}. Note further that a corollary of Theorem 4) is that the exchangeable bootstrap is also valid for the inference of the QTE process $\widehat{\Delta}(\tau)$, $\tau \in \mathcal{T}$.

5 Testing With Discrete Instruments and General Covariates

We in this section discuss the finite-sample analog of T_1^X and T_2^X. For this purpose, we need to estimate $F_{Y_d|X}(\cdot|x)$ and $F_{Y_d|X}^k(\cdot|x)$, $k = 1, \cdots, K-1$. We still estimate $F_{Y_d|X}(\cdot|x)$ by the IV-QRE whose moment conditions are specified as

$$E\left[\left(\begin{array}{c} 1 \\ p(X,Z) \end{array}\right)(\tau - 1(Y \leq Q_0^*(\tau|X) + D \cdot \Delta^*(\tau|X)))\middle| X = x\right] = 0, \tag{14}$$

where $Q_0^*(\cdot|\cdot)$ and $\Delta^*(\cdot|\cdot)$ are similarly defined as in (12). As shown in Yu (2016), the moment conditions imply

$$F_{Y_1|X}^*(y_1|x) = \sum_{k=0}^{K}(p_{x,k+1} - p_{xk})\left\{F_{Y_1|X}^k(y_1|x)\left[1 - F_{p(X,Z)|X}(p_{xk}|x)\right] + F_{Y_0|X}^k\left(\widetilde{F}_{Y_0|X}^{-1}\left(\widetilde{F}_{Y_1|X}(y_1|x)\middle| x\right)\middle| x\right)F_{p(X,Z)|X}(p_{xk}|x)\right\},$$

and

$$F_{Y_0|X}^*(y_0|x) = \sum_{k=0}^{K}(p_{x,k+1} - p_{xk})\left\{F_{Y_0|X}^k(y_0|x)F_{p(X,Z)|X}(p_{xk}|x) + F_{Y_1|X}^k\left(\widetilde{F}_{Y_1|X}^{-1}\left(\widetilde{F}_{Y_0|X}(y_0|x)\middle| x\right)\middle| x\right)\left[1 - F_{p(X,Z)|X}(p_{xk}|x)\right]\right\},$$

where $p_{x0} = 0$, $p_{x,K+1} = 1$, $F_{p(X,Z)|X}(p_{x0}|x) = 0$, $F_{p(X,Z)|X}(p_{xK}|x) = 1$,

$$F_{Y_1|X}^0(y_1|x) = \frac{E\left[1(Y \leq y_1)D|X = x, Z = z_1\right]}{P(D = 1|X = x, Z = z_1)}$$

is the cdf of Y_1 for always-takers $\mathcal{A}_x$,

$$F_{Y_0|X}^K(y_0|x) = \frac{E\left[1(Y \leq y_0)(1 - D)|X = x, Z = z_K\right]}{P(D = 0|X = x, Z = z_K)}$$

is the cdf of Y_0 for never-takers $\mathcal{N}_x$,

$$\widetilde{F}_{Y_d|X}(y_d|x) = \sum_{k=1}^{K-1}(p_{x,k+1} - p_{xk})\,h(p_{x,k+1})F_{Y_d|X}^k(y_d|x),$$

with

$$h(p_{x,k+1}) = \frac{Cov\left(p(X,Z), 1(p(X,Z) \geq p_{x,k+1})|X = x\right)}{Var\left(p(X,Z)|X = x\right)},$$

and

$$F_{Y_d|X}^k(y_d|x) = \frac{E[1(Y \leq y_d)\cdot 1(D=d)|X=x,Z=z_{k+1}] - E[1(Y \leq y_d)\cdot 1(D=d)|X=x,Z=z_k]}{P(D=d|X=x,Z=z_{k+1}) - P(D=d|X=x,Z=z_k)}, \tag{15}$$

$k = 1, \cdots, K-1$, being the cdf of Y_d for compliers $\mathcal{C}_{xk}$. Consequently, we can estimate $F_{Y_d|X}^*(y_d|x)$ by estimating $F_{Y_1|X}^k(y_1|x)$, $k = 0, 1, \cdots, K-1$ and $F_{Y_0|X}^k(y_0|x)$, $k = 1, \cdots, K$ and then plugging in the above formulas.

5.1 Construction of Test Statistics

Our test statistics are based on the distribution regression (DR) proposed by Foresi and Peracchi (1995) and extended by CFM. The details are described as follows.

Step 1: Let

$$\begin{aligned}
\widehat{F}_{Y|X,p(X,Z),D}(y_d|x,p,d) &= \Lambda\left(T\left(x,p\right)'\widehat{\beta}_d(y_d)\right), \\
\widehat{p}_{xk} &\equiv \widehat{p}\left(x,z_k\right)=\Lambda\left(R(x,z_k)'\widehat{\gamma}\right), k=1,\cdots,K, \\
\widehat{q}_{xk} &\equiv \widehat{q}_k\left(x\right)=q_k\left(B(x),\widehat{\eta}\right), k=1,\cdots,K-1, \\
\widehat{q}_{xK} &= 1-\sum\nolimits_{l=1}^{K-1}\widehat{q}_{xl},
\end{aligned}$$

where $y_d \in \mathcal{Y}_d$ with $\mathcal{Y}_d$ being a compact set in $\mathbb{R}$, $\Lambda(\cdot)$ is a link function, $q_k\left(x\right)=P\left(Z=z_k|X=x\right)$, $T\left(X,p(X,Z)\right)=\sum_{k=1}^{K}1\left(Z=z_k\right)T_k\left(X,p(X,z_k)\right)$ is a vector of transformation of X and p and T_k may be different for different k, $R(X,Z)=\sum_{k=1}^{K}1\left(Z=z_k\right)R_k(X)$ is a vector of transformations of X and Z and R_k may be different for different k, $B(X)$ is a vector of transformation of X,

$$\widehat{\beta}_d(y_d)=\arg\max_{\beta}\sum_{i=1}^{n}1(D_i=d)\left[1\left(Y_i\leq y_d\right)\ln\Lambda\left(T\left(X_i,\widehat{p}_i\right)'\beta\right)+1\left(Y_i>y_d\right)\ln\left(1-\Lambda\left(T\left(X_i,\widehat{p}_i\right)'\beta\right)\right)\right] \tag{16}$$

with $\widehat{p}_i=\widehat{p}\left(X_i,Z_i\right)$,

$$\widehat{\gamma}=\arg\max_{\gamma}\sum_{i=1}^{n}\left[D_i\ln\Lambda\left(R(X_i,Z_i)'\gamma\right)+\left(1-D_i\right)\ln\left(1-\Lambda\left(R(X_i,Z_i)'\gamma\right)\right)\right], \tag{17}$$

and

$$\widehat{\eta}=\arg\max_{\eta}\sum_{i=1}^{n}\sum_{k=1}^{K}1\left(Z_i=z_k\right)\ln q_k\left(B(X_i),\eta\right),$$

with $q_K=1-\sum_{l=1}^{K-1}q_l$.

Step 2: Let

$$\begin{aligned}
\widehat{F}^{k}_{Y_1|X}(y_1|x) &= \frac{\widehat{F}_{Y|X,p(X,Z),D}(y_1|x,\widehat{p}_{x,k+1},1)\widehat{p}_{x,k+1}-\widehat{F}_{Y|X,p(X,Z),D}(y_1|x,\widehat{p}_{xk},1)\widehat{p}_{xk}}{\widehat{p}_{x,k+1}-\widehat{p}_{xk}}, \\
\widehat{F}^{k}_{Y_0|X}(y_0|x) &= \frac{\widehat{F}_{Y|X,p(X,Z),D}(y_0|x,\widehat{p}_{xk},0)\left(1-\widehat{p}_{xk}\right)-\widehat{F}_{Y|X,p(X,Z),D}(y_0|x,\widehat{p}_{x,k+1},0)\left(1-\widehat{p}_{x,k+1}\right)}{\widehat{p}_{x,k+1}-\widehat{p}_{xk}},
\end{aligned}$$

for $k=1,\cdots,K-1$,

$$\begin{aligned}
\widehat{F}^{0}_{Y_1|X}(y_1|x) &= \widehat{F}_{Y|X,p(X,Z),D}(y_1|x,\widehat{p}_{x1},1), \\
\widehat{F}^{K}_{Y_0|X}(y_0|x) &= \widehat{F}_{Y|X,p(X,Z),D}(y_0|x,\widehat{p}_{xK},0),
\end{aligned}$$

and

$$\widehat{h}(\widehat{p}_{x,k+1})=\frac{\sum_{l=k+1}^{K}\widehat{q}_{xl}\left(\widehat{p}_{xl}-\overline{\widehat{p}}_x\right)}{\sum_{l=1}^{K}\widehat{q}_{xl}\left(\widehat{p}_{xl}-\overline{\widehat{p}}_x\right)^2}$$

where $\overline{\widehat{p}}_x=\sum_{l=1}^{K}\widehat{q}_{xl}\widehat{p}_{xl}$. Conduct rearrangement if $\widehat{F}^{k}_{Y_d|X}(y_d|x)$ is not monotone.

Step 3: Let

$$\widehat{F}_{Y_1|X}(y_1|x) = \sum_{k=0}^{K} (\widehat{p}_{x,k+1} - \widehat{p}_{xk}) \left\{ \widehat{F}^k_{Y_1|X}(y_1|x) \left[1 - \widehat{F}_{Z|X}(z_k|x)\right] + \widehat{F}^k_{Y_0|X}\left(\widehat{\widehat{F}}^{-1}_{Y_0|X}\left(\widehat{\widehat{F}}_{Y_1|X}\left(y_1|x\right)\Big| x\right)\Big| x\right) \widehat{F}_{Z|X}(z_k|x) \right\},$$

and

$$\widehat{F}_{Y_0|X}(y_0|x) = \sum_{k=0}^{K} (\widehat{p}_{x,k+1} - \widehat{p}_{xk}) \left\{ \widehat{F}^k_{Y_0|X}(y_0|x)\widehat{F}_{Z|X}(z_k|x) + \widehat{F}^k_{Y_1|X}\left(\widehat{\widehat{F}}^{-1}_{Y_1|X}\left(\widehat{\widehat{F}}_{Y_0|X}\left(y_0|x\right)\Big| x\right)\Big| x\right) \left[1 - \widehat{F}_{Z|X}(z_k|x)\right] \right\},$$

which are consistent to $F^*_{Y_d|X}(y_d|x)$, where

$$\widehat{F}_{Z|X}\left(z_k|x\right) = \sum_{l=1}^{k} \widehat{q}_{xl}, \widehat{\widehat{F}}_{Y_d|X}\left(y_d|x\right) = \sum_{k=1}^{K-1} (\widehat{p}_{x,k+1} - \widehat{p}_{xk})\, \widehat{h}(\widehat{p}_{x,k+1}) \widehat{F}^k_{Y_d|X}(y_d|x).$$

Step 4: Let

$$T^X_{1n} = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K-1} \int_{\mathcal{T}} \left[\widehat{F}^k_{Y_1|X}\left(\widehat{F}^{-1}_{Y_1|X}(\tau|X_i)\Big| X_i\right) - \widehat{F}^k_{Y_0|X}\left(\widehat{F}^{-1}_{Y_0|X}(\tau|X_i)\Big| X_i\right)\right]^2 d\tau$$

and

$$T^X_{2n} = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=2}^{K-1} \int_{\mathcal{T}} \left[\widehat{F}^k_{Y_1|X}\left(\widehat{Q}^1_{Y_1|X}(\tau|X_i)\Big| X_i\right) - \widehat{F}^k_{Y_0|X}\left(\widehat{Q}^1_{Y_0|X}(\tau|X_i)\Big| X_i\right)\right]^2 d\tau,$$

where $\widehat{Q}^1_{Y_d|X}$ is the inverse function of $\widehat{F}^1_{Y_d|X}$, $w_k(x) = F_X\left(x\right)$ independent of k and a uniform prior on $\mathcal{T}$ is used for simplicity.

We provide a few comments on the procedure above. First, when X is discrete and takes J values $\{x_1, \cdots, x_J\}$, $J \geq 2$, the estimation procedure can be simplified. Actually, the three-step procedure in Section 4 can be applied conditional on each x_j cell. This is equivalent to using saturated specification in $T\left(\cdot,\cdot\right)$, $R\left(\cdot,\cdot\right)$ and $B\left(\cdot\right)$. In practice, when J is large such that the data size in some cells is limited, the estimation procedure as above is preferred because it imposes some restriction on the relationship among cells and has less coefficients to estimate. Second, although $T\left(\cdot,\cdot\right)$, $R\left(\cdot,\cdot\right)$ and $B\left(\cdot\right)$ can take quite flexible functional forms, e.g., polynomials, B-splines, trigonometric polynomials, wavelets, etc, the number of terms in $T\left(\cdot,\cdot\right)$, $R\left(\cdot,\cdot\right)$ and $B\left(\cdot\right)$ does not depend on n, which facilitates the asymptotic inference.[13] Third, as mentioned at the end of Section 4.2, we can construct the counterparts of T'_{1n} and T'_{2n}, say T'^X_{1n} and T'^X_{2n}, to improve the finite-sample performance. Although we can allow ω_{kl} in (13) to depend on x, it seems more convenient in practice not to allow so. Specifically, we can choose ω_k, $k = 1, \cdots, K-2$, as the first $K-2$ maximizers of $\max_{\omega_l} \sum_{i=1}^{n} \|\widehat{\omega}_{X_i} - \omega_l\|^2$, where ω_l is the lth standard base vector in $\mathbb{R}^{K-1}$, and

[13] Note that our $\beta_d(y_d)$ estimation is semi-parametric because $\beta_d(y_d)$ varies nonparametrically as y varies. As in the supplementary materials of Yu (2015a), we can combine a goodness of fit test of these semi-parametric specifications with our test, but a detailed analysis is beyond the scope of this paper. Also, it is clear that our test is an omnibus test for CRS, framework (5), Assumption M and the semi-parametric specification.

$\widehat{\omega}_x = \left((\widehat{p}_{x2} - \widehat{p}_{x1}) \widehat{h}(\widehat{p}_{x2}), \cdots, (\widehat{p}_{xK} - \widehat{p}_{x,K-1}) \widehat{h}(\widehat{p}_{xK}) \right)'$, and then construct

$$
\begin{aligned}
T_{1n}^{\prime X} &= \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K-2} \int_{\mathcal{T}} \left[\widehat{F}_{Y_1|X}^{(k)} \left(\widehat{F}_{Y_1|X}^{-1}(\tau|X_i) \middle| X_i \right) - \widehat{F}_{Y_0|X}^{(k)} \left(\widehat{F}_{Y_0|X}^{-1}(\tau|X_i) \middle| X_i \right) \right]^2 d\tau, \\
T_{2n}^{\prime X} &= \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K-2} \int_{\mathcal{T}} \left[\widehat{F}_{Y_1|X}^{(k)} \left(\widehat{\widehat{F}}_{Y_1|X}^{-1}(\tau|X_i) \middle| X_i \right) - \widehat{F}_{Y_0|X}^{(k)} \left(\widehat{\widehat{F}}_{Y_0|X}^{-1}(\tau|X_i) \middle| X_i \right) \right]^2 d\tau.
\end{aligned}
$$

where

$$
\widehat{F}_{Y_d|X}^{(k)}(y_d|x) = \sum\nolimits_{l=1}^{K-1} \omega_{kl} \widehat{F}_{Y_d|X}^{l}(y_d|x), k = 1, \cdots, K-2.
$$

Since the asymptotic properties of $T_{1n}^{\prime X}$ and $T_{2n}^{\prime X}$ can be similarly developed as in Corollary 5, we omit the details in the following subsection. Finally, unlike in the case of T_{1n}, when the support of X is large, it is not easy to plot $\widehat{F}_{Y_1|X}^{k} \left(\widehat{F}_{Y_1|X}^{-1}(\tau|X) \middle| X \right)$ and $\widehat{F}_{Y_0|X}^{k} \left(\widehat{F}_{Y_0|X}^{-1}(\tau|X) \middle| X \right)$ in T_{1n}^{X} as a function of τ to provide intuitions on how the null is violated. To summarize information, we can integrate X out in the construction of T_{1n}^{X}. From Frölich (2007), the unconditional cdfs of always-takers $\mathcal{A}$, compliers $\mathcal{C}_k$ and never-takers $\mathcal{N}$ can be obtained as follows:

$$
\begin{aligned}
\widehat{F}_{Y_1|\mathcal{A}}(y_1) &= \frac{n^{-1} \sum_{i=1}^{n} \widehat{p}_{X_i,1} \widehat{F}_{Y|X,p(X,Z),D}(y_1|X_i, \widehat{p}_{X_i,1}, 1)}{\widehat{p}_1}, \\
\widehat{F}_{Y_1|\mathcal{C}_k}(y_1) &= \frac{n^{-1} \sum_{i=1}^{n} \left[\widehat{F}_{Y|X,p(X,Z),D}(y_1|X_i, \widehat{p}_{X_i,k+1}, 1) \widehat{p}_{X_i,k+1} - \widehat{F}_{Y|X,p(X,Z),D}(y_1|X_i, \widehat{p}_{X_i,k}, 1) \widehat{p}_{X_i,k} \right]}{\widehat{p}_{k+1} - \widehat{p}_k}, \\
\widehat{F}_{Y_0|\mathcal{C}_k}(y_0) &= \frac{n^{-1} \sum_{i=1}^{n} \left[\widehat{F}_{Y|X,p(X,Z),D}(y_0|X_i, \widehat{p}_{X_i,k}, 0) (1 - \widehat{p}_{X_i,k}) - \widehat{F}_{Y|X,p(X,Z),D}(y_0|X_i, \widehat{p}_{X_i,k+1}, 0) (1 - \widehat{p}_{X_i,k+1}) \right]}{\widehat{p}_{k+1} - \widehat{p}_k},
\end{aligned}
$$

$$
\widehat{F}_{Y_0|\mathcal{N}}(y_0) = \frac{n^{-1} \sum_{i=1}^{n} (1 - \widehat{p}_{X_i,K}) \widehat{F}_{Y|X,p(X,Z),D}(y_0|X_i, \widehat{p}_{X_i,K}, 0)}{1 - \widehat{p}_K},
$$

where $\widehat{p}_k = n^{-1} \sum_{i=1}^{n} \widehat{p}_{X_i,k}$, $k = 1, \cdots, K$. Then follow the procedure in Section 4.1 to construct $\widehat{F}_{Y_1|\mathcal{C}_k} \left(\widehat{F}_1^{-1}(\tau) \right)$ and $\widehat{F}_{Y_0|\mathcal{C}_k} \left(\widehat{F}_0^{-1}(\tau) \right)$ and plot them as a function of τ to provide intuitions. Except for the purpose of providing intuitions, it is not suggested to use the unconditional counterpart of T_{1n}^{X} as the test statistic because it is expected to have less power than T_{1n}^{X}. Similar comments apply to T_{2n}^{X}, $T_{1n}^{\prime X}$ and $T_{2n}^{\prime X}$.

5.2 Asymptotics for T_{1n}^{X} and T_{2n}^{X}

The following theorems state the asymptotic distribution of T_{1n}^{X} and T_{2n}^{X}. We first impose the following regularity assumptions.

Assumption DR: (a) $p_{xk} \equiv p(x, z_k) = \Lambda \left(R(x, z_k)' \gamma \right)$, $k = 1, \cdots, K$, $F_{Y|X,p(X,Z),D}(y_d|x, p, d) = \Lambda \left(T(x, p)' \beta_d(y_d) \right)$, and $q_{xk} \equiv q_k(x) = q_k(B(x), \eta), k = 1, \cdots, K-1$, for all $y \in \mathcal{Y}_d$, $x \in \mathcal{X}$, $z \in \{z_1, \cdots, z_K\}$ and $p \in \{p_{x1}, \cdots, p_{xK}\}$, where Λ is either Probit or Logit link function, and q_k is derived from multinomial Logit, conditional Logit or multinomial Probit. (b) The region of interest $\mathcal{Y}_d$ is a compact interval in $\bigcap_{x \in \mathcal{X}} \mathcal{S}_{xd}$ and contains an ϵ-enlargement of the set $\left\{ Q_{Y_d|X}(\tau|x) | x \in \mathcal{X}, \tau \in \mathcal{T} \right\} \bigcup \left\{ Q_{Y_d|X}^{1}(\tau|x) | x \in \mathcal{X}, \tau \in \mathcal{T} \right\}$. The conditional density $f_{Y|X,Z,D}(y|x, z, d)$ exists, is uniformly bounded and uniformly continuous in (y, x) in the conditional support of (Y_d, X) given $z \in \{z_1, \cdots, z_K\}$. $\mathcal{X}$ is compact. (c) $E \left[\|(R, T, B)\|^2 \right] < \infty$ and the

minimum eigenvalue of

$$J_p \equiv E\left[\frac{\widetilde{\lambda}^2}{\widetilde{p}\left[1-\widetilde{p}\right]}RR'\right], J_q \equiv E\left[\sum\nolimits_{l=1}^{K}\frac{1(Z=z_k)}{q_k\left(B,\eta\right)^2}\frac{\partial q_k\left(B,\eta\right)}{\partial \eta}\frac{\partial q_k\left(B,\eta\right)}{\partial \eta'}\right]$$

and

$$J_d(y_d) \equiv E\left[1(D=d)\frac{\lambda_d(y_d)^2}{\Lambda_d\left(y_d\right)\left[1-\Lambda_d\left(y_d\right)\right]}TT'\right]$$

is bounded away from zero uniformly over $y \in \mathcal{Y}_d$, where λ is the derivative of Λ, $R = R(X,Z)$, $\widetilde{p} = \Lambda(R'\gamma)$, $\widetilde{\lambda} = \lambda(R'\gamma)$, $T = T\left(X, p(X,Z)\right)$, $\Lambda_d\left(y_d\right) = \Lambda\left(T'\beta_d(y_d)\right)$, $\lambda_d(y_d) = \lambda(T'\beta_d(y_d))$ and $B = B(X)$.

We also consider the local alternative

$$H_1^\delta : F_{U_1|X}^k(\tau|x) - F_{U_0|X}^k(\tau|x) = \frac{\delta^k\left(\tau|x\right)}{\sqrt{n}}$$

where $\delta\left(\tau|x\right) \equiv \left(\delta^k\left(\tau|x\right)|k=0,1,\cdots,K\right)$ falls in

$$\begin{aligned}\mathcal{F}_X \quad = \quad & \left\{ g(\cdot|\cdot) \in C\left([0,1]\,\mathcal{X}\right)^{K+1} \middle| \, g(0|x) = g(1|x) = \mathbf{0}, \right. \\ & \left. p_{x1}g_1\left(\tau|x\right) + \sum\nolimits_{k=1}^{K-1}\left(p_{x,k+1}-p_{xk}\right)g_{k+1}\left(\tau|x\right) + \left(1-p_K\right)g_{K+1}\left(\tau|x\right) = 0, \tau \in [0,1], x \in \mathcal{X} \right\}.\end{aligned}$$

Theorem 5 *In the framework (5), suppose Assumptions M and DR hold. Under H_1^δ and Assumption LA,*

$$nT_{1n}^X \rightsquigarrow \sum_{k=1}^{K-1}\int_{\mathcal{X}}\int_{\mathcal{T}}\left(Z_k^{(1)}\left(\tau,x\right) + b_k^{(1)}\left(\tau,x\right)\right)^2 d\tau dF_X(x),$$

where $Z_k^{(1)}\left(\tau,x\right)$ is defined in the proof, and

$$b_k^{(1)}\left(\tau,x\right) = \delta^k(\tau|x) - \sum\nolimits_{l=1}^{K-1}\left(p_{x,l+1}-p_{xl}\right)h(p_{x,l+1})\delta^l(\tau|x)$$

is a linear map of $\delta\left(\tau|x\right)$. Thus, for any $c>0$, $P\left(nT_{1n}^X > c|H_1^\delta\right) \geq P\left(nT_{1n}^X > c|H_0\right)$, where the equality holds if and only if $b_k^{(1)}\left(\tau,x\right) = 0$ for any $k = 1,\cdots,K-1$, $\tau \in \mathcal{T}$, and P_X almost sure x.

Theorem 6 *In the framework (5), suppose Assumptions M and DR hold. Under H_1^δ and Assumption LA,*

$$nT_{2n}^X \rightsquigarrow \sum_{k=2}^{K-1}\int_{\mathcal{X}}\int_{\mathcal{T}}\left(Z_k^{(2)}\left(\tau,x\right) + b_k^{(2)}\left(\tau,x\right)\right)^2 d\tau dF_X(x),$$

where $Z_k^{(2)}\left(\tau,x\right)$ is defined in the proof, and

$$b_k^{(2)}\left(\tau,x\right) = \delta^k(Q_{U|X}^1(\tau|x)|x) - \frac{f_{U|X}^k\left(Q_{U|X}^1(\tau|x)|x\right)}{f_{U|X}^1\left(Q_{U|X}^1(\tau|x)|x\right)}\delta^1(Q_{U|X}^1(\tau|x)|x)$$

is a linear map of $\delta\left(\tau|x\right)$, $Q_{U|X}^1(\tau|x)$ is the inverse function of $F_{U|X}^1(\tau|x)$ and $f_{U|X}^k\left(\tau|x\right)$ is the density of $F_{U|X}^k(\tau|x)$, $k=1,\cdots,K-1$. Thus, for any $c>0$, $P\left(nT_{2n}^X > c|H_1^\delta\right) \geq P\left(nT_{2n}^X > c|H_0\right)$, where the equality holds if and only if $b_k^{(2)}\left(\tau,x\right) = 0$ for any $k = 2,\cdots,K-1$, $\tau \in \mathcal{T}$, and P_X almost sure x.

The comments on Theorem 2 and 3 can be similarly applied here. The only difference is that the covariates X complicate the null distribution and local power. The exchangeable bootstrap is also valid as detailed in S.2.1.

6 Testing With General Instruments and General Covariates

We first explain why the form of test statistic (11) is not applicable in this general setup. As shown in Yu (2016), when Z includes continuous components, the moment conditions (14) imply

$$\begin{aligned}F^*_{Y_1|X}(y_1|x) = &\int_0^{\underline{p}_x} F_{Y_1|X,V}(y_1|x,v)dv + \int_{\overline{p}_x}^1 F_{Y_0|X,V}\left(\widetilde{F}^{-1}_{Y_0|X}\left(\widetilde{F}_{Y_1|X}\left(y_1|x\right)\middle| x\right)\middle| x,v\right)dv,\\ &+\int_{\underline{p}_x}^{\overline{p}_x}\left[F_{Y_1|X,V}(y_1|x,v)(1-F_{p(X,Z)|X}(v|x)) + F_{Y_0|X,V}\left(\widetilde{F}^{-1}_{Y_0|X}\left(\widetilde{F}_{Y_1|X}\left(y_1|x\right)\middle| x\right)\middle| x,v\right)F_{p(X,Z)|X}(v|x)\right]dv\end{aligned}$$

and

$$\begin{aligned}F^*_{Y_0|X}(y_0|x) = &\int_0^{\underline{p}_x} F_{Y_1|X,V}\left(\widetilde{F}^{-1}_{Y_1|X}\left(\widetilde{F}_{Y_0|X}\left(y_0|x\right)\middle| x\right)\middle| x,v\right)dv + \int_{\overline{p}_x}^1 F_{Y_0|X,V}(y_0|x,v)dv\\ &+\int_{\underline{p}_x}^{\overline{p}_x}\left[F_{Y_1|X,V}\left(\widetilde{F}^{-1}_{Y_1|X}\left(\widetilde{F}_{Y_0|X}\left(y_0|x\right)\middle| x\right)\middle| x,v\right)(1-F_{p(X,Z)|X}(v|x)) + F_{Y_0|X,V}(y_0|x,v)F_{p(X,Z)|X}(v|x)\right]dv,\end{aligned}$$

where

$$\widetilde{F}_{Y_1|X}\left(y_1|x\right) = \int_{\underline{p}_x}^{\overline{p}_x} F_{Y_1|X,V}(y_1|x,v)h\left(v|x\right)dv,\ \widetilde{F}_{Y_0|X}\left(y_0|x\right) = \int_{\underline{p}_x}^{\overline{p}_x} F_{Y_0|X,V}(y_0|x,v)h\left(v|x\right)dv,$$

and

$$h(v|x) = \frac{Cov\left(p(X,Z), 1\left(p(X,Z)\geq v\right)|X=x\right)}{Var\left(p(X,Z)|X=x\right)}.$$

From Yu (2014a), we can identify $F_{Y_1|X,V}(y_1|x,v)$ and $F_{Y_0|X,V}(y_0|x,v)$ for $x\in\mathcal{X}$ and $v\in\mathcal{P}_x=\left[\underline{p}_x,\overline{p}_x\right]$ by

$$\begin{aligned}F_{Y_1|X,V}(y_1|x,v) &= \left.\frac{dE[1(Y\leq y_1)D|X=x,p(X,Z)=p]}{dp}\right|_{p=v} = F_{Y|X,p(X,Z),D}\left(y_1|x,v,1\right) + v\left.\frac{dF_{Y|X,p(X,Z),D}(y_1|x,p,1)}{dp}\right|_{p=v},\\ F_{Y_0|X,V}(y_0|x,v) &= -\left.\frac{dE[1(Y\leq y_0)(1-D)|X=x,p(Z)=p]}{dp}\right|_{p=v} = F_{Y|X,p(X,Z),D}\left(y_0|x,v,0\right) - (1-v)\left.\frac{dF_{Y|X,p(X,Z),D}(y_0|x,p,0)}{dp}\right|_{p=v},\end{aligned} \tag{18}$$

so that $\widetilde{F}_{Y_d|X}\left(y_1|x\right)$ and the $\int_{\underline{p}_x}^{\overline{p}_x}$ term of $F^*_{Y_d|X}(y_1|x)$ can be identified. Also,

$$\begin{aligned}&\int_0^{\underline{p}_x} F_{Y_1|X,V}(y_1|x,v)dv = E\left[1\left(Y\leq y_1\right)D|X=x,p(X,Z)=\underline{p}_x\right] = \underline{p}_x F_{Y|X,p(X,Z),D}(y_1|x,\underline{p}_x,1),\\ &\int_{\overline{p}_x}^1 F_{Y_0|X,V}(y_0|x,v)dv = E\left[1\left(Y\leq y_0\right)(1-D)|X=x,p(X,Z)=\overline{p}_x\right] = (1-\overline{p}_x)F_{Y|X,p(X,Z),D}(y_0|x,\overline{p}_x,0)\end{aligned}$$

are also identifiable. The difficulty in the inference comes from the estimation of $\underline{p}_x$ and $\overline{p}_x$. We can estimate $\underline{p}_x$ and $\overline{p}_x$ by $\widehat{\underline{p}}_x = \min_{1\leq i\leq n}\left(\widehat{p}\left(x,Z_i\right)\right), \widehat{\overline{p}}_x = \max_{1\leq i\leq n}\left(\widehat{p}\left(x,Z_i\right)\right)$ for some estimator of $p(\cdot,\cdot)$, where all $\widehat{p}\left(x,Z_i\right)$ are used in estimation because we assume $\mathcal{Z}$ does not depend on x (otherwise, only $\widehat{p}\left(x,Z_i\right)$ in the neighborhood of x can be used). However, even if $p\left(x,Z_i\right)$ were observable, the convergence rates and asymptotic distributions of $\widehat{\underline{p}}_x$ and $\widehat{\overline{p}}_x$ depend on the tail properties of $p(x,Z)$; even if we assume the density of $p(x,Z)$ is strictly positive on $\mathcal{P}_x$ so that $\min_{1\leq i\leq n}\left(p\left(x,Z_i\right)\right)$ and $\max_{1\leq i\leq n}\left(p\left(x,Z_i\right)\right)$ are n-consistent and do not affect the asymptotic distribution of the test statistic, the convergence rates (and asymptotic distributions) of $\widehat{\underline{p}}_x$ and $\widehat{\overline{p}}_x$ are still unknown if $p\left(x,Z_i\right)$ is replaced by $\widehat{p}\left(x,Z_i\right)$. Of course, we can avoid deriving the asymptotic distribution if the bootstrap is valid; however, the bootstrap validity is not warranted in the boundary estimation (see, e.g., Yu (2014b)). This difficulty does not exist when Z is discrete; this is

also why we switch to tests based on T_3^X and T_4^X where only $F_{Y_1|X,V}(y_1|x,v)$ and $F_{Y_0|X,V}(y_0|x,v)$ for $x \in \mathcal{X}$ and $v \in \mathcal{P}_x$ need to be estimated.

6.1 Construction of Test Statistics

From (18), to estimate $F_{Y_d|X,V}(y_d|x,v)$, we need to estimate $F_{Y|X,p(X,Z),D}(y_d|x,p,d)$. Our estimation is based on distribution regression as in Section 5.1; see also Yu (2014a). More specifically, we use the following three-step procedure.

Step 1: Let

$$\begin{aligned} \widehat{F}_{Y|X,p(X,Z),D}(y_d|x,p,d) &= \Lambda\left(T(x,p)'\widehat{\beta}_d(y_d)\right), \\ \widehat{p}(x,z) &= \Lambda\left(R(x,z)'\widehat{\gamma}\right), \end{aligned}$$

where $y_d \in \mathcal{Y}_d$ with $\mathcal{Y}_d$ being a compact set in $\mathbb{R}$, $\Lambda(\cdot)$ is a link function, and $T(X,p)$ and $R(X,Z)$ are vectors of general transformations of (X,p) and (X,Z) respectively,

$$\widehat{\beta}_d(y_d) = \arg\max_{\beta} \sum_{i=1}^{n} 1(D_i = d)\left[1(Y_i \le y_d)\ln\Lambda\left(T(X_i,\widehat{p}_i)'\beta\right) + 1(Y_i > y_d)\ln\left(1-\Lambda\left(T(X_i,\widehat{p}_i)'\beta\right)\right)\right]$$

with $\widehat{p}_i = \widehat{p}(X_i,Z_i)$, and

$$\widehat{\gamma} = \arg\max_{\gamma} \sum_{i=1}^{n} \left[D_i \ln\Lambda\left(R(X_i,Z_i)'\gamma\right) + (1-D_i)\ln\left(1-\Lambda\left(R(X_i,Z_i)'\gamma\right)\right)\right].$$

Step 2: Let

$$\widehat{F}_{Y_1|X,V}(y_1|x,v) = \widehat{F}_{Y|X,p(X,Z),D}(y_1|x,v,1) + v\left.\frac{\partial T(x,p)'}{\partial p}\right|_{p=v} \widehat{\beta}_1(y_1)\cdot\lambda\left(T(x,v)'\widehat{\beta}_1(y_1)\right),$$

and

$$\widehat{F}_{Y_0|X,V}(y_0|x,v) = \widehat{F}_{Y|X,p(X,Z),D}(y_0|x,v,0) - (1-v)\left.\frac{\partial T(x,p)'}{\partial p}\right|_{p=v} \widehat{\beta}_0(y_0)\cdot\lambda\left(T(x,v)'\widehat{\beta}_0(y_0)\right).$$

Conduct rearrangement if $\widehat{F}_{Y_d|X,V}(y_d|x,v)$ is not monotone.

Step 3: Let

$$T_{3n}^X = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j:\widehat{p}_{ij}\neq v_o}\int_{\mathcal{T}}\left[\widehat{F}_{Y_1|X,V}\left(\widehat{F}^{-1}_{Y_1|X,V}(\tau|X_i,v_o)\middle|X_i,\widehat{p}_{ij}\right) - \widehat{F}_{Y_0|X,V}\left(\widehat{F}^{-1}_{Y_0|X,V}(\tau|X_i,v_o)\middle|X_i,\widehat{p}_{ij}\right)\right]^2 d\tau$$

and

$$\begin{aligned} T_{4n}^X = \max\Bigg\{ &\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} 1(\widehat{p}_{ij}\ge v_o)\int_{\mathcal{T}}\left[\widehat{F}_{Y_1|X,V}\left(\widehat{F}^{-1}_{Y_1|X,V}(\tau|X_i,v_1)\middle|X_i,\widehat{p}_{ij}\right) - \widehat{F}_{Y_0|X,V}\left(\widehat{F}^{-1}_{Y_0|X,V}(\tau|X_i,v_1)\middle|X_i,\widehat{p}_{ij}\right)\right]^2 d\tau, \\ &\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} 1(\widehat{p}_{ij}\le v_o)\int_{\mathcal{T}}\left[\widehat{F}_{Y_1|X,V}\left(\widehat{F}^{-1}_{Y_1|X,V}(\tau|X_i,v_2)\middle|X_i,\widehat{p}_{ij}\right) - \widehat{F}_{Y_0|X,V}\left(\widehat{F}^{-1}_{Y_0|X,V}(\tau|X_i,v_2)\middle|X_i,\widehat{p}_{ij}\right)\right]^2 d\tau\Bigg\}. \end{aligned}$$

Here, $\widehat{p}_{ij} = \widehat{p}(X_i,Z_j)$, and (v_1,v_o,v_2) can be chosen as the first quartile, median and third quartile of

$\{\widehat{p}_{ij}\}_{j=1}^{n}$; implicitly, $w(x,v) = F_{X,p(X,Z)}(x,v)$ and a uniform prior on $\mathcal{T}$ are used in T_{3n}^{X} and T_{4n}^{X} for simplicity.

As mentioned at the end of Section 4.2, we can use a different hub from v_o, v_1 and v_2 to improve the performance. For example, as in T_{2n}', we can use $\widetilde{F}_d$ as the hub and employ different mixed v-subpopulations for comparison. First, let

$$\begin{aligned}\widehat{\widetilde{F}}_{Y_d|X}(y_d|X_i) &= \sum_{j=1}^{n-1}(\widehat{p}_{i,j+1}-\widehat{p}_{ij})\,\widehat{h}(\widehat{p}_{i,j+1})\frac{\widehat{F}_{Y_d|X,V}(y_d|X_i,\widehat{p}_{ij})+\widehat{F}_{Y_d|X,V}(y_d|X_i,\widehat{p}_{i,j+1})}{2}\\ &\equiv \sum_{j=1}^{n-1}(\widehat{p}_{i,j+1}-\widehat{p}_{ij})\,\widehat{h}(\widehat{p}_{i,j+1})\widehat{F}^{j}_{Y_d|X}(y_d|X_i),\end{aligned}$$

where suppose $\widehat{p}_{i1}\le\widehat{p}_{i2}\le\cdots\le\widehat{p}_{in}$,

$$\widehat{h}(\widehat{p}_{i,j+1}) = \frac{\widehat{E}\left[p(X,Z)1\left(p(X,Z)\ge\widehat{p}_{i,j+1}\right)|X=X_i\right]-\widehat{E}\left[p(X,Z)|X=X_i\right]\widehat{E}\left[1\left(p(X,Z)\ge\widehat{p}_{i,j+1}\right)|X=X_i\right]}{\widehat{E}\left[p(X,Z)^2|X=X_i\right]-\widehat{E}\left[p(X,Z)|X=X_i\right]^2} \tag{19}$$

with

$$\widehat{E}\left[g(X,Z)|X=x\right] = H(x)\left(\textstyle\sum_{i=1}^{n}H(X_i)\,H(X_i)'\right)^{-1}\left(\sum_{i=1}^{n}H(X_i)\,\widehat{g}(X_i,Z_i)\right)$$

for various g functions, $\widehat{g}$ is g replacing p by $\widehat{p}$, and $H(X)$ is a vector of transformations of X. Second, choose ω_k, $k=1,\cdots,K_n$, as the first K_n maximizers of $\max_{\omega_l}\sum_{i=1}^{n}\|\widehat{\omega}_{X_i}-\omega_l\|^2$ such that $\|\omega_k-\omega_j\|\ge c_n$, $j<k$, where $\omega_k=(\omega_{k1},\cdots,\omega_{k,n-1})\in\mathbb{S}_{n-2}$, $\widehat{\omega}_{X_i}=\left((\widehat{p}_{i2}-\widehat{p}_{i1})\,\widehat{h}(\widehat{p}_{i2}),\cdots,(\widehat{p}_{in}-\widehat{p}_{i,n-1})\,\widehat{h}(\widehat{p}_{in})\right)'$, and $c_n>0$ is to make sure the weights ω_k are distinct enough. Finally, construct

$$T_{3n}'^{X} = \frac{1}{nK_n}\sum_{i=1}^{n}\sum_{k=1}^{K_n}\int_{\mathcal{T}}\left[\widehat{F}^{(k)}_{Y_1|X}\left(\widehat{\widetilde{F}}^{-1}_{Y_1|X}(\tau|X_i)\middle|X_i\right)-\widehat{F}^{(k)}_{Y_0|X}\left(\widehat{\widetilde{F}}^{-1}_{Y_0|X}(\tau|X_i)\middle|X_i\right)\right]^2 d\tau.$$

where

$$\widehat{F}^{(k)}_{Y_d|X}(y_d|x) = \sum\nolimits_{l=1}^{n-1}\omega_{kl}\widehat{F}^{l}_{Y_d|X}(y_d|x),\, k=1,\cdots,K_n.$$

In practice, we can set c_n as $c\cdot\sqrt{\sum_{i=1}^{n}\|\widehat{\omega}_{X_i}\|^2}$ for c being a small positive number, e.g., 0.1. Also, we can monitor the p-value of $T_{3n}'^{X}$ as a function of K_n to choose K_n, e.g., choose K_n as the smallest positive integer such that the p-values get "stable".

To avoid the estimation of $\widehat{h}(\widehat{p}_{i,j+1})$ in (19), we may use other mixed v-subpopulations as the hub. For example, choose $[\underline{p},\overline{p}]$ as a compact subset of $\bigcap_{x\in\mathcal{X}}\mathcal{P}_x$, and use the equally weighted v-subpopulations on $[\underline{p},\overline{p}]$ as the hub; denote its cdfs as $\overleftrightarrow{F}_{Y_d|X}$. This hub is very convenient since

$$\begin{aligned}\widehat{\overleftrightarrow{F}}_{Y_1|X}(y_1|x) &= \frac{\overline{p}\widehat{F}_{Y|X,p(X,Z),D}(y_1|x,\overline{p},1)-\underline{p}\widehat{F}_{Y|X,p(X,Z),D}(y_1|x,\underline{p},1)}{\overline{p}-\underline{p}},\\ \widehat{\overleftrightarrow{F}}_{Y_0|X}(y_0|x) &= \frac{\left(1-\underline{p}\right)\widehat{F}_{Y|X,p(X,Z),D}(y_0|x,\underline{p},0)-\left(1-\overline{p}\right)\widehat{F}_{Y|X,p(X,Z),D}(y_0|x,\overline{p},0)}{\overline{p}-\underline{p}},\end{aligned}$$

and no differentiation as in $\widehat{F}_{Y_d|X,V}(y_d|x,v)$ is required. Proceed to select $\left[\underline{p}_k,\overline{p}_k\right]$, $k=1,\cdots,K_n$, as the first K_n maximizers of $\int_{\underline{p}}^{\overline{p}}\left|\frac{1}{\overline{p}-\underline{p}}-\frac{1\left(\underline{p}_k<v\le\overline{p}_k\right)}{\overline{p}_k-\underline{p}_k}\right|dv$ such that $\left\|\overline{p}_k-\underline{p}_k\right\|\ge p_n$ and $\int_{\underline{p}}^{\overline{p}}\left|\frac{1\left(\underline{p}_k<v\le\overline{p}_k\right)}{\overline{p}_k-\underline{p}_k}-\frac{1\left(\underline{p}_j<v\le\overline{p}_j\right)}{\overline{p}_j-\underline{p}_j}\right|dv\ge$

c_n, $j < k$; then obtain

$$\begin{aligned}
\overset{\frown}{\underset{\frown}{F}}{}^{(k)}_{Y_1|X}(y_1|x) &= \frac{\overline{p}_k \widehat{F}_{Y|X,p(X,Z),D}(y_1|x,\overline{p}_k,1) - \underline{p}_k \widehat{F}_{Y|X,p(X,Z),D}(y_1|x,\underline{p}_k,1)}{\overline{p}_k - \underline{p}_k}, \\
\overset{\frown}{\underset{\frown}{F}}{}^{(k)}_{Y_0|X}(y_0|x) &= \frac{\left(1-\underline{p}_k\right) \widehat{F}_{Y|X,p(X,Z),D}(y_0|x,\underline{p}_k,0) - (1-\overline{p}_k) \widehat{F}_{Y|X,p(X,Z),D}(y_0|x,\overline{p}_k,0)}{\overline{p}_k - \underline{p}_k},
\end{aligned}$$

where $p_n > 0$ is to guarantee enough data points are used in the estimation of $\overset{\frown}{\underset{\frown}{F}}{}^{(k)}_{Y_d|X}(y_1|x)$, and $c_n > 0$ is to guarantee the uniform weights on $\left[\underline{p}_k, \overline{p}_k\right]$, $k = 1, \cdots, K_n$, are distinct enough. In practice, we may choose p_n such that $n \cdot p_n = 100$, i.e., at least about 100 data points are used in the estimation of $\overset{\frown}{\underset{\frown}{F}}{}^{(k)}_{Y_d|X}(y_d|x)$, and choose c_n to be a small positive number, e.g., $c_n = 0.1$. Finally, construct the test statistic

$$T^{\prime X}_{4n} = \frac{1}{nK_n} \sum_{i=1}^{n} \sum_{k=1}^{K_n} \int_{\mathcal{T}} \left[\overset{\frown}{\underset{\frown}{F}}{}^{(k)}_{Y_1|X} \left(\overset{\frown}{\underset{\frown}{F}}{}^{-1}_{Y_1|X}(\tau|X_i) \middle| X_i \right) - \overset{\frown}{\underset{\frown}{F}}{}^{(k)}_{Y_0|X} \left(\overset{\frown}{\underset{\frown}{F}}{}^{-1}_{Y_0|X}(\tau|X_i) \middle| X_i \right) \right]^2 d\tau.$$

As in $T^{\prime X}_{1n}$ and $T^{\prime X}_{2n}$, we neglect the asymptotic properties of $T^{\prime X}_{3n}$ and $T^{\prime X}_{4n}$. Also, as mentioned at the end of Section 5.1, we can integrate X out to aid intuition.

6.2 Asymptotics for T^X_{3n} and T^X_{4n}

The following theorems state the asymptotic distribution of T^X_{3n} and T^X_{4n}. We first impose the following regularity assumptions.

Assumption DR′: (a) $p(x,z) = \Lambda\left(R(x,z)'\gamma\right)$ and $F_{Y|X,p(X,Z),D}(y_d|x,p,d) = \Lambda\left(T\left(x,p\right)'\beta_d(y_d)\right)$ for all $y \in \mathcal{Y}_d$, $x \in \mathcal{X}$, $z \in \mathcal{Z}$ and $p \in \mathcal{P}_x$, where Λ is either Probit or Logit link function. (b) The region of interest $\mathcal{Y}_d$ is a compact interval in $\bigcap_{x\in\mathcal{X}} \mathcal{S}_{xd}$ and contains an ϵ-enlargement of the set $\left\{Q_{Y_d|X,V}(\tau|x,v)|x \in \mathcal{X}, \tau \in \mathcal{T}, v \in \mathcal{P}_x\right\}$. The conditional density $f_{Y|X,Z,D}(y|x,z,d)$ exists, is uniformly bounded and uniformly continuous in (y,x,z) in the support of (Y_d,X,Z). $\mathcal{X}\mathcal{Z}$ is compact. (c) $E\left[\|(R,T)\|^2\right] < \infty$ and the minimum eigenvalue of

$$J_p \equiv E\left[\frac{\widetilde{\lambda}^2}{\widetilde{p}\left[1-\widetilde{p}\right]} RR'\right] \text{ and } J_d(y_d) \equiv E\left[1(D=d)\frac{\lambda_d(y_d)^2}{\Lambda_d\left(y_d\right)\left[1-\Lambda_d\left(y_d\right)\right]} TT'\right]$$

is bounded away from zero uniformly over $y \in \mathcal{Y}_d$, where R, $\widetilde{p}$, $\widetilde{\lambda}$, T, $\Lambda_d\left(y_d\right)$ and $\lambda_d(y_d)$ are similarly defined as in Assumption DR.

We also consider the local alternative

$$H_1^\delta : F_{U_1|X,V}(\tau|x,v) - F_{U_0|X,V}(\tau|x,v) = \frac{\delta\left(\tau|x,v\right)}{\sqrt{n}},$$

where $\delta\left(\tau|x,v\right)$ falls in

$$\mathcal{F}'_X = \left\{ g(\cdot|\cdot,\cdot) \in C\left(\left[0,1\right]^2 \mathcal{X}\right) \middle| g(0|x,v) = g(1|x,v) = 0, \int_0^1 g(u|x,v)dv = 0, u,v \in [0,1], x \in \mathcal{X} \right\}.$$

Theorem 7 *In the framework (5), suppose Assumptions M, DR′ and P hold. Under H_1^δ and Assumption*

LA,

$$nT_{3n}^{X} \rightsquigarrow \int_{\mathcal{X}} \int_{\mathcal{P}_x} \int_{\mathcal{T}} \left(Z^{(3)}(\tau, x, v) + b^{(3)}(\tau, x, v) \right)^2 d\tau F_{X,p(X,Z)}(x, v),$$

where $Z^{(3)}(\tau, x, v)$ *is defined in the proof, and*

$$b^{(3)}(\tau, x, v) = \delta(Q_{U|X,V}(\tau|x, v_o)|x, v) - \frac{f_{U|X,V}\left(Q_{U|X,V}(\tau|x, v_o)|x, v\right)}{f_{U|X,V}\left(Q_{U|X,V}(\tau|x, v_o)|x, v_o\right)} \delta(Q_{U|X,V}(\tau|x, v_o)|x, v_o)$$

is a linear map of $\delta(\tau|x, v)$*. Thus, for any* $c > 0$*,* $P\left(nT_{3n}^{X} > c|H_1^{\delta}\right) \geq P\left(nT_{3n}^{X} > c|H_0\right)$*, where the equality holds if and only if* $b^{(3)}(\tau, x, v) = 0$ *for any* $\tau \in \mathcal{T}$ *and* $P_{X,p(X,Z)}$ *almost surely* (x, v)*.*

Theorem 8 *In the framework (5), suppose Assumptions M, DR′ and P hold. Under* H_1^{δ} *and Assumption LA,*

$$\begin{aligned} nT_{4n}^{X} \quad \rightsquigarrow \quad \max \Bigg\{ & \int_{\mathcal{X}} \int_{v_o}^{\overline{p}_x} \int_{\mathcal{T}} \left(Z_1^{(4)}(\tau, x, v) + b_1^{(4)}(\tau, x, v) \right)^2 d\tau dF_{X,p(X,Z)}(x, v), \\ & \int_{\mathcal{X}} \int_{\underline{p}_x}^{v_o} \int_{\mathcal{T}} \left(Z_2^{(4)}(\tau, x, v) + b_2^{(4)}(\tau, x, v) \right)^2 d\tau dF_{X,p(X,Z)}(x, v) \Bigg\}, \end{aligned}$$

where $Z_1^{(4)}(\tau, x, v)$ *and* $Z_2^{(4)}(\tau, x, v)$ *are defined in the proof,*

$$b_1^{(4)}(\tau, x, v) = \delta(Q_{U|X,V}(\tau|x, v_1)|x, v) - \frac{f_{U|X,V}\left(Q_{U|X,V}(\tau|x, v_1)|x, v\right)}{f_{U|X,V}\left(Q_{U|X,V}(\tau|x, v_1)|x, v_1\right)} \delta(Q_{U|X,V}(\tau|x, v_1)|x, v_1)$$

and

$$b_2^{(4)}(\tau, x, v) = \delta(Q_{U|X,V}(\tau|x, v_2)|x, v) - \frac{f_{U|X,V}\left(Q_{U|X,V}(\tau|x, v_2)|x, v\right)}{f_{U|X,V}\left(Q_{U|X,V}(\tau|x, v_2)|x, v_2\right)} \delta(Q_{U|X,V}(\tau|x, v_2)|x, v_2)$$

are linear maps of $\delta(\tau|x, v)$*. Thus, for any* $c > 0$*,* $P\left(nT_{4n}^{X} > c|H_1^{\delta}\right) \geq P\left(nT_{2n}^{X} > c|H_0\right)$*, where the equality holds if and only if for any* $\tau \in \mathcal{T}$*,* $b_1^{(4)}(\tau, x, v) = 0$ *for* $P_{X,p(X,Z)}$ *almost surely* $(x, v) \in \mathcal{X} \times [v_o, \overline{p}_x]$ *and* $b_2^{(4)}(\tau, x, v) = 0$ *for* $P_{X,p(X,Z)}$ *almost surely* $(x, v) \in \mathcal{X} \times \left[\underline{p}_x, v_o\right]$*.*

We can still use the exchangeable bootstrap to obtain critical values for T_{3n}^{X} and T_{4n}^{X}; see S.2.2 for the details.

7 Discussion

We in this section compare the tests in this paper with two groups of tests. The first group of tests are the tests of correlated random coefficient models in Heckman et al. (2010) and Heckman and Schmierer (2010). The second group of tests are the URP tests of Yu (2015a).

To test the correlated random coefficient model, represent Y_i as $Y_i = \alpha_i + \beta_i D_i$, where $\alpha_i = Y_{0i}$ and $\beta_i = Y_{1i} - Y_{0i}$. The target is to test $H_0 : Cov(D_i, \alpha_i) = 0$ vs. $H_1 : Cov(D_i, \alpha_i) \neq 0$. If HV's framework is maintained, then this is essentially testing the presence of essential heterogeneity - $E[U_1|V] = E[U_0|V]$. Although Yu (2016) shows that $E[U_1|V] = E[U_0|V]$ and $F_{U_1|V}(u|v) = F_{U_0|V}(u|v)$ do not imply each other, the similarity of form between them makes the latter a natural counterpart of the former in the quantile context. Heckman et al. develop two kinds of tests in HV's framework. The first kind of test is based on the observation that under H_0, two different instrumental variable estimators are both consistent to the average

treatment effect, so significant difference between two such estimators indicates violation of H_0. This idea cannot be applied in the context of this paper although it is somewhat like the cross matching idea in T_2. As shown in Yu (2016), if $p(Z)$ is replaced by a general instrument, say $J(Z)$, in (12), then $\widetilde{F}_1$ and $\widetilde{F}_0$ need not be genuine cdfs and thus $\widetilde{F}_1^{-1}$ and $\widetilde{F}_0^{-1}$ are not well defined. As a result, $\widehat{F}_1(\cdot)$ and $\widehat{F}_0(\cdot)$ cannot be constructed, and so does the test statistic based on their difference. The second kind of test is based on the observation that $E[Y|p(Z)=p] = a + bp$ is a linear function of p under H_0, so any violation of the linearity is an indicator of the falsity of H_0. However, as shown in Yu (2014a),

$$\begin{aligned}\frac{\partial P(Y \leq y|p(Z)=p)}{\partial p} &= F_{Y_1|V}(y|p) - F_{Y_0|V}(y|p) \\ &= F_{U_1|V}(F_1(y)|p) - F_{U_0|V}(F_0(y)|p) = F_{U|V}(F_1(y)|p) - F_{U|V}(F_0(y)|p)\end{aligned}$$

under our H_0, where the right hand side generally depends on p (and y), so $P(Y \leq y|p(Z)=p)$ is not linear in p in general. Nevertheless, from (18),

$$\begin{aligned}\frac{\partial E[1(Y \leq y)D|p(Z)=p]}{\partial p} &= F_{U_1|V}(F_1(y)|p), \\ \frac{\partial E[1(Y \leq y)(1-D)|p(Z)=p]}{\partial p} &= -F_{U_0|V}(F_0(y)|p),\end{aligned}$$

so

$$\frac{\partial E\left[1(Y \leq F_1^{-1}(\tau))D|p(Z)=p\right]}{\partial p} + \frac{\partial E\left[1(Y \leq F_0^{-1}(\tau))(1-D)|p(Z)=p\right]}{\partial p} = F_{U_1|V}(\tau|p) - F_{U_0|V}(\tau|p) = 0$$

under H_0. In other words,

$$\begin{aligned}& E\left[1(Y \leq F_1^{-1}(\tau))D|p(Z)=p\right] + E\left[1(Y \leq F_0^{-1}(\tau))(1-D)|p(Z)=p\right] \\ = & \int_0^p F_{U|V}(\tau|v)\,dv + \int_p^1 F_{U|V}(\tau|v)\,dv = F_U(\tau) = \tau\end{aligned}$$

does not depend on p. Hence the test statistic can be based on

$$E\left[1(Y \leq F_1^{-1}(\tau))D + 1(Y \leq F_0^{-1}(\tau))(1-D) - \tau|p(Z)=p\right] = 0. \tag{20}$$

As in Heckman et al. (2010), we can use the Wald test or Bierens conditional moment test to carry out the test. But as argued in Section 6, the difficulty in our framework is the estimation of $F_d(\cdot)$ when Z includes continuous components. When Z is discrete, (20) is equivalent to

$$E\left[1(Y \leq F_1^{-1}(\tau))D + 1(Y \leq F_0^{-1}(\tau))(1-D) - \tau|Z=z_k\right] = 0,$$

$k = 1, \cdots, K$. To cancel τ, we take difference for $Z = z_{k+1}$ and z_k to have

$$\begin{aligned}& E\left[1(Y \leq F_1^{-1}(\tau))D|Z=z_{k+1}\right] - E\left[1(Y \leq F_1^{-1}(\tau))D|Z=z_k\right] \\ = & E\left[1(Y \leq F_0^{-1}(\tau))(1-D)|Z=z_k\right] - E\left[1(Y \leq F_0^{-1}(\tau))(1-D)|Z=z_{k+1}\right],\end{aligned}$$

which is exactly the testing idea of T_1 by noticing that

$$\begin{aligned} F_{Y_1|\mathcal{C}_k}(F_1^{-1}(\tau)) &= \frac{E\left[1\left(Y \leq F_1^{-1}(\tau)\right) D|Z=z_{k+1}\right] - E\left[1\left(Y \leq F_1^{-1}(\tau)\right) D|Z=z_k\right]}{P(D=1|Z=z_{k+1}) - P(D=1|Z=z_k)}, \\ F_{Y_0|\mathcal{C}_k}(F_0^{-1}(\tau)) &= \frac{E\left[1(Y \leq F_0^{-1}(\tau))(1-D)|Z=z_k\right] - E\left[1(Y \leq F_0^{-1}(\tau))(1-D)|Z=z_{k+1}\right]}{P(D=0|Z=z_k) - P(D=0|Z=z_{k+1})} \end{aligned}$$

and $P(D=1|Z=z_{k+1}) - P(D=1|Z=z_k) = P(D=0|Z=z_k) - P(D=0|Z=z_{k+1})$.

Although the testing ideas of Heckman et al. cannot be generally applied in the quantile context of this paper, there is indeed some similarity. First of all, both Heckman et al.'s tests and our tests are overidentification tests. As argued in Section 4 of Heckman et al. (2010), their two kinds of tests can be treated as special cases of conditional moment tests based on

$$E\left[\begin{pmatrix} 1 \\ J_k(Z) \end{pmatrix}(Y - a - b \cdot p(Z))\right] = 0, k = 1, \cdots, K,$$

where $K \geq 2$ implies overidentification information. Consequently, their tests require multiple instruments or multiple (more than two) values if only one instrument is available. For comparison, our tests are based on moment conditions

$$F_{Y_1|\mathcal{C}_k}(Q_1(\tau)) - F_{Y_0|\mathcal{C}_k}(Q_0(\tau)) = 0, k = 1, \cdots, K-1, \tau \in \mathcal{T},$$

or

$$F_{Y_1|\mathcal{C}_k}\left(Q_{Y_1|\mathcal{C}_1}(\tau)\right) - F_{Y_0|\mathcal{C}_k}\left(Q_{Y_0|\mathcal{C}_1}(\tau)\right) = 0, k = 2, \cdots, K-1, \tau \in \mathcal{T}.$$

As in the comments on Theorem 2, we need $K - 1 \geq 2$ to generate power, which requires Z to take at least three values. The power of our tests also originates from the overidentification information - cross mismatching.

We next compare the tests in this paper and the URP tests in Yu (2015a). As argued in Yu (2015a), the power of URP tests also originates from some overidentification information; however, the overidentification information there is different from that in this paper. To be precise, let U_1 and U_0 be the unconditional ranks in the two treatment statuses; then the null of URP tests is $U_1 = U_0$. This null implies the following moment conditions,

$$Q_{Y_1|X}(F_{Y_0|X}(Q_0(U_0)|x)|x) - Q_1(U_0) = 0, x \in \mathcal{X},$$

where we follow the notational convention of this paper. The intuition for this moment condition is that under the null, the counterfactual Y_1 of $Y_0 = Q_0(U_0)$ when the conditional (on $X = x$) rank is preserved, $Q_{Y_1|X}(F_{Y_0|X}(Q_0(U_0)|x)|x)$, must equal the counterfactual Y_1 of Y_0 when the unconditional rank is preserved, $Q_1(U_0)$. Obviously, if $\mathcal{X}$ includes only a single point, then $Q_{Y_1|X}(F_{Y_0|X}(Q_0(U_0)|x)|x) = Q_1(F_0(Q_0(U_0))) = Q_1(U_0)$ by definition and no testing power is possible. So the overidentification information (or power) of URP tests comes from the multiple (more than one) values of X, while the power of tests in this paper comes from the multiple (more than two) values of Z and no X is required.

8 Simulations

We conduct some simulations in this section to assess the performance of our tests in the benchmark case - Case (i). We will compare the performances of T_{1n}, T_{2n} with different hubs, T'_{1n} and T'_{2n}. $N = 500$

replications of two experiments with sample size 1000 and 2000 are considered. In bootstrapping critical values, the repetition number $B = 399$ for $n = 1000$ and $B = 199$ for $n = 2000$. The significance level α is set at 5%. $\mathcal{T} = [0.2, 0.8]$ and 61 uniformly distributed points on $\mathcal{T}$ are used to approximate the integration on $\mathcal{T}$. Our simulation study is quite limited due to computational cost since a bootstrap cyle is embedded inside a Monte Carlo cycle.[14]

Our data generating process (DGP) takes the form

$$\begin{aligned} Y_d &= 0.5\Phi^{-1}(U_d), \\ D &= 1(p(Z) - V \geq 0), V|Z \sim U(0,1), \end{aligned}$$

where Z takes $K = 3$ values, $z_1, z_2, z_3 = -1, 0, 1$, with equal probability, $p_1, p_2, p_3 = 0.1, 0.5, 0.9$, respectively, and

$$\begin{pmatrix} \Phi^{-1}(U_1) \\ \Phi^{-1}(U_0) \\ \Phi^{-1}(V) \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & a & b \\ a & 1 & c \\ b & c & 1 \end{pmatrix} \right), a, b, c \in [0,1].$$

Our DGP implies

$$\begin{aligned} F_{U_1|V}(u|v) &= P(U_1 \leq u | V = v) = P\left(\Phi^{-1}(U_1) \leq \Phi^{-1}(u) | \Phi^{-1}(V) = \Phi^{-1}(v)\right) \\ &= P\left(N\left(b\Phi^{-1}(v), 1 - b^2\right) \leq \Phi^{-1}(u)\right) = \Phi\left(\frac{\Phi^{-1}(u) - b\Phi^{-1}(v)}{\sqrt{1-b^2}}\right), \\ F_{U_0|V}(u|v) &= \Phi\left(\frac{\Phi^{-1}(u) - c\Phi^{-1}(v)}{\sqrt{1-c^2}}\right), \frac{\delta(u|v)}{\sqrt{n}} = F_{U_1|V}(u|v) - F_{U_0|V}(u|v). \end{aligned}$$

Neither H_0 nor H_1 involves a, a correlation measure between U_1 and U_0. Under H_0, $b = c$ and under H_1, $b \neq c$. To guarantee the covariance matrix to be positive semi-definite, a needs to satisfy $a^2 + b^2 + c^2 - 2abc - 1 \leq 0$. So under H_0, $a^2 + 2b^2 - 2ab^2 - 1 \leq 0$, which does not exclude $a = 0$ (i.e., U_1 and U_0 can be independent as mentioned in Section 2) if $b \leq 1/\sqrt{2}$ or $a = 1$ (i.e., $U_1 = U_0$). Under H_1, if $a = 1$, then $b^2 + c^2 - 2bc \leq 0$, which is impossible if $b \neq c$, but a can be zero if $b^2 + c^2 \leq 1$. To signify H_0 and H_1 by a scalar, set $a = c = 0$ and $b \in [0, 1)$, which implies U_0 is exogenous, so we use a scalar b to control the level of both endogeneity and violation of H_0. We will consider four b values, $0, 0.3, 0.6$ and 0.9, indicating the null, small, medium and large local alternatives, respectively. The two upper panels of Figure 1 show $F_{U_d|\text{type}}(\tau)$ under H_0 and H_1^δ and the implied $\delta_{\text{type}}(\tau)/\sqrt{n}$ when $b = 0.9$. Recall that the power comes only from $\delta_{\mathcal{C}_1}(\tau)/\sqrt{n}$ and $\delta_{\mathcal{C}_2}(\tau)/\sqrt{n}$.

Under our DGP, $P(\mathcal{C}_1) = p_2 - p_1 = 0.4 = p_3 - p_2 = P(\mathcal{C}_2)$. Since

$$\begin{aligned} P(D &= 1|\mathcal{C}_k) = P(V \leq p(Z) | p_k < V \leq p_{k+1}) = \frac{\sum_{l=1}^K P(V \leq p_l, p_k < V \leq p_{k+1}) P(Z = z_l)}{p_{k+1} - p_k} \\ &= \frac{\sum_{l=k+1}^K P(p_k < V \leq p_{k+1}) P(Z = z_l)}{p_{k+1} - p_k} = \sum\nolimits_{l=k+1}^K P(Z = z_l), \end{aligned}$$

we have

$$\begin{aligned} P(D &= 1, \mathcal{C}_1) = (p_2 - p_1) \sum\nolimits_{l=2}^3 P(Z = z_l) = 0.4 \times 2/3 = 0.27, \\ P(D &= 1, \mathcal{C}_2) = (p_3 - p_2) P(Z = z_3) = 0.4 \times 1/3 = 0.13, \end{aligned}$$

[14] I did not consider the with-covariate case because it is very time-consuming, e.g., the computation in the application of next section takes more than one week on my PC, where $N = 1$.

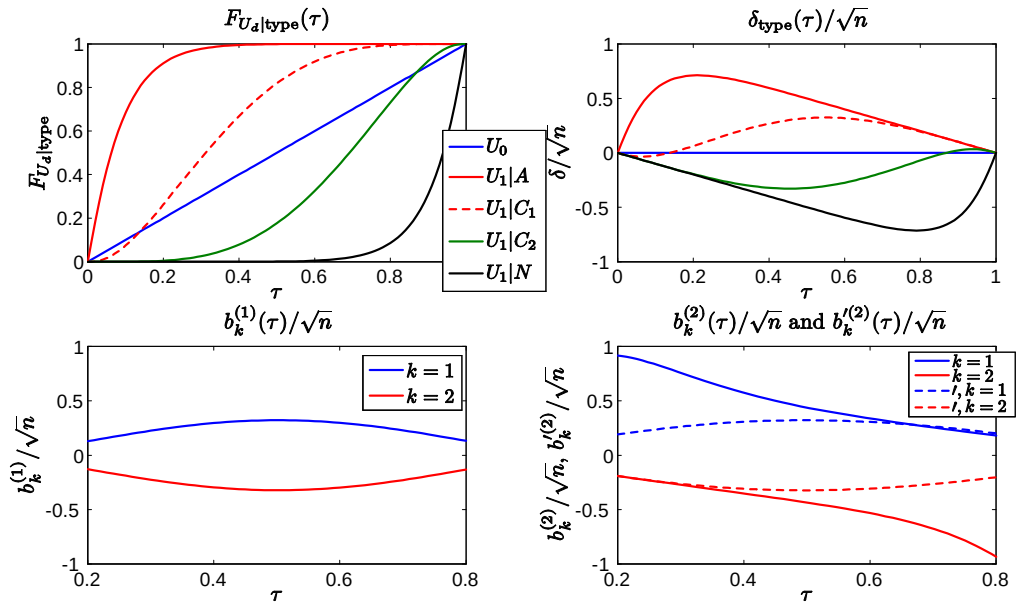

Figure 1: $F_{U_d|\text{type}}$ under H_0 and H_1^δ with Implied $\delta_{\text{type}}/\sqrt{n}$, $b_k^{(1)}, b_k^{(2)}$ and $b_k^{\prime(2)}$: $b = 0.9$

and similarly

$$P(D=0,\mathcal{C}_1) = 0.13, P(D=0,\mathcal{C}_2) = 0.27.$$

In other words, in construction of $\widehat{F}_{Y_0|\mathcal{C}_1}$ and $\widehat{F}_{Y_1|\mathcal{C}_2}$, we are using only about 130 data points when $n = 1000$. The specification of $p(Z)$ and the distribution of Z imply $E[D] = 0.5$, and

$$\begin{aligned} h(p_2) &= \frac{Cov\left(p(Z), 1\left(p(Z)\geq p_2\right)\right)}{Var\left(p(Z)\right)} = \frac{\left(p_2-0.5\right)/3+\left(p_3-0.5\right)/3}{\left(p_1-0.5\right)^2/3+\left(p_2-0.5\right)^2/3+\left(p_3-0.5\right)^2/3} = \frac{5}{4}, \\ h(p_3) &= \frac{Cov\left(p(Z), 1\left(p(Z)\geq p_3\right)\right)}{Var\left(p(Z)\right)} = \frac{\left(p_3-0.5\right)/3}{\left(p_1-0.5\right)^2/3+\left(p_2-0.5\right)^2/3+\left(p_3-0.5\right)^2/3} = \frac{5}{4}, \end{aligned}$$

so

$$\begin{aligned} \widetilde{F}_1(y_1) &= 0.4\times\frac{5}{4}\times F_{Y_1|\mathcal{C}_1}(y_1) + 0.4\times\frac{5}{4}\times F_{Y_1|\mathcal{C}_2}(y_1) \\ &= \frac{0.5}{0.5-0.1}\int_{0.1}^{0.5}\Phi\left(\frac{y_1-b\Phi^{-1}(v)}{\sqrt{1-b^2}}\right)dv + \frac{0.5}{0.9-0.5}\int_{0.5}^{0.9}\Phi\left(\frac{y_1-b\Phi^{-1}(v)}{\sqrt{1-b^2}}\right)dv \\ &= \frac{5}{4}\int_{0.1}^{0.9}\Phi\left(\frac{y_1-b\Phi^{-1}(v)}{\sqrt{1-b^2}}\right)dv, \\ \widetilde{F}_0(y_0) &= \Phi(y_0). \end{aligned}$$

In T_{2n}, we use either $\mathcal{C}_2$ or $\mathcal{C}_1$ as the hub, and the resulting test statistics and $b_k^{(2)}(\tau)$ are denoted as $T_{2n}^{(1)}$, $T_{2n}^{(2)}$, $b_1^{(2)}(\tau)$ and $b_2^{(2)}(\tau)$, respectively. We consider also T_{1n}' and T_{2n}' using the weight selection procedure suggested at the end of Section 4.2. To assess the performance of our weight selection procedure, we consider also $T_{1n}^{(1)}, T_{1n}^{(2)}, T_{2n}^{\prime(1)}$ and $T_{2n}^{\prime(2)}$. The first two are variants of T_{1n} which use only $\mathcal{C}_1$ and $\mathcal{C}_2$ in the construction of T_{1n}, and the last two are variants of T_{2n}' which use directly, rather than select, $\mathcal{C}_1$ and $\mathcal{C}_2$ in T_{2n}'. So totally, we will consider nine test statistics, $T_{1n}, T_{2n}^{(1)}, T_{2n}^{(2)}, T_{1n}', T_{1n}^{(1)}, T_{1n}^{(2)}, T_{2n}', T_{2n}^{\prime(1)}$ and $T_{2n}^{\prime(2)}$. Note

that in these nine statistics, except T_{1n}, the summation with respect to k includes only one term. Since $(p_2 - p_1)\, h(p_2) = (p_3 - p_2)\, h(p_3) = 0.5$, we select $\mathcal{C}_1$ and $\mathcal{C}_2$ with equal probability in T'_{1n} and T'_{2n}. As a result, $b_k^{\prime(1)}$ is $b_1^{(1)}$ or $b_2^{(1)}$ with equal probability in T'_{1n}, and $b_k^{\prime(2)}$ is $b_1^{\prime(2)}$ or $b_2^{\prime(2)}$ with equal probability in T'_{2n}, where

$$b_k^{\prime(2)}(\tau) = \delta_{\mathcal{C}_k}(\widetilde{F}_U^{-1}(\tau)) - \frac{f_{U|\mathcal{C}_k}\left(\widetilde{F}_U^{-1}(\tau)\right)}{\widetilde{f}_U\left(\widetilde{F}_U^{-1}(\tau)\right)}\widetilde{\delta}(\widetilde{F}_U^{-1}(\tau)), k = 1, 2,$$

with $\widetilde{F}, \widetilde{f}$ and $\widetilde{\delta}$ being equally weighted averages of the counterparts for $\mathcal{C}_1$ and $\mathcal{C}_2$. Here, we assume the null rank cdf $F_{U|\mathcal{C}_k}$ satisfies $F_{U_1|\mathcal{C}_k}(\tau) + F_{U_0|\mathcal{C}_k}(\tau) = 2F_{U|\mathcal{C}_k}(\tau)$, $k = 1, 2$. For example,

$$\begin{aligned} F_{U|\mathcal{C}_1}(u) &= \frac{1}{2(0.5 - 0.1)} \int_{0.1}^{0.5} \Phi\left(\frac{\Phi^{-1}(u) - b\Phi^{-1}(v)}{\sqrt{1 - b^2}}\right) dv + \frac{u}{2}, \\ F_{U|\mathcal{C}_2}(u) &= \frac{1}{2(0.9 - 0.5)} \int_{0.5}^{0.9} \Phi\left(\frac{\Phi^{-1}(u) - b\Phi^{-1}(v)}{\sqrt{1 - b^2}}\right) dv + \frac{u}{2}, \end{aligned}$$

and

$$\begin{aligned} f_{U|\mathcal{C}_1}(u) &= \frac{1}{2(0.5 - 0.1)} \int_{0.1}^{0.5} \phi\left(\frac{\Phi^{-1}(u) - b\Phi^{-1}(v)}{\sqrt{1 - b^2}}\right) \frac{1}{\sqrt{1 - b^2}\phi(\Phi^{-1}(u))} dv + \frac{1}{2}, \\ f_{U|\mathcal{C}_2}(u) &= \frac{1}{2(0.9 - 0.5)} \int_{0.5}^{0.9} \phi\left(\frac{\Phi^{-1}(u) - b\Phi^{-1}(v)}{\sqrt{1 - b^2}}\right) \frac{1}{\sqrt{1 - b^2}\phi(\Phi^{-1}(u))} dv + \frac{1}{2}. \end{aligned}$$

To intuitively illustrate the formation of local power, we plot $b_1^{(1)}/\sqrt{n}, b_2^{(1)}/\sqrt{n}, b_1^{(2)}/\sqrt{n}, b_2^{(2)}/\sqrt{n}, b_1^{\prime(2)}/\sqrt{n}$ and $b_2^{\prime(2)}/\sqrt{n}$ for $\tau \in \mathcal{T}$ in the two lower panels of Figure 1. Note that $0.5 \times b_1^{(1)} + 0.5 \times b_2^{(1)} = 0$ as expected.

$n \to$	1000				2000			
$b \to$	0	0.3	0.6	0.9	0	0.3	0.6	0.9
T_{1n}	0.044	0.164	0.550	0.982	0.050	0.268	0.856	1.000
$T_{2n}^{(1)}$	0.054	0.172	0.558	0.986	0.052	0.272	0.860	1.000
$T_{2n}^{(2)}$	0.050	0.172	0.554	0.970	0.050	0.274	0.848	1.000
T'_{1n}	0.050	0.122	0.570	0.978	0.048	0.256	0.820	1.000
$T_{1n}^{(1)}$	0.040	0.164	0.562	0.984	0.050	0.282	0.858	1.000
$T_{1n}^{(2)}$	0.044	0.154	0.512	0.972	0.054	0.242	0.830	1.000
T'_{2n}	0.050	0.112	0.568	0.980	0.048	0.250	0.830	1.000
$T_{2n}^{(1)\prime}$	0.054	0.126	0.592	0.984	0.046	0.264	0.840	1.000
$T_{2n}^{(2)\prime}$	0.044	0.112	0.532	0.970	0.048	0.248	0.816	1.000

Table 1: Size and Power of Various Forms of T_{1n}, T'_{1n}, T_{2n} and T'_{2n}: $\alpha = 0.05, N = 500$

Table 1 summarizes the simulation results. From Table 1, a few conclusions can be drawn. First, all tests perform satisfactorily well - the sizes are close to the nominal level 5%, and the powers are reasonably high; as expected, the powers when $n = 2000$ are better than those when $n = 1000$. Second, the power of T_{1n} is between those of $T_{1n}^{(1)}$ and $T_{1n}^{(2)}$. Third, the powers of $T_{2n}^{(1)}$ ($T_{2n}^{(2)}$) are generally better than those of $T_{1n}^{(1)}$ ($T_{1n}^{(2)}$) and $T_{2n}^{(1)\prime}$ ($T_{2n}^{(2)\prime}$), which matches the local power functions shown in the two lower panels of Figure 1. Fourth, the performance of T'_{2n} is between those of $T_{2n}^{(1)\prime}$ and $T_{2n}^{(2)\prime}$, while the relative performance of T'_{1n} compared to $T_{1n}^{(1)}$ and $T_{1n}^{(2)}$ depends on b and n. Basically, adaptively selecting the weights will introduce extra uncertainty to the test statistics, so the relative performances of the tests using adaptive weights depend on the trade-off

between the power benefit and the extra uncertainty. This is not a problem in practice since we have only one dataset, so we can just report the performances of $T_{1n}^{(1)}$, $T_{1n}^{(2)}$, $T_{2n}^{(1)\prime}$ and $T_{2n}^{(2)\prime}$ and check the sensitivity of the testing results. Of course, when K is large, it is still meaningful to report only T_{1n}' and T_{2n}' for simplicity.

9 Application

We use the dataset of Angrist and Krueger (1991) to illustrate some main points of this paper. This dataset is also used in the empirical study of Chernozhukov and Hansen (2006) as an illustration of IV-QRE, so the tests in this section can serve as pretests for their estimation. Angrist and Krueger (1991) estimate schooling coefficients using quarter of birth as instrument in a sample of 329509 men born 1930-39 from the 1980 census. Quarter of birth is correlated with educational attainment because of a mechanical interaction between compulsory school attendance laws and age at school entry. See the appendix to Angrist and Krueger (1991) for a detailed description of the data.

To fit the data into the framework of this paper, define $D = 1(S > 12)$ to be the indicator of a high school graduate, where $S \in \{0, 1, \cdots, 20\}$ is the years of schooling. Y is the log weekly wage, and X is a vector of covariates consisting of state and year of birth fixed effects.[15] Z can take four values, indicating the four quarters of birth. However, the marginal information when Z increases from one to four is decreasing, so we combine the third and fourth quarter to condense information. More specifically, if Z takes four values, $\widehat{p}_k = 0.3881, 0.3962, 0.4023, 0.4054$ for $k = 1, 2, 3, 4$, so $\widehat{p}_2 - \widehat{p}_1 = 0.008 > \widehat{p}_3 - \widehat{p}_2 = 0.006 > \widehat{p}_4 - \widehat{p}_3 = 0.003$. If Z takes $K = 3$ values, then $\widehat{p}_k = 0.3881, 0.3962, 0.4038$ for $k = 1, 2, 3$, so $\widehat{p}_2 - \widehat{p}_1 = 0.008$ is comparable to $\widehat{p}_3 - \widehat{p}_2 = 0.0075$. Obviously, the instrument is quite weak, but since $n = 329509$ is very large, $n(\widehat{p}_2 - \widehat{p}_1) = 2677$ and $n(\widehat{p}_3 - \widehat{p}_2) = 2494$, and we have enough data points to estimate the cdfs for compliers. Another reason to combine $Z = 3$ and $Z = 4$ is that $\widehat{F}_{Y_d|\mathcal{C}_3}$ and $\widehat{F}_{Y_d|\mathcal{C}_4}$ are not very stable if Z takes four values. One key assumption imposed in this paper is the monotonicity assumption whose validity is shown in Angrist and Imbens (1995).

Although X is discrete, it can take 510 possible values with the minimal cell containing only 3 data points (and the maximal cell containing 3203 data points), so it is better not to use the saturated specification of X but to impose some restriction on the relationship among cells. Specifically, our conditional distribution of the returns to schooling is specified as

$$P(Y \leq y|X, p(X,Z) = p, D = d) = \Lambda\left(\left(p, p^2\right)\alpha_d(y) + X'\beta_d(y)\right),^{16}$$

and the treatment status is determined by

$$D = 1(V \leq X'\gamma_1 + Z'\gamma_2),$$

where $\Lambda(\cdot)$ is the cdf of the standard normal, and V follows a standard normal distribution. Unlike in Chernozhukov and Hansen (2006), D is binary rather than the years of schooling S to fit in the framework of this paper. Also, we use dummies for the three quarters of birth rather than both the linear projection of S onto X and the three dummies as instruments. The distribution regression is conducted by the matlab function glmfit; we estimate $\widehat{F}_{Y|X,p(X,Z),D}(y_d|x, p, d)$ at 200 y_d values uniformly from $[\widehat{q}_d(0.01), \widehat{q}_d(0.99)]$ and

[15] The state and year of birth fixed effects include 59 dimensions. We add a constant in X in the outcome equation but exclude the constant in the participation equation due to the specification of Z.

[16] We also tried cubic polynomials of p, and the results are qualitatively similar. Note here that even if we use the saturated specification for X, the specificaiton is not fully saturated to both X and p. The fully saturated model should include the interaction terms of p and X. We neglect such interaction terms because $\mathcal{P}_x$ includes only three points and does not include much variation given that the instruments are relatively weak.

interpolate other y_d values, where $\widehat{q}_d(\tau)$ is the τth sample quantile of Y with $D=d$. All components of $\widehat{\gamma}_2$ are positive, which guarantees the propensity score is increasing in Z for any X value. Among $\widehat{\gamma}_1$, the coefficients on the year of birth dummies are increasing with the year, indicating that attending college is more popular in later years. In the algorithm in Section 5.1, we need also estimate $q_k(x)=P(Z=z_k|X=x)$. We tried the multinomial Logit model by using the matlab function mnrfit, where

$$q_k(B(x),\eta)=\frac{\exp\{x'\beta_k\}}{\sum_{l=1}^{K}\exp\{x'\beta_l\}}$$

with $\eta=\left(\beta_1',\cdots,\beta_K'\right)'$, but almost all coefficients are insignificant at the 5% level.[17] This is, of course, because Z (i.e., the quarter of birth) is independent of any other variables. As a result, we let $q_k(x)=q_k$ independent of x. Another justification of $Z\perp X$ is that in the sample,

$$\int p(x,Z)dF_X(x)=\int E[D|X=x,Z]\,dF_X(x)=\int E[D|X=x,Z]\,dF_{X|Z}(x|Z)=E[D|Z]\equiv p(Z)$$

hold perfectly.

Tests	$T_{2n}^{X(1)}$	$T_{2n}^{X(2)}$	$T_{2n}^{\prime X(1)}$	$T_{2n}^{\prime X(2)}$
Test Stat.	7.801×10^{-4}	0.0012	7.947×10^{-4}	7.591×10^{-4}
p-values	1.000	0.995	1.000	1.000

Table 2: Test Statistics and p-values for Four Tests

As suggested by the simulation study in Section 8, we check the performance of $T_{1n}, T_{1n}^{(1)}, T_{1n}^{(2)}, T_{2n}^{(1)}, T_{2n}^{(2)}, T_{2n}^{\prime(1)}$ and $T_{2n}^{\prime(2)}$ with covariates. However, the first three test statistics perform badly. This is because $\widehat{F}_{Y_0|\mathcal{N}}$ statistically dominates $\widehat{\widetilde{F}}_0$ (for each X value) such that $\widehat{F}_{Y_0|\mathcal{N}}\left(\widehat{\widetilde{F}}_0^{-1}(\cdot)\right)$ takes very small values. Given that the probability of never takers is large (e.g., in the unconditional model, this probability is about 0.6), $\widehat{F}_{Y_1}$ is quite small (e.g., $\max\left(\widehat{F}_{Y_1}(\cdot)\right)$ can be around 0.2 for some X values) such that $\widehat{F}_{Y_1}^{-1}$ in the first three test statistics does not perform well, which significantly affects their performances. As a result, we report only the last four test statistics in Table 2. In construction of the test statistics, $\mathcal{T}=[0.2,0.8]$ and 61 uniformly distributed points on $\mathcal{T}$ are used to approximate the integration on $\mathcal{T}$. In bootstrapping critical values, the repetition number $B=199$. From Table 2, the dominating conclusion from all tests is that we cannot reject the null, i.e., the IV-QRE in Chernozhukov and Hansen (2006) is justified at least based on our current tests. Note also that the optimal complier based on $\max_{l=1,2}\left\{\sum_{j=1}^{J}\left\|\widehat{\omega}_{x_j}-\omega_l\right\|^2\widehat{p}_{x_j}\right\}$, where $\widehat{\omega}_x=\left(\left(\widehat{p}_{x2}-\widehat{p}_{x1}\right)\widehat{h}(\widehat{p}_{x2}),\left(\widehat{p}_{x3}-\widehat{p}_{x2}\right)\widehat{h}(\widehat{p}_{x3})\right)'$, $\widehat{p}_{x_j}=n^{-1}\sum_{i=1}^{n}1(X_i=x_j)$ and $J=510$, as suggested in Section 5.1 is $\mathcal{C}_1$.

To provide more intuition on the testing result, we plot the unconditional cdfs in Figure 2 as suggested at the end of Section 5.1. Some interesting conclusions can be drawn from Figure 2. First, from the upper left panel, (i) $F_{Y_1|\mathcal{C}_1}$ and $F_{Y_1|\mathcal{C}_2}$ are almost the same, while $F_{Y_0|\mathcal{C}_2}$ stochastically dominates $F_{Y_0|\mathcal{C}_1}$, so if the rank is preserved, then the QTEs for $\mathcal{C}_1$ are larger than those for $\mathcal{C}_2$. This explains why $\mathcal{C}_1$ is more eager to enter college than $\mathcal{C}_2$. (ii) $F_{Y_1|\mathcal{A}}$ tends to be stochastically dominated by $F_{Y_1|\mathcal{C}_1}$ and $F_{Y_1|\mathcal{C}_2}$ at most y values, i.e., Y_1 for always-takers tends to be lower than compliers; parallely, Y_0 for never-takers tends to be higher than compliers. It is possible that Y_0 for always-takers is even lower (e.g., fewer chances if not going to college), so the return to college is much higher for them than compliers; parallely, since Y_0 for never-takers is already

[17] We can also try the multinomial Probit model, but it is quite time-consuming since usually the simulation method such as the GHK simulator would be employed.

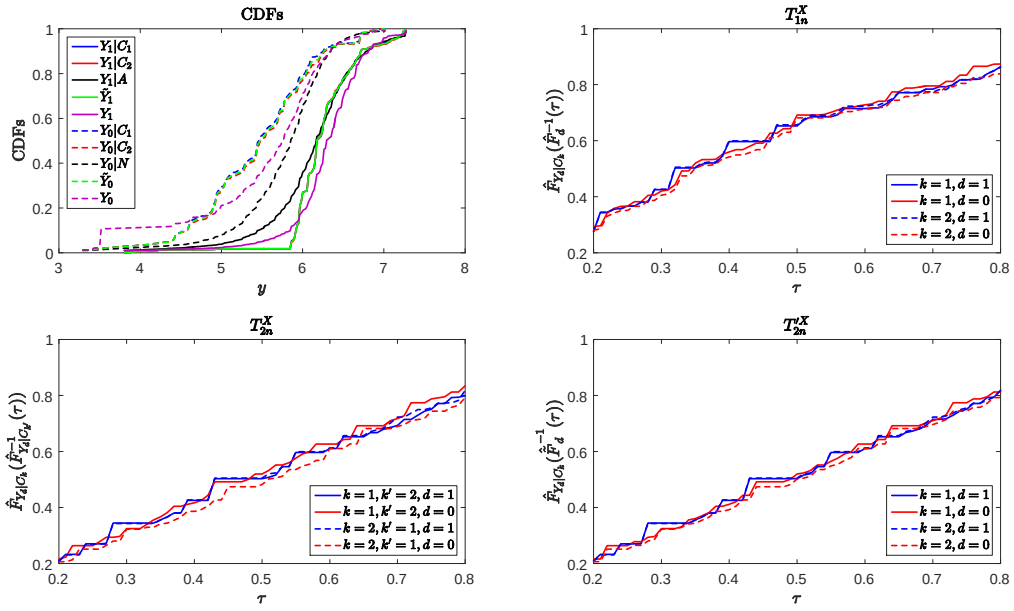

Figure 2: Unconditional Versions of Various Forms of T_{1n}^X, T_{2n}^X and $T_{2n}^{\prime X}$

high (e.g., more chances due to better family backgrounds), it is not necessary for them to go to college. (iii) As a weighted average of $F_{Y_d|\mathcal{C}_1}$ and $F_{Y_d|\mathcal{C}_2}$, $\widetilde{F}_{Y_d}$ stays between $F_{Y_d|\mathcal{C}_1}$ and $F_{Y_d|\mathcal{C}_2}$, but F_1^* does not stay among $F_{Y_1|\mathcal{A}}, F_{Y_1|\mathcal{C}_1}$ and $F_{Y_1|\mathcal{C}_2}$, and F_0^* does not stay among $F_{Y_0|\mathcal{C}_1}$, $F_{Y_0|\mathcal{C}_2}$ and $F_{Y_1|\mathcal{N}}$. Nevertheless, $\widetilde{F}_{Y_1}$ and F_1^* stochastically dominate $\widetilde{F}_{Y_0}$ and F_0^*, respectively. Second, from the other three panels, we can see that $\widehat{F}_{Y_1|\mathcal{C}_k}\left(\widehat{F}_1^{-1}(\tau)\right)$ and $\widehat{F}_{Y_0|\mathcal{C}_k}\left(\widehat{F}_0^{-1}(\tau)\right)$ for $k=1,2$, $\widehat{F}_{Y_1|\mathcal{C}_1}\left(\widehat{F}_{Y_1|\mathcal{C}_2}^{-1}(\tau)\right)$ and $\widehat{F}_{Y_0|\mathcal{C}_1}\left(\widehat{F}_{Y_0|\mathcal{C}_2}^{-1}(\tau)\right)$, $\widehat{F}_{Y_1|\mathcal{C}_2}\left(\widehat{F}_{Y_1|\mathcal{C}_1}^{-1}(\tau)\right)$ and $\widehat{F}_{Y_0|\mathcal{C}_2}\left(\widehat{F}_{Y_0|\mathcal{C}_1}^{-1}(\tau)\right)$, $\widehat{F}_{Y_1|\mathcal{C}_k}\left(\widehat{\widetilde{F}}_1^{-1}(\tau)\right)$ and $\widehat{F}_{Y_0|\mathcal{C}_k}\left(\widehat{\widetilde{F}}_0^{-1}(\tau)\right)$ for $k=1,2$, are all close to each other, which explains why we cannot reject the null.

10 Conclusion

In this paper, we test the conditional rank similarity assumption of CH which is a key identification assumption for the IV-QRE. Different from the unconditional rank similarity test, no covariates are required here. We test this assumption in the framework of HV, and consider three cases, covering discrete/continuous instruments with/without covariates. For each case, we propose two tests and two extensions, and we also suggest to use the bootstrap to obtain critical values.

There are some problems not covered in this paper. First, our simulation and application concentrate on the discrete instrument case, and more simulation studies and applications on the continuous instrument case would provide a more complete picture on the performance of our tests. Second, one key assumption of this paper is the monotonicity assumption of the participation equation. This assumption facilitates our testing procedure but is not required for the consistency of IV-QRE. One possible test when this assumption is relaxed is the Hausman-type test: use two different sets of instruments to construct two inverse quantile regression estimators of QTE as in Chernozhukov and Hansen (2006) and use their difference as an indicator of violation of the null. Third, we did not discuss the power optimality of our testing procedure. For example,

in the formulation of T_1 and T_2, it is unknown which w_k and $\mu_k(\tau)$ would provide the optimal power. In T'_{1n} and T'_{2n}, we propose some simple procedures to improve power; however, the optimal ω_k, $k = 1, \cdots, K-2$, and the optimal hub depend on not only $\widehat{\omega}$ but $\lambda_{ki}^{(1)}$ and the correlation structure of $\varepsilon_{ki}^{(1)}$ for T'_{1n} and $\lambda_{ki}^{(2)}$ and the correlation structure of $\varepsilon_{ki}^{(2)}$ for T'_{2n}, and seem hard to develop.

References

Abadie, A., 2002, Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models, *Journal of the American Statistical Association*, 97, 284-292.

Abadie, A., 2003, Semiparametric Instrumental Variable Estimation of Treatment Response Models, *Journal of Econometrics*, 113, 231-263.

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.

Abbring, J.H. and J.J. Heckman, 2007, Econometric Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, and General Equilibrium Policy Evaluation, *Handbook of Econometrics*, Vol. 6B, J.J. Heckman and E.E. Leamer, eds., Amsterdam: Elsevier Science B.V., Ch. 72, 5145-5303.

Angrist, J.D. and G.W. Imbens, 1995, Two-Stage Least Squares Estimation of Average Causal Effects in Models With Variable Treatment Intensity, *Journal of the American Statistical Association*, 90, 431-442.

Angrist, J.D. and A.B. Krueger, 1991, Does Compulsory School Attendance Affect Schooling and Earnings, *Quarterly Journal of Economics*, 106, 979-1014.

Athey, S. and G.W. Imbens, 2006, Identification and Inference in Nonlinear Difference-In-Differences Models, *Econometrica*, 74, 431–497.

Bickel, P. and J.-J. Ren, 2001, The Bootstrap in Hypothesis Testings, *Lecture Notes - Monograph Series*, 36, 91-112.

Bierens, H.J. and W. Ploberger, 1997, Asymptotic Theory of Integrated Conditional Moment Tests, *Econometrica*, 65, 1129-1151.

Blundell, R. and J.L. Horowitz, 2007, A Non-Parametric Test of Exogeneity, *Review of Economic Studies*, 74, 1035-1058.

Breusch, T.S., 1986, Hypothesis Testing in Unidentified Models, *Review of Economic Studies*, 53, 635-651.

Carneiro, P. and S. Lee, 2009, Estimating Distributions of Potential Outcomes Using Local Instrumental Variables with an Application to Changes in College Enrollment and Wage Inequality, *Journal of Econometrics*, 149, 191-209.

Chernozhukov, V., 2005, Extremal Quantile Regression, *Annals of Statistics*, 33, 806-839.

Chernozhukov, V. and I. Fernández-Val, 2011, Inference for Extremal Conditional Quantile Models, with an Application to Market and Birthweight Risks, *Review of Economic Studies*, 78, 559-589.

Chernozhukov, V., I. Fernández-Val and A. Galichon, 2010, Quantile and Probability Curves without Crossing, *Econometrica*, 78, 1093-1125.

Chernozhukov, V., I. Fernández-Val and B. Melly, 2013, Inference on Counterfactual Distributions, *Econometrica*, 81, 2205–2268.

Chernozhukov, V. and C. Hansen, 2005, An IV Model of Quantile Treatment Effects, *Econometrica*, 73, 245-261.

Chernozhukov, V. and C. Hansen, 2006, Instrumental Quantile Regression Inference for Structural and Treatment Effect Models, *Journal of Econometrics*, 132, 491-525.

Chernozhukov, V. and C. Hansen, 2013, Quantile Models with Endogeneity, *Annual Review of Economics*, 5, 57-81.

Dawid, A.P., 1979, Conditional Independence in Statistical Theory (with discussion), *Journal of the Royal Statistical Society, Series B*, 41, 1-31.

Donald, S.G. and Y.-C. Hsu, 2014, Estimation and Inference for Distribution Functions and Quantile Functions in Treatment Effect Models, *Journal of Econometrics*, 178, 383-397.

Dong, Y.-Y. and S. Shen, 2015, Testing for Rank Invariance or Similarity in Program Evaluation: The Effect of Training on Earnings Revisited, mimeo, UC-Irvine.

Firpo, S., 2007, Efficient Semiparametric Estimation of Quantile Treatment Effects, *Econometrica*, 259-276.

Foresi, S. and F. Peracchi, 1995, The Conditional Distribution of Excess Returns: An Empirical Analysis, *Journal of the American Statistical Association*, 90, 451-466.

Frandsen, B.R. and L.J. Lefgren, 2015, Testing Rank Similarity, mimeo, Brigham Young.

Frölich, M., 2007, Nonparametric IV Estimation of Local Average Treatment Effects with Covariates, *Journal of Econometrics*, 139, 35-75.

Kitagawa, T., 2015, A Test for Instrument Validity, *Econometrica*, 83, 2043-2063.

Heckman, J.J. and D. Schmierer, 2010, Tests of Hypotheses Arising in the Correlated Random Coefficient Model, *Economic Modelling*, 27, 1355-1367.

Heckman, J.J., D. Schmierer and S. Urzua, 2010, Testing the Correlated Random Coefficient Model, *Journal of Econometrics*, 158, 177-203.

Heckman, J.J. and E.J. Vytlacil (HV), 2005, Structural Equations, Treatment Effects, and Econometric Policy Evaluation, *Econometrica*, 73, 669-738.

Horowitz, J.L., 2006, Testing a Parametric Model Against a Nonparametric Alternative with Identification through Instrumental Variables, *Econometrica*, 74, 521-538.

Imbens, G.W. and J.D. Angrist (IA), 1994, Identification and Estimation of Local Average Treatment Effects, *Econometrica*, 62, 467-475.

Imbens, G.W. and Rubin, D.B., 1997, Estimating Outcome Distributions for Compliers in Instrumental Variables Models, *Review of Economic Studies*, 64, 555-574.

Kim, J.H. and B. Park, 2016, Testing for Rank Similarity in Semiparametric Models, mimeo, UNC.

van der Vaart, A.W. and J.A. Wellner, 1996, *Weak Convergence and Empirical Processes: With Applications to Statistics*, New York: Springer.

Vuong, Q. and Xu, H., 2016, Counterfactual Mapping and Individual Treatment Effects in Nonseparable Models with Discrete Endogeneity, *Quantitative Economics*, forthcoming.

Vytlacil, E.J., 2002, Independence, Monotonicity, and Latent Index Models: An Equivalence Result, *Econometrica*, 70, 331-341.

Wüthrich, K., 2015a, Semiparametric Estimation of Quantile Treatment Effects with Endogeneity, mimeo, Bern.

Wüthrich, K., 2015b, A Comparison of Two Quantile Models with Endogeneity, mimeo, Bern.

Yu, P., 2014a, Marginal Quantile Treatment Effect, mimeo, HKU.

Yu, P., 2014b, The Bootstrap in Threshold Regression, *Econometric Theory*, 30, 676-714.

Yu, P., 2015a, Testing Unconditional Rank Preservation Under Unconfoundedness, mimeo, HKU.

Yu, P., 2015b, Understanding the Identification of Conditional and Unonditional Local Treatment Effects, mimeo, HKU.

Yu, P., 2016, Treatment Effects Estimators Under Misspecification, mimeo, HKU.

Supplementary Material S.1

We first collect the notations that will be used in the following proofs. $\mathbb{G}_n(f(y)) = \sqrt{n}(\mathbb{P}_n - P)f(y)$ with $\mathbb{P}_n$ being the empirical measure is the empirical process indexed by $\{f(y)|y \in \mathcal{Y}\}$ for a compact set $\mathcal{Y}$, and $\mathbb{G}(f(y))$ is a zero-mean Gaussian process with the covariance function $E[\mathbb{G}(f(y))\mathbb{G}(f(y'))] = E[f(y)f(y')] - E[f(y)]E[f(y')]$. For a parameter θ, d_θ means its dimension.

Proof of Corollary 1. We show first $Q_{Y_1|V}(F_{Y_0|V}(y_0|v)|v) = Q_1(F_0(y_0))$; $Q_{Y_0|V}(F_{Y_1|V}(y_1|v)|v) = Q_0(F_1(y_1))$ can be similarly proved. Note that $F_{Y_1|V}(Q_1(\tau)|v) = F_{Y_0|V}(Q_0(\tau)|v)$ implies

$$Q_{Y_1|V}(\tau|v) = Q_1\left(F_0\left(Q_{Y_0|V}(\tau|v)\right)\right). \tag{21}$$

Letting $\tau = F_{Y_0|V}(y_0|v)$, we have

$$Q_{Y_1|V}(F_{Y_0|V}(y_0|v)|v) = Q_1\left(F_0\left(Q_{Y_0|V}(F_{Y_0|V}(y_0|v)|v)\right)\right) = Q_1(F_0(y_0))$$

as required, where the last equality uses $Q_{Y_0|V}(F_{Y_0|V}(y_0|v)|v) = y_0$. Conversely, if $Q_{Y_1|V}(F_{Y_0|V}(y_0|v)|v) = Q_1(F_0(y_0))$, then $F_{Y_1|V}(y_1|v) = F_{Y_0|V}(Q_0F_1(y_1)|v)$. So

$$F_{Y_1|V}(Q_1(\tau)|v) = F_{Y_0|V}(Q_0F_1(Q_1(\tau))|v) = F_{Y_0|V}(Q_0(\tau)|v).$$

We next show $F_{Y_1|V}(Q_{Y_1|V}(\tau|v)|v') = F_{Y_0|V}(Q_{Y_0|V}(\tau|v)|v')$ for arbitrary $v' \neq v$. First, $F_{Y_1|V}(Q_1(\tau)|v) = F_{Y_0|V}(Q_0(\tau)|v)$ implies

$$F_{Y_1|V}(y_1|x,v) = F_{Y_0|V}(Q_0(F_1(y_1))|v).$$

So combined with (21),

$$\begin{aligned} F_{Y_1|V}(Q_{Y_1|V}(\tau|v)|v') &= F_{Y_0|V}\left(Q_0\left(F_1(Q_{Y_1|V}(\tau|v))\right)|v'\right) \\ &= F_{Y_0|V}\left(Q_0\left(F_1(Q_1\left(F_0\left(Q_{Y_0|V}(\tau|v)\right)\right))\right)|v'\right) = F_{Y_0|V}(Q_{Y_0|V}(\tau|v)|v'). \end{aligned}$$

Conversely, if $F_{Y_1|V}(Q_{Y_1|V}(\tau|v)|v') = F_{Y_0|V}(Q_{Y_0|V}(\tau|v)|v')$ for arbitrary $v' \neq v$, then $F_{Y_1|V}(y_1|v') = F_{Y_0|V}(Q_{Y_0|V}\left(F_{Y_1|V}(y_1|v)|v\right)|v')$, so

$$F_1(y_1) = \int F_{Y_0|V}(Q_{Y_0|V}\left(F_{Y_1|V}(y_1|v)|v\right)|v')dv' = F_0\left(Q_{Y_0|V}\left(F_{Y_1|V}(y_1|v)|v\right)\right).$$

As a result, $Q_{Y_0|V}(F_{Y_1|V}(y_1|v)|v) = Q_0(F_1(y_1))$, which has been shown to imply $F_{Y_1|V}(Q_1(\tau)|v) = F_{Y_0|V}(Q_0(\tau)|v)$.

Finally, we show that fixed v or arbitrary v does not matter. For this purpose, we need only show that (8) holds for fixed v implies it holds for arbitrary v. Suppose $F_{Y_1|V}(Q_{Y_1|V}(\tau|v)|v') = F_{Y_0|V}(Q_{Y_0|V}(\tau|v)|v')$ for a fixed v and arbitrary $v' \neq v$; we try to show that $F_{Y_1|V}(Q_{Y_1|V}(\tau|v'')|v') = F_{Y_0|V}(Q_{Y_0|V}(\tau|v'')|v')$ for any $v'' \neq v$ and $v' \neq v''$. First, $F_{Y_1|V}(Q_{Y_1|V}(\tau|v)|v') = F_{Y_0|V}(Q_{Y_0|V}(\tau|v)|v')$ implies $F_{Y_1|V}(y_1|v') = F_{Y_0|V}(Q_{Y_0|V}\left(F_{Y_1|V}(y_1|v)|v\right)|v')$ and $Q_{Y_1|V}(\tau|v') = Q_{Y_1|V}(F_{Y_0|V}\left(Q_{Y_0|V}(\tau|v')|v\right)|v)$ for any $v' \neq v$. If $v' \neq v''$ but equals v, then

$$F_{Y_1|V}(Q_{Y_1|V}(\tau|v'')|v) = F_{Y_1|V}(Q_{Y_1|V}(F_{Y_0|V}\left(Q_{Y_0|V}(\tau|v'')|v\right)|v)|v) = F_{Y_0|V}\left(Q_{Y_0|V}(\tau|v'')|v\right).$$

If $v' \neq v''$ and v, then

$$\begin{aligned} F_{Y_1|V}(Q_{Y_1|V}(\tau|v'')|v') &= F_{Y_0|V}(Q_{Y_0|V}\left(F_{Y_1|V}\left(Q_{Y_1|V}(\tau|v'')|v\right)|v\right)|v') \\ &= F_{Y_0|V}(Q_{Y_0|V}\left(F_{Y_1|V}\left(Q_{Y_1|V}(F_{Y_0|V}\left(Q_{Y_0|V}(\tau|v'')|v\right)|v)|v\right)|v\right)|v') = F_{Y_0|V}\left(Q_{Y_0|V}(\tau|v'')|v'\right). \end{aligned}$$

■

Proof of Proposition 1. Since Z is independent of (Y_1, Y_0, V), we can specify the distributions of (Y_1, Y_0, V) and Z separately, where the distribution of Z is implied by the observable data distribution. Specify the joint distribution of (Y_1, Y_0, V) as

$$F_{Y_1,Y_0,V}(y_1, y_0, v) = F_{Y_1|V}(y_1) F_{Y_0|V}(y_0) v,$$

i.e., Y_1 and Y_0 are conditionally independent given V, where $F_V(v) = v$ because the distribution of V is uniform on $(0,1)$. The joint data distribution of (Y, D, Z) is

$$F_{Y,D,Z}(y, d, z) = F_{Y,D|Z}(y|d, z) F_Z(z) = \begin{cases} \int_0^{p(z)} F_{Y_1|V}(y|v)dv F_Z(z), & \text{if } d = 1, \\ \int_{p(z)}^1 F_{Y_0|V}(y|v)dv F_Z(z), & \text{if } d = 0, \end{cases}$$

in the framework (5) and under Assumption M. In summary, we need to specify $F_{Y_1|V}(y_1)$ and $F_{Y_0|V}(y_0)$ such that

$$\begin{aligned} \int_0^{p(z)} F_{Y_1|V}(y|v)dv &= F_{Y,D|Z}(y, 1|z), y \in \mathcal{S}_1, z \in \{z_1, \cdots, z_K\}, \\ \int_{p(z)}^1 F_{Y_0|V}(y|v)dv &= F_{Y,D|Z}(y, 0|z), y \in \mathcal{S}_0, z \in \{z_1, \cdots, z_K\}, \\ \frac{1}{p_{k+1} - p_k} \int_{p_k}^{p_{k+1}} F_{Y_d|V}(Q_d(\tau)|v) dv &= F_{Y_d|\mathcal{C}_k}(Q_d(\tau)), \tau \in (0,1), k = 1, \cdots, K-1, \\ \text{with } Q_d^{-1}(y) &= \int_0^1 F_{Y_d|V}(y|v)dv = \sum_{k=0}^{K} \int_{p_k}^{p_{k+1}} F_{Y_d|V}(y|v)dv, y \in \mathcal{S}_d, \\ F_{Y_1|V}(Q_1(\tau)|v) &= F_{Y_0|V}(Q_0(\tau)|v), \tau \in (0,1), v \in (0,1), \\ \text{with } F_{Y_1|\mathcal{C}_k}(Q_1(\tau)) &= F_{Y_0|\mathcal{C}_k}(Q_0(\tau)), \tau \in (0,1), k = 1, \cdots, K-1, \end{aligned}$$

where p_k, $k = 1, \cdots, K$, is implied by the the observable data distribution, the first and second equations are to match the joint distribution of (Y, D, Z), the third equation is to match $F_{Y_d|\mathcal{C}_k}(Q_d(\tau))$ which is exactly the definition of $F_{Y_d|\mathcal{C}_k}(\cdot)$ in our framework and contains no further information, and the last equation is to match H_0 with (9) holds. The question is whether we can find some $F_{Y_1|V}(y_1)$ and $F_{Y_0|V}(y_0)$ to satisfy these equations simultaneously. Obviously, we need only find feasible

$$\int_{p_k}^{p_{k+1}} F_{Y_d|V}(y|v)dv, d = 0, 1, k = 0, 1, \cdots, K.$$

$\int_{p_k}^{p_{k+1}} F_{Y_d|V}(y|v)dv$, $k = 1, \cdots, K-1$, is determined by $F_{Y,D|Z}(y, d|z_{k+1}) - F_{Y,D|Z}(y, d|z_k)$, $\int_0^{p_1} F_{Y_1|V}(y|v)dv$ is determined by $F_{Y,D|Z}(y, 1|z_1)$, and $\int_{p_K}^1 F_{Y_0|V}(y|v)dv$ is determined by $F_{Y,D|Z}(y, 0|z_K)$, so the only remaining freedom is $\int_0^{p_1} F_{Y_0|V}(y|v)dv$ and $\int_{p_K}^1 F_{Y_1|V}(y|v)dv$ or $F_{Y_0|\mathcal{A}}(y_0)$ and $F_{Y_1|\mathcal{N}}(y_1)$.

In summary, the question is whether there exist $F_{U|\mathcal{C}_k}, k = 1, \cdots, K-1$, $F_{U|\mathcal{A}}$, $F_{U|\mathcal{N}}$, $F_{Y_1|\mathcal{N}}$ and $F_{Y_0|\mathcal{A}}$

such that

$$\begin{aligned}
\boxed{F_{U|\mathcal{C}_k}}\left(F_1\left(y_1\right)\right) &= F_{Y_1|\mathcal{C}_k}\left(y_1\right), k=1,\cdots,K-1,\\
\boxed{F_{U|\mathcal{C}_k}}\left(F_0\left(y_0\right)\right) &= F_{Y_0|\mathcal{C}_k}\left(y_0\right), k=1,\cdots,K-1,\\
\boxed{F_{U|\mathcal{A}}}\left(F_1\left(y_1\right)\right) &= F_{Y_1|\mathcal{A}}\left(y_1\right), \boxed{F_{U|\mathcal{A}}}\left(F_0\left(y_0\right)\right) = \boxed{F_{Y_0|\mathcal{A}}}\left(y_0\right),\\
\boxed{F_{U|\mathcal{N}}}\left(F_0\left(y_0\right)\right) &= F_{Y_0|\mathcal{N}}\left(y_0\right), \boxed{F_{U|\mathcal{N}}}\left(F_1\left(y_1\right)\right) = \boxed{F_{Y_1|\mathcal{N}}}\left(y_1\right)
\end{aligned}$$

with

$$\begin{aligned}
F_1\left(y_1\right) &= p_1 F_{Y_1|\mathcal{A}}\left(y_1\right) + \sum_{k=1}^{K-1}\left(p_{k+1}-p_k\right)F_{Y_1|\mathcal{C}_k}\left(y_1\right) + \left(1-p_K\right)\boxed{F_{Y_1|\mathcal{N}}}\left(y_1\right),\\
F_0\left(y_0\right) &= p_1\boxed{F_{Y_0|\mathcal{A}}}\left(y_0\right) + \sum_{k=1}^{K-1}\left(p_{k+1}-p_k\right)F_{Y_0|\mathcal{C}_k}\left(y_0\right) + \left(1-p_K\right)F_{Y_0|\mathcal{N}}\left(y_0\right).
\end{aligned}$$

Equivalently, we need to find $F_{U|\mathcal{C}_k}, k=1,\cdots,K-1$, $F_{U|\mathcal{A}}$, $F_{U|\mathcal{N}}$, $F_{Y_1|\mathcal{N}}$ and $F_{Y_0|\mathcal{A}}$ such that for any data distribution of $F_{Y_d|\mathcal{C}_k}$, $k=1,\cdots,K-1$, $F_{Y_1|\mathcal{A}}$ and $F_{Y_0|\mathcal{N}}$,

$$\begin{aligned}
p_1 F_{Y_1|\mathcal{A}}\left(y_1\right) + \sum_{k=1}^{K-1}\left(p_{k+1}-p_k\right)F_{Y_1|\mathcal{C}_k}\left(y_1\right) + \left(1-p_K\right)\boxed{F_{Y_1|\mathcal{N}}}\left(y_1\right) &= \boxed{F^{-1}_{U|\mathcal{C}_k}}\left(F_{Y_1|\mathcal{C}_k}\left(y_1\right)\right), k=1,\cdots,K-1,\\
&= \boxed{F^{-1}_{U|\mathcal{A}}}\left(F_{Y_1|\mathcal{A}}\left(y_1\right)\right)\\
&= \boxed{F^{-1}_{U|\mathcal{N}}}\left(\boxed{F_{Y_1|\mathcal{N}}}\left(y_1\right)\right), y_1\in\mathcal{S}_1,\\
p_1\boxed{F_{Y_0|\mathcal{A}}}\left(y_0\right) + \sum_{k=1}^{K-1}\left(p_{k+1}-p_k\right)F_{Y_0|\mathcal{C}_k}\left(y_0\right) + \left(1-p_K\right)F_{Y_0|\mathcal{N}}\left(y_0\right) &= \boxed{F^{-1}_{U|\mathcal{C}_k}}\left(F_{Y_0|\mathcal{C}_k}\left(y_0\right)\right), k=1,\cdots,K-1,\\
&= \boxed{F^{-1}_{U|\mathcal{A}}}\left(\boxed{F_{Y_0|\mathcal{A}}}\left(y_0\right)\right)\\
&= \boxed{F^{-1}_{U|\mathcal{N}}}\left(F_{Y_0|\mathcal{N}}\left(y_0\right)\right), y_0\in\mathcal{S}_0.
\end{aligned}$$

Take any $k=1,\cdots,K-1$ and $\tau\in(0,1)$, we must allow $F^{-1}_{U|\mathcal{C}_k}\left(\tau\right)$ to satisfy two equations

$$\begin{aligned}
p_1 F_{Y_1|\mathcal{A}}\left(F^{-1}_{Y_1|\mathcal{C}_k}\left(\tau\right)\right) + \sum_{j=1}^{K-1}\left(p_{j+1}-p_j\right)F_{Y_1|\mathcal{C}_j}\left(F^{-1}_{Y_1|\mathcal{C}_k}\left(\tau\right)\right) + \left(1-p_K\right)\boxed{F_{Y_1|\mathcal{N}}}\left(F^{-1}_{Y_1|\mathcal{C}_k}\left(\tau\right)\right) &= \boxed{F^{-1}_{U|\mathcal{C}_k}}\left(\tau\right),\\
p_1\boxed{F_{Y_0|\mathcal{A}}}\left(F^{-1}_{Y_0|\mathcal{C}_k}\left(\tau\right)\right) + \sum_{j=1}^{K-1}\left(p_{j+1}-p_j\right)F_{Y_0|\mathcal{C}_j}\left(F^{-1}_{Y_0|\mathcal{C}_k}\left(\tau\right)\right) + \left(1-p_K\right)F_{Y_0|\mathcal{N}}\left(F^{-1}_{Y_0|\mathcal{C}_k}\left(\tau\right)\right) &= \boxed{F^{-1}_{U|\mathcal{C}_k}}\left(\tau\right),
\end{aligned}$$

$F^{-1}_{U|\mathcal{A}}$ to satisfy

$$\begin{aligned}
p_1\tau + \sum_{j=1}^{K-1}\left(p_{j+1}-p_j\right)F_{Y_1|\mathcal{C}_j}\left(F^{-1}_{Y_1|\mathcal{A}}\left(\tau\right)\right) + \left(1-p_K\right)\boxed{F_{Y_1|\mathcal{N}}}\left(F^{-1}_{Y_1|\mathcal{A}}\left(\tau\right)\right) &= \boxed{F^{-1}_{U|\mathcal{A}}}\left(\tau\right),\\
p_1\tau + \sum_{j=1}^{K-1}\left(p_{j+1}-p_j\right)F_{Y_0|\mathcal{C}_j}\left(\boxed{F^{-1}_{Y_0|\mathcal{A}}}\left(\tau\right)\right) + \left(1-p_K\right)F_{Y_0|\mathcal{N}}\left(\boxed{F^{-1}_{Y_0|\mathcal{A}}}\left(\tau\right)\right) &= \boxed{F^{-1}_{U|\mathcal{A}}}\left(\tau\right),
\end{aligned}$$

and $F_{U|\mathcal{N}}^{-1}$ to satisfy

$$\begin{aligned} p_1 F_{Y_1|\mathcal{A}}\left(\boxed{F_{Y_1|\mathcal{N}}^{-1}}(\tau)\right)+\sum_{j=1}^{K-1}(p_{j+1}-p_j)\,F_{Y_1|\mathcal{C}_j}\left(\boxed{F_{Y_1|\mathcal{N}}^{-1}}(\tau)\right)+(1-p_K)\,\tau &= \boxed{F_{U|\mathcal{N}}^{-1}}(\tau), \\ p_1\boxed{F_{Y_0|\mathcal{A}}}\left(F_{Y_0|\mathcal{N}}^{-1}(\tau)\right)+\sum_{j=1}^{K-1}(p_{j+1}-p_j)\,F_{Y_0|\mathcal{C}_j}\left(F_{Y_0|\mathcal{N}}^{-1}(\tau)\right)+(1-p_K)\,\tau &= \boxed{F_{U|\mathcal{N}}^{-1}}(\tau) \end{aligned}$$

simultaneously. This is possible if we let $F_{Y_1|V}\left(Q_1\left(\tau\right)|v\right)=F_{Y_0|V}\left(Q_0\left(\tau\right)|v\right)=F_{U|V}\left(\tau|v\right)$ since then $F_{Y_1|\mathcal{A}}\left(F_{Y_1|\mathcal{C}_k}^{-1}\left(\tau\right)\right)=\boxed{F_{Y_0|\mathcal{A}}}\left(F_{Y_0|\mathcal{C}_k}^{-1}\left(\tau\right)\right)=F_{U|\mathcal{A}}\left(F_{U|\mathcal{C}_k}^{-1}\left(\tau\right)\right)$, etc. ■

Proof of Proposition 2. From the proof of Proposition 1, only cross matching is required. ■

Proof of Theorem 2. (i) The test statistic involves the following processes $(\{\widehat{F}_{Y_1|\mathcal{C}_k}\left(y_1\right),\widehat{F}_{Y_0|\mathcal{C}_k}\left(y_0\right)\}_{k=1}^{K-1}$, $\widehat{Q}_1(\tau)$, $\widehat{Q}_0(\tau))$, $\tau\in\mathcal{T}$, $y_1\in\mathcal{Y}_1\equiv\{Q_1(\tau)|\tau\in\mathcal{T}\}$ and $y_0\in\mathcal{Y}_0\equiv\{Q_0(\tau)|\tau\in\mathcal{T}\}$, where $(\widehat{Q}_1(\tau),\widehat{Q}_0(\tau))$ involves $(\widehat{F}_{Y_1|\mathcal{A}}(y_1),\{\widehat{F}_{Y_1|\mathcal{C}_k}(y_1),\ \widehat{F}_{Y_0|\mathcal{C}_k}\left(y_0\right)\}_{k=1}^{K-1},\widehat{F}_{Y_0|\mathcal{N}}(y_0),\{\widehat{p}_l,\widehat{q}_l\}_{l=1}^{K})$ with $\widehat{q}_l=n_l/n$. Lemma 1 derives the weak limit of these components. Now, since $(\left\{F_{Y_1|\mathcal{C}_k}\left(y_1\right),F_{Y_0|\mathcal{C}_k}\left(y_0\right)\right\}_{k=1}^{K-1},F_1(\tau),F_0(\tau))$ is a Hadamard differentiable map of $(F_{Y_1|\mathcal{A}}(y_1),\{F_{Y_1|\mathcal{C}_k}(y_1),\ F_{Y_0|\mathcal{C}_k}\left(y_0\right)\}_{k=1}^{K-1},F_{Y_0|\mathcal{N}}(y_0),\{p_l,q_l\}_{l=1}^{K})$, by Lemma 1 and the functional Delta method, we have

$$\sqrt{n}\begin{pmatrix}\widehat{F}_{Y_1|\mathcal{C}_k}\left(y_1\right)-F_{Y_1|\mathcal{C}_k}\left(y_1\right)\\ \widehat{F}_{Y_0|\mathcal{C}_k}\left(y_0\right)-F_{Y_0|\mathcal{C}_k}\left(y_0\right)\\ \widehat{F}_1(y_1)-F_1(y_1)\\ \widehat{F}_0(y_0)-F_0(y_0)\end{pmatrix}\rightsquigarrow\Psi_y^{(2)}\left(\Psi_y^{(1)}\left(\mathbb{G}\left(\varphi_y(W)\right)\right)\right)\text{ in }\ell^\infty\left(\mathcal{Y}\right)^{2K}, \tag{22}$$

where $\Psi_y^{(1)}\left(\mathbb{G}\left(\varphi_y(W)\right)\right)\equiv\mathbb{G}(\psi_1^{\mathcal{A}}(W,y_1),\{\psi_1^k(W,y_1),\psi_0^k(W,y_0)\}_{k=1}^{K-1},\psi_0^{\mathcal{N}}(W,y_0),\{\phi_l\left(D,Z\right),\varphi_q\left(Z\right)\}_{l=1}^{K})$ is defined in Lemma 1. The operation of the linear map $\Psi_y^{(2)}\left(\cdot\right):C\left(\mathcal{Y}\right)^{4K}\to\ell^\infty\left(\mathcal{Y}\right)^{2K}$ on $\alpha\left(y\right)=(\alpha_1^{\mathcal{A}}(y_1),$ $\{\alpha_1^k(y_1),\ \alpha_0^k(y_0)\}_{k=1}^{K-1},\ \alpha_0^{\mathcal{N}}(y_0),\ \{\beta_l,\gamma_l\}_{l=1}^{K})'\in C\left(\mathcal{Y}\right)^{4K}$ is explained as follows. The first $2\left(K-1\right)$ elements of $\Psi_y^{(2)}\left(\alpha\left(y\right)\right)$ is straightforward, so we concentrate on the last two elements. First, since $h(p_{k+1})=\frac{\sum_{l=k+1}^{K}q_l\left(p_l-\sum_{\ell=1}^{K}p_\ell q_\ell\right)}{\sum_{l=1}^{K}q_lp_l^2-\left(\sum_{l=1}^{K}q_lp_l\right)^2}$ is a differentiable function of $\{p_l,q_l\}_{l=1}^{K}$, define $\phi_h^{k+1}\left(\beta,\gamma\right)=h_p^{k+1\prime}\beta+h_q^{k+1\prime}\gamma$, where h_p^{k+1} and h_q^{k+1} are the partial derivatives of $h(p_{k+1})$ with respect to p and q respectively, $p=(p_1,\cdots,p_K)'$ and $q=(q_1,\cdots,q_K)'$. Second, since $\widetilde{F}_d\left(y_d\right)=\sum_{k=1}^{K-1}\left(p_{k+1}-p_k\right)h(p_{k+1})F_{Y_d|\mathcal{C}_k}(y_d)$, define

$$\begin{aligned}\widetilde{\psi}_d\left(\left\{\alpha_d^k(y_d)\right\}_{k=1}^{K-1},\beta,\gamma\right)=&\textstyle\sum_{k=1}^{K-1}\left(p_{k+1}-p_k\right)h(p_{k+1})\alpha_d^k(y_d)+\sum_{k=1}^{K-1}\left(p_{k+1}-p_k\right)F_{Y_d|\mathcal{C}_k}(y_d)\phi_h^{k+1}\left(\beta,\gamma\right)\\&+\textstyle\sum_{k=1}^{K-1}h(p_{k+1})F_{Y_d|\mathcal{C}_k}(y_d)\left(\beta_{k+1}-\beta_k\right)\end{aligned} \tag{23}$$

as a linear map from $C\left(\mathcal{Y}\right)^{3K-1}\to\ell^\infty\left(\mathcal{Y}\right)$. Third, since $F_1^*(y_1)=p_1F_{Y_1|\mathcal{A}}\left(y_1\right)+\sum_{k=1}^{K-1}\left(p_{k+1}-p_k\right)\{F_{Y_1|\mathcal{C}_k}(y_1)(1-\sum_{l=1}^{k}q_l)+F_{Y_0|\mathcal{C}_k}(\widetilde{F}_0^{-1}\widetilde{F}_1\left(y_1\right))\sum_{l=1}^{k}q_l\}+\left(1-p_K\right)F_{Y_0|\mathcal{N}}(\widetilde{F}_0^{-1}\widetilde{F}_1\left(y_1\right))$, the second-to-last component of $\Psi_y^{(2)}\left(\alpha\left(y\right)\right)$ is

$$\begin{aligned}&p_1\alpha_1^{\mathcal{A}}(y_1)+\textstyle\sum_{k=1}^{K-1}\left(p_{k+1}-p_k\right)\left(1-\sum_{l=1}^{k}q_l\right)\alpha_1^k(y_1)+\sum_{k=1}^{K-1}\left(p_{k+1}-p_k\right)\left(\sum_{l=1}^{k}q_l\right)\\&\cdot\left[\alpha_0^k(\widetilde{F}_0^{-1}\widetilde{F}_1\left(y_1\right))-\frac{f_{Y_0|\mathcal{C}_k}(\widetilde{F}_0^{-1}\widetilde{F}_1\left(y_1\right))}{\widetilde{f}_0\left(\widetilde{F}_0^{-1}\widetilde{F}_1\left(y_1\right)\right)}\left(\widetilde{\psi}_0\left(\left\{\alpha_0^k\left(\widetilde{F}_0^{-1}\widetilde{F}_1\left(y_1\right)\right)\right\}_{k=1}^{K-1},\beta,\gamma\right)-\widetilde{\psi}_1\left(\left\{\alpha_1^k\left(y_1\right)\right\}_{k=1}^{K-1},\beta,\gamma\right)\right)\right]\\&+\left(1-p_K\right)\left[\alpha_0^{\mathcal{N}}(\widetilde{F}_0^{-1}\widetilde{F}_1\left(y_1\right))-\frac{f_{Y_0|\mathcal{N}}(\widetilde{F}_0^{-1}\widetilde{F}_1\left(y_1\right))}{\widetilde{f}_0\left(\widetilde{F}_0^{-1}\widetilde{F}_1\left(y_1\right)\right)}\left(\widetilde{\psi}_0\left(\left\{\alpha_0^k(\widetilde{F}_0^{-1}\widetilde{F}_1\left(y_1\right))\right\}_{k=1}^{K-1},\beta,\gamma\right)-\widetilde{\psi}_1\left(\left\{\alpha_1^k\left(y_1\right)\right\}_{k=1}^{K-1},\beta,\gamma\right)\right)\right]\end{aligned}$$

$$+\beta_1 F_{Y_1|\mathcal{A}}(y_1) + \sum_{k=1}^{K-1}(\beta_{k+1}-\beta_k) F_{Y_1|\mathcal{C}_k}(y_1)\left(1-\sum_{l=1}^{k} q_l\right) - \sum_{k=1}^{K-1}(p_{k+1}-p_k)\left\{F_{Y_1|\mathcal{C}_k}(y_1)\sum_{l=1}^{k}\gamma_l\right.$$
$$\left.+F_{Y_0|\mathcal{C}_k}(\widetilde{F}_0^{-1}\widetilde{F}_1(y_1))\sum_{l=1}^{k}\gamma_l\right\} + (1-\beta_K) F_{Y_0|\mathcal{N}}(\widetilde{F}_0^{-1}\widetilde{F}_1(y_1)),$$

where the f function is the pdf of the corresponding F function. The last component of $\Psi_y^{(2)}(\alpha(y))$ can be similarly derived. Note that since $p_1, (p_{k+1}-p_k)$ and $(1-p_K)$ will offset the denominator of $F_{Y_1|\mathcal{A}}(y_1)$, $F_{Y_d|\mathcal{C}_k}(y_d)$ and $F_{Y_0|\mathcal{N}}$ in $\widetilde{F}_d(y_d)$ and $\widetilde{F}_d(y_d)$, some terms involving β will offset each other.

Next, since $\left\{F_{Y_1|\mathcal{C}_k}\left(F_1^{-1}(\tau)\right), F_{Y_0|\mathcal{C}_k}\left(F_0^{-1}(\tau)\right)\right\}_{k=1}^{K-1}$ is a Hadamard differentiable map of $(\{F_{Y_1|\mathcal{C}_k}(y_1), F_{Y_0|\mathcal{C}_k}(y_0)\}_{k=1}^{K-1}, F_1(\tau), F_0(\tau))$, we can apply the functional Delta method again to have

$$\sqrt{n}\begin{pmatrix} \widehat{F}_{Y_1|\mathcal{C}_k}\left(\widehat{F}_1^{-1}(\tau)\right) - F_{U_1|\mathcal{C}_k}(\tau) \\ \widehat{F}_{Y_0|\mathcal{C}_k}\left(\widehat{F}_0^{-1}(\tau)\right) - F_{U_0|\mathcal{C}_k}(\tau) \end{pmatrix} \rightsquigarrow \Psi_\tau^{(3)}\left(\Psi_y^{(2)}\left(\Psi_y^{(1)}\left(\mathbb{G}\left(\varphi_y(W)\right)\right)\right)\right)\Big|_{y_1=F_1^{-1}(\tau), y_0=F_0^{-1}(\tau)} \text{ in } \ell^\infty(\mathcal{T})^{2(K-1)}.$$

Here, for $\alpha(y) = \left(\{\alpha_1^k(y_1), \alpha_0^k(y_0)\}_{k=1}^{K-1}, \alpha_1(y_1), \alpha_0(y_0)\right)' \in C(\mathcal{Y})^{2K}$, the linear map $\Psi_\tau^{(3)}(\cdot) : C(\mathcal{Y})^{2K} \to \ell^\infty(\mathcal{T})^{2(K-1)}$ evaluated at $\alpha(y)$ is defined as follows. The term associated with $\widehat{F}_{Y_1|\mathcal{C}_k}\left(\widehat{F}_1^{-1}(\tau)\right)$ is

$$\alpha_1^k\left(F_1^{-1}(\tau)\right) - \frac{f_{Y_1|\mathcal{C}_k}\left(F_1^{-1}(\tau)\right)}{f_1\left(F_1^{-1}(\tau)\right)}\alpha_1(F_1^{-1}(\tau)),$$

and the term associated with $\widehat{F}_{Y_0|\mathcal{C}_k}\left(\widehat{F}_0^{-1}(\tau)\right)$ is

$$\alpha_0^k\left(F_0^{-1}(\tau)\right) - \frac{f_{Y_0|\mathcal{C}_k}\left(F_0^{-1}(\tau)\right)}{f_0\left(F_0^{-1}(\tau)\right)}\alpha_0(F_0^{-1}(\tau)).$$

Finally, $T_1 = \sum_{k=1}^{K-1}\int_{\mathcal{T}}\left[F_{Y_1|\mathcal{C}_k}(Q_1(\tau)) - F_{Y_0|\mathcal{C}_k}(Q_0(\tau))\right]^2 d\tau$ is a continuous functional of $\{F_{Y_1|\mathcal{C}_k}\left(F_1^{-1}(\tau)\right), F_{Y_0|\mathcal{C}_k}\left(F_0^{-1}(\tau)\right)\}_{k=1}^{K-1}$, so by the continuous mapping theorem,

$$\begin{aligned} nT_{1n} &\equiv \sum_{k=1}^{K-1}\int_{\mathcal{T}} Z_{kn}^{(1)}(\tau)^2 d\tau \\ &\rightsquigarrow \sum_{k=1}^{K-1}\int_{\mathcal{T}}\left[\begin{array}{c} \Psi_{k1\tau}^{(3)}\left(\Psi_{F_1^{-1}(\tau)}^{(2)}\left(\Psi_{F_1^{-1}(\tau)}^{(1)}\left(\mathbb{G}\left(\varphi_{F_1^{-1}(\tau)}(W)\right)\right)\right)\right) \\ -\Psi_{k2\tau}^{(3)}\left(\Psi_{F_0^{-1}(\tau)}^{(2)}\left(\Psi_{F_0^{-1}(\tau)}^{(1)}\left(\mathbb{G}\left(\varphi_{F_0^{-1}(\tau)}(W)\right)\right)\right)\right) \end{array}\right]^2 d\tau \\ &\equiv \sum_{k=1}^{K-1}\int_{\mathcal{T}} Z_k^{(1)}(\tau)^2 d\tau, \end{aligned}$$

where $\Psi_\tau^{(3)} = \left\{\Psi_{k\tau}^{(3)}\right\}_{k=1}^{K-1}$ and $\Psi_{k\tau}^{(3)} = \left(\Psi_{k1\tau}^{(3)}, \Psi_{k2\tau}^{(3)}\right)$, the index of y in $\Psi_{k1\tau}^{(3)}$ is replaced by $F_1^{-1}(\tau)$ since only processes indexed by y_1 are involved, and similarly for the index of y in $\Psi_{k2\tau}^{(3)}$. We can equivalently express $\Psi_{k1\tau}^{(3)}\left(\Psi_{F_1^{-1}(\tau)}^{(2)}\left(\Psi_{F_1^{-1}(\tau)}^{(1)}\left(\mathbb{G}\left(\varphi_{F_1^{-1}(\tau)}(W)\right)\right)\right)\right)$ as $\mathbb{G}\left(\Psi_{k1\tau}^{(3)}\left(\Psi_{F_1^{-1}(\tau)}^{(2)}\left(\Psi_{F_1^{-1}(\tau)}^{(1)}\left(\varphi_{F_1^{-1}(\tau)}(W)\right)\right)\right)\right)$; similarly for $\Psi_{k2\tau}^{(3)}\left(\Psi_{F_0^{-1}(\tau)}^{(2)}\left(\Psi_{F_0^{-1}(\tau)}^{(1)}\left(\mathbb{G}\left(\varphi_{F_0^{-1}(\tau)}(W)\right)\right)\right)\right)$.

Since $Z_k(\tau)$, $k = 1, \cdots, K-1$, are correlated Gaussian processes, we must extend Mercer's theorem (see, e.g., Lemma 1 of Bierens and Ploberger (1997)) to express $\sum_{k=1}^{K-1}\int_{\mathcal{T}} Z_k(\tau)^2 d\tau$ as a mixed chi-square

distribution. First, by Mercer's theorem

$$\int_{\mathcal{T}} Z_k^{(1)}(\tau)^2 \, d\tau \sim \sum_{i=1}^{\infty} \lambda_{ki}^{(1)} \varepsilon_{ki}^{(1)2},$$

where $\varepsilon_{ki}^{(1)}$'s and $\lambda_{ki}^{(1)}$'s are defined in the theorem. For $k \neq l$, we must study the correlation between $\varepsilon_{ki}^{(1)}$ and $\varepsilon_{lj}^{(1)}$. Since

$$\varepsilon_{ki}^{(1)} = \int_{\mathcal{T}} Z_k^{(1)}(\tau) \varphi_i(\tau) \, d\tau \text{ and } \varepsilon_{lj}^{(1)} = \int_{\mathcal{T}} Z_l^{(1)}(\tau) \varphi_j(\tau) \, d\tau,$$

the covariance of $\varepsilon_{ki}^{(1)}$ and $\varepsilon_{lj}^{(1)}$ is

$$E\left[\int_{\mathcal{T}} Z_k^{(1)}(\tau) \varphi_i(\tau) \, d\tau \int_{\mathcal{T}} Z_l^{(1)}(\tau) \varphi_j(\tau) \, d\tau\right] = \int_{\mathcal{T}} \int_{\mathcal{T}} \Sigma_{kl}^{(1)}(\tau_1, \tau_2) \varphi_i(\tau_1) \varphi_j(\tau_2) \, d\tau_1 d\tau_2,$$

where $\Sigma_{kl}^{(1)}(\tau_1, \tau_2) = E\left[Z_k^{(1)}(\tau_1) Z_l^{(1)}(\tau_2)\right]$ need not be a positive semi-definite continuous function on $\mathcal{T} \times \mathcal{T}$, so $\int_{\mathcal{T}} \Sigma_{kl}^{(1)}(\tau_1, \tau_2) \varphi_j(\tau_1) \, d\tau_1$ need not be a multiple of $\varphi_j(\tau_2)$ such that this covariance depends on four indices (k, l, i, j).

(ii) Denote $F_{Y_d|\mathcal{C}_k}$ and F_d under H_1^δ as $F_{Y_d|\mathcal{C}_k}^n$ and F_d^n respectively. Under Assumption LA, we know by Lemma 2.8.7 of VW (p. 174) that

$$\begin{aligned} & \sqrt{n}\left(\widehat{F}_{Y_1|\mathcal{C}_k}\left(\widehat{F}_1^{-1}(\tau)\right) - F_{Y_1|\mathcal{C}_k}^n\left(F_1^{n-1}(\tau)\right)\right) - \sqrt{n}\left(\widehat{F}_{Y_0|\mathcal{C}_k}\left(\widehat{F}_0^{-1}(\tau)\right) - F_{Y_0|\mathcal{C}_k}^n\left(F_0^{n-1}(\tau)\right)\right) \\ = & \sqrt{n}\left(\widehat{F}_{Y_1|\mathcal{C}_k}\left(\widehat{F}_1^{-1}(\tau)\right) - \widehat{F}_{Y_0|\mathcal{C}_k}\left(\widehat{F}_0^{-1}(\tau)\right)\right) - \sqrt{n}\left(F_{Y_1|\mathcal{C}_k}^n\left(F_1^{n-1}(\tau)\right) - F_{Y_0|\mathcal{C}_k}^n\left(F_0^{n-1}(\tau)\right)\right) \rightsquigarrow Z_k(\tau), \end{aligned}$$

so it remains to derive $\sqrt{n}\left(F_{Y_1|\mathcal{C}_k}^n\left(F_1^{n-1}(\tau)\right) - F_{Y_0|\mathcal{C}_k}^n\left(F_0^{n-1}(\tau)\right)\right)$. From Yu (2016),

$$\begin{aligned} \sqrt{n}\left(F_1^{n-1}(\tau) - F_1^{-1}(\tau)\right) & \to -\frac{D_{F_1}(\delta)(\tau)}{f_1\left(F_1^{-1}(\tau)\right)}, \\ \sqrt{n}\left(F_0^{n-1}(\tau) - F_0^{-1}(\tau)\right) & \to -\frac{D_{F_0}(\delta)(\tau)}{f_0\left(F_0^{-1}(\tau)\right)}, \end{aligned}$$

where

$$\begin{aligned} D_{F_1}(\delta)(\tau) = & \frac{\sum_{k=1}^{K-1}(p_{k+1} - p_k) F_{p(Z)}(p_k) f_{U_0|\mathcal{C}_k}(\tau) + (1 - p_K) f_{U_0|\mathcal{N}}(\tau)}{\sum_{k=1}^{K-1}(p_{k+1} - p_k) h(p_{k+1}) f_{U_0|\mathcal{C}_k}(\tau)} \\ & \cdot \sum_{k=1}^{K-1}(p_{k+1} - p_k) h(p_{k+1}) \delta_{\mathcal{C}_k}(\tau) \\ & - \left[\sum_{k=1}^{K-1}(p_{k+1} - p_k) F_{p(Z)}(p_k) \delta_{\mathcal{C}_k}(\tau) + (1 - p_K) \delta_{\mathcal{N}}(\tau)\right], \\ D_{F_0}(\delta)(\tau) = & -\frac{p_1 f_{U_1|\mathcal{A}}(\tau) + \sum_{k=1}^{K-1}(p_{k+1} - p_k)\left(1 - F_{p(Z)}(p_k)\right) f_{U_1|\mathcal{C}_k}(\tau)}{\sum_{k=1}^{K-1}(p_{k+1} - p_k) h(p_{k+1}) f_{U_1|\mathcal{C}_k}(\tau)} \\ & \cdot \sum_{k=1}^{K-1}(p_{k+1} - p_k) h(p_{k+1}) \delta_{\mathcal{C}_k}(\tau) \\ & + \left[p_1 \delta_{\mathcal{A}}(\tau) + \sum_{k=1}^{K-1}(p_{k+1} - p_k)\left(1 - F_{p(Z)}(p_k)\right) \delta_{\mathcal{C}_k}(\tau)\right], \end{aligned}$$

with

$$\begin{aligned} f_{U_1|\mathcal{A}}(\tau) &= \frac{1}{p_1}\int_0^{p_1} f_{U_1|V}(\tau|v)\,dv, \\ f_{U_d|\mathcal{C}_k}(\tau) &= \frac{1}{p_2-p_1}\int_{p_1}^{p_2} f_{U_d|V}(\tau|v)\,dv, \\ f_{U_0|\mathcal{N}}(\tau) &= \frac{1}{1-p_2}\int_{p_2}^{1} f_{U_0|V}(\tau|v)\,dv, \end{aligned}$$

so

$$\begin{aligned} & \sqrt{n}\left(F^n_{Y_1|\mathcal{C}_k}\left(F_1^{n-1}(\tau)\right) - F^n_{Y_0|\mathcal{C}_k}\left(F_0^{n-1}(\tau)\right)\right) \\ = \;& \sqrt{n}\left(F^n_{U|\mathcal{C}_k}\left(F_1\left(F_1^{n-1}(\tau)\right)\right) - F_{U|\mathcal{C}_k}\left(F_0\left(F_0^{n-1}(\tau)\right)\right)\right) \\ = \;& \sqrt{n}\left(F^n_{U|\mathcal{C}_k}\left(F_1\left(F_1^{n-1}(\tau)\right)\right) - F_{U|\mathcal{C}_k}\left(F_1\left(F_1^{n-1}(\tau)\right)\right)\right) \\ & +\sqrt{n}\left(F_{U|\mathcal{C}_k}\left(F_1\left(F_1^{n-1}(\tau)\right)\right) - F_{U|\mathcal{C}_k}\left(F_1\left(F_1^{-1}(\tau)\right)\right)\right) \\ & +\sqrt{n}\left(F_{U|\mathcal{C}_k}\left(F_0\left(F_0^{-1}(\tau)\right)\right) - F_{U|\mathcal{C}_k}\left(F_0\left(F_0^{n-1}(\tau)\right)\right)\right) \\ \to \;& \delta_{\mathcal{C}_k}(\tau) - f_{U|\mathcal{C}_k}(\tau) f_1\left(F_1^{-1}(\tau)\right)\frac{D_{F_1}(\delta)(\tau)}{f_1\left(F_1^{-1}(\tau)\right)} + f_{U|\mathcal{C}_k}(\tau) f_0\left(F_0^{-1}(\tau)\right)\frac{D_{F_0}(\delta)(\tau)}{f_0\left(F_0^{-1}(\tau)\right)} \\ = \;& \delta_{\mathcal{C}_k}(\tau) - f_{U|\mathcal{C}_k}(\tau)\left[D_{F_1}(\delta)(\tau) - D_{F_0}(\delta)(\tau)\right] \equiv b_k^{(1)}(\tau), \end{aligned}$$

where $F^n_{U|\mathcal{C}_k} = F_{U|\mathcal{C}_k} + \delta_{\mathcal{C}_k}/\sqrt{n}$, $F_{U|\mathcal{C}_k}(u) = F_{U_0|\mathcal{C}_k}(u) = \frac{1}{p_2-p_1}\int_{p_1}^{p_2} F_{U_0|V}(u|v)\,dv$, and $f_{U|\mathcal{C}_k}(\cdot)$ is similarly defined. $b_k^{(1)}(\tau)$ can be further simplified:

$$\begin{aligned} b_k^{(1)}(\tau) &= \delta_{\mathcal{C}_k}(\tau) - f_{U|\mathcal{C}_k}(\tau)\left[\begin{array}{l} \frac{\sum_{k=1}^{K-1}(p_{k+1}-p_k)F_{p(Z)}(p_k)f_{U_0|\mathcal{C}_k}(\tau)+(1-p_K)f_{U_0|\mathcal{N}}(\tau)}{\sum_{k=1}^{K-1}(p_{k+1}-p_k)h(p_{k+1})f_{U_0|\mathcal{C}_k}(\tau)}\cdot\sum_{k=1}^{K-1}(p_{k+1}-p_k)h(p_{k+1})\delta_{\mathcal{C}_k}(\tau) \\ +\frac{p_1 f_{U_1|\mathcal{A}}(\tau)+\sum_{k=1}^{K-1}(p_{k+1}-p_k)\left(1-F_{p(Z)}(p_k)\right)f_{U_1|\mathcal{C}_k}(\tau)}{\sum_{k=1}^{K-1}(p_{k+1}-p_k)h(p_{k+1})f_{U_1|\mathcal{C}_k}(\tau)}\cdot\sum_{k=1}^{K-1}(p_{k+1}-p_k)h(p_{k+1})\delta_{\mathcal{C}_k}(\tau) \\ -\left[\sum_{k=1}^{K-1}(p_{k+1}-p_k)F_{p(Z)}(p_k)\delta_{\mathcal{C}_k}(\tau)+(1-p_K)\delta_{\mathcal{N}}(\tau)\right] \\ -\left[p_1\delta_{\mathcal{A}}(\tau)+\sum_{k=1}^{K-1}(p_{k+1}-p_k)\left(1-F_{p(Z)}(p_k)\right)\delta_{\mathcal{C}_k}(\tau)\right] \end{array}\right] \\ &= \delta_{\mathcal{C}_k}(\tau) - \frac{p_1 f_{U|\mathcal{A}}(\tau)+\sum_{k=1}^{K-1}(p_{k+1}-p_k)f_{U|\mathcal{C}_k}(\tau)+(1-p_K)f_{U|\mathcal{N}}(\tau)}{\sum_{k=1}^{K-1}(p_{k+1}-p_k)h(p_{k+1})} \\ & \quad\cdot\sum_{k=1}^{K-1}(p_{k+1}-p_k)h(p_{k+1})\delta_{\mathcal{C}_k}(\tau) + \left[p_1\delta_{\mathcal{A}}(\tau)+\sum_{k=1}^{K-1}(p_{k+1}-p_k)\delta_{\mathcal{C}_k}(\tau)+(1-p_K)\delta_{\mathcal{N}}(\tau)\right]f_{U|\mathcal{C}_k}(\tau) \\ &= \delta_{\mathcal{C}_k}(\tau) - \sum_{l=1}^{K-1}(p_{l+1}-p_l)h(p_{l+1})\delta_{\mathcal{C}_l}(\tau), \end{aligned}$$

where the second equality uses $f_{U_1|\text{type}} = f_{U_0|\text{type}} = f_{U|\text{type}}$ in the limit, and the last equality uses $p_1 f_{U|\mathcal{A}}(\tau) + \sum_{k=1}^{K-1}(p_{k+1}-p_k)f_{U|\mathcal{C}_k}(\tau) + (1-p_K)f_{U|\mathcal{N}}(\tau) = 1$, $p_1\delta_{\mathcal{A}}(\tau) + \sum_{k=1}^{K-1}(p_{k+1}-p_k)\delta_{\mathcal{C}_k}(\tau) + (1-p_K)\delta_{\mathcal{N}}(\tau) = 0$, and $\sum_{k=1}^{K-1}(p_{k+1}-p_k)h(p_{k+1}) = 1$. As a result,

$$T_{1n} \rightsquigarrow \sum_{k=1}^{K-1}\int_{\mathcal{T}}\left(Z_k^{(1)}(\tau) + b_k^{(1)}(\tau)\right)^2 d\tau = \sum_{k=1}^{K-1}\sum_{i=1}^{\infty}\left(\sqrt{\lambda_{ki}^{(1)}}\varepsilon_{ki}^{(1)} + b_{ki}^{(1)}\right)^2$$

under H_1^δ, where $b_{ki}^{(1)} = \int_{\mathcal{T}} b_k^{(1)}(\tau)\,\varphi_i^{(1)}(\tau)\,d\tau$.

Define $T_{ki}^{(1)} = \sum_{k=1}^{K-1}\sum_{j\neq i}(\sqrt{\lambda_{ki}^{(1)}}\varepsilon_{ki}^{(1)} + b_{ki}^{(1)})^2$. By repeated application of $P(\sum_{k=1}^{K-1}\sum_{i=1}^{\infty}(\sqrt{\lambda_{ki}^{(1)}}\varepsilon_{ki}^{(1)} + b_{ki}^{(1)})^2 > c) \geq (PT_{ki}^{(1)} + \lambda_{ki}^{(1)}\varepsilon_{ki}^{(1)2} > c)$, we get the result, where the inequality is strict if $b_{ki}^{(1)} \neq 0$. So unless

$b_{ki}^{(1)} = 0$ for all k and i, which is equivalent to $b_k^{(1)}(\tau) = 0$ for any k and τ, the strict inequality holds.

(iii) From the analysis in (ii),

$$nT_{1n} \approx \sum_{l=1}^{K-1} \int_{\mathcal{T}} \left[\widetilde{Z}_k^{(1)}(\tau) + \sqrt{n} \left(F_{Y_1|\mathcal{C}_k} \left(F_1^{*-1}(\tau) \right) - F_{Y_0|\mathcal{C}_k} \left(F_0^{*-1}(\tau) \right) \right) \right]^2 d\tau = O_p(n)..$$

Here, $\widetilde{Z}_k^{(1)}(\tau)$ depends on H_1; it may be different from $Z_k^{(1)}(\tau)$ (equals $Z_k^{(1)}(\tau)$ under H_0), but is a tight mean zero Gaussian process. So nT_{1n} is dominated by $\sum_{l=1}^{K-1} \int_{\mathcal{T}} \left[\sqrt{n} \left(F_{Y_1|\mathcal{C}_k} \left(F_1^{*-1}(\tau) \right) - F_{Y_0|\mathcal{C}_k} \left(F_0^{*-1}(\tau) \right) \right) \right]^2 d\tau$ which is $O(n)$. ■

Proof of Corollary 4. From the proof of Theorem 2, $Z_1(y_1)$ is the second-to-last component of $\Psi_y^{(2)} \left(\Psi_y^{(1)} \left(\mathbb{G} \left(\varphi_y(W) \right) \right) \right)$ in (22) and $Z_0(y_0)$ is the last component of $\Psi_y^{(2)} \left(\Psi_y^{(1)} \left(\mathbb{G} \left(\varphi_y(W) \right) \right) \right)$. Then by the functional Delta method, it is not hard to get the weak limit of $\widehat{\Delta}(\tau)$. ■

Proof of Theorem 3. The proof is quite similar to the Theorem 2, so we only outline the differences. (i) In $\Psi_y^{(1)} \left(\mathbb{G} \left(\varphi_y(W) \right) \right)$, we need only terms associated with $\left\{ F_{Y_1|\mathcal{C}_k}(y_1), F_{Y_0|\mathcal{C}_k}(y_0) \right\}_{k=1}^{K-1}$; in other words, define

$$\Psi_y^{(1')} \left(\mathbb{G} \left(\varphi_y(W) \right) \right) = \mathbb{G} \left(\{ \psi_1^k(W, y_1), \psi_0^k(W, y_0) \}_{k=1}^{K-1} \right).$$

We do not need $\Psi_y^{(2)}$, and change $\Psi_\tau^{(3)}$ to $\Psi_\tau^{(4)}$ below. The linear map $\Psi_\tau^{(4)} : C(\mathcal{Y})^{2(K-1)} \to \ell^\infty(\mathcal{T})^{2(K-2)}$ is defined as follows. For $\alpha(y) = \left(\{ \alpha_1^k(y_1), \alpha_0^k(y_0) \}_{k=1}^{K-1} \right)' \in C(\mathcal{Y})^{2(K-1)}$, the term associated with $\widehat{F}_{Y_1|\mathcal{C}_k} \left(\widehat{F}_{Y_1|\mathcal{C}_1}^{-1}(\tau) \right)$ in the image of $\Psi_\tau^{(4)}(\alpha(y))$ is

$$f_{Y_1|\mathcal{C}_k} \left(F_{Y_1|\mathcal{C}_1}^{-1}(\tau) \right) \alpha_1^k \left(F_{Y_1|\mathcal{C}_1}^{-1}(\tau) \right) - \frac{f_{Y_1|\mathcal{C}_k} \left(F_{Y_1|\mathcal{C}_1}^{-1}(\tau) \right)}{f_{Y_1|\mathcal{C}_1} \left(F_{Y_1|\mathcal{C}_1}^{-1}(\tau) \right)} \alpha_1^1(F_{Y_1|\mathcal{C}_1}^{-1}(\tau)),$$

and the term associated with $\widehat{F}_{Y_0|\mathcal{C}_k} \left(\widehat{F}_{Y_0|\mathcal{C}_1}^{-1}(\tau) \right)$ is

$$f_{Y_0|\mathcal{C}_k} \left(F_{Y_0|\mathcal{C}_1}^{-1}(\tau) \right) \alpha_0^k \left(F_{Y_0|\mathcal{C}_1}^{-1}(\tau) \right) - \frac{f_{Y_0|\mathcal{C}_k} \left(F_{Y_0|\mathcal{C}_1}^{-1}(\tau) \right)}{f_{Y_0|\mathcal{C}_1} \left(F_{Y_0|\mathcal{C}_1}^{-1}(\tau) \right)} \alpha_0^1(F_{Y_0|\mathcal{C}_1}^{-1}(\tau)).$$

In summary,

$$\begin{aligned} T_{2n} &\equiv \sum_{k=1}^{K-1} \int_{\mathcal{T}} Z_{kn}^{(2)}(\tau)^2 d\tau \\ &\rightsquigarrow \int_{\mathcal{T}} \left[\Psi_{k1\tau}^{(4)} \left(\left(\Psi_{F_{Y_1|\mathcal{C}_1}^{-1}(\tau)}^{(1')} \left(\mathbb{G} \left(\varphi_{F_{Y_1|\mathcal{C}_1}^{-1}(\tau)}(W) \right) \right) \right) \right) - \Psi_{k2\tau}^{(4)} \left(\Psi_{F_{Y_0|\mathcal{C}_1}^{-1}(\tau)}^{(1')} \left(\mathbb{G} \left(\varphi_{F_{Y_0|\mathcal{C}_1}^{-1}(\tau)}(W) \right) \right) \right) \right]^2 d\tau \\ &\equiv \sum_{k=1}^{K-1} \int_{\mathcal{T}} Z_k^{(2)}(\tau)^2 d\tau, \end{aligned}$$

where $\Psi_\tau^{(4)} = \left\{ \Psi_{k\tau}^{(4)} \right\}_{k=2}^{K-1}$ and $\Psi_{k\tau}^{(4)} = \left(\Psi_{k1\tau}^{(4)}, \Psi_{k2\tau}^{(4)} \right)$.

(ii) We need only to calculate the limit of $\sqrt{n} \left(F_{Y_1|\mathcal{C}_k}^n \left(F_{Y_1|\mathcal{C}_1}^{n-1}(\tau) \right) - F_{Y_0|\mathcal{C}_k}^n \left(F_{Y_0|\mathcal{C}_1}^{n-1}(\tau) \right) \right)$ to determine the

local power.

$$\begin{aligned}
& \sqrt{n}\left(F^{n}_{Y_1|\mathcal{C}_k}\left(F^{n-1}_{Y_1|\mathcal{C}_1}(\tau)\right)-F^{n}_{Y_0|\mathcal{C}_k}\left(F^{n-1}_{Y_0|\mathcal{C}_1}(\tau)\right)\right)\\
= \;& \sqrt{n}\left(F^{n}_{U|\mathcal{C}_k}\left(F_1\left(F_1^{-1}F^{n-1}_{U|\mathcal{C}_1}(\tau)\right)\right)-F_{U|\mathcal{C}_k}\left(F_0\left(F_0^{-1}F^{-1}_{U|\mathcal{C}_1}(\tau)\right)\right)\right)\\
= \;& \sqrt{n}\left(F^{n}_{U|\mathcal{C}_k}\left(F^{n-1}_{U|\mathcal{C}_1}(\tau)\right)-F_{U|\mathcal{C}_k}\left(F^{-1}_{U|\mathcal{C}_1}(\tau)\right)\right)\\
= \;& \sqrt{n}\left(F^{n}_{U|\mathcal{C}_k}\left(F^{n-1}_{U|\mathcal{C}_1}(\tau)\right)-F_{U|\mathcal{C}_k}\left(F^{n-1}_{U|\mathcal{C}_1}(\tau)\right)\right)+\sqrt{n}\left(F_{U|\mathcal{C}_k}\left(F^{n-1}_{U|\mathcal{C}_1}(\tau)\right)-F_{U|\mathcal{C}_k}\left(F^{-1}_{U|\mathcal{C}_1}(\tau)\right)\right)\\
\to \;& \delta_{\mathcal{C}_k}(F^{-1}_{U|\mathcal{C}_1}(\tau))-\frac{f_{U|\mathcal{C}_k}\left(F^{-1}_{U|\mathcal{C}_1}(\tau)\right)}{f_{U|\mathcal{C}_1}\left(F^{-1}_{U|\mathcal{C}_1}(\tau)\right)}\delta_{\mathcal{C}_1}(F^{-1}_{U|\mathcal{C}_1}(\tau)).
\end{aligned}$$

■

Proof of Corollary 5. From the construction of T'_{1n}, it is not hard to see that

$$nT'_{1n} \rightsquigarrow \sum\nolimits_{k=1}^{K-2}\int_{\mathcal{T}}\left[\sum\nolimits_{l=1}^{K-1}\omega_{kl}\left(Z^{(1)}_{l}(\tau)+b^{(1)}_{k}(\tau)\right)\right]^2 d\tau.$$

By Mercer's theorem, we can represent the weak limit in the form of Corollary 5. As to T'_{2n}, first note that the weak limit of $\widehat{\widetilde{F}}_d(y_d)$ is stated in (23) and the weak limit of $\widehat{F}_{Y_d|\mathcal{C}_k}(y_d)$ is stated in Lemma 1. For future reference, we denote them as $\mathbb{G}\left(\widetilde{\psi}_d(W,y_d)\right)$ and $\mathbb{G}\left(\psi^k_d(W,y_d)\right)$, respectively. Next, since $\left\{F^{(k)}_{Y_1|\mathcal{C}}\left(\widetilde{F}_1^{-1}(\tau)\right),F^{(k)}_{Y_0|\mathcal{C}}\left(\widetilde{F}_0^{-1}(\tau)\right)\right\}_{k=1}^{K-2}$ is a Hadamard differentiable map of $(\{F_{Y_1|\mathcal{C}_k}(y_1),F_{Y_0|\mathcal{C}_k}(y_0)\}_{k=1}^{K-1}$, $\widetilde{F}_1(\tau),\widetilde{F}_0(\tau))$, we can apply the functional Delta method to have

$$\sqrt{n}\begin{pmatrix}\widehat{F}^{(k)}_{Y_1|\mathcal{C}}\left(\widehat{\widetilde{F}}_1^{-1}(\tau)\right)-F^{(k)}_{U_1|\mathcal{C}}(\tau)\\ \widehat{F}^{(k)}_{Y_0|\mathcal{C}_k}\left(\widehat{F}_0^{-1}(\tau)\right)-F^{(k)}_{U_0|\mathcal{C}}(\tau)\end{pmatrix}\rightsquigarrow\left.\begin{pmatrix}\widetilde{\Psi}^{(k)}_{1\tau}\left(\mathbb{G}\left(\widetilde{\psi}_1(W,y_1)\right),\left\{\mathbb{G}\left(\psi^l_1(W,y_1)\right)\right\}_{l=1}^{K-1}\right)\\ \widetilde{\Psi}^{(k)}_{0\tau}\left(\mathbb{G}\left(\widetilde{\psi}_0(W,y_0)\right),\left\{\mathbb{G}\left(\psi^l_0(W,y_0)\right)\right\}_{l=1}^{K-1}\right)\end{pmatrix}\right|_{y_1=\widetilde{F}_1^{-1}(\tau),y_0=\widetilde{F}_0^{-1}(\tau)}$$

in $\ell^\infty(\mathcal{T})^{2(K-2)}$, where

$$F^{(k)}_{U_d|\mathcal{C}}(\tau)=\sum\nolimits_{l=1}^{K-1}\omega_{kl}F_{U_d|\mathcal{C}_l}\left(\widetilde{F}^{-1}_{U_d}(\tau)\right)\text{ with }\widetilde{F}_{U_d}(u)=\sum\nolimits_{l=1}^{K-1}(p_{l+1}-p_l)\,h(p_{l+1})F_{U_d|\mathcal{C}_l}(u),\tag{24}$$

and for $\alpha_d(y_d)=\left(\widetilde{\alpha}_d(y_d),\left\{\alpha^l_d(y_d)\right\}_{l=1}^{K-1}\right)'\in C(\mathcal{Y})^K$, the linear map $\widetilde{\Psi}^{(k)}_{d\tau}(\cdot):C(\mathcal{Y})^K\to\ell^\infty(\mathcal{T})$ evaluated at $\alpha(y)$ is defined as follows,

$$\begin{aligned}
\widetilde{\Psi}^{(k)}_{1\tau}(\alpha_1(y_1)) &= \sum\nolimits_{l=1}^{K-1}\omega_{kl}\alpha^l_1\left(\widetilde{F}_1^{-1}(\tau)\right)-\frac{\sum_{l=1}^{K-1}\omega_{kl}f_{Y_1|\mathcal{C}_l}\left(\widetilde{F}_1^{-1}(\tau)\right)}{\widetilde{f}_1\left(\widetilde{F}_1^{-1}(\tau)\right)}\widetilde{\alpha}_1\left(\widetilde{F}_1^{-1}(\tau)\right),\\
\widetilde{\Psi}^{(k)}_{0\tau}(\alpha_0(y_0)) &= \sum\nolimits_{l=1}^{K-1}\omega_{kl}\alpha^l_0\left(\widetilde{F}_0^{-1}(\tau)\right)-\frac{\sum_{l=1}^{K-1}\omega_{kl}f_{Y_0|\mathcal{C}_l}\left(\widetilde{F}_0^{-1}(\tau)\right)}{\widetilde{f}_0\left(\widetilde{F}_0^{-1}(\tau)\right)}\widetilde{\alpha}_0\left(\widetilde{F}_0^{-1}(\tau)\right).
\end{aligned}$$

As to $b_k^{\prime(2)}(\tau)$, by a similar analysis as in the proof of Theorem 3(ii), it is not hard to see that

$$b_k^{\prime(2)}(\tau) = \sum\nolimits_{l=1}^{K-1} \omega_{kl}\delta_{\mathcal{C}_k}(\widetilde{F}_U^{-1}(\tau)) - \frac{\sum_{l=1}^{K-1}\omega_{kl}f_{U|\mathcal{C}_k}\left(\widetilde{F}_U^{-1}(\tau)\right)}{\widetilde{f}_U\left(\widetilde{F}_U^{-1}(\tau)\right)}\widetilde{\delta}(\widetilde{F}_U^{-1}(\tau)),$$

where $\widetilde{F}_U(u)$ is defined in (24), $\widetilde{f}_U(u)$ is similarly defined, and $\widetilde{\delta}(u) = \sum_{k=1}^{K-1}(p_{k+1}-p_k)\,h(p_{k+1})f_{U|\mathcal{C}_k}(u)$. Finally, applying the continuous mapping theorem, we have

$$nT_{2n}' \rightsquigarrow \sum\nolimits_{k=1}^{K-2}\int_{\mathcal{T}}\left(Z_k^{\prime(2)}(\tau) + b_k^{\prime(2)}(\tau)\right)^2 d\tau,$$

where $Z_k^{\prime(2)}(\tau) = \widetilde{\Psi}_{1\tau}^{(k)}(\mathbb{G}(\widetilde{\psi}_1(W,\widetilde{F}_1^{-1}(\tau))),\{\mathbb{G}(\psi_1^l(W,\widetilde{F}_1^{-1}(\tau)))\}_{l=1}^{K-1}) - \widetilde{\Psi}_{0\tau}^{(k)}(\mathbb{G}(\widetilde{\psi}_0(W,\widetilde{F}_0^{-1}(\tau))),\{\mathbb{G}(\psi_0^l(W,\widetilde{F}_0^{-1}(\tau)))\}_{l=1}^{K-1})$. By Mercer's theorem, we can represent this weak limit in the form of Corollary 5. ■

Proof of Theorem 4. (i) First,

$$\mathbb{G}_n\left(\varphi_y(W)\right) \rightsquigarrow \mathbb{G}\left(\varphi_y(W)\right) \text{ in } \ell^\infty(\mathcal{Y})^{5K}$$

by Lemma 1. From the proof of Theorem 2, the process $\left(F_{Y_1|\mathcal{C}_k}\left(F_1^{-1}(\tau)\right), F_{Y_0|\mathcal{C}_k}\left(F_0^{-1}(\tau)\right)\right)$, $\tau \in \mathcal{T}$, is Hadamard differentiable at $E\left[\varphi_y(W)\right]$ tangentially to $C(\mathcal{Y})^{5K}$, so by the functional Delta method for the bootstrap (see, e.g., Theorem 3.9.11 of VW),

$$\sqrt{n}\begin{pmatrix} \widehat{F}^*_{Y_1|\mathcal{C}_k}\left(\widehat{F}_1^{*-1}(\tau)\right) - \widehat{F}_{Y_1|\mathcal{C}_k}\left(\widehat{F}_1^{-1}(\tau)\right) \\ \widehat{F}^*_{Y_0|\mathcal{C}_k}\left(\widehat{F}_0^{*-1}(\tau)\right) - \widehat{F}_{Y_0|\mathcal{C}_k}\left(\widehat{F}_0^{-1}(\tau)\right) \end{pmatrix} \overset{*}{\rightsquigarrow} \Psi_\tau^{(3)}\left(\Psi_y^{(2)}\left(\Psi_y^{(1)}\left(\mathbb{G}\left(\varphi_y(W)\right)\right)\right)\right)\Big|_{y_1=F_1^{-1}(\tau),y_0=F_0^{-1}(\tau)}$$

in $\ell^\infty(\mathcal{T})^{2(K-1)}$. Finally, by the continuous mapping theorem,

$$\begin{aligned} nT_{1n}^* &= \sum_{k=1}^{K-1}\int_{\mathcal{T}}\left[\sqrt{n}\left(\widehat{F}^*_{Y_1|\mathcal{C}_k}\left(\widehat{F}_1^{*-1}(\tau)\right) - \widehat{F}_{Y_1|\mathcal{C}_k}\left(\widehat{F}_1^{-1}(\tau)\right)\right) - \sqrt{n}\left(\widehat{F}^*_{Y_0|\mathcal{C}_k}\left(\widehat{F}_0^{*-1}(\tau)\right) - \widehat{F}_{Y_0|\mathcal{C}_k}\left(\widehat{F}_0^{-1}(\tau)\right)\right)\right]^2 d\tau \\ &\overset{*}{\rightsquigarrow} \sum\nolimits_{k=1}^{K-1}\int_{\mathcal{T}} Z_k^{(1)}(\tau)^2\, d\tau \end{aligned}$$

as desired. It follows that $c_{1n}^*(\alpha) = c_1(\alpha) + o_p(1)$ under H_0, where $c_1(\alpha)$ is the $(1-\alpha)$th quantile of the asymptotic distribution of nT_{1n}. This implies that nT_{1n} and $nT_{1n} - (c_{1n}^*(\alpha) - c_1(\alpha))$ converges to the same limiting distribution as $n \to \infty$, and hence we have that $P\left(nT_{1n} > c_{1n}^*(\alpha)\right) = \alpha + o(1)$.

(ii) By Corollary 2.1 of Bickel and Ren (2001, p. 97), the bootstrap is valid for $\{\widehat{F}_{Y_1|\mathcal{C}_k}(\widehat{F}_1^{-1}(\tau)), \widehat{F}_{Y_0|\mathcal{C}_k}(\widehat{F}_0^{-1}(\tau))\}_{k=1}^{K-1}$ if H_1^δ is contiguous to H_0, and thus the arguments in (i) can still go through to show that $c_{1n}^*(\alpha) = c_1(\alpha) + o_p(1)$ under H_1^δ. By Theorem 2(ii), the result follows.

(iii) Under a fixed alternative, $nT_{1n}^* \overset{*}{\rightsquigarrow} \sum_{l=1}^{K-1}\int_{\mathcal{T}}\left[\widetilde{Z}_k^{(1)}(\tau)\right]^2 d\tau$, where $\left\{\widetilde{Z}_k^{(1)}(\tau)\right\}_{k=1}^{K-1}$ are defined in the proof of Theorem 2(iii), and thus $c_{1n}^*(\alpha) = O_p(1)$. As a result, for any $\epsilon > 0$, there exists a constant M such that $P(c_{1n}^*(\alpha) > M) < \epsilon + o(1)$. Using elementary inequalities, we also have that

$$\begin{aligned} P(nT_{1n} &\leq c_{1n}^*(\alpha)) = P(nT_{1n} \leq c_{1n}^*(\alpha), c_{1n}^*(\alpha) \leq M) + P(nT_{1n} \leq c_{1n}^*(\alpha), c_{1n}^*(\alpha) > M) \\ &\leq P(nT_{1n} \leq M) + P(c_{1n}^*(\alpha) > M). \end{aligned}$$

From Theorem 2(iii), we know that $P(nT_{1n} \leq M) = o(1)$, and thus $P(nT_{1n} \leq c_{1n}^*(\alpha)) < \epsilon + o(1)$, which implies the statement of the theorem since ϵ can be chosen arbitrarily small.

(iv) The proof is parallel to that for T_{1n}. ■

Proof of Theorem 5. The test statistic involves the following processes $(\{\widehat{F}^k_{Y_1|X}(y_1|x), \widehat{F}^k_{Y_0|X}(y_0|x)\}_{k=1}^{K-1}$, $\widehat{F}^{-1}_{Y_1|X}(\tau|x), \widehat{F}^{-1}_{Y_0|X}(\tau|x), \widehat{F}_X(x))$, $\tau \in \mathcal{T}$, $y_d \in \mathcal{Y}_d$, $x \in \mathcal{X}$, where $(\widehat{F}^{-1}_{Y_1|X}(\tau|x), \widehat{F}^{-1}_{Y_0|X}(\tau|x))$ involves $(\widehat{F}^0_{Y_1|X}(y_1|x)$, $\{\widehat{F}^k_{Y_1|X}(y_1|x), \widehat{F}^k_{Y_0|X}(y_1|x)\}_{k=1}^{K-1}, \widehat{F}^K_{Y_0|X}(y_0|x), \{\widehat{p}_{xl}, \widehat{q}_{xl}\}_{l=1}^K)$. Lemma 2 derives the weak limit of these components. Now, since $(\{F^k_{Y_1|X}(y_1|x), F^k_{Y_0|X}(y_0|x)\}_{k=1}^{K-1}$, $F_{Y_1|X}(y_1|x), F_{Y_0|X}(y_0|x))$ is a Hadamard differentiable map of $(\{F^k_{Y_1|X}(y_1|x)\}_{k=0}^{K-1}, \{F^k_{Y_0|X}(y_0|x)\}_{k=1}^K, \{p_{xl}, q_{xl}\}_{l=1}^K)$, by Lemma 2 and the functional Delta method, we have

$$\sqrt{n}\begin{pmatrix} \widehat{F}^k_{Y_1|X}(y_1|x) - F^k_{Y_1|X}(y_1|x) \\ \widehat{F}^k_{Y_0|X}(y_0|x) - F^k_{Y_0|X}(y_0|x) \\ \widehat{F}_{Y_1|X}(y_1|x) - F_{Y_1|X}(y_1|x) \\ \widehat{F}_{Y_0|X}(y_0|x) - F_{Y_0|X}(y_0|x) \end{pmatrix} \rightsquigarrow \Psi^{(3)}_{yx}\left(\Psi^{(2)}_{yx}\left(\Psi^{(1)}_{yxp}\left(-J^{-1}(y)\mathbb{G}\left(\varphi_{y,\theta}\right)\right)\right)\right) \text{ in } \ell^\infty\left(\mathcal{YX}\right)^{2K},$$

where $\Psi^{(2)}_{yx}\left(\Psi^{(1)}_{yxp}\left(-J^{-1}(y)\mathbb{G}\left(\varphi_{y,\theta}\right)\right)\right)$ is defined in Lemma 2, and $\Psi^{(3)}_{yx}(\cdot) : C\left(\mathcal{YX}\right)^{4K} \to \ell^\infty\left(\mathcal{YX}\right)^{2K}$ is similarly derived as in the proof of Theorem 2 due to similar structures. Next, since $\{F^k_{Y_1|X}\left(F^{-1}_{Y_1|X}(\tau|x)|x\right)$, $F^k_{Y_0|X}\left(F^{-1}_{Y_0|X}(\tau|x)|x\right)\}_{k=1}^{K-1}$ is a Hadamard differentiable map of $(\{F^k_{Y_1|X}(y_1|x), F^k_{Y_0|X}(y_0|x)\}_{k=1}^{K-1}, F_{Y_1|X}(y_1|x)$, $F_{Y_0|X}(y_0|x))$, we can apply the functional Delta method again to have

$$\sqrt{n}\begin{pmatrix} \widehat{F}^k_{Y_1|X}\left(\widehat{F}^{-1}_{Y_1|X}(\tau|x)|x\right) - F^k_{U_1|X}(\tau|x) \\ \widehat{F}^k_{Y_0|X}\left(\widehat{F}^{-1}_{Y_0|X}(\tau|x)|x\right) - F^k_{U_0|X}(\tau|x) \end{pmatrix} \rightsquigarrow \Psi^{(4)}_{\tau x}\left(\Psi^{(3)}_{yx}\left(\Psi^{(2)}_{yx}\left(\Psi^{(1)}_{yxp}\left(-J^{-1}(y)\mathbb{G}\left(\varphi_{y,\theta}\right)\right)\right)\right)\right)\Big|_{y_1=Q_{Y_1|X}(\tau|x), y_0=Q_{Y_0|X}(\tau|x)} \tag{25}$$

in $\ell^\infty\left(\mathcal{TX}\right)^{2(K-1)}$, where $\Psi^{(4)}_{\tau x} : C\left(\mathcal{YX}\right)^{2K} \to \ell^\infty\left(\mathcal{TX}\right)^{2(K-1)}$ is similarly derived as in the proof of Theorem 2 due to similar structures. Finally, by the continuous mapping theorem and Slutsky theorem, we have under H_0,

$$\begin{aligned} nT^X_{1n} &\equiv \sum\nolimits_{k=1}^{K-1}\int_{\mathcal{X}}\int_{\mathcal{T}} Z^{(1)}_{kn}(\tau,x)^2\, d\tau d\widehat{F}_X(x) \\ &= \sum\nolimits_{k=1}^{K-1}\int_{\mathcal{X}}\int_{\mathcal{T}} Z^{(1)}_{kn}(\tau,x)^2\, d\tau dF_X(x) + o_p(1) \\ &\rightsquigarrow \sum\nolimits_{k=1}^{K-1}\int_{\mathcal{X}}\int_{\mathcal{T}} \Big[\Psi^{(4)}_{k1\tau x}\left(\Psi^{(3)}_{Q_{Y_1|X}(\tau|x)x}\left(\Psi^{(2)}_{Q_{Y_1|X}(\tau|x)x}\left(\Psi^{(1)}_{Q_{Y_1|X}(\tau|x)xp}\left(-J^{-1}(Q_{Y_1|X}(\tau|x))\mathbb{G}\left(\varphi_{Q_{Y_1|X}(\tau|x),\theta}(W)\right)\right)\right)\right)\right) \\ &\quad -\Psi^{(4)}_{k2\tau x}\left(\Psi^{(3)}_{Q_{Y_0|X}(\tau|x)x}\left(\Psi^{(2)}_{Q_{Y_0|X}(\tau|x)x}\left(\Psi^{(1)}_{Q_{Y_0|X}(\tau|x)xp}\left(-J^{-1}(Q_{Y_0|X}(\tau|x))\mathbb{G}\left(\varphi_{Q_{Y_0|X}(\tau|x),\theta}(W)\right)\right)\right)\right)\right)\Big]^2 d\tau dF_X(x) \\ &\equiv \sum\nolimits_{k=1}^{K-1}\int_{\mathcal{X}}\int_{\mathcal{T}} Z^{(1)}_{k}(\tau,x)^2\, d\tau dF_X(x), \end{aligned}$$

where $\Psi^{(4)}_{\tau x} = \left\{\Psi^{(4)}_{k\tau x}\right\}_{k=1}^{K-1}$ with $\Psi^{(4)}_{k\tau x} = \left(\Psi^{(4)}_{k1\tau x}, \Psi^{(4)}_{k2\tau x}\right)$, and the second equality is due to $\sup_{x\in\mathcal{X}}\left|\widehat{F}_X(x) - F_X(x)\right| = o_p(1)$ and $\int_{\mathcal{T}} Z^{(1)}_{kn}(\tau,x)^2\, d\tau$ is tight in $\ell^\infty(\mathcal{X})$ under H_0. The proof under H^δ_1 is completely the same as that of Theorem 2(ii) except that the calculation is conditional on $X = x$. ■

Proof of Theorem 6. We only derive $Z^{(2)}_k(\tau,x)$ to signify the difference from Theorem 5. In $\Psi^{(2)}_{yx}(\Psi^{(1)}_{yxp}(-J^{-1}(y)\mathbb{G}(\varphi_{y,\theta})))$, we need only terms associated with $\{F^k_{Y_1|X}(y_1|x), F^k_{Y_0|X}(y_0|k)\}_{k=1}^{K-1}$, so define $\Psi^{(2')}_{yx}\left(\Psi^{(1)}_{yxp}\left(-J^{-1}(y)\mathbb{G}\left(\varphi_{y,\theta}\right)\right)\right) \in \ell^\infty\left(\mathcal{YX}\right)^{2(K-1)}$ as the corresponding elements of $\Psi^{(2)}_{yx}\left(\Psi^{(1)}_{yxp}\left(-J^{-1}(y)\mathbb{G}\left(\varphi_{y,\theta}\right)\right)\right)$. We do not need $\Psi^{(3)}_{yx}$, and change $\Psi^{(4)}_{\tau x}$ to $\Psi^{(5)}_{\tau x}$ below. The linear map $\Psi^{(5)}_{\tau x} : C\left(\mathcal{YX}\right)^{2(K-1)} \to \ell^\infty\left(\mathcal{TX}\right)^{2(K-2)}$ is defined as follows. For $\alpha(y,x) = \left(\{\alpha^k_1(y_1,x), \alpha^k_0(y_0,x)\}_{k=1}^{K-1}\right)' \in C\left(\mathcal{YX}\right)^{2(K-1)}$, the term associated with $\widehat{F}^k_{Y_1|X}\left(\widehat{Q}^1_{Y_1|X}(\tau|x)|x\right)$ in

the image of $\Psi_{\tau x}^{(5)}(\alpha(y,x))$ is

$$f_{Y_1|X}^k\left(Q_{Y_1|X}^1(\tau|x)|x\right)\alpha_1^k\left(Q_{Y_1|X}^1(\tau|x),x\right)-\frac{f_{Y_1|X}^k\left(Q_{Y_1|X}^1(\tau|x)|x\right)}{f_{Y_1|X}^1\left(Q_{Y_1|X}^1(\tau|x)|x\right)}\alpha_1^1(Q_{Y_1|X}^1(\tau|x),x),$$

and the term associated with $\widehat{F}_{Y_0|X}^k\left(\widehat{Q}_{Y_0|X}^1(\tau|x)|x\right)$ is

$$f_{Y_0|\mathcal{C}_k}\left(Q_{Y_0|X}^1(\tau|x)|x\right)\alpha_0^k\left(F_{Y_0|\mathcal{C}_1}^{-1}(\tau),x\right)-\frac{f_{Y_0|X}^k\left(Q_{Y_0|X}^1(\tau|x)|x\right)}{f_{Y_0|X}^1\left(Q_{Y_0|X}^1(\tau|x)|x\right)}\alpha_0^1(Q_{Y_0|X}^1(\tau|x),x).$$

In summary,

$$\sqrt{n}\begin{pmatrix}\widehat{F}_{Y_1|X}^k\left(\widehat{Q}_{Y_1|X}^1(\tau|x)|x\right)-F_{U_1|X}^k(\tau|x)\\ \widehat{F}_{Y_0|X}^k\left(\widehat{Q}_{Y_0|X}^1(\tau|x)|x\right)-F_{U_0|X}^k(\tau|x)\end{pmatrix}\rightsquigarrow\Psi_{\tau x}^{(5)}\left(\left(\Psi_{yx}^{(2')}\left(\Psi_{yxp}^{(1)}\left(-J^{-1}(y)\mathbb{G}\left(\varphi_{y,\theta}\right)\right)\right)\right)\right)\Big|_{y_1=Q_{Y_1|X}^1(\tau|x),y_0=Q_{Y_0|X}^1(\tau|x)}$$

in $\ell^\infty\left(\mathcal{T}\mathcal{X}\right)^{2(K-2)}$, and

$$\begin{aligned}nT_{2n}^X &\equiv \sum_{k=2}^{K-1}\int_{\mathcal{X}}\int_{\mathcal{T}}Z_{kn}^{(2)}(\tau,x)^2\,d\tau d\widehat{F}_X(x)\\ &\rightsquigarrow \sum_{k=2}^{K-1}\int_{\mathcal{X}}\int_{\mathcal{T}}\left[\Psi_{k1\tau x}^{(5)}\left(\Psi_{Q_{Y_1|X}^1(\tau|x)x}^{(2')}\left(\Psi_{Q_{Y_1|X}^1(\tau|x)xp}^{(1)}\left(-J^{-1}(Q_{Y_1|X}^1(\tau|x))\mathbb{G}\left(\varphi_{Q_{Y_1|X}^1(\tau|x),\theta}(W)\right)\right)\right)\right)\right.\\ &\quad\left.-\Psi_{k2\tau x}^{(5)}\left(\Psi_{Q_{Y_0|X}^1(\tau|x)x}^{(2)}\left(\Psi_{Q_{Y_0|X}^1(\tau|x)xp}^{(1)}\left(-J^{-1}(Q_{Y_0|X}^1(\tau|x))\mathbb{G}\left(\varphi_{Q_{Y_0|X}^1(\tau|x),\theta}(W)\right)\right)\right)\right)\right]^2 d\tau dF_X(x)\\ &\equiv \sum_{k=2}^{K-1}\int_{\mathcal{X}}\int_{\mathcal{T}}Z_k^{(2)}(\tau,x)^2\,d\tau dF_X(x),\end{aligned}$$

where $\Psi_{\tau x}^{(5)}=\left\{\Psi_{k\tau x}^{(5)}\right\}_{k=2}^{K-1}$ with $\Psi_{k\tau x}^{(5)}=\left(\Psi_{k1\tau x}^{(5)},\Psi_{k2\tau x}^{(5)}\right)$. ■

Proof of Theorem 7. From Yu (2014a),

$$\sqrt{n}\begin{pmatrix}\widehat{F}_{Y_1|X,V}(y_1|x,p)\\ \widehat{F}_{Y_0|X,V}(y_0|x,p)\end{pmatrix}\rightsquigarrow\Psi_{yxp}^{(1)}\left(W_{\beta(y)}\right)\text{ in }\ell^\infty\left(\mathcal{Y}\mathcal{X}\mathcal{P}_x\right)^2,$$

where $\Psi_{yxp}^{(1)}:C\left(\mathcal{Y}\right)^{d_\beta}\longrightarrow\ell^\infty\left(\mathcal{Y}\mathcal{X}\mathcal{P}_x\right)^2$ is a linear map, and $W_{\beta(y)}=-\begin{pmatrix}J_0(y_0)^{-1}\left(J_{0p}(y_0)J_p^{-1}W_\gamma+W_0(y_0)\right)\\ J_1(y_1)^{-1}\left(J_{1p}(y_1)J_p^{-1}W_\gamma+W_1(y_1)\right)\end{pmatrix}$ is the weak limit of $\widehat{\beta}(y)$ as derived in (26). For $\alpha(y)=(\alpha_1(y_1),\alpha_0(y_0))\in C\left(\mathcal{Y}\right)^{d_\beta}$, the first element of $\Psi_{yxp}^{(1)}(\alpha(y))$ is defined as

$$\begin{aligned}&\left[\lambda\left(T(x,p)'\beta_1(y_1)\right)+p\frac{\partial T(x,p)'}{\partial p}\beta_1(y_1)\cdot\lambda'\left(T(x,p)'\beta_1(y_1)\right)\right]T(x,p)'\alpha_1(y_1)\\ &\qquad+p\frac{\partial T(x,p)'}{\partial p}\alpha_1(y_1)\cdot\lambda\left(T(x,p)'\beta_1(y_1)\right),\end{aligned}$$

and the second element is defined as

$$\left[\lambda\left(T\left(x,p\right)'\beta_0(y_0)\right)-(1-p)\frac{\partial T\left(x,p\right)'}{\partial p}\beta_0(y_0)\cdot\lambda'\left(T\left(x,p\right)'\beta_0(y_0)\right)\right]T(x,p)'\alpha_0(y_0)$$
$$-(1-p)\frac{\partial T\left(x,p\right)'}{\partial p}\alpha_0(y_0)\cdot\lambda\left(T\left(x,p\right)'\beta_0(y_0)\right).$$

Next, applying the functional Delta method again to have

$$\sqrt{n}\left(\begin{array}{c}\widehat{F}_{Y_1|X,V}\left(\widehat{F}^{-1}_{Y_1|X,V}(\tau|x,v_o)\middle|x,v\right)-F_{Y_1|X,V}\left(F^{-1}_{Y_1|X,V}(\tau|x,v_o)\middle|x,v\right)\\ \widehat{F}_{Y_0|X,V}\left(\widehat{F}^{-1}_{Y_0|X,V}\left(\tau|x,v_o\right)\middle|x,v\right)-F_{Y_0|X,V}\left(F^{-1}_{Y_0|X,V}(\tau|x,v_o)\middle|x,v\right)\end{array}\right)\rightsquigarrow\Psi^{(2)}_{\tau xv}\left(\Psi^{(1)}_{yxp}\left(W_{\beta(y)}\right)\right),$$

in $\ell^\infty\left(\mathcal{T}\mathcal{X}\mathcal{P}_x\right)^2$, where $\Psi^{(2)}_{\tau xv}:C\left(\mathcal{Y}\mathcal{X}\mathcal{P}_x\right)^2\to\ell^\infty\left(\mathcal{T}\mathcal{X}\mathcal{P}_x\right)^2$ is a linear map. For $\alpha\left(y,x,p\right)=\left(\alpha_1\left(y_1,x,p\right),\alpha_0\left(y_0,x,p\right)\right)\in C\left(\mathcal{Y}\mathcal{X}\mathcal{P}_x\right)^2$, the first element of $\Psi^{(2)}_{\tau xv}(\alpha\left(y,x,p\right))$ is defined as

$$f_{Y_1|X,V}\left(F^{-1}_{Y_1|X,V}(\tau|x,v_o)|x,v\right)\alpha_1\left(F^{-1}_{Y_1|X,V}(\tau|x,v_o),x,v\right)-\frac{f_{Y_1|X,V}\left(F^{-1}_{Y_1|X,V}(\tau|x,v_o)|x,v\right)}{f_{Y_1|X,V}\left(F^{-1}_{Y_1|X,V}(\tau|x,v_o)|x,v_o\right)}\alpha_1\left(F^{-1}_{Y_1|X,V}(\tau|x,v_o),x,v_o\right),$$

and the second element is defined as

$$f_{Y_0|X,V}\left(F^{-1}_{Y_0|X,V}(\tau|x,v_o)|x,v\right)\alpha_0\left(F^{-1}_{Y_0|X,V}(\tau|x,v_o),x,v\right)-\frac{f_{Y_0|X,V}\left(F^{-1}_{Y_0|X,V}(\tau|x,v_o)|x,v\right)}{f_{Y_0|X,V}\left(F^{-1}_{Y_0|X,V}(\tau|x,v_o)|x,v_o\right)}\alpha_0\left(F^{-1}_{Y_0|X,V}(\tau|x,v_o),x,v_o\right).$$

Finally, since $\mathcal{X}\mathcal{Z}$ is compact, $\{\widehat{p}_{ij}\}_{j=1}^n$ converges uniformly to $\{p_{ij}\}_{j=1}^n$ in probability. As a result, the empirical distribution of $\{X_i,\widehat{p}_{ij}\}$, say $\widehat{F}_{X,p(X,Z)}(x,v)$, converges uniformly to $F_{X,p(X,Z)}(x,v)$ in probability. Note that

$$\begin{aligned}nT^X_{3n}&=\int_{\mathcal{X}}\int_{\mathcal{P}_x}\int_{\mathcal{T}}\left[\sqrt{n}\left(\widehat{F}_{Y_1|X,V}\left(\widehat{F}^{-1}_{Y_1|X,V}(\tau|x,v_o)\middle|x,v\right)-\widehat{F}_{Y_0|X,V}\left(\widehat{F}^{-1}_{Y_0|X,V}\left(\tau|x,v_o\right)\middle|x,v\right)\right)\right]^2d\tau d\widehat{F}_{X,p(X,Z)}(x,v)+o_p(1)\\&=\int_{\mathcal{X}}\int_{\mathcal{P}_x}\int_{\mathcal{T}}\left[\sqrt{n}\left(\widehat{F}_{Y_1|X,V}\left(\widehat{F}^{-1}_{Y_1|X,V}(\tau|x,v_o)\middle|x,v\right)-\widehat{F}_{Y_0|X,V}\left(\widehat{F}^{-1}_{Y_0|X,V}\left(\tau|x,v_o\right)\middle|x,v\right)\right)\right]^2d\tau dF_{X,p(X,Z)}(x,v)+o_p(1),\end{aligned}$$

where the first $o_p(1)$ is due to $\frac{n}{n^2}\sum_{i=1}^n\int_{\mathcal{T}}[\widehat{F}_{Y_1|X,V}(\widehat{F}^{-1}_{Y_1|X,V}(\tau|X_i,v_o)|X_i,v_o)-\widehat{F}_{Y_0|X,V}(\widehat{F}^{-1}_{Y_0|X,V}\left(\tau|X_i,v_o\right)|X_i,v_o)]^2d\tau=o_p(1)$ uniformly in X_i, and the second $o_p(1)$ is due to $\int_{\mathcal{T}}[\widehat{F}_{Y_1|X,V}(\widehat{F}^{-1}_{Y_1|X,V}(\tau|x,v_o)|x,v)-\widehat{F}_{Y_0|X,V}(\widehat{F}^{-1}_{Y_0|X,V}\left(\tau|x,v_o\right)|x,v)]^2d\tau$ is tight in $\ell^\infty\left(\mathcal{X}\mathcal{P}_x\right)$ under H_0. By the continuous mapping theorem and Slutsky theorem,

$$nT^X_{3n}\rightsquigarrow\int_{\mathcal{X}}\int_{\mathcal{P}_x}\int_{\mathcal{T}}Z^{(3)}\left(\tau,x,v\right)^2d\tau dF_{X,p(X,Z)}(x,v),$$

where $Z^{(3)}\left(\tau,x,v\right)=\Psi^{(2)}_{1\tau xv}\left(\Psi^{(1)}_{yxp}\left(W_{\beta(y)}\right)\right)-\Psi^{(2)}_{2\tau xv}\left(\Psi^{(1)}_{yxp}\left(W_{\beta(y)}\right)\right)$ with $\Psi^{(2)}_{\tau xv}=\left(\Psi^{(2)}_{1\tau xv},\Psi^{(2)}_{2\tau xv}\right)$.

Under H_1^δ, we need only calculate $b^{(3)}\left(\tau,x,v\right)$. The calculation is an extension of the proof of Theorem 3(ii). ■

Proof of Theorem 8. We only outline the difference from the proof of Theorem 7. First define $Z_1^{(4)}\left(\tau,x,v\right)$ and $Z_2^{(4)}\left(\tau,x,v\right)$. They are the same as $Z^{(3)}\left(\tau,x,v\right)$ in the proof of Theorem 7 except that v_o in $Z^{(3)}\left(\tau,x,v\right)$

is replaced by v_1 for $Z_1^{(4)}(\tau,x,v)$ and is replaced by v_2 for $Z_2^{(4)}(\tau,x,v)$. Then

$$nT_{4n}^X \rightsquigarrow \max\left\{\int_{\mathcal{X}}\int_{v_o}^{\overline{p}_x}\int_{\mathcal{T}} Z_1^{(4)}(\tau,x,v)^2\, d\tau dF_{X,p(X,Z)}(x,v), \int_{\mathcal{X}}\int_{\underline{p}_x}^{v_o}\int_{\mathcal{T}} Z_2^{(4)}(\tau,x,v)^2\, d\tau dF_{X,p(X,Z)}(x,v)\right\}$$

under H_0. The derivation of the local power is also similar to that in the proof of Theorem 7(ii) except replacing v_o by v_1 and v_2 at suitable places. ■

Lemma 1 *Under the assumptions of Corollary 2,*

$$\begin{aligned}
\sqrt{n}\left(\widehat{F}_{Y_1|\mathcal{A}}(y_1) - F_{Y_1|\mathcal{A}}(y_1)\right) &\rightsquigarrow \mathbb{G}\left(\psi_1^{\mathcal{A}}(W,y_1)\right) \text{ in } \ell^\infty(\mathcal{Y}_1),\\
\sqrt{n}\left(\widehat{F}_{Y_d|\mathcal{C}_k}(y_d) - F_{Y_d|\mathcal{C}}(y_d)\right) &\rightsquigarrow \mathbb{G}\left(\psi_d^k(W,y_d)\right) \text{ in } \ell^\infty(\mathcal{Y}_d),\\
\sqrt{n}\left(\widehat{F}_{Y_0|\mathcal{N}}(y_0) - F_{Y_0|\mathcal{N}}(y_0)\right) &\rightsquigarrow \mathbb{G}\left(\psi_0^{\mathcal{N}}(W,y_0)\right) \text{ in } \ell^\infty(\mathcal{Y}_0),
\end{aligned}$$

and

$$\sqrt{n}\left(\widehat{p}_l - p_l\right) \rightsquigarrow \mathbb{G}\left(\phi_l(D,Z)\right), \sqrt{n}\left(\widehat{q}_l - q_l\right) \rightsquigarrow \mathbb{G}\left(\varphi_l(Z)\right).$$

where the weak convergence holds jointly, and the functions $\left(\psi_1^{\mathcal{A}}, \left\{\psi_1^k, \psi_0^k\right\}_{k=1}^{K-1}, \psi_1^{\mathcal{C}_k}, \{\phi_l, \varphi_l\}_{l=1}^K\right)$ *are defined in the proof.*

Proof. The building blocks for all these objects are a class of functions $\left\{\varphi_y(W) | y \in \mathcal{Y}_1 \times \mathcal{Y}_0\right\}$, where

$$\varphi_y(W) = \begin{pmatrix} 1(Y \le y_1) D1(Z = z_k) \\ 1(Y \le y_0)(1-D)1(Z = z_k) \\ D1(Z = z_k) \\ (1-D)1(Z = z_k) \\ 1(Z = z_k) \end{pmatrix}, k = 1, \cdots, K,$$

with $W = (D, Y, Z)$ and $y = (y_1, y_0)$ being the index, and φ_y is a $5K \times 1$ vector by stacking the z_k blocks.[18] Since the class of functions $\left\{\varphi_y(D,Y,Z) | y \in \mathcal{Y}_1 \times \mathcal{Y}_0\right\}$ is VC,

$$\mathbb{G}_n\left(\varphi_y(W)\right) \rightsquigarrow \mathbb{G}\left(\varphi_y(W)\right) \text{ in } \ell^\infty(\mathcal{Y})^{5K}$$

by Donsker's theorem, where $\mathcal{Y} \equiv \mathcal{Y}_1 \times \mathcal{Y}_0$. Now, by the functional Delta method (see, e.g., Theorem 3.9.4 of VW), we get

$$\sqrt{n}\begin{pmatrix} \widehat{F}_{Y_1|\mathcal{A}}(y_1) - F_{Y_1|\mathcal{A}}(y_1) \\ \widehat{F}_{Y_1|\mathcal{C}_k}(y_1) - F_{Y_1|\mathcal{C}_k}(y_1) \\ \widehat{F}_{Y_0|\mathcal{C}_k}(y_0) - F_{Y_0|\mathcal{C}_k}(y_0) \\ \widehat{F}_{Y_0|\mathcal{N}}(y_0) - F_{Y_0|\mathcal{N}}(y_0) \\ \widehat{p}_l - p_l \\ \widehat{q}_l - q_l \end{pmatrix} \rightsquigarrow \Psi_y^{(1)}\left(\mathbb{G}\left(\varphi_y(W)\right)\right) \text{ in } \ell^\infty(\mathcal{Y})^{4K},$$

where $\Psi_y^{(1)} : C(\mathcal{Y})^{5K} \to \ell^\infty(\mathcal{Y})^{4K}$ is the Hadamard derivative of $(F_{Y_1|\mathcal{A}}(y_1), \left\{F_{Y_1|\mathcal{C}_k}(y_1), F_{Y_0|\mathcal{C}_k}(y_0)\right\}_{k=1}^{K-1}$, $F_{Y_0|\mathcal{N}}(y_0), \{p_l, q_l\}_{l=1}^K)$ with respect to $E\left[\varphi_y\right]$. Note that $\widehat{F}_{Y_1|\mathcal{A}}(y_1)$ uses only $1(Y \le y_1) D1(Z = z_1)$ and

[18] In φ_y, some randomness is redundant, e.g., $D1(Z = z_k) + (1-D)1(Z = z_k) = 1(Z = z_k)$, but this form of φ_y is easier to use below.

$D1(Z=z_1)$, $F_{Y_1|\mathcal{C}_k}(y_1)$ uses only $1(Y\leq y_1)D1(Z=z_k), 1(Y\leq y_1)D1(Z=z_{k+1}), D1(Z=z_k)$ and $D1(Z=z_{k+1})$, $F_{Y_0|\mathcal{C}_k}(y_0)$ uses only $1(Y\leq y_0)(1-D)1(Z=z_k), 1(Y\leq y_0)(1-D)1(Z=z_{k+1}), (1-D)1(Z=z_k)$ and $(1-D)1(Z=z_{k+1})$, $F_{Y_0|\mathcal{N}}(y_0)$ uses only $1(Y\leq y_0)(1-D)1(Z=z_K)$ and $(1-D)1(Z=z_K)$, p_l uses only $D1(Z=z_l)$ and $1(Z=z_l)$, and q_l uses only $1(Z=z_k)$. This usage of information determines the structure of correlation between each pair of processes, e.g., p_l is only correlated with processes involving $D1(Z=z_l)$ and $1(Z=z_l)$ and q_l is only correlated with processes involving $1(Z=z_l)$. Given the correlation structure, we can state the weak limit of each process separately. Specifically,

$$\begin{aligned}
\psi_1^{\mathcal{A}}(W,y_1) &= \frac{\left[1(Y\leq y_1)-F_{Y_1|\mathcal{A}}(y_1)\right]}{E\left[D1(Z=z_1)\right]}D1(Z=z_1),\\
\psi_1^{\mathcal{C}_k}(W,y_1) &= \frac{\left(1(Y\leq y_1)-F_{Y_1|\mathcal{C}_k}(y_1)\right)D-E\left[\left(1(Y\leq y_1)-F_{Y_1|\mathcal{C}_k}(y_1)\right)D|Z=z_{k+1}\right]}{P(\mathcal{C}_k)q_{k+1}}1(Z=z_{k+1})\\
&\quad-\frac{\left(1(Y\leq y_1)-F_{Y_1|\mathcal{C}_k}(y_1)\right)D-E\left[\left(1(Y\leq y_1)-F_{Y_1|\mathcal{C}_k}(y_1)\right)D|Z=z_k\right]}{P(\mathcal{C}_k)q_k}1(Z=z_k)\\
\psi_0^{\mathcal{C}_k}(W,y_0) &= \frac{\left(1(Y\leq y_0)-F_{Y_0|\mathcal{C}_k}(y_0)\right)(1-D)-E\left[\left(1(Y\leq y_0)-F_{Y_0|\mathcal{C}_k}(y_0)\right)(1-D)|Z=z_k\right]}{P(\mathcal{C}_k)q_k}1(Z=z_k)\\
&\quad-\frac{\left(1(Y\leq y_0)-F_{Y_0|\mathcal{C}_k}(y_0)\right)(1-D)-E\left[\left(1(Y\leq y_0)-F_{Y_0|\mathcal{C}_k}(y_0)\right)(1-D)|Z=z_{k+1}\right]}{P(\mathcal{C}_k)q_{k+1}}1(Z=z_{k+1})
\end{aligned}$$

$$\psi_0^{\mathcal{N}}(W,y_0)=\frac{1(Y\leq y_0)-F_{Y_0|\mathcal{N}}(y_0)}{E\left[(1-D)1(Z=z_K)\right]}(1-D)1(Z=z_K),$$

and

$$\phi_l(D,Z)=\frac{D-p_l}{P(Z=z_l)}1(Z=z_l),\varphi_l(Z)=1(Z=z_l)-q_l,$$

where each term has mean zero, and the two terms in $\psi_d^{\mathcal{C}_k}(W,y_d)$ are uncorrelated. ■

Lemma 2 *Under the assumptions of Theorem 5,*

$$\begin{aligned}
&\left(\widehat{F}^0_{Y_1|X}(y_1|x),\left\{\widehat{F}^k_{Y_1|X}(y_1|x),\widehat{F}^k_{Y_0|X}(y_0|x)\right\}_{k=1}^{K-1},\widehat{F}^K_{Y_0|X}(y_0|x),\{\widehat{p}_{xk},\widehat{q}_{xk}\}_{k=1}^K\right)\\
&-\left(F^0_{Y_1|X}(y_1|x),\left\{F^k_{Y_1|X}(y_1|x),F^k_{Y_0|X}(y_0|x)\right\}_{k=1}^{K-1},F^K_{Y_0|X}(y_0|x),\{p_{xk},q_{xk}\}_{k=1}^K\right)\\
\rightsquigarrow\ &\Psi^{(2)}_{yx}\left(\Psi^{(1)}_{yxp}\left(-J^{-1}(y)\mathbb{G}\left(\varphi_{y,\theta}\right)\right)\right)\ \text{in } \ell^\infty(\mathcal{YX})^{4K},
\end{aligned}$$

where $\Psi^{(2)}_{yx},\Psi^{(1)}_{yxp},J(y)$ and $\varphi_{y,\theta}$ are defined in the proof.

Proof. The building blocks for all these objects are $\left\{\widehat{\beta}_d(y_d),\widehat{\gamma},\widehat{\eta}\right\}$. We apply Lemma E.3 of CFM to derive their weak limits. If we use the notation of CFM, $u=y\equiv(y_0,y_1)',\theta(u)=(\beta_1(y)',\beta_0(y)',\gamma',\eta')'\equiv(\beta(y)',\gamma',\eta')'\in\mathbb{R}^{d_\theta}$, and $\mathcal{U}=\mathcal{Y}\equiv\mathcal{Y}_1\mathcal{Y}_0=\{(y_1,y_0)|y_1\in\mathcal{Y}_1,y_0\in\mathcal{Y}_0\}$. Let

$$\varphi_{y,\theta}(W)=\begin{pmatrix}
D\left[\Lambda(T'\beta_1)-1(Y\leq y_1)\right]H(T'\beta_1)T,\\
(1-D)\left[\Lambda(T'\beta_0)-1(Y\leq y_0)\right]H(T'\beta_0)T,\\
(\widetilde{p}-D)H(R'\gamma)R,\\
\sum_{l=1}^K 1(Z=z_k)\frac{\partial q_k(B(X),\eta)/\partial\eta}{q_k(B(X),\eta)}
\end{pmatrix}$$

where $H(\cdot)=\lambda(\cdot)/\{\Lambda(\cdot)[1-\Lambda(\cdot)]\}$. Let $\Psi(\theta,y)=P\left(\varphi_{y,\theta}\right)$ and $\widehat{\Psi}(\theta,y)=\mathbb{P}_n\left(\varphi_{y,\theta}\right)$. From the first order conditions, $\widehat{\theta}(y)=\phi(\widehat{\Psi}(\theta,y),0)$ for each $y\in\mathcal{Y}$, where ϕ is the Z-map defined in Appendix E.1 of CFM.

Applying Lemma E.3 of CFM, we have

$$\sqrt{n}\left(\widehat{\theta}(y)-\theta(y)\right) \rightsquigarrow -J^{-1}(y)\mathbb{G}\left(\varphi_{y,\theta}\right) \text{ in } \ell^{\infty}\left(\mathcal{Y}\right)^{d_\theta},$$

where the four components of $\mathbb{G}\left(\varphi_{y,\theta}\right) \equiv \left(W_1(y_1)', W_0(y_0)', W_\gamma', W_\eta'\right)'$ are independent of each other, and

$$\dot{\Psi}_{\theta(y),y} = \begin{pmatrix} J_1(y_0) & \mathbf{0} & J_{1p}(y_0) & \mathbf{0} \\ \mathbf{0} & J_0(y_1) & J_{0p}(y_1) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & J_p & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & J_q \end{pmatrix} \equiv J(y),$$

with

$$J_{1p}(y_1) = E\left[\widetilde{p}\widetilde{\lambda}\lambda_1(y_1)H_1(y_1)\frac{\partial T(X,\widetilde{p})'\beta_1(y_1)}{\partial p}TR'\right],$$
$$J_{0p}(y_0) = E\left[(1-\widetilde{p})\widetilde{\lambda}\lambda_0(y_0)H_0(y_0)\frac{\partial T(X,\widetilde{p})'\beta_0(y_0)}{\partial p}TR'\right],$$

$H_d(y_d) = H(T'\beta_d(y_d))$ and other components of $J(y)$ being defined in the main text. The verification of conditions of Lemma E.3 is similar to that in the proof of Theorem 2 of Yu (2014a). $\widehat{\eta}$ is asymptotically independent of $\left(\widehat{\beta}_1(y_1), \widehat{\beta}_0(y_0), \widehat{\gamma}\right)$ which are dependent of each other. It can be further shown that

$$\sqrt{n}\left(\widehat{\beta}(y)-\beta(y)\right) \rightsquigarrow -\begin{pmatrix} J_1(y_1)^{-1}\left(J_{1p}(y_1)J_p^{-1}W_\gamma + W_1(y_1)\right) \\ J_0(y_0)^{-1}\left(J_{0p}(y_0)J_p^{-1}W_\gamma + W_0(y_0)\right) \end{pmatrix} \text{ in } \ell^{\infty}\left(\mathcal{Y}\right)^{d_\beta}. \tag{26}$$

Next, since $F_{Y|X,p(X,Z),D}(y_d|x,p,d) = \Lambda\left(T\left(x,p\right)'\beta_d(y_d)\right)$, $p_{xk} = \Lambda\left(R_k(x)'\gamma\right)$, $q_{xk} = q_k\left(B(x),\eta\right)$, $k = 1,\cdots,K-1$, and $q_K(x) = 1-\sum_{l=1}^{K-1}q_l(x)$, by the functional Delta method,

$$\sqrt{n}\begin{pmatrix} \widehat{F}_{Y|X,p(X,Z),D}(y_d|x,p,d) - F_{Y|X,p(X,Z),D}(y_d|x,p,d) \\ \widehat{p}_{xk} - p_{xk} \\ \widehat{q}_{xk} - q_{xk} \end{pmatrix} \rightsquigarrow \Psi^{(1)}_{yxp}\left(-J^{-1}(y)\mathbb{G}\left(\varphi_{y,\theta}\right)\right) \text{ in } \ell^{\infty}\left(\mathcal{YXP}_x\right)^{2K+2},$$

where the linear map $\Psi^{(1)}_{yxp}: C\left(\mathcal{Y}\right)^{d_\theta} \to \ell^{\infty}\left(\mathcal{YXP}_x\right)^{2K+2}$ is the Hadamard derivative of $(F_{Y|X,p(X,Z),D}(y_1|x,p,1), F_{Y|X,p(X,Z),D}(y_0|x,p,0), \{p_{xk}\}_{k=1}^K, \{q_k\left(x\right)\}_{k=1}^K)$ with respect to $\theta(y)$. More specifically, for $\alpha\left(y\right) = (\alpha_1(y_1)', \alpha_0(y_0)', \alpha_\gamma', \alpha_\eta') \in C\left(\mathcal{Y}\right)^{d_\theta}$, the first block of $\Psi^{(1)}_{yxp}\left(\alpha\left(y\right)\right)$ is $\lambda\left(T\left(x,p\right)'\beta_d(y_d)\right)T\left(x,p\right)'\alpha_d(y_d)$, the second block is $\Lambda\left(R(x,z_k)'\gamma\right)R(x,z_k)'\alpha_\gamma$, $k = 1,\cdots,K$, the third block is $\frac{\partial q_k(B(x),\eta)}{\partial\eta'}\alpha_\eta$, $k = 1,\cdots,K-1$, $-\sum_{l=1}^{K-1}\frac{\partial q_l(B(x),\eta)}{\partial\eta'}\alpha_\eta$, $k = K$. The block associated with $\{\widehat{q}_{xk}\}_{k=1}^K$ is dependent within the block but independent of other blocks which are dependent within and across blocks.

Finally, since $\left(F^0_{Y_1|X}(y_1|x), \left\{F^k_{Y_1|X}(y_1|x), F^k_{Y_0|X}(y_0|x)\right\}_{k=1}^{K-1}, F^K_{Y_0|X}(y_0|x), \{p_{xk}, q_{xk}\}_{k=1}^K\right)$ is Hadamard differentiable with respect to $(F_{Y|X,p(X,Z),D}(y_1|x,p,1),\ F_{Y|X,p(X,Z),D}(y_0|x,p,0), \{p_{xk}\}_{k=1}^K, \{q_{xk}\}_{k=1}^K)$, we can apply the functional Delta method again to derive the target weak limit. The remaining is to derive the Hadamard derivative, say, $\Psi^{(2)}_{yx}: C\left(\mathcal{YXP}_x\right)^{2K+2} \to \ell^{\infty}\left(\mathcal{YX}\right)^{4K}$. The Hadamard derivative corresponding to $\{p_{xk}, q_{xk}\}_{k=1}^K$ is trivial, so we concentrate on the F functions. For $\alpha\left(y,x,p\right) = (\alpha_1(y_1,x,p), \alpha_0(y_0,x,p), \{\alpha_{pk}\}_{k=1}^K,$ $\{\alpha_{qk}\}_{k=1}^K) \in C\left(\mathcal{YXP}_x\right)^{2K+2}$, the element of $\Psi^{(2)}_{yx}\left(\alpha\left(y,x,p\right)\right)$ corresponding to $F^0_{Y_1|X}(y_1|x)$ is

$$\alpha_1(y_1,x,p_{x1}) + \lambda\left(T\left(x,p_{x1}\right)'\beta_1(y_1)\right)\frac{\partial T\left(x,p_{x1}\right)'}{\partial p}\beta_1(y_1)\alpha_{p1},$$

the element corresponding to $F^k_{Y_1|X}(y_1|x)$, $k=1,\cdots,K-1$, is

$$\frac{p_{x,k+1}\alpha_1(y_1,x,p_{x,k+1})-p_{xk}\alpha_0(y_0,x,p_{xk})}{p_{x,k+1}-p_{xk}}+\frac{F_{Y|X,p(X,Z),D}(y_1|x,p_{x,k+1},1)+p_{x,k+1}\lambda\left(T(x,p_{x,k+1})'\beta_1(y_1)\right)\frac{\partial T\left(x,p_{x,k+1}\right)'}{\partial p}\beta_1(y_1)-F^k_{Y_1|X}(y_1|x)}{p_{x,k+1}-p_{xk}}\alpha_{p,k+1}$$
$$-\frac{F_{Y|X,p(X,Z),D}(y_1|x,p_{xk},1)+p_{xk}\lambda\left(T(x,p_{xk})'\beta_1(y_1)\right)\frac{\partial T(x,p_{xk})'}{\partial p}\beta_1(y_1)-F^k_{Y_1|X}(y_1|x)}{p_{x,k+1}-p_{xk}}\alpha_{pk},$$

the element corresponding to $F^k_{Y_0|X}(y_0|x)$, $k=1,\cdots,K-1$, is

$$\frac{(1-p_{xk})\alpha_0(y_0,x,p_{xk})-(1-p_{x,k+1})\alpha_0(y_0,x,p_{x,k+1})}{p_{x,k+1}-p_{xk}}-\frac{F_{Y|X,p(X,Z),D}(y_0|x,p_{xk},0)-(1-p_{xk})\lambda\left(T(x,p_{xk})'\beta_0(y_0)\right)\frac{\partial T(x,p_{xk})'}{\partial p}\beta_0(y_0)-F^k_{Y_0|X}(y_0|x)}{p_{x,k+1}-p_{xk}}\alpha_{pk}$$
$$+\frac{F_{Y|X,p(X,Z),D}(y_0|x,p_{x,k+1},0)-(1-p_{x,k+1})\lambda\left(T(x,p_{x,k+1})'\beta_0(y_0)\right)\frac{\partial T\left(x,p_{x,k+1}\right)'}{\partial p}\beta_0(y_0)-F^k_{Y_0|X}(y_0|x)}{p_{x,k+1}-p_{xk}}\alpha_{p,k+1},$$

and the element corresponding to $F^K_{Y_0|X}(y_0|x)$ is

$$\alpha_0(y_0,x,p_{xK})+\lambda\left(T\left(x,p_{xK}\right)'\beta_0(y_0)\right)\frac{\partial T\left(x,p_{xK}\right)'}{\partial p}\beta_0(y_0)\alpha_{pK}.$$

Different from the no-covariate case, except $\{\widehat{q}_{xk}\}_{k=1}^K$ all other estimators are correlated with each other. ■

Supplementary Material S.2

S.2.1 Bootstrapping Critical Values of T^X_{1n} and T^X_{2n}

We still suggest to use the exchangeable bootstrap to obtain the critical values for T^X_{1n} and T^X_{2n}. To ease implementation, we provide the detailed bootstrap procedure below; all notational conventions follow Section 4.3.

Step 1: Let

$$\begin{aligned}
\widehat{F}^*_{Y|X,p(X,Z),D}(y_d|x,p,d) &= \Lambda\left(T\left(x,p\right)'\widehat{\beta}^*_d(y_d)\right),\\
\widehat{p}^*_{xk} &\equiv \widehat{p}^*\left(x,z_k\right)=\Lambda\left(R(x,z_k)'\widehat{\gamma}^*\right),k=1,\cdots,K,\\
\widehat{q}^*_{xk} &\equiv \widehat{q}^*_k\left(x\right)=q_k\left(B(x),\widehat{\eta}^*\right),k=1,\cdots,K-1,\\
\widehat{q}^*_{xK} &= 1-\sum\nolimits_{l=1}^{K-1}\widehat{q}^*_{xl},
\end{aligned}$$

where

$$\widehat{\beta}^*_d(y_d)=\arg\max_{\beta}\sum_{i=1}^n\omega_i 1(D_i=d)\left[1\left(Y_i\le y_d\right)\ln\Lambda\left(T\left(X_i,\widehat{p}^*_i\right)'\beta\right)+1\left(Y_i>y_d\right)\ln\left(1-\Lambda\left(T\left(X_i,\widehat{p}^*_i\right)'\beta\right)\right)\right]$$

with $\widehat{p}^*_i=\widehat{p}^*\left(X_i,Z_i\right)$,

$$\widehat{\gamma}^*=\arg\max_{\gamma}\sum_{i=1}^n\omega_i\left[D_i\ln\Lambda\left(R(X_i,Z_i)'\gamma\right)+\left(1-D_i\right)\ln\left(1-\Lambda\left(R(X_i,Z_i)'\gamma\right)\right)\right],$$

and

$$\widehat{\eta}^*=\arg\max_{\eta}\sum_{i=1}^n\sum_{k=1}^K\omega_i 1\left(Z_i=z_k\right)\ln q_k\left(B(X_i),\eta\right),$$

with $q_K = 1 - \sum_{l=1}^{K-1} q_l$.

Step 2: Let

$$\begin{aligned}
\widehat{F}^{k*}_{Y_1|X}(y_1|x) &= \frac{\widehat{F}^{*}_{Y|X,p(X,Z),D}(y_1|x,\widehat{p}^{*}_{x,k+1},1)\widehat{p}^{*}_{x,k+1} - \widehat{F}^{*}_{Y|X,p(X,Z),D}(y_1|x,\widehat{p}^{*}_{xk},1)\widehat{p}^{*}_{xk}}{\widehat{p}^{*}_{x,k+1} - \widehat{p}^{*}_{xk}}, \\
\widehat{F}^{k*}_{Y_0|X}(y_0|x) &= \frac{\widehat{F}^{*}_{Y|X,p(X,Z),D}(y_0|x,\widehat{p}^{*}_{xk},0)\left(1-\widehat{p}^{*}_{xk}\right) - \widehat{F}^{*}_{Y|X,p(X,Z),D}(y_0|x,\widehat{p}^{*}_{x,k+1},0)\left(1-\widehat{p}^{*}_{x,k+1}\right)}{\widehat{p}^{*}_{x,k+1} - \widehat{p}^{*}_{xk}},
\end{aligned}$$

for $k = 1, \cdots, K-1$,

$$\begin{aligned}
\widehat{F}^{0*}_{Y_1|X}(y_1|x) &= \widehat{F}^{*}_{Y|X,p(X,Z),D}(y_1|x,\widehat{p}^{*}_{x1},1), \\
\widehat{F}^{K*}_{Y_0|X}(y_0|x) &= \widehat{F}^{*}_{Y|X,p(X,Z),D}(y_0|x,\widehat{p}^{*}_{xK},0),
\end{aligned}$$

and

$$\widehat{h}^{*}(\widehat{p}^{*}_{x,k+1}) = \frac{\sum_{l=k+1}^{K}\widehat{q}^{*}_{xl}\left(\widehat{p}^{*}_{xl} - \overline{\widehat{p}}^{*}_{x}\right)}{\sum_{l=1}^{K}\widehat{q}^{*}_{xl}\left(\widehat{p}^{*}_{xl} - \overline{\widehat{p}}^{*}_{x}\right)^2}$$

where $\overline{\widehat{p}}^{*}_{x} = \sum_{l=1}^{K}\widehat{q}^{*}_{xl}\widehat{p}^{*}_{xl}$. Conduct rearrangement if $\widehat{F}^{k*}_{Y_d|X}(y_d|x)$ is not monotone.

Step 3: Let

$$\begin{aligned}
\widehat{F}^{*}_{Y_1|X}(y_1|x) = \sum_{k=0}^{K}\left(\widehat{p}^{*}_{x,k+1} - \widehat{p}^{*}_{xk}\right)\Big\{&\widehat{F}^{k*}_{Y_1|X}(y_1|x)\left[1 - \widehat{F}^{*}_{Z|X}(z_k|x)\right] + \\
&\widehat{F}^{k*}_{Y_0|X}\left(\widehat{\widetilde{F}}^{*-1}_{Y_0|X}\left(\widehat{\widetilde{F}}^{*}_{Y_1|X}\left(y_1|x\right)\Big| x\right)\Big| x\right)\widehat{F}^{*}_{Z|X}(z_k|x)\Big\},
\end{aligned}$$

and

$$\begin{aligned}
\widehat{F}^{*}_{Y_0|X}(y_0|x) = \sum_{k=0}^{K}\left(\widehat{p}^{*}_{x,k+1} - \widehat{p}^{*}_{xk}\right)\Big\{&\widehat{F}^{k*}_{Y_0|X}(y_0|x)\widehat{F}^{*}_{Z|X}(z_k|x) + \\
&\widehat{F}^{k*}_{Y_1|X}\left(\widehat{\widetilde{F}}^{*-1}_{Y_1|X}\left(\widehat{\widetilde{F}}^{*}_{Y_0|X}\left(y_0|x\right)\Big| x\right)\Big| x\right)\left[1 - \widehat{F}^{*}_{Z|X}(z_k|x)\right]\Big\},
\end{aligned}$$

which are consistent to $\widehat{F}_{Y_d|X}(y_d|x)$, where

$$\widehat{F}^{*}_{Z|X}\left(z_k|x\right) = \sum_{l=1}^{k}\widehat{q}^{*}_{xl}, \widehat{\widetilde{F}}^{*}_{Y_d|X}\left(y_d|x\right) = \sum_{k=1}^{K-1}\left(\widehat{p}^{*}_{x,k+1} - \widehat{p}^{*}_{xk}\right)\widehat{h}^{*}(\widehat{p}^{*}_{x,k+1})\widehat{F}^{k*}_{Y_d|X}(y_d|x).$$

Step 4: Let

$$\begin{aligned}
T^{X*}_{1n} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K-1}\int_{\mathcal{T}}\Big[&\left(\widehat{F}^{k*}_{Y_1|X}\left(\widehat{F}^{*-1}_{Y_1|X}(\tau|X_i)|X_i\right) - \widehat{F}^{k}_{Y_1|X}\left(\widehat{F}^{-1}_{Y_1|X}(\tau|X_i)\Big| X_i\right)\right) \\
&- \left(\widehat{F}^{k*}_{Y_0|X}\left(\widehat{F}^{*-1}_{Y_0|X}(\tau|X_i)|X_i\right) - \widehat{F}^{k}_{Y_0|X}\left(\widehat{F}^{-1}_{Y_0|X}(\tau|X_i)\Big| X_i\right)\right)\Big]^2 d\tau
\end{aligned}$$

and

$$\begin{aligned}
T^{X*}_{2n} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=2}^{K-1}\int_{\mathcal{T}}\Big[&\widehat{F}^{k*}_{Y_1|X}\left(\widehat{Q}^{1*}_{Y_1|X}(\tau|X_i)|X_i\right) - \widehat{F}^{k}_{Y_1|X}\left(\widehat{Q}^{1}_{Y_1|X}(\tau|X_i)\Big| X_i\right) \\
&- \left(\widehat{F}^{k*}_{Y_0|X}\left(\widehat{Q}^{1*}_{Y_0|X}(\tau|X_i)|X_i\right) - \widehat{F}^{k}_{Y_0|X}\left(\widehat{Q}^{1}_{Y_0|X}(\tau|X_i)\Big| X_i\right)\right)\Big]^2 d\tau,
\end{aligned}$$

where X_i can be replaced by X_i^* (i.e., the weight ω_i can be imposed on X_i). Use the $(1-\alpha)$th quantile of $n^* T_{1n}^{X*}$, say $c_{1n}^{*X}(\alpha)$, and $(1-\alpha)$th quantile of $n^* T_{2n}^{X*}$, say $c_{2n}^{*X}(\alpha)$, as the critical values for nT_{1n}^X and nT_{2n}^X, respectively.

For completeness, we state the validity of the bootstrap procedure above without proof.

Theorem 9 *In the framework (5), suppose Assumptions M, DR and EB hold.*

(i) *Under H_0,*

$$\lim_{n\to\infty} P\left(nT_{1n}^X > c_{1n}^{*X}(\alpha)\right) = \alpha.$$

(ii) *Under H_1^δ and Assumption LA,*

$$\lim_{n\to\infty} P\left(nT_{1n}^X > c_{1n}^{*X}(\alpha)\right) \geq \alpha.$$

(iii) *Under the fixed alternative H_1 with $T_1^X = p\lim_{n\to\infty} T_{1n}^X > 0$,*

$$\lim_{n\to\infty} P\left(nT_{1n}^X > c_{1n}^{*X}(\alpha)\right) = 1.$$

(iv) *(i)-(iii) hold also for T_{2n}^X and $c_{2n}^{*X}(\alpha)$.*

S.2.2 Bootstrapping Critical Values of T_{3n}^X and T_{4n}^X

We still suggest to use the exchangeable bootstrap to obtain the critical values for T_{3n}^X and T_{4n}^X. To ease implementation, we provide the detailed bootstrap procedure below; all notational conventions follow Section 4.3.

Step 1: Let

$$\begin{aligned}\widehat{F}^*_{Y|X,p(X,Z),D}(y_d|x,p,d) &= \Lambda\left(T(x,p)'\,\widehat{\beta}_d^*(y_d)\right),\\ \widehat{p}^*(x,z) &= \Lambda\left(R(x,z)'\widehat{\gamma}^*\right),\end{aligned}$$

where

$$\widehat{\beta}_d^*(y_d) = \arg\max_\beta \sum_{i=1}^n \omega_i 1(D_i=d)\left[1(Y_i\leq y_d)\ln\Lambda\left(T(X_i,\widehat{p}_i^*)'\beta\right) + 1(Y_i>y_d)\ln\left(1-\Lambda\left(T(X_i,\widehat{p}_i^*)'\beta\right)\right)\right]$$

with $\widehat{p}_i^* = \widehat{p}^*(X_i,Z_i)$, and

$$\widehat{\gamma}^* = \arg\max_\gamma \sum_{i=1}^n \omega_i\left[D_i\ln\Lambda\left(R(X_i,Z_i)'\gamma\right) + (1-D_i)\ln\left(1-\Lambda\left(R(X_i,Z_i)'\gamma\right)\right)\right].$$

Step 2: Let

$$\widehat{F}^*_{Y_1|X,V}(y_1|x,v) = \widehat{F}^*_{Y|X,p(X,Z),D}(y_1|x,v,1) + v\left.\frac{\partial T(x,p)'}{\partial p}\right|_{p=v}\widehat{\beta}_1^*(y_1)\cdot\lambda\left(T(x,v)'\,\widehat{\beta}_1^*(y_1)\right),$$

and

$$\widehat{F}^*_{Y_0|X,V}(y_0|x,v) = \widehat{F}^*_{Y|X,p(X,Z),D}(y_0|x,v,0) - (1-v)\left.\frac{\partial T(x,p)'}{\partial p}\right|_{p=v}\widehat{\beta}_0^*(y_0)\cdot\lambda\left(T(x,v)'\,\widehat{\beta}_0^*(y_0)\right).$$

Conduct rearrangement if $\widehat{F}^*_{Y_d|X,V}(y_d|x,v)$ is not monotone.

Step 3: Let

$$T_{3n}^{X*} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j:\widehat{p}_{ij}\neq v_o}\int_{\mathcal{T}}\Big[\widehat{F}^*_{Y_1|X,V}\left(\widehat{F}^{*-1}_{Y_1|X,V}(\tau|X_i,v_o)\Big|\,X_i,\widehat{p}_{ij}\right)-\widehat{F}_{Y_1|X,V}\left(\widehat{F}^{-1}_{Y_1|X,V}(\tau|X_i,v_o)\Big|\,X_i,\widehat{p}_{ij}\right)$$
$$-\left(\widehat{F}^*_{Y_0|X,V}\left(\widehat{F}^{*-1}_{Y_0|X,V}(\tau|X_i,v_o)\Big|\,X_i,\widehat{p}_{ij}\right)-\widehat{F}_{Y_0|X,V}\left(\widehat{F}^{-1}_{Y_0|X,V}(\tau|X_i,v_o)\Big|\,X_i,\widehat{p}_{ij}\right)\right)\Big]^2 d\tau$$

and

$$T_{4n}^{X*}=\max\Bigg\{\tfrac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}1(\widehat{p}_{ij}\geq v_o)\int_{\mathcal{T}}\Big[\widehat{F}^*_{Y_1|X,V}\left(\widehat{F}^{*-1}_{Y_1|X,V}(\tau|X_i,v_1)\Big|\,X_i,\widehat{p}_{ij}\right)-\widehat{F}_{Y_1|X,V}\left(\widehat{F}^{-1}_{Y_1|X,V}(\tau|X_i,v_1)\Big|\,X_i,\widehat{p}_{ij}\right)$$
$$-\left(\widehat{F}_{Y_0|X,V}\left(\widehat{F}^{-1}_{Y_0|X,V}(\tau|X_i,v_1)\Big|\,X_i,\widehat{p}_{ij}\right)-\widehat{F}^*_{Y_0|X,V}\left(\widehat{F}^{*-1}_{Y_0|X,V}(\tau|X_i,v_1)\Big|\,X_i,\widehat{p}_{ij}\right)\right)\Big]^2 d\tau,$$
$$\tfrac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}1(\widehat{p}_{ij}\leq v_o)\int_{\mathcal{T}}\Big[\widehat{F}^*_{Y_1|X,V}\left(\widehat{F}^{*-1}_{Y_1|X,V}(\tau|X_i,v_2)\Big|\,X_i,\widehat{p}_{ij}\right)-\widehat{F}_{Y_1|X,V}\left(\widehat{F}^{-1}_{Y_1|X,V}(\tau|X_i,v_2)\Big|\,X_i,\widehat{p}_{ij}\right)$$
$$-\left(\widehat{F}^*_{Y_0|X,V}\left(\widehat{F}^{*-1}_{Y_0|X,V}(\tau|X_i,v_2)\Big|\,X_i,\widehat{p}_{ij}\right)-\widehat{F}_{Y_0|X,V}\left(\widehat{F}^{-1}_{Y_0|X,V}(\tau|X_i,v_2)\Big|\,X_i,\widehat{p}_{ij}\right)\right)\Big]^2 d\tau\Bigg\},$$

where $(X_i,\widehat{p}_{ij})$ can be replaced by $\left(X_i^*,\widehat{p}_{ij}^*\right)$ with $\widehat{p}_{ij}^*=\widehat{p}^*\left(X_i,Z_j\right)$. Use the $(1-\alpha)$th quantile of $n^*T_{3n}^{X*}$, say $c_{3n}^{*X}(\alpha)$, and $(1-\alpha)$th quantile of $n^*T_{4n}^{X*}$, say $c_{4n}^{*X}(\alpha)$, as the critical values for nT_{3n}^X and nT_{4n}^X, respectively.

For completeness, we state the validity of the bootstrap procedure above without proof.

Theorem 10 *In the framework (5), suppose Assumptions M, DR′, P and EB hold.*

(i) *Under H_0,*

$$\lim_{n\to\infty}P\left(nT_{3n}^X>c_{3n}^{*X}(\alpha)\right)=\alpha.$$

(ii) *Under H_1^δ and Assumption LA,*

$$\lim_{n\to\infty}P\left(nT_{3n}^X>c_{3n}^{*X}(\alpha)\right)\geq\alpha.$$

(iii) *Under the fixed alternative H_1 with $T_3^X=p\lim_{n\to\infty}T_{3n}^X>0$,*

$$\lim_{n\to\infty}P\left(nT_{3n}^X>c_{3n}^{*X}(\alpha)\right)=1.$$

(iv) *(i)-(iii) hold also for T_{4n}^X and $c_{4n}^{*X}(\alpha)$.*