Permutation Test for Heterogeneous Treatment Effects with a Nuisance Parameter

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

Mauricio Olivares
Department of Economics
UIUC
lvrsgnz2@illinois.edu

January 22, 2020

Abstract

This paper proposes an asymptotically valid permutation test for heterogeneous treatment effects in the presence of an estimated nuisance parameter. Not accounting for the estimation error of the nuisance parameter results in statistics that depend on the particulars of the data generating process, and the resulting permutation test fails to control the Type 1 error, even asymptotically.

In this paper we consider a permutation test based on the martingale transformation of the empirical process to render an asymptotically pivotal statistic, effectively nullifying the effect associated with the estimation error on the limiting distribution of the statistic. Under weak conditions, we show that the permutation test based on the martingale-transformed statistic results in the asymptotic rejection probability of α in general while retaining the exact control of the test level when testing for the more restrictive sharp null. We also show how our martingale-based permutation test extends to testing whether there exists treatment effect heterogeneity within subgroups defined by observable covariates. Our approach comprises testing the joint null hypothesis that treatment effects are constant within mutually exclusive subgroups while allowing the treatment effects to vary across subgroups.

Monte Carlo simulations show that the permutation test presented here performs well in finite samples, and is comparable to those existing in the literature. To gain further understanding of the test to practical problems, we investigate the gift exchange hypothesis in the context of two field experiments from Gneezy and List (2006). Lastly, we provide the companion `RATest` `R` package to facilitate and encourage the application of our test in empirical research.

Keywords: Permutation Test, Heterogeneous Treatment Effect, Empirical Process, Martingale Transformation, Multiple hypothesis testing, Westfall–Young.

JEL Classification: C12, C14, C46.

[†]A previous version of this paper was circulated under the title "Non-Parametric Hypothesis Testing with a Nuisance Parameter: A Permutation Test Approach." All errors are our own.

1 Introduction

The main goal of this paper is to test whether the treatment effect is heterogeneous in the presence of an estimated nuisance parameter. In particular, we propose a permutation test approach to conduct inference under minimal assumptions in situations where randomization ideas apply, such as randomized experiments.

The statistical problem we examine has the following structure. Consider two real-valued random variables $Y(0)$ and $Y(1)$ representing the potential outcomes from a randomized trial (Neyman, 1990; Rubin, 1974), with distribution functions $F_0(\cdot)$ and $F_1(\cdot)$, respectively. This paper focuses on the following type of null hypothesis:

$$H_0 : F_1(y + \delta) = F_0(y) \ \forall\, y, \quad \text{for some} \quad \delta \ ,$$

based on two independent samples from their respective distributions. In other words, we want to test the null hypothesis of whether the corresponding treatment induces a constant shift in the potential outcome distribution.

Permutation tests are known to have attractive properties under the randomization hypothesis (Lehmann and Romano, 2005). As long as the permuted sample has the same joint distribution as the original sample under the null hypothesis, permutation tests control size in finite samples, *i.e.* the rejection probability under the null hypothesis is *exactly* the nominal level α. Besides, they are nonparametric in the sense that they can be applied without any parametric assumptions about the underlying distribution that generates the data. Moreover, the general construction of a permutation test does not depend on the specific form of the test statistic, though some statistics will be more suitable and will have better power performance for a specific null hypothesis. Finally, Hoeffding (1952) showed that for many interesting problems, permutation tests are asymptotically as powerful as standard optimal procedures. These features make them desirable for analyzing randomized experiments.

However, these classical properties of the permutation tests do not apply to the testing problem at hand when δ is unknown and thus becomes an unknown nuisance parameter—the error involved in the estimation of δ renders a statistic whose limiting distribution depends on the underlying data generating process. Consequently, the resulting permutation test based on naively plugging in the estimated parameter fails to control Type 1 error even asymptotically since the statistic is no longer asymptotically pivotal.

We propose a novel asymptotically valid permutation test for testing heterogeneous treatment effect in the presence of an estimated nuisance parameter. Our approach exploits the martingale transformation of the empirical process introduced by Khmaladze (1981) in the two-sample case[1]. The idea behind the Khmaladze transformation is to modify the empirical

[1]There is a rich literature on using the martingale transformation method to obtain asymptotically distribution-free tests (see Li, 2009, for a thorough review). Notable examples in econometrics include the

process so that the resulting statistic becomes asymptotically pivotal. More specifically, the Khmaladze transformation clears the empirical process out from the nuisance parameters by decomposing it into two parts—a martingale with a standard Brownian motion limiting behavior, and a second part that vanishes in the limit as the sample size increases. This strategy leaves us with an asymptotically distribution-free empirical process, a property that carries over the sup-norm functionals of it. We show in this paper that a permutation test based on this martingale-transformed statistic controls the limiting rejection probability, restoring the asymptotic validity of the permutation test.[2] We extend the proposed method to test whether there exists treatment effect heterogeneity within subgroups defined by observable covariates. Our approach boils down to jointly testing the null hypotheses that treatment effects are constant within mutually exclusive subgroups while allowing them to be different across subgroups. A byproduct of this extension is that we are also able to determine for which groups, if any, there is a heterogeneous treatment effect. Lastly, we provide the companion `RATest` `R` package, available on CRAN, to simplify and encourage the application of our test in empirical research.

More broadly, the problem of nonstandard distributions for sup-norm tests, or procedures based on sup-norm functionals like the permutation test presented here, falls into the classical goodness-of-fit problem with estimated nuisance parameter. The martingale transformation of Khmaladze (1981, 1993) in this paper is just one way to generate asymptotically distribution-free tests, but other approaches are available. Durbin (1973, 1975, 1985) and Parker (2013) methods conduct distributionally dependent inference based on Fourier inversion and boundary-crossing probabilities, whereas Chernozhukov and Fernández-Val (2005) and Linton et al. (2005) propose resampling methods to determine critical values.

Detecting treatment effect heterogeneity among individuals plays a key role in the design and successful evaluation of a social program using randomized experiments[3]. For example, an individual may benefit or suffer greatly from a policy intervention while another individual

pioneering works of Bai and Ng (2001) on conditional symmetry in time series, Koenker and Xiao (2002) for the quantile regression process, and testing parametric conditional distributions by Bai (2003). This martingale approach has been generalized by Song (2010) to include semiparametric models such as single index restrictions, partially parametric regressions, and partially parametric quantile regressions. Other extensions include nonlinear regression (Stute et al., 1998; Khmaladze and Koul, 2004, 2009), specification tests for autoregressive processes (Koul and Stute, 1999; Delgado et al., 2005; Delgado and Stute, 2008), or tests for parametric volatility function of a diffusion model (Chen et al., 2015) are also readily available.

[2]Restoring asymptotic validity of the permutation test by modifying the statistic that is based upon (so that it is asymptotically pivotal) can be found in the literature, including the pioneering papers of Neuhaus (1993) in the context of censoring models, or equality of univariate means and statistical functionals (Janssen, 1997, 1999). More generally, the asymptotic theory in Chung and Romano (2013) allows to handle general univariate testing problems. See Chung and Romano (2016) and references therein for more examples of the same idea.

[3]The 2019 Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Novel exemplifies the importance of this claim—in the fight to alleviate poverty, The Royal Swedish Academy of Sciences argues, "questions are often best answered via carefully designed experiments among the people who are most affected" (Nobel Media AB, 2019). This careful design of experiments depends to a large extent on our ability to comprehend the potential heterogeneity in the treatment effect.

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 thoroughly and comprehensively.

In order to detect whether there is heterogeneity in the treatment effect, 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. Härdle and Marron, 1990; Neumeyer and Dette, 2003; Crump et al., 2008; Imai and Ratkovic, 2013; Wager and Athey, 2018). Even though 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[4].

We follow a different route, and look at the entire outcome distributions. There is already a body of research that devotes considerable attention to comparing distributions to overcome the limitations resulting from solely looking at the average treatment effects. Notable examples comparing the marginal distribution functions of the potential outcomes include the randomization test of Ding et al. (2015), and the multiple-testing approach of Goldman and Kaplan (2018) to determine where the distributions differ. Quantile-based inference, analogously, investigates heterogeneity across individuals conditional on the quantile of the outcome distribution (Lehmann, 1974; Doksum, 1974; Koenker and Xiao, 2002; Chernozhukov and Fernández-Val, 2005) by exploiting the correspondence between quantiles and distribution functions.

Among all the aforementioned papers, our work is most closely related to Ding et al. (2015), but differs substantially in two important ways when there is an unknown nuisance parameter. First, our test is asymptotic in nature—our permutation test is based on a martingale transformation of the empirical process to obtain a pivotal statistic. The permutation test proposed by Ding et al. (2015), on the other hand, relies on constructing a confidence interval for the unknown nuisance parameter, repeating the permutation test pointwise over the interval, and then taking the maximum p-value. Second, our procedure controls the limiting rejection probability asymptotically. Meanwhile, though the pointwise procedure of Ding et al. (2015) yields a valid permutation test, it is conservative because it considers the maximum p-value. Our Monte Carlo experiments show that our proposed method delivers a better size control, confirming this observation.

The layout of the article is organized as follows. Section 2 presents an overview of the statistical problem at hand, highlighting its main theoretical challenges. Section 2.1–2.2 introduce the basic setting for permutation tests for the sharp null, where the permutation test retains an exact control in finite samples. We show in Section 2.3 that the permutation test based on the test statistic with estimated nuisance parameter fails to control the rejection probability even asymptotically. To address this issue, in Section 3 we apply the martingale

[4]See Bitler et al. (2006, 2017) and Xiao and Xu (2019) for a good exposition about the limitations of mean impacts and subgroup variation.

transformation, yielding an empirical process that is asymptotically pivotal. Under weak assumptions that make this transformation possible, we show that the permutation test based on this martingale-transformed statistic controls the limiting rejection probability. In Section 4 we extend the proposed method to conduct inference about heterogeneity in the treatment effect for specific subgroups defined by observable covariates, approaching this testing problem as a multiple hypothesis testing problem. Numerical results, simulations and computational results of our paper and competing alternatives can be found in Section 5. Section 6 is dedicated to the empirical illustration of the proposed method, where we apply our test to investigate the gift exchange hypothesis in the context of two field experiments from Gneezy and List (2006). Lastly, conclusions are collected in Section 7. Proofs, auxiliary lemmas and additional material are contained in Appendices A–D.

2 Statistical Environment

Consider the following randomized experiment model, where Y_i denotes the (observed) outcome of interest for the unit ith, and D_i is a binary treatment indicating whether the ith unit is treated or not. As usual, if the unit is treated, $D_i = 1$ and we will say it belongs to the treatment group, otherwise $D_i = 0$ and it belongs to the control group. Let $Y_i(1)$ be the potential outcome of the ith unit if treated, and $Y_i(0)$ the potential outcome of the ith unit if not treated. The observed outcome of interest and the potential outcomes are related to treatment assignment by the relationship

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

Our object of interest is the treatment effect, defined to be the difference between potential outcomes of the ith unit, $\delta_i = Y_i(1) - Y_i(0)$. The treatment effect is **constant** if $\delta_i = \delta$, otherwise we say the treatment effect is **heterogeneous**. The constant treatment effect null hypothesis is then

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

If δ were to be known, then (1) becomes a sharp null[5]. Hypotheses like (1) are, however, not directly testable because we happen to observe at most one potential outcome for each unit (the so-called fundamental problem of causal inference (e.g. Holland, 1986)). A different but testable hypothesis is available if we consider the marginal distributions of the observed outcomes for units that were treated and units who were not.

[5]Hypotheses like (1) are sharp because under this hypothesis all potential outcomes are known exactly—it is specified for all units. Fisher's original formulation assumes the sharp null of zero effect i.e. $\delta = 0$ for all i.

More formally, let $Y_i(1)$ and $Y_i(0)$ be two independent real-valued random variables having distribution functions $F_1(\cdot)$ and $F_0(\cdot)$. The (testable) constant treatment effect null hypothesis becomes:

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

Note that (2) embeds the null hypothesis (1), and therefore a test that rejects H_0 implies rejecting the more restrictive null hypothesis H_0^s by necessity, but not the other way around.

Remark 1. Under the null hypothesis (2), the distribution functions (CDF) of the potential outcomes of treatment and control groups, $F_1(y)$ and $F_0(y)$, are a constant shift apart. Therefore, the means of the outcomes under treatment and control satisfy $\int y dF_1(y) = \int y dF_0(y) + \delta$. This implies that δ is identified and $\sqrt{n}$-consistently estimable as the difference in sample means from both groups. ■

Remark 2. Constant treatment effect null hypotheses may be equivalently formulated in terms of quantiles, rather than CDFs, by adopting the Doksum–Lehmann quantile treatment model (Doksum, 1974; Lehmann, 1974). Thus by changing variables so $\tau = F_0(y)$, we obtain the *quantile treatment effect*

$$\delta(\tau) = F_1^{-1}(\tau) - F_0^{-1}(\tau) \ , \tag{3}$$

where $F^{-1} = \inf\{y : F(y) \geq \tau\}$, as usual. As a result, the constant treatment effect null hypothesis boils down to suppress the dependency of δ on τ so $\delta(\tau) = \delta$ for all $\tau \in [0,1]$. Examples of this approach are found in Koenker and Xiao (2002) and Chernozhukov and Fernández-Val (2005). We are *not* adopting this formulation and hence we are dealing with CDFs. For more on the quantile treatment effects, see Doksum and Sievers (1976). ■

We now discuss two assumptions that are relevant throughout the paper:

A. 1. *Let $n \to \infty$, $m \to \infty$, with $N = n + m$, $p_m = m/N$, and $p_m \to p \in (0,1)$ with $p_m - p = \mathcal{O}(N^{-1/2})$.*

A. 2. *F_1 and F_0 are absolutely continuous, with densities, f_1 and f_0 respectively. Furthermore, F_0 and F_1 as well as their densities are continuously differentiable with respect to δ.*

Assumption A.1 is standard for the asymptotic results. However, its relevance will become more palpable when we investigate the asymptotic behavior of the permutation distribution because, as we will show, it behaves like the unconditional distribution of the test statistic when all N observations are i.i.d. from the mixture distribution $pF_1 + (1-p)F_0$.

Assumption A.2, on the other hand, will be key to establishing the properties of the permutation test as a result of estimating the nuisance parameter δ. In particular, *i)* it allows us to expand the empirical process around the nuisance parameter δ, *ii)* it guarantees that the mixture distribution is absolutely continuous as well, and *iii)* it ensures the transformation of the uniform empirical process into an innovation martingale.

2.1 Test Statistic

Let $Y_1(1), \ldots, Y_m(1)$ and $Y_1(0), \ldots, Y_n(0)$ be two independent random samples from F_1 and F_0, and collect them in Z as follows

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

Under the null hypothesis $H_0 : F_1(y + \delta) = F_0(y)$, so a natural candidate for a test statistic for the hypothesis (2) is to compare empirical CDFs

$$\hat{F}_1(y + \hat{\delta}) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\{Y_i(1) \leq y + \hat{\delta}\}} \quad \text{and} \quad \hat{F}_0(y) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{\{Y_j(0) \leq y\}}$$

where $\hat{\delta}$ is given by

$$\hat{\delta} = \frac{1}{m} \sum_{i=1}^{m} Y_i(1) - \frac{1}{n} \sum_{j=1}^{n} Y_j(0) \ .$$

This gives rise to the two-sample Kolmogorov–Smirnov statistic:

$$K_{m,n,\hat{\delta}}(Z) = \sup_y \left| V_{m,n}(y, \hat{\delta}) \right| \ , \tag{4}$$

where

$$V_{m,n}(y, \hat{\delta}) = \sqrt{\frac{mn}{N}} \left(\hat{F}_1(y + \hat{\delta}) - \hat{F}_0(y) \right) \tag{5}$$

is the two-sample empirical process. We may equivalently consider the the following transformation of the two-sample empirical process via the change of variable $y \mapsto F_0^{-1}(t)$ and work with

$$\begin{aligned} \upsilon_{m,n}(t, \hat{\delta}) &= \sqrt{\frac{mn}{N}} \left(\frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\{Y_i(1) - \hat{\delta} \leq F_0^{-1}(t)\}} - \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{Y_i(0) \leq F_0^{-1}(t)\}} \right) \\ &= \sqrt{\frac{mn}{N}} \left(\hat{F}_1(F_0^{-1}(t) + \hat{\delta}) - \hat{F}_0(F_0^{-1}(t)) \right) \\ &= V_{m,n}(F_0^{-1}(t), \hat{\delta}) \ . \end{aligned} \tag{6}$$

2.2 Permutation Test under the Sharp Null

We begin the study of the properties of the permutation test in the case when δ is *known* as a stepping stone to the more challenging case with estimated $\hat{\delta}$. This case corresponds with the sharp null, and we are going to refer to it as *classical*.[6]

[6]For a more thorough appraisal of the sharp null hypothesis in connection with the permutation tests, see Rosenbaum (2002); Caughey et al. (2017).

In the classical case with δ known, we are able to determine all potential outcomes as well as the exact null distribution. As a result, the permutation test is exact level α for a fixed sample size—the null hypothesis can be tested by comparing the observed statistic with the its distribution calculated by considering alternative permutations of the treatment.

To see why this construction works, let us introduce further notation. First, note that if δ were known, we could recenter the observations from the treatment group by δ. More specifically, let $Z^* = (Z_1^*, \ldots, Z_N^*)$ be given by

$$Z^* = \Big(Y_1(1) - \delta, \ldots, Y_m(1) - \delta, Y_1(0), \ldots, Y_n(0)\Big) \,, \tag{7}$$

and consider the classical two-sample Kolmogorov–Smirnov statistic:

$$K_{m,n,\delta}(Z^*) = \sup_y |V_{m,n}(y,\delta)| \tag{8}$$

where

$$V_{m,n}(y,\delta) = \sqrt{\frac{mn}{N}} \left(\frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\{Z_i^* \le y\}} - \frac{1}{n} \sum_{i=m+1}^{N} \mathbb{1}_{\{Z_i^* \le y\}} \right) \tag{9}$$

is the two-sample classical empirical process. Denote $\mathbf{G}_N$ as the set of all permutations π of $\{1, \ldots, N\}$, with $|\mathbf{G}_N| = N!$. Given $Z^* = z$, recompute $K_{m,n,\delta}(z)$ for all permutations $\pi \in \mathbf{G}_N$ and denote by

$$K_{m,n,\delta}^{(1)}(z) \le K_{m,n,\delta}^{(2)}(z) \le \cdots \le K_{m,n,\delta}^{(N!)}(z) \,,$$

the ordered values of $\{K_{m,n,\delta}(z_\pi) : \pi \in \mathbf{G}_N\}$, where z_π denotes the action of $\pi \in \mathbf{G}_N$ on $z \in \mathbb{R}$. Let $k = N! - \lfloor N!\,\alpha \rfloor$ and define

$$\begin{aligned} M^+(z) &= \Big| \{1 \le j \le N! : K_{m,n,\delta}^{(j)}(z) > K_{m,n,\delta}^{(k)}(z)\} \Big| \\ M^0(z) &= \Big| \{1 \le j \le N! : K_{m,n,\delta}^{(j)}(z) = K_{m,n,\delta}^{(k)}(z)\} \Big| \,. \end{aligned}$$

Using this notation, the permutation test is given by

$$\phi(z) = \begin{cases} 1 & K_{m,n,\delta}(z) > K_{m,n,\delta}^{(k)}(z) \\ a(z) & K_{m,n,\delta}(z) = K_{m,n,\delta}^{(k)}(z) \\ 0 & K_{m,n,\delta}(z) < K_{m,n,\delta}^{(k)}(z) \end{cases} \,, \tag{10}$$

where

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

Observe that for every $\pi \in \mathbf{G}_N$, the joint distribution of $(Z_1^*, \ldots, Z_N^*)$ is the same as $(Z_{\pi(1)}^*, \ldots, Z_{\pi(N)}^*)$ under the null hypothesis (2). This invariance property under the null hypothesis, the so-called randomization hypothesis, guarantees the finite-sample validity of the

permutation test. More formally, the permutation test (10) for the sharp null hypothesis satisfies

$$\mathbb{E}[\phi(z)] = \alpha\,, \quad \text{for any} \quad \alpha \in (0,1) \tag{11}$$

under the null hypothesis (Theorem 15.2.1, Lehmann and Romano, 2005). In other words, the true false-rejection probability of the permutation test is exactly equal to significance level α under the sharp null when δ is known.

Remark 3. Consider the same construction of the permutation test but replacing $K_{m,n,\delta}$ with

$$K^u_{m,n,\delta}(Z^*) = \sup_{0\le t\le 1} |\upsilon_{m,n}(t,\delta)|\,, \tag{12}$$

where $\upsilon_{m,n}(t,\delta) = V_{m,n}(F_0^{-1}(t),\delta)$. A remarkable feature of the permutation test is that they are level α test tests for *any* test statistic, as long as the randomization hypothesis holds. As a result, the finite-sample exactness of the permutation test under the sharp null still holds if we consider (12) instead. ■

Remark 4. Permutation inference requires recalculating the test statistic as π varies in $\mathbf{G}_N$. It often is the case in practice that $\mathbf{G}_N$ is too large ($N!$), which makes the calculation of the permutation test computationally expensive. In such cases, we can restore to a stochastic approximation without affecting the finte-sample validity of the test. Let $\hat{\mathbf{G}} = \{g_1, \ldots, g_B\}$ where g_1 is the identity permutation and $g_2, \ldots, g_B$ are i.i.d. uniform on $\mathbf{G}_N$. The test may again be used by replacing $\mathbf{G}_N$ with $\hat{\mathbf{G}}$, and this approximation can be made arbitrarily close for B sufficiently large (Section 4 Romano, 1989). Consequently, we will focus solely on $\mathbf{G}_N$ while keeping in mind that in practice we will resort to $\hat{\mathbf{G}}$. ■

2.3 Challenges for a Permutation Test with estimated δ

What happens to the permutation test if we replace δ by the sample estimate $\hat{\delta}$? The permutation test based on (4) differs from the classical case in several important ways. First, the finite-sample results are compromised since we do not know δ and therefore we cannot guarantee that the randomization hypothesis holds when δ is replaced with estimated $\hat{\delta}$. Second, while the permutation test in the classical case is also asymptotically valid , as we show in Theorems A.1 and A.2 in Appendix A, this is not the case when δ is unknown and needs to be estimated. Intuitively, the necessity of estimating δ introduces an additional component to the limit distribution of $V_{m,n}(\cdot,\hat{\delta})$, which no longer is the simple Brownian bridge as in the classical case. Instead, we now obtain a Gaussian process with covariance structure that depends on the particulars of the data generating process.

To formalize the ongoing discussion, we introduce further notation. Denote $\mathbb{G}$ the F_0-Brownian bridge, and let $\mathbb{S}$ be a Gaussian process with mean zero and covariance structure

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

where $\sigma_0^2 = \mathbb{V}(Y_i(0))$. Consider the process the process $\mathbb{B} = \mathbb{G} + \mathbb{S}$ with covariance structure

$$\mathbb{C}(\mathbb{G}(x), \mathbb{S}(y)) = f_0(y)F_0(x)\left(1 - F_0(x)\right)\left\{\mathbb{E}(Y(0)|Y(0) \leq x) - \mathbb{E}(Y(0)|Y(0) > x)\right\} \ . \tag{13}$$

The following theorem establishes the asymptotic behavior of the two-sample Kolmogorov–Smirnov statistic. It is due to Theorem 4 of Ding et al. (2015) for a suitably scaled variation of their test statistic, but we include here for completeness.

Theorem 1. *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. according to a probability distribution F_1. Consider testing the hypothesis (2) for some unknown δ based on the test statistic (4). Under conditions A.1–A.2, $K_{m,n,\hat{\delta}}$ converges weakly under the null hypothesis to*

$$K_1 \equiv \sup_y |\mathbb{B}(y)| \tag{14}$$

where $\mathbb{B}$ is given by $\mathbb{B} = \mathbb{G} + \mathbb{S}$, and whose marginal distributions are zero-mean normal with covariance structure (13).

The preceding theorem illustrates what Koenker and Xiao (2002) dub as the Durbin problem—the complexity arising from the estimated nuisance parameter, rendering the asymptotic null distribution intractable. The practical consequence of this complexity is to make it difficult, if not impossible, to obtain critical values.

Remark 5. We now illustrate the effect of the estimated nuisance parameter on the limiting distribution. As we show in the proof of Theorem 1 in the Appendix A, the smoothness condition A. 2 allows us to expand $V_{m,n}(y, \hat{\delta})$ around δ to obtain

$$V_{m,n}(y, \hat{\delta}) = \underbrace{V_{m,n}(y, \delta)}_{\xrightarrow{\text{d}} \mathbb{G}(y)} + \underbrace{\sqrt{\frac{mn}{N}}\left(f_0(y)(\hat{\delta} - \delta)\right)}_{\xrightarrow{\text{d}} \mathbb{S}(y)} + o_p(1) \ .$$

Observe that the first summand is the classical two-sample Kolmogorov–Smirnov statistic (5), whose weak limit distribution is the Brownian bridge $\mathbb{G}$ (see Theorem A.1 in Appendix A). However, the asymptotically distribution-free property of the classical two-sample Kolmogorov–Smirnov statistic is jeopardized due to the introduction of the drift $\mathbb{S}$—we now obtain a more complicated Gaussian process $\mathbb{B}$ whose covariance structure depends on the underlying data generating process, rather than $\mathbb{G}$. ■

Before formally stating the asymptotic properties of the permutation test based on $K_{m,n,\hat{\delta}}$, it might be helpful to consider an alternative description of the permutation test. More specifically, the permutation test rejects the null hypothesis (2) if $K_{m,n,\hat{\delta}}(z)$ exceeds the $1-\alpha$ quantile of the permutation distribution:

$$\hat{R}_{m,n}^{K(\hat{\delta})}(t) = \frac{1}{N!}\sum_{\pi \in \mathbf{G}_N} \mathbb{1}_{\{K_{m,n,\hat{\delta}}(z_{\pi(1)}, \ldots, z_{\pi(N)}) \leq t\}} \ . \tag{15}$$

The permutation distribution can be seen as the conditional distribution of $K_{m,n,\hat{\delta}}(Z_\pi)$ given Z, where π a random permutation uniformly distributed over $\mathbf{G}_N$. This is so because, conditionally on Z, $K_{m,n,\hat{\delta}}(Z_\pi)$ is equally likely to be any of $K_{m,n,\hat{\delta}}(Z_\pi)$ among $\pi \in \mathbf{G}_N$.

One can deduce from Theorem 1 that $K_{m,n,\hat{\delta}}$ is not asymptotically pivotal, thus the corresponding permutation test fails to control the Type 1 error even asymptotically. This is an immediate consequence of the fact that the permutation distribution based on the two-sample Kolmogorov–Smirnov statistic, $K_{m,n,\hat{\delta}}$, does not behave like the true unconditional limiting distribution asymptotically, as shown in the following theorem.

Theorem 2. *Assume the premises of Theorem 1. Then the permutation distribution* (15) *satisfies*

$$\sup_y \left| \hat{R}^{K(\hat{\delta})}_{m,n}(y) - J_0(y) \right| \xrightarrow{\text{p}} 0 \ ,$$

where $J_0(\cdot)$ *is the CDF of* $K_0 \equiv \sup_y |\mathbb{G}(y)|$.

This discrepancy between the permutation distribution and the true unconditional limiting sampling distribution breaks the asymptotic validity of the permutation test for testing constant the treatment effect—the limiting rejection probability tends to a value different than the nominal level α. As a result, one may have underrejection or overrejection under H_0, with the latter being more problematic. We confirm this phenomenon in the simulation studies presented in Section 5.

Remark 6. As a matter of fact, the permutation distribution based on either $K_{m,n,\hat{\delta}}$ or $K_{m,n,\delta}$ is mimicking the unconditional limiting distribution as if δ were known, completely ignoring the complexity introduced by the estimation error.[7] Intuitively, this resemblance occurs because in both cases, the permutation distribution is treating the observations as if they were i.i.d. More concretely, the construction that leads to the two-sample Kolmogorov–Smirnov statistic shifts the observations from the treatment group as in display (7), with δ replaced by $\hat{\delta}$. ■

3 Valid Permutation Test

Section 2.3 concludes that the introduction of the drift term $\mathbb{S}$ in $\mathbb{B}$ implies that the limiting behavior of the statistic based on the empirical process (4) is no longer asymptotically distribution-free. A direct consequence of this is that it invalidates permutation inference. To address this issue, Khmaladze (1981) employs a Doob–Meyer decomposition of the uniform empirical process in order to restore the asymptotically distribution-free nature of the Kolmogorov–Smirnov statistic in the one-sample case. This section extends Khmaladze's result to the two-sample case and presents the asymptotically valid permutation test based on the martingale-transformed statistic.

[7] See Theorems A.1–2 in Appendix for asymptotic results when δ is known.

3.1 Martingale Transformation

We briefly review relevant concepts from Khmaladze (1981) that will be important for our main result. We begin by introducing further notation. Define the function $g(s) = (g_1(s), g_2(s)) = (s, f_0(F_0^{-1}(s)))'$ on $[0,1]$, and $\dot{g}(s) = (\dot{g}_1(s), \dot{g}_2(s))$ so that $\dot{g}$ is the derivative of g. Therefore $\dot{g}(s) = (1, \dot{f}_0(F_0^{-1}(s))/f(F_0^{-1}(s)))$. Function g previously defined is closely connected with the score function. As a matter of fact, it can be shown that g is the integrated score function of the model (see remarks after assumption A2 in Bai (2003) and Section 4 in Parker (2013)).

Let $D[0,1]$ be the space of càdlàg functions on $[0,1]$, and denote by $\psi_g(h)(\cdot)$ the compensator of h, $\psi_g : D[0,1] \to D[0,1]$ given by

$$\psi_g(h)(t) = \int_0^t \left[\dot{g}(s)' C^{-1}(s) \int_s^1 \dot{g}(r) dh(r) \right] ds \tag{16}$$

where $C(s) = \int_s^1 \dot{g}(t)\dot{g}(t)' dt$. Arguing as in Parker (2013), we can think of equation (16) as the functional equivalent of the fitted values in a linear regression, where the extended score $\dot{g}(s)$ acts as the regressor, and $C^{-1}(s) \int_s^1 \dot{g}(r) dh(r)$ as the OLS estimator. This is the insight behind the numerical calculation of the compensator in Section 3.3.

Remark 7. Existence of $C(s)^{-1}$ for all $s < 1$ follows by Assumption A.2 (see also Theorem 3.3, Khmaladze, 1981). To see why, observe that Assumption A.2 implies that *(i)* the functions $\dot{g}_1(s)$ and $\dot{g}_2(s)$ belong to $L_2[0,1]$, the equivalence class of square-integrable functions on $[0,1]$, and *(ii)* the functions $\dot{g}_1(s)$ and $\dot{g}_2(s)$ are linearly independent in the neighborhood of $s = 1$. As a result, $\dot{g}_1$ and $\dot{g}_2$ form an orthonormal system of functions in $L_2[0,1]$, which ensures the transformation of the uniform empirical process into an innovation martingale (see remarks that follow after Theorem 3.2, Khmaladze, 1981). ■

The Khmaladze transformation of the two-sample empirical process (6) is given by

$$\begin{aligned} \tilde{\upsilon}_{m,n}(t, \hat{\delta}) &= \upsilon_{m,n}(t, \hat{\delta}) - \int_0^t \left[\dot{g}(s)' C^{-1}(s) \int_s^1 \dot{g}(r) d\upsilon_{m,n}(r, \hat{\delta}) \right] ds \\ &= \upsilon_{m,n}(t, \hat{\delta}) - \psi_g(\upsilon_{m,n}(t, \hat{\delta})) \ . \end{aligned} \tag{17}$$

The two-sample martingale-transformed version of the two-sample Kolmogorov–Smirnov statistics is

$$\tilde{K}_{m,n,\hat{\delta}}(Z) = \sup_{0 \le t \le 1} \left| \tilde{\upsilon}_{m,n}(t, \hat{\delta}) \right| \ . \tag{18}$$

The martingale-transformed statistic (18) is asymptotically pivotal and this is the key input for the asymptotic validity of the permutation test. The asymptotic behavior of the permutation distribution is obtained in the next Section.

3.2 Main Results

We now turn to our main theoretical result—the permutation test based on the martingale-transformed statistic behaves asymptotically like the true unconditional limiting sampling distribution. We break this result down into two pieces. First, we establish the limit behavior of (18), and then we show the asymptotic behavior of the proposed permutation test.

The following theorem states the limit behavior of (18). It essentially follows from an extension of Khmaladze (1981) to the two-sample case, where we show that $\tilde{\upsilon}_{m,n}(\cdot, \hat{\delta})$ converges weakly to a Brownian motion process $\mathbb{M}$, effectively nullifying the effect of the estimated nuisance parameter $\hat{\delta}$.

Theorem 3. *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. according to a probability distribution F_1. Consider testing the hypothesis (2) for some δ based on the test statistic (18). Under conditions A.1–A.2, $\tilde{K}_{m,n,\hat{\delta}}$ converges weakly under the null hypothesis to*

$$K_2 \equiv \sup_{0 \leq t \leq 1} | \, \mathbb{M}(t) \, | \tag{19}$$

where $\mathbb{M} = \mathbb{U} - \psi_g(\mathbb{U})$ is the standard Brownian motion, and $\mathbb{U}$ is the standard (uniform) Brownian bridge on $[0, 1]$.

To gain further intuition as to why this transformation works, observe that the mapping $\psi_g(h)(\cdot)$, the so-called compensator of h (Khmaladze, 1981), is a linear mapping with respect to h, and satisfies $\psi_g(cg) = cg$ for a constant or random variable c (Bai, 2003). These properties combined with Remark 5, allow us to write (17) as

$$\tilde{\upsilon}_{m,n}(t, \hat{\delta}) = \upsilon_{m,n}(t, \hat{\delta}) - \psi_g(\upsilon_{m,n}(t, \hat{\delta})) = \underbrace{\upsilon_{m,n}(t, \delta)}_{\xrightarrow{\mathrm{d}} \mathbb{U}(t)} - \underbrace{\psi_g(\upsilon_{m,n}(t, \delta))}_{\xrightarrow{\mathrm{d}} \psi(\mathbb{U})(t)} + o_p(1) \ .$$

From here it is easy to see that we may express the martingale-transformed two-sample empirical process as if δ were known, plus some term that is asymptotically negligible. This implies that the limit distribution is asymptotically distribution-free (see the proof of Theorem 3 in Appendix A for more details).

The following theorem shows that the proposed test is asymptotically valid *i.e.* the permutation distribution based on the martingale-transformed version of the two-sample Kolmogorov–Smirnov statistic behaves like the true unconditional limiting distribution of $\tilde{K}_{m,n,\hat{\delta}}$.

Theorem 4. *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. according to a probability distribution F_1. Consider testing the hypothesis (2) for some δ based on the test statistic (18). Under conditions A.1–A.2,*

the permutation distribution (15) *based on the Khmaladze transformed statistic* $\tilde{K}_{m,n,\hat{\delta}}$ *is such that*

$$\sup_{0\le t\le 1}\left|\hat{R}^{\tilde{K}(\hat{\delta})}_{m,n}(t) - J_2(t)\right| \xrightarrow{\text{p}} 0 \ ,$$

where $J_2(\cdot)$ *is the CDF of* K_2 *defined in equation* (19).

Thus the permutation distribution behaves asymptotically like the true unconditional limiting distribution. The relevance of Theorem 4 is that it asymptotically justifies the use of the proposed permutation test for testing the null hypothesis of constant treatment effects.

Remark 8. There is no loss in power in using permutation critical values. To see why, let $r_{m,n}$ be the $1-\alpha$ quantile of the distribution of $\tilde{K}_{m,n,\hat{\delta}}$. Typically the Kolmogorov–Smirnov test rejects when $\tilde{K}_{m,n,\hat{\delta}} > r_{m,n}$, where $r_{m,n}$ is nonrandom. We have that $r_{m,n} \to K_2^{-1}(1-\alpha)$. Assume that $\tilde{K}_{m,n,\hat{\delta}}$ weakly converges to some limit law $K_2'(\cdot)$ under some sequence of alternatives that are contiguous to some distribution satisfying the null hypothesis. Then the power of the test would tend to $1-K_2'(K_2^{-1}(1-\alpha))$. The premises of Theorems 3 and 4 imply that the permutation test based on a random critical value $\tilde{r}_{m,n}$, obtained from the permutation distribution, satisfies $\tilde{r}_{m,n} \xrightarrow{\text{p}} K_2^{-1}(1-\alpha)$. The same result follows under a sequence of contiguous alternatives, thus implying that the permutation test has the same limiting local power as the Kolmogorov–Smirnov test which uses nonrandom critical values. ■

In the next two subsections, we illustrate the mechanics behind the Khmaladze transformation, as well as the numerical calculation of it.

3.3 Khmaladze Transformation as a Continuous-time Detrending Operation

To gain further insight as to how the transformation works, we follow Bai (2003) and Parker (2013), and we consider (17) 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 (17) in differentiation form

$$d\tilde{\upsilon}_{m,n}(t,\hat{\delta}) = d\upsilon_{m,n}(t,\hat{\delta}) - \dot{g}(t)'C^{-1}(t)\int_t^1 \dot{g}(r)d\upsilon_{m,n}(r,\hat{\delta})dt \ . \tag{20}$$

Define $dt_i = t_{i+1} - t_i$, and let

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

then the right hand side of (20) can be interpreted as the recursive residuals:

$$y_i - x_i' \left(\sum_{k=i}^{m+1} x_k x_k' \right)^{-1} \sum_{k=i}^{m+1} x_k y_k = y_i - x_i' \hat{\boldsymbol{\beta}}_i \tag{21}$$

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.

3.4 Numerical Computation of the Khmaladze Transformation

In order to facilitate the numerical calculation of our test, we develop the `R` package `RATest` (Olivares and Sarmiento, 2017). For completeness, we now show how the `RATest` package calculates the compensator, as well as the martingale-transformed version of the two-sample Kolmogorov–Smirnov statistic in practice.

The computation of the compensator involves numerical integration. Therefore, we 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

$$\mathbf{X}_i = \sqrt{\frac{1}{m}} \begin{pmatrix} \dot{g}_1(t_{m+1}) & \dot{g}_2(t_{m+1}) \\ \dot{g}_1(t_m) & \dot{g}_2(t_m) \\ \vdots & \vdots \\ \dot{g}_1(t_i) & \dot{g}_2(t_i) \end{pmatrix}, \quad \mathbf{y}_i = \sqrt{m} \begin{pmatrix} \upsilon_{m,n}(t_{m+1}, \hat{\delta}) & - & \upsilon_{m,n}(t_m, \hat{\delta}) \\ \upsilon_{m,n}(t_m, \hat{\delta}) & - & \upsilon_{m,n}(t_{m-1}, \hat{\delta}) \\ & \vdots & \\ \upsilon_{m,n}(t_i, \hat{\delta}) & - & \upsilon_{m,n}(t_{i-1}, \hat{\delta}) \end{pmatrix},$$

where $\dot{g}_1(s) = 1$ and $\dot{g}_2(s) = \dot{f}_0(F_0^{-1}(s))/f(F_0^{-1}(s))$. The OLS estimator based on the last $m - i + 2$ observations described on right hand side of (21) 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 (17) 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

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

Observe that the computation of the compensator relies on the true density and score functions. Following Bai (2003) and Koenker and Xiao (2002), we assume that $g_2(s)$ and $\dot{g}_2(s)$ can be replaced by an estimators $g_{2,n}(s)$ and $\dot{g}_{2,n}(s)$, respectively, such that

$$\sup_{0\leq t\leq 1}|g_{2,n}(t)-g_2(t)|=o_p(1) \quad \text{and} \tag{22}$$

$$\sup_{0\leq t\leq 1}|\dot{g}_{2,n}(t)-\dot{g}_2(t)|=o_p(1) . \tag{23}$$

The conclusions of Theorems 3 and 4 follow if we replace $g_2(s)$ and $\dot{g}_2(s)$ by the uniformly consistent estimators $g_{2,n}(s)$ and $\dot{g}_{2,n}(s)$ by the same arguments as in (Bai, 2003, Theorem 4).

Remark 9. The implementation in `RATest` estimates both functions using the univariate adaptive kernel density estimation *á la Silverman* (e.g. Portnoy and Koenker, 1989; Koenker and Xiao, 2002), which satisfies the uniform requirements in (22)–(23) (Portnoy and Koenker, 1989, Lemma 3.2). ■

4 Within-group Treatment Effect Heterogeneity

One conventional approach to investigating the potential heterogeneity in the treatment effect involves estimating average treatment effects for subgroups defined by observable covariates, such as demographic or pre-intervention characteristics. The underlying modeling assumption of this approach treats mean impacts constant within subgroups while allowing them to vary across subgroups[8]. Then, one may characterize treatment effect heterogeneity by testing whether the existing differences vary significantly across subgroups.

The martingale transformed permutation test proposed here can be implemented to test whether there exists within-group treatment effect heterogeneity. In essence, we propose a test method for jointly testing the null hypotheses that treatment effects are constant *within* mutually exclusive subgroups while allowing them to be different *across* subgroups.

To formalize the ongoing discussion, let us introduce further notation. Throughout we will assume that the mutually exclusive subgroups are formed from observed covariates, and are

[8]Notwithstanding the simplicity of this approach, it has been shown that it fails to describe the heterogeneity in the treatment effect in some empirical examples, where it performs poorly relatively to other methods such as quantile treatment effects models. This point is well developed and documented in Bitler et al. (2017), where they analyze the effects of the Connecticut's Jobs First welfare reform on earnings.

taken as given. Denote $\mathcal{J}$ the total number of such subgroups. Let $F_0^j(y)$ and $F_1^j(y)$ be the CDFs of the control and treatment group for subgroup $1 \leq j \leq \mathcal{J}$. The null hypothesis of interest is given by the joint hypothesis

$$\mathbf{H}_0 : F_1^j(y + \delta_j) = F_0^j(y) \ , \text{ for all mutually exclusive } j \in \{1, \ldots, \mathcal{J}\} \text{, for some } \delta_j \ . \tag{24}$$

This section treats the testing problem (24) as a multiple testing problem in which every individual hypothesis $j \in \{1, \ldots, \mathcal{J}\}$, given by

$$H_{0,j} : F_1^j(y + \delta_j) = F_0^j(y) \ , \text{ for some } \delta_j \ , \tag{25}$$

specifies whether the treatment effect is heterogeneous for a particular subgroup[9].

In order to achieve control of the family-wise error rate (FWER), we propose a stepwise multiple testing procedure based on the Westfall–Young algorithm (Westfall and Young, 1993). Similar adjustments for multiple testing are also available, but we opt for the Westfall–Young due to its asymptotic optimality properties (Meinshausen et al., 2011), and its ability to incorporate the dependence structure of the individual tests.[10]

If each individual test can be summarized by a p-value, the following min p algorithm yields adjusted p-values that allow us to control the test FWER level (see Westfall and Young, 1993, Chapter 2). Observed data for each mutually exclusive subgroup is given by

$$Z_j = \left(Y_{j_1}(1), \ldots, Y_{j_{m_j}}(1), Y_{j_1}(0), \ldots, Y_{j_{n_j}}(0)\right), \quad \text{for all} \quad 1 \leq j \leq \mathcal{J} \ ,$$

where every subgroup Z_j, $1 \leq j \leq \mathcal{J}$ has $m_j + n_j$ elements such that $\sum_j n_j = n$ and $\sum_j m_j = m$. Denote $p_1, \ldots, p_{\mathcal{J}}$ the p-values of the $\mathcal{J}$ individual permutation tests for (25), and the ordered p-values $p_{r_1} \leq \cdots \leq p_{r_{\mathcal{J}}}$, with their respective associated hypotheses of the form (25) given by $H_{0,r_1}, \ldots, H_{0,r_{\mathcal{J}}}$. Define $\mathcal{T}_j = \{r_j, r_{j+1} \ldots, r_{\mathcal{J}}\}$ and let $g_{b,j}$ for $1 \leq j \leq \mathcal{J}$ be a random permutation of $\{1, \ldots, m_j + n_j\}$.

Algorithm 1 (Westfall–Young)

1. *For each permutation* $b = 1, \ldots, B < \min_{1 \leq j \leq \mathcal{J}}\{(m_j + n_j)!\}$*:*

   (i) Apply action $g_{b,j}$ *to every subgroup* Z_j, $1 \leq j \leq \mathcal{J}$*:* $(g_{b,1}Z_1, \ldots, g_{b,\mathcal{J}}Z_{\mathcal{J}})$*, with corresponding p-values* $p_j^{(b)}$ *for* $1 \leq j \leq \mathcal{J}$*.*

   (ii) Let

$$\tilde{p}_{r_1}^{(b)} = \min_{j \in \mathcal{T}_1} p_j^{(b)} \ , \ \tilde{p}_{r_2}^{(b)} = \min_{j \in \mathcal{T}_2} p_j^{(b)} \ , \ \ldots \ , \ \tilde{p}_{r_{\mathcal{J}}}^{(b)} = p_{r_{\mathcal{J}}}^{(b)} \ .$$

[9] Naively testing for treatment effect variation for each subgroup at level α may lead us to flawed inference though. With such a procedure the probability of one or more false rejections rapidly increases with the number of subgroups. To put it in other words, the probability of falsely claiming that the treatment effect is heterogeneous for some subgroup may be greater than α.

[10] We include an alternative procedure based on Holm (1979). See Appendix D for more details.

2. *Define*

$$\mathcal{L}_1 = \#\{p_{r_1} \geq \tilde{p}_{r_1}^{(b)} : 1 \leq b \leq B\} \ , \ \ldots \ , \ \mathcal{L}_{\mathcal{J}} = \#\{p_{r_{\mathcal{J}}} \geq \tilde{p}_{r_{\mathcal{J}}}^{(b)} : 1 \leq b \leq B\} \ .$$

3. *The adjusted p-values are given by*

$$p_{r_1}^* = \frac{\mathcal{L}_1}{B} \ , \ p_{r_2}^* = \max\left\{p_{r_1}^*, \frac{\mathcal{L}_2}{B}\right\} \ , \ \ldots \ , \ p_{r_{\mathcal{J}}}^* = \max\left\{p_{r_{\mathcal{J}-1}}^*, \frac{\mathcal{L}_{\mathcal{J}}}{B}\right\} \ .$$

In order to control the test FWER at level α, each adjusted p-value $p_{r_j}^*$, with associated hypothesis H_{0,r_j}, needs to be now compared with α, for $1 \leq j \leq \mathcal{J}$. Moreover, we reject the null hypothesis $\mathbf{H}_0$ if any one null hypothesis for a subgroup $j \in \{1, \ldots, \mathcal{J}\}$ is rejected.

Remark 10. A noteworthy byproduct of the testing problem we describe in the joint null hypothesis $\mathbf{H}_0$ is that we can also declare **for which subgroups**, if any, there is heterogeneity in the treatment effect. This is an immediate consequence of the step-down procedure we present since we can now determine which hypothesis H_{0,r_j} is rejected. Investigating which subgroups respond differentially to the treatment effect might be of particular interest, e.g. when deciding whether to scale the experiment up. ■

Remark 11. One of the main drawbacks of the min p method is that it is computationally intensive since the adjusted p-values arise from two levels of permutations—one from the permutation test, and one from the adjustment method itself. For this matter, we also consider two alternative procedures—the max T (Algorithm 2) and Holm (Algorithm 3) procedures—which control the family-wise error rate without incurring in such computational cost. See Appendix D for details. ■

Remark 12. Multiple testing approaches to treatment effect heterogeneity are also addressed in Lee and Shaikh (2014); List et al. (2016); Bitler et al. (2017). Our approach differs from theirs in several important ways. First, the handling of an estimated nuisance parameter. Neither Lee and Shaikh (2014) nor List et al. (2016) conduct inference based on empirical processes with estimated nuisance parameters, and while Bitler et al. (2017) mention that their method is valid in the presence of estimated nuisance parameters, their theoretical arguments are fundamentally different than ours—their approach is based on constructing what they call the "simulated-outcomes distribution." Second, we propose a stepwise multiple testing procedure based on the Westfall–Young adjustment. Lee and Shaikh (2014) and List et al. (2016) exploit similar yet different stepwise procedures (Romano and Wolf, 2005, 2010, respectively), and Bitler et al. (2017) adjustment is more conservative for they use Bonferroni bounds. Lastly, the choice of the statistic. Our approach is based on the two-sample, martingale-transformed empirical process. On the other hand, Lee and Shaikh (2014) and List et al. (2016) work with a statistic based on the p-values that arise from an underlying "difference-in-means" statistic. Meanwhile, Bitler et al. (2017) test for equality of distributions between their simulated outcomes and the actual observed data. ■

5 Monte Carlo Simulations

We present several Monte Carlo experiments to examine the finite sample performance of the proposed test in comparison to other methods. We adhere to the design 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). For $1 \leq i \leq N$, potential outcomes in the simulation study are generated according to the relationship

$$\begin{aligned} Y_i(0) &= \varepsilon_i, \;\; \delta_i = \delta + \sigma_\delta Y_i(0) \\ Y_i(1) &= \delta_i + Y_i(0) \;, \end{aligned}$$

where σ_δ denotes the different levels of heterogeneity, and $\sigma_\delta = 0$ induces a constant treatment effect. 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. In each of the following specifications ε_i, $1 \leq i \leq N$ are i.i.d. according to one of the following probability distributions: standard normal, lognormal, Student's t distribution with 5 degrees of freedom, and $N \in \{13, 50, 80, 200\}$. Rejection probabilities are computed using 5000 replications across Monte Carlo Experiments.

In the simulation results presented in Tables 1 and 2, we compare the proposed permutation test based on the martingale-transformed two-sample Kolmogorov–Smirnov statistic (denoted **mtPermTest**), which we calculate using the `R` package `RATest`, and the following five alternative tests:

Classic KM: This test is the permutation test based on the classical two-sample Kolmogorov–Smirnov statistic of Section 2.2. We present this test to serve as a benchmark of the ideal scenario.

Naive KS: This is the permutation test based on the two-sample Kolmogorov–Smirnov statistic of Section 2.3. We call it naive because it ignores the effect that the estimated nuisance parameter has on the limiting distribution.

FRT CI: This test is the Fisher's randomization test confidence interval method of Ding et al. (2015). Their approach finds the maximum p-value over a $(1-\gamma)$-level confidence interval for δ, CI_γ

$$p_\gamma = \sup_{\delta' \in CI_\gamma} p(\delta') + \gamma \;,$$

where $p(\delta')$ is obtained by performing the permutation test under the sharp null hypothesis (1). Following their numerical study, we take $\gamma = 0.01$.

Subsampling: This test is proposed by Chernozhukov and Fernández-Val (2005). It is based on subsampling the appropriately recentered empirical quantile regression process

$$\sup_{\tau \in \mathcal{T} \subset [0,1]} \left| \hat{\delta}(\tau) - \hat{\delta} \right| \;,$$

where $\hat{\delta}(\tau)$ is an estimator of $\delta(\tau)$ in (3) given by $\hat{\delta}(\tau) = \hat{F}_1^{-1}(\tau) - \hat{F}_0^{-1}(\tau)$, $\hat{F}^{-1} = \inf\{y : \hat{F}(y) \geq \tau\}$, and $\hat{F}$ is the empirical CDF. We use subsampling block size $b = 20 + N^{1/4}$ (see Section 3.4 in Chernozhukov and Fernández-Val, 2005).

Bootstrap: This test is proposed by (Linton et al., 2005, Section 6) and Chernozhukov and Fernández-Val (2005). It is based on the full-sample bootstrap approximation of the sampling distribution of the two-sample Kolmogorov–Smirnov statistic (4). Arguing as in Ding et al. (2015), we recenter treatment and control groups, and sample with replacement from the pooled vector of residuals.

Table 1 reports rejection probabilities under the null hypothesis of constant treatment effect ($\sigma_\delta = 0$)[11]. As a benchmark, it also reports the rejection probabilities of the classical Kolmogorov–Smirnov test, taking δ as given. As expected in the light of Theorem 2, we see that this permutation test in the classical case has rejection probabilities under the null hypothesis very close to the nominal level for all specifications and sample sizes we consider in the numerical experiments. These conclusions, however, do not carry over into the naive case when δ is unknown. When δ is unknwon and therefore becomes a nuisance parameter, the permutation test applied to the two-sample Kolmogorov–Smirnov statistic may under-reject (e.g., normal and t distributions) or over-reject (e.g. lognormal distribution) under the null hypothesis, which illustrates the complexity arising from the estimated nuisance parameter, and the challenges for permutation inference in this scenario.

Our proposed test performs fairly well across specifications. Interestingly, even though the density and score functions are estimated non-parametrically with considerably small sample sizes, the rejection probabilities only exceed the nominal level once (5.9%), though it is frequently much less than the nominal level (e.g. $N = 13$, or $\varepsilon \sim \mathcal{N}(0, 1)$).

FRT CI yields severely conservative rejection probabilities in all specifications considered here, especially for small sample sizes ($N \leq 50$). This feature seems to disappear as sample size increases, a situation when its rejection probability under the null hypothesis is comparable to our proposed test (up to simulation error). Subsampling delivers rejection probabilities under the null hypothesis less than the nominal level in all specifications although it is hyper-conservative. Finally, the bootstrap approach over-rejects severely for the symmetric normal and t distributions.

Table 2 reports the rejection probabilities for several levels of heterogeneity σ_δ and $\delta = 1$. In here, we only consider our proposed test, the FRT CI, and subsampling, leaving the other tests out due to their infeasibility (classical KS) or their inability to control rejection probabilities under the null for some specifications (naive KS and Bootstrap). In virtually all specifications, our proposed test has the highest rejection probabilities under the alternative hypothesis ($\sigma_\delta > 0$). This difference in power is more pronounced in situations when sample sizes are relatively

[11] Simulation results using the true density and score functions are similar in magnitude and therefore not shown in here, though available upon request.

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

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

Rejection probability for the six tests defined in the text, for three different data generating processes, and four different sample sizes.

small. FRT CI appears to be generally less powerful than our proposed test, though it delivers much greater rejection rates than subsampling, which has the lowest rejection probability under the alternative among the three methods considered here.

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

Rejection probability for the six tests defined in the text, for three different data generating processes, and four different sample sizes..

6 Empirical Application

We briefly revisit an experiment by Gneezy and List (2006), also considered in Goldman and Kaplan (2018), on the effects of gift exchange on worker effort, the so-called *gift exchange hypothesis*. The underlying idea behind this model is the assumption that there exists a positive relationship between wages and worker effort levels. Under this hypothesis, equilibrium unemployment arises as a result of workers putting more effort when paid above their opportunity cost, and firms pay above market wages (Akerlof, 1982). To assess this hypothesis, the authors conducted two field experiments.

In the first experiment, experimental subjects are required to computerize the holdings of a library at an hourly wage of \$12. Once the task is explained to every participant, individuals in the treatment group are informed that they would be paid \$20 rather than the \$12 rate originally advertised. Individuals in the control group only observe the \$12 rate. In line with the gift exchange model, individuals exhibited higher effort in the first period (first 90 min)—on average workers in the treatment group logged 51.7 books, whereas an average of only 40.7 books were logged by workers in the control group, yielding a statistically significant difference of almost 25 percent (see second column, first row in Table 3). The increased effort levels between control and treatment groups, however, disappears in subsequent periods, where the differences are not statistically significant.

In the second experiment, the participants were asked to engage in a door-to-door fundraising drive. In the same spirit as the first experiment, the displayed hourly wage was \$10, but treatment units were informed that they would get a \$20 wage instead. Analogously,

their empirical findings show that the individuals in the treatment group raised significantly more money in the first 3-hour window (before lunch) than solicitors in the control group—an average total collection of $33 ($11 per hour) in the treatment group, whereas in the control group solicitors raised an average total of $19.2 ($6.4 per hour), yielding a statistically significant mean difference of $13.80 total ($4.6 per hour), a difference of 70 per cent. This effect, however, disappears in the second 3-hour window (after lunch), where the difference is not statistically significant (see sixth column in Table 3).

In order to complement their findings, we test for heterogeneity in the responses in the first period in both experiments as well as the consecutive time periods, both individually and jointly, accounting for multiple hypothesis problem.

Table 3: Testing for Heterogeneity in the Treatment Effect of Gift Exchanges

	Library Task				Fundraising Task			
Time Period	Mean $T-C$ Difference	Test Statistic	unadjusted p-value	adjusted p-value	Mean $T-C$ Difference	Test Statistic	unadjusted p-value	adjusted p-value
1	10.96**	0.73	0.24	0.47	13.80**	0.76	0.88	0.84
2	4.38	0.73	0.28	0.47	1.17	1.09	0.085	0.27
3	0.46	0.66	0.98	0.67				
4	0.73	0.68	0.92	0.63				

This table reports treatment effect differences in effort levels as a result of a gift exchange in the two experiments described in Gneezy and List (2006). The sample sizes of the library task for control and treatment groups are $n = 10$ and $m = 9$, respectively. Similarly, the samples for fund-raising task consisted of $n = 10$ individuals in the control group, and $m = 13$ in the treatment group. Column 1 shows the different time periods for both experiments. In the library task, each period corresponds to a 90-minute interval, whereas in the fund-raising task periods 1 and 2 reflect three-hour periods (before/after lunch). Inference for the mean difference in columns 2 and 5 was carried out using a one-tailed, right handed Wilcoxon (Mann-Whitney) nonparametric test.
Significance at $p < 0.1$ and $p < 0.05$ is denoted with * and **, respectively.

Table 3 shows the results from our test using the `R` package `RATest`. Columns 3 and 6 report the Khmaladze transformed test statistic (18), with corresponding $p-$values. The labels "unadjusted" and "adjusted" represent whether the $p-$values account for multiple hypothesis testing (adjusted) or not (unadjusted). The adjusted p-values were calculated using max T Westfall–Young procedure (Algorithm 2) with $B = 200$. Stochastic approximations for the computation of $p-$values were calculated using 999 permutations (see Remark 4).

Our empirical results show that for the first period of the library experiment, we do not reject the null hypothesis that the treatment effect is constant (unadjusted p-val= 0.24/adjusted p-val= 0.47). This conclusion is also reached in Goldman and Kaplan (2018), although their analysis finds almost rejection in upper quantiles[12]. Furthermore, the same conclusion holds

[12]Even though Goldman and Kaplan (2018) are also testing for equality at each point in the distribution, they cast this question as a multiple hypothesis testing of a continuum of single hypotheses for the CDFs.

when we look at the subsequent periods — we do not have enough evidence in favor of treatment effect heterogeneity (adjusted p-values are $p = 0.47, p = 0.67$, and $p = 0.63$). The adjusted p-values of the individual tests shed some light into the general problem of simultaneously testing the constant treatment effect hypothesis for every period (subgroup). In particular, our test does not reject the joint null hypothesis of constant treatment effect for the library task.

In like manner, our martingale-transformed permutation test does not reject the null hypothesis that the treatment effect is constant in both the pre-lunch period of the fund-raising experiment (adjusted p-val= 0.84), and the post-lunch period (adjusted p-val= 0.27). It is worth mentioning that not accounting for the multiple testing may lead to flawed inference, like we argue in Section 4. More specifically, if we naively apply the individual test to each period in the fund-raising task, ignoring multiple testing, one would conclude that the treatment (gift) had a heterogeneous effect at a 10% level in the second period (unadjusted p-value= 0.085 vs adjusted p-value= 0.27). Similar to the library task, our test does not reject the joint null hypothesis of constant treatment effect when simultaneously testing across pre/post lunch periods.

Without additional information, it is hard to draw a definite conclusion on the heterogeneity in the treatment effect and its channels, but our results can complement those of Gneezy and List (2006) and Goldman and Kaplan (2018), as well as serving as a vehicle for a more systematic future investigation of the gift exchange hypothesis.

7 Conclusions

This paper proposes a permutation test for heterogeneous treatment effects in the presence of an estimated nuisance parameter. Our method is based on the martingale transformation of the empirical process to render an asymptotically pivotal statistic, effectively killing the effect associated with the estimation error on the limiting distribution of the statistic. We show that the permutation test based on the martingale-transformed statistic results in the asymptotic rejection probability of α in general while retaining the exact control of the test level when testing for the more restrictive sharp null. We carry out Monte Carlo experiments to investigate the finite sample performance of the proposed test in comparison with other candidate methods. Numerical evidence suggests that our method is comparable to alternative methods, complementing these alternatives.

To account for the fact that the treatment effect may vary concerning observable characteristics, we extend the new method to test whether there exists treatment effect heterogeneity within subgroups defined by observable covariates. This boils down to jointly testing the null hypotheses that treatment effects are constant within mutually exclusive subgroups while allowing them to be different across subgroups. A byproduct of this extension is that we are also able to determine for which groups, if any, there is a heterogeneous treatment effect. Lastly,

we introduce the RATest R package and apply the proposed method to an investigation of the gift exchange hypothesis in two field experiments. We illustrate how to apply our proposed test to determine whether the treatment effect is heterogeneous across and within time periods. Similar to earlier studies, we find evidence in favor of a constant treatment effect as opposed to compared results that do not adjust for multiple testing.

Acknowledgments

We are very grateful to the Associate Editor and two referees for their careful and detailed comments that led to considerable improvement of the paper. We would also like to thank Roger Koenker, Jose Luis Montiel-Olea, Joseph Romano, and seminar participants at various institutions for useful comments and feedback on this paper.

References

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

Akerlof, G. A. (1982). Labor contracts as partial gift exchange. *The quarterly journal of economics*, 97(4):543–569.

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

Bai, J. and Ng, S. (2001). A consistent test for conditional symmetry in time series models. *Journal of Econometrics*, 103(1-2):225–258.

Bitler, M. P., Gelbach, J. B., and Hoynes, H. W. (2006). What mean impacts miss: Distributional effects of welfare reform experiments. *American Economic Review*, 96(4):988–1012.

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. (2017). Beyond the sharp null: Randomization inference, bounded null hypotheses, and confidence intervals for maximum effects. *arXiv preprint arXiv:1709.07339*.

Chen, Q., Zheng, X., and Pan, Z. (2015). Asymptotically distribution-free tests for the volatility function of a diffusion. *Journal of econometrics*, 184(1):124–144.

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.

Chung, E. and Romano, J. P. (2016). Multivariate and multiple permutation tests. *Journal of Econometrics*, 193(1):76–91.

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.

Delgado, M. A., Hidalgo, J., and Velasco, C. (2005). Distribution free goodness-of-fit tests for linear processes. *The Annals of Statistics*, 33(6):2568–2609.

Delgado, M. A. and Stute, W. (2008). Distribution-free specification tests of conditional models. *Journal of Econometrics*, 143(1):37–55.

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

Doksum, K. (1974). Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *The annals of statistics*, pages 267–277.

Doksum, K. A. and Sievers, G. L. (1976). Plotting with confidence: Graphical comparisons of two populations. *Biometrika*, 63(3):421–434.

Donsker, M. D. (1952). Justification and extension of doob's heuristic approach to the kolmogorov-smirnov theorems. *The Annals of mathematical statistics*, pages 277–281.

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

Durbin, J. (1975). Kolmogorov-smirnov tests when parameters are estimated with applications to tests of exponentiality and tests on spacings. *Biometrika*, pages 5–22.

Durbin, J. (1985). The first-passage density of a continuous gaussian process to a general boundary. *Journal of Applied Probability*, 22(1):99–122.

Gneezy, U. and List, J. A. (2006). Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. *Econometrica*, 74(5):1365–1384.

Goldman, M. and Kaplan, D. M. (2018). Comparing distributions by multiple testing across quantiles or cdf values. *Journal of Econometrics*.

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

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *The Annals of Mathematical Statistics*, pages 169–192.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

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

Janssen, A. (1997). Studentized permutation tests for non-iid hypotheses and the generalized behrens-fisher problem. *Statistics & probability letters*, 36(1):9–21.

Janssen, A. (1999). Testing nonparametric statistical functionals with applications to rank tests. *Journal of Statistical Planning and Inference*, 81(1):71–93.

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

Khmaladze, E. V. (1993). Goodness of fit problem and scanning innovation martingales. *The Annals of Statistics*, 21(2):798–829.

Khmaladze, E. V. and Koul, H. L. (2004). Martingale transforms goodness-of-fit tests in regression models. *The Annals of Statistics*, 32(3):995–1034.

Khmaladze, E. V. and Koul, H. L. (2009). Goodness-of-fit problem for errors in nonparametric regression: Distribution free approach. *The Annals of Statistics*, 37(6A):3165–3185.

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

Koul, H. L. and Stute, W. (1999). Nonparametric model checks for time series. *The Annals of Statistics*, 27(1):204–236.

Lee, S. and Shaikh, A. M. (2014). Multiple testing and heterogeneous treatment effects: re-evaluating the effect of progresa on school enrollment. *Journal of Applied Econometrics*, 29(4):612–626.

Lehmann, E. L. (1974). *Nonparametrics: statistical methods based on ranks*. San Francisco: Holden-Day.

Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses*. Springer Science & Business Media.

Li, B. (2009). Asymptotically distribution-free goodness-of-fit testing: A unifying view. *Econometric Reviews*, 28(6):632–657.

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.

List, J. A., Shaikh, A. M., and Xu, Y. (2016). Multiple hypothesis testing in experimental economics. *Experimental Economics*, pages 1–21.

Meinshausen, N., Maathuis, M. H., and Bühlmann, P. (2011). Asymptotic optimality of the westfall–young permutation procedure for multiple testing under dependence. *The Annals of Statistics*, 39(6):3369–3391.

Neuhaus, G. (1993). Conditional rank tests for the two-sample problem under random censorship. *The Annals of Statistics*, 21(4):1760–1779.

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

Neyman, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Originally published in 1923 and then translated by Dabrowska, Dorota M and Speed, TP. *Statistical Science*, pages 465–472.

Nobel Media AB (2019). The prize in economic sciences 2019. NobelPrize.org. Press Release. Retrieved from https://www.nobelprize.org/prizes/economic-sciences/2019/press-release.

Olivares, M. and Sarmiento, I. (2017). *RATest: Randomization Tests*. R package version 0.1.6.

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

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

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

Romano, J. P. and Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282.

Romano, J. P. and Wolf, M. (2010). Balanced control of generalized error rates. *The Annals of Statistics*, 38(1):598–633.

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

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Song, K. (2010). Testing semiparametric conditional moment restrictions using conditional martingale transforms. *Journal of Econometrics*, 154(1):74–84.

Stute, W., Thies, S., and Zhu, L.-X. (1998). Model checks for regression: an innovation process approach. *The Annals of Statistics*, 26(5):1916–1934.

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

Van der Vaart, A. W. and Wellner, J. (1996). *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Westfall, P. H. and Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*, volume 279. John Wiley & Sons.

Xiao, Z. and Xu, L. (2019). What do mean impacts miss? distributional effects of corporate diversification. *Journal of Econometrics*.

Appendix

Throughout we adopt the following notational conventions, not necessarily introduced in the main text. All limits are taken as $n \to \infty$ and $m \to \infty$. We use $\xrightarrow{\text{p}}$ to denote convergence in probability, and $\xrightarrow{\text{d}}$ to denote convergence in distribution, respectively. $\mathbb{N}$, $\mathbb{R}$, and $\mathbb{R}^k$ are used for the set of natural numbers, real numbers, and the k-dimensional Euclidean space, respectively. Symbols $\mathcal{O}_p(1)$ and $o_p(1)$ stand for being bounded in probability and convergence to zero in probability. Finally, symbols $\mathbb{E}$, $\mathbb{V}$, and $\mathbb{C}$ stand for expectation, variance and covariance, respectively. All vector are column vectors.

In addition, π and π' will denote two independent random permutations of $\{1, \dots, N\}$, and π_0 will denote the permutation that reorders observations in $\bar{Z}$, as described in Appendix C. In order to emphasize the data that are being used in the computation of the two-sample empirical processes, we will write $V_{m,n}(y, \hat{\delta}; Z_\pi)$ or $V_{m,n}(y, \hat{\delta}; \bar{Z}_\pi)$, meaning that $V_{m,n}(y, \hat{\delta})$ was calculated using sample $(Z_{\pi(1)}, \dots, Z_{\pi(N)})$ or $(\bar{Z}_{\pi(1)}, \dots, \bar{Z}_{\pi(N)})$, respectively. Analogously, $V_{m,n}(y, \hat{\delta}; \bar{Z}_{\pi'})$ is defined with π replaced by π'.

Finally, we list some of the symbols denoting stochastic processes, functionals on them, and distribution functions that we will employ in the proofs. Some of them were introduced in the

main text though included here for the sake of exposition:

$\mathbb{U}$	Standard (uniform) Brownian bridge on $[0,1]$.
$\mathbb{G}$	F_0-Brownian bridge. F_0-Brownian bridge is obtainable as $\mathbb{U} \circ F_0$.
$\mathbb{G}_1$	F_1-Brownian bridge. F_1-Brownian bridge is obtainable as $\mathbb{U} \circ F_1$.
$\widetilde{\mathbb{G}}$	For $p \in (0,1)$, $\widetilde{\mathbb{G}}(\cdot) = \sqrt{1-p}\,\mathbb{G}_1(\cdot) - \sqrt{p}\,\mathbb{G}(\cdot)$.
$\mathbb{S}$	Gaussian process with mean 0 and covariance structure $\mathbb{C}(\mathbb{S}(x), \mathbb{S}(y)) = \sigma_0^2 f_0(x) f_0(y)$.
$\mathbb{B}$	Gaussian process defined by $\mathbb{B} = \mathbb{G} + \mathbb{S}$.
$\mathbb{B}_1$	Gaussian process defined by $\mathbb{B} = \mathbb{G}_1 + \mathbb{S}$.
$\widetilde{\mathbb{B}}$	For $p \in (0,1)$, $\widetilde{\mathbb{B}}(\cdot) = \sqrt{1-p}\,\mathbb{B}_1(\cdot) - \sqrt{p}\,\mathbb{B}(\cdot)$.
$\mathbb{M}$	Standard Brownian motion given by $\mathbb{M} = \mathbb{U} + \psi_g(\mathbb{U})$.
K_0	For $y \in \mathbb{R}$, $K_0 = \sup_y \lvert\mathbb{G}(y)\rvert$
K_1	For $y \in \mathbb{R}$, $K_1 = \sup_y \lvert\mathbb{B}(y)\rvert$
K_2	For $y \in [0,1]$, $K_2 = \sup_y \lvert\mathbb{M}(y)\rvert$
J_a	For $a \in \{0,1,2\}$, the CDF of K_a

A Proof of the Main Results

In the next two theorems, the asymptotic behavior of the permutation test based on the classical two-sample Kolmogorov–Smirnov statistic is obtained. First, we state the true unconditional limiting distribution of $K_{m,n,\delta}$, which is a straightforward extension of the uniform central limit theorem, originally due to Donsker (1952), to the two-sample case. Second, we show that the the permutation distribution based on the classical two-sample Kolmogorov–Smirnov statistic asymtotically behaves like the true unconditional limiting distribution.

Theorem A. 1. *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. according to a probability distribution F_1. Consider testing the hypothesis (2) for some known δ based on the test statistic (8). Under condition A.1, $K_{m,n,\delta}$ converges weakly under the sharp null hypothesis to*

$$K_0 \equiv \sup_y |\mathbb{G}(y)| \ , \tag{A.1}$$

Moreover, if the test statistic is replaced by (12) and conditions A.1–A.2 hold, then $K^u_{m,n,\delta}(Z)$

converges weakly under the sharp null hypothesis to

$$K_0^u \equiv \sup_{0\leq t\leq 1} |\mathbb{U}(t)| \ , \tag{A.2}$$

Theorem A. 2. *Assume the premises of Theorem 1. Then the permutation distribution (15) based on $K_{m,n,\delta}$ is such that*

$$\sup_y \left| \hat{R}_{m,n}^{K(\delta)}(y) - J_0(y) \right| \xrightarrow{\text{p}} 0$$

where $J_0(\cdot)$ denotes the CDF of K_0.

Comment of Theorems A.1 and A.2: The permutation distribution based on the classical two-sample Kolmogorov–Smirnov statistic asymtotically behaves like the true unconditional limiting distribution.Consequently, the permutation test for the sharp null results in asymptotically valid inference, meaning that its limiting rejection probability under the sharp null hypothesis equals the nominal level α. It is worth mentioning that the asymptotic validity of the permutation test still holds if we replace $K_{m,n,\delta}$ with $K_{m,n,\delta}^u$ by further assuming condition A.2. Intuitively, If we assume condition A. 2, the process $\upsilon_{m,n}(\cdot,\delta)$ becomes the uniform empirical process, rendering the statistic $K_{m,n,\delta}^u$ a pivotal quantity. This latter property is the key to establishing the asymptotic behavior of the permutation test based on $K_{m,n,\delta}^u$.

A.1 Proof of Theorem A.1

Consider the following derivation

$$V_{m,n}(y,\delta) = \sqrt{\frac{mn}{N}} \left(\hat{F}_1(y+\delta) - \hat{F}_0(y) \right) = \sqrt{1-p_m}\ V_{1,m}(y) - \sqrt{p_m}\ V_{0,n}(y)$$

where we used that $p_m = m/N$ and the following definitions

$$V_{1,m}(y) = \sqrt{m} \left(\hat{F}_1(y+\delta) - F_1(y+\delta) \right)$$
$$V_{0,n}(y) = \sqrt{n} \left(\hat{F}_0(y) - F_0(y) \right)$$

Under our assumptions, $V_{1,m}$ and $V_{0,n}$ weakly converge to two tight, independent Gaussian processes, $\mathbb{G}$ and $\mathbb{G}_1$, respectively (Van der Vaart, 2000, Theorem 19.3). Independence follows by the independence of the empirical processes $V_{1,m}$ and $V_{0,n}$. Therefore, $V_{m,n}$ weakly converges to $\widetilde{\mathbb{G}}$.

By elementary properties of Gaussian processes, $\widetilde{\mathbb{G}}$ is another zero-mean Brownian bridge with covariance structure under the null hypothesis given by

$$\begin{aligned}
\mathbb{C}\left(\widetilde{\mathbb{G}}(s), \widetilde{\mathbb{G}}(t) \right) &= (1-p)\, \mathbb{C}\left(\mathbb{G}_1(s), \mathbb{G}_1(t) \right) + p\, \mathbb{C}\left(\mathbb{G}(s), \mathbb{G}(t) \right) \\
&= (1-p) \left(F_1(s \wedge t) - F_1(s) F_1(t) \right) + p \left(F_0(s \wedge t) - F_1(s) F_0(t) \right) \\
&= F_0(s \wedge t) - F_0(s) F_0(t)
\end{aligned}$$

which is the same covariance structure as $\mathbb{G}$. The desired conclusion follows from the usual continuous-mapping argument.

We next prove that $\upsilon_{m,n}(\cdot,\delta)$ weakly converges to $\mathbb{U}(\cdot)$. The proof follows closely the proof of weak convergence of $V_{m,n}$, we therefore omit some details. Start by writing $\upsilon_{m,n}$ as follows

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

Under our assumptions and the independence of the empirical processes $V_{1,m}$ and $V_{0,n}$, $\upsilon_{m,n}(\cdot,\delta)$ weakly converges to $(1-p)\mathbb{U}(\cdot) - p\mathbb{U}(\cdot) = \mathbb{U}(\cdot)$ (Van der Vaart, 2000, Theorem 19.3). The conclusion follows by a direct application of the continuous mapping theorem.

A.2 Proof of Theorem A.2

Below we assume without loss of generality that $\delta = 0$; the general case follows from the same arguments with $Y_i(1)$ replaced by $Y_i(1) - \delta$. Write

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

Thus under the null hypothesis $Z_1,\ldots,Z_N$ are i.i.d. according to a probability distribution 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 the same way with π replaced by π'.

We first argue that under our assumptions,

$$\left(K_{m,n,\delta}(Z_\pi), K_{m,n,\delta}(Z_{\pi'})\right) \xrightarrow{\mathrm{d}} (K_0, K_0') \tag{A.3}$$

where K_0 and K_0' are independent and with common c.d.f. $J_0(\cdot)$.

Step 1. We start the proof by showing that $(V_{m,n}(\cdot,\delta;Z_\pi), V_{m,n}(\cdot,\delta;Z_{\pi'}))$ weakly converges to $(\mathbb{G}(\cdot),\mathbb{G}'(\cdot))$. This result follows by verifying that the finite-dimensional distributions

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

converge weakly to the marginals

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

for all $k \in \mathbb{N}$, and $t_1,\ldots,t_k \in \mathbb{R}$, and that the process $V_{m,n}(\cdot,\delta;Z_\pi)$ is asymptotically tight (Van der Vaart and Wellner, 1996, Theorem 1.5.3). Convergence of the finite-dimensional

distributions follows by Lemma B.1, whereas asymptotic tightness follows from condition A.2 paired with the arguments in the proof of Theorem 3.7.1 in Van der Vaart and Wellner (1996).

Step 2. We now prove that

$$\Big(\mathbb{G}(t_1),\ldots,\mathbb{G}(t_k)\Big) \perp\!\!\!\perp \Big(\mathbb{G}'(t_1),\ldots,\mathbb{G}'(t_k)\Big) \tag{A.4}$$

for all $k \in \mathbb{N}$, and $t_1,\ldots,t_k \in \mathbb{R}$. Since the joint limit distribution $\Big(\mathbb{G}(t_1),\ldots,\mathbb{G}(t_k),\mathbb{G}'(t_1),\ldots,\mathbb{G}'(t_k)\Big)$ is a multivariate normal distribution, it suffices to show that the covariance matrix is block-diagonal. The proof of Lemma B.1 establishes this result.

Combining the weak convergence result from **Step 1** and A.4 with the regular continuous mapping theorem, we see that A.3 holds.

It now follows from Hoeffding's Condition (Chung and Romano, 2013, Lemma 5.1)

$$\sup_y \Big|\hat{R}^{K(\delta)}_{m,n}(y) - J_0(y)\Big| \xrightarrow{\text{p}} 0$$

completing the proof of the theorem.

A.3 Proof of Theorem 1

We begin the proof by noting some preliminary facts which will be useful in the analysis of the asymptotic behavior of $K_{m,n,\hat{\delta}}$. Specifically, under the null hypothesis we have $\mathbb{V}(Y(1)) = \mathbb{V}(Y(0)) = \sigma^2$, and if we further assume condition A.2, we obtain $f_1(y+\delta) = f_0(y)$. Furthermore, assumption A. 2 implies that F_1 and F_0 are Lipschitz continuous, then $\sup_y f_0(y) < \infty$.

Step 1. We start by writing $V_{m,n}(y,\hat{\delta})$ as

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

where the last equality follows due to the fact the last summand in first equality

$$\sqrt{\frac{mn}{N}}\Big\{\Big(\hat{F}_1(y+\hat{\delta}) - F_1(y+\hat{\delta})\Big) - \Big(\hat{F}_1(y+\delta)) - F_1(y+\delta)\Big)\Big\} = o_p(1) \tag{A.5}$$

by stochastic equicontinuity of $\{\hat{F}_1(y) - F_1(y) : y \in \mathbb{R}\}$.

Condition A.2 allows us to expand $F_1(y+\hat{\delta})$ around δ to obtain:

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

Observe

$$\begin{aligned} \sqrt{\frac{mn}{N}}(\hat{\delta}-\delta) &= \sqrt{\frac{mn}{N}}\left(\frac{1}{m}\sum_{i=1}^{m}\Big(Y_i(1)-\mathbb{E}(Y(1)\Big)-\frac{1}{n}\sum_{i=m+1}^{N}\Big(Y_i(0)-\mathbb{E}(Y(0))\Big)\right) \\ &= \sqrt{\frac{n}{N}}\left(\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\Big(Y_i(1)-\mathbb{E}(Y(1))\Big)\right)-\sqrt{\frac{m}{N}}\left(\frac{1}{\sqrt{n}}\sum_{i=m+1}^{N}\Big(Y_i(0)-\mathbb{E}(Y(0))\Big)\right) \end{aligned}$$

therefore

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

Step 2. We next prove that under the null hypothesis and condition A.1, $V_{m,n}(\cdot,\delta)$ weakly converges to $\widetilde{\mathbb{B}}(\cdot)$. To show this, we first argue that

$$B_{m,1}(y)=\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\left\{\mathbb{1}_{\{Y_i(1)\le y+\delta\}}-F_1(y+\delta)+f_0(y)\left(Y_i(1)-\mathbb{E}(Y(1))\right)\right\} \tag{A.6}$$

and

$$B_{m,0}(y)=\frac{1}{\sqrt{n}}\sum_{i=m+1}^{N}\left\{\mathbb{1}_{\{Y_i(0)\le y\}}-F_0(y)+f_0(y)\Big(Y_i(0)-\mathbb{E}(Y(0))\Big)\right\} \tag{A.7}$$

weakly converge to two independent processes $\mathbb{B}_1(\cdot)$ and $\mathbb{B}(\cdot)$, respectively. This is follows by modifying slightly the proof of Theorem A.1. The modification simply involves accounting for the additional summand, which is a sum of mean-zero, independent stochastic processes with uniformly bounded expectation.

Under the null hypothesis $F_1 = F_0$, the limit variable $\widetilde{\mathbb{B}} = \sqrt{1-p}\,\mathbb{B}_1 - \sqrt{p}\,\mathbb{B}$ possesses the same distribution as $\mathbb{B}$. The desired conclusion follows from the usual continuous-mapping argument.

A.4 Proof of Theorem 2

Throughout this proof, we will consider $\tilde{Y}_i(1) \equiv Y_i(1) - \hat{\delta}$, $1 \leq i \leq m$, where $\tilde{Y}(1)$ is distributed according to probability distribution $\tilde{F}_1$. In other words, $\tilde{Y}(1)$ is the recentered version of $Y(1)$, where the shift is given by $\hat{\delta}$.

Write

$$X = (X_1, \ldots, X_N) = (\tilde{Y}_1(1), \ldots, \tilde{Y}_m(1), Y_1(0), \ldots, Y_n(0)) \tag{A.8}$$

Independent of the Xs, let $(\pi(1), \ldots, \pi(N))$ and $(\pi'(1), \ldots, \pi'(N))$ be two independent random permutations of $\{1, \ldots, N\}$. We will denote $X_\pi = (X_{\pi(1)}, \ldots, X_{\pi(N)})$; $X_{\pi'}$ is defined the same way with π replaced by π'.

We first argue that under our assumptions,

$$\left(K_{m,n,\hat{\delta}}(X_\pi), K_{m,n,\hat{\delta}}(X_{\pi'})\right) \xrightarrow{\text{d}} (K_0, K_0') \tag{A.9}$$

where K_0 and K_0' are independent and with common c.d.f. $J_0(\cdot)$.

Step 1. We start the proof of (A.9) by applying the coupling construction and contiguity result of Chung and Romano (2013). More specifically, couple data X with an auxiliary sample of N i.i.d. observations $\bar{X} = (\bar{X}_1, \ldots, \bar{X}_N)$ from the mixture distribution $\bar{P} = p\tilde{F}_1 + (1-p)F_0$, where $p = \lim_{m\to\infty} m/N$ (see Appendix C in this paper).

Step 2. We now argue that the permutation distribution based on X should behave approximately like the behavior of the permutation distribution based on $\bar{X}$. Following the arguments in the proof of Lemma 5.1 in Chung and Romano (2013), it suffices to verify the following two conditions

$$\left(K_{m,n,\hat{\delta}}(\bar{X}_\pi), K_{m,n,\hat{\delta}}(\bar{X}_{\pi'})\right) \xrightarrow{\text{d}} (K_0, K_0') \tag{A.10}$$

$$K_{m,n,\hat{\delta}}(\bar{X}_{\pi,\pi_0}) - K_{m,n,\hat{\delta}}(X_\pi) \xrightarrow{\text{p}} 0 \tag{A.11}$$

Lemma B.2 establishes (A.10) and Lemma B.3 establishes A.11.

It now follows from Hoeffding's Condition (Chung and Romano, 2013, Lemma 5.1)

$$\sup_y \left| \hat{R}_{m,n}^{K(\hat{\delta})}(y)) - J_0(y) \right| \xrightarrow{\text{p}} 0$$

completing the proof of the theorem.

A.5 Proof of Theorem 3

We begin the proof by stating some facts which follow from the null hypothesis, appearing also in the proof of Theorem 1, namely δ is $\sqrt{N}$-consistently estimable, $\mathbb{V}(Y(1)) = \mathbb{V}(Y(0)) = \sigma^2$.

If we further assume condition A.2, then $f_1(y+\delta) = f_0(y)$ for all y under the null hypothesis. Lastly, recall $\psi_g(h)(\cdot)$ is a linear mapping with respect to h, and $\psi_g(cg) = cg$ for a constant or random variable c.

The Khmaladze transformation based on $\upsilon_{m,n}(t,\hat{\delta})$ is

$$\begin{aligned}\tilde{\upsilon}_{m,n}(t,\hat{\delta}) &= \upsilon_{m,n}(t,\hat{\delta}) - \int_0^t \left[\dot{g}(s)'C^{-1}(s)\int_s^1 \dot{g}(r)d\upsilon_{m,n}(r,\hat{\delta})\right] ds \\ &= \upsilon_{m,n}(t,\hat{\delta}) - \psi_g(\upsilon_{m,n}(t,\hat{\delta}))\end{aligned} \tag{A.12}$$

From Assumption A.2, expand $\upsilon_{m,n}(t,\hat{\delta})$ around δ to derive the following asymptotic representation under the null hypothesis:

$$\begin{aligned}\upsilon_{m,n}(t,\hat{\delta}) &= \sqrt{\frac{mn}{N}}\left\{\left(\hat{F}_1\left(F_0^{-1}(t)+\delta\right) + f_0\left(F_0^{-1}(t)\right)\left(\hat{\delta}-\delta\right) + o_p(1)\right) - \hat{F}_0(F_0^{-1}(t))\right\} \\ &= \upsilon_{m,n}(y,\delta) + \sqrt{\frac{mn}{N}}\left(f_0\left(F_0^{-1}(t)\right)(\hat{\delta}-\delta)\right) + o_p(1)\end{aligned} \tag{A.13}$$

Next, note that

$$\psi_g(\upsilon_{m,n}(t,\hat{\delta})) = \psi_g(\upsilon_{m,n}(t,\delta)) + \sqrt{\frac{mn}{N}}\left(f_0\left(F_0^{-1}(t)\right)(\hat{\delta}-\delta)\right) + o_p(1) \tag{A.14}$$

by properties of map ψ. Plug (A.13)-(A.14) into (A.12) to obtain

$$\tilde{\upsilon}_{m,n}(t,\hat{\delta}) = \upsilon_{m,n}(t,\delta) - \psi_g\left(\upsilon_{m,n}(t,\delta)\right) + o_p(1) \tag{A.15}$$

It follows from Theorem A.1 that $\upsilon_{m,n}(\cdot,\delta)$ converges weakly to $\mathbb{U}$. Therefore, $\tilde{\upsilon}_{m,n}(\cdot,\hat{\delta})$ weakly converges to $\mathbb{M}(\cdot)$ (Khmaladze, 1981, 4.3). Combining this last result with a direct application of the continuous mapping theorem, we see that $\tilde{K}_{m,n,\hat{\delta}}$ converges in distribution under the null hypothesis to K_2, completing the proof of the theorem.

A.6 Proof of Theorem 4

The proof follows closely the proof of Theorem 2, we therefore omit some details. In particular, let $\tilde{Y}(1)$, X, π, π', X_π, and the coupled data $\bar{X}$ as in the proof of Theorem 2.

We first argue that under our assumptions,

$$\left(\tilde{K}_{m,n,\hat{\delta}}(X_\pi), \tilde{K}_{m,n,\hat{\delta}}(X_{\pi'})\right) \xrightarrow{\mathrm{d}} (K_2, K_2') \tag{A.16}$$

where K_2 and K_2' are independent and with common c.d.f. $J_2(\cdot)$.

Following the arguments in the proof of Lemma 5.1 in Chung and Romano (2013), a sufficient condition for (A.16) is given by the following

$$\left(\tilde{K}_{m,n,\hat{\delta}}(\bar{X}_\pi), \tilde{K}_{m,n,\hat{\delta}}(\bar{X}_{\pi'})\right) \xrightarrow{\mathrm{d}} (K_2, K_2') \tag{A.17}$$

$$\tilde{K}_{m,n,\hat{\delta}}(\bar{X}_{\pi,\pi_0}) - \tilde{K}_{m,n,\hat{\delta}}(X_\pi) \xrightarrow{\mathrm{p}} 0 \tag{A.18}$$

Lemma B.4 establishes (A.17) and Lemma B.5 establishes A.18.

It now follows from Hoeffding's Condition (Chung and Romano, 2013, Lemma 5.1)

$$\sup_y \left| \hat{R}_{m,n}^{\tilde{K}(\hat{\delta})}(y)) - J_2(y) \right| \xrightarrow{\mathrm{p}} 0$$

completing the proof of the theorem.

B Auxiliary Results

Lemma B.1. *Assume the premises of Theorem A.1. Let Z be defined as in the proof of Theorem A.2. Independent of Z, let π and π' be two independent random permutations of $\{1, \ldots, N\}$, with Z_π and $Z_{\pi'}$ defined as in the proof of Theorem A.2. Denote the $(2 \times k)$ vector of finite-dimensional distributions, $\mathbf{V}(\delta)$, as*

$$\mathbf{V}(\delta) = (V_{m,n}(t_1, \delta; Z_\pi)), \ldots, V_{m,n}(t_k, \delta; Z_\pi)), V_{m,n}(t_1, \delta; Z_{\pi'})), \ldots, V_{m,n}(t_k, \delta; Z_{\pi'})))^\intercal$$

for all $k \in \mathbb{N}$, $t_1, \ldots, t_k \in \mathbb{R}$, and $\delta \in \mathbb{R}$. Then

$$\mathbf{V}(\delta) \xrightarrow{\mathrm{d}} (\mathbb{G}(t_1), \ldots, \mathbb{G}(t_k), \mathbb{G}'(t_1), \ldots, \mathbb{G}'(t_k))$$

Proof. We start the proof by defining

$$W_i = \begin{cases} 1 & \text{if } \pi(i) \in \{1, \ldots, m\} \\ -\frac{m}{n} & \text{if } \pi(i) \in \{m+1, \ldots, N\} \end{cases}$$

for $1 \leq i \leq N$, and W_i' is defined with π replaced by π'. Note that $\mathbb{E}(W_i) = \mathbb{E}(W_i') = 0$, and $\mathbb{E}(W_i^2) = \mathbb{E}(W_i'^2) = m/n$.

Consider the following derivation

$$\begin{aligned} \mathbf{V}(\delta) = a_m^{1/2} \Bigg(& \sum_{i=1}^N \left(\mathbb{1}_{\{Z_i \leq t_1\}} - F_0(t_k) \right) W_i, \ldots, \sum_{i=1}^N \left(\mathbb{1}_{\{Z_i \leq t_k\}} - F_0(t_k) \right) W_i, \\ & \sum_{i=1}^N \left(\mathbb{1}_{\{Z_i \leq t_1\}} - F_0(t_k) \right) W_i', \ldots, \sum_{i=1}^N \left(\mathbb{1}_{\{Z_i \leq t_k\}} - F_0(t_k) \right) W_i' \Bigg)^\intercal \end{aligned}$$

where $a_m = (1 - p_m)/m$. Observe that independence of π, π' from Z ensures that

$$\begin{aligned}
\mathbb{E}\left(\left(\mathbb{1}_{\{Z_i \le t_j\}} - F_0(t_j)\right) W_i\right) &= 0 \\
\mathbb{E}\left(\left(\mathbb{1}_{\{Z_i \le t_j\}} - F_0(t_j)\right)^2 W_i^2\right) &= F_0(t_j)\big(1 - F_0(t_j)\big) \\
\mathbb{E}\left(\left(\mathbb{1}_{\{Z_i \le t_j\}} - F_0(t_j)\right)\left(\mathbb{1}_{\{Z_i \le t_l\}} - F_0(t_l)\right) W_i^2\right) &= \frac{m}{n}\big(F_0(t_j \wedge t_l) - F_0(t_j)F_0(t_l)\big) \\
\mathbb{E}\left(\left(\mathbb{1}_{\{Z_i \le t_j\}} - F_0(t_j)\right)\left(\mathbb{1}_{\{Z_i \le t_l\}} - F_0(t_l)\right) W_i W_i'\right) &= 0
\end{aligned}$$

for $1 \le i \le N$, $1 \le j \le k$, $1 \le l \le k$, and $k \in \mathbb{N}$. Same equalities follow if we replace W_i by W_i'. Combining these facts, it is easy to check that $\mathbb{E}(\mathbf{V}(\delta)) = \mathbf{0}$, and block-diagonal covariance matrix $\mathbf{\Sigma} = \operatorname{diag}\{\mathbf{\Sigma}_i \mid i = 1, 2\}$, with

$$\mathbf{\Sigma}_i = \begin{pmatrix} F_0(t_1)(1 - F_0(t_1) & \dots & F_0(t_1 \wedge t_k) - F_0(t_1)F_0(t_k) \\ \vdots & \ddots & \vdots \\ F_0(t_k \wedge t_1) - F_0(t_k)F_0(t_1) & \cdots & F_0(t_k)(1 - F_0(t_k) \end{pmatrix}$$

We now claim the asymptotic normality of $\mathbf{V}(\delta)$. Using the Cramér–Wold device (Lehmann and Romano, 2005, Theorem 11.2.3), it suffices to show for vector $\mathbf{c} \in \mathbb{R}^{2k}$ that

$$\mathbf{c}^\intercal \mathbf{V}(\delta) \xrightarrow{\text{d}} c_1\mathbb{G}(t_1) + \cdots + c_k\mathbb{G}(t_k) + c_{k+1}\mathbb{G}'(t_1) + \cdots + c_{2k}\mathbb{G}'(t_k) \tag{B.1}$$

Write $\mathbf{c}^\intercal \mathbf{V}(\delta)$ as follows

$$a_m^{1/2} \sum_{j=1}^{k} \sum_{i=1}^{N} \left(\mathbb{1}_{\{Z_i \le t_j\}} - F_0(t_j)\right)(c_j W_i c_{k+j} W_i') \tag{B.2}$$

Conditionally on W_i and W_i', (B.2) is an independent sum of linear combinations of independent random variables:

$$\begin{aligned}
\mathbf{c}^\intercal \mathbf{V}(\delta) &= a_m^{1/2} \sum_{j=1}^{k} \left(\sum_{i=1}^{m} \left(\mathbb{1}_{\{Z_i \le t_j\}} - F_0(t_j)\right)(c_j W_i c_{k+j} W_i') + \sum_{i=m+1}^{N} \left(\mathbb{1}_{\{Z_i \le t_j\}} - F_0(t_j)\right)(c_j W_i c_{k+j} W_i') \right) \\
&= a_m^{1/2} \sum_{j=1}^{k} \left(\sum_{i=1}^{m} \left(\mathbb{1}_{\{Y_i(1) \le t_j\}} - F_0(t_j)\right)(c_j W_i c_{k+j} W_i') + \sum_{i=m+1}^{N} \left(\mathbb{1}_{\{Y_i(0) \le t_j\}} - F_0(t_j)\right)(c_j W_i c_{k+j} W_i') \right)
\end{aligned}$$

For every summand j above, we can show that

$$a_m^{-1/2} \left(\frac{\max_{i=1,\dots,N} (c_j W_i c_{k+j} W_i')}{\sum_{i=1}^{N} (c_j W_i c_{k+j} W_i')^2} \right) \xrightarrow{\text{p}} 0, \quad \text{as} \quad m, n \to \infty$$

by the arguments in Example 15.2.5 of Lehmann and Romano (2005). Apply this to every summand to conclude

$$\mathbf{c}^\intercal \mathbf{V}(\delta) \xrightarrow{\mathrm{d}} \sum_{j=1}^{k} \Big(c_j \mathbb{G}(t_j) + c_{k+j} \mathbb{G}(t_j) \Big)$$

This finishes the proof.

□

Lemma B.2. *Assume the premises and notation introduced in the proof of Theorem 2. Let* $\bar{X} = (\bar{X}_1, \ldots, \bar{X}_N)$ *be an i.i.d. sequence from the mixture distribution* $\bar{P} = p\tilde{F}_1 + (1-p)F_0$. *Then*

$$\Big(K_{m,n,\hat{\delta}}(\bar{X}_\pi), K_{m,n,\hat{\delta}}(\bar{X}_{\pi'}) \Big) \xrightarrow{\mathrm{d}} (K_0, K_0')$$

Proof. We start the proof by showing that $\Big(V_{m,n}(\cdot, \hat{\delta}; \bar{X}_\pi), V_{m,n}(\cdot, \hat{\delta}; \bar{X}_{\pi'}) \Big)$ weakly converges to $(\mathbb{G}(\cdot), \mathbb{G}'(\cdot))$. The proof of this result closely follows the proof of Lemma B.1 so we omit some details.

Step 1. For marginal convergence, it suffices to show that

$$\mathbf{V}(\hat{\delta}) = \Big(V_{m,n}(t_1, \hat{\delta}; \bar{X}_\pi), \ldots, V_{m,n}(t_k, \hat{\delta}; \bar{X}_\pi), V_{m,n}(t_1, \hat{\delta}; \bar{X}_{\pi'}), \ldots, V_{m,n}(t_k, \hat{\delta}; \bar{X}_{\pi'}) \Big)$$

converge weakly to the marginals

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

for all $k \in \mathbb{N}$, and $t_1, \ldots, t_k \in \mathbb{R}$. Let W_i and W_i' be defined as in the proof of Lemma B.1 and consider the following derivation

$$\begin{aligned} \mathbf{V}(\hat{\delta}) = a_m^{1/2} \Bigg(& \sum_{i=1}^{N} \Big(\mathbb{1}_{\{\bar{X}_i \le t_1\}} - F_0(t_k) \Big) W_i, \ldots, \sum_{i=1}^{N} \Big(\mathbb{1}_{\{\bar{X}_i \le t_k\}} - F_0(t_k) \Big) W_i, \\ & \sum_{i=1}^{N} \Big(\mathbb{1}_{\{\bar{X}_i \le t_1\}} - F_0(t_k) \Big) W_i', \ldots, \sum_{i=1}^{N} \Big(\mathbb{1}_{\{\bar{X}_i \le t_k\}} - F_0(t_k) \Big) W_i' \Bigg)^\intercal \end{aligned}$$

where $a_m = (1 - p_m)/m$. Observe that independence of π, π' from $\bar{X}$ ensures that

$$\begin{aligned} \mathbb{E}\left(\Big(\mathbb{1}_{\{\bar{X}_i \le t_j\}} - F_0(t_j) \Big) W_i \right) &= 0 \\ \mathbb{E}\left(\Big(\mathbb{1}_{\{\bar{X}_i \le t_j\}} - F_0(t_j) \Big)^2 W_i^2 \right) &= \frac{m}{n} \Big(F_0(t_j) \left(1 - F_0(t_j) + p(1 - 2F_0(t_j))(\tilde{F}_1 - F_0(t_j)) \right) \Big) \\ &= \frac{m}{n} \Big(F_0(t_j) \left(1 - F_0(t_j) \right) \Big) + o_p(1) \end{aligned}$$

whereas the expected value of the cross-products

$$\mathbb{E}\left(\left(\mathbb{1}_{\{\bar{X}_i \le t_j\}} - F_0(t_j)\right)\left(\mathbb{1}_{\{\bar{X}_i \le t_l\}} - F_0(t_l)\right) W_i^2\right) = \frac{m}{n}\Big(F_0(t_j \wedge t_l) - F_0(t_j)F_0(t_l)\Big) + o_p(1)$$
$$\mathbb{E}\left(\left(\mathbb{1}_{\{\bar{X}_i \le t_j\}} - F_0(t_j)\right)\left(\mathbb{1}_{\{\bar{X}_i \le t_l\}} - F_0(t_l)\right) W_i W_i'\right) = 0$$

for $1 \le i \le N$, $1 \le j \le k$, $1 \le l \le k$, and $k \in \mathbb{N}$. Same equalities follow if we replace W_i by W_i'. Combining these facts, it is easy to check that $\mathbb{E}(\mathbf{V}(\delta)) = \mathbf{0}$, and block-diagonal covariance matrix $\mathbf{\Sigma} = \operatorname{diag}\{\mathbf{\Sigma}_i \mid i = 1, 2\} + o_p(1)$, with

$$\mathbf{\Sigma}_i = \begin{pmatrix} F_0(t_1)(1 - F_0(t_1) & \dots & F_0(t_1 \wedge t_k) - F_0(t_1)F_0(t_k) \\ \vdots & \ddots & \vdots \\ F_0(t_k \wedge t_1) - F_0(t_k)F_0(t_1) & \cdots & F_0(t_k)(1 - F_0(t_k) \end{pmatrix}$$

We now claim the asymptotic normality of $\mathbf{V}(\hat{\delta})$. It suffices by the the Cramér–Wold device (Lehmann and Romano, 2005, Theorem 11.2.3) to show that for every $\mathbf{c} \in \mathbb{R}^{2k}$,

$$\mathbf{c}^\intercal \mathbf{V}(\hat{\delta}) \xrightarrow{\mathrm{d}} c_1\mathbb{G}(t_1) + \cdots + c_k\mathbb{G}(t_k) + c_{k+1}\mathbb{G}'(t_1) + \cdots + c_{2k}\mathbb{G}'(t_k) \tag{B.3}$$

The verification of (B.3) follows by the same arguments of the proof of Lemma B.1 with Z_i replaced by $\bar{X}_i$.

Step 2. Asymptotic tightness from the standard arguments in the classical empirical process since

$$\mathbb{1}_{\{\bar{X}_i \le t_j\}} - F_0(t_j) = \left(\mathbb{1}_{\{\bar{X}_i \le t_j\}} - \bar{P}(t_j)\right) + \underbrace{p\left(\tilde{F}_1(t_j) - F_0(t_j)\right)}_{o_p(1)}$$

for all $j = 1, \dots, k$ under the null hypothesis and the assumptions of Theorem 2.

Independence follows by joint Gaussianity combined by the block-diagonal covariance matrix. Combining these results with a direct application of the continuous mapping theorem, we see that (A.10) follows.

□

Lemma B.3. *Assume the premises and notation introduced in the proof of Theorem 2. Let* $\bar{X} = (\bar{X}_1, \dots, \bar{X}_N)$ *be an i.i.d. sequence from the mixture distribution* $\bar{P} = p\tilde{F}_1 + (1 - p)F_0$. *Then*

$$K_{m,n,\hat{\delta}}(\bar{X}_{\pi,\pi_0}) - K_{m,n,\hat{\delta}}(X_\pi) \xrightarrow{\mathrm{p}} 0$$

Proof. The proof boils down to showing convergence in probability by proving convergence in quadratic mean. Everything stated below is implicitly *conditioned* on π_0, but we omit it to avoid notation clutter.

We start the proof by showing that for a given π,

$$\left(\frac{mn}{N}\right)^{-1/2}\left(V_{m,n}(y,\hat{\delta};\bar{X}_{\pi\pi_0}))-V_{m,n}(y,\hat{\delta};X_\pi)\right)=\frac{1}{m}\sum_{i=1}^{m}(\mathbb{1}_{\{\bar{X}_{\pi\pi_0(i)}\le y\}}-\mathbb{1}_{\{X_{\pi(i)}\le y\}})$$
$$-\frac{1}{n}\sum_{j=m+1}^{N}(\mathbb{1}_{\{\bar{X}_{\pi\pi_0(j)}\le y\}}-\mathbb{1}_{\{X_{\pi(j)}\le y\}})$$

Observe that the way we constructed $\bar{X}$, we have that $X_i=\bar{X}_{\pi_0(i)}$ for indexes i except for at most D entries. This is so because $\bar{X}_{\pi_0}$ is either of the form

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

or it is of the form

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

Then all the above sums are zero except for at most D places. For all the indexes such that the differences $\mathbb{1}_{\{\bar{X}_{\pi\pi_0(i)}\le y\}}-\mathbb{1}_{\{X_{\pi(i)}\le y\}}$ and $\mathbb{1}_{\{\bar{X}_{\pi\pi_0(j)}\le y\}}-\mathbb{1}_{\{X_{\pi(j)}\le y\}}$ are not zero, observe that

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

Assumption A.2 allows us to expand $F_1(y+\hat{\delta})$ around δ to obtain

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

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

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

Furthermore, any nonzero term like $\mathbb{1}_{\{\bar{X}_{\pi\pi_0(j)}\leq y\}} - \mathbb{1}_{\{X_{\pi(j)}\leq y\}}$ has variance bounded above by

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

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

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

and therefore the unconditional variance is bounded above by

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

Therefore convergence in quadratic mean follows. □

Lemma B.4. *Assume the premises and notation introduced in the proof of Theorem 4. Let* $\bar{X} = (\bar{X}_1, \ldots, \bar{X}_N)$ *be an i.i.d. sequence from the mixture distribution* $\bar{P} = p\tilde{F}_1 + (1-p)F_0$. *Then*

$$\left(\tilde{K}_{m,n,\hat{\delta}}(\bar{X}_\pi), \tilde{K}_{m,n,\hat{\delta}}(\bar{X}_{\pi'})\right) \xrightarrow{\mathrm{d}} (K_2, K_2')$$

with K_2 *and* K_2' *independent random variables with common c.d.f.* $J_2(\cdot)$.

Proof. We start the proof by showing that

$$\left(\tilde{v}_{m,n}(\cdot,\hat{\delta};\hat{X}_\pi)), \tilde{v}_{m,n}(\cdot,\hat{\delta};\hat{X}_{\pi'}))\right) \tag{B.4}$$

weakly converges to $(\mathbb{M}, \mathbb{M}')$ under the null hypothesis, where the limit variable $\mathbb{M}$ possesses the same distribution as $\mathbb{M}'$. This follows by the arguments in the discussion of Condition E in Romano (1989). However, the differentiability condition needed in order to verify Condition E holds for the present case since the testing problem we are dealing with is essentially a two-sample test of homogeneity (Romano, 1989, example 4). Therefore the convergence part of (A.17) follows by a simple application of the continuous mapping theorem.

We now argue that $\mathbb{M}$ and $\mathbb{M}'$ are independent. Since $(\mathbb{M}, \mathbb{M}')$ is a Gaussian process, zero-covariance renders independence. It suffices to show that

$$\mathbb{C}\left(\tilde{v}_{m,n}(t,\hat{\delta};\bar{X}_\pi)), \tilde{v}_{m,n}(s,\hat{\delta};\bar{X}_{\pi'}))\right) = o_p(1) \tag{B.5}$$

for any $0 \leq s, t \leq 1$.

Step 1. We first show that $\tilde{\upsilon}_{m,n}(t, \hat{\delta}; \bar{X}_\pi)$ still admits the asymptotic representation (A.15) introduced in Theorem 5 under $\bar{X}_\pi$. More formally:

$$\tilde{\upsilon}_{m,n}(t, \hat{\delta}; \bar{X}_\pi) = \upsilon_{m,n}(t, \delta; \bar{X}_\pi) - \psi_g\left(\upsilon_{m,n}(t, \delta; \bar{X}_\pi)\right) + o_p(1) \tag{B.6}$$

for any $0 \leq t \leq 1$. This follows by Lemma B.6.

Step 2. Consider the following derivation, which follows from the asymptotic representation in (B.6) and the linearity of the map ψ_g:

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

Recall that $\upsilon_{m,n}(t, \hat{\delta}, \bar{X}_\pi) = V_{m,n}(F_0^{-1}(t), \delta; \bar{X}_\pi)$, hence

$$\mathbb{C}\left(\upsilon_{m,n}(t, \delta; \bar{X}_\pi), \upsilon_{m,n}(s, \delta; \bar{X}_{\pi'})\right) = o_p(1) \tag{B.7}$$

by the arguments in the proof of Lemma B.2. The proof of Lemma B.2 also states

$$\mathbb{E}\left(\left(\mathbb{1}_{\{\bar{X}_i \leq t_j\}} - F_0(t_j)\right)\left(\mathbb{1}_{\{\bar{X}_i \leq t_l\}} - F_0(t_l)\right) W_i W_i'\right) = 0 \tag{B.8}$$

for all $k \in \mathbb{N}$, and $t_1, \ldots, t_k \in \mathbb{R}$. . Combine (B.8) and $\upsilon_{m,n}(t, \hat{\delta}, \bar{X}_\pi) = V_{m,n}(F_0^{-1}(t), \delta; \bar{X}_\pi)$ to conclude

$$\begin{aligned}
\mathbb{C}\left(\psi_g\left(\upsilon_{m,n}(t, \delta; \bar{X}_\pi)\right), \psi_g\left(\upsilon_{m,n}(s, \delta; \bar{X}_{\pi'})\right)\right) &= o_p(1) \\
\mathbb{C}\left(\upsilon_{m,n}(t, \delta; \bar{X}_\pi), \psi_g\left(\upsilon_{m,n}(s, \delta; \bar{X}_{\pi'})\right)\right) &= o_p(1) \\
\mathbb{C}\left(\upsilon_{m,n}(s, \delta; \bar{X}_{\pi'}), \psi_g\left(\upsilon_{m,n}(t, \delta; \bar{X}_\pi)\right)\right) &= o_p(1)
\end{aligned}$$

by the linearity of the map ψ_g, which allows us to exchange the order of ψ_g and the expectation. Hence condition (B.5) holds as desired.

□

Lemma B.5. *Assume the premises and notation introduced in the proof of Theorem 4. Let $\bar{X} = (\bar{X}_1, \ldots, \bar{X}_N)$ be an i.i.d. sequence from the mixture distribution $\bar{P} = p\tilde{F}_1 + (1-p)F_0$. Then*

$$\tilde{K}_{m,n,\hat{\delta}}(\bar{X}_{\pi,\pi_0}) - \tilde{K}_{m,n,\hat{\delta}}(X_\pi) \xrightarrow{\mathrm{p}} 0$$

Proof. Everything stated below is implicitly conditioned on π_0, but we omit it to ease notation. We start the proof by arguing the asymptotic representation in the proof of Theorem 3 still applies under X_π.

We argue that the remainder, defined in equation (A.5) in the proof of Theorem 1, is still $o_p(1)$ under X_π. This follows by Lemma B.6.

We now show that for a fixed π

$$\tilde{\upsilon}_{m,n}(t,\hat{\delta};\bar{X}_{\pi,\pi 0}) - \tilde{\upsilon}_{m,n}(t,\hat{\delta};X_\pi) = \upsilon_{m,n}(t,\delta;\bar{X}_{\pi,\pi 0}) - \upsilon_{m,n}(t,\delta;X_\pi) - \\ \left(\psi_g\left(\upsilon_{m,n}(t,\delta;\bar{X}_{\pi,\pi 0})\right) - \psi_g\left(\upsilon_{m,n}(t,\delta;X_\pi)\right)\right) + o_p(1)$$

We will divide this into two separete steps, namely

$$\upsilon_{m,n}(y,\delta;\bar{X}_{\pi,\pi 0}) - \upsilon_{m,n}(y,\delta;X_\pi) = o_p(1) \tag{B.9}$$

$$\psi_g\left(\upsilon_{m,n}(t,\delta;\bar{X}_{\pi,\pi 0})\right) - \psi_g\left(\upsilon_{m,n}(t,\delta;X_\pi)\right) = o_p(1) \tag{B.10}$$

Condition (B.9) follows by the same arguments as in the proof of Lemma B.3 and Theorem 1. For the verification of condition (B.10), we note that the linear operator ψ_g is also a Fredholm operator (Koenker and Xiao, 2002) on a Banach space, hence 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, (B.10) follows by (B.9) and the continuous mapping theorem.

Consequently, (A.18) follows by a simple application of the continuous mapping theorem. This finishes the proof.

□

Lemma B.6. *Assume the premises and notation introduced in the proof of Theorem 4. Let* $\bar{X} = (\bar{X}_1,\ldots,\bar{X}_N)$ *be an i.i.d. sequence from the mixture distribution* $\bar{P} = p\tilde{F}_1 + (1-p)F_0$. *Then* $\tilde{\upsilon}_{m,n}(t,\hat{\delta};\bar{X}_\pi)$ *still admits the asymptotic representation* (A.15) *introduced in Theorem 5 under* $\bar{X}_\pi$, *i.e.*

$$\tilde{\upsilon}_{m,n}(t,\hat{\delta};\bar{X}_\pi) = \upsilon_{m,n}(t,\delta;\bar{X}_\pi) - \psi_g\left(\upsilon_{m,n}(t,\delta;\bar{X}_\pi)\right) + o_p(1) \tag{B.11}$$

for any $0 \le t \le 1$.

Proof. Observe that $\upsilon_{m,n}(t,\hat{\delta},\bar{X}_\pi) = V_{m,n}(F_0^{-1}(t),\delta;\bar{X}_\pi)$, therefore it suffices to show that the remainder defined in equation (A.5) in the proof of Theorem 1 is still $o_p(1)$ under X_π.

Following the contiguity result of Chung and Romano (2013), let $E_1, E_2, \ldots,$ be i.i.d. from the mixture distribution $\bar{P} = p\tilde{F}_1 + (1-p)F_0$. Observe that under the null hypothesis

$$\mathbb{1}_{\{E_i \le y\}} - F_1(y) = \left(\mathbb{1}_{\{E_i \le y\}} - \bar{P}(y)\right) + \underbrace{p\left(\tilde{F}_1(y) - F_0(y)\right)}_{o_p(1)}$$

Then the remainder satisfies

$$\sqrt{\frac{mn}{N}}\left\{\left(\frac{1}{m}\sum_{i=i}^{m}\mathbb{1}_{\{E_i\leq y+\hat{\delta}\}}-F_1(y+\hat{\delta})\right)-\left(\frac{1}{m}\sum_{i=i}^{m}\mathbb{1}_{\{E_i\leq y+\delta\}}-F_1(y+\delta)\right)\right\}\xrightarrow{\text{p}}0$$

by stochastic equicontinuity of $\{m^{-1}\sum_{i=i}^{m}\mathbb{1}_{\{E_i\leq y\}}-\bar{P}(y):y\in\mathbb{R}\}$. As a result

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

by Lemma 5.3 of Chung and Romano (2013). This finishes the proof. □

C 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. according to a probability distribution F_1, treatment group. Let $N=n+m$ and write

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

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 behavior of the permutation distribution based on Z should behave approximately like the permutation distribution based on a sample of N i.i.d. observations $\bar{Z}=(\bar{Z}_1,\ldots,\bar{Z}_N)$ from the mixture distribution $pF_1+(1-p)F_0$.

We would wish to compare

$$\bar{Z}=(\bar{Z}_1,\cdots,\bar{Z}_N)\quad\text{vs}\quad Z=(Z_1,\ldots,Z_N)$$

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 13. 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 $pF_1+(1-p)F_0$.

Remark 14. 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.

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 have 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 access this R file to see how this algorithm works.

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 indexes i except for D of them.

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 Kolmogorov–Smirnov statistic, $K_{m,n,\delta}(Z)$, such that the difference between $K_{m,n,\delta}(Z) - K_{m,n,\delta}(\bar{Z}_{\pi_0})$ 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 $pF_1 + (1-p)F_0$.

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

$$\left(K_{m,n,\delta}(\bar{Z}_\pi), K_{m,n,\delta}(\bar{Z}_{\pi'})\right) \xrightarrow{\text{d}} (T, T') \tag{C.2}$$

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 i.i.d. according to probability distribution $\bar{P}$. 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

$$K_{m,n,\delta}(\bar{Z}_{\pi\pi_0}) - K_{m,n,\delta}(\bar{Z}_\pi) \xrightarrow{\text{p}} 0 \tag{C.3}$$

then it also follows that

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

and by Slutsky's theorem

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

we can conclude that $(K_{m,n,\delta}(Z_\pi), K_{m,n,\delta}(Z_{\pi'})) \xrightarrow{\text{d}} (T, T')$. 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. The case when δ in unknown follows from the same arguments and we therefore omit the details.

D Multiple Testing Procedures

For completeness, we present the Westfall–Young max T, and the Holm's step-down algorithms as alternatives to the min P procedure for p-value multiple testing adjustment (see Westfall and Young, 1993, Chapter 2). We note that the max T Algorithm is computationally faster than the min P procedure since we do not need to calculate the p-values as in Algorithm 1, whereas the computation gains in Holm's procedure come from the fact we only have one level of permutation (the one needed for the calculation of the p-values).

Denote $p_1, \ldots, p_{\mathcal{J}}$ the p-values of the $\mathcal{J}$ individual permutation tests for (25) based on the martingale-transformed KS statistic $\tilde{K}^{m,n,\hat{\delta}}$, and the ordered values of the statistics $\tilde{K}_{r_1} \geq \cdots \geq \tilde{K}_{r_{\mathcal{J}}}$. Define $\mathcal{T}_j = \{r_j, r_{j+1} \ldots, r_{\mathcal{J}}\}$ and let $g_{b,j}$ for $1 \leq j \leq \mathcal{J}$ be a random permutation of $\{1, \ldots, m_j + n_j\}$.

Algorithm 2 (Westfall–Young's max T)

1. *For each permutation* $b = 1, \ldots, B < \min_{1 \leq j \leq \mathcal{J}}\{(m_j + n_j)!\}$*:*

   (i) *Apply action* $g_{b,j}$ *to every subgroup* Z_j, $1 \leq j \leq \mathcal{J}$*:* $(g_{b,1}Z_1, \ldots, g_{b,\mathcal{J}}Z_{\mathcal{J}})$*, with corresponding statistics* $\tilde{K}_j^{(b)}$ *for* $1 \leq j \leq \mathcal{J}$*.*

   (ii) *Let*

   $$\hat{K}_{r_1}^{(b)} = \max_{j \in \mathcal{T}_1} \tilde{K}_j^{(b)} \ , \ \hat{K}_{r_2}^{(b)} = \max_{j \in \mathcal{T}_2} \tilde{K}_j^{(b)} \ , \ \ldots, \ \hat{K}_{r_{\mathcal{J}}}^{(b)} = \tilde{K}_{r_{\mathcal{J}}}^{(b)} \ .$$

2. *Define*

   $$\mathcal{H}_1 = \#\{\tilde{K}_{r_1} \leq \hat{K}_{r_1}^{(b)} : 1 \leq b \leq B\} \ , \ \ldots, \ \mathcal{H}_{\mathcal{J}} = \#\{\tilde{K}_{r_{\mathcal{J}}} \leq \hat{K}_{r_{\mathcal{J}}}^{(b)} : 1 \leq b \leq B\} \ .$$

3. *The adjusted p-values are given by*

   $$p_{r_1}^* = \frac{\mathcal{H}_1}{B} \ , \ p_{r_2}^* = \max\left\{p_{r_1}^*, \frac{\mathcal{H}_2}{B}\right\} \ , \ \ldots, \ p_{r_{\mathcal{J}}}^* = \max\left\{p_{r_{\mathcal{J}-1}}^*, \frac{\mathcal{H}_{\mathcal{J}}}{B}\right\} \ ,$$

   In order to control the test FWER at level α*, each adjusted p-value needs to be now compared with* α*.*

Let $p_{r_1} \leq \cdots \leq p_{r_{\mathcal{J}}}$ be the ordered p-values, with their respective associated hypotheses $H_{0,r_1}, \ldots, H_{0,r_{\mathcal{J}}}$. The following stepdown algorithm, due to Holm (1979), can be described as follows:

Algorithm 3 (Holm)

1. *If* $p_{r_1} \geq \alpha/\mathcal{J}$, *accept* $H_{0,r_1}, \ldots, H_{0,r_\mathcal{J}}$ *and stop. If* $p_{r_1} < \alpha/\mathcal{J}$, *reject* H_{0,r_1} *and test the remaining* $\mathcal{J}-1$ *hypotheses at level* $\alpha/(\mathcal{J}-1)$.

2. *If* $p_{r_1} < \alpha/\mathcal{J}$, *but* $p_{r_2} \geq \alpha/(\mathcal{J}-1)$, *accept* $H_{0,r_2}, \ldots, H_{0,\mathcal{J}}$ *and stop. If* $p_{r_1} < \alpha/\mathcal{J}$ *and* $p_{r_2} < \alpha/(\mathcal{J}-1)$, *reject* H_{0,r_2} *and test the remaining* $\mathcal{J}-2$ *hypotheses at level* $\alpha/(\mathcal{J}-2)$.

$\vdots$

j. *If* $p_{r_1} < \alpha/\mathcal{J}, \ldots, p_{r_{j-1}} < \alpha/(\mathcal{J}-j+2)$, *but* $p_{r_j} \geq \alpha/(\mathcal{J}-j+1)$, *accept* $H_{0,r_j}, \ldots, H_{0,\mathcal{J}}$ *and stop. If* $p_{r_1} < \alpha/\mathcal{J}, \ldots, p_{r_j} < \alpha/(\mathcal{J}-j+1)$, *reject* H_{0,r_j} *and test the remaining* $\mathcal{J}-j$ *hypotheses at level* $\alpha/(\mathcal{J}-j)$.

$\vdots$

$\mathcal{J}$. *If* $p_{r_\mathcal{J}} \geq \alpha$, *accept* $H_{0,r_\mathcal{J}}$, *otherwise reject* $H_{0,r_\mathcal{J}}$.