

Article

Homogeneity Test of Multi-Sample Covariance Matrices in High Dimensions

Peng Sun ^{1,2,3}, Yincui Tang ^{1,2,*}  and Mingxiang Cao ^{4,*}¹ Department of Statistics, East China Normal University, Shanghai 200062, China² KLATASDS-MOE, School of Statistics, East China Normal University, Shanghai 200062, China³ Center for Quantitative Medicine, Duke-NUS Medical School, Singapore 169857, Singapore⁴ School of Mathematics and Statistics, Anhui Normal University, Anhui 241002, China

* Correspondence: yctang@stat.ecnu.edu.cn (Y.T.); caomingx@ahnu.edu.cn (M.C.)

Abstract: In this paper, a new test statistic based on the weighted Frobenius norm of covariance matrices is proposed to test the homogeneity of multi-group population covariance matrices. The asymptotic distributions of the proposed test under the null and the alternative hypotheses are derived, respectively. Simulation results show that the proposed test procedure tends to outperform some existing test procedures.

Keywords: high-dimensional data; weighted Frobenius norm; homogeneity test; martingale central limit theorem; asymptotic distributions

MSC: 62H15; 62E20



Citation: Sun, P.; Tang, Y.; Cao, M. Homogeneity Test of Multi-Sample Covariance Matrices in High Dimensions. *Mathematics* **2022**, *10*, 4339. <https://doi.org/10.3390/math10224339>

Academic Editor: Stelios Psarakis

Received: 13 October 2022

Accepted: 11 November 2022

Published: 18 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In the last thirty years, statistical methods have made great developments for high-dimensional data. One common feature of high-dimensional data is that the data dimension is larger or vastly larger than the total sample size. In high-dimensional settings, many classic methods are not well-defined or have poor performances. As a result, a great deal of statistical methods have been proposed to deal with high-dimensional data. As an important issue of statistical inference, hypothesis testing is being followed with interest by scholars. Testing the equality of several covariance matrices is a fundamental problem in multivariate statistical analysis and can be found in many relevant works, such as [1,2]. This problem will arise from the analysis of gene expression data. It was shown in [3] that many genes have different variances in gene expressions between disease states. For example, the dataset of the breast cancer of patient is classified in three groups based on their gene expression signatures: well-differentiated tumors, moderately differentiated tumors, and poorly differentiated and undifferentiated tumors. Refs. [4–6] respectively used artificial neural networks, feature selection and a Markov blanket-embedded genetic algorithm to investigate this dataset. This dataset includes 83 samples, which is greatly less than data dimension 2308. For this gene dataset, we are interested in checking whether the covariance matrices of the four groups are the same or not. It is noted that an important assumption for multivariate analysis of variance is that of equal covariance matrices in different groups. Therefore, testing the equality of several covariance matrices is needed. Specifically, assume $X_{i1}, \dots, X_{in_i}$ are independent and identically distributed (iid) with a p -dimensional distribution with mean vector μ_i and covariance matrix Σ_i for $i = 1, \dots, k$ with $k \geq 2$. We want to test the following hypothesis

$$H_0 : \Sigma_1 = \dots = \Sigma_k \text{ vs. } H_1 : H_0 \text{ is not true.} \quad (1)$$

A classical method is the likelihood ratio test proposed by [7], where the population distribution is normal. However, as the data dimension is larger or vastly larger than the

sample size, the likelihood ratio test is not well-defined because of the singularity of sample covariance matrices in probability one. Thus, it is vital to propose new procedures suitable to high-dimensional data. Owing to the curse of dimensionality, it becomes more challenging in high dimensions. For the hypothesis (1), there exists some test procedures in the literature. Based on the Frobenius norm, Refs. [8–11] proposed test statistics, respectively. Ref. [12] presented a test for high-dimensional longitudinal data. Ref. [13] proposed a power-enhancement high-dimensional test when the maximum eigenvalue of Σ_i is bounded or $\text{tr}(\Sigma_i^j) = O(p^j)$ for $j = 1, 2, 3, 4$. Ref. [14] presented a test based on the well-known Box's test ([15]) for high-dimensional normal data. It is noted that [9] imposed the conditions $0 < \lim_{p \rightarrow \infty} \text{tr}(\Sigma^i)/p < \infty$ for $i = 1, \dots, 8$ to obtain the asymptotical null distribution of his test under normality, where Σ is equal to the covariance matrix under the null hypothesis H_0 . Under high-dimensional normal data, [10] proposed a test as $0 < \lim_{p \rightarrow \infty} \text{tr}(\Sigma^i)/p < \infty$ for $i = 1, \dots, 4$. Ref. [8] imposed $0 < \text{tr}(\Sigma^2)/p < \infty$ and other conditions to build asymptotical properties of his test statistic. The test statistics in [8–10,13,14] imposed the condition about the relationship between the data dimension p and the sample size n_i or about the normal data. These conditions restrict the application of their test procedures. Ref. [11] gave a Frobenius norm-based test, which can be seen as an extension of that in [16], where the data dimension and the sample size can change arbitrarily.

The existing Frobenius norm-based tests for the hypothesis H_0 were almost constructed by $\sum_{i=1}^k \text{tr}(\Sigma_i - \bar{\Sigma})^2 = k \sum_{i=1}^k \text{tr}(\Sigma_i - \bar{\Sigma})^2$, where $\bar{\Sigma} = \sum_{i=1}^k \Sigma_i / k$ means average covariance matrix of the population covariance matrices $\Sigma_1, \dots, \Sigma_k$. The deviations of Σ_i from overall $\bar{\Sigma}$ is weighted by all equal weight $1/k$, which, however, cannot emphasize the deviations of populations with large sample sizes. Based on this, we here construct a new test statistic by a different Frobenius norm $\sum_{i=1}^k n_i \text{tr}(\Sigma_i - \Sigma^*)^2$ where $\Sigma^* = \sum_{i=1}^k n_i \Sigma_i / n$ and $n = \sum_{i=1}^k n_i$. Here, Σ^* is a weighted average of k covariance matrices, which can emphasize the deviations of populations with large sample sizes. On the other hand, it is evident that the null hypothesis H_0 holds if and only if $\sum_{i=1}^k n_i \text{tr}(\Sigma_i - \Sigma^*)^2 = 0$.

The main purpose of this paper is to develop a new method to test the homogeneity of k high-dimensional covariance matrices on the basis of the weighted Frobenius norm. For the hypothesis in (1), most existing methods ([8–10,13]) imposed normality or the explicit relationship of between the data dimension p and the sample size n , and may behave poorly under non-normal data or ultra-high dimensional data. However, our method does not require normality or an explicit relationship between p and n , which is similar with the method in [11]. Thus, our method can be applied to non-normal and ultra-high dimensional data. On the other hand, the difference of the method in [11] and ours is that we use a sample-sizes-based Frobenius norm to build test statistics so that the tests in this paper and [11] behave differently, as the sample sizes are unequal. When $k = 2$, our proposed test statistic can be seen as an extension of that [16] except a constant. Hence it can also be applied to two-sample data.

Simulation results show that the proposed test behaves differently from existing tests such as the tests in [11,13,14]. We will discuss the differences between the proposed test and the competing tests through numerical comparisons in various scenarios. We observe that the proposed test outperforms the competing tests in both size and power in many cases.

The remainder of the paper is organized as follows. Section 2 first presents the statistical model and the imposed conditions in order to construct our new test. In Section 3, a new test statistic is proposed and its asymptotic properties are also given. Section 4 presents a numerical study of the proposed test to compare three competing tests. Concluding remarks are provided in Section 5. The proofs of main results are arranged in Appendix A.

2. Preliminaries

We consider the following general multivariate model that is often used in literature:

$$\mathbf{X}_{ij} = \boldsymbol{\mu}_i + \boldsymbol{\Gamma}_i \mathbf{Z}_{ij}$$

for $j = 1, \dots, n_i$ and $i = 1, \dots, k$, where Γ_i is a $p \times r$ matrix for some $r \geq p$ such that $\Gamma_i \Gamma_i^T = \Sigma_i$ and $\mathbf{Z}_{i1}, \dots, \mathbf{Z}_{in_i}$ are r -variate iid random vectors with $E(\mathbf{Z}_{ij}) = 0$ and $\text{Var}(\mathbf{Z}_{i1}) = \mathbf{I}_r$. Furthermore, denote $\mathbf{Z}_{ij} = (z_{ij1}, \dots, z_{ijr})^T$ and $z_{ij1}, \dots, z_{ijr}$ have a finite eighth moment with $E(z_{ijl}^4) = \Delta_l$ where Δ_l is some constant. Moreover, for any positive integers q and α_v , there has

$$E(z_{ijv_1}^{\alpha_1} \cdots z_{ijv_q}^{\alpha_q}) = E(z_{ijv_1}^{\alpha_1}) \cdots E(z_{ijv_q}^{\alpha_q}) \quad (2)$$

as $\sum_{v=1}^q \alpha_v \leq 8$, where $v_1, \dots, v_q$ are distinct indices. It shows from (2) that $z_{ijv_1}, \dots, z_{ijv_q}$ are pseudo-independent, which is naturally satisfied when samples are generated from normal distribution.

In order to obtain the asymptotic distributions of new test statistic, some conditions are imposed as follows:

(C1) As $n \rightarrow \infty$, $n_i/n \rightarrow c_i \in (0, 1)$ for $i = 1, \dots, k$.

(C2) As $n \rightarrow \infty$, $p = p(n_1, \dots, n_k) \rightarrow \infty$ and for arbitrary $l, m, s, h \in \{1, \dots, k\}$, $\text{tr}(\Sigma_l \Sigma_m) \rightarrow \infty$ and $\text{tr}(\Sigma_l \Sigma_m \Sigma_s \Sigma_h) = o[\text{tr}(\Sigma_l \Sigma_m) \text{tr}(\Sigma_s \Sigma_h)]$.

(C1) implies that all sample sizes have the same increasing rate, except constant terms. (C2) can be seen as an extension of the condition A2 in [16] to the case of multi-groups. The two conditions are the same as those in [11]. A key aspect of (C2) is that it does not impose any explicit relationship between p and sample size n . It is noted that (C2) naturally holds when all eigenvalues of k covariance matrices are uniformly bounded. Next, we consider another set of covariance matrices satisfying (C2), namely spiked covariance structures. For convenience, we set $k = 4$ and let $\Sigma_i = \text{diag}(a_{i1} p^{\delta_{i1}}, \dots, a_{im_i} p^{\delta_{im_i}}, a_{i, m_i+1}, \dots, a_{ip})$, where a_{ij} s, δ_{ij} s and m_i s are fixed positive constants with $\delta_{i1} \geq \dots \geq \delta_{im_i}$ for $i = 1, 2, 3, 4$. Then, the main items of $\text{tr}(\Sigma_1 \Sigma_2 \Sigma_3 \Sigma_4)$, $\text{tr}(\Sigma_1 \Sigma_2)$ and $\text{tr}(\Sigma_3 \Sigma_4)$ is respectively $a_{11} a_{21} a_{31} a_{41} p^{\delta_{11} + \delta_{21} + \delta_{31} + \delta_{41}} + O(p)$, $a_{11} a_{21} p^{\delta_{11} + \delta_{21}} + O(p)$ and $a_{31} a_{41} p^{\delta_{31} + \delta_{41}} + O(p)$. As a result, (C2) holds if $\delta_{11} + \delta_{21} < 1$ and $\delta_{31} + \delta_{41} < 1$. Let λ_{ij} denote the j th largest eigenvalue of Σ_i , then if $\delta_{11}, \dots, \delta_{41}$ are less than 0.5, $\lambda_{i1}^2 / \text{tr}(\Sigma_i^2) \rightarrow 0$ for $i = 1, \dots, 4$, which is called a non-strongly spiked eigenvalue (NSSE) structure in [17]. Otherwise, if there exists some $\delta_{i1} \geq 0.5$, then $\lambda_{i1}^2 / \text{tr}(\Sigma_i^2) \not\rightarrow 0$, which is called a strongly spiked eigenvalue (SSE) structure in [17]. Therefore, (C2) holds when all of the covariance matrices have NSSE structures, or some of the covariance matrices have SSE structures.

To propose our test statistic, the hypothesis in (1) is rewritten as the following hypothesis

$$H_0 : \sum_{i=1}^k n_i \text{tr}(\Sigma_i - \Sigma^*)^2 = 0 \quad \text{vs.} \quad H_1 : \sum_{i=1}^k n_i \text{tr}(\Sigma_i - \Sigma^*)^2 > 0. \quad (3)$$

Then, we can construct a new test statistic based on $\sum_{i=1}^k n_i \text{tr}(\Sigma_i - \Sigma^*)^2$. Note that

$$\sum_{i=1}^k n_i \text{tr}(\Sigma_i - \Sigma^*)^2 = \sum_{i=1}^k \frac{n_i(n - n_i)}{n} \text{tr}(\Sigma_i^2) - \sum_{i \neq j}^k \frac{n_i n_j}{n} \text{tr}(\Sigma_i \Sigma_j).$$

Thus, a new test statistic can be given if unbiased estimators of $\text{tr}(\Sigma_i^2)$ and $\text{tr}(\Sigma_i \Sigma_j)$ are respectively obtained for $i = 1, \dots, k$.

3. Main Results

In this section, we propose a new test statistic for the hypothesis in (1) and give its asymptotic properties under conditions (C1) and (C2). According to the equivalent hypothesis in (3), our new test statistic is given as the following

$$T := \sum_{i=1}^k \frac{n_i(n - n_i)}{n} A_i - \sum_{i \neq j}^k \frac{n_i n_j}{n} C_{ij}, \quad (4)$$

where A_i and C_{ij} are respectively unbiased estimators of $\text{tr}(\Sigma_i^2)$ and $\text{tr}(\Sigma_i \Sigma_j)$, which are given as follows

$$A_i = \frac{1}{(n_i)_2} \sum_{j,l}^* (\mathbf{X}_{ij}^T \mathbf{X}_{il})^2 - \frac{2}{(n_i)_3} \sum_{j,l,f}^* (\mathbf{X}_{ij}^T \mathbf{X}_{il} \mathbf{X}_{ij}^T \mathbf{X}_{if}) + \frac{1}{(n_i)_4} \sum_{j,l,f,g}^* (\mathbf{X}_{ij}^T \mathbf{X}_{il} \mathbf{X}_{if}^T \mathbf{X}_{ig}) \quad (5)$$

and

$$C_{ij} = \frac{1}{n_i n_j} \sum_{l=1}^{n_i} \sum_{m=1}^{n_j} (\mathbf{X}_{il}^T \mathbf{X}_{jm})^2 - \frac{1}{n_i (n_j)_2} \sum_{l=1}^{n_i} \sum_{f,m}^* (\mathbf{X}_{il}^T \mathbf{X}_{jf} \mathbf{X}_{il}^T \mathbf{X}_{jm}) - \frac{1}{(n_i)_2 n_j} \sum_{l=1}^{n_i} \sum_{f,m}^* (\mathbf{X}_{jl}^T \mathbf{X}_{if} \mathbf{X}_{jl}^T \mathbf{X}_{im}) + \frac{1}{(n_i)_2 (n_j)_2} \sum_{l,m}^* \sum_{f,g}^* (\mathbf{X}_{il}^T \mathbf{X}_{jf} \mathbf{X}_{im}^T \mathbf{X}_{jg}). \quad (6)$$

Here $(n_i)_l = \frac{(n_i)!}{(n_i-l)!}$ and $\sum^*$ denotes the sum for all different indices. Note that T is an unbiased estimator of $\sum_{i=1}^k n_i \text{tr}(\Sigma_i - \Sigma^*)^2$. The above two unbiased estimators are used in [11,16,18,19], respectively. It is noted that we can assume, without loss of generality, $\mu_1 = \mu_2 = \dots = \mu_k$ because T is invariant under location transformation. Under this assumption, the leading terms in A_i and C_{ij} are respectively the first term since the last two terms in A_i and the last three terms in C_{ij} are respectively infinitesimals of higher order of the first term. As a result, we only treat the first term in A_i and C_{ij} to save computation time. Please see [11,18] for more details about this computation time.

It shows from conditions (C1) and (C2) that the variance of T is

$$\begin{aligned} \sigma^2 = & \frac{4}{n^2} \left\{ \sum_{s=1}^k \frac{n_s(n-n_s)^2}{(n_s-1)} \text{tr}^2(\Sigma_s^2) + \sum_{s \neq h}^k 2n_s n_h \text{tr}^2(\Sigma_s \Sigma_h) \right\} \\ & + \frac{4}{n^2} \sum_{s=1}^k \sum_{h \neq f}^k n_s n_h n_f \text{tr} \left\{ \Gamma_s^T (\Sigma_s - \Sigma_h) \Gamma_s \circ \Gamma_s^T (\Sigma_s - \Sigma_f) \Gamma_s \right\} \\ & + \frac{8}{n^2} \sum_{s=1}^k \sum_{h \neq f}^k n_s n_h n_f \text{tr} \left\{ \Sigma_s (\Sigma_s - \Sigma_h) \Sigma_s (\Sigma_s - \Sigma_f) \right\} + o[\text{tr}^2(\Sigma_s^2)]. \end{aligned}$$

Then, we can obtain the following asymptotic distribution of T .

Theorem 1. Under (C1) and (C2), as $\min\{p, n\} \rightarrow \infty$, we have

$$\frac{T - E(T)}{\sigma} \xrightarrow{d} N(0, 1).$$

Proof. See Appendix B. $\square$

It is clear that the variance of T under the null hypothesis H_0 is

$$\sigma_0^2 = \frac{4}{n^2} \left\{ \sum_{i=1}^k \frac{n_i(n-n_i)^2}{(n_i-1)} \text{tr}^2(\Sigma_i^2) + \sum_{i \neq j}^k 2n_i n_j \text{tr}^2(\Sigma_i \Sigma_j) \right\}.$$

As a result, to formulate the test procedure, we need to give a ratio-consistent estimator of σ_0^2 . In this paper, we use the unbiased estimators A_i and C_{ij} in (5) and (6) to respectively estimate $\text{tr}(\Sigma_i^2)$ and $\text{tr}(\Sigma_i \Sigma_j)$. The following lemma is from [11].

Lemma 1. Under (C1) and (C2), as $\min\{p, n\} \rightarrow \infty$,

$$\frac{A_i}{\text{tr}(\Sigma_i^2)} \xrightarrow{p} 1, \quad \frac{C_{ij}}{\text{tr}(\Sigma_i \Sigma_j)} \xrightarrow{p} 1.$$

On the basis of Lemma 1, a ratio-consistent estimator of σ_0 is given by

$$\hat{\sigma}_0^2 := \frac{4}{n^2} \left\{ \sum_{i=1}^k \frac{n_i(n - n_i)^2}{(n_i - 1)} A_i^2 + \sum_{i \neq j}^k 2n_i n_j C_{ij}^2 \right\}.$$

Then, our new test is proposed by $\hat{T} = T/\hat{\sigma}_0$. It follows from Theorem 1 and Lemma 1 that $\hat{T} \xrightarrow{d} N(E(T)/\sigma_0, \sigma^2/\sigma_0^2)$ as $\min\{p, n\} \rightarrow \infty$. We see especially that the proposed test $\hat{T}$ is asymptotically distributed with the standard norm distribution under the null hypothesis H_0 . As a result, we reject the null hypothesis when $\hat{T} \geq z_\alpha$, where z_α is the upper α quantile of standard normal distribution.

The following corollary is easy to be taken, which gives the asymptotic power function under the hypothesis H_1 .

Corollary 1. Under (C1) and (C2), as $\min\{p, n\} \rightarrow \infty$, we have

$$P(\hat{T} \geq z_\alpha) = \Phi[-z_\alpha \sigma_0/\sigma + E(T)/\sigma] + o(1).$$

4. Simulation Studies

In this section, we compare our proposed test with four existing tests by simulation. The four competing tests are denoted by T_{ZBHW} , T_{ZLGY} and $\rho L_k(y)$ from [11,13,14], respectively. Note that the authors used eight test statistics according to a different y in [14]. Here, we only make simulations for four test statistics to save space, namely $\rho L_k(y_i)$, $i = 1, 2, 3$ and 4. We obtain that z_{ijr} s are independently generated from the standard normal distribution $N(0, 1)$ and the centralized gamma distribution $\text{Gamma}(4, 2) - 2$, respectively, for $i = 1, \dots, k$, $j = 1, \dots, n_i$ and $r = 1, \dots, p$. We set $k = 3$ and $\Gamma_i = \Sigma_i^{1/2}$ for $i = 1, 2$ and 3, where the covariance matrices are considered respectively in the following four cases:

- Case 1: $H_0 : \Sigma_1 = \Sigma_2 = \Sigma_3 = \Lambda \mathbf{U}_0 \Lambda$, $H_1 : \Sigma_i = \Lambda \mathbf{U}_{i-1} \Lambda$, $i = 1, 2, 3$, where $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_p)$ with $\lambda_1, \dots, \lambda_p \stackrel{i.i.d.}{\sim} \text{Unif}(1, 5)$ and $\mathbf{U}_{i-1}$ is a $p \times p$ matrix with the (a, b) th element being $(-1)^{a+b} \left(\frac{5-2^{i-1}}{6} \right)^{|a-b|^{0.1}}$.
- Case 2: $H_0 : \Sigma_1 = \Sigma_2 = \Sigma_3 = \Lambda \mathbf{U}_0 \Lambda$, $H_1 : \Sigma_i = \Lambda \mathbf{U}_{i-1} \Lambda$, $i = 1, 2, 3$, where $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_p)$ with $\lambda_1, \dots, \lambda_p \stackrel{i.i.d.}{\sim} \text{Unif}(1, 5)$ and $\mathbf{U}_{i-1}$ is a $p \times p$ matrix with the (a, b) th element being $(-1)^{a+b} \left(\frac{4-i}{5} \right)^{|a-b|^{0.1}}$.
- Case 3: $H_0 : \Sigma_1 = \Sigma_2 = \Sigma_3 = \Lambda_0$, $H_1 : \Sigma_i = \Sigma_{i-1}$, $i = 1, 2, 3$, where $\Lambda_0 = \text{diag}(\lambda_1, \dots, \lambda_p)$ with $\lambda_1, \dots, \lambda_p \stackrel{i.i.d.}{\sim} \text{Unif}(1, 5)$. Λ_1 is a $p \times p$ symmetric matrix with the (a, a) th element being 1.01, the $(a, a+1)$ th elements being 0.1 and the rest being 0; Λ_2 is a $p \times p$ symmetric matrix with the (a, a) th element being 3, the $(a, a+1)$ th elements being 2, the $(a, a+2)$ th elements being 1 and the rest being 0.
- Case 4: $H_0 : \Sigma_1 = \Sigma_2 = \Sigma_3 = \text{diag}(\omega_1, \dots, \omega_p)$ where $\omega_1, \dots, \omega_p \stackrel{i.i.d.}{\sim} \text{Unif}(0.5, 10)$, $H_1 : \Sigma_i = \mathbf{U}_i \Lambda \mathbf{U}_i^T$, $i = 1, 2, 3$, where $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_p)$ with $\lambda_1, \dots, \lambda_p \stackrel{i.i.d.}{\sim} \text{Gamma}(4, 0.5)$, $\mathbf{U}_i = (\mathbf{W}_i^T \mathbf{W}_i)^{-1/2} \mathbf{W}_i^T$ and $\mathbf{W}_i$ is a $p \times p$ matrix whose entries are independently generated from the normal distribution $N(0, i)$.

Without loss of generality, all the population means are chosen to be 0. The sample size and data dimension are respectively $n = 45, 95$ and $p = 16, 32, 64, 128, 256$. Empirical sizes and powers are computed under the nominal level $\alpha = 0.05$ with 10,000 replications.

All the simulation results are reported in Tables 1–8 and Figures 1–8. Figures provide intuitive observation, and tables show the simulated values. First, Tables 1–8 show that the empirical sizes of the test T_{ZLGY} are seriously inflated, especially when $n = 95$ in all cases of simulation. For example, just as what is presented in Table 1, the empirical sizes of T_{ZLGY} are respectively 0.0719, 0.1230, 0.2814, 0.5641 and 0.7692 when $n = 95$ and $p = 16, 32, 64, 128, 256$. Thus, T_{ZLGY} cannot control the nominal size reasonably. Second, Tables 1–4 imply that the empirical sizes of tests $\rho L_k(y_i)$, $i = 1, 2, 3, 4$, are closer to a given size than those of T_{ZBHW} and $\hat{T}$, while T_{ZBHW} and $\hat{T}$ respectively obtains about 7% and 6% of empirical sizes as $n = 95$. Thus, $\rho L_k(y_i)$ outperforms T_{ZBHW} and $\hat{T}$ in size under Cases 1 and 2 when data are generated from standard normal distribution and gamma distribution. At the same time, our new test $\hat{T}$ has higher empirical powers than T_{ZBHW} and $\rho L_k(y_i)$. For example, as $n = 95$ and $p = 256$, the empirical powers of $\hat{T}$, T_{ZBHW} and $\rho L_k(y_i)$, $i = 1, 2, 3, 4$, are respectively 0.9966, 0.8974, 0.5509, 0.5560, 0.5359 and 0.3469 under Case 1 and normal distribution. Finally, Tables 5–8 show that the six tests T_{ZBHW} , $\rho L_k(y_i)$, $i = 1, 2, 3, 4$, and $\hat{T}$ have similar empirical sizes, which are all close to the nominal size. On the other hand, the six tests still have similar empirical powers under Case 3. However, under Case 4, the empirical powers of $\rho L_k(y_i)$ are extremely deflated, which are no more than 0.1600. Moreover, the empirical powers of our new test are respectively 0.04 and 0.16 more than those of T_{ZBHW} when $n = 45$ and 95.

In summary, our proposed test $\hat{T}$ can control a given size reasonably and has greater powers than competing tests in all cases of our simulation whenever samples are from the normal model or the non-normal model. However, T_{ZLGY} fails in controlling a given size. $\rho L_k(y_i)$, $i = 1, 2, 3, 4$, seriously deflates the empirical powers in some cases of our simulations. T_{ZBHW} can slightly inflate the empirical size in some cases of our simulation.

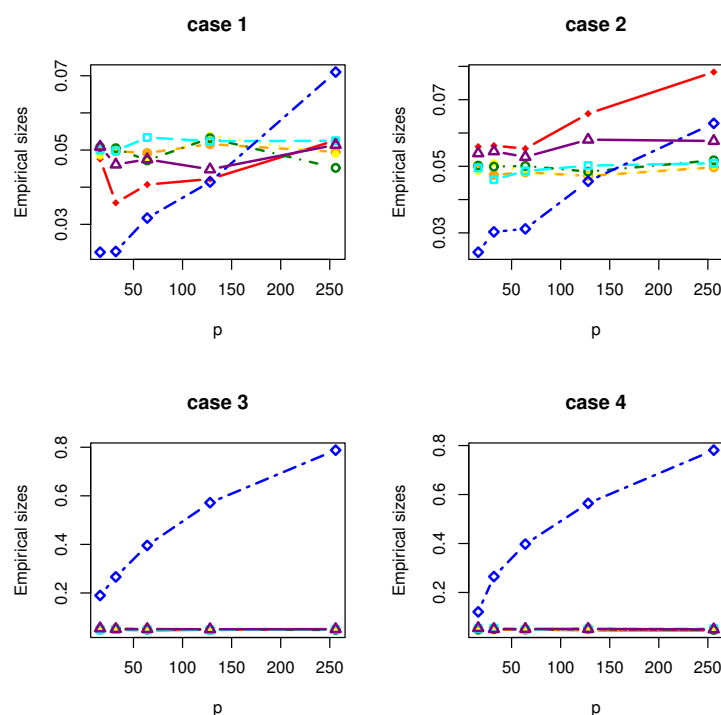


Figure 1. The Empirical sizes of T_{ZBHW} (red), $\rho L_k(y_1)$ (orange), $\rho L_k(y_2)$ (yellow), $\rho L_k(y_3)$ (green), $\rho L_k(y_4)$ (cyan), T_{ZLGY} (blue) and $\hat{T}$ (purple) under normal distribution.

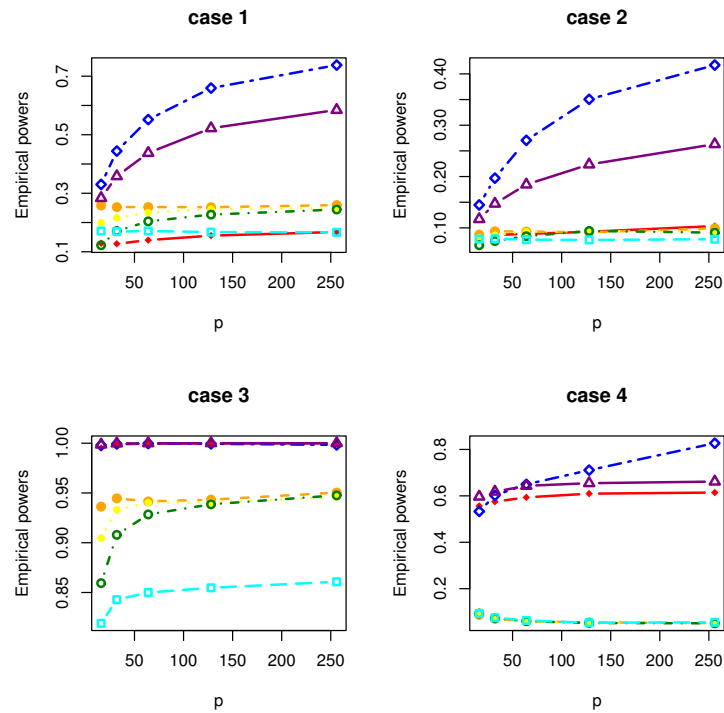


Figure 2. The Empirical powers of T_{ZBHW} (red), $\rho L_k(y_1)$ (orange), $\rho L_k(y_2)$ (yellow), $\rho L_k(y_3)$ (green), $\rho L_k(y_4)$ (cyan), T_{ZLGY} (blue) and $\hat{T}$ (purple) under normal distribution.

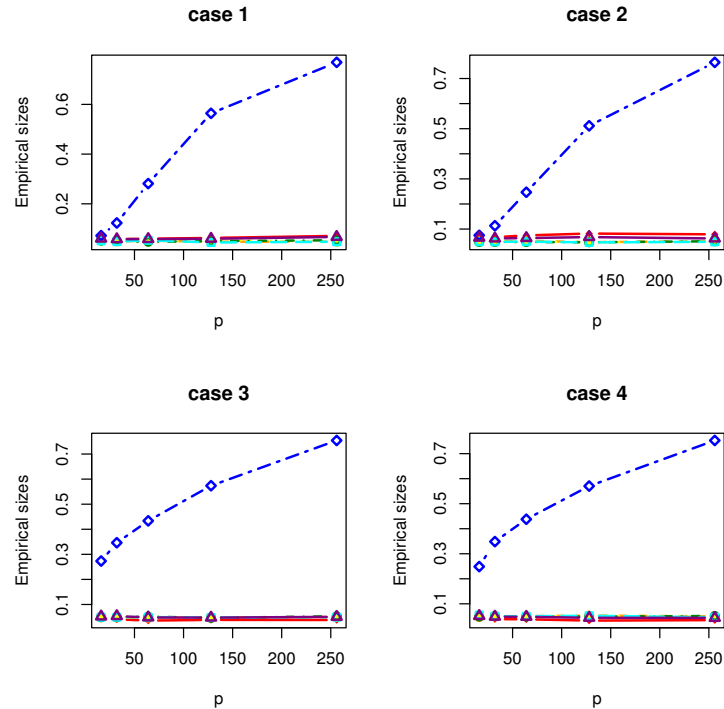


Figure 3. The Empirical sizes of T_{ZBHW} (red), $\rho L_k(y_1)$ (orange), $\rho L_k(y_2)$ (yellow), $\rho L_k(y_3)$ (green), $\rho L_k(y_4)$ (cyan), T_{ZLGY} (blue) and $\hat{T}$ (purple) under normal distribution.

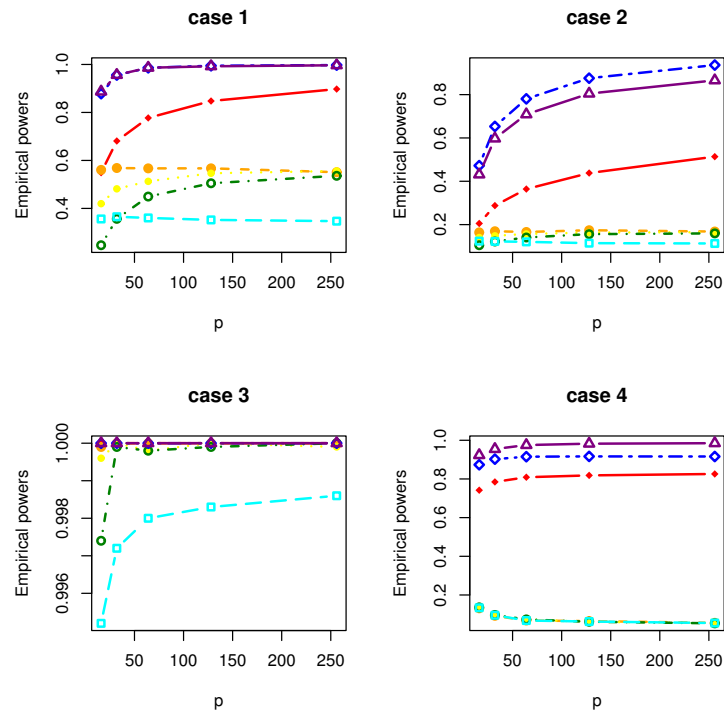


Figure 4. The Empirical powers of T_{ZBHW} (red), $\rho L_k(y_1)$ (orange), $\rho L_k(y_2)$ (yellow), $\rho L_k(y_3)$ (green), $\rho L_k(y_4)$ (cyan), T_{ZLGY} (blue) and $\hat{T}$ (purple) under normal distribution.

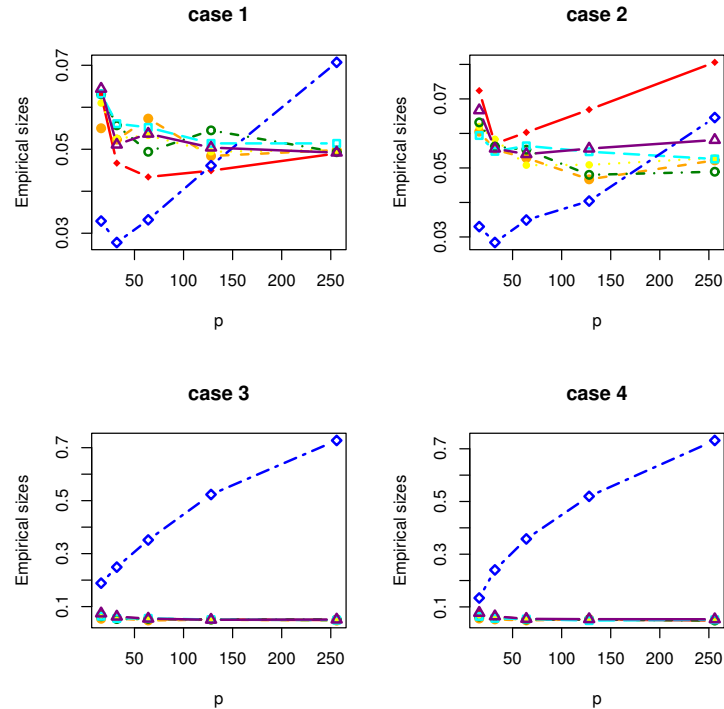


Figure 5. The Empirical sizes of T_{ZBHW} (red), $\rho L_k(y_1)$ (orange), $\rho L_k(y_2)$ (yellow), $\rho L_k(y_3)$ (green), $\rho L_k(y_4)$ (cyan), T_{ZLGY} (blue) and $\hat{T}$ (purple) under Gamma distribution.

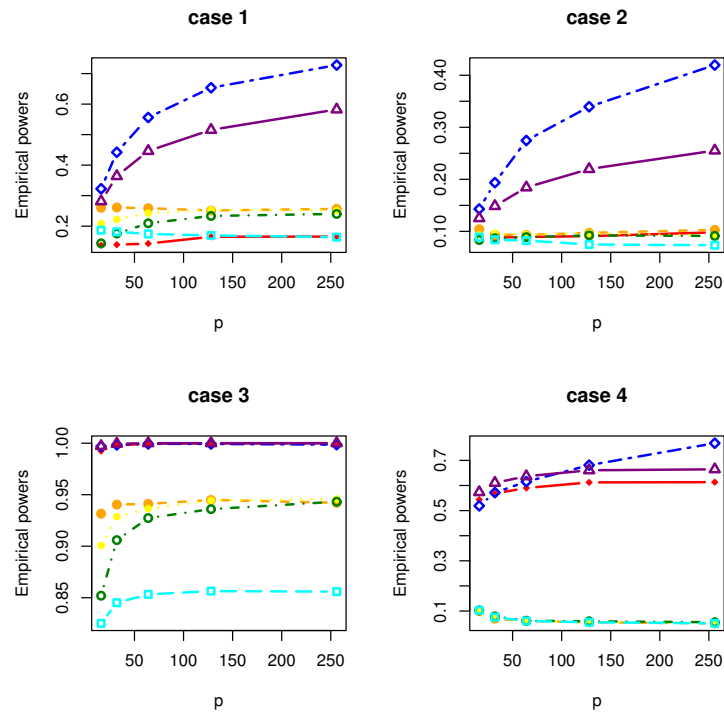


Figure 6. The Empirical powers of T_{ZBHW} (red), $\rho_{L_k}(y_1)$ (orange), $\rho_{L_k}(y_2)$ (yellow), $\rho_{L_k}(y_3)$ (green), $\rho_{L_k}(y_4)$ (cyan), T_{ZLGY} (blue) and $\hat{T}$ (purple) under Gamma distribution.

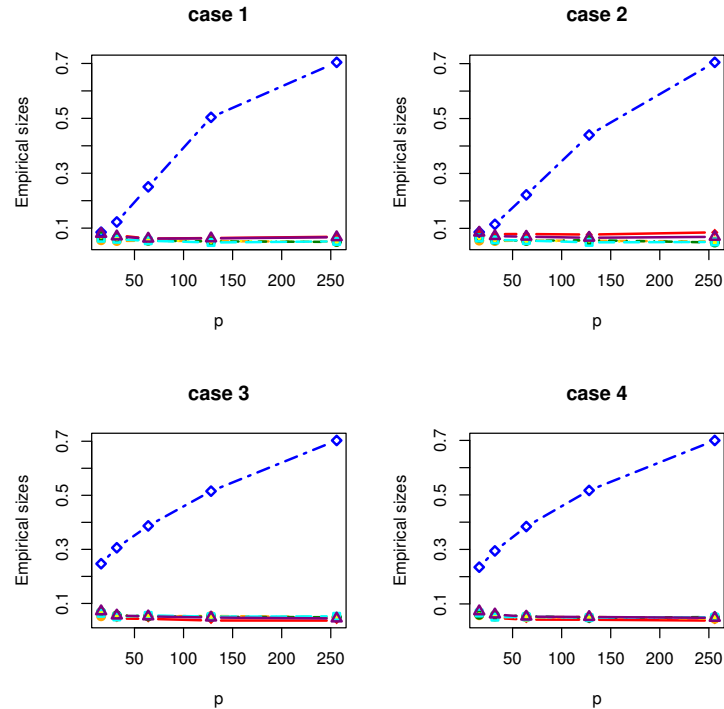


Figure 7. The Empirical sizes of T_{ZBHW} (red), $\rho_{L_k}(y_1)$ (orange), $\rho_{L_k}(y_2)$ (yellow), $\rho_{L_k}(y_3)$ (green), $\rho_{L_k}(y_4)$ (cyan), T_{ZLGY} (blue) and $\hat{T}$ (purple) under Gamma distribution.

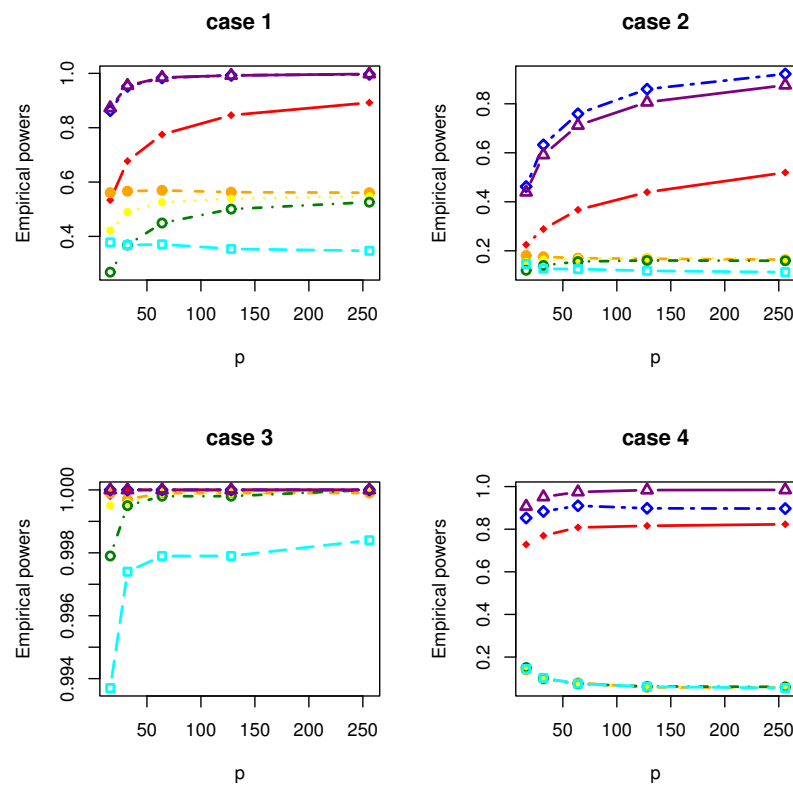


Figure 8. The Empirical powers of T_{ZBHW} (red), $\rho_{L_k}(y_1)$ (orange), $\rho_{L_k}(y_2)$ (yellow), $\rho_{L_k}(y_3)$ (green), $\rho_{L_k}(y_4)$ (cyan), T_{ZLGY} (blue) and $\hat{T}$ (purple) under Gamma distribution.

Table 1. Empirical sizes and powers of seven tests under Case 1 and normal distribution.

	p	T_{ZBHW}	$\rho_{L_k}(y_1)$	$\rho_{L_k}(y_2)$	$\rho_{L_k}(y_3)$	$\rho_{L_k}(y_4)$	T_{ZLGY}	$\hat{T}$
Size $n = 45$	16	0.0476	0.0501	0.0484	0.0508	0.0500	0.0225	0.0509
	32	0.0358	0.0497	0.0507	0.0505	0.0498	0.0227	0.0461
	64	0.0407	0.0492	0.0471	0.0472	0.0534	0.0317	0.0475
	128	0.0422	0.0516	0.0539	0.0532	0.0524	0.0414	0.0448
	256	0.0526	0.0495	0.0489	0.0452	0.0525	0.0710	0.0514
$n = 95$	16	0.0633	0.0524	0.0535	0.0522	0.0512	0.0719	0.0601
	32	0.0584	0.0488	0.0537	0.0515	0.0479	0.1230	0.0576
	64	0.0595	0.0518	0.0499	0.0472	0.0534	0.2814	0.0582
	128	0.0628	0.0477	0.0556	0.0499	0.0445	0.5641	0.0594
	256	0.0710	0.0478	0.0477	0.0534	0.0474	0.7692	0.0669
Power $n = 45$	16	0.1291	0.2582	0.1991	0.1224	0.1710	0.3301	0.2837
	32	0.1273	0.2526	0.2160	0.1709	0.1691	0.4441	0.3582
	64	0.1398	0.2528	0.2337	0.2039	0.1710	0.5516	0.4377
	128	0.1552	0.2523	0.2462	0.2270	0.1671	0.6594	0.5222
	256	0.1676	0.2595	0.2490	0.2444	0.1665	0.7380	0.5847
$n = 95$	16	0.5490	0.5605	0.4198	0.2466	0.3561	0.8783	0.8870
	32	0.6810	0.5684	0.4816	0.3566	0.3653	0.9559	0.9563
	64	0.7771	0.5669	0.5128	0.4494	0.3603	0.9861	0.9863
	128	0.8476	0.5670	0.5465	0.5048	0.3521	0.9946	0.9924
	256	0.8974	0.5509	0.5560	0.5359	0.3469	0.9974	0.9966

Table 2. Empirical sizes and powers of seven tests under Case 1 and Gamma distribution.

	p	T_{ZBHW}	$\rho L_k(y_1)$	$\rho L_k(y_2)$	$\rho L_k(y_3)$	$\rho L_k(y_4)$	T_{ZLGY}	$\hat{T}$
Size $n = 45$	16	0.0630	0.0550	0.0610	0.0631	0.0632	0.0329	0.0644
	32	0.0467	0.0523	0.0526	0.0557	0.0561	0.0278	0.0511
	64	0.0434	0.0573	0.0533	0.0494	0.0552	0.0332	0.0537
	128	0.0449	0.0484	0.0494	0.0545	0.0514	0.0461	0.0504
	256	0.0490	0.0498	0.0499	0.0493	0.0514	0.0707	0.0492
$n = 95$	16	0.0822	0.0560	0.0657	0.0699	0.0631	0.0851	0.0789
	32	0.0741	0.0531	0.0610	0.0629	0.0594	0.1224	0.0698
	64	0.0619	0.0544	0.0535	0.0545	0.0540	0.2505	0.0617
	128	0.0638	0.0524	0.0540	0.0517	0.0477	0.5037	0.0627
	256	0.0686	0.0508	0.0493	0.0498	0.0517	0.7041	0.0668
Power $n = 45$	16	0.1375	0.2602	0.2085	0.1437	0.1866	0.3225	0.2815
	32	0.1392	0.2613	0.2221	0.1760	0.1815	0.4421	0.3639
	64	0.1429	0.2587	0.2427	0.2094	0.1746	0.5561	0.4463
	128	0.1652	0.2513	0.2538	0.2334	0.1695	0.6536	0.5155
	256	0.1658	0.2565	0.2514	0.2401	0.1639	0.7279	0.5824
$n = 95$	16	0.5336	0.5610	0.4219	0.2685	0.3780	0.8629	0.8730
	32	0.6772	0.5664	0.4895	0.3683	0.3681	0.9523	0.9551
	64	0.7749	0.5695	0.5256	0.4493	0.3709	0.9834	0.9849
	128	0.8460	0.5628	0.5391	0.5005	0.3540	0.9929	0.9923
	256	0.8922	0.5608	0.5477	0.5260	0.3470	0.9966	0.9978

Table 3. Empirical sizes and powers of seven tests in Case 2 and normal distribution.

	p	T_{ZBHW}	$\rho L_k(y_1)$	$\rho L_k(y_2)$	$\rho L_k(y_3)$	$\rho L_k(y_4)$	T_{ZLGY}	$\hat{T}$
Size $n = 45$	16	0.0559	0.0501	0.0485	0.0502	0.0495	0.0242	0.0539
	32	0.0562	0.0474	0.0510	0.0499	0.0460	0.0303	0.0545
	64	0.0553	0.0482	0.0483	0.0501	0.0486	0.0312	0.0528
	128	0.0658	0.0471	0.0472	0.0484	0.0502	0.0455	0.0580
	256	0.0783	0.0497	0.0500	0.0518	0.0509	0.0629	0.0576
$n = 95$	16	0.0724	0.0511	0.0499	0.0484	0.0494	0.0750	0.0656
	32	0.0675	0.0492	0.0497	0.0486	0.0491	0.1132	0.0624
	64	0.0747	0.0532	0.0505	0.0483	0.0498	0.2468	0.0641
	128	0.0821	0.0479	0.0465	0.0473	0.0468	0.5114	0.0680
	256	0.0790	0.0514	0.0488	0.0514	0.0492	0.7645	0.0625
Power $n = 45$	16	0.0838	0.0869	0.0804	0.0662	0.0781	0.1447	0.1170
	32	0.0862	0.0933	0.0858	0.0743	0.0788	0.1969	0.1471
	64	0.0883	0.0928	0.0923	0.0838	0.0771	0.2706	0.1844
	128	0.0924	0.0915	0.0933	0.0940	0.0765	0.3505	0.2235
	256	0.1035	0.0989	0.0937	0.0907	0.0779	0.4172	0.2631
$n = 95$	16	0.2053	0.1640	0.1405	0.1048	0.1238	0.4722	0.4314
	32	0.2877	0.1693	0.1482	0.1220	0.1225	0.6532	0.5974
	64	0.3641	0.1659	0.1569	0.1405	0.1203	0.7805	0.7090
	128	0.4380	0.1740	0.1632	0.1559	0.1143	0.8758	0.8046
	256	0.5136	0.1680	0.1651	0.1595	0.1129	0.9364	0.8659

Table 4. Empirical sizes and powers of seven tests in Case 2 and Gamma distribution.

	p	T_{ZBHW}	$\rho L_k(y_1)$	$\rho L_k(y_2)$	$\rho L_k(y_3)$	$\rho L_k(y_4)$	T_{ZLGY}	$\hat{T}$
Size $n = 45$	16	0.0724	0.0605	0.0620	0.0632	0.0594	0.0330	0.0667
	32	0.0570	0.0553	0.0583	0.0562	0.0548	0.0284	0.0556
	64	0.0603	0.0529	0.0508	0.0553	0.0564	0.0349	0.0540
	128	0.0669	0.0467	0.0509	0.0480	0.0547	0.0404	0.0556
	256	0.0806	0.0522	0.0529	0.0489	0.0526	0.0646	0.0581
$n = 95$	16	0.0886	0.0556	0.0653	0.0660	0.0649	0.0865	0.0844
	32	0.0792	0.0539	0.0554	0.0580	0.0580	0.1146	0.0718
	64	0.0794	0.0557	0.0550	0.0545	0.0555	0.2220	0.0696
	128	0.0774	0.0549	0.0524	0.0570	0.0490	0.4400	0.0658
	256	0.0855	0.0523	0.0504	0.0486	0.0505	0.7047	0.0691
Power $n = 45$	16	0.0976	0.1041	0.0937	0.0834	0.0889	0.1430	0.1253
	32	0.0882	0.0934	0.0961	0.0870	0.0839	0.1936	0.1486
	64	0.0893	0.0931	0.0942	0.0890	0.0823	0.2746	0.1844
	128	0.0908	0.0972	0.0965	0.0922	0.0748	0.3395	0.2197
	256	0.0982	0.1029	0.0918	0.0913	0.0734	0.4195	0.2553
$n = 95$	16	0.2248	0.1807	0.1547	0.1205	0.1464	0.4621	0.4391
	32	0.2886	0.1757	0.1641	0.1403	0.1267	0.6325	0.5913
	64	0.3671	0.1699	0.1681	0.1560	0.1253	0.7592	0.7119
	128	0.4392	0.1675	0.1715	0.1605	0.1184	0.8598	0.8064
	256	0.5194	0.1641	0.1682	0.1593	0.1129	0.9217	0.8760

Table 5. Empirical sizes and powers of seven tests in Case 3 and normal distribution.

	p	T_{ZBHW}	$\rho L_k(y_1)$	$\rho L_k(y_2)$	$\rho L_k(y_3)$	$\rho L_k(y_4)$	T_{ZLGY}	$\hat{T}$
Size $n = 45$	16	0.0530	0.0517	0.0478	0.0495	0.0484	0.1897	0.0545
	32	0.0491	0.0539	0.0513	0.0514	0.0503	0.2663	0.0517
	64	0.0463	0.0521	0.0498	0.0487	0.0491	0.3956	0.0513
	128	0.0490	0.0481	0.0503	0.0496	0.0475	0.5715	0.0508
	256	0.0492	0.0512	0.0491	0.0478	0.0501	0.7879	0.0511
$n = 95$	16	0.0455	0.0489	0.0516	0.0493	0.0496	0.2732	0.0520
	32	0.0407	0.0484	0.0508	0.0493	0.0493	0.3462	0.0535
	64	0.0357	0.0471	0.0501	0.0516	0.0496	0.4335	0.0483
	128	0.0386	0.0483	0.0496	0.0465	0.0483	0.5736	0.0476
	256	0.0381	0.0497	0.0501	0.0532	0.0494	0.7541	0.0506
Power $n = 45$	16	0.9951	0.9363	0.9047	0.8594	0.8190	0.9981	0.9986
	32	0.9988	0.9446	0.9331	0.9079	0.8428	0.9993	0.9997
	64	0.9996	0.9415	0.9403	0.9284	0.8500	0.9997	0.9999
	128	0.9998	0.9433	0.9404	0.9386	0.8547	0.9994	1.0000
	256	0.9999	0.9504	0.9458	0.9476	0.8608	0.9983	1.0000
$n = 95$	16	1.0000	0.9999	0.9996	0.9974	0.9952	1.0000	1.0000
	32	1.0000	1.0000	1.0000	0.9999	0.9972	1.0000	1.0000
	64	1.0000	1.0000	0.9998	0.9998	0.9980	1.0000	1.0000
	128	1.0000	1.0000	1.0000	0.9999	0.9983	1.0000	1.0000
	256	1.0000	1.0000	0.9999	1.0000	0.9986	1.0000	1.0000

Table 6. Empirical sizes and powers of seven tests in Case 3 and Gamma distribution.

	p	T_{ZBHW}	$\rho L_k(y_1)$	$\rho L_k(y_2)$	$\rho L_k(y_3)$	$\rho L_k(y_4)$	T_{ZLGY}	$\hat{T}$
Size $n = 45$	16	0.0743	0.0540	0.0587	0.0633	0.0637	0.1885	0.0740
	32	0.0599	0.0553	0.0544	0.0530	0.0551	0.2490	0.0622
	64	0.0532	0.0476	0.0512	0.0536	0.0557	0.3514	0.0544
	128	0.0506	0.0508	0.0494	0.0511	0.0511	0.5232	0.0503
	256	0.0493	0.0471	0.0533	0.0516	0.0487	0.7273	0.0510
$n = 95$	16	0.0619	0.0526	0.0614	0.0679	0.0635	0.2469	0.0709
	32	0.0441	0.0513	0.0565	0.0562	0.0501	0.3059	0.0553
	64	0.0430	0.0526	0.0537	0.0524	0.0558	0.3869	0.0521
	128	0.0368	0.0534	0.0500	0.0506	0.0511	0.5153	0.0478
	256	0.0365	0.0494	0.0463	0.0489	0.0523	0.7029	0.0445
Power $n = 45$	16	0.9924	0.9316	0.9007	0.8519	0.8251	0.9968	0.9972
	32	0.9978	0.9404	0.9286	0.9059	0.8451	0.9984	0.9994
	64	0.9996	0.9411	0.9365	0.9273	0.8532	0.9994	1.0000
	128	0.9999	0.9449	0.9444	0.9360	0.8564	0.9993	1.0000
	256	1.0000	0.9421	0.9459	0.9432	0.8559	0.9986	1.0000
$n = 95$	16	0.9998	0.9999	0.9995	0.9979	0.9937	1.0000	1.0000
	32	1.0000	0.9997	0.9996	0.9995	0.9974	1.0000	1.0000
	64	1.0000	0.9999	0.9999	0.9998	0.9979	1.0000	1.0000
	128	1.0000	0.9999	0.9999	0.9998	0.9979	1.0000	1.0000
	256	1.0000	0.9999	1.0000	1.0000	0.9984	1.0000	1.0000

Table 7. Empirical sizes and powers of seven tests in Case 4 and normal distribution.

	p	T_{ZBHW}	$\rho L_k(y_1)$	$\rho L_k(y_2)$	$\rho L_k(y_3)$	$\rho L_k(y_4)$	T_{ZLGY}	$\hat{T}$
Size $n = 45$	16	0.0517	0.0490	0.0514	0.0484	0.0522	0.1209	0.0548
	32	0.0481	0.0496	0.0506	0.0507	0.0541	0.2649	0.0507
	64	0.0487	0.0480	0.0495	0.0493	0.0487	0.3975	0.0512
	128	0.0459	0.0496	0.0459	0.0515	0.0510	0.5640	0.0511
	256	0.0455	0.0504	0.0483	0.0469	0.0520	0.7811	0.0493
$n = 95$	16	0.0450	0.0470	0.0522	0.0489	0.0550	0.2487	0.0513
	32	0.0388	0.0504	0.0511	0.0514	0.0502	0.3486	0.0484
	64	0.0381	0.0504	0.0503	0.0488	0.0502	0.4379	0.0480
	128	0.0329	0.0530	0.0511	0.0487	0.0523	0.5704	0.0441
	256	0.0348	0.0498	0.0508	0.0514	0.0462	0.7527	0.0438
Power $n = 45$	16	0.5566	0.0872	0.0922	0.0930	0.0933	0.5331	0.5965
	32	0.5749	0.0710	0.0698	0.0724	0.0749	0.6018	0.6180
	64	0.5935	0.0596	0.0591	0.0601	0.0633	0.6495	0.6432
	128	0.6095	0.0547	0.0507	0.0522	0.0532	0.7108	0.6546
	256	0.6142	0.0516	0.0527	0.0508	0.0552	0.8269	0.6615
$n = 95$	16	0.7418	0.1311	0.1361	0.1359	0.1338	0.8735	0.9232
	32	0.7846	0.0923	0.0916	0.0966	0.0953	0.9024	0.9548
	64	0.8089	0.0680	0.0738	0.0735	0.0679	0.9151	0.9753
	128	0.8184	0.0650	0.0607	0.0615	0.0626	0.9166	0.9824
	256	0.8258	0.0560	0.0548	0.0533	0.0549	0.9161	0.9849

Table 8. Empirical sizes and powers of seven tests in Case 4 and Gamma distribution.

	p	T_{ZBHW}	$\rho L_k(y_1)$	$\rho L_k(y_2)$	$\rho L_k(y_3)$	$\rho L_k(y_4)$	T_{ZLGY}	$\hat{T}$
Size $n = 45$	16	0.0776	0.0559	0.0589	0.0632	0.0641	0.1338	0.0773
	32	0.0634	0.0526	0.0534	0.0584	0.0567	0.2403	0.0642
	64	0.0532	0.0488	0.0527	0.0527	0.0540	0.3579	0.0546
	128	0.0520	0.0533	0.0500	0.0517	0.0480	0.5195	0.0534
	256	0.0483	0.0495	0.0506	0.0482	0.0512	0.7316	0.0530
$n = 95$	16	0.0674	0.0586	0.0605	0.0605	0.0668	0.2348	0.0729
	32	0.0482	0.0528	0.0530	0.0586	0.0514	0.2945	0.0613
	64	0.0425	0.0511	0.0491	0.0551	0.0525	0.3839	0.0533
	128	0.0420	0.0523	0.0522	0.0491	0.0499	0.5164	0.0520
	256	0.0388	0.0456	0.0488	0.0519	0.0511	0.6998	0.0490
Power $n = 45$	16	0.5435	0.0993	0.1039	0.1028	0.1026	0.5191	0.5735
	32	0.5685	0.0697	0.0741	0.0787	0.0756	0.5717	0.6106
	64	0.5901	0.0606	0.0601	0.0615	0.0602	0.6153	0.6368
	128	0.6124	0.0551	0.0552	0.0592	0.0551	0.6810	0.6606
	256	0.6133	0.0511	0.0512	0.0554	0.0510	0.7685	0.6645
$n = 95$	16	0.7283	0.1401	0.1468	0.1506	0.1438	0.8526	0.9070
	32	0.7695	0.0973	0.0997	0.1000	0.1019	0.8832	0.9513
	64	0.8081	0.0790	0.0745	0.0738	0.0749	0.9100	0.9746
	128	0.8158	0.0577	0.0627	0.0621	0.0611	0.8980	0.9835
	256	0.8231	0.0570	0.0545	0.0601	0.0550	0.8966	0.9843

5. Real Data Analysis

This problem will arise from the analysis of gene expression data. It was shown in [3] that many genes have different variances in gene expressions between disease states. For example, the dataset of the breast cancer of patient is classified into three groups based on their gene expression signatures: well-differentiated tumors ($n_1 = 29$), moderately differentiated tumors ($n_2 = 136$), and poorly differentiated and undifferentiated tumors ($n_3 = 35$). The breast cancer microarray data sets, including patient outcome information, were downloaded from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) data repository and are accessible through GEO Series accession now. GSE11121 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE11121>, accessed on 26 June 2022). This dataset includes 200 samples, which is far fewer than data dimension 22,283. For this gene dataset, we are interested in checking whether the covariance matrices of the three groups are the same or not. It is noted that an important assumption for multivariate analysis of variance is that of equal covariance matrices in different groups. Therefore, testing the equality of several covariance matrices is necessary.

Since we used the raw breast cancer dataset (22,283 features), there are many features and no preprocessing. First, we preprocessed and filtered the data, including conventional preprocessing such as background adjustment, normalization, and summarization. Then, we performed feature screening, filtered out features whose coefficient of variation is out of range (0.25, 1.0) and controlled at least 5 samples to exceed 1320 count values. Finally, a data set of 200 samples and 1280 features was screened out, and then a high-dimensional hypothesis test problem was performed on the screened data set.

Compared with other methods and our method, the observed test statistics (p value) of the screened breast cancer dataset are $T_{ZBHW}(1.92, 2.75 \times 10^{-2})$, $\rho L_k(y_1)(0.46, 0.79)$, $\rho L_k(y_2)(0.4, 0.82)$, $\rho L_k(y_3)(0.91, 0.63)$, $\rho L_k(y_4)(0.27, 0.69)$, $T_{ZLGY}(2.47, 6.62 \times 10^{-3})$ and $\hat{T}(1.96, 2.52 \times 10^{-2})$. The p -values of the comparison method statistics T_{ZBHW} , T_{ZLGY} and our method's statistic are significant.

6. Concluding Remarks

In this paper, we propose a new test on the homogeneity of k -sample covariance matrices in a high-dimensional setting. The asymptotic properties of the proposed test are derived under some regularity conditions. We compare our new test with six competing

tests by simulation. Numerical results show that our proposed test can control the nominal size reasonably and that it has the highest empirical powers in our simulation scenario. However, just as what we showed in Section 2, the technique condition (C2) may not hold again when covariance matrices have some spiked eigenvalues. How to obtain the theoretical results of test statistics under spiked covariance matrices structure is an interested problem. We leave this problem as a future research direction.

Author Contributions: Conceptualization, P.S., Y.T. and M.C.; Simulation studies, P.S.; Formal analysis, P.S.; Funding acquisition, Y.T.; Investigation, P.S.; Methodology, P.S. and M.C.; Supervision, Y.T.; Software, P.S.; Visualization, P.S.; Validation, P.S. and M.C.; Writing—original draft, P.S.; Writing—review & editing, P.S. and M.C. All authors have read and agreed to the published version of the manuscript.

Funding: Tang's research is supported by the Natural Science Foundation of China (No. 12271168) and the 111 Project of China (No. B14019). Cao's research is supported by the Humanities and Social Sciences Foundation of Ministry of Education (No. 22YJC910001), the Natural Science Foundation of Anhui Province (No. 2108085MA09), the Foundation for Excellent Young Talents in College of Anhui Province (No. gxyqZD2021092) and the Program for Mathematical Statistics Research Team of Anhui Province (No. 2020jxtd102).

Data Availability Statement: Gene expression data generated for this manuscript is available to download and explore at <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE11121>, accessed on 26 June 2022. All raw data are available in GEO database under the accession number GEO: GSE11121.

Conflicts of Interest: On behalf of the authors, the corresponding author states that there is no conflict of interest.

Appendix A. Some Lemmas

As mentioned in Section 3, we can assume all the population means are 0. In this case, we only need to consider respectively the first terms $A_i^* = \frac{1}{(n_i)_2} \sum_{j,l}^* (\mathbf{X}_{ij}^T \mathbf{X}_{il})^2$ and $C_{ij}^* = \frac{1}{n_i n_j} \sum_{l=1}^{n_i} \sum_{m=1}^{n_j} (\mathbf{X}_{il}^T \mathbf{X}_{jm})^2$ of A_i and C_{ij} in (5) and (6). Let $T_1 := \sum_{i=1}^k \frac{n_i(n-n_i)}{n} A_i^* - \sum_{i \neq j}^k \frac{n_i n_j}{n} C_{ij}^*$, then we only need to prove the asymptotic normality of T_1 . To do this, we define $\mathbf{Z}_l = \mathbf{Z}_{si}$, where $l = i + \sum_{j=1}^{s-1} n_j$ and $\sum_{j=1}^0 n_j = 0$ for $i = 1, \dots, n_s$ and $s = 1, \dots, k$. A sequence of increasing σ -fields is defined as $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and $\mathcal{F}_j = \sigma\{\mathbf{Z}_1, \dots, \mathbf{Z}_j\}$ for $j = 1, \dots, n$. Let $E_j(\cdot)$ denote the conditional expectation with respect to $\mathcal{F}_j$ and $D_j := (E_j - E_{j-1})T_1$. It is easy to obtain $T_1 - E(T_1) = \sum_{j=1}^n D_j$.

Lemma A1. For any n , $\{D_j, \mathcal{F}_j\}_{j=1}^n$ is a square integrable martingale difference sequence.

Proof. The conclusion is obvious. Hence, the proof is omitted. $\square$

Therefore, to prove Theorem 1, we next apply the martingale central limit theorem.

Lemma A2. Under conditions (C1) and (C2), as $n, p \rightarrow \infty$,

$$\frac{\sum_{j=1}^n E(D_j^2 | \mathcal{F}_{j-1})}{\text{Var}(T_1)} \xrightarrow{p} 1.$$

Proof. By some calculations, we have

$$\begin{aligned}
 \sum_{j=1}^n E(D_j^2 | \mathcal{F}_{j-1}) &= \sum_{j=1}^n E_{j-1} [E_j(T_1) - E_{j-1}(T_1)]^2 \\
 &= \sum_{l=1}^k \sum_{j=1}^{n_l} E_{\Sigma_{i=1}^{l-1} n_i + j-1} \left[E_{\Sigma_{i=1}^{l-1} n_i + j} \left(\frac{n - n_l}{n} A_l^* - \sum_{h \neq l}^k \frac{n_h n_l}{n} C_{lh}^* \right) \right. \\
 &\quad \left. - E_{\Sigma_{i=1}^{l-1} n_i + j-1} \left(\frac{n - n_l}{n} A_l^* - \sum_{h \neq l}^k \frac{n_h n_l}{n} C_{lh}^* \right) \right]^2 \\
 &= \sum_{l=1}^k \sum_{j=1}^{n_l} E_{\Sigma_{i=1}^{l-1} n_i + j-1} \left[\frac{n - n_l}{n(n_l)^2} \sum_{i < j}^{n_l} \text{tr} \left[(\mathbf{X}_{li} \mathbf{X}_{li}^T - \Sigma_l)(\mathbf{X}_{lj} \mathbf{X}_{lj}^T - \Sigma_l) \right] \right. \\
 &\quad \left. + \frac{n - n_l}{n n_l} \text{tr} \left[(\mathbf{X}_{lj} \mathbf{X}_{lj}^T - \Sigma_l) \Sigma_l \right] - \frac{n_h}{n} \sum_{h > l}^k \text{tr} \left[(\mathbf{X}_{lj} \mathbf{X}_{lj}^T - \Sigma_l) \Sigma_h \right] \right. \\
 &\quad \left. - \frac{1}{n} \sum_{h < l}^k \sum_{i=1}^{n_h} (\mathbf{X}_{hi}^T (\mathbf{X}_{lj} \mathbf{X}_{lj}^T - \Sigma_l) \mathbf{X}_{hi}) \right]^2 \\
 &= \sum_{l=1}^k \frac{1}{n^2 n_l^2} \sum_{j=1}^{n_l} E_{\Sigma_{i=1}^{l-1} n_i + j-1} \left[\text{tr}((\mathbf{X}_{lj} \mathbf{X}_{lj}^T - \Sigma_l) \mathbf{Q}_{lj}) \right]^2 \\
 &= \sum_{l=1}^k \frac{1}{n^2 n_l^2} \sum_{j=1}^{n_l} \left\{ 2 \text{tr}(\mathbf{Q}_{lj} \Sigma_l)^2 + \Delta_l \text{tr}(\Gamma_l^T \mathbf{Q}_{lj} \Gamma_l \circ \Gamma_l^T \mathbf{Q}_{lj} \Gamma_l) \right\},
 \end{aligned}$$

where

$$\begin{aligned}
 \mathbf{Q}_{lj} &= \frac{n - n_l}{n_l - 1} \sum_{i < j}^{n_l} (\mathbf{X}_{li} \mathbf{X}_{li}^T - \Sigma_l) + (n - n_l) \Sigma_l + n_l \sum_{h \neq l}^k n_h \Sigma_h - n_l \sum_{h < l}^k \sum_i^{n_h} (\mathbf{X}_{hi} \mathbf{X}_{hi}^T - \Sigma_h) \\
 &=: \mathbf{Q}_{lj1} + \mathbf{Q}_{lj2} + \mathbf{Q}_{lj3} + \mathbf{Q}_{lj4}.
 \end{aligned}$$

Note that $E(\sum_{j=1}^n E(D_j^2 | \mathcal{F}_{j-1})) = \text{Var}(T_1)$. Next, we prove

$$\frac{\text{Var} \left(\sum_{j=1}^n E(D_j^2 | \mathcal{F}_{j-1}) \right)}{\text{Var}^2(T_1)} \rightarrow 0.$$

Note that

$$\begin{aligned}
 \text{tr}(\mathbf{Q}_{lj} \Sigma_l)^2 &= \text{tr}(\mathbf{Q}_{lj1} \Sigma_l)^2 + \text{tr}(\mathbf{Q}_{lj2} \Sigma_l)^2 + \text{tr}(\mathbf{Q}_{lj3} \Sigma_l)^2 + \text{tr}(\mathbf{Q}_{lj4} \Sigma_l)^2 \\
 &\quad + 2 \text{tr} \left[(\mathbf{Q}_{lj1} \Sigma_l) (\mathbf{Q}_{lj2} \Sigma_l) \right] + 2 \text{tr} \left[(\mathbf{Q}_{lj1} \Sigma_l) (\mathbf{Q}_{lj3} \Sigma_l) \right] + 2 \text{tr} \left[(\mathbf{Q}_{lj1} \Sigma_l) (\mathbf{Q}_{lj4} \Sigma_l) \right] \\
 &\quad + 2 \text{tr} \left[(\mathbf{Q}_{lj2} \Sigma_l) (\mathbf{Q}_{lj3} \Sigma_l) \right] + 2 \text{tr} \left[(\mathbf{Q}_{lj2} \Sigma_l) (\mathbf{Q}_{lj4} \Sigma_l) \right] + 2 \text{tr} \left[(\mathbf{Q}_{lj3} \Sigma_l) (\mathbf{Q}_{lj4} \Sigma_l) \right].
 \end{aligned}$$

It then follows that

$$\text{Var} \left(\sum_{l=1}^k \frac{1}{n^2 n_l^2} \sum_{j=1}^{n_l} \text{tr}(\mathbf{Q}_{lj} \Sigma_l)^2 \right) \leq 16 \sum_{s,t=1}^4 R_{st},$$

where $R_{st} = \text{Var}\left(\sum_{l=1}^k \frac{1}{n^2 n_l^2} \sum_{j=1}^{n_l} \text{tr}\left[E_{\sum_{i=1}^{l-1} n_i + j} \mathbf{Q}_{ljs} \boldsymbol{\Sigma}_l E_{\sum_{i=1}^{l-1} n_i + j} \mathbf{Q}_{ljt} \boldsymbol{\Sigma}_l\right]\right)$. In the following, we will prove $R_{st} = o\left[\text{Var}^2(T_1)\right]$ for $s, t = 1, 2, 3$ and 4.

$$\begin{aligned} R_{11} &= \text{Var}\left(\sum_{l=1}^k \frac{(n - n_l)^2}{n^2(n_l - 1)^2 n_l^2} \sum_{j_1, j_2=1}^{n_l} (n_l - j_1 \vee j_2) \text{tr}[(\mathbf{X}_{lj_1} \mathbf{X}_{lj_1}^T - \boldsymbol{\Sigma}_l)(\mathbf{X}_{lj_2} \mathbf{X}_{lj_2}^T - \boldsymbol{\Sigma}_l)]\right) \\ &= \sum_{l=1}^k \sum_{j_1, j_2=1}^{n_l} \frac{(n - n_l)^4 (n_l - j_1 \vee j_2)^2}{n^4 (n_l - 1)^4 n_l^4} \text{Var}\left(\text{tr}[(\mathbf{X}_{lj_1} \mathbf{X}_{lj_1}^T - \boldsymbol{\Sigma}_l)(\mathbf{X}_{lj_2} \mathbf{X}_{lj_2}^T - \boldsymbol{\Sigma}_l)]\right) \\ &= \sum_{l=1}^k \sum_{j_1=j_2=j}^{n_l} \frac{(n - n_l)^4 (n_l - j)^2}{n^4 (n_l)^4} \text{Var}\left(\text{tr}[(\mathbf{X}_{lj} \mathbf{X}_{lj}^T - \boldsymbol{\Sigma}_l)(\mathbf{X}_{lj} \mathbf{X}_{lj}^T - \boldsymbol{\Sigma}_l)]\right) \\ &\quad + \sum_{l=1}^k \sum_{j_1 \neq j_2}^{n_l} \frac{(n - n_l)^4 (n_l - j_1 \vee j_2)^2}{n^4 (n_l - 1)^4 n_l^4} \text{Var}\left(\text{tr}[(\mathbf{X}_{lj_1} \mathbf{X}_{lj_1}^T - \boldsymbol{\Sigma}_l)(\mathbf{X}_{lj_2} \mathbf{X}_{lj_2}^T - \boldsymbol{\Sigma}_l)]\right) \\ &\leq \sum_{l=1}^k \frac{1}{(n_l)^2} \text{tr}^4(\boldsymbol{\Sigma}_l^2) \left[o(1) + O\left(\frac{1}{n_l}\right)\right] = o\left[\text{Var}^2(T_1)\right]. \end{aligned}$$

Similarly,

$$\begin{aligned} R_{12} &\leq \sum_{l=1}^k \sum_{j=1}^{n_l} \frac{3(n - n_l)^2 (n_l - j)^2}{n^4 n_l^4 (n_l - 1)^2} \text{tr}\left[\left(\boldsymbol{\Sigma}_l \mathbf{Q}_{lj2} \boldsymbol{\Sigma}_l^2\right)^2\right] = \sum_{l=1}^k \frac{(n - n_l)^4 (2n_l - 1)}{2n^4 n_l^3 (n_l - 1)} \text{tr}[(\boldsymbol{\Sigma}_l^4)^2] \\ &\leq \sum_{l=1}^k \frac{(n - n_l)^4 (2n_l - 1)}{2n^4 n_l^3 (n_l - 1)} \text{tr}^2(\boldsymbol{\Sigma}_l^4) = o\left[\text{Var}^2(T_1)\right]. \end{aligned}$$

$$\begin{aligned} R_{13} &\leq \sum_{l=1}^k \sum_{j=1}^{n_l} \frac{3(n - n_l)^2 (n_l - j)^2}{n^4 n_l^4 (n_l - 1)^2} \text{tr}\left[\left(\boldsymbol{\Sigma}_l \mathbf{Q}_{lj3} \boldsymbol{\Sigma}_l^2\right)^2\right] \\ &= \sum_{l=1}^k \frac{(n - n_l)^2 (2n_l - 1)}{2n^4 n_l (n_l - 1)} \text{tr}\left(\boldsymbol{\Sigma}_l^3 n_l \sum_{h \neq l}^k n_h \boldsymbol{\Sigma}_h\right)^2 \\ &\leq \sum_{l=1}^k \frac{(n - n_l)^2 (2n_l - 1)}{2n^4 n_l (n_l - 1)} \text{tr}(\boldsymbol{\Sigma}_l^4) \text{tr}\left(\boldsymbol{\Sigma}_l \sum_{h \neq l}^k n_h \boldsymbol{\Sigma}_h\right)^2 = o\left[\text{Var}^2(T_1)\right]. \end{aligned}$$

We can obtain $R_{i4} = o\left[\text{Var}^2(T_1)\right]$ for $i = 2, 3$ and 4 by a similar method. It is noted that $R_{23} = R_{22} = R_{33} = 0$ since $\mathbf{Q}_{lj2}$ and $\mathbf{Q}_{lj3}$ are non random.

As a result, it follows from the above equalities that

$$\text{Var}\left(\sum_{l=1}^k \frac{1}{n^2 n_l^2} \sum_{j=1}^{n_l} 2\text{tr}(\mathbf{Q}_{lj} \boldsymbol{\Sigma}_l)^2\right) = o\left[\text{Var}^2(T_1)\right].$$

Finally, using similar calculations, we can obtain

$$\text{Var}\left(\sum_{l=1}^k \frac{1}{n^2 n_l^2} \sum_{j=1}^{n_l} \text{tr}(\boldsymbol{\Gamma}_l^T \mathbf{Q}_{lj} \boldsymbol{\Gamma}_l \circ \boldsymbol{\Gamma}_l^T \mathbf{Q}_{lj} \boldsymbol{\Gamma}_l)\right) = o\left[\text{Var}^2(T_1)\right].$$

This completes the proof of Lemma A2. $\square$

Lemma A3. Under the condition (C2), as $n, p \rightarrow \infty$,

$$\sum_{j=1}^n E(D_j^4) = o\left[\text{Var}^2(T_1)\right].$$

Proof. By some calculations, for some constants c_1, c_2, c_3 and c_4 , we can obtain

$$\sum_{j=1}^n E(D_j^4) \leq c_1 \sum_{l=1}^k \frac{1}{n_l^4} \text{tr}^4(\Sigma_l^2) + c_2 \sum_{l \neq h}^k \frac{n_h}{n_l^2} \text{tr}^4(\Sigma_l \Sigma_h)$$

and

$$\text{Var}^2(T_1) \geq c_3 \sum_{l=1}^k \frac{n_l^2(n - n_l)^4}{n^4(n_l - 1)^2} \text{tr}^4(\Sigma_l^2) + c_4 \sum_{l \neq h}^k \frac{n_l^2 n_h^2}{n^4} \text{tr}^4(\Sigma_l \Sigma_h).$$

As a result,

$$\frac{\sum_{j=1}^n E(D_j^4)}{\text{Var}^2(T_1)} \leq \sum_{l=1}^k \frac{1}{n_l} \rightarrow 0.$$

This completes the proof of Lemma A3. $\square$

Appendix B. Proof of Theorem 1

Proof. According Lemmas A1–A3, as $n, p \rightarrow \infty$, it is easy to complete the proof of Theorem 1. $\square$

References

1. Anderson, T.W. *An Introduction to Multivariate Statistical Analysis*, 3rd ed.; Wiley: Hoboken, NJ, USA, 2003.
2. Muirhead, R.J. *Aspects of Multivariate Statistical Theory*; Wiley: New York, NY, USA, 2005.
3. Schmidt, M.; Böhm, D.; Törne, C.; Steiner, E.; Puhl, A.; Pilch, H.; Lehr, H.; Hengstler, J.; Kölbl, H.; Gehrmann, M. The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Res.* **2008**, *68*, 5405–5413. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Khan, J.; Wei, J.S.; Ringner, M.; Saal, L.H.; Ladanyi, M.; Westermann, F.; Berthold, F.; Schwab, M.; Antonescu, C.R.; Peterson, C.; et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.* **2001**, *7*, 673–679. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Zhu, Z.; Ong, Y.S.; Dash, M. Markov blanket-embedded genetic algorithm for gene selection. *Pattern Recognit.* **2007**, *49*, 3236–3248. [\[CrossRef\]](#)
6. Li, T.; Zhang, C.; Ogihara, M. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics* **2004**, *20*, 2429–2437. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Wilks, S.S. Sample criteria for testing equality of means, equality of variances, and equality of covariances in a normal multivariate distribution. *Ann. Math. Stat.* **1946**, *17*, 257–281. [\[CrossRef\]](#)
8. Ahmad, M.R. Testing homogeneity of several covariance matrices and multi-sample sphericity for high-dimensional data under non-normality. *Commun. Stat. Theory Methods* **2017**, *46*, 3738–3753. [\[CrossRef\]](#)
9. Schott, J.R.A. Test for the equality of covariance matrices when the dimension is large relative to the sample sizes. *Comput. Stat. Data Anal.* **2007**, *51*, 6535–6542. [\[CrossRef\]](#)
10. Srivastava, M.S.; Yanagihara, H. Testing the equality of several covariance matrices with fewer observations than the dimension. *J. Multivar. Anal.* **2010**, *101*, 1319–1329. [\[CrossRef\]](#)
11. Zhang, C.; Bai, Z.; Hu, J.; Wang, C. Multi-sample test for high-dimensional covariance matrices. *Commun. Stat. Theory Methods* **2018**, *47*, 3161–3177. [\[CrossRef\]](#)
12. Zhong, P.; Li, R.; Shanto, S. Homogeneity tests of covariance matrices with high-dimensional longitudinal data. *Biometrika* **2019**, *106*, 619–634. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Zheng, S.; Lin, R.; Guo, J.; Yin, G. Testing homogeneity of high-dimensional covariance matrices. *Stat. Sin.* **2020**, *30*, 35–53. [\[CrossRef\]](#)
14. Qayed, A.; Han, D. Homogeneity test of several covariance matrices with high-dimensional data. *J. Biopharm. Stat.* **2021**, *31*, 523–540. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Box, G.E.P. A general distribution theory for a class of likelihood criteria. *Biometrika* **1949**, *36*, 317–346. [\[CrossRef\]](#)

16. Li, J.; Chen, S. Two sample tests for high-dimensional covariance matrices. *Ann. Stat.* **2012**, *40*, 908–940. [[CrossRef](#)]
17. Aoshima, M.; Yata, K. Two-sample tests for high-dimension, strongly spiked eigenvalue models. *Stat. Sin.* **2018**, *28*, 43–62. [[CrossRef](#)]
18. Jiang, Y.; Wen, C.; Jiang, Y.; Wang, X.; Zhang, H. Use of Random Integration to Test Equality of High Dimensional Covariance Matrices. *Stat. Sin.* **2020**. [[CrossRef](#)]
19. Chen, S.; Qin, Y. A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Stat.* **2010**, *38*, 808–835. [[CrossRef](#)]