Statistica Sinica Preprint No: SS-2017-0275	
Title	Testing homogeneity of high-dimensional covariance matrices
Manuscript ID	SS-2017-0275
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202017.0275
Complete List of Authors	Shurong Zheng Ruitao Lin Jianhua Guo and Guosheng Yin
Corresponding Author	Shurong Zheng
E-mail	zhengsr@nenu.edu.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0275

Statistica Sinica

TESTING HOMOGENEITY OF HIGH-DIMENSIONAL COVARIANCE MATRICES

Shurong Zheng[1], Ruitao Lin[2], Jianhua Guo[1] and Guosheng Yin[3]

[1]*Northeast Normal University,* [2]*The University of Texas MD Anderson Cancer Center and* [3]*The University of Hong Kong*

Abstract: Testing the homogeneity of multiple high-dimensional covariance matrices is becoming increasingly critical in multivariate statistical analyses owing to the emergence of big data. Many existing homogeneity tests for covariance matrices focus on two populations, and/or specific situations, sparse or dense alternatives. As a result, these methods are not suitable for general cases that include multiple groups. Therefore, we propose a power-enhancement high-dimensional test for multi-sample comparisons of covariance matrices, which includes homogeneity tests of two matrices as a special case. The proposed tests do not require a distributional assumption, and can handle both sparse and non-sparse structures. Based on random-matrix theory, the asymptotic normality properties of our tests are established under both the null and the alternative hypotheses. Numerical studies demonstrate the substantial gain in power for the proposed method. Furthermore, we illustrate the method using a gene expression data set from a breast cancer study.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0275

Key words and phrases: Asymptotic normality, high-dimensional covariance matrix, homogeneity test, multi-sample comparison, power enhancement.

2. Introduction

Covariance matrices play a fundamental role in multivariate statistical inferences. In various fields, including economics and biology, many modern statistical procedures require testing the equality of covariance matrices. Here, typical examples include the multivariate analysis of variance or Fisher's linear discriminant analysis. In conventional low-dimensional settings where the dimension of the variables is relatively small compared with the sample size, tests for the equality of covariance matrices have been studied extensively; for example, see Sugiura and Nagao (1968), Gupta and Giri (1973), Gupta and Tang (1984), and O'Brien (1992) for two populations; and Perlman (1980) and Anderson (2003) for multiple populations.

As a result of the rapid development of science and technology, the collection and storage of large amounts of data are becoming increasingly common. Often, this results in high dimensions for the observations when the number of variables is large relative to the sample size. Conventional methods for testing the equality of covariance matrices usually fail in high-dimensional settings, because the sample covariance matrix does not

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0275

converge to its population counterpart in such situations. For inferences on high-dimensional covariance matrices, extensive research has been conducted on analyzing the limiting distributions of the extreme eigenvalues of sample covariance matrix (Bai, 1993; Johnstone, 2001; EI Karoui, 2007; Johnstone and Lu, 2009; Bai and Silverstein, 2010), estimations of high-dimensional population covariance matrices (Bickel and Levina, 2008a,b; Fan, Fan and Lv, 2008; Rothman, Levina and Zhu, 2010; Cai and Ma, 2013), and one-sample tests for high-dimensional matrices (Bai et al., 2009; Chen, Zhang and Zhong, 2010; Jiang and Yang, 2013; Srivastava, Yanagihara and Kubokawa, 2014). However, few statistical methods have been proposed for testing two or more high-dimensional covariance matrices (Cai, 2017). Bai et al. (2009) and Jiang and Yang (2013) use likelihood ratio statistics to test the equality of two population covariance matrices when the dimension is smaller than the sample size. However, a likelihood ratio statistic cannot be defined when the dimension is greater than the sample size. For situations characterized by a "large p" and a "small n" where p is the dimension of data and n is the sample size, Schott (2007) utilized the trace $\mathrm{tr}(\mathbf{S}_1-\mathbf{S}_2)^2$ to quantify the difference between two matrices, where $\mathbf{S}_1$ and $\mathbf{S}_2$ are sample covariance matrices of the two groups being compared. Srivastava and Yanagihara (2010) proposed a test statistic based on a dis-

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0275

tance measure, $\mathrm{tr}\mathbf{S}_1^2/(\mathrm{tr}\mathbf{S}_1)^2 - \mathrm{tr}\mathbf{S}_2^2/(\mathrm{tr}\mathbf{S}_2)^2$. However, the theoretical results of these two methods are derived under high-dimensional Gaussian distributions, and thus cannot be applied to general populations. To accommodate both Gaussian and non-Gaussian populations, Li and Chen (2012) proposed a U-statistic and Cai, Liu and Xia (2013) introduced an extreme statistic for two samples. Although both tests are powerful and robust with respect to the population distributions, they have several limitations. For example, both approaches apply to two populations only, and thus are not valid for multiple populations (i.e., more than two populations). Furthermore, the method of Li and Chen (2012) focuses on nonsparse dense alternatives, where many small disturbances may exist. In contrast, the method of Cai, Liu and Xia (2013) focuses on sparse alternatives, that is, when the number of nonzero elements of the difference between the two covariance matrices is small. As a result, the test of Li and Chen (2012) may result in unsatisfactory performance under the sparse alternative, and that of Cai, Liu and Xia (2013) may not work well under the nonsparse dense alternative. This is because these two test procedures use only one type of norm to characterize the distance between the two sample covariance matrices: the former utilizes the Frobenius norm, and the latter uses the maximum norm. Yang and Pan (2017) proposed a weighted test statistic that is suitable for

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0275

both sparse and dense alternatives and is based on random-matrix theory. Their approach involves complicated two-dimensional contour integrals that usually do not have explicit expressions, making the method difficult to implement in practice. In addition, studies may require comparisons of more than two groups. However, few studies have investigated tests for (more than two) high-dimensional matrices. Schott (2007) and Srivastava and Yanagihara (2010) addressed the problem of comparing multiple high-dimensional covariance matrices. However, as mentioned earlier, such tests are strictly bound by the Gaussian population assumption.

We develop a new method for testing the homogeneity of several high-dimensional covariance matrices using a weighted statistic of the pairwise test statistics for testing two covariance matrices. The contributions of our method are as follows:

(i) Our test can be used for both two-sample and multisample comparisons. For testing the homogeneity of several high-dimensional covariance matrices, existing methods (Schott, 2007; Srivastava and Yanagihara, 2010) focus on Gaussian cases and, thus, may not work well in non-Gaussian situations. In contrast, our test statistic does not require distributional assumptions and demonstrates a substantial improvement over existing tests involving multiple groups. Deriving the

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0275

theoretical properties of the weighted statistic is not a trivial extension of the case of two populations. In particular, the mutual dependence between pairwise components poses major theoretical challenges.

(ii) Our test is suitable for both sparse and nonsparse alternatives. In practice, the structure of the difference between two covariance matrices is typically unknown. This poses problems when implementing existing methods, because they utilize only one type of norm to characterize the discrepancy between two samples, which is often inadequate. In contrast, our test statistic is composed of two terms: the main term quantifies the Frobenius norm, and thus captures the difference between two covariance matrices in a nonsparse setting; the second term utilizes a screening technique with a maximum norm to enhance the power under sparse alternatives. By combining these two terms, the proposed test can be used to test both sparse and nonsparse alternatives, or the mixture of the two. Our approach is much easier to implement than that of Yang and Pan (2017). We conduct extensive simulation studies to show that our test has comparable performance with that of existing methods for two-sample populations. In certain cases, the proposed test achieves significantly higher power.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0275

(iii) For the asymptotic normality of the L_2-norm statistic, our test requires conditions that are more relaxed than those of existing methods. For example, our test assumes only that the fourth moment of the samples exists, whereas many existing methods, such as that of Li and Chen (2012), are established upon the existence of the eighth moment. In addition, out test needs conditions on the covariance matrices that are less regularized than those of some existing tests. For example, the maximum eigenvalue of the covariance matrices can be unbounded for the proposed method.

The remainder of the paper proceeds as follows. Section 3 presents the test for the equality of multiple high-dimensional population covariance matrices. The asymptotic null and alternative distributions of the proposed test statistic are derived based on random-matrix theory. The theoretical power of the proposed test is also examined here. Section 4 presents simulation results to demonstrate the superiority of the proposed method when testing the homogeneity of two or more covariance matrices. Section 5 analyzes a real data set as an illustration of the proposed method. Section 6 concludes the paper. All technical details are presented in the Supplementary Material.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0275

3. Testing the homogeneity of multiple covariance matrices

3.1 The test statistic

For K groups, let $\{\mathbf{x}_{k1}, \ldots, \mathbf{x}_{kn_k}\}$ be independent samples from the kth p-dimensional population with a mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$, for $k = 1, \ldots, K$, where n_k is the sample size and $\mathbf{x}_{ki} = (x_{k1i}, \ldots, x_{kpi})^T$, with the super-index T as the transpose. We are interested in testing the equality of the population covariance matrices of these K groups:

$$H_{0K} : \boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_K = \boldsymbol{\Sigma} \; vs. \; H_{AK} : \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K \text{ are not all equal,} \quad (3.1)$$

where $\boldsymbol{\Sigma}$ is unknown and the subscript K in H_{0K} and H_{AK} indicates that K populations are being compared. The sample covariance matrix of $\boldsymbol{\Sigma}_k$ is given by

$$\mathbf{S}_k = (n_k - 1)^{-1} \sum_{i=1}^{n_k} (\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)^T, \quad k = 1, \ldots, K, \quad (3.2)$$

where $\bar{\mathbf{x}}_k = (\bar{x}_{k1}, \ldots, \bar{x}_{kp})^T = n_k^{-1} \sum_{i=1}^{n_k} \mathbf{x}_{ki}$ is the sample mean of the kth population.

In the existing literature, many tests for (3.1) focus on the case of $K = 2$, and are often established based on two types of norms of $\mathbf{S}_{k_1} - \mathbf{S}_{k_2}$, for $1 \leq k_1, k_2 \leq K$. For example, Li and Chen (2012) used the statistic $\text{tr}(\mathbf{S}_{k_1} - \mathbf{S}_{k_2})^2$, and Cai, Liu and Xia (2013) considered the maximum

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0275

statistic $\max\{\delta_{k_1k_2\ell_1\ell_2}, \ell_1, \ell_2 = 1, \ldots, p\}$, where

$$\delta_{k_1k_2\ell_1\ell_2} = \frac{(s_{k_1\ell_1\ell_2} - s_{k_2\ell_1\ell_2})^2}{n_{k_1}^{-1}\hat{\theta}_{k_1\ell_1\ell_2} + n_{k_2}^{-1}\hat{\theta}_{k_2\ell_1\ell_2}}, \tag{3.3}$$

with $\hat{\theta}_{k\ell_1\ell_2} = n_k^{-1}\sum_{i=1}^{n_k}\{(x_{k\ell_1 i} - \bar{x}_{k\ell_1})(x_{k\ell_2 i} - \bar{x}_{k\ell_2}) - s_{k\ell_1\ell_2}\}^2$ and $s_{k\ell_1\ell_2}$ being the (ℓ_1, ℓ_2)th entry of $\mathbf{S}_k$ for $\ell_1, \ell_2 = 1, \ldots, p$, and $k = k_1, k_2$. These two test statistics have advantages and disadvantages. For example, the first trace-based statistic can capture many small differences. Furthermore, it possesses high power when testing dense $\mathbf{\Sigma}_{k_1} - \mathbf{\Sigma}_{k_2}$, but usually incurs some power loss for sparse $\mathbf{\Sigma}_{k_1} - \mathbf{\Sigma}_{k_2}$. The second statistic is able to detect large disturbances when $\mathbf{\Sigma}_{k_1} - \mathbf{\Sigma}_{k_2}$ is sparse, but usually fails to achieve high power when testing dense alternatives. If we know *a priori* that $\mathbf{\Sigma}_{k_1} - \mathbf{\Sigma}_{k_2}$ possesses a dense or sparse structure, the test procedures of Li and Chen (2012) and Cai, Liu and Xia (2013) can be adapted to suit the respective targeted alternatives. However, in real applications, the structure of $\mathbf{\Sigma}_{k_1} - \mathbf{\Sigma}_{k_2}$ is typically unknown and may have a more complicated structure, such as a mixture of dense and sparse signals. As a result, we require a test statistic that attains desirable power under both dense and sparse cases. Using only one norm is inadequate for this purpose. A natural approach is to take a linear combination of the two aforementioned statistics, as in Yang and Pan (2017). However, this results in a complex limiting distribution owing to the correlation between the two statistics. In addition, few

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0275

methods exist for testing (3.1) with $K \geq 3$, and extensions to the methods of Li and Chen (2012) and Cai, Liu and Xia (2013) to multiple samples are nontrivial.

We are particularly interested in applications involving several covariance matrices, for which current methods do not work well. To incorporate the strength from the trace-based and maximum norms, we propose a new statistic, $T_K = T_{K1} + T_{K2}$, with

$$\begin{aligned} T_{K1} &= \sum_{1\leq k_1<k_2\leq K} \omega_{k_1k_2}\mathrm{tr}[(\mathbf{S}_{k_1}-\mathbf{S}_{k_2})^2], \\ T_{K2} &= K_0 \max_{1\leq k_1<k_2\leq K}[I\{\max_{1\leq \ell_1\leq \ell_2\leq p}\delta_{k_1k_2\ell_1\ell_2} > s(n_{k_1},n_{k_2},p)\}]. \end{aligned}$$

Here, $\{\omega_{k_1k_2}, 1 \leq k_1 \leq k_2 \leq K\}$ are the prespecified weights, where $\omega_{k_1k_2} \geq 0$ and $\sum_{1\leq k_1<k_2\leq K}\omega_{k_1k_2} = 1$, K_0 is a large positive number, $I\{\cdot\}$ is an indicator function, and $s(n_{k_1}, n_{k_2}, p)$ is a prespecified threshold that depends on the sample sizes n_{k_1} and n_{k_2} and the dimension p. When $K = 2$, the proposed procedure reduces to the homogeneity test for two covariance matrices, with the test statistic given by $T_2 = T_{21} + T_{22}$, where $T_{22} = K_0 I\{\max_{1\leq \ell_1\leq \ell_2\leq p}\delta_{12\ell_1\ell_2} > s(n_1, n_2, p)\}$ and $T_{21} = \mathrm{tr}[(\mathbf{S}_1 - \mathbf{S}_2)^2]$. As a result, the proposed statistic T_K can be treated as a weighted average of the statistics T_2 for all paired populations. However, it is challenging to establish the limiting distributions of this weighted statistic because its components are not independent. Using random-matrix theory, we derive

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0275

the limiting distributions of T_K under both the null hypothesis H_{0K} and the alternative hypothesis H_{AK}.

The proposed statistic is similar in spirit to the power-enhancement test statistic proposed by Fan, Liao and Yao (2015). The first term T_{K1} plays a dominant role when testing dense cases. With a properly chosen threshold $s(n_{k_1}, n_{k_2}, p)$, the second term T_{K2} serves a screening purpose, converging to zero under the null hypothesis, and converging to a large number if $\delta_{k_1k_2\ell_1\ell_2}$ exceeds the threshold $s(n_{k_1}, n_{k_2}, p)$. As a result, the proposed statistic T_K tends to become very large quite quickly, as long as sparse disturbances are detected by T_{K2} if K_0 is sufficiently large; For further discussion on the choice of K_0, see Fan, Liao and Yao (2015).

3.2 Limiting distributions

We first impose two assumptions commonly used in random-matrix theory.

(A1) The vector $\mathbf{x}_{ki}$ satisfies the independent component structure $\mathbf{x}_{ki} = \boldsymbol{\mu}_k + \boldsymbol{\Gamma}_k \mathbf{w}_{ki}$, where $\mathbf{w}_{ki} = (w_{k1i}, \ldots, w_{kpi})^T$ and the elements $\{w_{k\ell i}, k = 1, \ldots, K; \ell = 1, \ldots, p; i = 1, \ldots, n_k\}$ are independent, with $\mathrm{E}w_{k\ell i} = 0$, $\mathrm{E}(w_{k\ell i}^2) = 1$, and $\beta_k = \mathrm{E}(w_{k\ell i}^4) - 3$. Moreover, for each $k = 1, \ldots, K$, the maximum eigenvalue of $\boldsymbol{\Sigma}_k$ is bounded or $\mathrm{tr}(\boldsymbol{\Sigma}_k^q) = O(p^q)$ for $q = 1, 2, 3, 4$.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0275

(A2) The asymptotic regime is satisfied; that is, $p/n_k \to c_k \in (0, \infty)$.

Assumption (A1) requires that the populations have an independent component structure. The population fourth moment of $w_{k\ell i}$ is required to exist, but no other distributional assumptions are imposed. The dimension and the sample size are assumed to tend to infinity proportionally under Assumption (A2). These are regular assumptions in deriving the asymptotic distributions of high-dimensional statistics; for instance, see Bai and Silverstein (2004), Bai and Silverstein (2010), and Li and Chen (2012). The limiting null distributions of T_{K1} and T_K are established as follows.

Theorem 1. *Under H_{0K} and Assumptions (A1)–(A2), for multisample comparisons with $k = 1, \ldots, K$, we have*

$$\sigma_K^{-1}(T_{K1} - \hat{\mu}_{K1} - \mu_K) \xrightarrow{d} N(0,1), \quad \hat{\sigma}_K^{-1}(T_{K1} - \hat{\mu}_{K1} - \hat{\mu}_K) \xrightarrow{d} N(0,1).$$

Furthermore, if the threshold $s(n_{k_1}, n_{k_2}, p)$ satisfies $s(n_{k_1}, n_{k_2}, p) - 4\log p \geq 0$ for any $1 \leq k_1 < k_2 \leq K$, and the conditions (C1), (C2), and (C3) in Cai, Liu and Xia (2013) are satisfied, then*

$$\sigma_K^{-1}(T_K - \hat{\mu}_{K1} - \mu_K) \xrightarrow{d} N(0,1), \quad \hat{\sigma}_K^{-1}(T_K - \hat{\mu}_{K1} - \hat{\mu}_K) \xrightarrow{d} N(0,1),$$

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0275

where

$$
\begin{aligned}
\hat{\mu}_{K1} &= \sum_{1\le k_1<k_2\le K} \omega_{k_1k_2}\Big[\sum_{k=k_1,k_2}(n_k^2-n_k-1)n_k^{-1}(n_k-1)^{-2}(\mathrm{tr}\mathbf{S}_k)^2\Big],\\
\mu_K &= \sum_{1\le k_1<k_2\le K} \omega_{k_1,k_2}\Big\{\sum_{k=k_1,k_2}\Big[(n_k+1)(n_k-1)^{-2}\mathrm{tr}\boldsymbol{\Sigma}^2\\
&\qquad +\beta_k n_k(n_k-1)^{-2}\sum_{\ell=1}^{p}(\mathbf{e}_\ell^T\boldsymbol{\Sigma}\mathbf{e}_\ell)^2\Big]\Big\},\\
\sigma_K^2 &= 4\sum_{1\le k_1<k_2\le K}\omega_{k_1k_2}^2[(n_{k_1}-1)^{-1}+(n_{k_2}-1)^{-1}]^2[\mathrm{tr}(\boldsymbol{\Sigma}^2)]^2\\
&\quad +8\sum_{1\le k_1<k_2<k_3\le K}\omega_{k_1k_2}\omega_{k_2k_3}(n_{k_2}-1)^{-2}[\mathrm{tr}(\boldsymbol{\Sigma}^2)]^2\\
&\quad +8\sum_{1\le k_1<k_3<k_2\le K}\omega_{k_1k_2}\omega_{k_3k_2}(n_{k_2}-1)^{-2}[\mathrm{tr}(\boldsymbol{\Sigma}^2)]^2\\
&\quad +8\sum_{1\le k_2<k_1<k_3\le K}\omega_{k_2k_1}\omega_{k_2k_3}(n_{k_2}-1)^{-2}[\mathrm{tr}(\boldsymbol{\Sigma}^2)]^2,\\
\hat{\mu}_K &= \sum_{1\le k_1<k_2\le K}\omega_{k_1k_2}\sum_{k=k_1,k_2}\Big\{(n_k-2)^{-2}\sum_{i=1}^{n_k}[(\mathbf{x}_{ki}-\bar{\mathbf{x}}_k)^T(\mathbf{x}_{ki}-\bar{\mathbf{x}}_k)-\mathrm{tr}\mathbf{S}_k]^2\\
&\qquad -n_k(n_k+2)^{-2}[\mathrm{tr}(\mathbf{S}_k^2)-(n_k-2)^{-1}(\mathrm{tr}\mathbf{S}_k)^2]\Big\},
\end{aligned}
$$

and $\hat{\sigma}_K^2$ *is obtained by replacing* $\mathrm{tr}(\boldsymbol{\Sigma}^2)$ *by* $\mathrm{tr}(\mathbf{S}^2)-(n_1+\cdots+n_K-K)^{-1}(\mathrm{tr}\mathbf{S})^2$ *in* σ_K^2 *with* $\mathbf{S}=(n_1+\cdots+n_K-K)^{-1}\sum_{k=1}^K(n_k-1)\mathbf{S}_k$.

Theorem 1 establishes the asymptotic normality of T_K under the null hypothesis. The detailed proof is given in the Supplementary Material.

Remark 1. In deriving the central limit theorem for Theorem 1, we obtain that the variance term in the limiting distribution of $\mathrm{tr}[(\mathbf{S}_{k_1}-\mathbf{S}_{k_2})^2]$ under H_{0K} is $4[\mathrm{tr}(\boldsymbol{\Sigma}^2)]^2[(n_{k_1}-1)^{-1}+(n_{k_2}-1)^{-1}]^2$. Then, a reasonable weight is

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0275

given by

$$\omega_{k_1k_2} = \frac{[(n_{k_1}-1)^{-1} + (n_{k_2}-1)^{-1}]^{-1}}{\sum_{1\le i<j\le K}[(n_i-1)^{-1} + (n_j-1)^{-1}]^{-1}}, \quad 1 \le k_1 < k_2 \le K,$$

which shows that the weight $\omega_{k_1k_2}$ is large when the variance of $\mathrm{tr}[(\mathbf{S}_{k_1} - \mathbf{S}_{k_2})^2]$ is small for $1 \le k_1 < k_2 \le K$.

As a special case, when $K = 2$, the proposed test statistic T_2 is able to test the homogeneity of two high-dimensional covariance matrices. Its asymptotic null distribution follows immediately.

Proposition 1. *Under the conditions of Theorem 1, for the two-sample case with $k = 1, 2$, we have*

$$\hat{\sigma}_2^{-1}(T_{21} - \hat{\mu}_{21} - \hat{\mu}_2) \xrightarrow{d} N(0, 1).$$

Furthermore, if the threshold $s(n_1, n_2, p)$ satisfies $s(n_1, n_2, p) - 4\log p \ge 0$ and the conditions (C1), (C2), and (C3) in Cai, Liu and Xia (2013) are satisfied, then*

$$\hat{\sigma}_2^{-1}(T_2 - \hat{\mu}_{21} - \hat{\mu}_2) \xrightarrow{d} N(0, 1),$$

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0275

where

$$\begin{aligned}
\hat{\mu}_{21} &= \sum_{k=1,2}(n_k^2 - n_k - 1)n_k^{-1}(n_k-1)^{-2}(\mathrm{tr}\mathbf{S}_k)^2, \\
\hat{\mu}_2 &= \sum_{k=1,2}(n_k-2)^{-2}\sum_{i=1}^{n_k}[(\mathbf{x}_{ki}-\bar{\mathbf{x}}_k)^T(\mathbf{x}_{ki}-\bar{\mathbf{x}}_k) - \mathrm{tr}\mathbf{S}_k]^2 \\
&\quad - \sum_{k=1,2} n_k(n_k+2)^{-2}[\mathrm{tr}(\mathbf{S}_k^2) - (n_k-2)^{-1}(\mathrm{tr}\mathbf{S}_k)^2], \\
\hat{\sigma}_2^2 &= 4[(n_1-1)^{-1} + (n_2-1)^{-1}]^2[\mathrm{tr}(\mathbf{S}^2) - (n_1+n_2-2)^{-1}(\mathrm{tr}\mathbf{S})^2]^2,
\end{aligned}$$

with $\mathbf{S} = [(n_1-1)\mathbf{S}_1 + (n_2-1)\mathbf{S}_2]/(n_1+n_2-2)$.

Remark 2. In practice, to apply the proposed test to two groups with $K = 2$, we need to specify the value of $s(n_1, n_2, p)$. There are many choices for the threshold, as long as under H_{02}, $P(\max_{1\le\ell_1\le\ell_2\le p}\delta_{12\ell_1\ell_2} \le s(n_1,n_2,p))$ converges to one as $n_1, n_2 \to \infty$. For simplicity, the threshold is set as

$$s(n_1,n_2,p) = [\{\log\log(n_1/2+n_2/2)-1\}^2/4+1](4\log p - \log\log p) + q,$$

where $\exp\{-(8\pi)^{-1/2}\exp(-q/2)\} = 0.985$ and $\max_{1\le\ell_1\le\ell_2\le p}\delta_{12\ell_1\ell_2} - 4\log p + \log\log p$ converges to a type I extreme value distribution under the null hypothesis and some proper conditions (Cai, Liu and Xia, 2013). For multiple populations with $K \ge 3$, owing to multiple pairwise comparisons, we set the threshold as

$$s(n_{k_1},n_{k_2},p) = [\{\log\log(n_{k_1}/2+n_{k_2}/2)-1\}^2/4+1](4\log p - \log\log p) + q,$$

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0275

where $\exp\{-(8\pi)^{-1/2}\exp(-q/2)\} = 1-0.015/[K(K-1)/2]$, based on Bonferroni's correction, to control the inflation of the type I error rate. It is obvious that the thresholds $s(n_{k_1}, n_{k_2}, p)$ and $s(n_1, n_2, p)$ both satisfy the condition that $s(n_{k_1}, n_{k_2}, p) - 4\log p \geq 0$. The choice of K_0 is discussed extensively in Fan, Liao and Yao (2015). In general, K_0 should be large enough to reject the null once sparse signals are detected; here, we set $K_0 = p^2$.

3.3 Power comparison

According to Theorem 1, the acceptance region of the statistic T_K with respect to the nominal size α is

$$\{(\mathbf{x}_{k1}, \ldots, \mathbf{x}_{kn_k}, k = 1, \ldots, K) : T_K - \hat{\mu}_{K1} - \hat{\mu}_K \leq z_{1-\alpha}\hat{\sigma}_K\},$$

where $z_{1-\alpha}$ is the $(1-\alpha)$th quantile of the standard normal distribution $N(0,1)$. Therefore, the power function for testing (3.1) is

$$\begin{aligned} g_K(\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_K) &= P_{H_{AK}}(T_K - \hat{\mu}_{K1} - \hat{\mu}_K > z_{1-\alpha}\hat{\sigma}_K) \\ &\geq P_{H_{AK}}(T_{K1} - \hat{\mu}_{K1} - \hat{\mu}_K > z_{1-\alpha}\hat{\sigma}_K), \end{aligned}$$

because $T_K = T_{K1} + T_{K2}$ with $T_{K2} \geq 0$.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0275

To investigate the power of the proposed test, let

$$
\begin{aligned}
&\mu_{Ak_1k_2} = \mathrm{tr}[(\mathbf{\Sigma}_{k_1} - \mathbf{\Sigma}_{k_2})^2] \\
&+[(n_{k_1}+1)(n_{k_1}-1)^{-2}\mathrm{tr}(\mathbf{\Sigma}_{k_1}^2) + \beta_{k_1} n_{k_1}(n_{k_1}-1)^{-2}\sum_{\ell=1}^{p}(\mathbf{e}_\ell^T\mathbf{\Sigma}_{k_1}\mathbf{e}_\ell)^2] \\
&+[(n_{k_2}+1)(n_{k_2}-1)^{-2}\mathrm{tr}(\mathbf{\Sigma}_{k_2}^2) + \beta_{k_2} n_{k_2}(n_{k_2}-1)^{-2}\sum_{\ell=1}^{p}(\mathbf{e}_\ell^T\mathbf{\Sigma}_{k_2}\mathbf{e}_\ell)^2], \\
&\sigma^2_{Ak_1k_2} \\
= \;& 4[(n_{k_1}-1)^{-1}\mathrm{tr}(\mathbf{\Sigma}_{k_1}^2) + (n_{k_2}-1)^{-1}\mathrm{tr}(\mathbf{\Sigma}_{k_2}^2)]^2 \\
&+8(n_{k_1}-1)^{-1}(n_{k_2}-1)^{-1}\{[\mathrm{tr}(\mathbf{\Sigma}_{k_1}\mathbf{\Sigma}_{k_2})]^2 - \mathrm{tr}(\mathbf{\Sigma}_{k_1}^2)\mathrm{tr}(\mathbf{\Sigma}_{k_2}^2)\} \\
&+4(n_{k_1}-1)^{-1}\{2\mathrm{tr}[\mathbf{\Sigma}_{k_1}(\mathbf{\Sigma}_{k_1} - \mathbf{\Sigma}_{k_2})]^2 + \beta_{k_1}\sum_{\ell=1}^{p}[\mathbf{e}_\ell^T\mathbf{\Sigma}_{k_1}^{1/2}(\mathbf{\Sigma}_{k_1} - \mathbf{\Sigma}_{k_2})\mathbf{\Sigma}_{k_1}^{1/2}\mathbf{e}_\ell]^2\} \\
&+4(n_{k_2}-1)^{-1}\{2\mathrm{tr}[\mathbf{\Sigma}_{k_2}(\mathbf{\Sigma}_{k_1} - \mathbf{\Sigma}_{k_2})]^2 + \beta_{k_2}\sum_{\ell=1}^{p}[\mathbf{e}_\ell^T\mathbf{\Sigma}_{k_2}^{1/2}(\mathbf{\Sigma}_{k_1} - \mathbf{\Sigma}_{k_2})\mathbf{\Sigma}_{k_2}^{1/2}\mathbf{e}_\ell]^2\}, \\
&\sigma^2_{Ak_1k_2k_2k_3} \\
= \;& 4[n_{k_2}^{-1}\mathrm{tr}(\mathbf{\Sigma}_{k_2}^2)]^2 + 4(n_{k_2}-1)^{-1}[2\mathrm{tr}(\mathbf{\Sigma}_{k_2}^4) + \beta_{k_2}\sum_{\ell=1}^{p}(\mathbf{e}_\ell^T\mathbf{\Sigma}_{k_2}^2\mathbf{e}_\ell)^2] \\
&-4(n_{k_2}-1)^{-1}[2\mathrm{tr}(\mathbf{\Sigma}_{k_2}^3\mathbf{\Sigma}_{k_3}) + \beta_{k_2}\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{\Sigma}_{k_2}^{1/2}\mathbf{\Sigma}_{k_3}\mathbf{\Sigma}_{k_2}^{1/2}\mathbf{e}_\ell\mathbf{e}_\ell^T\mathbf{\Sigma}_{k_2}^2\mathbf{e}_\ell] \\
&-4(n_{k_2}-1)^{-1}[2\mathrm{tr}(\mathbf{\Sigma}_{k_2}^3\mathbf{\Sigma}_{k_1}) + \beta_{k_2}\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{\Sigma}_{k_2}^{1/2}\mathbf{\Sigma}_{k_1}\mathbf{\Sigma}_{k_2}^{1/2}\mathbf{e}_\ell\mathbf{e}_\ell^T\mathbf{\Sigma}_{k_2}^2\mathbf{e}_\ell] \\
&+4(n_{k_2}-1)^{-1}[2\mathrm{tr}(\mathbf{\Sigma}_{k_1}\mathbf{\Sigma}_{k_2}\mathbf{\Sigma}_{k_3}\mathbf{\Sigma}_{k_2}) \\
&\qquad\qquad +\beta_{k_2}\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{\Sigma}_{k_2}^{1/2}\mathbf{\Sigma}_{k_1}\mathbf{\Sigma}_{k_2}^{1/2}\mathbf{e}_\ell\mathbf{e}_\ell^T\mathbf{\Sigma}_{k_2}^{1/2}\mathbf{\Sigma}_{k_3}\mathbf{\Sigma}_{k_2}^{1/2}\mathbf{e}_\ell],
\end{aligned}
$$

for $1 \le k_1, k_2, k_3 \le K$. The limiting distributions of T_{K1} and T_K under the alternative hypothesis H_{AK} are given as follows.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0275

Theorem 2. *Under Assumptions (A1)–(A2) and letting* $\mathbf{A}_{k_1k_2} = \mathbf{\Sigma}_{k_1} - \mathbf{\Sigma}_{k_2}$ *for* $1 \le k_1 < k_2 \le K$*, for multisample comparisons with* K *groups, we have*

$$\sigma_{AK}^{-1}(T_{K1} - \hat{\mu}_{K1} - \mu_{AK}) \xrightarrow{d} N(0,1),$$

where $\mu_{AK} = \sum_{1\le k_1<k_2\le K} \omega_{k_1k_2}\mu_{Ak_1k_2}$ *and*

$$\begin{aligned}
&\sigma_{AK}^2 \\
=\ & \sum_{1\le k_1<k_2\le K} \omega_{k_1k_2}^2\sigma_{Ak_1k_2}^2 + 2\sum_{1\le k_1<k_2<k_3\le K} \omega_{k_1k_2}\omega_{k_2k_3}\sigma_{Ak_1k_2k_2k_3}^2 \\
& +2\sum_{1\le k_1<k_3<k_2\le K} \omega_{k_1k_2}\omega_{k_2k_3}\sigma_{Ak_1k_2k_2k_3}^2 + 2\sum_{1\le k_2<k_1<k_3\le K} \omega_{k_1k_2}\omega_{k_2k_3}\sigma_{Ak_1k_2k_2k_3}^2,
\end{aligned}$$

with weights $\{\omega_{k_1k_2}, 1 \le k_1, k_2 \le K\}$ *and* $\omega_{k_1k_2} = \omega_{k_2k_1}$*. Under the condition that* $\mathrm{E}|(x_{k\ell i} - \mathrm{E}x_{k\ell i})/\sqrt{\sigma_{k\ell\ell}}|^{8+\epsilon} < c_0$ *for some positive* c_0*, and* $\min_{1\le\ell_1\le\ell_2\le p} \theta_{k\ell 1\ell_2}(\sigma_{k\ell_1\ell_1}\sigma_{k\ell_2\ell_2})^{-1} \ge \tau_k$*, where* $\mathbf{\Sigma}_k = (\sigma_{k\ell_1\ell_2})_{\ell_1,\ell_2=1}^p$ *and* $\theta_{k\ell_1\ell_2} = \mathrm{Var}[(x_{k\ell_1 i} - \mathrm{E}x_{k\ell_1 i})(x_{k\ell_2 i} - \mathrm{E}x_{k\ell_2 i})]$*,* c_0*,* ϵ*, and* τ_k *are some positive constants for* $1 \le \ell, \ell_1, \ell_2 \le p$*,* $i = 1, \ldots, n_k$*, and* $k = 1, \ldots, K$*. Then we have*

$$T_{K2} - K_0 \xrightarrow{a.s.} 0$$

and

$$\sigma_{AK}^{-1}(T_K - K_0 - \hat{\mu}_{K1} - \mu_{AK}) \xrightarrow{d} N(0,1)$$

if there exists a pair of (k_1, k_2) *satisfying* $s(n_{k_1}, n_{k_2}, p) \le \max_{1\le i\le j\le p}[(\sigma_{k_1ij} - \sigma_{k_2ij})^2(\theta_{k_1ij}/n_{k_1} + \theta_{k_2ij}/n_{k_2})^{-1}]$.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0275

Based on the asymptotic normality of T_{K1} and T_K, as shown in Theorem 2, we obtain the following corollary on the power of the proposed test.

Corollary 1. *Under the conditions of Theorem 2, the following three results hold.*

(i) When $n_1, \ldots, n_K, p$ are sufficiently large, we have $g_K(\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_K) \geq \alpha$ with the nominal size α. In particular, when $\max\{\mathrm{tr}(\mathbf{A}_{ij}^2), 1 \leq i < j \leq K\} > c_1$ for a small positive constant c_1, we have $g_K(\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_K) > \alpha$.

(ii) When $\mathrm{tr}(\mathbf{A}_{ij}^2) \to \infty$ for some $1 \leq i < j \leq K$, we have $g_K(\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_K) \to 1$ as $n_1, \ldots, n_K, p \to \infty$.

(iii) When there exists a pair of (k_1, k_2) satisfying $s(n_{k_1}, n_{k_2}, p) + 4\log p \leq 0.5 \max_{1 \leq i \leq j \leq p}[(\sigma_{k_1 ij} - \sigma_{k_2 ij})^2(\theta_{k_1 ij}/n_{k_1} + \theta_{k_2 ij}/n_{k_2})^{-1}]$, we have

$$g_K(\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_K) \to 1$$

as $n_1, \ldots, n_K, p \to \infty$.

Corollary 1 shows that the proposed test T_K is asymptotically unbiased. As long as there exists a pair i, j with $\mathrm{tr}(\mathbf{A}_{ij}^2) > c_1$, the power function is greater than the nominal size. In addition, if $\mathrm{tr}(\mathbf{A}_{ij}^2) \to \infty$, the power function tends to one. Theorem 2 and Corollary 1 with $K = 2$ facilitate a

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0275

power comparison between the proposed tests T_2 and T_{21} and those of Li and Chen (2012) and Cai, Liu and Xia (2013). In particular, we define the power function of the statistic T_{21} as

$$g_{21}(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) = P_{H_{A2}}(T_{21} - \hat{\mu}_{21} > \hat{\mu}_2 + z_{1-\alpha}\hat{\sigma}_2),$$

and denote those of Li and Chen (2012) and Cai, Liu and Xia (2013) as $g_{\mathrm{LC}}(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2)$ and $g_{\mathrm{CLX}}(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2)$, respectively.

First, because $T_2 = T_{21} + T_{22}$ with $T_{22} \geq 0$, we have $g_2(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) \geq g_{21}(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2)$. In other words, the power of T_2 is greater than or equal to that of T_{21} owing to the positivity of T_{22}. Typically, when $\mathrm{tr}(\mathbf{A}_{12}^2) \to 0$, but there is at least one entry (ℓ_1, ℓ_2) of $\mathbf{A}_{12}$ greater than $4\sqrt{(\theta_{1\ell_1\ell_2}/n_1 + \theta_{2\ell_1\ell_2}/n_2)\log p}$, we have $P(\max_{1\leq\ell_1\leq\ell_2\leq p}\delta_{12\ell_1\ell_2} > s(n_1, n_2, p)) \to 1$ under the (C2*) condition in Cai, Liu and Xia (2013), which leads to $T_{22} \to K_0$, almost surely. As a result, the power functions of $g_{\mathrm{CLX}}(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2)$ and $g_2(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2)$ converge to 1 as $n_1, n_2, p \to \infty$ if K_0 is sufficiently large. On the other hand, in this situation, $g_{21}(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) \to \alpha$ and $g_{\mathrm{LC}}(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) \to \alpha$ (Cai, Liu and Xia, 2013), demonstrating that the screening term T_{22} enhances the power of T_{21}. This property of the proposed test T_2 is confirmed by the simulation results of Scenario 3 in the simulation study.

Second, if $\mathrm{tr}\mathbf{A}_{12}^2 \to \infty$, none of the absolute entries of $\mathbf{A}_{12}$ are greater than $[\min\{n_1, n_2\}]^{-(1+\epsilon)}\log p$ with $\epsilon > 0$, and $\theta_{k\ell_1\ell_2}$ has uniform positive

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0275

lower and upper bounds, then we have

$$g_{\mathrm{CLX}}(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) \to \alpha, \quad g_{\mathrm{LC}}(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) \to 1, \quad g_2(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) \to 1, \quad g_{21}(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) \to 1,$$

as $T_{22} \to 0$. That is, when all entries of $\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2$ are small nonzeros, the power of the test in Cai, Liu and Xia (2013) may be relatively small. However, our test T_2 and that in Li and Chen (2012) can discriminate between $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ with high power, which corresponds to Scenarios 1 and 2 in the simulation study.

Third, in some situations, when $\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2$ is composed of a mixture of dense and sparse signals, the terms T_{21} and T_{22} both contribute to detecting disturbances between $\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2$. As a result, the proposed method may deliver higher power than those of Li and Chen (2012) and Cai, Liu and Xia (2013). We examine this situation in Scenario 4 in the simulation studies.

4. Simulation studies

We evaluate the finite-sample performance of the proposed test with two populations and three populations using extensive simulation studies. To test the equality of two population covariance matrices, we compare our test T_2 with five existing methods: Yang and Pan (2017) (YP), Li and Chen (2012) (LC), Cai, Liu and Xia (2013) (CLX), Schott (2007) (SC),

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0275

and SY (Srivastava and Yanagihara, 2010). For the three-sample covariance matrix testing problems, we consider three methods: T_3, SC, and SY. The sample sizes are taken as $n_k = 60, 100, 200, 300$ for $k = 1, 2, 3$, and the dimension p is 100 or 300. The observations are drawn from $\mathbf{x}_{ki} = \mathbf{\Gamma}_k \mathbf{w}_{ki}$, where $\{\mathbf{w}_{k\ell i}, i = 1, \ldots, n_k, \ell = 1, \ldots, p, k = 1, 2, 3\}$ are independent and identically distributed (i.i.d.) from the standard normal (Gaussian) distribution $N(0, 1)$ or the shifted gamma distribution gamma$(4, 2) - 2$. The nominal test size is 5%, a nd we conduct 5000 replications to summarize the empirical proportion of rejecting the null hypothesis under each case. Four scenarios are considered for the comparison.

*Scenario 1.*Let $\mathbf{\Sigma}_1 = \mathbf{I}_p$ and $\mathbf{\Sigma}_k = \mathbf{\Gamma}_k \mathbf{\Gamma}_k^T$, where $\mathbf{\Gamma}_k = \mathbf{I}_p + \theta_k \left(u_{kij}\right)_{i,j=1}^{p}$ for $k = 2, \ldots, K$, with $\mathbf{I}_p$ being the $p \times p$ identity matrix. We consider $\{u_{2ij}, i, j = 1, \ldots, p\} \overset{\text{i.i.d.}}{\sim} \text{Unif}(-n_1^{-0.75}, n_1^{0.75})$ and $\{u_{3ij}, i, j = 1, \ldots, p\} \overset{\text{i.i.d.}}{\sim} \text{Unif}(-n_1^{-0.9}, n_1^{0.9})$. To test $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ with $K = 2$, we evaluate the empirical test size using $\theta_1 = \theta_2 = 0$ and the empirical power using $(\theta_1, \theta_2) = (0, 1)$. To test the equality of three covariance matrices, $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}_3$ with $K = 3$, we evaluate the empirical test size using $\theta_1 = \theta_2 = \theta_3 = 0$ and the empirical power using $(\theta_1, \theta_2, \theta_3) = (0, 1, 1)$. This configuration enables us to examine the performance of the proposed test under the dense alternative.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0275

Scenario 2. The observation $\mathbf{x}_{ki} = (x_{k1i}, \ldots, x_{kpi})^T$ is generated from $x_{kji} = w_{kji} + 2w_{k,j+1,i} + \theta_k w_{k,j+2,i}$ for $k = 1, \ldots, K$ (Li and Chen, 2012). To test $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ with $K = 2$, we evaluate the empirical test size using $\theta_1 = \theta_2 = 0$ and the empirical power using $(\theta_1, \theta_2) = (0, 0.6)$. To test the equality of three covariance matrices, $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}_3$ with $K = 3$, we evaluate the empirical test size using $\theta_1 = \theta_2 = \theta_3 = 0$ and the empirical power using $(\theta_1, \theta_2, \theta_3) = (0, 0.4, 0.6)$. Scenario 2 corresponds to the case with a relatively sparse alternative.

Scenario 3. The covariance matrix is $\mathbf{\Sigma}_k = \mathbf{C} + \delta_0 \mathbf{I}_p + \theta_k \mathbf{U}_k$ and $\mathbf{\Gamma}_k = \mathbf{\Sigma}_k^{1/2}$ for $k = 1, \ldots, K$, where $\mathbf{C} = (0.2^{I\{|i-j|>0\}})_{i,j=1}^p$. In addition, $\mathbf{U}_k$ is a $p \times p$ symmetric matrix with four nonzero entries from Unif$(0, 2)$ randomly located in the upper triangle, and another four located in the lower triangle, by symmetry (Cai, Liu and Xia, 2013). As a result, the differences between $\mathbf{\Sigma}_k$ are extremely sparse. To test $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ with $K = 2$, $\delta_0 = |\min\{\lambda_{\min}(\mathbf{C} + \mathbf{U}_2), \lambda_{\min}(\mathbf{C})\}| + 0.05$, and we evaluate the empirical test size using $\theta_1 = \theta_2 = 0$ and the empirical power using $(\theta_1, \theta_2) = (0, 1)$. To test $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}_3$ with $K = 3$, $\delta_0 = |\min\{\lambda_{\min}(\mathbf{C} + \mathbf{U}_2), \lambda_{\min}(\mathbf{C} + \mathbf{U}_3), \lambda_{\min}(\mathbf{C})\}| + 0.05$, and we evaluate the empirical test size using $\theta_1 = \theta_2 = \theta_3 = 0$ and the empirical power using $(\theta_1, \theta_2, \theta_3) = (0, 1, 1)$.

Scenario 4. The covariance matrix is $\mathbf{\Sigma}_k = (\sigma_{kij})_{i,j=1}^p$ and $\mathbf{\Gamma}_k = \mathbf{\Sigma}_k^{1/2}$

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0275

with $\sigma_{kij} = \theta_k[I\{k=2\}(2\log p/3)\mathbf{E}_{11}+I\{k=3\}(\log p/2)\mathbf{E}_{22}+I\{k=2\}u_{kij}]+(0.1^{|i-j|}+0.2^{|i-j|})/2$, where $\mathbf{E}_{ij}$ is the matrix with the (i,j)th entry equal to one, and the rest equal to zero. In addition, $u_{kij} \sim \text{Unif}(-n^{-0.8}, n^{-0.8})$, with $n = \sum_{k=1}^{K} n_k$ for $k = 1, \ldots, K$. To test $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ with $K = 2$, we evaluate the empirical test size with $\theta_1 = \theta_2 = 0$ and the empirical power using $(\theta_1, \theta_2) = (0, 1)$. To test $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}_3$ with $K = 3$, we evaluate the empirical test size using $\theta_1 = \theta_2 = \theta_3 = 0$ and the empirical power using $(\theta_1, \theta_2, \theta_3) = (0, 1, 1)$. Scenario 4 investigates the tests based on a mixture of sparse and dense alternatives.

The simulation results for the two-sample testing problems are summarized in Tables S1–S4 of the Supplementary Material. The results for three samples with Gaussian populations are presented in Figure S1 of the Supplementary Material, and those with gamma populations are provided in Figure S2 of the Supplementary Material. To test the equality of two covariance matrices, our test, LC, and CLX maintain empirical test sizes well for both the Gaussian and the gamma populations. In contrast, YP and SC work for the Gaussian population only. The performance of SY under the two-sample cases suffers from size distortion in some scenarios, especially when p is close to or larger than n. As a result, we do not present the results for SC under the gamma populations or those for SY under the

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0275

two-sample cases. For the alternatives, we compare the empirical power of our test, LC, and CLX. In Scenarios 1 and 2 (i.e., dense and relatively sparse cases, respectively), our test is as powerful as the LC method, and produces higher power than CLX does. In Scenario 3, for extremely sparse alternatives, the empirical power of the proposed method is slightly lower than that of CLX, but much higher than that of LC. In Scenario 4, where both large and small disturbances exist between the two population covariance matrices, our test outperforms the other methods. In addition, we consider the ultra high-dimensional setting with $p = 500$ and 1000. The simulation results of the new test are presented in Table S5 of the Supplementary Material. We conclude that our T_2 test still performs well when p is much larger than n. Moreover, if we change the threshold to $s(n_1, n_2, p) = [\{\log\log(n_1/2 + n_2/2) - 1\}^2 + 4](\log p - \log\log p/4) + q$, with $\exp\{-(8\pi)^{-1/2}\exp(-q/2)\} = 0.99$, the performance of T_2 is similar.

To test the homogeneity of three covariance matrices, we examine the performance of the statistic T_{31}, which is the first term of the proposed T_3. Our goal is to investigate the gain form the screening term T_{32}, especially in the sparse cases. The simulation results are similar to those in the two-sample cases. Moreover, the T_3 test demonstrates substantial advantages over the T_{31} test when sparse, but large disturbances exist under

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0275

the alternative hypothesis, and the empirical sizes for these two tests are comparable.

5. Real-data analysis

To illustrate of our proposed test, we present an analysis of the gene expression data set from the breast cancer study by Schmidt et al. (2008). The data are available from "Bioconductor", and include gene expression patterns of 200 tumors of patients who were not treated by systemic therapy after surgery. Patients were classified into three groups according to the tumor grade: group 1, with a well-differentiated tumor ($n_1 = 29$); group 2, with a moderately differentiated tumor ($n_2 = 137$); and group 3, with a poor differentiated/undifferentiated tumor ($n_3 = 35$). The heterogeneous nature of breast cancer facilitates the development of prognostic and predictive classification algorithms based on the related genes, and the choice of the classification methods relies on whether the covariance matrices are homogeneous. Hence, we are interested in testing the homogeneity of the variance–covariance matrices of these three groups.

The breast cancer data set contains 22283 features, yielding a high-dimensional hypothesis testing problem. To alleviate the computational burden, we perform a feature-screening procedure (`http://bioconductor.`

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0275

org/packages/release/bioc/html/genefilter.html) in which we filter out the features with coefficients of variation that fall outside the range $(0.25, 1.0)$ and control at least 30% of the selected features that have intensities above five. After the preliminary screening procedure, 1280 features are kept for the analysis. Let $\mathbf{\Sigma}_1, \mathbf{\Sigma}_2$, and $\mathbf{\Sigma}_3$ be the covariance matrices of these 1280 features in patients with tumor grades of 1, 2, and 3, respectively. To visualize the selected data set, we plot the values of $s_{k_1\ell_1\ell_2} - s_{k_2\ell_1\ell_2}$ for the pairwise comparisons in Figure 1. It is observed that $\mathbf{\Sigma}_2 - \mathbf{\Sigma}_1$ has more elements concentrated around zero than $\mathbf{\Sigma}_3 - \mathbf{\Sigma}_2$ does. In addition, one large disturbance (around the index of 20000) may exist between groups 1 and 2, whereas many moderate disturbances are present in $\mathbf{\Sigma}_3 - \mathbf{\Sigma}_2$.

We first apply the T_2, LC, and CLX methods to test the null hypotheses $H_{02}^{(1,2)} : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ and $H_{02}^{(2,3)} : \mathbf{\Sigma}_2 = \mathbf{\Sigma}_3$, separately. The nominal size is set at 5%. Our T_2 method rejects both null hypotheses $H_{02}^{(1,2)}$ and $H_{02}^{(2,3)}$, whereas LC and CLX reject only one of the two. Specifically, LC fails to detect the difference between $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ because there is only one large disturbance (feature 206023_at) between the two covariance matrices. On the other hand, CLX cannot detect the many small disturbances between $\mathbf{\Sigma}_2$ and $\mathbf{\Sigma}_3$. This example demonstrates that the structures of the differences between the two covariance matrices indeed affect the performance of the

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0275

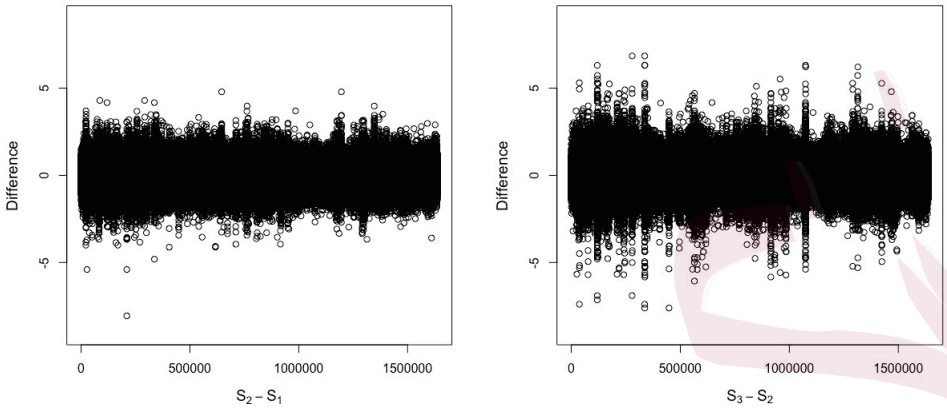

Figure 1: Plots of values of $s_{k_1\ell_1\ell_2} - s_{k_2\ell_1\ell_2}$ to quantify the difference between $\mathbf{\Sigma}_{k_1}$ and $\mathbf{\Sigma}_{k_2}$ for the breast cancer data set, for $\ell_1, \ell_2 = 1, \ldots, 1280$.

LC and CLX tests, because each is based on one type of norm statistic only. Without knowledge of the specific structure of the difference between the covariance matrices, the proposed test can identify both "few large" and "many tiny" disturbances, and thus leads to higher power in homogeneity testing problems.

Next, we consider $H_{03} : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}_3$ to test the equality of the covariance matrices of groups 1, 2, and 3. The observed test statistic based on the breast cancer data set is 116.9, with a p-value extremely close to zero; thus, we reject H_{03}. Moreover, the observed statistic of T_{31} is 5.3, with a p-value of 6.5×10^{-8}, indicating that both terms of the proposed test T_3

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0275

play a role in detecting the differences for dense and sparse alternatives.

6. Concluding remarks

We have proposed a new test for the homogeneity of multiple high-dimensional covariance matrices. In contrast to existing methods, which typically use only one type of norm statistic, our test statistic is composed of two different norms. The first detects a few strong signals, and the second detects many faint signals. By adaptively mixing the two norms, our test gains substantial power for different situations. More importantly, we do not need to know *a priori* which types of signals are present. The asymptotic properties of our tests are established using modern random-matrix theory, which demonstrates the elegance of the theoretical development.

The proposed statistic for testing several matrices is a weighted average of the pairwise test statistics, where the weight is proportional to the inverse of the sample size. A related question is to determine the optimal weight that maximizes the power. In fact, the power function can be represented

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0275

as

$$\begin{aligned}
& \Phi[(\mu_{AK} - \hat{\mu}_K - z_{1-\alpha}\hat{\sigma}_K)/\sigma_{AK}] \\
= \; & \Phi\left(\left\{\sum_{1\le k_1<k_1\le K} \omega_{k_1k_2}\mathrm{tr}[(\mathbf{\Sigma}_{k_1} - \mathbf{\Sigma}_{k_2})^2] - z_{1-\alpha}\hat{\sigma}_K\right\}/\sigma_{AK}\right) + o(1) \\
\ge \; & \Phi\left(\sum_{1\le k_1<k_1\le K} \omega_{k_1k_2}\mathrm{tr}[(\mathbf{\Sigma}_{k_1} - \mathbf{\Sigma}_{k_2})^2]/\sigma_{AK}\right) + o(1),
\end{aligned}$$

where $\Phi(\cdot)$ is the cumulative distribution function of $N(0,1)$, σ^2_{AK} is the quadratic form of $\{\sigma_{Ajj}, \sigma_{Aijjk}\}$, and $\mu_{AK} - \hat{\mu}_K = \sum_{1\le k_1<k_1\le K} \omega_{k_1k_2}\mathrm{tr}[(\mathbf{\Sigma}_{k_1} - \mathbf{\Sigma}_{k_2})^2] + o(1)$. As a result, the optimal weight $\{\omega_{k_1k_2}, 1 \le k_1 < k_2 \le K\}$ that maximizes the power function can be determined by solving the Markowitz portfolio problem (Markowitz, 1952), where $\sum_{1\le k_1<k_1\le K} \omega_{k_1k_2}\mathrm{tr}[(\mathbf{\Sigma}_{k_1} - \mathbf{\Sigma}_{k_2})^2]$ can be regarded as the return and σ^2_{AK} can be treated as the risk. In addition, an interesting extension of our method would be to the case of a large number of groups or covariance matrices, that is, K is also large.

Supplementary Material

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

Acknowledgments

We thank the associate editor, the referees, and the editor for their many constructive and insightful comments that have led to significant

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0275

improvements in the article. Zheng's research was supported by NSFC grant 11522105, Guo's research was supported by NSFC grants 11690012 and 11631003, and Yin's research was supported, in part, by a grant (grant number 17326316) from the Research Grants Council of Hong Kong.

References

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis* (3 ed.). John Wiley & Sons.

Bai, Z. D. (1993). Convergence rate of expected spectral distributions of large random matrices. II. Sample covariance matrices. *Ann. Probab. 21*, pp. 649–672.

Bai, Z. D., Jiang, D., Yao, J. F. and Zheng, S. (2009). Corrections to LRT on large dimensional covariance matrix by RMT. *Ann. Stat. 37*, pp. 3822–3840.

Bai, Z. D. and Silverstein, J. W. (2004). CLT for linear spectral statistics of large dimensional sample covariance matrices. *Ann. Probab. 32*, pp. 553–605.

Bai, Z. D. and Silverstein, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*. Beijing: Science Press.

Bickel, P. J. and Levina, E. (2008a). Regularized estimation of large covariance matrices. *Ann. Stat. 36*, pp. 199–227.

Bickel, P. J. and Levina, E. (2008b). Covariance regularization by thresholding. *Ann. Stat. 36*, pp. 2577–2604.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0275

Cai, T. T., Liu, W. D. and Xia, Y. (2013). Two-sample covariance matrix testing and support recovery in high dimensional and sparse settings. *J. Am. Stat. Assoc. 108*, pp. 265–277.

Cai, T. T. and Ma, Z. (2013). Optimal hypothesis testing for high dimensional covariance matrices. *Bernoulli 19*, pp. 2359–2388.

Cai T. T. (2017). Global testing and large-scale multiple testing for high-dimensional covariance structures. *Annu. Rev. Stat. Appl. 4*, pp. 423–446.

Chen, S. X., Zhang, L. X. and Zhong, P. S. (2010). Testing high dimensional covariance matrices. *J. Am. Stat. Assoc. 105*, pp. 810–819.

El Karoui, N. (2007). Tracy–Widom limit for the largest eigenvalue of a large class of complex sample covariance matrices. *Ann. Probab. 35*, pp. 663–714.

Fan, J., Fan, Y. and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *J. Econom. 147*, pp. 186–197.

Fan, J. Q., Liao, Y. and Yao, J. W. (2015). Power enhancement in high-dimensional cross-sectional tests. *Econometrica 83*, pp. 1497–1541.

Gupta, D. S. and Giri, N. (1973). Properties of tests concerning covariance matrices of normal distributions. *Ann. Stat. 6*, pp. 1222–1224.

Gupta, A. K. and Tang, J. (1984). Distribution of likelihood ratio statistic for testing equality of covariance matrices of multivariate Gaussian models. *Biometrika 71*, pp. 555–559.

Jiang, T. F. and Yang, F. (2013). Central limit theorems for classical likelihood ratio tests for

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0275

high-dimensional normal distributions. *Ann. Stat. 41*, pp. 2029–2074.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal com- ponents analysis. *Ann. Stat. 29*, pp. 295–327.

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Am. Stat. Assoc. 104*, pp. 682–693.

Li, J. and Chen, S. X. (2012). Two-sample tests for high dimensional covariance matrices. *Ann. Stat. 40*, pp. 908-940.

Markowitz, H. (1952). Portfolio selection. *J. Finance 7*, pp. 77–91.

O'Brien (1992). Roubst procedures for testing equality of covariance matrices. *Biometrics 48*, pp. 819–827.

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

Rothman, A. J., Levina, E. and Zhu, J. (2010). A new approach to Cholesky-based covariance regularization in high dimensions. *Biometrika 97*, pp. 539–550.

Schmidt, M., Böhm, D., von Törne, C., Steiner, E., Puhl, A., Pilch, H., Lehr, H., Hengstler, J. G., Kölbl, H., and Gehrmann, M. (2008). The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Res. 68*, pp. 5405–5413.

Schott, J. R. (2007). A test for the equality of covariance matrices when the dimension is large

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0275

relative to the sample sizes. *Comput. Stat. Data Anal. 51*, pp. 6535–6542.

Srivastava, M. S. and Yanagihara, H. (2010). Testing the equality of several covariance matrices with fewer observations than the dimension. *J. Multivar. Anal. 101*, pp. 1319–1329.

Srivastava, M. S., Yanagihara, H. and Kubokawa, T. (2014). Tests for covariance matrices in high dimension with less sample size. *J. Multivar. Anal. 130*, pp. 289–309.

Sugiura, N. and Nagao, H. (1968). Unbiasedness of some test criteria for the equality of one or two covariance matrice. *Ann. Math. Stat. 39*, pp. 1682–1692.

Yang, Q. and Pan, G. (2017). Weighted statistic in detecting faint and sparse alternatives for high-dimensional covariance matrices. *J. Am. Stat. Assoc. 517*, pp. 188–200.

School of Mathematics & Statistics and KLAS, Northeast Normal University, China

E-mail: zhengsr@nenu.edu.cn

Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, U.S.A.

E-mail: ruitaolin@gmail.com

School of Mathematics & Statistics and KLAS, Northeast Normal University, China

E-mail: jhguo@nenu.edu.cn

Department of Statistics and Actuarial Science, University of Hong Kong, Hong Kong, China

E-mail: gyin@hku.hk