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

August 17, 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 control and experimental outcomes from a randomized trial, 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 (2016b) and references therein for more examples of the same idea.

[3] The 2019 Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 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. (2016), 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. (2016), 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. (2016), 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. (2016) 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. Throughout the paper, we only consider completely randomized experiments, *i.e.*, covariates are not used to inform the treatment assignment.[5] 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.[6] 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

[5] A more detailed review of different randomization schemes can be found in Hu et al. (2014). See also Bugni et al. (2018); Ma et al. (2019) and the references therein.

[6] 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.

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.

More formally, Let $Y_{1,1}, \ldots, Y_{1,m}$ and $Y_{0,1}, \ldots, Y_{0,n}$ be two independent random samples having distribution functions $F_1(\cdot)$ and $F_0(\cdot)$, respectively.[7] The (testable) constant treatment effect null hypothesis becomes:

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

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 observations belonging to treatment and control groups, $F_1(\cdot)$ and $F_0(\cdot)$, 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 $F^{-1}(\tau) = \inf\{y : F(y) \geq \tau\}$, as usual. As a result, the constant treatment effect null hypothesis boils down to suppressing 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

[7] Thus, $Y_{1,i} = Y_i$ among the treated, and $Y_{0,i} = Y_i$ among the non-treated.

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.

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

One natural candidate for a test statistic for hypothesis (2) is to compare the empirical CDFs based on two independent random samples. To fix notation, suppose that outcome Y_1 is drawn from a distribution with CDF F_1, and similarly, Y_0 is drawn from a distribution with CDF F_0.

Let $Y_{1,1}, \ldots, Y_{1,m}$ and $Y_{0,1}, \ldots, Y_{0,n}$ be two independent random samples from F_1 and F_0. Collect the observed data in $Z = (Z_1, \ldots, Z_N)$ as follows

$$Z = (Y_{1,1}, \ldots, Y_{1,m}, Y_{0,1}, \ldots, Y_{0,n}) \ .$$

Consider the empirical CDFs

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

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

$$\hat{\delta} = \frac{1}{m} \sum_{i=1}^{m} Y_{1,i} - \frac{1}{n} \sum_{i=1}^{n} Y_{0,i} \ .$$

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

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

where

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

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

$$K^u_{m,n,\hat{\delta}}(Z) = \sup_{0 \leq t \leq 1} \left| \upsilon_{m,n}(t, \hat{\delta}; Z) \right| \ , \tag{6}$$

where

$$\upsilon_{m,n}(t,\hat{\delta};Z) = \sqrt{\frac{mn}{N}}\left(\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}_{\left\{Y_{1,i}-\hat{\delta}\le F_0^{-1}(t)\right\}} - \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\left\{Y_{0,i}\le F_0^{-1}(t)\right\}}\right) \tag{7}$$
$$= \sqrt{\frac{mn}{N}}\left(\hat{F}_1(F_0^{-1}(t)+\hat{\delta}) - \hat{F}_0(F_0^{-1}(t))\right)$$
$$= V_{m,n}(F_0^{-1}(t),\hat{\delta};Z)\ .$$

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*.[8]

In the classical case with δ known, we are able to determine all potential outcomes as well as the exact null distribution. With this in mind, we calculate the two-sample Kolmogorov–Smirnov statistic and then permute the data many times, computing the statistic on each permutation. The empirical distribution of the values of the statistic recalculated over these permutations of the data serves as a null distribution; this leads to a permutation test that is exact level α in finite samples.

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

$$Z^* = \left(Y_{1,1}-\delta,\ldots,Y_{1,m}-\delta,Y_{0,1},\ldots,Y_{0,n}\right)\ , \tag{8}$$

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

$$K_{m,n,\delta}(Z^*) = \sup_y |V_{m,n}(y,\delta;Z^*)|\ , \tag{9}$$

where

$$V_{m,n}(y,\delta;Z^*) = \sqrt{\frac{mn}{N}}\left(\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}_{\{Z_i^*\le y\}} - \frac{1}{n}\sum_{i=m+1}^{N}\mathbb{1}_{\{Z_i^*\le y\}}\right) \tag{10}$$

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

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

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

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

$$M^+(z) = \left|\{1 \le j \le N! : K^{(j)}_{m,n,\delta}(z) > K^{(k)}_{m,n,\delta}(z)\}\right|$$
$$M^0(z) = \left|\{1 \le j \le N! : K^{(j)}_{m,n,\delta}(z) = K^{(k)}_{m,n,\delta}(z)\}\right| .$$

Using this notation, the permutation test is given by

$$\phi(z) = \begin{cases} 1 & K_{m,n,\delta}(z) > K^{(k)}_{m,n,\delta}(z) \\ a(z) & K_{m,n,\delta}(z) = K^{(k)}_{m,n,\delta}(z) \\ 0 & K_{m,n,\delta}(z) < K^{(k)}_{m,n,\delta}(z) \end{cases} , \tag{11}$$

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 (11) for the sharp null hypothesis satisfies

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

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

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

where $\upsilon_{m,n}(t, \delta; Z^*) = V_{m,n}(F_0^{-1}(t), \delta; Z^*)$. 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 finite-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 (Romano, 1989, Section 4). 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 the randomization hypothesis does not hold 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 $V_{m,n}(\cdot, \hat{\delta}; Z)$, 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 $\sigma_0^2 = \mathbb{V}(Y_{0,i}) < \infty$. Consider the process $\mathbb{B} = \mathbb{G} + \mathbb{S}$ with covariance structure

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

The following theorem establishes the asymptotic behavior of the two-sample Kolmogorov–Smirnov statistic. It is due to Theorem 4 of Ding et al. (2016) for a suitably scaled variation of their test statistic, but we include here for completeness.

Theorem 1. *Assume $Y_{0,1}, \ldots, Y_{0,n}$ are i.i.d. according to a probability distribution F_0, and independently $Y_{1,1}, \ldots, Y_{1,m}$ 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 assumptions A.1–A.2, $K_{m,n,\hat{\delta}}$ converges weakly under the null hypothesis to*

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

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 (see also Lemma B.3), the smoothness condition A. 2 allows us to expand $V_{m,n}(y, \hat{\delta}; Z)$ around δ to obtain

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

Observe that the first summand is the classical two-sample empirical process, 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. ■

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

$$\hat{R}^{K(\hat{\delta})}_{m,n}(t) = \frac{1}{N!} \sum_{\pi \in \mathbf{G}_N} \mathbb{1}_{\{K^u_{m,n,\hat{\delta}}(z_{\pi(1)}, \ldots, z_{\pi(N)}) \leq t\}} \ . \tag{14}$$

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

Since $K^u_{m,n,\hat{\delta}}$ is not asymptotically pivotal as shown in Theorem 1, one can deduce that 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 $K^u_{m,n,\hat{\delta}}$, does not behave like the true unconditional limiting distribution asymptotically, as shown in the following theorem. Note that the null hypothesis is not assumed.

Theorem 2. *Assume $Y_{0,1}, \ldots, Y_{0,n}$ are i.i.d. according to a probability distribution F_0, and independently $Y_{1,1}, \ldots, Y_{1,m}$ are i.i.d. according to a probability distribution F_1. Consider testing the hypothesis (2) for some δ based on the test statistic (6). If assumptions A.1–A.2 hold, then the permutation distribution (14) based on $K^u_{m,n,\hat{\delta}}(Z)$ is such that*

$$\sup_t \left| \hat{R}^{K(\hat{\delta})}_{m,n}(t) - J_0(t) \right| \xrightarrow{\mathrm{p}} 0 \ ,$$

where $J_0(\cdot)$ denotes the CDF of $K^u_0 \equiv \sup_t |\mathbb{U}(t)|$. Here $\mathbb{U}(\cdot)$ is the uniform Brownian bridge on $[0, 1]$.

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 $K^u_{m,n,\hat{\delta}}$ when δ is estimates is asymptotically equivalent to the permutation distribution based on $K^u_{m,n,\delta}$ when δ is known.[9] Intuitively, this resemblance occurs because in both cases, the permutation distribution is treating the observations as if they were i.i.d. ■

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.

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 $\dot{g}(s) = (1, \dot{f}_0(F_0^{-1}(s))/f_0(F_0^{-1}(s)))$. 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

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

where $C(s) = \int_s^1 \dot{g}(t)\dot{g}(t)' dt$. Arguing as in Parker (2013), we can think of $\psi_g(h)(\cdot)$ as the functional equivalent of the fitted values in a linear regression, where the extended score $\dot{g}(s)$

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

acts as the regressor, and $C(s)^{-1}\int_s^1 \dot{g}(r)dh(r)$ 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 (7) is given by

$$\begin{aligned}\tilde{\upsilon}_{m,n}(t,\hat{\delta};Z) &= \upsilon_{m,n}(t,\hat{\delta};Z) - \int_0^t \left[\dot{g}(s)'C(s)^{-1}\int_s^1 \dot{g}(r)d\upsilon_{m,n}(r,\hat{\delta};Z)\right]ds \\ &= \upsilon_{m,n}(t,\hat{\delta};Z) - \psi_g(\upsilon_{m,n}(t,\hat{\delta};Z)) \ . \end{aligned} \tag{15}$$

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

$$\tilde{K}_{m,n,\hat{\delta}}(Z) = \sup_{0\le t\le 1}\left|\tilde{\upsilon}_{m,n}(t,\hat{\delta};Z)\right| \ . \tag{16}$$

The martingale-transformed statistic (16) 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 (16), and then we show the asymptotic behavior of the proposed permutation test.

The following theorem states the limit behavior of (16). It essentially follows from an extension of Khmaladze (1981) to the two-sample case, where we show that $\tilde{\upsilon}_{m,n}(\cdot,\hat{\delta};Z)$ converges weakly to a Brownian motion process $\mathbb{M}$ under the null hypothesis, effectively nullifying the effect of the estimated nuisance parameter $\hat{\delta}$.

Theorem 3. *Assume $Y_{0,1},\ldots,Y_{0,n}$ are i.i.d. according to a probability distribution F_0, and independently $Y_{1,1},\ldots,Y_{1,m}$ are i.i.d. according to a probability distribution F_1. Consider* *testing the hypothesis* (2) *for some* δ *based on the test statistic* (16)*. Under assumptions A.1–A.2,* $\tilde{K}_{m,n,\hat{\delta}}$ *converges weakly under the null hypothesis to*

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

with CDF denoted by $J_2(\cdot)$*. Here* $\mathbb{M}$ *is the standard Brownian motion given by* $\mathbb{M} = \mathbb{U} - \psi_g(\mathbb{U})$*, and* $\mathbb{U}$ *is the standard Brownian bridge on* $[0,1]$*.*

To gain further intuition as to why this transformation works, observe that the mapping $\psi_g(h)(\cdot)$ 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 (15) as

$$\begin{aligned}\tilde{\upsilon}_{m,n}(t,\hat{\delta};Z) &= \upsilon_{m,n}(t,\hat{\delta};Z) - \psi_g(\upsilon_{m,n})(t,\hat{\delta};Z)\\ &= \underbrace{\upsilon_{m,n}(t,\delta;Z^*)}_{\xrightarrow{\mathrm{d}}\mathbb{U}(t)} - \underbrace{\psi_g(\upsilon_{m,n})(t,\delta;Z^*)}_{\xrightarrow{\mathrm{d}}\psi_g(\mathbb{U})(t)} + o_p(1)\ .\end{aligned}$$

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).

We seek the limiting behavior of $\hat{R}^{\tilde{K}(\hat{\delta})}_{m,n}$—the permutation distribution (14) based on the Khmaladze transformed statistic $\tilde{K}_{m,n,\hat{\delta}}$—and its upper α-quantile, which we now denote $\hat{r}_{m,n}$, where

$$\hat{r}_{m,n}(1-\alpha) = \inf\{t : \hat{R}^{\tilde{K}(\hat{\delta})}_{m,n}(t) \ge 1-\alpha\}\ .$$

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 $\tilde{K}_{m,n,\hat{\delta}}$. Consequently, the α-upper quantiles $\hat{r}_{m,n}$ can be used as "critical values" for $\tilde{K}_{m,n,\hat{\delta}}$. Note that the null hypothesis is not assumed.

Theorem 4. *Assume* $Y_{0,1},\ldots,Y_{0,n}$ *are i.i.d. according to a probability distribution* F_0*, and independently* $Y_{1,1},\ldots,Y_{1,m}$ *are i.i.d. according to a probability distribution* F_1*. Consider testing the hypothesis (2) for some* δ *based on the test statistic* (16)*. Under assumptions A.1–A.2, the permutation distribution* (14) *based on the Khmaladze transformed statistic* $\tilde{K}_{m,n,\hat{\delta}}$ *is such that*

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

where $J_2(\cdot)$ *is the CDF of* K_2 *defined in Theorem 3. Let* $r(1-\alpha) = \inf\{t : J_2(t) \ge 1-\alpha\}$*. Then*

$$\hat{r}_{m,n}(1-\alpha) \xrightarrow{\mathrm{p}} r(1-\alpha)\ .$$

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 $r_{m,n} \to r(1-\alpha) = J_2^{-1}(1-\alpha)$. Assume that $\tilde{K}_{m,n,\hat{\delta}}$ weakly converges to some limit law $J_2'(\cdot)$ under some sequence of alternatives that are contiguous to some distribution satisfying the null hypothesis. Then the power of the test would tend to $1 - J_2'(J_2^{-1}(1-\alpha))$. Thus, under the premises of the preceding Theorems 3 and 4, we have that $\hat{r}_{m,n}$, obtained from the permutation distribution, satisfies $\hat{r}_{m,n} \xrightarrow{\mathrm{P}} J_2^{-1}(1-\alpha)$. 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. ■

Remark 9. From the construction of the permutation test in (11) based on $\tilde{K}_{m,n,\hat{\delta}}$, we have

$$\Pr\left\{\tilde{K}_{m,n,\hat{\delta}} > \hat{r}_{m,n}\right\} \leq \mathbb{E}\left[\phi(Z)\right] \leq \Pr\left\{\tilde{K}_{m,n,\hat{\delta}} \geq \hat{r}_{m,n}\right\} \ .$$

Hence it follows that Theorem 4 implies $\mathbb{E}\left[\phi(Z)\right] \to \alpha$. See Lehmann and Romano (2005, Section 15.2.2). ■

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 (15) with t taking discrete values, replacing integral with sums. For instance, suppose $0 = t_0 < t_1 < \cdots < t_q < t_{q+1} = 1$ is a partition of the interval $[0,1]$ and that t takes on values on $t_1, t_2, ..., t_q$. Write (15) in differentiation form

$$d\tilde{v}_{m,n}(t,\hat{\delta};Z) = dv_{m,n}(t,\hat{\delta};Z) - \dot{g}(t)'C(t)^{-1}\int_t^1 \dot{g}(r)dv_{m,n}(r,\hat{\delta};Z)dt \ . \tag{17}$$

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

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

then the right hand side of (17) can be interpreted as the recursive residuals:

$$y_i - x_i' \left(\sum_{k=i}^{q+1} x_k x_k' \right)^{-1} \sum_{k=i}^{q+1} x_k y_k = y_i - x_i' \hat{\boldsymbol{\beta}}_i \ , \tag{18}$$

where $\hat{\boldsymbol{\beta}}_i$ is the OLS estimator based on the last $q - i + 2$ observations. The cumulative sum (integration from $[0, t_i)$) of the 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 q. Stack y_i and x_i in the following manner

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

where $\dot{g}_1(s) = 1$ and $\dot{g}_2(s) = \dot{f}_0(F_0^{-1}(s))/f_0(F_0^{-1}(s))$. The OLS estimator based on the last $q - i + 2$ observations described on right hand side of (18) can be written as

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

which implies that the Khmaladze transformation of the empirical process in (15) can be obtained by numerically integrating from $[0, t_i)$, i.e.

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

and therefore the test statistic can be calculated as

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

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

$$\sup_{0\leq t\leq 1} |g_{2,n}(t) - g_2(t)| = o_p(1) \quad \text{and} \tag{19}$$

$$\sup_{0\leq t\leq 1} |\dot{g}_{2,n}(t) - \dot{g}_2(t)| = o_p(1) \; . \tag{20}$$

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

Remark 10. 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 (19)–(20) (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[10]. 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 assume that the mutually exclusive subgroups are formed from observed covariates, and are

[10] 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 $F_0^j(\cdot)$ and $F_1^j(\cdot)$ be the CDFs of the observations in control and treatment groups for subgroup $1 \leq j \leq \mathcal{J}$. The null hypothesis of interest is given by the joint hypothesis

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

This section treats the testing of $\mathbf{H}_0$ as a multiple testing problem in which every individual hypothesis $j \in \{1, \ldots, \mathcal{J}\}$, given by

$$H_{0,j} : F_1^j(y + \delta_j) = F_0^j(y) \ , \ \text{for some } \delta_j \ , \tag{21}$$

specifies whether the treatment effect is heterogeneous for a particular subgroup[11]. We reject the null hypothesis $\mathbf{H}_0$ if any one of the null hypotheses for a subgroup $j \in \{1, \ldots, \mathcal{J}\}$ is rejected.

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.[12]

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's FWER level (see Westfall and Young, 1993, Chapter 2). Observed data for each mutually exclusive subgroup is given by

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

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 (21), and the ordered p-values $p_{r_1} \leq \cdots \leq p_{r_{\mathcal{J}}}$, with their respective associated hypotheses of the form (21) 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* $1 \leq j \leq \mathcal{J}$*.*

[11] 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.

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

(ii) Let

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

2. *Define*

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

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

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

4. *Each adjusted p-value $p^*_{r_j}$—with associated hypothesis H_{0,r_j}—is now compared with α, for $1 \leq j \leq \mathcal{J}$, i.e., if $p^*_{r_j} \geq \alpha$ then we fail to reject, otherwise reject H_{0,r_j}.*

We reject the null hypothesis $\mathbf{H}_0$ if any one of the null hypotheses for a subgroup $j \in \{1, \ldots, \mathcal{J}\}$ is rejected.

Remark 11. 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 12. 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 13. 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 lies at the center of our testing procedure. 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), and Bitler et al. (2017) adjustment is more conservative for they use Bonferroni bounds. Lastly, our approach is based on the two-sample, martingale-transformed empirical process. 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. (2016). 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 it 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 $N \in \{13, 50, 80, 200, 800, 1000\}$. 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 is the permutation test based on the classical two-sample Kolmogorov–Smirnov statistic of Section 2.2. Even though this is an infeasible test—we do not know the true value of δ in practice—we present it here 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. (2016). 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 introduced 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. (2016), 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 ($\sigma_\delta = 0$)[13]. 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 Section 2.2, 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 unknown 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 sizes increase. Subsampling delivers rejection probabilities under the null hypothesis less than

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

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.[14]

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 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.

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 assumption in this model is 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 fund-

[14] For a comparison between the bootstrap and subsampling tests, see Linton et al. (2005). The authors note that the bootstrap mimics the asymptotic null distribution in the least favorable case, which is a subset of the boundary of the null where the marginal distribution functions are equal. However, the boundary is composite, implying that tests based on the approximation of the least favorable case are not asymptotically similar on this boundary. Subsampling, on the other hand, approximates the true sampling distribution under the composite null hypothesis and thus these tests are asymptotically similar on the boundary, resulting in an asymptotically more powerful test for some local alternatives.

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
	Classic KS	0.0511	0.0514	0.0518
$N = 800$	Naive KS	0.0000	0.4340	0.0045
$n = 500$	FRT CI	0.0398	0.0405	0.0350
$m = 300$	Subsampling	0.0480	0.0048	0.0125
	Bootstrap	0.0908	0.0656	0.0865
	mtPermTest	0.0288	0.0470	0.0438
	Classic KS	0.0498	0.0498	0.0476
$N = 1000$	Naive KS	0.0004	0.4362	0.0045
$n = 600$	FRT CI	0.0348	0.0458	0.0555
$m = 400$	Subsampling	0.0452	0.0070	0.0104
	Bootstrap	0.0920	0.0680	0.0824
	mtPermTest	0.0292	0.0480	0.0474

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

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

raising 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 7 report the Khmaladze transformed test statistic (16), 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[15]. Furthermore, the same conclusion holds 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

[15] 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.

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.

Beran, R. and Millar, P. (1986). Confidence sets for a multivariate distribution. *The Annals of Statistics*, pages 431–443.

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.

Bugni, F. A., Canay, I. A., and Shaikh, A. M. (2018). Inference under covariate-adaptive randomization. *Journal of the American Statistical Association*, 113(524):1784–1796.

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. (2016a). Asymptotically valid and exact permutation tests based on two-sample u-statistics. *Journal of Statistical Planning and Inference*, 168:97–105.

Chung, E. and Romano, J. P. (2016b). 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. (2016). 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.

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.

Hu, F., Hu, Y., Ma, Z., and Rosenberger, W. F. (2014). Adaptive randomization for balancing over covariates. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(4):288–303.

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.

Ma, W., Qin, Y., Li, Y., and Hu, F. (2019). Statistical inference for covariate-adaptive randomization procedures. *Journal of the American Statistical Association*, (just-accepted):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.

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 0.1.7.

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

Pollard, D. (1984). *Convergence of stochastic processes*. Springer Science & Business Media.

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

Præstgaard, J. T. (1995). Permutation and bootstrap kolmogorov-smirnov tests for the equality of two distributions. *Scandinavian Journal of Statistics*, pages 305–322.

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.

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

The classes $\mathcal{F}$ in all of the applications in this Appendix are collections of indicator functions of lower rectangles in $\mathbb{R}$. Thus, the empirical processes in this paper can be viewed as random maps into $\ell^\infty(\mathcal{F})$—the space of all bounded functions on $\mathbb{R}$ equipped with the uniform norm—and weak convergence is understood as convergence in distribution in $\ell^\infty(\mathcal{F})$. We are going to assume that the class $\mathcal{F}$ is pointwise measurable (Van der Vaart and Wellner, 1996, Example 2.3.4), ruling out measurability problems with regards suprema.

Throughout this appendix, if ξ is a random variable defined on a probability space $(\Omega, \mathcal{B}, P)$, it is assumed that $\xi_1, \ldots, \xi_N$ are coordinate projections on the product space $(\Omega^N, \mathcal{B}^N, P^N)$, and the expectations are computed for P^N. If auxiliary variables, independent of the ξs, are involved—as in Lemma B.1—we use a similar convention. In that case, the underlying probability space is assumed to be of the form $(\Omega^N, \mathcal{B}^N, P^N) \times (\mathcal{Z}, \mathcal{C}, Q)$, with $\xi_1, \ldots, \xi_N$ equal to the coordinate projections on the first N coordinates and the additional variables depending only on the $N+1$st coordinate.

Symbols $\mathcal{O}_p(1)$ and $o_p(1)$ stand for being bounded in probability and convergence to zero in probability, respectively. All vector are column vectors. We use $\lfloor \cdot \rfloor$ to denote the largest smaller integer, and $a \wedge b = \min\{a, b\}$. We use $\xrightarrow{\text{p}}$ to denote convergence in probability, and $\xrightarrow{\text{d}}$ to denote convergence in distribution, respectively. For two random variables ξ and η, write $\xi \overset{\text{d}}{=} \eta$ if they have the same distribution. 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 $\mathbb{B} = \mathbb{G} + \mathbb{S}$. Similarly, $\mathbb{B}_1 = \mathbb{G}_1 + \mathbb{S}$.
$\widetilde{\mathbb{B}}$	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_0^u	For $t \in [0,1]$, $K_0^u = \sup_t \lvert \mathbb{U}(t) \rvert$. Its CDF is given by $J_0(\cdot)$
K_1	For $y \in \mathbb{R}$, $K_1 = \sup_y \lvert \mathbb{B}(y) \rvert$.
K_2	For $t \in [0,1]$, $K_2 = \sup_t \lvert \mathbb{M}(t) \rvert$. Its CDF is given by $J_2(\cdot)$

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 $K_{m,n,\delta}$. Second, we show that the the permutation distribution based on the classical two-sample Kolmogorov–Smirnov statistic asymptotically behaves like the true unconditional limiting distribution. Note that the null hypothesis is not assumed in the second theorem. In order to deduce this second result, we follow Hoeffding (1952) approach. See also Lehmann and Romano (2005, Theorem 15.2.3) and Chung and Romano (2013, Lemma 5.1).

Theorem A. 1. *Assume $Y_{0,1}, \ldots, Y_{0,n}$ are i.i.d. according to a probability distribution F_0, and independently $Y_{1,1}, \ldots, Y_{1,m}$ 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 (9). Under assumption A.1, $K_{m,n,\delta}$ converges weakly under the sharp null hypothesis to*

$$K_0 \equiv \sup_y |\mathbb{G}(y)| \ .$$

Moreover, if the test statistic is replaced by (12) and assumptions A.1–A.2 hold, then $K^u_{m,n,\delta}(Z)$ converges weakly under the sharp null hypothesis to

$$K^u_0 \equiv \sup_{0 \le t \le 1} |\mathbb{U}(t)| \ .$$

Theorem A. 2. *Assume $Y_{0,1}, \ldots, Y_{0,n}$ are i.i.d. according to a probability distribution F_0, and independently $Y_{1,1}, \ldots, Y_{1,m}$ 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 (12). If assumptions A.1–A.2 hold, then the permutation distribution (14) based on $K^u_{m,n,\delta}(Z^*)$ is such that*

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

where $J_0(\cdot)$ denotes the CDF of K^u_0 defined in Theorem A.1.

Remark A.1. The permutation distribution based on $K^u_{m,n,\delta}$ asymptotically 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. Intuitively, if we assume condition A. 2 the process $\upsilon_{m,n}(\cdot, \delta; Z^*)$ becomes the uniform empirical process, rendering the statistic $K^u_{m,n,\delta}$ a pivotal quantity. This latter property is the key to establishing the asymptotic validity of the permutation test based on $K^u_{m,n,\delta}$. ■

A.1 Proof of Theorem A.1

We first prove $K_{m,n,\delta}$ converges weakly under the sharp null hypothesis to K_0. Note that the maps $z \to \|z\|$ from $\ell^\infty(\mathcal{F})$ into $\mathbb{R}$ are continuous with respect to the supremum norm. Thus by virtue of the continuous mapping theorem, it suffices to show that $V_{m,n}(\cdot,\delta;Z^*)$ converges weakly in $\ell^\infty(\mathcal{F})$ to $\mathbb{G}$ under the null hypothesis.

Consider the following derivation

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

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

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

It is known from classical results that the class $\mathcal{F}$ of lower rectangles is Donsker (Van der Vaart and Wellner, 1996, Section 2.1). Then under our assumptions, $V_{1,m}$ and $V_{0,n}$ converge weakly in $\ell^\infty(\mathcal{F})$ to two Gaussian processes, $\mathbb{G}$ and $\mathbb{G}_1$, respectively (Van der Vaart, 2000, Theorem 19.3). Since $V_{1,m}$ and $V_{0,n}$ are uncorrelated and the joint marginals converge to multivariate normal distributions, the sequence converges jointly to a vector of independent (tight) Brownian bridges $\mathbb{G}$ and $\mathbb{G}_1$. Then, $V_{m,n}$ weakly converges to $\widetilde{\mathbb{G}}$, where the limit variable $\widetilde{\mathbb{G}}$ possesses the same distribution as $\mathbb{G}$ under the null hypothesis. This concludes the proof of the first part of the Theorem.

We next prove that $\upsilon_{m,n}(\cdot,\delta;Z^*)$ 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 noting F_0^{-1} is well defined by assumption A. 2, and write $\upsilon_{m,n}$ as follows

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

Under our assumptions and the independence of the empirical processes $V_{1,m}$ and $V_{0,n}$, $\upsilon_{m,n}(\cdot,\delta;Z^*)$ 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

Independent of the Z^*s, 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 seek to show that

$$\left(K^u_{m,n,\delta}(Z^*_\pi), K^u_{m,n,\delta}(Z^*_{\pi'})\right) \xrightarrow{\mathrm{d}} \left(K^u_0, K^{u'}_0\right) \ , \tag{A.1}$$

where K^u_0 and $K^{u'}_0$ are independent with common CDF $J_0(\cdot)$. Then Hoeffding's Condition (Lehmann and Romano, 2005, Theorem 15.2.3) implies that

$$\sup_t \left| \hat{R}^{K(\delta)}_{m,n}(t) - J_0(t) \right| \xrightarrow{\mathrm{p}} 0 \ ,$$

completing the proof of the theorem. In the following, we prove (A.1) in two steps.

Step 1. Apply the coupling construction of Chung and Romano (2013). More specifically, couple data Z^* with an auxiliary sample of N i.i.d. observations $\bar{Z} = (\bar{Z}_1, \ldots, \bar{Z}_N)$ from the mixture distribution with CDF $\bar{P} = pF^\delta_1 + (1-p)F_0$, where $p = \lim_{m\to\infty} m/N$ and F^δ_1 is given by $F^\delta_1(y) = F_1(y+\delta)$. See Appendix C for a detailed exposition of the coupling construction.

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

$$\left(K^u_{m,n,\delta}(\bar{Z}_\pi), K^u_{m,n,\delta}(\bar{Z}_{\pi'})\right) \xrightarrow{\mathrm{d}} \left(K^u_0, K^{u'}_0\right) \tag{A.2}$$

$$K^u_{m,n,\delta}(\bar{Z}_{\pi,\pi_0}) - K^u_{m,n,\delta}(Z^*_\pi) \xrightarrow{\mathrm{p}} 0 \ , \tag{A.3}$$

where the permutation π_0 is properly defined in Appendix C. Lemma B.1 establishes (A.2), where K^u_0, $K^{u'}_0$ are independent with common CDF $J_0(\cdot)$, whereas (A.3) is the content of Lemma B.2.

A.3 Proof of Theorem 1

We start by proving that $V_{m,n}(y, \hat{\delta}; Z)$ converges weakly in $\ell^\infty(\mathcal{F})$ to a (tight) Gaussian process $\mathbb{B}$ given by $\mathbb{B} = \mathbb{G} + \mathbb{S}$ with covariance structure as in (13). We break this claim into four steps.

Step 1. Given our assumptions, we show in Lemma B.3 that $V_{m,n}(y, \hat{\delta}; Z)$ has the following asymptotic representation

$$V_{m,n}(y, \hat{\delta}; Z) = \sqrt{1-p_m}\, B_{m,1}(y) - \sqrt{p_m}\, B_{n,0}(y) + o_p(1) \ ,$$

where the $o_p(1)$ term holds uniformly over $y \in \mathbb{R}$, $p_m = m/N$, and

$$B_{m,1}(y) = \frac{1}{\sqrt{m}} \sum_{i=1}^m \left\{ \mathbb{1}_{\{Y_{1,i} \le y+\delta\}} - F_1(y+\delta) + f_0(y)\Big(Y_{1,i} - \mathbb{E}(Y_{1,i})\Big) \right\} \tag{A.4}$$

$$B_{n,0}(y) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \mathbb{1}_{\{Y_{0,i} \le y\}} - F_0(y) + f_0(y)\Big(Y_{0,i} - \mathbb{E}(Y_{0,i})\Big) \right\} \ . \tag{A.5}$$

Step 2. The sequences $(B_{m,1}(y_1),\ldots,B_{m,1}(y_k))$ and $(B_{m,0}(y_1),\ldots,B_{m,0}(y_k))$ converge weakly to the marginals $(\mathbb{B}_1(y_1),\ldots,\mathbb{B}_1(y_k))$ and $(\mathbb{B}(y_1),\ldots,\mathbb{B}(y_k))$, respectively, for all $k\in\mathbb{N}$, and $y_1,\ldots,y_k\in\mathbb{R}$. It suffices by the Cramér–Wold device (Lehmann and Romano, 2005, Theorem 11.2.3) to show that, for every y

$$B_{m,1}(y)\xrightarrow{\mathrm{d}}\mathcal{N}(0,F_0(y)[1-F_0(y)]+f_0^2(y)\sigma^2+2F_0(y)\left\{\mathbb{E}[Y_{0,i}\mathbb{1}_{\{Y_{0,i}\le y\}}]-\mathbb{E}[Y_{0,i}]F_0(y)\right\})$$

under the null hypothesis. This follows from the Lindeberg–Lévy central limit theorem. Repeat an analogous argument for $B_{n,0}(y)$ to reach the desired convergence in distribution result.

Step 3. In order to show the process $\{B_{n,0}(y)\ :\ y\in\mathbb{R}\}$, is asymptotically equicontinuous, observe that

$$B_{n,0}(y)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\mathbb{1}_{\{Y_{0,i}\le y\}}-F_0(y)\Big)+\frac{f_0(y)}{\sqrt{n}}\sum_{i=1}^{n}\Big(Y_{0,i}-\mathbb{E}(Y_{0,i})\Big)\ .$$

The first term on the right converges weakly to a (tight) Gaussian process $\mathbb{G}$ by Theorem A. 1; the second one is in $\ell^\infty(\mathcal{F})$ if and only if $\|f_0\|<\infty$. This follows from Assumption A.2, which implies that F_1 and F_0 are Lipschitz continuous, then $\sup_y|f_0(y)|<\infty$. Then the second term converges too. By Van der Vaart and Wellner (1996, Theorem 1.5.4), both terms in the sum of the last display are asymptotically equicontinuous. Repeat an analogous argument for $B_{m,1}$ to conclude the proof of asymptotic equicontinuity.

Step 4. Combine the previous steps to conclude by Van der Vaart and Wellner (1996, Theorem 1.5.4) that the processes (A.4)–(A.5) converge weakly in $\ell^\infty(\mathcal{F})$ to two (tight) processes $\mathbb{B}_1(\cdot)$ and $\mathbb{B}(\cdot)$, respectively, where the processes $\mathbb{B}_1(\cdot)$ and $\mathbb{B}(\cdot)$ are independent. We have that $p_m\to p\in(0,1)$ by Assumption A. 1. Then, under the null hypothesis the limit variable $\widetilde{\mathbb{B}}=\sqrt{1-p}\,\mathbb{B}_1-\sqrt{p}\,\mathbb{B}$ possesses the same distribution as $\mathbb{B}$.

Then, the conclusion of the Theorem follows by the regular continuous mapping theorem.

A.4 Proof of Theorem 2

Since δ is unknown, we cannot shift data by δ as in the proof of Theorem A.2. Let $\tilde{Y}_{1,i}\equiv Y_{1,i}-\hat{\delta}$, $1\le i\le m$ and write $X=(X_1,\ldots,X_N)=(\tilde{Y}_{1,1},\ldots,\tilde{Y}_{1,m},Y_{0,1},\ldots,Y_{0,n})$. In other words, $\tilde{Y}_{1,i}$ is the recentered version of $Y_{1,i}$, where the shift is now given by $\hat{\delta}$. Independent of data, 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'.

In the same spirit as in (A.1), we seek to show that

$$\left(K^u_{m,n,\hat{\delta}}(X_\pi),K^u_{m,n,\hat{\delta}}(X_{\pi'})\right)\xrightarrow{\mathrm{d}}\left(K^u_0,K^{u'}_0\right)\ , \tag{A.6}$$

where K_0^u and $K_0^{u'}$ are independent and with common CDF $J_0(\cdot)$. If the convergence result in (A.6) holds, then

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

by Hoeffding's Condition (Lehmann and Romano, 2005, Theorem 15.2.3), finishing the proof of the Theorem. Since the joint distribution of $\left(K^u_{m,n,\hat{\delta}}(X_\pi), K^u_{m,n,\hat{\delta}}(X_{\pi'})\right)$ is the joint distribution of $\left(\sup_t \left|\upsilon_{m,n}(t,\hat{\delta};X_\pi)\right|, \sup_t \left|\upsilon_{m,n}(t,\hat{\delta};X_{\pi'})\right|\right)$, it suffices to investigate the asymptotic behavior of

$$\left(V_{m,n}(\cdot,\hat{\delta};X_\pi), V_{m,n}(\cdot,\hat{\delta};X_{\pi'})\right) \ ,$$

where

$$V_{m,n}(y,\hat{\delta};X_\pi) = \sqrt{\frac{mn}{N}} \left(\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\{X_{\pi(i)} \le y\}} - \frac{1}{n}\sum_{i=m+1}^{N} \mathbb{1}_{\{X_{\pi(i)} \le y\}} \right) \ .$$

To this end, we argue that **(1)** the process $\left(V_{m,n}(\cdot,\hat{\delta};X_\pi), V_{m,n}(\cdot,\hat{\delta};X_{\pi'})\right)$ converges weakly to a tight process $\left(\mathbb{G}_{\bar{P}}(\cdot), \mathbb{G}'_{\bar{P}}(\cdot)\right)$ in $\ell^\infty(\mathcal{F}) \times \ell^\infty(\mathcal{F})$, and **(2)** the process $\left(\mathbb{G}_{\bar{P}}(\cdot), \mathbb{G}'_{\bar{P}}(\cdot)\right)$ is a vector of two independent $\bar{P}$-Brownian bridges. In what follows we break the proof of these requirements in two steps.

Step 1 To show weak convergence, we need to verify marginal convergence and stochastic equicontinuity (Van der Vaart and Wellner, 1996, Theorem 1.5.4). For marginal convergence it suffices by the Cramér–Wold device (Lehmann and Romano, 2005, Theorem 11.2.3) to determine the joint limiting behavior of

$$\left(V_{m,n}(y,\hat{\delta};X_\pi), V_{m,n}(y,\hat{\delta};X_{\pi'})\right) = \sqrt{\frac{n}{mN}} \left(\sum_{i=1}^{N} \mathbb{1}_{\{X_i \le y\}} W_i, \sum_{i=1}^{N} \mathbb{1}_{\{X_i \le y\}} W'_i \right)$$

for every $y \in \mathbb{R}$, where W_i and W'_i are defined as in Lemma B.1.

Observe that $\mathbb{E}(\mathbb{1}_{\{X_i \le y\}} W_i) = \mathbb{E}(\mathbb{1}_{\{X_i \le y\}} W'_i) = 0$ by independence between data and W_i, W'_i. Set $\mathcal{S}$ as the number of positive integers $i \le m$ with $W_i = 1$, which follows a hypergeometric distribution with $\mathbb{E}(S) = m^2/N$ and $\mathbb{V}(S) = (mn/N)^2/(N-1)$. Then by the law of total variance

$$\mathbb{V}\left(V_{m,n}(y,\hat{\delta};X_\pi)\right) = \mathbb{E}\left[\mathbb{V}\left(V_{m,n}(y,\hat{\delta};X_\pi)|\mathcal{S}\right)\right] + \mathbb{V}\left[\mathbb{E}\left(V_{m,n}(y,\hat{\delta};X_\pi)|\mathcal{S}\right)\right] \ . \tag{A.7}$$

Simple algebra gives

$$\begin{aligned}\mathbb{E}\left(V_{m,n}(y,\hat{\delta};X_\pi)|\mathcal{S}\right) &= \sqrt{\frac{n}{mN}}\Bigg\{\frac{SN}{n}\left[\mathbb{E}\left(\mathbb{1}_{\{\tilde{Y}_{1,i}\leq y\}}\right)-\mathbb{E}\left(\mathbb{1}_{\{Y_{0,i}\leq y\}}\right)\right] \\ &\quad -\frac{m^2}{n}\left[\mathbb{E}\left(\mathbb{1}_{\{\tilde{Y}_{1,i}\leq y\}}\right)-\mathbb{E}\left(\mathbb{1}_{\{Y_{0,i}\leq y\}}\right)\right]\Bigg\} \qquad (A.8)\\ \mathbb{V}\left(V_{m,n}(y,\hat{\delta};X_\pi)|\mathcal{S}\right) &= \frac{n}{mN}\Bigg\{S\,\mathbb{V}\left(\mathbb{1}_{\{\tilde{Y}_{1,i}\leq y\}}\right)+(m-S)\,\mathbb{V}\left(\mathbb{1}_{\{Y_{0,i}\leq y\}}\right)+(m-S)\left(\frac{m}{n}\right)^2\mathbb{V}\left(\mathbb{1}_{\{\tilde{Y}_{1,i}\leq y\}}\right) \\ &\quad +(n-m+S)\left(\frac{m}{n}\right)^2\mathbb{V}\left(\mathbb{1}_{\{Y_{0,i}\leq y\}}\right)\Bigg\}+o(1) \qquad (A.9)\end{aligned}$$

Note that

$$\begin{aligned}\mathbb{E}\left(\mathbb{1}_{\{\tilde{Y}_{1,i}\leq y\}}\right) &= F_1(y+\delta)+o(1)\\ \mathbb{V}\left(\mathbb{1}_{\{\tilde{Y}_{1,i}\leq y\}}\right) &= F_1(y+\delta)(1-F_1(y+\delta))+o(1)\ .\end{aligned}$$

Plugging the above expressions into Eq. (A.8)–(A.9) gives

$$\begin{aligned}\mathbb{V}\left[\mathbb{E}\left(V_{m,n}(y,\hat{\delta};X_\pi)|\mathcal{S}\right)\right] &= \frac{n}{mN}\left\{\frac{m^2}{N-1}\Big(F_1(y+\delta)-F_0(y)\Big)^2\right\}+o(1)\\ \mathbb{E}\left[\mathbb{V}\left(V_{m,n}(y,\hat{\delta};X_\pi)|\mathcal{S}\right)\right] &= \frac{n}{mN}\Bigg\{\left(\frac{m^2}{N}+\frac{m^3}{n^2}-\frac{m^2}{n^2}\frac{m^2}{N}\right)\mathbb{V}\left(\mathbb{1}_{\{Y_{1,i}\leq y\}}\right)\\ &\quad +\left(m-\frac{m^2}{N}+\frac{m^2}{n^2}(n-m)+\frac{m^2}{n^2}\frac{m^2}{N}\right)\mathbb{V}\left(\mathbb{1}_{\{Y_{0,i}\leq y\}}\right)\Bigg\}+o(1)\ .\end{aligned}$$

With this in mind, we can conclude that (A.8) reduces to

$$\begin{aligned}\mathbb{V}\left(V_{m,n}(y,\hat{\delta};X_\pi)\right) &= \frac{m}{N}F_1(y+\delta)\left(1-F_1(y+\delta)\right)+\frac{n}{N}F_0(y)\left(1-F_0(y)\right)\\ &\quad +\left(\frac{nm}{N(N-1)}\right)\Big(F_1(y+\delta)-F_0(y)\Big)^2+o(1)\ .\end{aligned}$$

Lastly, observe

$$\mathbb{C}\left(V_{m,n}(y,\hat{\delta};X_\pi),V_{m,n}(y,\hat{\delta};X_{\pi'})\right)=\frac{n}{mN}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}\left(\mathbb{1}_{\{X_i\leq y\}}\mathbb{1}_{\{X_j\leq y\}}W_iW_j'\right)=0\ ,$$

by independence of W_i and W_i' and the fact that $\mathbb{E}(W_i)=0$ for all i. If assumption A.1 holds, then

$$\left(V_{m,n}(y,\hat{\delta};X_\pi),V_{m,n}(y,\hat{\delta};X_{\pi'})\right)\xrightarrow{\mathrm{d}}\mathcal{N}\left(\mathbf{0},\begin{pmatrix}\bar{P}(y)(1-\bar{P}(y)) & 0\\ 0 & \bar{P}(y)(1-\bar{P}(y))\end{pmatrix}\right)$$

by the same arguments as used in the proof of Lemma B.1. This finishes the proof of marginal convergence. Asymptotic equicontinuity of the process $\{V_{m,n}(y,\hat{\delta};X_\pi)\ :\ y\in\mathbb{R}\}$ follows by the same arguments as used in the proof of (Præstgaard, 1995, Theorem 1), and thus leads to the desired results in ***Step 1***.

Step 2 We now prove that

$$\left(\mathbb{G}_{\bar{P}}(y_1),\ldots,\mathbb{G}_{\bar{P}}(y_k)\right) \perp\!\!\!\perp \left(\mathbb{G}'_{\bar{P}}(y_1),\ldots,\mathbb{G}'_{\bar{P}}(y_k)\right)$$

for all $k\in\mathbb{N}$, and $y_1,\ldots,y_k\in\mathbb{R}$. By ***Step 1***, we know that the joint marginals converge to a multivariate normal distribution whose covariance matrix is block-diagonal, then independence follows by zero correlation.

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 $\mathbb{V}(Y_{1,i})=\mathbb{V}(Y_{0,i})=\sigma^2<\infty$. 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.

First note that the Khmaladze transformation based on $\upsilon_{m,n}(t,\hat{\delta};Z)$ is

$$\begin{aligned}\tilde{\upsilon}_{m,n}(t,\hat{\delta};Z) &= \upsilon_{m,n}(t,\hat{\delta};Z)-\int_0^t\left[\dot{g}(s)'C(s)^{-1}\int_s^1\dot{g}(r)d\upsilon_{m,n}(r,\hat{\delta};Z)\right]ds\\ &= \upsilon_{m,n}(t,\hat{\delta};Z)-\psi_g\left(\upsilon_{m,n}\right)(t,\hat{\delta};Z)\ .\end{aligned} \tag{A.10}$$

If Assumption A.2 holds, then use Lemma B.3 and Remark 5 to see that

$$\begin{aligned}\upsilon_{m,n}(t,\hat{\delta};Z) &= V_{m,n}(F_0^{-1}(t),\hat{\delta};Z)\\ &= V_{m,n}(F_0^{-1}(t),\delta;Z^*)+\sqrt{\frac{mn}{N}}\left\{f_0\left(F_0^{-1}(t)\right)(\hat{\delta}-\delta)\right\}+o_p(1)\\ &= \upsilon_{m,n}(t,\delta;Z^*)+\sqrt{\frac{mn}{N}}\left\{f_0\left(F_0^{-1}(t)\right)(\hat{\delta}-\delta)\right\}+o_p(1)\ ,\end{aligned} \tag{A.11}$$

where the $o_p(1)$ term holds uniformly over $0\le t\le 1$. Next, note that

$$\psi_g\left(\upsilon_{m,n}\right)(t,\hat{\delta};Z)=\psi_g(\upsilon_{m,n})(t,\delta;Z^*)+\sqrt{\frac{mn}{N}}\left\{f_0\left(F_0^{-1}(t)\right)(\hat{\delta}-\delta)\right\}+o_p(1) \tag{A.12}$$

by properties of map ψ_g. Plug (A.11)-(A.12) into (A.10) to obtain

$$\tilde{\upsilon}_{m,n}(t,\hat{\delta};Z)=\upsilon_{m,n}(t,\delta;Z^*)-\psi_g\left(\upsilon_{m,n}\right)(t,\delta;Z^*)+o_p(1)\ . \tag{A.13}$$

It follows from Theorem A.1 that $\upsilon_{m,n}(\cdot,\delta;Z^*)$ converges weakly to $\mathbb{U}$. Further, we note that the linear operator ψ_g is 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). Then $\psi_g\left(\upsilon_{m,n}\right)(\cdot,\delta;Z^*)$ converges weakly to $\psi_g(\mathbb{U})$ and thus $\tilde{\upsilon}_{m,n}(\cdot,\hat{\delta};Z)$ converges weakly to $\mathbb{M}(\cdot)$ (Khmaladze, 1981, 4.3).

Next apply the usual continuous mapping theorem to conclude that $\tilde{K}_{m,n,\hat{\delta}}$ converges in distribution to K_2 under the null hypothesis, completing the proof of the theorem.

A.6 Proof of Theorem 4

We begin the proof by establishing the limiting behavior of $\hat{R}^{\tilde{K}(\hat{\delta})}_{m,n}$. Recall from Section 3.2 that

$$\begin{aligned}\tilde{\upsilon}_{m,n}(t,\hat{\delta};Z) &= \upsilon_{m,n}(t,\delta;Z^*) - \psi_g(\upsilon_{m,n})(t,\delta;Z^*) + o_p(1)\\ &= \tilde{\upsilon}_{m,n}(t,\delta;Z^*) + o_p(1)\ . \end{aligned} \tag{A.14}$$

We derive the limiting behavior of $\hat{R}^{\tilde{K}(\hat{\delta})}_{m,n}$ in three steps.

Step 1 We begin by determining the behavior of the permutation distribution based on $\tilde{\upsilon}_{m,n}(t,\delta;Z^*)$. To this end, note that we have established the asymptotic behavior of the permutation distribution based on $\upsilon_{m,n}(t,\delta;Z^*)$ in Theorem A.2. Moreover, we know that $\tilde{\upsilon}_{m,n}(t,\delta;Z^*)$ is a continuous mapping by the arguments in the proof of Theorem 3. Therefore,

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

by the continuous mapping theorem for randomization distributions, Chung and Romano (2016a, Lemma A.6), thus finishing the proof of the claim.

Step 2 We now prove that (A.14) holds under permutations, *i.e.*,

$$\tilde{\upsilon}_{m,n}(t,\hat{\delta};Z_\pi) - \tilde{\upsilon}_{m,n}(t,\delta;Z_\pi) \xrightarrow{\text{P}} 0\ . \tag{A.15}$$

In view of the contiguity result in Chung and Romano (2013, Lemma 5.3), we can deduce (A.15) from the basic assumption of how it behaves under i.i.d. observations from the mixture distribution $\bar{P}$. However, we know from Theorem A.2 and the Khmaladze transformation that $\tilde{\upsilon}_{m,n}(t,\hat{\delta};\bar{Z}_\pi) - \tilde{\upsilon}_{m,n}(t,\delta;\bar{Z}_\pi) \xrightarrow{\text{P}} 0$, where $\bar{Z} = \bar{Z}_1,\ldots,\bar{Z}_N$ is an i.i.d. sequence from the mixture distribution, so the desired conclusion follows.

Step 3 Combine ***Steps 1*** and ***2*** with the Slutsky's Theorem for randomization distributions (Chung and Romano, 2013, Theorem 5.2) to conclude

$$\sup_t \left| \hat{R}^{\tilde{K}(\hat{\delta})}_{m,n}(t) - J_2(t)\right| \xrightarrow{\text{P}} 0\ .$$

For the second part, we note that the distribution of K_2, *i.e.*, the distribution of the norm of a tight Brownian motion process, is strictly increasing and absolutely continuous with a positive density (Beran and Millar, 1986, Proposition 2). Thus, under the conditions of the Theorem,

$$\hat{r}_{m,n}(1-\alpha) \xrightarrow{\text{p}} r(1-\alpha) = \inf\{t : J_2(t) \geq 1-\alpha\}$$

by Lehmann and Romano (2005, Lemma 11.2.1 (ii)), concluding the proof of the theorem.

B Auxiliary Lemmas

Throughout this appendix, if ξ is a random variable defined on a probability space $(\Omega, \mathcal{B}, P)$, it is assumed that $\xi_1, \dots, \xi_N$ are coordinate projections on the product space $(\Omega^N, \mathcal{B}^N, P^N)$, and the expectations are computed for P^N. If auxiliary variables, independent of the ξs, are involved—as in Lemma B.1—we use a similar convention. In that case, the underlying probability space is assumed to be of the form $(\Omega^N, \mathcal{B}^N, P^N) \times (\mathcal{Z}, \mathcal{C}, Q)$, with $\xi_1, \dots, \xi_N$ equal to the coordinate projections on the first N coordinates and the additional variables depending only on the N+1st coordinate. For example, the coordinate projections on the first N coordinates in Lemma B.1 have distribution function $\bar{P}$.

Lemma B.1. *Let $\bar{Z}_1, \dots, \bar{Z}_N$ be an i.i.d. sequence from the mixture distribution with CDF $\bar{P} = pF_1^\delta + (1-p)F_0$, with $F_1^\delta(y) = F_1(y+\delta)$. Independent of the $\bar{Z}$s, let $(\pi(1), \dots, \pi(N))$ and $(\pi'(1), \dots, \pi'(N))$ be two independent random permutations of $\{1, \dots, N\}$. Set $\bar{Z}_\pi = (\bar{Z}_{\pi(1)}, \dots, \bar{Z}_{\pi(N)})$; $\bar{Z}_{\pi'}$ is defined the same way with π replaced by π'. If conditions A.1–A.2 hold, then*

$$\left(K^u_{m,n,\delta}(\bar{Z}_\pi), K^u_{m,n,\delta}(\bar{Z}_{\pi'})\right) \xrightarrow{\text{d}} \left(K^u_0, K^{u'}_0\right) \ ,$$

where K^u_0 and $K^{u'}_0$ are independent with common CDF $J_0(\cdot)$.

Proof. We can deduce the asymptotic behavior of $K^u_{m,n,\delta}(\bar{Z}_\pi)$ from the asymptotic behavior of $K_{m,n,\delta}(\bar{Z}_\pi)$ via the change of variable $y \mapsto \bar{P}^{-1}(t)$ and noting

$$\begin{aligned}
\upsilon^{\bar{P}}_{m,n}(t, \delta; \bar{Z}_\pi) &= V_{m,n}(\bar{P}^{-1}(t), \delta; \bar{Z}_\pi) \\
&= \sqrt{\frac{mn}{N}} \left\{ \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\{\bar{Z}_{\pi(i)} \leq \bar{P}^{-1}(t)\}} - \frac{1}{n} \sum_{i=m+1}^{N} \mathbb{1}_{\{\bar{Z}_{\pi(i)} \leq \bar{P}^{-1}(t)\}} \right\} \\
K^u_{m,n,\delta}(\bar{Z}_\pi) &= \sup_t \left| \upsilon^{\bar{P}}_{m,n}(t, \delta; \bar{Z}_\pi) \right| = \sup_y \left| V_{m,n}(y, \delta; \bar{Z}_\pi) \right| = K_{m,n,\delta}(\bar{Z}_\pi) \ .
\end{aligned}$$

By the usual continuous-mapping argument, the desired conclusion follows if we can prove that **(1)** the process $\left(V_{m,n}(\cdot, \delta; \bar{Z}_\pi), V_{m,n}(\cdot, \delta; \bar{Z}_{\pi'})\right)$ converges weakly to a tight process $\left(\mathbb{G}_{\bar{P}}(\cdot), \mathbb{G}'_{\bar{P}}(\cdot)\right)$

in $\ell^\infty(\mathcal{F}) \times \ell^\infty(\mathcal{F})$, and **(2)** the process $\left(\mathbb{G}_{\bar{P}}(\cdot), \mathbb{G}'_{\bar{P}}(\cdot)\right)$ is a vector of two independent $\bar{P}$-Brownian bridges. In the following, we prove the requirements **(1)**–**(2)** in two steps.

Step 1 In order to show weak convergence, we need to establish marginal convergence convergence and asymptotic equicontinuity (Van der Vaart and Wellner, 1996, Theorem 1.5.4). Marginal convergence follows by verifying that the marginals

$$\mathbf{V}(\delta) = \left(V_{m,n}(y_1, \delta; \bar{Z}_\pi), \ldots, V_{m,n}(y_k, \delta; \bar{Z}_\pi), V_{m,n}(y_1, \delta; \bar{Z}_{\pi'}), \ldots, V_{m,n}(y_k, \delta; \bar{Z}_{\pi'})\right)$$

converge weakly to the marginals

$$\left(\mathbb{G}_{\bar{P}}(y_1), \ldots, \mathbb{G}_{\bar{P}}(y_k), \mathbb{G}'_{\bar{P}}(y_1), \ldots, \mathbb{G}'_{\bar{P}}(y_k)\right)$$

for all $k \in \mathbb{N}$, and $y_1, \ldots, y_k \in \mathbb{R}$. To this end, define

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

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 $\mathbb{E}(W_i^2) = \mathbb{E}(W_i'^{\,2}) = m/n$. With this in mind, rewrite $\mathbf{V}(\delta)$ as

$$\mathbf{V}(\delta) = a_m^{1/2} \left(\sum_{i=1}^N \mathbb{1}_{\left\{\bar{Z}_i \leq y_1\right\}} W_i, \ldots, \sum_{i=1}^N \mathbb{1}_{\left\{\bar{Z}_i \leq y_k\right\}} W_i, \sum_{i=1}^N \mathbb{1}_{\left\{\bar{Z}_i \leq y_1\right\}} W'_i, \ldots, \sum_{i=1}^N \mathbb{1}_{\left\{\bar{Z}_i \leq y_k\right\}} W'_i \right)^\intercal ,$$

where $a_m = n/Nm$. Observe that independence of π, π' from $\bar{Z}$ ensures that

$$\begin{aligned} \mathbb{E}\left(\mathbb{1}_{\left\{\bar{Z}_i \leq y_j\right\}} W_i\right) &= 0 \\ \mathbb{V}\left(\mathbb{1}_{\left\{\bar{Z}_i \leq y_j\right\}} W_i\right) &= \frac{m}{n} \bar{P}(y_j)\big(1 - \bar{P}(y_j)\big) \\ \mathbb{C}\left(\mathbb{1}_{\left\{\bar{Z}_i \leq y_j\right\}} W_i, \mathbb{1}_{\left\{\bar{Z}_i \leq y_l\right\}} W_i\right) &= \frac{m}{n}\big(\bar{P}(y_j \wedge y_l) - \bar{P}(y_j)\bar{P}(y_l)\big) \\ \mathbb{E}\left(\mathbb{1}_{\left\{\bar{Z}_i \leq y_j\right\}} \mathbb{1}_{\left\{\bar{Z}_i \leq y_l\right\}} W_i W'_i\right) &= 0 , \end{aligned}$$

for $1 \leq i \leq N$, $1 \leq j \leq k$, $1 \leq l \leq k$, 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 $\mathbb{V}(\mathbf{V}(\delta)) = \operatorname{diag}\{\mathbf{\Sigma}_i \mid i = 1, 2\}$, with

$$\mathbf{\Sigma}_i = \begin{pmatrix} \bar{P}(y_1)(1 - \bar{P}(y_1)) & \ldots & \bar{P}(y_1 \wedge y_k) - \bar{P}(y_1)\bar{P}(y_k) \\ \vdots & \ddots & \vdots \\ \bar{P}(y_k \wedge y_1) - \bar{P}(y_k)\bar{P}(y_1) & \cdots & \bar{P}(y_k)(1 - \bar{P}(y_k)) \end{pmatrix} .$$

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 that for vector $\mathbf{c} \in \mathbb{R}^{2k}$,

$$\mathbf{c}^\intercal \mathbf{V}(\delta) \xrightarrow{\mathrm{d}} c_1 \mathbb{G}_{\bar{P}}(y_1) + \cdots + c_k \mathbb{G}_{\bar{P}}(y_k) + c_{k+1} \mathbb{G}'_{\bar{P}}(y_1) + \cdots + c_{2k} \mathbb{G}'_{\bar{P}}(y_k) \ .$$

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

$$a_m^{1/2} \sum_{j=1}^{k} \left(\sum_{i=1}^{N} \mathbb{1}_{\left\{\bar{Z}_i \leq y_j\right\}} \left(c_j W_i c_{k+j} W'_i\right) \right) \ . \tag{B.1}$$

Conditionally on W_i and W'_i, (B.1) is an independent sum of linear combinations of independent random variables. For every summand j above, we can show that

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

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

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

This finishes the proof of marginal convergence. For convergence in $\ell^\infty(\mathcal{F}) \times \ell^\infty(\mathcal{F})$, it suffices to check asymptotic equicontinuity of the process $\{V_{m,n}(y, \delta; \bar{Z}_\pi) \ : y \in \mathbb{R}\}$. The proof follows by the same arguments as used in the proof of Van der Vaart and Wellner (1996, Theorem 3.7.1) and omitted.

Step 2 We now prove that

$$\Big(\mathbb{G}_{\bar{P}}(y_1), \ldots, \mathbb{G}_{\bar{P}}(y_k)\Big) \perp\!\!\!\perp \Big(\mathbb{G}'_{\bar{P}}(y_1), \ldots, \mathbb{G}'_{\bar{P}}(y_k)\Big)$$

for all $k \in \mathbb{N}$, and $y_1, \ldots, y_k \in \mathbb{R}$. By ***Step 1***, we know that the joint marginals converge to a multivariate normal distribution whose covariance matrix is block-diagonal. Then the sequence converges to a vector of two independent $\bar{P}$-Brownian bridges with the same distribution.

□

Lemma B.2. *Consider the setting described in Lemma B.1. If conditions A.1–A.2 hold, then*

$$K^u_{m,n,\delta}(\bar{Z}_{\pi,\pi_0}) - K^u_{m,n,\delta}(Z^*_\pi) \xrightarrow{\mathrm{p}} 0 \ .$$

Proof. To appreciate what is in the verification of the Theorem, we apply the coupling construction in Appendix C, where π_0 and D are properly defined. For notational convenience, abbreviate $S_{m,n} = V_{m,n}(y, \delta; \bar{Z}_{\pi\pi_0}) - V_{m,n}(y, \delta; Z^*_\pi)$ for every y, and observe that

$$S_{m,n} = \sqrt{\frac{n}{mN}} \left\{ \sum_{i=1}^{N} \left(\mathbb{1}_{\{\bar{Z}_{\pi_0(i)} \le y\}} - \mathbb{1}_{\{Z^*_i \le y\}} \right) W_{\pi(i)} \right\} , \tag{B.2}$$

where W_i is defined as in Lemma B.1. It is straightforward to see that $\mathbb{E}\left[S_{m,n}\right] = 0$ by independence of data and $W_{\pi(i)}$. To investigate the variance, observe that the elements in $\bar{Z}_{\pi_0}$ and Z^* are the same except for D of them. This makes all the terms in the difference $S_{m,n}$ zero, except for at most D of them. Conditioning on the random drawing of indices in the coupling construction—hence conditioning on D and π_0—and on the permutation π, the variance of $S_{m,n}$ is determined by

$$\mathbb{V}\left[S_{m,n}\right] = \mathbb{E}\left[\mathbb{V}\left(S_{m,n} \middle| D, \pi, \pi_0\right)\right] + \mathbb{V}\left[\mathbb{E}\left(S_{m,n} \middle| D, \pi, \pi_0\right)\right] \tag{B.3}$$

by the law of total variance. We claim that both terms in previous display are zero, asymptotically. Note that the conditional variance in the first term in (B.3) is bounded above

$$\mathbb{V}\left[S_{m,n} \middle| D, \pi, \pi_0\right] = \frac{n}{Nm} D\, \mathbb{V}\left[W_{\pi(i)} \left(\mathbb{1}_{\{\bar{Z}_{\pi_0(i)} \le y\}} - \mathbb{1}_{\{Z^*_i \le y\}} \right) \middle| D, \pi, \pi_0 \right] \le \frac{n}{m} \frac{D}{N} \mathcal{O}(1) \ .$$

In view of Chung and Romano (2013, Section 5.3), $\mathbb{E}(D/N) \le N^{-1/2}$ and so the first term on the right hand side of (B.3) converges to 0. Another application of the law of total variances applied to the second term in (B.3) yields

$$\mathbb{V}\left[\mathbb{E}\left(S_{m,n} \middle| D, \pi, \pi_0\right)\right] = \mathbb{E}\left\{ \mathbb{V}\left[\mathbb{E}\left(S_{m,n} \middle| D, \pi, \pi_0\right) \middle| D, \pi_0 \right] \right\} + \mathbb{V}\left\{ \mathbb{E}\left[\mathbb{E}\left(S_{m,n} \middle| D, \pi, \pi_0\right) \middle| D, \pi_0 \right] \right\} .$$

Let S be the number of observations among those D observations that have $W_{\pi(i)} = 1$. Conditioning on the random drawing of indices in the coupling construction—hence conditioning on D and π_0—, the distribution of S is hypergeometric with D draws out of N elements, among which m have $W_{\pi(i)} = 1$. This gives

$$\mathbb{E}[S|D, \pi_0] = D\left(\frac{m}{N}\right), \quad \text{and} \quad \mathbb{V}[S|D, \pi_0] = D\left(\frac{m}{N}\right)\left(\frac{n}{N}\right)\left(\frac{N-D}{N-1}\right) \ .$$

With this in mind, it can be shown that

$$\mathbb{E}\left\{ \mathbb{V}\left[\mathbb{E}\left(S_{m,n} \middle| D, \pi, \pi_0\right) \middle| D, \pi_0 \right] \right\} = \frac{1}{N-1}\left[\mathbb{E}(D) - \mathbb{E}(D^2)\left(\frac{1}{N}\right) \right] \mathcal{O}(1) = o(1)$$

$$\mathbb{V}\left\{ \mathbb{E}\left[\mathbb{E}\left(S_{m,n} \middle| D, \pi, \pi_0\right) \middle| D, \pi_0 \right] \right\} = 0 \ .$$

Then (B.2) converges to 0 in quadratic mean. Since both processes defining $S_{m,n}$ are asymptotically equicontinuous, the convergence in probability holds uniformly. This finishes the proof of the lemma.

□

Lemma B.3. *Assume $Y_{0,1},\dots,Y_{0,n}$ are i.i.d. according to a probability distribution F_0, and independently $Y_{1,1},\dots,Y_{1,m}$ are i.i.d. according to a probability distribution F_1. If conditions A.1–A.2 hold, then the process $V_{m,n}(y,\hat{\delta};Z)$ admits the following asymptotic representation*

$$V_{m,n}(y,\hat{\delta};Z) = V_{m,n}(y,\delta;Z^*) + \sqrt{\frac{mn}{N}}\Big(f_1(y+\delta)(\hat{\delta}-\delta)\Big) + o_p(1) \ ,$$

where the $o_p(1)$ term holds uniformly over $y\in\mathbb{R}$. Moreover, suppose that the null hypothesis holds, then

$$\begin{aligned} V_{m,n}(y,\hat{\delta};Z) &= \sqrt{1-p_m}\left(\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\Big\{\mathbb{1}_{\{Y_{1,i}\le y+\delta\}} - F_1(y+\delta) + f_1(y+\delta)\Big(Y_{1,i}-\mathbb{E}(Y_{1,i})\Big)\Big\}\right) \\ &\quad - \sqrt{p_m}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{\mathbb{1}_{\{Y_{0,i}\le y\}} - F_0(y) + f_1(y+\delta)\Big(Y_{0,i}-\mathbb{E}(Y_{0,i}\Big)\Big\}\right) + o_p(1) \ , \end{aligned}$$

where $p_m = m/N$ and the $o_p(1)$ term holds uniformly over $y\in\mathbb{R}$.

Proof. Simple algebra allows us to write $V_{m,n}(y,\hat{\delta};Z)$ as

$$\begin{aligned} \sqrt{\frac{mn}{N}}\Big\{\hat{F}_1(y+\hat{\delta}) - \hat{F}_0(y)\Big\} &= \sqrt{\frac{mn}{N}}\Big\{\hat{F}_1(y+\delta) - \hat{F}_0(y)\Big\} + \sqrt{\frac{mn}{N}}\Big\{F_1(y+\hat{\delta}) - F_1(y+\delta)\Big\} \\ &\quad + \sqrt{\frac{mn}{N}}\Big\{\Big(\hat{F}_1(y+\hat{\delta}) - F_1(y+\hat{\delta})\Big) - \Big(\hat{F}_1(y+\delta)) - F_1(y+\delta)\Big)\Big\} \\ &= V_{m,n}(y,\delta;Z^*) + \sqrt{\frac{mn}{N}}\Big\{F_1(y+\hat{\delta}) - F_1(y+\delta)\Big\} + o_p(1) \ . \end{aligned}$$

The last equality follows due to the fact that

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

by stochastic equicontinuity of $\Big\{\sqrt{m}\Big(\hat{F}_1(y) - F_1(y)\Big) : y\in\mathbb{R}\Big\}$ and the arguments in Pollard (Chapter VII.1 1984, pp. 139–140). In view of Condition A.2, we expand $F_1(y+\hat{\delta})$ around δ to obtain:

$$\begin{aligned} V_{m,n}(y,\hat{\delta};Z) &= V_{m,n}(y,\delta;Z^*) + \sqrt{\frac{mn}{N}}\Big\{\Big(F_1(y+\delta) + f_1(y+\delta)(\hat{\delta}-\delta)\Big) - F_1(y+\delta)\Big\} + o_p(1) \\ &= V_{m,n}(y,\delta;Z^*) + \sqrt{\frac{mn}{N}}\Big\{f_1(y+\delta)(\hat{\delta}-\delta)\Big\} + o_p(1) \ . \end{aligned}$$

This finishes the first part of the proof. Suppose further that the null hypothesis holds. Then

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

and so

$$\begin{aligned}V_{m,n}(y,\hat{\delta};Z) &= \sqrt{\frac{mn}{N}}\left\{\Big(\hat{F}_1(y+\delta)-F_1(y+\delta)\Big)-\Big(\hat{F}_0(y)-F_0(y)\Big)\right\}\\ &+\sqrt{\frac{mn}{N}}\left\{f_1(y+\delta)(\hat{\delta}-\delta)\right\}+o_p(1)\\ &= \sqrt{1-p_m}\left\{\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\Big[\mathbb{1}_{\{Y_{1,i}\le y+\delta\}}-F_1(y+\delta)+f_1(y+\delta)\Big(Y_{1,i}-\mathbb{E}(Y_{1,i})\Big)\Big]\right\}\\ &-\sqrt{p_m}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[\mathbb{1}_{\{Y_{0,i}\le y\}}-F_0(y)+f_1(y+\delta)\Big(Y_{0,i}-\mathbb{E}(Y_{0,i})\Big)\Big]\right\}+o_p(1) ,\end{aligned}$$

as desired. □

C Coupling Construction

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 $\bar{P}=pF_1^{\delta}+(1-p)F_0$, where $F_1^{\delta}(y)=F_1(y+\delta)$.

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

The algorithm

Except for ordering, we can construct $\bar{Z}$ to include almost the same set of observations as in Z^*. First draw an index j from $\{0,1\}$ with probability $\mathbb{P}(j=1)=p$. Then, conditionally on the outcome being $j=1$, set $\bar{Z}_1=Y_{1,1}-\delta$. Next, draw another index i from $\{0,1\}$ at random with probability $\mathbb{P}(i=1)=p$. If $i=0$, then $\bar{Z}_2=Y_{0,1}$; otherwise if $i=1$ as in the previous step, then $\bar{Z}_2=Y_{1,2}-\delta$. Keep repeating this process, noting that there will probably be a point in which you exhaust all the m observations governed by F_1^{δ}. If this happens and another index $j=1$ is drawn again, then just sample a new observation from F_1^{δ}, and analogously if the observations you have exhausted are from population with CDF F_0. 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—this is why this method is called "coupling." The number of observations in which Z^* and $\bar{Z}$ differ, say D, is the (random) number of added observations required to fill up $\bar{Z}$.

Remark C.1. Observe that the number of observations from F_1^δ in Z^* is exactly m, whereas the number of observations $\bar{Z}_i$ out of N which are from population F_1^δ follows a Binomial (N, p) distribution with mean pN, which is approximately m. Thus—except for the fact that the ordering in Z^* is such that the first m observations are coming from F_1^δ, and the last n are coming from F_0—the original sampling scheme is still only approximately like that of sampling from $\bar{P} = pF_1^\delta + (1-p)F_0$. ■

Reordering according to π_0

We can reorder the observations in $\bar{Z}$ by a permutation π_0 so that Z_i^* and $\bar{Z}_{\pi_0(i)}$ agree for all i except for some hopefully small (random) number D. Recall that Z^* has the observations in order, that is, the first m observations arose from F_1^δ, while the last n observations are distributed according to F_0. Thus, to couple $\bar{Z}$ with Z^*, put all observation in $\bar{Z}$ that came from F_1^δ in the first up to m. If the number of observations from F_1^δ is *greater than or equal to m* (recall that this is a possibility), then $\bar{Z}_{\pi(i)}$ for $i = 1, \ldots, m$ are filled according to the observations in $\bar{Z}$ which came from F_1^δ, 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_1^δ is *less* than m, fill up as many of $\bar{Z}$ from F_1^δ as possible, and leave the rest of the blank spots for now.

Next, move onto the observations in $\bar{Z}$ that came from F_0 and repeat the above procedure for $m+1, m+2, \ldots, m+n$ 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 $\bar{Z}_{\pi_0}$ differ 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^δ from the behavior of the permutation distribution when all N observations come from the mixture distribution.

Suppose π and π' are independent random permutations of $\{1, \ldots, N\}$, and independent of the Z_i^* and $\bar{Z}_i$. Suppose we can show that

$$\left(K_{m,n,\delta}(\bar{Z}_\pi), K_{m,n,\delta}(\bar{Z}_{\pi'})\right) \xrightarrow{\text{d}} (K_0, K_0') \ , \tag{C.1}$$

where K_0 and K_0' are independent with common CDF $J_1(\cdot)$. Then by Chung and Romano (2013, Theorem 5.1), the randomization distribution based on $K_{m,n}$ converges in probability to $J_1(\cdot)$ 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

$$\left(K_{m,n,\delta}(\bar{Z}_{\pi\pi_0}), K_{m,n,\delta}(\bar{Z}_{\pi'\pi_0})\right) \xrightarrow{\text{d}} (K_0, K_0') \ .$$

Using the coupling construction above, suppose it can be shown that $K_{m,n,\delta}(\bar{Z}_{\pi\pi_0}) - K_{m,n,\delta}(\bar{Z}_\pi)$ converges to 0 in probability. Then it also follows that $K_{m,n,\delta}(\bar{Z}_{\pi'\pi_0}) - K_{m,n,\delta}(\bar{Z}_{\pi'}) \xrightarrow{\text{p}} 0$. Therefore, we can conclude that $(K_{m,n,\delta}(Z_\pi), K_{m,n,\delta}(Z_{\pi'})) \xrightarrow{\text{d}} (K_0, K_0')$ by Slutsky's theorem. Another application of Chung and Romano (2013, Theorem 5.1) allows us to conclude that the permutation distribution also converges in probability to $J_1(\cdot)$ under the original model of two samples from possibly different distributions.

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 (21) based on the martingale-transformed Kolmogorov–Smirnov 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

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

2. *Define*

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

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

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

4. *Each adjusted p-value $p^*_{r_j}$—with associated hypothesis H_{0,r_j}—is now compared with α, for $1 \leq j \leq \mathcal{J}$, i.e., if $p^*_{r_j} \geq \alpha$ then we fail to reject, otherwise reject H_{0,r_j}.*

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 $\alpha/(\mathcal{J} - 2)$.*

⋮

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 $H_{0,r_j}, \ldots, H_{0,r_\mathcal{J}}$ 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 $\alpha/(\mathcal{J} - j)$.*

⋮

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