Permutation Tests for Heterogeneous Treatment Effect

EunYi Chung[†]
Department of Economics
UIUC
eunyi@illinois.edu

Mauricio Olivares-González
Department of Economics
UIUC
lvrsgnz2@illinois.edu

May 20, 2018

Abstract

Detecting treatment effect heterogeneity among individuals plays a key role in any successful evaluation of a social program using randomized experiments. In this paper, we propose a permutation test for testing the null hypothesis that the *distributions* of the treated and control groups are a *constant shift apart*. When the constant shift is known, the permutation test is exact in finite samples. However, when it is unknown and thus is a nuisance parameter, the permutation test based on the plug-in test statistic may fail to control a Type 1 error. We overcome this so-called Durbin problem by implementing the martingale transformation, as proposed by Khmaladze (1981). As a result, the transformed test statistic becomes asymptotically pivotal and thus the permutation test based on this transformed statistic will be asymptotically valid while still providing an exact error control in finite samples when the average treatment effect is known. Moreover, our method can be extended to testing the joint null hypothesis of constant treatment effects within individual subgroups while allowing the treatment effects to vary across subgroups. We contrast our procedure against other methods in this context using a Monte Carlo simulation study.

Keywords: Heterogeneous Treatment Effect, Permutation Test, Empirical Process, Martingale Transformation.

1 Introduction

Detecting treatment effect heterogeneity among individuals plays a key role in any successful evaluation of a social program using randomized experiments. For example, a student may benefit or suffer greatly from a policy intervention while another student may experience little to no effect. Understanding heterogeneity in treatment effects might help researchers or policy makers design or extend social programs better since the full treatment effect can be investigated in a thorough and comprehensive way.

However, the original experiment is oftentimes not designed to assess the heterogeneity in the treatment effect. This may occur either because of the complexity of the design, or simply because it is not clear to the researchers what the channels are through which the treatment affects the outcome.

[†]All errors are our own.

As a result, many applied researchers compare the *average* treatment effects conditional on covariates, which has led to the development of nonparametric tests for the null hypothesis that the *average* treatment effects, conditional on covariates, are zero (or identical) across all subgroups (e.g., Hardle and Marron (1990), Neumeyer et al. (2003), Crump et al. (2008), Imai et al. (2013)). Although these approaches will detect some forms of treatment effect variation, their scope is limited in the sense that they only look at one aspect of the distribution, namely the mean. Only accounting for constant *average* treatment effects across subgroups while ignoring within-group heterogeneity can be misleading in understanding treatment effect heterogeneity. (See Bitler et al. (2017) for details discussion.)

In this paper, we propose a permutation test for testing the null hypothesis that the *distributions* of the treated and control groups are a *constant shift apart*. In other words, under the null, the treatment effect is constant. Moreover, the proposed method can be extended to testing the joint null hypothesis that treatment effects are constant within individual subgroups, while allowing for varying *average* treatment effects across subgroups. This test will be able to detect treatment effect heterogeneity within individual subgroups even if the average treatment effects are identical across subgroups.

Permutation tests are known to have attractive properties under the randomization hypothesis (Lehmann and Romano (2006)). As long as the permuted sample has the same joint distribution as the original sample under the null, permutation tests control a Type 1 error in finite samples: the rejection probability under the null is *exactly* the nominal level α. That is, one does not need to rely on asymptotics when making inferences. Moreover, they are nonparametric in the sense that they can be applied without any parametric assumptions of the underlying distribution that generates the data. Also, the general construction of a permutation test does not depend on the specific form of the test statistic, although some test statistics will be more suitable for a specific null hypothesis for better power performance. These features make them desirable for experimental studies, where the treatment is randomly assigned to units in possibly complex designs.

In the case of a known constant shift (or the average treatment effect), the randomization hypothesis holds, and thus, permutation tests offer a powerful approach to testing the null hypothesis of a constant treatment effect. However, when the constant shift is unknown and thus needs to be estimated, it becomes a nuisance parameter. The presence of a nuisance parameter under the null renders a major drawback: naively plugging an estimate into the test statistic makes the test statistic non-pivotal and the permutation test based on the plug-in test statistic may fail to control the Type 1 error even asymptotically. In other words, the asymptotic distribution depends on the unknown underlying distributions, offsetting their usefulness in empirical work when testing for heterogeneous treatment effects.

To overcome this so-called Durbin problem (Durbin (1973)), this paper proposes a novel permutation test by extending the martingale transformation of the empirical process introduced by Khmaladze (1981) to the two-sample case. This procedure debugs the empirical process of the nuisance parameters by decomposing the empirical process into two parts - a martingale that has a standard Brownian motion behavior, and a second part that vanishes as the sample size grows large. This strategy leaves us with a distribution-free Kolmogorov-Smirnov type test. Therefore, the permutation distribution based on the transformed test statistic inherits a pivotal limiting law, which restores the validity of the permutation test for the null hypothesis of a constant treatment effect.

Papers that are closely related to our method,in regards to testing for treatment effect heterogeneity by comparing distributions in the presence of nuisance parameters, include Koenker and Xiao (2002), Abadie (2002), Chernozhukov and Fernández-Val (2005), Linton et al. (2005), and Ding et al. (2015). Abadie (2002) compares the corresponding CDFs in the context of instrumental variables, where he obtains the critical values via bootstrap. Koenker and Xiao

(2002), Linton et al. (2005) and Chernozhukov and Fernández-Val (2005), on the other hand, exploit the relationship between CDFs and quantiles, and test for treatment effect variation using the quantile process rather than the empirical process. Another point of divergence with the theory presented here is that Linton et al. (2005) and Chernozhukov and Fernández-Val (2005) propose resampling methods to overcome the Durbin problem. Koenker and Xiao (2002), on the other hand, use the Khmaladze decomposition to restore the asymptotically distribution free nature of the test, an approach we are implementing here. Nevertheless, we are working with a permutation test which is based on a test statistic that takes as input the empirical process.

Perhaps the most related paper to ours is Ding et al. (2015), who used a Fisher randomization test based on the comparisons of CDFs using a Kolmogorov-Smirnov statistic. But, there is one key distinction that fundamentally differentiates both tests. Our test relies on a martingale decomposition of the empirical process that renders an *asymptotically* pivotal test. Ding et al. (2015), on the other hand, yield valid inference by constructing a confidence interval for the constant shift, repeating the test procedure pointwise over that interval, and taking the maximum p-value. Hence, their method does not rely on asymptotic methods. However, our Monte Carlo simulation results show that our methods outperform theirs in terms of size control and power performance in finite samples.

The rest of the paper is organized as follows. Section 2 describes the testable hypothesis in the context of the potential outcomes model. Section 3 briefly reviews the basics of permutation tests. Section 4 studies asymptotic behavior of permutation distributions based on the two-sample empirical process when δ is known as well as when δ is unknown. Section 5 contains the main theoretical results. Monte Carlo experiments are detailed in Section 6. A short discussion on testing the null hypothesis of constant treatment effects within subgroups while allowing the treatment effects to vary across subgroups can be found in Section 7. Section 8 concludes. The proofs of the main results and the coupling construction are collected in Appendix A and B.

2 Testable Hypothesis

2.1 Potential Outcomes

Consider the simplest model for a randomized experiment with subject i's (continuous) response Y_i to a binary treatment D_i. Assume we have a sample of size N and we randomly assign treatment to $m < N$ of them, while the remaining $n = N - m$ subjects are not exposed to such treatment. We will denote the m individuals in the first group as *treatment group* while the second group of size n will be the *control group*.

For every subject i, there are two mutually exclusive potential outcomes - either subject gets treated or not. If subject i were to receive the treatment ($D_i = 1$), the potential outcome that could be observed is denoted by $Y_i(1)$. Similarly, the potential outcome $Y_i(0)$ is defined if the subject i were not to be exposed to the treatment. Given D_i, one of them is observed and the other is the counterfactual outcome we would have observed under the other treatment level (1-D_i). To put it in a more compact way, we say individual i's observed outcome, Y_i^* is:

$$Y_i^* = Y_i(0) + (Y_i(1) - Y_i(0))D_i \ .$$

The treatment effect is defined by the difference between potential outcomes, i.e., individual i's treatment effect is $\delta_i = Y_i(1) - Y_i(0)$, for all $i = 1, \ldots, N$. The treatment effect is constant if $\delta_i = \delta$ for all i, otherwise we say the treatment effect is heterogeneous in the sense that it

varies across subjects. As a result, the hypothesis of constant effect is

$$H_0 : Y_i(1) - Y_i(0) = \delta \ \ \forall \ i \quad \text{for some} \quad \delta \ . \tag{1}$$

This hypothesis, however, is not directly testable because we happen to observe at most one potential outcome for each unit. An alternative testable hypothesis is available if we consider the marginal distributions of the observed outcomes for each group. Specifically, let $F_0(y)$ and $F_1(y)$ be the cumulative distribution functions (CDFs) of the control and treatment group, respectively. Then, we can cast the constant treatment effect hypothesis as

$$H_0 : F_1(y + \delta) = F_0(y) \quad \text{for some} \quad \delta \ . \tag{2}$$

In other words, $F_1(\cdot)$ and $F_0(\cdot)$ are a constant shift apart.

2.2 Test Statistic

Given the aforementioned hypothesis (2), a natural candidate for a test statistic of the measure of discrepancy is the Kolmogorov-Smirnov or Cramér-von Mises type test; we adopt the former. Assume $Y_1(0), \ldots, Y_n(0)$ are i.i.d. according to a probability distribution F_0, the control group, and independently $Y_1(1), \ldots, Y_m(1)$ are i.i.d. F_1, treatment group. Let $N = n + m$ and write

$$Z = (Z_1, \ldots, Z_N) = (Y_1(1), \ldots, Y_m(1), Y_1(0), \ldots, Y_n(0)) \ .$$

The empirical CDFs are denoted $F_0(y)$ and $\hat{F}_1(y)$, respectively. Thus, $\hat{F}_1(y) = m^{-1} \sum_{i=1}^{m} \mathbf{1}_{\{Z_i \leq y\}}$, and similarly $\hat{F}_0(y) = n^{-1} \sum_{i=m+1}^{N} \mathbf{1}_{\{Z_i \leq y\}}$. Consider the *(classical) two-sample empirical process*

$$V_{m,n}(y, \delta; Z) = \sqrt{\frac{mn}{N}} \left(\hat{F}_1(y + \delta) - \hat{F}_0(y) \right) \ . \tag{3}$$

Except when it's crucial to stress the dependency on data Z, we will drop Z from (3) to ease notation. In practice, we rarely know δ nonetheless. Instead, we estimate it by simply computing the difference in sample means $\hat{\delta} = \mu(\hat{F}_1) - \mu(\hat{F}_0)$, where $\mu(\hat{F}_1)$ and $\mu(\hat{F}_0)$ are plug-in estimators of $\mu(F_1)$ and $\mu(F_0)$ respectively. This gives rise to the *two-sample empirical process*

$$V_{m,n}(y, \hat{\delta}) = \sqrt{\frac{mn}{N}} \left(\hat{F}_1(y + \hat{\delta}) - \hat{F}_0(y) \right) \ .$$

From the two-sample empirical process, we can define the *"classical" Kolmogorov-Smirnov test statistic*

$$K_{m,n,\delta}(Z) = \sup_y \parallel V_{m,n}(y, \delta) \parallel \tag{4}$$

and the *"shifted" Kolmogorov-Smirnov test statistic*

$$K_{m,n,\hat{\delta}}(Z) = \sup_y \parallel V_{m,n}(y, \hat{\delta}) \parallel \ . \tag{5}$$

3 Permutation Test

Let $\mathbf{G}_N$ be the set of all permutations π of $\{1, \ldots, N\}$, and $\Omega_0 = \{(P, Q) : P = Q\}$, where P and Q are probability distributions defined on a sample space $\mathcal{X}$. Then if $(F_1, F_0) \in \Omega_0$, then the joint distribution of $(Z_1, \ldots, Z_N)$ is the same as $(Z_{\pi(1)}, \ldots, Z_{\pi(N)})$ for any permutation

$\pi(1), \ldots, \pi(N)$. [1]. Thus, if $F_1 = F_0$ holds, then an exact level α test can be constructed by a permutation test. To see how, consider any test statistic $T_{m,n}$. Given the test statistics $T_{m,n}$, recompute $T_{m,n}$ for all permutations π, i.e. calculate $T_{m,n}(z_{\pi(1)}, \ldots, z_{\pi(N)})$ for all $\pi \in \mathbf{G}_N$. Order these values

$$T_{m,n}^{(1)} \leq T_{m,n}^{(2)} \leq \cdots \leq T_{m,n}^{(N!)}$$

and fix a nominal level $\alpha \in (0,1)$. Define $k = N! - \lfloor N!\alpha \rfloor$ where $\lfloor \nu \rfloor$ is the largest integer less than or equal to ν. Let $M^+(z)$ and $M^0(z)$ be the number of values $K_{m,n,\delta}^{(j)}(z)$, $j = 1, \ldots, N!$, which are greater than $T_{m,n}^{(k)}(z)$ and equal to $T_{m,n}^{(k)}(z)$ respectively. Set

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

Define the randomization test function $\varphi(z)$ as

$$\varphi(z) = \begin{cases} 1 & T_{m,n}(z) > T_{m,n}^{(k)}(z) \\ a(z) & T_{m,n}(z) = T_{m,n}^{(k)}(z) \\ 0 & T_{m,n}(z) < T_{m,n}^{(k)}(z) \ . \end{cases}$$

Then, under $F_1 = F_0$, the resulting permutation test is exact level α (see theorem 15.2.1 in Lehmann and Romano (2006)). In other words, under $F_1 = F_0$,

$$\mathbb{E}[\varphi(Y_1(1), \ldots, Y_m(1), Y_1(0), \ldots, Y_n(0))] = \alpha \ .$$

Moreover, define the randomization distribution based on the test statistic $T_{m,n}$ as

$$\hat{R}_N(t) = \frac{1}{N!} \sum_{\pi \in \mathbf{G}_N} I\{T_{m,n}(z_{\pi(1)}, \ldots, z_{\pi(N)}) \leq t\} \tag{6}$$

Hence, the permutation test rejects the null hypothesis (2) if $T_{m,n}(z)$ is bigger than the $1-\alpha$ quantile of the randomization distribution (6).

It is important to emphasize that the construction of an *exact* level α test by a permutation test heavily hinges on the fact that the underlying distributions F_1 and F_0 are identical under the null. In other words, if the null hypothesis of intest does not imply $F_1 = F_0$, the rejection probability under the null may not be α even asymptotically. We illustrate this point in the following section.

4 Permutation Test based on the Empirical Process

4.1 Known Average Treatment Effect

In this subsection, we will assume that δ is known. If δ were known, under the null hypothesis (2), the shifted CDFs are the same and then samples are generated from the same probability law, so we can i) invoke the randomization hypothesis, ii) calculate the test statistic (4) over all possible permutations, and iii) construct an exact α level test.

Our goal in this section is to determine the *limiting* behavior of the randomization distribution (6). Following the coupling argument in Chung and Romano (2013), it suffices to show

[1] This is the so called *randomization hypothesis* (see chapter 15 in Lehmann and Romano (2006)), which establishes that the null hypothesis parameter space $\Omega_0 \subset \Omega$ remains invariant under $\pi \in \mathbf{G}_N$. Here Ω represents the class of all pairs (P, Q) of probability distributions defined on $\mathcal{X}$. Under this randomization hypothesis assumption, observations can be permuted and the resulting distribution is the same as that of the original samples.

that the permutation distribution (6) behaves like the unconditional distribution of the test statistic (4) when all N observations are iid from a mixture distribution $\bar{P} = pF_1 + (1-p)F_0$ where $p = \lim m/N$ when $m \to \infty$[2].

Let $\mathbb{G}$ be a Gaussian process whose marginal distributions are zero-mean with covariance structure

$$\mathbb{E}\,\mathbb{G}(s)\mathbb{G}(t) = F_0(s \wedge t) - F_0(s)F_0(t)$$

In other words, $\mathbb{G}$ is an F_0-Brownian Bridge process. First, we present the standard convergence result for the two-sample KS statistic in the following Proposition

Proposition 4.1. *(Donsker). Assume $Y_1(0), \ldots, Y_n(0)$ are i.i.d. according to a probability distribution F_0, and independently $Y_1(1), \ldots, Y_m(1)$ are i.i.d. F_1. Assume the CDFs F_1 and F_0, as well as their densities, f_1 and f_0 respectively, are continuously differentiable with respect to δ. Consider testing the hypothesis (2) for some δ known based on the test statistic*

$$K_{m,n,\delta}(Z) = \sup_y \parallel V_{m,n}(y,\delta) \parallel = \sqrt{\frac{mn}{N}} \sup_y \parallel \hat{F}_1(y,\delta) - \hat{F}_0(y) \parallel$$

Let $n \to \infty$, $m \to \infty$, with $N = n + m$, $p_m = m/N$, and $p_m \top \in (0,1)$ with

$$p_m - p = \mathcal{O}(N^{-1/2})$$

Then $K_{m,n,\delta}$ converges weakly under the null hypothesis to

$$J_0(y) \equiv \sup_y \parallel \mathbb{G}(y) \parallel \quad .$$

Remark 4.1. *Exploiting the fact that the underlying distributions are absolutely continuous, it can be readily shown that the limiting distribution above is pivotal. The change of variable $y \mapsto F^{-1}(t;\delta)$ renders uniform empirical processes,*

$$\upsilon_{m,n}(t,\delta) = \sqrt{\frac{mn}{N}} \left(\hat{G}_1(t,\delta) - \hat{G}_0(t) \right)$$

and

$$\upsilon_{m,n}(t,\hat{\delta}) = \sqrt{\frac{mn}{N}} \left(\hat{G}_1(t,\hat{\delta}) - \hat{G}_0(y) \right) \tag{7}$$

where $\hat{G}_1(t,\delta) = m^{-1} \sum_{i=1}^m \mathbf{1}_{\{F_1(Z_i;\delta) \le t\}}$, $\hat{G}_0(t) = n^{-1} \sum_{i=1}^n \mathbf{1}_{\{F_0(Z_i) \le t\}}$, and $\hat{G}_1(t,\hat{\delta})$ is $\hat{G}_1(t,\delta)$ with δ replaced with $\hat{\delta}$. In other words, the empirical process defined in (3) is equivalent to a process based on N i.i.d. uniform variables and its limiting distribution is a Brownian Bridge $\mathbb{B}^0$ on $[0,1]$.

Remark 4.2. *Proposition 4.1 shows that a test based on the uniform empirical process is particularly attractive because it is asymptotically distribution-free i.e. when δ is known, the limiting distribution of the Kolmogorov-Smirnov test statistic is the same regardless of the underlying distribution generating the data. Furthermore, it follows that if the null hypothesis holds, so that $\hat{F}_0$ and $\hat{F}_1$ are independent empirical distribution functions from the same continuous distribution function, then the classical KS statistic converges weakly to the same limit distribution as in the one-sample two-sided case.*

In the next Proposition, the limiting behavior of the permutation distribution is obtained.

[2]See Appendix 8 for further discussion about this construction.

Proposition 4.2. *(Bickel) Assume the premises of Proposition 4.1. Then the permutation distribution based on $K_{m,n,\delta}$ given by (6) is such that*

$$\sup_t \parallel \hat{R}_N(t) - J_0(t) \parallel \xrightarrow{\text{P}} 0,$$

where $J_0(\cdot)$ denotes the c.d.f. of $\sup \parallel \mathbb{G} \parallel$.

Remark 4.3. *As noted by Einmahl and Khmaladze (2001), Bickel (1969) was the first to motivate and provide a methodology to systematically study permutation tests based on the two-sample empirical process (3). See Raghavachari (1973) and DasGupta (2008), chapter 26, for further discussion.*

Remark 4.4. *Under H_0 given by (2) when δ is known, the permutation distribution behaves asymptotically like the supremum of a Brownian Bridge given by $\mathbb{G}$, as is the true unconditional limiting distribution of the classical KS statistic.*

4.2 Unknown Average Treatment Effect

When δ is unknown, it becomes a nuisance parameter. If one uses a plug-in test statistic given by (5), what happens to the limit distribution of the permutation test?

We showed in Proposition 4.2 that when δ is known, the asymptotic distribution of the two-sample empirical process converges weakly to the Brownian Bridge $\mathbb{G}$, which does not depend on the true underlying distribution $F_0(y)$. However, when δ is unknown and we plug in $\hat{\delta}$ for δ, the resulting limiting distribution differs from $\mathbb{G}$ and is no longer distribution-free. The asymptotic behavior of the Kolmogorov's goodness-of-fit test in the presence of nuisance parameters dates back to Durbin (1973) and this situation of jeopardizing the distribution free character is called the Durbin problem.

When δ is unknown, we will show that the shifted KS statistic (5) converges weakly to the supremum of the *Brownian Bridge with drift*, $\mathbb{B}$. More formally, let $\xi(\cdot)$ be a Gaussian process with mean 0 and covariance structure

$$\mathbb{C}(\xi(x), \xi(y)) = \sigma_0^2 f_0(x) f_0(y)$$

where $\sigma_0^2 = \sigma^2(F_0)$, and f_0 is the density of F_0. Then, the Brownian Bridge with drift is given by

$$\mathbb{B}(\cdot) = \mathbb{G}(\cdot) + \xi(\cdot) \tag{8}$$

with covariance structure

$$\mathbb{C}(\mathbb{G}(x), \xi(y)) = f_0(y) F_0(x) \left(1 - F_0(x)\right) \left\{\mathbb{E}(Y(0)|Y(0) \leq x) - \mathbb{E}(Y(0)|Y(0) > x)\right\} .$$

It is because of this dependency between $\mathbb{G}$ and ξ that the asymptotic distribution is no longer asymptotically independent of the hypothetical F_0. This is formally established in the following Proposition.

Proposition 4.3. *(Ding, Feller, Miratrix). Assume $Y_1(0), \ldots, Y_n(0)$ are i.i.d. according to a probability distribution F_0, and independently $Y_1(1), \ldots, Y_m(1)$ are i.i.d. F_1. Assume the CDFs F_1 and F_0, as well as their densities, f_1 and f_0 respectively, are continuously differentiable with respect to δ. Consider testing the hypothesis (2) for some δ based on the test statistic*

$$K_{m,n,\hat{\delta}}(Z) = \sup_y \parallel V_{m,n}(y, \hat{\delta}) \parallel = \sqrt{\frac{mn}{N}} \sup_y \parallel \hat{F}_1(y + \hat{\delta}) - \hat{F}_0(y) \parallel$$

Let $n \to \infty$, $m \to \infty$, *with* $N = n + m$, $p_m = m/N$, *and* $p_m \top \in (0, 1)$ *with*

$$p_m - p = \mathcal{O}(N^{-1/2})$$

Then $K_{m,n,\hat{\delta}}$ *converges weakly under the null to*

$$J_1(y) \equiv \sup_y \parallel \mathbb{B}(y) \parallel$$

where $\mathbb{B}(\cdot)$ *is given by (8).*

The following proposition shows that the limiting behavior of the permutation distribution based on the shifted K-S statistic given by **(5)** is different from that of the unconditional true sampling distribution.

Proposition 4.4. *Assume the premises of Proposition 4.3. Then the permutation distribution (6) based on* $K_{m,n,\hat{\delta}}$ *satisfies*

$$\sup_t \parallel \hat{R}_N(t) - J_0(t) \parallel \xrightarrow{\text{P}} 0,$$

where $J_0(\cdot)$ *denotes the c.d.f. of* $\sup \parallel \mathbb{G} \parallel$.

Remark 4.5. *Under the hypothesis (2), the true unconditional sampling distribution of* $K_{m,n,\hat{\delta}}$ *is given by* $J_1(\cdot)$ *in Proposition 4.3, which does not equal* $J_0(\cdot)$ *in general. Then, the permutation distribution and the true unconditional sampling distribution behave differently asymptotically in the presence of nuisance parameters. Hence, the permutation test for the hypothesis (2) fails to control the size.*

5 Martingale Transformation

We concluded in Section 4.2 that the consequence of the drift term implied the test statistic based on the shifted empirical process is no longer distribution-free. Khmaladze (1981) proposed an approach to this problem in the one sample case, which boils down to a Doob-Meyer decomposition of the uniform empirical process. We're going to extend Khmaladze's result to the two-sample case and work with the two sample uniform empirical process (7).

More specifically, let the real-valued function $g(s) = (s, f_0(s))'$ on $[0, 1]$ be bounded and continuous in its arguments, and $\dot{g}(s) = (1, \dot{f}_0(s))'$, where $\dot{g}$ is the derivative of g. Define $C(s) = \int_s^1 \dot{g}(t)\dot{g}(t)'dt$, and assume it is invertible for $s \in [0, 1)$. Then the Khmaladze transformation of the parametric empirical process (7) is given by

$$\tilde{\upsilon}_{m,n}(t, \hat{\delta}) = \upsilon_{m,n}(t, \hat{\delta}) - \int_0^y \left[\dot{g}(s)'C^{-1}(s) \int_s^1 \dot{g}(r)d\upsilon_{m,n}(r, \hat{\delta})\right] ds \ . \tag{9}$$

Khmaladze (1981) showed that (9) converges weakly to a Brownian motion process, effectively nullifying the effect of the estimated nuisance parameter. Define the map $\phi_g : D[0, 1] \to D[0, 1]$ such that

$$\phi_g(h)(t) = \int_0^t \left[\dot{g}(s)'C^{-1}(s) \int_s^1 \dot{g}(r)dh(r)\right] ds \ , \tag{10}$$

where this is defined as the compensator of h (see Parker (2013)). Moreover, as noted in Bai (2003), ϕ_g is a linear mapping and $\phi_g(cg) = cg$ for a constant or random variable c. This allows us to write (9) as

$$\tilde{\upsilon}_{m,n}(t, \hat{\delta}) = \upsilon_{m,n}(t, \hat{\delta}) - \phi_g(\upsilon_{m,n}(t, \hat{\delta})) = \upsilon_{m,n}(t, \delta) - \phi_g(\upsilon_{m,n}(t, \delta)) + o_P(1) \ .$$

The following proposition shows the Khmaladze transformation removes the effect of $\hat{\delta}$ on the limiting process.

Proposition 5.1. *(Khmaladze) Assume $Y_1(0), \ldots, Y_n(0)$ are i.i.d. according to a probability distribution F_0, and independently $Y_1(1), \ldots, Y_m(1)$ are i.i.d. F_1. Assume the CDFs F_1 and F_0, as well as their densities, f_1 and f_0 respectively, are continuously differentiable with respect to δ. Consider testing the hypothesis (2) for some δ known based on the test statistic*

$$\tilde{K}_{m,n,\hat{\delta}}(Z) = \sup_t \parallel \tilde{v}_{m,n}(t, \hat{\delta}) \parallel$$

where $\tilde{v}_{m,n}(t, \hat{\delta})$ is the Khmaladze transformation in (9). Let $n \to \infty$, $m \to \infty$, with $N = n+m$, $p_m = m/N$, and $p_m \top \in (0, 1)$ with

$$p_m - p = \mathcal{O}(N^{-1/2})$$

Then the limiting distribution of $\tilde{K}_{m,n,\hat{\delta}}$ is

$$J_2(y) \equiv \sup_t \parallel BM(t) \parallel$$

where $BM(\cdot) = \mathbb{B}^0(\cdot) - \phi_g(\mathbb{B}^0(\cdot))$ is the standard Brownian Motion.

5.1 Khmaladze Transformation as a Continuous-time Detrending Operation

To gain further insight as to why the transformation works, we follow Bai (2003) and Parker (2013), and we consider (9) with y taking discrete values, replacing integral with sums. For instance, suppose $0 = t_0 < t_1 < \cdots < t_m < t_{m+1} = 1$ is a partition of the interval $[0, 1]$ and that y takes on values on $t_1, t_2, ..., t_m$. Write (9) in differentiation form

$$d\tilde{v}_{m,n}(t, \hat{\delta}) = dv_{m,n}(t, \hat{\delta}) - \dot{g}(t)' C^{-1}(t) \int_t^1 \dot{g}(r) dv_{m,n}(r, \hat{\delta}) dt \tag{11}$$

let

$$\begin{aligned} y_i &= dv_{m,n}(t_i, \hat{\delta}) \\ \dot{g}(t_i)' dt_i &= x_i \\ C(t_i) &= \sum_{k=i}^{m+1} x_k x_k' \\ \int_y^1 \dot{g}(r) dv_{m,n}(r, \hat{\delta}) &= \sum_{k=i}^{m+1} x_k y_k \end{aligned}$$

then the right hand side of (11) can be interpreted as the recursive residuals:

$$y_i - x_i' \left(\sum_{k=i}^{m+1} x_k x_k' \right)^{-1} \sum_{k=i}^{m+1} x_k y_k = y_i - x_i' \hat{\boldsymbol{\beta}}_i \tag{12}$$

where $\hat{\boldsymbol{\beta}}_i$ is the OLS estimator based on the last $m - i + 2$ observations. The cumulative sum (integration from $[0, t_i)$) of above expression gives rise to a Brownian motion process.

5.2 Numerical Computation of the Khmaladze Transformation

Computationally, we will integrate numerically so we typically assume the partition $\{t_i\}_i$ is evenly spaced, with the accuracy of the method depending on the number of points m. Stack

y_i and x_i in the following manner

$$\mathbf{X}_i = \begin{pmatrix} \sqrt{\frac{1}{m}} & \sqrt{\frac{1}{m}}\dot{f}_0(t_{m+1}) \\ \sqrt{\frac{1}{m}} & \sqrt{\frac{1}{m}}\dot{f}_0(t_{m+1}) \\ \vdots & \vdots \\ \sqrt{\frac{1}{m}} & \sqrt{\frac{1}{m}}\dot{f}_0(t_i) \end{pmatrix}, \quad \mathbf{y}_i = \begin{pmatrix} \sqrt{m}\left(\upsilon_{m,n}(t_{m+1},\hat{\delta}) - \upsilon_{m,n}(t_m,\hat{\delta})\right) \\ \sqrt{m}\left(\upsilon_{m,n}(t_m,\hat{\delta}) - \upsilon_{m,n}(t_{m-1},\hat{\delta})\right) \\ \vdots \\ \sqrt{m}\left(\upsilon_{m,n}(t_i,\hat{\delta}) - \upsilon_{m,n}(t_{i-1},\hat{\delta})\right) \end{pmatrix}$$

then the OLS estimator based on the last $m-i+2$ observations described on right hand side of (12) can be written as

$$\hat{\boldsymbol{\beta}}_i = \left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\mathbf{X}_i'\mathbf{y}_i$$

which implies that the Khmaladze transformation of the empirical process in (9) can be obtained by numerically integrating from $[0, t_i)$, i.e.

$$\upsilon_{m,n}(t_i,\hat{\delta}) - \frac{1}{m}\sum_{j=1}^{i} x_j'\hat{\boldsymbol{\beta}}_j$$

and therefore the test statistic can be calculated as

$$\max_{1\le i\le 1}\left|\upsilon_{m,n}(t_i,\hat{\delta}) - \frac{1}{m}\sum_{j=1}^{i} x_j'\hat{\boldsymbol{\beta}}_j\right|$$

5.3 Main Result

The following proposition shows that the permutation test based on the Khmaladze transformed test statistic leads to an asymptotically valid test.

Proposition 5.2. *Assume $Y_1(0),\ldots,Y_n(0)$ are i.i.d. according to a probability distribution F_0, and independently $Y_1(1),\ldots,Y_m(1)$ are i.i.d. F_1. Assume the CDFs F_1 and F_0, as well as their densities, f_1 and f_0 respectively, are continuously differentiable with respect to δ. Consider testing the hypothesis (2) based on the test statistic*

$$\tilde{K}_{m,n,\hat{\delta}}(Z) = \sup_t \parallel \tilde{\upsilon}_{m,n}(t,\hat{\delta}) \parallel$$

where $\tilde{\upsilon}_{m,n}(t,\hat{\delta})$ is the Khmaladze transformation in (9). Then the permutation distribution (6) based on the Khmaladze transformed statistic $\tilde{K}_{m,n,\hat{\delta}}(Z)$

$$\sup_t \parallel \hat{R}_N(t) - J_2(t) \parallel \xrightarrow{\mathrm{P}} 0,$$

where $J_2(\cdot)$ denotes the c.d.f. of $\sup \parallel BM \parallel$, where BM is a Brownian motion on $[0,1]$.

6 Monte Carlo

6.1 Implementation

The martingale transformation described in 5 uses the true density and score functions. In the Monte Carlo experiments of section 6.1.1, both functions were estimated employing the univariate adaptive kernel density estimation (e.g. Portnoy and Koenker, 1989; Koenker and Xiao, 2002), and the results were obtained directly from the **R** package **quantreg** (Koenker (2016)). Simulation results using the true density and score functions were similar in magnitude and therefore not shown in here, though available upon request.

6.1.1 Validity of Permutation Test

We are interested in comparing the rejection probabilities of α size permutation tests based on different test statistics: classic Kolmogorov-Smirnov (δ is known), the shifted Kolmogorov-Smirnov (a naive approach where we calculate the usual KS p-value assuming that the estimated treatment is in fact the true treatment effect), and the Khmaladze martingale transformation of the empirical process based on Kolmogorov-Smirnov test. Moreover, we consider three additional methods against which we compare our approach: the Fisher Randomization Tests (FRT) in Ding et al. (2015), and the subsampling and bootstrap methods from Chernozhukov and Fernández-Val (2005).

Table 1 contains the rejection probability results of the Monte Carlo simulations. We generated samples from three different distributions: standard normal, lognormal, and student's t distribution with 5 degrees of freedom. In this experiment sample sizes vary between groups[3]. We considered the sequence of total sample size $N \in \{13, 50, 80, 200\}$, and for each sample size and distribution, a constant treatment effect $\delta = 1$ was assigned[4]. We ran these simulations with 5000 replications across Monte Carlo Experiments.

[3]In the context of test for the ATE, Caughey et al. (2016) pointed out the dominance of the permutation test compared to the t-test when sample sizes between groups differ mightily (1000 vs 30) and the distributions are skewed. In this paper we worked with less accentuated differences. Simulations with alternative choices of samples sizes are also available though not included in this text.

[4]Similar results were obtained when we allow for different treatment effects.

Table 1: Size of $\alpha = 0.05$ tests H_0 : Constant ($\delta = 1$) Treatment Effect Effect.

		Distributions		
N	Method	Normal	Lognormal	t_5
	Classic KS	0.0494	0.0482	0.0522
$N = 13$	Naive KS	0.0000	0.0298	0.0002
$n = 8$	FRTI CI	0.0000	0.0004	0.0000
$m = 5$	Subsampling	0.0004	0.0050	0.0016
	Bootstrap	0.0742	0.0314	0.0658
	Khmaladze	0.0000	0.0472	0.0118
	Classic KS	0.0528	0.0506	0.0460
$N = 50$	Naive KS	0.0002	0.3116	0.0014
$n = 30$	FRTI CI	0.0064	0.0222	0.0062
$m = 20$	Subsampling	0.0062	0.0108	0.0102
	Bootstrap	0.0330	0.0480	0.0360
	Khmaladze	0.0266	0.0354	0.0472
	Classic KS	0.0452	0.0516	0.0510
$N = 80$	Naive KS	0.0000	0.3244	0.0016
$n = 50$	FRTI CI	0.0122	0.0280	0.0148
$m = 30$	Subsampling	0.0206	0.0062	0.0066
	Bootstrap	0.0818	0.0414	0.0894
	Khmaladze	0.0236	0.0590	0.0354
	Classic KS	0.0472	0.0548	0.0486
$N = 200$	Naive KS	0.0004	0.3912	0.0032
$n = 120$	FRTI CI	0.0290	0.0334	0.0250
$m = 80$	Subsampling	0.0344	0.0062	0.0124
	Bootstrap	0.0926	0.0622	0.0864
	Khmaladze	0.0236	0.0354	0.0428

For the FRT CI we used 99.99% CI_γ for $\hat{\tau}$. We followed the suggested subsampling size is $b = 20 + n^{1/4}$.

The permutation test based on the martingale transformation *à la* Khmaladze is yielding considerably correct rejection rates in all cases regardless of the skewness of the distribution or the sample size. It is worth mentioning that our method outperforms all the others (except when δ is known) in terms of controlling the Type 1 error when sample sizes are small despite the fact that we estimate the ATE, and the density and score functions are also estimated nonparametrically.

Moreover, these experiments confirm the story of the theoretical results in section 4: the permutation test based on the (naive) shifted KS statistic fails to control the type I error, even in large samples. We argued that the permutation distribution based on the shifted KS statistic depends on the underlying law that generates the data and therefore, the permutation distribution is no longer asymptotically distribution free. Despite Ding et al. (2015) did not compute the permutation distribution using this naive KS statistic, the conclusions of their naive approach are similar to those found in here[5]. As shown in Table 1, the permutation test is either too conservative (normal and student's t) or it fails to control the size (lognormal). In the case of skewed distributions, the size of the test increases with the sample size.

[5]Their Monte Carlo experiment for the naive approach does not calculate the p-value that arises from the permutation distribution, but the p-value from the KS distribution.

Both the confidence interval FRT (FRT CI) by Ding et al. (2015) and subsampling by Chernozhukov and Fernández-Val (2005) control the rejection probabilities across different sample sizes and data generating processes, but in a rather conservative fashion nonetheless. For instance, when the total sample size is either 50 or 13, FRT CI test is hyper-conservative. We also show a similar conclusion regarding the bootstrap. More specifically, the Bootstrap is not valid across distributions for $N > 50$. This is not surprising since the bootstrap does not have the same generality as, say, subsampling.

6.1.2 Power of the test

To illustrate the power of the test, we adhere to the design shown in Koenker and Xiao (2002), which serves as the benchmark for the Monte Carlo experiments in Chernozhukov and Fernández-Val (2005) and Ding et al. (2015):

$$Y_i(0) = \varepsilon_i, \;\; \delta_i = \delta + \sigma_\delta Y_i(0)$$
$$Y_i(1) = \delta_i + Y_i(0)$$

where σ_δ denotes the different levels of heterogeneity. Effects that vary from person to person in this manner are broadly discussed in Rosenbaum (2002), although is worth mentioning the proposed test allows us to work under more general forms of heterogeneity.

We generate data according to this rule and we calculate the empirical rejection probabilities for 5% level our permutation test for the null hypothesis of constant treatment effect. For the sake of comparison, Table 2 also includes the performance of the FRT CI and Subsampling.

In this spirit, we consider the same data generating processes (ε_i follows a lognormal distribution) and several choices of heterogeneity ($\sigma_\delta \in \{0, 0.2, 0.5\}$). Since it is part of our interest to show the performance in small sample as well, we consider $N = 50$ in addition to the ones found in the papers mentioned above. These quantities are based on 5000 experiments.

Table 2: Power of $\alpha = 0.05$ tests for several levels of heterogeneity σ_δ, and $\delta = 1$

N	Results for Khmaladze			Results for FRT CI			Results for Subsampling		
$n = m$	$\sigma_\delta = 0$	$\sigma_\delta = 0.2$	$\sigma_\delta = 0.5$	$\sigma_\delta = 0$	$\sigma_\delta = 0.2$	$\sigma_\delta = 0.5$	$\sigma_\delta = 0$	$\sigma_\delta = 0.2$	$\sigma_\delta = 0.5$
Lognormal Outcomes									
50	0.0118	0.0354	0.1084	0.0194	0.0508	0.0218	0.0120	0.0318	0.0108
100	0.0120	0.0900	0.2320	0.0272	0.0550	0.1526	0.0124	0.0178	0.0590
400	0.0511	0.2910	0.8520	0.0438	0.1880	0.6616	0.0060	0.0340	0.3136
800	0.0440	0.6105	0.9901	0.0332	0.3522	0.9382	0.0064	0.0806	0.7172

For the FRT CI we used 99.99% CI_γ for $\hat{\tau}$. We followed the suggested subsampling size is $b = 20 + n^{1/4}$.

The power performance of our test illustrates that for the lognormal case, both our test and the FRT CI have greater rejection rates than subsampling, even in large samples. It is worth mentioning that FRT CI has higher rejection rates than the Khmaladze test presented here in small samples ($N = 50$), but this situation is reverse when the sample size increases, a situation where the asymptotic approximation works better.

7 Within-group Treatment Effect Heterogeneity

The permutation test proposed in the paper can be extended to testing the joint null hypothesis of constant treatment effects within individual subgroups while allowing the treatment effects to vary across subgroups. As Bitler et al. (2017) point out, only accounting for constant *average*

treatment effects across subgroups while ignoring within-group heterogeneity can be misleading to truly understand treatment effect heterogeneity.

In this section, we propose a test method for jointly testing null hypotheses that treatment effects are constant across mutually exclusive subgroups while the average treatment effects can vary across subgroups. Formally, the null hypothesis of interest is now

$$H_0^g : F_1^g(y + \delta_g) = F_0^g(y) \ , \quad \text{for all mutually exclusive subgroup } g$$

where $F_0^g(y)$ and $F_1^g(y)$ are the cumulative distribution functions (CDFs) of the control and treatment group, respectively, for subgroup g. Note that the nuisance parameter δ_g for subgroup g can vary across subgroups.

To this end, as will be explained in Algorithm 7.1, we will be testing as many multiple hypotheses simultaneously as the number of subgroups G. If one ignores the multiplicity issue and tests each hypothesis at level α, the probability of one or more false rejections may be much greater than α. Thus, we carefully conduct our test while controlling the familywise error rate (FWER) at level α using a Bonferroni method.

Following our results on the permutation test based on the Khmaladze transformation, an algorithm for testing the null hypothesis H_0^g is given by the following.

Algorithm 7.1. *(Testing Treatment Effect Heterogeneity Across Subgroups)*

1. *For each subgroup g, perform the permutation test based on the Khmaladze transformed $K-S$ statistic at level α/G, where G is the number of subgroups.*

2. *Reject the null H_0^g if any one null for a subgroup is rejected. In other words, reject the joint null hypothesis H_0^g if the observed test statistic $\tilde{v}_{m,n}$ is greater than[6] the lower $(1-\alpha/G)$ quantile of the permutation distribution for any subgroup g.*

8 Conclusions

Heterogeneity in the treatment effect in randomized experiments is of paramount importance to correctly evaluate a policy or clinical trial. While this task is oftentimes carried out by comparing the average treatment effects conditional on covariates, we propose a test procedure that allows one to compare the entire distributions of the control and treatment groups within individual subgroups. This can done by performing permutation tests that render asymptotically valid inference using the martingale decomposition of the empirical process *á la Khmaladze*. More specifically, the transformed test statistic becomes asymptotically pivotal and thus the permutation test based on this transformed statistic will be asymptotically valid in general while still providing an exact error control in finite samples when the average treatment effect is known. Furthermore, we confirm in a series of Monte Carlo experiments that the test displays not only a good size control relative to other tests proposed in the literature, but also fairly good power in certain scenarios.

[6]To be more precise, one can use randomization explained in the permutation construction described in Section 3.

References

Abadie, A. (2002). Bootstrap tests for distributional treatment effects in instrumental variable models. *Journal of the American statistical Association*, 97(457):284–292.

Abramovich, Y. A. and Aliprantis, C. D. (2002). *An invitation to operator theory*, volume 1. American Mathematical Soc.

Bai, J. (2003). Testing parametric conditional distributions of dynamic models. *Review of Economics and Statistics*, 85(3):531–549.

Bickel, P. J. (1969). A distribution free version of the smirnov two sample test in the p-variate case. *The Annals of Mathematical Statistics*, 40(1):1–23.

Bitler, M. P., Gelbach, J. B., and Hoynes, H. W. (2017). Can variation in subgroups' average treatment effects explain treatment effect heterogeneity? evidence from a social experiment. *Review of Economics and Statistics*, 99(4):683–697.

Caughey, D., Dafoe, A., and Miratrix, L. (2016). Beyond the sharp null: Permutation tests actually test heterogeneous effects. *Unpublished manuscript*.

Chernozhukov, V. and Fernández-Val, I. (2005). Subsampling inference on quantile regression processes. *Sankhyā: The Indian Journal of Statistics*, pages 253–276.

Chung, E. and Romano, J. P. (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics*, 41(2):484–507.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2008). Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics*, 90(3):389–405.

DasGupta, A. (2008). *Asymptotic theory of statistics and probability*. Springer Science & Business Media.

Ding, P., Feller, A., and Miratrix, L. (2015). Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Durbin, J. (1973). Weak convergence of the sample distribution function when parameters are estimated. *The Annals of Statistics*, pages 279–290.

Einmahl, J. and Khmaladze, E. (2001). The two-sample problem in and measure-valued martingales. *Lecture Notes-Monograph Series*, pages 434–463.

Hardle, W. and Marron, J. S. (1990). Semiparametric comparison of regression curves. *The Annals of Statistics*, pages 63–89.

Imai, K., Ratkovic, M., et al. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470.

Khmaladze, E. V. (1981). Martingale approach in the theory of goodness-of-fit tests. *Theory of Probability & Its Applications*, 26(2):240–257.

Koenker, R. (2016). *quantreg: Quantile Regression*. R package version 5.26.

Koenker, R. and Xiao, Z. (2002). Inference on the quantile regression process. *Econometrica*, 70(4):1583–1612.

Lehmann, E. L. and Romano, J. P. (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

Linton, O., Maasoumi, E., and Whang, Y.-J. (2005). Consistent testing for stochastic dominance under general sampling schemes. *The Review of Economic Studies*, 72(3):735–765.

Neumeyer, N., Dette, H., et al. (2003). Nonparametric comparison of regression curves: an empirical process approach. *The Annals of Statistics*, 31(3):880–920.

Parker, T. (2013). A comparison of alternative approaches to supremum-norm goodness-of-fit tests with estimated parameters. *Econometric Theory*, 29(05):969–1008.

Pollard, D. (2012). *Convergence of stochastic processes*. Springer Science & Business Media.

Portnoy, S. and Koenker, R. (1989). Adaptive l-estimation for linear models. *The Annals of Statistics*, pages 362–381.

Raghavachari, M. (1973). Limiting distributions of kolmogorov-smirnov type statistics under the alternative. *The Annals of Statistics*, pages 67–73.

Romano, J. P. (1989). Bootstrap and randomization tests of some nonparametric hypotheses. *The Annals of Statistics*, pages 141–159.

Rosenbaum, P. R. (2002). Observational studies. In *Observational studies*, pages 1–17. Springer.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Wellner, J. and Van der Vaart, A. W. (2013). *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media.

Appendix

Appendix A: Coupling Construction

Assume $Y_1(0), \ldots, Y_n(0)$ are i.i.d. according to a probability distribution F_0, the control group, and independently $Y_1(1), \ldots, Y_m(1)$ are i.i.d. F_1, treatment group. Let $N = n + m$ and write

$$Z = (Z_1, \ldots, Z_N) = (Y_1(1), \ldots, Y_m(1), Y_1(0), \ldots, Y_n(0)) \tag{13}$$

Moreover, suppose $\lim_{n\to\infty} n/N = p \in (0, 1)$ in such a way that

$$p - \frac{n}{N} = \mathcal{O}(N^{-1/2})$$

The main idea behind the coupling argument in Chung and Romano (2013) is that the limiting distribution of the the permutation distribution based on Z should behave approximately like the limit law of the permutation distribution based on a sample of N iid observations $\bar{Z} = (\bar{Z}_1, \ldots, \bar{Z}_N)$ from the mixture distribution $\bar{P} = pF_1 + (1 - p)F_0$.

We would wish to compare

$$\bar{Z} = (\bar{Z}_1, \cdots, \bar{Z}_N) \quad \text{vs} \quad Z = (Y_1(1), \ldots, Y_m(1), Y_1(0), \ldots, Y_n(0))$$

The basic intuition stems from the following. Since the permutation distribution considers the empirical distribution of a statistic evaluated at all possible permutations of the data, it clearly does not depend on the ordering of the observations.

Remark 8.1. *The elements of $\bar{Z}$ can be thought as the outcome of a compound lottery. First, draw a random index j from $\{0, 1\}$ with probability $\mathbb{P}(j = 0) = p$. Then, conditionally on the outcome being j, sample $\bar{Z}_i$ from F_0 if $j = 0$, and from F_1 otherwise.*

Except for the fact that the ordering in Z is such that the first n observations are coming from F_0, and the last m are coming from F_1, the original sampling scheme is still only approximately like that of sampling from $\bar{P}$.

Remark 8.2. *Recall the binomial distribution is used to model the number of successes m when sampling with replacement from a population of size N. Hence, the number of observations $\bar{Z}_i$ out of N which are from population F_0 follows the Binomial distribution with parameters N and p. This number has mean $Np \approx n$, whereas the exact number of observations from F_0 in Z is n.*

Let $\pi = (\pi(1), \ldots, \pi(N))$ be a random permutation of $\{1, \ldots, N\}$. Then, if we consider a random permutation of Z and $\bar{Z}$, the number of observations in the first n entries of Z which were $Y(0)$s has the hypergeometric distribution, while the number of observations in the first n entries of $\bar{Z}$ which were $Y(0)$s still has the binomial distribution.

8.0.1 The algorithm

First draw an index j from $\{0, 1\}$ with probability $\mathbb{P}(j = 0) = p$. Then, conditionally on the outcome being j, set $\bar{Z}_1 = Y_1(j)$. Next, draw another index i from $\{0, 1\}$ at random with probability $\mathbb{P}(i = 0) = p$. If $i = j$, set $\bar{Z}_2 = Y_2(j)$, otherwise $\bar{Z}_2 = Y_1(i)$. Keep repeating this process, noting that there will probably be a point in which you exhaust all the n observations from the control group governed by F_0. If this happens and another index $j = 1$ is drawn again, then just sample a new observation $Y_{n+1}(0)$ from F_0, and analogously if the observations

you've exhausted are from population F_1. Continue this way so that as many as possible of the original Z_i observations are used in the construction of $\bar{Z}$. After this, you will end up with Z and $\bar{Z}$, with many of their coordinates in common (and this is why this method is called "coupling," because we couple $\bar{Z}$ with Z). The number of observations which differs, say D, is the (random) number of added observations required to fill up $\bar{Z}$. You can acces this R file to see how this algorithm works.

8.0.2 Reordering according to π_0

Furthermore, we can reorder the observations in $\bar{Z}$ by a permutation π_0 so that Z_i and $Z_{\pi_0(i)}$ agree for all i except for some hopefully small (random) number D. Recall that Z has the observations in order, that is, the first n observations arose from F_0, while the last m observations are distributed according to F_1. Thus, to couple $\bar{Z}$ with Z, put all observation in $\bar{Z}$ that came from F_0 in the first up to n. If the number of observations from F_0 is *greater or equal* to n (recall that this is a possibility), then $\bar{Z}_{\pi(i)}$ for $i = 1, \ldots, n$ are filled according to the observations in $\bar{Z}$ which came from F_0, and if the number is greater, put them aside for now. On the other hand, if the number of observations in $\bar{Z}$ which came from F_0 is *less* than n, fill up as many of $\bar{Z}$ from F_0 as possible, and leave the rest of the blank spots for now.

Next, move onto the observations in $\bar{Z}$ that came from F_1 and repeat the above procedure for $n+1, n+2, \ldots, n+m$ spots in order to complete the observations in $\bar{Z}_{\pi(i)}$; simply fill up the empty spots with the remaining observations which were put aside (at this point the order does not matter, but chronological order is an option). This permutation of the observations in $\bar{Z}$ corresponds to a permutation π_0 and satisfies $Z_i = \bar{Z}_{\pi_0(i)}$ for indices i except for D of them.

8.0.3 Why does coupling work?

The number of observations D where Z and $\bar{Z}_{\pi_0}$ differs is random and it can be shown that

$$\mathbb{E}(D/N) \leq N^{-1/2}$$

Therefore, if the randomization distribution is based on the shifted Kolmogorov-Smirnov statistic in eq (4), $K_{m,n}(Z)$, such that the difference between $K_{m,n}(Z) - K_{m,n}(\bar{Z}_{\pi_0})$ is small in some sense whenever $\bar{Z}$ and $\bar{Z}_{\pi_0}$ mostly agree, then one should be able to deduce the behavior of the permutation distribution under samples from F_0, F_1 from the behavior of the permutation distribution when all N observations come from the same distribution $\bar{P}$.

Suppose π and π' are independent random permutations, and independent of the Z_i and $\bar{Z}_i$. Suppose we can show that

$$\left(K_{m,n}(\bar{Z}_\pi), K_{m,n}(\bar{Z}_{\pi'})\right) \xrightarrow{\text{d}} (T, T') \tag{14}$$

where T and T' are independent with common cdf $R(\cdot)$. Then by theorem 5.1 in Chung and Romano (2013), the randomization distribution based on $K_{m,n}$ converges in probability to $R(\cdot)$ when all observations are iid according to $\bar{P}$. But since $\pi\pi_0$ (meaning π composed with π_0, so π_0 is applied first) and $\pi'\pi_0$ are also independent random permutations. Then it also implies that

$$\left(K_{m,n}(\bar{Z}_{\pi\pi_0}), K_{m,n}(\bar{Z}_{\pi'\pi_0})\right) \xrightarrow{\text{d}} (T, T')$$

Using the coupling construction, suppose it can be shown that

$$K_{m,n}(\bar{Z}_{\pi\pi_0}) - K_{m,n}(\bar{Z}_\pi) \xrightarrow{\text{P}} 0 \tag{15}$$

then it also follows that

$$K_{m,n}(\bar{Z}_{\pi'\pi_0}) - K_{m,n}(\bar{Z}_{\pi'}) \xrightarrow{\text{P}} 0$$

and by Slutsky's theorem

$$\begin{aligned}(K_{m,n}(Z_\pi), K_{m,n}(Z_{\pi'})) &= (K_{m,n}(Z_\pi), K_{m,n}(Z_{\pi'})) + \left(K_{m,n}(\bar{Z}_{\pi\pi_0}), K_{m,n}(\bar{Z}_{\pi'\pi_0})\right) \\ &\quad - \left(K_{m,n}(\bar{Z}_{\pi\pi_0}), K_{m,n}(\bar{Z}_{\pi'\pi_0})\right) \\ &= -(\underbrace{K_{m,n}(Z_\pi) - K_{m,n}(\bar{Z}_{\pi\pi_0})}_{\xrightarrow{\mathrm{P}} 0}, \underbrace{K_{m,n}(\bar{Z}_{\pi'\pi_0}) - K_{m,n}(Z_{\pi'})}_{\xrightarrow{\mathrm{P}} 0}) \\ &\quad + \underbrace{\left(K_{m,n}(\bar{Z}_{\pi\pi_0}), K_{m,n}(\bar{Z}_{\pi'\pi_0})\right)}_{\xrightarrow{\mathrm{d}} (T,T')}\end{aligned}$$

we can conclude that $(K_{m,n}(Z_\pi), K_{m,n}(Z_{\pi'})) \xrightarrow{\mathrm{d}} (T, T')$. Another application of Theorem 5.1 allows us to conclude that the randomization distribution also converges in probability to $R(\cdot)$ under the original model of two samples from possibly different distributions.

Appendix B: Proofs

8.1 Proofs of section 4.1

Proof of Proposition 4.1[7]. Assume the premises of the proposition, and write $\hat{F}_1(y+\delta) - \hat{F}_0(y)$ as $(\hat{F}_1(y+\delta) - F_1(y+\delta)) - (\hat{F}_0(y) - F_0(y))$. Then

$$\upsilon_{m,n}(y,\delta) = \sqrt{\frac{mn}{N}}\left(\hat{F}_1(y+\delta) - \hat{F}_0(y)\right) = \sqrt{1-p_m}\upsilon_1 - \sqrt{p_m}\upsilon_0$$

where $\upsilon_0 = \sqrt{n}(\hat{F}_0(y) - F_0(y))$ and $\upsilon_1 = \sqrt{m}(\hat{F}_1(y+\delta) - F_1(y+\delta))$ are two independent empirical processes. By Donsker's theorem (Theorem 19.3 in Van der Vaart (2000)), both sequences υ_0 and υ_1 can be approximated by two independent F_0 and F_1 Brownian bridge processes, $\mathbb{G}_0$ and $\mathbb{G}_1$ respectively. We can take these Brownian bridges to be independent because the empirical processes are. Therefore, $V_{m,n}(y,\delta)$ converges weakly to

$$\sqrt{1-p}\mathbb{G}_1(y) - \sqrt{p}\mathbb{G}_0(y)$$

which is another Brownian Bridge. Therefore, by the usual continuous mapping theorem, the sequences of "classical" KS statistic $K_{m,n,\delta} = \sup_y \parallel V_{m,n}(y,\delta) \parallel$ converge under the null hypothesis to

$$J_0(y) \equiv \sup_y \parallel \mathbb{G}(y) \parallel$$

□

The outline of the proof of Proposition 4.2 is the following. From Hoeffding's Condition (See Theorem 5.1 of Chung and Romano (2013)), we must verify

$$(V_{m,n}(y,\delta; Z_\pi)), V_{m,n}(y,\delta; Z_{\pi'}))) \tag{16}$$

converges weakly to a tight process $(\mathbb{G}, \mathbb{G}')$, where $\mathbb{G}$ and $\mathbb{G}'$ are independent Brownian bridges, each with identically distributed marginals having mean zero and covariance structure

$$\mathbb{E}\,\mathbb{G}(s)\mathbb{G}(t) = F_0(s \wedge t) - F_0(s)F_0(t)$$

[7] See also theorem 19.3 in Van der Vaart (2000).

Proof of Proposition 4.2.

Throughout this proof we suppose δ is known, so let us recenter the m observations coming from F_1 as follows

$$\tilde{Y}_i(1) = Y_i(1) - \delta \quad \text{for all} \quad i = 1, \ldots, m$$

then $\tilde{Y}_i(1) \sim \tilde{F}_1$. Since this is an affine transformation of the continuously distributed $Y(1)$ with density function f_1, we have that $\tilde{Y}(1)$ has probability density function $\tilde{f}_1$ given by $\tilde{f}_1(y) = f_1(y + \delta)$. Write

$$Z = (Z_1, \ldots, Z_N) = (\tilde{Y}_1(1), \ldots, \tilde{Y}_m(1), Y_1(0), \ldots, Y_n(0))$$

Thus under the null hypothesis $Z_1, \ldots, Z_N$ are iid F_0, implying that the mixture distribution is essentially F_0. Independent of the Zs, let $(\pi(1), \ldots, \pi(N))$ and $(\pi'(1), \ldots, \pi'(N))$ be two independent random permutations of $\{1, \ldots, N\}$. We will denote $Z_\pi = (Z_{\pi(1)}, \ldots, Z_{\pi(N)})$; $Z_{\pi'}$ is defined with π replaced by π'.

Assume the premises of Proposition 4.1. By applying Theorem 1.5.4 in Wellner and Van der Vaart (2013), the proof comes down to showing the marginals

$$(V_{m,n}(t_1, \delta; Z_\pi)), \ldots, V_{m,n}(t_k, \delta; Z_\pi)), V_{m,n}(t_1, \delta; Z_{\pi'})), \ldots, V_{m,n}(t_k, \delta; Z_{\pi'})))$$

converge weakly to the marginals

$$(\mathbb{G}(t_1), \ldots, \mathbb{G}(t_k), \mathbb{G}'(t_1), \ldots, \mathbb{G}'(t_k)))$$

for all $k \in \mathbb{N}$, and $t_1, \ldots, t_k \in \mathbb{R}$. For the sake of exposition, we first restrict our attention to the scalar y. Under H_0, we observe that

$$\begin{aligned}(V_{m,n}(y, \delta; Z_\pi)), V_{m,n}(y, \delta; Z_{\pi'}))) &= (1 - p_m)^{1/2} m^{-1/2} \left(\sum_{i=1}^N X_i W_i, \sum_{i=1}^N X_i W_i' \right) \\ &= K(m) \left(\sum_{i=1}^N X_i W_i, \sum_{i=1}^N X_i W_i' \right)\end{aligned}$$

where $X_i = 1_{\{Z_i \le y\}} - F_0(y)$, and $W_i = 1$ if $\pi(i) \in I_1 = \{1, \ldots, m\}$, $W_i = -m/n$ otherwise, for all i. Analogously, W_i' is defined with π replaced by π'. It is easy to check $\mathbb{E}(W_i) = 1\,\mathbb{P}(\pi(i) \in I_1) - m/n\,\mathbb{P}(\pi(i) \notin I_1) = 0$, and $\mathbb{E}((1_{\{Z_{\pi(i)}\} \le y} - F_0(y))W_i) = 0$ since π is independent of Z. Same is true for W_i'.

Notice that under the null,

$$\begin{aligned}&\mathbb{E}\left(V_{m,n}(y, \delta; Z_\pi)\right) = 0 \\ &\mathbb{V}\left(V_{m,n}(y, \delta; Z_\pi)\right) = \frac{mn}{N} \left(\frac{F_0(y)(1 - F_0(y))}{m} + \frac{F_0(y)(1 - F_0(y))}{n} \right) = F_0(y)(1 - F_0(y))\end{aligned}$$

We claim the asymptotic normality of

$$K(m) \left(\sum_{i=1}^N X_i W_i, \sum_{i=1}^N X_i W_i' \right)$$

To do this, we use the Cramér-Wold device (Theorem 11.2.3 of Lehmann and Romano (2006)). Then, for any any a and b, we must verify the limiting distribution of

$$K(m) \sum_{i=1}^N (a X_i W_i + b X_i W_i')) = \sum_{i=1}^N C_{m,n,i} X_i \tag{17}$$

where

$$C_{m,n,i} = K(m)(aW_i + bW_i')$$

Condition on W_i and W_i', then (17) is a conditionally independent sum of linear combination of independent variables:

$$\sum_{i=1}^{m} C_{m,n,i}X_i + \sum_{j=m+1}^{N} C_{m,n,j}X_j = \sum_{i=1}^{m} C_{m,n,i}\left(1_{\{\tilde{Y}_i(1)\le y\}} - F_0(y)\right) + \sum_{j=1}^{n} C_{m,n,m+j}\left(1_{\{Y_j(0)\le y\}} - F_0(y)\right)$$

By the arguments in Example 15.2.5 of Lehmann and Romano (2006), we conclude that

$$\frac{\max_{i=1,\dots,N} C_{m,n,i}}{\sum_{i=1}^{N} C_{m,n,j}^2} \xrightarrow{\mathrm{P}} 0, \quad \text{as} \quad m,n \to \infty$$

and so

$$\sum_{i=1}^{m} C_{m,n,i}\left(1_{\{\tilde{Y}_i(1)\le y\}} - F_0(y)\right) + \sum_{j=1}^{n} C_{m,n,m+j}\left(1_{\{Y_j(0)\le y\}} - F_0(y)\right) \xrightarrow{\mathrm{d}} a\mathbb{G} + b\mathbb{G}'$$

therefore

$$(V_{m,n}(y,\delta;Z_\pi)), V_{m,n}(y,\delta;Z_{\pi'}))) \xrightarrow{\mathrm{d}} (\mathbb{G}(y),\mathbb{G}'(y))$$

where $\mathbb{G}(y)$ and $\mathbb{G}'(y)$ follow the same zero-mean Gaussian process with covariance function $F_0(y)(1-F_0)$. Finally, conditionally on Ws, we have

$$\begin{aligned}\mathbb{C}\left(V_{m,n}(y,\delta;Z_\pi)), V_{m,n}(y,\delta;Z_{\pi'}))\right) &= K^2(m)\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{C}\left(X_iW_i, X_jW_j'\right) \\ &= K^2(m)\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}\left(X_iW_iX_jW_j'\right) = 0\end{aligned}$$

because π, π' are independent of Z, and mutually independent from each other. It follows that $\mathbb{G}(y)$ and $\mathbb{G}'(y)$ are independent, as desired. The same reasoning and the multivariate CLT apply for arbitrary tuples $t_1,\dots,t_k \in \mathbb{R}$.

It now follows that $(K_{m,n,\delta}(Z_\pi), K_{m,n,\delta}(Z_{\pi'}))$ are asymptotically independent. By the regular the continuous mapping theorem,

$$(K_{m,n,\delta}(Z_\pi), K_{m,n,\delta}(Z_{\pi'}))$$

converges in distribution to the (J_0, J_0') process with independent, identically distributed marginals as described in Proposition 4.1. Then by Hoeffding's Condition (Theorem 5.1 of Chung and Romano (2013)),

$$\sup_t \parallel \hat{R}_N(t) - J_0(t) \parallel \xrightarrow{\mathrm{P}} 0$$

□

8.2 Proofs of section 4.2

Proof of Proposition 4.3

See Ding et al. (2015), *Theorem 4*, page 21. We include the proof here for the sake of completeness. The discussion and results in examples V.15 and V.23 in Pollard (2012) will be useful.

Under the null hypothesis (2), we know that for some δ, $\delta = \mu(F_1) - \mu(F_0)$, $\sigma^2(F_1) = \sigma^2(F_0) = \sigma^2$, and $f_1(y+\delta) = f_0(y)$. Then we develop $V_{m,n}(y, \hat{\delta})$ as

$$\begin{aligned}\sqrt{\frac{mn}{N}}\left\{\hat{F}_1(y+\hat{\delta}) - \hat{F}_0(y)\right\} &= \sqrt{\frac{mn}{N}}\left\{\hat{F}_1(y+\delta) - \hat{F}_0(y)\right\} + \sqrt{\frac{mn}{N}}\left\{F_1(y+\hat{\delta}) - F_1(y+\delta)\right\} \\ &+ \sqrt{\frac{mn}{N}}\left\{\hat{F}_1(y+\hat{\delta}) - F_1(y+\hat{\delta})\right\} - \sqrt{\frac{mn}{N}}\left\{\hat{F}_1(y+\delta) - F_1(y+\delta)\right\} \\ &= \sqrt{\frac{mn}{N}}\left\{\hat{F}_1(y+\delta) - \hat{F}_0(y)\right\} + \sqrt{\frac{mn}{N}}\left\{F_1(y+\hat{\delta}) - F_1(y+\delta)\right\} + o_p(1)\end{aligned}$$

due to the fact the last two summands

$$\sqrt{\frac{mn}{N}}\left\{(\hat{F}_1(y+\hat{\delta}) - F_1(y+\hat{\delta})) - (\hat{F}_1(y+\delta)) - F_1(y+\delta)\right\} = o_p(1) \tag{18}$$

by stochastic equicontinuity of the indicator function. Now expand $F_1(y+\hat{\delta})$ around δ to obtain

$$\begin{aligned}V_{m,n}(y,\hat{\delta}) &= \sqrt{\frac{mn}{N}}\left\{\hat{F}_1(y+\delta) - \hat{F}_0(y)\right\} + \sqrt{\frac{mn}{N}}\left\{\left(F_1(y+\delta) + f_1(y+\delta)(\hat{\delta}-\delta)\right) - F_1(y+\delta)\right\} + o_p(1) \\ &= \sqrt{\frac{mn}{N}}(\hat{F}_1(y+\delta) - \hat{F}_0(y)) + \sqrt{\frac{mn}{N}}\left(f_0(y)(\hat{\delta}-\delta)\right) + o_p(1)\end{aligned}$$

Observe

$$\begin{aligned}\sqrt{\frac{mn}{N}}(\hat{\delta}-\delta) &= \sqrt{\frac{mn}{N}}\left((\mu(\hat{F}_1) - \mu(F_1)) - (\mu(\hat{F}_0) - \mu(F_0))\right) \\ &= \sqrt{\frac{mn}{N}}\left(\frac{1}{m}\sum_{i=1}^{m}(Y_i(1) - \mu(F_1)) - \frac{1}{n}\sum_{i=m+1}^{N}(Y_i(0) - \mu(F_0))\right) \\ &= \sqrt{\frac{n}{N}}\left(\frac{1}{\sqrt{m}}\sum_{i=1}^{m}(Y_i(1) - \mu(F_1))\right) - \sqrt{\frac{m}{N}}\left(\frac{1}{\sqrt{n}}\sum_{i=m+1}^{N}(Y_i(0) - \mu(F_0))\right)\end{aligned}$$

therefore

$$\begin{aligned}V_{m,n}(y,\hat{\delta}) &= \sqrt{\frac{mn}{N}}\left\{\hat{F}_1(y+\delta) - \hat{F}_0(y)\right\} + \sqrt{\frac{mn}{N}}\left(f_0(y)(\hat{\delta}-\delta)\right) + o_p(1) \\ &= \sqrt{\frac{mn}{N}}\left\{\left(\hat{F}_1(y+\delta) - F_1(y+\delta)\right) - \left(\hat{F}_0(y) - F_0(y)\right)\right\} + \sqrt{\frac{mn}{N}}\left(f_0(y)(\hat{\delta}-\delta)\right) + o_p(1) \\ &= \sqrt{1-p_n}\left(\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\left\{1_{\{Y_i(1)\le y+\delta\}} - F_1(y+\delta) + f_0(y)\left(Y_i(1) - \mu(F_1)\right)\right\}\right) \\ &- \sqrt{p_n}\left(\frac{1}{\sqrt{n}}\sum_{i=m+1}^{N}\left\{1_{\{Y_i(0)\le y\}} - F_0(y) + f_0(y)\left(Y_i(0) - \mu(F_0)\right)\right\}\right) + o_p(1)\end{aligned}$$

since both terms have the same limit distribution as *shifted* Brownian Bridges in Proposition 4.3, we have

$$V_{m,n}(y,\hat{\delta}) = \sqrt{\frac{mn}{N}}\left\{\hat{F}_0(y) - \hat{F}_1(y+\hat{\delta})\right\} \xrightarrow{\mathrm{d}} \mathbb{G}(y) - \xi(y)$$

and the final statement follows from the symmetry of the Brownian Bridge with drift, and the usual Continuous Mapping Theorem applied to it. □

Sketch of the proof of Proposition 4.4: Since we don't know δ, we cannot shift the observations as we did in the case of δ known. As a result, the general strategy must be based on the auxiliary results in section 5 of Chung and Romano (2013). In particular

(i) Let $\bar{Z}_1, \bar{Z}_2 \ldots,$ be iid from the mixture distribition $\bar{P}$. Stack in $\bar{Z} = (\bar{Z}_1, \ldots, \bar{Z}_N)$. Then show

(a)

$$\left(V_{m,n}(y, \hat{\delta}; \bar{Z}_\pi)), V_{m,n}(y, \hat{\delta}; \bar{Z}_{\pi'}))\right) \xrightarrow{\mathrm{d}} (\mathbb{G}, \mathbb{G}') \tag{19}$$

with $\mathbb{G}$ and $\mathbb{G}'$ independent with common CDF. Independence will follow from the zero-covariance argument, since the limits are Gaussian.

(b)

$$V_{m,n}(y, \hat{\delta}; \bar{Z}_{\pi,\pi 0}) - V_{m,n}(y, \hat{\delta}; Z_\pi) \xrightarrow{\mathrm{P}} 0 \tag{20}$$

(ii) Invoke Lemma 5.1 of Chung and Romano (2013) to conclude

$$\left(V_{m,n}(y, \hat{\delta}; Z_\pi)), V_{m,n}(y, \hat{\delta}; Z_{\pi'}))\right) \xrightarrow{\mathrm{d}} (\mathbb{G}, \mathbb{G}')$$

(iii) Apply Hoeffding's Condition (Theorem 5.1 of Chung and Romano (2013)) to conclude

$$\sup_t \parallel \hat{R}_N(t) - J_0(t) \parallel \xrightarrow{\mathrm{P}} 0$$

Proof of Proposition 4.3. Let $\hat{\delta} = \mu(\hat{F}_1) - \mu(\hat{F}_0)$ and recenter the m observations coming from F_1 as follows

$$\tilde{Y}_i(1) = Y_i(1) - \hat{\delta} \quad \text{for all} \quad i = 1, \ldots, m$$

where $\tilde{Y}_i(1) \sim \tilde{F}_1$. Write

$$Z = (Z_1, \ldots, Z_N) = (\tilde{Y}_1(1), \ldots, \tilde{Y}_m(1), Y_1(0), \ldots, Y_n(0))$$

Independent of the Zs, let $(\pi(1), \ldots, \pi(N))$ be an independent random permutation of $\{1, \ldots, N\}$. Let $\bar{Z}$ and π_0 be constructed by the coupling method of Chung and Romano (2013). We want to show Condition (20) first, i.e.

$$V_{m,n}(y, \hat{\delta}; \bar{Z}_{\pi,\pi 0}) - V_{m,n}(y, \hat{\delta}; Z_\pi) \xrightarrow{\mathrm{P}} 0$$

Everything stated below is implicitly conditioned on π_0, but we omit it to ease notation. For a given π,

$$\begin{aligned}\left(\frac{mn}{N}\right)^{-1/2} \left(V_{m,n}(y, \hat{\delta}; \bar{Z}_{\pi\pi_0})) - V_{m,n}(y, \hat{\delta}; Z_\pi)\right) = &\frac{1}{m}\sum_{i=1}^{m}(I\{\bar{Z}_{\pi\pi_0(i)} \le y\} - I\{Z_{\pi(i)} \le y\}) \\ &- \frac{1}{n}\sum_{j=m+1}^{N}(I\{\bar{Z}_{\pi\pi_0(j)} \le y\} - I\{Z_{\pi(j)} \le y\})\end{aligned}$$

and observe that the way we constructed $\bar{Z}$, we have that $Z_i = \bar{Z}_{\pi_0(i)}$ for indeces i except for at most D entries. This is so because $\bar{Z}_{\pi_0}$ is either of the form

$$(Z_{\pi_0(1)}, \ldots, Z_{\pi_0(N)}) = (\tilde{Y}_1(1), \ldots, \tilde{Y}_1(m), Y_1(0), \ldots, Y_{n-D}(0), \tilde{Y}_{m+1}(1), \ldots, \tilde{Y}_{m+D}(1))$$

or it is of the form

$$(Z_{\pi_0(1)}, \ldots, Z_{\pi_0(N)}) = (\tilde{Y}_1(1), \ldots, \tilde{Y}_{m-D}(1), Y_{n+1}(0), \ldots, Y_{n+D}(0), Y_0(1), \ldots, Y_0(n))$$

Then all the above sums are zero except for at most D places. For all the indices such that the differences $I\{\bar{Z}_{\pi\pi_0(i)} \leq y\} - I\{Z_{\pi(i)} \leq y\}$ and $I\{\bar{Z}_{\pi\pi_0(j)} \leq y\} - I\{Z_{\pi(j)} \leq y\}$ are not zero, observe that

$$\begin{aligned}\mathbb{E}\left(I\{\bar{Z}_{\pi\pi_0(i)} \leq y\} - I\{Z_{\pi(i)} \leq y\}\right) &= -\mathbb{E}\left(I\{\bar{Z}_{\pi\pi_0(j)} \leq y\} - I\{Z_{\pi(j)} \leq y\}\right) \\ &= p\tilde{F}_1(y) + (1-p)F_0(y) - F_0(y) \\ &= pF_1(y+\hat{\delta}) + (1-p)F_0(y) - F_0(y)\end{aligned}$$

Expand $F_1(y+\hat{\delta})$ around δ to obtain

$$\begin{aligned}\mathbb{E}\left(I\{\bar{Z}_{\pi\pi_0(i)} \leq y\} - I\{Z_{\pi(i)} \leq y\}\right) &= p\left(F_1(y+\delta) + f_1(y+\delta)(\hat{\delta}-\delta)\right) \\ &\quad + (1-p)F_0(y) - F_0(y) + o_p(1) = o_p(1)\end{aligned}$$

under the null hypothesis. Hence, conditionally on D and π,

$$\begin{aligned}\mathbb{E}\left(V_{m,n}(y,\hat{\delta};\bar{Z})) - V_{m,n}(y,\hat{\delta};Z)\right) &\leq \sqrt{\frac{mn}{N}}\left(\frac{D}{\min\{m,n\}}\right)\mathbb{E}\left(I\{\bar{Z}_{\pi\pi_0(i)} \leq y\} - I\{Z_{\pi(i)} \leq y\}\right) \\ &\leq \sqrt{\frac{mn}{N}}\left(\frac{\mathcal{O}(N^{1/2})}{\min\{m,n\}}\right)o_p(1) = o_p(1)\end{aligned}$$

Furthermore, any nonzero term like $I\{\bar{Z}_{\pi\pi_0(j)} \leq y\} - I\{Z_{\pi(j)} \leq y\}$ has variance bounded above by

$$\begin{aligned}\mathbb{V}\left(I\{\bar{Z}_{\pi\pi_0(j)} \leq y\} - I\{Z_{\pi(j)} \leq y\}\right) &= \mathbb{V}\left(I\{\bar{Z}_{\pi\pi_0(j)} \leq y\}\right) + \mathbb{V}\left(I\{Z_{\pi(j)} \leq y\}\right) \\ &= \mathbb{E}\left(I\{\bar{Z}_{\pi\pi_0(j)} \leq y\}\right)\left(1 - \mathbb{E}\left(I\{\bar{Z}_{\pi\pi_0(j)} \leq y\}\right)\right) \\ &\quad + \mathbb{E}\left(I\{Z_{\pi(j)} \leq y\}\right)\left(1 - E\left(I\{Z_{\pi(j)} \leq y\}\right)\right) \leq \frac{1}{2}\end{aligned}$$

Similarly, $\mathbb{V}\left(I\{\bar{Z}_{\pi\pi_0(i)} \leq y\} - I\{Z_{\pi(i)} \leq y\}\right) \leq 1/2$. Conditioning on D and π, the variance is bounded above in the sense:

$$\mathbb{V}\left(V_{m,n}(y,\hat{\delta};\bar{Z})) - V_{m,n}(y,\hat{\delta};Z)\right) \leq \frac{mn}{N}\left(D\left(\frac{1}{m^2} + \frac{1}{n^2}\right)\right) = \frac{mn}{N}\left(\frac{n^2+m^2}{n^2m^2}\right)D$$

and therefore the unconditional variance is bounded above by

$$\frac{mn}{N}\left(\frac{n^2+m^2}{n^2m^2}\right)\mathcal{O}(N^{1/2}) = \left(\frac{n}{m} + \frac{m}{n}\right)\mathcal{O}(N^{-1/2}) = \mathcal{O}(N^{-1/2}) = o(1)$$

and therefore condition (20) follows by convergence is quadratic mean.

In order to show (19), let $\bar{Z}_1, \bar{Z}_2, \ldots,$ be i.i.d. $\bar{P} = p\tilde{F}_1 + (1-p)F_0$, and stack the first N elements of the sequence into $\bar{Z}$ such that

$$\bar{Z} = (\bar{Z}_1, \ldots, \bar{Z}_N)$$

Independent of the $\bar{Z}$s, let $(\pi(1), \ldots, \pi(N))$ and $(\pi'(1), \ldots, \pi'(N))$ be two independent random permutations of $\{1, \ldots, N\}$. A direct application of Theorem 1.5.4 in Wellner and Van der Vaart (2013) implies we need to show the marginals

$$\left(V_{m,n}(t_1,\hat{\delta};\bar{Z}_\pi)), \ldots, V_{m,n}(t_k,\hat{\delta};\bar{Z}_\pi)), V_{m,n}(t_1,\hat{\delta};Z_{\pi'})), \ldots, V_{m,n}(t_k,\hat{\delta};\bar{Z}_{\pi'}))\right)$$

converge weakly to the marginals

$$(\mathbb{G}(t_1), \ldots, \mathbb{G}(t_k), \mathbb{G}'(t_1), \ldots, \mathbb{G}'(t_k)))$$

for all $k \in \mathbb{N}$, and $t_1, \ldots, t_k \in \mathbb{R}$. Just like we did in the proof of Proposition (4.4), we first restrict our attention to the scalar y. Under H_0, we observe that

$$\begin{aligned}\left(V_{m,n}(y, \hat{\delta}; \bar{Z}_\pi)), V_{m,n}(y, \hat{\delta}; \bar{Z}_{\pi'}))\right) &= (1-p_m)^{1/2} m^{-1/2} \left(\sum_{i=1}^N X_i W_i, \sum_{i=1}^N X_i W_i'\right) \\ &= K(m) \left(\sum_{i=1}^N X_i W_i, \sum_{i=1}^N X_i W_i'\right)\end{aligned}$$

where $X_i = 1_{\{\bar{Z}_i \le y\}} - F_0(y)$, and $W_i = 1$ if $\pi(i) \in I_1 = \{1, \ldots, m\}$, $W_i = -m/n$ otherwise, for all i. Analogously, W_i' is defined with π replaced by π'. It is easy to check $\mathbb{E}(W_i) = 1\,\mathbb{P}(\pi(i) \in I_1) - m/n\,\mathbb{P}(\pi(i) \notin I_1) = 0$, and $\mathbb{E}((1_{\{\bar{Z}_{\pi(i)}\} \le y} - F_0(y))W_i) = 0$ since π is independent of $\bar{Z}$. Same is true for W_i'.

Notice that under the null,

$$\begin{aligned}\mathbb{E}\left(V_{m,n}(y, \hat{\delta}; \bar{Z}_\pi)\right) &= p\tilde{F}_1(y) + (1-p)F_0 - F_0 \\ &= f_0(y)(\hat{\delta} - \delta) + o_p(1) = \hat{R}_{m,n} \\ \mathbb{V}\left(V_{m,n}(y, \hat{\delta}; \bar{Z}_\pi)\right) &= \frac{mn}{N}\left(\frac{\bar{P}(y)(1-\bar{P}(y))}{m} + \frac{\bar{P}(1-\bar{P}(y))}{n}\right) = \bar{P}(y)(1-\bar{P}(y)) \\ &= F_0(y)(1-F_0(y)) + \hat{R}_{m,n}(1 - \hat{R}_{m,n} - 2F_0(y))\end{aligned}$$

We claim the asymptotic normality of

$$K(m) \left(\sum_{i=1}^N X_i W_i, \sum_{i=1}^N X_i W_i'\right)$$

To do this, we use the Cramér-Wold device (Theorem 11.2.3 of Lehmann and Romano (2006)). Then, for any any a and b, we must verify the limiting distribution of

$$K(m) \sum_{i=1}^N (aX_i W_i + bX_i W_i')) = \sum_{i=1}^N C_{m,n,i} X_i \tag{21}$$

where

$$C_{m,n,i} = K(m)(aW_i + bW_i')$$

Condition on W_i and W_i', then (21) is a conditionally independent sum of linear combination of independent variables:

$$\sum_{i=1}^m C_{m,n,i} X_i + \sum_{j=m+1}^N C_{m,n,j} X_j = \sum_{i=1}^m C_{m,n,i}\left(1_{\{\bar{Z}_i \le y\}} - F_0(y)\right) + \sum_{j=1}^n C_{m,n,m+j}\left(1_{\{\bar{Z}_j \le y\}} - F_0(y)\right)$$

By the arguments in Example 15.2.5 of Lehmann and Romano (2006), we conclude that

$$\frac{\max_{i=1,\ldots,N} C_{m,n,i}}{\sum_{i=1}^N C_{m,n,j}^2} \xrightarrow{\mathrm{P}} 0, \quad \text{as} \quad m, n \to \infty$$

and so

$$\sum_{i=1}^m C_{m,n,i}\left(1_{\{\bar{Z}_i \le y\}} - F_0(y)\right) + \sum_{j=1}^n C_{m,n,m+j}\left(1_{\{\bar{Z}_j \le y\}} - F_0(y)\right) \xrightarrow{\mathrm{d}} a\mathbb{G} + b\mathbb{G}'$$

therefore

$$\left(V_{m,n}(y,\hat{\delta};\bar{Z}_{\pi})), V_{m,n}(y,\hat{\delta};\bar{Z}_{\pi'}))\right) \xrightarrow{\mathrm{d}} (\mathbb{G}(y),\mathbb{G}'(y))$$

where $\mathbb{G}(y)$ and $\mathbb{G}'(y)$ follow the same zero-mean Gaussian process with covariance function $F_0(y)(1-F_0)$. Finally, conditionally on Ws, we have

$$\begin{aligned}\mathbb{C}\left(V_{m,n}(y,\hat{\delta};\bar{Z}_{\pi})), V_{m,n}(y,\hat{\delta};\bar{Z}_{\pi'}))\right) &= K^2(m)\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{C}\left(X_iW_i, X_jW_j'\right)\\ &= K^2(m)\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}\left(X_iW_iX_jW_j'\right) = 0\end{aligned}$$

because π, π' are independent of $\bar{Z}$, and mutually independent from each other. It follows that $\mathbb{G}(y)$ and $\mathbb{G}'(y)$ are independent, as desired. The same reasoning and the multivariate CLT apply for arbitrary tuples $t_1,\ldots,t_k \in \mathbb{R}$.

It now follows that $\left(K_{m,n,\hat{\delta}}(\bar{Z}_{\pi}), K_{m,n,\hat{\delta}}(\bar{Z}_{\pi'})\right)$ are asymptotically independent. By the regular the continuous mapping theorem,

$$\left(K_{m,n,\hat{\delta}}(\bar{Z}_{\pi}), K_{m,n,\hat{\delta}}(\bar{Z}_{\pi'})\right)$$

converges in distribution to the (J_0, J_0') process with independent, identically distributed marginals as described in Proposition (4.3). Then by Hoeffding's Condition (Theorem 5.1 of Chung and Romano (2013)),

$$\sup_t \parallel \hat{R}_N(t) - J_0(t) \parallel \xrightarrow{\mathrm{P}} 0$$

where $\hat{R}_N$ is the permutation distribution (8) based on $K_{m,n,\hat{\delta}}$ as desired.

□

8.3 Proofs of section 5

Proof of Proposition 5.1 Assume the premises of Proposition 5.1. Consider the asymptotic representation

$$\begin{aligned}\upsilon_{m,n}(t,\hat{\delta}) &= \sqrt{\frac{mn}{N}}\left(\hat{G}_1(t,\delta) - \hat{G}_0(y)\right) + \sqrt{\frac{mn}{N}}\left(f_0\left(F_0^{-1}(t)\right)(\hat{\delta}-\delta)\right) + o_p(1)\\ &= \upsilon_{m,n}(y,\delta) + \sqrt{\frac{mn}{N}}\left(f_0\left(F_0^{-1}(t)\right)(\hat{\delta}-\delta)\right) + o_p(1)\end{aligned}$$

Using $g(r) = (r, f_0)'$, the Khmaladze transformation based on $\upsilon_{m,n}(y,\hat{\delta})$ is

$$\begin{aligned}\tilde{\upsilon}_{m,n}(t,\hat{\delta}) &= \upsilon_{m,n}(t,\hat{\delta}) - \int_0^t \left[\dot{g}(s)'C^{-1}(s)\int_s^1 \dot{g}(r)d\upsilon_{m,n}(r,\hat{\delta})\right]ds\\ &= \upsilon_{m,n}(t,\hat{\delta}) - \phi_g(\upsilon_{m,n}(t,\hat{\delta}))\end{aligned}$$

From the properties of the map ϕ, we have $\phi_g(cg) = cg$ for a constant or random variable c. Then, for $g(t) = (t, f_0(t))'$ we have $\phi_g(cf_0) = cf_0$. Replace

$$c = \sqrt{\frac{mn}{N}}\left(\hat{\delta}-\delta\right)$$

therefore

$$\begin{aligned}\upsilon_{m,n}(t,\hat{\delta}) - \phi_g(\upsilon_{m,n}(t,\hat{\delta})) &= \upsilon_{m,n}(t,\delta) + cf_0\left(F_0^{-1}(t)\right) - \phi_g\left(\upsilon_{m,n}(t,\delta)\right) - \phi_g(cf_0\left(F_0^{-1}(t)\right)) + o_p(1)\\ &= \upsilon_{m,n}(t,\delta) - \phi_g\left(\upsilon_{m,n}(t,\delta)\right) + o_p(1)\end{aligned}$$

Weak convergence of $\upsilon_{m,n}(t,\delta)$ to $\mathbb{B}^0$ was established in Remark 4.1. Thus, $\tilde{\upsilon}_{m,n}(t,\hat{\delta})$ weakly converges to the Brownian motion $\mathbb{B}^0(t) - \phi_g(\mathbb{B}^0(t))$, by 4.3 of Khmaladze (1981).

□

Sketch of the proof of Proposition 5.2: In a similar spirit as in the proof of Proposition 4.4 and the results in section 5 of Chung and Romano (2013),

(i) Let $\bar{Z}_1, \bar{Z}_2 \ldots,$ be iid from the mixture distribution $\bar{P}$. Stack in $\bar{Z} = (\bar{Z}_1, \ldots, \bar{Z}_N)$. Then show

(a)

$$\left(\tilde{\upsilon}_{m,n}(t,\hat{\delta};\bar{Z}_\pi)), \tilde{\upsilon}_{m,n}(t,\hat{\delta};\bar{Z}_{\pi'}))\right) \xrightarrow{\mathrm{d}} (BM, BM') \tag{22}$$

with $\mathbb{BM}$ and $\mathbb{BM}'$ independent with common CDF. Independence will follow from the zero-covariance argument, since the limits are Gaussian.

(b)

$$\tilde{\upsilon}_{m,n}(t,\hat{\delta};\bar{Z}_{\pi,\pi 0}) - \tilde{\upsilon}_{m,n}(t,\hat{\delta};Z_\pi) \xrightarrow{\mathrm{P}} 0 \tag{23}$$

(ii) Invoke Lemma 5.1 of Chung and Romano (2013) to conclude

$$\left(\tilde{\upsilon}_{m,n}(t,\hat{\delta};Z_\pi)), \tilde{\upsilon}_{m,n}(t,\hat{\delta};Z_{\pi'}))\right) \xrightarrow{\mathrm{d}} (BM, BM')$$

(iii) Apply Hoeffding's Condition (Theorem 5.1 of Chung and Romano (2013)) to conclude

$$\sup_t \parallel \hat{R}_N(t) - J_2(t) \parallel \xrightarrow{\mathrm{P}} 0$$

Proof of Proposition 5.2 Write

$$Z = (Z_1, \ldots, Z_N) = (Y_1(1), \ldots, Y_m(1), Y_1(0), \ldots, Y_n(0))$$

Independent of the Zs, let $(\pi(1), \ldots, \pi(N))$ be an independent random permutation of $\{1, \ldots, N\}$. Let $\bar{Z}$ and π_0 be constructed by the coupling method of Chung and Romano (2013). We want to show Condition (23) first, i.e.

$$\tilde{\upsilon}_{m,n}(t,\hat{\delta};\bar{Z}_{\pi,\pi 0}) - \tilde{\upsilon}_{m,n}(t,\hat{\delta};Z_\pi) \xrightarrow{\mathrm{P}} 0$$

Everything stated below is implicitly conditioned on π_0, but we omit it to ease notation. Fix π and use the asymptotic representation in proof of Proposition (5.1)

$$\begin{aligned}\tilde{\upsilon}_{m,n}(t,\hat{\delta};\bar{Z}_{\pi,\pi 0}) - \tilde{\upsilon}_{m,n}(t,\hat{\delta};Z_\pi) = \upsilon_{m,n}(t,\delta;\bar{Z}_{\pi,\pi 0}) - \upsilon_{m,n}(t,\delta;\bar{Z}_\pi) - \\ \left(\phi_g\left(\upsilon_{m,n}(t,\delta;\bar{Z}_{\pi,\pi 0})\right) - \phi_g\left(\upsilon_{m,n}(t,\delta;\bar{Z}_\pi)\right)\right) + o_p(1)\end{aligned}$$

We need to guarantee that the remainder, defined in eq (18) in the proof of Proposition 4.3, is still $o_p(1)$ under Z_π. We will use the contiguity result of Chung and Romano (2013); let $V_1, V_2, \ldots,$ be iid from the mixture distribution $\bar{P} = pF_1 + (1-p)F_0$, and observe the remainder satisfies

$$\sqrt{\frac{mn}{N}}\left\{\frac{1}{m}\sum_{i=i}^{m} 1_{\{V_i \le y+\hat{\delta}\}} - F_1(y+\hat{\delta})\right\} - \sqrt{\frac{mn}{N}}\left\{\frac{1}{m}\sum_{i=i}^{m} 1_{\{V_i \le y+\delta\}} - F_1(y+\delta)\right\} \xrightarrow{\mathrm{P}} 0$$

by stochastic equicontinuity of the indicator function. Then, by Lemma 5.3 of Chung and Romano (2013),

$$\sqrt{\frac{mn}{N}}\left\{\frac{1}{m}\sum_{i=i}^{m}1_{\{Z_{\pi(i)}\leq y+\hat{\delta}\}}-F_1(y+\hat{\delta})\right\}-\sqrt{\frac{mn}{N}}\left\{\frac{1}{m}\sum_{i=i}^{m}1_{\{Z_{\pi(i)}\leq y+\delta\}}-F_1(y+\delta)\right\}\xrightarrow{\mathrm{P}}0$$

as desired. Furthermore, by the arguments of Proposition 4.2 and Slutsky theorem with the change of variable described in Remark 4.1,

$$\upsilon_{m,n}(y,\delta;\bar{Z}_{\pi,\pi 0})-\upsilon_{m,n}(y,\delta;\bar{Z}_{\pi})=o_p(1)$$

The linear operator ϕ_g is also a Fredholm operator (Koenker and Xiao (2002)) on a Banch space, therefore it is a bounded operator. But an operator between normed spaces is bounded if and only if it is a continuous operator (Abramovich and Aliprantis (2002)). Therefore, by the Continuous Mapping Theorem,

$$\phi_g\left(\upsilon_{m,n}(t,\delta;\bar{Z}_{\pi,\pi 0})\right)-\phi_g\left(\upsilon_{m,n}(t,\delta;\bar{Z}_{\pi})\right)=o_p(1)$$

then

$$\tilde{\upsilon}_{m,n}(t,\hat{\delta};\bar{Z}_{\pi,\pi 0})-\tilde{\upsilon}_{m,n}(t,\hat{\delta};Z_{\pi})=o_p(1)$$

as desired.

Joint Gaussianity in Condition (22) follows from the discussion in Condition E of Romano (1989). More specifically, the differentiability condition needed in order to verify Condition E holds for the present case, since testing the null hypothesis (2) is essentially a two-sample test of homogeneity (see example 4 of Romano (1989)). Having shown the limits are Gaussian, zero-covariance renders independence. Then, it needs to be shown that

$$\mathbb{C}\left(\tilde{\upsilon}_{m,n}(t,\hat{\delta};Z_{\pi})),\tilde{\upsilon}_{m,n}(t,\hat{\delta};Z_{\pi'}))\right)=0$$

Notice that

$$\tilde{\upsilon}_{m,n}(t,\hat{\delta};Z_{\pi})=\upsilon_{m,n}(t,\delta;Z_{\pi})-\phi_g\left(\upsilon_{m,n}(t,\delta;Z_{\pi})\right)+o_p(1)$$

by exploiting the linearity of the map ϕ_g. Therefore

$$\begin{aligned}\mathbb{C}\left(\tilde{\upsilon}_{m,n}(t,\hat{\delta};Z_{\pi})),\tilde{\upsilon}_{m,n}(t,\hat{\delta};Z_{\pi'}))\right)&=\mathbb{C}\left(\upsilon_{m,n}(t,\delta;Z_{\pi}),\upsilon_{m,n}(t,\delta;Z_{\pi'})\right)\\&+\mathbb{C}\left(\phi_g\left(\upsilon_{m,n}(t,\delta;Z_{\pi})\right),\phi_g\left(\upsilon_{m,n}(t,\delta;Z_{\pi'})\right)\right)\\&-\mathbb{C}\left(\upsilon_{m,n}(t,\delta;Z_{\pi}),\phi_g\left(\upsilon_{m,n}(t,\delta;Z_{\pi'})\right)\right)\\&-\mathbb{C}\left(\upsilon_{m,n}(t,\delta;Z_{\pi'}),\phi_g\left(\upsilon_{m,n}(t,\delta;Z_{\pi})\right)\right)+o_p(1)\end{aligned}$$

It follows from the arguments in the proof of Proposition 4.2 that

$$\begin{aligned}V_{m,n}(y,\delta;Z_{\pi}))&=K(m)\sum_{i=1}^{N}X_iW_i\\\mathbb{C}\left(V_{m,n}(y,\delta;Z_{\pi}),V_{m,n}(y,\delta;Z_{\pi'})\right)&=0\end{aligned}$$

Use linearity of the map ϕ_g once again,

$$\phi_g\left(V_{m,n}(y,\delta;Z_{\pi}))\right)=\phi_g\left(K(m)\sum_{i=1}^{N}X_iW_i\right)=K(m)\sum_{i=1}^{N}\phi_g\left(X_iW_i\right)$$

Therefore, conditionally on Ws,

$$\begin{aligned}
\mathbb{C}\left(\phi_g\left(V_{m,n}(y,\delta;Z_\pi)\right),\phi_g\left(V_{m,n}(y,\delta;Z_{\pi'})\right)\right) &= K^2(m)\,\mathbb{C}\left(\sum_{i=1}^{N}\phi_g\left(X_iW_i\right),\sum_{i=1}^{N}\phi_g\left(X_iW_i'\right)\right)\\
&= K^2(m)\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}\left(\phi_g\left(X_iW_i\right)\phi_g\left(X_jW_j'\right)\right)=0\\
\mathbb{C}\left(V_{m,n}(y,\delta;Z_{\pi'}),\phi_g\left(V_{m,n}(y,\delta;Z_\pi)\right)\right) &= K^2(m)\,\mathbb{C}\left(\sum_{i=1}^{N}\phi_g\left(X_iW_i\right),\sum_{i=1}^{N}X_iW_i'\right)\\
&= K^2(m)\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}\left(\phi_g\left(X_iW_i\right)X_jW_j'\right)=0
\end{aligned}$$

because π, π' are independent of Z, and mutually independent from each other. Once again, Slutsky theorem with the change of variable described in Remark 4.1 implies

$$\mathbb{C}\left(\tilde{v}_{m,n}(t,\hat{\delta};Z_\pi)),\tilde{v}_{m,n}(t,\hat{\delta};Z_{\pi'}))\right)=o_p(1)$$

and by Hoeffding's Condition (Theorem 5.1 of Chung and Romano (2013)), we conclude

$$\sup_t \parallel \hat{R}_N(t)-J_2(t)\parallel \xrightarrow{\mathrm{P}} 0$$

□