Multivariate and Multiple Permutation Tests

EunYi Chung*

Department of Economics
University of Illinois at Urbana-Champaign

Joseph P. Romano†

Departments of Statistics and Economics
Stanford University

January 2016

Abstract

In this article, we consider the use of permutation tests for comparing multivariate parameters from two populations. First, the underlying properties of permutation tests when comparing parameter vectors from two distributions P and Q are developed. Although an exact level α test can be constructed by a permutation test when the fundamental assumption of identical underlying distributions holds, permutation tests have often been misused. Indeed, permutation tests have frequently been applied in cases where the underlying distributions need not be identical under the null hypothesis. In such cases, permutation tests fail to control the Type 1 error, even asymptotically. However, we provide valid procedures in the sense that even when the assumption of identical distributions fails, one can establish the asymptotic validity of permutation tests in general while retaining the exactness property when all the observations are i.i.d. In the multivariate testing problem for testing the global null hypothesis of equality of parameter vectors, a modified Hotelling's T^2 -statistic as well as tests based on the maximum of studentized absolute differences are considered. In the latter case, a bootstrap prepivoting test statistic is constructed, which leads to a bootstrapping after permuting algorithm. Then, these tests are applied as a basis for testing

*Research has been supported by B.F. Haley and E.S. Shaw Fellowship for Economics.

†Research has been supported by NSF Grant DMS-0707085.

multiple hypotheses simultaneously by invoking the closure method to control the Familywise Error Rate. Lastly, Monte Carlo simulation studies and an empirical example are presented.

KEY WORDS: Bootstrap; Familywise Error Rate; Multiple Tests; Permutation Test; Prepivoting

1 Introduction

In many empirical applications in economics and pretty much any scientific study, testing of several null hypotheses simultaneously is frequently performed. One such example includes evaluating a treatment or a program that has several outcomes and assessing which outcomes yield significant results. We first consider tests for the multivariate problems, which will serve as a foundation for the permutation tests in multiple testing.

Suppose $X_1, \dots, X_m$ are i.i.d. according to a probability distribution P , and independently, $Y_1, \dots, Y_n$ are i.i.d. Q . The space where P and Q lie is quite general, but we are especially interested in the cases where the observations are multivariate (or vectors). Let $N = m + n$, and by putting all the observations together, write the matrix

$$Z = (Z_1, \dots, Z_N) = (X_1, \dots, X_m, Y_1, \dots, Y_n) .$$

A fundamental tool for learning about differences between population distributions P and Q is based on sample comparison. As simple of an idea as this is, statistical theory is needed to assess whether sample differences are real. As such a tool, it is well-known that permutation tests can be constructed so as to be exact level α , as long as the fundamental assumption of identical underlying distributions holds. Under the assumption of identical distributions, any permuted sample has the same joint distribution as the original sample. Thus, the permutation distribution, which is the empirical c.d.f. of a given test statistic recomputed over all permutations of the data, serves as a valid null distribution, and one can achieve exact control of the Type 1 error even in finite samples. However, researchers are oftentimes interested in testing a particular parameter of the underlying distributions, such as testing equality of means or medians (as opposed to testing equality of distributions). Under such null hypotheses, the underlying distributions need not be the same (as equality of distributions is a stronger assumption). As a result, the logic upon which a permutation test is constructed is no longer valid and thus the permutation test fails to control the Type 1 error, even asymptotically¹. This paper

¹See Murphy 1967 and Romano 1990.

seeks to understand the underlying properties of permutation tests in multivariate cases and to provide appropriate procedures which possess valid error control. Based on such foundations, we further consider more complex settings where many tests need to be performed simultaneously. We apply multivariate permutation tests as a basis for testing multiple hypotheses by invoking the closure method to control the Familywise Error Rate. Lastly, Monte Carlo simulation studies and an empirical example are presented.

To understand the basic setting for permutation tests, assume $P = Q$. Then, for any permutation $(\pi(1), \dots, \pi(N))$ of $\{1, \dots, N\}$, the joint distribution of $(Z_{\pi(1)}, \dots, Z_{\pi(N)})$ is the same as that of the original data $(Z_1, \dots, Z_N)$. Thus, if $P = Q$ holds under the null hypothesis of interest, then an exact level α test can be constructed by a permutation test. (See Lehmann and Romano (2005) for more details.)

However, if the null hypothesis of interest does not imply $P = Q$, the rejection probability need not be α even asymptotically. Unfortunately, permutation tests are widely used in many applications of academic research even when this fundamental assumption of identical distributions need not hold, as examined in great detail in the case of univariate problems in Chung and Romano (2013, 2015). To be concrete, consider testing equality of means specified by

$$H_0 : \mu(P) = \mu(Q) . \tag{1}$$

In this case, $P = Q$ need not be implied. When using a permutation test based on the unstudentized difference of sample means, the limiting probability of the Type 1 error need not be α , and can even be near $1/2$ in an one-sided test or near 1 in a two-sided test. While control of Type 1 error is of paramount importance, the implications regarding both Type 2 and Type 3 errors should also be emphasized as well. For if one negates the lack of Type 1 error control by declaring one is really testing $P = Q$, then such a permutation test would have no power against alternatives with respect to $P \neq Q$ but $\mu(P) = \mu(Q)$, and so one should not use such a test statistic for that purpose. On the other hand, if the purpose is indeed to test H_0 defined in (??), then lack of Type 1 error control inevidently results in lack of Type 3 error, or directional error, control². For example, if $\bar{Y}_n > \bar{X}_m$, rejection of H_0 or even the stricter null hypothesis that $P = Q$ is invariably accompanied by an inference that $\mu(Q) > \mu(P)$. But, having established that the probability of a Type 1 error is, say $\gamma \gg \alpha$ under P and Q satisfying $\mu(P) = \mu(Q)$, it follows by continuity that there exist P' and Q with $\mu(Q) < \mu(P')$ but the chance that the permutation test rejects H_0 with the added inference that $\mu(Q) > \mu(P)$ has a Type 3 error with probability near γ . Clearly, rejection of the null in favor of a positive

²A Type 3 or directional error occurs if one declares $\mu(Q) > \mu(P)$ when in fact $\mu(P) > \mu(Q)$.

difference when the actual effect is negative is worrisome.

In addressing this problem, Neuhaus (1993) proposed a permutation test based on a studentized statistic in the context of a censoring model; by appropriately studentizing the test statistic, the permutation test can achieve asymptotic validity even when the underlying distributions are not identical. In other words, even if the underlying distributions are not the same under the null hypothesis, the asymptotic rejection probability of the test is the nominal level α . In addition, the test retains the exact control of the rejection probability α if the underlying distributions are the same. Janssen (1997, 1999) also applied this insightful idea to testing equality of univariate means and statistical functionals (when the population distributions can be different) and showed that by proper studentization of the test statistic, i.e., dividing the sample mean difference by an appropriate standard error, the permutation test yields asymptotically valid inferences even if the underlying distributions are not the same. The same idea has been extended to other applications by Janssen and Pauls (2003, 2005), Janssen (2005), Neubert and Brunner (2007), Pauly (2010), Konietzke and Pauly (2012), Omelka and Pauly (2012), Chung and Romano (2013, 2015), and Pauly et al. (2015). Chung and Romano (2013) provide very general asymptotic arguments to handle general univariate testing problems. In all of these cases, the main idea is that if a test statistic is chosen (or modified) to be asymptotically distribution-free, then the so-called permutation distribution asymptotically approximates the unconditional true sampling distribution of the test statistic. Indeed, the asymptotic arguments in this paper rely on the study of the permutation distribution, which is just the empirical distribution function of some test statistic recomputed over all permutations of the data. More formally, for a given (possibly multivariate) test statistic $T_{m,n}$, define the multivariate permutation distribution as

$$\hat{R}_{m,n}^T(t) = \frac{1}{N!} \sum_{\pi \in \mathbf{G}_N} I\{T_{m,n}(Z_{\pi(1)}, \dots, Z_{\pi(N)}) \leq t\}, \quad (2)$$

where $\mathbf{G}_N$ denotes the $N!$ permutations of $\{1, 2, \dots, N\}$ and $t = (t_1, \dots, t_d)' \in \mathbf{R}^d$. (Note that d need not be the same as the dimension of the data, e.g., d could be 1.)

This paper generalizes this phenomenon to multivariate and multiple testing problems, where, unlike the univariate case, test statistics need not be asymptotically normal and so a simple studentization is not available. We provide a framework under which permutation tests can achieve asymptotic control of the Type 1 error in general. Of course, other resampling methods such as the bootstrap (Efron, 1979) or subsampling (Politis, Romano, and Wolf, 1999) are valid alternatives to obtain the asymptotic result. However, permutation tests have an additional desired property that other resampling methods do not have; namely, the test is exact level α in finite samples in the case

of homogeneous populations. We demonstrate that by an appropriate choice of test statistic, permutation tests obtain both the asymptotic validity in general and the exactness property when $P = Q$. In addition, the key element of our results shows that the permutation distribution behaves like the unconditional true sampling distribution when *all* the observations are i.i.d. from the mixture distribution $\bar{P} = pP + (1 - p)Q$, where p is the limit of m/N . Indeed, this may be distinct from the true unconditional distribution when m observations are from P and n from Q . But, this leads to the observation that one way for the permutation distribution and the true unconditional sampling distribution to be asymptotically the same is to choose a test statistic which is asymptotically pivotal (generalizing the idea of studentizing).

The plan of the paper is as follows. In Section ??, the multivariate problem is introduced and it illustrates how the permutation test can fail to control the rejection probability even asymptotically. When the underlying distributions P and Q need not be identical under the null hypothesis, the permutation distribution behaves differently from the unconditional true sampling distribution. In Subsection ??, we consider the multivariate nonparametric Behrens-Fisher problem where we are interested in testing equality of means for multivariate populations (with possibly different covariance matrices). We show that the permutation test based on an asymptotically pivotal modified Hotelling's T^2 statistic for testing equality of means results in the asymptotic rejection probability of α in general while retaining the exact control of the test level when $P = Q$. For testing equality of means, one might instead be interested in using the maximum value of the mean absolute differences over all the components as a test statistic as non-zero maximum value means divergence from the null hypothesis. In this case, the test statistic is not asymptotically pivotal and one can deduce that the permutation test based on the maximum value will fail to control the rejection probability even asymptotically. To address this issue, in Subsection ?? we apply the “prepivoting” idea of Beran (1988a, 1988b) as an alternative way of rendering a test statistic asymptotically pivotal. A prepivoted statistic is a statistic transformed by a bootstrap estimate of its true sampling distribution and essentially converts a test statistic into a bootstrap p -value (or more precisely 1 minus a bootstrap p -value). By transforming the test statistic by its bootstrap c.d.f., the prepivoted test statistic converges in distribution to a uniform distribution. By using such an asymptotically pivotal statistic, the permutation test based on the prepivoted statistic achieves our desired results. Section ?? provides a generalization of Subsection ?? whereby the parameter of interest is not just a vector of means but a general vector parameter that depends on the underlying populations. Under weak assumptions that the parameters are asymptotically linear and that consistent covariance estimators are available, we provide a general framework whereby the permutation test can control the rejection probability while still retaining exact control

of the level in the case $P = Q$.

In Section ??, a further extension to the multiple testing problem is considered. By applying the closure method in multivariate cases, the familywise error rate (FWER), which is the probability of one or more false rejections, can be controlled at level α (in finite samples or asymptotically). Monte Carlo simulations studies based on the modified Hotelling's T^2 statistic are performed in Section ?. Lastly, an empirical study based on Charness and Gneezy (2009) is presented in Section ?. Charness and Gneezy (2009) study the effects of exercise in terms of seven biometric measures. The main tool these authors use to assess differences between groups is the classical Wilcoxon test, which implicitly is valid only when the underlying distributions are the same under the null hypothesis. In addition, they do not consider multiple testing, resulting in an inflated Type 1 error rate when testing several hypotheses simultaneously. We illustrate the performance of the permutation test based both on the modified Hotelling's T^2 statistic and on the prepivoted statistic while controlling the familywise error rate. All proofs are reserved for the appendix.

2 Multivariate Permutation Test

Consider the behavior of multivariate two-sample permutation tests when the assumption of identical distributions need not hold. Suppose $X_1, \dots, X_m$ are d -dimensional i.i.d. P , where $X_i = (X_{i,1}, \dots, X_{i,d})'$ for $i = 1, \dots, m$ with mean vector $(\mu_1(P), \dots, \mu_d(P))'$ and covariance matrix Σ_P , and independently, $Y_1, \dots, Y_n$ are d -dimensional i.i.d. Q , where $Y_j = (Y_{j,1}, \dots, Y_{j,d})'$ for $j = 1, \dots, n$ with mean vector $(\mu_1(Q), \dots, \mu_d(Q))'$ and covariance matrix Σ_Q . Let $N = m + n$, and write

$$Z = (Z_1, \dots, Z_N) = (X_1, \dots, X_m, Y_1, \dots, Y_n) .$$

Throughout this paper, assume that the dimension of the observations d is fixed and smaller than the numbers of observations m and n . In this section, permutation tests are studied when comparing means of multidimensional observations from two populations³ (though generalized to general parameters in Section ?). Specifically, consider testing the null hypothesis

$$H_0 : \mu_k(P) = \mu_k(Q) \text{ for all } k = 1, \dots, d , \quad (3)$$

³The results can be readily generalized to multiple samples with more than two populations, see Chung and Romano (2013) for univariate cases.

versus the alternative hypothesis

$$H_1 : \mu_k(P) \neq \mu_k(Q) \text{ for some } k = 1, \dots, d .$$

When $P = Q$, all the observations are i.i.d, and thus, an exact level α test can be constructed using a permutation test. However, if $P \neq Q$, the test may fail to control the probability of Type 1 error, even asymptotically. Our goal is to construct a procedure that allows for permutation tests to obtain asymptotic validity in general while retaining the exactness property in finite samples in the case of $P = Q$.

For now, attention focuses on the joint testing problem (??), but we will also treat the multiple testing problems based on the tests developed here in Section ?? . Consider a permutation test based on the difference of the sample mean vectors

$$T_{m,n} = (T_{m,n,1}, \dots, T_{m,n,d}) = m^{1/2} [\bar{X}_m - \bar{Y}_n] = m^{-1/2} \left[\sum_{i=1}^m X_i - \frac{m}{n} \sum_{j=1}^n Y_j \right] , \quad (4)$$

where $\bar{X}_m = (\bar{X}_{m,1}, \dots, \bar{X}_{m,d})'$ and $\bar{Y}_n = (\bar{Y}_{n,1}, \dots, \bar{Y}_{n,d})'$ with $\bar{X}_{m,k} = 1/m \sum_{i=1}^m X_{i,k}$ and $\bar{Y}_{n,k} = 1/n \sum_{j=1}^n Y_{j,k}$ for $k = 1, \dots, d$. First, we argue that the permutation distribution behaves asymptotically like the limiting unconditional sampling distribution of the statistic sequence when sampling i.i.d. observations from $\bar{P} = pP + (1-p)Q$, where $p = \lim \frac{m}{m+n}$. Specifically, in the case of comparing the means based on the multivariate statistic (??), the permutation distribution converges in probability to the d -variate normal distribution with mean 0 and variance

$$\bar{\Sigma} = \frac{p}{1-p} \Sigma_P + \Sigma_Q . \quad (5)$$

Note that this holds even if H_0 is not true. The theorem below states this formally.

Theorem 2.1. *Consider the above setup. Assume $E(X_k) = E(Y_k)$, for $k = 1, \dots, d$ with*

$$0 < (X_{i,k}) < \infty \text{ and } 0 < (Y_{j,k}) < \infty . \quad (6)$$

Let $m \rightarrow \infty$, $n \rightarrow \infty$. Assume $\bar{\Sigma}$ given in (??) is positive-definite. Consider the permutation distribution $\hat{R}_{m,n}^T$ defined in (??) based on the vector of sample mean difference $T_{m,n}$ given in (??). Then,

$$\sup_{t \in \mathbf{R}^d} \left| \hat{R}_{m,n}^T(t) - G(t) \right| \xrightarrow{P} 0^4, \quad (7)$$

⁴A sequence of random distributions $\hat{G}_n$ converges weakly in probability to a (nonrandom) distri-

where G denotes the d -variate normal distribution with mean 0 and variance $\bar{\Sigma}$ defined in (??).

Remark 2.1. Under H_0 , the true unconditional sampling distribution of $T_{m,n}$ is asymptotically normal with mean 0 and covariance matrix

$$\Sigma_P + \frac{p}{1-p} \Sigma_Q, \quad (8)$$

which does not equal (??) in general unless $\Sigma_P = \Sigma_Q$ or $m/n \rightarrow 1$ holds.

Remark 2.2. The result holds even when $p = 0$, i.e., $\frac{m}{N} \rightarrow 0$. Observe that in this case, the permutation distribution has covariance matrix $\bar{\Sigma} = \Sigma_Q$ while the unconditional sampling distribution has covariance matrix Σ_P . By interchanging the roles of the X s and the Y s, we can get a similar result for $p = 1$.

Remark 2.3. The scaling by $m^{1/2}$ in (??) in no way affects the inference based on permutation tests. In other words, the same inference would result if m were replaced by n or N . (However, the conditions for p changes: $p \in (0, 1]$ when the scaling factor is $n^{1/2}$, and $p \in (0, 1)$ if $N^{1/2}$ is used instead.) It only serves an asymptotic purpose in order to get a nondegenerate limiting distribution.

From Theorem ?? together with Remark ??, one can deduce that any continuous function (for instance, either the usual Euclidean norm or the maximum value over all components) of the multivariate permutation distribution based on (??) is not asymptotically distribution-free. (This requires a continuous mapping theorem for randomization distributions; see Lemma ??.) Thus, the corresponding permutation tests fail to control the Type 1 error even asymptotically. In general, the permutation distribution does not approximate the true unconditional distribution. However, for test statistics that are asymptotically distribution-free it is possible to control the asymptotic rejection probability even when the underlying distributions need not be identical under the null hypothesis while also achieving finite sample exactness when all the observations are i.i.d. The following subsections provide different methods to achieve the desired properties for different test statistics of interest.

bution G if $\rho(\hat{G}_n, G) \rightarrow 0$ in probability, where $\rho(\cdot, \cdot)$ is any metric which metrics weakly converges or equivalently convergence in distribution, e.g. the Levy metric (see Section 11.2 of Lehmann and Romano (2005)). It is equivalent to saying that for any continuity point t of G , $\hat{G}_n(t) \rightarrow G(t)$ in probability. Unless otherwise noted, we use “in probability” and “weakly in probability” interchangeably for permutation distributions.

2.1 Modified Hotelling's T^2 Statistic

The key element that will lead us to asymptotic validity of the permutation tests is using a test statistic that is asymptotically pivotal; that is, the limiting distribution of the test statistic does not depend on the underlying distributions. In this subsection, we consider a modified Hotelling's T^2 statistic defined in (??) below. We will show that this asymptotically pivotal statistic achieves the asymptotic rejection probability of α , while attaining the exact control when the underlying distributions are the same. First, the behavior of the multivariate (transformed) difference of means is studied.

Theorem 2.2. *Assume the setup and conditions of Theorem ?? . For $N = m + n$, further assume $p_m = m/N \rightarrow p \in [0, 1)$ with*

$$p_m - p = O(m^{-1/2}) . \quad (9)$$

Define the test statistic

$$S_{m,n} = \hat{\Sigma}^{-1/2} T_{m,n} = m^{-1/2} \hat{\Sigma}^{-1/2} \left[\sum_{i=1}^m X_i - \frac{m}{n} \sum_{j=1}^n Y_j \right] , \quad (10)$$

where

$$\hat{\Sigma} = \hat{\Sigma}_P + \frac{m}{n} \hat{\Sigma}_Q \quad (11)$$

and the matrices $\hat{\Sigma}_P$ and $\hat{\Sigma}_Q$ are consistent estimators of Σ_P and Σ_Q , defined having (r, s) component given by

$$\hat{\Sigma}_P^{r,s} = \frac{1}{m-1} \sum_{i=1}^m (X_{i,r} - \bar{X}_{m,r})(X_{i,s} - \bar{X}_{m,s}) \quad (12)$$

and

$$\hat{\Sigma}_Q^{r,s} = \frac{1}{n-1} \sum_{j=1}^n (Y_{j,r} - \bar{Y}_{n,r})(Y_{j,s} - \bar{Y}_{n,s}), \quad (13)$$

respectively. Then, the permutation distribution $\hat{R}_{m,n}^S$ of $S_{m,n}$ defined in (??) with T replaced by S satisfies

$$\sup_{t \in \mathbf{R}^d} \left| \hat{R}_{m,n}^S(t) - \Phi_d(t) \right| \xrightarrow{P} 0 ,$$

where Φ_d denotes the standard d -variate normal distribution with mean 0 and variance $\mathbf{I}_{d \times d}$.

Remark 2.4. Although, in principle, there are $N!$ permutations available to construct the permutation distribution, the exactness property can still be achieved with identical underlying distributions even if we only consider a finite number of randomly sampled permutations B ($< N!$) such that for a given level α , $(B + 1)\alpha$ is an integer (one extra permutation is the original sample).

Next, consider the modified Hotelling's T^2 statistic defined by

$$\tilde{S}_{m,n} = \|S_{m,n}\|^2 = T'_{m,n} \hat{\Sigma}^{-1} T_{m,n} , \quad (14)$$

where $S_{m,n}$ is d -dimensional as defined in (??), $\hat{\Sigma}$ is defined in (??), and $\|\cdot\|$ denotes the usual Euclidean norm. On the other hand, the classical Hotelling's T^2 statistic is defined by

$$\tilde{T}^2 = T'_{m,n} \tilde{\Sigma}^{-1} T_{m,n} ,$$

where

$$\tilde{\Sigma} = \left(\frac{m+n}{n} \right) \frac{(m-1)\hat{\Sigma}_P + (n-1)\hat{\Sigma}_Q}{m+n-2} .$$

Of course, $\tilde{T}^2$ is derived under normality of the underlying distributions and equality of covariance matrices. As such, a pooled estimator of covariance is used. In such a case, the limiting distribution of $\tilde{T}^2$ is not distribution free and the approach fails. In the theorem below, we do not assume normality nor equal covariance matrices.

Theorem 2.3. *Assume the setup and conditions of Theorem ?? and Theorem ?. Consider the modified Hotelling's T^2 statistic defined in (??). Then, for $t \in \mathbf{R}$, the permutation distribution $\hat{R}_{m,n}^{\tilde{S}}(t)$ of $\tilde{S}_{m,n}$ defined in (??) with T replaced by $\tilde{S}$ satisfies*

$$\left| \hat{R}_{m,n}^{\tilde{S}}(t) - \chi_d^2(t) \right| \xrightarrow{P} 0 ,$$

where χ_k^2 denotes the Chi-squared distribution with k degrees of freedom.

Remark 2.5. Since the test statistic (??) is based on a Euclidean norm, the resulting test is designed to test against multi-sided alternatives. Thus, for a given nominal level α , a test is rejected if the sample statistic based on $\tilde{S}_{m,n}$ lies in the upper $100\alpha\%$ of the permutation distribution.

2.2 Maximum Statistic

In this subsection, we consider the maximum of the sample mean absolute differences over all the components as an alternative test statistic to test the null hypothesis (??). This test statistic is a natural one in testing (3) as non-zero maximum value supports the alternative of having at least one unequal mean among all the components.

By adopting the “prepivoting” method proposed by Beran (1988a, 1988b), asymptotically pivotality will be achieved.

Theorem 2.4. *Assume the same setup and conditions of Theorem ?? . For $N = m + n$, further assume $p_m = m/N \rightarrow p \in [0, 1)$ with (??). Consider the permutation distribution $\hat{R}_{m,n}^M(\cdot)$ based on the test statistic*

$$M_{m,n} = \max_{k=1,\dots,d} (|T_{m,n,k}|) , \quad (15)$$

where $T_{m,n,k}$ denotes the k th component of $T_{m,n}$. Then, for $t \in \mathbf{R}$, the permutation distribution $\hat{R}_{m,n}^M$ of $M_{m,n}$ defined in (??) with T replaced by M satisfies

$$\left| \hat{R}_{m,n}^M(t) - F(t) \right| \xrightarrow{P} 0 ,$$

where $F(\cdot)$ is the c.d.f. of $\max_{k=1,\dots,d} (|G_1|, \dots, |G_d|)$, and $(G_1, \dots, G_d)$ is the multivariate normal with c.d.f. G given in (??).

Remark 2.6. This result is still true even if $\bar{\Sigma}$ is singular as we only need non-zero marginal variances as assumed in (??).

The maximum statistic (??) is not asymptotically distribution-free, because its limiting distribution depends on the underlying covariance matrices through $\bar{\Sigma}$. The idea is to modify the test statistic so that the resulting statistic becomes asymptotically pivotal. Before applying the ‘prepivoting’ method, we first consider dividing the test statistic by its marginal standard error. By studentizing each difference, the differences are placed on the same scale; also see Remark ?? . Although marginal studentization does result in the asymptotic marginal distributions of the studentized differences to be distribution free, the entire joint distribution of the studentized differences depends on the underlying covariance matrices (as well as $\lim \frac{m}{n}$).

Theorem 2.5. Assume the setup and conditions of Theorem ?? and Theorem ?. Consider the following statistic

$$\tilde{M}_{m,n} = \max_{1 \leq k \leq d} \left(\text{diag}(\hat{\Sigma})^{-1/2} |T_{m,n,k}| \right) = \sqrt{m} \max_{1 \leq k \leq d} \left(\frac{|\bar{X}_{m,k} - \bar{Y}_{n,k}|}{\hat{S}_{m,n,k}} \right), \quad (16)$$

where

$$\hat{\Sigma} = \hat{\Sigma}_P + \frac{m}{n} \hat{\Sigma}_Q$$

and matrices $\hat{\Sigma}_P$ and $\hat{\Sigma}_Q$ are consistent estimators of Σ_P and Σ_Q with (r, s) component defined in (??) and (??), respectively, and $\hat{S}_{m,n,k}$ denotes the k th element of the diagonal matrix $(\text{diag}(\hat{\Sigma}))^{1/2}$. Then, the permutation distribution $R_{m,n}^{\tilde{M}}$ of $\tilde{M}_{m,n}$ defined in (??) with T replaced by $\tilde{M}$ satisfies

$$\left| \hat{R}_{m,n}^{\tilde{M}}(t) - H(t) \right| \xrightarrow{P} 0, \quad (17)$$

where $H(\cdot)$ is the c.d.f. of $\max(|H_1|, \dots, |H_d|)$, where $(H_1, \dots, H_d)$ is the d -variate normal distribution with mean 0 and covariance matrix $\tilde{\Sigma}$ given by

$$\tilde{\Sigma} = (\tilde{\sigma}_{ij}) = \left(\frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}} \right) = (\text{diag}(\bar{\Sigma}))^{-1/2} \bar{\Sigma} (\text{diag}(\bar{\Sigma}))^{-1/2}. \quad (18)$$

The permutation test based on the maximum of the mean absolute difference $M_{m,n}$, or even after being divided by its marginal standard error $\tilde{M}_{m,n}$, fails to control the Type 1 error even asymptotically as neither $M_{m,n}$ nor $\tilde{M}_{m,n}$ is asymptotically pivotal. Here, we provide an alternative method based on “prepivotting” which transforms the test statistic so that it is asymptotically uniformly distributed on $[0, 1]$ and hence asymptotically pivotal. In fact, the transformed or ‘prepivotted’ test statistic converts the original statistic to one minus a bootstrap p -value. Thus, a permutation test based on the transformed or prepivotted test statistic produces results that are both exact and asymptotically robust for heterogeneous populations.

Before showing how the prepivotting method works, let $J_{m,n}(P, Q)$ be the distribution of $\tilde{M}_{m,n}$ under P and Q , and let $J_{m,n}(\cdot, P, Q)$ be its corresponding c.d.f. defined by

$$J_{m,n}(x, P, Q) = P_{P,Q} \left\{ \max_{1 \leq k \leq d} \sqrt{m} \frac{|\bar{X}_{m,k} - \mu_k(P) - (\bar{Y}_{n,k} - \mu_k(Q))|}{\hat{S}_{m,n,k}} \leq x \right\}, \quad (19)$$

where $\hat{S}_{m,n,k}$ is given in (??). The prepivotted statistic is then defined by $J_{m,n}(\tilde{M}_{m,n}, \hat{P}_m, \hat{Q}_n)$,

where $\hat{P}_m$ and $\hat{Q}_n$ are the empirical distributions of P and Q , respectively. In other words, the prepivoted statistic is a bootstrap estimate of $J_{m,n}(P, Q)$ evaluated at $\tilde{M}_{m,n}$. Moreover, $1 - J_{m,n}(\tilde{M}_{m,n}, \hat{P}_m, \hat{Q}_n)$ can be viewed as a bootstrap p -value for testing the joint null hypothesis of equality of means. The main idea here is that by transforming a given statistic by its bootstrap c.d.f., the prepivoted test statistic now becomes asymptotically pivotal. The prepivoting method involves bootstrapping for each permuted sample. An algorithm for the permutation test based on a prepivoted statistic of $\tilde{M}_{m,n}$ is given by the following.

Algorithm 2.1. (*Prepivoting Method based on $\tilde{M}_{m,n}$*)

1. For each permutation $\pi_s, s = 1, \dots, N!$, calculate the test statistics value $\tilde{M}_s(Z_{\pi_s})$ given by $\tilde{M}_s = \max_{1 \leq k \leq d} \left(\frac{|\frac{1}{m} \sum_{i=1}^m Z_{\pi_s(i),k} - \frac{1}{n} \sum_{j=m+1}^N Z_{\pi_s(j),k}|}{\hat{S}_{m,n,k}(Z_{\pi_s})} \right)$.
2. Given the permuted sample Z_{π_s} based on π_s , resample $x_b^* = (x_{b,1}^*, \dots, x_{b,m}^*)$ from the first m observations $Z_{\pi_s(1)}, \dots, Z_{\pi_s(m)}$ with replacement and $y_b^* = (y_{b,1}^*, \dots, y_{b,n}^*)$ from the last n observations $Z_{\pi_s(m+1)}, \dots, Z_{\pi_s(N)}$ with replacement, and recalculate the test statistic $\tilde{M}_{s,b}^*$ based on x_b^* and y_b^* .
3. Repeat step 2 B times, for $b = 1, \dots, B$.
4. Define the prepivoted test statistic $J_{m,n,s} = J_{m,n}(\tilde{M}_s, \hat{P}_m, \hat{Q}_n)$ to be the fraction of the values $\{\tilde{M}_{s,b}^* : 1 \leq b \leq B\}$ that are less than or equal to $\tilde{M}_s$. The empirical c.d.f. of the $\{J_{m,n,s} : 1 \leq s \leq N!\}$ approximates the permutation distribution of $J_{m,n}(\tilde{M}_{m,n}, \hat{P}_m, \hat{Q}_n)$.
5. In words, the permutation test rejects if $J_{m,n}(\tilde{M}_{m,n}, \hat{P}_m, \hat{Q}_n)$ exceeds the $1 - \alpha$ quantile of the permutation distribution.

The following theorem shows that the prepivoted test statistic achieves asymptotic validity.

Theorem 2.6. Assume the same setup and conditions of Theorem ?? and Theorem ??. Define the prepivoted test statistic

$$J_{m,n}(\tilde{M}_{m,n}, \hat{P}_m, \hat{Q}_n) = P_{\hat{P}_m, \hat{Q}_n} \left(\max_{1 \leq k \leq d} \sqrt{m} \frac{\left| \left(\bar{X}_{m,k} - \mu_k(\hat{P}_m) \right) - \left(\bar{Y}_{n,k} - \mu_k(\hat{Q}_n) \right) \right|}{\hat{S}'_{m,n,k}} \leq \tilde{M}_{m,n} \right). \quad (20)$$

Then, the permutation distribution $\hat{R}_{m,n}^J$ of $J_{m,n}(\tilde{M}_{m,n}, \hat{P}_m, \hat{Q}_n)$ defined in (??) with T replaced by J satisfies

$$\left| \hat{R}_{m,n}^J(t) - U(t) \right| \xrightarrow{P} 0, \quad (21)$$

where $U(\cdot)$ denotes the c.d.f. of the uniform distribution $U(0, 1)$.

Remark 2.7. Note that the prepivoting method still holds even if $J_{m,n}$ is defined to be the distribution of $M_{m,n}$ (before divided by its marginal standard error) and the limiting distribution of $J_{m,n}(M_{m,n}, \hat{P}_m, \hat{Q}_n)$ is uniform. However, using $\tilde{M}_{m,n}$ is advantageous because it is better “balanced” (by Beran (1988a)) in the sense that the limiting rejection probability due to the k th coordinate being the “largest” does not depend on k ; see Beran (1988a) and Romano and Wolf (2010).

3 Generalization: Testing Equality of Parameters

Consider now more general setting where the parameter of interest is not confined to be just a vector of means but a more general vector parameter that depends on the underlying distributions⁵ The inference problem consists of comparing multivariate parameters of two populations. Specifically, we are interested in testing the null hypothesis

$$H_0 : \theta_k(P) = \theta_k(Q), \quad \text{for all } k = 1, \dots, d, \quad (22)$$

versus the alternative hypothesis

$$H_1 : \theta_k(P) \neq \theta_k(Q), \quad \text{for some } k = 1, \dots, d,$$

where $\theta_k(\cdot)$ is a real-valued parameter, defined on some space of distributions $\mathcal{P}$. For example, we may be testing equality of mean vectors as before, or now median vectors. Alternatively, we may be testing equality of first and second moments, so that the form of θ_k may depend on k in (??). Just like before, if $P = Q$, then the permutation test can be constructed to have exact level α . On the contrary, if $P \neq Q$, the test in general fails to control the rejection probability at α even asymptotically. The objective here is to provide a general theory under weak assumptions whereby the permutation test obtains its asymptotic validity under weak conditions while maintaining the exact control of the rejection probability if $P = Q$ in this general setting.

⁵Testing functionals in univariate cases has been studied in the literature. Please see Janssen (1997, 1999, 2001, 2005), Janssen and Pauls (2005), and Konietzschke and Pauly (2012) among others.

Assume that available estimators are asymptotically linear. That is, under P , there exists an estimator $\hat{\theta}_m = (\hat{\theta}_{m,1}, \dots, \hat{\theta}_{m,d})'$, where, for $k = 1, \dots, d$, $\hat{\theta}_{m,k}(X_{1,k}, \dots, X_{m,k})$ satisfies

$$m^{1/2}[\hat{\theta}_{m,k} - \theta_k(P)] = \frac{1}{\sqrt{m}} \sum_{i=1}^m f_{P,k}(X_{i,k}) + o_P(1) . \quad (23)$$

Note that the influence function f_P , in (??) can depend on k . For example, one may compare means and variances simultaneously. Similarly, under Q , there exists an estimator $\hat{\theta}_n = (\hat{\theta}_{n,1}, \dots, \hat{\theta}_{n,d})'$, where, for $k = 1, \dots, d$, $\hat{\theta}_{n,k}(Y_{1,k}, \dots, Y_{n,k})$ satisfies

$$n^{1/2}[\hat{\theta}_{n,k} - \theta_k(Q)] = \frac{1}{\sqrt{n}} \sum_{j=1}^n f_{Q,k}(Y_{j,k}) + o_Q(1) . \quad (24)$$

Further assume that the expansion (??) is assumed to hold not only for i.i.d observations from P and Q , but also when i.i.d. observations are sampled from the mixture distribution $\bar{P} = pP + (1-p)Q$, where $m/N \rightarrow p$ as $\min(m, n) \rightarrow \infty$. Typically, $\hat{\theta}_{m,k}$ takes the form of an empirical estimator $\theta_k(\hat{P}_{m,k})$, where $\hat{P}_{m,k}$ is the empirical measure which assigns mass $\frac{1}{m}$ to each data point $X_{i,k}, i = 1, \dots, m$. Note that the expansions above do not require us to assume some form of differentiability of the functional $\theta_k(\cdot)$, such as compact differentiability. Although such a strong assumption is sufficient for the expansion (??), it is not necessary for our results; the assumption of asymptotic linearity of the estimators is all that is required to derive the asymptotic behavior of the permutation distribution. Based on such weak assumptions, we now extend the results in earlier subsections to this more general setting. As before, we first consider the behavior of the multivariate statistic of sample differences.

Theorem 3.1. *Assume the above setup. Let*

$$W_{m,n} = m^{1/2} \left[\hat{\theta}_m(X_1, \dots, X_m) - \hat{\theta}_n(Y_1, \dots, Y_n) \right] , \quad (25)$$

where the d -dimensional estimators $\hat{\theta}_m$ and $\hat{\theta}_n$ satisfy (??) and (??). Further assume, for $k = 1, \dots, d$, $E_P f_{P,k}(X_{i,k}) = E_Q f_{Q,k}(Y_{j,k}) = 0$ and

$$0 <_P (f_{P,k}(X_{i,k})) < \infty \quad \text{and} \quad 0 <_Q (f_{Q,k}(Y_{j,k})) < \infty \quad (26)$$

with $(f_{P,1}(X_{i,1}), \dots, f_{P,d}(X_{i,d}))$ and $(f_{Q,1}(Y_{j,1}), \dots, f_{Q,d}(Y_{j,d}))$ having its covariance matrix of Γ_P and Γ_Q , respectively. Let $m \rightarrow \infty$, $n \rightarrow \infty$, with $N = m + n$, $p_m = m/N$, and $p_m \rightarrow p \in [0, 1)$ with (??). Assume the estimator sequence also satisfies (??) with

P replaced by $\bar{P} = pP + (1 - p)Q$ with $E_{\bar{P}} f_{\bar{P},k}(\cdot) = 0$. Also, let

$$\bar{\Gamma} \equiv (\gamma_{ij}) \equiv \Gamma_{\bar{P}} + \frac{p}{1-p} \Gamma_P = \frac{p}{1-p} \Gamma_{\bar{P}} , \quad (27)$$

and assume $\bar{\Gamma}$ is positive definite. Then, for $t \in \mathbf{R}^d$, the permutation distribution $\hat{R}_{m,n}^W$ of $W_{m,n}$ defined in (??) with T replaced by W satisfies

$$\left| \hat{R}_{m,n}^W(t) - \tilde{L}(t) \right| \xrightarrow{P} 0 , \quad (28)$$

where $\tilde{L}$ denotes the d -variate normal distribution with mean 0 and covariance matrix $\bar{\Gamma}$ defined in (??).

Remark 3.1. Under H_0 , the true unconditional sampling distribution of $W_{m,n}$ is asymptotically normal with mean 0 and covariance matrix

$$\Gamma_P + \frac{p}{1-p} \Gamma_Q ,$$

which does not equal $\bar{\Gamma}$ defined by (??) in general.

The permutation test based on any function of $W_{m,n}$ in general fails to achieve the asymptotic rejection probability of α as the limiting distribution of the statistic $W_{m,n}$ depends on the underlying distributions P and Q . By multiplying by the inverse of the squared root of the estimated covariance matrix, this modified Hotelling's T^2 statistic becomes asymptotically pivotal and one can achieve the asymptotic validity of the permutation tests even when the underlying distributions are not identical under the null hypothesis, while still retaining the finite sample exactness in the case of homogeneous underlying populations.

Theorem 3.2. Assume the setup and conditions of Theorem ?? . Define the test statistic

$$\tilde{W}_{m,n} = (\hat{\Gamma})^{-1/2} W_{m,n} = m^{1/2} (\hat{\Gamma})^{-1/2} \left[\hat{\theta}_m(X_1, \dots, X_m) - \hat{\theta}_n(Y_1, \dots, Y_n) \right] , \quad (29)$$

where

$$\hat{\Gamma} = \hat{\Gamma}_P + \frac{m}{n} \hat{\Gamma}_Q$$

and matrices $\hat{\Gamma}_P$ and $\hat{\Gamma}_Q$ are consistent estimators of Γ_P and Γ_Q . That is $\hat{\Gamma}_P = \hat{\Gamma}_P(X_1, \dots, X_m)$ is a consistent estimator of Γ_P under P , and also of $\Gamma_{\bar{P}}$ under $\bar{P}$;

similarly, $\hat{\Gamma}_Q$ is a consistent estimator of Γ_Q under Q and of $\Gamma_{\bar{P}}$ under $\bar{P}$. Then, the permutation distribution $R_{m,n}^{\tilde{W}}$ of $\tilde{W}_{m,n}$ defined in (??) with T replaced by $\tilde{W}$ satisfies

$$\left| \hat{R}_{m,n}^{\tilde{W}}(t) - \Phi_d(t) \right| \xrightarrow{P} 0, \quad (30)$$

where Φ_d denotes the d -variate standard normal distribution.

Now, we generalize the modified Hotelling's T^2 statistic in the case of means to general parameters by considering the squared Euclidean norm of $\tilde{W}_{m,n}$ defined in (??) below.

Theorem 3.3. Assume the setup and conditions of Theorem ?? and Theorem ??. Let

$$A_{m,n} = \|\tilde{W}_{m,n}\|^2 = W_{m,n}' \hat{\Gamma}^{-1} W_{m,n}, \quad (31)$$

where $\tilde{W}_{m,n}$ is d -dimensional as defined in (??) and $\|\cdot\|$ denotes the usual Euclidean norm. Then, for $t \in \mathbf{R}$, the permutation distribution $\hat{R}_{m,n}^A(t)$ of $A_{m,n}$ defined in (??) with T replaced by A satisfies

$$\left| \hat{R}_{m,n}^A(t) - \chi_d^2(t) \right| \xrightarrow{P} 0,$$

where χ_d^2 denotes the Chi-squared distribution with d degrees of freedom.

Example 3.1. (Testing Equality of Median Vector) Suppose $X_1, \dots, X_m$ are d -dimensional i.i.d. P , where $X_i = (X_{i,1}, \dots, X_{i,d})'$ for $i = 1, \dots, m$ with median vector $(m_1(P), \dots, m_d(P))'$, and independently, $Y_1, \dots, Y_n$ are d -dimensional i.i.d. Q , where $Y_j = (Y_{j,1}, \dots, Y_{j,d})'$ for $j = 1, \dots, n$ with median vector $(m_1(Q), \dots, m_d(Q))'$. We are interested in testing the null hypothesis

$$H_0 : m_k(P) = m_k(Q) \text{ for all } k = 1, \dots, d,$$

versus the alternative hypothesis

$$H_1 : m_k(P) \neq m_k(Q) \text{ for some } k = 1, \dots, d.$$

Denote by $f_k(\cdot)$ the marginal density of X . Then, the variance-covariance matrix of the sample median vector $(m_1(P), \dots, m_d(P))'$ is given (Babu and Rao, 1988) by

$$\Gamma_P = \begin{pmatrix} \frac{\gamma_{11}}{f_1^2} & \frac{\gamma_{12}}{f_1 f_2} & \cdots & \frac{\gamma_{1d}}{f_1 f_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\gamma_{d1}}{f_d f_1} & \frac{\gamma_{d2}}{f_d f_2} & \cdots & \frac{\gamma_{dd}}{f_d^2} \end{pmatrix},$$

assuming the marginal density $f_k(m_k(P))$ at the median value m_k exists and is strictly positive, and

$$\gamma_{r,s} = P(X_r \leq m_r(P), X_s \leq m_s(P)) - \frac{1}{4}, \quad r, s = 1, \dots, d.$$

The unknown quantities $f_k(m_k(P))$ and γ_{rs} can be estimated as follows: The estimated marginal density $\hat{f}_k(\cdot)$ can be obtained by the kernel estimator (Devroye and Wagner, 1980), bootstrap estimator (Efron, 1979), or the smoothed bootstrap (Hall, DiCiccio, and Romano, 1989). Also, $\hat{\gamma}_{r,s}$, a consistent estimator of $\gamma_{r,s}$, can be calculated using the empirical joint c.d.f.

$$\hat{\gamma}_{r,s} = \frac{1}{m} \sum_{i=1}^m I(X_{ir} \leq \hat{m}_r(P), X_{is} \leq \hat{m}_s(P)) - \frac{1}{4},$$

where $(\hat{m}_1(P), \dots, \hat{m}_d(P))'$ is the sample median vector of $(m_1(P), \dots, m_d(P))'$. Such estimators are well-known to be consistent estimators (under weak conditions) under the underlying data distribution, whether it is P , Q , or $\bar{P}$, and so the conditions of Theorem 3.2 hold.

4 Multiple Testing Using the Closure Method

Thus far, we have considered the joint testing problem when the problem of interest is to test the single hypothesis (??) that all $\theta_k(P) = \theta_k(Q)$ for $k = 1, \dots, d$. In this section, we will examine the multiple testing problem where we are now interested in testing *which* hypotheses among the d null hypotheses are false, rather than just testing whether *any* component of the d null hypotheses is false. In other words, we would like to establish which differences $\theta_k(P) - \theta_k(Q)$ are nonzero. In doing so, it requires a careful assessment; unlike testing a single hypothesis, testing many multiple hypotheses simultaneously may cause problems due to the possibility of many Type 1 errors. If one ignores the multiplicity issue and tests each hypothesis at level α , the probability of one or more false rejections grows rapidly with the number of hypotheses d and may be much greater than α . In such a case, the claim that the procedure controls the probability of any false rejections at level α is untrue. We shall therefore restrict our attention to multiple testing methods that control the classical familywise error rate (FWER), which is the probability of one or more false rejections, at level α . That is, control of the

FWER at level α requires that

$$FWER = P\{\text{reject at least one true null hypothesis}\} \leq \alpha$$

for all P in the model $\mathcal{P}$, in finite samples or at least asymptotically.

In multiple testing, there exist various procedures that control the FWER at level α . (See Chapter 9 of Lehmann and Romano (2005) for details on these procedures.) Although any appropriate FWER-controlling method can be used, we will provide an algorithm based on the closure method proposed by Marcus et al. (1976), which reduces the problem of controlling the FWER to that of performing individual tests of single hypotheses which control the usual probability of the Type 1 error at level α . To carry out the closure method in multiple testing based on permutation tests, we can proceed in the following manner (Romano et al., 2011).

Algorithm 4.1. (*Closure Method Based on Permutation Test*)

1. For each given $K \subseteq \{1, \dots, d\}$, test H_K at level α using the permutation test based on an asymptotically pivotal statistic (either (??) or the prepivoted statistic defined in (??)). Reject H_K , if the observed test statistic $S_{m,n,K} > \hat{c}_{m,n,K}$ where $\hat{c}_{m,n,K}$ is the lower $(1 - \alpha)$ quantile of the permutation distribution.
2. By the closure method, for a given $i \in \{1, \dots, d\}$, reject H_i if and only if H_K is rejected at level α for all 2^{d-1} subsets K for which $i \in K$.⁶

Corollary 4.1. *Algorithm ?? controls the FWER exactly if the underlying distributions are identical and controls the FWER asymptotically as long as second moments are finite and $\bar{\Sigma}$ defined in (??) is nonsingular.*

Remark 4.1. Note that when calculating the exact permutation test, $\alpha \cdot N!$ need not be an integer, in which case, the rejection probability may be slightly less than α . However, we can still achieve the finite sample exactness via randomization by making each individual test exact.

⁶Although, in principle, one needs to execute as many as $O(2^d)$ tests to apply the closure method, this procedure only requires the individual tests to be of level α to control the FWER without any further assumptions. To illustrate this, let A be the event that any *true* hypothesis H_i is rejected by the closure method, and let B be the event that the intersection of all true hypotheses is rejected at level α . Since $A \subseteq B$,

$$FWER = P\{A\} \leq P\{B\} \leq \alpha .$$

5 Simulation Results

Monte Carlo simulation studies based on the modified Hotelling's T^2 statistic (MT2) and prepivoted statistics (PP) are performed to investigate their performance in finite samples in regard to controlling the Type 1 error under the null hypothesis (??) and the power of the statistics under the alternatives.

For the Type 1 error simulations, in addition to the two asymptotically valid procedures MT2 and PP, we also consider other pre-existing methods such as the classical Hotelling's T^2 test (T2) and robust tests based on trimming and winsorized covariance matrices (RT). The simulation results comparing the behavior of each test in controlling the Type 1 error in finite samples are displayed in Table ??, where the nominal level considered is $\alpha = 0.05$. As demonstrated in the first column of Table ??, we investigate both symmetric and asymmetric distributions based on pairs of multivariate ($d = 7$) normal distributions, pairs of multivariate t -distributions with five and ten degree of freedom, respectively, and a multivariate log normal distribution with a set of multivariate normal distributions with different covariance matrices. All pairs of distributions studied are chosen in such a way that they share the same mean vector (0, except the last case where the mean is vector of $e^{1/2}$) but with different covariance matrices. The covariance matrices used in the studies include the 7 by 7 identity matrix I_7 as well as Σ_1 , and Σ_2^7 given by

$$\Sigma_1 = \begin{pmatrix} 1 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 1 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 1 & 0.8 & 0.64 & 0.51 & 0.41 & 0.33 & 0.26 \\ 0.8 & 1 & 0.8 & 0.64 & 0.51 & 0.41 & 0.33 \\ 0.64 & 0.8 & 1 & 0.8 & 0.64 & 0.51 & 0.41 \\ 0.51 & 0.64 & 0.8 & 1 & 0.8 & 0.64 & 0.51 \\ 0.41 & 0.51 & 0.64 & 0.8 & 1 & 0.8 & 0.64 \\ 0.33 & 0.41 & 0.51 & 0.64 & 0.8 & 1 & 0.8 \\ 0.26 & 0.33 & 0.41 & 0.51 & 0.64 & 0.8 & 1 \end{pmatrix}$$

The sample sizes (m and n) are displayed in the first two rows of Table ?. For each pair of distributions considered, 10,000 simulations were performed where, for any simulation, 9,999 permutations were randomly sampled to calculate the permutation distributions. For the prepivoted statistic (??), 2,000 bootstrap samples were used.

As displayed in Table ??, the classical Hotelling's T^2 test, which requires multivariate normal distributions with the same covariance matrix for validity, is not robust to the violation of the underlying assumptions. Since the pairs of distributions we consider do

⁷Note that Σ_2 is a typical autoregressive structure with $\rho = 0.8$.

not share the same covariance matrix and some of them are not normally distributed, the classical Hotelling's T^2 test fails to control the Type 1 error. For example, when samples are drawn from $\mathcal{N}(0, \mathbf{I}_7)$ and $\mathcal{N}(0, \Sigma_1)$, despite the fact that they have the same mean vector of 0, the rejection probabilities of the classical Hotelling's T^2 for the various sample sizes considered range between 0.0737 and 0.1847.

Distributions	Tests	Sample Sizes (m / n)							
		10	30	50	50	100	100	200	300
		20	50	100	150	150	200	300	500
$\mathcal{N}(0, \mathbf{I}_7) \text{ \& } \mathcal{N}(0, \Sigma_1)$	MT2	0.0802	0.0593	0.0542	0.0529	0.0561	0.0565	0.0479	0.0509
	PP	0.0736	0.0560	0.0494	0.0506	0.0511	0.0514	0.0484	0.0496
	T2	0.1847	0.1052	0.1016	0.1407	0.0835	0.1034	0.0737	0.0849
	RT	0.9829	0.9044	0.9807	0.9865	0.9814	0.9810	0.9802	0.9811
$\mathcal{N}(0, \Sigma_1) \text{ \& } \mathcal{N}(0, \Sigma_2)$	MT2	0.0843	0.0591	0.0532	0.0580	0.0549	0.0537	0.0471	0.0511
	PP	0.0456	0.0484	0.0454	0.0428	0.0504	0.0481	0.0458	0.0488
	T2	0.1052	0.1078	0.1138	0.1717	0.2853	0.1093	0.0796	0.0896
	RT	0.9116	0.8854	0.7816	0.8001	0.8175	0.7764	0.7574	0.7125
$\mathcal{N}(0, \mathbf{I}_7) \text{ \& } \mathcal{N}(0, \Sigma_2)$	MT2	0.1224	0.0693	0.0614	0.0581	0.0538	0.0512	0.0516	0.0505
	PP	0.0424	0.0587	0.0564	0.0498	0.0449	0.0472	0.0486	0.0491
	T2	0.1824	0.1709	0.1634	0.2776	0.1033	0.1629	0.1036	0.1221
	RT	0.9936	0.9923	0.9941	0.9952	0.9936	0.9945	0.9938	0.9821
$t_5(0, \mathbf{I}_7) \text{ \& } t_5(0, \Sigma_1)$	MT2	0.0761	0.0578	0.0563	0.0594	0.0530	0.0536	0.0487	0.0507
	PP	0.0540	0.0522	0.0510	0.0504	0.0464	0.0450	0.0486	0.0498
	T2	0.0848	0.0913	0.0945	0.1362	0.0772	0.0940	0.0701	0.0842
	RT	0.9899	0.9801	0.9896	0.9907	0.9908	0.9912	0.9905	0.9879
$t_5(0, \Sigma_1) \text{ \& } t_5(0, \Sigma_2)$	MT2	0.0804	0.0558	0.0528	0.0534	0.0543	0.0473	0.0556	0.0504
	PP	0.0404	0.0504	0.0432	0.0346	0.0490	0.0406	0.0450	0.0478
	T2	0.0927	0.1009	0.1042	0.1059	0.2711	0.1058	0.0802	0.0812
	RT	0.9387	0.9102	0.8978	0.9902	0.8911	0.8936	0.8842	0.8920
$t_5(0, \mathbf{I}_7) \text{ \& } t_5(0, \Sigma_2)$	MT2	0.1014	0.0606	0.0585	0.0598	0.0525	0.055	0.0539	0.0499
	PP	0.0308	0.0454	0.0494	0.0544	0.0524	0.0466	0.0484	0.0494
	T2	0.1596	0.1621	0.1582	0.2628	0.1039	0.1587	0.1008	0.1197
	RT	0.9956	0.9987	0.9978	0.9960	0.9957	0.9977	0.9949	0.9974
$t_{10}(0, \mathbf{I}_7) \text{ \& } t_{10}(0, \Sigma_1)$	MT2	0.0785	0.0556	0.0544	0.053	0.0518	0.0514	0.0493	0.0508
	PP	0.0660	0.0566	0.0540	0.0492	0.0518	0.0476	0.0504	0.0494
	T2	0.0929	0.0979	0.0985	0.1414	0.0778	0.0944	0.0768	0.0845
	RT	0.9849	0.9852	0.9845	0.9856	0.9844	0.9851	0.9833	0.9858
$t_{10}(0, \Sigma_1) \text{ \& } t_{10}(0, \Sigma_2)$	MT2	0.0804	0.0548	0.0531	0.0527	0.0533	0.0518	0.0519	0.0502
	PP	0.0422	0.0502	0.0486	0.0438	0.0440	0.0442	0.0464	0.0488
	T2	0.1004	0.1102	0.1042	0.1059	0.2711	0.1058	0.0830	0.0832
	RT	0.9206	0.8996	0.8561	0.8697	0.8473	0.8591	0.8497	0.8488
$t_{10}(0, \mathbf{I}_7) \text{ \& } t_{10}(0, \Sigma_2)$	MT2	0.1163	0.0651	0.0611	0.0552	0.0497	0.0525	0.0515	0.0509
	PP	0.0350	0.0474	0.0558	0.0538	0.0459	0.0488	0.0494	0.0496
	T2	0.1649	0.1701	0.1371	0.0772	0.1506	0.0989	0.0722	0.0899
	RT	0.9952	0.9932	0.9960	0.9952	0.9948	0.9957	0.9959	0.9943
$\ln \mathcal{N}(0, \mathbf{I}_7) \text{ \& } \mathcal{N}(e^{0.5}, \mathbf{I}_7)$	MT2	0.0647	0.0778	0.0852	0.0951	0.0692	0.0638	0.0546	0.0546
	PP	0.0582	0.0496	0.0414	0.0492	0.0424	0.0468	0.0472	0.0486
	T2	0.2872	0.1987	0.1301	0.1328	0.1278	0.2190	0.1232	0.2137
	RT	0.9983	0.9979	0.9984	0.9978	0.9991	0.9989	0.9992	0.9990
$\ln \mathcal{N}(0, \mathbf{I}_7) \text{ \& } \mathcal{N}(e^{0.5}, \Sigma_1)$	MT2	0.0586	0.0816	0.0772	0.0928	0.0654	0.0603	0.0552	0.0533
	PP	0.0551	0.0560	0.0459	0.0512	0.0488	0.0454	0.0454	0.0478
	T2	0.3784	0.2034	0.1470	0.3774	0.1308	0.1374	0.1352	0.2320
	RT	0.9979	0.9976	0.9987	0.9989	0.9992	0.9990	0.9989	0.9995
$\ln \mathcal{N}(0, \mathbf{I}_7) \text{ \& } \mathcal{N}(e^{0.5}, \Sigma_2)$	MT2	0.0855	0.0935	0.0946	0.1041	0.0731	0.0682	0.0555	0.0540
	PP	0.0354	0.0484	0.0562	0.0502	0.0478	0.0494	0.0462	0.0484
	T2	0.3051	0.2892	0.1660	0.3869	0.1308	0.2679	0.1352	0.2320
	RT	0.9976	0.9977	0.9984	0.9989	0.9985	0.9993	0.9995	0.9989

Table 1: Rejection probabilities for Multivariate Permutation Test under (??) ($\alpha = 0.05$)

The robust T^2 test based on trimmed means and winsorized covariance matrices also performs poorly in controlling the Type 1 error. The rejection probabilities of the robust T^2 test across various distributions and sample sizes are (too) liberal, far exceeding the nominal level $\alpha = 0.05$. This result may explain why the robust estimators presented in Lix and Keselman (2004)⁸ have (seemingly) stronger power. Our simulation results shows that it is probably due to its liberal control of the Type 1 error. When the rejection probability under the null is large, by continuity, the rejection probabilities under some alternatives can also be very large. With the increased rejection probability, rejection of the test may be erroneously construed as rejection of $\mu(P) = \mu(Q)$ when in fact the rejection is caused by the inequality of distributions.

In contrast, our proposed methods seem to work well in controlling Type 1 error in both moderate and large sample sizes. The simulation results in Table ?? confirm that for each pair of distributions we consider, the permutation tests based on the modified Hotelling's T^2 test statistic and the prepivoted statistic are valid for testing equality of multivariate means since the rejection probabilities of both tests approximately attain the nominal level α in large samples. Although the rejection probabilities for MT2 in the case of the asymmetric distributions with log normal seem to converge to the nominal level α at a slower rate than those in the symmetric cases, they still tend to the nominal level α as sample sizes get bigger.

The powers of the two asymptotically valid procedures MT2 and PP are investigated for $\mathcal{N}(0, I_7)$ -distributed data in the first and $\mathcal{N}(\mu, \Sigma)$ -distributed data in the second group, where μ and Σ vary as depicted in the first two columns of Table ?. For simplicity, we use a balanced design with varying $m = n$ as displayed in the first row of the table. Both tests are consistent in the sense that the power is approaching one as sample sizes increase. However, the power for small to moderate sample sizes is quite different and seems to depend on the correlation structure and the considered alternatives. For a shift effect in all components and independent components ($\Sigma = I_7$), MT has considerably larger power than the PP procedure. In all other cases, it is the other way around.

⁸Among all the robust tests presented in Lix and Keselman (2004), we chose to perform the robust T^2 test so as to compare its performance of the Type 1 error control with the classical Hotelling's T^2 test and the permutation test based on the modified Hotelling's T^2 statistic.

Distributions		Tests	Sample Sizes ($m = n$)							
μ	Σ		20	30	50	100	150	200	300	500
$\begin{pmatrix} 0.3 \\ \vdots \\ 0.3 \end{pmatrix}$	I_7	MT2	0.3285	0.5218	0.8119	0.9937	1	1	1	1
		PP	0.2135	0.3545	0.5155	0.8751	0.9701	0.9960	1	1
	Σ_1	MT2	0.1476	0.2211	0.3733	0.7109	0.8992	0.9706	0.9987	1
		PP	0.2041	0.2775	0.4651	0.7605	0.9255	0.9705	0.9965	1
	Σ_2	MT2	0.1469	0.2109	0.3428	0.6796	0.8721	0.9579	0.9973	1
		PP	0.2395	0.2785	0.4431	0.7495	0.9181	0.9635	0.9975	1
$\begin{pmatrix} 0.5 \\ \vdots \\ 0.5 \end{pmatrix}$	I_7	MT2	0.8019	0.9641	0.9995	1	1	1	1	1
		PP	0.5530	0.8038	0.9544	1	1	1	1	1
	Σ_1	MT2	0.3665	0.5810	0.8552	0.9966	0.9999	1	1	1
		PP	0.4726	0.6608	0.8868	0.9956	0.9998	1	1	1
	Σ_2	MT2	0.3501	0.5469	0.8314	0.9932	1	1	1	1
		PP	0.5392	0.6464	0.8680	0.9968	1	1	1	1
$\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$	I_7	MT2	1	1	1	1	1	1	1	1
		PP	0.9968	1	1	1	1	1	1	1
	Σ_1	MT2	0.9535	0.9983	1	1	1	1	1	1
		PP	0.9766	0.9984	1	1	1	1	1	1
	Σ_2	MT2	0.9386	0.9969	1	1	1	1	1	1
		PP	0.9844	0.9976	1	1	1	1	1	1
$\begin{pmatrix} 0.1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$	I_7	MT2	0.5211	0.7578	0.9626	0.9997	1	1	1	1
		PP	0.6224	0.8538	0.9836	1	1	1	1	1
	Σ_1	MT2	0.6156	0.8504	0.9866	1	1	1	1	1
		PP	0.6160	0.8502	0.9848	0.9998	1	1	1	1
	Σ_2	MT2	0.6544	0.8724	0.9901	1	1	1	1	1
		PP	0.6982	0.8552	0.9818	1	1	1	1	1

Table 2: Power Comparison Between MT2 and PP for $\mathcal{N}(0, I_7)$ and $\mathcal{N}(\mu, \Sigma)$ ($\alpha = 0.05$)

6 Empirical Illustration

In this section, we illustrate empirical applications of multiple testing based on the multivariate permutation tests developed above to test the effects that exercise has on seven biometric measures. Charness and Gneezy (2009) conduct an experiment in which they randomly divide participants into three different groups. While there is no further requirement for the participants in the control group (C), the first treatment group members (T1) are paid \$25 to attend the gym once in a week, and the participants in the second treatment group (T2) are paid an additional \$100 to attend the gym eight more times in the following four weeks. Before, during, and for a period of seven weeks after the experiment period, the 39 members of the control group, the 57 members of the

first treatment group, and the 60 members of the second treatment group are measured on seven different health level indicators (body fat %, pulse rate, weight, BMI, waist, systolic blood pressure, and diastolic blood pressure). See Charness and Gneezy (2009) for more details.

Based on the marginal p -values from the two-sample Wilcoxon test, Charness and Gneezy conclude that “with the exception of the blood-pressure measures, we see that the biometric measures of the eight-times group improved significantly relative to both the control group and (with the further exception of the pulse rate) the one-time group. Thus, it appears that there are real health benefits that accrue from paying people to go to the gym eight times in a month.” However, their approach has two drawbacks. First, the two-sample Wilcoxon statistic is not a suitable test statistic for testing equality of means because it may fail to control the rejection probability at α unless a shift model is assumed. As argued in Chung and Romano (2015) in great detail, the Wilcoxon test is most suitable for testing $P(X \leq Y) = P(Y \leq X)$, when the X s are i.i.d. P and independently, the Y s are i.i.d. Q . Even in this case, however, the Wilcoxon statistic has to be appropriately modified so that the test statistic becomes asymptotically pivotal. (See Janssen (1999) or Chung and Romano (2015) for more details.) Moreover, testing seven measurements at the same time will cause the Familywise Error Rate to be greater than the nominal level α . In order to take into account the fact that seven individual tests are performed, a more careful assessment is required.

Our primary goal is to simultaneously test the effect of exercise based on *mean differences* of seven biometric measures while controlling the Familywise Error Rate. Instead of testing the effect of each biometric measure individually, we want to test whether there is an overall effect of exercise for each comparison group and to examine *which* hypotheses among the seven are to be rejected. To do so, we use the permutation tests based on the modified Hotelling’s T^2 statistic defined in (??) as well as the prepivoted statistic defined in (??) and apply the closure method explained above to control the FWER. The adjusted p -value for H_i is defined to be the smallest value α such that H_i is rejected by the multiple testing procedure which controls the FWER at level α . Thus, when applying the closure method, the adjusted p -value for H_i is defined to be the largest p -value among all the intersection hypotheses which contain i . The marginal p -values and the adjusted p -values for each comparison group are presented in Table ?? and Table ?? for the modified Hotelling’s T^2 statistic and the prepivoted statistic, respectively. The results below are based on 99,999 number of permutations for the modified Hotelling’s T^2 statistic and 1,000 number of bootstrap samples and 9,999 number of permutations for the prepivoted statistic.

	C-T1		C-T2		T1 - T2	
	Marginal	Adjusted	Marginal	Adjusted	Marginal	Adjusted
Body Fat %	0.05441	0.45582	0.00001	0.00105	0.00316	0.09155
Pulse Rate	0.0225	0.35441	0.06451	0.15424	0.74411	0.91710
Weight (kg)	0.99397	0.99397	0.13194	0.32724	0.00651	0.13261
BMI	0.9169	0.93457	0.09447	0.25627	0.0054	0.13261
Waist (in.)	0.71694	0.94425	0.06466	0.18402	0.07568	0.41272
Systolic BP	0.26476	0.77049	0.19305	0.41282	0.82418	0.91710
Diastolic BP	0.30853	0.79320	0.87274	0.87274	0.41255	0.83281

Table 3: Marginal and Adjusted p -values for Testing Equality of Means based on the Modified Hotelling’s Statistic (??)

For C-T2 group, Body Fat % is significant based on both Bonferroni and the closure method. However, for T1-T2 comparison, Body Fat %, Weight, and BMI are significant based on Bonferroni or Holm’s procedure while the closure method does not reject any hypothesis.

	C-T1		C-T2		T1 - T2	
	Marginal	Adjusted	Marginal	Adjusted	Marginal	Adjusted
Body Fat %	0.0641	0.2716	0.0004	0.0020	0.0044	0.0122
Pulse Rate	0.0248	0.1246	0.0581	0.3110	0.7202	0.9371
Weight (kg)	0.9999	0.9999	0.1457	0.4045	0.0066	0.0172
BMI	0.9398	0.9692	0.1032	0.3176	0.0062	0.0156
Waist (in.)	0.6973	0.9492	0.0725	0.3110	0.0706	0.2255
Systolic BP	0.2530	0.7417	0.1963	0.4045	0.8246	0.9371
Diastolic BP	0.3081	0.7417	0.9175	0.9175	0.4313	0.7960

Table 4: Marginal and Adjusted p -values for Testing Equality of Means based on the Prepivoted Statistic (??)

According to the Bonferroni or the Holm (1979) procedure, only Body Fat % is significant for both C-T2 and T1-T2 groups. As explained earlier, these procedures are

quite conservative. If we apply the closure method based on Algorithm ?? using the prepivoted statistic (??), more hypotheses are rejected. Not only is Body Fat % for C-T2 and T1-T2 significant, but both Weight and BMI for T1-T2 show significant results.

We suggest to use the maximum statistic (prepivoted statistic) over the modified Hotelling's T^2 in the cases where we expect to see a few strong effects. On the other hand, when there are minor effects across all cases, the modified Hotelling's T^2 , which captures the overall effects together, would work better than the maximum statistic.

Overall, we conclude that exercise is beneficial as do the previous authors. However, the basis for such claims is more statistically sound when one accounts for doing many tests and by making sure each test is valid in some finite sample or asymptotic sense.

Remark 6.1. It may be surprising to see the cases where the Bonferroni procedure leads to a rejection but the closure method does not. In fact, neither dominates the other. Consider the simplified setting where for $i = 1, \dots, d$, $X_i \sim N(\theta_i, 1)$ is independent of each other and $H_i : \theta_i = 0$ for $i = 1, \dots, d$. For the intersection hypothesis $H_I : \theta_i = 0 \ i \in I$, use the test statistic $T_I = \sum_{i \in I} X_i^2$, which follows $\chi_{|I|}^2$. For example, when $s = 10$ and $\theta_1 = 3$, $\theta_2 = 1$, $\theta_3 = 2$, and $\theta_i = 0$ for $i = 4, \dots, 10$, the marginal p -value for θ_1 is 0.006046 while the adjusted p -value based on the closure method is 0.145849. On the other hand, if the test statistic for H_I is the minimum p -value over $i \in I$, the closure method will reject at least as many hypotheses as the Bonferroni method. In fact, the Holm-Bonferroni procedure is a special case of a closure method where the closure method is applied to the Bonferroni test. Closure methods based on different constructions of the tests for the intersections can lead to different conclusions.

7 Conclusion

Permutation tests has been quite popular among academic researchers due to their simplicity and the exact finite sample Type 1 error control property that other resampling methods such as bootstrap or subsampling lack. However, for testing parameters, conducting a permutation test requires a careful treatment as misuse can result in losing control of the rejection probability even asymptotically. The fundamental motivation behind the permutation tests hinges on the fact that when all the observations are i.i.d., any permuted sample has the same distribution as the original sample. If observations are sampled from heterogeneous populations, however, the justification of permutation tests no longer holds and permutation tests lose their asymptotic validity, even in the simple case of testing equality of means. We provide a framework whereby the permutation test can asymptotically attain the rejection probability at α even with

heterogeneous populations while retaining their exactness property in finite samples in the case of homogeneous populations.

To summarize, if one is interested in testing equality of means of multivariate populations based on the mean differences or the maximum of the sample mean absolute differences, permutation tests based on appropriate statistics, namely asymptotically pivotal statistics, can serve the purpose of error control. In addition, if the maximum value of the mean differences over all the components is of interest, one can transform a test statistic using the prepivoting method in order to achieve asymptotic validity in general, while maintaining the exactness property when $P = Q$. By studying the behavior of the permutation test, we learn that the permutation distribution behaves like the unconditional true sampling distribution when all the observations are sampled from the mixture distribution $\bar{P} = pP + (1 - p)Q$, where p is the limit of the fraction of observations from P . If in our case permutation tests are constructed based on a test statistic that is asymptotically distribution-free, asymptotic justification of the test can be achieved.

Moreover, in dealing with multivariate cases, a careful assessment is required as the rejection probability can be inflated by performing many tests simultaneously. By applying the closure method in the multiple testing setting, one can control the familywise error rate at the nominal level α in a systematic way. We show in the empirical illustration that the closure method tends to reject more hypotheses than Bonferroni or Holm when testing based on the maximum statistic. Thus, when we expect to have a few strong effects, we suggest to use the closure method based on the maximum statistic. On the other hand, when the test statistic is designed to capture the overall effects together, the modified Hotelling's T^2 statistic may perform better. Neither test dominates the other in terms of its ability to find true rejections. However, our analysis shows that both offer valid error control.

A Useful Lemmas for Multivariate Cases

Suppose data $X^n = (X_1, \dots, X_n)$ has distribution P_n in $\mathcal{X}_n$ and let $\mathbf{G}_n$ be a finite group of transformations of $\mathcal{X}_n$ onto itself. For a given test statistic $T_n = T_n(X^n)$, let $\hat{R}_n^T(\cdot)$ denote the randomization distribution of T_n , defined by

$$\hat{R}_n^T(t) = \frac{1}{|\mathbf{G}_n|} \sum_{g \in \mathbf{G}_n} \mathbf{I}\{T_n(gX_n) \leq t\} . \quad (32)$$

Hoeffding (1952) gave a sufficient condition to derive the limiting behavior of the randomization distribution of a real-valued test statistic T_n . We generalize this result to multivariate cases where we investigate the limiting behavior of the multivariate randomization distribution based on a test statistic T_n that is a d -dimensional vector in $\mathbf{R}^d$. So, in (??), both T_n and t are vectors in $\mathbf{R}^d$. Note that $\hat{R}_n^T(\cdot)$ takes values in $\mathbf{R}$ for any t .

Lemma A.1. *Let G_n and G'_n be independent and uniformly distributed over $\mathbf{G}_n$ (and independent of X^n). Suppose, under P_n ,*

$$(T_n(G_n X^n), T_n(G'_n X^n)) \xrightarrow{d} (T, T'), \quad (33)$$

where T and T' are independent d -variate distributions, each with common multivariate c.d.f. $R^T(\cdot)$. Then, for all continuity points $t \in \mathbf{R}^d$ of $R^T(\cdot)$,

$$\hat{R}_n^T(t) R^T(t) . \quad (34)$$

Conversely, if (??) holds for some limiting c.d.f. R^T whenever t is a continuity point of $R^T(\cdot)$, then (??) holds.

PROOF OF LEMMA ??: This is a special case of Lemma 4.1 from Dümbgen and Zerial (2013). ■

We extend Slutsky's theorem for the randomization distributions given in Subsection 5.2 of Chung and Romano (2013) to the multivariate case.

Lemma A.2. *Suppose X^n has distribution P_n in $\mathcal{X}_n$, and $\mathbf{G}_n$ is a finite group of transformations g of $\mathcal{X}_n$ onto itself. Also, let G_n be a random variable that is uniform on $\mathbf{G}_n$. Assume X^n and G_n are mutually independent. Let $\hat{R}_n^B$ denotes the randomization distributions of a d -dimensional random vector B_n , defined by*

$$\hat{R}_n^B(t) = \frac{1}{|\mathbf{G}_n|} \sum_{g \in \mathbf{G}_n} \mathbf{I}\{B_n(gX^n) \leq t\}. \quad (35)$$

Suppose, under P_n ,

$$B_n(G_n X^n) \xrightarrow{P} b \quad (36)$$

for a constant $b \in \mathbf{R}^d$. Then, under P_n ,

$$\hat{R}_n^B(t) = \frac{1}{|\mathbf{G}_n|} \sum_{g \in \mathbf{G}_n} \mathbf{I}\{B_n(gX^n) \leq t\} \xrightarrow{P} \delta_b(t) \quad \text{if } t \neq b, \quad (37)$$

where $\delta_c(\cdot)$ denotes the distribution function corresponding to the point mass function at $c \in \mathbf{R}^d$.

PROOF OF LEMMA ???: Let G'_n have the same distribution as G_n and be independent from G_n and X_n . Since $B_n(G_n X^n)$ converges in probability to a constant b (i.e., each element of $B_n(G_n X^n)$ converges in probability to its corresponding element of b) and $B_n(G'_n X^n) \xrightarrow{P} b$, it follows that $(B_n(G_n X^n), B_n(G'_n X^n)) \xrightarrow{P} (b, b)$. Thus, the result follows from Lemma ??. ■

Lemma A.3. *Let B_n and T_n be sequences of d -dimensional random variables satisfying (??) and*

$$(T_n(G_n X^n), T_n(G'_n X^n)) \xrightarrow{d} (T, T'), \quad (38)$$

where T and T' are independent, each with common d -variate c.d.f. $R^T(\cdot)$. Let $\hat{R}_n^{T+B}(t)$ denote the randomization distribution of $T_n + B_n$, defined in (??) with B replaced by $T + B$. Then, $\hat{R}_n^{T+B}(t)$ converges to the c.d.f. of $T + b$ in probability. In other words,

$$\hat{R}_n^{T+B}(t) \equiv \frac{1}{|\mathbf{G}_n|} \sum_{g \in \mathbf{G}_n} \mathbf{I}\{T_n(gX^n) + B_n(gX^n) \leq t\} \xrightarrow{P} R^{T+b}(t),$$

if R^{T+b} is continuous at $t \in \mathbf{R}^d$, where $R^{T+b}(\cdot)$ denotes the corresponding d -variate c.d.f. of $T + b$. (Of course, $R^{T+b}(t) = R^T(t - b)$.)

PROOF OF LEMMA ???: Without loss of generality, assume $b_k = 0$ for $k = 1, \dots, d$. For any $\varepsilon \in \mathbf{R}^d$ with each component ε_k being positive for $k = 1, \dots, d$,

$$\begin{aligned} & \frac{1}{|\mathbf{G}_n|} \sum_{g \in \mathbf{G}_n} \mathbf{I}\{T_n(gX^n) \leq t - \varepsilon\} - \frac{1}{|\mathbf{G}_n|} \sum_{g \in \mathbf{G}_n} \mathbf{I}\{|B_n(gX^n)| > \varepsilon\} \\ & \leq \frac{1}{|\mathbf{G}_n|} \sum_{g \in \mathbf{G}_n} \mathbf{I}\{T_n(gX^n) + B_n(gX^n) \leq t\} \\ & \leq \frac{1}{|\mathbf{G}_n|} \sum_{g \in \mathbf{G}_n} \mathbf{I}\{T_n(gX^n) \leq t + \varepsilon\} + \frac{1}{|\mathbf{G}_n|} \sum_{g \in \mathbf{G}_n} \mathbf{I}\{|B_n(gX^n)| > \varepsilon\} \end{aligned}$$

since it holds for each k -component of $T_n(gX^n)$, $B_n(gX^n)$, t , and ε . First, note that, by Lemma ??, $\frac{1}{|\mathbf{G}_n|} \sum_{g \in \mathbf{G}_n} \mathbf{I}\{|B_n(gX^n)| > \varepsilon\}$ converges in probability to 0 for any $\varepsilon > 0$. Also, by Lemma ??, (??) implies

$$\hat{R}_n^T(t) = \frac{1}{|\mathbf{G}_n|} \sum_{g \in \mathbf{G}_n} \mathbf{I}\{T_n(gX^n) \leq t\} \xrightarrow{P} R^T(t) \quad (39)$$

if $R^T(\cdot)$ is continuous at $t \in \mathbf{R}^d$. Thus, if both $t - \varepsilon$ and $t + \varepsilon$ are continuity points of $R^T(\cdot)$, the first term of the first line and the first term of the third line converge in probability to $R^T(t - \varepsilon)$ and $R^T(t + \varepsilon)$, respectively. Therefore,

$$R^T(t - \varepsilon) \leq \hat{R}_n^{T+B}(t) \leq R^T(t + \varepsilon)$$

with probability tending to one, for continuity points $t - \varepsilon$ and $t + \varepsilon$ of $R^T(\cdot)$. Now, let $\varepsilon \downarrow 0$ through continuity points to deduce that

$$\hat{R}_n^{T+B}(t) \xrightarrow{P} R^T(t). \quad \blacksquare$$

Lemma A.4. *Let A_n and T_n , respectively, be a sequence of $d \times d$ nonsingular random matrices and a sequence of random variables satisfying the conditions*

$$A_n(G_n X^n) \xrightarrow{P} C$$

where C is a fixed $d \times d$ nonsingular matrix, and

$$(T_n(G_n X^n), T_n(G'_n X^n)) \xrightarrow{d} (T, T'),$$

where T and T' are independent, each with common d -variate c.d.f. $R^T(\cdot)$. Then, the randomization distribution of $A_n^{-1}T_n$ converges to the c.d.f. of $C^{-1}T$ in probability. In other words,

$$\hat{R}_n^{A^{-1}T}(t) \equiv \frac{1}{|\mathbf{G}_n|} \sum_{g \in \mathbf{G}_n} \mathbf{I}\{A_n(gX^n)^{-1}T_n(gX^n) \leq t\} \xrightarrow{P} R^{C^{-1}T}(t),$$

if $R^{C^{-1}T}$ is continuous at t , where $R^{C^{-1}T}(\cdot)$ denotes the corresponding c.d.f. of $C^{-1}T$. ■

PROOF OF LEMMA ?? : Write

$$A_n^{-1}T_n = C^{-1}T_n + (A_n^{-1} - C^{-1})T_n .$$

Then, we can apply Lemma ?? with $B_n = (A_n^{-1} - C^{-1})T_n$, if we can verify the condition $B_n(G_n X^n)0$. But,

$$B_n(G_n X^n) = [A_n(G_n X^n)^{-1} - C^{-1}]T_n(G_n X^n)0 \cdot T = 0 ,$$

by the usual multivariate Slutsky's Theorem. Finally, the behavior of $C^{-1}T_n$ follows trivially from that of T_n . $\blacksquare$

Lemma A.5. *Let G_n and G'_n be independent and uniformly distributed over $\mathbf{G}_n$ (and independent of X^n). Assume a d -dimensional random vector T_n satisfies (??). Also, assume $A_n(\cdot)$ is a sequence of $d \times d$ nonsingular random matrices such that*

$$A_n(G_n X^n)C \quad (40)$$

for a fixed $d \times d$ nonsingular matrix C , i.e., each element of $A_n(\cdot)$ converges in probability to the corresponding element of C . Further assume $B_n(\cdot)$ is a d -dimensional random vector such that

$$B_n(G_n X^n)b, \quad (41)$$

for a constant $b = (b_1, \dots, b_d)' \in \mathbf{R}^d$. Let $R^{C^{-1}T+b}(\cdot)$ denote the distribution of $C^{-1}T + b$, where T is the limiting random variable assumed in (??). Let $\hat{R}_n^{A^{-1}T+B}(\cdot)$ denote the randomization distribution corresponding to the statistic sequence $A_n^{-1}T_n + B_n$, i.e. replace T_n in (??) by $A_n^{-1}T_n + B_n$, so

$$\hat{R}_n^{A^{-1}T+B}(t) \equiv \frac{1}{|\mathbf{G}_n|} \sum_{g \in \mathbf{G}_n} \mathbf{I}\{A_n(gX^n)^{-1}T_n(gX^n) + B_n(gX^n) \leq t\} \quad (42)$$

Then,

$$\hat{R}_n^{A^{-1}T+B}(t)R^{C^{-1}T+b}(t),$$

if the distribution $R^{C^{-1}T+b}(\cdot)$ of $C^{-1}T + b$ is continuous at $t = (t_1, \dots, t_d) \in \mathbf{R}^d$. (Of course, $R^{C^{-1}T+b}(t) = R^T(C(t - b))$.)

PROOF OF LEMMA ???: The proof follows from Lemma ?? and Lemma ??. $\blacksquare$

The following lemma provides a generalization of the continuous mapping theorem for the randomization distributions in multivariate cases.

Lemma A.6. *Suppose the randomization distribution of a test statistic T_n converges to T in probability. In other words,*

$$\hat{R}_n^T(t) \equiv \frac{1}{|\mathbf{G}_n|} \sum_{g \in \mathbf{G}_n} \mathbf{I}\{T_n(gX^n) \leq t\} \xrightarrow{P} R^T(t),$$

if R^T is continuous at $t \in \mathbf{R}^d$, where $R^T(\cdot)$ denotes the corresponding c.d.f. of T . Let h be a measurable map from $\mathbf{R}^d$ to $\mathbf{R}^s$. Let C be the set of points in $\mathbf{R}^d$ for which h is

continuous. If $P(T \in C) = 1$, then the randomization distribution of $h(T_n)$ converges to $h(T)$ in probability.

PROOF OF LEMMA ?? : Since the randomization distribution of the test statistic T_n converges to T in probability, by Lemma ??, under P_n ,

$$(T_n(G_n X^n), T_n(G'_n X^n)) \xrightarrow{d} (T, T')$$

holds, where G_n and G'_n are independent and uniformly distributed over $\mathbf{G}_n$, and T and T' are independent d -variate distributions, with a common c.d.f. $R^T(\cdot)$. From the continuity assumption of $h(\cdot)$,

$$(h(T_n(G_n X^n)), h(T_n(G'_n X^n))) \xrightarrow{d} (h(T), h(T')) ,$$

is satisfied, where $h(T)$ and $h(T')$ are independent s -variate distributions, and thus, the result follows again from Lemma ??. $\blacksquare$

Lemma A.7. *Let the metric $d(P, Q)$ between P and Q be defined as*

$$d(P, Q) = \max \left(\sup_{t \in \mathbf{R}^d} |F_P(t) - F_Q(t)|, \|\Sigma_P - \Sigma_Q\| \right) ,$$

where $F_P(\cdot)$ and $F_Q(\cdot)$ denote the corresponding c.d.f.s of the probability distribution P and Q , respectively, and

$$\|\Sigma_P - \Sigma_Q\| = \max_{i,j} |\sigma_{i,j}(P) - \sigma_{i,j}(Q)| .$$

Assume $d(P_m, \bar{P}) \rightarrow 0$ and $d(Q_n, \bar{P}) \rightarrow 0$. Let $\bar{\Sigma} = \frac{p}{1-p}\Sigma_P + \Sigma_Q$.

Then, the distribution $L_{m,n}(P_m, Q_n)$ of $T_{m,n}$ defined in (??) under P_m and Q_n converges weakly to $L(\bar{P}, \bar{P})$, where $L(\bar{P}, \bar{P})$ is the multivariate normal distribution with mean zero and covariance matrix $\bar{\Sigma}$.

PROOF OF LEMMA ?? : The proof is implicit in the proof of Theorem 3.7 of Romano and Shaikh (2012), but we include a sketch (which is actually simpler as we are not considering a studentized root here). Consider any sequence P_m and Q_n satisfying $d(P_m, \bar{P}) \rightarrow 0$ and $d(Q_n, \bar{P}) \rightarrow 0$. It is easy to see that convergence of distribution and second moments implies convergence of first moments (see Problem 15.18 in Lehmann and Romano (2005)), so that $\mu(P_m) \rightarrow \mu(\bar{P})$ and $\mu(Q_n) \rightarrow \mu(\bar{P})$, where $\mu(\bar{P}) = p\mu(P) +$

$(1-p)\mu(Q)$ is the mean vector of $\bar{P}$. To show that $T_{m,n}$ defined in (4) converges as stated, it suffices to apply the Cramér-Wold device, which reduces the problem to convergence of a normalized average to a univariate normal distribution. But then, one can simply apply a triangular array result for univariate means under conditions of weak convergence and convergence of second moments, as given in Theorem 15.4.3(i) of Lehmann and Romano (2005). ■

Lemma A.8. *Assume the same setup and conditions of Lemma ???. Consider the distribution $J_{m,n}(P_m, Q_n)$ of $\tilde{M}'_{m,n}$ under P_m and Q_n defined in (??). Further assume that $\bar{\Sigma}$ contains at least one nonzero component. Then,*

$$J_{m,n}(P_m, Q_n) \xrightarrow{d} J(\bar{P}, \bar{P}) ,$$

where $J(\bar{P}, \bar{P})$ is the distribution of $\max |F|$ when F is the multivariate normal distribution with mean zero and covariance matrix $\tilde{\Sigma}$ defined in (??).

PROOF OF LEMMA ???: By Lemma ??, the Slutsky's Theorem, and the Continuous Mapping Theorem, it suffices to show $\hat{\Sigma}$ converges in probability to $\bar{\Sigma}$. But this follows from the Law of Large Numbers applied to each component. ■

B Proofs

PROOF OF THEOREM ???: Put all the $N = m + n$ observations and write

$$Z^N = (Z_1, \dots, Z_N)' = (X_1, \dots, X_m, Y_1, \dots, Y_n)' = \begin{pmatrix} X_{1,1} & \cdots & X_{1,d} \\ \vdots & \ddots & \vdots \\ X_{m,1} & \cdots & X_{m,d} \\ Y_{1,1} & \cdots & Y_{1,d} \\ \vdots & \ddots & \vdots \\ Y_{n,1} & \cdots & Y_{n,d} \end{pmatrix} .$$

Independent of the Z s, let $(\pi(1), \dots, \pi(N))$ and $(\pi'(1), \dots, \pi'(N))$ be independent random permutations of $\{1, \dots, N\}$. By Lemma ??, it suffices to show

$$(T_{m,n}(Z_\pi), T_{m,n}(Z_{\pi'})) \xrightarrow{d} (T, T') , \quad (43)$$

where T and T' are independent d -vectors each having the multivariate normal distribution with 0 and covariance matrix $\bar{\Sigma} = \frac{p}{1-p}\Sigma_P + \Sigma_Q$. However, by the Cramér-Wold de-

vice, a sufficient condition for (??) is that, for any choice of constants $t_1 = (t_{1,1}, \dots, t_{1,d})'$ and $t_2 = (t_{2,1}, \dots, t_{2,d})' \in \mathbf{R}^d$,

$$m^{-1/2} \left[\left(\sum_{i=1}^m \sum_{k=1}^d t_{1,k} X_{\pi(i),k} - \frac{m}{n} \sum_{j=1}^n \sum_{k=1}^d t_{1,k} Y_{\pi(m+j),k} \right) + \left(\sum_{i=1}^m \sum_{k=1}^d t_{2,k} X_{\pi'(i),k} - \frac{m}{n} \sum_{j=1}^n \sum_{k=1}^d t_{2,k} Y_{\pi'(m+j),k} \right) \right] \xrightarrow{d} N(0, t_1' \bar{\Sigma} t_1 + t_2' \bar{\Sigma} t_2) . \quad (44)$$

Let $W_i = \begin{cases} 1 & \text{if } \pi(i) \leq m \\ -m/n & \text{if } \pi(i) > m \end{cases}$ and W'_i is defined with π replaced by π' . Then, conditioning on the W_i and W'_i , the left side of (??) can be rewritten as

$$\begin{aligned} & m^{-1/2} \left[\sum_{i=1}^N \left(\sum_{k=1}^d t_{1,k} Z_{i,k} W_i + \sum_{k=1}^d t_{2,k} Z_{i,k} W'_i \right) \right] \\ &= m^{-1/2} \sum_{i=1}^m \left(\sum_{k=1}^d (t_{1,k} X_{k,i} W_i + t_{2,k} X_{k,i} W'_i) \right) + m^{-1/2} \sum_{j=1}^n \left(\sum_{k=1}^d (t_{1,k} Y_{k,j} W_{m+j} + t_{2,k} Y_{k,j} W'_{m+j}) \right) \\ &= m^{-1/2} \sum_{i=1}^m \tilde{X}_i + m^{-1/2} \sum_{j=1}^n \tilde{Y}_j , \end{aligned} \quad (45)$$

where $\tilde{X}_i \equiv t_1' X_i W_i + t_2' X_i W'_i$ and $\tilde{Y}_j \equiv t_1' Y_j W_{m+j} + t_2' Y_j W'_{m+j}$. Note that conditional on W and W' , (??) is an independent sum of a linear combination of independent variables. Define

$$\sigma_i^2 = (\tilde{X}_i | W_i, W'_i) = t_1' \Sigma_P t_1 W_i^2 + t_2' \Sigma_P t_2 W_i'^2 + 2t_1' \Sigma_P t_2 W_i W'_i$$

and let $s_m^2 = \sum_{i=1}^m \sigma_i^2$. If we can show that, for any subsequence m_j , there exists a further subsequence m_{j_k} such that the random variables $\tilde{X}_1, \dots, \tilde{X}_{m_{j_k}}$ under the conditional distribution given $W_{m_{j_k}}$ and $W'_{m_{j_k}}$ satisfy the Lindeberg condition with probability one, so that conditionally on $W_{m_{j_k}}$ and $W'_{m_{j_k}}$,

$$m_{j_k}^{-1/2} \sum_{i=1}^{m_{j_k}} \tilde{X}_i / s_{m_{j_k}} \xrightarrow{d} N(0, 1) \quad \text{with probability one}$$

then, unconditionally,

$$m^{-1/2} \sum_{i=1}^m \tilde{X}_i \xrightarrow{d} N \left(0, \frac{p}{1-p} (t_1' \Sigma_P t_1 + t_2' \Sigma_P t_2) \right) , \quad (46)$$

as

$$\frac{1}{m_{j_k}} s_{m_{j_k}}^2 \xrightarrow{P} \frac{p}{1-p} (t'_1 \Sigma_P t_1 + t'_2 \Sigma_P t_2) .$$

To verify the Lindeberg condition, observe that for each $\varepsilon > 0$, the (conditional) Lindeberg condition becomes

$$\begin{aligned} & \sum_{i=1}^{m_{j_k}} \frac{1}{s_{m_{j_k}}^2} \mathbb{E}[\tilde{X}_i^2 \mathbb{I}\{|\tilde{X}_i| > \varepsilon s_{m_{j_k}}\} | W_{m_{j_k}}, W'_{m_{j_k}}] \\ & \leq \frac{1}{s_{m_{j_k}}^2} m_{j_k} \max_{i=1, \dots, m_{j_k}} \sigma_i^2 \mathbb{E} \left[\frac{\tilde{X}_i^2}{\sigma_i^2} \mathbb{I} \left\{ \frac{\tilde{X}_i^2}{\sigma_i^2} > \frac{\varepsilon^2 s_{m_{j_k}}^2}{\sigma_i^2} \right\} \middle| W_{m_{j_k}}, W'_{m_{j_k}} \right] \\ & \leq \left(\frac{1}{s_{m_{j_k}}^2} m_{j_k} \max_{i=1, \dots, m_{j_k}} \sigma_i^2 \right) \sup_i \mathbb{E} \left[\frac{\tilde{X}_i^2}{\sigma_i^2} \mathbb{I} \left\{ \frac{\tilde{X}_i^2}{\sigma_i^2} > \frac{\varepsilon^2 s_{m_{j_k}}^2}{\max_{i=1, \dots, m_{j_k}} \sigma_i^2} \right\} \middle| W_{m_{j_k}}, W'_{m_{j_k}} \right] . \quad (47) \end{aligned}$$

Thus, in order to show $(??) \rightarrow 0$ with probability one, it suffices to show, conditional on W and W' ,

$$\left(\frac{\tilde{X}_i}{\sigma_i} \right)^2 \text{ is uniformly integrable ,} \quad (48)$$

and

$$m_{j_k} \frac{\max_i \sigma_i^2}{\sum_{i=1}^{m_{j_k}} \sigma_i^2} = \frac{\max_i \sigma_i^2}{\frac{1}{m_{j_k}} \sum_{i=1}^{m_{j_k}} \sigma_i^2} < \infty \quad (49)$$

as $m \rightarrow \infty$. The condition $(??)$ follows by the assumption $(??)$. Certainly,

$$\begin{aligned} & \max_{i=1, \dots, m} \sigma_i^2 = \max_{i=1, \dots, m} \left(t'_1 \Sigma_P t_1 W_i^2 + t'_2 \Sigma_P t_2 W_i'^2 + 2t'_1 \Sigma_P t_2 W_i W_i' \right) \\ & \leq t'_1 \Sigma_P t_1 \max_i W_i^2 + t'_2 \Sigma_P t_2 \max_i W_i'^2 + 2t'_1 \Sigma_P t_2 \max_i W_i W_i' = t'_1 \Sigma_P t_1 + t'_2 \Sigma_P t_2 + 2t'_1 \Sigma_P t_2 < \infty . \end{aligned}$$

Furthermore, note that

$$\begin{aligned} & \mathbb{E}(W_i) = \mathbb{E}(W_i') = 0 , \\ & \mathbb{E}(W_i^2) = \mathbb{E}(W_i'^2) = \frac{m}{n} \rightarrow \frac{p}{1-p} , \quad \mathbb{E}(W_i W_i') = \mathbb{E}(W_i W_i') = 0 , \\ & \mathbb{E}(W_i^4) = \mathbb{E}(W_i'^4) = \frac{m}{N} + \frac{m^4}{n^4} \frac{n}{N} \rightarrow p + \frac{p^4}{(1-p)^3} . \end{aligned}$$

Also, note further that for $i \neq j$,

$$W_i W_j = \begin{cases} 1 & \text{with probability } \frac{m(m-1)}{N(N-1)}, \\ -\frac{m}{n} & \text{with probability } 2\frac{mn}{N(N-1)}, \\ \frac{m^2}{n^2} & \text{with probability } \frac{n(n-1)}{N(N-1)}, \end{cases}$$

and similarly,

$$W_i^2 W_j^2 = \begin{cases} 1 & \text{with probability } \frac{m(m-1)}{N(N-1)}, \\ \frac{m^2}{n^2} & \text{with probability } 2\frac{mn}{N(N-1)}, \\ \frac{m^4}{n^4} & \text{with probability } \frac{n(n-1)}{N(N-1)}. \end{cases}$$

Hence, for $i \neq j$,

$$E(W_i W_j) = \frac{1}{N(N-1)} \left[m(m-1) - 2\frac{m}{n}mn + \frac{m^2}{n^2}n(n-1) \right] = -\frac{m}{n(N-1)} \rightarrow 0,$$

and

$$\begin{aligned} E(W_i^2 W_j^2) &= \frac{1}{N(N-1)} \left[m(m-1) + 2\frac{m^2}{n^2}mn + \frac{m^4}{n^4}n(n-1) \right] \\ &= \frac{m}{n^3 N(N-1)} [nm(m+n)^2 - (n^3 - m^3)] \\ &= \frac{N}{N-1} \frac{m^2}{n^2} - \frac{m(n^2 - mn + m^2)}{n^3(N-1)} \rightarrow \frac{p^2}{(1-p)^2}, \end{aligned}$$

which implies

$$(W_i^2, W_j^2) \rightarrow 0.$$

Based on these facts, it can be readily shown that

$$E(\sigma_i^2) = \frac{m}{n} (t'_1 \Sigma_P t_1 + t'_2 \Sigma_P t_2) \rightarrow \frac{p}{1-p} (t'_1 \Sigma_P t_1 + t'_2 \Sigma_P t_2),$$

$$E(\sigma_i^4) \rightarrow \left(p + \frac{p^4}{(1-p)^3} \right) ((t'_1 \Sigma_P t_1)^2 + (t'_2 \Sigma_P t_2)^2) + 2\frac{p^2}{(1-p)^2} ((t'_1 \Sigma_P t_1)(t'_2 \Sigma_P t_2) + 2(t_1 \Sigma_P t_2)^2),$$

and for $i \neq j$,

$$E(\sigma_i^2 \sigma_j^2) \rightarrow \frac{p^2}{(1-p)^2} (t'_1 \Sigma_P t_1 + t'_2 \Sigma_P t_2)^2.$$

Therefore, we now have

$$E\left(\frac{1}{m} \sum_{i=1}^m \sigma_i^2\right) = \frac{m}{n} (t'_1 \Sigma_P t_1 + t'_2 \Sigma_P t_2) \rightarrow \frac{p}{1-p} (t'_1 \Sigma_P t_1 + t'_2 \Sigma_P t_2)$$

and

$$\begin{aligned}
\left(\frac{1}{m} \sum_{i=1}^m \sigma_i^2 \right) &= \frac{1}{m^2} \mathbb{E} \left(\sum_{i=1}^m \sigma_i^2 \right)^2 - \frac{m^2}{n^2} (t'_1 \Sigma_P t_1 + t'_2 \Sigma_P t_2)^2 \\
&= \frac{1}{m^2} \left[\mathbb{E} \left(\sum_{i=1}^m \sigma_i^4 \right) + \sum_{i \neq j} \mathbb{E} (\sigma_i^2 \sigma_j^2) \right] - \frac{m^2}{n^2} (t'_1 \Sigma_P t_1 + t'_2 \Sigma_P t_2)^2 \\
&\rightarrow \frac{p^2}{(1-p)^2} (t'_1 \Sigma_P t_1 + t'_2 \Sigma_P t_2)^2 - \frac{p^2}{(1-p)^2} (t'_1 \Sigma_P t_1 + t'_2 \Sigma_P t_2)^2 = 0
\end{aligned}$$

implying

$$\left(\frac{1}{m} \sum_{i=1}^m \sigma_i^2 \right) \xrightarrow{P} \frac{p}{1-p} (t'_1 \Sigma_P t_1 + t'_2 \Sigma_P t_2) .$$

Consequently, the condition (??) holds, and thus, along the subsequence m_{j_k} , (??) converges to zero with probability one, implying (??). Using a similar argument, the limit of the second term in (??) can be readily shown to be $N(0, t'_1 \Sigma_Q t_1 + t'_2 \Sigma_Q t_2)$ and thus, by the multivariate Polya's theorem (Chandra, 1988), the result follows immediately. $\blacksquare$

PROOF OF THEOREM ?? : Write $\hat{\Sigma} = \hat{\Sigma}(Z_1, \dots, Z_N)$ and let $(\pi(1), \dots, \pi(N))$ denote a random permutation of $\{1, \dots, N\}$. We first will show that

$$\hat{\Sigma} (Z_{\pi(1)}, \dots, Z_{\pi(N)}) \xrightarrow{P} \bar{\Sigma} ,$$

where

$$\bar{\Sigma} = \frac{p}{1-p} \Sigma_P + \Sigma_Q .$$

To do this, it suffices to show that

$$\hat{\Sigma}_P (Z_{\pi(1)}, \dots, Z_{\pi(m)}) \xrightarrow{P} p \Sigma_P + (1-p) \Sigma_Q \tag{50}$$

and

$$\hat{\Sigma}_Q (Z_{\pi(m+1)}, \dots, Z_{\pi(N)}) \xrightarrow{P} p \Sigma_P + (1-p) \Sigma_Q . \tag{51}$$

However, since we know that, for i.i.d. observations $\bar{Z}_1, \dots, \bar{Z}_N$ from the mixture distribution $\bar{P}$,

$$\hat{\Sigma}_P (\bar{Z}_1, \dots, \bar{Z}_m) \xrightarrow{P} p \Sigma_P + (1-p) \Sigma_Q \quad \text{and} \quad \hat{\Sigma}_Q (\bar{Z}_1, \dots, \bar{Z}_m) \xrightarrow{P} p \Sigma_P + (1-p) \Sigma_Q ,$$

by the contiguity results between multinomial and multivariate hypergeometric distri-

butions (see Subsection 5.4 of Chung and Romano (2013)), both (??) and (??) hold. Thus, we can use Theorem ?? and apply Lemma ?? to conclude that the permutation distribution of the studentized test statistic $S_{m,n}$ behaves as in the stated result. ■

PROOF OF THEOREM ?? AND THEOREM ??: Both results follow from the continuous mapping theorem for the randomization distribution in multivariate cases given in Lemma ??. ■

PROOF OF THEOREM ??: As in the proof of Theorem ??, we have already shown (??) and (??) hold. Thus, we can use Theorem ?? together with Lemma ?? and Lemma ?? to conclude that the permutation distribution of the test statistic $\tilde{M}_{m,n}$ behaves as in the stated result. ■

To investigate the permutation distribution of the prepivoted statistics, we shall define an appropriate metric on the space of probabilities. For probability distributions $P, Q \in \mathbf{R}^d$ with finite covariance matrices Σ_P and Σ_Q , let the metric $d(P, Q)$ between P and Q be defined as

$$d(P, Q) = \max \left(\sup_{t \in \mathbf{R}^d} |F_P(t) - F_Q(t)|, \|\Sigma_P - \Sigma_Q\| \right), \quad (52)$$

where $F_P(\cdot)$ and $F_Q(\cdot)$ denote the corresponding c.d.f.s of the probability distribution P and Q , respectively, and $\|\Sigma_P - \Sigma_Q\| = \max_{i,j} |\sigma_{i,j}(P) - \sigma_{i,j}(Q)|$.

PROOF OF THEOREM ??: By definition, the permutation distribution $\hat{R}_{m,n}^J$ of $J_{m,n}(\tilde{M}_{m,n}, \hat{P}_m, \hat{Q}_n)$ is the empirical distribution of $J_{m,n}(\tilde{M}_{m,n}(\pi(i)), \hat{P}_m(\pi(i)), \hat{Q}_n(\pi(i)))$, i.e.,

$$\hat{R}_{m,n}^J(t) = \frac{1}{N!} \sum_{\pi(i) \in \mathbf{G}_N} \mathbf{I} \left\{ J_{m,n} \left(\tilde{M}_{m,n}(\pi(i)), \hat{P}_m(\pi(i)), \hat{Q}_n(\pi(i)) \right) \leq t \right\}.$$

Fix $\delta > 0$ and divide the permutation distribution into two parts where $i \in \hat{I} \equiv \left\{ i : d(\hat{P}_m(\pi(i)), \bar{P}) \leq \delta, d(\hat{Q}_n(\pi(i)), \bar{P}) \leq \delta \right\}$ and $i \in \hat{I}^c$. Thus, the permutation distribution $\hat{R}_{m,n}^J(t)$ can be rewritten as

$$\begin{aligned} \hat{R}_{m,n}^J(t) &= \frac{1}{N!} \sum_{i \in \hat{I}} \mathbf{I} \left\{ J_{m,n} \left(\tilde{M}_{m,n}(\pi(i)), \hat{P}_m(\pi(i)), \hat{Q}_n(\pi(i)) \right) \leq t \right\} \\ &\quad + \frac{1}{N!} \sum_{i \in \hat{I}^c} \mathbf{I} \left\{ J_{m,n} \left(\tilde{M}_{m,n}(\pi(i)), \hat{P}_m(\pi(i)), \hat{Q}_n(\pi(i)) \right) \leq t \right\}. \end{aligned}$$

We shall first show that $\frac{1}{N!}|\hat{I}| \xrightarrow{P} 1$, where $|\hat{I}|$ denotes the cardinality of $\hat{I}$. It suffices to show

$$\frac{1}{N!} \sum_i \mathbb{I} \left\{ d(\hat{P}_m(\pi(i)), \bar{P}) \leq \delta \right\} \xrightarrow{P} 1. \quad (53)$$

and similarly for $d(\hat{Q}_n(\pi(i)), \bar{P})$. To show (??), it is sufficient to show that

$$\frac{1}{N!} \sum_i \mathbb{P} \left\{ d(\hat{P}_m(\pi(i)), \bar{P}) \leq \delta \right\} \rightarrow 1 ,$$

or equivalently

$$W_n(Z_{\pi(1)}, \dots, Z_{\pi(m)}) \equiv \mathbb{P} \left\{ d(\hat{P}_m(\Pi), \bar{P}) \leq \delta \right\} \rightarrow 1 . \quad (54)$$

However, by the contiguity results in Subsection 5.4 of Chung and Romano (2013), if, for $V_1, \dots, V_m$ i.i.d $\bar{P}$, one can show

$$W_n(V_1, \dots, V_m) \equiv P \left\{ \max \left(\sup_{t \in \mathbf{R}^d} |F_{\hat{P}_m}(t) - F_{\bar{P}}(t)|, ||\Sigma_{\hat{P}_m} - \bar{\Sigma}|| \right) \leq \delta \right\} \rightarrow 1 , \quad (55)$$

then (??) is satisfied. For the first component in (??), for any δ ,

$$\mathbb{P} \left(\sup_{t \in \mathbf{R}^d} |F_{\hat{P}_m(V_1, \dots, V_m)}(t) - F_{\bar{P}}(t)| \leq \delta \right) \rightarrow 1$$

by the Glivenko-Cantelli Theorem. Also, by the Strong Law of Large Numbers,

$$\Sigma_{\hat{P}_m(V_1, \dots, V_m)} \xrightarrow{P} \bar{\Sigma} \quad \text{with probability one.}$$

Thus, it follows that (??) holds and similarly, it can be shown that

$$\mathbb{P} \left\{ d(\hat{Q}_n(\Pi), \bar{P}) \leq \delta \right\} \rightarrow 1.$$

Knowing that $\frac{1}{N!}|\hat{I}^c| \xrightarrow{P} 0$, we now have

$$\hat{R}_{m,n}^J(t) = \frac{1}{N!} \sum_{i \in \hat{I}} \mathbb{I} \left\{ J_{m,n} \left(\tilde{M}_{m,n}(\pi(i)), \hat{P}_m(\pi(i)), \hat{Q}_n(\pi(i)) \right) \leq t \right\} + o_P(1) .$$

For any $\varepsilon > 0$, it follows by Lemma ?? that the first term on the right hand side is

bounded as follows:

$$\begin{aligned} \frac{1}{N!} \sum_{i \in \hat{I}} \mathbb{I} \left\{ J \left(\tilde{M}_{m,n}(\pi(i)), \bar{P}, \bar{P} \right) \leq t - \varepsilon \right\} &\leq \frac{1}{N!} \sum_{i \in \hat{I}} \mathbb{I} \left\{ J_{m,n} \left(\tilde{M}_{m,n}(\pi(i)), \hat{P}_m(\pi(i)), \hat{Q}_n(\pi(i)) \right) \leq t \right\} \\ &\leq \frac{1}{N!} \sum_{i \in \hat{I}} \mathbb{I} \left\{ J \left(\tilde{M}_{m,n}(\pi(i)), \bar{P}, \bar{P} \right) \leq t + \varepsilon \right\} \quad \text{with probability tending to one.} \end{aligned}$$

Note that we know from Theorem ?? that the permutation distribution of $\tilde{M}_{m,n}$ converges in probability to $F'(\cdot) = J(\cdot, \bar{P}, \bar{P})$, which is continuous and strictly increasing at $J^{-1}(\cdot, \bar{P}, \bar{P})$. Applying the continuous mapping theorem, we obtain that

$$\frac{1}{N!} \sum_{i \in \hat{I}} \mathbb{I} \left\{ J \left(\tilde{M}_{m,n}(\pi(i)), \bar{P}, \bar{P} \right) \leq t - \varepsilon \right\} \xrightarrow{P} t - \varepsilon$$

and similarly

$$\frac{1}{N!} \sum_{i \in \hat{I}} \mathbb{I} \left\{ J \left(\tilde{M}_{m,n}(\pi(i)), \bar{P}, \bar{P} \right) \leq t + \varepsilon \right\} \xrightarrow{P} t + \varepsilon ,$$

implying that for any $\varepsilon > 0$,

$$t - \varepsilon \leq \frac{1}{N!} \sum_{i \in \hat{I}} \mathbb{I} \left\{ J_{m,n} \left(\tilde{M}_{m,n}(\pi(i)), \hat{P}_m(\pi(i)), \hat{Q}_n(\pi(i)) \right) \leq t \right\} \leq t + \varepsilon .$$

Since $\varepsilon > 0$ was arbitrary, the result (??) is proved. ■

PROOF OF THEOREM ?? : Put all the $N = m + n$ observations together and write

$$Z^N = (Z_1, \dots, Z_N)' = (X_1, \dots, X_m, Y_1, \dots, Y_n)' = \begin{pmatrix} X_{1,1} & \cdots & X_{1,d} \\ \vdots & \ddots & \vdots \\ X_{m,1} & \cdots & X_{m,d} \\ Y_{1,1} & \cdots & Y_{1,d} \\ \vdots & \ddots & \vdots \\ Y_{n,1} & \cdots & Y_{n,d} \end{pmatrix} .$$

Let $V_1, \dots, V_N$ be i.i.d. $\bar{P}$. Then, by assumption,

$$m^{1/2} \left[\hat{\theta}_m(V_1, \dots, V_m) - \theta(\bar{P}) \right] - m^{-1/2} \sum_{i=1}^m f_{\bar{P}}(V_i) \xrightarrow{P} 0 ,$$

where $f_{\bar{P}}(\cdot) = (f_{\bar{P},1}(\cdot), \dots, f_{\bar{P},d}(\cdot))'$. Using this fact after applying the contiguity result

from Lemma 5.3 of Chung and Romano (2013) element by element, we now have, for a permutation π of $\{1, \dots, N\}$,

$$\varepsilon_m(Z_{\pi(1)}, \dots, Z_{\pi(m)}) \equiv m^{1/2} \left[\hat{\theta}_m(Z_{\pi(1)}, \dots, Z_{\pi(m)}) - \theta(\bar{P}) \right] - m^{-1/2} \sum_{i=1}^m f_{\bar{P}}(Z_{\pi(i)}) \xrightarrow{P} 0 \quad (56)$$

and

$$\varepsilon_n(Z_{\pi(m+1)}, \dots, Z_{\pi(N)}) \equiv n^{1/2} \left[\hat{\theta}_n(Z_{\pi(m+1)}, \dots, Z_{\pi(N)}) - \theta(\bar{P}) \right] - n^{-1/2} \sum_{j=1}^n f_{\bar{P}}(Z_{\pi(m+j)}) \xrightarrow{P} 0. \quad (57)$$

Thus, we can write

$$\begin{aligned} W_{m,n}(Z_{\pi(1), \dots, \pi(N)}) &= m^{1/2} \left[\hat{\theta}_m(Z_{\pi(1)}, \dots, Z_{\pi(m)}) - \hat{\theta}_n(Z_{\pi(m+1)}, \dots, Z_{\pi(N)}) \right] \\ &= m^{1/2} \left[\frac{1}{m} \sum_{i=1}^m f_{\bar{P}}(Z_{\pi(i)}) - \frac{1}{n} \sum_{j=1}^n f_{\bar{P}}(Z_{\pi(m+j)}) \right] \\ &\quad + \varepsilon_m(Z_{\pi(1)}, \dots, Z_{\pi(m)}) + \left(\frac{m}{n} \right)^{1/2} \varepsilon_n(Z_{\pi(m+1)}, \dots, Z_{\pi(N)}) . \end{aligned}$$

Note that the last two terms converge in probability to zero by (??) and (??). Therefore, we can apply Slutsky's Theorem for multivariate randomization distributions Lemma ??; that is, it suffices to determine the limit behavior of

$$m^{1/2} \left[\frac{1}{m} \sum_{i=1}^m f_{\bar{P}}(Z_{\pi(i)}) - \frac{1}{n} \sum_{j=1}^n f_{\bar{P}}(Z_{\pi(m+j)}) \right]. \quad (58)$$

Independent of the Z s, let $(\pi(1), \dots, \pi(N))$ and $(\pi'(1), \dots, \pi'(N))$ be independent random permutations of $\{1, \dots, N\}$. By Lemma ?? together with (??), it suffices to show

$$\begin{aligned} &\left(m^{-1/2} \left[\sum_{i=1}^m f_{\bar{P}}(Z_{\pi(i)}) - \frac{m}{n} \sum_{j=1}^n f_{\bar{P}}(Z_{\pi(m+j)}) \right], m^{-1/2} \left[\sum_{i=1}^m f_{\bar{P}}(Z_{\pi'(i)}) - \frac{m}{n} \sum_{j=1}^n f_{\bar{P}}(Z_{\pi'(m+j)}) \right] \right) \\ &\quad \xrightarrow{d} (T, T') , \end{aligned}$$

where T and T' are independent with each of T and T' being d -vectors that have multivariate normal distributions with 0 and variance $\bar{\Gamma}$. However, this reduces the problem to the mean case in Theorem ??.

PROOF OF THEOREM ??: Write $\hat{\Gamma} = \hat{\Gamma}(Z_1, \dots, Z_N)$ and let $(\pi(1), \dots, \pi(N))$ denote a

random permutation of $\{1, \dots, N\}$. We will show that

$$\hat{\Gamma}(Z_{\pi(1)}, \dots, Z_{\pi(N)}) \xrightarrow{P} \bar{\Gamma},$$

where $\bar{\Gamma}$ is given in (??). To do this, since $m/n \rightarrow p/(1-p)$, it suffices to show that

$$\hat{\Gamma}_P(Z_{\pi(1)}, \dots, Z_{\pi(m)}) \xrightarrow{P} \Gamma_{\bar{P}} \quad (59)$$

and

$$\hat{\Gamma}_Q(Z_{\pi(m+1)}, \dots, Z_{\pi(N)}) \xrightarrow{P} \Gamma_{\bar{P}}. \quad (60)$$

Since $\hat{\Gamma}_P$ is a consistent estimator of $\Gamma_{\bar{P}}$ under the mixture model, (??) follows from contiguity results between multinomial and multivariate hypergeometric distributions (see Subsection 5.4 of Chung and Romano (2013)). Thus, we can use Theorem ?? and Lemma ?? to conclude that the permutation distribution of the test statistic $\tilde{W}_{m,n}$ satisfies the result. ■

References

- Babu G. J. and Rao C. R. (1988) Joint Asymptotic Distribution of Marginal Quantiles and Quantile Functions in Samples from a Multivariate Population. *Journal of Multivariate Analysis* **27**, 15-23.
- Beran, R. (1988a). Balanced Simultaneous Confidence Sets. *Journal of American Statistical Association* **83**, 679–686.
- Beran, R. (1988b). Prepivoting Test Statistics: A Bootstrap View of Asymptotic Refinements. *Journal of American Statistical Association* **83**, 687–697.
- Bickel, P. J., and van Zwet, W. R. (1978). Asymptotic Expansions for the Power of Distributionfree Tests in the Two-Sample Problem. *Annals of Statistics* **6**, 937–1004.
- Chandra, P. T. (1989). Multidimensional Polya’s Theorem. *Bulletin of the Calcutta Mathematical Society* **81**, 227–231.
- Charness G., and Gneezy, U. (2009). Incentives to Exercise. *Econometrica* **77**, 909-931.
- Chung, E., and Romano, P. J. (2013). Exact and Asymptotically Robust Permutation Tests. *Annals of Statistics* **41**, 484-507.
- Chung, E., and Romano, P. J. (2015). Asymptotically Valid and Exact Permutation Tests Based on Two-sample U -Statistics. *Journal of Statistical Planning and Inference*, <http://dx.doi.org/10.1016/j.jspi.2015.07.004>.

- Devroye, L., and Wagner, T.J. (1980). The strong uniform consistency of kernel density estimates. *Multivariate Analysis V* (P.R. Krishnaiah, *ed.*). North Holland, 59–77.
- Dümbgen L., and Zerial P. (2013). On low-dimensional projections of high-dimensional distributions. *IMS Collections* **9**, 91–104.
- Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *Annals of Statistics* **7**, 1–26.
- Hall, P., DiCiccio, T., and Romano, J. (1989). On Smoothing and the Bootstrap. *Annals of Statistics* **17**, 692–704.
- Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *The Annals of Mathematical Statistics* **23**, 169–192.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65–70.
- Janssen, A. (1997). Studentized permutation tests for non-i.i.d. hypotheses and the generalized Behrens-Fisher problem. *Statistics and Probability Letters* **36**, 9–21.
- Janssen, A. (1999). Testing nonparametric statistical functionals with applications to rank tests. *Journal of Statistical Planning and Inference* **81**, 71–93. Erratum (2001) **92**, 297.
- Janssen, A. (2005). Resampling student’s t-type statistics. *Annals of the Institute of Statistical Mathematics* **57**, 507–529.
- Janssen, A. and Pauls, T. (2003). How do bootstrap and permutation tests work? *Annals of Statistics* **31**, 768–806.
- Janssen, A. and Pauls, T. (2005). A Monte Carlo comparison of studentized bootstrap and permutation tests for heteroscedastic two-sample problems. *Computational Statistics* **20**, 369–383.
- Konietschke, F. and Pauly, M. (2012). A studentized permutation test for the nonparametric Behrens-Fisher problem in paired data. *Electronic Journal of Statistics* **6**, 1358–1372.
- Lehmann, E. L. and Romano, J. (2005). *Testing Statistical Hypotheses*. 3rd edition, Springer-Verlag, New York.
- Lix, L. M. and Keselman, H. J. (2004). Multivariate tests of means in independent groups designs effects of covariance heterogeneity and nonnormality. *Evaluation and the health professions* **27**, 45–69.
- Marcus, R., Peritz, E., and Gabriel, K. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655–660.
- Murphy, B. P. (1967). Some two-sample tests when the variances are unequal: a simulation study. *Biometrika* **54**, 679–683.
- Neubert, K. and Brunner, E. (2007). A Studentized permutation test for the nonparametric Behrens-Fisher problem. *Computational Statistics & Data Analysis* **51**,

5192–5204.

- Neuhaus, G. (1993). Conditional rank tests for the two-sample problem under random censorship. *Annals of Statistics* **21**, 1760–1779.
- Omelka, M. and Pauly, M. (2012). Testing equality of correlation coefficients in an potentially unbalanced two-sample problem via permutation methods. *Journal of Statistical Planning and Inference* **142**, 1396–1406.
- Pauly, M. (2010). Discussion about the quality of F-ratio resampling tests for comparing variances. *TEST*, 1–17.
- Pauly, M., Brunner, E., and Konietzschke, F. (2015). Asymptotic Permutation Tests in General Factorial Designs. *Journal of the Royal Statistical Society Series: Series B* **77**, 461–473.
- Politis, D., Romano, J. and Wolf, M. (1999). *Subsampling*. Springer-Verlag, New York.
- Romano, J. (1990). On the behavior of randomization tests without a group invariance assumption. *Journal of the American Statistical Association* **85**, 686–692.
- Romano, J. and Shaikh, A. (2012). On the Uniform Asymptotic Validity of Subsampling and the Bootstrap. *Annals of Statistics* **40**, 2798–2822.
- Romano, J., Shaikh, A. and Wolf, M. (2011). Consonance and the Closure Method in Multiple Testing. *International Journal of Biostatistics* **7**, Article 12.
- Romano, J. and Wolf, M. (2010). Balanced control of generalized error rates. *Annals of Statistics* **38**, 598–633.
- Simes, R. J. (1986) An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 751–754.

ADDRESS:

EunYi Chung: Department of Economics, University of Illinois at Urbana-Champaign, Urbana, IL 61801; eunyi@illinois.edu

Joseph P. Romano: Departments of Statistics and Economics, Stanford University, Stanford, CA 94305-4065; romano@stanford.edu