Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 13 (2): 270-277, 2013
ISSN 1812-5654 / DOI: 10.3923/jas.2013.270.277

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

Saowapha 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 $x_{ij} = (x_{ij1}, ..., x_{ijp})'$, $j = 1, ..., n_i$, $i = 1,2$, be random samples drawn from independent multivariate normal populations $N_p(\mu_i, \Sigma_i)$, 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 Σ 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; Sebastiani *et al*., 2006; Huang *et al*., 2009). When the data have $p \geq n_i$, 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 Yanagihara (2010). To have more powerful choice of test statistic for testing H_0 against H_1 when $p \geq n_i$, 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_i.

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:

$$(A1)\ \lim_{(p,n)\to\infty} p/n = c \in (0,\infty)$$

$$(A2)\ \lim_{(p,n_i)\to\infty} p/n_i = c_i \in (0,\infty),\ i = 1,2$$

$$(A3)\ \lim_{p\to\infty} a_k = \alpha_k \in (0,\infty),\ k = 1,...,16$$

$$(A4)\ \lim_{p\to\infty} a_{ji} = \alpha_{ji} \in (0,\infty),\ i = 1,2,\ j = 1,...,8$$

where, $a_k = (tr\Sigma^k)/p$ and $a_{ji} = (tr\Sigma_i^j)/p$.
Let:

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

$$A_i = \sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)',\ i = 1,2$$

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

Corresponding Author: S. Chaipitak, School of Applied Statistics, National Institute of Development Administration, Bangkapi District, Bangkok, 10240, Thailand Tel: 668-1711-0104

$$\hat{a}_{1i} = \frac{1}{p}trS_i,\ i = 1, 2$$

and

$$\hat{a}_{2i} = \frac{(n_i - 1)^2}{p(n_i - 2)(n_i + 1)}\left\{trS_i^2 - \frac{1}{n_i - 1}(trS_i)^2\right\},\ i = 1, 2 \qquad (1)$$

Since S_1 and S_2 are independent estimates of the covariance matrices Σ_1 and Σ_2, respectively, with $(n_i-1)\ S_i \sim W_p\ (\Sigma_i, n_i-1)$, i = 1, 2, where $W_p\ (\Sigma_i, n_i-1)$ is a Wishart distribution with n_i-1 degree of freedom and covariance matrix Σ_i then the common covariance matrix Σ can be estimated by:

$$\hat{\Sigma} = \frac{1}{n}A = \frac{1}{n_1 + n_2 - 2}(A_1 + A_2) \equiv S$$

Let:

$$\hat{a}_1 = \frac{1}{p}trS$$

and

$$\hat{a}_2 = \frac{n^2}{p(n-1)(n+2)}\left\{trS^2 - \frac{1}{n}(trS)^2\right\} \qquad (2)$$

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

$$L = |S_1|^{(n_1-1)/2}|S_2|^{(n_2-1)/2}/|S|^{n/2}$$

and is valid when $p<n_i$. In particularly, if p is fixed, the asymptotic null distribution of -2 log L, as $n_i \to \infty$, 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 $p \geq n_i$, 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 Σ_1-Σ_2, namely $tr(\Sigma_1-\Sigma_2)^2$ his test statistic is given by:

$$T_J = \frac{(n_1 - 1)(n_2 - 1)}{2(n_1 + n_2 - 2)\hat{a}_2}\left(\hat{a}_{21} + \hat{a}_{22} - \frac{2}{p}tr(S_1S_2)\right)$$

Under the null hypothesis, T_J is asymptotically distributed as N (0, 1) as $(p, n_1, n_2) \to \infty$.

Srivastava (2007) proposed a test based on a lower bound on Frobenius norm. It is given by:

$$T_S = \frac{\hat{a}_{21} - \hat{a}_{22}}{\sqrt{\hat{\eta}_1^2 + \hat{\eta}_2^2}}$$

where:

$$\hat{\eta}_i^2 = \frac{4}{(n_i - 1)^2}\hat{a}_2^2\left(1 + \frac{2(n_i - 1)\hat{a}_4}{p\hat{a}_2^2}\right),\ i = 1, 2$$

and

$$\hat{a}_4 = \frac{1}{c_0}\left(\frac{1}{p}trA^4 - pc_1\hat{a}_1 - p^2c_2\hat{a}_1^2\hat{a}_2 - pc_3\hat{a}_2^2 - np^3\hat{a}_1^4\right)$$

where, $c_0 = n\ (n^3+6n^2+21n+18)$, $c_1 = 2n\ (2n^2+6n+9)$, $c_2 = 2n\ (3n+2)$ and $c_3 = n\ (2n^2+5n+7)$. Under the null hypothesis T_S 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 $\gamma_1 - \gamma_2 = tr\sum_1^2/(tr\sum_1)^2 - tr\sum_2^2/(tr\sum_2)^2$, where $\gamma_i = tr\sum_i^2/(tr\sum_i)^2$, i = 1, 2. The consistent estimators of γ_i are given by $\hat{\gamma}_i, \hat{\gamma}_i = \hat{a}_{2i}/\hat{a}_{1i}^2, i = 1, 2$. The test statistic is given by:

$$T_{SY} = \frac{\hat{\gamma}_1 - \hat{\gamma}_2}{\sqrt{\hat{\xi}_1^2 + \hat{\xi}_2^2}}$$

where:

$$\hat{\xi}_i^2 = \frac{4}{(n_i - 1)^2}\left\{\frac{\hat{a}_2^2}{\hat{a}_1^4} + \frac{2(n_i - 1)}{p}\left(\frac{\hat{a}_2^3}{\hat{a}_1^6} - \frac{2\hat{a}_2\hat{a}_3}{\hat{a}_1^5} + \frac{\hat{a}_4}{\hat{a}_1^4}\right)\right\},\ i = 1, 2$$

and

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

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

THE PROPOSED STATISTIC

To test the hypothesis H_0: $\Sigma_1 = \Sigma_2 = \Sigma$ against H_1: $\Sigma_1 \neq \Sigma_2$ for $p \geq n_i$ it is observed that if $\Sigma_1 = \Sigma_2$, then $tr\sum_i^2 = tr\sum_2^2$. Thus under the null hypothesis, the measurement $b = tr\sum_1^2/tr\sum_2^2 = 1$. Using lemma A3 extended from lemma A1 obtained from Srivastava (2005) in the Appendix, a consistent estimator of b can be estimated by $\hat{b} = \hat{a}_{21}/\hat{a}_{22}$. The following lemma gives the asymptotic distribution of the consistent estimators.

Lemma 1: Let $(n_i-1)\,S_i \sim W_p(\Sigma_i, n_i-1)$, $\hat{a}_{2i}$, $i = 1, 2$, as defined in (1) and $a_{ji} = (tr\sum_i^j)/p, i = 1, 2, j = 1, \ldots 4$, then under the assumptions (A2) and (A4):

$$\begin{pmatrix}\hat{a}_{21}\\ \hat{a}_{22}\end{pmatrix} \xrightarrow{D} N_2\left(\begin{pmatrix}a_{21}\\ a_{22}\end{pmatrix}, \begin{pmatrix}\frac{4}{(n_1-1)p}\left(2a_{41}+\frac{pa_{21}^2}{n_1-1}\right) & 0\\ 0 & \frac{4}{(n_2-1)p}\left(2a_{42}+\frac{pa_{22}^2}{n_2-1}\right)\end{pmatrix}\right)$$

where, $x \xrightarrow{D} y$ denotes x converges in distribution to y.

Proof: Since random samples x_{ij}, $j = 1, \ldots, n_i$, $i = 1,2$ are drawn from two independent populations and sample covariance matrices S_i are calculated from corresponding independent random samples x_{1j} and x_{2j} thus, S_1 and S_2 must be independent of each other. In fact, the statistic $\hat{a}_{21}$ is a function of S_1 alone while the statistic $\hat{a}_{22}$ is also a function of S_2 alone. Thus $\hat{a}_{21}$ and $\hat{a}_{22}$ are also independent and then it makes COV $(\hat{a}_{21}, \hat{a}_{22}) = 0$. By lemma A4 in the Appendix, $\hat{a}_{2i}$, $i = 1, 2$ are asymptotically normally distributed with mean a_{2i} and variance:

$$\eta_i^2 = \frac{4}{(n_i-1)p}\left(2a_{4i}+\frac{pa_{2i}^2}{n_i-1}\right)$$

and the fact that the covariance between $\hat{a}_{21}$ and $\hat{a}_{22}$ is zero, it follows that the jointly asymptotic distribution of statistics $\hat{a}_{21}$ and $\hat{a}_{22}$ are the bivariate normal distribution with mean vector and covariance matrix as given above. The proof is completed.

Note that $\hat{b} = \hat{a}_{21}/\hat{a}_{22}$ 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), $\hat{b} \xrightarrow{D} N(b, \delta^2)$, where:

$$\delta^2 = \frac{4}{pa_{22}^2}\left\{\frac{1}{(n_1-1)}\left(2a_{41}+\frac{pa_{21}^2}{n_1-1}\right)+\frac{a_{21}^2}{(n_2-1)a_{22}^2}\left(2a_{42}+\frac{pa_{22}^2}{n_2-1}\right)\right\}$$

Proof: We note that $\hat{b} = \hat{a}_{21}/\hat{a}_{22}$. Hence the partial derivative of $\hat{b}(\hat{a}_{21}, \hat{a}_{22})$ with respect to $\hat{a}_{21}$ is:

$$\left(\frac{\partial\hat{b}}{\partial\hat{a}_{21}}\right) = \frac{1}{\hat{a}_{22}}$$

Similarly, the partial derivative of $\hat{b}(\hat{a}_{21}, \hat{a}_{22})$ with respect to $\hat{a}_{22}$ is:

$$\left(\frac{\partial\hat{b}}{\partial\hat{a}_{22}}\right) = -\frac{\hat{a}_{21}}{\hat{a}_{22}^2}$$

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

$$\begin{aligned}\delta^2 &= \begin{pmatrix}\frac{1}{a_{22}} & -\frac{a_{21}}{a_{22}^2}\end{pmatrix}\begin{pmatrix}\frac{4}{(n_1-1)p}\left(2a_{41}+\frac{pa_{21}^2}{n_1-1}\right) & 0\\ 0 & \frac{4}{(n_2-1)p}\left(2a_{42}+\frac{pa_{22}^2}{n_2-1}\right)\end{pmatrix}\begin{pmatrix}\frac{1}{a_{22}}\\ -\frac{a_{21}}{a_{22}^2}\end{pmatrix}\\ &= \frac{1}{(n_1-1)pa_{22}^2}4\left(2a_{41}+\frac{pa_{21}^2}{n_1-1}\right)+\frac{a_{21}^2}{(n_2-1)pa_{22}^4}4\left(2a_{42}+\frac{pa_{22}^2}{n_2-1}\right)\\ &= \frac{4}{pa_{22}^2}\left\{\frac{1}{(n_1-1)}\left(2a_{41}+\frac{pa_{21}^2}{n_1-1}\right)+\frac{a_{21}^2}{(n_2-1)a_{22}^2}\left(2a_{42}+\frac{pa_{22}^2}{n_2-1}\right)\right\}\end{aligned}$$

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:

$$T = \frac{\hat{b}-1}{2\left\{\frac{1}{p}\left[\frac{2a_4}{a_2^2}\sum_{i=1}^{2}\frac{1}{(n_i-1)}+\sum_{i=1}^{2}\frac{p}{(n_i-1)^2}\right]\right\}^{\frac{1}{2}}} \xrightarrow{D} N(0,1)$$

Proof: Under H_0, then $a_{21} = a_{22} = a_2$ and $a_{41} = a_{42} = a_4$. Thus:

$$\delta^2 \overset{H_0}{=} \frac{4}{p}\left\{\frac{2a_4}{a_2^2}\sum_{i=1}^{2}\frac{1}{(n_i-1)}+\sum_{i=1}^{2}\frac{p}{(n_i-1)^2}\right\}$$

It follows Theorem 1, then the proof is completed.

In order to use T in practice, we have to estimate δ^2 involving estimate of a_2 and a_4. Under the null hypothesis and by using consistent estimators of a_2 and a_4 as $\hat{a}_2$ and $\hat{a}_4^*$ as given in lemmas A1 and A5 in the Appendix, respectively and by assumption (A2), we obtained a corresponding consistent estimator of δ^2 namely $\hat{\delta}^2$ as:

$$\hat{\delta}^2 = \frac{4}{p}\left\{\frac{2\hat{a}_4^*}{\hat{a}_2^2}\sum_{i=1}^{2}\frac{1}{(n_i-1)}+\sum_{i=1}^{2}\frac{p}{(n_i-1)^2}\right\} = 4\left\{\frac{2\hat{a}_4^*}{p\hat{a}_2^2}\sum_{i=1}^{2}\frac{1}{(n_i-1)}+\sum_{i=1}^{2}\frac{1}{(n_i-1)^2}\right\}$$

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 α 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_J, T_S and T_{SY} 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 (RNMVN) subroutine of Fortran programming language (FORTRAN). The nominal significance level α used was 0.05. Under the null hypothesis, the test statistics T*, T_J, T_S and T_{SY} 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 $\Sigma = (\sigma_{ij})^p_{i,j=1}$. We considered the hypothesis as follows:
  - $H_0^1 : \Sigma_1 = \Sigma_2 = U_0$ against
  - $H_1^1 : \Sigma_1 = U_0$ and $\Sigma_2 = U_1$

  where, $U_0 = (\sigma_{ij})^p_{i,j=1}$ where $\sigma_{ij} = 1$ (if i = j); $\sigma_{ij} = (-1)^{i+j}(0.10i)/j$ (if $i \neq j$) and $U_1 = (\sigma_{ij})^p_{i,j=1}$ where $\sigma_{ij} = 1$ (if i = j); $\sigma_{ij} = (-1)^{i+j}(0.05i)/j$ (if $i \neq j$)
- **Compound Symmetry pattern (CS):** It is defined as $\Sigma = \sigma^2 I_p + k 1_p 1'_p$, where $\sigma^2 > 0$, k is appropriate constant, I_p denotes the p×p identity matrix and 1_p denotes the p×1 vector of ones. The hypothesis was set as:
  - $H_0^2 : \Sigma_1 = \Sigma_2 = C_0 = 0.99 I_p + (0.01) 1_p 1'_p$ against
  - $H_1^2 : \Sigma_1 = C_0$ and $\Sigma_2 = C_1 = 0.95 I_p + (0.05) 1_p 1'_p$
- **Heterogeneous compound symmetry pattern (CSH):** It is defined as $\Sigma = (\sigma_{ij})^p_{i,j=1}$ where $\sigma_{ij} = \sigma^2_i > 0$ (if i = j); $\sigma_{ij} = \sigma_i \sigma_j \rho$ (if $i \neq j$), where ρ is the correlation parameter satisfying $|\rho| < 1$. The hypothesis was set as:
  - $H_0^3 : \Sigma_1 = \Sigma_2 = M_0$ where M_0 is matrix in CSH with $\sigma_{ij} \sim U(5,6)$ (if i = j), $\rho = 0.5$, against
  - $H_1^3 : \Sigma_1 = M_0$ and $\Sigma_2 = M_1$ where M_1 is matrix in CSH with $\sigma_{ij} \sim U(4,5)$ (if i = j), $\rho = 0.4$.
- **Simple pattern (SIM):** It is defined as $\Sigma = \sigma^2 I$ We set the hypothesis testing according to:
  - $H_0^4 : \Sigma_1 = \Sigma_2 = 2I$ against $H_1^4 : \Sigma_1 = 2I$ and $\Sigma_2 = 1.5I$
  - $H_0^5 : \Sigma_1 = \Sigma_2 = \Sigma = I_p$ against $H_1^5 : \Sigma_1 = I_p$ and $\Sigma_2 = \text{Diag}\ (1,1,1,2, \ldots, 1,1,1,2)$

RESULTS AND DISCUSSIONS

Table 1 presents the ASL and empirical powers of T_J, T_S, T_{SY} and T* when all covariance matrices in the hypothesis were under unstructured pattern (UN). The ASL of the tests T_S and T_{SY} were not close to the nominal significance level 0.05 and much lower than it for all cases considered. The test T_J generally yielded the ASL not close to 0.05 for all cases considered. Moreover, the test T_J gave the ASL around 0.060 when the sample sizes were small, $n_1 = n_2 = 20$ here and tended to increase when the sample sizes became larger for any p. For instance, when p = 80, the ASL of T_J was 0.057 (at $n_1 = n_2 = 20$) and increased to 0.061 (at $n_1 = n_2 = 80$). 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_J, T_S and T_{SY} were not reasonable tests whereas the proposed test T* were. Considering the power of the test, since the competitive tests T_J, T_S and T_{SY} 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 $n_1 = n_2 = 20$) and increased to 0.944 (at $n_1 = n_2 = 160$).

Table 1: ASL of T_J, T_S, T_{SY} and T* under $H_0^1 : \Sigma_1 = \Sigma_2 = U_0$ and their empirical powers under $H_1^1 : \Sigma_1 = U_0$ and $\Sigma_2 = U_1$ applied at $\alpha = 0.05$

		ASL				Empirical power			
p	$n_1 = n_2$	T_J	T_S	T_{SY}	T*	T_J	T_S	T_{SY}	T*
20	20	0.061	0.052	0.048	0.062	0.070	0.067	0.065	0.078
40	20	0.055	0.027	0.019	0.058	0.075	0.048	0.045	0.103
	40	0.090	0.027	0.012	0.053	0.107	0.059	0.066	0.128
80	20	0.057	0.014	0.007	0.062	0.099	0.026	0.022	0.167
	40	0.061	0.013	0.005	0.054	0.161	0.079	0.087	0.241
	80	0.061	0.025	0.015	0.053	0.329	0.266	0.431	0.428
160	20	0.064	0.001	0.000	0.071	0.150	0.012	0.006	0.288
	40	0.064	0.008	0.003	0.057	0.268	0.127	0.139	0.446
	80	0.066	0.019	0.013	0.056	0.537	0.485	0.627	0.719
	160	0.066	0.031	0.024	0.053	0.867	0.885	0.976	0.944

ASL: The attained significance level, U_0 and U_1: The matrices defined under unstructured pattern (UN), α: The nominal significance level

Table 2: ASL of T_J, T_S, T_{SY} and T* under $H_0^2 : \sum_1 = \sum_2 = C_0$ and their empirical powers under $H_1^2 : \sum_1 = C_0$ and $\sum_2 = C_1$ applied at $\alpha = 0.05$

		ASL				Empirical power			
p	$n_1 = n_2$	T_J	T_S	T_{SY}	T*	T_J	T_S	T_{SY}	T*
20	20	0.057	0.081	0.084	0.056	0.079	0.081	0.082	0.080
40	20	0.055	0.086	0.081	0.052	0.099	0.076	0.077	0.114
	40	0.052	0.053	0.078	0.050	0.169	0.086	0.131	0.145
80	20	0.047	0.082	0.081	0.056	0.165	0.065	0.068	0.210
	40	0.051	0.052	0.059	0.052	0.324	0.176	0.249	0.333
	80	0.055	0.049	0.045	0.053	0.654	0.473	0.778	0.585
160	20	0.048	0.059	0.057	0.052	0.286	0.047	0.046	0.387
	40	0.053	0.041	0.040	0.052	0.589	0.361	0.454	0.670
	80	0.050	0.040	0.035	0.051	0.918	0.860	0.966	0.932
	160	0.051	0.043	0.031	0.052	0.998	0.997	1.000	0.998

ASL: The attained significance level, C_0 and C_1: The matrices defined under compound symmetry pattern (CS), α: The nominal significance level

Table 2 reports the ASL and empirical powers of T_J, T_S, T_{SY} and T* when all covariance matrices in the hypothesis were set under compound symmetry pattern (CS). Both tests T* and T_J gave the satisfactory ASL which quite controlled 0.05 for all cases considered whereas those of T_S and T_{SY} were not close to 0.05. As seen in the table, the ASL of T_S and T_{SY} decreased as p and the sample sizes increased. For instance, when $p = n_1 = n_2 = 80$, the ASL of T_S and T_{SY} were 0.049 and 0.045, respectively and both decreased to 0.043 and 0.031 when $p = n_1 = n_2 = 160$, respectively. Moreover, when p is fixed, the ASL of T_S and T_{SY} were dropped when the sample sizes increased. For example, when p = 160 ASL of T_S and T_{SY} were 0.059 and 0.057 (at $n_1 = n_2 = 20$), respectively and both values decreased to 0.043 and 0.031, respectively (at $n_1 = n_2 = 160$). This indicates that the tests T_S and T_{SY} were not suitable whereas the proposed test T* and T_J test were appropriate. This table reports that the empirical powers of the proposed test T* and T_J test were quite high and rapidly tended to one. Moreover, the empirical powers of both tests were quite responsive to the increase of p and the sample sizes. Furthermore, the empirical powers of the proposed test T* were slightly higher than those of the test T_J in cases considered.

Table 3 displays the ASL and empirical powers of tests T_J, T_S, T_{SY} and T* when all covariance matrices in the hypothesis were set under heterogeneous compound symmetry pattern (CSH). We observed that the ASL of all tests under this CSH pattern were similar formats to the ASL obtained under UN pattern provided in Table 1. The ASL of the tests T_J, T_S and T_{SY} were not close to 0.05 whereas that of the proposed test T* well approximate 0.05 as p and the sample sizes increased. It can be observed that the ASL of the tests T_S and T_{SY} from this table were lower than those from Table 1 for all cases considered. This indicates that the convergences of the tests T_S and T_{SY} to the standard normal distribution were very slow and not accomplished when the common covariance matrix was 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 3: ASL of T_J, T_S, T_{SY} and T* under $H_0^3 : \sum_1 = \sum_2 = M_0$ and their empirical powers under $H_1^3 : \sum_1 = M_0$ and $\sum_2 = M_1$ applied at $\alpha = 0.05$

		ASL				Empirical power			
p	$n_1 = n_2$	T_J	T_S	T_{SY}	T*	T_J	T_S	T_{SY}	T*
20	20	0.060	0.004	0.000	0.058	0.088	0.070	0.000	0.478
40	20	0.055	0.001	0.000	0.058	0.078	0.054	0.000	0.551
	40	0.054	0.005	0.000	0.053	0.115	0.480	0.000	0.851
80	20	0.052	0.001	0.000	0.060	0.076	0.032	0.000	0.579
	40	0.056	0.002	0.000	0.052	0.121	0.435	0.000	0.862
	80	0.059	0.011	0.001	0.049	0.219	0.943	0.001	0.989
160	20	0.056	0.000	0.000	0.057	0.077	0.009	0.000	0.528
	40	0.058	0.002	0.001	0.052	0.117	0.253	0.001	0.755
	80	0.058	0.010	0.005	0.050	0.217	0.801	0.005	0.946
	160	0.062	0.024	0.015	0.051	0.466	0.992	0.015	0.998

ASL: The attained significance level, M_0 and M_1: The matrices defined under heterogeneous compound symmetry pattern (CSH), α: The nominal significance level

Table 4: ASL of T_J, T_S, T_{SY} and T* under $H_0^4 : \sum_1 = \sum_2 = 2I$ and their empirical powers under $H_1^4 : \sum_1 = 2I$ and $\sum_2 = 1.5I$ applied at $\alpha = 0.05$

		ASL				Empirical power			
p	$n_1 = n_2$	T_J	T_S	T_{SY}	T*	T_J	T_S	T_{SY}	T*
20	20	0.057	0.005	0.000	0.058	0.117	0.268	0.000	0.757
40	20	0.054	0.001	0.000	0.051	0.105	0.309	0.000	0.870
	40	0.054	0.003	0.000	0.052	0.205	0.962	0.000	0.998
80	20	0.048	0.000	0.000	0.056	0.096	0.297	0.000	0.936
	40	0.051	0.001	0.000	0.051	0.198	0.992	0.000	1.000
	80	0.056	0.005	0.000	0.053	0.495	1.000	0.000	1.000
160	20	0.049	0.000	0.000	0.052	0.086	0.211	0.000	0.964
	40	0.053	0.005	0.000	0.050	0.185	0.999	0.000	1.000
	80	0.048	0.002	0.000	0.051	0.931	1.000	0.000	1.000
	160	0.051	0.004	0.000	0.050	1.000	1.000	0.000	1.000

ASL: The attained significance level, α: The nominal significance level

Table 4 reports the ASL of tests T_J, T_S, T_{SY} and T* under the common covariance matrix $\Sigma = 2I$ (simple pattern) and their empirical powers under $\Sigma_1 = 2I$ and $\Sigma_2 = 1.5I$. As expected, the ASL of T* and T_J were quite close to 0.05 for all cases considered while those of T_S and T_{SY} seemed to be zero for all cases considered. The empirical powers of the proposed test T* and T_J 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_J, especially when $n_1 = n_2 \leq 40$ for all p. For example, when $p = n_1 = n_2 = 20$, the empirical powers of T* and T_J were 0.757 and 0.117, respectively. This indicates that, under simple pattern, T* was reasonable test and more powerful than T_J test, particularly in case of small samples.

Table 5 presents the ASL of tests T_J, T_S, T_{SY} and T* under the common covariance matrix $\Sigma = I$ (simple pattern) and empirical powers under $\Sigma_1 = I$ and a certain matrix $\Sigma_2 = \text{Diag}(1,1,1,2, \ldots, 1,1,1,2)$. As displayed in this table,

Table 5: ASL of T_J, T_S, T_{SY} and T* under $H_0^S: \Sigma_1 = \Sigma_2 = \Sigma = I_p$ and their empirical powers under $H_1^S: \Sigma_1 = I_p$ and Σ_2 = Diag(1,1,1,2,...,1,1,1,2) applied at $\alpha = 0.05$

		ASL				Empirical power			
p	$n_1 = n_2$	T_J	T_S	T_{SY}	T*	T_J	T_S	T_{SY}	T*
20	20	0.057	0.082	0.084	0.058	0.243	0.422	0.015	0.230
40	20	0.054	0.090	0.087	0.051	0.241	0.533	0.009	0.502
	40	0.054	0.054	0.084	0.052	0.564	0.976	0.051	0.978
80	20	0.048	0.093	0.090	0.055	0.223	0.581	0.003	0.708
	40	0.051	0.061	0.073	0.051	0.573	0.997	0.026	0.999
	80	0.056	0.054	0.057	0.052	0.969	1.000	0.355	1.000
160	20	0.049	0.082	0.075	0.052	0.213	0.538	0.001	0.821
	40	0.053	0.062	0.071	0.050	0.561	0.999	0.029	1.000
	80	0.048	0.052	0.055	0.051	0.971	1.000	0.469	1.000
	160	0.051	0.049	0.051	0.050	1.000	1.000	0.999	1.000

ASL: The attained significance level, α: The nominal significance level

Table 6: ASL of T_J, T_S, T_{SY} and T* T* under $H_0^S: \Sigma_1 = \Sigma_2 = \Sigma = I_p$ and their empirical powers under $H_1^S: \Sigma_1 = I_p$ and Σ_2 = Diag(1,1,1,2,...,1,1,1,2) when $n_2 = 2n_1$ applied at $\alpha = 0.05$

			ASL				Empirical power			
p	n_1	$n_2 = 2n_1$	T_J	T_S	T_{SY}	T*	T_J	T_S	T_{SY}	T*
40	20	40	0.055	0.064	0.079	0.052	0.265	0.659	0.004	0.743
80	20	40	0.053	0.067	0.090	0.052	0.284	0.784	0.004	0.907
	40	80	0.053	0.051	0.056	0.049	0.751	1.000	0.079	1.000
160	20	40	0.049	0.060	0.067	0.047	0.254	0.846	0.002	0.972
	40	80	0.052	0.052	0.056	0.050	0.747	1.000	0.007	1.000
	80	160	0.055	0.050	0.054	0.051	0.999	1.000	0.768	1.000

ASL: The attained significance level, α: The nominal significance level

the ASL of the proposed test T* and T_J 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 σ^2 defined in simple pattern, from $\sigma^2 = 2$ became $\sigma^2 = 1$, is not effected to the convergence of the asymptotic normality of the proposed test T* and T_J test. But it is greatly effected to the convergence of the asymptotic normality of T_S and T_{SY} 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_S and T_{SY} 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_J 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_J 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_J, T_S, T_{SY} and T* under the null hypothesis $H_0^S: \Sigma_1 = \Sigma_2 = \Sigma = I_p$. Corresponding empirical powers of these tests were also manipulated under the alternative hypothesis $H_1^S: \Sigma_1 = I_p$ and Σ_2 = Diag (1,1,1,2, ..., 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_S and T_{SY} 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_J 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_J test.

APPLICATION

In this section, the dataset from Notterman *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_J = 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 $(p, n_1, n_2) \to \infty$ even if $p/n_i \to c_i \in (0, \infty)$, 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_J, which is a special case of the test for testing the equality of several covariance matrices proposed by Schott (2007), T_S and T_{SY} 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 $nS \sim W_p(\Sigma,n)$ and $a_k = (tr\Sigma^k)/p$, $k = 1,...,4$. Then under the assumptions (A1) and (A3), unbiased and consistent estimators of a_2 as $(p, n)\rightarrow\infty$ is given by $\hat{a}_2$.

Lemma A2: Let $nS \sim W_p(\Sigma,n)$, $\hat{a}_2$ as defined in (2) and $a_k = (tr\Sigma^k)/p$, $k = 1,...,4$. Then under the assumptions (A1) and (A3):

$$\lim_{(p,n)\rightarrow\infty} P\left[(\hat{a}_2 - a_2)/\eta \le x\right] = \Phi(x)$$

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

$$\eta^2 = \frac{4}{np}\left(2a_4 + \frac{pa_2^2}{n}\right)$$

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

Lemma A3: Let $(n_i-1)S_i \sim W_p(\sum_i, n_i-1)$ and $a_{ji} = (tr\sum_i^j)/p$ $i = 1,2$, $j = 1,...,4$. Then under the assumptions (A2) and (A4), unbiased and consistent estimators of a_{2i} as $(p, n_i) \rightarrow \infty$ are given by $\hat{a}_{2i}$, $i = 1,2$.

Lemma A4: Let $(n_i-1)S_i \sim W_p(\sum_i, n_i-1)$ $\hat{a}_{2i}$, $i = 1,2$, as defined in (1) and $a_{ji} = (tr\sum_i^j)/p$, $i = 1,2$, $j = 1,..., 4$. Then under the assumptions (A2) and (A4):

$$\lim_{(p,n_i)\rightarrow\infty} P\left[(\hat{a}_{2i} - a_{2i})/\eta_i \le x\right] = \Phi(x)$$

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

$$\eta_i^2 = \frac{4}{(n_i-1)p}\left(2a_{4i} + \frac{pa_{2i}^2}{n_i-1}\right)$$

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

Lemma A5: Let $nS \sim W_p(\Sigma, n)$ and $a_k = (tr\Sigma^k)/p$, $k = 1,...16$. Then under the assumptions (A1) and (A3), unbiased and consistent estimators of a_4 as $(p,n)\rightarrow\infty$ is given by $\hat{a}_4^*$ defined as:

$$\hat{a}_4^* = \frac{\tau}{p}\left[trS^4 + b^*trS^3trS + c^*(trS^2)^2 + dtrS^2(trS)^2 + e(trS)^4\right]$$

where:

$$\tau = \frac{n^5(n^2+n+2)}{(n+1)(n+2)(n+4)(n+6)(n-1)(n-2)(n-3)}$$

$$b^* = -\frac{4}{n}, \quad c^* = -\frac{2n^2+3n-6}{n(n^2+n+2)}, \quad d = \frac{2(5n+6)}{n(n^2+n+2)}$$

and

$$e = -\frac{5n+6}{n^2(n^2+n+2)}$$

REFERENCES

Bartlett, M.S., 1937. Properties of sufficiency and statistical tests. Proc. R. Soc. Lond. A, 160: 268-282.

Dudoit, S., J. Fridlyand and T.P. Speed, 2002. Comparison of discrimination methods for the classification of tumors using gene expression data. J. Am. Statist. Assoc., 97: 77-87.

Fisher, T., X. Sun and C.M. Gallagher, 2010. A new test for sphericity of the covariance matrix for high dimensional data. J. Multivariate Anal., 101: 2554-2570.

Fujikoshi, Y., V.V. Ulyanov and R. Shimizu, 2010. Multivariate Statistics: High-Dimensional and Large-Sample Approximations. John Wiley and Sons, New Jersey, ISBN-13: 9780470411698.

Gamage, J. and T. Mathew, 2008. Inference on mean sub-vectors of two multivariate normal populations with unequal covariance matrices. Stat. Probabil. Lett., 78: 420-425.

Huang, D., Y. Quan, M. He and B. Zhou, 2009. Comparison of linear discriminant analysis methods for the classification of cancer based on gene expression data. J. Exp. Clin. Cancer Res., 28: 149.

Ibrahim, J.G., M. Chen and R.J. Gray, 2002. Baysian models for gene expression with DNA microarray data. J. Am. Statist. Assoc., 97: 88-99.

Johnson, D.E., 1998. Applied Multivariate Methods for Data Analysis. Duxbury Press, New York, USA.

Krzanowski, W.J., 2000. Principles of Multivariate Analysis: A User's Perspective. 1st Edn., Oxford University Press, USA., ISBN-10: 0198507089, Pages: 608.

Lehmann, E.L. and J.P. Romano, 2005. Testing Statistical Hypotheses. 3rd Edn., Springer, New York, ISBN-13: 9780387988641, Pages: 786.

Notterman, D.A., U. Alon, A.J. Sierk and A.J. Levine, 2001. Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma and normal tissue examined by oligonucleotide arrays. Cancer Res., 61: 3124-3130.

Schott, J.R., 2007. A test for the equality of covariance matrices when the dimension is large relative to the sample size. Comput. Stat. Data Anal., 51: 6535-6542.

Sebastiani, P., H. Xie and M.F. Ramoni, 2006. Baysian analysis of comparative microarray experiments by model averaging. Bayesian Anal., 1: 107-732.

Srivastava, M.S., 2002. Methods of Multivariate Statistics. John Wiley and Sons, New York, ISBN-13: 9780471223818, Pages: 728.

Srivastava, M.S., 2005. Some tests concerning the covariance matrix in high dimensional data. J. Japan Statist. Soc., 35: 251-272.

Srivastava, M.S., 2007. Testing the equality of two covariance matrices and testing the independence of two subvectors with fewer observations than the dimension. Proceedings of the International Conference on Advances in Interdisciplinary Statistics and Combinatorics, October 12-14, 2007, Greensboro, North Carolina, USA.

Srivastava, M.S. and H. Yanagihara, 2010. Testing the equality of several covariance matrices with fewer observations than the dimension. J. Multivariate Anal., 101: 1319-1329.