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
Notice: Accepted version subject to English editing.	

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

Testing Homogeneity of High-dimensional Covariance Matrices

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

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

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

[3]Department of Statistics and Actuarial Science, University of Hong Kong, Hong Kong, China

Abstract

Testing homogeneity of multiple high-dimensional covariance matrices is becoming more critical in multivariate statistical analysis owing to the emergence of big data. Many existing homogeneity tests for high-dimensional covariance matrices mainly focus on two populations, and they often target at some specific situations, for example, either sparse alternatives or dense alternatives, thus the available methods are not suitable for general cases with multiple groups. To accommodate various situations, we propose a power-enhancement high-dimensional test for multi-sample comparisons of covariance matrices, which includes the homogeneity testing of two matrices as a special case. Not only do the proposed tests require no distributional assumption, but they can also handle both sparsity and non-sparsity structures. Based on random

matrix theories, the asymptotic normality properties of our tests are established under both the null and alternative hypotheses. Numerical studies demonstrate substantial gain in power for our proposal, and the new method is illustrated with a gene expression dataset from the breast cancer study.

Key words: Asymptotic normality; High-dimensional covariance matrix; Homogeneity test; Multi-sample comparison; Power enhancement.

1 Introduction

Covariance matrices play a fundamental role in multivariate statistical inference. In various fields, such as economy and biology, many modern statistical procedures require testing the equality of covariance matrices; for example, multivariate analysis of variance or Fisher's linear discriminant analysis, and so on. In the conventional low-dimensional setting where the dimension of variables is relatively small compared with the sample size, tests for 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.

Due to the rapid development of science and technology, a large amount of data can be collected and stored at an ever-increasing speed and capacity, which usually results in high dimensions for the observations when the number of variables is large relative to the sample size. Conventional methods for testing equality of covariance matrices usually fail in high-dimensional settings, as the sample covariance matric does not converge to the population counterpart in high-dimensional situations. For inference on high-dimensional covariance matrices, extensive research has been conducted in analyzing the limiting distributions of extreme eigenvalues of the sample covariance matrix (Bai, 1993; Johnstone, 2001; El Karoui, 2007; Johnstone and Lu, 2009; Bai and Silverstein, 2010), estimation 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 cutting-edge statistical methods have been proposed for testing two or more high-dimensional covariance matrices (Cai, 2017). For example, Bai et al. (2009) and Jiang and Yang (2013) considered the likelihood ratio statistics to test the equality of two population covariance matrices when the dimension is smaller than the sample size. On the other hand, the likelihood ratio statistic cannot be defined under the circumstance where the dimension is greater than the sample size. Under the "large p, small n" situations, 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 under comparison. Srivastava and Yanagihara (2010) proposed a test statistic based on a distance 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 various cases including 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, some limitations are acknowledged as follows. For example, both approaches are developed for two populations only and thus are not valid for multiple populations (more than two populations). Furthermore, the method in Li and Chen (2012) aims for non-sparse dense alternatives, where many small disturbances may exist, and that of Cai, Liu and Xia (2013) focuses on the sparse alternatives, i.e., the number of non-zero 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 non-sparse dense alternative. This is due to the fact

that these two test procedures consider only one type of norm to characterize the distance between the two sample covariance matrices: the former utilizes the Frobenius norm while the latter suggests the maximum norm. Yang and Pan (2017) proposed a weighted test statistic that is suitable for both sparse and dense alternatives based on random matrix theories, while their approach involves complicated two-dimensional contour integrals that usually do not have explicit expressions, and thus the method is difficult to implement in practice. In addition, there may involve more than two groups for comparison in real studies, and research in testing several (more than two) high-dimensional matrices is rather limited. Schott (2007) and Srivastava and Yanagihara (2010) addressed the problem of comparing multiple high-dimensional covariance matrices, yet as aforementioned, such tests are strictly bundled with the Gaussian population assumption.

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

(i) Our test is applicable for both two-sample and multiple-sample comparisons. For testing homogeneity of several high-dimensional covariance matrices, existing methods (Schott, 2007; Srivastava and Yanagihara, 2010) focus on Gaussian cases and thus they may not work well in non-Gaussian situations. On the other side, our test statistic does not require distributional assumptions, and it demonstrates substantial improvement over the existing tests in the settings with multiple groups. Deriving the theoretical properties of the weighted statistic is not a trivial extension of the case with two populations and, in particular, the mutual dependence among the pairwise components poses major theoretical challenges.

(ii) Our test is suitable for both sparse and non-sparse alternatives. In practice, the structure of the difference between two covariance matrices is typically unknown, which

casts obstacles to implementation of the existing methods, as they utilize only one type of norm to characterize the discrepancy between two samples which is often inadequate. On the contrary, our test statistic is composed of two terms: the first term, which acts as the main one, quantifies the Frobenius norm and thus can capture the difference between two covariance matrices in the non-sparse setting; and the second term utilizes the screening technique with a maximum norm to enhance the power under sparse alternatives. By combining these two terms, the proposed test is broadly applicable in testing various alternatives including both sparse and non-sparse alternatives or the mixture of them. Our approach is much easier to implement than the one by Yang and Pan (2017). We conduct extensive simulation studies to show that our proposed test has comparable performance with the existing methods for two-sample populations. In certain cases, our proposed test can achieve much higher power.

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

The arrangement of this paper is as follows. Section 2 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 theories. Theoretical power of the proposed test is also studied. Section 3 presents various simulation results to demonstrate the superiority of the proposed methods for testing homogeneity of two or more covariance matrices. Section 4 analyzes a real dataset as an illustration of the

proposed testing method. Section 5 provides some concluding remarks, and all technical details are presented in the Supplementary Material.

2 Testing homogeneity of multiple covariance matrices

2.1 The test statistic

Considering K groups, let $\{\mathbf{x}_{k1}, \ldots, \mathbf{x}_{kn_k}\}$ be independent samples from the kth p-dimensional population with the 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 population covariance matrices of these K groups,

$$H_{0K} : \boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_K = \boldsymbol{\Sigma} \quad \text{versus} \quad H_{AK} : \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K \text{ are not all equal} \tag{2.1}$$

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

$$\mathbf{S}_k = (n_k - 1)^{-1} \sum_{i=1}^{n_k} (\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)^T, \quad k = 1, \ldots, K \tag{2.2}$$

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 (2.1) focus on the case of $K = 2$, and they are often established based on two types of norms of $\mathbf{S}_{k_1} - \mathbf{S}_{k_2}$, $1 \le k_1, k_2 \le K$. For example, Li and Chen (2012) used the statistic $\mathrm{tr}(\mathbf{S}_{k_1} - \mathbf{S}_{k_2})^2$, and Cai, Liu and Xia (2013) considered the maximum statistic, $\max\{\delta_{k_1 k_2 \ell_1 \ell_2}, \ell_1, \ell_2 = 1, \ldots, p\}$, where

$$\delta_{k_1 k_2 \ell_1 \ell_2} = \frac{(s_{k_1 \ell_1 \ell_2} - s_{k_2 \ell_1 \ell_2})^2}{n_{k_1}^{-1} \hat{\theta}_{k_1 \ell_1 \ell_2} + n_{k_2}^{-1} \hat{\theta}_{k_2 \ell_1 \ell_2}} \tag{2.3}$$

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 their own advantages and disadvantages. For example, the first trace-based statistic can capture many

small differences, and it possesses high power for testing dense $\mathbf{\Sigma}_{k_1} - \mathbf{\Sigma}_{k_2}$, while there usually incurs some power loss for sparse $\mathbf{\Sigma}_{k_1} - \mathbf{\Sigma}_{k_2}$. On the other side, the second statistic is able to detect large disturbances when $\mathbf{\Sigma}_{k_1} - \mathbf{\Sigma}_{k_2}$ is sparse, but it 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 adaptively chosen 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 sometimes it may possess even more complicated structures such as a mixture of dense and sparse signals. As a result, it is more sensible to develop a test statistic that attains desirable power under both dense and sparse cases, while using only one norm is inadequate for this purpose. A natural approach is to take a linear combination of the aforementioned two statistics as in Yang and Pan (2017), but its limiting distribution is much more complicated due to the correlation between the two statistics. In addition, existing methods for testing (2.1) with $K \geq 3$ are limited, and extensions 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 the current available methods do not work well. To borrow the strength from the trace-based and maximum norms, we propose a new statistic, $T_K = T_{K1} + T_{K2}$, with

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

where $\{\omega_{k_1k_2}, 1 \leq k_1 \leq k_2 \leq K\}$ are the prespecified weights, $\omega_{k_1k_2} \geq 0$ 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 depending on sample sizes n_{k_1}, n_{k_2} and dimension p. When $K = 2$, the proposed procedure reduces to the homogeneity test of 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 as its components are not independent. Based on random matrix theories, the limiting distributions of T_K are derived under both the null hypothesis H_{0K} and the alternative hypothesis H_{AK}.

In fact, the proposed statistic is in a similar spirit to the power-enhancement test statistic proposed by Fan, Liao and Yao (2015). The first term T_{K1} plays a dominant role for testing dense cases. With a properly chosen threshold $s(n_{k_1}, n_{k_2}, p)$, the second term T_{K2} serves as the screening purpose, which converges to zero under the null hypothesie, and converges 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 be very large quickly as long as sparse disturbances are detected by T_{K2} if K_0 is large enough; for more discussion on the choice of K_0, see Fan, Liao and Yao (2015).

2.2 Limiting distributions

We first impose two assumptions that are commonly used in random matrix theories.

(A1) The vector $\mathbf{x}_{ki}$ satisfies the independent component structure $\mathbf{x}_{ki} = \boldsymbol{\mu}_k + \boldsymbol{\Gamma}_k \mathbf{w}_{ki}$, where $\mathbf{w}_{ki} = (w_{k1i}, \ldots, w_{kpi})^T$, 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 $q = 1, 2, 3, 4$.

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

Assumption (A1) requires the independent component structure of the populations. The population fourth moment of $w_{k\ell i}$ is required to exist, while no other distributional assumptions are imposed. The dimension and the sample size are assumed to tend to infinity

proportionally under Assumption (A2). These two assumptions are regular ones 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 multiple-sample comparisons with $k = 1, \ldots, K$, we have*

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

and if further 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*

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

where

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

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, for which the detailed proof is given in the Supplementary Material.

Remark 1 In deriving the CLT of 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 $4[\mathrm{tr}(\mathbf{\Sigma}^2)]^2[(n_{k_1}-1)^{-1}+(n_{k_2}-1)^{-1}]^2$, then a reasonable weight is 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, and its asymptotic null distribution can be immediately obtained.

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

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

and if further 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*

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

where

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

$$\hat{\sigma}_2^2 \;=\; 4[(n_1-1)^{-1}+(n_2-1)^{-1}]^2[\mathrm{tr}(\mathbf{S}^2)-(n_1+n_2-2)^{-1}(\mathrm{tr}\mathbf{S})^2]^2,$$

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

Remark 2 To apply the proposed test to two groups with $K=2$ in practice, 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} it satisfies $P(\max_{1\le\ell_1\le\ell_2\le p}\delta_{12\ell_1\ell_2}\le s(n_1,n_2,p))$ 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 $K\ge 3$, due to multiple pairwise comparisons, we set the threshold to be

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

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 specified thresholds $s(n_{k_1},n_{k_2},p)$ or $s(n_1,n_2,p)$ both satisfy the condition that $s(n_{k_1},n_{k_2},p)-4\log p\ge 0$. The choice of K_0 has been discussed extensively in Fan, Liao and Yao (2015). In general, K_0 should be large enough in order to reject the null once the sparse signals are detected, and we take $K_0=p^2$.

2.3 Power comparison

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

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

where $z_{1-\alpha}$ is the $(1-\alpha)$th quantile of the standard normal distribution $N(0,1)$. Therefore, the power function for testing (2.1) is

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

due to the fact that $T_K = T_{K1} + T_{K2}$ with $T_{K2} \geq 0$.

To investigate the power of the proposed test, let

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

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

Theorem 2 *Under Assumptions (A1)–(A2) and let* $\mathbf{A}_{k_1k_2} = \boldsymbol{\Sigma}_{k_1} - \boldsymbol{\Sigma}_{k_2}$ *for* $1 \leq k_1 < k_2 \leq K$, *for multiple-sample 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*

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

with the 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* $\min_{1\le\ell_1\le\ell_2\le p} \theta_{k\ell1\ell_2}(\sigma_{k\ell_1\ell_1}\sigma_{k\ell_2\ell_2})^{-1} \ge \tau_k$ *where* $\mathbf{\Sigma}_k = (\sigma_{k\ell_1\ell_2})^p_{\ell_1,\ell_2=1}$, $\theta_{k\ell_1\ell_2} = \mathrm{Var}[(x_{k\ell_1i} - \mathrm{E}x_{k\ell_1i})(x_{k\ell_2i} - \mathrm{E}x_{k\ell_2i})]$, 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

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

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

Based on the asymptotic normality of T_{K1} as well as T_K as shown in Theorem 2, we obtain a 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 large enough, we have* $g_K(\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_K) \ge \alpha$ *with the nominal size* α *and, in particular, when* $\max\{\mathrm{tr}(\mathbf{A}^2_{ij}), 1 \le i < j \le K\} > c_1$ *for small positive constant* c_1, *we have* $g_K(\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_K) > \alpha$.

(ii) When $\mathrm{tr}(\mathbf{A}^2_{ij}) \to \infty$ *for some* $1 \le i < j \le K$, *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* $s(n_{k_1}, n_{k_2}, p) + 4\log p \le 0.5\max_{1\le i\le j\le p}[(\sigma_{k_1ij} - \sigma_{k_2ij})^2(\theta_{k_1ij}/n_{k_1} + \theta_{k_2ij}/n_{k_2})^{-1}]$, *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 the power comparison between the proposed tests T_2, T_{21} with the existing ones of Li and Chen (2012) and Cai, Liu and Xia (2013). In particular, 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 similarly denote those of Li and Chen (2012) and Cai, Liu and Xia (2013) by $g_{\mathrm{LC}}(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2)$ and $g_{\mathrm{CLX}}(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2)$, respectively.

First of all, since $T_2 = T_{21} + T_{22}$ 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} due 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 one as $n_1, n_2, p \to \infty$ if K_0 is large enough. On the other hand, it can be shown under this situation that $g_{21}(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) \to \alpha$ and $g_{\mathrm{LC}}(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) \to \alpha$ (Cai, Liu and Xia, 2013), which in turn demonstrates that the screening term T_{22} can indeed enhance the power of T_{21}. Such a property of the proposed T_2 test will be confirmed by simulation results of Scenario 3 in the simulation study.

Second, if all the absolute entries of $\mathbf{A}_{12}$ are not greater than $[\min\{n_1, n_2\}]^{-1-\epsilon}\log p$ and $\mathrm{tr}\mathbf{A}_{12}^2 \to \infty$ with $\epsilon > 0$, and $\theta_{k\ell_1\ell_2}$ has uniform positive lower and upper bounds, then we have

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

with $T_{22} \to 0$. That is, when all the entries of $\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2$ are small nonzeros, the power of the test in Cai, Liu and Xia (2013) may be relatively small but our test T_2 as well as that in Li

and Chen (2012) can discriminate between $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ with high power, which corresponds to Scenarios 1 and 2 in the simulation study.

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

3 Simulation studies

We evaluate the finite-sample performance of the proposed test with two populations and three populations by extensive simulation studies. For testing equality of two population covariance matrices, we compare our test T_2 with four existing methods of Yang and Pan (2017) (YP), Li and Chen (2012) (LC), Cai, Liu and Xia (2013) (CLX), Schott (2007) (SC) and SY (Srivastava and Yanagihara, 2010). For three-sample covariance matrix testing problems, we consider three methods including T_3, SC and SY. The sample sizes are taken to be $n_k = 60, 100, 200, 300$ 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 $\text{Gamma}(4, 2) - 2$. The nominal test size is 5%, and we conduct 5000 replications to summarize the empirical proportion of rejecting the null hypothesis under each case. Four scenarios are considered for 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 $\{u_{3ij}, i, j = 1, \ldots, p\} \overset{\text{i.i.d.}}{\sim} \text{Unif}(-n_1^{-0.9}, n_1^{0.9})$. For testing $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ with $K = 2$, we evaluate the empirical test size with $\theta_1 = \theta_2 = 0$ and empirical power with

$(\theta_1, \theta_2) = (0, 1)$. For testing equality of three covariance matrices, $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}_3$ with $K = 3$, we evaluate the empirical test size with $\theta_1 = \theta_2 = \theta_3 = 0$ and empirical power with $(\theta_1, \theta_2, \theta_3) = (0, 1, 1)$. Such a configuration is to examine the performance of the proposed test under the dense alternative.

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). For testing $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ with $K = 2$, we evaluate the empirical test size with $\theta_1 = \theta_2 = 0$ and empirical power with $(\theta_1, \theta_2) = (0, 0.6)$. For testing equality of three covariance matrices, $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}_3$ with $K = 3$, we evaluate the empirical test size with $\theta_1 = \theta_2 = \theta_3 = 0$ and empirical power with $(\theta_1, \theta_2, \theta_3) = (0, 0.4, 0.6)$. 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 $\mathbf{C} = (0.2^{I\{|i-j|>0\}})_{i,j=1}^p$ and $\mathbf{U}_k$ 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 among $\mathbf{\Sigma}_k$'s are extremely sparse. For testing $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ 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 with $\theta_1 = \theta_2 = 0$ and empirical power with $(\theta_1, \theta_2) = (0, 1)$. For testing $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}_3$ 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 with $\theta_1 = \theta_2 = \theta_3 = 0$ and empirical power with $(\theta_1, \theta_2, \theta_3) = (0, 1, 1)$.

Scenario 4. The covariance matrix is $\mathbf{\Sigma}_k = (\sigma_{kij})_{i,j=1}^p$ and $\mathbf{\Gamma}_k = \mathbf{\Sigma}_k^{1/2}$ with $\sigma_{kij} = (0.1^{|i-j|}+0.2^{|i-j|})/2+\theta_k[I\{k = 2\}(2 \log p/3)\mathbf{E}_{11}+I\{k = 3\}(\log p/2)\mathbf{E}_{22}+I\{k = 2\}u_{kij}]$, where $\mathbf{E}_{ij}$ is the matrix with the (i, j)th entry being one and the rest being zero, and $u_{kij} \sim$ Unif$(-n^{-0.8}, n^{-0.8})$ with $n = \sum_{k=1}^K n_k$ for $k = 1, \ldots, K$. For testing $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ with $K = 2$, we evaluate the empirical test size with $\theta_1 = \theta_2 = 0$ and empirical power with $(\theta_1, \theta_2) = (0, 1)$. For testing $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}_3$ with $K = 3$, we evaluate the empirical test size with

$\theta_1 = \theta_2 = \theta_3 = 0$ and empirical power with $(\theta_1, \theta_2, \theta_3) = (0, 1, 1)$. Scenario 4 investigates the tests based on a mixture of sparse and dense alternatives.

The simulation results for two-sample testing problems are summarized in Tables 1–4. The results for three samples with Gaussian populations are presented in Figure 1, while those with Gamma populations are provided in Supplementary Material. For testing equality of two covariance matrices, it is observed that our test, LC and CLX can maintain empirical test sizes well for both Gaussian and Gamma populations. By contrast, YP and SC only work for the Gaussian population. 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 of SC under the Gamma populations and those of SY under the two-sample cases. Under the alternatives, we mainly focus on comparing empirical power of our test, LC and CLX. In Scenarios 1 and 2 for dense or relatively sparse cases, our test is as powerful as the LC method, and it produces higher power than CLX. 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 also consider the ultra high-dimensional setting with $p = 500$ and 1000, and report the simulation results of the new test in Table 5. We conclude that our T_2 test still enjoys desirable performance when p is much larger than n. Moreover, we change the threshold to be $s(n_1, n_2, p) = [\{\log\log(n_1/2 + n_2/2) - 1\}^2 + 4](\log p - \log\log p/4) + q$, with $\exp\{-(8\pi)^{-1/2}\exp(-q/2)\} = 0.99$, and the performance of T_2 is similar.

For testing the homogeneity of three covariance matrices, we additionally 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 of the screening term T_{32}, especially in the sparse cases. Based on the simulation results, similar phenomena as those in the two-sample cases are observed. Moreover, the T_3 test demonstrates substantial advantages over the T_{31} test when sparse but

large disturbances exist under the alternative hypothesis, while the empirical sizes for these two tests are comparable.

4 Real data analysis

For illustration of our proposed test, we present an analysis of a gene expression dataset from the breast cancer study by Schmidt et al. (2008). The data are available from "Bioconductor", consisting of 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 well differentiated tumor ($n_1 = 29$), group 2 with moderately differentiated tumor ($n_2 = 137$), and group 3 with poor/undifferentiated tumor ($n_3 = 35$). The heterogeneous nature of breast cancer facilitates development of some 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 dataset contains 22283 features, which indeed leads to a high-dimensional hypothesis testing problem. To alleviate the computational burden, we perform a feature-screening procedure (http://bioconductor.org/packages/release/bioc/html/genefilter.html) by filtering out the features with the coefficients of variation falling outside of the range $(0.25, 1.0)$ and in addition controlling at least 30% of the selected features possessing intensities above five. Through such a preliminary screening procedure, a total number of 1280 features are kept for 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 dataset, 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 2. It is observed that $\mathbf{\Sigma}_2 - \mathbf{\Sigma}_1$ has elements more concentrated around zero than $\mathbf{\Sigma}_3 - \mathbf{\Sigma}_2$ and

one large disturbance (around the index of 20000) may exist between groups 1 and 2, while 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 of $H_{02}^{(1,2)}$ and $H_{02}^{(2,3)}$, while LC and CLX reject only one of these two. Particularly, LC fails to detect the difference between $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ due to the fact that there exists only one large disturbance (feature 206023_at) between the two covariance matrices. On the other hand, CLX cannot distinguish those many small disturbances between $\mathbf{\Sigma}_2$ and $\mathbf{\Sigma}_3$. Such an example demonstrates that the structures of the differences between the two covariance matrices indeed affect the performance of the LC and CLX tests, as each is solely based on one type of norm statistic. Without knowledge of the specific structure of the difference between covariance matrices, the proposed test can identify both "few large" and "many tiny" disturbances, and thus leads to higher power in homogeneity testing problems.

We then 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 this breast cancer dataset is 116.9 with a p-value extremely close to zero, thus leading to the rejection of H_{03}. Moreover, it is noted that the observed statistic of T_{31} is 5.3 with a p-value of 6.5×10^{-8}, which thereby indicates that both terms of the proposed T_3 take effect in detecting the differences for dense and sparse alternatives.

5 Concluding remarks

We have proposed a new test for homogeneity of multiple high-dimensional covariance matrices. In contrast to existing methods that typically use only one type of norm statistic, our test statistic is composed of two different norms with one for strong but few signals and the other for faint but many signals. By adaptively mixing the two norms, our test gains

substantial power for different situations and, more importantly, it is not required to know *a priori* which types of signals are present. The asymptotic properties of our tests are established using modern random matrix theories, which demonstrate the elegancy of theoretical development.

The proposed statistic for testing several matrices is a weighted average of the pairwise testing statistics where the weight is proportional to the inverse of the sample size. A further related question is the determination of the optimal weight that can maximize the power. In fact, the power function can be represented 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_{AK}^2 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 similarly as 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_{AK}^2 can be treated as the risk. In addition, when there is a large number of groups or covariance matrices, i.e., K is also large, it would be of interest to extend our work to accommodate such a situation.

Supplementary material

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

Acknowledgement

We thank the associate editor, the referees, and the editor for their many constructive and insightful comments that have led to significant 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*. 3rd Edition. John Wiley & Sons.

Bai, Z. D. (1993). Convergence rate of expected spectral distributions of large random matrices. II. Sample covariance matrices. *Ann. Probab.*, **21**, 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**, 3822-3840.

Bai, Z. D. and Silverstein, J. W. (2004). CLT for linear spectral statistics of large dimensional sample covariance matrices. *Ann. Probab.*, **32**, 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 matri- ces. *Ann. Stat.*, **36**, 199–227.

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

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**, 265-277.

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

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

Chen, S. X., Zhang, L. X. and Zhong, P/ S. (2010). Testing high dimensional covariance matrices. *J. Am. Stat. Assoc.*, **105**, 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**, 663–714.

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

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

Gupta, D. S. and Giri, N. (1973). Properties of tests concerning covariance matrices of normal distributions. *Ann. Stat.*, **6**, 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**, 555-559.

Jiang, T. F. and Yang, F. (2013). Central limit theorems for classical likelihood ratio tests for high-dimensional normal distributions. *Ann. Stat.*, **41**, 2029–2074.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Stat.*, **29**, 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**, 682–693.

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

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

O'Brien (1992). Roubst procedures for testing equality of covariance matrices. *Biometrics*, **48**, 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**, 247-263.

Rothman, A. J., Levina, E. and Zhu, J. (2010). A new approach to Cholesky-based covariance regularization in high dimensions. *Biometrika*, **97**, 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**, 5405–5413.

Schott, J. R. (2007). A test for the equality of covariance matrices when the dimension is large relative to the sample sizes. *Comput. Stat. Data Anal.*, **51**, 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**, 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**, 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**, 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**, 188–200.

Table 1: Comparison of empirical sizes and power (in percentage) of the proposed two-sample test T_2 with four existing methods under Scenario 1.

			Gaussian					Gamma			
p	n_1	n_2	T_2	CLX	LC	SC	YP	T_2	CLX	LC	YP
							Size (%)				
100	60	60	6.1	5.3	4.9	5.2	7.1	5.8	3.3	6.3	21.4
	100	200	5.8	5.5	4.2	5.0	—	5.4	3.6	6.6	—
	200	200	6.3	4.9	3.8	5.5	5.0	5.7	3.4	4.2	13.5
	300	300	5.9	4.6	5.4	5.4	6.0	5.9	3.8	5.2	11.9
300	60	60	5.9	6.0	4.2	4.9	10.5	5.0	3.4	4.6	22.5
	100	200	5.2	4.5	7.8	5.2	—	4.9	3.2	5.2	—
	200	200	5.6	4.4	4.2	5.2	4.1	5.4	2.8	6.0	13.9
	300	300	5.6	4.2	5.6	5.4	6.7	5.4	3.2	4.0	13.4
							Power (%)				
100	60	60	63.8	6.9	62.2	66.2	94.6	63.3	4.7	62.8	95.6
	100	200	68.7	8.7	68.8	70.2	—	67.1	6.2	66.4	—
	200	200	32.1	6.2	30.8	32.0	36.8	31.9	4.4	29.8	44.0
	300	300	26.1	5.6	27.2	26.1	24.2	26.1	4.7	26.8	36.0
300	60	60	99.9	7.5	100.0	100.0	99.8	99.9	4.9	100.0	97.5
	100	200	100.0	9.9	100.0	100.0	—	100.0	7.4	100.0	—
	200	200	96.4	5.6	95.8	96.6	100.0	95.8	3.5	96.0	100.0
	300	300	87.4	5.8	88.6	87.7	99.9	87.6	3.8	86.8	100.0

Note: Four existing tests include Yang and Pan (2017) (YP), Li and Chen (2012) (LC), Cai, Liu and Xia (2013) (CLX) and Schott (2007) (SC), and "—" denotes "not applicable".

Table 2: Comparison of empirical sizes and power (in percentage) of the proposed two-sample test $T_2$with four existing methods under Scenario 2.

			Gaussian					Gamma			
p	n_1	n_2	T_2	CLX	LC	SC	YP	T_2	CLX	LC	YP
						Size (%)					
100	60	60	6.6	5.7	5.0	5.7	100.0	6.0	3.8	6.1	100.0
	100	200	5.5	5.9	4.8	4.4	—	5.8	4.4	5.6	—
	200	200	6.3	4.9	4.0	5.7	99.8	6.1	3.7	6.6	97.7
	300	300	5.6	4.7	2.6	5.3	70.6	5.4	3.3	5.4	96.3
300	60	60	5.8	5.4	5.6	4.9	100.0	5.1	3.4	4.9	100.0
	100	200	5.8	4.9	5.6	5.1	—	6.5	3.8	4.4	—
	200	200	6.1	4.3	4.4	6.0	100.0	5.8	3.2	4.8	100.0
	300	300	5.4	4.6	6.8	5.1	53.5	5.0	3.7	4.4	99.9
						Power (%)					
100	60	60	37.9	8.4	38.5	38.9	100.0	38.0	6.5	40.4	100.0
	100	200	90.0	56.3	89.8	88.7	—	88.8	53.5	86.2	—
	200	200	99.3	53.7	99.6	99.3	100.0	99.1	42.8	99.2	99.9
	300	300	100.0	91.8	100.0	100.0	99.8	100.0	81.0	100.0	95.3
300	60	60	36.2	6.4	37.8	39.0	100.0	36.8	4.8	38.4	100.0
	100	200	91.5	53.7	85.0	90.5	—	91.4	47.9	89.4	—
	200	200	99.7	44.0	99.8	99.7	100.0	99.7	31.4	99.2	100.0
	300	300	100.0	90.3	100.0	100.0	100.0	100.0	75.7	100.0	99.7

Table 3: Comparison of empirical sizes and power (in percentage) of the proposed two-sample test T_2 with four existing methods under Scenario 3.

			Gaussian					Gamma			
p	n_1	n_2	T_2	CLX	LC	SC	YP	T_2	CLX	LC	YP
						Size (%)					
100	60	60	6.3	5.2	5.6	5.2	100.0	5.5	3.3	4.5	100.0
	100	200	5.9	4.5	5.0	5.3	—	5.7	3.5	7.0	—
	200	200	5.8	4.6	6.8	5.6	50.0	5.8	3.1	5.4	99.5
	300	300	5.6	4.4	5.4	5.1	9.4	5.5	4.0	3.6	30.8
300	60	60	5.4	5.4	5.1	4.8	99.7	4.6	3.1	5.2	100.0
	100	200	5.6	5.4	6.2	5.0	—	5.0	3.5	5.8	—
	200	200	5.5	4.4	4.4	4.9	77.3	5.2	2.6	6.4	9.5
	300	300	5.2	4.7	5.6	5.0	23.9	5.5	3.5	4.6	80.6
						Power (%)					
100	60	60	43.5	53.8	7.9	11.5	100.0	37.2	42.9	16.0	100.0
	100	200	99.0	99.6	13.4	24.5	—	96.1	98.3	29.8	—
	200	200	100.0	100.0	43.2	41.2	100.0	99.8	99.9	36.6	100.0
	300	300	100.0	100.0	85.4	66.2	100.0	100.0	100.0	88.6	100.0
300	60	60	24.2	33.3	5.9	6.5	100.0	20.0	25.7	6.6	100.0
	100	200	95.8	98.5	13.6	8.9	—	89.9	95.4	11.6	—
	200	200	99.9	100.0	6.8	12.0	99.9	99.1	99.7	11.0	100.0
	300	300	100.0	100.0	27.0	18.2	100.0	100.0	100.0	12.6	100.0

Table 4: Comparison of empirical sizes and power (in percentage) of the proposed two-sample test T_2 with four existing methods under Scenario 4.

			Gaussian					Gamma			
p	n_1	n_2	T_2	CLX	LC	SC	YP	T_2	CLX	LC	YP
							Size (%)				
100	60	60	6.0	5.4	4.5	5.3	8.0	5.0	3.8	5.5	17.8
	100	200	5.5	4.5	5.0	5.2	—	6.2	3.9	6.0	—
	200	200	5.8	4.2	7.2	5.2	6.5	6.1	3.3	6.0	11.6
	300	300	5.1	4.7	4.4	4.8	5.5	5.5	3.8	5.0	12.8
300	60	60	6.0	5.4	4.7	4.9	7.7	4.9	3.0	6.6	20.1
	100	200	5.2	5.5	4.4	4.6	—	5.3	3.5	4.6	—
	200	200	4.8	5.1	4.8	4.6	5.8	5.4	2.7	5.2	11.3
	300	300	5.3	4.4	5.4	5.0	5.7	5.4	3.3	4.8	13.7
							Power (%)				
100	60	60	36.2	17.6	31.0	33.8	86.7	32.9	8.7	34.4	84.9
	100	200	96.1	98.1	47.2	48.7	—	71.7	69.4	48.3	—
	200	200	99.2	99.8	67.2	70.3	99.7	85.5	79.6	65.2	97.5
	300	300	100.0	100.0	87.2	89.5	99.9	98.1	97.3	83.0	99.2
300	60	60	65.1	16.1	62.0	67.2	96.7	65.0	7.7	65.4	79.3
	100	200	99.3	99.7	57.0	55.5	—	81.9	76.5	80.3	—
	200	200	99.8	99.9	68.0	68.0	100.0	89.2	83.8	67.2	99.8
	300	300	100.0	100.0	78.5	79.7	100.0	98.6	98.1	75.6	100.0

Table 5: Empirical sizes and empirical power (in percentage) of the proposed two-sample test T_2 for ultra high-dimensional cases under Scenarios 1–4.

			Size (%)				Power (%)			
p	n_1	n_2	1	2	3	4	1	2	3	4
500	60	60	5.5	5.8	5.9	5.5	100.0	35.5	18.8	84.7
	100	200	5.0	5.2	5.5	5.6	100.0	91.8	93.7	99.7
	200	200	4.9	5.5	5.4	5.0	100.0	99.8	99.8	99.9
	300	300	5.5	5.2	5.4	5.1	100.0	100.0	100.0	100.0
1000	60	60	4.8	5.2	5.3	4.5	100.0	34.5	14.1	98.1
	100	200	5.1	5.3	4.9	5.3	100.0	92.4	90.0	100.0
	200	200	5.7	5.4	4.3	4.7	100.0	99.5	99.4	100.0
	300	300	4.9	5.2	5.4	4.9	100.0	100.0	100.0	100.0

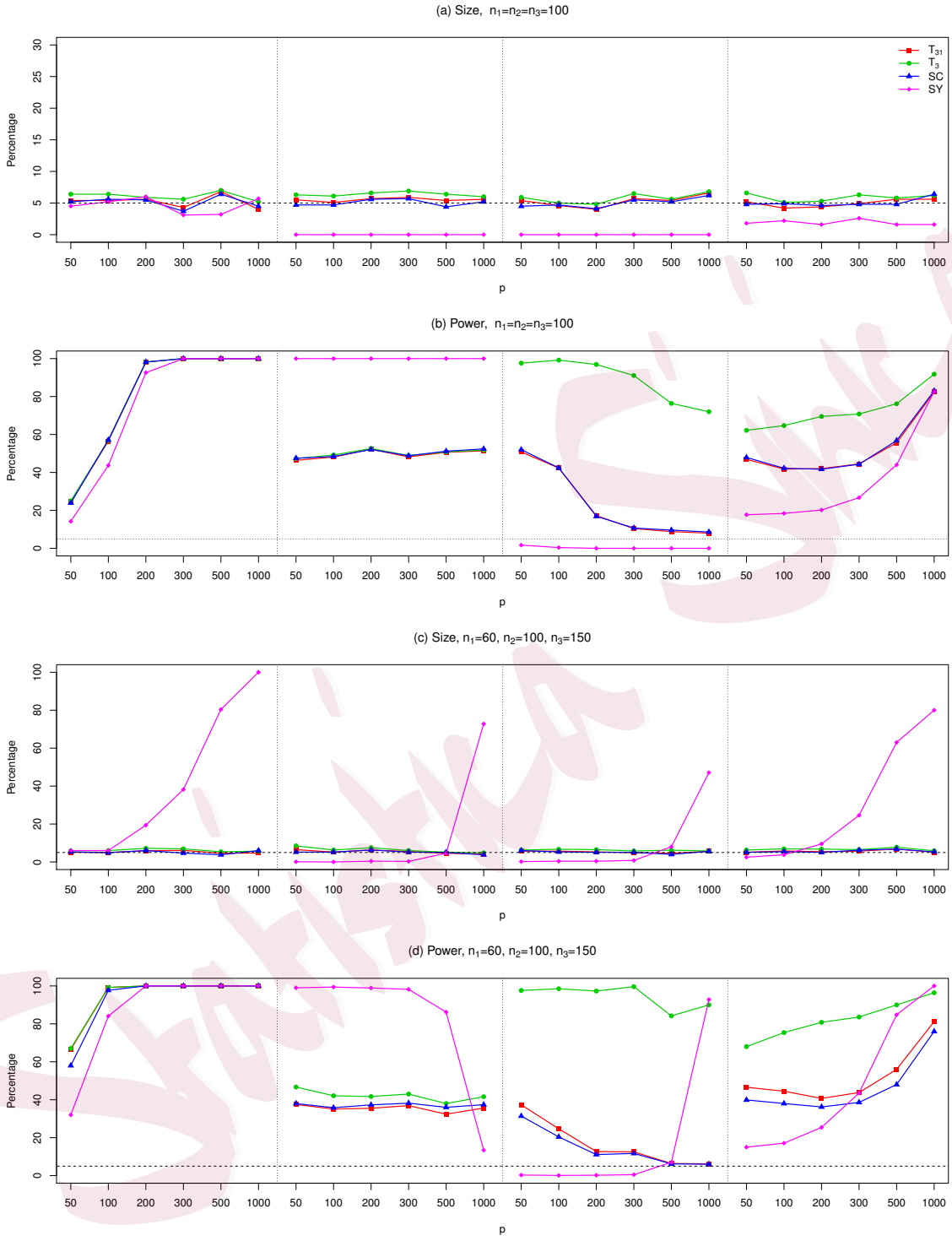

Figure 1: Simulation results for testing the equality of three covariance matrices with Gaussian populations under Scenarios 1–4 in comparison with two existing tests of Schott (2007) (SC) and Srivastava and Yanagihara (2010) (SY).

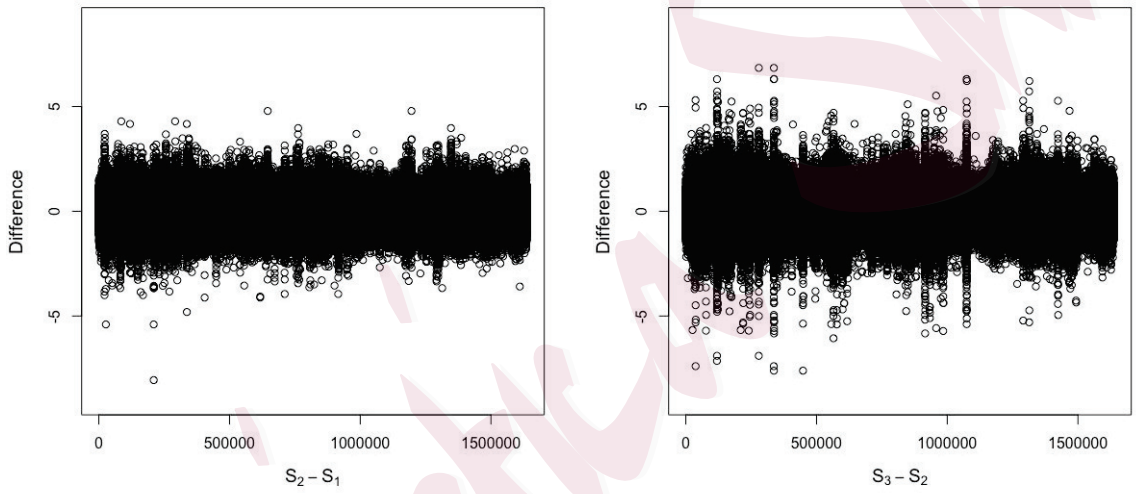

Figure 2: Plots of values of $s_{k_1\ell_1\ell_2} - s_{k_2\ell_1\ell_2}$ to quantify the difference between $\mathbf{\Sigma}_{k_1}$ and $\mathbf{\Sigma}_{k_2}$ for the breast cancer dataset, $\ell_1, \ell_2 = 1, \ldots, 1280$.