Distribution-free Tests of Mean Vectors and Covariance Matrices for Multivariate Paired Data

Erning Li, Johan Lim, Kyunga Kim, and Shin-Jae Lee *

Abstract

We study a permutation procedure to test the equality of mean vectors, homogeneity of covariance matrices, or simultaneous equality of both mean vectors and covariance matrices in multivariate paired data. We propose to use two test statistics for the equality of mean vectors and the homogeneity of covariance matrices, respectively, and combine them to test the simultaneous equality of both mean vectors and covariance matrices. Since the combined test has composite null hypothesis, we control its type I error probability and theoretically prove the asymptotic unbiasedness and consistency of the combined test. The new procedure requires no structural assumption on the covariances. No distributional assumption is imposed on the data, except that the permutation test for mean vector equality assumes symmetric joint distribution of the paired data. We illustrate the good performance of the proposed approach with comparison to competing methods via simulations. We apply the proposed method to testing the symmetry of tooth size in a dental study and to finding differentially expressed gene sets with dependent structures in a microarray study of prostate cancer.

Key words: Multivariate paired data; Permutation; Equality of mean vectors; Homogeneity of covariance matrices.

*Erning Li is in the Department of Statistics, Texas A&M University, College Station, TX 77843, USA; Johan Lim is in the Department of Statistics, Seoul National University, Seoul, 151-747, Korea; Kyunga Kim is in the Department of Statistics, Sookmyung Women's University, Korea; Shin-Jae Lee is in the School of Dentistry, Seoul National University, Seoul, Korea. Erning Li is the corresponding author, email: erningli.work@gmail.com, phone: (979)845-3141, fax: (979)845-3144.

1 Introduction

Multivariate paired data (paired multivariate observations from the same group of experimental units) are common in many studies, and it is often of interest to test qualitative properties of population (or treatment) mean vectors and/or covariance matrices. Obviously, multivariate responses within each specific experimental unit are likely to be dependent. The two samples are correlated due to the fact of having paired observations. Moreover, in some cases, data may demonstrate distributions other than multivariate normal. The main contribution of the presented paper is the development of distribution-free tests that do not require stringent assumptions, their supporting theory, and a permutation procedure for implementation that take these considerations into account when testing the equality of mean vectors, equality of covariance matrices, and simultaneous equality of both mean vectors and covariance matrices between multivariate paired samples.

This work is motivated by two studies. One is a study on human tooth size. The data set contains 179 male adults who had natural normal occlusion with age range of 17–24 through a community dental health survey between 1999 and 2002 in Seoul, Korea, and is part of a standard occlusion study that has been undergoing since 1997 (Kim et al., 2005; Lee et al., 2007). A human adult normally has 14 permanent teeth in either maxilla (upper jaw) or mandible (lower jaw) with 7 teeth (central incisor to second molar) on either left or right side. The tooth sizes in terms of mesiodistal diameter of teeth of the patients were measured using digital Vernier calipers with sharpened points. Investigating whether tooth size profile is the same for the left and right sides around central incisors in terms of both mean values and variation would be helpful to understand the developmental biology of teeth and gain important insight into normative data of human tooth size, diagnostic criteria for malocclusion, and dental treatment planning (Garn et al., 2002; Uysal et al., 2005). Notice that the tooth sizes on the left and right sides from the same patient are naturally paired. In statistical aspects, however, previous analysis mainly focused on the asymmetry in mean

vectors between left and right sides, while little has been done to justify the asymmetry between the covariance parameters, letting alone the possible difference between the underlying distributions. Among the subjects in the data set, we found several teeth exhibited excessive sample skewness and/or kurtosis, implying some violations of the usual multivariate normality assumption. We thus study a test procedure to address all these issues.

Another example is to test the asymmetry in either mean vectors or covariance matrices in multivariate paired observations in a microarray study. In cancer microarray experiments, it is a common practice to collect the tumor tissue from the same individual from whom the normal tissue is taken. Such experiments produce multivariate paired observations. The microarray data analysis focuses on finding differentially expressed genes between the control and the treatment groups. In recent years, much efforts have been given to identifying sets of genes that are significantly differentially expressed (e.g., Efron and Tibshirani, 2007; Newton et al., 2007), in which differentially expressed genes are identified by testing the asymmetry (or equivalently, inequality) between the means of normal and tumor tissues. On the other hand, in gene tests, the non-zero entries of the concentration matrix (i.e., the inverse covariance matrix) imply conditional dependence between corresponding genes given the rest of genes (Lauritzen, 2004). Testing the symmetry or equality of the covariance matrices or the concentration matrices can elucidate the changes in the dependent structure (or the regulatory network) among genes due to a treatment.

We consider three notions of asymmetry for multivariate paired data: mean vectors, covariance matrices, and both. To be specific, suppose $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T$ and $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{ip})^T$ are p-variate vectors of observations of variables $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, from the ith subject, $i = 1, \ldots, n$. $\boldsymbol{X}_i$ and $\boldsymbol{Y}_i$ are paired as they come from the same subject. It is commonly assumed that subjects are independent. Let the mean vector and covariance

matrix of the vector $\boldsymbol{Z}_i = (\boldsymbol{X}_i^T, \boldsymbol{Y}_i^T)^T$ be denoted by

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y \end{pmatrix}, \qquad \boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_X, & \boldsymbol{\Omega}_{XY} \\ \boldsymbol{\Omega}_{YX}, & \boldsymbol{\Omega}_Y \end{pmatrix},$$

respectively. Then,

(i) the equality hypothesis of the mean vectors means $\mathcal{H}_0 : \boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$,

(ii) the hypothesis of homogeneity in the covariance matrices implies $\mathcal{H}_0 : \boldsymbol{\Omega}_X = \boldsymbol{\Omega}_Y$, and

(iii) the hypothesis of equality in both is equivalent to $\mathcal{H}_0 : \boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$ and $\boldsymbol{\Omega}_X = \boldsymbol{\Omega}_Y$.

Notice that no particular distributional assumption is imposed on $\boldsymbol{X}$ or $\boldsymbol{Y}$. The test of (iii) is the main theme of this paper since it is the most challenging, although all three notions of asymmetry are of interest and addressed.

The test (i) of mean inequality has been well studied in the literature (Anderson, 2003, Chapter 8). Under the usual multivariate normality assumption, the problem is equivalent to testing the linear hypothesis $\boldsymbol{\mu}_X - \boldsymbol{\mu}_Y = 0$ and the likelihood ratio test (LRT) has been widely used. Moreover, group invariant tests, particularly a permutation invariant test — Hotelling's T^2 statistic (Hotelling, 1931), have received much attention. Other test procedures are also available for the case $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent to each other, that is the case $\boldsymbol{\Omega}_{XY} = \boldsymbol{\Omega}_{YX} = \boldsymbol{0}$; for examples, Aslan and Zech, 2005 and de Leon, 2007. Pesarin (2001) provides a detailed exposition of permutation tests and discusses permutation tests of treatment effects for multivariate paired observations irrespective of the underlying dependence and unknown distributions. Under the assumption that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, the test (ii) of covariance matrix equality has been extensively studied (Szatrowski, 1979; Perlman, 1980; Conover, Johnson and Johnson, 1981; Anderson, 2003, Chapter 10). Under certain structural assumptions, Han (1968), Choi and Wette (1972), and Harris (1985), among others, studied the testing problem (ii). Lim et al. (2010) studied the LRTs for correlated multivariate samples under the multivariate normality assumption and characterized the finite-sample

distribution of the LRT statistic for the hypothesis (ii) to be a function of Wishart random variables which depends on the unknown true overall covariance matrix $\mathbf{\Omega}$, of which for implementation they employed a parametric bootstrap procedure. However, LRTs can be sensitive to non-normality in data (e.g., Olson, 1974). Assuming normality for paired data, Bradley and Blackwood (1989) considered an F-test for the hypothesis (iii) for univariate case based on a regression context between the differences and sums of paired data.

In this paper, we propose a procedure for the three tests that allows the paired samples to be correlated without any structural assumption. The proposed procedure also requires no distributional assumption, except the symmetric paired data joint distribution assumption for the permutation test for hypothesis (i) of mean equality. In particular, we utilize the well-known Hotelling's T^2 statistic, denoted by $\mathbf{T}_1$, for testing the mean vector equality and another statistic, denoted by $\mathbf{T}_2$ and described in the next Section, for testing the covariance matrix equality. We show that the null distributions of $\mathbf{T}_1$ and $\mathbf{T}_2$ are invariant to the permutation between $\boldsymbol{X}_i$ and $\boldsymbol{Y}_i$ and the permutation between $\boldsymbol{X}_i - \boldsymbol{\mu}_X$ and $\boldsymbol{Y}_i - \boldsymbol{\mu}_Y$, respectively, which is independently done for every subject i. A major challenge rests in combining $\mathbf{T}_1$ and $\mathbf{T}_2$ to test the simultaneous equality of both mean vectors and covariance matrices, in other words, to obtain the rejection region of the form $\{\mathbf{T}_1 > c_1 \text{ or } \mathbf{T}_2 > c_2\}$. To resolve the difficulty, we restrict our class of rejection regions to sets with two "tuning" parameters chosen by the investigator and approximate the probability of type I error of the combined test using a permutation procedure. We demonstrate in theory that the combined test is asymptotically unbiased and consistent.

The proposed method, the theoretical results, and the permutation procedure for implementation are described in Section 2. We report empirical size and power of the proposed method and compare them to those of the competing methods via simulations in Section 3. The application of the methods to the aforementioned data examples are provided in Section 4. Section 5 concludes the paper.

2 Proposed Method

Our goal is to develop a procedure to test the three notions of asymmetry, particularly the asymmetry of both mean vectors and covariance matrices. We consider two test statistics for mean vector equality and for covariance matrix equality, and then combine them to test the equality of both.

2.1 Test Statistics

Based on the same notations as in Section 1, the random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ comes from a p-variate distribution with mean vector $\boldsymbol{\mu}_X$ and covariance matrix $\boldsymbol{\Omega}_X$, and the random sample $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ follows another p-variate distribution with mean vector $\boldsymbol{\mu}_Y$ and covariance matrix $\boldsymbol{\Omega}_Y$. The paired sample $\boldsymbol{Z}_i = (\boldsymbol{X}_i^T, \boldsymbol{Y}_i^T)^T$, $i = 1, \ldots, n$, is a random sample assumed to be independent and identically distributed with mean vector $\boldsymbol{\mu} = (\boldsymbol{\mu}_X^T, \boldsymbol{\mu}_Y^T)^T$ and covariance matrix $\boldsymbol{\Omega}$ whose $p \times p$ diagonal matrices are $\boldsymbol{\Omega}_X$ and $\boldsymbol{\Omega}_Y$. The paired samples are dependent in the sense that $\boldsymbol{\Omega}$ has upper $p \times p$ off-diagonal matrix $\boldsymbol{\Omega}_{XY}$ and lower off-diagonal matrix $\boldsymbol{\Omega}_{YX}$. Alternatively, $\boldsymbol{X}_i$ and $\boldsymbol{Y}_i$ can be viewed as following general models with additive errors, i.e., $\boldsymbol{X}_i = \boldsymbol{\mu}_X + \boldsymbol{e}_{Xi}$ and $\boldsymbol{Y}_i = \boldsymbol{\mu}_Y + \boldsymbol{e}_{Yi}$, respectively, where $\boldsymbol{\mu}_X$ and $\boldsymbol{\mu}_Y$ are mean vectors or profiles of $\boldsymbol{X}$ and $\boldsymbol{Y}$, $\boldsymbol{e}_{Xi}$ and $\boldsymbol{e}_{Yi}$ are mean zero error terms with $\text{var}(\boldsymbol{e}_{Xi}) = \boldsymbol{\Omega}_X$, $\text{var}(\boldsymbol{e}_{Yi}) = \boldsymbol{\Omega}_Y$, $\text{cov}(\boldsymbol{e}_{Xi}, \boldsymbol{e}_{Yi}) = \boldsymbol{\Omega}_{XY}$, and $\text{cov}(\boldsymbol{e}_{Yi}, \boldsymbol{e}_{Xi}) = \boldsymbol{\Omega}_{YX}$. Write the sample mean vectors $\overline{\boldsymbol{X}} = n^{-1} \sum_{i=1}^n \boldsymbol{X}_i$, $\overline{\boldsymbol{Y}} = n^{-1} \sum_{i=1}^n \boldsymbol{Y}_i$, and $\overline{\boldsymbol{Z}} = n^{-1} \sum_{i=1}^n \boldsymbol{Z}_i$. Representing the entire paired sample data with the $n \times 2p$ data matrix $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)^T$, we can easily compute the $2p \times 2p$ sample covariance matrix using $S_Z = (n-1)^{-1} \sum_{i=1}^n (\boldsymbol{Z}_i - \overline{\boldsymbol{Z}})(\boldsymbol{Z}_i - \overline{\boldsymbol{Z}})^T = (n-1)^{-1} \left(\boldsymbol{Z} - n^{-1} J_n \boldsymbol{Z}\right)^T \left(\boldsymbol{Z} - n^{-1} J_n \boldsymbol{Z}\right)$, where J_n stands for the $n \times n$ matrix whose all elements are 1. Let S_X and S_Y be the $p \times p$ diagonal matrices and S_{XY} and S_{YX} be the $p \times p$ off-diagonal matrices of S_Z. It is straightforward to see that S_X and S_Y are the sample covariance matrices of X and Y samples, respectively. In a permutation framework, the factor $(n-1)^{-1}$ in S_Z may be omitted.

Hotelling's T^2 test is the most common and popular test for the equality of mean vectors in two sample problems. It is also asymptotically the most powerful invariant test when homoscedastic data are normally distributed. Since we have paired data, we consider the Hotelling's T^2 statistic to test $\mathcal{H}_0 : \boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$ based on the one-sample $\boldsymbol{D}_i = \boldsymbol{X}_i - \boldsymbol{Y}_i$, $i = 1, 2, \ldots, n$. Write $\overline{\boldsymbol{D}} = n^{-1} \sum_{i=1}^{n} \boldsymbol{D}_i$, and $S_D = (n-1)^{-1} \sum_{i=1}^{n} (\boldsymbol{D}_i - \overline{\boldsymbol{D}})(\boldsymbol{D}_i - \overline{\boldsymbol{D}})^T$. The statistic for testing mean equality is defined as

$$\mathbf{T}_1 \;=\; \overline{\boldsymbol{D}}^T S_D^{-1} \overline{\boldsymbol{D}} = \left(\overline{\boldsymbol{X}} - \overline{\boldsymbol{Y}}\right)^T \left(S_X + S_Y - S_{XY} - S_{YX}\right)^{-1} \left(\overline{\boldsymbol{X}} - \overline{\boldsymbol{Y}}\right). \tag{1}$$

For the test of hypothesis (i), we assume that the joint distribution of $(\boldsymbol{X}, \boldsymbol{Y})$ is symmetric around $(\boldsymbol{\mu}_X, \boldsymbol{\mu}_Y)$. Under this assumption, the distribution of $\boldsymbol{D}_i$ (and hence the distribution of $\mathbf{T}_1$) is invariant to the permutation of $\boldsymbol{X}_i$ and $\boldsymbol{Y}_i$ when $\mathcal{H}_0 : \boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$ is true, as shown in Section 2.3. This assumption is not needed for the other two tests. It is easy to see that the distribution of $\mathbf{T}_1$ is invariant with respect to affine transformations $\tilde{\boldsymbol{X}} = A\boldsymbol{X} + b$ and $\tilde{\boldsymbol{Y}} = A\boldsymbol{Y} + b$ for a fixed $p \times p$ matrix A of real constants with a nonzero determinant and a $p \times 1$ vector b of constants (Anderson, 2003). For normal data, $\mathbf{T}_1$ follows an F-distribution.

We straightforwardly test $\mathcal{H}_0 : \boldsymbol{\Omega}_X = \boldsymbol{\Omega}_Y$ by comparing the sample covariance matrices and thus propose the test statistic

$$\mathbf{T}_2 = \big| \log |S_X| - \log |S_Y| \big|, \tag{2}$$

where $|A|$ denotes the determinant of A. The null distribution of $\mathbf{T}_2$ is invariant to the permutation of $\boldsymbol{X}_i - \boldsymbol{\mu}_X$ and $\boldsymbol{Y}_i - \boldsymbol{\mu}_Y$ and an approximate permutation procedure, where population mean vectors are replaced by original sample mean vectors, is proposed in Section 2.3 to compute $\mathbf{T}_2$.

We propose to combine $\mathbf{T}_1$ and $\mathbf{T}_2$ to test for $\mathcal{H}_0 : \boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$ and $\boldsymbol{\Omega}_X = \boldsymbol{\Omega}_Y$, and we consider the Tippett type combing function (see Chung and Fraser, 1958; Hirotsu, 1986, 1998; Pesarin, 2001). Let λ_1 and λ_2 be the p-values of the tests $\mathbf{T}_1$ and $\mathbf{T}_2$, respectively. One challenge is to find the combined rejection region of the two tests: $\{\mathbf{T}_1 > c_1\} \cup \{\mathbf{T}_2 > c_2\}$.

We resolve the difficulty by restricting our class of rejection regions to the sets that satisfy $\mathbb{P}(\mathbf{T}_1 > c_1) = k_1\gamma$ and $\mathbb{P}(\mathbf{T}_1 > c_1) = k_2\gamma$ for fixed k_1 and k_2. The coefficients k_1 and k_2 can be viewed as "tuning" parameters chosen by investigators. To define γ, suppose we observe test statistics $\mathbf{T}_1^o$ and $\mathbf{T}_2^o$, then we combine the p-values of $\mathbf{T}_1$ and $\mathbf{T}_2$ using

$$\gamma = \min\left\{\mathbb{P}(\mathbf{T}_1 \geqslant \mathbf{T}_1^o)/k_1,\ \mathbb{P}(\mathbf{T}_2 \geqslant \mathbf{T}_2^o)/k_2\right\} = \min\left(\lambda_1/k_1,\ \lambda_2/k_2\right).$$

We reject $\mathcal{H}_0 : \boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$ and $\boldsymbol{\Omega}_X = \boldsymbol{\Omega}_Y$ if $\gamma \leqslant \delta$. The problem thus becomes finding suitable δ such that the significance level, say α, can be attained.

When data are normally distributed and $\mathcal{H}_0 : \boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$ is true, testing statistics $\mathbf{T}_1$ and $\mathbf{T}_2$ are independent of each other, since $\mathbf{T}_1$ is essentially the multivariate one-sample test statistic that follows an F-distribution. Thus the p-values $\lambda_1 = \mathbb{P}(\mathbf{T}_1 \geqslant \mathbf{T}_1^o)$ and $\lambda_2 = \mathbb{P}(\mathbf{T}_2 \geqslant \mathbf{T}_2^o)$ have independent uniform distributions over the unit interval $[0, 1]$ under the null hypothesis $\mathcal{H}_0 : \boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$ and $\boldsymbol{\Omega}_X = \boldsymbol{\Omega}_Y$ for the combined test and under normality. However, data may not always be normal in reality, and when this happens, λ_1 and λ_2 are uniformly distributed but no longer independent of each other under the $\mathcal{H}_0$. Thus,

$$\begin{aligned}\mathbb{P}(\gamma \leqslant \delta) &= \mathbb{P}(\lambda_1 \leqslant k_1\delta \text{ or } \lambda_2 \leqslant k_2\delta) = \mathbb{P}(\lambda_1 \leqslant k_1\delta) + \mathbb{P}(\lambda_2 \leqslant k_2\delta) - \mathbb{P}(\lambda_1 \leqslant k_1\delta, \lambda_2 \leqslant k_2\delta) \\ &= (k_1 + k_2)\delta - \tau k_1 k_2 \delta^2,\end{aligned}$$

where $\tau = \mathbb{P}(\lambda_1 \leqslant k_1\delta, \lambda_2 \leqslant k_2\delta)\big/\{\mathbb{P}(\lambda_1 \leqslant k_1\delta) \cdot \mathbb{P}\big(\lambda_2 \leqslant k_2\delta)\}$. Obviously, $\tau = 1$ whenλ_1 and λ_2 are independent. If $\mathbb{P}\big(\lambda_1 \leqslant k_1\delta) = 0$ or $\mathbb{P}\big(\lambda_2 \leqslant k_2\delta) = 0$, $\tau = 0$. Hence, to achieve the significance level α, the choice of δ is

$$\delta = \left[k_1 + k_2 - \left\{(k_1 + k_2)^2 - 4\alpha\tau k_1 k_2\right\}^{1/2}\right] \Big/ (2\tau k_1 k_2).$$

In Section 2.3, we propose a permutation procedure to estimate δ and τ. The null hypothesis $\mathcal{H}_0 : \boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$ and $\boldsymbol{\Omega}_X = \boldsymbol{\Omega}_Y$ is rejected if $\gamma \leqslant \delta$, or equivalently, if the corresponding p-value is less than or equal to a pre-specified significance level of the test α, i.e., $(k_1+k_2)\gamma - \tau k_1 k_2 \gamma^2 \leqslant \alpha$. We denote this proposed combined test by $\mathbf{CT}(k_1, k_2)$.

We choose $(k_1, k_2) = (1, 1)$ to test the simultaneous equalities of both mean vectors and covariance matrices, so we obtain $\gamma = \min(\lambda_1, \lambda_2)$ and the p-value of the test $2\gamma - \gamma^2\tau$. In addition, more general setup of the hypothesis (iii) can be accommodated in the proposed combined test framework and other values of (k_1, k_2) may also be selected based on the hypothesis of interest. For example, one may specify a larger value for k_1 than k_2 if the hypothesis (iii) is specified such that a more stringent criterion is imposed on the mean equality than the sameness of covariance matrices. Furthermore, $\mathbf{T}_1$ and $\mathbf{T}_2$ may be individually included in the combined framework as two extreme cases. We can choose $(k_1, k_2) = (1, 0)$ to test the equality of mean vectors; in this case both γ and the p-value equal to λ_1, and the test is essentially $\mathbf{T}_1$. Similarly, we can set $(k_1, k_2) = (0, 1)$ to test the homogeneity of covariance matrices; then both γ and the p-value are λ_2, and the test is the same as $\mathbf{T}_2$. Obviously, neither $\mathbf{T}_1$ nor $\mathbf{T}_2$ has existing δ.

2.2 Unbiasedness and Consistency

In this section, we aim to show theoretical properties, the asymptotic unbiasedness and the consistency, of the proposed combined test **CT**. The theoretical derivation partially relies on the Central Limit Theorem of sample mean vector and sample covariance matrix, which requires the existence of the fourth moment of $\boldsymbol{Z}_i = (\boldsymbol{X}_i^T, \boldsymbol{Y}_i^T)^T$. Let θ be the parameters of the model, which include $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$. Let $\phi_n(\boldsymbol{Z}_n) : \mathcal{R}^{n\times(2p)} \to \{0, 1\}$ be the proposed combined test function for testing $\mathcal{H}_0 : \boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$ and $\boldsymbol{\Omega}_X = \boldsymbol{\Omega}_Y$ against its alternative hypothesis $\mathcal{H}_1$ that the null hypothesis is not true; that is, $\phi_n(\boldsymbol{Z}_n)$ is 1 if $\boldsymbol{Z}_n$ is in the rejection region, or equivalently the observed p-value is less than or equal to α, and 0 otherwise. We prove the following two properties of the test $\phi_n(\boldsymbol{Z}_n)$

- asymptotic unbiasedness: $\sup_{\theta\in\mathcal{H}_0} \mathbb{E}_\theta \{\phi_n(\boldsymbol{Z}_n)\} \leqslant \alpha \leqslant \inf_{\theta\in\mathcal{H}_1} \mathbb{E}_\theta \{\phi_n(\boldsymbol{Z}_n)\}$, that is, the power of the test achieves its minimum at the null hypothesis, as $n \to \infty$; and

- consistency: if $\theta \in \mathcal{H}_1$, $\lim_{n\to\infty} \xi_{\theta,n} = 0$, where $\xi_{\theta,n} = 1 - \mathbb{E}_\theta \{\phi_n(\boldsymbol{Z}_n) = 1\}$, that is, the test reaches its critical region with probability one if the alternative hypothesis is true, as $n \to \infty$.

We start by showing that the tests $\mathbf{T}_{1,n}$ and $\mathbf{T}_{2,n}$ are consistent. We assume in Theorems 1 and 2 that the fourth moment of $\boldsymbol{Z}_i = \left(\boldsymbol{X}_i^T, \boldsymbol{Y}_i^T\right)^T$ exists.

Theorem 1. *Tests $\mathbf{T}_{1,n}$ and $\mathbf{T}_{2,n}$ defined in Section 2.1 are consistent.*

Proof. We first prove the consistency of $\mathbf{T}_{1,n}$ for testing $\mathcal{H}_0 : \boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$ against $\mathcal{H}_1 : \boldsymbol{\mu}_X \neq \boldsymbol{\mu}_Y$. Let $\boldsymbol{D}_{i,n} = \boldsymbol{X}_{i,n} - \boldsymbol{Y}_{i,n}$. Then the sample mean vector and covariance matrix of the $\boldsymbol{D}_n$ sample are $\overline{\boldsymbol{D}}_n = \overline{\boldsymbol{X}}_n - \overline{\boldsymbol{Y}}_n$ and $S_{D_n} = S_{X_n} + S_{Y_n} - S_{X_nY_n} - S_{Y_nX_n}$, respectively. Here we assume S_{D_n} to be positive definite for any sample size n and sample data $(\boldsymbol{X}, \boldsymbol{Y})$. Write $\boldsymbol{\mu}_D = \boldsymbol{\mu}_X - \boldsymbol{\mu}_Y$. From the Central Limit Theorem, we have

$$S_{D_n}^{-1/2} \sqrt{n} \left(\overline{\boldsymbol{D}}_n - \boldsymbol{\mu}_D\right) \tag{3}$$

converges in distribution to the standard p-dimensional multivariate normal distribution as $n \to \infty$. Under $\mathcal{H}_1 : \boldsymbol{\mu}_X \neq \boldsymbol{\mu}_Y$, $n\mathbf{T}_{1,n}$ has an asymptotically noncentral chi-squared distribution with degrees of freedom p and non-centrality parameter $\Delta_n = n\, \boldsymbol{\mu}_D^T \left(\boldsymbol{\Omega}_X + \boldsymbol{\Omega}_Y - \boldsymbol{\Omega}_{XY} - \boldsymbol{\Omega}_{YX}\right)^{-1} \boldsymbol{\mu}_D$. The probability

$$\begin{aligned} \xi_{\theta,n}^{(1)} &= \mathbb{P}\left(n\mathbf{T}_{1,n} \leqslant \chi^2_{\alpha,p} \middle| \theta \in \mathcal{H}_1\right) \\ &= \mathbb{P}\left(n\mathbf{T}_{1,n} + \Delta_n \leqslant \chi^2_{\alpha,p} \middle| \theta \in \mathcal{H}_0\right) \end{aligned} \tag{4}$$

converges to 0 as $n \to \infty$. Therefore, $\mathbf{T}_{1,n}$ is consistent.

Next we prove the consistency of $\mathbf{T}_{2,n}$ for testing $\mathcal{H}_0 : \boldsymbol{\Omega}_X = \boldsymbol{\Omega}_Y$ against $\mathcal{H}_1 : \boldsymbol{\Omega}_X \neq \boldsymbol{\Omega}_Y$. Notice that $(n-1)S_{Z_n} = \left(\boldsymbol{Z}_n - n^{-1}J_n\boldsymbol{Z}_n\right)^T \left(\boldsymbol{Z}_n - n^{-1}J_n\boldsymbol{Z}_n\right) = \boldsymbol{Z}_n^T(I_n - n^{-1}J_n)\boldsymbol{Z}_n$ where I_n is the $n \times n$ identity matrix. Since the matrix $I_n - n^{-1}J_n$ is an idempotent matrix with rank $n-1$, the sum of squares $(n-1)S_{Z_n}$ has the Wishart distribution with parameters $\boldsymbol{\Omega}$, $n-1$

and $2p$, in asymptotic. Thus, $\sqrt{n}S_{Z_n}$ converges in distribution to the normal distribution with mean matrix $\mathbf{\Omega}$ and finite covariance matrix, say Ψ, as $n \to \infty$. By the continuity mapping theorem, we have that $g(S_{Z_n}) = \sqrt{n}\mathbf{T}_{2,n} = \sqrt{n}\big|\log|S_{X_n}| - \log|S_{Y_n}|\big|$ converges in distribution to the normal distribution with mean $g(\mathbf{\Omega}) = \sqrt{n}\big|\log|\mathbf{\Omega}_X| - \log|\mathbf{\Omega}_Y|\big|$ and covariance matrix $Q = \{\partial g(\mathbf{\Omega})/\partial\mathbf{\Omega}\}^T \Psi \{\partial g(\mathbf{\Omega})/\partial\mathbf{\Omega}\}$ which is assumed finite. Under $\mathcal{H}_1$, $g(\mathbf{\Omega}) \sim O(\sqrt{n})$. Thus, the probability

$$\begin{aligned} \xi^{(2)}_{\theta,n} &= \mathbb{P}\left(\sqrt{n}\mathbf{T}_{2n} \leqslant C_{\alpha,n} \big| \theta \in \mathcal{H}_1\right) \\ &= \mathbb{P}\left\{\sqrt{n}\mathbf{T}_{2,n} + g(\mathbf{\Omega}) \leqslant C_{\alpha,n} \big| \theta \in \mathcal{H}_0\right\}, \end{aligned} \tag{5}$$

where $C_{\alpha,n}$ is the critical value at sample size n and is $O(1)$, converges to 0 as $n \to \infty$. So, $\mathbf{T}_{2,n}$ is consistent. □

Theorem 2. *The combined test* $\mathbf{CT}(k_1, k_2)$ *defined in Section 2.1 is asymptotically unbiased and consistent.*

Proof. At first, we introduce three conditions that the combined test statistic, or equivalently its p-value should satisfy for asymptotic unbiasedness and consistency. Recall the p-value of the suggested combined test $\mathbf{CT}(k_1, k_2)$ is $\gamma = \gamma(\lambda_1, \lambda_2) = \min(\lambda_1/k_1, \lambda_2/k_2)$, where λ_1 and λ_2 are the p-values of the tests $\mathbf{T}_1$ and $\mathbf{T}_2$, and k_1 and k_2 are given constants. It is easy to justify that the combing function $\gamma(\lambda_1, \lambda_2)$ satisfies the following properties:

(P1) $\gamma(\lambda_1, \lambda_2)$ is a non-decreasing function of each argument.

(P2) $\gamma(\lambda_1, \lambda_2)$ decreases to 0, if one of $\lambda_l, l = 1, 2$, decreases to 0.

(P3) The p-values are well defined and non-trivial. In other words, the critical value is finite and non-trivial for every choice of significance level.

Similar properties of combining functions have been discussed by Goutis et al. (1996) and Pesarin (2001), among others.

We first prove that $\mathbf{CT}(k_1, k_2)$ is asymptotically unbiased. Let $\boldsymbol{Z}_n(\Delta_n)$ be random variables from model (3) with non-centrality Δ_n, and let $\lambda_1\{\boldsymbol{Z}_n(\Delta_n)\}$ be the p-value using the testing statistic $\mathbf{T}_{1,n}$. Then, (3) and (4) imply the stochastic ordering in $\lambda_1\{\boldsymbol{Z}_n(\Delta_n)\}$ for sufficiently large n; that is, if $\Delta_n \leqslant \Delta_n'$,

$$\mathbb{P}\left[\lambda_1\{\boldsymbol{Z}_n(\Delta_n)\} \leqslant z\right] \leqslant \mathbb{P}\left[\lambda_1\{\boldsymbol{Z}_n(\Delta_n')\} \leqslant z\right], \tag{6}$$

for every $z \in (0, 1)$. Thus,

$$\alpha = \mathbb{P}\left[\lambda_1\{\boldsymbol{Z}_n(0)\} \leqslant \alpha\right] \leqslant \mathbb{P}\left[\lambda_1\{\boldsymbol{Z}_n(\Delta_n)\} \leqslant \alpha\right].$$

Hence $\mathbf{T}_{1,n}$ is unbiased in asymptotic. Likewise, we have similar stochastic ordering to (6) for $\mathbf{T}_{2,n}$ with Δ_n replaced by $g(\boldsymbol{\Omega})$ in the p-value $\lambda_2\left[\boldsymbol{Z}_n\{g(\boldsymbol{\Omega})\}\right]$. This stochastic ordering relationship implies the asymptotic unbiasedness of $\mathbf{T}_{2,n}$. Since $\mathbf{T}_{1,n}$ and $\mathbf{T}_{2,n}$ are marginally unbiased in asymptotic, by the non-decreasing property of the combination function $\gamma(\lambda_1, \lambda_2)$ (P1), we have $\gamma\left[\lambda_1\{\boldsymbol{Z}_n(\Delta_n)\}, \lambda_2\right]$ is stochastically larger than $\gamma\left[\lambda_1\{\boldsymbol{Z}_n(0)\}, \lambda_2\right]$, and similarly, $\gamma\left(\lambda_1, \lambda_2\left[\boldsymbol{Z}_n\{g(\boldsymbol{\Omega})\}\right]\right)$ is stochastically larger than $\gamma\left[\lambda_1, \lambda_2\left\{\boldsymbol{Z}_n(0)\right\}\right]$. Hence, the asymptotic unbiasedness of $\mathbf{CT}(k_1, k_2)$ is achieved.

Next, we show consistency of $\mathbf{CT}(\lambda_1, \lambda_2)$. To be consistent, it must reach its critical region with probability one, if at least one of the two sub-alternative hypotheses $\mathcal{H}_1^{(1)} : \boldsymbol{\mu}_X \neq \boldsymbol{\mu}_Y$ and $\mathcal{H}_1^{(2)} : \boldsymbol{\Omega}_X \neq \boldsymbol{\Omega}_Y$ is true. By the consistency of the partial tests $\mathbf{T}_{1,n}$ and $\mathbf{T}_{2,n}$ as stated in Theorem 1, if either $\mathcal{H}_1^{(l)}$ is true, λ_l converges to 0 with probability one, for $l = 1, 2$. Thus by the properties (P2) and (P3), $\gamma(\lambda_1, \lambda_2)$ converges to 0 with probability one as $n \to \infty$. □

2.3 Permutation procedure

Without imposing distributional assumptions on $\boldsymbol{X}$ or $\boldsymbol{Y}$, we apply a nonparametric method — a permutation procedure to obtain reference distributions for the tests and to estimate their finite-sample p-values. Notice that an assumption behind a permutation test is that the

observations are exchangeable under the null hypothesis. We first discuss how the framework is set up to satisfy this requirement and then describe the proposed permutation procedure.

Let $\pi = \{\pi(1), \ldots, \pi(n)\}$, where each $\pi(i)$ is 0 or 1, $i = 1, \ldots, n$. Let Π denote the collection of all possible such π. For each random permutation π, the permuted paired sample $\{\boldsymbol{X}(\pi), \boldsymbol{Y}(\pi)\}$ is defined as that the observations for the $i$the subject become

$$\{\boldsymbol{X}(\pi)_i, \boldsymbol{Y}(\pi)_i\} = \begin{cases} (\boldsymbol{X}_i, \boldsymbol{Y}_i) & \text{if} \quad \pi(i) = 0, \\ (\boldsymbol{Y}_i, \boldsymbol{X}_i) & \text{if} \quad \pi(i) = 1. \end{cases}$$

For each permuted sample $\{\boldsymbol{X}(\pi), \boldsymbol{Y}(\pi)\}$, we compute the test statistics $\mathbf{T}_1(\pi)$ and $\mathbf{T}_2(\pi)$. When the global null hypothesis $\mathcal{H}_0 : \boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$ and $\boldsymbol{\Omega}_X = \boldsymbol{\Omega}_Y$ is true, data exchangeability is satisfied.

Recall $\mathbf{T}_1$ depends only on the differences $\boldsymbol{D}_i$ rather than on the paired data $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$ directly. For $\mathbf{T}_1$, instead of imposing $\boldsymbol{\Omega}_X = \boldsymbol{\Omega}_Y$, we preserve the exchangeability condition by assuming symmetric joint distribution of paired data. As proven in the Appendix A.1, under $\mathcal{H}_0 : \boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$, the null distribution of the test statistic $\mathbf{T}_1$ is invariant to the permutation π when the joint distribution of $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$ is symmetric.

The proposed permutation test for $\mathbf{T}_2$ is based on the permutation of $\boldsymbol{X}_i - \boldsymbol{\mu}_X$ and $\boldsymbol{Y}_i - \boldsymbol{\mu}_Y$, without requiring $\boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$. Appendix A.2 provides the proof that under $\mathcal{H}_0 : \boldsymbol{\Omega}_X = \boldsymbol{\Omega}_Y$, the null distribution of $\mathbf{T}_2$ is invariant to the permutation π. In practice, the mean vectors $\boldsymbol{\mu}_X$ and $\boldsymbol{\mu}_Y$ are unknown. We propose to approximate the permutation test $\mathbf{T}_2$ by substituting the population means with their original sample mean vectors. To be specific, for each permutation $\pi \in \Pi$, we approximate $\boldsymbol{X}(\pi)_i - \boldsymbol{\mu}_X$ and $\boldsymbol{Y}(\pi)_i - \boldsymbol{\mu}_Y$ as $\boldsymbol{X}^*(\pi)_i = \boldsymbol{X}(\pi)_i - \overline{\boldsymbol{X}}^o$ and $\boldsymbol{Y}^*(\pi)_i = \boldsymbol{Y}(\pi)_i - \overline{\boldsymbol{Y}}^o$ with $\overline{\boldsymbol{X}}^o$ and $\overline{\boldsymbol{Y}}^o$ being the sample mean vectors of the original paired data, respectively, when computing $\mathbf{T}_2(\pi)$.

The proposed permutation procedure is summarized as follows:

(a) Compute the observed test statistics $\mathbf{T}_1^o$ and $\mathbf{T}_2^o$ from the original data.

(b) For each permutation $\pi \in \Pi$, compute $\mathbf{T}_1(\pi)$ of the permuted sample $\{\boldsymbol{X}(\pi), \boldsymbol{Y}(\pi)\}$

and compute $\mathbf{T}_2(\pi)$ based on the sample $\{\boldsymbol{X}^*(\pi), \boldsymbol{Y}^*(\pi)\}$ with $\boldsymbol{X}^*(\pi)_i = \boldsymbol{X}(\pi)_i - \overline{\boldsymbol{X}}^o$ and $\boldsymbol{Y}^*(\pi)_i = \boldsymbol{Y}(\pi)_i - \overline{\boldsymbol{Y}}^o$, where $\overline{\boldsymbol{X}}^o$ and $\overline{\boldsymbol{Y}}^o$ are from the original paired samples and thus fixed.

(c) The p-value of the test $\mathbf{T}_1$ is estimated by $\hat{\lambda}_1 = \sum_{\pi\in\Pi} \mathbf{1}\left\{\mathbf{T}_1(\pi) \geqslant \mathbf{T}_1^o\right\}/|\Pi|$, where $\mathbf{1}(\cdot)$ stands for the indicator function and $|\Pi|$ is the number of all possible permutations π in Π. The p-value of the test $\mathbf{T}_2$ is estimated by $\hat{\lambda}_2 = \sum_{\pi\in\Pi} \mathbf{1}\left\{\mathbf{T}_2(\pi) \geqslant \mathbf{T}_2^o\right\}/|\Pi|$. The parameter τ is approximated by

$$\hat{\tau} = |\Pi| \frac{\sum_{\pi\in\Pi} \mathbf{1}\left\{\mathbf{T}_1(\pi) \geqslant \mathbf{T}_1^o, \mathbf{T}_2(\pi) \geqslant \mathbf{T}_2^o\right\}}{\sum_{\pi\in\Pi} \mathbf{1}\left\{\mathbf{T}_1(\pi) \geqslant \mathbf{T}_1^o\right\} \cdot \sum_{\pi\in\Pi} \mathbf{1}\left\{\mathbf{T}_2(\pi) \geqslant \mathbf{T}_2^o\right\}}.$$

Finally, the p-value of the combined test $\mathbf{CT}(k_1, k_2)$ is estimated by $(k_1+k_2)\hat{\gamma} - k_1k_2\hat{\tau}\hat{\gamma}^2$, where $\hat{\gamma} = \min(\hat{\lambda}_1/k_1,\ \hat{\lambda}_2/k_2)$.

In practice, $|\Pi| = 2^n$ can be very large if n is large. When there are too many possible orderings of the data to conveniently allow complete enumeration, the computational intensity may be unfeasible. To reduce the computational burden when this happens, we modify the procedure using an asymptotically equivalent technique by Monte Carlo sampling:

(b)-1 Generate K random permutations using Bernoulli random numbers. That is, to obtain a permuted sample, generate n independent random numbers $b_i, i = 1, \ldots, n$, from the Bernoulli distribution with probability 0.5, and permute $\boldsymbol{X}_i$ and $\boldsymbol{Y}_i$ only if $b_i = 1$. Let $\pi^{(k)}$ label the kth such generated permutation, $k = 1, \ldots, K$.

(c)-1 Calculate $\hat{\lambda}_1 = \sum_{k=1}^{K} \mathbf{1}\left\{\mathbf{T}_1(\pi^{(k)}) \geqslant \mathbf{T}_1^o\right\}/K$, $\hat{\lambda}_2 = \sum_{k=1}^{K} \mathbf{1}\left\{\mathbf{T}_2(\pi^{(k)}) \geqslant \mathbf{T}_2^o\right\}/K$, and

$$\hat{\tau} = K\ \frac{\sum_{k=1}^{K} \mathbf{1}\left\{\mathbf{T}_1(\pi^{(k)}) \geqslant \mathbf{T}_1^o, \mathbf{T}_2(\pi^{(k)}) \geqslant \mathbf{T}_2^o\right\}}{\sum_{k=1}^{K} \mathbf{1}\left\{\mathbf{T}_1(\pi^{(k)}) \geqslant \mathbf{T}_1^o\right\} \cdot \sum_{k=1}^{K} \mathbf{1}\left\{\mathbf{T}_2(\pi^{(k)}) \geqslant \mathbf{T}_2^o\right\}}.$$

Here, K is small, relative to $|\Pi|$. If the p-value is 0.05, the approximate standard error of the estimated p-value from $K = 10,000$ random permutations is 0.0022.

3 Simulation

Monte Carlo samples were generated to evaluate the performances of the proposed test procedure, including the size (i.e., type I error probability) and power (i.e., one minus type II error probability) of the three tests. We compare the proposed tests to their counterparts using likelihood ratio tests which are derived under the assumption that all the data are normally distributed. Lim, et al. (2010) showed the finite-sample distributions of the likelihood ratio test (LRT) statistics and proposed a parametric bootstrap procedure for implementation without structural assumptions on the covariance matrices. They also showed that when the normality assumption is met, the LRT statistics have the usual asymptotic Chi-squared distributions under the null hypotheses. In the comparisons, we include both the finite-sample LRTs using the re-sampling procedure and the asymptotic Chi-squared tests, with the normality assumption imposed. To investigate performance under different true distributions, we considered multivariate normal distributions and two non-normal scenarios: multivariate bimodal mixture of two normals with mixing proportion 50% and multivariate student's t distributions with 10 degrees of freedom.

We generated paired samples $\boldsymbol{Z}_i = (\boldsymbol{X}_i^T, \boldsymbol{Y}_i^{\mathrm{T}})^T$, $i = 1, \ldots, n$, from a $(2p)$-variate distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Omega}$, where both $\boldsymbol{X}$ and $\boldsymbol{Y}$ are either normal, or bimodal mixture of normals, or t_{10}. We let $\boldsymbol{\mu} = (\boldsymbol{\mu}_X^T, \boldsymbol{\mu}_Y^T)^T$ with $\boldsymbol{\mu}_X = \mu_X \mathbf{1}_p$ and $\boldsymbol{\mu}_Y = \mu_Y \mathbf{1}_p$ for $\mathbf{1}_p$ being the $p \times 1$ vector of 1's; and $\boldsymbol{\Omega}$ with diagonal matrices that have compound symmetry structures $\boldsymbol{\Omega}_X = \sigma_X^2 \{(1-\rho_X) I_p + \rho_X J_p\}$ and $\boldsymbol{\Omega}_Y = \sigma_Y^2 \{(1-\rho_Y) I_p + \rho_Y J_p\}$ and off-diagonal matrices $\boldsymbol{\Omega}_{XY} = \rho_{XY} J_p$ and $\boldsymbol{\Omega}_{YX} = \rho_{YX} J_p$, for J_p denoting the $p \times p$ matrix with all entries equal to 1. In such simulation setup, $\mu_X = \mu_Y$ implies $\boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$, and $\boldsymbol{\Omega}_X = \boldsymbol{\Omega}_Y$ happens when both $\sigma_X^2 = \sigma_Y^2$ and $\rho_X = \rho_Y$ are satisfied. The within-population correlation is determined by ρ_X and ρ_Y in $\boldsymbol{\Omega}_X$ and $\boldsymbol{\Omega}_Y$, respectively. The parameters ρ_{XY} and ρ_{YX} control the dependence between the two samples which are uncorrelated only when $\rho_{XY} = \rho_{YX} = 0$. In the simulation results reported here, $p = 5$, $\mu_X = 0$, $\sigma_X^2 = 1$, $\rho_X = \rho_Y = 0.5$, and

$\rho_{XY} = \rho_{YX} = 0.3$. We investigated the performances of the tests for various null and alternative situations created by different values of $\mu_Y - \mu_X$ and different values of $\sigma_Y^2 - \sigma_X^2$. We also examined the outcomes for different sample sizes. We generated 500 Monte Carlo samples for each setup with a choice of $\mu_Y (= 0, 0.5, 1)$, a choice of $\sigma_Y^2 (= 1, 1.5, 2)$, and a choice of sample size $n (= 15, 25, 50)$. For each data set we tested the three hypotheses (i)–(iii) respectively using the proposed $\mathbf{T}_1$, $\mathbf{T}_2$ and $\mathbf{CT}$ as described in Section 2.3, as well as their corresponding finite-sample LRT and asymptotic Chi-squared test based on LRT (Lim, et al., 2010). As $p = 5$, the asymptotic Chi-squared tests have degrees of freedom 5, 15 and 20 for the hypotheses (i)–(iii), respectively. A null hypothesis is rejected if a p-value is no more than the significance level $\alpha = 0.05$. The empirical rejection probability of a test was calculated as the proportion of rejections from 500 replicates.

Table 1 provides the empirical rejection probabilities of the $\mathbf{T}_1$, LRT and χ_5^2 for testing $\mathcal{H}_0 : \boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$. When the null hypothesis is true ($\mu_Y - \mu_X = 0$), regardless of the actual distribution of data and sample size, the empirical sizes of $\mathbf{T}_1$ are close to the nominal 0.05 level; while the LRT severely underestimates the 0.05 level when sample size is very small and the $\mathbf{T}_1$ still slightly outperforms the LRT for larger n. As the difference between μ_X and μ_Y increases, the $\mathbf{T}_1$ demonstrates more power in detecting the inequality between the mean vectors and is more powerful than the LRT. From Table 2 which displays the empirical rejection probabilities of the $\mathbf{T}_2$, LRT and χ_{15}^2 for testing $\mathcal{H}_0 : \boldsymbol{\Omega}_X = \boldsymbol{\Omega}_Y$, the $\mathbf{T}_2$ attains the nominal 0.05 level under the null hypothesis ($r = \sigma_Y^2/\sigma_X^2 = 1$), no matter data are normal, bimodal, or have heavy tails and regardless of the sample size. The empirical size of the LRT is way below the nominal level for $n = 15$ regardless of the underlying distribution. Under departures from normality, especially when data present heavy tails, the LRT can show degradation of performance with empirical size above the nominal level for larger n, verifying its sensitivity to non-normality. The greater difference between $\boldsymbol{\Omega}_X$ and $\boldsymbol{\Omega}_Y$ (deviation of r from 1), the more powerful the $\mathbf{T}_2$ is for detecting the heterogeneity

of covariance matrices in general. When data are normally distributed under which the LRT is valid, $\mathbf{T}_2$ are substantially more powerful than the LRT. Table 3 gives the empirical rejection probabilities of the **CT**, LRT and χ^2_{20} for testing $\mathcal{H}_0 : \boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$ and $\boldsymbol{\Omega}_X = \boldsymbol{\Omega}_Y$. As expected, **CT** achieves the nominal level when the null hypothesis is true ($\mu_Y - \mu_X = 0$ and $r = \sigma^2_Y/\sigma^2_X = 1$), in spite of data distribution and sample size; while the LRT underestimates the nominal level when $n = 15$. Again, the LRT is sensitive to violations of normality in data, evident by having remarkably liberal empirical size for t distributions for larger n. Under normality, the **CT** shows better power than the LRT in detecting an alternative. As the alternative hypothesis becomes more pronounced, the **CT** gains greater power. From all of the three Tables, if a null hypothesis is true, the empirical size of the Chi-squared test is generally fairly larger than 0.05, and its performance worsens with more liberal empirical size especially for t distributions. The power of the Chi-squared tests is generally unreliable due to its unacceptably liberal empirical size. Similar conclusions have been drawn for $n = 100$ with improved performances of the LRTs and Chi-squared tests as expected. In another simulation not reported here where both $\boldsymbol{X}$ and $\boldsymbol{Y}$ samples were generated from multivariate asymmetric bimodal mixture of two normals with mixing proportion 70%, all three proposed tests also had satisfactory performance similar to the above.

Overall, $\mathbf{T}_1$, $\mathbf{T}_2$ and **CT** demonstrate good performance regardless of the data distribution. Since their empirical power increases as the distance between the alternative and the null hypotheses gets larger, they have monotone power functions, which implies their unbiasedness. Their power also increases with sample size, suggesting the consistency of the tests. These numerical results complement the theory in Section 2.2.

4 Applications

4.1 Tooth size example

We apply the proposed tests to the dental data set introduced in Section 1. We focus on the tooth size measurements from 179 men who had natural normal occlusion. The goal is to assess whether the tooth size is symmetric between the left and right sides around central incisors in either maxilla or mandible using the proposed tests.

Denote the maxilla tooth size measurements of subject i by $\boldsymbol{X}_i^u = (X_{i1}^u, \ldots, X_{i7}^u)^T$ and $\boldsymbol{Y}_i^u = (Y_{i1}^u, \ldots, Y_{i7}^u)^T$ for the left and right sides respectively from central incisor to second molar in the upper jaw, $i = 1, \ldots, 179$. Let $\boldsymbol{\mu}_X^u$ and $\boldsymbol{\Omega}_X^u$ be the population mean and covariance for $\boldsymbol{X}^u$, and $\boldsymbol{\mu}_Y^u$ and $\boldsymbol{\Omega}_Y^u$ be the population mean and covariance for $\boldsymbol{Y}^u$. The three symmetry hypotheses to be tested are $\boldsymbol{\mu}_X^u = \boldsymbol{\mu}_Y^u$, $\boldsymbol{\Omega}_X^u = \boldsymbol{\Omega}_Y^u$, and simultaneous $\boldsymbol{\mu}_X^u = \boldsymbol{\mu}_Y^u$ and $\boldsymbol{\Omega}_X^u = \boldsymbol{\Omega}_Y^u$. Similarly, we denote mandibular tooth size measurements by $\boldsymbol{X}^l$ and $\boldsymbol{Y}^l$ and test the symmetry hypotheses $\boldsymbol{\mu}_X^l = \boldsymbol{\mu}_Y^l$, $\boldsymbol{\Omega}_X^l = \boldsymbol{\Omega}_Y^l$, and simultaneous $\boldsymbol{\mu}_X^l = \boldsymbol{\mu}_Y^l$ and $\boldsymbol{\Omega}_X^l = \boldsymbol{\Omega}_Y^l$. We apply the tests to the maxilla and mandible data sets separately.

The estimated p-values of $\mathbf{T}_1$, $\mathbf{T}_2$ and $\mathbf{CT}$ implemented using the proposed permutation procedure are reported in Table 4, along with corresponding LRT and asymptotic Chi-squared tests. The proposed test $\mathbf{T}_1$ suggests that $\boldsymbol{\mu}_X^u \neq \boldsymbol{\mu}_Y^u$ and $\boldsymbol{\mu}_X^l \neq \boldsymbol{\mu}_Y^l$; the LRT and Chi-squared tests give similar conclusions. At the significance level 0.05, the proposed test $\mathbf{T}_2$ suggests $\boldsymbol{\Omega}_X^u = \boldsymbol{\Omega}_Y^u$ and $\boldsymbol{\Omega}_X^l = \boldsymbol{\Omega}_Y^l$; but the LRT and Chi-squared tests show a disagreement with the suggestion of $\boldsymbol{\Omega}_X^l \neq \boldsymbol{\Omega}_Y^l$. Complementing the results of $\mathbf{T}_1$, the combined test $\mathbf{CT}$ provide significant p-values for both maxilla and mandible; while the LRT and Chi-squared tests draw a similar conclusion for mandible, but are insignificant when testing for simultaneous equalities of mean vectors and covariance matrices for maxilla.

The LRT and Chi-squared tests rely on normality of data. Their disagreement with $\mathbf{T}_2$ is due to departures from normality in the data. For example, in mandible, the sample skewness

and kurtosis of left central incisors (X^l_{i1}) are 4.15 and 37.78, respectively. The right first molars in mandible (Y^l_{i6}) have sample skewness -1.46 and kurtosis 8.29. These violations of normality cause the LRT and Chi-squared tests to be liberal and reject $\mathcal{H}_0 : \mathbf{\Omega}^l_X = \mathbf{\Omega}^l_Y$, conflicting with the test $\mathbf{T}_2$ which is insensitive to departures from normality. Such behavior of normality violation is less pronounced for the teeth size measurements in maxilla, except a mild departure in right first molars (Y^u_{i6}) whose sample skewness and kurtosis are -1.06 and 6.24, respectively. This explains that the conclusions of the LRT and Chi-squared tests coincide with the $\mathbf{T}_2$ in general for maxilla, although $\mathbf{T}_2$ shows a borderline significant p-value of 0.0722.

Overall, use of the proposed tests, unlike the competitors, offers the analyst assurance of credible results. Based on the proposed tests $\mathbf{T}_1$, $\mathbf{T}_2$ and $\mathbf{CT}$, we conclude that the mean tooth size profiles differ between the left and right sides in either maxilla or mandible of these men. The dispersions of tooth sizes on the two sides are the same in mandible, and the dispersions are nearly the same between left and right sides in maxilla. If we presume that the tooth size should be genetically symmetric, the asymmetry in mean tooth size profiles in both maxilla and mandible of male adults might be conjectured to be influenced by environmental effects and this surely requires further investigation to justify.

4.2 Gene example

To illustrate the usefulness of the proposed method for testing differentially structured gene sets in microarray studies discussed in Section 1, we apply the proposed permutation method to a microarray expression data set with paired tumor and normal tissue samples from 53 primary prostate cancer patients. This data set was originally generated on Affymetrix Human Genome U95C Arrays to study gene expression alteration in prostate cancer (see Yu et al., 2004, for the experimental details), and is downloadable with the accession number GSE6919 via the NCBI Gene Expression Omnibus repository (Barrett et al., 2005). From

the Molecular Signatures Database (`http://www.broadinstitute.org/gsea/msigdb`), we downloaded three rather small-size gene sets that were defined by mining large collections of cancer-oriented microarray data (Subramanian et al. 2005). The sizes of the gene sets vary from $n = 24$ to 39, and the expression levels range from $p = 9$ to 12 genes. Table 5 reports detailed information on these gene sets.

For a given gene set, denote the expression level of subject (or array) i by $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T$ and $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{ip})^T$ for the normal and tumor tissues, respectively. Let $\boldsymbol{\mu}_X$ and $\boldsymbol{\Omega}_X$ be the population mean and covariance for $\boldsymbol{X}$, and $\boldsymbol{\mu}_Y$ and $\boldsymbol{\Omega}_Y$ be the population mean and covariance for $\boldsymbol{Y}$. We test the three symmetry hypotheses: $\boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$, $\boldsymbol{\Omega}_X = \boldsymbol{\Omega}_Y$, and simultaneous $\boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$ and $\boldsymbol{\Omega}_X = \boldsymbol{\Omega}_Y$. Because these gene sets are known to be related with prostrate cancer and be differentially expressed in mean between normal and tumor tissues, our main interest is in testing $\boldsymbol{\Omega}_X = \boldsymbol{\Omega}_Y$, which will elucidate the difference between dependent structures. The test results in Table 6 show that the dependent structures differ between normal and tumor tissues in gene sets GCM_FANCL and GCM_CASP2, while there is boarder-line differences between the normal and tumor tissue dependent structures in gene set GNF2_ICAM3. However, the results based on LRT and asymptotic Chi-squared tests are quite liberal.

The estimated sample concentration matrices for gene set GCM_FANCL are

$$\widehat{\mathbf{\Omega}}_X^{-1} =$$

$$\begin{pmatrix}
539.1 & -6.3 & -38.0 & -69.9 & 56.0 & -104.8 & 69.8 & -47.6 & -68.9 & 229.8 \\
-6.3 & 837.2 & 60.7 & -25.1 & -185.9 & 280.8 & -93.5 & -37.3 & -218.5 & -380.2 \\
-38.5 & 60.7 & 1097.0 & -59.8 & -9.1 & -105.3 & -42.4 & -421.5 & -21.2 & -244.2 \\
-69.9 & -25.1 & -59.8 & 630.4 & -88.6 & -72.9 & -98.4 & 79.0 & 115.0 & -423.9 \\
56.6 & -185.9 & -9.1 & -88.6 & 698.5 & -357.6 & 138.6 & -158.7 & -41.3 & 357.1 \\
-104.8 & 280.8 & -105.3 & -72.9 & -357.6 & 951.4 & -255.5 & 134.0 & 114.7 & -491.3 \\
69.8 & -93.5 & -42.4 & -98.4 & 138.6 & -255.5 & 1051.9 & -105.6 & -439.0 & -162.1 \\
-47.6 & -37.3 & -421.5 & 79.0 & -158.7 & 134.0 & -105.6 & 723.7 & -84.1 & 260.7 \\
-68.9 & -218.5 & -21.2 & 115.0 & -41.3 & 114.7 & -439.0 & -84.1 & 1723.7 & -930.6 \\
229.8 & -380.2 & -244.2 & -423.9 & 357.1 & -491.3 & -162.1 & 260.7 & -930.6 & 3154.5
\end{pmatrix}$$

and

$$\widehat{\mathbf{\Omega}}_Y^{-1} =$$

$$\begin{pmatrix}
399.7 & -168.0 & 150.8 & -6.1 & -6.3 & -43.5 & 121.7 & 106.7 & -195.9 & 23.1 \\
-168.0 & 3399.1 & 395.6 & -263.6 & 69.1 & -253.0 & -375.4 & 194.3 & -1968.1 & -526.6 \\
150.8 & 395.6 & 3276.4 & -93.0 & 13.1 & 34.5 & 151.7 & -130.7 & -3115.1 & 108.5 \\
-6.1 & -263.6 & -93.0 & 368.6 & -0.4 & 40.0 & 53.7 & -107.6 & 189.0 & -11.6 \\
-6.3 & 69.1 & 13.1 & -0.4 & 372.0 & -131.7 & 108.6 & 68.2 & -164.8 & 76.2 \\
-43.5 & -253.0 & 34.5 & 40.0 & -131.7 & 808.0 & -331.7 & -263.2 & 243.2 & 17.7 \\
121.7 & -375.4 & 151.7 & 53.7 & 108.6 & -331.7 & 1935.2 & 297.2 & -1541.7 & 138.5 \\
106.7 & 194.3 & -130.7 & -107.6 & 68.2 & -263.2 & 297.2 & 958.2 & -920.7 & 45.1 \\
-195.9 & -1968.1 & -3115.1 & 189.0 & -164.8 & 243.2 & -1541.7 & -920.7 & 6582.1 & 31.4 \\
23.1 & -526.6 & 108.5 & -11.6 & 76.2 & 17.7 & 138.5 & 45.1 & 31.4 & 299.8
\end{pmatrix}.$$

It is easy to see that the estimates of $\mathbf{\Omega}_Y$ elements are larger than those of $\mathbf{\Omega}_X$ in general. This implies changes in dependent structures between normal and tumor tissues, and is consistent with the significance of the proposed test $\mathbf{T}_2$. Moreover, to illustrate the differences in dependent structure, we applied the graphical lasso method studied by Yuan and Lin (2007) and Friedman et al. (2008) to estimating the sparse concentration matrices of the normal

and tumor genes in gene set GCM_FANCL. The tuning parameter in the graphical lasso was chosen such that the BIC-type criterion of Yuan and Lin (2007) was minimized. Figure 1 plots the estimated dependent structures of the normal and the tumor genes in gene set GCM_FANCL, where the 10 genes in the gene set are represented by the 10 nodes, and the solid lines between nodes indicate two genes are conditionally dependent to each other given all other genes. It is shown that many new edges appear in the graph of tumor genes, which is in consistency with the findings from the proposed tests and with the observation of estimated concentration matrices.

5 Conclusion

We have presented tests for three hypotheses in multivariate paired data: equality of mean vectors, sameness of covariance matrices, and simultaneous equalities of mean vectors and covariance matrices. A main feature of the proposed tests is that they need neither a covariance structural nor a distributional assumption on the data, except that the permutation test for mean vector equality needs the joint distribution of the paired data to be symmetric. We have proposed to combine the two partial tests into a test for simultaneous equalities of mean vectors and covariance matrices. The combined test has been proven to be asymptotically unbiased and consistent in theory because the partial tests are consistent and independent in asymptotic. A permutation procedure has been proposed to estimate the reference distributions, which is easy to implement. In contrast to competing tests that rely on the normality assumption of data and may provide misleading conclusions when the assumption is violated, the proposed test yield sound performance in terms of attaining valid nominal type I error probabilities under the null hypotheses and achieving appreciable power in detecting various alternative hypotheses regardless of departures from normality. Although the choice of the rejection region for the combined hypothesis is somewhat *ad hoc*, the resulting test has demonstrated good numerical performances indicating its power and

usefulness in practice. Its alternatives and properties are included in our future work. Reliable and comprehensive conclusions can be made when all three proposed tests are applied together to a data set, as illustrated by simulation studies and the applications. In addition, the proposed tests are constructed without the requirement that the correlation between the two paired populations be symmetric, i.e., $\mathbf{\Omega}_{XY}$ and $\mathbf{\Omega}_{YX}$ are allowed to differ. This may lead to the usefulness of adapting the proposed tests to situations, such as some microarray studies where the assumption $\mathbf{\Omega}_{XY} = \mathbf{\Omega}_{YX}$ have been found invalid.

The permutation test of mean equality, $\mathbf{T}_1$, requires symmetric joint distribution of $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$, while the permutation tests for sameness of covariance matrices and for simultaneous equalities of mean vectors and covariance matrices do not need such requirement. A referee has helpfully showed in a small simulation that the test $\mathbf{T}_1$ can be problematic under a violation of this assumption since the permutation principle of data exchangeability is invalid. It may be interesting to study the sensitivity of $\mathbf{T}_1$ to departures from this assumption. In this paper, we do not assume any structured covariance model for multivariate paired observations within a subject. However, when repeated multivariate measurements (or longitudinal data) for each subject are available, more detailed structured models such as multivariate mixed effects models could be posited. Our strategy may be extended to such structured covariance models for developing distribution-free tests, although the development seems non-trivial and some efforts will be needed. These are included in our future work.

Acknowledgements

We are grateful to Mr. Donghyeon Yu for help with the graphical lasso method. We thank two referees and the Editor for their helpful comments and suggestions. Johan Lim's research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by Ministry of Education, Science and Technology(No.

2010-0011448).

References

Anderson, T.W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd edition. New York: Wiley.

Aslan, B. and Zech, G. (2005). New test for the multivariate two-sample problem based on the concept of minimum energy. *Journal of Statistical Computation and Simulation* **75**, 109-119.

Barrett, T., Suzek, T.O., Troup, D.B., Wilhite, S.E., Ngau, W.C., Ledoux, P., Rudnev, D., Lash, A.E., Fujibuchi, W., Edgar, R. (2005). NCBI GEO: mining millions of expression profiles–database and tools. *Nucleic Acids Res 33(Database issue)*:D562-566.

Bradley, E. and Blackwood, L.G. (1989). Comparing paired data: a simultaneous test for means and variances. *The American Statistician* **43**, 234–235.

Chung, J.H. and Fraser, D.A.S. (1958). Randomization tests for a multivariate two-sample problem. *Journal of the American Statistical Association* **53**, 729–735.

Choi, S.C. and Wette, R. (1972). A test for homogeneity of variances among correlated variables. *Biometrics* **28**, 589–592.

Conover, W.J., Johnson, M.E. and Johnson, M.M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics* **23**, 351–361.

de Leon, A. (2007). One-sample likelihood ratio tests for mixed data. *Communications in Statistics. A. Theory and Methods* **36**, 129–141.

Efron, B. and Tibshirani, R. (2007) On testing the significance of sets of genes. *The Annals of Applied Statistics*, **1**, 107-129.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432-441.

Garn, S.M., Lewis, A.B., and Kerewsky, R.S. (2002). The meaning of bilateral asymmetry in the permanent dentition. *American Journal of Orthodontics & Dentofacial Orthopedics* **36**, 55–62.

Goutis, C., Casella, G., and Wells, M.T. (1996). Assessing evidence in multiple hypotheses. *Journal of the American Statistical Association* **91**, 1268–1277.

Han, C.P. (1968). Testing the homogeneity of a test of correlated variances. *Biometrika* **55**, 317–326.

Harris, P. (1985). Testing for variance homogeneity of correlated variables. *Biometrika* **72**, 103–107.

Hirotsu, C. (1986). Cummulative chi-squared statistic or a tool for testing goodness of fit. *Biometrika* **73**, 165–173.

Hirotsu, C. (1998). Max t test for analysing a dose-response relationship - an efficient algorithm for p value calculation. In L. Pronzato, *Volume of Abstracts of MODA-5 (5th International Conference on Advances in Model Oriented Data Analysis and Experimental Design)*. CIRM, Marseille.

Hotelling, H. (1931). The generalization of Student's ratio. *The Annals of Mathematical Statistics* **2**, 360-378.

Kariya, T. (1981). A ronustness propertity of Hotelling's T^2-test. *The Annals of Statistics* **9**, 211–214.

Kim, J.Y., Lee, S.J., Kim, T.W., Nahm, D.S., and Chang Y.I. (2005). Classification of the skeletal variation in normal occlusion. *Angle Orthodontics* **75**, 303–311.

Lauritzen, S.L. (2004) *Graphical Models.* Oxford University Press. New York.

Lee, S.J., Lee, S., Lim, J., Ahn, S.J., and Kim, T.W. (2007). Cluster analysis of human tooth size in subjects with normal occlusion. *American Journal of Orthodontics & Dentofacial Orthopedics* **132**, 796–800.

Lim, J., Li, E., and Lee, S.J. (2010). Likelihood ratio tests of correlated multivariate samples. *Journal of Multivariate Analysis* **101**, 541–554.

Newton, M., Quintana, F., Den Boon, J., Sengupta, S. and Ahlquist, P. (2007) Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *The Annals of Applied Statistics*, **1**, 85-106.

Olson, C.L. (1974). Comparative robustness of six tests in multivariate analysis of variance. *Journal of the American Statistical Association* **69**, 894–908.

Perlman, M.D. (1980). Unbiasedness of the likelihood ratio tests for equality of several covariance matrices and equality of several multivariate normal populations. *The Annals of Statistics* **8**, 247–263.

Pesarin, F. (2001). *Multivariate Permutation Tests With Applications in Biostatistics*. New York: Wiley.

Srivastava, M.S. and Awan, H.M. (1982). On the robustness of Hotelling T^2-test and distribution of linear and quadratic forms in sampling from a mixture of two multivariate normal populations. *Communications in statistics - Theory and Methods* **11**, 81–107.

Subramanian, A., Tamayoa, P., Moothaa, V.K., Mukherjeed, S., Eberta, B.L., Gillettea, M.A., Paulovichg, A., Pomeroy, S.L., Goluba T.R., Landera, E.S., and Mesirova,

J.P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of National Academy of Science, U.S.A.*, **102**, 15545-15550.

Szatrowski, T.H. (1979). Asymptotic nonnull distributions for likelihood ratio statistics in the multivariate normal patterned mean and covariance matrix testing problem. *The Annals of Statistics* **7**, 823–837.

Uysal, T., Sari, Z., Basciftci, F.A., and Memili, B. (2005). Intermaxillary tooth size discrepancy and malocclusion: is there a relation? *Angle Orthodontics* **75**, 204–209.

Yu, Y.P., Landsittel, D., Jing, L., Nelson, J., Ren, B., Liu, L., McDonald, C., Thomas, R., Dhir, R., Finkelstein, S., Michalopoulos, G., Becich, M., Luo, J.H. (2004). Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. *1: Journal of Clinical Oncology*, **22**, 2790-2799.

Yuan, M. and Lin., Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika*, **94**, 19-35.

Appendix

A.1 Permutation invariance of $\mathbf{T}_1$

Recall $\boldsymbol{D}_i = \boldsymbol{X}_i - \boldsymbol{Y}_i$. Define $\left(\widetilde{\boldsymbol{X}}_i, \widetilde{\boldsymbol{Y}}_i\right) \equiv (-\boldsymbol{X}_i, -\boldsymbol{Y}_i)$ and $\widetilde{\boldsymbol{D}}_i = \widetilde{\boldsymbol{X}}_i - \widetilde{\boldsymbol{Y}}_i$, then it is straightforward that $\widetilde{\boldsymbol{D}}_i = \boldsymbol{Y}_i - \boldsymbol{X}_i$. For each permutation $\pi \in \Pi$, $\{\boldsymbol{X}(\pi)_i, \boldsymbol{Y}(\pi)_i\}$ is the permutated paired sample for the ith subject. Let $\boldsymbol{D}(\pi)_i = \boldsymbol{X}(\pi)_i - \boldsymbol{Y}(\pi)_i$, then

$$\boldsymbol{D}(\pi)_i = \begin{cases} \boldsymbol{X}_i - \boldsymbol{Y}_i \ = \boldsymbol{D}_i & \text{if } \pi(i) = 0, \\ \boldsymbol{Y}_i - \boldsymbol{X}_i \ = \widetilde{\boldsymbol{D}}_i & \text{if } \pi(i) = 1. \end{cases}$$

Suppose the joint distribution of $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$ is symmetric around its mean $(\boldsymbol{\mu}_X, \boldsymbol{\mu}_Y)$ and the null hypothesis $\mathcal{H}_0 : \boldsymbol{\mu}_X = \boldsymbol{\mu}_Y (= \boldsymbol{\mu})$ is true. The joint distribution of $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$ is symmetric

about $(\boldsymbol{\mu}, \boldsymbol{\mu})$. Then the distribution of $(\boldsymbol{X}_i - \boldsymbol{\mu}, \boldsymbol{Y}_i - \boldsymbol{\mu})$ is symmetric about $(\mathbf{0}, \mathbf{0})$. It follows that $(\boldsymbol{X}_i - \boldsymbol{\mu}, \boldsymbol{Y}_i - \boldsymbol{\mu})$ has the same distribution as $\{-(\boldsymbol{X}_i - \boldsymbol{\mu}), -(\boldsymbol{Y}_i - \boldsymbol{\mu})\} = \left(\widetilde{\boldsymbol{X}}_i + \boldsymbol{\mu}, \widetilde{\boldsymbol{Y}}_i + \boldsymbol{\mu}\right)$. This implies that the difference $(\boldsymbol{X}_i - \boldsymbol{\mu}) - (\boldsymbol{Y}_i - \boldsymbol{\mu}) = \boldsymbol{X}_i - \boldsymbol{Y}_i$ and the difference $(\widetilde{\boldsymbol{X}}_i + \boldsymbol{\mu}) - (\widetilde{\boldsymbol{Y}}_i + \boldsymbol{\mu}) = \widetilde{\boldsymbol{X}}_i - \widetilde{\boldsymbol{Y}}_i$ have the same distribution, i.e., $\boldsymbol{D}_i$ and $\widetilde{\boldsymbol{D}}_i$ have the same distribution. The above can also be carried out by assuming $\boldsymbol{\mu} = \mathbf{0}$ without loss of generality. Therefore, $\boldsymbol{D}(\pi)_i$ has the same distribution as $\boldsymbol{D}_i$. The test statistic $\mathbf{T}_1$ depends only on $\boldsymbol{D}_i$. It concludes that the distribution of $\mathbf{T}_1$ is invariant to the permutation π.

A.2 Permutation invariance of $\mathbf{T}_2$

The test $\mathbf{T}_2$ is based on the permutation of $\boldsymbol{X}_i - \boldsymbol{\mu}_X$ and $\boldsymbol{Y}_i - \boldsymbol{\mu}_Y$. Let $\boldsymbol{\mu}_X = \boldsymbol{\mu}_Y = \mathbf{0}$ without loss of generality for notational simplicity. Let $\mathrm{I}_1(\pi) = \{i : \pi(i) = 1\}$ be the index set of subjects whose $\boldsymbol{X}_i$ and $\boldsymbol{Y}_i$ are permuted, and $\mathrm{I}_0(\pi) = \{i : \pi(i) = 0\}$ analogous for those not permuted. Define the random variables

$$\mathbf{U}_1 = \begin{pmatrix} \sum_{i\in \mathrm{I}_1} \boldsymbol{X}_i \boldsymbol{X}_i^T & \sum_{i\in \mathrm{I}_1} \boldsymbol{X}_i \boldsymbol{Y}_i^T \\ \sum_{i\in \mathrm{I}_1} \boldsymbol{Y}_i \boldsymbol{X}_i^T & \sum_{i\in \mathrm{I}_1} \boldsymbol{Y}_i \boldsymbol{Y}_i^T \end{pmatrix},$$

$$\mathbf{U}_1(\pi) = \begin{pmatrix} \sum_{i\in \mathrm{I}_1} \boldsymbol{X}(\pi)_i \boldsymbol{X}(\pi)_i^T & \sum_{i\in \mathrm{I}_1} \boldsymbol{X}(\pi)_i \boldsymbol{Y}(\pi)_i^T \\ \sum_{i\in \mathrm{I}_1} \boldsymbol{Y}(\pi)_i \boldsymbol{X}(\pi)_i^T & \sum_{i\in \mathrm{I}_1} \boldsymbol{Y}(\pi)_i \boldsymbol{Y}(\pi)_i^T \end{pmatrix} = \begin{pmatrix} \sum_{i\in \mathrm{I}_1} \boldsymbol{Y}_i \boldsymbol{Y}_i^T & \sum_{i\in \mathrm{I}_1} \boldsymbol{Y}_i \boldsymbol{X}_i^T \\ \sum_{i\in \mathrm{I}_1} \boldsymbol{X}_i \boldsymbol{Y}_i^T & \sum_{i\in \mathrm{I}_1} \boldsymbol{X}_i \boldsymbol{X}_i^T \end{pmatrix}.$$

Under $\mathcal{H}_0 : \boldsymbol{\Omega}_X = \boldsymbol{\Omega}_Y$, we know that $\sum_{i\in \mathrm{I}_1} \boldsymbol{X}_i \boldsymbol{X}_i^T$ and $\sum_{i\in \mathrm{I}_1} \boldsymbol{Y}_i \boldsymbol{Y}_i^T$ have the same distribution, thus $\mathbf{U}_1(\pi)$ has same distribution as

$$\begin{pmatrix} \sum_{i\in \mathrm{I}_1} \boldsymbol{X}_i \boldsymbol{X}_i^T & \sum_{i\in \mathrm{I}_1} \boldsymbol{Y}_i \boldsymbol{X}_i^T \\ \sum_{i\in \mathrm{I}_1} \boldsymbol{X}_i \boldsymbol{Y}_i^T & \sum_{i\in \mathrm{I}_1} \boldsymbol{Y}_i \boldsymbol{Y}_i^T \end{pmatrix}.$$

We define $\mathbf{U}_0$ and $\mathbf{U}_0(\pi)$ for I_0 analogously. It is obvious that the distribution of $\mathbf{U}_0(\pi)$ is identical to that of $\mathbf{U}_0$. Consequently, $\mathbf{U}_0(\pi) + \mathbf{U}_1(\pi)$ has the same null distribution as

$$\begin{pmatrix} \sum_i \boldsymbol{X}_i \boldsymbol{X}_i^T & \sum_{i\in \mathrm{I}_0} \boldsymbol{X}_i \boldsymbol{Y}_i^T + \sum_{i\in \mathrm{I}_1} \boldsymbol{Y}_i \boldsymbol{X}_i^T \\ \sum_{i\in \mathrm{I}_0} \boldsymbol{Y}_i \boldsymbol{X}_i^T + \sum_{i\in \mathrm{I}_1} \boldsymbol{X}_i \boldsymbol{Y}_i^T & \sum_i \boldsymbol{Y}_i \boldsymbol{Y}_i^T \end{pmatrix}.$$

Since $\mathbf{T}_2(\pi)$ involves only the diagonal matrices of $\mathbf{U}_0(\pi) + \mathbf{U}_1(\pi)$, it follows that the null distribution of $\mathbf{T}_2$ is invariant to the permutation π.

Table 1: Empirical rejection probabilities for testing the equality of mean vectors ($\mathcal{H}_0 : \boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$) of multivariate paired populations at level $\alpha = 0.05$. $\mu_X = 0$; $\sigma_X^2 = 1$; $\rho_X = \rho_Y = 0.5$; $\rho_{XY} = \rho_{YX} = 0.3$; n is sample size; $\mathbf{T}_1$: proposed permutation test for equality of mean vectors; LRT: finite-sample likelihood ratio test; χ_5^2: asymptotic Chi-squared test of the likelihood ratio statistic with 5 degrees of freedom.

		$\mu_Y - \mu_X = 0$			$\mu_Y - \mu_X = 0.5$			$\mu_Y - \mu_X = 1$		
($r = \sigma_Y^2/\sigma_X^2$)		$r = 1$	$r = 1.5$	$r = 2$	$r = 1$	$r = 1.5$	$r = 2$	$r = 1$	$r = 1.5$	$r = 2$
		Multivariate Normal distributions								
$n = 15$	$\mathbf{T}_1$	0.05	0.04	0.05	0.25	0.14	0.16	0.80	0.55	0.51
	LRT	0.00	0.01	0.01	0.07	0.04	0.04	0.29	0.13	0.13
	χ_5^2	0.45	0.40	0.50	0.78	0.72	0.77	0.99	0.97	0.95
$n = 25$	$\mathbf{T}_1$	0.05	0.07	0.04	0.55	0.36	0.27	1.00	0.95	0.89
	LRT	0.02	0.02	0.01	0.31	0.22	0.14	0.97	0.80	0.67
	χ_5^2	0.20	0.20	0.23	0.82	0.69	0.66	1.00	1.00	0.99
$n = 50$	$\mathbf{T}_1$	0.05	0.05	0.04	0.93	0.78	0.63	1.00	1.00	1.00
	LRT	0.03	0.03	0.02	0.89	0.70	0.53	1.00	1.00	1.00
	χ_5^2	0.10	0.13	0.12	0.97	0.89	0.82	1.00	1.00	1.00
		Multivariate Bimodal Mixture of Normals								
$n = 15$	$\mathbf{T}_1$	0.03	0.04	0.02	0.20	0.17	0.14	0.78	0.61	0.48
	LRT	0.03	0.02	0.00	0.05	0.04	0.03	0.30	0.20	0.10
	χ_5^2	0.47	0.42	0.43	0.76	0.72	0.71	0.99	0.96	0.95
$n = 25$	$\mathbf{T}_1$	0.05	0.06	0.04	0.56	0.34	0.29	1.00	0.97	0.89
	LRT	0.02	0.03	0.01	0.32	0.16	0.16	0.96	0.82	0.76
	χ_5^2	0.20	0.20	0.24	0.85	0.68	0.65	1.00	0.99	0.98
$n = 50$	$\mathbf{T}_1$	0.06	0.05	0.05	0.92	0.77	0.62	1.00	1.00	1.00
	LRT	0.04	0.04	0.03	0.88	0.70	0.52	1.00	1.00	1.00
	χ_5^2	0.10	0.12	0.11	0.96	0.88	0.81	1.00	1.00	1.00
		Multivariate t_{10} distributions								
$n = 15$	$\mathbf{T}_1$	0.06	0.05	0.06	0.24	0.22	0.15	0.63	0.62	0.41
	LRT	0.01	0.00	0.01	0.05	0.06	0.05	0.27	0.25	0.12
	χ_5^2	0.47	0.49	0.56	0.80	0.75	0.75	0.94	0.92	0.89
$n = 25$	$\mathbf{T}_1$	0.07	0.04	0.03	0.33	0.30	0.27	0.68	0.60	0.58
	LRT	0.02	0.01	0.01	0.19	0.18	0.13	0.59	0.47	0.42
	χ_5^2	0.27	0.29	0.33	0.61	0.59	0.61	0.86	0.78	0.77
$n = 50$	$\mathbf{T}_1$	0.05	0.04	0.05	0.30	0.23	0.17	0.50	0.47	0.39
	LRT	0.03	0.01	0.02	0.24	0.15	0.12	0.44	0.39	0.32
	χ_5^2	0.13	0.20	0.24	0.43	0.42	0.39	0.66	0.59	0.57

Table 2: Empirical rejection probabilities for testing the equality of covariance matrices ($\mathcal{H}_0 : \mathbf{\Omega}_X = \mathbf{\Omega}_Y$) of multivariate paired populations at level $\alpha = 0.05$. $\mu_X = 0$; $\sigma_X^2 = 1$; $\rho_X = \rho_Y = 0.5$; $\rho_{XY} = \rho_{YX} = 0.3$; n is sample size; $\mathbf{T}_2$: proposed permutation test for equality of covariance matrices; LRT: finite-sample likelihood ratio test; χ^2_{15}: asymptotic Chi-squared test of the likelihood ratio statistic with 15 degrees of freedom.

		$\mu_Y - \mu_X = 0$			$\mu_Y - \mu_X = 0.5$			$\mu_Y - \mu_X = 1$		
($r = \sigma_Y^2/\sigma_X^2$)		$r = 1$	$r = 1.5$	$r = 2$	$r = 1$	$r = 1.5$	$r = 2$	$r = 1$	$r = 1.5$	$r = 2$
		Multivariate Normal distributions								
$n = 15$	$\mathbf{T}_2$	0.04	0.29	0.71	0.06	0.34	0.70	0.06	0.25	0.73
	LRT	0.00	0.00	0.09	0.00	0.00	0.01	0.00	0.00	0.01
	χ^2_{15}	0.40	0.51	0.68	0.36	0.51	0.67	0.39	0.45	0.71
$n = 25$	$\mathbf{T}_2$	0.06	0.53	0.94	0.04	0.54	0.94	0.05	0.52	0.93
	LRT	0.09	0.23	0.56	0.07	0.22	0.58	0.08	0.24	0.58
	χ^2_{15}	0.16	0.35	0.74	0.13	0.36	0.74	0.17	0.37	0.75
$n = 50$	$\mathbf{T}_2$	0.05	0.87	1.00	0.04	0.88	1.00	0.06	0.90	1.00
	LRT	0.08	0.47	0.95	0.06	0.49	0.95	0.06	0.50	0.96
	χ^2_{15}	0.08	0.51	0.96	0.08	0.53	0.96	0.08	0.53	0.97
		Multivariate Bimodal Mixture of Normals								
$n = 15$	$\mathbf{T}_2$	0.04	0.35	0.76	0.05	0.37	0.77	0.04	0.30	0.77
	LRT	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
	χ^2_{15}	0.36	0.45	0.66	0.36	0.46	0.67	0.30	0.48	0.66
$n = 25$	$\mathbf{T}_2$	0.07	0.63	0.95	0.07	0.63	0.96	0.03	0.63	0.98
	LRT	0.06	0.17	0.51	0.06	0.19	0.51	0.05	0.20	0.54
	χ^2_{15}	0.11	0.30	0.70	0.12	0.33	0.67	0.11	0.33	0.71
$n = 50$	$\mathbf{T}_2$	0.04	0.91	1.00	0.05	0.93	1.00	0.05	0.95	1.00
	LRT	0.05	0.38	0.96	0.05	0.45	0.95	0.04	0.40	0.96
	χ^2_{15}	0.06	0.41	0.98	0.05	0.48	0.97	0.06	0.43	0.98
		Multivariate t_{10} distributions								
$n = 15$	$\mathbf{T}_2$	0.03	0.25	0.61	0.07	0.30	0.65	0.05	0.26	0.67
	LRT	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.01
	χ^2_{15}	0.47	0.50	0.70	0.51	0.58	0.67	0.49	0.60	0.70
$n = 25$	$\mathbf{T}_2$	0.07	0.40	0.81	0.06	0.47	0.85	0.07	0.46	0.85
	LRT	0.15	0.27	0.50	0.13	0.25	0.47	0.14	0.22	0.48
	χ^2_{15}	0.34	0.52	0.72	0.31	0.52	0.74	0.34	0.47	0.73
$n = 50$	$\mathbf{T}_2$	0.08	0.55	0.91	0.06	0.57	0.90	0.06	0.54	0.91
	LRT	0.38	0.59	0.82	0.35	0.57	0.82	0.42	0.62	0.83
	χ^2_{15}	0.50	0.71	0.92	0.45	0.70	0.92	0.52	0.72	0.93

Table 3: Empirical rejection probabilities for testing the simultaneous equality of both mean vectors and covariance matrices ($\mathcal{H}_0 : \boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$ and $\boldsymbol{\Omega}_X = \boldsymbol{\Omega}_Y$) of multivariate paired populations at level $\alpha = 0.05$. $\mu_X = 0$; $\sigma_X^2 = 1$; $\rho_X = \rho_Y = 0.5$; $\rho_{XY} = \rho_{YX} = 0.3$; n is sample size; **CT**: proposed permutation test for simultaneous equality of both mean vectors and covariance matrices; LRT: finite-sample likelihood ratio test; χ^2_{20}: asymptotic Chi-squared test of the likelihood ratio statistic with 20 degrees of freedom.

		$\mu_Y - \mu_X = 0$			$\mu_Y - \mu_X = 0.5$			$\mu_Y - \mu_X = 1$		
($r = \sigma_Y^2/\sigma_X^2$)		$r = 1$	$r = 1.5$	$r = 2$	$r = 1$	$r = 1.5$	$r = 2$	$r = 1$	$r = 1.5$	$r = 2$
		Multivariate Normal distributions								
$n = 15$	**CT**	0.05	0.19	0.55	0.19	0.30	0.60	0.48	0.50	0.72
	LRT	0.00	0.02	0.02	0.01	0.03	0.04	0.04	0.03	0.10
	χ^2_{20}	0.49	0.56	0.74	0.66	0.71	0.85	0.92	0.90	0.93
$n = 25$	**CT**	0.06	0.43	0.88	0.43	0.58	0.89	0.99	0.94	0.97
	LRT	0.08	0.21	0.50	0.31	0.42	0.70	0.85	0.77	0.91
	χ^2_{20}	0.19	0.39	0.74	0.54	0.63	0.87	0.96	0.95	0.97
$n = 50$	**CT**	0.05	0.80	1.00	0.87	0.94	1.00	1.00	1.00	1.00
	LRT	0.06	0.43	0.92	0.69	0.83	0.99	1.00	1.00	1.00
	χ^2_{20}	0.06	0.48	0.96	0.75	0.89	0.99	1.00	1.00	1.00
		Multivariate Bimodal Mixture of Normals								
$n = 15$	**CT**	0.05	0.28	0.66	0.15	0.31	0.67	0.65	0.59	0.77
	LRT	0.00	0.00	0.01	0.01	0.02	0.02	0.02	0.05	0.05
	χ^2_{20}	0.46	0.53	0.72	0.66	0.67	0.80	0.89	0.88	0.92
$n = 25$	**CT**	0.07	0.50	0.93	0.42	0.61	0.94	1.00	0.97	1.00
	LRT	0.06	0.18	0.49	0.27	0.36	0.62	0.85	0.82	0.92
	χ^2_{20}	0.14	0.31	0.71	0.52	0.59	0.81	0.96	0.96	0.98
$n = 50$	**CT**	0.05	0.89	1.00	0.87	0.97	1.00	1.00	1.00	1.00
	LRT	0.06	0.34	0.95	0.67	0.83	0.98	1.00	1.00	1.00
	χ^2_{20}	0.07	0.39	0.96	0.72	0.87	0.99	1.00	1.00	1.00
		Multivariate t_{10} distributions								
$n = 15$	**CT**	0.04	0.20	0.49	0.21	0.29	0.55	0.54	0.60	0.68
	LRT	0.00	0.01	0.03	0.01	0.00	0.04	0.05	0.03	0.08
	χ^2_{20}	0.57	0.63	0.77	0.75	0.79	0.79	0.88	0.94	0.91
$n = 25$	**CT**	0.08	0.32	0.77	0.27	0.52	0.83	0.64	0.70	0.87
	LRT	0.20	0.27	0.51	0.33	0.38	0.59	0.56	0.58	0.70
	χ^2_{20}	0.36	0.54	0.76	0.57	0.69	0.83	0.81	0.83	0.91
$n = 50$	**CT**	0.09	0.48	0.89	0.27	0.54	0.86	0.46	0.65	0.89
	LRT	0.37	0.60	0.83	0.49	0.63	0.86	0.67	0.77	0.88
	χ^2_{20}	0.53	0.73	0.92	0.62	0.77	0.93	0.79	0.86	0.96

Table 4: Data analysis results of p-values for the tooth size study. $\mathbf{T}_1$: proposed permutation test for equality of mean vectors; $\mathbf{T}_2$: proposed permutation test for equality of covariance matrices; **CT**: proposed permutation test for simultaneous equality of both mean vectors and covariance matrices; LRT: finite-sample likelihood ratio test; χ^2_d: asymptotic Chi-squared test of the likelihood ratio statistic with d degrees of freedom.

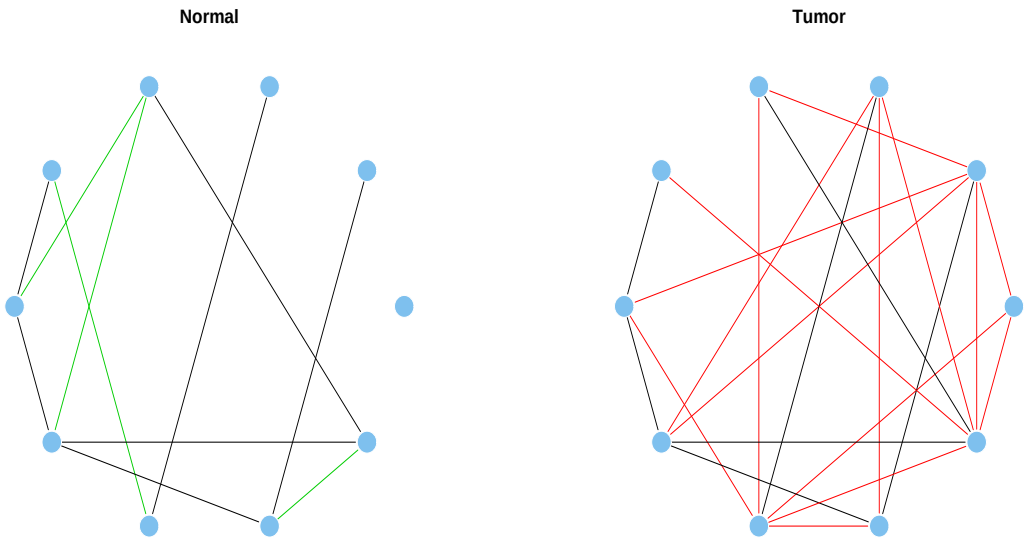
	$\mathcal{H}_0 : \boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$			$\mathcal{H}_0 : \boldsymbol{\Omega}_X = \boldsymbol{\Omega}_Y$			$\mathcal{H}_0 : \boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$ and $\boldsymbol{\Omega}_X = \boldsymbol{\Omega}_Y$		
p-value	$\mathbf{T}_1$	LRT	χ^2_7	$\mathbf{T}_2$	LRT	χ^2_{28}	**CT**	LRT	χ^2_{35}
Maxilla	.0003	.0013	.0002	.0722	.3067	.9538	.0006	.0802	.1368
Mandible	.0003	.0019	.0003	.6696	.0104	.0004	.0006	.0004	.0000

Table 5: Information on gene sets in the analysis

Gene set name	Description	Genes in prostate cancer data
GCM_FANCL	Neighborhood of FANCL (Fanconi anemia, complementation group L) in the GCM expression compendium	ZMYM2, ZFP14, CRNKL1, C12orf30, ZNF655, ANKRD17, DMTF1, NCOA5, USP37, REPIN1
GCM_CASP2	Neighborhood of CASP2 (caspase 2, apoptosis-related cysteine peptidase, neural precursor cell expressed, developmentally down-regulated 2) in the GCM expression compendium	METT11D1, DHX37, USP39, CCAR1, MLL, C17orf42, CS, SPPL2B, THRAP3
GNF2_ICAM3	Neighborhood of ICAM3 (intercellular adhesion molecule 3) in the GNF2 expression compendium	ITGB2, DAZAP2, TMEM127, GIT2, ZCCHC6,TXNIP, PTPRC, WAS, HLA-F, ARRB2, SH3BGRL3, RIN3

Table 6: Data analysis results of p-values for the gene study. $\mathbf{T}_1$: proposed permutation test for equality of mean vectors; $\mathbf{T}_2$: proposed permutation test for equality of covariance matrices; **CT**: proposed permutation test for simultaneous equality of both mean vectors and covariance matrices; LRT: finite-sample likelihood ratio test; χ^2: asymptotic Chi-squared test of the likelihood ratio statistic.

	$\mathcal{H}_0 : \boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$			$\mathcal{H}_0 : \boldsymbol{\Omega}_X = \boldsymbol{\Omega}_Y$			$\mathcal{H}_0 : \boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$ and $\boldsymbol{\Omega}_X = \boldsymbol{\Omega}_Y$		
p-value	$\mathbf{T}_1$	LRT	χ^2	$\mathbf{T}_2$	LRT	χ^2	**CT**	LRT	χ^2
GCM_FANCL	.0000	.0000	.0000	.0178	.0000	.0000	.0000	.0000	.0000
GCM_CASP2	.0000	.0000	.0000	.0219	.0000	.0000	.0000	.0000	.0000
GNF2_ICAM3	.0000	.0000	.0000	.1328	.0362	.0000	.0000	.0001	.0000



(a) normal (b) tumor

Figure 1: Estimated dependent structures of the 10 genes in gene set GCM_FANCL. The genes reported in Table 5 are ordered counter-clockwise as the nodes with the far right node being the first gene in the table. A line between two nodes means the two genes are conditionally dependent given all other genes.