A Test for Testing the Equality of Two Covariance Matrices for High-dimensional Data

Saowapcha Chaipitak and Samruam Chongcharoen
School of Applied Statistics, National Institute of Development Administration,
Bangkapi District, Bangkok, 10240, Thailand

Abstract: This study proposed a test for the equality of two covariance matrices from two independent multivariate normal populations with high-dimensional data. The test statistic is based on unbiased and consistent estimator of the ratio between the sums of squares of covariance matrix elements. Under the null hypothesis, the proposed test statistic is asymptotically standard normal distributed when the number of variables and the sample sizes go together to infinity. Simulation study is conducted to investigate the performance of the proposed test statistic. The results showed that the proposed test is superior to the other three tests appeared in the literature for various patterns of common covariance matrix. Finally, two real data sets are analyzed to illustrate the application of our theoretical results.

Key words: Asymptotic normality, consistent estimator, hypothesis testing, multivariate normal population, significance level

INTRODUCTION

Let $x_{ij} = (x_{ij1}, \ldots, x_{ijn})'$, $j = 1, \ldots, n_i$, $i = 1, 2$, be random samples drawn from independent multivariate normal populations $N_p(\mu_i, \Sigma)$, where all the parameters are unknown. It is a requirement in many statistical techniques, such as in discriminant analysis, testing the equality of two mean vectors, testing the equality of two mean sub-vectors, to know whether covariance matrices of the two populations are equal or not (Johnson, 1998; Krzanowski, 2000; Srivastava, 2002; Gamage and Mathew, 2008; Fujikoshi et al., 2010). Before applying any further analysis, this equality must be tested. The widely used traditional technique for testing the hypothesis that $H_0: \Sigma_1 = \Sigma_2 = \Sigma$ against $H_1: \Sigma_1 \neq \Sigma_2$, where $\Sigma$ is the common unknown covariance matrix of the two populations, when the sample sizes $n_i$ larger than the number of variables, $p$, is the modified likelihood ratio test. However, in applications concerning modern sciences and economics, the data consist of very large number of variables taken from small samples. For instance, DNA microarrays typically measure thousands to millions of gene expressions on the small sample sizes (Dudoit et al., 2002; Ibrahim et al., 2002; Sebastián et al., 2006; Huang et al., 2009). When the data have $p \gg n$, called high-dimensional data, the sample covariance matrices $S_i$ are singular making the modified likelihood ratio test is not valid. The tests under this problem were recently worked by Schott (2007), Srivastava (2007) and Srivastava and Yamagihara (2010). To have more powerful choice of test statistic for testing $H_0$ against $H_1$ when $p \geq n$, a new test statistic is proposed. It is shown that this proposed test statistic is asymptotically distributed as the standard normal distribution for any type of common covariance matrix considered with large $p, n$.

Let $n_i$ be the sample size drawn from population $i$, $i = 1, 2$ and $n = n_1 + n_2 - 2$, the following assumptions are made:

$$\lim_{p \to \infty} p/n_i = c \in (0, \infty)$$

$$\lim_{p \to \infty} p/n_i = c_i \in (0, \infty), i = 1, 2$$

$$\lim_{p \to \infty} a_{p, k} = a_{k} \in (0, \infty), k = 1, \ldots, 16$$

$$\lim_{p \to \infty} a_{p, j} = a_{j} \in (0, \infty), i = 1, 2, j = 1, \ldots, 8$$

where, $a_{p} = \frac{(\text{tr} \Sigma)^2}{p}$ and $a_{j} = \frac{\text{tr} \Sigma_j}{p}$. Let:

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, i = 1, 2$$

$$A_i = \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)'$$

$$S_i = \frac{1}{n_i - 1} A_i, i = 1, 2$$

Corresponding Author: S. Chaipitak, School of Applied Statistics, National Institute of Development Administration, Bangkapi District, Bangkok, 10240, Thailand Tel: 668-1711-0104
\[ \hat{a}_i = \frac{1}{p} \text{tr} S_i, \quad i = 1, 2 \]

and

\[ \hat{a}_3 = \frac{1}{p} \text{tr} S_i - 2, \quad i = 1, 2 \]

\[ r^2 = \frac{1}{n} - \frac{1}{n_1 + n_2} = S \]

Let:

\[ \hat{a}_i = \frac{1}{p} \text{tr} S_i \]

and

\[ \hat{a}_3 = \frac{1}{p} \left( \frac{n^2}{(n-1)(n+2)} \text{tr} S_i - \frac{1}{n} \text{tr} S_i^2 \right) \]

The modified likelihood ratio test suggested by Bartlett (1937) on an intuitive ground is based on the statistic:

\[ L = |S|^{1/2} |S_i|^{1/2} |S_i^2|^{1/2} \]

and is valid when \( p < n \). In particular, if \( p \) is fixed, the asymptotic null distribution of \(-2 \log L\), as \( n \to \infty \), for \( i = 1, 2 \), is chi-squared distribution with \( p(p+1)/2 \) degree of freedom.

Because of the unavailability of the modified likelihood ratio test \( L \) when \( p > n \), Schott (2007) proposed a test for the equality of several covariance matrices. This study then considers Schott’s test statistic only for the case of two covariance matrices. Based on the consistent estimator of the square of Frobenius norm of \( \Sigma_i - \Sigma_2 \), namely \( \text{tr}(\Sigma_i - \Sigma_2)^2 \) his test statistic is given by:

\[ T_2 = \frac{\hat{a}_2 - \hat{a}_3}{\sqrt{\hat{a}_2^2 + \hat{a}_3^2}} \]

where:

\[ \hat{a}_2 = \frac{1}{p} \text{tr} S_i - \text{tr} S_i^2 \]

and

\[ \hat{a}_3 = \frac{1}{c_0} \left( \frac{1}{p} \text{tr} A^4 - p \text{tr} A^2 \text{tr} S_i - p \text{tr} S_i^2 \right) \]

where, \( c_0 = n(n^2 - 6n^2 + 21n + 18) \), \( c_1 = 2n(2n^2 + 6n + 9) \), \( c_2 = 2n(3n^2 + 2) \) and \( c_3 = n(2n^2 + 5n + 7) \). Under the null hypothesis \( T_2 \) is asymptotically distributed as \( N(0, 1) \) as \( (p, n) \to \infty \).

Srivastava and Yanagihara (2010) proposed an alternative test based on a consistent estimator of a measure of distance by \( \gamma_i^2 = \text{tr} A^2 \text{tr} S_i + \text{tr} S_i^2 \), where \( \gamma_i = n \sum_j \text{tr} S_i^j \), \( i = 1, 2 \). The consistent statistic is given by:

\[ T_{\gamma(i)} = \frac{\gamma_i^2 - \gamma_1}{\sqrt{\gamma_2^2 + \gamma_3^2}} \]

where:

\[ \gamma_i = \frac{1}{n(n^2 + 3n + 4)} \left( \frac{1}{p} \text{tr} A^4 - 3n(1+n+4) \text{tr} A^2 \text{tr} S_i - n p \text{tr} S_i^2 \right) \]

and

\[ \hat{a}_i = \frac{1}{n(n^2 + 3n + 4)} \left( \frac{1}{p} \text{tr} A^4 - 3n(1+n+4) \text{tr} A^2 \text{tr} S_i - n p \text{tr} S_i^2 \right) \]

Under the null hypothesis, \( T_{\gamma(i)} \) is asymptotically distributed as \( N(0, 1) \) as \( (p, n) \to \infty \).

**THE PROPOSED STATISTIC**

To test the hypothesis \( H_i: \Sigma_i = \Sigma \) against \( H_i: \Sigma_i \neq \Sigma \) for \( p > n \), it is observed that if \( \Sigma_1 = \Sigma_2 \) then \( \text{tr} \Sigma_i^2 - \text{tr} \Sigma_i^2 = 0 \). Thus under the null hypothesis, the measurement \( b - n \text{tr} \Sigma_i^2 \) is unbiased. Using lemma A3 extended from lemma A1 obtained from Srivastava (2005) in the Appendix, a consistent estimator of \( b \) can be estimated by \( b - \hat{a}_i \). The following lemma gives the asymptotic distribution of the consistent estimators.
Lemma 1: Let \((n-1)S_i - W_i (\Sigma_r n_i -1), \hat{a}_{i1}, i = 1, 2\), as defined in (1) and \(a_i - (\sum_{j=1}^{r} n_j p_j, i = 1, 2, j = 1 \ldots, 4\), then under the assumptions (A2) and (A4):

\[
\begin{pmatrix}
\hat{a}_{i1} \\
\hat{a}_{i2}
\end{pmatrix} \xrightarrow{\text{D}} \begin{pmatrix}
\mathbf{a}_{i1} \\
\mathbf{a}_{i2}
\end{pmatrix} \left( \begin{pmatrix}
\frac{4}{(n_1-1)p_1} \left( 2a_{i1} + \frac{p_{i1}}{n_i-1} \right) & 0 \\
0 & \frac{4}{(n_2-1)p_2} \left( 2a_{i2} + \frac{p_{i2}}{n_i-1} \right)
\end{pmatrix} \right)
\]

where, \(\xrightarrow{\text{D}}\) denotes \(x\) converges in distribution to \(y\).

Proof: Since random samples \(x_j, j = 1, \ldots, n_i, i = 1, 2\) are drawn from two independent populations and sample covariance matrices \(S_r\) are calculated from corresponding independent random samples \(x_j\) and \(x_i\) thus, \(S_1\) and \(S_2\) must be independent of each other. In fact, the statistic \(\hat{a}_{i1}\) is a function of \(S_1\) alone while the statistic \(\hat{a}_{i2}\) is also a function of \(S_2\) alone. Thus \(\hat{a}_{i1}\) and \(\hat{a}_{i2}\) are also independent and then it makes \(\text{COV} (\hat{a}_{i1}, \hat{a}_{i2}) = 0\). By lemma A4 in the Appendix, \(\hat{a}_{i1}, i = 1, 2\) are asymptotically normally distributed with mean \(a_{i1}\) and variance:

\[n_{i1} = \frac{4}{(n_i-1)p_i} \left( 2a_{i1} + \frac{p_{i1}}{n_i-1} \right)\]

and the fact that the covariance between \(\hat{a}_{i1}\) and \(\hat{a}_{i2}\) is zero, it follows that the jointly asymptotic distribution of statistics \(\hat{a}_{i1}\) and \(\hat{a}_{i2}\) are the bivariate normal distribution with mean vector and covariance matrix as given above. The proof is completed.

Note that \(\hat{b} = \frac{\hat{a}_{i1}}{\hat{a}_{i2}}\) is a ratio of two uncorrelated estimators. By the delta method (Lehmann and Romano, 2005), it ensures that a function of two random variables can be approximated as normal distribution. The following theorem establishes the asymptotic normality of the statistic \(\hat{b}\).

Theorem 1: Let \(b\) and \(\hat{b}\) be as defined above. Then, under the assumptions (A1)-(A4), \(\hat{b} \sim \mathcal{N}(b, \delta_i^2)\), where:

\[\delta_i^2 = \frac{1}{\frac{4}{(n_i-1)p_i} \left( 2a_{i1} + \frac{p_{i1}}{n_i-1} \right) + \frac{p_{i2}}{n_i} \left( 2a_{i2} + \frac{p_{i2}}{n_i-1} \right)}\]

Proof: We note that \(\hat{b} = \frac{\hat{a}_{i1}}{\hat{a}_{i2}}\). Hence the partial derivative of \(b(\hat{a}_{i1}, \hat{a}_{i2})\) with respect to \(\hat{a}_{i1}\) is:

\[\left( \frac{\partial b}{\partial \hat{a}_{i1}} \right) = \frac{1}{\hat{a}_{i2}}\]

Similarly, the partial derivative of \(b(\hat{a}_{i1}, \hat{a}_{i2})\) with respect to \(\hat{a}_{i2}\) is:

\[\left( \frac{\partial b}{\partial \hat{a}_{i2}} \right) = \frac{1}{\hat{a}_{i2}}\]

Thus, by applying the delta method, \(\hat{b} \sim \mathcal{N}(b, \delta^2)\) asymptotically with:

\[\delta^2 = \frac{1}{\frac{4}{(n_i-1)p_i} \left( 2a_{i1} + \frac{p_{i1}}{n_i-1} \right) + \frac{p_{i2}}{n_i} \left( 2a_{i2} + \frac{p_{i2}}{n_i-1} \right)}\]

The proof is completed.

Corollary 1: Let \(\hat{b}\) be as defined above. Under \(H_0: \Sigma_1 = \Sigma_2 = \Sigma\) and the assumptions (A1)-(A4), then:

\[T = \frac{\hat{b} - 1}{\sqrt{\frac{\delta^2}{\frac{4}{(n_1-1)p_1} \left( 2a_{i1} + \frac{p_{i1}}{n_i-1} \right) + \frac{p_{i2}}{n_i} \left( 2a_{i2} + \frac{p_{i2}}{n_i-1} \right)}}} \sim \mathcal{N}(0,1)\]

Proof: Under \(H_0\), then \(a_{i1} = a_{i2} = a_1\) and \(a_{i4} = a_{i3} = a_4\). Thus:

\[\delta^2 = \frac{1}{\frac{4}{(n_1-1)p_1} \left( 2a_{i1} + \frac{p_{i1}}{n_i-1} \right) + \frac{p_{i2}}{n_i} \left( 2a_{i2} + \frac{p_{i2}}{n_i-1} \right)}\]

It follows Theorem 1, then the proof is completed.

In order to use \(T\) in practice, we have to estimate \(\delta^2\) involving estimate of \(a_{i1}\) and \(a_{i2}\). Under the null hypothesis and by using consistent estimators of \(a_{i1}\) and \(a_{i2}\) as \(\hat{a}_{i1}\) and \(\hat{a}_{i2}\) as given in lemmas A1 and A5 in the Appendix, respectively and by assumption (A2), we obtained a corresponding consistent estimator of \(\delta^2\) namely \(\hat{\delta}^2\) as:

\[\hat{\delta}^2 = \frac{1}{\frac{4}{(n_1-1)p_1} \left( 2a_{i1} + \frac{p_{i1}}{n_i-1} \right) + \frac{p_{i2}}{n_i} \left( 2a_{i2} + \frac{p_{i2}}{n_i-1} \right)}\]

Thus a test of \(H_0\) is based on the statistic:

\[T = \frac{\hat{b} - 1}{\hat{\delta}}\]

and also its asymptotic null distribution is the standard normal. The proposed test statistic \(T^*\) with \(\alpha\) level of significance rejects \(H_0\) if \(|T^*| > Z_{\alpha/2}\), where \(Z_{\alpha/2}\) denotes the upper \(\alpha/2\) quantile of the standard normal distribution.
SIMULATION STUDY

Here, the performance of the proposed test statistic $T^*$ compared to three tests $T_b$, $T_s$, and $T_{sv}$ was shown through numerical simulation technique. In order to assess being normality of the tests, the Attained Significance Level (ASL) of these tests were simulated and expected to be close to the nominal significance level setting. The empirical powers of these tests in different situations were also performed.

Parameter selection: Independent 10,000 replications of the multivariate normal random datasets were generated using International Mathematics and Statistics Library (IMSL) with multivariate normal random number generator (RNORM) subroutine of Fortran programming language (FORTRAN). The nominal significance level $\alpha$ used was 0.05. Under the null hypothesis, the test statistics $T^*$, $T_b$, $T_s$ and $T_{sv}$ were computed and the proportions of rejection of test statistics under the null hypothesis were recorded, called the Attained Significance Level (ASL). In our work presented here, the ASL under the null hypothesis and corresponding empirical power under the alternative hypothesis were manipulated for following hypotheses in different patterns of covariance matrix setup as follow:

- **Unstructured pattern (UN):** It is defined as $\Sigma = (\sigma_{ij})_{p \times p}$. We considered the hypothesis as follows:
  - $H_0^U$: $\Sigma = \Sigma_0 = U$, against $H_1^U$: $\Sigma = \Sigma_1 = U_1$

where $U_0 = (\sigma_{ij})_{p \times p}$ where $\sigma_{ij} = 1$ if $i = j$; $\sigma_{ij} = (-1)^i \times (0.5)^j$ if $i \neq j$ and $U_1 = (\sigma_{ij})_{p \times p}$ where $\sigma_{ij} = 1$ if $i = j$; $\sigma_{ij} = (-1)^j \times (0.05)^i$ if $i \neq j$.

- **Compound Symmetry pattern (CS):** It is defined as $\Sigma = \sigma \mathbb{I}_p + kl$, where $\sigma > 0$, $k$ is an appropriate constant, $\mathbb{I}_p$ denotes the $p \times p$ identity matrix and $l_p$ denotes the $p \times 1$ vector of ones. The hypothesis was set as:
  - $H_0^c$: $\Sigma = \Sigma_0 = \Sigma_0 = C_0 = 0.99 \mathbb{I}_p + (0.01) \mathbb{I}_p$, against $H_1^c$: $\Sigma = \Sigma_1 = C_0, \Sigma_2 = C_1 = 0.95 \mathbb{I}_p + (0.05) \mathbb{I}_p$

- **Heterogeneous compound symmetry pattern (HCS):** It is defined as $\Sigma = (\sigma_{ij})_{p \times p}$, where $\sigma_{ij} = \sigma^2 > 0$ if $i = j$; $\sigma_{ij} = 0, \rho \rho$ if $i \neq j$, where $\rho$ is the correlation parameter satisfying $|\rho| < 1$. The hypothesis was set as:
  - $H_0^h$: $\Sigma = \Sigma_0 = M_0$, where $M_0$ is matrix in CSH with $\sigma_{ij} = U(5,6)$ if $i = j$, $\rho = 0.5$, against $H_1^h$: $\Sigma = \Sigma_1 = M_1, \Sigma_2 = M_2$, where $M_1$ is matrix in CSH with $\sigma_{ij} = U(4,5)$ if $i = j$, $\rho = 0.4$.

- **Simple pattern (SIM):** It is defined as $\Sigma = \sigma \mathbb{I}_p$. We set the hypothesis testing according to:
  - $H_0^s$: $\Sigma_1 = \Sigma_2 = 2 \mathbb{I}_p$ against $H_1^s$: $\Sigma_1 = 2 \mathbb{I}_p$ and $\Sigma_2 = 1.5 \mathbb{I}_p$
  - $H_0^s$: $\Sigma_1 = \Sigma_2 = \mathbb{I}_p$, against $H_1^s$: $\Sigma_1 = \mathbb{I}_p$ and $\Sigma_2 = \text{Diag}(1,1,1,2, ..., 1,1,1,2)$

RESULTS AND DISCUSSIONS

Table 1 presents the ASL and empirical powers of $T_b$, $T_s$, $T_{sv}$ and $T^*$ when all covariance matrices in the hypothesis were under unstructured pattern (UN). The ASL of the tests $T_b$, $T_s$, and $T_{sv}$ were not close to the nominal significance level 0.05 and much lower than it for all cases considered. The test $T_b$ generally yielded the ASL not close to 0.05 for all cases considered. Moreover, the test $T_s$ gave the ASL around 0.060 when the sample sizes were small, $n_b = n_s = 20$ here and tended to increase when the sample sizes became larger for any $p$. For instance, when $p = 80$, the ASL of $T_s$ was 0.057 (at $n_b = n_s = 20$) and increased to 0.061 (at $n_b = n_s = 80$). From this table, it is observed that the ASL of the proposed test $T^*$ were reasonably close to 0.05 and get better when $p$ and the sample sizes increased. This is clear that the tests $T_b$, $T_s$, and $T_{sv}$ were not reasonable tests whereas the proposed test $T^*$ were. Considering the power of the test, since the competitive tests $T_b$, $T_s$, and $T_{sv}$ were not reasonable tests at this situation, then their empirical powers provided in Table 1 will be skipped. As shown from this table, the empirical powers of the proposed test $T^*$ increased to one when $p$ and the sample sizes increased. In addition, the empirical powers of the proposed test $T^*$ increased for increasing the sample sizes when $p$ is fixed. For instance, when $p = 160$, the empirical power of the proposed test $T^*$ was 0.288 (at $n_b = n_s = 20$) and increased to 0.944 (at $n_b = n_s = 160$).

<table>
<thead>
<tr>
<th>$p$</th>
<th>$n_b = n_s$</th>
<th>$T_b$</th>
<th>$T_s$</th>
<th>$T_{sv}$</th>
<th>$T^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.061</td>
<td>0.055</td>
<td>0.062</td>
<td>0.058</td>
<td>0.0617</td>
</tr>
<tr>
<td>40</td>
<td>0.064</td>
<td>0.059</td>
<td>0.064</td>
<td>0.060</td>
<td>0.0657</td>
</tr>
<tr>
<td>80</td>
<td>0.065</td>
<td>0.060</td>
<td>0.069</td>
<td>0.062</td>
<td>0.0656</td>
</tr>
<tr>
<td>160</td>
<td>0.066</td>
<td>0.065</td>
<td>0.071</td>
<td>0.071</td>
<td>0.0657</td>
</tr>
</tbody>
</table>

Table 1: ASL of $T_b$, $T_s$, $T_{sv}$ and $T^*$ under $H_0^U_i \Sigma_1 = \Sigma_2 = \mathbb{I}_p$, and their empirical powers under $H_1^U_i \Sigma_1 = \mathbb{I}_p$, and $\Sigma_2 = \mathbb{I}_p$ applied at $\alpha = 0.05$.
Table 2: ASL of $T_2$, $T_3$, $T_{st}$ and $T^*$ under $H_0: \sum_{i=1}^n b_i = c_0$ and their empirical powers under $H_1: \sum_{i=1}^n b_i \neq c_0$ and $\sum_{i=1}^n b_i > c_0$ applied at $\alpha = 0.05$

<table>
<thead>
<tr>
<th>$p$</th>
<th>$n_1 = n_2$</th>
<th>$T_2$</th>
<th>$T_3$</th>
<th>$T_{st}$</th>
<th>$T^*$</th>
<th>$T_2$</th>
<th>$T_3$</th>
<th>$T_{st}$</th>
<th>$T^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.057</td>
<td>0.061</td>
<td>0.084</td>
<td>0.056</td>
<td>0.079</td>
<td>0.081</td>
<td>0.082</td>
<td>0.080</td>
<td>0.080</td>
</tr>
<tr>
<td>40</td>
<td>0.055</td>
<td>0.086</td>
<td>0.081</td>
<td>0.052</td>
<td>0.099</td>
<td>0.076</td>
<td>0.077</td>
<td>0.114</td>
<td>0.114</td>
</tr>
<tr>
<td>60</td>
<td>0.052</td>
<td>0.053</td>
<td>0.078</td>
<td>0.050</td>
<td>0.109</td>
<td>0.086</td>
<td>0.103</td>
<td>0.145</td>
<td>0.145</td>
</tr>
<tr>
<td>80</td>
<td>0.047</td>
<td>0.082</td>
<td>0.081</td>
<td>0.056</td>
<td>0.163</td>
<td>0.065</td>
<td>0.068</td>
<td>0.210</td>
<td>0.210</td>
</tr>
<tr>
<td>100</td>
<td>0.051</td>
<td>0.052</td>
<td>0.059</td>
<td>0.052</td>
<td>0.324</td>
<td>0.176</td>
<td>0.249</td>
<td>0.333</td>
<td>0.333</td>
</tr>
<tr>
<td>120</td>
<td>0.053</td>
<td>0.049</td>
<td>0.045</td>
<td>0.053</td>
<td>0.654</td>
<td>0.473</td>
<td>0.778</td>
<td>0.588</td>
<td>0.588</td>
</tr>
<tr>
<td>160</td>
<td>0.048</td>
<td>0.059</td>
<td>0.057</td>
<td>0.052</td>
<td>0.286</td>
<td>0.047</td>
<td>0.046</td>
<td>0.387</td>
<td>0.387</td>
</tr>
</tbody>
</table>

ASL: The attained significance level, $C_1$ and $C_2$: The matrices defined under compound symmetry pattern (CS): $\alpha$: The nominal significance level.

Table 3: ASL of $T_2$, $T_3$, $T_{st}$ and $T^*$ under $H_0: \sum_{i=1}^n b_i = c_0$ and their empirical powers under $H_1: \sum_{i=1}^n b_i < c_0$ and $\sum_{i=1}^n b_i > c_0$ applied at $\alpha = 0.05$

<table>
<thead>
<tr>
<th>$p$</th>
<th>$n_1 = n_2$</th>
<th>$T_2$</th>
<th>$T_3$</th>
<th>$T_{st}$</th>
<th>$T^*$</th>
<th>$T_2$</th>
<th>$T_3$</th>
<th>$T_{st}$</th>
<th>$T^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.060</td>
<td>0.004</td>
<td>0.000</td>
<td>0.058</td>
<td>0.088</td>
<td>0.070</td>
<td>0.000</td>
<td>0.478</td>
<td>0.478</td>
</tr>
<tr>
<td>40</td>
<td>0.055</td>
<td>0.001</td>
<td>0.000</td>
<td>0.058</td>
<td>0.078</td>
<td>0.054</td>
<td>0.000</td>
<td>0.551</td>
<td>0.551</td>
</tr>
<tr>
<td>60</td>
<td>0.054</td>
<td>0.005</td>
<td>0.000</td>
<td>0.053</td>
<td>0.115</td>
<td>0.480</td>
<td>0.000</td>
<td>0.851</td>
<td>0.851</td>
</tr>
<tr>
<td>80</td>
<td>0.052</td>
<td>0.001</td>
<td>0.000</td>
<td>0.060</td>
<td>0.076</td>
<td>0.032</td>
<td>0.000</td>
<td>0.579</td>
<td>0.579</td>
</tr>
<tr>
<td>100</td>
<td>0.056</td>
<td>0.002</td>
<td>0.000</td>
<td>0.052</td>
<td>0.121</td>
<td>0.435</td>
<td>0.000</td>
<td>0.862</td>
<td>0.862</td>
</tr>
<tr>
<td>120</td>
<td>0.059</td>
<td>0.011</td>
<td>0.000</td>
<td>0.049</td>
<td>0.219</td>
<td>0.943</td>
<td>0.000</td>
<td>0.989</td>
<td>0.989</td>
</tr>
<tr>
<td>160</td>
<td>0.058</td>
<td>0.010</td>
<td>0.005</td>
<td>0.050</td>
<td>0.217</td>
<td>0.801</td>
<td>0.005</td>
<td>0.946</td>
<td>0.946</td>
</tr>
</tbody>
</table>

ASL: The attained significance level, $C_1$ and $C_2$: The matrices defined under heterogeneous compound symmetry pattern (CSH): $\alpha$: The nominal significance level.

Table 4: ASL of $T_2$, $T_3$, $T_{st}$ and $T^*$ under $H_0: \sum_{i=1}^n b_i = 21$ and their empirical powers under $H_1: \sum_{i=1}^n b_i \neq 21$ applied at $\alpha = 0.05$

<table>
<thead>
<tr>
<th>$p$</th>
<th>$n_1 = n_2$</th>
<th>$T_2$</th>
<th>$T_3$</th>
<th>$T_{st}$</th>
<th>$T^*$</th>
<th>$T_2$</th>
<th>$T_3$</th>
<th>$T_{st}$</th>
<th>$T^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.057</td>
<td>0.005</td>
<td>0.000</td>
<td>0.058</td>
<td>0.117</td>
<td>0.268</td>
<td>0.000</td>
<td>0.757</td>
<td>0.757</td>
</tr>
<tr>
<td>40</td>
<td>0.054</td>
<td>0.001</td>
<td>0.000</td>
<td>0.051</td>
<td>0.105</td>
<td>0.309</td>
<td>0.000</td>
<td>0.870</td>
<td>0.870</td>
</tr>
<tr>
<td>60</td>
<td>0.054</td>
<td>0.003</td>
<td>0.000</td>
<td>0.052</td>
<td>0.205</td>
<td>0.962</td>
<td>0.000</td>
<td>0.988</td>
<td>0.988</td>
</tr>
<tr>
<td>80</td>
<td>0.048</td>
<td>0.000</td>
<td>0.000</td>
<td>0.056</td>
<td>0.086</td>
<td>0.297</td>
<td>0.000</td>
<td>0.936</td>
<td>0.936</td>
</tr>
<tr>
<td>100</td>
<td>0.051</td>
<td>0.001</td>
<td>0.000</td>
<td>0.051</td>
<td>0.198</td>
<td>0.992</td>
<td>0.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>120</td>
<td>0.056</td>
<td>0.005</td>
<td>0.000</td>
<td>0.053</td>
<td>0.495</td>
<td>1.000</td>
<td>0.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>160</td>
<td>0.049</td>
<td>0.000</td>
<td>0.000</td>
<td>0.052</td>
<td>0.086</td>
<td>0.211</td>
<td>0.000</td>
<td>0.964</td>
<td>0.964</td>
</tr>
<tr>
<td>200</td>
<td>0.053</td>
<td>0.005</td>
<td>0.000</td>
<td>0.050</td>
<td>0.185</td>
<td>0.999</td>
<td>0.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>240</td>
<td>0.058</td>
<td>0.000</td>
<td>0.000</td>
<td>0.051</td>
<td>0.391</td>
<td>1.000</td>
<td>0.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

ASL: The attained significance level, $\alpha$: The nominal significance level.

The covariance matrix was used under CSH and UN patterns. As displayed in this table, the empirical powers of the proposed test $T^*$ rapidly converged to one when $p$ and the sample sizes increased.

Table 4 reports the ASL of tests $T_2$, $T_3$, $T_{st}$ and $T^*$ under the common covariance matrix $\Sigma = 2I$ (simple pattern) and their empirical powers under $\Sigma = 2I$ and $\Sigma = 1.5I$. As expected, the ASL of $T^*$ and $T_2$ were quite close to 0.05 for all cases considered while those of $T_3$ and $T_{st}$ seemed to be zero for all cases considered. The empirical powers of the proposed test $T^*$ and $T_2$ test converged to one as $p$ and the sample sizes increased. The convergence to one of the empirical powers of the proposed test $T^*$ was extremely faster than that of $T_2$, especially when $n_1 = n_2 = 40$ for all $p$. For example, when $p = n_1 = n_2 = 20$, the empirical powers of $T^*$ and $T_2$ were 0.757 and 0.117, respectively. This indicates that, under simple pattern, $T^*$ was a reasonable test and more powerful than $T_2$ test, particularly in case of small samples.

Table 5 presents the ASL of $T_2$, $T_3$, $T_{st}$ and $T^*$ under the common covariance matrix $\Sigma = I$ (simple pattern) and empirical powers under $\Sigma = I$ and a certain matrix $\Sigma = \text{Diag}(1,1,1,2,\ldots,1,1,1,2)$. As displayed in this table,
Table 5: ASL of $T_1$, $T_2$, $T_{27}$ and $T^*$ under $H_0: \Sigma_1 = \Sigma_2 = \Sigma = I$ and their empirical powers under $H_1: \Sigma_1 = I_2$ and $\Sigma = \text{Diag}(1,1,1,2,\ldots,1,1,1,2)$ applied at $\alpha = 0.05$

<table>
<thead>
<tr>
<th>p</th>
<th>$n_1$</th>
<th>$n_2$</th>
<th>$T_1$</th>
<th>$T_2$</th>
<th>$T_{27}$</th>
<th>$T^*$</th>
<th>Empirical power</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>20</td>
<td>0.057</td>
<td>0.092</td>
<td>0.084</td>
<td>0.058</td>
<td>0.243</td>
<td>0.422</td>
</tr>
<tr>
<td>40</td>
<td>20</td>
<td>0.054</td>
<td>0.090</td>
<td>0.087</td>
<td>0.051</td>
<td>0.241</td>
<td>0.533</td>
</tr>
<tr>
<td>40</td>
<td>40</td>
<td>0.054</td>
<td>0.084</td>
<td>0.084</td>
<td>0.052</td>
<td>0.564</td>
<td>0.976</td>
</tr>
<tr>
<td>80</td>
<td>20</td>
<td>0.048</td>
<td>0.093</td>
<td>0.090</td>
<td>0.055</td>
<td>0.223</td>
<td>0.581</td>
</tr>
<tr>
<td>40</td>
<td>40</td>
<td>0.051</td>
<td>0.061</td>
<td>0.073</td>
<td>0.051</td>
<td>0.573</td>
<td>0.997</td>
</tr>
<tr>
<td>80</td>
<td>40</td>
<td>0.056</td>
<td>0.054</td>
<td>0.057</td>
<td>0.052</td>
<td>0.969</td>
<td>1.000</td>
</tr>
<tr>
<td>160</td>
<td>20</td>
<td>0.049</td>
<td>0.082</td>
<td>0.075</td>
<td>0.052</td>
<td>0.213</td>
<td>0.538</td>
</tr>
<tr>
<td>40</td>
<td>40</td>
<td>0.053</td>
<td>0.062</td>
<td>0.071</td>
<td>0.050</td>
<td>0.561</td>
<td>0.999</td>
</tr>
<tr>
<td>80</td>
<td>40</td>
<td>0.048</td>
<td>0.052</td>
<td>0.055</td>
<td>0.051</td>
<td>0.971</td>
<td>1.000</td>
</tr>
<tr>
<td>160</td>
<td>40</td>
<td>0.051</td>
<td>0.049</td>
<td>0.051</td>
<td>0.050</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

ASL: The attained significance level, $\alpha$: The nominal significance level

Table 6: ASL of $T_1$, $T_2$, $T_{27}$ and $T^*$ under $H_0: \Sigma_1 = \Sigma_2 = \Sigma = I$ and $\Sigma = \text{Diag}(1,1,1,2,\ldots,1,1,1,2)$ and their empirical powers under $H_1: \Sigma_2 = I_2$ and $\Sigma = \text{Diag}(1,1,1,2,\ldots,1,1,1,2)$ applied at $\alpha = 0.05$

<table>
<thead>
<tr>
<th>p</th>
<th>$n_1$</th>
<th>$n_2$</th>
<th>$T_1$</th>
<th>$T_2$</th>
<th>$T_{27}$</th>
<th>$T^*$</th>
<th>Empirical power</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>20</td>
<td>0.055</td>
<td>0.064</td>
<td>0.079</td>
<td>0.052</td>
<td>0.265</td>
<td>0.659</td>
</tr>
<tr>
<td>80</td>
<td>20</td>
<td>0.053</td>
<td>0.087</td>
<td>0.096</td>
<td>0.052</td>
<td>0.284</td>
<td>0.784</td>
</tr>
<tr>
<td>40</td>
<td>80</td>
<td>0.053</td>
<td>0.051</td>
<td>0.056</td>
<td>0.049</td>
<td>0.751</td>
<td>1.000</td>
</tr>
<tr>
<td>160</td>
<td>40</td>
<td>0.049</td>
<td>0.060</td>
<td>0.067</td>
<td>0.047</td>
<td>0.254</td>
<td>0.846</td>
</tr>
<tr>
<td>40</td>
<td>80</td>
<td>0.052</td>
<td>0.052</td>
<td>0.056</td>
<td>0.050</td>
<td>0.747</td>
<td>1.000</td>
</tr>
<tr>
<td>80</td>
<td>160</td>
<td>0.055</td>
<td>0.050</td>
<td>0.054</td>
<td>0.051</td>
<td>0.999</td>
<td>1.000</td>
</tr>
</tbody>
</table>

ASL: The attained significance level, $\alpha$: The nominal significance level

the ASL of the proposed test $T^*$ and $T_1$ test were similar to those from Table 4 and reasonable approximate 0.05 for all cases of $p$ and the sample sizes. This means that changing the scalar $\sigma^2$ defined in simple pattern, from $\sigma^2 = 2$ became $\sigma^2 = 1$, is not affected to the convergence of the asymptotic normality of the proposed test $T^*$ and $T_1$ test. But it is greatly affected to the convergence of the asymptotic normality of $T_2$ and $T_{27}$ because ASL of both tests when $\Sigma - I$ were much better than those when $\Sigma = 2I$ for all case considered. However, the ASL of the tests $T_1$ and $T_{27}$ mainly were still not control 0.05, particularly when the sample sizes were less than or equal to 40 for any $p$. As expected, the empirical powers of the tests $T_1$ and $T^*$ quickly tended to one as $p$ and the sample sizes increased. In addition, the proposed test $T^*$ generally gave the higher power than $T_1$ test.

We carried out additional simulations for the case that the sample sizes were not equal ($n_1 \neq n_2$) choosing $n_2 = 2n_1$ of four tests $T_1$, $T_2$, $T_{27}$ and $T^*$ under the null hypothesis $H_0: \Sigma_1 = \Sigma_2 = \Sigma = I_2$. Corresponding empirical powers of these tests were also manipulated under the alternative hypothesis $H_1: \Sigma_2 = I_2$ and $\Sigma = \text{Diag}(1,1,1,2,\ldots,1,1,1,2)$. The results are provided in Table 6.

Table 6 presents that both ASL and empirical powers of these tests were not substantially different from those given in Table 5. It appears that the tests $T_2$ and $T_{27}$ still had the ASL not close to 0.05, particularly for the small sample sizes, $n_1 = 20$ and $n_2 = 40$ here. The proposed test statistic $T^*$ and $T_1$ test remained appropriate even the sample sizes were not the same. The empirical powers of the proposed test $T^*$ maintained better than and converged to one faster than those of $T_1$ test.

APPLICATION

In this section, the dataset from Noterman et al. (2001) is online at http://genomics-pubs.princeton.edu/oncology/Data/CarcinomaNormalDatasetCancerResearch.txt (last accessed: 9 October 2012). Two groups of colon tissues (adenocarcinoma and adenoma) were examined by oligonucleotide arrays. The expression levels about 6500 human genes were probed in 18 colon adenocarcinomas and 4 colon adenomas. We restricted attention to a subset of all gene expressions of 100 expression levels on 4 colon adenocarcinomas and 4 colon adenomas. Thus we had $n_1 = 4$, $n_2 = 4$ and $p = 100$. We examined whether the covariance matrices of the two groups are equal. The data presented the observed test statistic values of $T_1 = 0.908$ and $T^* = -0.636$. Corresponding p-values were 0.182 and 0.524 indicating the hypothesis of equality of such two covariance matrices of these data was not rejected at any reasonable significance level.

CONCLUSIONS

In this study, we proposed an alternative test statistic for testing the equality of two covariance matrices for two independent multivariate normal data with $p \geq n_i$, $i = 1,2$. The test statistic $T^*$ based on the consistent estimators is introduced. Its asymptotic distribution approximately follows the standard normal distribution as $p/n_i \to \infty$ even if $p/n_i \sim c \in (0, \infty)$, $i = 1,2$. The simulation results strongly supported the performance of the proposed test statistic $T^*$ that it accurately control size of test and not greatly affected by changing the common covariance matrix appearing in the null hypothesis. As seen in the simulation study, the proposed test statistic $T^*$ has the highest power among competitive test statistics, $T_1$, which is a special case of the test for testing the equality of several covariance matrices proposed by Schott (2007), $T_2$ and $T_{27}$ given by Srivastava (2007) and Srivastava and Yanagihara (2010).

ACKNOWLEDGMENT

We would like to thank the Commission on Higher Education (CHE) of Thailand for financial support through a grant fund under the Strategic Scholarships Fellowships Frontier Research Networks.
APPENDIX

Most of work in this study could be viewed as an extension some results of Srivastava (2005) and Fisher et al. (2010). In order to proof Lemma 1 we have taken the following two useful lemmas (lemma A1 and A2) from Srivastava (2005).

Lemma A1: Let \( nS=W_{p}(\Sigma,n) \) and \( a_{k}=(\tau \Sigma^{k})/\rho \), \( k=1,...,4 \). Then under the assumptions (A1) and (A3), unbiased and consistent estimators of \( a_{k} \) as \( (p,n) \rightarrow \infty \) is given by \( \hat{a}_{k} \).

Lemma A2: Let \( nS=W_{p}(\Sigma,n) \), \( \hat{a}_{k} \) as defined in (2) and \( a_{k}=(\tau \Sigma^{k})/\rho \), \( k=1,...,4 \). Then under the assumptions (A1) and (A3):

\[
\lim_{(p,n) \rightarrow \infty} P[(\hat{a}_{k} - a_{k})/\eta_{k} \leq x] = \Phi(x)
\]

where, \( \Phi(x) \) denotes the cumulative distribution function of a standard normal random variable and:

\[
\eta_{k} = \frac{4}{np} \left( 2a_{k} + \frac{p\hat{a}_{k}^{2}}{a_{k}} \right)
\]

The extensions of lemmas A1 and A2 can be obtained without proofs as follow:

Lemma A3: Let \( \hat{a}_{k}=(nS_{j}-W_{p}(\Sigma,n)) \) and \( a_{k}=(\tau \Sigma^{k})/\rho \), \( i=1,2 \), \( j=1,...,4 \). Then under the assumptions (A2) and (A4), unbiased and consistent estimators of \( a_{k} \) as \( (p,n) \rightarrow \infty \) are given by \( \hat{a}_{k,i} \), \( i=1,2 \).

Lemma A4: Let \( \hat{a}_{k}=(nS_{j}-W_{p}(\Sigma,n)) \) and \( a_{k}=(\tau \Sigma^{k})/\rho \), \( i=1,2 \), \( j=1,...,4 \). Then under the assumptions (A2) and (A4):

\[
\lim_{(p,n) \rightarrow \infty} P[(\hat{a}_{k} - a_{k})/\eta_{k} \leq x] = \Phi(x)
\]

where \( \Phi(x) \) denotes the cumulative distribution function of a standard normal random variable and:

\[
\eta_{k} = \frac{4}{(n_{i}-1)p} \left( 2a_{k} + \frac{p\hat{a}_{k}^{2}}{a_{k}} \right)
\]

The following lemma is taken from Fisher et al. (2010). Thus it also is presented without proof.

Lemma A5: Let \( nS=W_{p}(\Sigma,n) \) and \( a_{k}=(\tau \Sigma^{k})/\rho \), \( k=1,...,16 \). Then under the assumptions (A1) and (A3), unbiased and consistent estimators of \( a_{k} \) as \( (p,n) \rightarrow \infty \) is given by \( \hat{a}_{k} \) defined as:

\[
\hat{a}_{k} = \frac{\tau}{p} \left[ \tau a_{k}^{2} + b_{k} \tau a_{k}^{3} + c_{k} (\tau a_{k})^{2} + d_k (\tau a_{k}^{2} + e_k (\tau a_{k})^{2} \right]
\]

where:

\[
\tau = \frac{n^{2}(n^{2}+n+2)}{(n+1)(n+2)(n+4)(n+6)(n-1)(n-2)(n-3)}
\]

\[
b_{k} = \frac{2^{k} - 3}{n} \sqrt{\frac{n(n+2)}{n^{2}+n+2}} \]

\[
c_{k} = \frac{2^{k} - 3n}{n(n+2)}
\]

\[
d_{k} = \frac{2^{k} - 3n}{n(n+2)}
\]

\[
e_{k} = \frac{2^{k} - 3n}{n(n+2)}
\]

\[
\tau = \frac{n^{2}+n+2}{n^{2}+n+2}
\]

REFERENCES


