Statistica Sinica Preprint No: SS-2020-0486	
Title	Use of Random Integration to Test Equality of High Dimensional Covariance Matrices
Manuscript ID	SS-2020-0486
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202020.0486
Complete List of Authors	Yunlu Jiang, Canhong Wen, Yukang Jiang, Xueqin Wang and Heping Zhang
Corresponding Author	Heping Zhang
E-mail	heping.zhang@yale.edu
Notice: Accepted version subject to English editing.	

Use of random integration to test equality of high dimensional covariance matrices

Yunlu Jiang, Canhong Wen, Yukang Jiang, Xueqin Wang and Heping Zhang

Jinan University, University of Science and Technology of China,

Sun Yat-Sen University, Yale University

Abstract: Testing the equality of two covariance matrices is a fundamental problem in statistics, and especially challenging when the data are high-dimensional. Through a novel use of random integration, we can test the equality of high-dimensional covariance matrices without assuming parametric distributions for the two underlying populations, even if the dimension is much larger than the sample size. The asymptotic properties of our test for arbitrary number of covariates and sample size are studied in depth under a general multivariate model. The finite-sample performance of our test is evaluated through numerical studies. The empirical results demonstrate that our test is highly competitive with existing tests in a wide range of settings. In particular, our proposed test is distinctly powerful under different settings when there exist a few large or many small diagonal disturbances between the two covariance matrices.

Key words and phrases: High-dimensional data, Covariance matrix, Random integration.

1. Introduction

Testing the equality of two covariance matrices arises from many important problems including both the classic experimental designs and the analysis of high throughput omic data. For example, gene expression data have often been explored to classify disease types. Igolkina et al. (2018) pointed out that variance in gene expression is an important characteristic of schizophrenia. Roberts et al. (2018) showed that many genes differ in the variances of their gene expressions between disease states. Comparison between two covariance matrices is therefore essential to analyze gene expression microarray data of two different groups. This comparison becomes difficult because the number of the samples is usually much smaller than the number of genes. In the literature, many methods have been proposed to test the equality of two covariance matrices, but perform poorly in practice, especially when there are a few large or many small diagonal disturbances between the two covariance matrices.

Let $\mathbf{X}$ and $\mathbf{Y}$ be the p -dimensional random variables with covariance matrices Σ_1 and Σ_2 , respectively. Given independent samples $\mathcal{X}_m = \{\mathbf{X}_1, \dots, \mathbf{X}_m\}$ from $\mathbf{X}$ and $\mathcal{Y}_n = \{\mathbf{Y}_1, \dots, \mathbf{Y}_n\}$ from $\mathbf{Y}$, we want to test:

$$H_0 : \Sigma_1 = \Sigma_2 \quad vs \quad H_1 : \Sigma_1 \neq \Sigma_2. \quad (1.1)$$

In the classic low-dimensional setting, Anderson, T. W. (2003) proposed a like-

likelihood ratio test (LRT) statistic and showed that the proposed LRT statistic is asymptotically χ^2 -distribution with degrees of freedom $p(p + 1)/2$ under the multivariate normality assumption and H_0 when p is fixed.

In recent applications from gene expression (Pan et al., 2018), to neuroimaging (Le et al., 2001), and to risk management (Bollerslev, 2019), the dimension p can be much larger than the sample size n , e.g., “large p small n ” or “large p large n .” In this setting, the sample covariance matrix does not converge to the population counterpart (Bai et al, 2009), and the aforementioned classical methods for the low dimensional case either are not applicable or perform poorly. As well summarized by Cai (2017), this problem is so important and difficult that it has attracted a great deal of attention. We briefly review some of the methods below.

The modified LRT based methods have been considered by Bai et al (2009), Jiang et al. (2013), and Jiang and Qi (2015). Because $\Sigma_1 = \Sigma_2$ is equivalent to the Frobenius norm $tr(\Sigma_1 - \Sigma_2)^2 = 0$, the Frobenius norm-based tests have also been proposed (Schott, 2007; Srivastava and Yanagihara, 2010). However, The modified LRT based and Frobenius norm based methods assume multivariate normality. Li and Chen (2012) removed the normality assumption by using a linear combination of three one-sample U statistics. The test proposed by Li and Chen (2012) can be applied without assuming parametric distributions for

the two populations and is very powerful when there are many small differences between two population covariance matrices, but this test was found not powerful against sparse alternative (Yang and Pan, 2017) or when the two covariance matrices differ slightly only in the diagonal (Wu and Li, 2015). Using random projection, Wu and Li (2015) constructed a test statistic to improve the power when there are many small diagonal disturbances between the two covariance matrices, but again assuming the normality. He et al. (2021) introduced the adaptive testing to combine the finite-order U-statistics that includes the variants of Frobenius norm based statistics.

Assuming sparsity for $\Sigma_1 - \Sigma_2$ under the alternative hypothesis, Cai et al. (2013) introduced an extreme statistic, which is robust with respect to the population distributions, and is very powerful when there are only several large disturbances between two population covariance matrices, and Chang et al. (2017) proposed a computationally fast procedure. Zhu (2017) constructed a sparse-Leading-Eigenvalue-Driven (sLED) test to deal with the scenario in which the signals are both sparse and weak, and proved that sLED can achieve full power asymptotically when the sparse signal is strong enough. However, these tests are either not powerful when there are many small disturbances between Σ_1 and Σ_2 or the theoretical properties required an explicit relationship between n and p or too complicated for practical use.

Above mentioned tests often targeted at some specific situations. To accommodate various situations, some weighted combination tests are proposed. For example, Yang and Pan (2017) proposed a weighted test statistic based on random matrix theories. Zheng et al. (2020) proposed a power-enhancement high-dimensional test. These tests can handle both sparsity and non-sparsity structures. However, these tests depend on the proper choice of weights, which is a very challenging task. Furthermore, these procedures also required an explicit relationship between n and p . Yu et al. (2020) proposed a scale-invariant power enhancement test based on Fisher's method, but again assuming the normality.

In short, although many methods exist, they are limited by different reasons. Through a novel use of random integration, we propose a method to test equality of two high dimensional covariance matrices. Specifically, the our proposed test possesses the following three merits at the same time:

1. Our test does not require the distributional assumption.
2. Our test works well for the paradigm of “large p ”, even when there exist a few large or many small diagonal disturbances between the two covariance matrices.
3. The asymptotic theory is established under a general multivariate model

with certain moment conditions, but without requiring an explicit relationship between p and n .

The rest of the paper is organized as follows. In Section 2, we introduce our test statistic via the random integration technique, establish the asymptotic properties. In Section 3, simulation studies are conducted to evaluate the finite sample performance of the proposed test. In Section 4, a real data set is analyzed to compare the proposed test with some existing methods. We conclude with some remarks in Section 5 and all technical proofs are presented in the Supplementary Material.

2. Methodology and main results

To introduce a proper statistic to test (1.1), especially when there are many small diagonal disturbances between the two covariance matrices, it would be helpful if we can strengthen the information of diagonal disturbances for $\Sigma_1 - \Sigma_2$. Unlike the random matrix projection method, which needs the normality assumption to strengthen the information on a line, we develop the random integration technique to strengthen the information through integrations. To this end, we first denote $\mathbf{X}^c = \mathbf{X} - \mu_1$ and $\mathbf{Y}^c = \mathbf{Y} - \mu_2$, where $\mu_1 = E\mathbf{X}$ and $\mu_2 = E\mathbf{Y}$. Then,

$\Sigma_1 = E\mathbf{X}^c\mathbf{X}^{c\top}$ and $\Sigma_2 = E\mathbf{Y}^c\mathbf{Y}^{c\top}$. Note the following equivalences

$$\begin{aligned}\Sigma_1 = \Sigma_2 &\Leftrightarrow E\mathbf{X}^c\mathbf{X}^{c\top} = E\mathbf{Y}^c\mathbf{Y}^{c\top} \\ &\Leftrightarrow \boldsymbol{\alpha}^\top E\mathbf{X}^c\mathbf{X}^{c\top}\boldsymbol{\alpha} = \boldsymbol{\alpha}^\top E\mathbf{Y}^c\mathbf{Y}^{c\top}\boldsymbol{\alpha}, \text{ for any } \boldsymbol{\alpha} \in R^p \\ &\Leftrightarrow E[\boldsymbol{\alpha}^\top (\mathbf{X}^c\mathbf{X}^{c\top} - \mathbf{Y}^c\mathbf{Y}^{c\top})\boldsymbol{\alpha}] = 0, \text{ for any } \boldsymbol{\alpha} \in R^p.\end{aligned}$$

Thus, testing whether Σ_1 and Σ_2 amounts to testing whether

$$RI_w(\mathbf{X}, \mathbf{Y}) \triangleq \int E^2[\boldsymbol{\alpha}^\top (\mathbf{X}^c\mathbf{X}^{c\top} - \mathbf{Y}^c\mathbf{Y}^{c\top})\boldsymbol{\alpha}] w(\boldsymbol{\alpha}) d\boldsymbol{\alpha} = 0, \quad (2.1)$$

where $w(\boldsymbol{\alpha})$ is a positive weight function. A critical observation is that $RI_w(\mathbf{X}, \mathbf{Y})$ may be evaluated easily for certain properly chosen w . The following Theorem 1 enables us to derive an explicit form for (2.1) and obtain our difference measure of two covariances.

Theorem 1. *If $w(\boldsymbol{\alpha})$ is a p -dimensional standard normal density function, then*

$$RI(\mathbf{X}, \mathbf{Y}) \triangleq RI_w(\mathbf{X}, \mathbf{Y}) = [\text{tr}(\Sigma_1 - \Sigma_2)]^2 + 2\text{tr}(\Sigma_1 - \Sigma_2)^2, \quad (2.2)$$

and $RI(\mathbf{X}, \mathbf{Y}) \geq 0$ with the equality holds if and only if $\Sigma_1 = \Sigma_2$.

Remark 1. The p -dimensional standard normal random vector can be expressed as the independent product between a uniformly distributed random vector on the unit sphere S_1^{p-1} and the radial random variable. We choose the standard normal random density function to treat every unit vector equally in the integration.

Moreover, the multivariate normal distribution is a special case of a multivariate stable distribution, and multivariate stable distributions have been used as the weight functions by Chen et al. (2019). As future work, it may be worthwhile to consider other weight functions such as uniform distribution (Zhu et al., 2017; Kim et al., 2020) and Bernoulli distribution (Qiu et al., 2021).

Note that the second term on the right hand side of (2.2) is the Frobenius norm of the difference between the two covariance matrix. the test can be designed to be powerful when there are many small disturbances between Σ_1 and Σ_2 . Meanwhile, unlike Li and Chen (2012), the same test can be powerful when the two covariance matrices only differ by a small amount in the diagonal, due to the first term on the right hand side of (2.2). It represents the square of the difference between the diagonal elements of the two covariance matrix. If nonzero signals of the difference between the diagonal elements of Σ_1 and Σ_2 are weakly dense with almost the same sign, $RI(\mathbf{X}, \mathbf{Y})$ should be higher powerful than a different test statistic in which $[tr(\Sigma_1 - \Sigma_2)]^2$ is replaced by $\sum_{k=1}^p (\Sigma_{1,kk} - \Sigma_{2,kk})^2$. In addition, the estimation of $RI(\mathbf{X}, \mathbf{Y})$ does not need the consistent estimates of the covariance matrices since $RI(\mathbf{X}, \mathbf{Y})$ is based on the trace of matrices. In the following, we will obtain an unbiased estimator of $RI(\mathbf{X}, \mathbf{Y})$ which is our desired test statistic to test (1.1) .

Denote

$$\begin{aligned}
 A_m^1 &= \frac{1}{m(m-1)} \sum_{i \neq j} (\mathbf{X}_i^\top \mathbf{X}_i)(\mathbf{X}_j^\top \mathbf{X}_j) - \frac{2}{m(m-1)(m-2)} \sum_{i,j,k}^* \mathbf{X}_i^\top \mathbf{X}_j \mathbf{X}_k^\top \mathbf{X}_k \\
 &\quad + \frac{1}{m(m-1)(m-2)(m-3)} \sum_{i,j,k,l}^* \mathbf{X}_i^\top \mathbf{X}_j \mathbf{X}_k^\top \mathbf{X}_l, \\
 B_n^1 &= \frac{1}{n(n-1)} \sum_{i \neq j} (\mathbf{Y}_i^\top \mathbf{Y}_i)(\mathbf{Y}_j^\top \mathbf{Y}_j) - \frac{2}{n(n-1)(n-2)} \sum_{i,j,k}^* \mathbf{Y}_i^\top \mathbf{Y}_j \mathbf{Y}_k^\top \mathbf{Y}_k \\
 &\quad + \frac{1}{n(n-1)(n-2)(n-3)} \sum_{i,j,k,l}^* \mathbf{Y}_i^\top \mathbf{Y}_j \mathbf{Y}_k^\top \mathbf{Y}_l, \\
 C_{m,n}^1 &= \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n (\mathbf{X}_i^\top \mathbf{X}_i)(\mathbf{Y}_j^\top \mathbf{Y}_j) - \frac{1}{nm(m-1)} \sum_{i,k}^* \sum_{j=1}^n \mathbf{Y}_j^\top \mathbf{Y}_j \mathbf{X}_i^\top \mathbf{X}_k \\
 &\quad - \frac{1}{mn(n-1)} \sum_{i,k}^* \sum_{j=1}^m \mathbf{X}_j^\top \mathbf{X}_j \mathbf{Y}_i^\top \mathbf{Y}_k + \frac{1}{m(m-1)n(n-1)} \sum_{i,k}^* \sum_{j,l}^* \mathbf{X}_i^\top \mathbf{X}_k \mathbf{Y}_j^\top \mathbf{Y}_l, \\
 A_m^2 &= \frac{1}{m(m-1)} \sum_{i \neq j} (\mathbf{X}_i^\top \mathbf{X}_j)^2 - \frac{2}{m(m-1)(m-2)} \sum_{i,j,k}^* \mathbf{X}_i^\top \mathbf{X}_j \mathbf{X}_j^\top \mathbf{X}_k \\
 &\quad + \frac{1}{m(m-1)(m-2)(m-3)} \sum_{i,j,k,l}^* \mathbf{X}_i^\top \mathbf{X}_j \mathbf{X}_k^\top \mathbf{X}_l, \\
 B_n^2 &= \frac{1}{m(m-1)} \sum_{i \neq j} (\mathbf{Y}_i^\top \mathbf{Y}_j)^2 - \frac{2}{m(m-1)(m-2)} \sum_{i,j,k}^* \mathbf{Y}_i^\top \mathbf{Y}_j \mathbf{Y}_j^\top \mathbf{Y}_k \\
 &\quad + \frac{1}{m(m-1)(m-2)(m-3)} \sum_{i,j,k,l}^* \mathbf{Y}_i^\top \mathbf{Y}_j \mathbf{Y}_k^\top \mathbf{Y}_l, \\
 C_{m,n}^2 &= \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n (\mathbf{X}_i^\top \mathbf{Y}_j)^2 - \frac{1}{nm(m-1)} \sum_{i,k}^* \sum_{j=1}^n \mathbf{X}_i^\top \mathbf{Y}_j \mathbf{Y}_j^\top \mathbf{X}_k \\
 &\quad - \frac{1}{mn(n-1)} \sum_{i,k}^* \sum_{j=1}^m \mathbf{Y}_i^\top \mathbf{X}_j \mathbf{X}_j^\top \mathbf{Y}_k + \frac{1}{m(m-1)n(n-1)} \sum_{i,k}^* \sum_{j,l}^* \mathbf{X}_i^\top \mathbf{Y}_j \mathbf{X}_k^\top \mathbf{Y}_l.
 \end{aligned}$$

where $\sum^*$ denotes summation over mutually distinct indices. Then the proposed

sample test statistic is

$$RI_{m,n} = A_m^1 - 2C_{m,n}^1 + B_n^1 + 2(A_m^2 - 2C_{m,n}^2 + B_n^2), \quad (2.3)$$

which is an unbiased estimator of $RI(\mathbf{X}, \mathbf{Y})$.

Remark 2. The computation cost is at the order of $\max\{pn^4, pm^4\}$ if computed $RI_{m,n}$ directly. In fact, the computation burden comes from the last two sums in $A_m^1, B_n^1, A_m^2, B_n^2$ and the last three in $C_{m,n}^1, C_{m,n}^2$. Since $RI_{m,n}$ is invariant under the location shift, we can assume without loss of generality that $\mu_1 = \mu_2 = 0$. Under this assumption, the last two sums in $A_m^1, B_n^1, A_m^2, B_n^2$ and the last three in $C_{m,n}^1, C_{m,n}^2$ are all of smaller order than the first. This indicates that we can first transform data $\mathbf{X}_i$ to $\mathbf{X}_i - \bar{\mathbf{X}}$ and $\mathbf{Y}_j$ to $\mathbf{Y}_j - \bar{\mathbf{Y}}$, and then compute only the first term in $A_m^1, B_n^1, A_m^2, B_n^2, C_{m,n}^1, C_{m,n}^2$. These will reduce the computation cost to the order of $\max\{pn^2, pm^2\}$ without affecting the asymptotic properties of our proposed test. Meanwhile, by similar techniques in Zhong et al. (2019), we can reduce the computation cost for the third terms in A_m^1, B_n^1 , the last two sums in A_m^2, B_n^2 , and the last three in $C_{m,n}^2$ to the order of $\max\{pn^2, pm^2\}$ and can maintain their unbiasedness. However, the second terms in A_m^1, B_n^1 and the second and the last three in $C_{m,n}^1$ can not reduce the order of $\max\{pn^2, pm^2\}$ to maintain their unbiasedness by this techniques.

2.1 Asymptotic properties

2.1 Asymptotic properties

To establish the limiting distribution of $\text{RI}_{m,n}$, we assume the following three conditions:

E1. There exist a $p \times m_1$ matrix Γ_1 , a $p \times m_2$ matrix Γ_2 , m_1 -dimensional random vectors $\{\mathbf{Z}_{1j}\}_{j=1}^m$, and m_2 -dimensional random vectors $\{\mathbf{Z}_{2j}\}_{j=1}^n$, such that $\mathbf{X}_j = \boldsymbol{\mu}_1 + \Gamma_1 \mathbf{Z}_{1j}$ for $j = 1, \dots, m$, and $\mathbf{Y}_j = \boldsymbol{\mu}_2 + \Gamma_2 \mathbf{Z}_{2j}$ for $j = 1, \dots, n$. And $\Gamma_i, i = 1, 2$, and $\mathbf{Z}_{ij} = (\mathbf{Z}_{ij1}, \dots, \mathbf{Z}_{ijm_i})^\top$ for $i = 1, j = 1, \dots, m$ and $i = 2, j = 1, \dots, n$ satisfy:

1. $\Gamma_1 \Gamma_1^\top = \Sigma_1$ and $\Gamma_2 \Gamma_2^\top = \Sigma_2$ with $\min\{m_1, m_2\} \geq p$.
2. $\{\mathbf{Z}_{1j}\}_{j=1}^m$ and $\{\mathbf{Z}_{2j}\}_{j=1}^n$ are independent and identically distributed, respectively, with $E\mathbf{Z}_{1j} = \mathbf{0}$, $\text{Var}(\mathbf{Z}_{1j}) = \mathbf{I}_{m_1}$, and $E\mathbf{Z}_{2j} = \mathbf{0}$, $\text{Var}(\mathbf{Z}_{2j}) = \mathbf{I}_{m_2}$, where $\mathbf{I}_{m_i}$ is the $m_i \times m_i$ identity matrix.
3. $\sup_{i,k} E|\mathbf{Z}_{ijk}|^8 < \infty$ and $E\mathbf{Z}_{ijk}^4 = 3 + \Delta_i$ for some constant Δ_i . Also,

$$E\left(\mathbf{Z}_{ijl_1}^{\varsigma_1} \cdots \mathbf{Z}_{ijl_q}^{\varsigma_q}\right) = E\left(\mathbf{Z}_{ijl_1}^{\varsigma_1}\right) \cdots E\left(\mathbf{Z}_{ijl_q}^{\varsigma_q}\right) \quad (2.4)$$

for any positive integers q and ς_l such that $\sum_{l=1}^q \varsigma_l \leq 8$, and $l_1, l_2, \dots, l_q$ are distinct indices.

E2. As $\min\{m, n\} \rightarrow \infty$, $p \rightarrow \infty$, and for any $i, j, k, l \in \{1, 2\}$, $\text{tr}(\Sigma_i \Sigma_j) \rightarrow \infty$ and

$$\text{tr}\{(\Sigma_i \Sigma_j)(\Sigma_k \Sigma_l)\} = o\{\text{tr}(\Sigma_i \Sigma_j) \text{tr}(\Sigma_k \Sigma_l)\}. \quad (2.5)$$

E3. As $\min\{m, n\} \rightarrow \infty$, $m/(m+n) \rightarrow \tau \in (0, 1)$.

2.1 Asymptotic properties

Remark 3. Condition E1 gives rise to a general multivariate model for high-dimensional data analysis, which includes commonly used distributions such as the multivariate normal distribution (Chen et al., 2010b; Srivastava and Yanagihara, 2010; Li and Chen, 2012). According to Chen and Qin (2010a), $\min\{m_1, m_2\} \geq p$ means that the rank and eigenvalues of Σ_1 or Σ_2 are not affected by the transformation. According to Chen and Qin (2010a); He and Chen (2018), (2.4) can be viewed as a pseudo-independent condition of $\mathbf{Z}_{ij}$. namely, a relaxed independence relation that allows some margin over probabilities (Kim and Lesser, 2008). Obviously, if $\mathbf{Z}_{ij}$ has independent components, then (2.4) is true.

In Condition E2, we do not require a direct relationship between p and n . We know cases where E2 does not imply a relation between p and n . For example, if all the eigenvalues of Σ_i are bounded away from zero and infinity, E2 holds. Meanwhile, some of the commonly encountered covariance structures satisfy Condition E2 (Chen et al., 2010b).

Condition E3 is a standard regularity assumption in two-sample problems, which guarantees that m and n go to infinity proportionally.

Theorem 2. *Under Conditions E1-E3, as $\min\{m, n\} \rightarrow \infty$, we have*

$$\frac{RI_{m,n} - RI(\mathbf{X}, \mathbf{Y})}{\sigma_{m,n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1),$$

where $\sigma_{m,n}^2$ is defined in (A.2) in Supplementary Material.

2.1 Asymptotic properties

Under H_0 , we can obtain $\text{RI}(\mathbf{X}, \mathbf{Y}) = 0$, and

$$\sigma_{0,m,n}^2 = 24 \left(\frac{1}{m} + \frac{1}{n} \right)^2 \text{tr}^2(\Sigma^2).$$

Therefore, we obtain the following corollary.

Corollary 1. *Under Conditions E1-E3 and $H_0 : \Sigma_1 = \Sigma_2 = \Sigma$, as $\min\{m, n\} \rightarrow \infty$, we have*

$$RI_{m,n}/\sigma_{0,m,n} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1).$$

To formulate a test procedure, we need to estimate $\sigma_{0,m,n}$. Since $EA_m^2 = \text{tr}(\Sigma_1^2)$, $EB_n^2 = \text{tr}(\Sigma_2^2)$, the following is a consistent estimate $\hat{\sigma}_{0,m,n}$ of $\sigma_{0,m,n}$ under H_0 :

$$\hat{\sigma}_{0,m,n} = 2\sqrt{6} \left(\frac{1}{m} A_m^2 + \frac{1}{n} B_n^2 \right).$$

Furthermore, the following theorem assures that $\hat{\sigma}_{0,m,n}$ is ratio-consistent to $\sigma_{0,m,n}$.

Theorem 3. *Under Conditions E1-E3 and $H_0 : \Sigma_1 = \Sigma_2 = \Sigma$, as $\min\{m, n\} \rightarrow \infty$, we have*

$$RI_{m,n}/\hat{\sigma}_{0,m,n} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1). \quad (2.6)$$

As we shall derive in the Supplementary Material,

$$\frac{A_m^2}{\text{tr}(\Sigma_1^2)} \xrightarrow{P} 1, \quad \frac{B_n^2}{\text{tr}(\Sigma_2^2)} \xrightarrow{P} 1, \quad \text{and} \quad \frac{\hat{\sigma}_{0,m,n}}{\sigma_{0,m,n}} \xrightarrow{P} 1.$$

Theorem 3 follows from Corollary 1 and the Slutsky's theorem. Therefore, the proposed test with a nominal θ level of significance rejects H_0 if $RI_{m,n} \geq \hat{\sigma}_{0,m,n} z_\theta$,

2.1 Asymptotic properties

where z_θ is the upper- θ quantile of $\mathcal{N}(0, 1)$. The approximation result in the Supplementary Material indicates that the standard normal distribution can adequately substitute for the null distribution of $\text{RI}_{m,n}/\hat{\sigma}_{0,m,n}$.

Next, we study the power of our proposed test. Let $g_{m,n}(\Sigma_1, \Sigma_2; \theta) = P(\text{RI}_{m,n} \geq \hat{\sigma}_{0,m,n} z_\theta | H_1)$ be the power of the proposed test under $H_1 : \Sigma_1 \neq \Sigma_2$. Let $\text{SNR}_{m,n}(\Sigma_1, \Sigma_2) = \text{RI}(\mathbf{X}, \mathbf{Y})/\sigma_{m,n}$ and $\gamma_{m,n} = \left[\frac{\text{tr}(\Sigma_1^2)}{m} + \frac{\text{tr}(\Sigma_2^2)}{n} \right] / \text{RI}(\mathbf{X}, \mathbf{Y})$, then we obtain the following Theorem 4.

Theorem 4. *Under Conditions E1-E3 and $H_1 : \Sigma_1 \neq \Sigma_2$, we have*

$$\lim_{m,n \rightarrow \infty} g_{m,n}(\Sigma_1, \Sigma_2; \theta) \geq \lim_{m,n \rightarrow \infty} \Phi\left(-\sqrt{2}z_\theta + \text{SNR}_{m,n}(\Sigma_1, \Sigma_2)\right),$$

where $\Phi(\cdot)$ is the cumulative standard normal distribution function.

Theorem 4 indicates that the power of our proposed test is bounded from below. Meanwhile, the power is mainly determined by $\text{SNR}_{m,n}(\Sigma_1, \Sigma_2)$. From (A.2) in Supplementary Material, we have

$$\sigma_{m,n}^2 \leq 24 \left[\frac{\text{tr}(\Sigma_1^2)}{m} + \frac{\text{tr}(\Sigma_2^2)}{n} \right]^2 + 20 \max\{2 + \Delta_1, 2 + \Delta_2\} \left[\frac{\text{tr}(\Sigma_1^2)}{m} + \frac{\text{tr}(\Sigma_2^2)}{n} \right] \text{RI}(\mathbf{X}, \mathbf{Y}),$$

that is,

$$\text{SNR}_{m,n}(\Sigma_1, \Sigma_2) \geq \left[24\gamma_{m,n}^2 + 20 \max\{2 + \Delta_1, 2 + \Delta_2\} \gamma_{m,n} \right]^{-1/2}.$$

Therefore, when $\gamma_{m,n} \rightarrow 0$ as $\min\{m, n\} \rightarrow \infty$, we have $\text{SNR}_{m,n}(\Sigma_1, \Sigma_2) \rightarrow \infty$.

Thus, we have

$$\lim_{m,n \rightarrow \infty} g_{m,n}(\Sigma_1, \Sigma_2; \theta) = 1.$$

2.1 Asymptotic properties

Especially, we consider the following three cases.

Case I: Let $\Sigma_1 = r\mathbf{I}_p + AR(0.1)$ and $\Sigma_2 = AR(0.1)$, where $AR(\rho) = (a_{ij})_{p \times p}$ is a covariance matrix with $a_{ij} = \rho^{|i-j|}$ for $i, j = 1, \dots, p$. Then, we have the following Corollary 2.

Corollary 2. *Under Conditions E1-E3, and $\lim_{m,n,p \rightarrow \infty} \sqrt{p}(m+n)r^2 = c$, where $0 < c < \infty$, then we have*

$$\lim_{m,n \rightarrow \infty} g_{m,n}(\Sigma_1, \Sigma_2; \theta) = 1.$$

Corollary 2 shows that our proposed test is very powerful when there are many small diagonal disturbances between the two covariance matrices. In addition, under the conditions in Corollary 2, the signal-to-noise ratio of the method proposed by Li and Chen (2012) diminishes to 0. Therefore, the power of the test proposed by Li and Chen (2012) has a low bounded from below.

Case II: Let $\Sigma_1 = \mathbf{I}_p$, $\Sigma_2 = \mathbf{I}_p + H(\varpi_0, \varpi_1, p_0)$, where $H(\varpi_0, \varpi_1, p_0) = (h_{ij})_{p \times p}$ with $h_{ij} = 0$ except $h_{ii} = \varpi_0, i = 1, \dots, p_0$ and $h_{i,i+1} = h_{i+1,i} = \varpi_1, i = 1, \dots, p_0 - 1$. Then, we have the following Corollary 3.

Corollary 3. *Under Conditions E1-E3, $\frac{p}{(m+n)p_0^2\varpi_0^2} = o(1)$ and $np_0\varpi_0 \rightarrow \infty$, then we have*

$$\lim_{m,n \rightarrow \infty} g_{m,n}(\Sigma_1, \Sigma_2; \theta) = 1.$$

According to Cai et al. (2013), we take $\varpi_0 = O(\sqrt{p/(m+n)})$ and $p_0 = p^{1/4}$.

Corollary 3 shows that the proposed test is powerful. Meanwhile, the method proposed by Cai et al. (2013) is also powerful under this case.

Case III: Let $\Sigma_1 = \mathbf{I}_p$, $\Sigma_2 = \mathbf{I}_p + M$, where M is a $p \times p$ matrix with $M_{ii} = 0$ and $M_{ij} = \omega_1$ for $i \neq j$. Then, we have the following Corollary 4.

Corollary 4. *Under Conditions E1-E3, $(m+n)p\omega_1^2 \rightarrow \infty$ as $m, n, p \rightarrow \infty$, then we have*

$$\lim_{m,n \rightarrow \infty} g_{m,n}(\Sigma_1, \Sigma_2; \theta) = 1.$$

Corollary 4 indicates that our proposed test is also very powerful under some conditions when the diagonals are the same and there are many small non-diagonal disturbances between the two covariance matrices.

3. Simulation Studies

We carry out numerical simulations to investigate finite-sample performance of our proposed method. We consider our proposed method (RI) with the method proposed by Li and Chen (2012) (LC), the method proposed by Cai et al. (2013) (Cai), the method proposed by Zhu (2017) (sLED), the method proposed by Chang et al. (2017) ($\Psi_{B,\alpha}$), T2 method proposed by Zheng et al. (2020), and a scale-invariant power enhancement test based on Fisher's method (Yu et al., 2020) ($F_{m,n}$). We set the nominal level of significance at 0.05. We choose the

sample sizes as $m = n = 60, 100$, and $m = 200, n = 60$, and the dimension as $p = 300, 500, 800, 1000, 1200, 1500$. All empirical sizes and powers are calculated from 1000 replications. For Σ_1 and Σ_2 , we consider the following six scenarios:

Scenario 1: $\Sigma_1^{(1)} = \mathbf{I}_p$, the identity matrix, $\Sigma_2^{(1)} = \mathbf{I}_p + H(0.04, 0.2, \lfloor 0.3p \rfloor)$, where $H(\varpi_0, \varpi_1, k) = (h_{ij})_{p \times p}$ with $h_{ij} = 0$ except $h_{ii} = \varpi_0, i = 1, \dots, k$ and $h_{i,i+1} = h_{i+1,i} = \varpi_1, i = 1, \dots, k-1$;

Scenario 2: $\Sigma_1^{(2)} = \mathbf{I}_p, \Sigma_2^{(2)} = \mathbf{I}_p + H(0.04, 0.2, p)$;

Scenario 3: $\Sigma_1^{(3)} = \mathbf{I}_p, \Sigma_2^{(3)} = \mathbf{I}_p + \Sigma_*^{(3)}$, where $\Sigma_*^{(3)} = (\sigma_{ij,*}^{(3)})$ is a $p \times p$ matrix with $\sigma_{ii,*}^{(3)} = 0$ for $i = 1, \dots, p$ and $\sigma_{ij,*}^{(3)} = 1/\sqrt{p}$ for $i \neq j$.

Scenario 4: $\Sigma_1^{(4)} = \mathbf{I}_p, \Sigma_2^{(4)} = \mathbf{I}_p + H(4, 0.05, \lfloor 0.02p \rfloor)$;

Scenario 5: $\Sigma_1^{(5)} = \Sigma_*^{(5)} + \delta_0 \mathbf{I}_p, \Sigma_2^{(5)} = \Sigma_*^{(5)} + \delta_0 \mathbf{I}_p + U$, where $\Sigma_*^{(5)} = (\sigma_{ij,*}^{(5)}) = D^{1/2} C D^{1/2}$, $D = \text{diag}(d_1, \dots, d_p)$ and $d_1, \dots, d_p \sim^{i.i.d.} \text{Unif}(0.5, 2.5)$, $C = (c_{ij})$ with $c_{ii} = 1, c_{ij} = 0.5$ for $5(k-1) + 1 \leq i \neq j \leq 5k$, where $k = 1, \dots, \lfloor p/5 \rfloor$ and $c_{ij} = 0$ otherwise. U is a $p \times p$ symmetric matrix with four nonzero entries from $\text{Unif}(0, 4) \times \max_{1 \leq j \leq p} \sigma_{jj,*}^{(5)}$ randomly located in the upper triangle, and another four located in the lower triangle by symmetry. $\delta_0 = |\min\{\lambda_{\min}(\Sigma_*^{(5)} + U), \lambda_{\min}(\Sigma_*^{(5)})\}| + 0.05$, where $\lambda_{\min}(A)$ denotes the minimum eigenvalue of a symmetric matrix A .

Scenario 6: $\Sigma_1^{(6)} = 0.2 \times \mathbf{I}_p + AR(0.1), \Sigma_2^{(6)} = AR(0.1)$, where $AR(\rho) = (a_{ij})_{p \times p}$ is a covariance matrix with $a_{ii} = 1$ and $a_{ij} = \rho^{|i-j|}$ for $i \neq j$;

Scenario 7: $\Sigma_1^{(7)} = 0.1 \times \mathbf{I}_p + AR(0.2)$, $\Sigma_2^{(7)} = AR(0.24)$.

Scenario 8: $\Sigma_1^{(8)} = \Sigma_*^{(8)} + \lambda_0 \mathbf{I}_p$, $\Sigma_2^{(8)} = \Sigma_*^{(8)} + Q + \lambda_0 \mathbf{I}_p$, where $\Sigma_*^{(8)} = (\sigma_{ij,*}^{(8)})_{1 \leq i,j \leq p}$ with i.i.d $\sigma_{ii,*}^{(8)} \sim \text{Unif}(1, 2)$ and $\sigma_{ij,*}^{(8)} = \{(|i-j|+1)^{2H} + (|i-j|-1)^{2H} - 2(|i-j|)^{2H}\}/2$ with $H = 0.85$ for $i \neq j$. A perturbation matrix Q has $\lfloor 0.05p \rfloor$ random non-zero elements in diagonal and in non-diagonal, respectively. $\lfloor 0.05p \rfloor/2$ non-zero elements are randomly allocated in the upper triangle part of Q and the others are in its lower triangle part by symmetry. The magnitudes of non-zero elements are randomly generated from $\text{Unif}(\tau/2, 3\tau/2)$ with $\tau = 8 \max\{\max_{1 \leq i \leq p} \sigma_{ii,*}^{(8)}, (\log p)^{1/2}\}$, and $\lambda_0 = |\min\{\lambda_{\min}(\Sigma_*^{(8)} + Q), \lambda_{\min}(\Sigma_*^{(8)})\}| + 0.05$.

Finally, the data are generated by $\mathbf{X}_i = \Sigma_1^{1/2} \mathbf{Z}_i$ for $i = 1, \dots, m$ and $\mathbf{Y}_l = \Sigma_2^{1/2} \mathbf{Z}_{m+l}$ for $l = 1, \dots, n$, where $\{\mathbf{Z}_i : i = 1, \dots, m+n\}$ are independent p -dimensional random variables with i.i.d. coordinates Z_{ij} , $j = 1, \dots, p$. We consider the following four distributions for Z_{ij} :

1. The standard normal $\mathcal{N}(0, 1)$;
2. The t -distribution with degrees of freedom 15, i.e., $t(15)$;
3. The centralized Gamma distribution with $a = 16, b = 0.25$, i.e., $\Gamma(16, 0.25)$ –4;
4. The discrete distribution has five possible values $-2, -1, 0, 3/2, 4$ with

probabilities $1/12, 4/25, 13/24, 16/75, 1/600$, i.e.,

$$Z_{ij} \sim \frac{1}{12}\delta_{-2} + \frac{4}{25}\delta_{-1} + \frac{13}{24}\delta_0 + \frac{16}{75}\delta_{3/2} + \frac{1}{600}\delta_4.$$

This distribution is used in Yang and Pan (2017), and they showed that the first four moments of Z_{ij} are the same as those of $\mathcal{N}(0, 1)$.

There are reasonably small disturbances between Σ_1 and Σ_2 , and these two covariance matrices are reasonably sparse in **Scenario 1**. This is similar to the case considered by Yang and Pan (2017). In **Scenario 2**, there are many small disturbances between Σ_1 and Σ_2 . This case was also considered by Yang and Pan (2017) and Li and Chen (2012). In **Scenario 3**, the two matrices differ only in off-diagonal entries with many weakly dense signals. **Scenario 4** deals with the sparse situation, with $\lfloor 0.02p \rfloor$ features exerting larger signals. **Scenario 5** is for the extremely sparse situation, with 8 nonzero signals in the non-diagonal entries. **Scenarios 5** was considered by Cai et al. (2013). In **Scenario 6**, the two covariance matrices only differ in the diagonal. The two covariance matrices are entirely different by a larger amount in **Scenario 7**. **Scenarios 6** and **7** were also considered by Wu and Li (2015). **Scenarios 8** was studied by Chang et al. (2017) except that a perturbation matrix Q adds some non-zero elements in diagonal.

In all scenarios, we first calculate the empirical p -values when $\Sigma_1 = \Sigma_2 = \Sigma^{(i)}$ for $i = 1, \dots, 8$. The corresponding results are given in Tables 1-15 of the

Supplementary Material. It can be seen from Tables 1-15 that the estimated p -values of the proposed RI method and the other six methods are controlled fairly well around 0.05 for all cases except the method proposed by Cai et al. (2013) for the discrete distribution.

Since the empirical powers for $m = n = 100$ and $m = 200, n = 60$ are very similar to those for $m = n = 60$, we only present the empirical powers with $\Sigma_1 = \Sigma_1^{(i)}$ and $\Sigma_2 = \Sigma_2^{(i)}$, $i = 1, \dots, 8$ for $m = n = 60$ in Figures 1-4. The empirical powers for $m = n = 100$ and $m = 200, n = 60$ are included in the Supplementary Material.

For **Scenarios 1-2**, we have the following findings:

1. Our proposed RI test is considerably more powerful than the other methods. LC is the second most powerful, suggesting its ability in detecting the covariance differences with many small disturbances, as also observed by Yang and Pan (2017). T2 is the third most powerful since it is a weighted statistic. The Cai, sLED and $\Psi_{B,\alpha}$ methods have poor power, below 0.20 in almost all cases.
2. The powers of the RI, LC, T2, and $F_{m,n}$ methods increase from **Scenario 1** to **Scenario 2**, which is expected because we increase the small disturbances in the differences between the two covariance matrices. So, RI, LC, T2, and $F_{m,n}$ gain power as the amount of deviations increases, even

if these deviations are not large. This is expected because they are defined from the Frobenius norm of the difference in the two covariance matrices. However, the other three methods are little affected as we increase the differences.

3. It is not surprising but reassuring that the power of our proposed method increases as the sample sizes n, m increases.

In **Scenario 3**, the two covariance matrices have many weakly dease signals in the non-diagonal entries. The power of RI, LC, sLED, T2, and $F_{m,n}$ is close to 1, but the power of Cai test and $\Psi_{B,\alpha}$ is lower. The results are verified by Corollary 4.

Scenario 4 contains strong and sparse signals in the main diagonal of the difference between the two covariance matrices. All seven methods perform quite well, especially when the sample size m or n is large. Meanwhile, the power of our proposed RI method is close to 1.

Scenario 5 contains the extremely sparse, strong signals in the non-diagonal differences between the two covariance matrices. The power of Cai test, T2, $\Psi_{B,\alpha}$, and $F_{m,n}$ is high, but the power of RI, LC and sLED are lower.

When the two covariance matrices differ in the main diagonal with sparsely weak signals in **Scenario 6**, our proposed RI method still has the perfect power in all cases. The powers of the other six methods are poor. The results are

verified by Corollary 2.

In **Scenario 7**, the two covariance matrices are entirely different to a large degree. The power of the RI method is close to 1, whereas the other six methods are much less powerful and not competitive with RI.

For **Scenario 8**, the two covariance matrices have long-range dependence according to Chang et al. (2017). The power of the RI method is more higher than other six methods for large p .

In conclusion, the RI method has considerably higher and more stable power than the six existing methods in a wide range of settings. Not only can it deal with the cases with many small deviations in difference of the covariance matrices, but also handle the cases with sparsely strong or sparsely weak signals. Thus it is applicable in broader applications in testing the difference between covariance matrices.

Figure 1: Empirical powers for Scenarios 1-8 with Z_{ij} follows $\mathcal{N}(0, 1)$ and $m = n = 60$.

Figure 2: Empirical powers for Scenarios 1-8 with Z_{ij} follows $t(15)$ and $m = n =$

Figure 3: Empirical powers for Scenarios 1-8 with Z_{ij} follows $\Gamma(16, 0.25) - 4$ and $m = n = 60$.

Figure 4: Empirical powers for Scenarios 1-8 with Z_{ij} follows the discrete dis-

tribution $\frac{1}{12}\delta_{-2} + \frac{4}{25}\delta_{-1} + \frac{13}{24}\delta_0 + \frac{16}{75}\delta_{3/2} + \frac{1}{600}\delta_4$ and $m = n = 60$.

4. Real data Analysis

In this section, we apply the proposed method to analyze a gene expression dataset on breast cancer from a study reported by Schemidt et al. (2008). We downloaded the dataset from

<http://bioconductor.org/packages/release/data/experiment/html/breastCancerMAINZ.html>.

The dataset contains gene expression patterns from 200 tumors of patients who were not treated by systemic therapy after surgery, and consists of 22,283 features. There were three groups of different tumor grades: 29 well differentiated tumors (group 1), 136 moderately differentiated tumors (group 2), and 35 poor/undifferentiated tumors (group 3). This dataset has been analyzed in the literature under the assumption that the two covariance matrices are equal; e.g., Teschendorff and Caldas (2008) and Haibe-Kains et al. (2012). The equality of the two covariance matrices is a very important assumption for the validity of the reported findings. We test whether this assumption is valid.

Following Gentleman et al. (2011), for quality control and computational burden considerations, we first select the features for which more than 50% of their intensities were greater than 5 and their coefficients of variation (CV) fall inside of the range (0.22, 1.0). The intensities are gene expressions as measured by Affymetrix hgu133a technology, and the coefficient of variation is the

standard deviation divided by the absolute value of the mean. We also screen the features by predetermined cutoffs, which remove low-quality features while retaining high-quality features as previously done in this dataset (Sherafatian, 2018; Chong et al., 2018; Schiffman et al., 2008). After these selection steps, 1193 features remain in our analysis. Let Σ_1 , Σ_2 and Σ_3 be the covariance matrices of these 1193 features with groups 1, 2 and 3, respectively. We standardized each of the 1193 features so that its mean is zero. Then, we apply the LC, Cai, sLED, T2, $\Psi_{B,\alpha}$, $F_{m,n}$ and RI methods to test separately (a) $H_0^{(1,2)} : \Sigma_1 = \Sigma_2$, and (b) $H_0^{(2,3)} : \Sigma_2 = \Sigma_3$. The p -values for all these methods are reported in Table 1. From Table 1, the RI method rejects both $H_0^{(1,2)}$ and $H_0^{(2,3)}$ at the significant level of 0.05, but the LC, sLED, T2, and $F_{m,n}$ methods reject $H_0^{(2,3)}$ only.

To visualize the comparison among the different methods, we plot the heat maps of $(\hat{\Sigma}_1 - \hat{\Sigma}_2)$ and $(\hat{\Sigma}_2 - \hat{\Sigma}_3)$ for the top 100 features with the largest absolute values of the two-sample t statistics, respectively, where $\hat{\Sigma}_1, \hat{\Sigma}_2, \hat{\Sigma}_3$ are the sample covariance matrices based on the selected 100 features. The corresponding results are shown in Figure 5. It is observed from Figure 5 that $(\hat{\Sigma}_2 - \hat{\Sigma}_3)$ has stronger signals than $(\hat{\Sigma}_1 - \hat{\Sigma}_2)$. Meanwhile, many moderate disturbances are present in $(\hat{\Sigma}_2 - \hat{\Sigma}_3)$. The maximum absolute value in the elements of the estimator of $(\hat{\Sigma}_2 - \hat{\Sigma}_3)$ for the 1193 features is 8.972, which is larger than the maximum absolute value in $(\hat{\Sigma}_1 - \hat{\Sigma}_2)$, i.e., 4.648. However, the diagonal in $(\hat{\Sigma}_1 - \hat{\Sigma}_2)$ has

Table 1: The corresponding p -value of the LC, Cai, sLED, T2, $\Psi_{B,\alpha}$, $F_{m,n}$, and RI methods for a gene expression dataset

Method	LC	Cai	sLED	T2	$\Psi_{B,\alpha}$	$F_{m,n}$	RI
$H_0^{(1,2)}$	0.105	0.104	0.330	0.555	0.056	0.061	1.110×10^{-16}
$H_0^{(2,3)}$	3.050×10^{-6}	0.093	0.000	8.157×10^{-9}	0.052	4.577×10^{-6}	3.030×10^{-9}

much stronger signals than those in $(\hat{\Sigma}_2 - \hat{\Sigma}_3)$.

Both Equation (2.2) and the simulation results of **Scenarios 4-5** reveal that that the our proposed method is more powerful than the other three methods when there are more signals in the diagonal for the difference of two covariance matrices. This explains why our proposed method rejects $H_0^{(1,2)}$ whereas the others did not. Consequently, our method is only one that is able to detect a more subtle but very important difference in this commonly analyzed dataset.

Figure 5: (a) the heat map of $(\hat{\Sigma}_1 - \hat{\Sigma}_2)$ for the selected 100 features; (b) the heat map of $(\hat{\Sigma}_2 - \hat{\Sigma}_3)$ for the selected 100 features.

5. Discussion

Conducting inference for high-dimensional covariance matrix is highly challenging. In this paper, we make a novel use of a random integration technique to develop a two-covariance matrix test statistic. This test can be performed without estimating the covariance matrices, which is known to be extremely difficult in high dimension data. We investigate both theoretical properties and numerical performance of our method, and find that it is not only competitive but also often much more powerful than the existing methods in both simulation studies and a real data analysis when there are many small diagonal disturbances between the two covariance matrices.

There are several issues that warrant further investigation. First, a general multivariate model is applied to obtain asymptotic results. Although it is a common assumption in the literature, it would be useful to investigate the asymptotic properties of our proposed method in weaker conditions, for example, assumption 2.1 in Han and Wu (2020). Second, we consider a two-sample test for high-dimensional covariance matrices. As in Zheng et al. (2020), it would be interesting to extend our method for testing the equality of more than two covariance matrices. Third, we will extend the proposed method to test for high dimensional correlation matrices, which is a more difficult task than testing the **covariance matrices** (Zheng et al., 2019). Finally, we use the standard multivari-

ate normal density function as the weight function during the construction of our test statistic. This is a common practice, but it is worthy investigating other choices that may perform better under various settings.

Supplementary Materials

The Supplementary Material includes detailed proofs of the theoretical results and additional simulation results.

Acknowledgements

Wang and Zhang contributed equally to this article. Jiang's research is partially supported by NSFC(12171203) and the Natural Science Foundation of Guangdong (2019A1515011830). Wen's research is partially supported by NSFC(12171449, 11801540) and Natural Science Foundation of Anhui Province (BJ2040170017). Wang's research is partially supported by NSFC(72171216, 71991474), the International Science & Technology cooperation program of Guangdong, China(2016B050502007), The Key Research and Development Program of Guangdong, China(2019B020228001), and Science and Technology Program of Guangzhou, China (202002030129). Zhang's research is partially supported by U.S. National Institutes of Health (R01HG010171, R01MH116527) and National Science Foundation (DMS1722544).

REFERENCES

References

- Anderson, T. W.(2003). An introduction to multivariate statistical analysis(3rd ed.). *The Annals of Applied Statistics*. New York: Wiley-Interscience.
- Bai, Z., Jiang, D., Yao, J.-F. and Zheng, S. (2009). Corrections to LRT on large-dimensional covariance matrix by RMT. *The Annals of Statistics* **37**, 3822 – 3840.
- Bollerslev, T., Meddahi, N. and Nyawa, S. (2019). High-dimensional multivariate realized volatility estimation. *Journal of Econometrics* **212**, 116 – 136.
- Cai, T., Liu, W. and Xia, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *Journal of the American Statistical Association* **108**, 265 – 277.
- Cai, T. T. (2017). Global testing and large-scale multiple testing for high-dimensional covariance structures. *Annals Review of Statistics and Its Application* **4**, 423– 446.
- Chang, J., Zhou, W., Zhou, W.-X. and Wang, L. (2017). Comparing large covariance matrices under weak conditions on the dependence structure and its application to gene clustering. *Biometrics* **73**, 31 – 4.
- Chen, S. X., Qin, Y. L. (2010a). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics* **38**, 808 – 835.
- Chen, S. X., Zhang, L. X. and Zhong, P. S. (2010b). Tests for high-dimensional covariance matrices. *Journal of the American Statistical Association* **105**, 810 – 819.
- Chen, F., Meintanis, S. G. and Zhu, L. (2019). On some characterizations and multidimensional criteria for testing homogeneity, symmetry and independence. *Journal of Multivariate Analysis* **173**, 125 – 144.

REFERENCES

- Chong, J., Soufan, O., Li, C., Caraus, I., Li, S., Bourque, G., Wishart, D. and Xia, J. (2018). Metaboanalyst 4.0: towards more transparent and integrative metabolomics analysis. *Nucleic Acids Research* **46**, 486 – 494.
- Gentleman, R., Carey, V., Huber, W. and Hahne, F. (2011). Genefilter: Methods for filtering genes from microarray experiments. *R package version 1*.
- Haibe-Kains, B., Desmedt, C., Loi, S., Culhane, A. C., Bontempi, G., Quackenbush, J. and Sotiriou, C. (2012). A three-gene model to robustly identify breast cancer molecular subtypes. *Journal of the National Cancer Institute* **104**, 311 – 325.
- He, J. and Chen, S. X. (2018). High-dimensional two-sample covariance matrix testing via super-diagonals. *Statistica Sinica* **28**, 2671–2696.
- Han, Y. and Wu, W.B. (2020). Test for high dimensional covariance matrices. *The Annals of Statistics* **48**, 3565 – 3588.
- He, Y., Xu, G., Wu, C. and Pan, W. (2021). Asymptotically independent u-statistics in high-dimensional testing. *The Annals of Statistics* **49**, 154 – 181.
- Igolkina, A., Aromokus, C., Newman, J., Evgrafov, O., McIntyre, L., Nuzhdin, S. and Samsonova, M.(2018). Analysis of gene expression variance in schizophrenia using structural equation modeling. *Frontiers in Molecular Neuroscience* **11**, 192 – 192.
- Jiang, T. and Qi, Y. (2015). Likelihood ratio tests for high-dimensional normal distributions. *Scandinavian Journal of Statistics* **42**, 988 – 1009.

REFERENCES

- Jiang, T. and Yang, F. (2013). Central limit theorems for classical likelihood ratio tests for high-dimensional normal distributions. *The Annals of Statistics* **41**, 2029 – 2074.
- Kim, Y. and Lesser, V. (2008). Finding minimum data requirements using pseudo-independence. In *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* **2**, 57–64.
- Kim, I., Balakrishnan, S. and Wasserman, L. (2020). Robust multivariate nonparametric tests via projection averaging. *The Annals of Statistics* **48**, 3417 – 3441.
- Le Bihan, D., Mangin, J.-F., Poupon, C., Clark, C. A., Pappata, S., Molko, N. and Chabrita, H.(2001), Diffusion tensor imaging: concepts and applications. *Journal of Magnetic Resonance Imaging* **13**, 534 – 546.
- Li, J. and Chen, S. X. (2012). Two sample tests for high-dimensional covariance matrices. *The Annals of Statistics* **40**, 908 – 940.
- Pan, W., Tian, Y., Wang, X. and Zhang, H. (2018). Ball divergence: nonparametric two sample test. *The Annals of Statistics* **46**, 1109 – 1137.
- Qiu, T., Xu, W. and Zhu, L. (2021). Two-sample test in high dimensions through random selection. *Computational Statistics & Data Analysis* **106**, 107218.
- Roberts, A. G., Catchpole, D. R. and Kennedy, P. J. (2018). Variance-based feature selection for classification of cancer subtypes using gene expression data. In *2018 International Joint Conference on Neural Networks(IJCNN)*. IEEE.

REFERENCES

- Schiffman, C., Peterick, L., Perttula, K., Yano, Y., Carlsson, H., Whitehead, T., Metayer, C., Hayes, J., Rapaport, S. and Dudoit, S. (2008). Filtering procedures for untargeted lc-ms metabolomics data. *BMC bioinformatics* **20**, 334 – 334.
- Schmidt, M., Böhm, D., Von Törne, C., Steiner, E., Puhl, A., Pilch, H., Lehr, H.-A., Hengstler, J. G., Kölbl, H. and Gehrmann, M. (2008). The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Research* **68**, 5405 – 5413.
- Schott, J. R. (2007). A test for the equality of covariance matrices when the dimension is large relative to the sample sizes. *Computational Statistics & Data Analysis* **51**, 6335 – 6542.
- Sherafatian, M. (2018). Tree-based machine learning algorithms identified minimal set of mirna biomarkers for breast cancer diagnosis and molecular subtyping. *Gene* **677**, 111 – 118.
- Srivastava, M. S. and Yanagihara, H. (2010). Testing the equality of several covariance matrices with fewer observations than the dimension. *Journal of Multivariate Analysis* **101**, 1319 – 1329.
- Teschendorff, A. E. and Caldas, C. (2008). A robust classifier of high predictive value to identify good prognosis patients in er-negative breast cancer *Breast Cancer Research and Treatment* **4**, 1 – 11.
- Wu, T.-L. and Li, P. (2015). Tests for high-dimensional covariance matrices using random matrix projection. *arXiv preprint arXiv: 1511.01611*.
- Yang, Q. and Pan, G. (2017). Weighted statistic in detecting faint and sparse alternatives for high-dimensional covariance matrices. *Journal of the American Statistical Association* **112**, 118 – 200.
- Yu, X., Li, D. and Xue, L. (2020). Fisher’s combined probability test for high-dimensional covariance

REFERENCES

matrices. *arXiv preprint arXiv: 2006.00426*.

Zheng, S., Lin, R., Guo, J. and Yin, G. (2020). Testing homogeneity of high-dimensional covariance matrices. *Statistical Sinica* **30**, 35 – 53.

Zhu, L., Lei, J., Devlin, B. and Roeder, K. (2017). Testing high-dimensional covariance matrices, with application to detecting schizophrenia risk genes. *Annals of Applied Statistics* **11**, 1810 – 1831.

Zhu, L., Xu, K., Li, R. and Zhong, W. (2017). Projection correlation between two random vectors. *Biometrika* **104**, 829 – 843.

Zheng, S., Cheng, G., Guo, J. and Zhu, H. (2019). Test for high dimensional correlation matrices. *Annals of Statistics* **47**, 2887 – 2921.

Zhong, P. S., Li, R. and Santo, S. (2019). Homogeneity tests of covariance matrices with high-dimensional longitudinal data. *Biometrika* **106**, 619 – 634.

Department of Statistics, College of Economics, Jinan University, Guangzhou, GD 510632, China.

E-mail: tjiangyl@jnu.edu.cn

Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei, AH 230026, China.

E-mail: wench@ustc.edu.cn

Department of Statistical Science, School of Mathematics, Sun Yat-Sen University, Guangzhou, GD 510275, China.

E-mail: jiangyk3@mail2.sysu.edu.cn

REFERENCES

Department of Statistics and Finance, School of Management, University of Science and Technology of

China, Hefei, AH 230026, China.

E-mail: wangxq20@ustc.edu.cn

Department of Biostatistics, School of Public Health, Yale University, New Haven, CT 06520-8034, U.S.A.

E-mail: heping.zhang@yale.edu