Mathematics and Statistics 12(5): 455-464, 2024
DOI: 10.13189/ms.2024.120507 http://www.hrpub.org

# A New Test for Equality of Two Covariance Matrices in High-Dimensional Data

**Saowapa Chaipitak[1], Boonyarit Choopradit[2,*]**

[1]Department of Statistics, Faculty of Science, Kasetsart University, Thailand
[2]Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Thailand

*Received May 10, 2024; Revised August 1, 2024; Accepted September 19, 2024*

***Cite This Paper in the Following Citation Styles***

***(a):*** *[1] Saowapa Chaipitak, Boonyarit Choopradit , "A New Test for Equality of Two Covariance Matrices in High-Dimensional Data," Mathematics and Statistics, Vol. 12, No. 5, pp. 455 - 464, 2024. DOI: 10.13189/ms.2024.120507.*

***(b):*** *Saowapa Chaipitak, Boonyarit Choopradit (2024). A New Test for Equality of Two Covariance Matrices in High-Dimensional Data. Mathematics and Statistics, 12(5), 455 - 464. DOI: 10.13189/ms.2024.120507.*



**Abstract** High-dimensional data, characterized by datasets with many variables (dimension) relative to the number of observations, is growing in prominence owing to advances in data collection and storage capabilities. This data type is widespread across various fields. While it presents unique challenges and opportunities, robust statistical approaches are crucial for effectively leveraging the potential of high-dimensional data. In multivariate statistical analysis, the likelihood ratio test (LRT) is frequently used for evaluating the equality of two covariance matrices. Nevertheless, the LRT lacks definition in high-dimensional data contexts. This paper aims to introduce a new test for ascertaining the equality of two covariance matrices in high-dimensional data, especially for datasets that follow a multivariate normal distribution. The test ($T_{new}$) proposed in this study utilizes consistent estimators in quadratic and symmetric bilinear forms. As the dimension and sample sizes approach infinity, the asymptotic null distribution of the test converges to the standard normal distribution. A simulation study was conducted to assess the performance of the proposed test compared to three existing tests. The existing tests were proposed by Schott in 2007, Srivastava and Yanagihara in 2010, and Li and Chen in 2012. The focus was on type I error rates and the test's power under spherical and Toeplitz covariance matrix structures. The simulation results demonstrate that $T_{new}$ outperforms the tests proposed by Srivastava and Yanagihara, as well as Li and Chen, in all scenarios evaluated. Moreover, it performs comparably to Schott's test. Additionally, $T_{new}$ demonstrates remarkable stability even when faced with alterations in the covariance matrix structure. This robust performance of $T_{new}$ suggests that it can serve as a reliable tool for statistical inference in multivariate analyses, especially in high-dimensional contexts. For practitioners, the use of the proposed test could mean more accurate decision-making in scientific research and policymaking, where precision and reliability are paramount.

**Keywords** Multivariate Normal Distribution, High-dimensional Covariance Matrix, DNA Microarray Data, Toeplitz Structure, Spherical Structure

## 1. Introduction

Testing the equality of two population covariance matrices is essential in multivariate analysis. Many commonly used multivariate techniques and tests rely on determining whether the population covariance matrices are identical. For instance, if two population covariance matrices are equal, computing the asymptotic distribution of Hotelling's $T^2$ statistic, used for testing the equality of means, becomes straightforward [1,2]. Similarly, verifying the equality of population covariance matrices in applications like classification is crucial to correctly applying linear discriminant analysis (LDA) [1].

Let $\mathbf{X}_{ik} = (X_{ik1}, ..., X_{ikp})'$, $k = 1, 2, \ldots, N_i$, and $i = 1, 2$, be independent and identically distributed multivariate normal random vectors with unknown mean vectors $\boldsymbol{\mu}_i$ and unknown covariance matrices $\boldsymbol{\Sigma}_i$, where $\boldsymbol{\Sigma}_i$ are positive definite matrices, i.e., $\mathbf{X}_{ik} : N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$. The

objective of this study is to test the hypothesis of equality between two covariance matrices:

$$H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma} \quad \text{versus} \quad H_1 : \boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2, \tag{1}$$

when the number of variables $p$ is greater than or equal to the sample sizes $N_i$, i.e., $p \geq N_i$. Here, $\boldsymbol{\Sigma}$ represents the common unknown covariance matrix of the two populations under the null hypothesis.

For testing (1), the likelihood ratio test (LRT) is commonly used and is as follows:

$$\Lambda = \frac{\prod_{i=1}^{2} |\mathbf{V}_i|^{N_i/2}}{|\mathbf{V}|^{N/2}} \frac{N^{Np/2}}{\prod_{i=1}^{2} N_i^{N_i p/2}},$$

where $\mathbf{V}_i = \sum_{k=1}^{N_i} \left(\mathbf{X}_{ik} - \bar{\mathbf{X}}_i\right)\left(\mathbf{X}_{ik} - \bar{\mathbf{X}}_i\right)'$, $\bar{\mathbf{X}}_i = (1/N_i)\sum_{k=1}^{N_i} \mathbf{X}_{ik}$, $\mathbf{V} = \mathbf{V}_1 + \mathbf{V}_2$ and $N = N_1 + N_2$. The test statistic $\Lambda$ is non-degenerate only if each $\mathbf{V}_i$ is a non-singular matrix (full rank matrix), which means that $p < N_i$ is required for $i = 1, 2$ [3].

Current trends are towards increasing the number of observations and expanding the number of variables in data analysis. This is made possible by the automated collection of detailed information about each observation. Various fields, such as genomics, economics, bioinformatics, engineering, social networks, econometrics, and meteorology, commonly encounter datasets where a single observation may contain thousands of dimensions. However, the total number of observations may be relatively low, ranging in the tens or hundreds. This phenomenon is referred to as "high-dimensional data". For instance, DNA microarray data often involves thousands of gene expressions with only a few observations in each group. This poses challenges for traditional methods in understanding, classifying, and selecting gene associations across different phenotypes. The two-sample test for covariance matrices is crucial in gene expression data analysis, as highlighted by Hu et al. [4] and Cai et al. [5]. Many researchers have studied this issue, such as Schott [6], Srivastava and Du [7], Chen and Qin [8], Chen et al. [9], and Chaipitak and Chongcharoen [10].

The LRT becomes undefined in high-dimensional settings. This occurs when the number of variables is greater than or equal to the sample sizes. In such cases, $\mathbf{V}_i$ becomes singular. A singular matrix poses a fundamental problem because the LRT relies on the determinants of these matrices, and the determinant of a singular matrix is zero. As a result, the LR statistic becomes infinite or undefined, altering its asymptotic behavior [1,11].

To test (1) in high-dimensional settings, numerous researchers have developed alternative testing methods. For example, Schott [12] introduced a test based on the Frobenius norm of $\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2$. Srivastava and Yanagihara [13] developed a test based on the distance measurement $\text{tr}(\boldsymbol{\Sigma}_1^2)/[\text{tr}(\boldsymbol{\Sigma}_1)]^2$ - $\text{tr}(\boldsymbol{\Sigma}_2^2)/[\text{tr}(\boldsymbol{\Sigma}_2)]^2$. Li and Chen [14] introduced a test using a combination of *U*-statistics, which was also motivated by an unbiased estimator of the Frobenius norm of $\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2$. Jiang et al. [15] proposed the LRT for comparing two normal distributions with similar covariance matrices. Cai et al. [5] introduced a test using the maximal standardized differences ($\ell_{\max}$-type formulation) between two sample covariance matrix entries. Chaipitak and Chongcharoen [16] proposed a test based on an unbiased and consistent estimator of the ratio between the sum of squares of covariance matrix elements. Ahmad and Rosen [17] formulated a test statistic for evaluating a high-dimensional covariance matrix from a population distributed as a multivariate normal using the theory of *U*-statistics. Zhang et al. [18] and Bai et al. [19] proposed a test statistic that generalizes the test proposed by Li and Chen [14]. In recent years, He et al. [20] introduced a family of *U*-statistics and integrated them with an $\ell_{\max}$-type statistic. Zi and Chen [21] proposed a test based on spatial sign function. Yu et al. [22] suggested combining a thresholding statistic with a $\ell_2$-statistic to test high-dimensional means and covariances matrix. Chen et al. [23] proposed a test based on a multi-level thresholding procedure. Ding et al. [24] introduced a methodology that relies on a data-splitting procedure. Yu et al. [25] introduced a test based on Fisher's method to combine the *p*-values of quadratic and maximum form statistics.

Motivated by the mentioned issue, this paper introduces a new test statistic for testing (1) in a high-dimensional setting. We utilize unbiased and consistent estimators based on *U*-statistics developed by Ahmad and Rosen [17].

The organization of this paper is as follows: Section 2 introduces materials and methods and reviews existing tests. Section 3 presents the new test statistic and discusses its asymptotic normality. Section 4 evaluates the proposed test's performance through simulations, comparing it with existing tests from Schott [12], Srivastava and Yanagihara [13], and Li and Chen [14]. Section 5 applies the test to DNA microarray data. Finally, Section 6 concludes the paper, and the results obtained by Ahmad and Rosen [17] are provided in the appendix.

## 2. Materials and Methods

Define

$$\mathbf{S} = (1/n)\mathbf{V}, \quad n = n_1 + n_2,$$

$$\hat{a}_1 = (1/p)tr\mathbf{S}, \tag{2}$$

$$\hat{a}_2 = \frac{n^2}{p(n-1)(n+2)}\left\{tr\mathbf{S}^2 - \frac{1}{n}(tr\mathbf{S})^2\right\}, \tag{3}$$

$$\mathbf{S}_i = (1/n_i)\mathbf{V}_i,\ n_i = N_i - 1,\ \ i = 1, 2,$$

$$\hat{a}_{1i} = (1/p)tr\mathbf{S}_i,\ i = 1, 2,$$

$$\hat{a}_{2i} = \frac{n_i^2}{p(n_i-1)(n_i+2)}\left\{tr\mathbf{S}_i^2 - \frac{1}{n_i}(tr\mathbf{S}_i)^2\right\},\ \ i = 1, 2,$$

$$\hat{a}_3 = \frac{1}{n(n^2+3n+4)}\left\{\frac{1}{p}tr\mathbf{V}^3 - 3n(n+1)p\hat{a}_2\hat{a}_1 - np^2\hat{a}_1^3\right\},$$

$$\hat{a}_4 = \frac{1}{d_0}\left(\frac{1}{p}tr\mathbf{V}^4 - pd_1\hat{a}_1 - p^2d_2\hat{a}_1^2\hat{a}_2 - pd_3\hat{a}_2^2 - np^3\hat{a}_1^4\right), \quad (4)$$

where $d_0 = n(n^3+6n^2+21n+18)$, $d_1 = 2n(2n^2+6n+9)$, $d_2 = 2n(3n+2)$, and $d_3 = n(2n^2+5n+7)$.

The pooled sample covariance matrix $\mathbf{S}$ estimates the common covariance matrix $\mathbf{\Sigma}$ [1].

Let $a_r = (1/p)tr\mathbf{\Sigma}^r$, $r$ = 1, 2, 3, 4. Srivastava [26] demonstrated that $\hat{a}_1$ and $\hat{a}_2$ are unbiased and consistent estimators of $a_1$ and $a_2$, respectively. Furthermore, Srivastava and Yanagihara [13] established that $\hat{a}_3$ and $\hat{a}_4$ are unbiased and consistent estimators of $a_3$ and $a_4$, respectively.

Schott [12] proposed a test statistic $T_{SC}$, which is based on a consistent estimator of the sum of squared differences between elements of two covariance matrices, represented by the squared Frobenius norm of $\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2$, i.e., $\|\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2\|_F^2$. The test statistic is defined as follows:

$$T_{SC} = \frac{n_1n_2}{2(n_1+n_2)\hat{a}_2}\left(\hat{a}_{21} + \hat{a}_{22} - \frac{2}{p}tr(\mathbf{S}_1\mathbf{S}_2)\right). \quad (5)$$

Under the null hypothesis $H_0$, the test statistic $T_{SC}$ follows the standard normal as $(p, n_i) \to \infty$, $i$ = 1, 2. The null hypothesis is rejected if $|T_{SC}| > Z_{\alpha/2}$ where $Z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution.

Srivastava and Yanagihara [13] considered testing $H_0$ with the test statistic given by

$$T_{SY} = \frac{\hat{\gamma}_1 - \hat{\gamma}_2}{\sqrt{\hat{\xi}_1^2 + \hat{\xi}_2^2}}, \quad (6)$$

where $\hat{\gamma}_i = \hat{a}_{2i}/\hat{a}_{1i}^2$ which is a consistent estimator of $a_{2i}/a_{1i}^2$ and $\hat{\xi}_i^2 = \frac{4}{n_i^2}\left\{\frac{\hat{a}_2^2}{\hat{a}_1^4} + \frac{2n_i}{p}\left(\frac{\hat{a}_2^3}{\hat{a}_1^6} - \frac{2\hat{a}_2\hat{a}_3}{\hat{a}_1^5} + \frac{\hat{a}_4}{\hat{a}_1^4}\right)\right\}$, $i$ = 1, 2. Under the null hypothesis $H_0$, $T_{SY}$ converges in distribution to the standard normal as $(p, n_i) \to \infty$, $i$ = 1, 2, and the null hypothesis is rejected if $|T_{SY}| > Z_{\alpha/2}$.

Li and Chen [14] proposed a test that does not rely on specific distributional assumptions for the two populations and accommodates dimensions substantially exceeding the sample sizes. The test statistic is defined as follows:

$$T_{LC} = A_{N_1} + A_{N_2} - 2C_{N_1,N_2} \quad (7)$$

where

$$A_{N_i} = \frac{1}{N_i(N_i-1)}\sum_{j\neq k}^{N_i}\left(\boldsymbol{X}_{ij}'\boldsymbol{X}_{ik}\right)^2 - \frac{2}{N_i(N_i-1)(N_i-2)}\sum_{j\neq k\neq m}^{N_i}\boldsymbol{X}_{ij}'\boldsymbol{X}_{ik}\boldsymbol{X}_{ik}'\boldsymbol{X}_{im} + \frac{1}{N_i(N_i-1)(N_i-2)(N_i-3)}\sum_{j\neq k\neq m\neq n}^{N_i}\boldsymbol{X}_{ij}'\boldsymbol{X}_{ik}\boldsymbol{X}_{im}'\boldsymbol{X}_{in}$$

and

$$C_{N_1,N_2} = \frac{1}{N_1N_2}\sum_{j=1}^{N_1}\sum_{k=1}^{N_2}\left(\boldsymbol{X}_{1j}'\boldsymbol{X}_{2k}\right)^2 - \frac{1}{N_1N_2(N_1-1)}\sum_{j\neq m}^{N_1}\sum_{k=1}^{N_2}\boldsymbol{X}_{1j}'\boldsymbol{X}_{2k}\boldsymbol{X}_{2k}'\boldsymbol{X}_{1m} - \frac{1}{N_1N_2(N_2-1)}\sum_{j\neq m}^{N_2}\sum_{k=1}^{N_1}\boldsymbol{X}_{2j}'\boldsymbol{X}_{1k}\boldsymbol{X}_{1k}'\boldsymbol{X}_{2m} + \frac{1}{N_1N_2(N_1-1)(N_2-1)}\sum_{j\neq m}^{N_1}\sum_{k\neq n}^{N_2}\boldsymbol{X}_{1j}'\boldsymbol{X}_{2k}\boldsymbol{X}_{1m}'\boldsymbol{X}_{2n}.$$

The statistics $A_{N_i}$ and $C_{N_1,N_2}$ are unbiased estimators of $tr(\mathbf{\Sigma}_i^2)$ and $tr(\mathbf{\Sigma}_1\mathbf{\Sigma}_2)$, respectively. Under the null hypothesis $H_0$, the test statistic $T_{LC}$ is distributed as the standard normal as $\min\{N_1, N_2\} \to \infty$, and the null hypothesis is rejected if $|T_{LC}| > Z_{\alpha/2}$.

# 3. Main Results

For $\mathbf{X}_{ik}$, $k = 1, ..., N_i$, and $i$ = 1, 2, define $A_{ik} = \mathbf{X}_{ik}'\mathbf{X}_{ik}$, as a quadratic form, and $A_{ikl} = \mathbf{X}_{ik}'\mathbf{X}_{il}$, $k \neq l$, as a symmetric bilinear form. Therefore, similar to the case of one population (refer to the Appendix), unbiased and consistent estimators of $tr\mathbf{\Sigma}_i$, $(tr\mathbf{\Sigma}_i)^2$, and $tr\mathbf{\Sigma}_i^2$, $i$ = 1, 2, are defined as

$$C_{1i} = \frac{1}{N_i(N_i-1)}\sum_{k=1}^{N_i}A_{ik},$$

$$C_{2i} = \frac{1}{N_i(N_i-1)}\sum_{k\neq l}^{N_i}A_{ik}A_{il}, \quad (8)$$

and

$$C_{3i} = \frac{1}{N_i(N_i-1)}\sum_{k\neq l}^{N_i}A_{ikl}^2, \quad (9)$$

respectively. The consistency remains unchanged even

when the number of variables exceeds the sample size.

For testing (1), we note that when $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, $tr\boldsymbol{\Sigma}_1 = tr\boldsymbol{\Sigma}_2$, $tr\boldsymbol{\Sigma}_1^2 = tr\boldsymbol{\Sigma}_2^2$, etc. Thus, under the null hypothesis, we must have $\psi_1 = \psi_2$, where $\psi_i = \dfrac{tr\boldsymbol{\Sigma}_i^2 / p}{\left(tr\boldsymbol{\Sigma}_i / p\right)^2} = ptr\boldsymbol{\Sigma}_i^2 / \left(tr\boldsymbol{\Sigma}_i\right)^2$, $i$ = 1, 2. Therefore, we use the consistent estimators presented in equations (8) and (9) to define $\hat{\psi}_i = pC_{3i} / C_{2i}$, $i$ = 1, 2. Our focus is on the measure derived by concentrating on $\psi_1 - \psi_2$. With this in mind, we present the asymptotic normality of $\hat{\psi}_1 - \hat{\psi}_2$. For a valid application of Lemma 2 in the Appendix, we shall make the following assumptions.

(A1) For $p \to \infty$, let $(1/p)tr\boldsymbol{\Sigma}_i = O(1)$, $i$ = 1, 2,

(A2) $tr\boldsymbol{\Sigma}_i^2 / p^2 = \delta_i$, where $0 < \delta_i \le \infty$, $i$ = 1, 2,

(A3) For $N_i$, $p \to \infty$, $p / N_i \to c_i$, $c_i \in (0, \infty)$, $i$ = 1, 2,

(A4) $\inf_p (tr\boldsymbol{\Sigma}_i^4 / (tr\boldsymbol{\Sigma}_i^2)^2) > 0$, $i$ = 1, 2.

**Theorem 1.** Let $\hat{\psi}_i$ and $\psi_i$ be as defined above and $a_{ri} = (1/p)tr\boldsymbol{\Sigma}_i^r$, $r$ = 1, …, 4, for $i$ = 1, 2. Then, under Assumptions (A1) - (A4),

$$\hat{\psi}_1 - \hat{\psi}_2 \xrightarrow{d} N\left(\psi_1 - \psi_2, \beta_1^2 + \beta_2^2\right),$$

as $(p, N_i) \to \infty$,

where $\beta_i^2 = \dfrac{4m_{1i}}{N_i\left(N_i - 1\right)}\left[\left(2N_i - 1\right)m_{2i} + 1\right]$, $m_{1i} = a_{2i}^2 / a_{1i}^4$ and $m_{2i} = a_{4i} / pa_{2i}^2$, $i$ = 1, 2. The symbol "$\xrightarrow{d}$" stands for "converges in distribution."

**Proof of Theorem 1.** By employing Lemma 2 given in the Appendix, each $\hat{\psi}_i$ is a random variable asymptotically distributed as a normal distribution with mean $\psi_i$ and variance $\beta_i^2$. Consequently, the asymptotic distribution of $\hat{\psi}_1 - \hat{\psi}_2$ follows a normal distribution, which represents a linear function of independent normal random variables with mean $\psi_1 - \psi_2$ and variance $\beta_1^2 + \beta_2^2$. This completes the proof.

**Corollary 1.** If $H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ and under Assumptions (A1) - (A4), then

$$\frac{\hat{\psi}_1 - \hat{\psi}_2}{\eta} \xrightarrow{d} N\left(0,1\right), \tag{10}$$

as $(p, N_i) \to \infty$, where

$$\eta^2 = \left[4m_1\left(\sum_{i=1}^{2} N_i\left(N_i - 1\right) + m_2\sum_{i \ne j}^{2}\left(2N_i - 1\right)N_j\left(N_j - 1\right)\right)\right] / u,$$

$m_1 = a_2^2 / a_1^4$, $m_2 = a_4 / pa_2^2$, and $u = \prod_{i=1}^{2} N_i\left(N_i - 1\right)$.

**Proof of Corollary 1.** Under $H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, then $a_{11} = a_{12}$, $a_{21} = a_{22}$, and $a_{41} = a_{42}$. It leads to that $m_{11} = m_{12} \equiv m_1$, and $m_{21} = m_{22} \equiv m_2$. Therefore, by applying Theorem 1 and invoking Slutsky's theorem, the proof is completed.

In practice, the variance $\eta^2$ can be estimated by replacing the consistent estimators of $a_1$, $a_2$, and $a_4$, as specified in equation (10), with the consistent estimators $\hat{a}_1$, $\hat{a}_2$, and $\hat{a}_4$, outlined in equations (2) through (4), respectively. Consequently, based on $\hat{m}_1 = \hat{a}_2^2 / \hat{a}_1^4$ and $\hat{m}_2 = \hat{a}_4 / p\hat{a}_2^2$, the new test statistic under the null hypothesis $H_0$ is defined as

$$T_{new} = \frac{\hat{\psi}_1 - \hat{\psi}_2}{\hat{\eta}}, \tag{11}$$

where

$$\hat{\eta}^2 = \left[4\hat{m}_1\left(\sum_{i=1}^{2} N_i\left(N_i - 1\right) + \hat{m}_2\sum_{i \ne j}^{2}\left(2N_i - 1\right)N_j\left(N_j - 1\right)\right)\right] / u.$$

# 4. Simulation Studies

In this section, a Monte Carlo simulation was conducted to evaluate the efficiency of the proposed test $T_{new}$ as defined in equation (11) and compare it to three existing tests: $T_{SC}$ in equation (5), $T_{SY}$ in equation (6), and $T_{LC}$ in equation (7). Using the FORTRAN programming language and the RNMVN subroutine from the IMSL library, we generated two independent $p$-variate normal datasets over 5,000 iterations for different scenarios, with $p \ge N_i$, $i$ = 1, 2. Both $p$ and $N_i$ varied across values of 10 (very small), 20 (small), 30 (moderate), 60 (large), and 180 (very large). The simulation study employed a nominal significance level of 0.05.

Under the null hypothesis $H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, the empirical type I error rate was evaluated under two distinct forms of covariance matrix structures for the common covariance matrix $\boldsymbol{\Sigma}$, which were considered spherical and Toeplitz. The first one was a spherical structure (or simple structure), $\boldsymbol{\Sigma} = \mathbf{I}_p$, where $\mathbf{I}_p$ is the $p \times p$ identity matrix, while the second one was set as a Toeplitz structure, $\boldsymbol{\Sigma} = \mathbf{T}_{(q)}$, where $\mathbf{T}_{(q)}$ is a $p \times p$ matrix in the form of a Toeplitz structure: $\mathbf{T}_{(q)} = (\sigma_{ij})_{p\times p} = (\sigma_{|i-j|})_{p\times p}$ for which elements $\sigma_0 = 1$, $\sigma_1 = q = -0.50$ while the remaining elements are all equal to zero. Tables 1 and 2 present the results of the empirical type I error rates of all four tests. Additionally, to facilitate a clearer comparison of the empirical type I error rates among all four tests as sample sizes increase, these empirical type I error rates are illustrated in Figures 1 and 2, respectively.

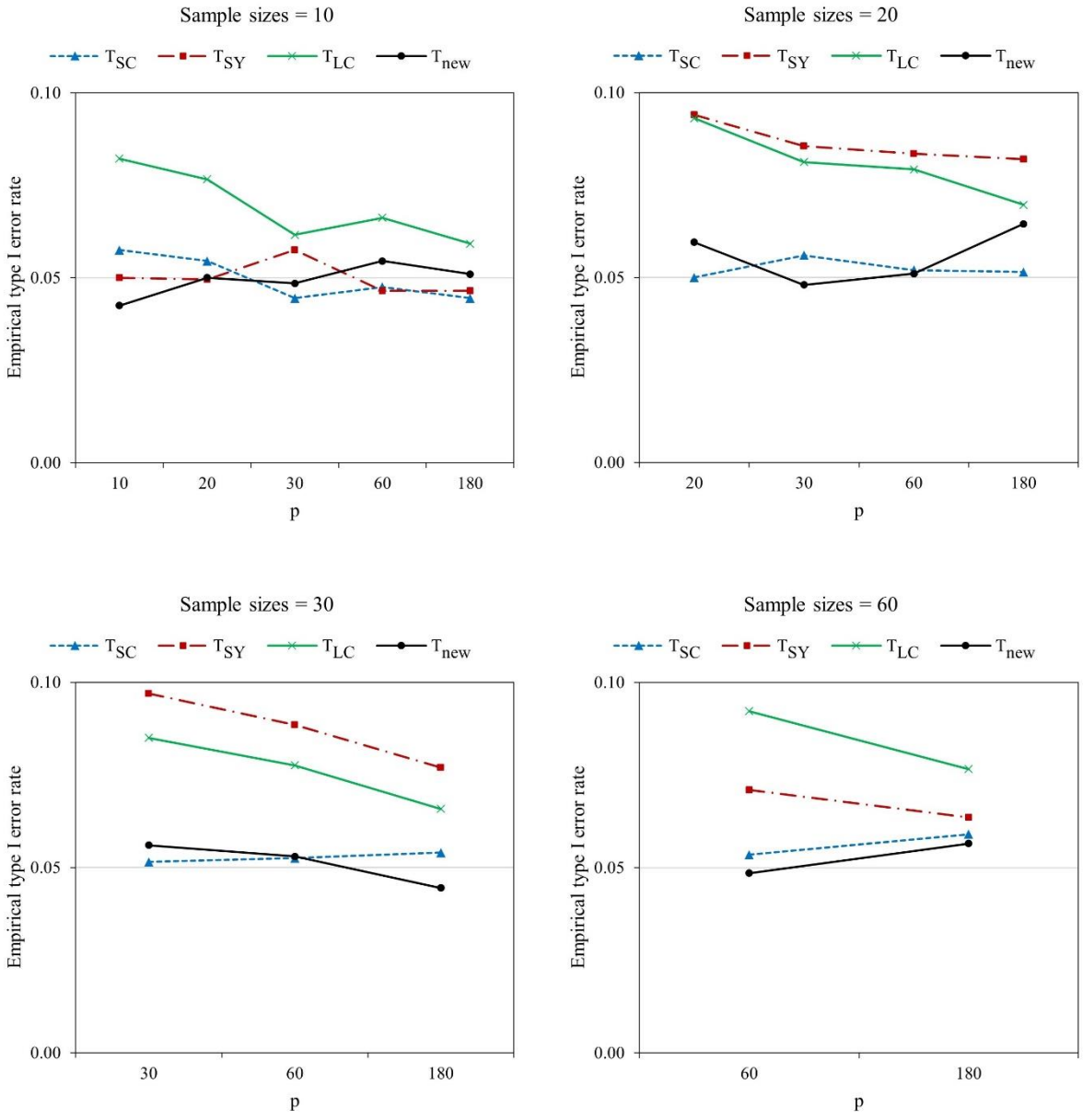


**Figure 1.** Empirical type I error rates plot for $T_{SC}$, $T_{SY}$, $T_{LC}$, and $T_{new}$ under $H_0 : \Sigma_1 = \Sigma_2 = \mathbf{I}_p$ across sample sizes of 10, 20, 30, and 60, maintaining the nominal significance level of 0.05

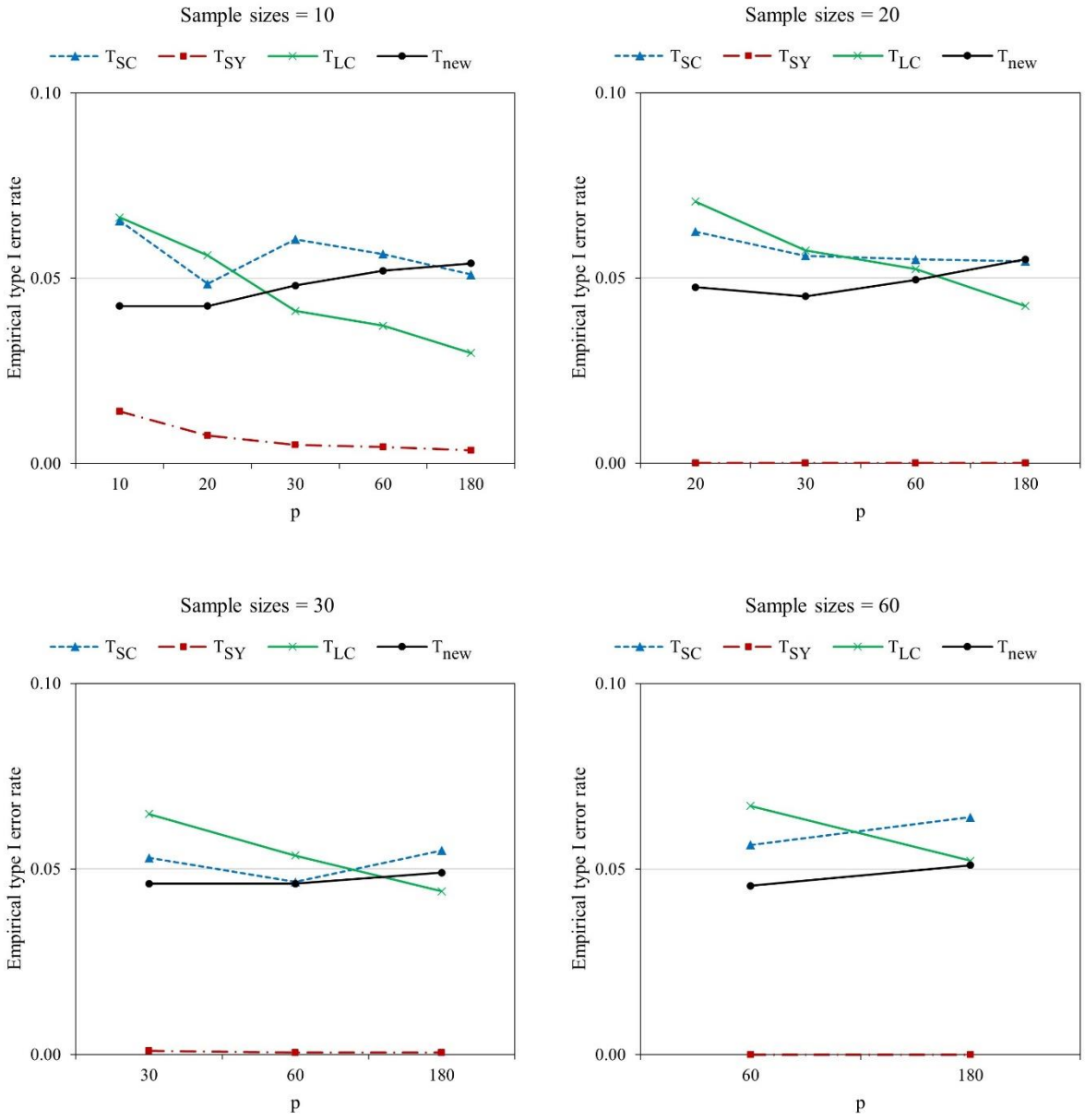


**Figure 2.** Empirical type I error rates plot for $T_{SC}$, $T_{SY}$, $T_{LC}$, and $T_{new}$ under $H_0 : \Sigma_1 = \Sigma_2 = \mathbf{T}_{(-0.50)}$ across sample sizes of 10, 20, 30, and 60, maintaining the nominal significance level of 0.05

We assessed the empirical powers of all tests that successfully controlled the type I error rate. A test's power is defined as the probability of correctly rejecting a false null hypothesis $H_0$, with a more powerful test demonstrating a higher empirical power. Under the alternative hypothesis $H_1 : \mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$, additional simulations were performed to compare the powers of $T_{SC}$, $T_{LC}$, $T_{SY}$, and $T_{new}$. Two different covariance matrices, $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$, were established in distinct settings. The first setting was that $\mathbf{\Sigma}_1 = \mathbf{I}_p$ and $\mathbf{\Sigma}_2 = \mathbf{T}_{(-0.50)}$. The second setting was that $\mathbf{\Sigma}_1 = \mathbf{T}_{(-0.50)}$ and $\mathbf{\Sigma}_2 = \mathbf{T}_{(-0.01)}$. The empirical power results are displayed in Tables 1 and 2, respectively.

**Table 1.** Empirical type I error rates under $H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{I}_p$ and empirical powers under $H_1 : \mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$, where $\mathbf{\Sigma}_1 = \mathbf{I}_p$ and $\mathbf{\Sigma}_2 = \mathbf{T}_{(-0.50)}$

| | | Empirical type I error rate | | | | Empirical power | | | |
|---|---|---|---|---|---|---|---|---|---|
| $p$ | $N_1 = N_2$ | $T_{SC}$ | $T_{SY}$ | $T_{LC}$ | $T_{new}$ | $T_{SC}$ | $T_{SY}$ | $T_{LC}$ | $T_{new}$ |
| 10 | 10 | 0.0575 | 0.0500 | 0.0822 | 0.0425 | 0.2315 | - | - | 0.3045 |
| 20 | 10 | 0.0545 | 0.0495 | 0.0766 | 0.0500 | 0.2545 | - | - | 0.3375 |
| | 20 | 0.0500 | 0.0940 | 0.0930 | 0.0595 | 0.6270 | - | - | 0.8235 |
| 30 | 10 | 0.0445 | 0.0575 | 0.0616 | 0.0485 | 0.2440 | - | - | 0.3575 |
| | 20 | 0.0560 | 0.0855 | 0.0812 | 0.0480 | 0.6375 | - | - | 0.8365 |
| | 30 | 0.0515 | 0.0970 | 0.0850 | 0.0560 | 0.9030 | - | - | 0.9885 |
| 60 | 10 | 0.0475 | 0.0465 | 0.0662 | 0.0545 | 0.2620 | - | - | 0.3735 |
| | 20 | 0.0520 | 0.0835 | 0.0792 | 0.0510 | 0.6510 | - | - | 0.8650 |
| | 30 | 0.0525 | 0.0885 | 0.0776 | 0.0530 | 0.9235 | - | - | 0.9945 |
| | 60 | 0.0535 | 0.0710 | 0.0922 | 0.0485 | 0.9995 | - | - | 1.0000 |
| 180 | 10 | 0.0445 | 0.0465 | 0.0592 | 0.0510 | 0.2565 | - | - | 0.3960 |
| | 20 | 0.0515 | 0.0820 | 0.0696 | 0.0645 | 0.7025 | - | - | 0.8795 |
| | 30 | 0.0540 | 0.0770 | 0.0658 | 0.0445 | 0.9375 | - | - | 0.9965 |
| | 60 | 0.0590 | 0.0635 | 0.0766 | 0.0565 | 1.0000 | - | - | 1.0000 |
| | 180 | 0.0460 | 0.0605 | 0.0845 | 0.0530 | 1.0000 | - | - | 1.0000 |

Note: "-" means do not report the power of a test that fails to control the type I error rate.

**Table 2.** Empirical type I error rates under $H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{T}_{(-0.50)}$ and empirical powers under $H_1 : \mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$, where $\mathbf{\Sigma}_1 = \mathbf{T}_{(-0.50)}$ and $\mathbf{\Sigma}_2 = \mathbf{T}_{(-0.01)}$

| | | Empirical type I error rate | | | | Empirical power | | | |
|---|---|---|---|---|---|---|---|---|---|
| $p$ | $N_1 = N_2$ | $T_{SC}$ | $T_{SY}$ | $T_{LC}$ | $T_{new}$ | $T_{SC}$ | $T_{SY}$ | $T_{LC}$ | $T_{new}$ |
| 10 | 10 | 0.0655 | 0.0140 | 0.0664 | 0.0425 | 0.2245 | - | - | 0.0720 |
| 20 | 10 | 0.0485 | 0.0075 | 0.0562 | 0.0425 | 0.2425 | - | - | 0.1160 |
| | 20 | 0.0625 | 0.0000 | 0.0706 | 0.0475 | 0.5500 | - | - | 0.4645 |
| 30 | 10 | 0.0605 | 0.0050 | 0.0412 | 0.0480 | 0.2470 | - | - | 0.1275 |
| | 20 | 0.0560 | 0.0000 | 0.0574 | 0.0450 | 0.6040 | - | - | 0.5330 |
| | 30 | 0.0530 | 0.0010 | 0.0648 | 0.0460 | 0.8670 | - | - | 0.8955 |
| 60 | 10 | 0.0565 | 0.0045 | 0.0372 | 0.0520 | 0.2365 | - | - | 0.1450 |
| | 20 | 0.0550 | 0.0000 | 0.0524 | 0.0495 | 0.6190 | - | - | 0.5830 |
| | 30 | 0.0465 | 0.0005 | 0.0536 | 0.0460 | 0.9090 | - | - | 0.9385 |
| | 60 | 0.0565 | 0.0000 | 0.0670 | 0.0455 | 1.0000 | - | - | 1.0000 |
| 180 | 10 | 0.0510 | 0.0035 | 0.0298 | 0.0540 | 0.2195 | - | - | 0.1575 |
| | 20 | 0.0545 | 0.0000 | 0.0424 | 0.0550 | 0.6370 | - | - | 0.6160 |
| | 30 | 0.0550 | 0.0005 | 0.0440 | 0.0490 | 0.9155 | - | - | 0.9530 |
| | 60 | 0.0640 | 0.0000 | 0.0522 | 0.0510 | 1.0000 | - | - | 1.0000 |
| | 180 | 0.0530 | 0.0000 | 0.0630 | 0.0530 | 1.0000 | - | - | 1.0000 |

Note: "-" means do not report the power of a test that fails to control the type I error rate.

According to Table 1, under setting the null hypothesis $H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{I}_\mathbf{p}$, the empirical type I error rates of $T_{SC}$ and the new test $T_{new}$ approximate the nominal significance level 0.05 reasonably well in all cases while $T_{LC}$ and $T_{SY}$ exhibit inflated empirical type I error rates in almost all scenarios, except when the sample sizes $N_i$ are very small (10 here) and equal to $p$ (also 10 here), where $T_{SY}$ effectively controls the nominal significance level. Furthermore, Figure 1 illustrates that for very small sample sizes ($N_i$ =10), as $p$ increases, the empirical type I error rates of three tests—$T_{SC}$, $T_{SY}$, and $T_{new}$ —converge toward the nominal significance level of 0.05. This convergence aligns with the findings associated with each test statistic. However, $T_{LC}$'s type I error rate convergence is comparatively slower. As the sample size increases, only two statistics, $T_{SC}$ and $T_{new}$, effectively maintain the nominal significance level. When evaluating the testing power of the test statistics, comparisons are limited to those

capable of controlling type I error rates. Consequently, only two statistics, $T_{SC}$ and $T_{new}$, are considered in this scenario. As shown in Table 1, $T_{new}$ consistently demonstrates substantially greater empirical powers than $T_{SC}$ in all cases. Especially for the very small sample size category, the empirical powers of $T_{new}$ are between 31.53% and 54.39% higher than those of $T_{SC}$, with an average increase of 41.52%. In scenarios where the sample sizes are at least moderate (30 and above), the empirical powers of $T_{new}$ become more comparable. However, $T_{new}$ remains slightly superior with power advantages of less than 9.47%. As expected, both tests approach a power of one as both $p$ and $N_i$ increase significantly.

As shown in Table 2, under setting $H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{T}_{(-0.50)}$, when considering only the $T_{SC}$ and $T_{new}$ tests, the type I error rates are effectively maintained close to the nominal significance level of 0.05 across all scenarios. However, $T_{LC}$ and $T_{SY}$ do not approximate the nominal significance level well. $T_{SY}$ 's empirical type I error rates are notably lower than the expected 0.05, approaching zero instead. Figure 2 demonstrates that $T_{SY}$ and $T_{LC}$ cannot effectively control the type I error rate in the cases under consideration, leading to the decision not to report their empirical powers for this setting. Furthermore, Figure 2 indicates that as $N_i$ increases, the empirical type I error rates for our $T_{new}$ approximate the nominal significance level 0.05 more closely than those of $T_{SC}$. According to Table 2, $T_{SC}$ exhibits greater power compared to $T_{new}$ for very small ($N_i$ = 10) and small ($N_i$ = 20) sample sizes, with $T_{SC}$'s empirical powers ranging from 3.41 to 211.81%, averaging 62.04% higher than those of $T_{new}$. Conversely, for moderate sample sizes ($N_i$ = 30), $T_{new}$ demonstrates higher power, with its powers being 3.25 to 4.10% greater than $T_{SC}$, averaging 3.54% higher. As anticipated, both $T_{SC}$ and $T_{new}$ approach an empirical power of one as both $p$ and $N_i$ increase.

# 5. Application

This section focuses on applying statistical tests capable of controlling type I error rates to real-world data. Only two test statistics, $T_{SC}$ and $T_{new}$, were deemed suitable in the previous section and thus utilized here. We used these tests on a dataset provided by Notterman et al. [27]. This dataset consists of 18 colon adenocarcinomas and 18 normal colon tissues that were analyzed using oligonucleotide arrays ($N_i$=18 for each group). The expression levels of 6,500 human genes were measured per tissue. However, due to the program's processing capacity limitations, only the first 250 gene expressions (p = 250, where p > $N_i$) were considered. We evaluated the equality of the two population covariance matrices, each with a dimension of 250×250. The observed test statistics resulted in values of 3.5160 for $T_{SC}$ and 5.3239 for $T_{new}$. Since both test statistics had $p$-values below 0.05, we rejected the hypothesis of equality between the two covariance matrices. Therefore, both tests concluded that the two population covariance matrices were unequal at a significance level of 0.05.

# 6. Conclusions

In this study, we introduce a new test statistic called $T_{new}$. This statistic is specifically designed to assess the equality of two covariance matrices for two independent multivariate normal datasets in high-dimensional settings in which the number of variables $p$ equals or exceeds the sample size $N_i$. The test statistic $T_{new}$ is based on the quadratic and symmetric bilinear forms-based statistics developed by Ahmad and Rosen [17]. As $p$ and $N_i$ increase, $T_{new}$ asymptotically follows the standard normal distribution. A simulation study was conducted to evaluate $T_{new}$'s performance and compare it to the three other tests: $T_{SC}$ by Schott [12], $T_{SY}$ by Srivastava and Yanagihara [13], and $T_{LC}$ by Li and Chen [14]. The results demonstrate that $T_{new}$ performs comparably to Schott's test, $T_{SC}$, which is superior to the tests by Srivastava and Yanagihara [13] and Li and Chen [14] in all scenarios considered. Under two different structures of covariance matrices (spherical and Toeplitz), only $T_{SC}$ and $T_{new}$ consistently controlled the type I error rate and were not adversely affected by structural alterations, unlike $T_{SY}$ and $T_{LC}$, which showed significant susceptibility to changes in the covariance matrix structure. This attribute is particularly valuable in real-world data analysis, where the assumption of a consistent covariance structure often does not hold. Nevertheless, this study examined two types of covariance matrix structures; future research could enhance the utility of $T_{new}$ by adapting it for use with larger datasets, possibly including thousands of variables. Additionally, testing of $T_{new}$ across a wider range of real-world scenarios and various disciplines could also reveal potential limitations or necessary adjustments to its algorithm. Researchers can expand this test statistic to assess the equality of multiple covariance matrices.

# Appendix

This section presents preliminary results from the research conducted by Ahmad and Rosen [17] focusing on one population.

Let $\mathbf{X}_k = \left(X_{k1}, \mathrm{K}, X_{kp}\right)'$, $k = 1, \ldots, N$, be $N$ independent and identically distributed multivariate normal random vectors with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$, i.e., $\mathbf{X}_k : N(\boldsymbol{\mu}, \mathbf{\Sigma})$. For $\mathbf{X}_k, k = 1, \ldots, N$, define $A_k = \mathbf{X}_k' \mathbf{X}_k$ as a quadratic form, and $A_{kl} = \mathbf{X}_k' \mathbf{X}_l$, $k \neq l$, as a symmetric bilinear form. Following Ahmad and Rosen [17], unbiased and consistent estimators of $tr\mathbf{\Sigma}$, $(tr\mathbf{\Sigma})^2$, and $tr\mathbf{\Sigma}^2$, are

$$C_1 = \frac{1}{N} \sum_{k=1}^{N} A_k,$$

$$C_2 = \frac{1}{N(N-1)}\sum_{k\neq l}^{N} A_k A_l,$$

and $C_3 = \frac{1}{N(N-1)}\sum_{k\neq l}^{N} A_{kl}^2$, respectively. It is worth noting that the estimators $C_1$, $C_2$, and $C_3$ serve as natural representations of $U$-statistics. These estimators demonstrate consistency in a high-dimensional setup [17,28,29].

Define $T_1 = pC_3 / C_2 - 1 = \hat{\psi} - 1$, where $\hat{\psi} = pC_3 / C_2$ estimates $\psi = ptr(\mathbf{\Sigma}^2)/\left[tr(\mathbf{\Sigma})\right]^2$. The following lemma was derived by Ahmad and Rosen [17]. The proof can be found in their work. The required assumptions are as follows:

(A5) For $p \to \infty$, let $(1/p)tr\mathbf{\Sigma} = O(1)$,

(A6) $tr\mathbf{\Sigma}^2 / p^2 = \delta$, where $0 < \delta \leq \infty$,

(A7) For $N, p \to \infty$, $p/N \to c$, $c \in (0, \infty)$,

(A8) $\inf_p \left[ tr\mathbf{\Sigma}^4 / (tr\mathbf{\Sigma}^2)^2 \right] > 0$.

**Lemma 1.** Let $T_1$ be as defined above. Then, under Assumptions (A5) - (A8),

$$\sigma_{T_1}^{-1}\left(\frac{T_1+1}{\psi} - 1\right) \xrightarrow{d} N(0,1),$$

as $(p, N) \to \infty$, where $\sigma_{T_1}^2$ is the asymptotic variance of $T_1$ which is given by

$$\sigma_{T_1}^2 = \frac{4}{N(N-1)}\left[\frac{(2N-1)tr\mathbf{\Sigma}^4}{\left(tr\mathbf{\Sigma}^2\right)^2} + 1\right].$$

By applying Slutsky's theorem to Lemma 1, we obtain the asymptotic normality of $\hat{\psi}$, as stated in Lemma 2.

**Lemma 2.** Let $\psi$ and $\hat{\psi}$ be as defined above and $a_r = (1/p)tr\Sigma^r$, $r = 1, \ldots, 4$. Then, under Assumptions (A5) - (A8),

$$\hat{\psi} \xrightarrow{d} N(\psi, \beta^2),$$

as $(p, N) \to \infty$, where $\beta^2 = \frac{4m_1}{N(N-1)}\left[(2N-1)m_2 + 1\right]$, $m_1 = a_2^2 / a_1^4$ and $m_2 = a_4 / pa_2^2$.

## Acknowledgements

This work was supported by the International SciKU Branding (ISB), Faculty of Science, Kasetsart University, and the Department of Statistics, Faculty of Science, Kasetsart University, Thailand. We would like to express our gratitude to the referees for their valuable suggestions, which greatly improved the quality of this paper.

## REFERENCES

[1] Anderson T. W., An introduction to multivariate statistical analysis, 3rd ed. New York: John Wiley and Sons, 2003.

[2] Yao J., Zheng S., Z. Bai, Large sample covariance matrices and high-dimensional data analysis. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2015.

[3] Guo. W., Y. Qi, "Asymptotic distributions for likelihood ratio tests for the equality of covariance matrices," Metrika, Vol. 87, pp. 247–279, 2024. doi: 10.1007/s00184-023-00912-6.

[4] Hu R., Qiu X., G. Glazko, "A new gene selection procedure based on the covariance distance," Bioinformatics, Vol. 26, No. 3, pp. 348–354, 2010. doi: 10.1093/bioinformatics/btp672.

[5] Cai T., Liu W., Y. Xia, "Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings," Journal of the American Statistical Association, Vol. 108, No. 501, pp. 265–277, 2013. doi: 10.1080/01621459.2012.758041.

[6] Schott J. R., "Some high-dimensional tests for a one-way MANOVA," Journal of Multivariate Analysis, Vol. 98, No. 9, pp. 1825–1839, 2007b. doi: 10.1016/J.JMVA.2006.11.007.

[7] Srivastava M. S., M. Du, "A test for the mean vector with fewer observations than the dimension," Journal of Multivariate Analysis, Vol. 99, No. 3, pp. 386–402, 2008. doi: 10.1016/J.JMVA.2006.11.002.

[8] Chen S. X., Y.-L. Qin, "A two-sample test for high-dimensional data with applications to gene-set testing," The Annals of Statistics, Vol. 38, No. 2, pp. 808–835, 2010. doi: 10.1214/09-AOS716.

[9] Chen S. X., Zhang L.-X., P.-S. Zhong, "Tests for high-dimensional covariance matrices," Journal of the American Statistical Association, Vol. 105, No. 490, pp. 810–819, 2010. doi: 10.1198/jasa.2010.tm09560.

[10] Chaipitak S., S. Chongcharoen, "A covariance matrix test for high-dimensional data," Songklanakarin Journal of Science and Technology, Vol. 38, pp. 521–535, 2016.

[11] Fujikoshi Y., Ulyanov V. V., R. Shimizu, Multivariate statistics high-dimensional and large-sample approximations. New York: John Wiley and Sons, 2010.

[12] Schott J. R., "A test for the equality of covariance matrices when the dimension is large relative to the sample sizes," Computational Statistics and Data Analysis, Vol. 51, No. 12, pp. 6535–6542, 2007a. doi: 10.1016/J.CSDA.2007.03.004.

[13] Srivastava M. S., H. Yanagihara, "Testing the equality of several covariance matrices with fewer observations than the dimension," Journal of Multivariate Analysis, Vol. 101, No. 6, pp. 1319–1329, 2010. doi: 10.1016/j.jmva.2009.12.010.

[14] Li J., S. X. Chen, "Two sample tests for high-dimensional

covariance matrices," The Annals of Statistics, Vol. 40, No. 2, pp. 908–940, 2012. doi: 10.1214/12-AOS993.

[15] Jiang D., Jiang T., F. Yang, "Likelihood ratio tests for covariance matrices of high-dimensional normal distributions," Journal of Statistical Planning and Inference, Vol. 142, No. 8, pp. 2241–2256, 2012. doi: 10.1016/j.jspi.2012.02.057.

[16] Chaipitak S., S. Chongcharoen, "A test for testing the equality of two covariance matrices for high-dimensional data," Journal of Applied Sciences, Vol. 13, No. 2, pp. 270–277, 2013. doi: 10.3923/jas.2013.270.277.

[17] Ahmad M. R., D. von Rosen, "Tests for high-dimensional covariance matrices using the theory of U -statistics," Journal of Statistical Computation and Simulation, Vol. 85, No. 13, pp. 2619–2631, 2015. doi: 10.1080/00949655.2014.948441.

[18] Zhang C., Bai Z., Hu J., C. Wang, "Multi-sample test for high-dimensional covariance matrices," Communications in Statistics - Theory and Methods, Vol. 47, No. 13, pp. 3161–3177, 2018. doi: 10.1080/03610926.2017.1350272.

[19] Bai Z., Hu J., Wang C., C. Zhang, "Test on the linear combinations of covariance matrices in high-dimensional data," Statistical Papers, Vol. 62, No. 2, pp. 701–719, 2021. doi: 10.1007/s00362-019-01110-1.

[20] He, Y., Xu, G., Wu, C., W. Pan, "Asymptotically independent U-Statistics in high-dimensional testing," The Annals of Statistics, Vol. 49, No. 1, pp. 154–181, 2021. doi: 10.1214/20-aos1951.

[21] Zi, X., H. Chen, "Robust tests of the equality of two high-dimensional covariance matrices," Communications in Statistics - Theory and Methods, Vol. 51, No. 10, pp. 3120–3141, 2022. doi: 10.1080/03610926.2020.1788085.

[22] Yu, X., Li, D., Xue, L., R. Li, "Power-enhanced simultaneous test of high-dimensional mean vectors and covariance matrices with application to gene-set testing," Journal of the American Statistical Association, Vol. 118, pp. 2548–2561, 2022. doi: 10.1080/01621459.2022.2061354.

[23] Chen, S. X., Guo, B., Y. Qiu, Y., "Testing and signal identification for two-sample high-dimensional covariances via multi-level thresholding," Journal of Econometrics, Vol. 235, pp. 1337–1354, 2023. doi: 10.1016/j.jeconom.2022.10.008.

[24] Ding, X., Hu, Y., Z. Wang, "Two sample test for covariance matrices in ultra-high dimension," 2023. arxiv.org/abs/2312.10796v1.

[25] Yu, X., Li, D., L. Xue, "Fisher's combined probability test for high-dimensional covariance matrices," Journal of the American Statistical Association, Vol. 119, No. 545, pp. 511–524, 2024. doi: 10.1080/01621459.2022.2126781.

[26] Srivastava M. S., "Some tests concerning the covariance matrix in high dimensional data," Journal of the Japan Statistical Society, Vol. 35, No. 2, pp. 251–272, 2005. doi: 10.14490/jjss.35.251.

[27] Notterman D. A., Alon U., Sierk A. J., A. J. Levine, "Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays.," Cancer Res, Vol. 61, No. 7, pp. 3124–3130, Apr. 2001. Online available from http://genomics-pubs.princeton.edu/oncology/. (accessed January 9, 2023).

[28] Hoeffding E., "A class of statistics with asymptotically normal distribution," The Annals of Mathematical Statistics, Vol. 19, No. 3, pp. 293–325, 1948. doi: 10.1214/aoms/1177730196.

[29] Ahmad M. R., Werner C., E. Brunner., "Analysis of high dimensional repeated measures designs: the one sample case," Ph.D. thesis. Göttingen, Germany: Cuvillier Verlag, 2008.