Statistica Sinica **34** (2024), 229-256
doi:https://doi.org/10.5705/ss.202021.0369

A PAIRWISE HOTELLING METHOD FOR TESTING HIGH-DIMENSIONAL MEAN VECTORS

Zongliang Hu, Tiejun Tong* and Marc G. Genton

Shenzhen University, Hong Kong Baptist University and King Abdullah University of Science and Technology

Abstract: For high-dimensional data with a small sample size, we cannot use Hotelling's T^2 test to test the mean vectors because of the singularity problem in the sample covariance matrix. To overcome this problem, there are three main approaches but each has limitations and only works well in certain situations. Inspired by this, we propose a pairwise Hotelling method for testing high-dimensional mean vectors that provides a good balance between the existing approaches. To use the correlation information efficiently, we construct the new test statistics as the sum of Hotelling's test statistics for the covariate pairs with strong correlations and the squared t-statistics for the individual covariates that have little correlation with others. We further derive the asymptotic null distributions and power functions for the proposed tests under some regularity conditions. Numerical results show that our tests are able to control the type-I error rates and achieve a higher statistical power than that of existing methods, especially when the covariates are highly correlated. Two real-data examples are used to demonstrate the efficacy of our pairwise Hotelling's tests.

Key words and phrases: High-dimensional data, Hotelling's test, pairwise correlation, screening, statistical power, type-I error rate.

1. Introduction

A fundamental problem in multivariate statistics is to test whether a mean vector is equal to a given vector for the one-sample test, or to test whether two mean vectors are equal for the two-sample test. To start with, let $\boldsymbol{\mu}$ and Σ be the mean vector and covariance matrix, respectively, of a random vector $\boldsymbol{X}$. For the one-sample case, we are interested in testing the hypothesis

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \qquad \text{versus} \qquad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0, \tag{1.1}$$

where $\boldsymbol{\mu}_0 = (\mu_{01}, \ldots, \mu_{0p})^T$ is a given vector, p is the dimension, and the superscript T denotes the transpose of a vector or a matrix. Assume that $\boldsymbol{X}_k = (X_{k1}, \ldots, X_{kp})^T$, for $k = 1, \ldots, n$, are independent copies of $\boldsymbol{X} = (X_1, \ldots, X_p)^T$, where n is the sample size. Then to test hypothesis (1.1) under the assumption

*Corresponding author.

of data normality, the classical Hotelling's T^2 test (Hotelling (1931)) is

$$T^2 = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)^T S^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0),$$

where $\bar{\boldsymbol{X}} = \sum_{k=1}^{n} \boldsymbol{X}_k/n$ is the sample mean vector, and $S = \sum_{k=1}^{n}(\boldsymbol{X}_k - \bar{\boldsymbol{X}})(\boldsymbol{X}_k - \bar{\boldsymbol{X}})^T/(n-1)$ is the sample covariance matrix.

The era of big data has witnessed an increase in high-dimensional data, where the dimension is usually larger or much larger than the sample size. The resulting "large p small n" paradigm poses new challenges for testing problem (1.1). For example, when testing whether two gene sets, or pathways, have equal expression levels under two experimental conditions the number of genes (p) may be much larger than the number of samples (n). For high-dimensional data with a small sample size, Bai and Saranadasa (1996) show that we cannot apply Hotelling's T^2 test because of the singularity problem in the sample covariance matrix.

Several methods have been proposed to overcome this problem. There are three categories of approaches for handling the noninvertible sample covariance matrix:

(a) In the first category, researchers substitute the sample covariance matrix S with the $p \times p$ identity matrix I_p, leading to the *unscaled Hotelling's tests* (UHTs), with the test statistic

$$T^2_{\mathrm{UHT}} = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)^T(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0);$$

see, for example, Bai and Saranadasa (1996), Chen and Qin (2010), Wang, Peng and Qi (2013), Ahmad (2014), Ayyala, Park and Roy (2017), and Zhang et al. (2020). In addition, Xu et al. (2016) consider an adaptive testing procedure with the test statistic $T(\gamma) = \sum_{j=1}^{p}(\bar{X}_j - \mu_{0j})^{\gamma}$, and He et al. (2021) use the idea of a UHT to develop a unified U-statistic for testing mean vectors, covariance matrices, and regression coefficients.

(b) In the second category, researchers replace the sample covariance matrix with a diagonal covariance matrix, yielding the *diagonal Hotelling's tests* (DHTs). Specifically, by letting $D = \mathrm{diag}(S)$ be the diagonal covariance matrix, Wu, Genton and Stefanski (2006) introduce the test statistic

$$T^2_{\mathrm{DHT}} = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)^T D^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0),$$

and Srivastava and Du (2008) study the limiting behaviors of this test statistic under data normality. Cai, Liu and Xia (2014) consider a test based on the maximum of the squared marginal t-statistics, and Hu, Tong and Genton (2019) propose a likelihood ratio test based on a diagonal covariance matrix structure. Feng et al. (2017) assume a block diagonal structure for the covariance matrix, and apply Hotelling's T^2 within each block. Further studies on DHT include those of Srivastava (2009), Park and Ayyala (2013), Srivastava, Katayama and Kano (2013), Feng and Sun (2015), Feng et al. (2015), Gregory et al. (2015), Dong et al. (2016), Cao, Lin and Li (2018), Chen, Li and Zhong (2019), and Jiang and Li (2021).

(c) In the third category, researchers apply regularization methods to estimate the covariance matrix to overcome the singularity problem in the sample covariance matrix, yielding the *regularized Hotelling's tests* (RHTs). Here, Chen et al. (2011) propose a ridge-type regularization with the test statistic

$$T^2_{\mathrm{RHT},1} = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)^T (S + \lambda I_p)^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0).$$

This test statistic is also considered by Li et al. (2020) for the two-sample testing problem. Lopes, Jacob and Wainwright (2011) propose another regularized test statistic based on the random projection technique, namely,

$$T^2_{\mathrm{RHT},2} = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)^T P_R^T (P_R S P_R^T)^{-1} P_R (\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0),$$

where P_R is a random matrix of size $k \times p$. Further developments on projection-based techniques include, for example, Thulin (2014), Srivastava, Li and Ruppert (2016), and Zoh et al. (2018).

The tests in the first two categories do not account for correlations between the covariates, and thus may not provide a valid test with a controlled type-I error rate and/or acceptable statistical power. In contrast, the RHT in the third category is a universal method that attempts to account for all correlations within the covariance matrix. In other words, the ridge-type and projection-based statistics do not consider the sparsity of the covariance matrix. Consequently, the RHT may not provide satisfactory performance when the sample size n is small relative to the dimension p (Dong et al. (2016)). Li (2017) considers a composite Hotelling's test (CHT) to account for the correlations. The author extracts two-dimensional pairs $(X_i, X_j)^T$, with $i < j$, from the p-dimensional vector $\boldsymbol{X}$, and then takes the average of the classical Hotelling's test statistics for all the bivariate sub-vectors. When the covariance matrix is sparse and the sample size is small, a CHT may not provide satisfactory performance either. This is confirmed by Bickel and Levina (2004), who find that if the estimated correlations are very noisy because of the small sample size, it is probably better not to estimate them at all.

To overcome the drawbacks of the aforementioned tests, we propose a new category of testing methods to further advance the existing literature on testing high-dimensional mean vectors. Our main idea is to find a good balance between the second and third categories by leveraging the advantages of both of them. Specifically, to use the correlation information efficiently, we first construct the classical Hotelling's statistics for the covariate pairs with strong correlations. For individual covariates that have little correlation with others, we apply the squared t-statistics to account for their respective contributions to the multivariate testing problem. Our new test statistics are summations over all of the Hotelling's statistics and squared t-statistics. Consequently, they capture sufficient dependence information among the components, while also accounting for the sparsity of the covariance matrices. We further derive the asymptotic null distributions and power functions of the proposed statistics, and investigate the regularity conditions needed to establish their asymptotic results. The results of simulation studies and real-data analyses show that our proposed tests outperform existing methods in a wide range of settings.

The rest of the paper is organized as follows. In Section 2, we propose the pairwise Hotelling's testing method for the one-sample test, and derive the asymptotic distributions of the test statistic under the null and local alternative hypotheses. In Section 3, we propose the pairwise Hotelling's testing method for the two-sample test, and derive the asymptotic results, including the asymptotic null distribution and the power function. In Section 4, we conduct simulation studies to evaluate the proposed tests and compare them with existing methods. We then apply the proposed tests to two real-data examples in Section 5, and conclude the paper in Section 6 with a brief summary and suggestions for possible future work. All technical details are provided in online Supplementary Material.

2. One-Sample Test

In this section, we consider the one-sample testing problem (1.1) under the "large p small n" paradigm. Recall that we cannot use Hotelling's T^2 test when the dimension is larger than the sample size. To overcome the singularity problem, one possible approach is to downsize the dimension of the sample covariance matrix.

To achieve this, we decompose the p-dimensional vector $\boldsymbol{X}$ into a series of bivariate sub-vectors $(X_i, X_j)^T$, with $i < j$. We then apply the bivariate Hotelling's test statistic to account for their pairwise correlation as

$$\begin{aligned} T_{ij}^2 &= (\bar{X}_i - \mu_{0i}, \bar{X}_j - \mu_{0j}) \begin{pmatrix} s_{ii} & s_{ij} \\ s_{ji} & s_{jj} \end{pmatrix}^{-1} (\bar{X}_i - \mu_{0i}, \bar{X}_j - \mu_{0j})^T \\ &= (\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)^T P_{ij}^T (P_{ij} S P_{ij}^T)^{-1} P_{ij} (\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0), \end{aligned}$$

where $\bar{X}_i = \sum_{k=1}^n X_{ki}/n$ is the sample mean of the ith covariate, s_{ij} is the sample covariance of the ith and jth covariates, and $P_{ij} = \begin{pmatrix} 0 \cdots 1 \cdots 0 \cdots 0 \\ 0 \cdots 0 \cdots 1 \cdots 0 \end{pmatrix}$ is a $2 \times p$ matrix with the $(1, i)$ and $(2, j)$ components being one and all others being zero. Finally, we can apply the following U-type test statistic to accumulate all the pairwise correlations between the covariates:

$$W_1 = n\sum_{j=2}^{p}\sum_{i=1}^{j-1} T_{ij}^2 = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)^T \left(\sum_{j=2}^{p}\sum_{i=1}^{j-1} P_{ij}^T(P_{ij}SP_{ij}^T)^{-1}P_{ij}\right)(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0). \quad (2.1)$$

The pairwise idea W_1 can be traced back to the pairwise likelihood methods. For likelihood-based inference involving distributions with high-dimensional dependencies, applying the approximate likelihoods based on the bivariate marginal distributions can be a powerful approach (Cox and Reid (2004), Varin, Reid and Firth (2011), Li (2017)). Note that, as long as $n \geq 3$, the pairwise method in (2.1) is always applicable, and so it resolves the singularity problem in the original Hotelling's T^2 test.

2.1. Pairwise Hotelling's test statistic

For high-dimensional data, the covariance matrix is often sparse, with only a small proportion of non-zero correlations. In such settings, the U-type test statistic W_1 will include many noisy terms, and the test may not provide sufficiently large power, particularly when n is small relative to p.

To further improve the test statistic (2.1), we propose a thresholding method that shrinks the small estimates of correlations to zero to reduce the noise level in W_1. Specifically, we consider a screening procedure based on Kendall's tau correlation matrix, mainly because it is more robust than Pearson's correlation. Moreover, Kendall's tau correlation is a U-statistic, and so by Hoeffding's inequality, it can guarantee higher screening accuracy (Li et al. (2012); Zhang (2021)). Let $R = (r_{ij})_{1\leq i,j\leq p} \in \mathbb{R}^{p\times p}$ be Kendall's tau correlation matrix, and $\Gamma = (\tau_{ij})_{1\leq i,j\leq p} \in \mathbb{R}^{p\times p}$, with $\tau_{ij} = |r_{ij}|$, where $|\cdot|$ is the absolute value function. In addition, let

$$A_1 = \{(i,j) : \tau_{ij} > \tau_0 \text{ and } i < j\} \quad \text{and} \quad A_2 = \{i : \tau_{ij} < \tau_0 \text{ for all } j \neq i\}$$

be two sets of indices, where $\tau_0 \in [0,1]$ is a prespecified threshold. Clearly, covariate pairs with strong correlations fall into A_1, and individual covariates with little correlation with others fall into A_2. In practice, R, A_1, and A_2 are all unknown, and thus need to be estimated from the sample data.

Assume that $\hat{R}$ is Kendall's tau sample correlation matrix. Then, with a given τ_0, the sample estimates of A_1 and A_2 are, respectively,

$$\hat{A}_1 = \{(i,j) : \hat{\tau}_{ij} > \tau_0 \text{ and } i < j\} \quad \text{and} \quad \hat{A}_2 = \{i : \hat{\tau}_{ij} < \tau_0 \text{ for all } j \neq i\},$$

where $\hat{\tau}_{ij} = |\hat{r}_{ij}|$. In addition, let $\boldsymbol{X}_{ij;k} = (X_{ki}, X_{kj})^T \in \mathbb{R}^2$ be the kth sample of $(X_i, X_j)^T$, $\bar{\boldsymbol{X}}_{\{i,j\}}$ be the sample mean vector, and $S_{\{i,j\}}$ be the sample covariance matrix of $\boldsymbol{X}_{ij;k}$. Then, to test hypothesis (1.1), the thresholding test statistic can be represented as

$$W_1(\tau_0) = n \sum_{(i,j)\in\hat{A}_1} \left(\bar{\boldsymbol{X}}_{\{i,j\}} - \boldsymbol{\mu}_{0,\{i,j\}}\right)^T S_{\{i,j\}}^{-1} \left(\bar{\boldsymbol{X}}_{\{i,j\}} - \boldsymbol{\mu}_{0,\{i,j\}}\right) + n \sum_{i\in\hat{A}_2} \frac{(\bar{x}_i - \mu_{0i})^2}{s_{ii}},$$

where $\boldsymbol{\mu}_{0,\{i,j\}} = (\mu_{0i}, \mu_{0j})^T$. The test statistic $W_1(\tau_0)$ fully takes into account the pairwise correlations between the covariates. Specifically, we apply Hotelling's test statistics to account for the contributions from the covariate pairs with strong correlations (i.e., for all $(i,j) \in \hat{A}_1$), and apply squared t-statistics to account for the contributions from the individual covariates with little correlation with others (i.e., for all $i \in \hat{A}_2$).

Let $P_i = (0, \ldots, 1, \ldots, 0)$, where the ith component is one and all others are zero. Let $\widehat{P}_{\mathcal{O}} = \sum_{(i,j)\in\hat{A}_1} P_{ij}^T (P_{ij} S P_{ij}^T)^{-1} P_{ij} + \sum_{i\in\hat{A}_2} P_i^T (P_i S P_i^T)^{-1} P_i$. Using the new notation, we can rewrite $W_1(\tau_0)$ as

$$W_1(\tau_0) = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)^T \widehat{P}_{\mathcal{O}} (\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0).$$

For simplicity, we also let $P_{\mathcal{O}} = \sum_{(i,j)\in A_1} P_{ij}^T (P_{ij} \Sigma P_{ij}^T)^{-1} P_{ij} + \sum_{i\in A_2} P_i^T (P_i \Sigma P_i^T)^{-1} P_i$ be the unknown population value of $\widehat{P}_{\mathcal{O}}$. Note that $W_1(\tau_0)$ involves the terms $(\boldsymbol{X}_s - \boldsymbol{\mu}_0)^T \widehat{P}_{\mathcal{O}} (\boldsymbol{X}_s - \boldsymbol{\mu}_0)$, for $s = 1, \ldots, n$, which introduce higher-order moments in the centering and scaling parameters when establishing the limiting distributions. Hence, to stabilize the test statistic, we apply the leave-one-out method, as in Chen and Qin (2010), and propose the new test statistic

$$T_1(\tau_0) = \frac{1}{n(n-1)} \sum_{s=1}^{n} \sum_{t\neq s}^{n} (\boldsymbol{X}_s - \boldsymbol{\mu}_0)^T \widehat{P}_{\mathcal{O}}^{(s,t)} (\boldsymbol{X}_t - \boldsymbol{\mu}_0), \tag{2.2}$$

where $\widehat{P}_{\mathcal{O}}^{(s,t)} = \sum_{(i,j)\in\hat{A}_1} P_{ij}^T (P_{ij} S^{(s,t)} P_{ij}^T)^{-1} P_{ij} + \sum_{i\in\hat{A}_2} P_i^T (P_i S^{(s,t)} P_i^T)^{-1} P_i$, and $S^{(s,t)}$ is the sample covariance matrix without observations $\boldsymbol{X}_s$ and $\boldsymbol{X}_t$. We refer to the test statistic in (2.2) as the pairwise Hotelling's test (PHT) statistic. As a special case, if we set $\tau_0 = 1$, then $\hat{A}_1 = \emptyset$ and $\hat{A}_2 = \{1, \ldots, p\}$, in which case, the PHT statistic reduces to the diagonal Hotelling's test of Park and Ayyala (2013). On the other hand, if we set $\tau_0 = 0$, then $\hat{A}_1 = \{(i,j) : i < j\}$, for $i, j = 1, \ldots, p$, and $\hat{A}_2 = \emptyset$; that is, the PHT statistic accounts for all correlations in the covariance matrix, making it the same as W_1 in (2.1).

2.2. Asymptotic results

First, we show that the selected sets $\hat{A}_1$ and $\hat{A}_2$ based on the sample data are consistent estimates of A_1 and A_2, respectively, when the sample sizes tend to infinity; the proof is given in Appendix C.1.

Theorem 1. *Assume that* τ_0 *satisfies* $\liminf_{i,j=1,\ldots,p}\{\tau_{ij} | \tau_{ij} > \tau_0\} > \tau_0$ *and* $\limsup_{i,j=1,\ldots,p}\{\tau_{ij} | \tau_{ij} < \tau_0\} < \tau_0$. *Let* $\hat{A}_1$ *and* $\hat{A}_2$ *be the two sets based on the* *threshold τ_0 in the screening procedure. Then, for any given positive integer m_0, if $p = O(n^{m_0})$, we have*

$$P(\hat{A}_2 = A_2) \geq P(\hat{A}_1 = A_1) \to 1 \quad \text{as} \quad (n, p) \to \infty.$$

Next, following the assumptions in Chen and Qin (2010), we assume that the random vector $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ follows the linear model

$$\boldsymbol{X} = C_1 \boldsymbol{Z} + \boldsymbol{\mu}, \tag{2.3}$$

where $C_1 \in \mathbb{R}^{p \times q}$, with $q \geq p$, such that $\Sigma = C_1 C_1^T$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^T$, and the random vector $\boldsymbol{Z}$ satisfies $E(\boldsymbol{Z}) = \boldsymbol{0}$ and $\mathrm{Var}(\boldsymbol{Z}) = I_q$. In addition for $\boldsymbol{Z} = (Z_1, \ldots, Z_q)^T$, we assume that the following moment conditions hold: $E(Z_j^4) = 3 + \Delta_1 < \infty$, where Δ_1 is a positive constant, and

$$E(Z_{l_1}^{\alpha_1} Z_{l_2}^{\alpha_2} \cdots Z_{l_k}^{\alpha_k}) = E(Z_{l_1}^{\alpha_1}) E(Z_{l_2}^{\alpha_2}) \cdots E(Z_{l_k}^{\alpha_k}),$$

where k is a positive integer such that $\alpha_1 + \cdots + \alpha_k \leq 8$, and $l_1 \neq l_2 \neq \cdots \neq l_k$.

We further assume that $\{(X_i, X_j) : i, j = 1, 2, \ldots, p \text{ with } i \neq j\}$ is a two-dimensional random field, and define the ρ-mixing coefficient for $X = \{X_j, j = 1, 2, \ldots, p\}$ as

$$\rho(s) = \sup\Big\{|\mathrm{Corr}(g_1, g_2)| : g_1 \in \mathcal{L}_2(X(A_3)), g_2 \in \mathcal{L}_2(X(A_4)), \mathrm{dist}(A_3, A_4) \geq s\Big\}$$

over any possible sets $A_3, A_4 \subset \{1, 2, \ldots, p\}$, with $\mathrm{card}(A_3) \leq 2$ and $\mathrm{card}(A_4) \leq 2$, where $\mathrm{card}(\cdot)$ is an operator that counts the number of elements in a given set, $\mathrm{dist}(A_3, A_4) = \min_{i \in A_3, j \in A_4} |i - j|$ is the distance between A_3 and A_4, $\mathrm{Corr}(g_1, g_2)$ is the correlation between g_1 and g_2, and $\mathcal{L}_2\{X(E)\}$ is the set of all measurable functions defined on the σ-algebra generated by X over $E \subset \{1, 2, \ldots, p\}$ with the existence of the second moment.

To establish the asymptotic null and alternative distributions of the proposed test statistic, we also need the following conditions:

(C1) There exists a finite positive number $\bar{K}_1$ such that $1/\bar{K}_1 \leq \lambda_p(\Sigma) \leq \cdots \leq \lambda_1(\Sigma) \leq \bar{K}_1$, where $\lambda_i(\Sigma)$ is the ith largest eigenvalue of Σ.

(C2) Assume that $\{X_j : j \geq 1\}$ is a ρ-mixing sequence such that $\rho(s) \leq \varpi_0 \exp(-s)$, where $\varpi_0 > 0$ is a constant.

(C3) There exists an oracle constant $\tau^* \in (0, 1)$ such that, for a finite positive integer K_0, $\sup_{i \leq p} \mathrm{card}(A_i^*) \leq K_0$, where $A_i^* = \{j : \tau_{ij} > \tau^*\}$. In addition, we assume that $\liminf_{i,j=1,\ldots,p}\{\tau_{ij} | \tau_{ij} > \tau^*\} > \tau^*$ and $\limsup_{i,j=1,\ldots,p}\{\tau_{ij} | \tau_{ij} < \tau^*\} < \tau^*$.

(C4) There exists a positive integer $m_0 > 4$ such that the higher-order moments $E(X_1^{4m_0+2}), \ldots, E(X_p^{4m_0+2})$ are bounded uniformly, indicating that there exists a constant $\varpi_1 > 0$ such that $E(X_{kj}^{4m_0+2}) < \varpi_1$ holds for $j = 1, \ldots, p$. In addition, we assume that $E\big\|S_{\{i,j\}}^{-1}\big\|^8$ for $(i,j) \in A_1$ and $E(s_{jj}^{-8})$ for $j \in A_2$ are bounded uniformly, where $\|\cdot\|$ is the Frobenius norm.

(C5) Assume that $\boldsymbol{\mu}^T P_{\mathcal{O}} \boldsymbol{\mu} = o(\sqrt{p/n})$. There exists a constant $\varpi_2 > 0$ such that $|\mu_j - \mu_{0j}|^2 \le \varpi_2/\sqrt{n}$.

Condition (C1) assumes that the eigenvalues are bounded uniformly away from zero and infinity, which is the same condition as in Cai, Liu and Xia (2014) and Xu et al. (2016). Condition (C2) is the so-called ρ-mixing condition, which follows from Lin and Lu (1997) and implies a weak dependence structure of the data, commonly assumed in many genome-wide association studies. For example, single nucleotide polymorphisms (SNPs) have a local dependence structure in which the correlations between SNPs often decay rapidly as the distances between the gene loci increase. Condition (C3) assumes that our PHT statistic allows the number of covariate pairs with strong correlations to increase at the same order of p. Conditions (C4) and (C5) are technical conditions needed to derive the asymptotic results of the proposed test statistic.

Theorem 2. *Assume that* $\tau_0 \ge \tau^*$, $\liminf_{i,j=1,\ldots,p}\{\tau_{ij}|\tau_{ij} > \tau_0\} > \tau_0$, *and* $\limsup_{i,j=1,\ldots,p}\{\tau_{ij}|\tau_{ij} < \tau_0\} < \tau_0$. *Then, under model* (2.3) *and conditions* (C1)–(C5), *if* $p = o(n^{(m_0-3)/2})$ *with* m_0, *as defined in* (C4), *we have*

$$\frac{T_1(\tau_0) - \boldsymbol{\delta}_1^T P_{\mathcal{O}} \boldsymbol{\delta}_1}{\sqrt{2n^{-2}\mathrm{tr}(\Lambda_1^2)}} \xrightarrow{D} N(0,1) \quad \text{as} \quad (n,p) \to \infty,$$

where $\boldsymbol{\delta}_1 = \boldsymbol{\mu} - \boldsymbol{\mu}_0$, $\Lambda_1 = \Sigma^{1/2} P_{\mathcal{O}} \Sigma^{1/2}$, $\mathrm{tr}(\cdot)$ *is the trace function, and* $\xrightarrow{D}$ *denotes convergence in distribution.*

The proof of Theorem 2 is given in Appendix C.2. This theorem shows that, despite not knowing the exact threshold τ^*, we can select a larger threshold $\tau_0 > \tau^*$, such that if τ_0 satisfies $\liminf_{i,j=1,\ldots,p}\{\tau_{ij}|\tau_{ij} > \tau_0\} > \tau_0$ and $\liminf_{i,j=1,\ldots,p}\{\tau_{ij}|\tau_{ij} < \tau_0\} < \tau_0$, then the test statistic $T_1(\tau_0)$ still converges to the standard normal distribution after proper centering and scaling.

To apply Theorem 2, we need a ratio-consistent estimator for the unknown $\mathrm{tr}(\Lambda_1^2)$. For this purpose, we establish the following lemma, with the proof provided in Appendix C.3.

Lemma 1. *Assume that* τ_0 *satisfies the assumptions in Theorem* 2. *Then, under model* (2.3) *and conditions* (C1)–(C5), *we have that*

(*i*) *if* $p = o(n^3)$, *then*

$$\widehat{\mathrm{tr}(\Lambda_1^2)} = \frac{1}{n(n-1)} \sum_{s \ne t}^{n} (\boldsymbol{X}_s - \bar{\boldsymbol{X}}^{(s,t)})^T \widehat{P}_{\mathcal{O}}^{(s,t)} \boldsymbol{X}_t (\boldsymbol{X}_t - \bar{\boldsymbol{X}}^{(s,t)})^T \widehat{P}_{\mathcal{O}}^{(s,t)} \boldsymbol{X}_s$$

is a ratio-consistent estimator of $\mathrm{tr}(\Lambda_1^2)$*, where* $\bar{\boldsymbol{X}}^{(s,t)}$ *is the sample mean vector without observations* $\boldsymbol{X}_s$ *and* $\boldsymbol{X}_t$*;*

(*ii*) *if* $p = o(\min(n^3, n^{(m_0-3)/2}))$*, then under the null hypothesis of* (1.1)*,*

$$\frac{T_1(\tau_0)}{\sqrt{2n^{-2}\widehat{\mathrm{tr}(\Lambda_1^2)}}} \xrightarrow{D} N(0,1) \quad \text{as} \quad (n,p) \to \infty.$$

By Theorem 2, the power function of the PHT statistic for the one-sample test is

$$\mathrm{Power}(\boldsymbol{\delta}_1) = \Phi\biggl(-z_\alpha + \frac{\boldsymbol{\delta}_1^T P_{\mathcal{O}} \boldsymbol{\delta}_1}{\sqrt{2n^{-2}\mathrm{tr}(\Lambda_1^2)}}\biggr), \tag{2.4}$$

where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution. The performance of the new test depends on the quantities $\boldsymbol{\delta}_1^T P_{\mathcal{O}} \boldsymbol{\delta}_1$ and $\mathrm{tr}(\Lambda_1^2)$. Theoretically, a reasonable choice of the threshold τ_0 maximizes $\mathrm{Power}(\boldsymbol{\delta}_1)$ so that PHT achieves the highest asymptotic power. However, this maximization procedure is infeasible in practice, because $\boldsymbol{\delta}_1^T P_{\mathcal{O}} \boldsymbol{\delta}_1/\sqrt{\mathrm{tr}(\Lambda_1^2)}$ involves unknown quantities $\boldsymbol{\delta}_1$ and Σ. We further examine a practical choice of τ_0 in Section 4.3.

3. Two-Sample Test

This section considers the two-sample test for mean vectors with equal covariance matrices. Let $\{\boldsymbol{X}_s = (X_{s1}, \ldots, X_{sp})^T\}_{s=1}^{n_1}$ and $\{\boldsymbol{Y}_t = (Y_{t1}, \ldots, Y_{tp})^T\}_{t=1}^{n_2}$ be two groups of independent and identically distributed (i.i.d.) random vectors from two independent multivariate populations. Furthermore, let $E(\boldsymbol{X}_s) = \boldsymbol{\mu}_1 = (\mu_{11}, \ldots, \mu_{1p})^T$ be the mean vector of the first population, $E(\boldsymbol{Y}_t) = \boldsymbol{\mu}_2 = (\mu_{21}, \ldots, \mu_{2p})^T$ be the mean vector of the second population, and Σ be the common covariance matrix for both populations. For the two-sample test, we are interested in testing the hypothesis

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{versus} \quad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2. \tag{3.1}$$

3.1. Pairwise Hotelling's test statistic

Following similar notation as that for the one-sample test, we let Kendall's tau correlation matrix be $R = (r_{ij})_{1\le i,j\le p} \in \mathbb{R}^{p\times p}$, and $\Gamma = (\tau_{ij})_{1\le i,j\le p} \in \mathbb{R}^{p\times p}$, with $\tau_{ij} = |r_{ij}|$. Furthermore, let

$$A_1 = \{(i,j) : \tau_{ij} > \tau_0 \text{ and } i < j\} \quad \text{and} \quad A_2 = \{i : \tau_{ij} < \tau_0 \text{ for all } j \neq i\}$$

be two sets of indices, where $\tau_0 \in [0,1]$ is a prespecified threshold, and denote

$$P_{\mathcal{O}} = \sum_{(i,j)\in A_1} P_{ij}^T (P_{ij}\Sigma P_{ij}^T)^{-1} P_{ij} + \sum_{i\in A_2} P_i^T (P_i \Sigma P_i^T)^{-1} P_i.$$

Assume that $\hat{R}_1 = (\hat{r}_{ij,1})_{1\le i,j\le p} \in \mathbb{R}^{p\times p}$ and $\hat{R}_2 = (\hat{r}_{ij,2})_{1\le i,j\le p} \in \mathbb{R}^{p\times p}$ are the respective Kendall's tau sample correlation matrices for the two groups. For simplicity, let $N = n_1 + n_2$, and assume $n_1/N \to \varphi_0 \in (0, 1)$ as $N \to \infty$. Then, with a given τ_0, the sample estimates of A_1 and A_2 are, respectively,

$$\hat{A}_1 = \{(i, j) : \hat{\tau}_{ij} > \tau_0 \text{ and } i < j\} \quad \text{and} \quad \hat{A}_2 = \{i : \hat{\tau}_{ij} < \tau_0 \text{ for all } j \neq i\},$$

where $\hat{\tau}_{ij} = (n_1\hat{\tau}_{ij,1} + n_2\hat{\tau}_{ij,2})/N$, $\hat{\tau}_{ij,1} = |\hat{r}_{ij,1}|$, and $\hat{\tau}_{ij,2} = |\hat{r}_{ij,2}|$. In addition, we need the following notation related to the sample covariance matrices:

(a) Let S_1 (or S_2) be the sample covariance matrix of group 1 (or group 2), $S_1^{(s)}$ (or $S_2^{(s)}$) be the sample covariance matrix of group 1 (or group 2) without observation $\boldsymbol{X}_s$ (or $\boldsymbol{Y}_s$), and $S_1^{(s,t)}$ (or $S_2^{(s,t)}$) be the sample covariance matrix of group 1 (or group 2) without observations $\boldsymbol{X}_s$ and $\boldsymbol{X}_t$ (or $\boldsymbol{Y}_s$ and $\boldsymbol{Y}_t$).

(b) Let $s_{1,jj}$ (or $s_{2,jj}$) be the sample variance of X_{kj} (or Y_{kj}), and $S_{1,\{ij\}}$ and $S_{2,\{ij\}}$ be the sample covariance matrices of $(X_{ki}, X_{kj})^T$ and $(Y_{ki}, Y_{kj})^T$, respectively. Furthermore, let $s_{1,jj}^{(s,t)}$ (or $s_{2,jj}^{(s,t)}$) be the sample variance of X_{kj} (or Y_{kj}) without observations X_{sj} and X_{tj} (or Y_{sj} and Y_{tj}).

(c) Let $S_{1*}^{(s,t)} = [(n_1 - 2)S_1^{(s,t)} + n_2 S_2]/(N - 2)$ be the pooled sample covariance matrix without observations $\boldsymbol{X}_s$ and $\boldsymbol{X}_t$ in group 1, and $S_{2*}^{(s,t)} = [n_1 S_1 + (n_2 - 2)S_2^{(s,t)}]/(N - 2)$ be the pooled sample covariance matrix without observations $\boldsymbol{Y}_s$ and $\boldsymbol{Y}_t$ in group 2.

(d) Let $S_{12} = \{(n_1 - 1)S_1 + (n_2 - 1)S_2\}/(N - 2)$ be the pooled sample covariance matrix of the two groups, and $S_{12,*}^{(s,t)} = \{(n_1 - 1)S_1^{(s)} + (n_2 - 1)S_2^{(t)}\}/(N - 2)$ be the pooled sample covariance matrices without $\boldsymbol{X}_s$ and $\boldsymbol{Y}_t$ in groups 1 and 2, respectively.

Following similar arguments as in (2.1), we propose the following U-type test statistic for the two-sample test:

$$W_2 = \frac{n_1 + n_2}{n_1 n_2}(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})^T \left\{ \sum_{j=2}^{p} \sum_{i=1}^{j-1} P_{ij}^T (P_{ij} S_{12} P_{ij}^T)^{-1} P_{ij} \right\} (\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}), \tag{3.2}$$

where $\bar{\boldsymbol{X}}$ and $\bar{\boldsymbol{Y}}$ are the sample mean vectors of the two groups. In addition, using the screening procedure and the leave-one-out method, our PHT statistic for the two-sample test is given by

$$\begin{aligned} T_2(\tau_0) = {} & \frac{1}{n_1(n_1 - 1)} \sum_{s=1}^{n_1} \sum_{t\neq s}^{n_1} \boldsymbol{X}_s^T \widehat{P}_{1,\mathcal{O}}^{(s,t)} \boldsymbol{X}_t + \frac{1}{n_2(n_2 - 1)} \sum_{s=1}^{n_2} \sum_{t\neq s}^{n_2} \boldsymbol{Y}_s^T \widehat{P}_{2,\mathcal{O}}^{(s,t)} \boldsymbol{Y}_t \\ & - \frac{2}{n_1 n_2} \sum_{s=1}^{n_1} \sum_{t=1}^{n_2} \boldsymbol{X}_s^T \widehat{P}_{12,\mathcal{O}}^{(s,t)} \boldsymbol{Y}_t, \end{aligned} \tag{3.3}$$

where $\widehat{P}^{(s,t)}_{1,\mathcal{O}}$, $\widehat{P}^{(s,t)}_{2,\mathcal{O}}$, and $\widehat{P}^{(s,t)}_{12,\mathcal{O}}$ are three sample-based estimates of $P_{\mathcal{O}}$, with

$$\widehat{P}^{(s,t)}_{1,\mathcal{O}} = \sum_{(i,j)\in\hat{A}_1} P_{ij}^T(P_{ij}S^{(s,t)}_{1*}P_{ij}^T)^{-1}P_{ij} + \sum_{i\in\hat{A}_2} P_i^T(P_iS^{(s,t)}_{1*}P_i^T)^{-1}P_i,$$

$$\widehat{P}^{(s,t)}_{2,\mathcal{O}} = \sum_{(i,j)\in\hat{A}_1} P_{ij}^T(P_{ij}S^{(s,t)}_{2*}P_{ij}^T)^{-1}P_{ij} + \sum_{i\in\hat{A}_2} P_i^T(P_iS^{(s,t)}_{2*}P_i^T)^{-1}P_i,$$

$$\widehat{P}^{(s,t)}_{12,\mathcal{O}} = \sum_{(i,j)\in\hat{A}_1} P_{ij}^T(P_{ij}S^{(s,t)}_{12,*}P_{ij}^T)^{-1}P_{ij} + \sum_{i\in\hat{A}_2} P_i^T(P_iS^{(s,t)}_{12,*}P_i^T)^{-1}P_i.$$

When $\tau_0 = 1$, we have $\hat{A}_1 = \emptyset$ and $\hat{A}_2 = \{1,\ldots,p\}$, so that the PHT statistic reduces to the diagonal Hotelling's test in Park and Ayyala (2013). In contrast, when $\tau_0 = 0$, we have $\hat{A}_1 = \{(i,j) : i < j\}$, for $i,j = 1,\ldots,p$ and $\hat{A}_2 = \emptyset$. Thus, the PHT statistic is the U-type test statistic (3.2) for the two-sample test.

3.2. Asymptotic results

First, we show that the selected sets $\hat{A}_1$ and $\hat{A}_2$ based on the sample data converge to A_1 and A_2, respectively, when the sample sizes tend to infinity; see Appendix D.1 for the proof.

Theorem 3. *Assume that τ_0 satisfies $\liminf_{i,j=1,\ldots,p}\{\tau_{ij}|\tau_{ij} > \tau_0\} > \tau_0$ and $\limsup_{i,j=1,\ldots,p}\{\tau_{ij}|\tau_{ij} < \tau_0\} < \tau_0$. Let $\hat{A}_1$ and $\hat{A}_2$ be the two sets based on the threshold τ_0 in the screening procedure. Then, for any given positive integer m_0, if $p = O(N^{m_0})$, we have*

$$P(\hat{A}_2 = A_2) \geq P(\hat{A}_1 = A_1) \to 1 \quad \text{as} \quad (N,p) \to \infty.$$

For ease of notation, we assume that the random vectors $\boldsymbol{X} = (X_1,\ldots,X_p)^T$ and $\boldsymbol{Y} = (Y_1,\ldots,Y_p)^T$ follow the two models

$$\boldsymbol{X} = C_2\boldsymbol{Z}^{(1)} + \boldsymbol{\mu}_1 \quad \text{and} \quad \boldsymbol{Y} = C_2\boldsymbol{Z}^{(2)} + \boldsymbol{\mu}_2, \tag{3.4}$$

respectively, where $C_2 \in \mathbb{R}^{p\times q}$, with $q \geq p$, such that $\Sigma = C_2C_2^T$, and the random vector $\boldsymbol{Z}^{(i)}$ satisfies $E(\boldsymbol{Z}^{(i)}) = \boldsymbol{0}$ and $\mathrm{Var}(\boldsymbol{Z}^{(i)}) = I_q$, for $i = 1,2$. In addition, we assume that the following moment conditions hold: $E\big(Z_j^{(i)}\big)^4 = 3 + \Delta_2 < \infty$, where Δ_2 is a positive constant, and

$$E\left\{(Z^{(i)}_{l_1})^{\alpha_1}(Z^{(i)}_{l_2})^{\alpha_2}\cdots(Z^{(i)}_{l_k})^{\alpha_k}\right\} = E\Big\{(Z^{(i)}_{l_1})^{\alpha_1}\Big\}E\Big\{(Z^{(i)}_{l_2})^{\alpha_2}\Big\}\cdots E\Big\{(Z^{(i)}_{l_k})^{\alpha_k}\Big\}, \tag{3.5}$$

where k a positive integer such that $\alpha_1 + \cdots + \alpha_k \leq 8$, and $l_1 \neq l_2 \neq \cdots \neq l_k$.

We further assume that $\{(X_i,X_j) : i,j = 1,2,\ldots,p \text{ with } i \neq j\}$ and $\{(Y_i,Y_j) : i,j = 1,2,\ldots,p \text{ with } i \neq j\}$ are two random fields. Analogous to conditions (C1)–(C5), to derive the asymptotic properties of the two-sample PHT statistic, we need the following conditions:

(C1$'$) There exists a finite positive number $\bar{K}_2$ such that $1/\bar{K}_2 \leq \lambda_p(\Sigma) \leq \cdots \leq \lambda_1(\Sigma) \leq \bar{K}_2$.

(C2$'$) Assume that $\{X_j : j \geq 1\}$ and $\{Y_j : j \geq 1\}$ are two ρ-mixing sequences, with the corresponding ρ-mixing coefficients $\rho_X(s)$ and $\rho_Y(s)$, respectively. There exists a constant $\varpi_3 > 0$ such that $\rho_X(s) \leq \varpi_3 \exp(-s)$ and $\rho_Y(s) \leq \varpi_3 \exp(-s)$.

(C3$'$) There exists an oracle constant $\tau^* > 0$ such that, for a finite positive integer K_0, $\sup_{i \leq p} \text{card}(A_i^*) \leq K_0$, where $A_i^* = \{j : \tau_{ij} > \tau^*\}$. In addition, we assume that $\liminf_{i,j=1,\ldots,p}\{\tau_{ij} | \tau_{ij} > \tau^*\} > \tau^*$ and $\limsup_{i,j=1,\ldots,p}\{\tau_{ij} | \tau_{ij} < \tau^*\} < \tau^*$.

(C4$'$) There exists a positive integer $m_0 > 4$ such that the higher-order moments $E(X_j^{4m_0+2})$ and $E(Y_j^{4m_0+2})$ are bounded uniformly for $j = 1, \ldots, p$. In addition, we assume that $E\big\|S_{1,\{ij\}}^{-1}\big\|^8$ and $E\big\|S_{2,\{ij\}}^{-1}\big\|^8$ are bounded uniformly for $(i,j) \in A_1$, and $E(s_{1,jj}^{-8})$ and $E(s_{2,jj}^{-8})$ are bounded uniformly for $j \in A_2$.

(C5$'$) Assume that $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T P_{\mathcal{O}} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = o(\sqrt{p/N})$ and $\boldsymbol{\mu}_1^T P_{\mathcal{O}} \boldsymbol{\mu}_1 = o(\sqrt{p/N})$. There exists a constant $\varpi_4 > 0$ such that $\mu_{1j}^2 + \mu_{2j}^2 \leq \varpi_4/\sqrt{N}$.

Note that conditions (C1$'$)–(C5$'$) are analogous to conditions (C1)–(C5), respectively. Condition (C1$'$) assumes that the eigenvalues are bounded uniformly away from zero and infinity. Condition (C2$'$) implies a weak dependence structure among the data. Condition (C3$'$) assumes that our PHT statistic allows the number of covariate pairs with strong correlations to increase at the same order of p. Conditions (C4$'$) and (C5$'$) are technical conditions to derive the asymptotic results of the proposed test statistic.

Theorem 4. *Assume that $\tau_0 \geq \tau^*$, $\liminf_{i,j=1,\ldots,p}\{\tau_{ij} | \tau_{ij} > \tau_0\} > \tau_0$, and $\limsup_{i,j=1,\ldots,p}\{\tau_{ij} | \tau_{ij} < \tau_0\} < \tau_0$. Then, under model* (3.4) *and conditions* (C1$'$)–(C5$'$), *if $p = o(N^{(m_0-3)/2})$ with m_0 as defined in* (C4$'$), *we have*

$$\frac{T_2(\tau_0) - \boldsymbol{\delta}_2^T P_{\mathcal{O}} \boldsymbol{\delta}_2}{\sqrt{\phi(n_1, n_2)\text{tr}(\Lambda_1^2)}} \xrightarrow{D} N(0,1) \quad \text{as} \quad (N, p) \to \infty,$$

where $\boldsymbol{\delta}_2 = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$ and $\phi(n_1, n_2) = 2/\{n_1(n_1-1)\} + 2/\{n_2(n_2-1)\} + 4/(n_1 n_2)$.

The proof of Theorem 4 is given in Appendix D.2. This theorem shows that, for a larger threshold $\tau_0 > \tau^*$, if τ_0 satisfies $\liminf_{i,j=1,\ldots,p}\{\tau_{ij} | \tau_{ij} > \tau_0\} > \tau_0$ and $\liminf_{i,j=1,\ldots,p}\{\tau_{ij} | \tau_{ij} < \tau_0\} < \tau_0$, then the test statistic $T_2(\tau_0)$ still converges to the standard normal distribution, after proper centering and scaling. Hence, despite not knowing the exact threshold τ^* that satisfies condition (C3$'$), we can always select a larger threshold when performing the test.

To apply Theorem 4, we have the following lemma that derives a ratio-consistent estimator for $\mathrm{tr}(\Lambda_1^2)$; the proof is given in Appendix D.3.

Lemma 2. *Assume that τ_0 satisfies the assumptions in Theorem 4. Under model (3.4) and conditions (C1′)–(C5′), we have that*

(i) if $p = o(N^3)$, then

$$\widehat{\mathrm{tr}(\Lambda_1^2)} = \frac{1}{2n_1(n_1-1)} \sum_{s=1}^{n_1} \sum_{t \neq s}^{n_1} (\boldsymbol{X}_s - \bar{\boldsymbol{X}}^{(s,t)})^T \widehat{P}_{1,\mathcal{O}}^{(s,t)} \boldsymbol{X}_t (\boldsymbol{X}_t - \bar{\boldsymbol{X}}^{(s,t)})^T \widehat{P}_{1,\mathcal{O}}^{(s,t)} \boldsymbol{X}_s$$
$$+ \frac{1}{2n_2(n_2-1)} \sum_{s=1}^{n_2} \sum_{t \neq s}^{n_2} (\boldsymbol{Y}_s - \bar{\boldsymbol{Y}}^{(s,t)})^T \widehat{P}_{2,\mathcal{O}}^{(s,t)} \boldsymbol{Y}_t (\boldsymbol{Y}_t - \bar{\boldsymbol{Y}}^{(s,t)})^T \widehat{P}_{2,\mathcal{O}}^{(s,t)} \boldsymbol{Y}_s$$

is a ratio-consistent estimator of $\mathrm{tr}(\Lambda_1^2)$, where $\bar{\boldsymbol{X}}^{(s,t)}$ (or $\bar{\boldsymbol{Y}}^{(s,t)}$) is the sample mean vector of group 1 (or group 2) without observations $\boldsymbol{X}_s$ and $\boldsymbol{X}_t$ (or $\boldsymbol{Y}_s$ and $\boldsymbol{Y}_t$);

(ii) if $p = o\{\min(N^3, N^{(m_0-3)/2})\}$, then under the null hypothesis of (3.1),

$$\frac{T_2(\tau_0)}{\sqrt{\phi(n_1,n_2)\widehat{\mathrm{tr}(\Lambda_1^2)}}} \xrightarrow{D} N(0,1) \quad \text{as} \quad (N,p) \to \infty.$$

By Theorem 4, the power function of the PHT statistic for the two-sample test is

$$\mathrm{Power}(\boldsymbol{\delta}_2) = \Phi\left\{ -z_\alpha + \frac{\boldsymbol{\delta}_2^T P_{\mathcal{O}} \boldsymbol{\delta}_2}{\sqrt{\phi(n_1,n_2)\mathrm{tr}(\Lambda_1^2)}} \right\}. \tag{3.6}$$

The performance of the new test depends on the quantities $\boldsymbol{\delta}_2^T P_{\mathcal{O}} \boldsymbol{\delta}_2$ and $\mathrm{tr}(\Lambda_1^2)$. Theoretically, for the PHT statistic to achieve the highest asymptotic power, a reasonable choice for the threshold τ_0 is to maximize $\mathrm{Power}(\boldsymbol{\delta}_2)$. However, this maximization procedure may not be feasible in practice, because $\boldsymbol{\delta}_2^T P_{\mathcal{O}} \boldsymbol{\delta}_2 / \sqrt{\mathrm{tr}(\Lambda_1^2)}$ involves unknown quantities, including $\boldsymbol{\delta}_2$ and Σ. In Section 4.3, we provide a data-driven procedure for selecting τ_0 when there is no prior information available for the signals or the structure of the covariance matrix. Additional results on the power analysis are available in Appendix A.

4. Monte Carlo Simulation Studies

In this section, we assess the finite-sample performance of our proposed testing method. For ease of presentation, we conduct simulation studies for the two-sample test only. We also consider eight other tests for comparison: the unscaled Hotelling's tests CQ of Chen and Qin (2010) and aSUP of Xu et al. (2016), the diagonal Hotelling's tests PA of Park and Ayyala (2013), GCT of Gregory et al. (2015), and DLRT of Hu, Tong and Genton (2019), the composite

Hotelling's test CHT of Li (2017), and the regularized Hotelling's tests RMPBT of Zoh et al. (2018) and RHT of Li et al. (2020).

For each simulation, we generate observations $\boldsymbol{X}_s$, for $s = 1, \ldots, n_1$, and $\boldsymbol{Y}_t$, for $t = 1, \ldots, n_2$, from model (3.4). Without loss of generality, we let $\boldsymbol{\mu}_1 = \mathbf{0}$ and $\Sigma \in \mathbb{R}^{p\times p}$ be the common covariance matrix. Then, $\boldsymbol{X}_s = \Sigma^{1/2}\boldsymbol{Z}_s^{(1)}$ and $\boldsymbol{Y}_t = \Sigma^{1/2}\boldsymbol{Z}_t^{(2)} + \boldsymbol{\mu}_2$, where all the components of $\boldsymbol{Z}_s^{(1)}$ and $\boldsymbol{Z}_t^{(2)}$ are i.i.d. random variables with zero mean and unit variance. Under the null hypothesis, we set $\boldsymbol{\mu}_2 = \mathbf{0}$. Under the alternative hypothesis, we set $\boldsymbol{\mu}_2 = (\mu_{21}, \ldots, \mu_{2p_0}, 0, \ldots, 0)^T$, where $p_0 = \lfloor \beta p \rfloor$, with $\beta \in [0, 1]$ a tuning parameter that controls the degree of sparsity in the signals, and $\lfloor x \rfloor$ is the largest integer equal to or less than x.

4.1. Normal data

In the first simulation, we generate $\boldsymbol{Z}_s^{(1)}$ and $\boldsymbol{Z}_t^{(2)}$ from the p-dimensional multivariate normal distribution $N_p(\mathbf{0}, I_p)$. Let $D_p = \mathrm{diag}(d_{11}^2, \ldots, d_{pp}^2)$ be a diagonal matrix, with d_{ii} sampled randomly from the uniform distribution on $[0.5, 1.5]$. For the common covariance matrix Σ, we consider the following four structures:

(M1) $\Sigma_1 = D_p^{1/2} R_1 D_p^{1/2}$, where $R_1 = (0.9^{|i-j|})_{p\times p}$;

(M2) $\Sigma_2 = D_p^{1/2} R_2 D_p^{1/2}$, where $R_2 = ((-0.9)^{|i-j|})_{p\times p}$;

(M3) $\Sigma_3 = D_p^{1/2} R_3 D_p^{1/2}$, where R_3 is a block diagonal matrix with the same block as $B = (0.9^{I(i\neq j)})_{5\times 5}$, and $I(\cdot)$ is the indicator function;

(M4) $\Sigma_4 = D_p^{1/2} R_4 D_p^{1/2}$, where $R_4 = I_p$ is the identity matrix.

Table 1 summarizes the empirical size for the nine tests over 2,000 simulations with the given covariance matrices. The threshold for PHT is set as $\tau_0 = 0.8$. As shown in Table 1, PHT, aSUP, and RHT provide a more stable test statistic with a better controlled type-I error rate under most settings. When the dimension is large and the correlations between the covariates are strong, DLRT, GCT, and RMPBT suffer from significantly inflated type-I error rates compared with those of CQ and PA. When the covariates are weakly correlated, for example, the diagonal structure, most tests have a reasonable type-I error rate, except for CHT. CHT always risks an inflated type-I error rate compared with the nominal level at $\alpha = 0.05$, and so may not provide a perfect test.

To assess the power performance of the nine tests, we set the jth nonzero component in $\boldsymbol{\mu}_2$ as $\mu_{2j} = \kappa\delta_j$, where κ controls the signal strength, and $\delta_j \sim N(1.5, 1)$, for $j = 1, \ldots, p_0$. The other parameters are $n_1 = 30, n_2 = 25$, and $(\kappa = 0.1, p = 100)$ or $(\kappa = 0.075, p = 500)$. We then randomly generate 1,000 data sets under each scenario, and plot the simulation results in Figures 1 and 2.

As shown in the figures, when the true covariance matrix has a complex structure (including Σ_1, Σ_2, and Σ_3), our proposed PHT exhibits a significant

Table 1. Type-I error rates for PHT and eight competitors with normal data, where the sample sizes are $n_1 = 30$ and $n_2 = 25$, and the nominal level is $\alpha = 0.05$.

	p	PHT	DLRT	GCT	PA	RMPBT	aSUP	CQ	CHT	RHT
Σ_1	100	0.066	0.134	0.231	0.072	0.083	0.047	0.057	0.289	0.056
	500	0.061	0.142	0.170	0.060	0.162	0.056	0.062	0.376	0.064
Σ_2	100	0.065	0.127	0.257	0.068	0.095	0.055	0.086	0.296	0.059
	500	0.058	0.148	0.160	0.067	0.163	0.070	0.067	0.369	0.068
Σ_3	100	0.052	0.077	0.172	0.076	0.116	0.056	0.074	0.342	0.055
	500	0.045	0.093	0.068	0.059	0.181	0.054	0.053	0.408	0.068
Σ_4	100	0.056	0.073	0.107	0.056	0.057	0.079	0.072	0.375	0.059
	500	0.057	0.045	0.068	0.048	0.072	0.078	0.053	0.373	0.066

improvement in terms of power performance. Specifically, as long as the signals are not too sparse, PHT always has higher power than that of the other tests. When the covariates are independent of each other, aSUP achieves the highest power when the dimension is large. PHT also exhibits high power for detection that is nearly as good as that of PA. RMPBT shows good power performance when the dimension is not large. However, if the dimension becomes large, RMPBT suffers from low power, especially when the covariance matrix follows a diagonal structure. DLRT, GCT, and CQ also suffer from low power for detection especially when some covariates are highly correlated. Finally, RHT is not able to provide stable and comparable power compared with that of PHT and aSUP, especially when the dimension is large.

4.2. Heavy-tailed data

In the second simulation, we generate $\boldsymbol{Z}_s^{(1)}$ and $\boldsymbol{Z}_t^{(2)}$ from a heavy-tailed distribution to examine the robustness of the proposed tests. Following Gregory et al. (2015) and Hu, Tong and Genton (2019), we consider a "double" Pareto distribution with parameters $a > 0$ and $b > 0$. The detailed algorithm is as follows:

Step 1: Generate two independent random variables U and V, where U is from the Pareto distribution with the cumulative distribution function $F(x) = 1 - (1 + x/b)^{-a}$, for $x \geq 0$, and V is a binary random variable with $P(V = 1) = P(V = -1) = 0.5$. Then, $Z = UV$ follows the double Pareto distribution with parameters a and b.

Step 2: Generate random vectors $\widetilde{\boldsymbol{Z}}_s^{(1)} = (\tilde{z}_{s1}^{(1)}, \tilde{z}_{s2}^{(1)}, \ldots, \tilde{z}_{sp}^{(1)})^T$, for $s = 1, \ldots, n_1$, and $\widetilde{\boldsymbol{Z}}_t^{(2)} = (\tilde{z}_{t1}^{(2)}, \tilde{z}_{t2}^{(2)}, \ldots, \tilde{z}_{tp}^{(2)})^T$, for $t = 1, \ldots, n_2$, where all the components of $\widetilde{\boldsymbol{Z}}_s^{(1)}$ and $\widetilde{\boldsymbol{Z}}_t^{(2)}$ are sampled independently from the double Pareto distribution with parameters $a = 16.5$ and $b = 8$.

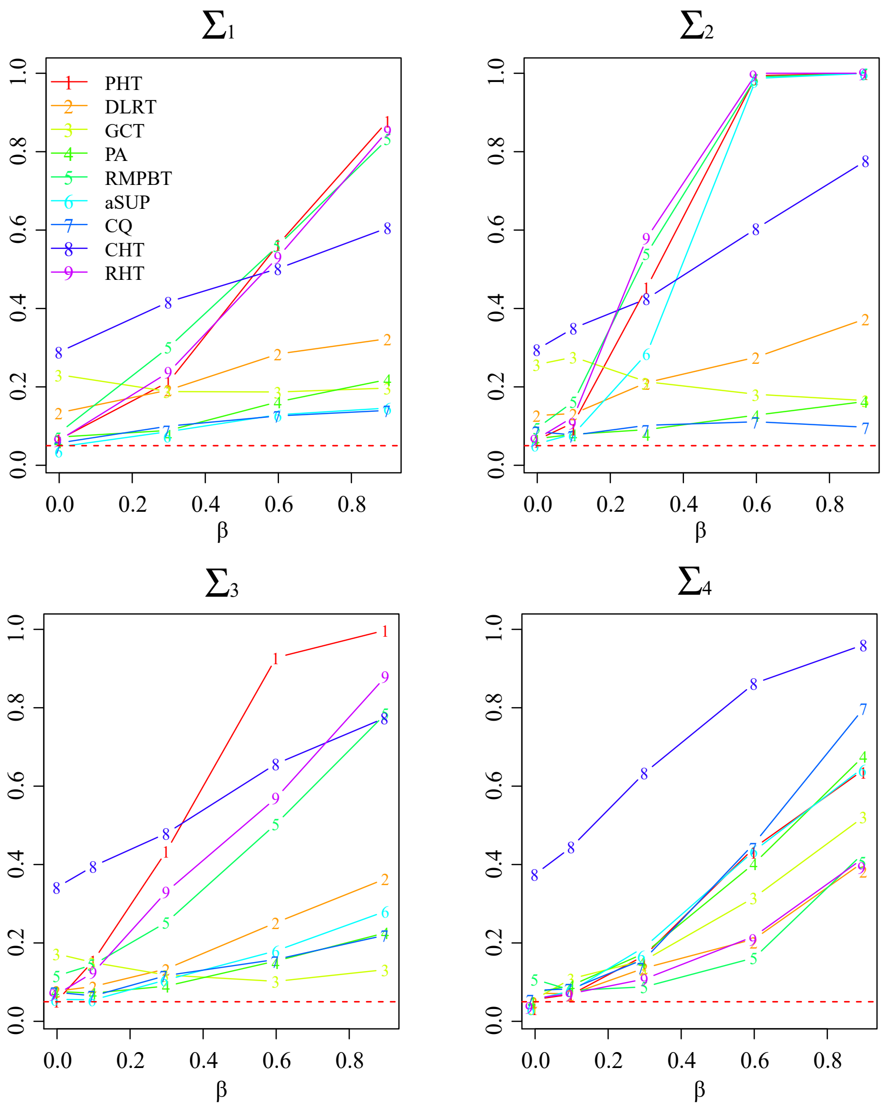

Figure 1. Power comparison between PHT and eight competitors, with $n_1 = 30, n_2 = 25$, and $p = 100$. The horizontal dashed lines represent the nominal level of $\alpha = 0.05$, and the results are based on normal data.

Step 3: Let $\boldsymbol{Z}_s^{(1)} = \widetilde{\boldsymbol{Z}}_s^{(1)}/c_0$ and $\boldsymbol{Z}_t^{(2)} = \widetilde{\boldsymbol{Z}}_t^{(2)}/c_0$, where $c_0^2 = 512/899$ is the variance of the double Pareto distribution with parameters $a = 16.5$ and $b = 8$.

Given $\boldsymbol{Z}_s^{(1)}$ and $\boldsymbol{Z}_t^{(2)}$, we use the same settings as those in Section 4.1 to generate the observations of $\boldsymbol{X}_s$ and $\boldsymbol{Y}_t$ for each simulation.

Table 2 and Figures 3 and 4 present the empirical size and power for the nine tests with heavy-tailed data at the nominal level of $\alpha = 0.05$. The

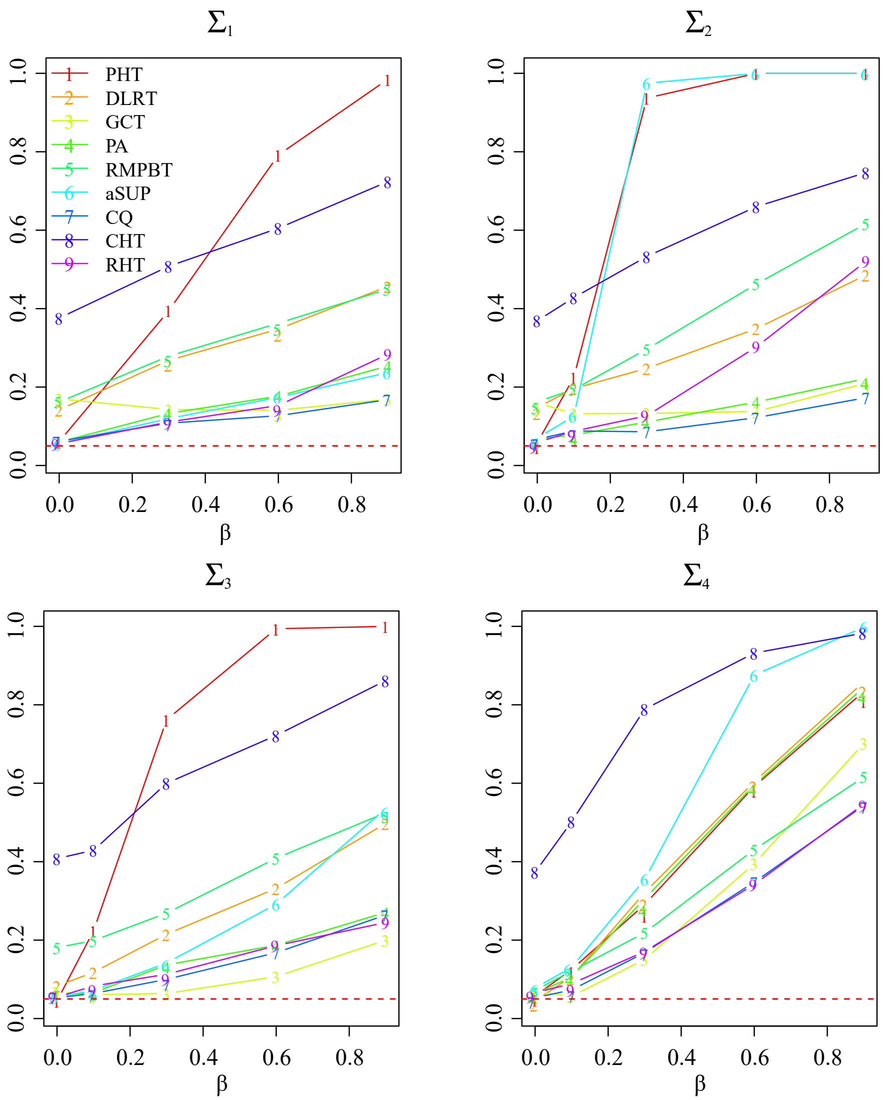

Figure 2. Power comparison between PHT and eight competitors, with $n_1 = 30, n_2 = 25$, and $p = 500$. The horizontal dashed lines represent the nominal level of $\alpha = 0.05$, and the results are based on normal data.

simulations used to compute the empirical size and power are over 2,000 and 1,000 simulations, respectively. In particular, when the dimension is large and the correlations between the covariates are strong, PHT controls the type-I error rate, and achieves a higher power for detection. RMPBT exhibits good power performance when the dimension is not large and the covariance matrix has a complex structure, but it suffers from a slightly inflated type-I error rate. When the dimension is large and the covariance matrix has a complex structure (including Σ_1, Σ_2, and Σ_3), RMPBT exhibits a substantially inflated type-I error

Table 2. Type-I error rates for PHT and eight competitors, with heavy-tailed data, where the sample sizes are $n_1 = 30$ and $n_2 = 25$, and the nominal level is $\alpha = 0.05$.

	p	PHT	DLRT	GCT	PA	RMPBT	aSUP	CQ	CHT	RHT
Σ_1	100	0.064	0.145	0.257	0.067	0.105	0.062	0.071	0.305	0.078
	500	0.051	0.153	0.145	0.061	0.189	0.057	0.069	0.364	0.075
Σ_2	100	0.071	0.116	0.254	0.074	0.080	0.050	0.069	0.308	0.074
	500	0.061	0.139	0.176	0.069	0.180	0.048	0.053	0.363	0.070
Σ_3	100	0.060	0.082	0.180	0.079	0.093	0.050	0.070	0.382	0.065
	500	0.051	0.076	0.096	0.054	0.137	0.048	0.046	0.390	0.069
Σ_4	100	0.056	0.058	0.141	0.077	0.052	0.051	0.057	0.383	0.059
	500	0.055	0.050	0.113	0.051	0.085	0.051	0.059	0.356	0.062

rate, and suffers from low power. RHT exhibits similar power performance to that of RMPBT, but is inferior to PHT, especially when the dimension is large and the correlations between the covariates are strong. In addition, aSUP exhibits a well-controlled type-I error rate in most settings. However, its power performance may be sensitive to the structure of the covariance matrix; for example, it suffers from low power under Σ_1 and Σ_3, but exhibits good power under Σ_2. DLRT has a well-controlled type-I error rate under the diagonal covariance matrix, but suffers from low power. Finally, GCT and CHT always suffer from a significantly inflated type-I error rate.

4.3. A data-driven threshold for τ_0

In this section, we provide a data-driven method for selecting the threshold τ_0. When there is no prior information on the covariance matrix structure, a reasonable choice for the threshold τ can be to maximize the empirical estimator for the signal-to-noise ratio that determines the power of the PHT statistic. From (2.4) and (3.6), we have $\mathrm{SNR}_1(\tau_0) = (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T P_{\mathcal{O}}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)/\sqrt{\mathrm{tr}(\Lambda_1^2)}$ and $\mathrm{SNR}_2(\tau_0) = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T P_{\mathcal{O}}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)/\sqrt{\mathrm{tr}(\Lambda_1^2)}$ for the one- and two-sample tests, respectively. We then estimate the two ratios by

$$\widehat{\mathrm{SNR}}_1(\tau_0) = \frac{T_1(\tau_0)}{\sqrt{\widehat{\mathrm{tr}(\Lambda_1^2)}}} \quad \text{and} \quad \widehat{\mathrm{SNR}}_2(\tau_0) = \frac{T_2(\tau_0)}{\sqrt{\widehat{\mathrm{tr}(\Lambda_1^2)}}}.$$

From Lemmas 1 and 2, we have $\widehat{\mathrm{SNR}}_1(\tau_0) \xrightarrow{P} \mathrm{SNR}_1(\tau_0)$ as $n \to \infty$, and $\widehat{\mathrm{SNR}}_2(\tau_0) \xrightarrow{P} \mathrm{SNR}_2(\tau_0)$ as $N \to \infty$, where $\xrightarrow{P}$ denotes convergence in probability. For simplicity, we present the selection procedure of the threshold τ_0 for the two-sample test only. The same procedure can be readily adapted for the one-sample test.

Step 1: Randomly generate two subsets $\mathrm{Set}_X^* = \{\boldsymbol{X}_k, k = 1, \ldots, n_1^*\}$ and $\mathrm{Set}_Y^* = \{\boldsymbol{Y}_l, l = 1, \ldots, n_2^*\}$, where $n_1^* < n_1$ and $n_2^* < n_2$, and $\boldsymbol{X}_l^*$ and

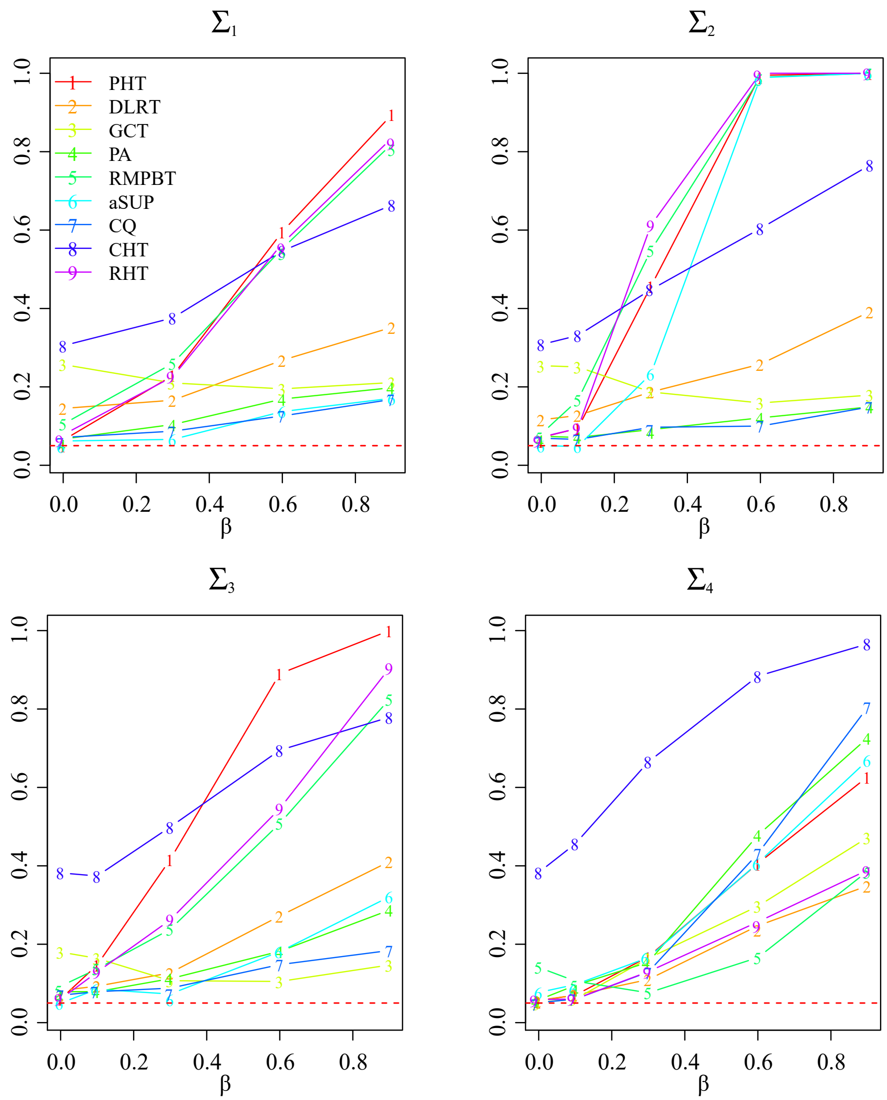

Figure 3. Power comparison between PHT and eight competitors, with $n_1 = 30, n_2 = 25$, and $p = 100$. The horizontal dashed lines represent the nominal level of $\alpha = 0.05$, and the results are based on heavy-tailed data.

$\boldsymbol{Y}_k^*$ are selected randomly without replacement from $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_1}\}$ and $\{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_2}\}$, respectively.

Step 2: Given the grid points $\mathcal{T}_{\tau_0} = \{\tau_{01}, \ldots, \tau_{0H}\}$, for each point $\tau_{0h} \in \mathcal{T}_{\tau_0}$, compute $\widehat{\mathrm{SNR}}_2(\tau_{0h})$ using Set_X^* and Set_Y^*, and then select $\hat{\tau}_0 = \mathrm{argmax}_{\tau_{0h} \in \mathcal{T}_{\tau_0}} \widehat{\mathrm{SNR}}_2(\tau_{0h})$.

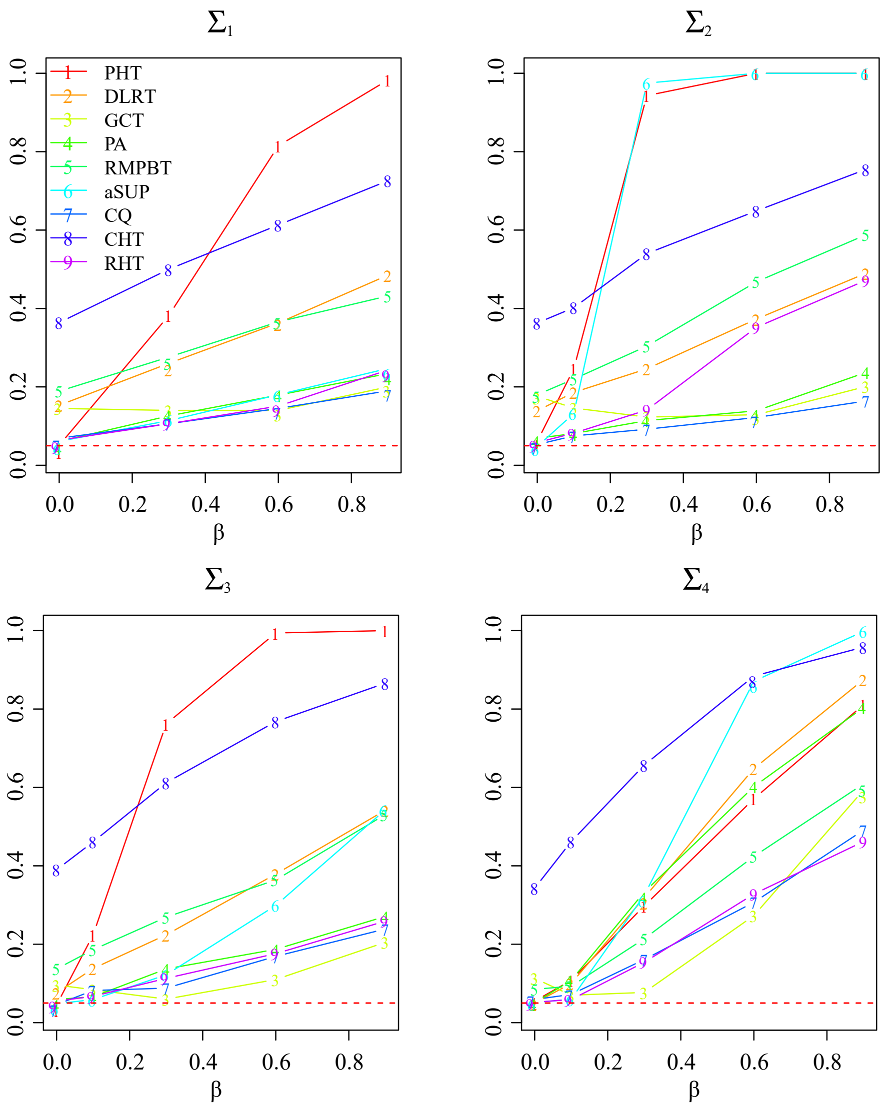

Figure 4. Power comparison between PHT and eight competitors, with $n_1 = 30, n_2 = 25$, and $p = 500$. The horizontal dashed lines represent the nominal level of $\alpha = 0.05$, and the results are based on heavy-tailed data.

Step 3: Repeat Steps 1–2 for B times, and denote the selected $\hat{\tau}_0$ as $\hat{\tau}_0^{(b)}$ for the bth time. The optimal τ_0 is defined as the median of $\{\hat{\tau}_0^{(1)}, \ldots, \hat{\tau}_0^{(B)}\}$.

When the sample size is not large, our simulations show that the median of $\{\hat{\tau}_0^{(1)}, \ldots, \hat{\tau}_0^{(B)}\}$ provides a more robust estimate than the mode does for the true value that maximizes the signal-to-noise ratio. In addition, to balance the computation time and the detection ability of our PHT statistic, we recommend to use $n_1^* = \lfloor 2n_1/3 \rfloor$, $n_2^* = \lfloor 2n_2/3 \rfloor$, $B = 10$, and $\mathcal{T}_{\tau_0} = \{0.7, 0.8, 0.9, 1\}$.

Table 3. Type-I error rates for PHT and eight competitors, with normal data, where the sample sizes are $n_1 = 30$ and $n_2 = 25$, and the nominal level is $\alpha = 0.05$.

	p	PHT	DLRT	GCT	PA	RMPBT	aSUP	CQ	CHT	RHT
Σ_1	100	0.085	0.134	0.231	0.072	0.083	0.047	0.057	0.289	0.056
	500	0.092	0.142	0.170	0.060	0.162	0.056	0.062	0.376	0.064
Σ_2	100	0.081	0.127	0.257	0.068	0.095	0.055	0.086	0.296	0.059
	500	0.094	0.148	0.160	0.067	0.163	0.070	0.067	0.369	0.068
Σ_3	100	0.081	0.077	0.172	0.076	0.116	0.056	0.074	0.342	0.055
	500	0.088	0.093	0.068	0.059	0.181	0.054	0.053	0.408	0.068
Σ_4	100	0.049	0.073	0.107	0.056	0.057	0.079	0.072	0.375	0.059
	500	0.059	0.045	0.068	0.048	0.072	0.078	0.053	0.373	0.066

To assess the usefulness of the selection procedure for τ_0, we compare the results of PHT with those of the other tests. For the common covariance matrix, we also consider the four structures Σ_1, Σ_2, Σ_3, and Σ_4. The other parameters are the same as in the previous simulations. Table 3 summarizes the empirical size for the nine tests over 2,000 simulations with the given covariance matrices. When the correlations between the covariates are strong, PHT exhibits some inflated type-I error rates compared with CQ, aSUP, and RHT. This may be the price that PHT pays for the unknown prior information of the covariance matrix, or perhaps a better estimate of the optimal threshold is required.

Figures 5 and 6 display the power performance for the nine tests with normal data at the nominal level of $\alpha = 0.05$. Specifically, if the covariance matrix has a complex structure (including Σ_1, Σ_2, and Σ_3), PHT always possesses higher power than that of the other methods, as long as the signals are not too sparse. When the covariance matrix follows a diagonal structure, aSUP achieves the highest power as the dimension becomes large; PHT also exhibits high power for detection that is nearly the same as that of PA. PMPBT and RHT suffer from low power when the dimension is large. In addition, when the correlations between the covariates are strong, DLRT, GCT, and CQ usually also suffer from low power for detection.

5. Applications

5.1. Small round blue-cell tumor data

We apply our proposed PHT to analyze two microarray data sets. The first contains data on the small round blue-cell tumors (SRBCTs), studied by Khan et al. (2001), including 2,308 genes for four types of childhood tumors. The data set is from `http://www.biolab.si/supp/bi-cancer/projections/info/SRBCT.html`. As in Zoh et al. (2018), we are interested in testing the differential expression of genes between the Burkitt lymphoma (BL) tumor and the neuroblastoma (NB) tumor. The sample sizes of the BL and NB tumors

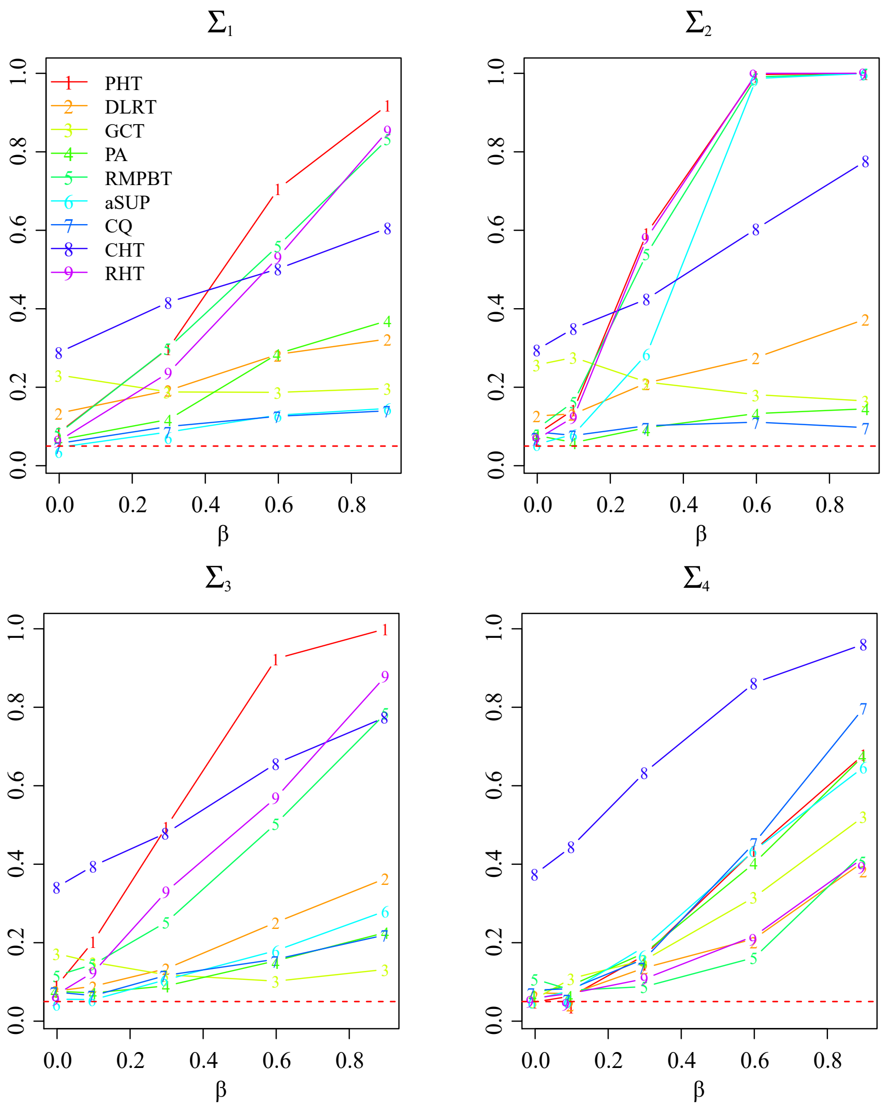

Figure 5. Power comparison between PHT and eight competitors, with $n_1 = 30, n_2 = 25$, and $p = 100$. The horizontal dashed lines represent the nominal level of $\alpha = 0.05$, and the results are based on normal data.

are 11 and 18, respectively. Owing to the small sample sizes, we perform PHT with a fixed threshold of $\tau_0 = 0.8$, and then compare the results with those of DLRT, GCT, PA, RMPBT, aSUP, CQ, CHT, and RHT. The p-values of the nine tests are all smaller than 0.0001. Thus, all the tests significantly reject the null hypothesis of the two-sample test at the nominal level of $\alpha = 0.05$.

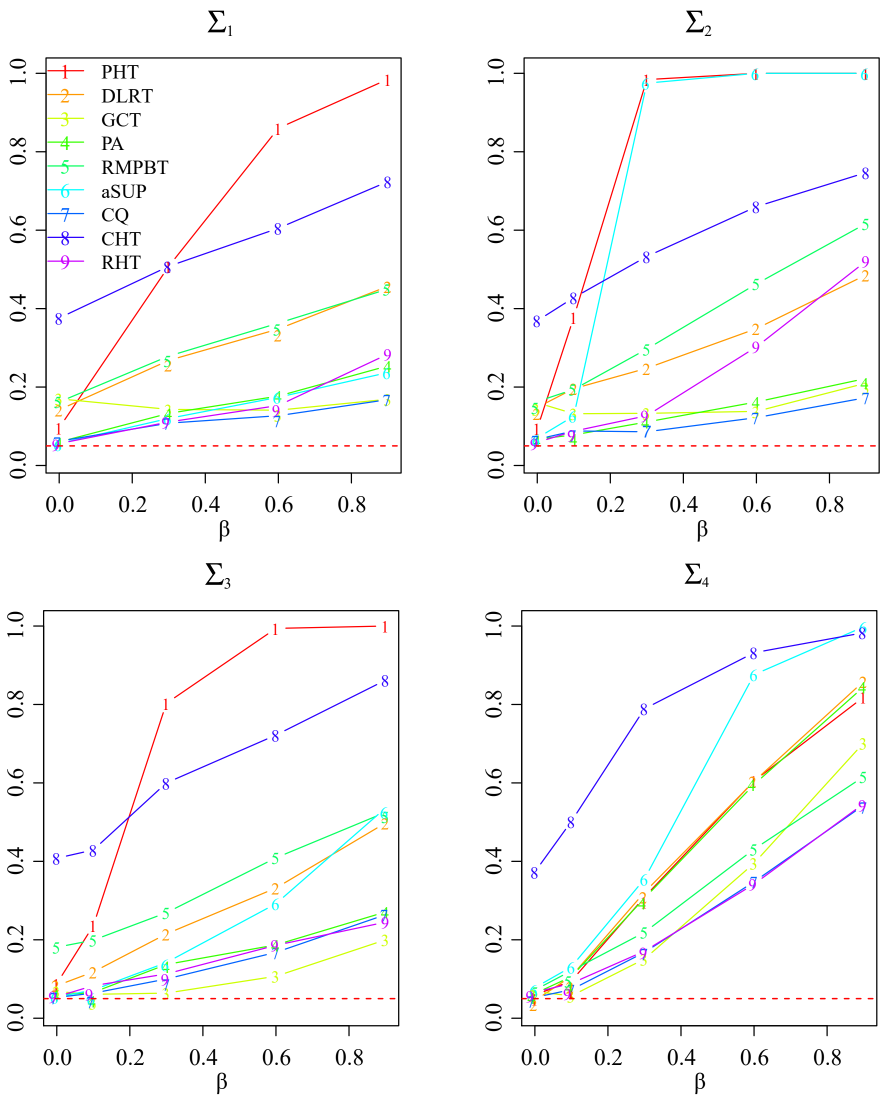

Figure 6. Power comparison between PHT and eight competitors, with $n_1 = 30, n_2 = 25$, and $p = 500$. The horizontal dashed lines represent the nominal level of $\alpha = 0.05$, and the results are based on normal data.

5.2. Leukemia data

The second data set contains leukemia data from two groups of patients, namely, those with acute lymphoblastic leukemia (ALL), and those with acute myeloid leukemia (AML). The data set contains 7,129 genes and 72 samples, with 47 ALL patients and 25 AML patients, and is publicly available in the R package "golubEsets". To compare the performance of the tests, we first perform two-sample t-tests to screen the top 250 significant genes. We then apply our

Table 4. False (F) and true (T) positive rates of our data-driven PHT and eight competitors for leukemia data at the nominal level of 0.05.

	PHT	DLRT	GCT	PA	RMPBT	aSUP	CQ	CHT	RHT
F	0.089	0.229	0.587	0.082	0.101	0.221	0.078	0.573	0.056
T	0.887	0.968	0.027	0.841	0.998	0.972	0.773	1.000	0.802

data-driven PHT to the selected gene set with the threshold on the grid points $\{0.7, 0.8, 0.9, 1\}$, and compare the results with those of the other eight tests. The p-values of the nine tests are all smaller than 0.0001, indicating that the mean expression levels of the gene set between the ALL and AML groups are significantly different.

To further compare the performance of the tests, we select the top 50 significant genes and the last 200 nonsignificant genes to form a new gene set. The signal strength of the new gene set is weaker than that with the top 250 significant genes. We then apply the permutation method to create two artificial groups for the new gene set to mimic the null and alternative hypotheses, respectively. Specifically, we randomly sample two distinct subclasses, without replacement, from the pooled data with sample sizes 30 and 17, respectively. Because both classes are partitioned from the pooled data, the null hypothesis can be regarded as true. Finally, we repeat the procedure 1,000 times, and perform the nine tests at the nominal level of 0.05. The rejection rate is computed to represent the false positive rate. Similarly, to mimic the alternative hypothesis, we randomly sample one class from the ALL group with sample size 30, and another class from the AML group with sample size 17. Then, based on 1,000 simulations for each test method, we compute the rejection rates at the nominal level of 0.05. For each test method, the rejection rate is computed to represent the true positive rate.

Table 4 shows that DLRT, GCT, RMPBT, aSUP, and CHT suffer from inflated false positive rates, particularly GCT and CHT. In contrast, PHT, PA, CQ, and RHT provide a reasonable type-I error rate and can serve as valid tests.

6. Conclusion

We provide a pairwise Hotelling method for testing whether a mean vector is equal to a given vector for a one-sample test, or testing whether two mean vectors are equal for a two-sample test in a high-dimensional setting with a low sample size. Our proposed PHT statistics differ from those of existing tests, including UHT, DHT, and RHT. Specifically, UHT and DHT both ignore correlations between covariates. When some covariates exhibit strong correlations in the data, neither of the two methods provide satisfactory performance. In contrast, RHT does account for the correlations, but it involves a regularized covariance matrix. When the sample size is small relative to the dimension, the regularized covariance matrix can be very noisy, especially when the covariance matrix is sparse.

Consequently, the test statistics involving the sample covariance matrix may lead to inflated type-I error rates and/or suffer from low statistical power. Our proposed pairwise Hotelling method overcomes the drawbacks of DHT and RHT. Specifically, we first perform a screening procedure to identify covariate pairs that exhibit strong correlations, and then construct the classic Hotelling's test statistics for these covariate pairs. For the remaining covariates that are weakly correlated with others, we construct the squares of the componentwise t-statistics for each of the individual covariates. Our proposed PHT statistics are then the sum of all the Hotelling's test statistics and the squared t-statistics. Simulation results show that our new tests improve the statistical power significantly when some covariates are highly correlated. Furthermore, even when most covariates are weakly correlated, our proposed tests still maintain high power compared with that of the existing tests in the literature.

Here, we assume that the eigenvalues of the covariance matrix are bounded by constants through $1/\bar{K}_1$ and $\bar{K}_1$. This assumption is widely adopted in the literature; see, for example, Cai, Liu and Xia (2014), Xu et al. (2016), and Cui et al. (2020). For a fair comparison, we follow the same condition. However, allowing $\bar{K}_1$ to grow with p is an interesting topic and deserves further research. In particular, when some correlations go to one as p increases, or when some eigenvalues of the covariance matrix are large, the covariance matrix will tend to be a singular matrix, or even a spiked matrix (Johnstone (2001)). Under such a structure, Aoshima and Yata (2018) and Xie, Zeng and Zhu (2021) show that additional restrictive conditions on the eigenvalues are required in order to ensure the convergence of the test statistics.

Finally, we note that, when the covariance matrices of the two samples are unequal, our PHT statistic will encounter the high-dimensional Behrens–Fisher problem, as in Feng et al. (2015). To deal with this situation, one may adopt the ideas of Anderson (2003) and Ishii, Yata and Aoshima (2019) to construct a two-sample PHT statistic. Without loss of generality, we assume $n_1 \leq n_2$ and let $\boldsymbol{V}_i = \boldsymbol{X}_i - \boldsymbol{Y}_i$, for $i = 1, \dots, n_1$, with $\boldsymbol{\mu}_V = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. Then, to test $H_0 : \boldsymbol{\mu}_V = \mathbf{0}$ versus $H_1 : \boldsymbol{\mu}_V \neq \mathbf{0}$, we can apply our newly proposed one-sample PHT based on the data $\boldsymbol{V}_1, \dots, \boldsymbol{V}_{n_1}$. The limitation of this method is that the remaining samples $\boldsymbol{Y}_{n_1+1}, \dots, \boldsymbol{Y}_{n_2}$ are ignored when the sample sizes are not balanced. To conclude, it remains challenging to construct efficient tests for solving the high-dimensional Behrens–Fisher problem. We leave this to future research.

Supplementary Material

This supplement contains 4 web appendices. In Appendix A, we provide some additional comparisons on the statistical power. In Appendix B, we provide some useful lemmas as the preliminary results. In Appendix C, we provide the proofs of Theorems 1 and 2, and Lemma 1. In Appendix D, we provide the proofs of

Theorems 3 and 4, and Lemma 2.

Acknowledgments

The authors thank the editor, associate editor, and two anonymous reviewers for their insightful comments and suggestions. Zongliang Hu's research was supported by National Natural Science Foundation of China (12001378), Guangdong Basic and Applied Basic Research Foundation (2019A1515110449), and Natural Science Foundation of Gangdong Province (2023A1515010027). Tiejun Tong's research was supported by General Research Funds (HKBU12303918, HKBU12303421, HKBU12300123), Initiation Grant for Faculty Niche Research Areas (RC-FNRA-IG/20-21/SCI/03) of Hong Kong Baptist University, and National Natural Science Foundation of China (1207010822). Marc G. Genton's research was supported by the King Abdullah University of Science and Technology (KAUST).

References

Ahmad, M. R. (2014). A U-statistic approach for a high-dimensional two-sample mean testing problem under non-normality and Behrens-Fisher setting. *The Annals of the Institute of Statistical Mathematics* **66**, 33–61.

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

Aoshima, M. and Yata, K. (2018). Two-sample tests for high-dimension, strongly spiked eigenvalue models. *Statistica Sinica* **28**, 43–62.

Ayyala, D. N., Park, J. and Roy, A. (2017). Mean vector testing for high-dimensional dependent observations. *Journal of Multivariate Analysis* **153**, 136–155.

Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: By an example of a two sample problem. *Statistica Sinica* **6**, 311–329.

Bickel, P. J. and Levina, E. (2004). Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli* **10**, 989–1010.

Cai, T., Liu, W. and Xia, Y. (2014). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 349–372.

Cao, Y., Lin, W. and Li, H. (2018). Two-sample tests of high-dimensional means for compositional data. *Biometrika* **105**, 115–132.

Chen, L., Paul, D., Prentice, R. L. and Wang, P. (2011). A regularized Hotelling's T^2 test for pathway analysis in proteomic studies. *Journal of the American Statistical Association* **106**, 1345–1360.

Chen, S., Li, J. and Zhong, P. (2019). Two-sample and ANOVA tests for high dimensional means. *The Annals of Statistics* **47**, 1443–1474.

Chen, S. and Qin, Y. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics* **38**, 808–835.

Cox, D. R. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika* **91**, 729–737.

Cui, X., Li, R., Yang, G. and Zhou, W. (2020). Empirical likelihood test for a large-dimensional mean vector. *Biometrika* **107**, 591–607.

Dong, K., Pang, H., Tong, T. and Genton, M. G. (2016). Shrinkage-based diagonal Hotelling's tests for high-dimensional small sample size data. *Journal of Multivariate Analysis* **143**, 127–142.

Feng, L. and Sun, F. (2015). A note on high-dimensional two-sample test. *Statistics & Probability Letters* **105**, 29–36.

Feng, L., Zou, C., Wang, Z. and Zhu, L. (2015). Two-sample Behrens-Fisher problem for high-dimensional data. *Statistica Sinica* **25**, 1297–1312.

Feng, L., Zou, C., Wang, Z. and Zhu, L. (2017). Composite T^2 test for high-dimensional data. *Statistica Sinica* **27**, 1419–1436.

Gregory, K. B., Carroll, R. J., Baladandayuthapani, V. and Lahiri, S. N. (2015). A two-sample test for equality of means in high dimension. *Journal of the American Statistical Association* **110**, 837–849.

He, Y., Xu, G., Wu, C. and Pan, W. (2021). Asymptotically independent U-statistics in high-dimensional testing. *The Annals of Statistics* **49**, 154–184.

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

Hu, Z., Tong, T. and Genton, M. G. (2019). Diagonal likelihood ratio test for equality of mean vectors in high-dimensional data. *Biometrics* **75**, 256–267.

Ishii, A., Yata, K. and Aoshima, M. (2019). Inference on high-dimensional mean vectors under the strongly spiked eigenvalue model. *Japanese Journal of Statistics and Data Sciences* **2**, 105–128.

Jiang, T. and Li, P. (2021). Mean test with fewer observation than dimension and ratio unbiased estimator for correlation matrix. *arXiv:2108.06892*.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics* **29**, 295–327.

Khan, J., Wei, J., Ringner, M., Saal, L., Ladanyi, M., Westermann, F. et al. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine* **7**, 673–679.

Li, G., Peng, H., Zhang, J. and Zhu, L. (2012). Robust rank correlation based screening. *The Annals of Statistics* **40**, 1846–1877.

Li, H., Aue, A., Paul, D., Peng, J. and Wang, P. (2020). An adaptable generalization of Hotelling's T^2 test in high dimension. *The Annals of Statistics* **48**, 1815–1847.

Li, T. (2017). Composite Hotelling's T-square test for high-dimensional data. *Chinese Journal of Applied Probablility and Statistics* **33**, 349–368.

Lin, Z. and Lu, C. (1997). *Limit Theory for Mixing Dependent Random Variables*. Springer, New York.

Lopes, M. E., Jacob, L. and Wainwright, M. J. (2011). A more powerful two-sample test in high dimensions using random projection. *Advances in Neural Information Processing Systems* **24**, 1206–1214.

Park, J. and Ayyala, D. N. (2013). A test for the mean vector in large dimension and small samples. *Journal of Statistical Planning and Inference* **143**, 929–943.

Srivastava, M. S. (2009). A test for the mean vector with fewer observations than the dimension under non-normality. *Journal of Multivariate Analysis* **100**, 518–532.

Srivastava, M. S. and Du, M. (2008). A test for the mean vector with fewer observations than the dimension. *Journal of Multivariate Analysis* **99**, 386–402.

Srivastava, M. S., Katayama, S. and Kano, Y. (2013). A two sample test in high dimensional data. *Journal of Multivariate Analysis* **114**, 349–358.

Srivastava, R., Li, P. and Ruppert, D. (2016). RAPTT: An exact two-sample test in high dimensions using random projections. *Journal of Computational and Graphical Statistics* **23**, 954–970.

Thulin, M. (2014). A high-dimensional two-sample test for the mean using random subspaces. *Computational Statistics and Data Analysis* **74**, 26–38.

Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica* **21**, 5–42.

Wang, R., Peng, L. and Qi, Y. (2013). Jackknife empirical likelihood test for equality of two high dimensional means. *Statistica Sinica* **23**, 667–690.

Wu, Y., Genton, M. G. and Stefanski, L. A. (2006). A multivariate two-sample mean test for small sample size and missing data. *Biometrics* **62**, 877–885.

Xie, J., Zeng, Y. and Zhu, L. (2021). Limiting laws for extreme eigenvalues of large-dimensional spiked Fisher matrices with a divergent number of spikes. *Journal of Multivariate Analysis* **184**, 104742.

Xu, G., Lin, L., Wei, P. and Pan, W. (2016). An adaptive two-sample test for high-dimensional means. *Biometrika* **103**, 609–624.

Zhang, D. (2021). Robust estimation of the mean and covariance matrix for high dimensional time series. *Statistica Sinica* **31**, 797–820.

Zhang, J. T., Guo, J., Zhou, B. and Cheng, M. Y. (2020). A simple two-sample test in high dimensions based on L^2-norm. *Journal of the American Statistical Association* **115**, 1011–1027.

Zoh, R. S., Sarkar, A., Carroll, R. J. and Mallick, B. K. (2018). A powerful Bayesian test for equality of means in high dimensions. *Journal of the American Statistical Association* **113**, 1733–1741.

Zongliang Hu

School of Mathematical Sciences, Shenzhen University, Shenzhen, China.

E-mail: zlhu@szu.edu.cn

Tiejun Tong

Department of Mathematics, Hong Kong Baptist University, Hong Kong.

E-mail: tongt@hkbu.edu.hk

Marc G. Genton

Statistics Program, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia.

E-mail: marc.genton@kaust.edu.sa

(Received October 2021; accepted May 2022)