Testing Unconditional Rank Preservation
Under Unconfoundedness[*]

Ping Yu[†]

The University of Hong Kong

Started: July 2013

First Version: April 2015

This Version: February 2017

Abstract

The assumption of rank preservation is important for causal interpretation of quantile treatment effects, but it is commonly believed that this assumption cannot be tested due to the missing data problem of causal inference. In this paper, we propose Hausman-type tests to test unconditional rank preservation under unconfoundedness when covariates are available. One key advantage of our tests is that the powers can be intuitively detected by figures. The basic idea is that unconditional rank preservation implies conditional rank preservation but the converse is not true, so significant difference between two statistics with one preserving conditional rank and the other preserving unconditional rank is an indicator of rank nonpreservation. In other words, we are testing rank preservation across covariate values rather than within a covariate value. We develop both parametric and nonparametric tests for both the overall quantile treatment effect and the quantile treatment effect on the treated. Since the asymptotic null distributions are nonstandard, we suggest to use the exchangeable bootstrap in the parametric tests and simulation in the nonparametric tests to obtain critical values. We illustrate our tests on data from the National Supported Work Program.

KEYWORDS: rank preservation, quantile treatment effects, Hausman test, degenerate U-statistic, bootstrap, simulation

JEL-CLASSIFICATION: C12, C21

1 Introduction

Quantile treatment effects (QTEs), as an alternative of average treatment effects, have attracted much attention in recent developments of program evaluation; see Abbring and Heckman (2007) and Yu (2014) for a summary of relevant literature. Let Y_1 and Y_0 be the potential outcomes under the treatment status and the control status, respectively; then the QTE is the quantile of $Y_1 - Y_0$, which involves the joint distribution of Y_1 and Y_0. However, due to the fundamental problem of causal inference (Holland, 1986), i.e., Y_1 and Y_0 cannot be observed simultaneously, the QTE cannot be point identified generally. Instead of studying the *quantile of differences* of Y_1 and Y_0, Lehmann (1974) and Doksum (1974) suggested to study the *difference of quantiles* of Y_1 and Y_0, which requires only the marginal distributions of Y_1 and Y_0 and so can be point identified. This solution is built on a key assumption - the ranks of Y_1 and Y_0 are the same; we label this assumption as the *rank preservation* (RP) assumption. Although the marginal distributions of Y_1 and Y_0 have their independent roles in treatment effects evaluation, e.g., testing stochastic dominance, it should be emphasized that only under the RP assumption, the difference of quantiles of Y_1 and Y_0 has a causal interpretation. This paper is about testing this key assumption.

It is commonly believed that the RP assumption cannot be tested. This is partially true because Y_1 and Y_0 cannot be observed simultaneously neither can their ranks. Due to this missing data problem, it is hard to judge whether the rank is preserved (see, e.g., Heckman et al., 1997). Figure 1 illustrates this point intuitively. In Figure 1, the solid points represent observables, the circles represent unobservables, the arrows match the identities of Y_0 and Y_1, and supp(Y_1) =supp(Y_0) = $\{0,1,2,3\}$, where supp(Y_d), $d=0,1$, is the support of Y_d. Obviously, given the same data set, the rank can be preserved, unpreserved, or unpreserved for the population but preserved for the treated. Nevertheless, we show in this paper that when covariates are available, as in most applications, we can test whether the rank is preserved in the population or among the treated under unconfoundedness. In other words, our tests are *unconditional* tests rather than *conditional* tests given each covariate value because the latter tests are not possible from Figure 1.
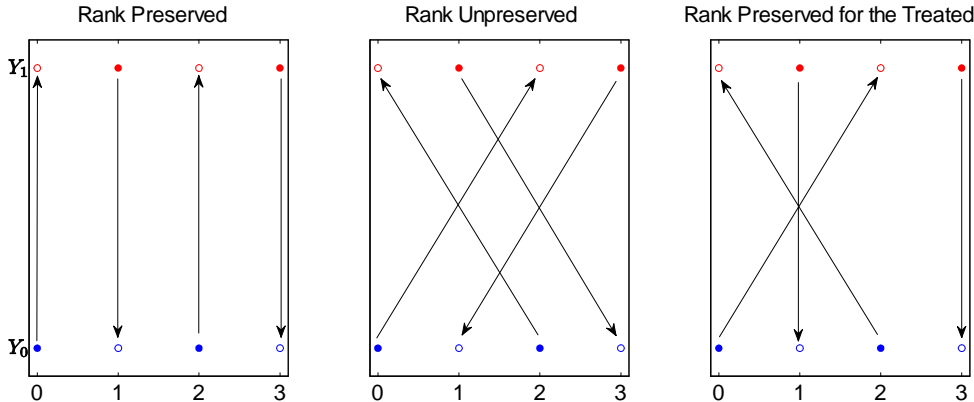

Figure 1: Rank can be either Preserved or Unpreserved or Only Preserved for the Treated: $Y_d \in \{0,1,2,3\}$

We first in Section 2 explain the role of the RP assumption in treatment effects evaluation. It turns out that combined with unconfoundedness, the RP assumption can fully determine the joint distribution of (Y_1, Y_0, D), where D is the treatment status. More importantly, the RP assumption is not only sufficient but necessary for the quantile of the impact distribution $Q_{Y_1-Y_0}(\cdot)$ to be identified solely from the difference of marginal quantiles $Q_{Y_1}(\cdot) - Q_{Y_0}(\cdot)$, where for a random variable Y, $Q_Y(\tau)$ means its τth quantile. Further,

even if $Q_{Y_1-Y_0}(\tau)$ can be identified from $Q_{Y_1}(\cdot) - Q_{Y_0}(\cdot)$, it need not equal $Q_{Y_1}(\tau) - Q_{Y_0}(\tau)$.

We then in Section 3 overview our testing ideas. To facilitate our discussion, we introduce some notations here. By the Skorohod representation, we can represent

$$Y_d = Q_d(U_d^X|X), d = 0, 1,$$

where $Q_d(\tau|x)$ is the τth conditional quantile of Y_d given $X = x$, and U_d^x is the rank variable of Y_d for the subpopulation $X = x$. U_d^x represents some unobserved characteristic of $Y_d|_{X=x}$, e.g., ability or proneness, and $U_d^x \sim U(0,1)$ for any $x \in \text{supp}(X)$, where $Y_d|_{X=x}$ represents the Y_d values for the subpopulation $X = x$.[1] Here, U_0^X need not equal U_1^X if the conditional rank is unpreserved. We can also represent Y_d in the unconditional form,

$$Y_d = q_d(U_d),$$

where $q_d(\tau)$ is the τth unconditional quantile of Y_d, and $U_d \sim U(0,1)$ is the rank variable of Y_d.[2] In other words, we sort individuals among the whole population rather than within $X = x$. Given these notations, we can state the null hypothesis of RP for the whole population as

$$H_0 : U_0 = U_1$$

and the alternative hypothesis as

$$H_1 : U_0 \neq U_1,$$

where the equality and inequality are understood to hold only almost surely.

Suppose X is a set of observable covariates with support $\text{supp}(X)$. If the ranks of Y_1 and Y_0 are preserved for the whole population, then the ranks must be preserved for any subpopulation $X = x$, $x \in \text{supp}(X)$. The converse is not true, i.e., if the ranks for any subpopulation $X = x$ are preserved, the unconditional rank need not be preserved. This is the basis of the tests developed in this paper. As argued above, the RP *within* the subpopulation $X = x$ (or the *conditional* RP $U_0^x = U_1^x$) cannot be tested, so any violation of RP we can identify is only the violation *across* different subpopulations rather than within any subpopulation. Actually, if we maintain conditional RP, then we can show that the unconditional RP is equivalent to RP across X values. Given this observation, we construct a Hausman-type test statistic. Specifically, we construct the counterfactuals of Y_0 in the treatment status under $U_0 = U_1$ and under $U_0^x = U_1^x$, respectively. The former is valid under both H_0 and H_1 while the latter is valid only under H_1, so significance of their difference is an indicator of the violation of H_0. Moreover, we can show that the population version of our test statistic equals zero if and only if the rank is preserved across X values. We further extend this testing idea to test RP for the treated.

Note that our null hypothesis is refutable but nonverifiable; see Breusch (1986) for a general discussion and Kitagawa (2015) for a recent example of this kind of hypothesis. In other words, if the null is rejected, then we are sure that the unconditional rank is not preserved; while if the null is not rejected, the unconditional rank may or may not be preserved (of course, the reason of nonpreservation must be attributed to within-X nonpreservation rather than across-X nonpreservation).

Our test statistic involves estimators of the counterfactuals of Y_0 under $U_0 = U_1$ and under $U_0^x = U_1^x$. In Sections 4 and 5, we provide parametric and nonparametric estimators for these counterfactuals, respectively. Both parametric and nonparametric test statistics are third-order degenerate V-statistics under H_0, but

---

[1] Note that $Y_d|_{X=x}$ is a conditional random variable. Since X is not independent of Y_d, this notation is different from the potential outcome Y_d when X is exogenously assigned as x, which is usually denoted as Y_{dx}.

[2] Be cautious that $U_d^x \neq U_d|_{X=x}$ because the latter generally does not follow the $U(0,1)$ distribution. $U_d|_{X=x}$ includes more information than U_d^x because it also includes the relative location of an individual in the whole population.

their null asymptotic distributions are different. The parametric test statistic follows a mixed chi-square distribution asymptotically, while the nonparametric test statistic is asymptotically normal with a positive mean. The critical values for the former are hard to obtain, so we suggest to use the exchangeable bootstrap to get them. The convergence rate to normal distribution in the latter statistic is slow, so we suggest to use the simulation method to obtain its critical values. We also show that our bootstrap and simulation schemes are valid. In Section 6, we extend our tests in Sections 4 and 5 to test RP among the treated. Section 7 includes some further discussions on our tests, e.g., overidentification interpretation of our tests, modifications, alternative forms, and extensions. Section 8 provides some simulation results, Section 9 applies our tests to a dataset from the National Supported Work Program, 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 the associated lemmas, and S.2 contains some discussions that we do not want to expand in the main text.

We close this introduction by discussing some recent related literature on RP testing. First, we distinguish two terms - *rank invariance* and *rank similarity* (RS). These two terms are introduced by Chernozhukov and Hansen (2005) to identify the QTE when endogeneity is present. The rank invariance assumption is our RP assumption, and the RS assumption is a weaker version of the rank invariance assumption. The term "rank preservation" is borrowed from Firpo (2007), where the author studies the estimation and inference of the unconditional QTE and unconditional quantile treatment effect on the treated (QTT) under the "unconditional" RP assumption and unconfoundedness, so the tests developed in this paper can serve as pretests to Firpo's estimation. On the other hand, Chernozhukov and Hansen (2005) show that the "conditional" RS assumption is enough to identify the marginal quantiles of Y_1 and Y_0 in the presence of endogeneity. To verify the conditional RS assumption, a parallel paper by Yu (2016) proposes two tests where, unlike in this paper, *no* covariates are required. These tests can serve as pretests to Chernozhukov and Hansen's estimation. Recently, an independent paper by Dong and Shen (2015) tests "unconditional" RS under both unconfoundedness and confoundedness (i.e., endogeneity); see also Frandsen and Lefgren (2015) for a regression-based implementation. Note that the unconditional RS assumption is weaker than the unconditional RP assumption and is stronger than the conditional RS assumption. Since "unconditional" RS is tested, covariates are also required as in this paper. As argued above, to have causal interpretation for unconditional QTE, rank preservation or rank invariance is a must. So these authors are testing a weaker version or an implication of the required condition, which is definitely interesting but different from the goals of this paper and Yu (2016).[3] Since the motivations and testing procedures of this paper and Yu (2016) are different from those of these authors, our results are more complements than substitutes to theirs.

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. 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 two random vectors, Y and X, $Y|_{X=x}$ is the random vector Y restricted at $X = x$, F_{YX} is the joint cumulative distribution function (cdf) of (Y, X), and $F_{Y|X}$ is the conditional cdf of Y given X. For three random vectors, X, Y and Z, $X \perp Y | Z$ means X is independent of Y conditional on Z, where "$\perp$" denotes independence (c.f., Dawid, 1979) and variables to the right of "|" are the conditioning variables. $\mathcal{T}$ with a subscript or superscript is a compact subset of $[\epsilon, 1 - \epsilon]$ for some $\epsilon > 0$. $p(\cdot)$ is the propensity score. $f_d(\cdot)$, $F_d(\cdot)$ and $q_d(\cdot)$ are the unconditional probability density function (pdf), cdf and quantile function, and $f_d(\cdot|\cdot)$, $F_d(\cdot|\cdot)$ and $Q_d(\cdot|\cdot)$ are the conditional pdf, cdf and quantile function, respectively. f_d, F_d and q_d with

---

[3] They did not consider the unconditional rank similarity for the treated. Their tests can serve as "weaker" pretests to the QTE estimation of Firpo (2007) under unconfoundedness and the QTE estimation of Frölich and Melly (2013a) under confoundedness.

a superscript t represent the counterparts of f_d, F_d and q_d for the treated. $\Phi(\cdot)$ 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, and objects with a superscript $*$ indicate the samples or estimators based on the bootstrap measure.[4] VW is short for van der Vaart and Wellner (1996), HIR for Hirano, Imbens and Ridder (2003), CFM for Chernozhukov, Fernández-Val and Melly (2013), and DH for Donald and Hsu (2014).

2 The Role of Rank Preservation in Treatment Effects Evaluation

To understand the role of rank preservation, we neglect the covariates X and assume Y is continuously distributed in this section to ease the discussion. Recall that the potential outcome approach of Rubin (1974, 1978) augments the observed data (Y, D) by two potential outcomes (Y_1, Y_0). Since $Y = DY_1 + (1-D)Y_0$,

$$F_Y(y) = F_{Y_1|D}(y|1)\,p + F_{Y_0|D}(y|0)(1-p),$$

which is a mixture distribution of $F_{Y_1|D}(y|1)$ and $F_{Y_0|D}(y|0)$, where $p = P(D=1)$. In other words, the distribution of Y is determined by that of (D, Y_1, Y_0), so the "net" randomness in this framework comes from (D, Y_1, Y_0). The joint distribution of (D, Y_1, Y_0) is

$$F_{Y_1,Y_0,D}(y_1, y_0, d) = \begin{cases} F_{(Y_1,Y_0)|D}(y_1, y_0|1)\,p, & \text{if } d = 1, \\ F_{(Y_1,Y_0)|D}(y_1, y_0|0)(1-p), & \text{if } d = 0, \end{cases} \tag{1}$$

The unknown in $F_{Y_1,Y_0,D}(y_1, y_0, d)$ is the conditional distribution $F_{(Y_1,Y_0)|D}(y_1, y_0|d)$. By Sklar's theorem, $F_{(Y_1,Y_0)|D}(y_1, y_0|d) = C_d(F_{Y_1|D}(y_1|d), F_{Y_0|D}(y_0|d))$, where $C_d(\cdot, \cdot)$ is the copula for $F_{(Y_1,Y_0)|D}(y_1, y_0|d)$. $F_{Y_1|D}(y_1|1)$ and $F_{Y_0|D}(y_0|0)$ can be identified from the data, so the unidentified are $C_d(\cdot, \cdot)$, $F_{Y_1|D}(y_1|0)$ and $F_{Y_0|D}(y_0|1)$.

The combination of unconfoundedness and rank preservation settles down the ambiguity in $C_d(\cdot, \cdot)$, $F_{Y_1|D}(y_1|0)$ and $F_{Y_0|D}(y_0|1)$. Recall that unconfoundedness assumes $(Y_1, Y_0) \perp D$, i.e., $F_{(Y_1,Y_0)|D}(y_1, y_0|1) = F_{(Y_1,Y_0)|D}(y_1, y_0|0)$, which implies $F_{Y_1|D}(y_1|0) = F_{Y_1|D}(y_1|1)$, $F_{Y_0|D}(y_0|1) = F_{Y_0|D}(y_0|0)$ and $C_1(\cdot, \cdot) = C_0(\cdot, \cdot)$. To be more precise, $F_{Y_1|D}(y_1|0) = F_{Y_1|D}(y_1|1)$ comes from $Y_1 \perp D$, $F_{Y_0|D}(y_0|1) = F_{Y_0|D}(y_0|0)$ comes from $Y_0 \perp D$, and $C_1(\cdot, \cdot) = C_0(\cdot, \cdot)$ comes from the independence between D and the relationship of Y_1 and Y_0. So the marginal distribution of Y_1, $F_{Y_1}(y_1) = F_{Y_1|D}(y_1|1)p + F_{Y_1|D}(y_1|0)(1-p) = F_{Y_1|D}(y_1|1) = F_{Y_1|D}(y_1|0)$, and similarly, $F_{Y_0}(y_0) = F_{Y_0|D}(y_0|1) = F_{Y_0|D}(y_0|0)$.[5] As a result, unconfoundedness implies

$$F_{Y_1,Y_0,D}(y_1, y_0, d) = C(F_{Y_1}(y_1), F_{Y_0}(y_0))p^d(1-p)^{1-d} = F_{Y_1,Y_0}(y_1, y_0)P(D=d),$$

where $C(\cdot, \cdot)$ is the common copula. So the only unidentified in $F_{Y_1,Y_0,D}(y_1, y_0, d)$ is $C(\cdot, \cdot)$; this is where rank preservation plays the role. In one word, unconfoundedness is about the relationship between D and (Y_1, Y_0), while rank preservation is about the relationship between Y_1 and Y_0 and does not involve D.

Rank preservation implies $C(u_1, u_0) = \min(u_1, u_0)$, i.e., Y_1 and Y_0 are comonotone random variables. Although other copula functions can be employed to identify $F_{Y_1,Y_0,D}(y_1, y_0, d)$, $\min(u_1, u_0)$ is very special in the sense of the following proposition.

Proposition 1 *Under unconfoundedness, the distribution of the treatment effect, $F_{Y_1-Y_0}(\cdot)$, can be identified*

---

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

[5] This implies $F_Y(y) = F_{Y_1}(y)p + F_{Y_0}(y)(1-p)$.

solely from $\Delta\left(\cdot\right)\equiv Q_{Y_1}\left(\cdot\right)-Q_{Y_0}\left(\cdot\right)$ if and only if $C\left(u_1,u_0\right)=\min\left(u_1,u_0\right)$ or the rank is preserved.

Proof. For a general copula, denote $C_{U_1|U_0}\left(u_1|u_0\right)$ as the conditional distribution of $U_1=F_{Y_1}\left(Y_1\right)$ given $U_0=F_{Y_0}\left(Y_0\right)$. Then $F_{Y_1-Y_0}\left(y\right)=\int_0^1\int_0^1 1\left(Q_{Y_1}\left(u_1\right)-Q_{Y_0}\left(u_0\right)\le y\right)dC_{U_1|U_0}\left(u_1|u_0\right)du_0$.
Sufficiency: If the rank is preserved, $C_{U_1|U_0}\left(u_1|u_0\right)$ is a point mass at u_0, so $F_{Y_1-Y_0}\left(y\right)=\int_0^1 1\left(\Delta\left(\tau\right)\le y\right)d\tau$.
Necessity: If the rank is not preserved, the identification of $F_{Y_1-Y_0}\left(y\right)$ requires the knowledge of $C_{U_1|U_0}\left(u_1|u_0\right)$ (i.e., the mapping of ranks under the control and treatmed statuses) and also $Q_{Y_1}\left(\cdot\right)$ and $Q_{Y_0}\left(\cdot\right)$ separately.
∎

To illustrate the unidentifiability of $F_{Y_1-Y_0}\left(\cdot\right)$ by $\Delta\left(\cdot\right)$ without RP and how it can be identified by $\Delta\left(\cdot\right)$ under RP, consider the joint normality case of $\left(Y_1,Y_0\right)$.

Example 1 *Suppose $\left(Y_1,Y_0\right)\sim N\left(\mu,\Sigma\right)$, where*

$$\mu=\begin{pmatrix}0\\0\end{pmatrix}\ and\ \Sigma=\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_0\\ \rho\sigma_1\sigma_0 & \sigma_0^2\end{pmatrix}.$$

Consider three cases. (i) $\sigma_1^2=\sigma_0^2=1$. If $\rho\ne0$, then

$$F_{Y_1-Y_0}\left(y\right)=P\left(Y_1-Y_0\le y\right)=\Phi\left(\frac{y}{\sqrt{2-2\rho}}\right),$$

while $\Delta\left(\tau\right)=0$ for any τ, i.e., Y_1-Y_0 is a point mass at zero. Unless $\rho=1$, $\Phi\left(\frac{y}{\sqrt{2-2\rho}}\right)$ cannot degenerate to $1\left(y\ge0\right)$. (ii) $\sigma_1^2=4$ and $\sigma_0^2=1$. If $\rho\ne0$, then

$$F_{Y_1-Y_0}\left(y\right)=P\left(Y_1-Y_0\le y\right)=\Phi\left(\frac{y}{\sqrt{5-4\rho}}\right),$$

while $\Delta\left(\tau\right)=z_\tau$, where $z_\tau=\Phi^{-1}\left(\tau\right)$ is the τth quantile of a standard normal. Unless $\rho=1$, $\Phi\left(\frac{y}{\sqrt{5-4\rho}}\right)\ne\Phi\left(y\right)=\int_0^1 1\left(\Delta\left(\tau\right)\le y\right)d\tau$. If $\rho=1$, $Q_{Y_1-Y_0}\left(\tau\right)=\Delta\left(\tau\right)$. (iii) $\sigma_1^2=1$ and $\sigma_0^2=4$. If $\rho\ne0$, then

$$F_{Y_1-Y_0}\left(y\right)=P\left(Y_1-Y_0\le y\right)=\Phi\left(\frac{y}{\sqrt{5-4\rho}}\right),$$

while $\Delta\left(\tau\right)=-z_\tau$. Unless $\rho=1$, $\Phi\left(\frac{y}{\sqrt{5-4\rho}}\right)\ne\Phi\left(y\right)=\int_0^1 1\left(\Delta\left(\tau\right)\le y\right)d\tau$. Even if $\rho=1$, $Q_{Y_1-Y_0}\left(\tau\right)=-\Delta\left(\tau\right)\ne\Delta\left(\tau\right)$.

Case (i) with $\rho=1$ is the constant treatment effect case, where $Q_{Y_1-Y_0}\left(\cdot\right)$ can be identified from $\Delta\left(\tau\right)$ for any τ. Case (ii) with $\rho=1$ has a variable treatment effect, and $Q_{Y_1-Y_0}\left(\tau\right)$ can be identified as $\Delta\left(\tau\right)$. Case (iii) with $\rho=1$ also has a variable treatment effect, but $Q_{Y_1-Y_0}\left(\tau\right)$ cannot be identified as $\Delta\left(\tau\right)$. In short, even under RP, $Q_{Y_1-Y_0}\left(\tau\right)$ need not equal $Q_{Y_1}\left(\tau\right)-Q_{Y_0}\left(\tau\right)$ unless $Q_{Y_1}\left(\tau\right)-Q_{Y_0}\left(\tau\right)$ is an increasing function of τ.

Now, we turn the way around - first impose the RP assumption and then the unconfoundedness assumption. It is better now to decompose $F_{Y_1,Y_0,D}\left(y_1,y_0,d\right)$ as

$$
\begin{aligned}
F_{Y_1,Y_0,D}\left(y_1,y_0,d\right) &= F_{Y_1,Y_0}\left(y_1,y_0\right)P_{D|Y_1,Y_0}\left(d|y_1,y_0\right)\\
&= C\left(F_{Y_1}\left(y_1\right),F_{Y_0}\left(y_0\right)\right)\begin{cases}\frac{pF_{(Y_1,Y_0)|D}(y_1,y_0|1)}{pF_{(Y_1,Y_0)|D}(y_1,y_0|1)+(1-p)F_{(Y_1,Y_0)|D}(y_1,y_0|0)}, & \text{if }d=1,\\ \frac{(1-p)F_{(Y_1,Y_0)|D}(y_1,y_0|0)}{pF_{(Y_1,Y_0)|D}(y_1,y_0|1)+(1-p)F_{(Y_1,Y_0)|D}(y_1,y_0|0)}, & \text{if }d=0,\end{cases}
\end{aligned}
$$

where the second equality uses Bayes's rule. The RP assumption reduces $C(F_{Y_1}(y_1), F_{Y_0}(y_0))$ to $\min\left(F_{Y_1}(y_1), F_{Y_0}(y_0)\right)$, while the unconfoundedness assumption reduces $F_{(Y_1, Y_0)|D}(y_1, y_0|1)$ and $F_{(Y_1, Y_0)|D}(y_1, y_0|0)$ to $F_{(Y_1, Y_0)}(y_1, y_0)$ such that $P_{D|Y_1, Y_0}(d|y_1, y_0) = P_D(d) = p^d (1-p)^{1-d}$.

We close this section by a delicate distinction between unconfoundedness and rank preservation. In Matzkin (2003) and Imbens and Newey (2009), a key identification assumption is $Y = h(D, U)$, where $h(D, U)$ is strictly increasing in the scalar error U for any value of D, and $U \perp D$. To simplify our discussion, suppose D is binary. This assumption is implied by unconfoundedness but does not require rank preservation. To see why, note that by the Skorohod representation,

$$Y = h(D, DU_1 + (1-D)U_0),$$

i.e., $U = DU_1 + (1-D)U_0$, where $U_d \sim U[0,1]$. Since U involves D, it seems that U and D cannot be independent unless assuming $U_1 = U_0$ so that D disappears in U. This is not the case thanks to the special form of U as a function of D and the identity of the distributions of U_1 and U_0. First, given $D = d$, $Y = h(d, U_d)$, so $h(d, \cdot)$ is strictly increasing in the second argument. Second, $F_{U|D}(u|d) = F_{U_d|D}(u|d)$, so if $Y_d \perp D$, $F_{U_d|D}(u|d) = F_{U_d}(u) = u$ does not depend on d, or, $U \perp D$. Here we require only $Y_1 \perp D$ and $Y_0 \perp D$ rather than $(Y_1, Y_0) \perp D$.[6] In other words, the assignment of D is allowed to depend on the rank correlation between Y_1 and Y_0. Also, rank preservation does not play any role here. If only rank preservation is imposed, then $U_1 = U_0 = U$ and $Y = h(D, U)$. Although the monotonicity condition is satisfied, D may depend on U. An archetype of $Y = h(D, U)$ is $Y = \Lambda^{-1}(D\beta + U)$ as in Horowitz (1996), where $\Lambda(\cdot)$ is an unknown, strictly increasing function, and $U \perp D$. This model imposes both unconfoundedness and rank preservation.

3   An Overview of Testing Ideas

We introduce more notations to facilitate our discussion. Define

$$\widetilde{Y}_1(U_0) = Q_1(F_0(q_0(U_0)|X)|X),$$

which is the counterfactual Y_1 of $Y_0 = q_0(U_0)$ when the conditional rank is preserved (i.e., $U_0^X = U_1^X$). To see why, note that $F_0(q_0(U_0)|X)$ is the conditional rank U_0^X of $q_0(U_0)$, so if $U_0^X = U_1^X$, then $\widetilde{Y}_1(U_0) = Q_1(U_1^X|X) = Y_1$. It is clear that unless the conditional rank is preserved, $\widetilde{Y}_1 \neq Y_1$. $\widetilde{Y}_1$ can be treated as a monotone rearrangement of Y_1 within each X value according to the rank of Y_0, so is only a partial rearrangement of Y_1; whereas

$$\overline{Y}_1(U_0) = q_1(U_0)$$

is the full rearrangement. $\widetilde{Y}_1 \neq \overline{Y}_1$ unless the unconditional rank is preserved. Note here that $\widetilde{Y}_1(U_0)$ need not be unique for a given value of U_0 because $X|_{U_0}$ may be a genuine random variable. We therefore define

$$\widetilde{q}_1(\tau) = E\left[\widetilde{Y}_1(U_0)|U_0 = \tau\right].$$

Of course, if $X|_{U_0}$ is a point mass, i.e., the U_0 value can uniquely determine the X value, then $\widetilde{Y}_1(U_0)$ is uniquely defined and $\widetilde{Y}_1(U_0) = \widetilde{q}_1(U_0)$.

$\widetilde{Y}_1 \neq \overline{Y}_1$ is the basis of our tests. The rough idea is that if the unconditional rank is not preserved,

---

[6]To be more precise, we require only $U_1 \perp D$ and $U_0 \perp D$ since the value information in Y_1 and Y_0 has been embodied in $h(\cdot, \cdot)$.

full monotone rearrangement and partial rearrangement of Y_1 according to U_0 will generate different results. Especially, the full rearrangement quantile $q_1(U_0)$ is a *further* monotone rearrangement of $\widetilde{Y}_1(U_0)$ (across different X values), so is different from $\widetilde{Y}_1(U_0)$ if the unconditional rank is not preserved. In other words, the powers of our tests come from the rank nonpreservation across different X values rather than within each X value.

3.1 Motivating Examples

We in this subsection use two examples to illustrate that when the unconditional rank is not preserved, $\widetilde{Y}_1 \neq \overline{Y}_1$. Suppose X follows a Bernoulli distribution with the success probability $1/2$, and $Y_0 = X + \epsilon$ with ϵ independent of X. We assume the rank is preserved conditional on X, but the unconditional rank is not preserved. In this case, let $Y_1 = (2-a)X + a \cdot \epsilon$, $a \in (1, 2]$, where $a = 1$ corresponds to the case with rank fully preserved, and $a = 2$ corresponds to the case with rank fully unpreserved. In Example 2, $\epsilon \sim U(0, 1)$ and in Example 3, $\epsilon \sim N(0, 1)$. Note that Y_d in Exmaple 3 follows a mixed normal distribution unconditionally.

Example 2 *In this example, $\widetilde{Y}_1(U_0) = \widetilde{q}_1(U_0)$ since U_0 can uniquely determine X. It can be shown that*

$$
q_0(\tau) = 2\tau,
$$

$$
q_1(\tau) = \begin{cases} 2a\tau, & \text{if } 0 < \tau < \frac{1}{a} - \frac{1}{2}, \\ a\tau - a/2 + 1, & \text{if } \frac{1}{a} - \frac{1}{2} \leq \tau < \frac{3}{2} - \frac{1}{a}, \\ 2a\tau + 2 - 2a, & \text{if } \frac{3}{2} - \frac{1}{a} \leq \tau < 1, \end{cases}
$$

and

$$
\widetilde{q}_1(\tau) = \begin{cases} 2a\tau, & \text{if } 0 < \tau \leq 0.5, \\ 2a\tau + 2 - 2a, & \text{if } 0.5 < \tau \leq 1. \end{cases}
$$

The three lower panels of Figure 2 show the three functions when $a = 1, 1.5$ and 2, where we also show the support of $Y_d|_{X=x}$ for $x = 0, 1$ and $d = 0, 1$ in the three upper panels. Obviously, when the unconditional rank is not preserved, $\widetilde{q}_1(\tau)$ and $q_1(\tau)$ are different. Note here that when $a = 2$, $q_0(\tau) = q_1(\tau)$, but the rank is mostly unpreserved.

From Example 2, we can base our test on

$$
T_o \equiv E\left[\left(Q_1(F_0(q_0(U_0)|X)|X) - q_1(U_0) \right)^2 \right], \tag{2}
$$

where the subscript o is for "oracle". Under H_0, $Q_1(F_0(q_0(U_0)|X)|X) = q_1(F_0(q_0(U_0)))$ for any value of U_0 and $X|_{U_0}$, so $T_o = 0$. Under H_1, $Q_1(F_0(q_0(U_0)|X)|X) \neq q_1(F_0(q_0(U_0)))$ for some U_0 and $X|_{U_0}$ values, so $T_o > 0$ which generates power.

In Example 2, $\widetilde{q}_1(\tau)$ is not monotone when $a \in (1, 2]$ while $q_1(\tau)$ is strictly increasing, so we expect T_o to be strictly positive. Actually, this result is generally true.

Proposition 2 *Suppose $q_0(\tau)$ and $q_1(\tau)$ are strictly increasing measurable functions on $[0, 1]$, and $q_1(\tau)$ and $Q_1(F_0(q_0(\tau)|x)|x)$ are bounded almost surely for $\tau \in [0, 1]$ and $x \in supp(X)$. If there exist regions $\mathcal{T}_0$ and $\mathcal{T}_0'$ in $[0, 1]$, each of Lebesgue measure greater than $\delta > 0$, such that for all $\tau \in \mathcal{T}_0$ and $\tau' \in \mathcal{T}_0'$ we have that (i) $\tau' > \tau$, (ii) $\widetilde{q}_1(\tau) > \widetilde{q}_1(\tau') + \epsilon$ and (iii) $q_1(\tau') > q_1(\tau) + \epsilon$ for some $\epsilon > 0$, then $T_o > 0$.*

Proof. From Part 2 of Proposition 1 of Chernozhukov et al. (2009), we know

$$
E\left[\left(\widetilde{q}_1(U_0) - q_1(U_0) \right)^2 \right] > E\left[\left(\bar{q}_1(U_0) - q_1(U_0) \right)^2 \right] \geq 0,
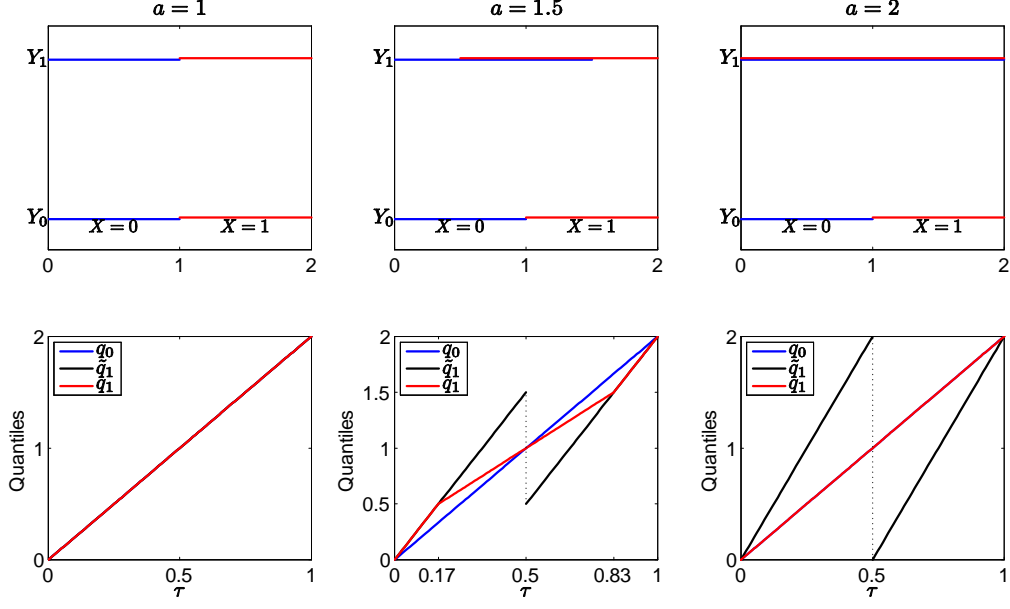$$

7

Figure 2: $q_0(\tau), \widetilde{q}_1(\tau)$ and $q_1(\tau)$ in Example 2 when $a = 1, 1.5$ and 2

where $\overline{q}_1(\tau)$ is the rearranged $\widetilde{q}_1(\tau)$. By Jensen's inequality, for each $\tau \in [0, 1]$,

$$(\widetilde{q}_1(\tau) - q_1(\tau))^2 \leq E\left[(Q_1(F_0(q_0(U_0)|X)|X) - q_1(U_0))^2\Big|U_0 = \tau\right].$$

As a result,

$$0 < E\left[(\widetilde{q}_1(U_0) - q_1(U_0))^2\right] \leq T_o.$$

∎

This proposition provides an intuitive method to detect rank nonpreservation. In practice, we can plot an estimate of $q_1(\tau)$ and $\widetilde{q}_1(\tau)$ and check whether they have significant differences on some region of τ.

T_o is actually more powerful than the situation stated in Proposition 2. Even if $\widetilde{q}_1(\tau)$ is monotone and close to $q_1(\tau)$, T_o will still have power because T_o is based on $\widetilde{Y}_1$ rather than $\widetilde{q}_1$. Example 3 illustrates this point.

Example 3 *For $\tau \in (0, 1)$, $q_d(\tau) = F_d^{-1}(\tau)$, where*

$$
\begin{aligned}
F_0(y) &= 0.5\Phi(y) + 0.5\Phi(y - 1), \\
F_1(y) &= 0.5\Phi\left(\frac{y}{a}\right) + 0.5\Phi\left(\frac{y - (2 - a)}{a}\right).
\end{aligned}
$$

Given $U_0 = \tau$, X can be either 0 or 1. By Bayes' rule,

$$P(X = 0|U_0 = \tau) = \frac{P(Y_0 = q_0(\tau)|X = 0)P(X = 0)}{P(Y_0 = q_0(\tau)|X = 0)P(X = 0) + P(Y_0 = q_0(\tau)|X = 1)P(X = 1)} = \frac{\phi(q_0(\tau))}{\phi(q_0(\tau)) + \phi(q_0(\tau) - 1)}.$$
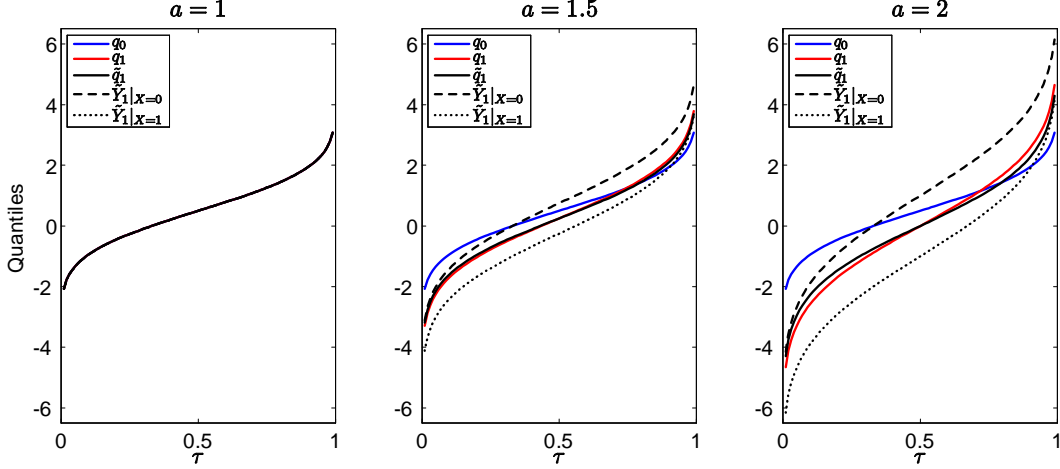
Figure 3: $q_0(\tau), q_1(\tau), \widetilde{q}_1(\tau)$ and $\widetilde{Y}_1(\tau)$ in Example 3 when $a = 1, 1.5$ and 2

It can be shown that

$$\widetilde{Y}_1(\tau) = \begin{cases} aq_0(\tau), & if\ X = 0, \\ aq_0(\tau) + 2 - 2a, & if\ X = 1, \end{cases}$$

so

$$\widetilde{q}_1(\tau) = \frac{aq_0(\tau) \cdot \phi(q_0(\tau))}{\phi(q_0(\tau)) + \phi(q_0(\tau) - 1)} + \frac{[aq_0(\tau) + 2 - 2a]\,\phi(q_0(\tau) - 1)}{\phi(q_0(\tau)) + \phi(q_0(\tau) - 1)}.$$

Figure 3 shows $q_0(\tau), q_1(\tau), \widetilde{q}_1(\tau)$ and $\widetilde{Y}_1(\tau)$ when $a = 1, 1.5$ and 2. In this example, $\widetilde{q}_1(\tau)$ is monotone. When $a = 1.5$ and 2, although $\widetilde{q}_1$ and q_1 are quite close, $\widetilde{Y}_1$ and q_1 are still far apart, which generates power.

3.2 Testing Ideas for the QTE

To rigorously state the testing idea implied in Examples 2 and 3, we first define rank preservation across X values and decompose $U_1 = U_0$ into within-X-value rank preservation $U_1^x = U_0^x$ and across-X-value rank preservation.

Definition 1 *The ranks of Y_1 and Y_0 are said to be preserved across X values if $F_{X|U_1}(x|u) = F_{X|U_0}(x|u)$ for P_{XU} almost sure (x, u), where P_{XU} is the common distribution of (X, U_1) and (X, U_0).*

$F_{X|U_1}(x|u) = F_{X|U_0}(x|u)$ means that the unconditional uth quantile individuals in the treatment and control statuses are distributed balancedly across X values. Our definition of RP across X values allows the common support of (X, U_1) and (X, U_0) to be different from supp$(X) \times [0, 1]$ as in Example 2. It also implies the joint distributions of (X, U_1) and (X, U_0) are the same. To see why, note that

$$F_{XU_1}(u, x) = F_{X|U_1}(x|u)F_{U_1}(u) = F_{X|U_1}(x|u)u = F_{X|U_0}(x|u)F_{U_0}(u) = F_{XU_0}(u, x),$$

where the second equality is from the fact that the marginal distribution of U_d is uniform, and the third equality is from Definition 1.

Suppose D does not affect X, or $X_1 = X_0$,[7] where X_d denotes a potential value of X if D is set to d; then $F_{X|U_1}(x|u) = F_{X|U_0}(x|u)$ combined with $U_1^x = U_0^x$ implies $U_1 = U_0$.

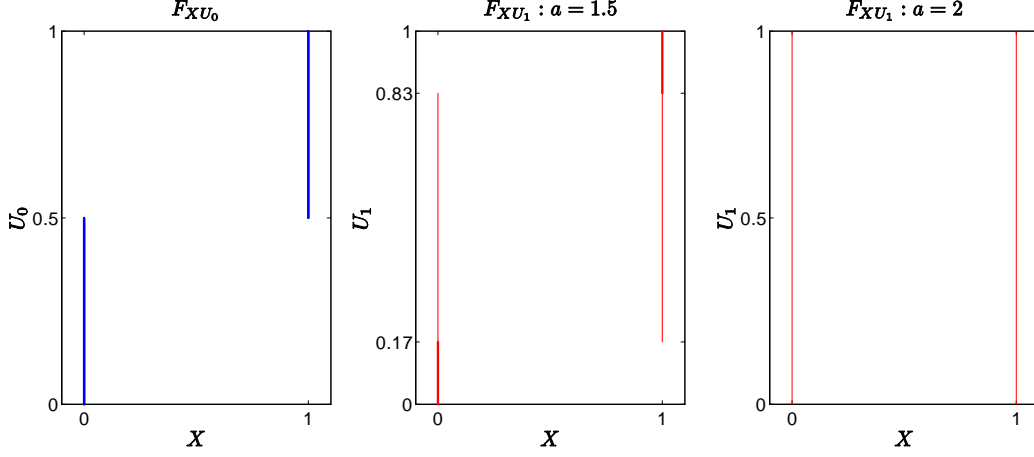[7] This is the "no feedback" condition of Heckman and Vytlacil (2005). Such X is often called "concomitants".

Figure 4: U_0 and U_1 are Unbalanced Across $X = 0$ and 1 in Example 2 when $a = 1.5$ and 2

Theorem 1 *If $X_1 = X_0$, then $U_1 = U_0$ almost surely if and only if $U_1^x = U_0^x$ for P_X almost sure x and $F_{X|U_1}(x|u) = F_{X|U_0}(x|u)$ for P_{XU} almost sure (x, u).*

Proof. First note that the joint distribution of (X, U_d),

$$F_{XU_d}(u, x) = F_{U_d|X}(u|x)F_{X_d}(x) = F_{U_d|X}(u|x)F_X(x) = F_{X|U_d}(x|u)F_{U_d}(u) = F_{X|U_d}(x|u)u, \qquad (3)$$

where the second equality is from $X_1 = X_0$. So $F_{U_d|X}(u|x) = \frac{F_{X|U_d}(x|u)u}{F_X(x)}$, i.e., $F_{U_1|X}(u|x) = F_{U_0|X}(u|x)$ if and only if $F_{X|U_1}(x|u) = F_{X|U_0}(x|u)$.

Second, note that $U_1 = U_0$ is equivalent to $U_1^x = U_0^x$ and $U_1|_{X=x} = U_0|_{X=x}$ for P_X almost sure x. Given that $F_{U_d|X}(u|x)$ is the cdf of $U_d|_{X=x}$, $U_1|_{X=x} = U_0|_{X=x}$ is stronger than $F_{U_1|X}(u|x) = F_{U_0|X}(u|x)$. However, if $U_1^x = U_0^x$ is maintained, then $F_{U_1|X}(u|x) = F_{U_0|X}(u|x)$ implies $U_1|_{X=x} = U_0|_{X=x}$. This is because $F_{U_1|X}(u|x) = F_{U_0|X}(u|x)$ means the same (maybe unsorted) parts of U_1 and U_0 are allocated to $X = x$, and $U_1^x = U_0^x$ means these parts are then sorted, which is exactly the meaning of $U_1|_{X=x} = U_0|_{X=x}$.

Combining these two points, we can conclude the result in the theorem. ■

This theorem shows that if we maintain the following Assumption M, then testing $U_1 = U_0$ is equivalent to testing $F_{X|U_1}(x|u) = F_{X|U_0}(x|u)$ or $F_{U_1|X}(u|x) = F_{U_0|X}(u|x)$.

Assumption M: $X_1 = X_0$, $U_1^x = U_0^x$ for P_X almost sure x.

Figure 4 shows how the distributions of U_0 and U_1 are unbalanced across X values when $a = 1.5$ and 2 in Example 2, where the thickness of lines represents the magnitude of density.

From the discussion in the last subsection, we can use T_o to test the RP hypothesis. A natural question is whether T_o exhausts the information in $U_1 = U_0$. The following theorem gives an affirmative answer.

Theorem 2 *Under Assumption M, $F_{X|U_1}(x|u) = F_{X|U_0}(x|u)$ for P_{XU} almost sure (x, u) if and only if $T_o = 0$.*

Proof. First, recall from Theorem 1, under Assumption M, $F_{X|U_1}(x|u) = F_{X|U_0}(x|u)$ or $F_{U_1|X}(u|x) = F_{U_0|X}(u|x)$ for P_{XU} almost sure (x, u) is equivalent to $U_1 = U_0$.

The necessity is obvious, so we concentrate on the proof of sufficiency. First note that $F_{U_d|X}(u|x)$ is the value that U_d^x will take when $U_d|_{X=x} = u$. The opposite statement of $F_{U_1|X}(u|x) = F_{U_0|X}(u|x)$ for P_{XU}

10

almost sure (x,u) is $F_{U_1|X}(u|x) \neq F_{U_0|X}(u|x)$ on a P_{XU_0} positive set of (x,u). Since $Q_1(F_{U_1|X}(u|x)|x) = q_1(u)$, $F_{U_1|X}(u|x) \neq F_{U_0|X}(u|x)$ on a P_{XU_0} positive set of (x,u) implies $Q_1(F_{U_0|X}(u|x)|x) \neq q_1(u)$ on this set. Equivalently, $E\left[(Q_1(F_0(q_0(U_0)|X)|X) - q_1(U_0))^2\right] \neq 0$. This contradiction implies the result to prove. ∎

A straightforward coroallry of Theorem 1 and 2 is that the test based on T_o is optimal in the sense that it exhausts the testable implication of the data distribution. In other words, if $T_o = 0$, then any other feature of the data distribution cannot contribute to invalidate H_0 further. Such an optimality problem is unsolved in Dong and Shen (2015) and Frandsen and Lefgren (2015).

Corollary 1 *Under Assumption M, if $T_o = 0$, then there exists a joint distribution of (Y_1, Y_0, D, X) that satisfies H_0 and generates the joint data distribution of (Y, D, X).*

Proof. From Theorem 2, under Assumption M, $T_o = 0$ is equivalent to $F_{U_1|X}(u|x) = F_{U_0|X}(u|x) = F_U(u|x)$ for P_{XU} almost sure (x,u). The joint distribution of (Y_1, Y_0, D, X) is

$$
\begin{aligned}
F_{Y_1,Y_0,D,X}(y_1,y_0,d,x) &= F_{(Y_1,Y_0)|D,X}(y_1,y_0|d,x)\,p(x)^d(1-p(x))^{1-d}F_X(x) \\
&= F_{(Y_1,Y_0)|X}(y_1,y_0|x)\,p(x)^d(1-p(x))^{1-d}F_X(x),
\end{aligned}
$$

and the joint data distribution of (Y, D, X) is

$$
\begin{aligned}
F_{Y,D,X}(y,d,x) &= F_{Y|D,X}(y|d,x)\,p(x)^d(1-p(x))^{1-d}F_X(x) \\
&= F_{Y_1|X}(y|x)^d F_{Y_0|X}(y_0|x)^{1-d}p(x)^d(1-p(x))^{1-d}F_X(x).
\end{aligned}
$$

Since $p(x)$ and $F_X(x)$ are implied by the observable data distribution, we need only specify $F_{(Y_1,Y_0)|X}(y_1,y_0|x)$ to generate $F_{Y_1|X}(y_1|x)$ and $F_{Y_0|X}(y_0|x)$ and satisfy H_0. From the proof of Theorem 1, Assumption M plus $F_{U_1|X}(u|x) = F_{U_0|X}(u|x) = F_U(u|x)$ for P_{XU} almost sure (x,u) imply H_0, so we can construct

$$
F_{(Y_1,Y_0)|X}(y_1,y_0|x) = \min\left(F_{Y_1|X}(y_1|x), F_{Y_0|X}(y_0|x)\right),
$$

which satisfies $F_{(Y_1,Y_0)|X}(y_1,\infty|x) = F_{Y_1|X}(y_1|x)$ and $F_{(Y_1,Y_0)|X}(\infty,y_0|x) = F_{Y_0|X}(y_0|x)$. ∎

Our formal test is based on a truncated version of T_o

$$
T \equiv E\left[(Q_1(F_0(q_0(U_0)|X)|X) - q_1(U_0))^2\,1(U_0 \in \mathcal{T}_0)\right].
$$

where $\mathcal{T}_0$, as mentioned at the end of Introduction, is a truncation set of quantile index.[8] 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 techincal 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 RP assumption is easier to hold. For example, the extreme rich (poor) tends to be extremely rich (poor) after 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.[9]

T involves the joint distribution of (X, U_0) which is unobservable since U_0 is unobservable. We attack this problem in two steps. First, replace U_0 by $F_0(Y_0)$; second, replace Y_0 by Y with an adjustment factor

---

[8] Note that $\mathcal{T}_0$ can include only a single (e.g., 0.5) or a few quantile indices (e.g., the middle four quintiles).

[9] If Y_0 is known to be censored at a "middle" quantile τ, we can also kick a small neighborhood of τ out of $\mathcal{T}_0$.

$\frac{1-D}{1-p(X)}$ as suggested in HIR. Specifically,

$$
\begin{aligned}
T &= E\left[(Q_1(F_0(Y_0|X)|X) - q_1\left(F_0\left(Y_0\right)\right))^2 \, 1(Y_0 \in \mathcal{Y}_0)\right] \\
&= E\left[\frac{1-D}{1-p(X)}\left(Q_1(F_0(Y|X)|X) - q_1\left(F_0\left(Y\right)\right)\right)^2 \, 1(Y \in \mathcal{Y}_0)\right],
\end{aligned}
$$

where $\mathcal{Y}_0 = \{q_0(\tau)|\tau \in \mathcal{T}_0\}$. Objects such as $q_1(F_0(\cdot))$ also appear in, e.g., Theorem 3.1 of Athey and Imbens (2006), but they are estimating the quantile treatment effects under the RP assumption whereas we are testing the RP assumption. Given the data $\{W_i\}_{i=1}^n$ with $W = (Y, D, X)$, the sample analog of T is

$$
T_n = \frac{1}{n}\sum_{i=1}^n 1(Y_i \in \mathcal{Y}_0)1(X_i \in \mathcal{X})\frac{1-D_i}{1-\widehat{p}(X_i)}\left[\widehat{Q}_1\left(\left.\widehat{F}_0(Y_i|X_i)\right|X_i\right) - \widehat{q}_1\left(\widehat{F}_0(Y_i)\right)\right]^2, \tag{4}
$$

where the objects with hat are estimators of population counterparts. Allowance of truncating X_i on a compact set $\mathcal{X} \subseteq \operatorname{supp}(X)$ is for practical convenience. More generally, we can truncate $(Y_i, X_i) \in \mathcal{A}$, where $\mathcal{A}$ need not be a cartesian product of two sets such as $\mathcal{Y}_0 \times \mathcal{X}$. This general truncation scheme makes sense in, e.g., Example 2 where there does not exist any product set on which (Y_0, X) has a positive density. Such a general truncation scheme affects the asymptotics of T_n in a minor way, so we do not explicitly pursue it in the main text. Note further that T_n implicitly employs the joint distribution of (X, Y_0) because (X_i, Y_i) is from the same individual i.

3.3 Testing Ideas for the QTT

Practitioners are often interested in testing only RP for the treated. Such a RP hypothesis can be less stringent, i.e., even if the rank is not preserved for the whole population, it may be preserved for the treated. The corresponding null hypothesis is

$$
H_0^t : U_0^t = U_1^t,
$$

and the alternative is

$$
H_1^t : U_0^t \neq U_1^t,
$$

where $U_d^t = F_d^t(Y_d^t) \sim U(0,1)$ is the counterpart of U_d for the treated, $Y_d^t = Y_d|_{D=1}$, and $F_d^t(y) = P(Y_d \leq y|D = 1)$ is the cdf of Y_d^t. Note here that $U_d^t \neq U_d|_{D=1}$ because $U_d|_{D=1}$ does not follow the uniform distribution in general unless the assignment of D is random (i.e., $p(x)$ is constant).

The testing ideas for the QTE can also be applied for the QTT; the only difference is that the population under consideration is not the whole population but the treated. The counterpart of Assumption M is

Assumption M': $X_1 = X_0$, $U_1^{tx} = U_0^{tx}$ for P_X almost sure x, where $U_d^{tx} = F_d^{tx}(Y_d^{tx}) \sim U(0,1)$ with $F_d^{tx}(y) = P(Y_d \leq y|X = x, D = 1)$ and $Y_d^{tx} = Y_d|_{D=1, X=x}$.

Under Assumption M', we can parallelly show that $U_1^t = U_0^t$ if and only if $F_{X|U_1^t}(x|u) = F_{X|U_0^t}(x|u)$ for P_{XU^t} almost sure (x,u), where P_{XU^t} is the common distribution of (X, U_1^t) and (X, U_0^t). Also, $F_{X|U_1^t}(x|u) = F_{X|U_0^t}(x|u)$ for P_{XU^t} almost sure (x,u) if and only if

$$
T_o^t \equiv E\left[\left(Q_1(F_0(q_0^t(U_0^t)|X)|X) - q_1^t\left(U_0^t\right)\right)^2\right] = 0,
$$

and the test based on T_o^t is optimal. Moreover, the intuitive method in Proposition 2 can be applied with

12

$q_d(\tau)$ replaced by $q_d^t(\tau)$ and $\widetilde{q}_1(\tau)$ replaced by

$$\widetilde{q}_1^t(\tau) = E\left[\widetilde{Y}_1^t(U_0^t)|U_0^t = \tau\right].$$

where $\widetilde{Y}_o^t(U_0^t) \equiv Q_1(F_0(q_0^t(U_0^t)|X)|X)$. As in T, we truncate the quantile index of T_o^t and apply the inverse probability weighting scheme of HIR to get

$$
\begin{aligned}
T^t &= E\left[\left(Q_1(F_0(q_0^t(U_0^t)|X)|X) - q_1^t\left(U_0^t\right)\right)^2 1(U_0^t \in \mathcal{T}_0^t)\middle| D = 1\right] \\
&= E\left[\left(Q_1(F_0(Y_0|X)|X) - q_1^t\left(F_0^t(Y_0)\right)\right)^2 1(Y_0 \in \mathcal{Y}_0^t)\middle| D = 1\right] \\
&= E\left[\frac{p(X)}{E[D]}\frac{1-D}{1-p(X)}\left(Q_1(F_0(Y|X)|X) - q_1^t\left(F_0^t(Y)\right)\right)^2 1(Y \in \mathcal{Y}_0^t)\right],
\end{aligned}
$$

where $\mathcal{Y}_0^t = \{q_0^t(\tau)|\tau \in \mathcal{T}_0^t\}$. The sample analog of T^t is

$$T_n^t = \frac{1}{\sum_{i=1}^n D_i}\sum_{i=1}^n 1(Y_i \in \mathcal{Y}_0^t)1(X_i \in \mathcal{X})\frac{\widehat{p}(X_i)\,(1-D_i)}{1-\widehat{p}(X_i)}\left[\widehat{Q}_1\left(\widehat{F}_0(Y_i|X_i)|X_i\right) - \widehat{q}_1^t\left(\widehat{F}_0^t(Y_i)\right)\right]^2, \qquad (5)$$

where as in T_n, we allow truncation on X_i.

In the following three sections, we will propose both parametric and nonparametric forms of $\widehat{p}(\cdot), \widehat{F}_0(\cdot), \widehat{q}_1(\cdot)$, $\widehat{F}_0^t(\cdot), \widehat{q}_1^t(\cdot), \widehat{F}_0(\cdot|\cdot)$ and $\widehat{Q}_1(\cdot|\cdot)$. The parametric test parallels Bierens (1982), while the nonparametric test parallels Härdle and Mammen (1993). In practice, when the distribution of X is complicated, $\dim(X)$ is large and/or the sample size n is not large, we suggest to use the parametric test; otherwise, when there are only one or two continuous covariates and the sample size is large, the nonparametric test is preferable.

4 Parametric Test of Rank Preservation For the QTE

In the parametric setup, we estimate $p(\cdot)$ by

$$\widehat{p}(x) = \Lambda\left(x'\widehat{\gamma}\right),$$

where $\Lambda(\cdot)$ is a known link function such as the logistic cdf $L(\cdot) = \exp(\cdot)/(1 + \exp(\cdot))$, and

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

$F_0(\cdot)$ and $q_1(\cdot)$ are estimated based on the inverse probability weighted (IPW) method of DH. Specifically,

$$
\begin{aligned}
\widehat{F}_0(y) &= \frac{1}{n}\sum_{i=1}^n \frac{1-D_i}{1-\widehat{p}(X_i)}1(Y_i \le y), \\
\widehat{F}_1(y) &= \frac{1}{n}\sum_{i=1}^n \frac{D_i}{\widehat{p}(X_i)}1(Y_i \le y),
\end{aligned}
$$

where we can reweight $\widehat{F}_0(y)$ and $\widehat{F}_1(y)$ by $n^{-1}\sum_{i=1}^n \frac{1-D_i}{1-\widehat{p}(X_i)}$ and $n^{-1}\sum_{i=1}^n \frac{D_i}{\widehat{p}(X_i)}$ without affecting their asymptotic properties.[10] Given that $\widehat{p}(X_i) \in (0,1)$, $\widehat{F}_1(y)$ is automatically (weakly) increasing with jumps

---

[10]This reweighting can guarantee that $\widehat{F}_d(y) \in [0,1]$. T_n and T_n^t can also be reweighted by $n^{-1}\sum_{i=1}^n \frac{1-D_i}{1-\widehat{p}(X_i)}$ and

at Y_{1i}'s. So we define

$$\widehat{q}_1(\tau) = \inf\left\{y|\widehat{F}_1(y) \geq \tau\right\}.$$

As to $F_0(\cdot|\cdot)$ and $Q_1(\cdot|\cdot)$, our estimation scheme is based on the distribution regression (DR) proposed by Foresi and Peracchi (1995) and extended by CFM. Specifically, we estimate $F_d(y|x)$ by $\widehat{F}_d(y|x) \equiv \Lambda\left(x'\widehat{\beta}_d(y)\right)$, where for $y \in \mathcal{Y}_d$,

$$\widehat{\beta}_d(y) = \arg\max_\beta \sum_{i=1}^n 1(D_i = d)\left[1\left(Y_i \leq y\right)\ln \Lambda\left(X_i'\beta\right) + 1\left(Y_i > y\right)\ln\left(1 - \Lambda\left(X_i'\beta\right)\right)\right].^{11} \qquad (6)$$

We estimate $Q_1(\tau|x)$ by inverting $\widehat{F}_1(y|x)$. Because $\widehat{F}_1(y|x)$ need not be monotone in y, following Chernozhukov et al. (2010), we first monotonely rearrange it before the inversion. Note that for inversion, the set of y values at which $F_1(y)$ and $\beta_1(y)$ are estimated need not be the same as $\mathcal{Y}_1$.

In summary, our test statistic is

$$T_n = \frac{1}{n}\sum_{i=1}^n 1(Y_i \in \mathcal{Y}_0)1(X_i \in \mathcal{X})\frac{1 - D_i}{1 - \widehat{p}(X_i)}\left[\widehat{Q}_1\left(\Lambda\left(X_i'\widehat{\beta}_0(Y_i)\right)|X_i\right) - \widehat{q}_1\left(\widehat{F}_0(Y_i)\right)\right]^2.$$

This is a Hausman-type test statistic because both $\widehat{Q}_1\left(\Lambda\left(X_i'\widehat{\beta}_0(Y_{0i})\right)|X_i\right)$ and $\widehat{q}_1\left(\widehat{F}_0(Y_{0i})\right)$ estimate the counterfactual of Y_{0i} in the treatment status under the null, but only $\widehat{Q}_1\left(\Lambda\left(X_i'\widehat{\beta}_0(Y_i)\right)|X_i\right)$ estimates it under the alternative.

Before developing the asymptotic properties of T_n, we provide two comments on the construction procedure of T_n above. First, in estimating $p(\cdot), F_0(\cdot|\cdot)$ and $F_1(\cdot|\cdot)$, we use a parametric setup for the conditional cdfs. As in HIR, we can use the Series Logit Estimator (SLE) to estimate $p(\cdot)$. Suppose the power series are used. For $K = 1, 2, \cdots$, let $R^K(x) = (r_{1K}(x), r_{2K}(x), \cdots, r_{KK}(x))'$ be a K-vector of functions. Let $\lambda = (\lambda_1, \cdots, \lambda_r)'$ be an r-dimensional vector of nonnegative integers (multi-indices), with norm $|\lambda| = \sum_{j=1}^r \lambda_j$, let $(\lambda(k))_{k=1}^\infty$ be a sequence that includes all distinct multi-indices and satisfies $|\lambda(k)| \leq |\lambda(k+1)|$, and let $x^\lambda = \prod_{j=1}^r x_j^{\lambda_j}$. For a sequence $\lambda(k)$ the series $r_{kK}(x) = x^{\lambda(k)}.^{12}$ Similarly, we can replace the regressors in (6) by a transformation of x such as polynomials or B-splines. Our test in this section can be extended to the case where the number of series terms K is fixed, but not to the case where $K \to \infty$ as $n \to \infty$ due to technical complications.^{1314} As an alternative to the nonparametric series estimator, we suggest to use the goodness of fit tests to control for the misspecification bias in $p(\cdot), F_0(\cdot|\cdot)$ and $F_1(\cdot|\cdot)$;^{15} see the supplementary materials for more discussions and Horowitz (2011, p. 349) for a critical view of this method. Second, the estimation of $F_d(\cdot|\cdot)$ and $F_d(\cdot)$ can be based on other methods. For exam-

---

$\sum_{i=1}^n \frac{\widehat{p}(X_i)(1 - D_i)}{1 - \widehat{p}(X_i)}$, respectively; see more discussions in Section 7.

[11] Note that for $d = 1$, $P(D = 1, Y \leq y|X = x) = p(x)F_1(y|x)$ and $P(D = 1, Y > y|X = x) = p(x)(1 - F_1(y|x))$. So $\Lambda\left(X_i'\beta\right)$ and $1 - \Lambda\left(X_i'\beta\right)$ in equation (6) should be replaced by $p(X_i'\gamma)\Lambda\left(X_i'\beta\right)$ and $p(X_i'\gamma)\left(1 - \Lambda\left(X_i'\beta\right)\right)$, respectively. However, since γ and β are separable, we can estimate β by the objective function in (6) without loss of generality.

[12] When X is continuous, $r_{1K}(x)$ is usually set as 1; when the support of X is finite, $R^K(x)$ just contains indicators of all points of supp(X) and K is fixed. Also from footnote 12 of CFM, the choice of link function $\Lambda(\cdot)$ has some freedom, e.g., probit, linear, log-log and Gosset functions are all fine as long as K is large enough.

[13] Of course, $p(\cdot)$ can be estimated by the SLE with $K \to \infty$, but because this section concentrates on the parametric test, we use a parametric estimator of $p(\cdot)$ here. The next section will extend to the nonparametric form of $\widehat{p}(\cdot)$. Note also that even if $p(X)$ is correctly parametrized, the efficient estimators of $F_d(y)$ require nonparametric estimates of $p(\cdot)$; see, e.g., MaCurdy et al. (2011).

[14] Nonparametric tests such as Hong and White (1995) and Härdle and Mammen (1993) in the classical specification testing are based on the linear series and the kernel estimator, respectively. To my best knowledge, there is no literature on how to extend the nonparametric test to base on nonlinear estimation procedures as in our case.

[15] In some special cases, the goodness of fit tests are not necessary; e.g., when X is discrete, a fully saturated linear model does not involve any misspecification error.

ple, $q_0(\cdot)$ and $q_1(\cdot)$ can be estimated by the IPW method of Firpo (2007), and $F_0(\cdot)$ is then estimated by inverting the estimator of $q_0(\cdot)$.[16] Similarly, $F_0(\cdot|\cdot)$ and $Q_1(\cdot|\cdot)$ can be estimated based on quantile regression rather than distribution regression, or $F_0(\cdot|\cdot)$ is estimated by distribution regression while $Q_1(\cdot|\cdot)$ is estimated by quantile regression.[17] Furthermore, $F_d(\cdot)$ can be estimated by integrating $F_d(\cdot|\cdot)$ rather than based on the IPW method.[18] Our choice of estimation methods is based mainly on technical convenience,[19] and by no means indicates better performances in finite samples.

4.1 Asymptotics for T_n

To derive the asymptotic distribution of T_n, we first impose the following assumptions.

Assumption U (Unconfoundedness): $(Y_0, Y_1) \perp D | X$.

Assumption X (Distribution of X): $\text{supp}(X) \subset \mathbb{R}^r$ is compact.

Assumption Y (Distributions of Y_0 and Y_1): (i) $f_0(y)$ is bounded, positive and continuous on $\mathcal{Y}_0$, where $\mathcal{Y}_0$ is compact. (ii) $f_1(y)$ is bounded, positive and continuously differentiable on $\mathcal{Y}_1$ which is compact and contains an ϵ-enlargement of the set $\{q_1(F_0(y)) : y \in \mathcal{Y}_0\}$.

Assumption P (Propensity Score): $p(x) = \Lambda(x'\gamma_0)$ for all $x \in \mathcal{X}$, where Λ is either the probit or logit link function, and $p(x)$ is bounded away from zero and one: $0 < \underline{p} \leq p(x) \leq \overline{p} < 1$.

Assumption DR: (i) $F_d(y|x) = \Lambda(x'\beta_d(y))$ for all $y \in \mathcal{Y}_d$ and $x \in \mathcal{X}$, where Λ is either the probit or logit link function, $\mathcal{Y}_d\mathcal{X}$ is compact, and $\mathcal{Y}_1$ contains an ϵ-enlargement of the set $\{Q_1(F_0(y|x)|x)|x \in \mathcal{X}, y \in \mathcal{Y}_0\}$. (ii) $f_d(y|x)$ is uniformly bounded, and is uniformly continuous for $(y, x) \in \mathcal{Y}_d\mathcal{X}$, $f_1(y|x)$ is continuously differentiable in y, and $f_1(y|x) > 0$ for $(y, x) \in \mathcal{Y}_1\mathcal{X}$. (iii) the minimum eigenvalue of

$$J_d(y) \equiv E\left[\frac{1(D=d)\lambda(X'\beta_d(y))^2}{\Lambda(X'\beta_d(y))(1 - \Lambda(X'\beta_d(y)))}XX'\right]$$

is bounded away from zero uniformly over $y \in \mathcal{Y}_d$, where λ is the derivative of Λ.

We also consider the local alternative H_1^δ:

$$\begin{aligned} U_{0n}^x &= U_{1n}^x \text{ for all } x \in \mathcal{X}, \\ p_n(x) &= (1 - \delta_\gamma/\sqrt{n})p_*(x) + (\delta_\gamma/\sqrt{n})\wp(x), \\ F_n^d(y|x) &= (1 - \delta_d/\sqrt{n})F_*^d(y|x) + (\delta_d/\sqrt{n})\mathfrak{F}^d(y|x), \end{aligned}$$

where $p_n(x)$ satisfies Assumption P and $F_n^d(y|x)$ satisfies Assumption DR, and U_{dn}^x is the counterpart of U_d^x under F_n^d. $p_*(x)$ and $F_*^d(y|x)$ satisfy $Q_1(F_0(y|x)|x) - q_1(F_0(y)) = 0$ for $yx \in \mathcal{Y}_0\mathcal{X}$, but $\wp(x)$ and $\mathfrak{F}^d(y|x)$ do not. This specification of local alternative is indirect but more intuitive than the direct specification $F_{U_{1n}|X}(u|x) - F_{U_{0n}|X}(u|x) = \delta(u|x)/\sqrt{n}$ because the unconditional rank U_{dn} can only be generated from $F_n^d(y|x)$. We impose the following assumption on the data distribution under H_1^δ.

---

[16] Specifically, $\widehat{q}_0(\tau) = \arg\min_{q_0} n^{-1}\sum_{i=1}^n \frac{1-D_i}{1-\widehat{p}(X_i)}\rho_\tau(Y_i - q_0)$, $\widehat{q}_1(\tau) = \arg\min_{q_1} n^{-1}\sum_{i=1}^n \frac{D_i}{\widehat{p}(X_i)}\rho_\tau(Y_i - q_1)$ and $\widehat{F}_0(y) = \varepsilon + \int_\varepsilon^{1-\varepsilon} 1(\widehat{q}_0(\tau) \leq y)d\tau$, where $\rho_\tau(u) = u \cdot 1(\tau - 1(u < 0))$ is the check function, and ε is a specified small positive number to avoid estimating tail quantiles.

[17] Specifically, $\widehat{F}_0(y|x) = \varepsilon + \int_\varepsilon^{1-\varepsilon} 1(x'\widehat{\beta}_0(\tau) \leq y)d\tau$, and $\widehat{Q}_1(\tau|x) = x'\widehat{\beta}_1(\tau)$, where $\widehat{\beta}_d(\tau) = \arg\min_{\beta_d}\sum_{i=1}^n 1(D_i = d)\rho_\tau(Y_i - X_i'\beta_d)$.

[18] Specifically, $\widehat{F}_d(y) = n^{-1}\sum_{i=1}^n \widehat{F}_d(y|X_i)$.

[19] For example, the uniform consistency of the IPW estimator of $q_d(\cdot)$ is not established in Firpo (2007) but the corresponding result for the IPW estimator of $F_d(\cdot)$ is established in DH. Actually, because the former is numerically equivalent to the inverse of the latter, uniform consistency of Firpo's estimator of $q_d(\cdot)$ is implied by DH.

15

Assumption LA: The joint distribution of W implied by the local alternative is contiguous to that implied by $p_*(\cdot)$ and $F_*^d(\cdot|\cdot)$.

We provide a few comments on the assumptions above. First, unconfoundedness is a strong assumption but extensively used in theoretical analysis and applications; see Imbens (2004) for a summary of relevant literature. Second, the distribution of X can be either continuous or discrete. Also, the support of X can be unbounded as long as we add $E\left[\|X\|^2\right] < \infty$ in Assumption DR and specify $\mathcal{X}$ as a compact subset of $\operatorname{supp}(X)$. Third, as noted by Khan and Tamer (2010), the assumption that the propensity score is bounded away from zero and one plays an important role in determining the convergence rate of IPW estimators. Fourth, Assumption DR is borrowed from CFM to guarantee the validity of uniform inference for DR estimators. Fifth, our local alternative preserves the conditional rank, but violates the null through perturbing the conditional distribution of (D, Y_0, Y_1). For future reference, define $F_*^d(y) = E\left[F_*^d(y|X)\right]$, $q_*^d(\cdot)$ is the inverse function of $F_*^d(y)$, $Q_*^d(\cdot|x)$ is the inverse function of $F_*^d(\cdot|x)$, $f_*^d(\cdot|x)$ is the density associated with $F_*^d(\cdot|x)$, and $f_*^d(\cdot)$ is the density associated with $F_*^d(\cdot)$. The corresponding objects associated with $\mathfrak{F}^d$ are similarly defined. Sixth, the requirement for the contiguity of the local alternative to the null is standard in analyzing the local power. A sufficient condition for contiguity is that

$$\sup_{(y,x):f_*^d(y|x)>0} \mathfrak{f}^d(y|x)/f_*^d(y|x) < \infty,$$

where $\mathfrak{f}^d(\cdot|\cdot)$ is the conditional density associated with $\mathfrak{F}^d(\cdot|\cdot)$. Intuitively, this would be the case when $\mathfrak{F}^d$ has lighter tails than F_*^d.

The asymptotic null distribution of T_n is quite complicated due to the plug-in estimators $\widehat{p}(\cdot), \widehat{Q}_1(\cdot|\cdot), \widehat{F}_0(\cdot|\cdot), \widehat{q}_1(\cdot)$ and $\widehat{F}_0(\cdot)$. To facilitate the statement of the asymptotic distribution of T_n, define $v = (y, x)$, and $Z(v)$ as a mean zero Gaussian process on $\mathcal{Y}_0\mathcal{X}$ with the covariance function

$$\Sigma(v_1, v_2) = E\left[Z(v_1)Z(v_2)\right] = E\left[\left(\Psi_c\left(W, x_1, y_1\right) - \Psi_u\left(W, y_1\right)\right)\left(\Psi_c\left(W, x_2, y_2\right) - \Psi_u\left(W, y_2\right)\right)\right],$$

where

$$
\begin{aligned}
\Psi_c\left(W, x, y\right) &= \frac{\phi_0\left(W, x, y\right) - \phi_1\left(W, x, Q_1\left(F_0(y|x)|x\right)\right)}{f_1\left(Q_1\left(F_0(y|x)|x\right)|x\right)}, \\
\Psi_u\left(W, y\right) &= \frac{\psi_0(W, y) - \psi_1(W, q_1(F_0(y)))}{f_1(q_1(F_0(y)))},
\end{aligned}
$$

and

$$
\begin{aligned}
\phi_0(W, x, y) &= \lambda\left(x'\beta_0(y)\right)x'J_0^{-1}(y)\frac{(1-D)\left[1(Y \le y) - \Lambda\left(X'\beta_0(y)\right)\right]}{\Lambda\left(X'\beta_0(y)\right)\left(1 - \Lambda\left(X'\beta_0(y)\right)\right)}\lambda\left(X'\beta_0(y)\right)X, \\
\phi_1(W, x, y) &= \lambda\left(x'\beta_1(y)\right)x'J_1^{-1}(y)\frac{D\left[1(Y \le y) - \Lambda\left(X'\beta_1(y)\right)\right]}{\Lambda\left(X'\beta_1(y)\right)\left(1 - \Lambda\left(X'\beta_1(y)\right)\right)}\lambda\left(X'\beta_1(y)\right)X, \\
\psi_0(W, y) &= \frac{1-D}{1-p(X)}1(Y \le y) - F_0(y) \\
&\quad + E\left[\frac{\lambda(X'\gamma_0)}{1-p(X)}F_0(y|X)X'\right]E\left[\frac{\lambda(X'\gamma_0)^2}{p(X)(1-p(X))}XX'\right]^{-1}X\frac{\lambda(X'\gamma_0)}{p(X)}\frac{D-p(X)}{1-p(X)}, \\
\psi_1(W, y) &= \frac{D}{p(X)}1(Y \le y) - F_1(y) \\
&\quad - E\left[\frac{\lambda(X'\gamma_0)}{p(X)}F_1(y|X)X'\right]E\left[\frac{\lambda(X'\gamma_0)^2}{p(X)(1-p(X))}XX'\right]^{-1}X\frac{\lambda(X'\gamma_0)}{1-p(X)}\frac{D-p(X)}{p(X)}.
\end{aligned}
$$

The terms associated with ϕ_1, ϕ_0, ψ_1 and ψ_0 are the contribution of $\widehat{Q}_1(\cdot|\cdot), \widehat{F}_0(\cdot|\cdot), \widehat{q}_1(\cdot)$ and $\widehat{F}_0(\cdot)$, respectively. The effects of $\widehat{p}(\cdot)$ on the influence function include two parts: (i) the direct effect, i.e., the effect of $\widehat{p}(\cdot)$ that appears in T_n, which is asymptotically neglectable; (ii) the indirect effects, i.e., the effects of $\widehat{p}(\cdot)$ in $\widehat{F}_0(\cdot)$ and $\widehat{q}_1(\cdot)$, are included in ψ_0 and ψ_1 as the terms associated with $D - p(X)$. Define λ_i's as the eigenvalues of $\Sigma(v_1, v_2)$; by Mercer's theorem (see Lemma 1 of Bierens and Ploberger, 1997), there exist orthonormal eigenfunctions $\varphi_i(v)$ such that

$$\int \Sigma(v_1, v_2)\varphi_i(v_2)\, d\mu(v_2) = \lambda_i\varphi_i(v_1),$$

where $\mu(\cdot)$ is a measure on $\mathcal{Y}_0\mathcal{X}$ such that for any measurable set A in $\text{supp}(Y_0X)$, $\mu(A) = \int_{A\cap\mathcal{Y}_0\mathcal{X}} dF_{Y_0|X}(y|x)dF_X(x),$[20] $\lambda_i \geq 0$ need not be distinct, and $\sum_{i=1}^{\infty} \lambda_i < \infty$.

Theorem 3 *Under Assumptions DR, P, U, X and Y, the following statements hold:*

(i) *Under H_0,*

$$nT_n \rightsquigarrow \sum_{i=1}^{\infty} \lambda_i\chi_{1i}^2,$$

where χ_{1i}^2's are iid χ_1^2 random variables, and λ_i's are eigenvalues of $\Sigma(v_1, v_2)$.

(ii) *Under H_1^δ and Assumption LA,*

$$nT_n \rightsquigarrow \sum_{i=1}^{\infty} \left(b_i + \varepsilon_i\sqrt{\lambda_i}\right)^2 = \sum_{i=1}^{\infty} \lambda_i\chi_{1i}^2\left(b_i^2/\lambda_i\right),$$

where the ε_i are iid $N(0,1)$, $b_i = \int b(v)\varphi_i(v)\, d\mu(v)$ with

$$
\begin{aligned}
b(y, x) &= \frac{\delta_0\left[\mathfrak{F}^0(y|x) - F_*^0(y|x)\right] - \delta_1\left[\mathfrak{F}^1(Q_*^1(F_*^0(y|x)|x)|x) - F_*^0(y|x)\right]}{f_*^1(Q_*^1(F_*^0(y|x)|x)|x)} \\
&\quad - \frac{\delta_0\left[\mathfrak{F}^0(y) - F_*^0(y)\right] - \delta_1\left[\mathfrak{F}^1(q_*^1(F_*^0(y))) - F_*^0(y)\right]}{f_*^1(q_*^1(F_*^0(y)))},
\end{aligned}
$$

and $\chi_{1i}^2\left(b_i^2/\lambda_i\right)$'s are independent noncentral χ_1^2 random variables with noncentral parameters b_i^2/λ_i. Thus, for any $c > 0$, $P\left(nT_n > c|H_1^\delta\right) \geq P\left(nT_n > c|H_0\right)$, where the equality holds if and only if $b_i = 0$ for any i.

(iii) *Under the fixed alternative H_1 with $\text{plim}_{n\to\infty} T_n > 0$,*

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

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

We provide a few comments on Theorem 3. First, although T_n takes an average form, its asymptotic distribution is not normal since T_n is nonnegative for any n. Actually, from the proof of Theorem 3, T_n is asymptotically equivalent to a third-order V-statistic. However, this V-statistic is degenerate because the first term of Hoeffding projection is zero. We must rely on the second term of Hoeffding projection to obtain

---

[20]$\mu(\cdot)$ is not a probability measure but a truncated measure on $\mathcal{Y}_0\mathcal{X}$ since $\mu(\mathcal{Y}_0\mathcal{X}) = \int_{\mathcal{Y}_0\mathcal{X}} dF_{Y_0|X}(y|x)dF_X(x) < 1$, but Mercer's theorem can still be applied. $\mu(y, x)$ is understood as $\mu(A_{yx})$ with $A_{yx} = \{(\mathsf{y}, \mathsf{x}) \in \mathcal{Y}_0\mathcal{X}|\mathsf{y} \leq y, \mathsf{x} \leq x\}$. We can actually normalize $\mu(\cdot)$ to be a probability measure; see the discussion in Section 7.

a nondegenerate asymptotic distribution; this is why the mixed chi-square distribution emerges. Second, the power comes from the perturbation of $F_d(y|x)$ such that the conditional quantile and unconditional quantile are not the same or U_0 and U_1 are distributed among X values in an unbalanced way. Specifically, $\frac{\mathfrak{F}^0(y|x)-F_*^0(y|x)}{f_*^1(Q_*^1(F^0(y|x)|x)|x)} \neq \frac{\mathfrak{F}^0(y)-F_*^0(y)}{f_*^1(q_*^1(F_*^0(y)))}$ and $\frac{\mathfrak{F}^1(Q_*^1(F_*^0(y|x)|x)|x)-F_*^0(y|x)}{f_*^1(Q_*^1(F^0(y|x)|x)|x)} \neq \frac{\mathfrak{F}^1(q_*^1(F_*^0(y)))-F_*^0(y)}{f_*^1(q_*^1(F_*^0(y)))}$. Of course, if these differences disappear in some sense after averaging over $(y,x) \in \mathcal{Y}_0\mathcal{X}$, our test will not have power. Third, letting c in (ii) and c_n in (iii) be the critical value of our test, then (ii) implies that our test is asymptotically locally unbiased and (iii) implies that our test is consistent.

4.2 Bootstrapping the Critical Values of T_n

The eigenvalues λ_i are necessary inputs to determine the critical values of T_n, but they depend on the data-generating process under the null and are hard to estimate.[21] To make our testing procedure more applicable, we suggest to use the exchangeable bootstrap to obtain the critical values. We formally summarize the bootstrap procedure in the following Algorithm B. First, 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 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 samples and estimators.

Algorithm B:

Step 1: Define

$$\widehat{F}_0^*(y) = \frac{1}{n^*}\sum_{i=1}^n \omega_i \frac{1-D_i}{1-\widehat{p}^*(X_i)}1(Y_i \leq y),$$

$$\widehat{F}_1^*(y) = \frac{1}{n^*}\sum_{i=1}^n \omega_i \frac{D_i}{\widehat{p}^*(X_i)}1(Y_i \leq y),$$

where $n^* = \sum_{i=1}^n \omega_i$, and

$$\widehat{p}^*(x) = \Lambda\left(x'\widehat{\gamma}^*\right)$$

with

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

Then

$$\widehat{q}_d^*(\tau) = \inf\left\{y|\widehat{F}_d^*(y) \geq \tau\right\}.$$

Step 2: Define

$$\widehat{F}_d^*(y|x) = \Lambda\left(x'\widehat{\beta}_d^*(y)\right),$$

$$\widehat{Q}_d^*(\tau|x) = \inf\left\{y\left|\widehat{F}_d^*(y|x) \geq \tau\right.\right\},$$

where

$$\widehat{\beta}_d^*(y) = \arg\max_\beta \sum_{i=1}^n \omega_i 1(D_i = d) \left[1\left(Y_i \le y\right) \ln \Lambda\left(X_i'\beta\right) + 1\left(Y_i > y\right) \ln\left(1 - \Lambda\left(X_i'\beta\right)\right) \right].$$

Step 3: Define the bootstrap counterpart of T_n as

$$T_n^* = \frac{1}{n^*} \sum_{i=1}^n \omega_i 1(Y_i \in \mathcal{Y}_0) 1(X_i \in \mathcal{X}) \frac{1 - D_i}{1 - \widehat{p}^*(X_i)}$$
$$\cdot \left[\left(\widehat{Q}_1^* \left(\widehat{F}_0^*(Y_i|X_i)|X_i \right) - \widehat{Q}_1 \left(\widehat{F}_0(Y_i|X_i)|X_i \right) \right) - \left(\widehat{q}_1^* \left(\widehat{F}_0^*(Y_i) \right) - \widehat{q}_1 \left(\widehat{F}_0(Y_i) \right) \right) \right]^2.$$

Step 4: Simulate T_n^* B times to get $\{T_{nb}^*\}_{b=1}^B$ for B large enough, and then reject H_0 if $nT_n > \widehat{c}_n^*(\alpha)$, where $\widehat{c}_n^*(\alpha)$ is the $(1-\alpha)$th quantile of $\{n^* T_{nb}^*\}_{b=1}^B$ which approximates the $(1-\alpha)$th quantile of $n^* T_n^*$, say, $c_n^*(\alpha)$. Of course, we can also check whether the p-value $B^{-1} \sum_{b=1}^B 1(n^* T_{nb}^* \ge nT_n)$ is less than α to decide whether to reject H_0.

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 $\epsilon > 0$,

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

where $\xrightarrow{P^*}$ signifies the convergence in the probability of bootstrap measure.[22]

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 the case where $(\omega_1, \cdots, \omega_n)$ is equal to $\sqrt{n/m}$ times a multinomial vector with parameter m and probabilities $(1/n, \cdots, 1/n)$. The subsampling bootstrap corresponds to the case where $(\omega_1, \cdots, \omega_n)$ is 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.

The following theorem states the validity of the above bootstrap procedure.

Theorem 4 *Under Assumptions DR, EB, P, U, X and Y, the following statements hold:*

(i) *Under H_0,*

$$\lim_{n\to\infty} P\left(nT_n > c_n^*(\alpha)\right) = \alpha.$$

(ii) *Under H_1^δ and Assumption LA,*

$$\lim_{n\to\infty} P\left(nT_n > c_n^*(\alpha)\right) \ge \alpha.$$

---

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

(iii) *Under the fixed alternative H_1 with $\text{plim}_{n\to\infty} T_n > 0$,*

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

(i) implies that under H_0, $c_n^*(\alpha) \xrightarrow{p} c(\alpha)$, where $c(\alpha)$ is the $(1-\alpha)$th quantile of the asymptotic distribution of nT_n, 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_n using $c_n^*(\alpha)$ as the critical value is asymptotically locally unbiased. (iii) is a corollary of Theorem 3(iii) since $c_n^*(\alpha)$ is bounded in probability under the fixed alternative. Finally, it can be shown that the quantiles of

$$\frac{1}{n}\sum_{i=1}^{n} 1(Y_i \in \mathcal{Y}_0)1(X_i \in \mathcal{X})\frac{1-D_i}{1-\widehat{p}(X_i)}\left[\left(\widehat{Q}_1^*\left(\widehat{F}_0^*(Y_i|X_i)|X_i\right) - \widehat{Q}_1\left(\widehat{F}_0(Y_i|X_i)|X_i\right)\right) - \left(\widehat{q}_1^*\left(\widehat{F}_0^*(Y_i)\right) - \widehat{q}_1\left(\widehat{F}_0(Y_i)\right)\right)\right]^2$$

can also serve as valid critical values of T_n. This new bootstrap statistic is like the wild bootstrap statistic; essentially, we do not bootstrap the measure $\mu(\cdot)$ here.

5   Nonparametric Test of Rank Preservation For the QTE

In the nonparametric setup, we estimate $p(\cdot), F_0(\cdot|\cdot), Q_1(\cdot|\cdot)$ nonparametrically and $F_0(\cdot), q_1(\cdot)$ semiparametrically. Specifically, let $K(\cdot)$ be a r-dimentional kernel function having support on $[-1,1]^r$, $h = h_n$ be a sequence of bandwidth, and $K_h(u) = K(u/h)$. Then for $x \in \mathcal{X}$ and $y \in \mathcal{Y}_0$

$$
\begin{aligned}
\widehat{F}_0(y|x) &= \frac{(nh^r)^{-1}\sum_{j=1}^{n}(1-D_j)\,1(Y_j \leq y)K_h(X_j - x)}{(nh^r)^{-1}\sum_{j=1}^{n}(1-D_j)\,K_h(X_j - x)}, \\
\widehat{F}_1(y|x) &= \frac{(nh^r)^{-1}\sum_{j=1}^{n}D_j 1(Y_j \leq y)K_h(X_j - x)}{(nh^r)^{-1}\sum_{j=1}^{n}D_j K_h(X_j - x)}, \\
\widehat{Q}_1(\tau|x) &= \inf\left\{y|\widehat{F}_1(y|x) \geq \tau\right\}.
\end{aligned}
$$

The estimates of $F_d(\cdot)$ are the same as in the last section except that we use the SLE of HIR to estimate $p(\cdot)$, where the number of series terms $K \to \infty$ as $n \to \infty$. Of course, other estimators of $p(\cdot)$ such as the local polynomial estimator (LPE) in Ichimura and Linton (2005) and the higher order kernel estimator in Abrevaya et al. (2015) can also be used, but the SLE seems most convenient in practice. We provide a few comments on our nonparametric test statistic. First, we estimate $F_d(\cdot|\cdot)$ by the local constant estimator (LCE) to guarantee that $\widehat{F}_d(y|x)$ is monotone in y, which seems convenient to obtain the quantile functions. Theoretically, the LPE or the LCE based on a higher order kernel can also be used although then $\widehat{F}_d(y|x)$ may not stay in $(0,1)$ and/or be monotone in y in finite samples.[23] Second, $1(Y_j \leq y)$ in $\widehat{F}_d(y|x)$ can be replaced by $G\left(\frac{y-Y_j}{h_0}\right)$ where $G(\cdot)$ is a kernel CDF such as the standard normal CDF, and h_0 may be different from h. Third, we estimate $p(\cdot)$ by the SLE to guarantee $\widehat{p}(x) \in (0,1)$. Although the kernel-type estimator can be used, it seems that the LPE or the LCE based on a higher order kernel must be used to control the bias in semiparametric estimation, which may make $\widehat{p}(x)$ out of $(0,1)$ for some x values; see more discussions in Donald et al. (2014) on why the SLE is preferred and how to specify the series when there are discrete covariates.

Another specification in T_n is $\mathcal{X}$. We replace Assumption X by

---

[23]Alternative estimators that can guarantee these two conditions can be found in Section 6.2 of Li and Racine (2007).

Assumption X′ (Distribution of X): (i) $\text{supp}(X) = \prod_{j=1}^{r}[x_{lj}, x_{uj}]$ is Cartesian product of compact intervals; (ii) the density of X, $f(x)$, is twice continuously differentiable, bounded, and bounded away from 0, on $\mathcal{X}$.

Since $\text{supp}(X)$ is compact, we need to use a boundary kernel in the estimation of $F_d(y|x)$ when x falls in the h neighborhood of a boundary point of $\text{supp}(X)$. Also, a different bandwidth should be used when x is near the boundary of $\text{supp}(X)$. To avoid such complications in practice, we may let $\mathcal{X}$ be a subset of $\mathcal{X}_h \equiv \prod_{j=1}^{r}[x_{lj} + h, x_{uj} - h].$[24] In Assumption X′, we assume X to be continuous, but our test can very easily handle inclusion of discrete regressors as well. The nonparametric distribution function estimation would simply include both types of regressors, either by doing a separate kernel regression for each discrete cell, or by smoothing over cells as in, e.g., Li and Racine (2008). The rest of the test would then proceed exactly as before.

5.1 Asymptotics for T_n

We first impose some assumptions on the kernel and bandwidth.

Assumption K (Kernel): $K(\cdot)$ is a nonnegative, bounded, symmetric, twice continuously differentiable function, zero outside a bounded set, and $\int K(u)du = 1$.

Assumption H (Bandwidth): $h \to 0, nh^r/\ln n \to \infty, nh^{r/2}h^4 \to 0$.

In Assumption H, $nh^r/\ln n \to \infty$ guarantees the uniform linear approximation of $\widehat{F}_d(y|x)$ to be possible; $nh^{r/2}h^4 \to 0$ guarantees the bias of $\widehat{F}_d(y|x)$ to be asymptotically neglectable in T_n. These two restrictions on h cannot be satisfied simultaneously unless $r < 8$; if $r \geq 8$, our nonparametric test does not seem practical. In theory, when $r \geq 8$, the LPE or the LCE based on a higher order kernel can be used. In this case, the results in Kong et al. (2010) can be applied to guarantee the uniform linear approximation of $\widehat{F}_d(y|x)$.

We next impose some assumptions on the propensity score and the conditional CDF of Y_d, which are relaxation of Assumptions P and DR in Section 4.1.

Assumption P′ (Propensity Score): For all $x \in \text{supp}(X)$, $p(x)$ is continuously differentiable of order $s \geq 7r$, and $p(x)$ is bounded away from zero and one: $0 < \underline{p} \leq p(x) \leq \overline{p} < 1$.

Assumption Fd (Conditional CDF): (i) $F_d(y|x)$ is twice continuously differentiable on $\mathcal{X}$ uniformly in $y \in \mathcal{Y}_d$, where $\mathcal{Y}_d\mathcal{X}$ is compact, and $\mathcal{Y}_1$ contains an ϵ-enlargement of the set $\{Q_1(F_0(y|x)|x)|x \in \mathcal{X}, y \in \mathcal{Y}_0\}$. (ii) $f_d(y|x)$ is uniformly bounded, and is uniformly continuous for $(y, x) \in \mathcal{Y}_d\mathcal{X}$, $f_1(y|x)$ is continuously differentiable in y, and $f_1(y|x) > 0$ for $(y, x) \in \mathcal{Y}_1\mathcal{X}$.

We finally impose some assumptions on the SLE of $p(\cdot)$ which are borrowed from HIR.

Assumption S (SLE): The SLE of $p(x)$ uses a power series with $K = an^\nu$ for some $a > 0$ and $r/4(s-r) < \nu < 1/9$.

As in Section 4.1, we also consider the local alternative $H_1^{\delta\prime}$:

$$
\begin{aligned}
U_{0n}^x &= U_{1n}^x \text{ for all } x \in \mathcal{X}, \\
p_n(x) &= (1 - \delta_p/\sqrt{nh^{r/2}})p_*(x) + \left(\delta_p/\sqrt{nh^{r/2}}\right)\wp(x), \\
F_n^d(y|x) &= (1 - \delta_d/\sqrt{nh^{r/2}})F_*^d(y|x) + \left(\delta_d/\sqrt{nh^{r/2}}\right)\mathfrak{F}^d(y|x),
\end{aligned}
$$

---

[24]If we let $\mathcal{X} = \mathcal{X}_h$, then since $\mathcal{X}_h$ converges to $\text{supp}(X)$ as $n \to \infty$, the asymptotic distributions of our test statistics are the same as when $1(X_i \in \text{supp}(X))$ is used in the test statistic construction. Similarly, when boundary kernels are used as X_i is close to the boundary of $\text{supp}(X)$, the asymptotic distributions are not affected because this part of data points are neglectable asymptotically.

where both $p_*(x)$ and $\wp(x)$ satisfy Assumption P', both $F_*^d(y|x)$ and $\mathfrak{F}^d(y|x)$ satisfy Assumption Fd, and U_{dn}^x is the counterpart of U_d^x under F_n^d. $F_*^d(y|x)$ satisfies $Q_1(F_0(y|x)|x) - q_1(F_0(y)) = 0$ for $yx \in \mathcal{Y}_0\mathcal{X}$, but $\mathfrak{F}^d(y|x)$ does not.

To facilitate the statement of the asymptotic distribution of T_n, define

$$g_{yx}^0(U_{yi}^0, X_i) = \frac{1}{h^{r/2}(1 - p(x))f(x)} U_{yi}^0 K_h(X_i - x),$$

$$g_{yx}^1(U_{yi}^1, X_i) = \frac{1}{h^{r/2}p(x)f(x)} U_{yi}^1 K_h(X_i - x),$$

$$g_{yx}^{01}(U_{yi}^0, U_{Q_1(F_0(y|x)|x)i}^1, X_i) = \frac{g_{yx}^0(U_{yi}^0, X_i) - g_{Q_1(F_0(y|x)|x)x}^1\left(U_{Q_1(F_0(y|x)|x)i}^1, X_i\right)}{f_1(Q_1(F_0(y|x)|x)|x)},$$

where

$$U_{yi}^1 = D_i\left[1(Y_i \le y) - F_1(y|X_i)\right], U_{yi}^0 = (1 - D_i)\left[1(Y_i \le y) - F_0(y|X_i)\right].$$

Note that g_{yx}^{01} is the influence function of $\sqrt{nh^r}\left(\widehat{Q}_1\left(\widehat{F}_0(y|x)|x\right) - Q_1(F_0(y|x)|x)\right)$ uniformly for $x \in \mathcal{X}, y \in \mathcal{Y}_0$. The sample analog of g_{yx}^{01} is

$$\widehat{g}_{yx}^{01}(\widehat{U}_{yi}^0, \widehat{U}_{\widehat{Q}_1(\widehat{F}_0(y|x)|x)i}^1, X_i) = \frac{\widehat{g}_{yx}^0(\widehat{U}_{yi}^0, X_i) - \widehat{g}_{\widehat{Q}_1(\widehat{F}_0(y|x)|x)x}^1(\widehat{U}_{yi}^1, X_i)}{\widehat{f}_1\left(\widehat{Q}_1\left(\widehat{F}_0(y|x)|x\right)\Big| x\right)}, \tag{7}$$

where

$$\widehat{g}_{yx}^0(\widehat{U}_{yi}^0, X_i) = \frac{1}{h^{r/2}(1 - \widehat{p}(x))\widehat{f}(x)} \widehat{U}_{yi}^0 K_h(X_i - x),$$

$$\widehat{g}_{yx}^1(\widehat{U}_{yi}^1, X_i) = \frac{1}{h^{r/2}\widehat{p}(x)\widehat{f}(x)} \widehat{U}_{yi}^1 K_h(X_i - x),$$

$$\widehat{f}_1(y|x) = \frac{(nh^{r+1})^{-1}\sum_{j=1}^n D_j K_h(Y_j - y)K_h(X_j - x)}{(nh^r)^{-1}\sum_{j=1}^n D_j K_h(X_j - x)}$$

with

$$(1 - \widehat{p}(x))\widehat{f}(x) = \frac{1}{nh^r}\sum_{j=1}^n (1 - D_j)K_h(X_j - x), \widehat{p}(x)\widehat{f}(x) = \frac{1}{nh^r}\sum_{j=1}^n D_j K_h(X_j - x),$$

$$\widehat{U}_{yi}^0 = (1 - D_i)\left[1(Y_i \le y) - \widehat{F}_0(y|X_i)\right], \widehat{U}_{yi}^1 = D_i\left[1(Y_i \le y) - \widehat{F}_1(y|X_i)\right].$$

Define

$$a^2 = \int_{\mathcal{X}}\left(\int_{\mathcal{Y}_0}\frac{1}{f_1(Q_1(F_0(y|X_i)|X_i)|X_i)^2}E\left[\left(\frac{U_{yi}^0}{1-p(X_i)} - \frac{U_{Q_1(F_0(y|X_i)|X_i)i}^1}{p(X_i)}\right)^2\Bigg| X_i\right]dF_{Y_0|X}(y|X_i)\right)^2 dX_i,$$

$$b = \int_{\mathcal{X}}\int_{\mathcal{Y}_0}\frac{1}{f_1(Q_1(F_0(y|X_i)|X_i)|X_i)^2}E\left[\left(\frac{U_{yi}^0}{1-p(X_i)} - \frac{U_{Q_1(F_0(y|X_i)|X_i)i}^1}{p(X_i)}\right)^2\Bigg| X_i\right]dF_{Y_0|X}(y|X_i)dX_i.$$

Theorem 5 *Under Assumptions Fd, K, H, P', S, U, X' and Y, the following statements hold:*

(i) *Under H_0,*

$$nh^{r/2}T_n - B_h \rightsquigarrow N\left(0, \sigma^2\right),$$

22

where $\sigma^2 = 2 \int \overline{K}(v)^2 dv \cdot a^2$ with $\overline{K}(v) = \int K(u) K(v+u) du$ being the twofold convolution kernel derived from $K(\cdot)$, and $B_h = h^{-r/2}\overline{K}(0)b$ can be consistently estimated by

$$\widehat{v}_n^2 = \frac{2}{n^2 h^r} \sum_{j=1}^{n} \sum_{l \neq j} \widehat{w}_{jl}^2 \ \ and \ \ \widehat{b}_n = \frac{1}{nh^{r/2}} \sum_{j=1}^{n} \widehat{w}_{jj}$$

with

$$\widehat{w}_{jl} = n^{-1} \sum_{i=1}^{n} 1(Y_i \in \mathcal{Y}_0, X_i \in \mathcal{X}) \frac{1 - D_i}{1 - \widehat{p}(X_i)}$$
$$\cdot \widehat{g}_{Y_i X_i}^{01} \left(\widehat{U}_{Y_i j}^0, \widehat{U}_{\widehat{Q}_1(\widehat{F}_0(Y_i|X_i)|X_i)j}^1, X_j \right) \widehat{g}_{Y_i X_i}^{01} \left(\widehat{U}_{Y_i l}^0, \widehat{U}_{\widehat{Q}_1(\widehat{F}_0(Y_i|X_i)|X_i)l}^1, X_l \right).$$

As a result, the test based on the studentized test statistic $1\left(\frac{nh^{r/2}T_n - \widehat{b}_n}{\widehat{v}_n} > z_\alpha\right)$ has the significance level α, where z_α is the $1 - \alpha$ quantile of the standard normal distribution.[25]

(ii) *Under $H_1^{\delta'}$ and Assumption LA,*

$$nh^{r/2}T_n - B_h \rightsquigarrow N\left(\Delta, \sigma^2\right), \ \frac{nh^{r/2}T_n - \widehat{b}_n}{\widehat{v}_n} \rightsquigarrow N\left(\frac{\Delta}{\sigma}, 1\right)$$

where

$$\Delta = \int b(y, x)^2 d\mu(y, x)$$

$b(y, x)$ and $\mu(y, x)$ are defined in Theorem 3(ii).

(iii) *Under the fixed alternative H_1 with $\text{plim}_{n \to \infty} T_n > 0$,*

$$\lim_{n \to \infty} P\left(\frac{nh^{r/2}T_n - \widehat{b}_n}{\widehat{v}_n} > c_n\right) = 1$$

for any nonstochastic constant $c_n = o(nh^{r/2})$.

We provide a few comments on Theorem 5. First, the effect of $\widehat{q}_1\left(\widehat{F}_0(y)\right)$ is asymptotically neglectable because as a semiparametric estimator it is $\sqrt{n}$-consistent. From DH, the influence function of $\sqrt{n}\left(\widehat{q}_1\left(\widehat{F}_0(y)\right) - q_1\left(F_0(y)\right)\right)$ uniformly in $y \in \mathcal{Y}_0$ is

$$\Psi_y^u(W) = \frac{\psi_y^0(W) - \psi_{q_1(F_0(y))}^1(W)}{f_1(q_1(F_0(y)))},$$

where

$$\psi_y^0(W) = \frac{1 - D}{1 - p(X)}1(Y \leq y) - F_0(y) + F_0(y|X)\frac{D - p(X)}{1 - p(X)},$$
$$\psi_y^1(W) = \frac{D}{p(X)}1(Y \leq y) - F_1(y) - F_1(y|X)\frac{D - p(X)}{p(X)}.$$

Comparing with the parametric case, $E\left[\frac{\lambda(X'\gamma_0)}{1-p(X)}F_0(y|X)X'\right] E\left[\frac{\lambda(X'\gamma_0)^2}{p(X)(1-p(X))}XX'\right]^{-1} X\frac{\lambda(X'\gamma_0)}{p(X)}$ in $\psi_0(W, y)$ is replaced by $F_0(y|X)$. In other words, the parametric case projects $F_0(y|X)$ on the space spanned by $\frac{\lambda(X'\gamma_0)}{p(X)}X$ along the orthogonal space spanned by $\frac{\lambda(X'\gamma_0)}{1-p(X)}X$ (see Chapter 3 of Ruud (2000) for the definition

---

[25] This is a one-side test because T_n is based on the L^2-distance between $Q_1(F_0(Y_0|X)|X)$ and $q_1(F_0(Y_0))$.

23

and characterization of projection along a subspace). Similarly, the parametric case projects $F_1(y|X)$ on the space spanned by $\frac{\lambda(X'\gamma_0)}{1-p(X)}X$ along the orthogonal space spanned by $\frac{\lambda(X'\gamma_0)}{p(X)}X$. It is well known that $F_d(y|X)$ is the projection of $1(Y_d \leq y)$ on the space spanned by all functions of X that are square integrable, so the parametric case projects $F_d(y|X)$ further on a finite-dimensional space along some specific direction. Second, as in Proposition 1 of Härdle and Mammen (1993), $nh^{r/2}T_n$ has a positive bias $B_h = O(h^{-r/2})$ because $T_n > 0$ for any n. Third, different from Proposition 2 of Härdle and Mammen (1993), the bias in estimating $\widehat{F}_d(y|x)$ will not contribute to the local power due to undersmoothing. This not only simplifies the local power expression, but simplifies the simulation of critical values as discussed in the next subsection; such undersmoothing techniques are also used in Blundell and Horowitz (2007) and Chernozhukov et al. (2013) for similar reasons. Finally, in both Theorems 3 and 5, the asymptotic distribution under H_0 in (i) is a special case of that under the local alternative, and the consistency of the test statistics in (iii) is well understood, so we will state only the asymptotic distribution under the local alternative in the RP tests for the QTT in the next section.

5.2   Simulating the Critical Values of T_n

As in the classical specification testing (see, e.g., Li and Wang, 1998), the convergence rate of T_n to the normal distribution is quite slow ($O(h^{r/2})$). To approximate the critical value more accurately, the usual literature suggests to use the wild bootstrap (see, e.g., Härdle and Mammen, 1993). In our case, however, it is hard to impose the null if a bootstrap scheme is used. As an alternative, we suggest to use a simulation method to approximate the critical values. We formally summarize the simulation procedure in the following Algorithm S.

Algorithm S:

Step 1: Define $\widehat{g}_{yx}^{01}$ as in (7) with

$$
\begin{aligned}
&(nh^r)^{-1/2}\frac{1}{\sqrt{n}}\widehat{g}_{Y_iX_i}^{01}\left(\widehat{U}_{Y_ij}^0, \widehat{U}_{\widehat{Q}_1\left(\widehat{F}_0(Y_i|X_i)|X_i\right)j}^1, X_j\right)\\
&= \frac{1}{\widehat{f}_1\left(\widehat{Q}_1\left(\widehat{F}_0(Y_i|X_i)|X_i\right)|X_i\right)}\left\{\frac{[1(Y_j \leq Y_i)-\widehat{F}_0(Y_i|X_j)](1-D_j)}{\sum_{l=1}^n(1-D_l)K_h(X_l-X_i)}K_h(X_j-X_i)\right.\\
&\left.\quad - \frac{[1(Y_j \leq \widehat{Q}_1\left(\widehat{F}_0(Y_i|X_i)|X_i\right))-\widehat{F}_1\left(\widehat{Q}_1\left(\widehat{F}_0(Y_i|X_i)|X_i\right)|X_j\right)]D_j}{\sum_{l=1}^n D_l K_h(X_l-X_i)}K_h(X_j-X_i)\right\}.
\end{aligned}
\tag{8}
$$

Step 2: Define

$$
T_n^\xi = \frac{1}{n}\sum_{i=1}^n 1(Y_i \in \mathcal{Y}_0, X_i \in \mathcal{X})\frac{1-D_i}{1-\widehat{p}(X_i)}\left[(nh^r)^{-1/2}\frac{1}{\sqrt{n}}\sum_{j=1}^n \xi_j\widehat{g}_{Y_iX_i}^{01}\left(\widehat{U}_{Y_ij}^0, \widehat{U}_{\widehat{Q}_1\left(\widehat{F}_0(Y_i|X_i)|X_i\right)j}^1, X_j\right)\right]^2,
$$

where ξ_j are iid $N(0,1)$, independent of the original data.

Step 3: Simulate T_n^ξ B times to get $\left\{T_{nb}^\xi\right\}_{b=1}^B$ for B large enough, and then reject H_0 if $T_n > \widehat{c}_n^\xi(\alpha)$, where $\widehat{c}_n^\xi(\alpha)$ is the $(1-\alpha)$th quantile of $\left\{T_{nb}^\xi\right\}_{b=1}^B$ which approximates the $(1-\alpha)$th quantile of T_n^ξ, say, $c_n^\xi(\alpha)$. Of course, we can also check whether the p-value $B^{-1}\sum_{b=1}^B 1(T_{nb}^\xi \geq T_n)$ is less than α to decide whether to reject H_0.

The following theorem states the validity of approximating the critical value of T_n by the quantile of T_n^ξ.

Theorem 6 *Under Assumptions Fd, K, H, P′, S, U, X′ and Y,*

$$\sup_{z\in\mathbb{R}}\left|P\left(\frac{nh^{r/2}T_n^\xi-\widehat{b}_n}{\widehat{v}_n}\leq z\,\middle|\,\mathcal{F}_n\right)-\Phi(z)\right|=o_p(1),$$

where $\widehat{b}_n$ and $\widehat{v}_n$ are defined in Theorem 5(i), and $\mathcal{F}_n=\{W_i\}_{i=1}^n$.

We provide a few comments on this simulation procedure. First, since the effect of $\widehat{q}_1\left(\widehat{F}_0(y)\right)$ is asymptotically neglectable in T_n, we do not need to simulate its influence function.[26] Second, although we did not explicitly impose the null in Algorithm S, the simulation procedure is valid. This is because when we simulate g_{yx}^{01}, the influence function of $\sqrt{nh^r}\left(\widehat{Q}_1\left(\widehat{F}_0(y|x)|x\right)-Q_1\left(F_0(y|x)|x\right)\right)$, we implicitly impose the null. Of course, if the original data are from the alternative, the probability limit of $\widehat{b}_n$ and $\widehat{v}_n$ may not be the same as in Theorem 5(i) because the distribution of W would be different. However, studentization of T_n^ξ ensures its asymptotic distribution invariant to the distribution of W just as in the studentization of T_n in Theorem 5(i). Third, from the proof of Theorem 6, we can replace $\widehat{b}_n$ and $\widehat{v}_n^2$ in the theorem by $\widehat{b}_n^\xi\equiv\frac{1}{nh^{r/2}}\sum_{j=1}^n\xi_j^2\widehat{w}_{jj}$ and $\widehat{v}_n^{\xi2}\equiv\frac{2}{n^2h^r}\sum_{j=1}^n\sum_{l\neq j}\xi_j^2\xi_l^2\widehat{w}_{jl}^2$, respectively. Then we need to modify T_n^ξ to $\frac{nh^{r/2}T_n^\xi-\widehat{b}_n^\xi}{\widehat{v}_n^\xi}$ and compare it with $\frac{nh^{r/2}T_n-\widehat{b}_n}{\widehat{v}_n}$. Our formulation of the theorem avoids estimation of $\widehat{b}_n,\widehat{v}_n^2,\widehat{b}_n^\xi$ and $\widehat{v}_n^{\xi2}$. Fourth, in nonparametric estimation of $F_0(y|x)$ and $F_1(y|x)$, all bandwidths are the same. We can allow the bandwidth vectors h_0 in $\widehat{F}_0(y|x)$ and h_1 in $\widehat{F}_1(y|x)$ to be different without difficulty. The only difference in Algorithm S is that the h in the first term in the brace of (8) is changed to h_0 and the h in the second term is changed to h_1.

6 Testing Rank Preservation For the QTT

Parametric and nonparametric constructions of $\widehat{p}(\cdot)$, $\widehat{Q}_1(\cdot|\cdot)$ and $\widehat{F}_0(\cdot|\cdot)$ have been discussed in the last two sections, so we concentrate on the construction of $\widehat{q}_1^t(\cdot)$ and $\widehat{F}_0^t(\cdot)$ here. Following DH,

$$\widehat{F}_0^t(y)\;=\;\frac{1}{n}\sum_{i=1}^n\frac{\widehat{p}(X_i)\,(1-D_i)}{1-\widehat{p}(X_i)}1(Y_i\leq y)\,\middle/\,\widehat{p}_0,$$

$$\widehat{F}_1^t(y)\;=\;\frac{1}{n}\sum_{i=1}^n D_i1(Y_i\leq y)\,\middle/\,\widehat{p}_1,$$

where $\widehat{p}_0$ can be either $n^{-1}\sum_{i=1}^n\frac{\widehat{p}(X_i)(1-D_i)}{1-\widehat{p}(X_i)}$ or $n^{-1}\sum_{i=1}^n D_i$, and $\widehat{p}_1$ can be either $n^{-1}\sum_{i=1}^n D_i$ or $n^{-1}\sum_{i=1}^n\widehat{p}(X_i)$. For the parametric test, $\widehat{p}(\cdot)$ in $\widehat{F}_d^t(\cdot)$ takes the parametric form in Section 4 and for the nonparametric test, it takes the nonparametric form in Section 5. Note that $\widehat{F}_1^t(y)$ is automatically (weakly) increasing with jumps at Y_{1i}'s, so we define

$$\widehat{q}_1^t(\tau)=\inf\left\{y|\widehat{F}_1^t(y)\geq\tau\right\}.$$

To derive the asymptotic distribution of T_n^t, we replace Assumption Y by the following Assumption Y′.

Assumption Y′ (Distributions of Y_0 and Y_1 on the Treated): (i) $f_0^t(y)$ is bounded, positive and continuous on $\mathcal{Y}_0^t$, where $\mathcal{Y}_0^t$ is compact. (ii) $f_1^t(y)$ is bounded, positive and continuously differentiable on $\mathcal{Y}_1^t$ which is compact and contains an ϵ-enlargement of the set $\{q_1^t(F_0^t(y)):y\in\mathcal{Y}_0^t\}$.

---

[26]Based on limited simulation results (using the setup in the second simulation of Section 8), the performance of T_n^ξ with $\Psi_y^u(W)$ also simulated is worse.

6.1 Asymptotics for the Parametric Test

To facilitate the statement of the asymptotic distribution of T_n^t, define $Z^t(v)$ as a mean zero Gaussian process on $\mathcal{Y}_0^t \mathcal{X}$ with the covariance function

$$\Sigma^t(v_1, v_2) = E\left[Z^t(v_1) Z^t(v_2)\right] = E\left[\left(\Psi_c(W, x_1, y_1) - \Psi_u^t(W, y_1)\right)\left(\Psi_c(W, x_2, y_2) - \Psi_u^t(W, y_2)\right)\right],$$

where Ψ_c is defined in Section 4.1,

$$\Psi_u^t(W, y) = \frac{\psi_0^t(W, y) - \psi_1^t(W, q_1^t(F_0^t(y)))}{f_1^t(q_1^t(F_0^t(y)))},$$

with

$$
\begin{aligned}
\psi_0^t(W, y) &= \frac{1}{E[D]}\left\{ E\left[\frac{\lambda(X'\gamma_0)}{1-p(X)} F_0(y|X) X'\right] E\left[\frac{\lambda(X'\gamma_0)^2}{p(X)(1-p(X))} XX'\right]^{-1} X \frac{\lambda(X'\gamma_0)}{p(X)} \frac{D-p(X)}{1-p(X)} \right. \\
&\quad \left. + \frac{p(X)(1-D)}{1-p(X)} 1(Y \le y) - F_0^t(y)D \right\}, \\
\psi_1^t(W, y) &= \frac{D}{E[D]}\left[1(Y \le y) - F_1^t(y)\right].
\end{aligned}
$$

The terms associated with ψ_1^t and ψ_0^t are the contribution of $\hat{q}_1^t(\cdot)$ and $\widehat{F}_0^t(\cdot)$, respectively. Define λ_i^t's as the eigenvalues of $\Sigma^t(v_1, v_2)$, i.e., for orthonormal eigenfunctions $\varphi_i^t(v)$,

$$\int \Sigma^t(v_1, v_2) \varphi_i^t(v_2)\, d\mu^t(v_2) = \lambda_i^t \varphi_i^t(v_1),$$

where $\mu^t(\cdot)$ is a measure on $\mathcal{Y}_0^t \mathcal{X}$ such that for any measurable set A in $\mathrm{supp}(Y_0 X)$, $\mu^t(A) = \int_{A \cap \mathcal{Y}_0^t \mathcal{X}} \frac{p(x)}{E[D]} dF_{Y_0|X}(y|x) dF_X(x)$,[27] and $\sum_{i=1}^{\infty} \lambda_i^t < \infty$.

Corollary 2 *Suppose Assumptions DR, P, U, X and Y' hold. Then under H_1^δ and Assumption LA,*

$$nT_n^t \rightsquigarrow \sum_{i=1}^{\infty}\left(b_i^t + \varepsilon_i \sqrt{\lambda_i^t}\right)^2 = \sum_{i=1}^{\infty} \lambda_i \chi_{1i}^2\left(\left(b_i^t\right)^2/\lambda_i^t\right),$$

where the ε_i are iid $N(0,1)$, λ_i^t's are eigenvalues of $\Sigma^t(v_1, v_2)$, $b_i^t = \int b^t(v)\, \varphi_i^t(v)\, d\mu^t(v)$ with

$$b^t(y, x) = \frac{\delta_0\left[\mathfrak{F}^0(y|x) - F_*^0(y|x)\right] - \delta_1\left[\mathfrak{F}^1(Q_*^1(F_*^0(y|x)|x)|x) - F_*^0(y|x)\right]}{f_*^1(Q_*^1(F_*^0(y|x)|x)|x)} - \frac{\Delta_F^{0t}(y) - \Delta_F^{1t}(q_*^{1t}(F_*^{0t}(y)))}{f_*^{1t}(q_*^{1t}(F_*^{0t}(y)))}$$

and

$$
\begin{aligned}
\Delta_F^{dt}(y) &= \frac{\delta_\gamma}{E[p_*(X)]}\left\{ E\left[(\wp(X) - p_*(X)) F_*^d(y|X)\right] - \frac{E[\wp(X) - p_*(X)]}{E[p_*(X)]} E\left[p_*(X) F_*^d(y|X)\right]\right\} \\
&\quad + \frac{\delta_d}{E[p_*(X)]} E\left[p_*(X)\left(\mathfrak{F}^d(y|X) - F_*^d(y|X)\right)\right], d = 0, 1,
\end{aligned}
$$

and $\chi_{1i}^2\left(\left(b_i^t\right)^2/\lambda_i^t\right)$'s are independent noncentral χ^2 random variables with noncentral parameters $(b_i^t)^2/\lambda_i^t$. Thus, for any $c > 0$, $P\left(nT_n^t > c|H_1^\delta\right) \ge P\left(nT_n^t > c|H_0^t\right)$, where the equality holds if and only if $b_i^t = 0$ for

---

[27] If define $A_x = \{y \in \mathcal{Y}_0^t | (y, x) \in A\}$, then $\mu^t(A) = E\left[P\left(Y_0 \in A_X | X\right) | D = 1\right]$. $\mu^t(y, x)$ is understood as $\mu^t(A_{yx})$ with $A_{yx} = \{(y, x) \in \mathcal{Y}_0^t \mathcal{X} | y \le y, x \le x\}$.

any i.

The contribution of the perturbation of $F^d(y|x)$ to the local power can be similarly analyzed as in T_n. However, different from T_n, the local power of T_n^t also depends on the perturbation of the propensity score. This is obviously due to the further weight $p(X)/E[D]$ in estimating $F_d^t(y)$. To check the pure effect of $p(\cdot)$ on the local power, suppose $\delta_0 = \delta_1 = 0$. After some manipulation, we can show

$$
\begin{aligned}
& b^t(y,x) \cdot f_*^{1t}(q_*^{1t}(F_*^{0t}(y))) \cdot E[p_*(X)]^2/\delta_\gamma \\
= \ & Cov\left(p_*(X)\left(F_*^1(q_*^{1t}(F_*^{0t}(y))|X) - F_*^0(y|X)\right), \wp(X) - p_*(X)\right) \\
& - Cov\left(p_*(X), (\wp(X) - p_*(X))\left(F_*^1(q_*^{1t}(F_*^{0t}(y))|X) - F_*^0(y|X)\right)\right).
\end{aligned}
$$

Recall that $q_*^{1t}(F_*^{0t}(y)) = Q_*^1(F_*^0(y|x)|x)$ for $x \in \mathrm{supp}(X|(Y_0^t = y))$ such that $F_*^1(q_*^{1t}(F_*^{0t}(y))|x) = F_*^0(y|x)$ for such x, so the power is contributed by the covariance difference which is due to $x \in (\mathrm{supp}(X) \setminus \mathrm{supp}(X|(Y_0^t = y))) \cap \mathcal{X}$. If $\mathrm{supp}(X) = \mathrm{supp}(X|(Y_0^t = y))$ for all $y \in \mathcal{Y}_0^t$, then misspecification in $p(\cdot)$ will not contribute to the power.

We can also show that the exchangeable bootstrap is valid for T_n^t. To be specific, define the bootstrap counterpart of T_n^t as

$$
\begin{aligned}
T_n^{t*} = \ & \frac{1}{\sum_{i=1}^n \omega_i D_i} \sum_{i=1}^n \omega_i 1(Y_i \in \mathcal{Y}_0^t) \frac{\widehat{p}^*(X_i)(1-D_i)}{1-\widehat{p}^*(X_i)} \\
& \cdot \left[\widehat{Q}_1^*\left(\widehat{F}_0^*(Y_i|X_i)|X_i\right) - \widehat{Q}_1\left(\widehat{F}_0(Y_i|X_i)|X_i\right) - \left(\widehat{q}_1^{t*}\left(\widehat{F}_0^{t*}(Y_i)\right) - \widehat{q}_1^t\left(\widehat{F}_0^t(Y_i)\right)\right)\right]^2,
\end{aligned}
$$

where $\widehat{F}_0^*(\cdot|\cdot)$ and $\widehat{Q}_1^*(\cdot|\cdot)$ are defined in Section 4.2, and $\widehat{F}_0^{t*}(\cdot)$ and $\widehat{q}_1^{t*}(\cdot)$ are defined as follows:

$$
\widehat{F}_0^{t*}(y) = \frac{1}{n}\sum_{i=1}^n \omega_i \frac{\widehat{p}^*(X_i)(1-D_i)}{1-\widehat{p}^*(X_i)} 1(Y_i \le y) \bigg/ \frac{1}{n}\sum_{i=1}^n \omega_i D_i
$$

with $\widehat{p}^*$ defined in Section 4.2, and

$$
\widehat{q}_1^{t*}(\cdot) = \inf\left\{y \left| \widehat{F}_1^{t*}(y) \ge \tau\right.\right\}
$$

with

$$
\widehat{F}_1^{t*}(y) = \frac{1}{n}\sum_{i=1}^n \omega_i D_i 1(Y_i \le y) \bigg/ \frac{1}{n}\sum_{i=1}^n \omega_i D_i.
$$

Define $\widehat{c}_n^{*t}(\alpha)$ as the $(1-\alpha)$th quantile of $\{n^* T_{nb}^{t*}\}_{b=1}^B$, which approximates the $(1-\alpha)$th quantile of $n^* T_n^{t*}$, say, $c_n^{*t}(\alpha)$, and then reject H_0^t if $nT_n^t > \widehat{c}_n^{*t}(\alpha)$, where T_{nb}^{t*} is the bth resample of T_n^{t*} and B is large enough. Alternatively, if the p-value $B^{-1}\sum_{b=1}^B 1(n^* T_{nb}^{t*} \ge nT_n^t)$ is less than α then reject H_0^t. The validity of bootstrap here can be parallelly stated as in Theorem 4, so omitted for simplicity.

6.2 Asymptotics for the Nonparametric Test

We state the asymptotic distribution of T_n^t without proof since the proof is similar to that of Theorem 5. We first define the counterparts of a^2 and b in Theorem 5 as

$$
(a^t)^2 = \int_{\mathcal{X}} \left(\int_{\mathcal{Y}_0^t} \frac{1}{f_1(Q_1(F_0(y|X_i)|X_i)|X_i)^2} E\left[\left(\frac{U_{yi}^0}{1-p(X_i)} - \frac{U_{Q_1(F_0(y|X_i)|X_i)i}^1}{p(X_i)}\right)^2 \bigg| X_i\right] \frac{p(X_i)}{E[D]} dF_{Y_0|X}(y|X_i)\right)^2 dX_i,
$$

$$
b^t = \int_{\mathcal{X}} \int_{\mathcal{Y}_0^t} \frac{1}{f_1(Q_1(F_0(y|X_i)|X_i)|X_i)^2} E\left[\left(\frac{U_{yi}^0}{1-p(X_i)} - \frac{U_{Q_1(F_0(y|X_i)|X_i)i}^1}{p(X_i)}\right)^2 \bigg| X_i\right] \frac{p(X_i)}{E[D]} dF_{Y_0|X}(y|X_i) dX_i.
$$

Corollary 3 *Suppose Assumptions Fd, K, H, P′, S, U, X′ and Y′ hold. Then under $H_1^{\delta\prime}$ and Assumption LA,*

$$nh^{r/2}T_n^t - B_h^t \rightsquigarrow N\left(\Delta^t, (\sigma^t)^2\right), \quad \frac{nh^{r/2}T_n^t - \widehat{b}_n^t}{\widehat{v}_n^t} \rightsquigarrow N\left(\frac{\Delta^t}{\sigma^t}, 1\right)$$

where $(\sigma^t)^2 = 2\int \overline{K}(v)^2 dv \cdot (a^t)^2$ with $\overline{K}(v)$ defined in Theorem 5(i), and $B_h^t = h^{-r/2}\overline{K}(0)b^t$ can be consistently estimated by

$$\left(\widehat{v}_n^t\right)^2 = \frac{2}{n^2 h^r}\sum_{j=1}^{n}\sum_{l\neq j}\left(\widehat{w}_{jl}^t\right)^2 \quad and \quad \widehat{b}_n^t = \frac{1}{nh^{r/2}}\sum_{j=1}^{n}\widehat{w}_{jj}^t$$

with

$$\begin{aligned}
\widehat{w}_{jl}^t &= \frac{1}{\sum_{i=1}^{n}D_i}\sum_{i=1}^{n}1(Y_i \in \mathcal{Y}_0^t)1(X_i \in \mathcal{X})\frac{\widehat{p}(X_i)(1-D_i)}{1-\widehat{p}(X_i)} \\
&\quad \cdot \widehat{g}_{Y_iX_i}^{01}\left(\widehat{U}_{Y_ij}^0, \widehat{U}_{\widehat{Q}_1\left(\widehat{F}_0(Y_i|X_i)|X_i\right)j}^1, X_j\right)\widehat{g}_{Y_iX_i}^{01}\left(\widehat{U}_{Y_il}^0, \widehat{U}_{\widehat{Q}_1\left(\widehat{F}_0(Y_i|X_i)|X_i\right)l}^1, X_l\right),
\end{aligned}$$

and

$$\Delta^t = \int b^t(y,x)^2 d\mu^t(y,x)$$

with $b^t(y,x)$ and $\mu^t(\cdot)$ defined in Corollary 2. As a result, the test based on the studentized test statistic $1\left(\frac{nh^{r/2}T_n^t - \widehat{b}_n^t}{\widehat{v}_n^t} > z_\alpha\right)$ has the significance level α, where z_α is the $1-\alpha$ quantile of the standard normal distribution.

As in T_n, the effect of $\widehat{q}_1^t\left(\widehat{F}_0^t(y)\right)$ is asymptotically neglectable. From DH, the influence function of $\sqrt{n}\left(\widehat{q}_1^t\left(\widehat{F}_0^t(y)\right) - q_1^t(F_0^t(y))\right)$ uniformly in $y \in \mathcal{Y}_0^t$ is

$$\Psi_y^{tu}(W) = \frac{\psi_y^{t0}(W) - \psi_{q_1^t(F_0^t(y))}^{t1}(W)}{f_1^t(q_1^t(F_0^t(y)))},$$

where

$$\begin{aligned}
\psi_y^{t0}(W) &= \frac{1}{E[D]}\left\{\frac{p(X)(1-D)}{1-p(X)}1(Y \leq y) + F_0(y|X)\frac{D-p(X)}{1-p(X)} - F_0^t(y)D\right\}, \\
\psi_y^{t1}(W) &= \frac{D}{E[D]}\left[1(Y \leq y) - F_1^t(y)\right].
\end{aligned}$$

The only difference between the asymptotic distribution of T_n^t and T_n is the weight $p(X_i)/E[D]$ in B_h^t and $(\sigma^t)^2$ which is inherited from the definition of $\mu^t(\cdot)$.

We can also simulate the critical values of T_n^t as in Section 5.2. More specifically, define

$$T_n^{t\xi} = \frac{1}{\sum_{i=1}^{n}D_i}\sum_{i=1}^{n}1(Y_i \in \mathcal{Y}_0^t)1(X_i \in \mathcal{X})\frac{\widehat{p}(X_i)(1-D_i)}{1-\widehat{p}(X_i)}\left[(nh^r)^{-1/2}\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\xi_j\widehat{g}_{Y_iX_i}^{01}\left((\widehat{U}_{Y_ij}^0, \widehat{U}_{\widehat{Q}_1\left(\widehat{F}_0(Y_i|X_i)|X_i\right)j}^1, X_j)\right)\right]^2,$$

and adjust Algorithm S correspondingly, where ξ_j and $\widehat{g}_{yx}^{01}$ are defined in T_n^ξ. The validity of simulation here can be parallelly stated as in Theorem 6, so omitted for simplicity.

7 Discussion

We in this section provide more comments on our tests. First, where is the power of our tests from? It turns out that our tests are in fact overidentification tests. From the proof of Theorem 2, $U_0 = U_1$ implies the following moment conditions,

$$Q_{Y_1|X}(F_{Y_0|X}(q_0(U_0)|x)|x) - q_1(U_0) = 0, x \in \text{supp}(X).$$

Obviously, if $\text{supp}(X)$ 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 power of our RP tests for the QTE comes from the overidentification information which originates from multiple (more than one) values of X. Similarly, $U_0^t = U_1^t$ implies $Q_1(F_0(q_0^t(U_0^t)|x)|x) - q_1^t(U_0^t)$ for $x \in \text{supp}(X)$. If $\text{supp}(X)$ includes only a single point, then $Q_1(F_0(q_0^t(U_0^t)|x)|x) = q_1^t(F_0^t(q_0^t(U_0^t))) = q_1^t(U_0^t)$ and no power can be achieved.

Second, we discuss some modifications of our tests. Recall that the reweightings $\frac{1-D}{1-p(X)}$ in T and $\frac{p(X)}{E[D]}\frac{1-D}{1-p(X)}$ in T^t are used to generate the distribution $F_0(\cdot)$ and $F_0^t(\cdot)$ respectively, but they may induce inverse effects in practice since $\widehat{p}(X_i)$ in T_n and T_n^t may be close to 1. To avoid such effects, we can replace both reweightings by $1 - D$ without changing the essential aspects of our tests; especially, $\text{supp}(Y_0X)$ is still recovered. The differences in the asymptotic distributions are (i) $\mu(\cdot)$ in Theorem 3 is redefined as $\mu(A) = \int_{A \cap \mathcal{Y}_0 \mathcal{X}}(1 - p(x))dF_{Y_0|X}(y|x)dF_X(x)$ and $\mu^t(\cdot)$ in Corollary 2 is redefined as $\mu^t(A) = \int_{A \cap \mathcal{Y}_0^t \mathcal{X}}(1 - p(x))dF_{Y_0|X}(y|x)dF_X(x)$ for any measurable set A in $\text{supp}(Y_0X)$; (ii) the integrands of a^2 and b in Theorem 5 are multiplied by $1 - p(X_i)$, and $\frac{p(X_i)}{E[D]}$ in $(a^t)^2$ and b^t of Corollary 3 are replaced by $1 - p(X_i)$; (iii) $\frac{1-D_i}{1-\widehat{p}(X_i)}$ in $\widehat{w}_{jl}$ is replaced by $1 - D_i$ and $\frac{1}{\sum_{i=1}^n D_i}\frac{\widehat{p}(X_i)(1-D_i)}{1-\widehat{p}(X_i)}$ in $\widehat{w}_{jl}^t$ is replaced by $\frac{1-D_i}{n}$. In bootstrapping critical values, $\frac{1-D_i}{1-\widehat{p}^*(X_i)}$ in T_n^* is replaced by $1 - D_i$ and $\frac{1}{\sum_{i=1}^n \omega_i D_i}\frac{\widehat{p}^*(X_i)(1-D_i)}{1-\widehat{p}^*(X_i)}$ in T_n^{t*} is replaced by $\frac{1-D_i}{n^*}$. In simulating the critical values, $\frac{1-D_i}{1-\widehat{p}(X_i)}$ in T_n^ξ is replaced by $1 - D_i$ and $\frac{1}{\sum_{i=1}^n D_i}\frac{\widehat{p}(X_i)(1-D_i)}{1-\widehat{p}(X_i)}$ in $T_n^{t\xi}$ is replaced by $\frac{1-D_i}{n}$. Another modification is to normalize $\mu(\cdot)$ and $\mu^t(\cdot)$ as probability measures. For example, in T_n, we can replace $\frac{1}{n}$ by $1 \Big/ \sum_{i=1}^n 1(Y_i \in \mathcal{Y}_0)1(X_i \in \mathcal{X})\frac{1-D_i}{1-\widehat{p}(X_i)}$, and in T_n^t, replace $\frac{1}{\sum_{i=1}^n D_i}$ by $1 \Big/ \sum_{i=1}^n 1(Y_i \in \mathcal{Y}_0^t)1(X_i \in \mathcal{X})\frac{\widehat{p}(X_i)(1-D_i)}{1-\widehat{p}(X_i)}$. As a result, in the asymptotic distributions of T_n and T_n^t, $\mu(\cdot)$ and $\mu^t(\cdot)$ are replaced by their normalized counterparts, and the bootstrap (simulation) procedures are adjusted correspondingly.

Third, we discuss a few alternative forms of our tests. For this purpose, we first state the following corollary of Theorem 2.

Corollary 4 *Under Assumption M, $F_{X|U_1}(x|u) = F_{X|U_0}(x|u)$ for P_{XU} almost sure (x, u) if and only if $E\left[(Q_0(F_1(q_1(U_1)|X)|X) - q_0(U_1))^2\right] = 0$ or $E\left[(F_1(q_1(U_0)|X) - F_0(q_0(U_0)|X))^2\right] = 0$, where U_0 can be replaced by U_1.*

Proof. The proof for this corollary is parallel to that of Theorem 2, so omitted here. ∎

The first equivalent statement $E\left[(Q_0(F_1(q_1(U_1)|X)|X) - q_0(U_1))^2\right] = 0$ implies that our RP tests for the QTE can base on

$$T' = E\left[D\left(Q_0(F_1(Y|X)|X) - q_0(F_1(Y))\right)^2 1(Y \in \mathcal{Y}_1)\right],$$

where $\mathcal{Y}_1 = \{q_1(\tau)|\tau \in \mathcal{T}_1\}$ with $\mathcal{T}_1$ being a truncation set of quantile index, and the HIR weight $\frac{D}{p(X)}$ is

replaced by D as suggested in the second comment above. Similarly, the RP tests for the QTT can base on

$$T^{t\prime} = E\left[D\left(Q_0(F_1(Y|X)|X) - q_0^t\left(F_1^t(Y)\right)\right)^2 1(Y \in \mathcal{Y}_1^t)\right],$$

where $\mathcal{Y}_1^t = \{q_1^t(\tau)|\tau \in \mathcal{T}_1^t\}$. Their sample analogs T_n' and $T_n^{t\prime}$ can be constructed similarly as T_n and T_n^t. Although T and T' are theoretically equivalent (when no truncation is involved), the sample sizes in their truncation sets may be different, which may affect the finite-sample relative performance of T and T'. As a result, we can base our RP tests for the QTE on $\max\{T_n, T_n'\}$ to robustify power in finite samples. Note that to make T_n and T_n' comparable in magnitude, it is better to use normalized T_n and T_n' as mentioned in the second comment above. By the continuous mapping theorem, the asymptotic distribution of $\max\{T_n, T_n'\}$ is the maximum of the corresponding asymptotic distributions of T_n and T_n' which are correlated. The bootstrap and simulation procedures are adjusted correspondingly; only note that ω_i (ξ_j) used in T_n^* and $T_n^{*\prime}$ $(T_n^\xi$ and $T_n^{\xi\prime})$ are the same to accommodate the correlation between T_n and T_n'. Similarly, we can base the RP tests for the QTT on $\max\{T_n^t, T_n^{t\prime}\}$.

From the second equivalent statement $E\left[(F_1\left(q_1\left(U_0\right)|X\right) - F_0(q_0(U_0)|X))^2\right] = 0$, one may suggest to base our RP tests for the QTE on

$$\widetilde{S} = \int \int_0^1 E\left[\left(F_1\left(q_1\left(u\right)|x\right) - F_0\left(q_0\left(u\right)|x\right)\right)^2\right] du dF_X(x),$$

whose sample analog is

$$\widetilde{S}_n = \frac{1}{n}\sum_{i=1}^n \int_{\mathcal{T}} \left(\widehat{F}_1(\widehat{q}_1\left(u\right)|X_i) - \widehat{F}_0(\widehat{q}_0\left(u\right)|X_i)\right)^2 du,$$

where $\mathcal{T}$ is a truncation set of quantile index. Since $\widetilde{S}_n$ involves the basic ingredients of T_n, $\widehat{F}_d\left(\cdot|\cdot\right)$ and $\widehat{q}_d\left(\cdot\right)$, its asymptotic distribution can be derived based on the techniques used in deriving the asymptotic distribution of T_n. Similarly, the RP tests for the QTT can base on

$$\widetilde{S}^t = \int \int_0^1 E\left[\left(F_1\left(q_1^t\left(u\right)|X\right) - F_0\left(q_0^t\left(u\right)|X\right)\right)^2\right] du dF_X(x),$$

whose sample analog is

$$\widetilde{S}_n^t = \frac{1}{n}\sum_{i=1}^n \int_{\mathcal{T}^t} \left(\widehat{F}_1(\widehat{q}_1^t\left(u\right)|X_i) - \widehat{F}_0(\widehat{q}_0^t\left(u\right)|X_i)\right)^2 du,$$

where $\mathcal{T}^t$ is a truncation set of quantile index. We do not suggest to use $\widetilde{S}_n$ or $\widetilde{S}_n^t$ in practice for two reasons. To simplify discussion, we take $\widetilde{S}_n$ as an example and similar comments apply to $\widetilde{S}_n^t$. (i) it is not easy to intuitively detect the quantile indices at which the null is violated especially when the distribution of X is complicated. To see why, note first that $F_1\left(q_1\left(u\right)|x\right) = P\left(Y_1 \le q_1\left(u\right)|X = x\right) = P\left(U_1 \le u|X = x\right) = F_{U_1|X}(u|x)$ and similarly, $F_0\left(q_0\left(u\right)|x\right) = F_{U_0|X}(u|x)$, so $\widetilde{S}$ is comparing $F_{U_1|X}(u|x)$ and $F_{U_0|X}(u|x)$ for $u \in [0,1]$ and $x \in \text{supp}(X)$. From (3), we can plot $F_{X|U_1}\left(x|u\right)$ versus $F_{X|U_0}(x|u)$ as functions of u to detect the violated quantile indices, where $F_{X|U_d}$ can be estimated from $\left\{(X_i, \widehat{F}_d(Y_i))\right\}_{i=1}^n$ by noting that $\widehat{F}_d(Y_i)$ is unconditional rank of Y_i under the treatment state d. However, if the distribution of X is complicated (e.g., $\dim(X)$ is high and/or $\text{supp}(X)$ is large), we need to draw many pictures to detect the violation of H_0; also, it is not easy to integrate the information in all these pictures. On the contrary, as shown in Proposition 2, we can draw one picture (at most two if T' is also used) to detect the violated quantile indices. (ii) $\widetilde{S}$ implicitly assumes that U_1 and U_0 have the same support (which includes $\mathcal{T}$ as a subset) for all X values.

Under H_0, $U_1|_{X=x}$ and $U_0|_{X=x}$ have the same support, but need not share a common area for all X values; under H_1, $U_1|_{X=x}$ and $U_0|_{X=x}$ may not even have the same support; see Figure 4 for an intuitive illustration. One implied technical difficulty here is that $F_1(q_1(u)|x)$ and $F_0(q_0(u)|x)$ can take extreme values such as 0 and 1, which makes the asymptotic arguments quite hard. For example, in Example 2, $F_0(q_0(0.5)|0) = 1$ and $F_0(q_0(0.5)|1) = 0$. Actually,

$$
\begin{aligned}
E\left[(F_1(q_1(U_0)|X) - F_0(q_0(U_0)|X))^2\right] &= \int\int_0^1 [F_1(q_1(u)|x) - F_0(q_0(u)|x)]^2 \, dF_{U_0|X}(u|x)dF_X(x) \\
&\neq \int\int_0^1 E\left[(F_1(q_1(u)|x) - F_0(q_0(u)|x))^2\right] du dF_X(x)
\end{aligned}
$$

unless $U_0 \perp X$, so $\widetilde{S}$ is implicitly using this extra assumption.[28] To incorporate the joint distribution of (X, U_0), as in T, we replace $q_0(U_0)$ by Y_0 in $E\left[(F_1(q_1(U_0)|X) - F_0(q_0(U_0)|X))^2\right]$ and truncate Y_0 to get

$$
\begin{aligned}
S &= E\left[(F_1(q_1(F_0(Y_0))|X) - F_0(Y_0|X))^2 \, 1(Y_0 \in \mathcal{Y}_0)\right] \\
&= E\left[\frac{1-D}{1-p(X)}(F_1(q_1(F_0(Y))|X) - F_0(Y|X))^2 \, 1(Y_0 \in \mathcal{Y}_0)\right].
\end{aligned}
$$

Parallely, we can base the RP tests for the QTT on

$$
S^t = E\left[\frac{p(X)}{E[D]}\frac{1-D}{1-p(X)}\left(F_1\left(q_1^t\left(F_0^t(Y)\right)|X\right) - F_0(Y|X)\right)^2 1(Y \in \mathcal{Y}_0^t)\right].
$$

It is easy to notice the similarity between S and T and between S^t and T^t.

Fourth, the testing ideas in this paper can also be extended to test some forms of conditional rank preservation. As mentioned in the Introduction, the rank preservation within each X value cannot be tested. However, if we want to test rank preservation in a coarser partition of X values, e.g., for each X_1 value with X_1 being a subset of X, we can still apply our testing ideas and base the RP tests for the QTE on

$$
E\left[(1-D)(Q_1(F_0(Y|X)|X) - Q_1(F_0(Y|X_1)|X_1))^2 \, 1(Y \in \mathcal{Y}_0)\right],
$$

and base the RP tests for the QTT on

$$
E\left[(1-D)(Q_1(F_0(Y|X)|X) - Q_1^t(F_0^t(Y|X_1)|X_1))^2 \, 1(Y \in \mathcal{Y}_0^t)\right],
$$

where the formulas for $\widehat{F}_d(y)$ and $\widehat{F}_d^t(y)$ can still be used to estimate $F_d(Y|X_1 = x_1)$ and $F_d^t(Y|X_1 = x_1)$ except that the data employed are restricted to the $X_1 = x_1$ stratum; see, e.g., Abrevaya et al. (2015). However, it seems that unconditional quantile treatment effects are the most popular in practice, so it may be enough to consider only unconditional rank preservation tests for empirical purposes.

8 Simulations

In this section, we use two simple examples to illustrate the performances of our RP tests for the QTE and QTT. Suppose the relationship between Y_d and X is the same as that in the example of Section 3.1, i.e.,

---

[28] If the measure $F_{U_0|X}(\cdot|x)$ and the uniform measure on $[0,1]$ are dominated by each other for any $x \in \text{supp}(X)$, then $\int\int_0^1 E\left[(F_1(q_1(u)|x) - F_0(q_0(u)|x))^2\right] du dF_X(x)$ is also a valid test statistic. So the key point here is whether $\text{supp}(U_0|X = x) = [0,1]$ for any $x \in \text{supp}(X)$.

$Y_0 = X + \epsilon$, $Y_1 = (2 - a)X + a \cdot \epsilon$. To satisfy unconfoundedness, let $D = 1(X + \eta > 0)$, where X, η and ϵ are independent of each other, $\eta \sim N(0, 1)$, and $\epsilon \sim N(0, 1)$. We will consider four $a - 1$ values, $0, 0.3, 0.6$ and 0.9, indicating the null, small, medium and large local alternatives, respectively.

For the parametric tests, let $X \sim Bernoulli(0.5)$. Because X is binary, the parametric estimation of $p(\cdot), F_0(\cdot|\cdot)$ and $F_1(\cdot|\cdot)$ in Sections 4 and 6.1 does not include misspecification error; choice of the link function $\Lambda(\cdot)$ is irrelevant.

For the nonparametric tests, let $X \sim U[-1, 1]$. The kernel function is set as the quartic kernel $K(u) = \frac{15}{16}(1 - u^2)^2 1(|u| \leq 1)$. The bandwidth is set as in the conditional mean estimation in Section 7.3 of Chernozhukov et al. (2013).[29] As suggested at the end of Section 5.2, we use different bandwidths, h_0 and h_1, for $\widehat{F}_0(\cdot|\cdot)$ and $\widehat{F}_1(\cdot|\cdot)$, respectively. Take h_1 as an example. Now, the data used in estimation are $\{Y_i, X_i\}_{i=1}^{n_1}$ such that the associated $D_i = 1$, where $n_1 = \sum_{i=1}^n D_i$, and we assume the first n_1 individuals are treated. Then

$$h_1 = \widehat{h}_{1,ROT} \times \widehat{s}_1 \times n_1^{1/5} \times n_1^{-2/7},$$

where $\widehat{s}_1$ is the square root of the sample variance of the X_i and $\widehat{h}_{1,ROT}$, is the rule-of-thumb bandwidth for estimation of $m_1(x) \equiv E[Y_i|X_i = x, D_i = 1]$ with studentized X_i, as prescribed in Section 4.2 of Fan and Gijbels (1996). The exact form of $\widehat{h}_{1,ROT}$ is

$$\widehat{h}_{1,ROT} = 2.036 \left[\frac{\widetilde{\sigma}_1^2 \int w_1(x)dx}{n_1^{-1} \sum_{i=1}^{n_1} \left\{ \widetilde{m}_1^{(2)}(\widetilde{X}_i) \right\}^2 w_1(\widetilde{X}_i)} \right]^{1/5} n_1^{-1/5},$$

where $\widetilde{X}_i$'s are studentized X_i's, $\widetilde{m}_1^{(2)}(\cdot)$ is the second-order derivative of the global quartic parametric fit of $m_1(x)$ with studentized X_i, $\widetilde{\sigma}_1^2$ is the simple average of squared residuals from the parametric fit, and $w_1(\cdot)$ is a uniform weight function that has value 1 for any $\widetilde{X}_i$ between the 0.10 and 0.90 sample quantiles of $\widetilde{X}_i$. The factor $n_1^{1/5} \times n_1^{-2/7}$ is multiplied in h_1 to ensure that the bias is asymptotically negligible due to undersmoothing. In simulating the critical values, the bandwidths used in $\widehat{g}_{yx}^0$ and $\widehat{g}_{yx}^1$ are h_0 and h_1, respectively, and the estimation of $\widehat{f}_1(y|x)$ is based on the algorithm kde2d.m of Botev et al. (2010). In estimating the propensity score and the parametric $F_d(\cdot|\cdot)$, Λ is set as the probit link function to avoid model misspecification and the probit fit is conducted by the matlab function glmfit. To improve the power of nonparametric tests, we set $\mathcal{X} = supp(X)$. The simulation study in Müller (1991) shows that a bandwidth without boundary adjustment works well, and we therefore use the same bandwidth for both interior and boundary points.

In both experiments, we consider six test statistics for the QTE and five for the QTT, respectively. As suggested in the second and third comments of Section 7, in the RP tests for the QTE, we consider $T_n, T_n', \max\{T_n, T_n'\}$ and the no-$p(X)$-reweighting counterparts; in the RP tests for the QTT, we consider $T_n^t, T_n^{t'}, \max\{T_n^t, T_n^{t'}\}$ and the no-$p(X)$-reweighting counterparts, with $T_n^{t'}$ the same as its no-$p(X)$-reweighting counterpart. As suggested in the third comment of Section 7, all test statistics (i.e., $T_n, T_n', T_n^t, T_n^{t'}$ and their no-$p(X)$-reweighting counterparts) are normalized to make them comparable. $\mathcal{Y}_d$ and $\mathcal{Y}_d^t$ are chosen as $[\overline{q}_d(0.1), \overline{q}_d(0.9)]$, where $\overline{q}_d(\tau)$ is the τth sample quantile of Y_i with $D_i = d$. 500 replications of both experiments with sample size 400 and 1000 are considered. In bootstrapping or simulating the critical values, the repetition number $B = 399$ for $n = 400$ and $B = 199$ for $n = 1000$. The significance level α is set at 5%. Our simulation study is limited due to computational cost since a bootstrap or simulation cyle is embedded

---

[29]Ideally, the bandwidth in $\widehat{F}_d(y|x)$ should depend on y but selecting bandwidth in this way is too burdensome. Other methods such as cross-validation is too time-consuming. Chernozhukov et al.'s method is designed for the local linear conditional mean estimation. The purpose to use their bandwidth here is only to get the right rate such that undersmoothing is assured.

inside a Monte Carlo cycle.

n	400				1000			
$a-1$	0	0.3	0.6	0.9	0	0.3	0.6	0.9
T_n	0.028	0.298	0.732	0.924	0.052	0.720	0.988	1.000
T_n'	0.012	0.252	0.660	0.874	0.032	0.686	0.986	1.000
$\max\{T_n, T_n'\}$	0.008	0.274	0.726	0.922	0.036	0.708	0.988	1.000
T_n (no-$p(X)$-reweighting)	0.024	0.330	0.748	0.932	0.046	0.754	0.990	1.000
T_n'	0.010	0.242	0.634	0.844	0.034	0.676	0.986	1.000
$\max\{T_n, T_n'\}$	0.008	0.288	0.740	0.932	0.034	0.726	0.990	1.000
T_n^t	0.028	0.328	0.744	0.938	0.052	0.726	0.990	1.000
$T_n^{t'}$	0.008	0.238	0.624	0.836	0.034	0.676	0.984	1.000
$\max\{T_n^t, T_n^{t'}\}$	0.008	0.300	0.740	0.936	0.032	0.710	0.990	1.000
T_n^t (no-$p(X)$-reweighting)	0.028	0.374	0.770	0.946	0.050	0.762	0.992	1.000
$\max\{T_n^t, T_n^{t'}\}$	0.014	0.368	0.768	0.946	0.042	0.758	0.992	1.000

Table 1: Power of Parametric Rank Preservation Tests for the QTE and QTT

n	400				1000			
$a-1$	0	0.3	0.6	0.9	0	0.3	0.6	0.9
T_n	0.022	0.108	0.484	0.782	0.016	0.474	0.970	0.998
T_n'	0.022	0.214	0.608	0.850	0.020	0.574	0.986	1.000
$\max\{T_n, T_n'\}$	0.022	0.116	0.482	0.782	0.014	0.476	0.970	0.998
T_n (no-$p(X)$-reweighting)	0.020	0.168	0.566	0.836	0.026	0.530	0.968	1.000
T_n'	0.012	0.196	0.608	0.842	0.006	0.500	0.964	1.000
$\max\{T_n, T_n'\}$	0.014	0.168	0.564	0.834	0.014	0.530	0.968	1.000
T_n^t	0.018	0.060	0.358	0.646	0.006	0.350	0.904	0.992
$T_n^{t'}$	0.014	0.150	0.476	0.758	0.006	0.408	0.924	0.998
$\max\{T_n^t, T_n^{t'}\}$	0.006	0.064	0.358	0.646	0.002	0.348	0.904	0.992
T_n^t (no-$p(X)$-reweighting)	0.046	0.412	0.830	0.948	0.044	0.812	0.998	1.000
$\max\{T_n^t, T_n^{t'}\}$	0.036	0.394	0.828	0.948	0.030	0.804	0.998	1.000

Table 2: Power of Nonparametric Rank Preservation Tests for the QTE and QTT

The simulation results for parametric RP tests are summarized in Table 1. From Table 1, the following conclusions can be drawn. First, all tests perform satisfactorily well, which matches the prediction of Theorem 2 and Corollary 4; as expected, the size and power when $n = 1000$ are better than those when $n = 400$. Second, when $n = 400$, the sizes of all tests are smaller than the nominal level, so our tests are relatively conservative when the sample size is small.[30] This means that when the tests reject, it is a strong signal that the rank is not preserved. Third, T_n (T_n^t) performs relatively better than T_n' ($T_n^{t'}$), and the performance of $\max\{T_n, T_n'\}$ ($\max\{T_n^t, T_n^{t'}\}$) is in-between (closer to the better one), no matter $p(X)$ is included in reweighting or not. Fourth, no-$p(X)$-reweighting counterparts perform a little bit better than the original tests. Fifth, the powers of the tests for the QTE and for the QTT are similar. The simulation results for nonparametric RP tests are summarized in Table 2. The first, second and fourth conclusions from Table 1 can still be applied here; actually, even when $n = 1000$, there is still the under-sized problem. The third and fifth conclusions can be adjusted as follows. Third, T_n' ($T_n^{t'}$) performs relatively better than T_n

---

[30] Conservative size is not nonstandard in the literature, see, e.g., Bierens and Ploberger (1997) for misspecification testing, Wang and Zivot (1998) for inferences in the weak instruments case, and Abadie (2002) for the stochastic dominance tests.

(T_n^t), and the performance of $\max\{T_n, T_n'\}$ is closer to the worse one (except the no-$p(X)$-reweighting tests for the QTT). Fifth, the powers of the tests for the QTE are better than those for the QTT when $p(X)$ is included in reweighting but worse when $p(X)$ is excluded. Based on these simulation results, we provide two general suggestions to practitioners: (i) use no-$p(X)$-reweighting forms of our tests; (ii) conduct both T_n (T_n^t) and T_n' ($T_n^{t'}$) to check the sensitivity of our tests.

9   Application

We apply our tests to the dataset from the National Supported Work Program (NSW). This dataset was first analyzed by LaLonde (1986) and later by Heckman and Hotz (1989), Dehejia and Wahba (1999), Imbens (2003), Smith and Todd (2001, 2005), Firpo (2007) and Abadie and Imbens (2011) among others. We refer to LaLonde (1986) for detailed descriptions of this dataset. We actually use only subsamples of LaLonde's original sample, termed "RE74 subset" and "PSID-1" in Dehejia and Wahba (1999). RE74 subset contains an experimental sample from a randomized evaluation of the NSW program with 185 individuals treated and 260 untreated. PSID-1 contains the experimental participants in the RE74 subset and a non-experimental comparison group with 2490 individuals from the PSID. Summary statistics can be found in Table 1 of Abadie and Imbens (2011). Because the composition of covariates is quite complicated, we apply only the parametric tests to check rank preservation.

The outcome of interest Y is the earning in 1978 (in thousands of 1982 U.S. dollars). The treatment status D is an indicator for participating in the job training or not. As to the specification of X, we follow the suggestion of DH. Specifically, for RE74 subset, X can be two configurations: (i) a constant, age and the squared age; (ii) a constant, age, age squared, dummies for black, hispanic, married and high school dropout, and earnings in 1974 and 1975. For PSID-1, X can be three configurations: (1) same as configuration (ii) in RE74 subset; (2) configuration (ii) plus education, squared education, squared earnings in 1974 and 1975, and the interaction term between the dummy for black and the dummy for unemployed in 1974; (3) same as configuration (2) except that the interaction terms are replaced by marital status with earnings in 1974 and marital status with the dummy for unemployed in 1974.

In RE74 subset, because the treatment is randomly assigned, Assumption U is satisfied for any covariates X. Under random designs, there is no difference between T and T^t, so we apply only the rank preservation tests for the QTE to RE74 subset.[31] Also, we do not need to adjust the difference in the covariates distribution of the two groups, i.e., $\widehat{p}(X_i)$ in $\widehat{F}_d(\cdot)$ can be replaced by $n^{-1}\sum_{i=1}^n D_i$. 35.4% Y_{0i}'s are zero, so we specify $\mathcal{Y}_0$ as $[q_0(0.4), q_0(0.9)]$ which includes 130 Y_{0i}'s; 24.3% Y_{1i}'s are zero, so we specify $\mathcal{Y}_1$ as $[q_1(0.3), q_1(0.9)]$ which includes 112 Y_{1i}'s. Because both Y_0 and Y_1 have a point mass at zero, distribution regresssion is a more suitable method to estimate $Q_1(\cdot|\cdot)$ and $F_0(\cdot|\cdot)$ than quantile regression. For PSID-1, we apply only the test for the QTT since the non-experimental comparison group is essentially different from the treated group. The range of $F_0^t(Y_0)$ is $[0.358, 1]$, so we specify $\mathcal{Y}_0^t$ as $[q_0^t(0.4), q_0^t(0.9)]$ which includes 332 Y_{0i}'s, and specify $\mathcal{Y}_1^t = \mathcal{Y}_1$. In bootstrapping the critical values, the repetition number $B = 399$.

As suggested by the simulations in Section 8, we use the no-$p(X)$-reweighting normalized tests and check both T_n (T_n^t) and T_n' ($T_n^{t'}$). The results are summarized in Table 3. For RE74 subset, neither configuration of X rejects the null, so we report only the results for the second configuration (which is more general) for simplicity. For PSID-1, the algorithm for specification (3) is not very stable, so we believe the results from specifications (1) and (2) are more reliable. Since the results for these two specifications are similar, we report only the results for specification (2) here. Two general conclusions are (i) we cannot reject rank

---

[31] Note that $D \perp (X, Y_0, Y_1)$ does not restrict the relationship between X and (Y_0, Y_1), so $F_d(\cdot|X)$ need not equal $F_d(\cdot)$ and our tests still have power.

preservation for the QTE in RE74 at the 5% level; (ii) we can reject rank preservation for the QTT in PSID-1 at the 5% level. In other words, the QTE in the application of Firpo (2007) has a causal interpretation, while the QTT does not. More specifically, the p-value for T_n is smaller than that for T'_n, the p-value for $\max\{T_n, T'_n\}$ stays in-between, and all p-values are much larger than 5%. On the other hand, the p-value for T^t_n is larger than that for $T^{t\prime}_n$, the p-value for $\max\{T^t_n, T^{t\prime}_n\}$ is equal to that for T^t_n because T^t_n is much larger than $T^{t\prime}_n$, and all p-values are smaller than 5%. These results match those in the simulations of Section 8.

Datasets	RE74		PSID-1	
	Test Stat.	p-value	Test Stat.	p-value
T_n (T^t_n)	8.10	0.68	64.76	0.045
T'_n $(T^{t\prime}_n)$	3.28	0.93	23.03	0.030
$\max\{T_n, T'_n\}$ $(\max\{T^t_n, T^{t\prime}_n\})$	8.10	0.74	64.76	0.045
T_n $(\mathcal{T}_0 = [0.75, 0.9])$	12.48	0.52		

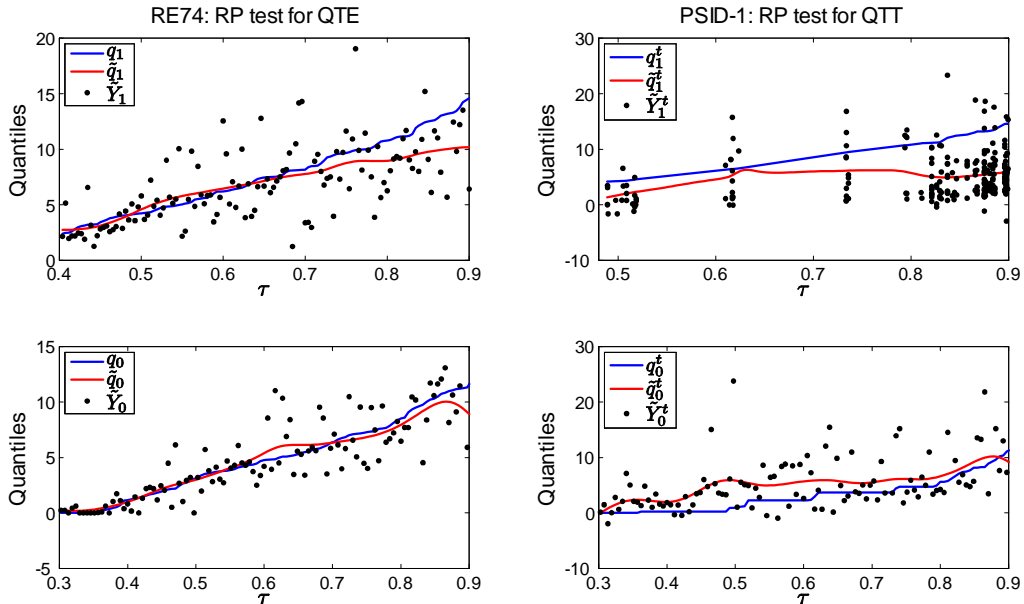Table 3: Parametric Rank Preservation Tests for the QTE in RE74 and for the QTT in PSID-1

Figure 5: $q_d(\tau), \widetilde{q}_d(\tau), \widetilde{Y}_d(\tau)$ in RE74 and $q^t_d(\tau), \widetilde{q}^t_d(\tau), \widetilde{Y}^t_d(\tau)$ in PSID-1

The testing results of Table 3 are intuitively illustrated in Figure 5. The upper-left panel of Figure 5 shows $q_1(\tau), \widetilde{q}_1(\tau)$ and $\widetilde{Y}_1(\tau)$ in T_n and the lower-left panel shows $q_0(\tau), \widetilde{q}_0(\tau)$ and $\widetilde{Y}_0(\tau)$ in T'_n, where $\widetilde{q}_d(\tau)$ is estimated by a local linear smoother. Similarly, the upper-right panel shows $q^t_1(\tau), \widetilde{q}^t_1(\tau)$ and $\widetilde{Y}^t_1(\tau)$ in T^t_n and the lower-right panel shows $q^t_0(\tau), \widetilde{q}^t_0(\tau)$ and $\widetilde{Y}^t_0(\tau)$ in $T^{t\prime}_n$. From Figure 5, $\widetilde{q}_d$ and q_d are close in RE74, while $\widetilde{q}^t_1$ is lower than q^t_1 and $\widetilde{q}^t_0$ is higher than q^t_0 in PSID-1. In RE74, $\widetilde{q}_1(\tau)$ is lower than $q_1(\tau)$ for $\tau \in [0.75, 0.9]$, but this part of information is dominated by the similarity of $\widetilde{q}_1(\tau)$ and $q_1(\tau)$ at other values of τ. To check whether rank preservation can be rejected for this range of τ, we conduct T_n with $\mathcal{T}_0 = [0.75, 0.9]$. The p-value is indeed smaller than that of T_n with $\mathcal{T}_0 = [0.4, 0.9]$, but is still much larger than 5%. Also, there is more randomness in $\widetilde{Y}_d(\tau)$ than in $\widetilde{q}_d(\tau)$, but all the randomness seems due to finite sample variations rather than violation of rank preservation. From the right two panels, we can understand why both T^t_n and $T^{t\prime}_n$ are

much larger than $\max\{T_n, T_n'\}$ and why T_n^t is much larger than $T_n^{t'}$. In summary, such figures can provide information that is buried in our test statistics, e.g., why the test statistics tend to be large, and which part of τ values contributes to the power.

Finally, let's reemphasize that our testing results should be interpreted cautiously. First, the non-rejection of rank preservation in RE74 may be due to the low power when the sample size is small as illustrated in the first simulation of Section 8. Second, our tests are checking whether the rank is preserved across covariate values, so non-rejection of the null does not exclude the possibility that the rank is unpreserved. However, our testing results indicate that if the rank were not preserved, the only possibility is that the rank is not preserved within some covariate values. Third, although the sample size of PSID-1 is so large that the rejection of rank preservation is quite conclusive, it is still possible that the powers of our tests originate from the violation of unconfoundedness because Y_0 comes from an observational comparison group rather than a random design.

10 Conclusion

Rank preservation is important for causal interpretation of quantile treatment effects. In this paper, we propose unconditional rank preservation tests for the QTE and QTT under unconfoundedness. Our tests are Hausman-type tests which are based on the observation that if the unconditional rank is preserved then the conditional rank is preserved but the converse is not true. One key advantage of our tests is that the powers can be intuitively detected by figures. We propose both the parametric and nonparametric tests. Since the asymptotic null distributions are nonstandard, we suggest to use the exchangeable bootstrap in the parametric tests and simulation in the nonparametric tests to obtain critical values. We apply our tests to a dataset from a job training program.

The testing ideas in this paper are not easy to apply to test rank preservation when unconfoundedness fails. For example, in the LATE framework, we may want to test whether the unconditional rank is preserved for the compliers (e.g., to give a causal interpretation for the unconditional quantile treatment effects under endogeneity in Frölich and Melly (2013a)). However, our tests require the knowledge of the identity of a subpopulation while the identity of compliers cannot be identified.[32] Another interesting problem that is not solved in this paper is the power-optimal testing procedure of rank preservation; note that the optimality in Corollary 1 is different from power optimality. Although the simulation studies in Section 8 provide some information on the relative performance of a few tests in finite samples, power-optimal tests in large samples remain a challenge.

---

[32] In the case of one-sided noncompliance (see, e.g., Frölich and Melly, 2013b), we can test rank preservation for the treated (or equivalently, the treated compliers) because the identity of this subpopulation can be identified. See also Dong and Shen (2015) and Frandsen and Lefgren (2015) for testing unconditional rank similarity for the compliers and Yu (2016) for testing conditional rank similarity when unconfoundedness fails.

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. and G. Imbens, 2011, Bias Corrected Matching Estimators for Average Treatment Effects, *Journal of Business and Economic Statistics*, 29, 1–11.

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.

Abrevaya, J., Y.-C. Hsu and R.P. Lieli, 2015, Estimating Conditional Average Treatment Effects, *Journal of Business and Economic Statistics*, 33, 485-505.

Andrews, D.W.K., 1994, Empirical Process Methods in Econometrics, *Handbook of Econometrics*, Vol. 4, R.F. Eagle and D.L. McFadden, eds., New York: Elsevier Science B.V., Ch. 37, 2247-2294.

Arcones, M.A. and E. Giné, 1993, Limit Theorems for U-Processes, *Annals of Probability*, 21, 1494-1542.

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., 1982, Consistent Model Specification Tests, *Journal of Econometrics*, 20, 105-134.

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.

Botev, Z.I., J.F. Grotowski and D.P. Kroese, 2010, Kernel Density Estimation via Diffusion, *Annals of Statistics*, 38, 2916-2957.

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

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, 2009, Improving Point and Interval Estimation of Monotone Functions by Rearrangement, *Biometrika*, 96, 3, 559-575.

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 (CFM), 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., S. Lee and A.M. Rosen, 2013, Intersection Bounds: Estimation and Inference, *Econometrica*, 81, 667-737.

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

Dehejia, R. and S. Wahba, 1999, Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs, *Journal of the American Statistical Association*, 94, 1053–1062.

de Jong, P., 1987, A Central Limit Theorem for Generalized Quadratic Forms, *Probablity and Related Fields*, 75, 261-277.

Doksum, K., 1974, Empirical Probability Plots and Statistical Inference for Nonlinear Models in the Two-Sample Case, *Annals of Statistics*, 2, 267-277.

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

Donald, S.G., Y.-C. Hsu and R.P. Lieli, 2014, Testing the Unconfoundedness Assumption via Inverse Probability Weighted Estimators of (L)ATT, *Journal of Business and Economic Statistics*, 32, 395-415.

Dong, Y. and S. Shen, Testing for Rank Invariance or Similarty in Program Evaluation: The Effect of Training on Earnings Revisited, mimeo, UCI.

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

Fan, Y. and Q. Li, 1996, Consistent Model Specification Tests: Omitted Variables and Semiparametric Functional Forms, *Econometrica*, 64, 865-890.

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. and B. Melly, 2013a, Unconditional Quantile Treatment Effects Under Endogeneity, *Journal of Business and Economic Statistics*, 31, 346-357.

Frölich, M. and B. Melly, 2013b, Identification of Treatment Effects on the Treated With One-Sided Non-compliance, *Econometric Reviews*, 32, 384-414.

Härdle, W. and E. Mammen, 1993, Comparing Nonparametric versus Parametric Regression Fits, *Annals of Statistics*, 21, 1926-1947.

Heckman, J. and J. Hotz, 1989, Choosing among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training, *Journal of the American Statistical Association*, 84, 862–874.

Heckman, J.J., J. Smith and N. Clements, 1997, Making the Most out of Programme Evaluations and Social Experiments Accounting for Heterogeneity in Programme Impacts, *Review of Economic Studies*, 64, 487-535.

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

Hirano, K., G.W. Imbens and G. Ridder (HIR), 2003, Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score, *Econometrica*, 71, 1161–1189.

Holland, P., 1986, Statistics and Causal Inference (with discussion), *Journal of the American Statistical Association*, 81, 945–970.

Hong, Y. and H. White, 1995, Consistent Specificaiton Testing via Nonparametric Series Regression, *Econometrica*, 63, 1133-1159.

Horowitz, J.L., 1996, Semiparametric Estimation of a Regression Model with an Unknown Transformation of the Dependent Variable, *Econometrica*, 64, 103–137.

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

Horowitz, J.L., 2011, Applied Nonparametric Instrumental Variables Estimation, *Econometrica*, 79, 347-394.

Ichimura, H., and O. Linton, 2005, Asymptotic Expansions for Some Semiparametric Program Evaluation Estimators, in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, ed. by D.W.K. Andrews and J.H. Stock, Cambridge, U.K.: Cambridge University Press, Chapter 6, 149-170.

Imbens, G.W., 2003, Sensitivity to Exogeneity Assumptions in Program Evaluation, *American Economic Review*, 93, 126–132.

Imbens, G.W., 2004, Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review, *Review of Economics and Statistics*, 86, 4-29.

Imbens, G.W. and W.K. Newey, 2009, Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity, *Econometrica*, 77, 1481-1512.

Khan, S. and E. Tamer, 2010. Irregular Identification, Support Conditions, and Inverse Weight Estimation, *Econometrica*, 78, 2021–2042.

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

Kong, E., Linton, O. and Y. Xia, 2010, Uniform Bahadur Representation for Local Polynomial Estimates of M-Regression and Its Application to the Additive Model, *Econometric Theory*, 26, 1529-1564.

Kosorok, M.R., 2008, *Introduction to Empirical Processes and Semiparametric Inference*, New York: Springer.

LaLonde, R., 1986, Evaluating the Econometric Evaluation of Training Programs with Experimental Data, *American Economic Review*, 76, 604-620.

Lehmann, E.L., 1974, *Nonparametrics: Statistical Methods Based on Ranks*, San Francisco: Holden-Day.

Li, Q. and J.S. Racine, 2007, *Nonparametric Econometrics: Theory and Practice*, Princeton, N.J. : Princeton University Press.

Li, Q. and J.S. Racine, 2008, Nonparametric Estimation of Conditional CDF and Quantile Functions With Mixed Categorical and Continuous Data, *Journal of Business and Economic Statistics*, 26, 423-434.

Li, Q. and S. Wang, 1998, A Simple Consistent Bootstrap Test for a Parametric Regression Function, *Journal of Econometrics*, 87, 145-165.

MaCurdy, T., X. Chen and H. Hong, 2011, Flexible Estimation of Treatment Effect Parameters, American Economic Review: Papers&Proceedings, 101, 544-551.

Matzkin, R.L., 2003, Nonparametric Estimation of Nonadditive Random Functions, *Econometrica*, 71, 1339-1375.

Müller, H.-G., 1991, Smooth Optimum Kernel Estimators Near Endpoints, *Biometrika*, 78, 521-530.

Newey, W.K., 1994, Kernel Estimation of Partial Means and a General Variance Estimator, *Econometric Theory*, 10, 233-253.

Rubin, D.B., 1974, Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies, *Journal of Educational Psychology*, 66, 688-701.

Rubin, D.B., 1978, Bayesian Inference for Causal Effects: The Role of Randomization, *Annals of Statistics*, 6, 34–58.

Ruud, P.A., 2000, *An Introduction to Classical Econometric Theory*, New York: Oxford University Press.

Shao, J., 2003, *Mathematical Statistics*, 2nd edition, New York: Springer.

Smith, J.A. and P.E. Todd, 2001, Reconciling Conflicting Evidence on the Performance of Propensity-Score Matching Methods, *American Economic Review*, 91, 112–118.

Smith, J.A. and P.E. Todd, 2005, Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators? *Journal of Econometrics*, 125, 305–353.

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

Wang, J. and E. Zivot, 1998, Inference on Structural Parameters in Instrumental Variables Regression with Weak Instruments, *Econometrica*, 66, 1389-1404.

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

Yu, P., 2016, Testing Conditional Rank Similarity With and Without Covariates, mimeo, HKU.

Supplementary Material S.1

S.1.1 Proofs

Proof of Theorem 3.

(i) Note that

$$nT_n = \frac{1}{n}\sum_{i=1}^{n} 1(Y_i \in \mathcal{Y}_0)1(X_i \in \mathcal{X})\frac{1-D_i}{1-\widehat{p}(X_i)}\left[\begin{array}{c}\sqrt{n}\left(\widehat{Q}_1\left(\Lambda\left(X_i'\widehat{\beta}_0(Y_i)\right)|X_i\right) - Q_1\left(\Lambda\left(X_i'\beta_0(Y_i)\right)|X_i\right)\right) \\ -\sqrt{n}\left(\widehat{q}_1\left(\widehat{F}_0(Y_i)\right) - q_1\left(F_0(Y_i)\right)\right)\end{array}\right]^2 .$$

We will show that nT_n with $\widehat{p}(X_i)$ replaced by $p(X_i)$ is $O_p(1)$. Given the uniform consistency of $\widehat{p}(\cdot)$, we can replace $\widehat{p}(X_i)$ by $p(X_i)$ without affecting the asymptotic distribution.

We first check the effect of $\widehat{q}_1(\cdot)$ and $\widehat{F}_0(\cdot)$. From Lemma 1,

$$\sqrt{n}\left(\widehat{F}_d(y) - F_d(y)\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_d(W_j, y) + o_p(1),$$

where $o_p(1)$ is uniform in $y \in \mathcal{Y}_0$. $q_1(\cdot)$, as the inverse function of $F_1(\cdot)$, is Hadamard-differentiable with the derivative being continuous, so by the second part of Theorem 3.9.4 of VW,

$$\sup_{Y_i\in\mathcal{Y}_0}\left|\sqrt{n}\left(\widehat{q}_1\left(\widehat{F}_0(Y_i)\right) - q_1\left(\widehat{F}_0(Y_i)\right)\right) - \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{-\psi_1(W_j, q_1(\widehat{F}_0(Y_i)))}{f_1(q_1(\widehat{F}_0(Y_i)))}\right| = o_p(1),$$

$$\sup_{Y_i\in\mathcal{Y}_0}\left|\sqrt{n}\left(q_1\left(\widehat{F}_0(Y_i)\right) - q_1\left(F_0(Y_i)\right)\right) - \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{\psi_0(W_j, Y_i)}{f_1(q_1(F_0(Y_i)))}\right| = o_p(1).$$

Here, note the difference in the Hadamard derivative of $q_1(\tau)$ with respect to $F_1(\tau)$ and τ; note also that due to the random variation in $\widehat{F}_0(Y_i)$, we require $f_1(y)$ to be bounded, positive and continuous on an interval $[a, b]$ containing an ϵ-enlargement of the set $\{q_1(F_0(y)) : y \in \mathcal{Y}_0\}$ as imposed in Assumption Y(ii). By Assumption Y and DR, Theorem 3 of Andrews (1994) implies that $\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{\psi_1(W_j, y)}{f_1(y)}$ is stochastically equicontinuous with respect to y. Combined with $\sup_{Y_i\in\mathcal{Y}_0}\left|\widehat{F}_0(Y_i) - F_0(Y_i)\right| = o_p(1)$ and the Hadamard differentiability of $q_1(\tau)$ with respect to τ, we have

$$\sup_{Y_i\in\mathcal{Y}_0}\left|\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{\psi_1(W_j, q_1(\widehat{F}_0(Y_i)))}{f_1(q_1(\widehat{F}_0(Y_i)))} - \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{\psi_1(W_j, q_1(F_0(Y_i)))}{f_1(q_1(F_0(Y_i)))}\right| = o_p(1).$$

In summary,

$$\begin{aligned}\sqrt{n}\left(\widehat{q}_1\left(\widehat{F}_0(Y_i)\right) - q_1\left(F_0(Y_i)\right)\right) &= \sqrt{n}\left(\widehat{q}_1\left(\widehat{F}_0(Y_i)\right) - q_1\left(\widehat{F}_0(Y_i)\right)\right) + \sqrt{n}\left(q_1\left(\widehat{F}_0(Y_i)\right) - q_1\left(F_0(Y_i)\right)\right) \\ &= -\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{\psi_1(W_j, q_1(F_0(Y_i)))}{f_1(q_1(F_0(Y_i)))} + \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{\psi_0(W_j, Y_i)}{f_1(q_1(F_0(Y_i)))} + o_p(1) \\ &= \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{\psi_0(W_j, Y_i) - \psi_1(W_j, q_1(F_0(Y_i)))}{f_1(q_1(F_0(Y_i)))} + o_p(1)\end{aligned}$$

1

$$\equiv \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \Psi_u\left(W_j, Y_i\right) + o_p(1),$$

where $o_p(1)$ is uniform in $Y_i \in \mathcal{Y}_0$.

We next check the effect of $\widehat{Q}_1(\cdot|\cdot)$ and $\widehat{F}_0(\cdot|\cdot)$. From Corollary 5.4 of CFM,

$$\sup_{X_i \in \mathcal{X}, Y_i \in \mathcal{Y}_0} \left| \sqrt{n}\left(\Lambda\left(X_i'\widehat{\beta}_0(Y_i)\right) - \Lambda\left(X_i'\beta_0(Y_i)\right)\right) - \frac{1}{\sqrt{n}}\sum_{j=1}^{n} \phi_0(W_j, X_i, Y_i) \right| = o_p(1),$$

$$\sup_{X_i \in \mathcal{X}, Y_i \in \mathcal{Y}_0} \left| \sqrt{n}\left(\Lambda\left(X_i'\widehat{\beta}_1(Y_i)\right) - \Lambda\left(X_i'\beta_1(Y_i)\right)\right) - \frac{1}{\sqrt{n}}\sum_{j=1}^{n} \phi_1(W_j, X_i, Y_i) \right| = o_p(1),$$

where

$$\phi_0(W_j, x, y) = \lambda\left(x'\beta_0(y)\right) x' J_0^{-1}(y) \frac{(1-D_j)\left[1(Y_j \le y) - \Lambda\left(X_j'\beta_0(y)\right)\right]}{\Lambda\left(X_j'\beta_0(y)\right)\left(1 - \Lambda\left(X_j'\beta_0(y)\right)\right)} \lambda\left(X_j'\beta_0(y)\right) X_j,$$

$$\phi_1(W_j, x, y) = \lambda\left(x'\beta_1(y)\right) x' J_1^{-1}(y) \frac{D_j\left[1(Y_j \le y) - \Lambda\left(X_j'\beta_1(y)\right)\right]}{\Lambda\left(X_j'\beta_1(y)\right)\left(1 - \Lambda\left(X_j'\beta_1(y)\right)\right)} \lambda\left(X_j'\beta_1(y)\right) X_j.$$

$Q_1(\cdot|\cdot)$, as the inverse function of $F_1(\cdot|\cdot)$, is Hadamard-differentiable with the derivative being continuous, so

$$\sup_{X_i \in \mathcal{X}, Y_i \in \mathcal{Y}_0} \left| \sqrt{n}\left(\widehat{Q}_1\left(\Lambda\left(X_i'\widehat{\beta}_0(Y_i)\right)|X_i\right) - Q_1\left(\Lambda\left(X_i'\widehat{\beta}_0(Y_i)\right)|X_i\right)\right) - \frac{1}{\sqrt{n}}\sum_{j=1}^{n} \frac{-\phi_1\left(W_j, X_i, Q_1\left(\Lambda\left(X_i'\widehat{\beta}_0(Y_i)\right)|X_i\right)\right)}{f_1\left(Q_1\left(\Lambda\left(X_i'\widehat{\beta}_0(Y_i)\right)|X_i\right)|X_i\right)} \right| = o_p(1),$$

$$\sup_{X_i \in \mathcal{X}, Y_i \in \mathcal{Y}_0} \left| \sqrt{n}\left(Q_1\left(\Lambda\left(X_i'\widehat{\beta}_0(Y_i)\right)|X_i\right) - Q_1\left(F_0(Y_i|X_i)|X_i\right)\right) - \frac{1}{\sqrt{n}}\sum_{j=1}^{n} \frac{\phi_0(W_j, X_i, Y_i)}{f_1\left(Q_1(F_0(Y_i|X_i)|X_i)||X_i\right)} \right| = o_p(1).$$

By the proof of Theorem 5.2 in CFM, $\beta_1(y)$ is continuously differentiable with bounded derivative on $\mathcal{Y}_0$. As a result, Theorem 3 of Andrews (1994) implies that $\frac{1}{\sqrt{n}}\sum_{j=1}^{n} \frac{\phi_1(W_j, x, y)}{f_1(y|x)}$ is stochastically equicontinuous with respect to y. Combined with $\sup_{X_i \in \mathcal{X}, Y_i \in \mathcal{Y}_0}\left|\Lambda\left(X_i'\widehat{\beta}_0(Y_i)\right) - F_0(Y_i|X_i)\right| = o_p(1)$ and the Hadamard differentiability of $Q_1(y|x)$ with respect to y, we have

$$\sup_{X_i \in \mathcal{X}, Y_i \in \mathcal{Y}_0} \left| \frac{1}{\sqrt{n}}\sum_{j=1}^{n} \frac{\phi_1\left(W_j, X_i, Q_1\left(\Lambda\left(X_i'\widehat{\beta}_0(Y_i)\right)|X_i\right)\right)}{f_1\left(Q_1\left(\Lambda\left(X_i'\widehat{\beta}_0(Y_i)\right)|X_i\right)|X_i\right)} - \frac{1}{\sqrt{n}}\sum_{j=1}^{n} \frac{\phi_1\left(W_j, X_i, Q_1(F_0(Y_i|X_i)|X_i)\right)}{f_1\left(Q_1(F_0(Y_i|X_i)|X_i)|X_i\right)} \right| = o_p(1).$$

In summary,

$$\sqrt{n}\left(\widehat{Q}_1\left(\Lambda\left(X_i'\widehat{\beta}_0(Y_i)\right)|X_i\right) - Q_1\left(F_0(Y_i|X_i)|X_i\right)\right)$$

$$= \sqrt{n}\left(\sqrt{n}\left(\widehat{Q}_1\left(\Lambda\left(X_i'\widehat{\beta}_0(Y_i)\right)|X_i\right) - Q_1\left(\Lambda\left(X_i'\widehat{\beta}_0(Y_i)\right)|X_i\right)\right)\right)$$

$$\quad + \sqrt{n}\left(Q_1\left(\Lambda\left(X_i'\widehat{\beta}_0(Y_i)\right)|X_i\right) - Q_1\left(F_0(Y_i|X_i)|X_i\right)\right)$$

$$= -\frac{1}{\sqrt{n}}\sum_{j=1}^{n} \frac{\phi_1\left(W_j, X_i, Q_1(F_0(Y_i|X_i)|X_i)\right)}{f_1\left(Q_1(F_0(Y_i|X_i)|X_i)|X_i\right)} + \frac{1}{\sqrt{n}}\sum_{j=1}^{n} \frac{\phi_0(W_j, X_i, Y_i)}{f_1\left(Q_1(F_0(Y_i|X_i)|X_i)|X_i\right)} + o_p(1)$$

$$= \frac{1}{\sqrt{n}}\sum_{j=1}^{n} \frac{\phi_0(W_j, X_i, Y_i) - \phi_1\left(W_j, X_i, Q_1(F_0(Y_i|X_i)|X_i)\right)}{f_1\left(Q_1(F_0(Y_i|X_i)|X_i)|X_i\right)} + o_p(1)$$

$$\equiv \frac{1}{\sqrt{n}}\sum_{j=1}^{n} \Psi_c\left(W_j, X_i, Y_i\right) + o_p(1),$$

where $o_p(1)$ is uniform in $X_i \in \mathcal{X}, Y_i \in \mathcal{Y}_0$.

Combining the approximations in the last two paragraphs, we have

$$
\begin{aligned}
T_n &= \frac{1}{n^2} \sum_{i=1}^n 1(Y_i \in \mathcal{Y}_0) 1(X_i \in \mathcal{X}) \frac{1 - D_i}{1 - p(X_i)} \left[\frac{1}{\sqrt{n}} \sum_{j=1}^n \left(\Psi_c(W_j, X_i, Y_i) - \Psi_u(W_j, Y_i) \right) \right]^2 + o_p(n^{-1}) \\
&= \frac{1}{n^3} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \kappa(W_i, W_j, W_k) + o_p(n^{-1}),
\end{aligned}
$$

which is a V-statistic, where

$$
\kappa(W_i, W_j, W_k) = 1(Y_i \in \mathcal{Y}_0) 1(X_i \in \mathcal{X}) \frac{1 - D_i}{1 - p(X_i)} \left(\Psi_c(W_j, X_i, Y_i) - \Psi_u(W_j, Y_i) \right) \left(\Psi_c(W_k, X_i, Y_i) - \Psi_u(W_k, Y_i) \right).
$$

From (3.57) of Shao (2003),

$$
\begin{aligned}
T_n &= \frac{1}{n^2} \sum_{j=1}^n \sum_{k=1}^n E\left[\kappa(W_i, W_j, W_k) | W_j, W_k \right] + o_p(n^{-1}) \\
&= \frac{1}{n^2} \sum_{j=1}^n \sum_{k=1}^n \int \left(\Psi_c(W_j, x, y) - \Psi_u(W_j, y) \right) \left(\Psi_c(W_k, x, y) - \Psi_u(W_k, y) \right) d\mu(y, x) + o_p(n^{-1}) \\
&= \frac{1}{n} \int \left[\frac{1}{\sqrt{n}} \sum_{j=1}^n \left(\Psi_c(W_j, x, y) - \Psi_u(W_j, y) \right) \right]^2 d\mu(y, x) + o_p(n^{-1})
\end{aligned}
$$

where $\mu(\cdot)$ is a measure on $\mathcal{Y}_0 \mathcal{X}$ defined in the main text. Because $n^{-1/2} \sum_{j=1}^n \left(\Psi_c(W_j, x, y) - \Psi_u(W_j, y) \right) \rightsquigarrow Z(y, x)$, by the continuous mapping theorem,

$$
n T_n \rightsquigarrow \int Z(x, y)^2 d\mu(y, x).
$$

Following the arguments on pages 1133-1135 of Bierens and Ploberger (1997), we get

$$
\int Z(y, x)^2 d\mu(y, x) \sim \sum_{i=1}^\infty \lambda_i \chi_{1i}^2,
$$

where χ_{1i}^2's are iid χ_1^2 random variables, and λ_i's are eigenvalues of $\Sigma(v_1, v_2)$ with $v = (y, x)$ and $\Sigma(v_1, v_2) = E\left[Z(v_1) Z(v_2) \right]$. Specifically, there exist orthonormal eigenfunctions $\varphi_i(v)$ such that

$$
\int \Sigma(v_1, v_2) \varphi_i(v_2) d\mu(v_2) = \lambda_i \varphi_i(v_1),
$$

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

(ii) Under H_1^δ, $\widehat{p}(\cdot)$ is uniformly consistent to $p_*(\cdot)$, so $n T_n$ still has the same weak limit as $n T_n$ with $\widehat{p}(X_i)$ replaced by $p(X_i)$. Applying Lemma 2.8.7 of VW (p. 174), we know

$$
\sqrt{n} \left(\widehat{Q}_1 \left(\Lambda \left(X_i' \widehat{\beta}_0(y) \right) | x \right) - Q_n^1 \left(F_n^0(y|x) | x \right) \right) - \sqrt{n} \left(\widehat{q}_1 \left(\widehat{F}_0(y) \right) - q_n^1 \left(F_n^0(y) \right) \right)
$$

has the same weak limit as $\frac{1}{\sqrt{n}} \sum_{j=1}^n \left(\Psi_c(W_j, x, y) - \Psi_u(W_j, y) \right)$. Repeating the analysis in (i), we can

3

show that nT_n has the same weak limit as

$$\int \left[\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left(\Psi_c \left(W_j, x, y \right) - \Psi_u \left(W_j, y \right) \right) + \sqrt{n} \left(Q_n^1(F_n^0(y|x)|x) - q_n^1 \left(F_n^0(y) \right) \right) \right]^2 d\mu(y, x).$$

It remains to derive $\sqrt{n} \left(Q_n^1(F_n^0(y|x)|x) - q_n^1 \left(F_n^0(y) \right) \right)$, so we need to find the relationship between $Q_n^1(\cdot|x)$ and $Q_*^1(\cdot|x)$, $F_n^0(\cdot)$ and $F_*^0(\cdot)$, and $q_n^1(\cdot)$ and $q_*^1(\cdot)$. Given that $F_n^1(y|x) = (1 - \delta_1/\sqrt{n})F_*^1(y|x) + (\delta_1/\sqrt{n}) \mathfrak{F}^1(y|x) = F_*^1(y|x) + n^{-1/2}\delta_1 \left[\mathfrak{F}^1(y|x) - F_*^1(y|x) \right]$,

$$Q_n^1(\tau|x) = Q_*^1(\tau|x) - n^{-1/2}\delta_1 \left[\mathfrak{F}^1(Q_*^1(\tau|x)|x) - F_*^1(Q_*^1(\tau|x)|x) \right] / f_*^1(Q_*^1(\tau|x)|x) + o\left(n^{-1/2} \right),$$

where $f_*^1(\cdot|\cdot)$ is the density associated with $F_*^1(\cdot|\cdot)$. Since $F_n^1(y) = E\left[F_n^1(y|X) \right] = F_*^1(y) + n^{-1/2}\delta_1 \left[\mathfrak{F}^1(y) - F_*^1(y) \right]$, where $\mathfrak{F}^1(y) = E\left[\mathfrak{F}^1(y|X) \right]$ and $F_*^1(y) = E\left[F_*^1(y|X) \right]$, we have

$$q_n^1(\tau) = q_*^1(\tau) - n^{-1/2}\delta_1 \left[\mathfrak{F}^1(q_*^1(\tau)) - F_*^1(q_*^1(\tau)) \right] / f_*^1(q_*^1(\tau)) + o\left(n^{-1/2} \right),$$

where $f_*^1(\cdot)$ is the density associated with $F_*^1(\cdot)$. As a result,

$$
\begin{aligned}
&\sqrt{n} \left(Q_n^1(F_n^0(y|x)|x) - q_n^1 \left(F_n^0(y) \right) \right) \\
=\ & \sqrt{n} \left[Q_n^1(F_n^0(y|x)|x) - Q_*^1(F_n^0(y|x)|x) + Q_*^1(F_n^0(y|x)|x) - Q_*^1(F_*^0(y|x)|x) \right] \\
& -\sqrt{n} \left[q_n^1 \left(F_n^0(y) \right) - q_*^1(F_n^0(y)) + q_*^1(F_n^0(y)) - q_*^1 \left(F_*^0(y) \right) \right] \\
\approx\ & \frac{\delta_0 \left[\mathfrak{F}^0(y|x) - F_*^0(y|x) \right] - \delta_1 \left[\mathfrak{F}^1(Q_*^1(F_*^0(y|x)|x)|x) - F_*^0(y|x) \right]}{f_*^1(Q_*^1(F_*^0(y|x)|x)|x)} \\
& - \frac{\delta_0 \left[\mathfrak{F}^0(y) - F_*^0(y) \right] - \delta_1 \left[\mathfrak{F}^1(q_*^1(F_*^0(y))) - F_*^0(y) \right]}{f_*^1(q_*^1(F_*^0(y)))} \\
\equiv\ & b(y, x),
\end{aligned}
$$

where $\approx$ means a higher order term is omitted. From the analysis on pages 1133-1135 of Bierens and Ploberger (1997),

$$nT_n \rightsquigarrow \sum_{i=1}^{\infty} \left(b_i + \varepsilon_i \sqrt{\lambda_i} \right)^2,$$

where the ε_i are iid $N(0, 1)$, and $b_i = \int b(v) \varphi_i(v) d\mu(v)$. The second part of the result follows from Corollary 1 of Bierens and Ploberger (1997).

(iii) This result is because from the analysis in (ii),

$$nT_n \approx \int \left[\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left(\widetilde{\Psi}_c \left(W_j, x, y \right) - \widetilde{\Psi}_u \left(W_j, y \right) \right) + \sqrt{n} \left(Q_1(F_0(y|x)|x) - q_1 \left(F_0(y) \right) \right) \right]^2 d\widetilde{\mu}(y, x) = O_p(n).$$

Here, $\widetilde{\Psi}_c$ and $\widetilde{\Psi}_u$ may be different from Ψ_c and Ψ_u due to the misspecification in H_1, and $n^{-1/2} \sum_{j=1}^{n} (\widetilde{\Psi}_c \left(W_j, x, y \right) - \widetilde{\Psi}_u \left(W_j, y \right)) \rightsquigarrow \widetilde{Z}(y, x)$, where $\widetilde{Z}(y, x)$ is a tight mean zero Gaussian process which is equal to $Z(y, x)$ under H_0. $\widetilde{\mu}(\cdot)$ is a finite measure on $\mathcal{Y}_0 \mathcal{X}$ which is equal to $\mu(\cdot)$ under H_0. $\int \left[(Q_1(F_0(y|x)|x) - q_1 \left(F_0(y) \right)) \right]^2 d\widetilde{\mu}(y, x) = \lim_{n \to \infty} T_n > 0$. So nT_n is dominated by $\int \left[\sqrt{n} \left(Q_1(F_0(y|x)|x) - q_1 \left(F_0(y) \right) \right) \right]^2 d\widetilde{\mu}(y, x)$ which is $O(n)$.

∎

Proof of Theorem 4. We first define $\ell^\infty(\mathcal{F})$ as the space of real-valued bounded functions on the index set equipped with the supremum norm $\|\cdot\|_{\ell^\infty(\mathcal{F})}$, and define $C(\mathcal{F})$ as the space of continuous function on $\mathcal{F}$.

(i) From Corollary 5.4 of CFM, the bootstrap is valid for $\widehat{F}_1(\cdot|\cdot)$ in $\ell^\infty(\mathcal{Y}_1\mathcal{X})$, and is valid for $\widehat{F}_0(\cdot|\cdot)$ in $\ell^\infty(\mathcal{Y}_0\mathcal{X})$. Also, the validity is joint for $\left(\widehat{F}_1(\cdot|\cdot), \widehat{F}_0(\cdot|\cdot)\right)$. From Lemma E.3 of CFM, the bootstrap is valid for $\widehat{F}_0(\cdot)$ in $\ell^\infty(\mathcal{Y}_0)$ and $\widehat{F}_1(\cdot)$ in $\ell^\infty(\mathcal{Y}_1)$ as they are Z-estimators. $Q_1(F_0(\cdot|\cdot)|\cdot)$ is Hadamard differentiable at $(F_1(\cdot|\cdot), F_0(\cdot|\cdot))$ tangentially to $C(\mathcal{Y}_1\mathcal{X}) \times C(\mathcal{Y}_0\mathcal{X})$, so by Theorem 3.9.11 of VW, the bootstrap is valid for $\widehat{Q}_1(\widehat{F}_0(\cdot|\cdot)|\cdot)$. Similarly, since $q_1(F_0(\cdot))$ is Hadamard differentiable at $(F_1(\cdot), F_0(\cdot))$ tangentially to $C(\mathcal{Y}_1) \times C(\mathcal{Y}_0)$, the bootstrap is valid for $\widehat{q}_1(\widehat{F}_0(\cdot))$. Note also that the bootstrap validity for $\widehat{Q}_1(\widehat{F}_0(\cdot|\cdot)|\cdot)$ and $\widehat{q}_1(\widehat{F}_0(\cdot))$ is joint.

To show the bootstrap is valid for T_n, we apply a generalized version of Proposition 7.27 of Kosorok (2008).[33] Specify in Proposition 7.27 that

$$X_n(y,x) = n^* \left[\left(\widehat{Q}_1^* \left(\widehat{F}_0^*(y|x)|x \right) - \widehat{Q}_1 \left(\widehat{F}_0(y|x)|x \right) \right) - \left(\widehat{q}_1^* \left(\widehat{F}_0^*(y) \right) - \widehat{q}_1 \left(\widehat{F}_0(y) \right) \right) \right]^2,$$

$$G_n(y,x) = \frac{1}{n^*} \sum_{i=1}^n \omega_i 1(Y_i \in \mathcal{Y}_0) 1(X_i \in \mathcal{X}) \frac{1 - D_i}{1 - \widehat{p}^*(X_i)} 1(X_i \le x, Y_i \le y),$$

and then

$$n^* T_n^* = \int X_n(y,x) dG_n(y,x).$$

By a multiplier Glivenko-Cantelli theorem (e.g., Lemma 3.6.16 of VW),

$$G_n(y,x) \xrightarrow{P^*} \mu(y,x) \text{ in } \ell^\infty(\mathcal{Y}_0\mathcal{X}).$$

By the continuous mapping theorem,

$$X_n(y,x) \overset{*}{\rightsquigarrow} Z(y,x)^2 \text{ in } \ell^\infty(\mathcal{Y}_0\mathcal{X}).$$

So by Proposition 7.27 of Kosorok (2008),

$$n^* T_n^* \overset{*}{\rightsquigarrow} \int Z(y,x)^2 d\mu(y,x)$$

as desired. It follows that $c_n^*(\alpha) = c(\alpha) + o_p(1)$ under H_0, where $c(\alpha)$ is the $(1-\alpha)$th quantile of $\sum_{i=1}^n \lambda_i \chi_{1i}^2$. This implies that nT_n and $nT_n - (c_n^*(\alpha) - c(\alpha))$ converges to the same limiting distribution as $n \to \infty$, and hence we have that $P(nT_n > c_n^*(\alpha)) = \alpha + o(1)$.

(ii) By Corollary 2.1 of Bickel and Ren (2001, p. 97), the bootstrap is valid for $\left(\widehat{F}_1(\cdot|\cdot), \widehat{F}_0(\cdot|\cdot)\right)$, $\left(\widehat{F}_1(\cdot), \widehat{F}_0(\cdot)\right)$ and $G_n(\cdot)$ if H_1^δ is contiguous to H_0, and thus the arguments in (i) can still go through to show that $c_n^*(\alpha) = c(\alpha) + o_p(1)$ under H_1^δ. By Theorem 3(ii), the result follows.

(iii) Under a fixed alternative, $X_n(y,x) \overset{*}{\rightsquigarrow} \widetilde{Z}(y,x)^2$ in $\ell^\infty(\mathcal{Y}_0\mathcal{X})$ and $G_n(y,x) \xrightarrow{P^*} \widetilde{\mu}(y,x)$ in $\ell^\infty(\mathcal{Y}_0\mathcal{X})$, where $\widetilde{Z}(y,x)$ and $\widetilde{\mu}(y,x)$ are defined in the proof of Theorem 3(iii). So $n^* T_n^* \overset{*}{\rightsquigarrow} \int \widetilde{Z}(y,x)^2 d\widetilde{\mu}(y,x)$, and thus $c_n^*(\alpha) = O_p(1)$. As a result, for any $\epsilon > 0$, there exists a constant M such that $P(c_n^*(\alpha) >$

---

[33] Proposition 7.27 of Kosorok (2008) is stated as follows: Let $X_n, G_n \in D[a,b]$ be stochastic processes with $X_n \rightsquigarrow X$ and $G_n \xrightarrow{p} G$ in $D[a,b]$, where X is bounded with continuous sample paths, G is fixed, and G_n and G have total variation bounded by $K < \infty$. Then $\int_a^{(\cdot)} X_n(s) dG_n(s) \rightsquigarrow \int_a^{(\cdot)} X(s) dG(s)$ in $D[a,b]$.

M) $< \epsilon + o(1)$. Using elementary inequalities, we also have that

$$P(nT_n \leq c_n^*(\alpha)) = P(nT_n \leq c_n^*(\alpha), c_n^*(\alpha) \leq M) + P(nT_n \leq c_n^*(\alpha), c_n^*(\alpha) > M) \leq P(nT_n \leq M) + P(c_n^*(\alpha) > M).$$

From Theorem 3(iii), we know that $P(nT_n \leq M) = o(1)$, and thus $P(nT_n \leq c_n^*(\alpha)) < \epsilon + o(1)$, which implies the statement of the theorem since ϵ can be chosen arbitrarily small.

■

Proof of Theorem 5.

(i) From Lemma 1 and 3, we can show that

$$nh^{r/2}T_n = h^{r/2} \sum_{i=1}^{n} 1(Y_i \in \mathcal{Y}_0, X_i \in \mathcal{X}) \frac{1-D_i}{1-p(X_i)}$$
$$\left[(nh^r)^{-1/2} \mathbb{G}_n(g_{Y_i X_i}^{01}) \left(1 + O_p\left(\sqrt{\ln n/nh^r}\right) \right) + O_p(h^2) - n^{-1/2} \mathbb{G}_n(\Psi_{Y_i}^u) + o_p(n^{-1/2}) \right]^2 + o_p(1)$$
$$= h^{r/2} \sum_{i=1}^{n} 1(Y_i \in \mathcal{Y}_0, X_i \in \mathcal{X}) \frac{1-D_i}{1-p(X_i)} \left(\frac{1}{nh^{r/2}} \sum_{j=1}^{n} g_{Y_i X_i}^{01}(U_{Y_i j}^0, U_{Q_1(F_0(Y_i|X_i)|X_i)j}^1, U_j, X_j)) \right)^2$$
$$+ h^{r/2} \sum_{i=1}^{n} 1(Y_i \in \mathcal{Y}_0, X_i \in \mathcal{X}) \frac{1-D_i}{1-p(X_i)} \left(\frac{1}{n} \sum_{j=1}^{n} \Psi_{Y_i}^u(W_j) \right)^2 + o_p(1)$$
$$+ 2h^{r/2} \sum_{i=1}^{n} 1(Y_i \in \mathcal{Y}_0, X_i \in \mathcal{X}) \frac{1-D_i}{1-p(X_i)} \left[\frac{1}{nh^{r/2}} \sum_{j=1}^{n} g_{Y_i X_i}^{01}(U_{Y_i j}^0, U_{Q_1(F_0(Y_i|X_i)|X_i)j}^1, U_j, X_j)) \right] \left[\frac{1}{n} \sum_{j=1}^{n} \Psi_{Y_i}^u(W_j) \right]$$
$$\equiv J_1 + J_2 + J_3 + o_p(1)$$

where the $o_p(1)$ in the first equality acccounts for the replacement of $\hat{p}(X_i)$ by $p(X_i)$, the $o_p(1)$ in the second equality accounts also for the $O_p\left(\sqrt{\ln n/nh^r}\right)$, $O_p(h^2)$ and $o_p(n^{-1/2})$ terms in the first equality, and g_{yx}^{01} and Ψ_y^u are defined in the lemmas. By Assumption H, the $O_p(h^2)$ term is $o_p(1)$. Since the $O_p\left(\sqrt{\ln n/nh^r}\right)$ term is $o_p(J_1)$ by Assumption H and the $o_p(n^{-1/2})$ term is $o_p(J_2)$, so we need only show that $J_1 = O_p(1)$ and $J_2 = O_p(1)$. By the proof in Theorem 3, $h^{-r/2}J_2 = O_p(1)$, so $J_2 = o_p(1)$. By the Cauchy-Schwarz inequality, $h^{-r/2}J_3 \leq C\sqrt{h^{-r/2}J_1 h^{-r/2}J_2}$, so $J_3 = O\left(\sqrt{J_1 J_2}\right)$. As will be shown below, $J_1 = O_p(1)$, while $J_2 = o_p(1)$, so $J_3 = o_p(1)$. The remaining is to show that $J_1 = O_p(1)$ and find the asymptotic distribution of J_1. It turns out that J_1 is a third-order degenerate V-statistic.

Note that

$$J_1 = \frac{1}{n^2 h^{r/2}} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{l=1}^{n} 1(Y_i \in \mathcal{Y}_0, X_i \in \mathcal{X}) \frac{1-D_i}{1-p(X_i)} g_{Y_i X_i}^{01}(U_{Y_i j}^0, U_{Q_1(F_0(Y_i|X_i)|X_i)j}^1, X_j)) g_{Y_i X_i}^{01}(U_{Y_i l}^0, U_{Q_1(F_0(Y_i|X_i)|X_i)l}^1, X_l))$$
$$= \frac{1}{n^2 h^{r/2}} \sum_{\substack{i \neq j \neq l}}^{n} 1(Y_i \in \mathcal{Y}_0, X_i \in \mathcal{X}) \frac{1-D_i}{1-p(X_i)} g_{Y_i X_i}^{01}(U_{Y_i j}^0, U_{Q_1(F_0(Y_i|X_i)|X_i)j}^1, X_j)) g_{Y_i X_i}^{01}(U_{Y_i l}^0, U_{Q_1(F_0(Y_i|X_i)|X_i)l}^1, X_l))$$
$$+ \frac{1}{n^2 h^{r/2}} \sum_{i=1}^{n} \sum_{j \neq i} 1(Y_i \in \mathcal{Y}_0, X_i \in \mathcal{X}) \frac{1-D_i}{1-p(X_i)} g_{Y_i X_i}^{01}(U_{Y_i j}^0, U_{Q_1(F_0(Y_i|X_i)|X_i)j}^1, X_j))^2 + o_p(1)$$
$$\equiv J_{11} + J_{12} + o_p(1),$$

where the $o_p(1)$ term includes the summands with subscripts $j \neq l$ but either j or l equal to i (which can be shown to be $O_p\left((nh^{r/2})^{-1}\right)$) and the summands with subscripts $i = j = l$ ((which can be shown to be $O_p\left((nh^r)^{-3/2}\right)$)). As long as we can show $J_{11} \xrightarrow{d} N(0, \sigma^2)$, and $J_{12} - B_h = o_p(1)$, the result follows.

6

To show $J_{11} \xrightarrow{d} N(0, \sigma^2)$, we will apply Lemma B.4 of Fan and Li (1996). Note that

$$J_{11} = \frac{\binom{n}{3}}{n^3 h^{r/2}} \left[n \binom{n}{3}^{-1} \sum_{1 \le i < j < l \le n} P_n\left(W_i, W_j, W_l\right) \right],$$

where

$$
\begin{aligned}
P_n\left(W_i, W_j, W_l\right) &= \sum_{3!} 1(Y_i \in \mathcal{Y}_0, X_i \in \mathcal{X}) \frac{1 - D_i}{1 - p(X_i)} \\
&\quad g_{Y_i X_i}^{01}(U_{Y_i j}^0, U_{Q_1(F_0(Y_i|X_i)|X_i)j}^1, X_j)) g_{Y_i X_i}^{01}(U_{Y_i l}^0, U_{Q_1(F_0(Y_i|X_i)|X_i)l}^1, X_l))
\end{aligned}
$$

with $\sum_{3!}$ extending over $3! = 6$ different permutations of i, j, l. Define $P_n(W_j, W_l) = E\left[P_n\left(W_i, W_j, W_l\right)|W_j, W_l\right]$. Then

$$P_n(W_j, W_l) = 2 \int g_{yx}^{01}(U_{yj}^0, U_{Q_1(F_0(y|x)|x)j}^1, X_j)) g_{yx}^{01}(U_{yl}^0, U_{Q_1(F_0(y|x)|x)l}^1, X_l)) d\mu(y, x). \tag{9}$$

Hence,

$$
\begin{aligned}
E\left[P_n^2(W_j, W_l)\right] &= 4E\left[\left(\int g_{yx}^{01}(U_{yj}^0, U_{Q_1(F_0(y|x)|x)j}^1, X_j)) g_{yx}^{01}(U_{yl}^0, U_{Q_1(F_0(y|x)|x)l}^1, X_l)) d\mu(y, x) \right)^2 \right] \\
&= 4E\left[\left(\int \int_{\mathcal{Y}_0} \frac{1}{f_1\left(Q_1\left(F_0(y|X_j^u)|X_j^u\right)|X_j^u\right)^2 f(X_j^u)^2} \left(\frac{U_{yj}^0}{1-p(X_j^u)} - \frac{U_{Q_1(F_0(y|X_j^u)|X_j^u)j}^1}{p(X_j^u)} \right) \left(\frac{U_{yl}^0}{1-p(X_j^u)} - \frac{U_{Q_1(F_0(y|X_j^u)|X_j^u)l}^1}{p(X_j^u)} \right) \right. \right. \\
&\qquad \left. \left. K(u) K\left(\frac{X_l - X_j}{h} + u\right) f_{Y_0|X}(y|X_j^u) f(X_j^u) 1(X_j^u \in \mathcal{X}) dy du \right)^2 \right] \\
&\approx 4 \int \left(\int \int_{\mathcal{Y}_0} \frac{1}{f_1\left(Q_1(F_0(y|X_j)|X_j)|X_j\right)^2 f(X_j)^2} \left(\frac{U_{yj}^0}{1-p(X_j)} - \frac{U_{Q_1(F_0(y|X_j)|X_j)j}^1}{p(X_j)} \right) \left(\frac{U_{yl}^0}{1-p(X_j)} - \frac{U_{Q_1(F_0(y|X_j)|X_j)l}^1}{p(X_j)} \right) \right. \\
&\qquad K(u) K\left(\frac{X_l - X_j}{h} + u\right) f_{Y_0|X}(y|X_j) f(X_j) 1(X_j \in \mathcal{X}) dy du \\
&\qquad \left. dF_{DY|X}(D_j, Y_j|X_j) dF_{DY|X}(D_l, Y_l|X_l) f(X_j) dX_j f(X_l) dX_l \right)^2 \\
&\approx 4 h^r \int \int_{\mathcal{X}} \int \left(\int \int_{\mathcal{Y}_0} \frac{1}{f_1\left(Q_1(F_0(y|X_j)|X_j)|X_j\right)^2 f(X_j)^2} \left(\frac{U_{yj}^0}{1-p(X_j)} - \frac{U_{Q_1(F_0(y|X_j)|X_j)j}^1}{p(X_j)} \right)^2 dF_{DY|X}(D_j, Y_j|X_j) \right. \\
&\qquad K(u) K(v+u) f_{Y_0|X}(y|X_j) f(X_j) dy du \\
&\qquad \left. f(X_j)^2 dX_j dv \right)^2 \\
&= 4 h^r \int_{\mathcal{X}} \left(\int \int_{\mathcal{Y}_0} \frac{1}{f_1\left(Q_1(F_0(y|X_j)|X_j)|X_j\right)^2} \left(\frac{U_{yj}^0}{1-p(X_j)} - \frac{U_{Q_1(F_0(y|X_j)|X_j)j}^1}{p(X_j)} \right)^2 dF_{Y_0|X}(y|X_j) dF_{DY|X}(D_j, Y_j|X_j) \right)^2 dX_j \\
&\qquad \int \overline{K}(v)^2 dv \\
&= 4 h^r \int \overline{K}(v)^2 dv \int_{\mathcal{X}} \left(\int_{\mathcal{Y}_0} \frac{1}{f_1\left(Q_1(F_0(y|X_j)|X_j)|X_j\right)^2} E\left[\left(\frac{U_{yj}^0}{1-p(X_j)} - \frac{U_{Q_1(F_0(y|X_j)|X_j)j}^1}{p(X_j)} \right)^2 \bigg| X_j \right] dF_{Y_0|X}(y|X_j) \right)^2 dX_j \\
&= 4 h^r \sigma_a^2,
\end{aligned}
$$

where in the second equality $X_j^u = X_j + uh$, and $\overline{K}(v) = \int K(u) K(v+u) du$. By tedious calculation, we can show

$$\frac{\binom{n}{3}^2}{n^6 h^r} E\left[P_n\left(W_i, W_j, W_l\right)^2\right] < \infty, \; E\left[P_n\left(W_i, W_j, W_l\right)^2\right] \bigg/ 4 h^r \sigma_a^2 = o(n),$$

$$\frac{E\left[(E[P_n(W_j, W_l) P_n(W_j, W_k)|W_l, W_k])^2\right] + n^{-1} E\left[P_n(W_j, W_l)^4\right]}{\sigma_a^4} \to 0,$$

so by Lemma B.4 of Fan and Li (1996),

$$J_{11} \rightsquigarrow N\left(0, \sigma^2\right),$$

where

$$\sigma^2 = \frac{\binom{n}{3}^2}{n^6 h^r} 4h^r 2^{-1} 3^2 (3-1)^2 \sigma_a^2 = 2\sigma_a^2.$$

To derive the probability limit of J_{12}, we need only calculate its mean and check its variance shrinks to zero. For simplicity, calculate the mean only:

$$E\left[\frac{1}{n^2 h^{r/2}} \sum_{i=1}^n \sum_{j=1}^n 1(Y_i \in \mathcal{Y}_0, X_i \in \mathcal{X}) \frac{1-D_i}{1-p(X_i)} g_{Y_i X_i}^{01}(U_{Y_i j}^0, U_{Q_1(F_0(Y_i|X_i)|X_i)j}^1, X_j))^2\right]$$

$$= h^{-r/2} \int E\left[g_{yx}^{01}(U_{yj}^0, U_{Q_1(F_0(y|x)|x)j}^1, X_j))^2\right] d\mu(y,x)$$

$$= h^{-r/2} \int \int_{\mathcal{Y}_0} E\left[\frac{1}{f_1\left(Q_1\left(F_0(y|X_j^u)|X_j^u\right)\middle| X_j^u\right)^2 f(X_j^u)^2}\left(\frac{U_{yj}^0}{1-p(X_j^u)} - \frac{U_{Q_1\left(F_0(y|X_j^u)|X_j^u\right)j}^1}{p(X_j^u)}\right)^2 K(u)^2\right.$$
$$\left. f_{Y_0|X}(y|X_j^u) f(X_j^u) 1(X_j^u \in \mathcal{X})\right] dy du$$

$$\approx h^{-r/2} \int_{\mathcal{X}} \int \int \int_{\mathcal{Y}_0} \frac{1}{f_1\left(Q_1\left(F_0(y|X_j)|X_j\right)\middle| X_j\right)^2 f(X_j)}\left(\frac{U_{yj}^0}{1-p(X_j)} - \frac{U_{Q_1\left(F_0(y|X_j)|X_j\right)j}^1}{p(X_j)}\right)^2 K(u)^2$$
$$f_{Y_0|X}(y|X_j) dy du dF_{DY|X}(D_j, Y_j|X_j) f(X_j) dX_j$$

$$= h^{-r/2} \overline{K}(0) \int_{\mathcal{X}} \int_{\mathcal{Y}_0} \frac{1}{f_1\left(Q_1\left(F_0(y|X_j)|X_j\right)\middle| X_j\right)^2} E\left[\left(\frac{U_{yj}^0}{1-p(X_j)} - \frac{U_{Q_1\left(F_0(y|X_j)|X_j\right)j}^1}{p(X_j)}\right)^2 \middle| X_j\right] dF_{Y_0|X}(y|X_j) dX_j$$

$$= B_h.$$

By Lemma 4 and 5, σ^2 and B_h can be consistently estimated by $\widehat{v}_n^2$ and $\widehat{b}_n$, respectively.

(ii) Note that

$$nh^{r/2} T_n$$
$$= h^{r/2} \sum_{i=1}^n 1(Y_i \in \mathcal{Y}_0, X_i \in \mathcal{X}) \frac{1-D_i}{1-\widehat{p}(X_i)}$$
$$\cdot \left[\widehat{Q}_1\left(\widehat{F}_0(Y_i|X_i)|X_i\right) - Q_n^1(F_n^0(Y_i|X_i)|X_i) - \left(\widehat{q}_1\left(\widehat{F}_0(Y_i)\right) - q_n^1\left(F_n^0(y)\right)\right) + \left(Q_n^1(F_n^0(Y_i|X_i)|X_i) - q_n^1\left(F_n^0(Y_i)\right)\right)\right]^2$$
$$= h^{r/2} \sum_{i=1}^n 1(Y_i \in \mathcal{Y}_0, X_i \in \mathcal{X}) \frac{1-D_i}{1-p(X_i)}$$
$$\cdot \left[\frac{1}{nh^{r/2}} \sum_{j=1}^n g_{Y_i X_i}^{01}(U_{Y_i j}^{n0}, U_{Q_1(F_0(Y_i|X_i)|X_i)j}^{n1}, X_j)) - \frac{1}{n} \sum_{j=1}^n \Psi_{Y_i}^{nu}(W_j) + \left(Q_n^1(F_n^0(Y_i|X_i)|X_i) - q_n^1\left(F_n^0(Y_i)\right)\right)\right]^2 + o_p(1),$$

where U_{yi}^{nd} is similarly defined as U_{yi}^d but replacing $F_d(y|X_i)$ by $F_n^d(y|X_i)$, and $\Psi_y^{nu}(W_i)$ is similarly defined as $\Psi_y^u(W_i)$ but replacing $F_d(y|X_i)$, $F_d(y)$ and $p(X_i)$ by $F_n^d(y|X_i)$, $F_n^d(y)$ and $p_n(X_i)$. $nh^{r/2} T_n$ includes three terms: the squares of $g_{Y_i X_i}^{01}$, the squares of $Q_n^1 - q_n^1$, and the cross term. By similar arguments as in (i), we can show that the first term has the same asymptotic distribution as under the null, where Assumption LA guarantees that the asymptotic variance and the bias B_h are the same as under the hull. Also, the cross term can be shown to be $o_p(1)$ by checking that its mean converges to

8

0 and its variance shrinks to 0. So the local power is only determined by

$$h^{r/2} \sum_{i=1}^{n} 1(Y_i \in \mathcal{Y}_0, X_i \in \mathcal{X}) \frac{1 - D_i}{1 - p(X_i)} \left[\left(Q_n^1(F_n^0(Y_i|X_i)|X_i) - q_n^1 \left(F_n^0(Y_i) \right) \right) \right]^2. \tag{10}$$

By similar calculations as in the proof of Theorem 3(ii), we can show (10) converges to $\int b(y,x)^2 d\mu(y,x)$, where $b(y,x)$ is defined in Theorem 3(ii).

(iii) Under H_1, following the analysis in (ii), $nh^{r/2}T_n$ minus the term (10) can be approximated by $N\left(B_h, \sigma^2\right)$, where B_h and σ^2 are defined using the distribution under H_1, but the fact that $B_h = O(h^{-r/2}) = o(nh^{r/2})$ and σ^2 is finite remains. Note that the term (10) is $O_p(nh^{r/2})$ under H_1 so it dominates $nh^{r/2}T_n$ and the result follows.

∎

Proof of Theorem 6. Note that $\Phi(z)$ is a continuous function. By Polya's theorem, it suffices to show for any fixed value of $z \in \mathbb{R}$, $\left| P\left(\frac{nh^{r/2}T_n^\xi - \widehat{b}_n}{\widehat{v}_n} \le z | \mathcal{F}_n \right) - \Phi(z) \right| = o_p(1)$.

$$nh^{r/2}T_n^\xi$$
$$\approx \frac{1}{nh^{r/2}} \sum_{j=1}^{n} \sum_{l=1}^{n} \xi_j \xi_l \left\{ \frac{1}{n} \sum_{i=1}^{n} 1(Y_i \in \mathcal{Y}_0, X_i \in \mathcal{X}) \frac{1 - D_i}{1 - \widehat{p}(X_i)} \right.$$
$$\left. \cdot \widehat{g}_{Y_i X_i}^{01} \left(\widehat{U}_{Y_i j}^0, \widehat{U}_{\widehat{Q}_1\left(\widehat{F}_0(Y_i|X_i)|X_i\right)j}^1, X_j \right) \widehat{g}_{Y_i X_i}^{01} \left(\widehat{U}_{Y_i l}^0, \widehat{U}_{\widehat{Q}_1\left(\widehat{F}_0(Y_i|X_i)|X_i\right)l}^1, X_l \right) \right\}$$
$$\equiv \frac{1}{nh^{r/2}} \sum_{j=1}^{n} \sum_{l \ne j} \xi_j \xi_l \widehat{w}_{jl} + \frac{1}{nh^{r/2}} \sum_{j=1}^{n} \xi_j^2 \widehat{w}_{jj} \equiv \sum_{1 \le j < l \le n} U_{n,jl}^\xi + \frac{1}{nh^{r/2}} \sum_{j=1}^{n} \xi_j^2 \widehat{w}_{jj}$$
$$\equiv J_1^\xi + J_2^\xi,$$

where $U_{n,jl}^\xi = 2\xi_j \xi_l \widehat{w}_{jl}/nh^{r/2}$, J_1^ξ is a second-order U-statistic as a function of ξ_j and ξ_l conditional on the original data. To prove the result, we need to show that $J_2^\xi - \widehat{b}_n \overset{*}{\rightsquigarrow} 0$, and $\frac{J_1^\xi}{\widehat{v}_n} \overset{*}{\rightsquigarrow} N(0,1)$. We will use Proposition 3.2 of de Jong (1987) to prove the result. For this purpose, we need to show that

$$G_I = \sum_i \sum_{j>i} E^* \left[U_{n,ij}^{\xi 4} \right] = o_p\left(\widehat{v}_n^4 \right),$$

$$G_{II} = \sum_i \sum_{j>i} \sum_{l>j>i} E^* \left[U_{n,ij}^{\xi 2} U_{n,il}^{\xi 2} + U_{n,ji}^{\xi 2} U_{n,jl}^{\xi 2} + U_{n,li}^{\xi 2} U_{n,lj}^{\xi 2} \right] = o_p\left(\widehat{v}_n^4 \right),$$

$$G_{IV} = \sum_i \sum_{j>i} \sum_{k>j>i} \sum_{l>k>j>i} E^* \left[U_{n,ij}^\xi U_{n,ik}^\xi U_{n,lj}^\xi U_{n,lk}^\xi + U_{n,ij}^\xi U_{n,il}^\xi U_{n,kj}^\xi U_{n,kl}^\xi + U_{n,ik}^\xi U_{n,il}^\xi U_{n,jk}^\xi U_{n,jl}^\xi \right] = o_p\left(\widehat{v}_n^4 \right).$$

where $E^*[\cdot] = E[\cdot|\mathcal{F}_n]$ is the expectation conditional on the original data, and

$$\widehat{v}_n^2 = E^* \left[\left(J_1^\xi \right)^2 \right] = \frac{2}{n^2 h^r} \sum_{j=1}^{n} \sum_{l \ne j} \widehat{w}_{jl}^2.$$

It is straightforward to show that

$$G_I^* = O_p\left(\left(n^2 h^r\right)^{-1} \right), G_{II}^* = O_p(n^{-1}), G_{IV}^* = O_p(h^r).$$

Since $\widehat{v}_n^2 = O_p(1)$, the result follows immediately.

9

It remains to show that $J_2^\xi - \widehat{b}_n \overset{*}{\rightsquigarrow} 0$. $E^* \left[J_2^\xi \right] = \widehat{b}_n$, and $Var^* \left(J_2^\xi \right) = o_p(1)$, which completes the proof.

∎

Proof of Corollary 2. We briefly describe the key steps of the proof given that it is similar to the proof of Theorem 3.

$$
\begin{aligned}
nT_n^t &= \frac{1}{\sum_{i=1}^n D_i} \sum_{i=1}^n 1(Y_i \in \mathcal{Y}_0^t) 1(X_i \in \mathcal{X}) \frac{\widehat{p}(X_i)(1-D_i)}{1-\widehat{p}(X_i)} \left[\sqrt{n} \left(\widehat{Q}_1 \left(\widehat{F}_0(Y_i|X_i)|X_i \right) - \widehat{q}_1^t \left(\widehat{F}_0^t(Y_i) \right) \right) \right]^2 \\
&\approx \frac{1}{n} \frac{1}{E[D]} \sum_{i=1}^n 1(Y_i \in \mathcal{Y}_0^t) 1(X_i \in \mathcal{X}) \frac{p(X_i)(1-D_i)}{1-p(X_i)} \left[\sqrt{n} \left(\widehat{Q}_1 \left(\widehat{F}_0(Y_i|X_i)|X_i \right) - \widehat{q}_1^t \left(\widehat{F}_0^t(Y_i) \right) \right) \right]^2 \\
&\approx \frac{1}{n} \frac{1}{E[D]} \sum_{i=1}^n 1(Y_i \in \mathcal{Y}_0^t) 1(X_i \in \mathcal{X}) \frac{p(X_i)(1-D_i)}{1-p(X_i)} \left[\frac{1}{\sqrt{n}} \sum_{j=1}^n \left(\Psi_c(W_j, X_i, Y_i) - \Psi_u^t(W_j, Y_i) \right) \right]^2 \\
&\approx \frac{1}{n} \int \left[\frac{1}{\sqrt{n}} \sum_{j=1}^n \left(\Psi_c(W_j, x, y) - \Psi_u^t(W_j, y) \right) \right]^2 d\mu^t(y, x) \\
&\rightsquigarrow \int Z^t(y, x)^2 d\mu^t(y, x),
\end{aligned}
$$

where the asymptotic expansion of $\widehat{q}_1^t \left(\widehat{F}_0^t(Y_i) \right)$ is similar as that in the proof of Theorem 3 combined with Lemma 2. By Mercer's theorem, the weak limit can be expressed in the form as stated in the corollary.

Under H_0^δ, we need to calculate $\sqrt{n} \left(Q_n^1(F_n^0(y|x)|x) - q_n^{1t}(F_n^{0t}(y)) \right)$. For this purpose, we must understand $F_n^{dt}(y)$. Recall that $p_n(x) = (1 - \delta_\gamma/\sqrt{n})p_*(x) + (\delta_\gamma/\sqrt{n}) \wp(x)$, so

$$
\begin{aligned}
F_n^{0t}(y) &= E_n \left[\frac{p_n(X)}{E_n[D]} 1(Y_0 \le y) \right] = E \left[\frac{p_n(X)}{E_n[D]} F_n^0(y|X) \right] \\
&= \int_{\mathcal{X}} \frac{p_*(x) + (\delta_\gamma/\sqrt{n})(\wp(x) - p_*(x))}{E_*[D] + (\delta_\gamma/\sqrt{n}) E[\wp(X) - p_*(X)]} \left[F_*^0(y|x) + (\delta_0/\sqrt{n})(\mathfrak{F}^0(y|x) - F_*^0(y|x)) \right] dF_X(x) \\
&= F_*^{0t}(y) + \frac{\delta_\gamma/\sqrt{n}}{E_*[D]} \left\{ E \left[(\wp(X) - p_*(X)) F_*^0(y|X) \right] - \frac{E[\wp(X) - p_*(X)]}{E_*[D]} E \left[p_*(X) F_*^0(y|X) \right] \right\} \\
&\quad + \frac{\delta_0/\sqrt{n}}{E_*[D]} E \left[p_*(X) (\mathfrak{F}^0(y|X) - F_*^0(y|X)) \right] \\
&= F_*^{0t}(y) + n^{-1/2} \Delta_F^{0t}(y),
\end{aligned}
$$

and

$$
\begin{aligned}
F_n^{1t}(y) &= E_n \left[\frac{p_n(X)}{E_n[D]} 1(Y_1 \le y) \right] \\
&= F_*^{1t}(y) + \frac{\delta_\gamma/\sqrt{n}}{E_*[D]} \left\{ E \left[(\wp(X) - p_*(X)) F_*^1(y|X) \right] - \frac{E[\wp(X) - p_*(X)]}{E_*[D]} E \left[p_*(X) F_*^1(y|X) \right] \right\} \\
&\quad + \frac{\delta_1/\sqrt{n}}{E_*[D]} E \left[p_*(X) (\mathfrak{F}^1(y|X) - F_*^1(y|X)) \right] \\
&= F_*^{1t}(y) + n^{-1/2} \Delta_F^{1t}(y).
\end{aligned}
$$

As a result,

$$
q_n^{1t}(\tau) = q_*^{1t}(\tau) - n^{-1/2} \frac{\Delta_F^{1t}(q_*^{1t}(\tau))}{f_*^{1t}(q_*^{1t}(\tau))}.
$$

10

In summary,

$$\sqrt{n}\left(Q_n^1(F_n^0(y|x)|x) - q_n^{1t}\left(F_n^{0t}(y)\right)\right)$$

$$= \sqrt{n}\left[Q_*^1(F_*^0(y|x)|x) - n^{-1/2}\frac{\delta_1\left[\mathfrak{F}^1(Q_*^1(F_*^0(y|x)|x)|x) - F_*^0(y|x)\right]}{f_*^1(Q_*^1(F_*^0(y|x)|x)|x)} + n^{-1/2}\frac{\delta_0\left[\mathfrak{F}^0(y|x) - F_*^0(y|x)\right]}{f_*^1(Q_*^1(F_*^0(y|x)|x)|x)}\right.$$

$$\left. - \left(q_*^{1t}(F_*^{0t}(y)) - n^{-1/2}\frac{\Delta_F^{1t}(q_*^{1t}(F_*^{0t}(y)))}{f_*^{1t}(q_*^{1t}(F_*^{0t}(y)))} + n^{-1/2}\frac{\Delta_F^{0t}(y)}{f_*^{1t}(q_*^{1t}(F_*^{0t}(y)))}\right)\right]$$

$$\rightarrow \frac{\delta_0\left[\mathfrak{F}^0(y|x) - F_*^0(y|x)\right] - \delta_1\left[\mathfrak{F}^1(Q_*^1(F_*^0(y|x)|x)|x) - F_*^0(y|x)\right]}{f_*^1(Q_*^1(F_*^0(y|x)|x)|x)} - \frac{\Delta_F^{0t}(y) - \Delta_F^{1t}(q_*^{1t}(F_*^{0t}(y)))}{f_*^{1t}(q_*^{1t}(F_*^{0t}(y)))}.$$

∎

S.1.2 Lemmas

Lemma 1 $\sqrt{n}\left(\widehat{F}_d(y) - F_d(y)\right) = \frac{1}{\sqrt{n}}\sum_{j=1}^n \psi_d(W_j, y) + o_p(1)$, where $o_p(1)$ is uniform in $y \in \mathcal{Y}_d$.

Proof. Note that

$$\sqrt{n}\left(\widehat{F}_0(y) - F_0(y)\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^n\left(\frac{1-D_i}{1-\widehat{p}(X_i)}1(Y_i \le y) - F_0(y)\right)$$

$$= \frac{1}{\sqrt{n}}\sum_{i=1}^n\left(\frac{1-D_i}{1-\widehat{p}(X_i)}1(Y_i \le y) - \frac{1-D_i}{1-p(X_i)}1(Y_i \le y)\right) + \frac{1}{\sqrt{n}}\sum_{i=1}^n\left(\frac{1-D_i}{1-p(X_i)}1(Y_i \le y) - F_0(y)\right)$$

$$= \frac{1}{n}\sum_{i=1}^n\left(\frac{\sqrt{n}(\widehat{p}(X_i)-p(X_i))}{(1-\widehat{p}(X_i))(1-p(X_i))}(1-D_i)1(Y_i \le y)\right) + \frac{1}{\sqrt{n}}\sum_{i=1}^n\left(\frac{1-D_i}{1-p(X_i)}1(Y_i \le y) - F_0(y)\right)$$

$$= \frac{1}{n\sqrt{n}}\sum_{i=1}^n\sum_{j=1}^n\left(\frac{\lambda(X_i'\gamma_0)X_i'}{(1-p(X_i))^2}(1-D_i)1(Y_i \le y)E\left[\lambda(X'\gamma_0)H(X'\gamma_0)XX'\right]^{-1}(D_j - p(X_j))H(X_j'\gamma_0)X_j\right)$$

$$+ \frac{1}{\sqrt{n}}\sum_{i=1}^n\left(\frac{1-D_i}{1-p(X_i)}1(Y_i \le y) - F_0(y)\right) + o_p(1),$$

where the last equality is because

$$\sqrt{n}\left(\widehat{p}(X_i) - p(X_i)\right) = \lambda(X_i'\overline{\gamma})X_i'\sqrt{n}\left(\widehat{\gamma} - \gamma_0\right)$$

$$= \lambda(X_i'\gamma_0)X_i'E\left[\lambda(X'\gamma_0)H(X'\gamma_0)XX'\right]^{-1}\frac{1}{\sqrt{n}}\sum_{j=1}^n(D_j - p(X_j))H(X_j'\gamma_0)X_j + o_p(1)$$

where $\overline{\gamma}$ is between $\widehat{\gamma}$ and γ_0, $H(\cdot) = \lambda(\cdot)/\{\Lambda(\cdot)[1 - \Lambda(\cdot)]\}$, and $o_p(1)$ is uniform in i given that supp(X) is compact. By Hoeffding's projection of U-statistic (see, e.g., Arcones and Giné, 1993),

$$\frac{1}{n\sqrt{n}}\sum_{i=1}^n\sum_{j=1}^n\left(\frac{\lambda(X_i'\gamma_0)X_i'}{(1-p(X_i))^2}(1-D_i)1(Y_i \le y)E\left[\lambda(X'\gamma_0)H(X'\gamma_0)XX'\right]^{-1}(D_j - p(X_j))H(X_j'\gamma_0)X_j\right)$$

$$= \frac{1}{\sqrt{n}}\sum_{j=1}^n E\left[\frac{\lambda(X_i'\gamma_0)X_i'}{(1-p(X_i))^2}(1-D_i)1(Y_i \le y)E\left[\lambda(X'\gamma_0)H(X'\gamma_0)XX'\right]^{-1}(D_j - p(X_j))H(X_j'\gamma_0)X_j\,\Big|\,W_j\right] + o_p(1)$$

$$= \frac{1}{\sqrt{n}}\sum_{j=1}^n E\left[\frac{\lambda(X_i'\gamma_0)}{(1-p(X_i))^2}(1-D_i)1(Y_i \le y)X_i'\right]E\left[\lambda(X'\gamma_0)H(X'\gamma_0)XX'\right]^{-1}(D_j - p(X_j))H(X_j'\gamma_0)X_j + o_p(1)$$

$$= \frac{1}{\sqrt{n}}\sum_{j=1}^n E\left[\frac{\lambda(X'\gamma_0)}{1-p(X)}F_0(y|X)X'\right]E\left[\lambda(X'\gamma_0)H(X'\gamma_0)XX'\right]^{-1}X_j\frac{\lambda(X_j'\gamma_0)}{\Lambda(X_j'\gamma_0)}\frac{D_j - p(X_j)}{1-p(X_j)} + o_p(1),$$

11

where $o_p(1)$ is uniform in $y \in \mathcal{Y}_0$. In summary,

$$
\begin{aligned}
\psi_0(W_j, y) &= \frac{1 - D_j}{1 - p(X_j)} 1(Y_j \leq y) - F_0(y) \\
&\quad + E\left[\frac{\lambda(X'\gamma_0)}{1 - p(X)} F_0(y|X) X'\right] E\left[\frac{\lambda(X'\gamma_0)^2}{p(X)(1 - p(X))} X X'\right]^{-1} X_j \frac{\lambda(X'_j\gamma_0)}{p(X_j)} \frac{D_j - p(X_j)}{1 - p(X_j)}.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
\psi_1(W_j, y) &= \frac{D_j}{p(X_j)} 1(Y_j \leq y) - F_1(y) \\
&\quad - E\left[\frac{\lambda(X'\gamma_0)}{p(X)} F_1(y|X) X'\right] E\left[\frac{\lambda(X'\gamma_0)^2}{p(X)(1 - p(X))} X X'\right]^{-1} X_j \frac{\lambda(X'_j\gamma_0)}{p(X_j)} \frac{D_j - p(X_j)}{1 - p(X_j)}.
\end{aligned}
$$

∎

Lemma 2 $\sqrt{n}\left(\widehat{F}^t_d(y) - F^t_d(y)\right) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \psi^t_d(W_j, y) + o_p(1)$, where $o_p(1)$ is uniform in $y \in \mathcal{Y}^t_d$.

Proof. Recall that

$$
\begin{aligned}
\widehat{F}^t_0(y) &= \frac{1}{n} \sum_{i=1}^{n} \frac{\widehat{p}(X_i)(1 - D_i)}{1 - \widehat{p}(X_i)} 1(Y_i \leq y) \bigg/ \widehat{p}_0, \\
\widehat{F}^t_1(y) &= \frac{1}{n} \sum_{i=1}^{n} D_i 1(Y_i \leq y) \bigg/ \widehat{p}_1,
\end{aligned}
$$

where $\widehat{p}_0 = \widehat{p}_1 = n^{-1} \sum_{i=1}^{n} D_i$.

$$
\begin{aligned}
\sqrt{n}\left(\widehat{F}^t_0(y) - F^t_0(y)\right) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{\frac{\widehat{p}(X_i)(1 - D_i)}{1 - \widehat{p}(X_i)} 1(Y_i \leq y)}{\widehat{p}_0} - \frac{F^t_0(y) E[D]}{E[D]}\right) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{E[D] \frac{\widehat{p}(X_i)(1 - D_i)}{1 - \widehat{p}(X_i)} 1(Y_i \leq y) - \widehat{p}_0 F^t_0(y) E[D]}{\widehat{p}_0 E[D]} \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{E[D]\left(\frac{\widehat{p}(X_i)(1 - D_i)}{1 - \widehat{p}(X_i)} 1(Y_i \leq y) - E[D] F^t_0(y)\right) - (\widehat{p}_0 - E[D]) F^t_0(y) E[D]}{\widehat{p}_0 E[D]} \\
&\approx \frac{1}{E[D]} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{\widehat{p}(X_i)(1 - D_i)}{1 - \widehat{p}(X_i)} 1(Y_i \leq y) - E[D] F^t_0(y)\right) - \frac{F^t_0(y)}{E[D]} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (D_i - E[D]),
\end{aligned}
$$

where the last equality is from Slutsky's theorem. We derive the asymptotic expansion for the first term by the Z-map technique of CFM. $n^{-1} \sum_{i=1}^{n} \frac{\widehat{p}(X_i)(1 - D_i)}{1 - \widehat{p}(X_i)} 1(Y_i \leq y)$ as an estimator of $E[D] F^t_0(y)$ is defined by the moment conditions

$$
E\left[\begin{array}{c} (\Lambda(X'\gamma) - D) H(X'\gamma) X \\ \frac{\Lambda(X'\gamma)(1 - D)}{1 - \Lambda(X'\gamma)} 1(Y \leq y) - E[D] F^t_0(y) \end{array}\right] = 0.
$$

The derivative of this Z-map with respect to $(\gamma, E[D] F^t_0(y))$ evaluated at the true value is

$$
E\left[\left(\begin{array}{cc} \lambda(X'\gamma_0) H(X'\gamma_0) X X' & 0 \\ \frac{\lambda(X'\gamma_0)(1 - D)}{(1 - \Lambda(X'\gamma_0))^2} X' 1(Y \leq y) & -1 \end{array}\right)\right] = E\left[\left(\begin{array}{cc} \lambda(X'\gamma_0) H(X'\gamma_0) X X' & 0 \\ \frac{\lambda(X'\gamma_0)}{1 - p(X)} F_0(y|X) X' & -1 \end{array}\right)\right],
$$

so by Lemma E.3 of CFM,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{\widehat{p}(X_i)(1-D_i)}{1-\widehat{p}(X_i)} 1(Y_i \leq y) - E[D]F_0^t(y) \right)$$
$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} E\left[\frac{\lambda(X'\gamma_0)}{1-p(X)} F_0(y|X)X' \right] E\left[\lambda(X'\gamma_0)H(X'\gamma_0)XX' \right]^{-1} X_i \frac{\lambda(X_i'\gamma_0)}{p(X_i)} \frac{D_i - p(X_i)}{1-p(X_i)}$$
$$+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{\Lambda(X_i'\gamma_0)(1-D_i)}{1-\Lambda(X_i'\gamma_0)} 1(Y_i \leq y) - E[D]F_0^t(y) \right) + o_p(1).$$

In summary,

$$\sqrt{n} \left(\widehat{F}_0^t(y) - F_0^t(y) \right) =$$
$$= \frac{1}{E[D]} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ E\left[\frac{\lambda(X'\gamma_0)}{1-p(X)} F_0(y|X)X' \right] E\left[\lambda(X'\gamma_0)H(X'\gamma_0)XX' \right]^{-1} X_i \frac{\lambda(X_i'\gamma_0)}{p(X_i)} \frac{D_i - p(X_i)}{1-p(X_i)} \right.$$
$$\left. + \frac{\Lambda(X_i'\gamma_0)(1-D_i)}{1-\Lambda(X_i'\gamma_0)} 1(Y_i \leq y) - E[D]F_0^t(y) - F_0^t(y)(D_i - E[D]) \right\} + o_p(1)$$
$$= \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \psi_0^t(W_j, y) + o_p(1).$$

$\widehat{F}_1^t(y)$ does not involves the estimation of $p(\cdot)$, so its asymptotic expansion is the same as that in Theorem 6.2 of DH. ∎

Lemma 3 *Under Assumptions K, H, P and Fd, $\widetilde{p}(x)$, $\widehat{F}_d(y|x)$ and $\widehat{Q}_1 \left(\widehat{F}_0(y|x)|x \right)$ have asymptotic linear expansions uniformly in $y \in \mathcal{Y}_d, x \in \mathcal{X}$,*

$$\widetilde{p}(x) - p(x) = (nh^r)^{-1/2} \mathbb{G}_n(g_x) \left(1 + O_p\left(\sqrt{\ln n / nh^r} \right) \right) + O_p(h^2),$$
$$\widehat{F}_0(y|x) - F_0(y|x) = (nh^r)^{-1/2} \mathbb{G}_n(g_{yx}^0) \left(1 + O_p\left(\sqrt{\ln n / nh^r} \right) \right) + O_p(h^2),$$
$$\widehat{F}_1(y|x) - F_1(y|x) = (nh^r)^{-1/2} \mathbb{G}_n(g_{yx}^1) \left(1 + O_p\left(\sqrt{\ln n / nh^r} \right) \right) + O_p(h^2),$$
$$\widehat{Q}_1 \left(\widehat{F}_0(y|x)|x \right) - Q_1 (F_0(y|x)|x) = (nh^r)^{-1/2} \mathbb{G}_n(g_{yx}^{01}) \left(1 + O_p\left(\sqrt{\ln n / nh^r} \right) \right) + O_p(h^2),$$

where

$$\widetilde{p}(x) = \frac{(nh^r)^{-1} \sum_{j=1}^{n} D_j K_h(X_j - x)}{(nh^r)^{-1} \sum_{j=1}^{n} K_h(X_j - x)}$$

and

$$g_x(U_i, X_i) = \frac{1}{h^{r/2} f(x)} U_i K_h(X_i - x),$$
$$g_{yx}^0(U_{yi}^0, X_i) = \frac{1}{h^{r/2} (1 - p(x)) f(x)} U_{yi}^0 K_h(X_i - x),$$
$$g_{yx}^1(U_{yi}^1, X_i) = \frac{1}{h^{r/2} p(x) f(x)} U_{yi}^1 K_h(X_i - x),$$
$$g_{yx}^{01}(U_{yi}^0, U_{Q_1(F_0(y|x)|x)i}^1, X_i) = \frac{g_{yx}^0 - g_{Q_1(F_0(y|x)|x)x}^1}{f_1 \left(Q_1 (F_0(y|x)|x)| x \right)},$$

with $U_i = D_i - p(X_i)$, $U_{yi}^1 = D_i \left[1(Y_i \leq y) - F_1(y|X_i) \right]$, and $U_{yi}^0 = (1 - D_i) \left[1(Y_i \leq y) - F_0(y|X_i) \right]$.

Proof. We first check $\widetilde{p}(x)$. Uniformly in $x \in \mathcal{X}$,

$$
\begin{aligned}
\widetilde{p}(x) - p(x) &= \frac{(nh^r)^{-1} \sum_{j=1}^{n} (D_j - p(X_j)) K_h(X_j - x)}{(nh^r)^{-1} \sum_{j=1}^{n} K_h(X_j - x)} + \frac{(nh^r)^{-1} \sum_{j=1}^{n} (p(X_j) - p(x)) K_h(X_j - x)}{(nh^r)^{-1} \sum_{j=1}^{n} K_h(X_j - x)} \\
&= \frac{f(x)}{\widehat{f}(x)} \frac{1}{nh^r f(x)} \sum_{j=1}^{n} U_j K_h(X_j - x) + O_p(h^2) \\
&= \frac{1}{nh^r f(x)} \sum_{j=1}^{n} U_j K_h(X_j - x) \left(1 + O_p\left(\sqrt{\ln n / nh^r}\right)\right) + O_p(h^2),
\end{aligned}
$$

where $O_p(h^2)$ in the second equality is the bias (uniformly in x), and $O_p\left(\sqrt{\ln n / nh^r}\right)$ in the third equality is the uniform approximation rate of $f(x)$ by $\widehat{f}(x)$ (see, e.g., Lemma B.1 of Newey, 1994).

The proof for $F_d(y|x)$ is similar; take $F_1(y|x)$ as an example.

$$
\begin{aligned}
\widehat{F}_1(y|x) - F_1(y|x) &= \widehat{F}_1(y|x) - \frac{F_1(y|x)(nh^r)^{-1} \sum_{j=1}^{n} D_j K_h(X_j - x)}{(nh^r)^{-1} \sum_{j=1}^{n} D_j K_h(X_j - x)} \\
&= \frac{(nh^r)^{-1} \sum_{j=1}^{n} D_j \left[1(Y_j \leq y) - F_1(y|x)\right] K_h(X_j - x)}{(nh^r)^{-1} \sum_{j=1}^{n} D_j K_h(X_j - x)} \\
&= \frac{(nh^r)^{-1} \sum_{j=1}^{n} D_j \left[1(Y_j \leq y) - F_1(y|x)\right] K_h(X_j - x) / (nh^r)^{-1} \sum_{j=1}^{n} K_h(X_j - x)}{(nh^r)^{-1} \sum_{j=1}^{n} D_j K_h(X_j - x) / (nh^r)^{-1} \sum_{j=1}^{n} K_h(X_j - x)} \\
&= \frac{p(x)}{\widetilde{p}(x)} \frac{1}{p(x)} \left[\frac{1}{nh^r f(x)} \sum_{j=1}^{n} D_j \left[1(Y_j \leq y) - F_1(y|X_j)\right] K_h(X_j - x) \left(1 + O_p\left(\sqrt{\ln n / nh^r}\right)\right) + O_p(h^2) \right] \\
&= \frac{1}{p(x)} \frac{1}{nh^r f(x)} \sum_{j=1}^{n} U_{yi}^1 K_h(X_j - x) \left(1 + O_p\left(\sqrt{\ln n / nh^r}\right)\right) + O_p(h^2),
\end{aligned}
$$

where the second to last equality is from a similar analysis as in $\widetilde{p}(x)$ and the last equality is from the approximation rate of $p(x)$ by $\widetilde{p}(x)$. If we write U_{yi}^1 as $[D_i 1(Y_i \leq y) - p(X_i) F_1(y|X_i)] - F_1(y|X_i) U_i$, then the first term is attributed to the variation in the numerator of $\widehat{F}_1(y|x)$ and the second term is attributed to the variation in the denominator.

Finally, we analyze the approximation of $\widehat{Q}_1\left(\widehat{F}_0(y|x)|x\right)$:

$$
\widehat{Q}_1\left(\widehat{F}_0(y|x)|x\right) - Q_1\left(F_0(y|x)|x\right) = \widehat{Q}_1\left(\widehat{F}_0(y|x)|x\right) - Q_1\left(\widehat{F}_0(y|x)|x\right) + Q_1\left(\widehat{F}_0(y|x)|x\right) - Q_1\left(F_0(y|x)|x\right).
$$

By Hadamard differentiability of $Q_1(\cdot|\cdot)$ at $F_1(Q_1(\cdot|\cdot)|\cdot)$, we have

$$
\begin{aligned}
\widehat{Q}_1\left(\widehat{F}_0(y|x)|x\right) - Q_1\left(\widehat{F}_0(y|x)|x\right) &\approx -\frac{1}{f_1\left(Q_1\left(\widehat{F}_0(y|x)|x\right)\Big|x\right)} \left(\widehat{F}_1\left(Q_1\left(\widehat{F}_0(y|x)|x\right)\Big|x\right) - F_1\left(Q_1\left(\widehat{F}_0(y|x)|x\right)\Big|x\right)\right) \\
&\approx -\frac{1}{f_1\left(Q_1\left(F_0(y|x)|x\right)|x\right)} \left(\widehat{F}_1\left(Q_1\left(F_0(y|x)|x\right)|x\right) - F_1\left(Q_1\left(F_0(y|x)|x\right)|x\right)\right) \\
&= (nh^r)^{-1/2} \mathbb{G}_n \left(\frac{-g_{Q_1(F_0(y|x)|x)x}^1}{f_1\left(Q_1\left(F_0(y|x)|x\right)|x\right)}\right) \left(1 + O_p\left(\sqrt{\ln n / nh^r}\right)\right) + O_p(h^2),
\end{aligned}
$$

where the first approximation is from the definition of Hadamard differentiability, the second approximation is from the uniform consistency of $\widehat{F}_0(y|x)$, and the equality is from the approximation of $F_1(y|x)$ by $\widehat{F}_1(y|x)$.

Similarly,

$$Q_1\left(\widehat{F}_0(y|x)|x\right) - Q_1\left(F_0(y|x)|x\right) = (nh^r)^{-1/2}\,\mathbb{G}_n\left(\frac{g_{yx}^0}{f_1\left(Q_1\left(F_0(y|x)|x\right)|x\right)}\right)\left(1 + O_p\left(\sqrt{\ln n/nh^r}\right)\right) + O_p(h^2).$$

Combining these two results, we finish the proof. ∎

Lemma 4 $\widehat{v}_n^2 = \frac{2}{n^2 h^r}\sum\limits_{j=1}^{n}\sum\limits_{l\neq j}\widehat{w}_{jl}^2 \xrightarrow{p} \sigma^2$, where $\widehat{w}_{jl}$ is defined in the main text.

Proof. We need only to show that $v_n^2 = \frac{2}{n^2 h^r}\sum\limits_{j=1}^{n}\sum\limits_{l\neq j}w_{jl}^2 \xrightarrow{p} \sigma^2$, where $w_{jl}\equiv n^{-1}\sum_{i=1}^{n}a_{ijl}$ is the same as $\widehat{w}_{jl}\equiv n^{-1}\sum_{i=1}^{n}\widehat{a}_{ijl}$ but replacing all hat estimators by their true values, because the difference between $\widehat{v}_n^2$ and v_n^2 is $o_p(1)$. As usual, we need show that $E\left[v_n^2\right]\to\sigma^2$ and $Var\left(v_n^2\right)\to 0$. For simplicity, we only check the former.

$$\begin{aligned}
E\left[v_n^2/2\right] &= E\left[w_{jl}^2\right]/h^r = \frac{1}{n^2 h^r}\sum_{i=1}^{n}\sum_{k=1}^{n}E\left[a_{ijl}a_{kjl}\right]\\
&\approx \frac{1}{n^2 h^r}\sum_{i=1}^{n}\sum_{k\neq i}E\left[a_{ijl}a_{kjl}\right],
\end{aligned}$$

where $l\neq j\neq k\neq i$, and the last approximation is because the rest terms contribute only $o(1)$ to $E\left[v_n^2/2\right]$. Now,

$$E\left[a_{ijl}a_{kjl}\right]/h^r = E\left[E\left[a_{ijl}|W_j,W_l\right]E\left[a_{kjl}|W_j,W_l\right]\right]/h^r = E\left[\left(P_n(W_j,W_l)/2\right)^2\right]/h^r \to \sigma_a^2,$$

where the first equality is due to the independence between W_i and W_k given W_j and W_l, $P_n(W_j,W_l)$ is defined in (9), the convergence is from the proof of Theorem 5(i), and $\sigma_a^2 = \sigma^2/2$. ∎

Lemma 5 $\widehat{b}_n - B_h = \frac{1}{nh^{r/2}}\sum\limits_{j=1}^{n}\widehat{w}_{jj} - B_h \xrightarrow{p} 0$, where $\widehat{w}_{jj}$ is defined in the main text.

Proof. As in the last lemma, we can replace $\widehat{w}_{jj}$ by $w_{jj} = n^{-1}\sum_{i=1}^{n}a_{ijj}$ without affecting the probability limit of $\widehat{b}_n$; denote $b_n = \left(nh^{r/2}\right)^{-1}\sum_{j=1}^{n}w_{jj} = \left(n^2 h^{r/2}\right)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ijj}$. Since the sum of the terms with $i=j$ is $o_p(1)$, $b_n = b_n' + o_p(1)$ with $b_n' = \left(n^2 h^{r/2}\right)^{-1}\sum_{i=1}^{n}\sum_{j\neq i}a_{ijj}$. We only show $E\left[b_n'\right] - B_h \longrightarrow 0$ since it can be shown that $Var\left(b_n'\right)$ converges to 0. Note that

$$E\left[b_n\right] \approx E\left[a_{ijj}\right]/h^{r/2} = h^{-r/2}\int E\left[g_{yx}^{01}(U_{yj}^0, U_{Q_1(F_0(y|x)|x)j}^1, X_j))^2\right]d\mu(y,x) \approx B_h$$

from the proof of Theorem 5(i). ∎

Supplementary Material S.2

Because our parametric test statistics are constructed under the parametric specification of propensity score and conditional cdfs, the power of our tests may come from the model misspecification and the size may not match the nominal level. We suggest here to combine our parametric tests with the goodness of fit tests to alleviate the effects of model misspecification.

Our goodness of fit tests for $p(\cdot)$ and $F_d(\cdot|\cdot)$ are based on Andrews (1997) and Rothe and Wied (2013). The former takes the Kolmogorov-Smirnov (KS) form and the latter takes the Cramer-von Mises (CM) form. For simplicity, we describe the KS-type statistic only:

$$
\begin{aligned}
KS_n &= \sqrt{n} \sup_{X_i \in \mathcal{X}} \left| \widehat{H}_n(X_i) - \widehat{F}_n(X_i) \right| + \sqrt{n} \sup_{(Y_i, X_i) \in \mathcal{Y}_1 \mathcal{X}: D_i=1} \left| \widehat{H}_n^1(Y_i, X_i) - \widehat{F}_n^1(Y_i, X_i) \right| \\
&\quad + \sqrt{n} \sup_{(Y_i, X_i) \in \mathcal{Y}_0 \mathcal{X}: D_i=0} \left| \widehat{H}_n^0(Y_i, X_i) - \widehat{F}_n^0(Y_i, X_i) \right|,
\end{aligned}
$$

where

$$
\begin{aligned}
&\widehat{H}_n(x) = n^{-1} \sum_{i=1}^n D_i 1(X_i \le x), \widehat{F}_n(x) = n^{-1} \sum_{i=1}^n \Lambda(X_i' \widehat{\gamma}) 1(X_i \le x), \\
&\widehat{H}_n^d(y, x) = n^{-1} \sum_{i=1}^n 1(D_i = d) 1(Y_i \le y) 1(X_i \le x), \\
&\widehat{F}_n^d(y, x) = n^{-1} \sum_{i=1}^n 1(D_i = d) \Lambda(X_i' \widehat{\beta}_d(y)) 1(X_i \le x).
\end{aligned} \tag{11}
$$

$\widehat{H}_n(x)$ and $\widehat{F}_n(x)$ are estimating $H(x) = P(D = 1, X \le x)$ without and with the parametric restriction; similarly, $\widehat{H}_n^d(y, x)$ and $\widehat{F}_n^d(y, x)$ are estimating $H^d(y, x) = P(D = d, Y \le y, X \le x)$ without and with the parametric restriction. If there is no misspecification, $H(x) = F(x)$ and $H^d(y, x) = F^d(y, x)$, where $F(x) = E[1(X \le x) \Lambda(X' \gamma_0)]$ and $F^d(y, x) = E[1(D = d) 1(X \le x) \Lambda(X' \beta_d(y))]$. To see why, note that without misspecification,

$$
\begin{aligned}
H^d(y, x) &= P(D = d, Y \le y, X \le x) \\
&= E[P(D = d, Y_d \le y | X) 1(X \le x)] \\
&= E[p(X) \Lambda(X' \beta_d(y)) 1(X \le x)] \\
&= E[1(D = d) 1(X \le x) \Lambda(X' \beta_d(y))] = F^d(y, x),
\end{aligned}
$$

where the third equality uses unconfoundedness and $P(Y_d \le y | X) = \Lambda(X' \beta_d(y))$. Similarly, we can show $H(x) = F(x)$ if without misspecification. In summary, this test is to check whether the joint distributions of (D, X) and (Y_d, X) are correctly specified.

The bootstrap can be used to obtain the critical values of KS_n. Specifically, the following Algorithm G can be used.

Algorithm G:

Step 1: Draw a bootstrap sample $\{X_i^*, 1 \le i \le n\}$ with replacement from the realized values $\{X_i, 1 \le i \le n\}$.

Step 2: For every $1 \le i \le n$, put D_i^* as a simulation from the Bernoulli distribution with success probability $\Lambda(X_i^{*\prime} \widehat{\gamma})$. For every i with $D_{i*} = d$, put

$$
Y_i^* = \begin{cases} \widehat{F}_d^{-1}(U_i^* | X_i^*), & \text{if } \widehat{F}_d^{-1}(U_i^* | X_i^*) \in \mathcal{Y}_d, \\ Y_{i*}, & \text{otherwise,} \end{cases}
$$

where (D_{i*}, Y_{i*}) is the (D_i, Y_i) corresponding to X_i^* in the original sample, $\{U_i^*, 1 \le i \le n\}$ is a simulated iid sequence of standard uniformly distributed random variables, and $\widehat{F}_d(\cdot|\cdot)$ is defined in Section 4.

Step 3: Use the bootstrap data $\{(Y_i^*, D_i^*, X_i^*), 1 \le i \le n\}$ to compute estimates $\widehat{H}_n^*$, $\widehat{F}_n^*$, $\widehat{H}_n^{d*}$ and $\widehat{F}_n^{d*}$

exactly as in (11), and compute the corresponding bootstrap realization of the test statistic:

$$
\begin{aligned}
KS_n^* &= \sqrt{n} \sup_{X_i \in \mathcal{X}} \left| \widehat{H}_n^*(X_i) - \widehat{F}_n^*(X_i) \right| + \sqrt{n} \sup_{(Y_i, X_i) \in \mathcal{Y}_1 \mathcal{X} : D_i = 1} \left| \widehat{H}_n^{1*}(Y_i, X_i) - \widehat{F}_n^{1*}(Y_i, X_i) \right| \\
&\quad + \sqrt{n} \sup_{(Y_i, X_i) \in \mathcal{Y}_0 \mathcal{X} : D_i = 0} \left| \widehat{H}_n^{0*}(Y_i, X_i) - \widehat{F}_n^{0*}(Y_i, X_i) \right|.
\end{aligned}
$$

Step 4: Repeat Step 1-3 B times to get $\left\{ KS_{nb}^* \right\}_{b=1}^B$ which approximate the bootstrap distribution of the test statistics, and use the $(1-\alpha)$th empirical quantile of $\left\{ T_{nb}^* \right\}_{b=1}^B$ to approximate the asymptotic critical value.

Since the bootstrap distribution in Step 2 mimics the distribution of (D, Y) under the null, the bootstrap procedure above is valid even though the data might be generated from an alternative distribution. If the original data set is too large, we can use subsampling to approximate the distribution of KS_n.

Eventually, in the parametric RP tests, we construct two test statistics, KS_n and T_n (or KS_n and T_n^t). The size of the joint test is less than the sum of the sizes of the two tests. By choosing the two sizes appropriately, we emphasize tests of misspecification or rank preservation. For example, when the total size $\alpha = 0.05$, the misspecification size $\alpha_{KS} = 0.01$ and the rank preservation size $\alpha_T = 0.04$ imply that we focus on testing rank preservation.

Finally, as mentioned in Section 4, the parametric specification allows functions of X_i as covariates, so we need a benchmark specification of the covariates for the misspecification test. Of course, we hope the model is correctly specified so we can concentrate on the rank preservation test. Suppose the candidate covariates are a subset of series sequence; then we need only select the number of series terms, K. As in Firpo (2007), we suggest to use Hall (1987)'s likelihood cross-validation to choose K in estimating $p(\cdot)$ and use least squares cross validation of the conditional mean estimation to choose K in estimating $F_d(\cdot|\cdot)$. Specifically, in estimating $p(\cdot)$, the benchmark K is chosen as the minimizer of

$$
CV(K) = \frac{1}{n} \sum_{i=1}^n \left[D_i \log \widehat{p}_{-i}^K(X_i) + (1 - D_i) \log(1 - \widehat{p}_{-i}^K(X_i)) \right],
$$

where $\widehat{p}_{-i}^K(X_i)$ is the SLE of $p(X_i)$ based on K series terms and with the ith data point deleted;[34] in estimating $F_d(y|x)$, the benchmark K is chosen as the minimizer of

$$
CV_d(K) = \sum_{i : D_i = d} (Y_i - \widehat{m}_{d,-i}^K(X_i))^2,
$$

where $\widehat{m}_{d,-i}^K(X_i)$ is the linear series estimator of $E[Y|X, D = d]$ based on K series terms and with the ith data point among $D = d$ deleted.

If $\dim(X)$ is too large or the distribution of X is too complicated (e.g., many discrete covariates in X such that each cell contains few observations), we can test whether the marginal distribution of (D, Y_0, Y_1) predicted by the specification matches that calculated directly from the data. In this case, the test statistic is modified as

$$
KS_n = \sqrt{n} \left| \widehat{H}_n - \widehat{F}_n \right| + \sqrt{n} \sup_{Y_i \in \mathcal{Y}_1 : D_i = 1} \left| \widehat{H}_n^1(Y_i) - \widehat{F}_n^1(Y_i) \right| + \sqrt{n} \sup_{Y_i \in \mathcal{Y}_0 : D_i = 0} \left| \widehat{H}_n^0(Y_i) - \widehat{F}_n^0(Y_i) \right|,
$$

---

[34] The K based on $CV(K)$ can also be used in the semiparametric estimation of $p(\cdot)$ in Section 5.

where

$$\widehat{H}_n(x) = n^{-1} \sum_{i=1}^n D_i, \widehat{F}_n = n^{-1} \sum_{i=1}^n \Lambda(X_i'\widehat{\gamma}),$$
$$\widehat{H}_n^d(y) = n^{-1} \sum_{i=1}^n 1\,(D_i = d)\,1(Y_i \leq y),$$
$$\widehat{F}_n^d(y, x) = n^{-1} \sum_{i=1}^n 1\,(D_i = d)\,\Lambda(X_i'\widehat{\beta}_d(y)).$$

The bootstrap procedure is adjusted correspondingly.

References

Andrews, D.W.K., 1997, A Conditional Kolmogorov Test, *Econometrica*, 65, 1097-1128.

Hall, P., 1987, On Kullback–Leibler Loss and Density Estimation, *Annals of Statistics*, 15, 1491–1519.

Rothe, C. and D. Wied, 2013, Misspecification Testing in a Class of Conditional Distributional Models, *Journal of the American Statistical Association*, 108, 314-324.