Non-Parametric Hypothesis Testing with a Nuisance Parameter: A Permutation Test Approach

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

Mauricio Olivares
Department of Economics
UIUC
lvrsgnz2@illinois.edu

February 11, 2019

Abstract

This paper studies a classical problem in statistics: testing goodness of fit in the presence of a nuisance parameter. The main contribution of this paper is a novel permutation test for this testing problem that is asymptotically valid under fairly weak assumptions, while still providing an exact error control in finite samples under more restrictive conditions. In addition, the permutation test presented here is shown to have finite- and large-sample properties comparable to those existing in the literature.

The main result relies on the martingale transformation of the empirical process introduced by Khmaladze (1981). This procedure clears the empirical process out from the effect of the nuisance parameter by decomposing it into two parts - a martingale that has a standard Brownian motion asymptotic behavior, and a second part that vanishes as the sample size grows large.

A noteworthy application of this testing problem is the one of testing for heterogeneous treatment effects in a randomized experiment. In this context, the null hypothesis implies that the *distribution* of the treatment and control groups are a constant shift apart. Moreover, the proposed method can be extended to testing the joint null hypothesis that treatment effects are constant within individual subgroups, while allowing for varying average treatment effects across subgroups. As a result, this test is able to detect treatment effect heterogeneity within individual subgroups even if the average treatment effects are different across subgroups.

To gain further understanding of the test to practical problems, we provide the companion `RATest` `R` package (Olivares and Sarmiento (2017)) and apply our test to investigate the gift exchange hypothesis in the context of two field experiments from Gneezy and List (2006). Our test rejects the null hypothesis in favor of the heterogeneity in the treatment effect, where solely looking at the average treatment effect does not provide evidence in favor of the gift exchange hypothesis.

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

[†]We would like to thank workshop participants at the University of Illinois Urbana-Champaign, ITAM, Seoul National University, and Norther Illinois University. We owe special thanks to Roger Koenker, Jose Luis Montiel-Olea, and Joseph Romano for extremely useful comments and feedback. All errors are our own. This version: February 2019.

1 Introduction

The main goal of this paper is to study the classical goodness-of-fit hypothesis testing problem with a nuisance parameter. In particular, we propose a permutation test approach to make inference under minimal assumptions in situations where randomization ideas apply.

The statistical hypothesis problem we examine has the following structure. Consider two real-valued random variables Y and X with probability distributions F_0 and F_1, respectively. We want to test whether these random variables differ in location:

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

based on two independent samples from F_0 and F_1. In other words, the CDFs are a *constant shift apart*. A common example of this situation is when the researcher is interested in testing the hypothesis of constant treatment effect in a randomized trial, where the aforementioned hypothesis essentially states that the treatment induces a shift in the potential outcomes.

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, permutation tests control Type 1 error in finite samples: the rejection probability under the null is *exactly* the nominal level α. Moreover, they are nonparametric in the sense that they can be applied without any parametric assumptions about the underlying distribution that generates the data. Also, the general construction of a permutation test does not depend on the specific form of the test statistic, though some test 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 experimental studies, where the treatment is randomly assigned to units in potentially complex designs.

These standard results only apply to scenarios in which the shift δ is known nonetheless. When the constant shift is unknown and thus needs to be estimated, it becomes a nuisance parameter. The presence of a nuisance parameter under the null renders a major drawback: naively plugging an estimate into the test statistic makes the test statistic non-pivotal, and the permutation test based on the plug-in test statistic may fail to control the Type 1 error even asymptotically. In other words, the asymptotic distribution will depend on the unknown underlying distributions, offsetting its practicality in empirical research.

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

Application: Testing for Heterogeneous Treatment Effect. A key example of a hypothesis testing problem with this structure is the constant treatment effect hypothesis. Let $Y \sim F_0$ and $X \sim F_1$ be two real-valued random variables representing the potential outcomes from a randomized trial following Rubin (1974). In other words, we can think of Y and X as the outcomes for control and treatment groups, respectively. The goal of this paper is to revisit this testing problem when δ is unknown, and thus needs to be estimated, showing that our approach renders an asymptotically valid procedure.

Detecting treatment effect heterogeneity among individuals plays a key role in any successful evaluation of a social program using randomized experiments. For example, a student may benefit or suffer greatly from a policy intervention while another student may experience little to no effect. Understanding heterogeneity in treatment effects might help researchers or policy makers design or extend social programs better since the full treatment effect can be investigated in a thorough and comprehensive way. 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., Hardle and Marron (1990), Neumeyer et al. (2003), Crump et al. (2008), Imai et al. (2013)). Notwithstanding these approaches will detect some forms of treatment effect variation, their scope is limited in the sense that they only look at one aspect of the distribution, namely the mean. Only accounting for constant *average* treatment effects across subgroups while ignoring within-group heterogeneity can be misleading and fails to account for treatment effect heterogeneity. A notable example about the limitations of this method to investigate treatment effect heterogeneity solely based on averages can be found in Bitler et al. (2017).

To accommodate this practice our method can be extended to testing the joint null hypothesis that treatment effects are constant within individual subgroups, while allowing for varying *average* treatment effects across subgroups. Our test will be able to detect treatment effect heterogeneity within individual subgroups even if the average treatment effects are different across subgroups. Furthermore, we provide the companion `RATest` `R` package, available on CRAN, to simplify and encourage the application of our test in empirical research.

Numerical evidence suggests that the performance of the new test when testing for heterogeneous treatment effects is comparable to that of Koenker and Xiao (2002), Chernozhukov and Fernández-Val (2005), Linton et al. (2005), and Ding et al. (2015), outperforming them in scenarios where others fail, such as unbalanced control/treatment sample sizes, or when sample size is small. In all these cases, however, there are substantial differences between their methods and the one presented in this paper. Koenker and Xiao (2002), Linton et al. (2005) and Chernozhukov and Fernández-Val (2005) exploit the relationship between CDFs and quantiles, and their testing approach is based on Kolmogorov-Smirnov or Cramér-von Mises test statistics which are defined on the *empirical quantile regression process*. Furthermore, Linton et al. (2005) and Chernozhukov and Fernández-Val (2005) propose resampling methods to overcome the effects of the estimated nuisance parameter in the limiting distribution, while Koenker and Xiao (2002), on the other hand, use the Khmaladze decomposition of the empirical quantile regression process to restore the asymptotically distribution free nature of the test. Despite the fact that we use a Khmaladze transformation of the empirical process and our test statistic is based on this transformation, this modified process is simply the input for the construction of an asymptotically valid permutation test.

Another technique to treatment effect heterogeneity that relies on the comparison of CDFs is the one in Goldman and Kaplan (2018) .It is worth mentioning that even though Goldman and Kaplan (2018) are testing for equality at each point in the distribution, they cast this question as a multiple hypothesis testing of a continuum of CDFs hypothesis, which is rather different from (3).

Perhaps the most related paper to ours is Ding et al. (2015), who use a Fisher randomization test based on the comparisons of CDFs using a Kolmogorov-Smirnov statistic. But, there is one key distinction that fundamentally differentiates both methods. Our test relies on a martingale decomposition of the empirical process that renders an *asymptotically* pivotal test. Ding et al. (2015), on the other hand, yield valid inference by constructing a confidence interval

for the constant shift, pointwise repeating the test procedure over that interval, and taking the maximum p-value. As a result, their method does not rely on asymptotic methods though it is more conservative than our approach, as we will show in the Monte Carlo exercise.

The remainder of this paper is organized as follows. Section 2 formally states the problem, the Kolmogorov-Smirnov type test statistic and the permutation test based on it, with emphasis on the application to testing for treatment effect heterogeneity. We first consider the case when the shift is known for the sake of exposition and to fix notation and key ideas. Then, we relax the assumption about knowing δ and develop the so-called Durbin problem with unknown δ, highlighting the main drawbacks of a naive approach which ignores the estimated nuisance parameter. Section 3 contains the main theoretical results, and takes up some technical and computational aspects of the Khmaladze transformation. A short discussion on testing the null hypothesis of constant treatment effects within subgroups while allowing the treatment effects to vary across subgroups can be found in Section 4. Section 5 presents the Monte Carlo designs and compares our proposed test to the existing approaches of Koenker and Xiao (2002), Chernozhukov and Fernández-Val (2005), Linton et al. (2005), and Ding et al. (2015). Section 6 contains the empirical application, and conclusions are left in Section 7. Auxiliary results appear in Appendix A. Some additional results regarding the asymptotic behavior of permutation distributions can be found in Appendix B. The proof of the main result is in Appendix C.

2 Testing Goodness of Fit

2.1 Set Up

Throughout this paper, we contemplate the following setting. Consider two real-valued random variables $Y \sim F_0$ and $X \sim F_1$. We want to test whether Y and X are a constant shift apart:

$$H_0 : F_1(t+\delta) = F_0(t) \ \forall\, t, \quad \text{for some} \quad \delta \tag{1}$$

based on two independent samples $Y_1, \ldots, Y_n$ and $X_1, \ldots, X_m$ from F_0 and F_1, respectively. We ultimately want to consider the case where δ is not specified, but first let us consider the case when the constant shift is *known* to ease the exposition.

RUNNING EXAMPLE: One way to clarify ideas and to gain further intuition about our testing procedure is to relate it to our empirical application. Consider the simplest model for a randomized experiment with subject i's (continuous) response Y_i to a binary treatment D_i. Assume we have a sample of size N and we randomly assign treatment to $m < N$ of them, while the remaining $n = N - m$ subjects are not exposed to such treatment. We will denote the m individuals in the first group as *treatment group* while the second group of size n will be the *control group*.

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

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

The treatment effect is defined by the difference between potential outcomes, *i.e.* individual i's treatment effect is $\delta_i = Y_i(1) - Y_i(0)$, for all $i = 1, \ldots, N$. The treatment effect is *constant* if

$\delta_i = \delta$ for all i, otherwise we say the treatment effect is *heterogeneous* in the sense that it varies across subjects. As a result, the hypothesis of constant effect is

$$H_0 : Y_i(1) - Y_i(0) = \delta \quad \forall \; i \quad \text{for some} \quad \delta \tag{2}$$

This hypothesis, however, is not directly testable because we happen to observe at most one potential outcome for each unit. A different but testable hypothesis is available if we consider the marginal distributions of the observed outcomes for each group. Therefore, the testable hypothesis becomes.

$$H_0 : F_1(y + \delta) = F_0(y) \;\; \forall \, y, \quad \text{for some} \quad \delta \tag{3}$$

based on two independent samples from F_0 and F_1. In what follows, we will adopt the potential outcomes notation.

We now discuss two assumptions:

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. *The CDFs 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 w.r.t δ.*

Assumption A.1 is standard for the asymptotic results. However, its relevance will become more palpable when we investigate the asymptotic behavior of the permutation distribution because, as we will show, it behaves like the unconditional distribution of the test statistic when all N observations are i.i.d. from the mixture distribution $pF_1 + (1 - p)F_0$. Assumption A.2, on the other hand, will be key to establishing the properties of the permutation test when when δ is unknown, which is our case of interest. In particular, we will require this smoothness condition to expand the empirical process around the nuisance parameter δ. More details in Section 2.4 and Appendix B, C.

2.2 Test Statistic

A natural candidate for a test statistic for the hypothesis (3) is to compare empirical CDFs

$$\hat{F}_1(y + \delta) = m^{-1} \sum_{i=1}^{m} \mathbf{1}_{\{Y_i(1) \leq y + \delta\}}, \;\; \hat{F}_0(y) = n^{-1} \sum_{j=1}^{n} \mathbf{1}_{\{Y_i(0) \leq y\}}$$

for some δ based on two independent samples from F_0 and F_1, that we collect in Z:

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

This gives rise to the *classical* Kolmogorov-Smirnov goodness of fit test statistic:

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

where

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

is the two-sample classical empirical process. A well-known result in the theory of stochastic processes states that the distribution of $K_{m,n,\delta}$ under H_0 is the same for all continuous F_0 and F_1. Let $c_{n,\alpha}$ be the $1 - \alpha$ quantile of the distribution of $K_{m,n,\delta}$ under any continuous F_0 and F_1. Then the Kolmogorov-Smirnov test rejects the null (3) for large values of $K_{m,n,\delta}$ i.e. if $K_{m,n,\delta} > c_{n,\alpha}$.

2.3 Permutation Test

Consider data Z taking values in a sample space $\mathcal{Z}$, and let $\Omega = \{(P, Q)\}$ be a family of pairs of probability distributions. Let $\bar{\Omega} = \{(P, Q) : P = Q\}$, and suppose we are testing the null hypothesis $H_0 : (P, Q) \in \Omega_0$, where $\Omega_0 \in \bar{\Omega}$. Let $\mathbf{G}_N$ be the set of all permutations π of $\{1, \ldots, N\}$. Consider the following assumption:

Randomization Hypothesis: (Lehmann and Romano (2005), p. 633) Under the null hypothesis, the distribution of Z is invariant under transformations in $\mathbf{G}_N$; that is, 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 this randomization hypothesis, observations can be permuted and the resulting distribution is the same as that of the original samples. Thus, under the randomization hypothesis an exact level α test can be constructed by a permutation test as follows. Consider any test statistic $T_{m,n}$. Given the test statistics $T_{m,n}$, recompute $T_{m,n}$ for all permutations π, i.e. calculate $T_{m,n}(z_{\pi(1)}, \ldots, z_{\pi(N)})$ for all $\pi \in \mathbf{G}_N$. Order these values

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

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

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

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

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

Then, under any $(P, Q) \in \Omega_0$, the resulting permutation test is exact level α:

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

In addition, define the *permutation distribution* based on the test statistic $T_{m,n}$ as

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

in other words, the permutation distribution is the empirical distribution of the $N!$ values resulting from recomputing the test statistic for all permutations of the sample.

RUNNING EXAMPLE — CTD: Under the null hypothesis (3), the CDFs are a constant shift apart for some δ, so the randomization hypothesis holds. Thus if we recenter the observations coming from the treatment group by the treatment effect, *i.e.* $Y_i(1) - \delta$, then we could permute the observations from control and treatment groups and have the same joint distribution. This will allow us to construct an exact α level test for the null hypothesis (3) based on the Kolmogorov-Smirnov test statistic (4). The permutation distribution based on (4) is given by

$$\hat{R}_{m,n}^{K(\delta)}(t) = \frac{1}{N!} \sum_{\pi \in \mathbf{G}_N} I\{K_{m,n,\delta}(z_{\pi(1)}, \ldots, z_{\pi(N)}) \leq t\} \tag{7}$$

Hence, the permutation test rejects the null hypothesis (3) if $K_{m,n,\delta}(z)$ is bigger than the $1 - \alpha$ quantile of the permutation distribution (7).

2.4 Durbin Problem

The sequences of empirical processes (5), viewed as random functions, converge in distribution[1] to a Gaussian process $\mathbb{G}$ whose marginal distributions are zero-mean with covariance structure

$$\mathbb{E}\,\mathbb{G}(s)\mathbb{G}(t) = F_0(s \wedge t) - F_0(s)F_0(t) \tag{8}$$

In other words, $\mathbb{G}$ is an F_0-Brownian Bridge process, and the limiting distribution of (4), which we will denote $J_0(\cdot)$, follows as stated in the following result.

Theorem 1. . *Assume $Y_1(0), \ldots, Y_n(0)$ are i.i.d. according to a probability distribution F_0, and independently $Y_1(1), \ldots, Y_m(1)$ are i.i.d. F_1. Consider testing the hypothesis (3) for some known δ known based on the test statistic (4). Under condition A.1, $K_{m,n,\delta}$ converges weakly under the null hypothesis to*

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

where $\mathbb{G}(\cdot)$ is a Gaussian process with covariance structure given by (8).

This uniform central limit theorem, originally due to Donsker (1952) based on previous work by Kolmogorov (1933) and Doob (1949), plays a key role in the theory of stochastic processes and goodness of fit problems.

Remark 1. Exploiting the fact that the underlying distributions are absolutely continuous, it can be readily shown that the limiting distribution above is pivotal. The change of variable $y \mapsto F_0^{-1}(t)$ renders uniform empirical processes,

$$\begin{aligned} \upsilon_{m,n}(t,\delta) &= \sqrt{\frac{mn}{N}}\left(\frac{1}{m}\sum_{i=1}^{m} 1_{\{Y_i(1)-\delta \le F_0^{-1}(t)\}} - \frac{1}{n}\sum_{i=1}^{n} 1_{\{Y_i(0)\le F_0^{-1}(t)\}}\right) \\ &= \sqrt{\frac{mn}{N}}\left(\hat{F}_1(F_0^{-1}(t)+\delta) - \hat{F}_0(F_0^{-1}(t))\right) \end{aligned} \tag{9}$$

In other words, the empirical process defined in (5) is equivalent to a process based on N i.i.d. uniform variables and its limiting distribution is a Brownian Bridge on $[0,1]$, which we will denote as $\mathbb{B}^0$. ■

Remark 2. Theorem 1 and Remark 1 show that a test based on the uniform empirical process is particularly attractive because it is asymptotically distribution-free *i.e.* when δ is known, the limiting distribution of the Kolmogorov-Smirnov test statistic is the same regardless of the underlying distribution generating the data — the supremum of a standard Brownian Bridge process. Furthermore, it follows that if the null hypothesis holds, so that $\hat{F}_0$ and $\hat{F}_1$ are independent empirical distribution functions from the same continuous distribution function, then the classical KS statistic converges weakly to the same limit distribution as in the one-sample two-sided case. ■

In addition, the following theorem shows that the permutation distribution (7) converges in probability to the same limit law as the true unconditional limiting distribution $J_0(\cdot)$.

Theorem 2. *Assume the premises of Theorem 1. Then the permutation distribution (6) based on $K_{m,n,\delta}$ is such that*

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

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

[1] Proofs and other asymptotic results in Sections 2.4 can be found in Appendix A (Asymptotic Results). We omitted them here for the sake of exposition.

So far, we have dealt with the case when δ is known. However, the location shift or treatment effect, is unknown in practice nonetheless. This is our case of interest. As a result, the hypothesis (3) makes this case a non-parametric hypothesis testing problem with an estimated nuisance parameter, and then limiting distribution of the test statistic (4) with estimated δ will depend on the underlying distribution F_0, jeopardizing the asymptotically distribution-free nature of the test, which is known as the Durbin problem (Durbin (1973)).

Following previous discussion, the straight forward approach would be to compare the CDFs with "plug-in"[2] δ:

$$\hat{F}_1(y+\hat{\delta}) = m^{-1}\sum_{i=1}^{m}\mathbf{1}_{\{Y_i(1)\leq y+\hat{\delta}\}},\quad \hat{F}_0(y) = n^{-1}\sum_{j=1}^{n}\mathbf{1}_{\{Y_i(0)\leq y\}}$$

We may equivalently consider the *shifted* Kolmogorov-Smirnov test statistic:

$$K_{m,n,\hat{\delta}}(Z) = \sup_y |V_{m,n}(y,\hat{\delta})| \ . \tag{10}$$

where

$$V_{m,n}(y,\hat{\delta}) = \sqrt{\frac{mn}{N}}\left(\hat{F}_1(y+\hat{\delta}) - \hat{F}_0(y)\right) \tag{11}$$

is the two-sample empirical process. Similarly, the permutation distribution with estimated $\hat{\delta}$ results from substituting the test statistic in (6) by (10). When δ is unknown, the empirical process (11) converges weakly to a Gaussian process $\mathbb{B}$ instead of $\mathbb{G}$. More formally, let $\xi(\cdot)$ be a Gaussian process with mean 0 and covariance structure

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

where $\sigma_0^2 = \sigma^2(F_0)$, and f_0 is the density of F_0. Then, $\mathbb{B}$ is defined as:

$$\mathbb{B}(\cdot) = \mathbb{G}(\cdot) + \xi(\cdot) \tag{12}$$

with covariance function

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

As a result, the limiting distribution of (10), denoted by $J_1(\cdot)$, will be different from $J_0(\cdot)$, the limiting distribution of the case when δ is known, as formalized in the next theorem.

Theorem 3. *Assume $Y_1(0),\ldots,Y_n(0)$ are i.i.d. according to a probability distribution F_0, and independently $Y_1(1),\ldots,Y_m(1)$ are i.i.d. F_1. Consider testing the hypothesis (3) for some unknown δ based on the test statistic (10). Under conditions A.1-A.2, $K_{m,n,\hat{\delta}}$ converges weakly under the null hypothesis to*

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

where $\mathbb{B}(\cdot)$ is given by (12).

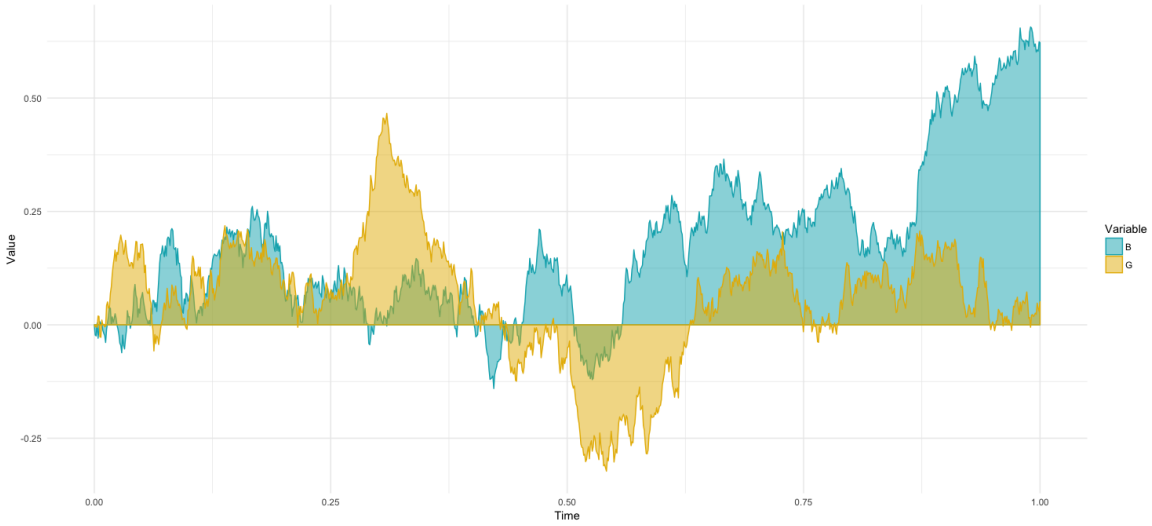
In other words, the necessity of estimating δ introduces a drift $\xi(\cdot)$, which prompts the limiting distribution of (10) to depend on F_0 and functionals of it, making the asymptotic null distribution intractable. The practical consequence of this is to make difficult, if not impossible, to obtain critical values. Figure 1 exemplifies this discrepancy when F_0 is the standard normal distribution[3].

The following theorem shows the limiting behavior of the permutation distribution based on the shifted K-S statistic given by (10)

[2] $\hat{\delta} = \mu(\hat{F}_1) - \mu(\hat{F}_0)$, where $\mu(\hat{F}_1)$ and $\mu(\hat{F}_0)$ are plug-in estimators of $\mu(F_1)$ and $\mu(F_0)$ respectively.

[3] When $Y(0)\sim\mathcal{N}(0,1)$, then $Y(0)$ conditional on $a<Y(0)<b$ has a truncated standard normal distribution.

Figure 1: Realizations of Classical and Shifted Empirical Processes



Sample paths are formed based on 1000 observations. Covariance structures calculated assuming F_0 follows the standard normal distribution.

Theorem 4. *Assume the premises of Theorem 3. Then the permutation distribution (6) based on $K_{m,n,\hat{\delta}}$ is such that*

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

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

Under the hypothesis (3), the true unconditional sampling distribution of $K_{m,n,\hat{\delta}}$ is given by $J_1(\cdot)$ in Theorem 3, which does not equal $J_0(\cdot)$ in general. Then, the permutation distribution and the true unconditional sampling distribution behave differently asymptotically in the presence of nuisance parameters. Hence, the permutation test for the hypothesis (3) fails to control the size.

Tests that are formulated as a function of (11), like the Kolmogorov-Smirnov or Cramér-von-Mises type tests, and don't take into account the dependency on F_0 (or functionals of it) may suffer from this problem and therefore, break their distribution-free character, even asymptotically. This is also the case for testing procedures defined on the *empirical quantile regression process* rather than the empirical process[4]. In order to circumvent this problem, one may adopt a resampling strategy to determine the critical values, or to bypass the nuisance parameter by removing the effect of the drift in large samples. For example, Chernozhukov and Fernández-Val (2005) and Linton et al. (2005) subsample appropriately recentered empirical quantile regression process to remove the effect of the estimated parameter. Meanwhile,

In our case, its truncated moments

$$\mathbb{E}(Y(0)|Y(0) \leq s) = -\frac{\phi(s)}{\Phi(s)} \quad \text{and} \quad \mathbb{E}(Y(0)|Y(0) > s) = \frac{\phi(s)}{1-\Phi(s)}$$

Then, the covariance function

$$\mathbb{C}(\mathbb{G}(s), \xi(t)) = \phi(t)\Phi(s)(1-\Phi(s))\left[\frac{\phi(s)}{1-\Phi(s)} + \frac{\phi(s)}{\Phi(s)}\right]$$

[4]These alternative formulations stem from the relationship between CDFs and quantiles. In the simplest case of no covariates, the quantile process for this problem is $\delta(\tau)$ such that $F_1^{-1}(\tau) = F_0^{-1}(\tau) + \delta(\tau)$.

Koenker and Xiao (2002) opt for a martingale decomposition of the quantile regression process which yields a martingale with a standard limit distribution.

In order to avoid asymptotics, Ding et al. (2015) sidestep the Durbin problem and restore valid inference on the basis of randomization alone. This finite sample alternative relies on constructing a confidence interval for the constant shift, pointwise repeating the test procedure over that interval, and taking the maximum p-value, yielding a valid yet conservative solution to the presence of a nuisance parameter.

Numerical evidence suggests that the proposed test is comparable to or outperforms the existing methods in situations considered in our simulations. Our strategy is similar to Koenker and Xiao (2002) in the sense that our permutation test will be based on the martingale transformation of the empirical process introduced by Khmaladze (1981), as we explain in the next section.

3 Permutation Test based on the Martingale Transformation

We concluded in Section 2.4 that the consequence of the drift term implied that the limiting behavior of the test statistic based on the empirical process (10) is no longer distribution-free. Khmaladze (1981) proposed a solution to this problem in the one sample case, which boils down to a Doob-Meyer decomposition of the uniform empirical process. We're going to extend Khmaladze's result to the two-sample case and work with the two sample uniform empirical process (5).

More specifically, let the real-valued function $g(s) = (s, f_0(s))'$ on $[0,1]$ be bounded and continuous in its arguments, and $\dot{g}(s) = (1, \dot{f}_0(s))'$, where $\dot{g}$ is the derivative of g. Define $C(s) = \int_s^1 \dot{g}(t)\dot{g}(t)'dt$, and assume it is invertible for $s \in [0,1)$. Then the Khmaladze transformation of the parametric empirical process (5) is given by

$$\tilde{\upsilon}_{m,n}(t,\hat{\delta}) = \upsilon_{m,n}(t,\hat{\delta}) - \int_0^y \left[\dot{g}(s)'C^{-1}(s)\int_s^1 \dot{g}(r)d\upsilon_{m,n}(r,\hat{\delta})\right] ds \ . \tag{13}$$

Khmaladze (1981) showed that (13) converges weakly to a Brownian motion process, effectively nullifying the effect of the estimated nuisance parameter $\hat{\delta}$. To gain further insight in regards the Khmaladze transformation, define the map $\phi_g : D[0,1] \to D[0,1]$ such that

$$\phi_g(h)(t) = \int_0^t \left[\dot{g}(s)'C^{-1}(s)\int_s^1 \dot{g}(r)dh(r)\right] ds \ , \tag{14}$$

Remark 3. $\phi_g(h)(\cdot)$ is the so-called compensator of h (see Parker (2013)). As noted in Bai (2003), ϕ_g is a linear mapping and $\phi_g(cg) = cg$ for a constant or random variable c. This allows us to write (13) as

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

■

In a similar fashion, we will define the *Khmaladze-transformed* version of the Kolmogorov-Smirnov test statistics

$$\tilde{K}_{m,n,\hat{\delta}}(Z) = \sup_t |\tilde{\upsilon}_{m,n}(t,\hat{\delta})| \tag{15}$$

where $\tilde{\upsilon}_{m,n}(t,\hat{\delta})$ is the Khmaladze transformation in (13). The following proposition shows the Khmaladze transformation removes the effect of $\hat{\delta}$ on the limit process.

Theorem 5. *Assume $Y_1(0), \ldots, Y_n(0)$ are i.i.d. according to a probability distribution F_0, and independently $Y_1(1), \ldots, Y_m(1)$ are i.i.d. F_1. Consider testing the hypothesis (3) for some δ based on the test statistic (15). Under conditions A.1-A.2, the limiting distribution of $\tilde{K}_{m,n,\hat{\delta}}$ is*

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

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

3.1 Khmaladze Transformation as a Continuous-time Detrending Operation

To gain further insight as to why the transformation works, we follow Bai (2003) and Parker (2013), and we consider (13) with y taking discrete values, replacing integral with sums. For instance, suppose $0 = t_0 < t_1 < \cdots < t_m < t_{m+1} = 1$ is a partition of the interval $[0, 1]$ and that y takes on values on $t_1, t_2, ..., t_m$. Write (13) in differentiation form

$$d\tilde{v}_{m,n}(t, \hat{\delta}) = dv_{m,n}(t, \hat{\delta}) - \dot{g}(t)' C^{-1}(t) \int_t^1 \dot{g}(r) dv_{m,n}(r, \hat{\delta}) dt \tag{16}$$

let

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

then the right hand side of (16) can be interpreted as the recursive residuals:

$$y_i - x_i' \left(\sum_{k=i}^{m+1} x_k x_k' \right)^{-1} \sum_{k=i}^{m+1} x_k y_k = y_i - x_i' \hat{\boldsymbol{\beta}}_i \tag{17}$$

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

3.2 Numerical Computation of the Khmaladze Transformation

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

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

then the OLS estimator based on the last $m - i + 2$ observations described on right hand side of (17) can be written as

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

which implies that the Khmaladze transformation of the empirical process in (13) can be obtained by numerically integrating from $[0, t_i)$, i.e.

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

and therefore the test statistic can be calculated as

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

3.3 Main Result

The following Theorem shows that the permutation distribution (6) based on the test statistic (15) converges in probability to the same limit law as the true unconditional limiting distribution $J_2(\cdot)$.

Theorem 6. *Assume $Y_1(0), \ldots, Y_n(0)$ are i.i.d. according to a probability distribution F_0, and independently $Y_1(1), \ldots, Y_m(1)$ are i.i.d. F_1. Consider testing the hypothesis (3) for some δ based on the test statistic (15). Under conditions A.1-A.2, the permutation distribution (6) based on the Khmaladze transformed statistic $\tilde{K}_{m,n,\hat{\delta}}$ is such that*

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

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

Thus the permutation distribution is asymptotically the supremum of a standard Brownian motion process, as is the true unconditional limiting distribution of the test statistic $\tilde{K}_{m,n,\hat{\delta}}$. Under fairly weak assumptions the theorem restores asymptotically valid inference when the shift or treatment effect δ is estimated.

Remark 4. Suppose P_0 satisfies the null hypothesis H_0, and consider a sequence of contiguous alternatives P_n to P_0. Assume that under P_0, the $(1-\alpha)$ quantile of the permutation distribution converges to the $1-\alpha$ quantile of R, where $R(\cdot)$ is the limiting distribution of the test statistic. Also assume the test statistic converges in law to R' under P_n. Then, the probabilities that the test rejects the null under P_n would tend to $1 - R'\left(R^{-1}(1-\alpha)\right)$. In other words, the power of the permutation test is essentially the same as the "asymptotic" test in large samples, meaning there is no loss in using a permutation critical value. Nevertheless, the permutation test is more robust against large nonparametric families of distributions, making this gains notable. ■

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[5]. Then, one may characterize treatment effect heterogeneity by testing whether the existing differences vary significantly across subgroups.

[5]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.

The permutation test proposed in the paper can be implemented to test the adequacy of this traditional approach that assumes a constant-treatment-effects model, *i.e.* we can 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 across mutually exclusive subgroups while the average treatment effects can vary across subgroups.

Remark 5. Our approach to test the usefulness of the constant-effects-model to characterize heterogeneity is similar to the one developed in Bitler et al. (2017). Their approach starts by assuming the constant-effects-model is correct and then constructing what they coined as a simulated-outcomes distribution. Once equipped with this auxiliary distribution, they then test for equality of distributions between the actual observed outcomes and the simulated outcomes. Here, we need not to construct such a simulated distribution and we test based on the actual observed distribution instead. As a result, these two ways to investigate the acceptability of the constant-effects-model involve quite different theoretical arguments. ■

To formalize the ongoing discussion, suppose we split data into G mutually exclusive subgroups defined by covariates. The null hypothesis of interest is now

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

where $F_0^g(y)$ and $F_1^g(y)$ are the cumulative distribution functions (CDFs) of the control and treatment group, respectively, for subgroup $g = 1, \ldots, G$. Note that the nuisance parameter δ_g for subgroup g can vary across subgroups.

Remark 6. To this end, as will be explained in Algorithm 1, we will be testing as many multiple hypotheses simultaneously as the number of subgroups G. If one ignores the multiplicity issue and tests each hypothesis at level α, the probability of one or more false rejections may be much greater than α. Thus, we carefully conduct our test while controlling the familywise error rate (FWER) at level α using a Bonferroni method. ■

Consider now for each mutually exclusive subgroup g,

$$Z_g = (Y_1(1), \ldots, Y_{m_g}(1), Y_1(0), \ldots, Y_{n_g}(0))$$

for all $g = 1, \ldots, G$ such that $\sum_g n_g = n$ and $\sum_g m_g = m$. Following our results on the permutation test based on the Khmaladze transformation, an algorithm for testing the null hypothesis H_0^g is given by the following.

Algorithm 1 (Testing Treatment Effect Heterogeneity Across Subgroups)

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

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

5 Monte Carlo Simulations

5.1 Implementation

The martingale transformation described in 3 uses the true density and score functions. In the Monte Carlo experiments of section 5.1.1, both functions were estimated employing the

[6]To be more precise, one can use randomization explained in the permutation construction described in Section 2.3.

univariate adaptive kernel density estimation (e.g. Portnoy and Koenker, 1989; Koenker and Xiao, 2002), and the estimates were obtained directly from the **R** package **quantreg** (Koenker (2016)). Simulation results using the true density and score functions were similar in magnitude and therefore not shown in here, though available upon request.

Even though the computation of the permutation distribution (6) would require the calculation of the test statistics for all $N!$ permutations of $\{1, \ldots, N\}$, one can approximate this distribution arbitrarily close by Monte-Carlo approximation — randomly sample permutations π from $\mathbf{G}_N$ without replacement and recompute the test statistic for these samples. The law of large numbers guarantees that the quantiles of this stochastic approximation converges to the quantiles of (6) (see Problem 15.4 in Lehmann and Romano (2005)).

5.1.1 Validity of Permutation Test

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

Table 1 contains the rejection probability results of the Monte Carlo simulations. We generated samples from three different distributions: standard normal, lognormal, and student's t distribution with 5 degrees of freedom. In this experiment sample sizes vary between groups[7]. We considered the sequence of total sample size $N \in \{13, 50, 80, 200\}$, and for each sample size and distribution, a constant treatment effect $\delta = 1$ was assigned[8]. We ran these simulations with 5000 replications across Monte Carlo Experiments.

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

Moreover, these experiments confirm the story of the theoretical results in section 2.4: the permutation test based on the (naive) shifted KS statistic fails to control the type I error, even in large samples. We argued that the permutation distribution based on the shifted KS statistic depends on the underlying law that generates the data and therefore, the permutation distribution is no longer asymptotically distribution free. Although Ding et al. (2015) did not compute the permutation distribution using this naive KS statistic, the conclusions of their naive approach are similar to those found in here[9]. As shown in Table 1, the permutation test is either too conservative (normal and student's t) or it fails to control the size (lognormal). In the case of skewed distributions, the size of the test increases with the sample size.

Both the confidence interval FRT (FRT CI) by Ding et al. (2015) and subsampling by Chernozhukov and Fernández-Val (2005) control the rejection probabilities across different sample

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

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

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

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

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

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

sizes and data generating processes, but in a rather conservative fashion nonetheless. For instance, when the total sample size is either 50 or 13, FRT CI test is hyper-conservative. We also show a similar conclusion regarding the bootstrap. More specifically, the Bootstrap is not valid across distributions for $N > 50$. This is not surprising since the bootstrap does not have the same generality as, say, subsampling.

5.1.2 Power of the test

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

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

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

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

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

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

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

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

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

6 Empirical Application

We briefly revisit an experiment by Gneezy and List (2006), also considered in Goldman and Kaplan (2018), on the effects of gift exchange on worker effort, the so-called *gift exchange hypothesis*. The underlying idea behind this model is the assumption that there exists a positive relationship between wages and worker effort levels. To assess this hypothesis, the authors conducted two field experiments.

In the first experiment, experimental subjects were required to computerize the holdings of a library at an hourly wage of \$12. Individuals in the treatment group were later informed that they would be paid \$20 instead. In line with the gift exchange model, individuals exhibited higher effort in the first period (first 90 min), and the effort levels between control and treatment groups were not statistically significant in subsequent periods. In the second experiment, the participants were asked to engage in a door-to-door fund-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 fist period (few hours before lunch) than solicitors in the control group, but this effect disappeared in the second period (few hours after lunch).

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, where the zero treatment effect null hypothesis is not rejected.

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

	Library Task			Fundraising Task		
Time Period	Mean $T-C$ Difference	Test Statistic	p-value	Mean $T-C$ Difference	Test Statistic	p-value
1	10.96^{**}	0.73	0.24	13.80^{**}	0.76	0.88
2	4.38	0.73	0.28	1.17^{*}	1.09	0.085
3	0.46	0.66	0.98			
4	0.73	0.68	0.92			

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 before and 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. Columns 3 and 6 report the Khmaladze transformed test statistic (15), with corresponding $p-$values in columns 4 and 7. Stochastic approximations for the computation of $p-$values were calculated using 999 permutations.
Significance at $p < 0.1$ and $p < 0.05$ is denoted with * and **, respectively.

Table 3 shows the results from our test. For the first period of the library experiment, we fail to reject the null hypothesis that this nearly 25 percent difference treatment effect of the gift exchange induces a constant shift between the distributions of control and treatment groups ($p = 0.24$). This conclusion is also reached in Goldman and Kaplan (2018), although their analysis finds almost rejection in upper quantiles[10]. Furthermore, the same conclusion holds when we look at the subsequent periods — we do not have enough evidence in favor of treatment effect heterogeneity ($p = 0.28, p = 0.98$, and $p = 0.92$).

In like manner, our Khmaladze transformed permutation test does not reject that effort CDFs between treatment and control groups are a constant shift apart in the pre-lunch period of the fund-raising experiment ($p = 0.88$). However, out test sheds some light in the second period for we reject the null hypothesis at a 10% level in favor of heterogeneity in the treatment

[10]It is worth mentioning that 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 CDFs hypothesis.

effect of the gift exchange. We therefore complement the results in Gneezy and List (2006), which reported no statistically significant effect after lunch, masking potential heterogeneity since they are only looking at one aspect of the distribution, namely the mean. Thus, data suggest that in the second period, the gift exchange has a heterogeneous effect on effort levels, in spite the average effect did not provide compelling evidence in favor of the gift exchange hypothesis.

7 Conclusions

This paper studied the classical goodness-of-fit hypothesis testing with a nuisance parameter. The leading example was testing for heterogeneity in the treatment effect in the context of a randomized experiment. The main result of this paper showed that an asymptotically valid permutation test is readily available for this problem. The permutation tests presented here exploits the martingale transformation of the empirical process to annihilate the effect of the estimated nuisance parameter and restore the validity of the permutation test. Numerical evidence suggests that the performance of the new test when testing for heterogeneous treatment effects is comparable to existing methods in literature, outperforming them in certain scenarios such as unbalanced control/treatment sample sizes, or when sample size is small.

We have developed the `R` package `RATest`, which simplifies the implementation of the test we propose in this paper for the empirically oriented researchers. We apply our test to investigate the gift exchange hypothesis in the context of two field experiments from Gneezy and List (2006). Our test complements their results in two ways. First, it fails to reject the null that the gift exchange effect induces a constant shift in the productivity distribution of those who were treated. Second, it rejects the null hypothesis in favor of the heterogeneity in the treatment effect where solely looking at the average treatment effect does not provide evidence in favor of the gift exchange hypothesis.

References

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

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

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

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

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

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

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

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

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

Doob, J. L. (1949). Heuristic approach to the kolmogorov-smirnov theorems. *The Annals of Mathematical Statistics*, pages 393–403.

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

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*.

Hardle, 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.

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

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

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

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

Kolmogorov, A. (1933). Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.*, 4:83–91.

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

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

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

Olivares, M. and Sarmiento, I. (2017). *RATest: Randomization Tests*. R package version 0.1.4.

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

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

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

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

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

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

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

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

Appendix

Appendix A: Auxiliary Results

The Coupling Construction

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

$$Z = (Z_1, \ldots, Z_N) = (Y_1(1), \ldots, Y_m(1), Y_1(0), \ldots, Y_n(0)) \tag{18}$$

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

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

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

We would wish to compare

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

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

Remark 7. The elements of $\bar{Z}$ can be thought as the outcome of a compound lottery. First, draw a random index j from $\{0, 1\}$ with probability $\mathbb{P}(j = 0) = p$. Then, conditionally on the outcome being j, sample $\bar{Z}_i$ from F_0 if $j = 0$, and from F_1 otherwise. ■

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

Remark 8. Recall the binomial distribution is used to model the number of successes m when sampling with replacement from a population of size N. Hence, the number of observations $\bar{Z}_i$ out of N which are from population F_0 follows the Binomial distribution with parameters N and p. This number has mean $Np \approx n$, whereas the exact number of observations from F_0 in Z is n. ■

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

The algorithm

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

again, then just sample a new observation $Y_{n+1}(0)$ from F_0, and analogously if the observations you've exhausted are from population F_1. Continue this way so that as many as possible of the original Z_i observations are used in the construction of $\bar{Z}$. After this, you will end up with Z and $\bar{Z}$, with many of their coordinates in common (and this is why this method is called "coupling," because we couple $\bar{Z}$ with Z). The number of observations which differs, say D, is the (random) number of added observations required to fill up $\bar{Z}$. You can access this R file to see how this algorithm works.

Reordering according to π_0

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

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

Why does coupling work?

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

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

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

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

$$\left(K_{m,n}(\bar{Z}_\pi), K_{m,n}(\bar{Z}_{\pi'})\right) \xrightarrow{\text{d}} (T, T') \tag{19}$$

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

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

Using the coupling construction, suppose it can be shown that

$$K_{m,n}(\bar{Z}_{\pi\pi_0}) - K_{m,n}(\bar{Z}_\pi) \xrightarrow{\text{P}} 0 \tag{20}$$

then it also follows that

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

and by Slutsky's theorem

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

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

Asymptotic Results

Theorems 1, 3, and 5, as well as their proofs are presented in this manuscript for the sake of completeness. The two-sample *Donsker*'s theorem (Theorem 1) is a straightforward extension of the one-sample case (for the one sample case see Theorem 19.3 in Van der Vaart (2000) . Theorem 3 is proven in the Appendix of Ding et al. (2015), Theorem 4, with one minor modification regarding the normalizing constants in the test statistic. For additional insights, see the discussion and results in examples V.15 and V.23 in Pollard (2012). Finally, Theorem 5 is the two-sample extension of Khmaladze (1981), Theorem 4.3.

Theorem 1. . *Assume $Y_1(0), \ldots, Y_n(0)$ are i.i.d. according to a probability distribution F_0, and independently $Y_1(1), \ldots, Y_m(1)$ are i.i.d. F_1. Consider testing the hypothesis (3) for some known δ known based on the test statistic (4). Under condition A.1, $K_{m,n,\delta}$ converges weakly under the null hypothesis to*

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

where $\mathbb{G}(\cdot)$ is a Gaussian process with covariance structure given by (8).

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

$$V_{m,n}(y,\delta) = \sqrt{\frac{mn}{N}}\left(\hat{F}_1(y+\delta) - \hat{F}_0(y)\right) = \sqrt{1-p_m}V_1 - \sqrt{p_m}V_0$$

where $V_0 = \sqrt{n}(\hat{F}_0(y) - F_0(y))$ and $V_1 = \sqrt{m}(\hat{F}_1(y+\delta) - F_1(y+\delta))$ are two independent empirical processes. By Donsker's theorem, both sequences V_0 and V_1 can be approximated by two independent F_0 and F_1 Brownian bridge processes, $\mathbb{G}_0$ and $\mathbb{G}_1$ respectively. We can take these Brownian bridges to be independent because the empirical processes are. Therefore, $V_{m,n}(y,\delta)$ converges weakly to

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

which is another zero-mean Brownian Bridge with the same covariance structure as $\mathbb{G}(\cdot)$. Therefore, by the usual continuous mapping theorem, the sequences of "classical" KS statistic $K_{m,n,\delta} = \sup_y |V_{m,n}(y,\delta)|$ converge under the null hypothesis to

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

□

The following theorem establishes the asymptotic behavior of the Kolmogorov-Smirnov test statistic when δ is replaced by $\hat{\delta}$. As a result of estimating the nuisance parameter, an additional smoothness condition is required as stated in assumption A.2. It is shown that the test statistic follows an asymptotic law $J_1(\cdot)$ which is different from $J_0(\cdot)$, the limiting distribution of the case when δ is known. See also Ding et al. (2015).

Theorem 3. *Assume $Y_1(0), \ldots, Y_n(0)$ are i.i.d. according to a probability distribution F_0, and independently $Y_1(1), \ldots, Y_m(1)$ are i.i.d. F_1. Consider testing the hypothesis (3) for some unknown δ based on the test statistic (10). Under conditions A.1-A.2, $K_{m,n,\hat{\delta}}$ converges weakly under the null hypothesis to*

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

where $\mathbb{B}(\cdot)$ is given by (12).

Proof. Under the null hypothesis (3), we know that for some δ, $\delta = \mu(F_1) - \mu(F_0)$, $\sigma^2(F_1) = \sigma^2(F_0) = \sigma^2$, and $f_1(y+\delta) = f_0(y)$. Then we develop $V_{m,n}(y, \hat{\delta})$ as

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

due to the fact the last two summands

$$\sqrt{\frac{mn}{N}}\left\{(\hat{F}_1(y+\hat{\delta}) - F_1(y+\hat{\delta})) - (\hat{F}_1(y+\delta)) - F_1(y+\delta)\right\} = o_p(1) \tag{21}$$

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

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

Observe

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

therefore

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

since both terms have the same limit distribution as *shifted* Brownian Bridges in Theorem 3, we have

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

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

The following theorem establishes the asymptotic behavior of the Kolmogorov-Smirnov test statistic based on the Khmaladze transformation of the empirical process. In particular, it is shown that the test statistic follows an asymptotic law that is the supremum of the standard Brownian motion. We considered a uniform empirical process in Remark 1, denoted $\upsilon_{m,n}(t,\delta)$. In a similar fashion, $\upsilon_{m,n}(t,\hat{\delta})$ will denote the uniform empirical process with estimated δ. See also Khmaladze (1981).

Theorem 5. *Assume $Y_1(0),\ldots,Y_n(0)$ are i.i.d. according to a probability distribution F_0, and independently $Y_1(1),\ldots,Y_m(1)$ are i.i.d. F_1. Consider testing the hypothesis (3) for some δ based on the test statistic (15). Under conditions A.1-A.2, the limiting distribution of $\tilde{K}_{m,n,\hat{\delta}}$ is*

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

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

Proof. The outline of the proof is the following. We work with the uniform empirical process of Remark 1 with estimated δ, and exploit the differentiability with respect to δ (condition A.1) to expand around it. We use the properties of the map ϕ_g and Khmaladze theorem to prove weak convergence to the standard Brownian motion.

Consider the asymptotic representation

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

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

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

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

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

therefore

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

Weak convergence of $\upsilon_{m,n}(t,\delta)$ to $\mathbb{B}^0$ was established in Remark 1. Thus, $\tilde{\upsilon}_{m,n}(t,\hat{\delta})$ weakly converges to the Brownian motion $\mathbb{B}^0(t) - \phi_g(\mathbb{B}^0(t))$, by 4.3 of Khmaladze (1981). The convergence of $\tilde{K}_{m,n,\hat{\delta}}$ follows by the usual continuous mapping theorem.

□

Appendix B: Asymptotic Behavior of the Permutation Distribution

NOTATION: In what follows, it should be understood that $\mathbb{G}$ refers to a zero-mean Gaussian process with covariance structure described by (8). In addition, π and π' will denote two independent random permutations of $\{1,\ldots,N\}$, and π_0 will denote the permutation that reorders observations in $\bar{Z}$, as described in Appendix A. In order to emphasize the data that are being used in the computation of the two-sample empirical processes, we will write $V_{m,n}(y,\hat{\delta};Z_\pi))$ or $V_{m,n}(y,\hat{\delta};\bar{Z}_\pi))$, meaning that $V_{m,n}(y,\hat{\delta})$ was calculated using sample $(Z_{\pi(1)},\ldots,Z_{\pi(N)})$ or $(\bar{Z}_{\pi(1)},\ldots,\bar{Z}_{\pi(N)})$, respectively. Analogously, $V_{m,n}(y,\hat{\delta};\bar{Z}_{\pi'}))$ is defined with π replaced by π'.

Theorem 2: Limiting Behavior of $\hat{R}_{m,n}^{K(\delta)}(t,\delta)$

DESCRIPTION: This theorem establishes the asymptotic behavior of the Permutation Distribution based on the Kolmogorov-Smirnov test statistic when the parameter δ is known. In particular, it is shown that the permutation distribution behaves asymptotically like the true unconditional limiting distribution of the classical KS statistic *i.e.* the supremum of a Gaussian process given by $\mathbb{G}$. This result follows from the arguments in Romano (1989). We decided to include a proof in this paper for completeness and because the proof strategy we follow will be useful for other results.

PRELIMINARIES: Let us recenter the m observations coming from F_1 as follows

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

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

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

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

Theorem 2. *Assume the premises of Theorem 1. Then the permutation distribution (6) based on $K_{m,n,\delta}$ is such that*

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

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

Proof. The outline of the proof is the following. We show the finite-dimensional distributions of $(V_{m,n}(y,\delta;Z_\pi)), V_{m,n}(y,\delta;Z_{\pi'})))$ converge weakly to the marginals $(\mathbb{G}(y),\mathbb{G}'(y))$, and that $\mathbb{G}(y)$ and $\mathbb{G}'(y)$ are independent. Then we use Theorem 1.5.4 in Wellner and Van der Vaart (2013) to establish weak convergence. The limiting distribution of $(K_{m,n,\delta}(Z_\pi), K_{m,n,\delta}(Z_{\pi'}))$ to (J_0, J_0') follows by the regular continuous mapping theorem. Finally, we use Hoeffding's Condition (Theorem 5.1 of Chung and Romano (2013)) to conclude that the permutation distribution converges in probability to the same limit law as the true unconditional limiting distribution.

Weak Convergence. We want to show that the marginals

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

converge weakly to the marginals

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

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

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

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

Notice that under the null,

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

We claim the asymptotic normality of

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

To do this, we use the Cramér-Wold device (Theorem 11.2.3 of Lehmann and Romano (2005)). Then, for any any a and b, we must verify the limiting distribution of

$$K(m)\sum_{i=1}^N (aX_iW_i + bX_iW_i')) = \sum_{i=1}^N C_{m,n,i}X_i \tag{22}$$

where

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

Condition on W_i and W_i', then (22) is a conditionally independent sum of linear combination of independent variables:

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

By the arguments in Example 15.2.5 of Lehmann and Romano (2005), we conclude that

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

and so

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

therefore

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

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

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

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

Limit Law of $\hat{R}_{m,n}^{K(\delta)}$. From the previous results, it now follows that $(K_{m,n,\delta}(Z_\pi), K_{m,n,\delta}(Z_{\pi'}))$ are asymptotically independent. By the regular the continuous mapping theorem,

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

converges in distribution to the (J_0, J_0') process with independent, identically distributed marginals as described in Theorem 1. Therefore, by Hoeffding's Condition (Theorem 5.1 of Chung and Romano (2013)),

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

□

Theorem 4: Limiting Behavior of $\hat{R}_{m,n}^{K(\hat{\delta})}(t, \delta)$

DESCRIPTION: This Proposition establishes the asymptotic behavior of the Permutation Distribution based on the Kolmogorov-Smirnov test statistic when the parameter δ is unknown. In particular, it is shown that the permutation distribution behaves asymptotically like the true unconditional limiting distribution of the classical KS statistic *i.e.* the supremum of a Gaussian process given by $\mathbb{G}$, which is in general different than $\mathbb{B}$.

PRELIMINARIES I: Since we don't know δ, we cannot shift the observations as we did in the case of δ known. Instead, we will recenter the m observations coming from F_1 using $\hat{\delta} = \mu(\hat{F}_1) - \mu(\hat{F}_0)$. More specifically, $\tilde{Y}_i(1) = Y_i(1) - \hat{\delta}$ for all $i = 1, \ldots, m$ where $\tilde{Y}_i(1) \sim \tilde{F}_1$.

PRELIMINARIES II: The general proof strategy will be based on the contiguity and coupling construction results in section 5 of Chung and Romano (2013) and Appendix A in this paper. The key idea is that the permutation distribution based on Z should behave approximately like the behavior of 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 $\bar{P}$. In order to establish this result, we will need the following two lemmas.

Lemma 1. *Under conditions A.1 and A.2, let $\bar{Z}_1, \bar{Z}_2 \ldots,$ be i.i.d. from the mixture distribution $\bar{P} = p\tilde{F}_1 + (1-p)F_0$, and denote $\bar{Z} = (\bar{Z}_1, \ldots, \bar{Z}_N)$. Let π and π' be independent of $\bar{Z}$. Then*

$$\Big(V_{m,n}(y, \hat{\delta}; \bar{Z}_\pi)), V_{m,n}(y, \hat{\delta}; \bar{Z}_{\pi'}))\Big)$$

converges weakly to $(\mathbb{G}, \mathbb{G}')$ with $\mathbb{G}$ and $\mathbb{G}'$ two independent Gaussian processes with common CDF.

Proof. The outline of the proof is similar to the proof of Theorem (2), *i.e.* we want to show that the finite-dimensional distributions of $\Big(V_{m,n}(y, \delta; \bar{Z}_\pi)), V_{m,n}(y, \delta; \bar{Z}_{\pi'}))\Big)$ converge weakly to the marginals $(\mathbb{G}(y), \mathbb{G}'(y))$, and that $\mathbb{G}(y)$ and $\mathbb{G}'(y)$ are independent with common CDF. Then a direct application of Theorem 1.5.4 in Wellner and Van der Vaart (2013) will establish weak convergence, finishing the proof.

We need to show the marginals

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

converge weakly to the marginals

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

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

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

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

$$\mathbb{E}(W_i) = 1\,\mathbb{P}(\pi(i) \in I_1) - m/n\,\mathbb{P}(\pi(i) \notin I_1) = 0$$

and $\mathbb{E}((1_{\{\bar{Z}_{\pi(i)\}} \le y} - F_0(y))W_i) = 0$ since π is independent of $\bar{Z}$. Same is true for W_i'.

Notice that under the null,

$$\begin{aligned}
\mathbb{E}\left(V_{m,n}(y,\hat{\delta};\bar{Z}_\pi)\right) &= p\tilde{F}_1(y) + (1-p)F_0 - F_0 \\
&= pf_0(y)(\hat{\delta}-\delta) + o_p(1) = \hat{R}_{m,n} \\
\mathbb{V}\left(V_{m,n}(y,\hat{\delta};\bar{Z}_\pi)\right) &= \frac{mn}{N}\left(\frac{\bar{P}(y)(1-\bar{P}(y))}{m} + \frac{\bar{P}(1-\bar{P}(y))}{n}\right) = \bar{P}(y)(1-\bar{P}(y)) \\
&= F_0(y)(1-F_0(y)) + \hat{R}_{m,n}(1-\hat{R}_{m,n}-2F_0(y)) \\
&= F_0(y)(1-F_0(y)) + o_p(1)
\end{aligned}$$

since $\hat{R}_{m,n} = o_p(1)$. We claim the asymptotic normality of

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

To do this, we use the Cramér-Wold device (Theorem 11.2.3 of Lehmann and Romano (2005)). Then, for any any a and b, we must verify the limiting distribution of

$$K(m)\sum_{i=1}^N (aX_iW_i + bX_iW_i')) = \sum_{i=1}^N C_{m,n,i}X_i \tag{23}$$

where

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

Condition on W_i and W_i', then (23) is a conditionally independent sum of linear combination of independent variables:

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

By the arguments in Example 15.2.5 of Lehmann and Romano (2005), we conclude that

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

and so

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

therefore

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

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

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

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

Lemma 2. *Under conditions A.1 and A.2, let $\bar{Z}_1, \bar{Z}_2 \ldots,$ be i.i.d. from the mixture distribution $\bar{P} = p\tilde{F}_1 + (1-p)F_0$, and denote $\bar{Z} = (\bar{Z}_1, \ldots, \bar{Z}_N)$. Let π and π' be independent of $\bar{Z}$. Then*

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

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

For a given π,

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

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

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

or it is of the form

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

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

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

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

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

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

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

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

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

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

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

and therefore the unconditional variance is bounded above by

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

and therefore convergence is quadratic mean follows. □

Theorem 4. *Assume the premises of Theorem 3. Then the permutation distribution (6) based on $K_{m,n,\hat{\delta}}$ is such that*

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

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

Proof. Lemma 1 and 2 imply that $\left(K_{m,n,\hat{\delta}}(\bar{Z}_\pi), K_{m,n,\hat{\delta}}(\bar{Z}_{\pi'})\right)$ are asymptotically independent. By the regular the continuous mapping theorem,

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

converges in distribution to the (J_0, J_0') process with independent, identically distributed marginals as described in Theorem 1. Then by Hoeffding's Condition (Theorem 5.1 of Chung and Romano (2013)),

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

where $\hat{R}^{K(\hat{\delta})}_{m,n}$ is the permutation distribution (7) based on $K_{m,n,\hat{\delta}}$ as desired. □

Appendix C: Proof of the Main Result

DESCRIPTION: process. In particular, it is shown that the test statistic follows an asymptotic law that is the supremum of the standard Brownian motion.

NOTATION: In what follows, it should be understood that BM refers to a standard Brownian motion process. In addition, π and π' will denote two independent random permutations of $\{1, \ldots, N\}$, and π_0 will denote the permutation that reorders observations in $\bar{Z}$, as described in Appendix A. In order to emphasize the data that are being used in the computation of the two-sample uniform empirical processes, we will write $\upsilon_{m,n}(t, \hat{\delta}; Z_\pi))$ or $\upsilon_{m,n}(t, \hat{\delta}; \bar{Z}_\pi))$, meaning that $\upsilon_{m,n}(t, \hat{\delta})$ was calculated using sample $(Z_{\pi(1)}, \ldots, Z_{\pi(N)})$ or $(\bar{Z}_{\pi(1)}, \ldots, \bar{Z}_{\pi(N)})$, respectively. Analogously, $\upsilon_{m,n}(t, \hat{\delta}; \bar{Z}_{\pi'}))$ is defined with π replaced by π'.

PRELIMINARIES: The general proof strategy will be based on the contiguity and coupling construction results in section 5 of Chung and Romano (2013) and Appendix A in this paper. The key idea is that the permutation distribution based on Z should behave approximately like the behavior of 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 $\bar{P}$. In order to establish this result, we will need the following two lemmas.

Lemma 3. *Under conditions A.1 and A.2, let $\bar{Z}_1, \bar{Z}_2 \ldots,$ be i.i.d. from the mixture distribution $\bar{P} = p\tilde{F}_1 + (1-p)F_0$, and denote $\bar{Z} = (\bar{Z}_1, \ldots, \bar{Z}_N)$. Let π and π' be independent of $\bar{Z}$. Then*

$$\left(\upsilon_{m,n}(t, \hat{\delta}; \bar{Z}_\pi)), \upsilon_{m,n}(t, \hat{\delta}; \bar{Z}_{\pi'}))\right)$$

converges weakly to (BM, BM') with BM and BM' two independent standard Brownian motion processes with common CDF.

Proof. Joint Gaussianity in follows from the discussion in Condition E of Romano (1989). More specifically, the differentiability condition needed in order to verify Condition E holds for the present case, since testing the null hypothesis (3) is essentially a two-sample test of homogeneity (see example 4 of Romano (1989)). Having shown the limits are Gaussian, zero-covariance renders independence. Then, it needs to be shown that

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

Notice that

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

by exploiting the linearity of the map ϕ_g. Therefore

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

It follows from the arguments in the proof of Theorem 4 that

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

Use linearity of the map ϕ_g once again,

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

Therefore, conditionally on Ws,

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

because π, π' are independent of Z, and mutually independent from each other. Once again, Slutsky theorem with the change of variable described in Remark 1 implies

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

□

Lemma 4. *Under conditions A.1 and A.2, let $\bar{Z}_1,\bar{Z}_2\ldots,$ be i.i.d. from the mixture distribution $\bar{P}=p\tilde{F}_1+(1-p)F_0$, and denote $\bar{Z}=(\bar{Z}_1,\ldots,\bar{Z}_N)$. Let π and π' be independent of $\bar{Z}$. Then*

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

Proof. Everything stated below is implicitly conditioned on π_0, but we omit it to ease notation. Fix π and use the asymptotic representation in proof of Theorem 5

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

We need to guarantee that the remainder, defined in eq (21) in the proof of Theorem 3, is still $o_p(1)$ under Z_π. We will use the contiguity result of Chung and Romano (2013); let $V_1,V_2,\ldots,$ be iid from the mixture distribution $\bar{P}=pF_1+(1-p)F_0$, and observe the remainder satisfies

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

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

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

as desired. Furthermore, by the arguments of Proposition 2 and Slutsky theorem with the change of variable described in Remark 1,

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

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

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

then

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

as desired.

□

Theorem 6. *Assume* $Y_1(0), \ldots, Y_n(0)$ *are i.i.d. according to a probability distribution* F_0*, and independently* $Y_1(1), \ldots, Y_m(1)$ *are i.i.d.* F_1*. Consider testing the hypothesis (3) for some* δ *based on the test statistic (15). Under conditions A.1-A.2, the permutation distribution (6) based on the Khmaladze transformed statistic* $\tilde{K}_{m,n,\hat{\delta}}$ *is such that*

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

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

Proof. Lemma 3 imply that $\left(\tilde{K}_{m,n,\hat{\delta}}(\bar{Z}_\pi), \tilde{K}_{m,n,\hat{\delta}}(\bar{Z}_{\pi'})\right)$ are asymptotically independent. Regular continuous mapping theorem implies that

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

converges in distribution to the (J_2, J_2') process with independent, identically distributed marginals as described in Theorem 5. By Lemmas 3 and 4 and Hoeffding's Condition (Theorem 5.1 of Chung and Romano (2013)), we have

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

where $\hat{R}_{m,n}^{\tilde{K}(\hat{\delta})}$ is the permutation distribution (6) based on $\tilde{K}_{m,n,\hat{\delta}}$ as desired.

□