Science and Technology
RMUTT Journal

Vol.3 No.2 (2013) : 45-50
www.sci.rmutt.ac.th/stj
Online ISSN 2229-1547

การแจกแจงของตัวประมาณคงเส้นคงวาของอัตราส่วนของเทรซของเมตริกซ์ของความแปรปรวนสองประชากร

The Distribution of a Consistent Estimator of the Traces Ratio of Two Population Covariance Matrices

Saowapha Chaipitak[1*] and Boonyarit Choopradit[2]

[1] Faculty of Science and Technology, Rajamangala University of Technology Thanyaburi,
Pathum Thani, Thailand
[2]Faculty of Sciences and Industrial Technology, Prince of Songkla University, Songkla, Thailand
*E-mail: s.chaipitak@yahoo.com

Abstract

In this paper, testing the hypothesis of the equality of two covariance matrices from independent multivariate normal populations is of interested. To test the hypothesis, a test statistic is proposed based on an unbiased and consistent estimator of the ratio between two traces of two population covariance matrices. The asymptotic distribution of the consistent estimator is investigated using the delta method. Finally, it converges in distribution to normal as the number of variables and sample sizes go forward to infinity.

Keyword: Multivariate Normal Distribution, Covariance Matrices, Sample Sizes, Converge in Distribution.

Received: Nov 18, 2013
Accepted: Dec 02, 2013

1. Introduction

Let $\mathbf{X}_{jk}, j=1,...,n_k;\ k=1,2,$ be random samples drawn from independent multivariate normal populations $N_p(\boldsymbol{\mu}_k, \Sigma_k),$ where mean vector $\boldsymbol{\mu}_k$ and covariance matrix Σ_k are unknown. It is one of the most important requirements in many statistical techniques to know whether covariance matrices of the two populations are equal or not, such as in discriminant analysis, testing the equality of two mean vectors, testing the equality of two mean sub-vectors, ([4], [5], [7], and [10]). Before applying any further analysis, this equality must be tested. The 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_k, larger than the number of variables, $p,$ is the modified likelihood ratio test. This test is not valid when $p \geq n_k$. In the present, many applications, such as economics or modern sciences, the data have very small sample sizes while the large number of the variables taken from each observation, i.e. $p > n_k$. For instance, DNA microarrays typically measure thousands to millions of gene expressions on the small sample sizes [6]. When the data fails into a $p \geq n_k$ situation, the sample covariance matrices $\mathbf{S}_k$ are singular making the modified likelihood ratio test is not available. The recent tests under this problem were worked by [2], [9], [11], and [12].

In order to test the hypothesis, a new parameter function is suggested and its consistent estimator is investigated. The distribution of the consistent estimator is asymptotically distributed as the standard normal distribution for large p, n_k.

The organization of this article is as follows. Section 2 gives some preliminary. A new parameter function and a consistent estimator are proposed in Section 3. The asymptotic distribution of the estimator is investigated in Section 4. The useful lemma is given in the Appendix.

2. Preliminaries

Let $\mathbf{X}_{jk}, j=1,...,n_k;\ k=1,2,$ be random samples drawn independently normally distributed populations $N_p(\boldsymbol{\mu}_k, \Sigma_k)$ where $\boldsymbol{\mu}_k$ denotes an unknown mean vector of the k^{th} population and Σ_k denotes an unknown positive definite covariance matrix of the k^{th} population.

Let

$$\overline{\mathbf{X}}_k = \frac{1}{n_k}\sum_{j=1}^{n_k}\mathbf{X}_{jk}, \quad k=1,2,$$

$$\mathbf{A}_k = \sum_{j=1}^{n_k}\left(\mathbf{X}_{jk} - \overline{\mathbf{X}}_k\right)\left(\mathbf{X}_{jk} - \overline{\mathbf{X}}_k\right)', \quad k=1,2,$$

$$\mathbf{S}_k = \frac{1}{n_k - 1}\mathbf{A}_k, \quad k = 1,2,$$

$$\hat{a}_{1k} = \frac{1}{p} tr\mathbf{S}_k, \quad k = 1,2, \qquad (2.1)$$

Suppose there are independent estimates $\mathbf{S}_1$, $\mathbf{S}_2$, the sample covariance matrices of the covariance matrices Σ_1 and Σ_2, respectively, with $(n_k - 1)\mathbf{S}_k \sim W_p(\Sigma_k, n_k - 1),\ k = 1,2,$ i.e. $(n_k - 1)\mathbf{S}_k$ has a Wishart distribution with n_k degrees of freedom and covariance matrix Σ_k, see [1]. The common covariance matrix Σ is estimated by the pooled sample covariance matrix

$$\hat{\Sigma} = \frac{1}{n-1}\mathbf{A} = \frac{1}{n-1}(\mathbf{A}_1 + \mathbf{A}_2) \equiv \mathbf{S},$$

where $n = n_1 + n_2 - 1.$ Note that $(n-1)\mathbf{S} \sim W_p(\Sigma, n-1),$ i.e. $(n-1)\mathbf{S}$ has a Wishart distribution with $n-1$ degrees of freedom and covariance matrix $\Sigma.$ Therefore, we let

$$\hat{a}_1 = \frac{1}{p} tr\mathbf{S}, \qquad \hat{a}_2 = \frac{(n-1)^2}{p(n-2)(n+1)}\left\{ tr\mathbf{S}^2 - \frac{1}{n-1}(tr\mathbf{S})^2 \right\}.$$

3. The Consistent Estimator

In order to test the hyphothesis $H_0 : \Sigma_1 = \Sigma_2 = \Sigma$, a test statistic is developed based on the fact that if the null hypothesis H_0 holds, i.e. $\Sigma_1 = \Sigma_2$, then $tr\Sigma_1 = tr\Sigma_2$. Thus, under the null hypothesis, we consider the parameter function as

$$b = \frac{tr\Sigma_1}{tr\Sigma_2} = \frac{a_{11}}{a_{12}} = 1,$$

where $a_{1k} = \frac{1}{p} tr\Sigma_k,\ k = 1,2.$

Therefore, the hypothesis $H_0 : b = 1$ can be tested against $H_1 : b \neq 1.$

The following assumptions are made:

(A1) As $(p,n) \to \infty, \frac{p}{n} \to c,\ c \in (0,\infty)$

(A2) As $(p,n_k) \to \infty, \frac{p}{n_k} \to c_k,\ c_k \in (1,\infty), \qquad k = 1,2$

(A3) As $p \to \infty,\ a_m \to \alpha_m,\ \alpha_m \in (0,\infty), \qquad m = 1,\ldots,16$

(A4) As $p \to \infty,\ a_{lk} \to \alpha_{lk},\ \alpha_{lk} \in (0,\infty), \qquad k = 1,2;\ l = 1,\ldots,8,$

where $a_m = \frac{1}{p} tr\Sigma^m$, and $a_{lk} = \frac{1}{p} tr\Sigma_k^l$

By applying the results provided by [3], from one population to the case of two populations (Thus it is presented without proof here), unbiased and consistent estimator of $a_{1k} = \frac{1}{p} tr(\Sigma_k)$ is given by $\hat{a}_{1k} = \frac{1}{p} tr(\mathbf{S}),\ k = 1,2.$ Thus the quantity b can be estimated by

$$\hat{b} = \frac{\hat{a}_{11}}{\hat{a}_{12}} = \frac{tr\mathbf{S}_1}{tr\mathbf{S}_2}.$$

4. The Distribution of the Consistent Estimator

It is provided in [3] that $\hat{a}_{1k},\ k = 1,2,$ is asymptotically normally distributed with mean a_{1k} and variance $\delta_k^2 = \frac{2a_{2k}}{n_k p}$.

The following lemma gives the asymptotic distribution of the consistent estimators of the parameters appeared in b.

Lemma 4.1 *Let* $(n_k - 1)\mathbf{S}_k \sim W_p(\Sigma_k, n_k - 1),\ \hat{a}_{1k},\ k = 1,2,$ *as defined in (2.1), and* $a_{lk} = (tr\Sigma_k^l)/p, k = 1,2,\ l = 1,\ldots,4,$ *then under the assumptions (A2) and (A4)*

$$\begin{pmatrix} \hat{a}_{11} \\ \hat{a}_{12} \end{pmatrix} \xrightarrow{D} N_2\left[\begin{pmatrix} a_{11} \\ a_{12} \end{pmatrix}, \begin{pmatrix} 2a_{21}/n_1 p & 0 \\ 0 & 2a_{22}/n_2 p \end{pmatrix}\right]$$

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

Proof Since random samples $\mathbf{X}_{jk},\ j = 1,\ldots,n_k;\ k = 1,2,$ are drawn from two independent populations and sample covariance matrices $\mathbf{S}_k$ are calculated from corresponding independent random samples $\mathbf{X}_{j1}$, and $\mathbf{X}_{j2}$, thus, $\mathbf{S}_1$ and $\mathbf{S}_2$ must be independent each other. In fact, the statistic $\hat{a}_{11}$ is a function of $\mathbf{S}_1$ alone while the statistic $\hat{a}_{12}$ is also a function of $\mathbf{S}_2$ alone. Thus $\hat{a}_{11}$ and $\hat{a}_{12}$ are also independent and then it makes $COV(\hat{a}_{11}, \hat{a}_{12}) = 0.$ Since $\hat{a}_{1k},\ k = 1,2$ are asymptotically normally distributed with mean a_{1k} and variance δ_k^2, and the fact that the covariance between $\hat{a}_{11}$ and $\hat{a}_{12}$ is zero, then the jointly asymptotic distribution of statistics $\hat{a}_{11}$ and $\hat{a}_{12}$ are the bivariate normal distribution with mean vector and covariance matrix as given above. The proof is completed.

By applying the delta method given in the Appendix to a function of two random variables, the following theorem provide an asymptotic distribution of the consistent estimator $\hat{b}$.

Theorem 4.1 *Let* b, *and* $\hat{b}$ *be as defined above. Then, under the assumptions (A1)-(A4),*

$$\hat{b} \xrightarrow{D} N(b, \delta^2),$$

where

$$\delta^2 = \frac{2}{p}\left(\frac{1}{n_1 a_{12}} + \frac{a_{22}a_{11}^2}{n_2 a_{12}^4}\right).$$

Proof We note that $\hat{b} = \hat{a}_{11} / \hat{a}_{12}$. Hence the partial derivative of $\hat{b}(\hat{a}_{11}, \hat{a}_{12})$ with respect to $\hat{a}_{11}$ is $\left(\frac{\partial \hat{b}}{\partial \hat{a}_{11}}\right) = \frac{1}{\hat{a}_{12}}$. Similarly, the partial derivative of $\hat{b}(\hat{a}_{11}, \hat{a}_{12})$ with respect to $\hat{a}_{12}$ is $\left(\frac{\partial \hat{b}}{\partial \hat{a}_{12}}\right) = -\frac{\hat{a}_{11}}{\hat{a}_{12}^2}$. Thus, by applying the delta method, $\hat{b} \sim N(b, \delta^2)$ asymptotically with

$$\delta^2 = \begin{pmatrix} \frac{1}{a_{12}} & -\frac{a_{11}}{a_{12}^2} \end{pmatrix} \begin{pmatrix} \frac{2a_{21}}{n_1 p} & 0 \\ 0 & \frac{2a_{22}}{n_2 p} \end{pmatrix} \begin{pmatrix} \frac{1}{a_{12}} \\ -\frac{a_{11}}{a_{12}^2} \end{pmatrix}$$

$$= \frac{2}{p}\left(\frac{1}{n_1 a_{12}} + \frac{a_{22}a_{11}^2}{n_2 a_{12}^4}\right).$$

The proof is completed.

Corollary 4.1 *Let $\hat{b}$ be as defined above. Under $H_0 : \Sigma_1 = \Sigma_2 = \Sigma$ and the assumptions (A1)-(A4), then*

$$T = \frac{\hat{b} - 1}{\left\{\frac{2}{p}\left(\frac{1}{n_1 a_{12}} + \frac{a_{22}a_{11}^2}{n_2 a_{12}^4}\right)\right\}^{\frac{1}{2}}} \xrightarrow{D} N(0,1).$$

Proof Under H_0, then $a_{12} = a_{12} = a_1$, and $a_{21} = a_{22} = a_2$.

Thus $\delta^2 \overset{H_0}{=} \frac{2}{p}\left(\frac{1}{n_1 a_1} + \frac{a_2 a_1^2}{n_2 a_1^4}\right) = \frac{2}{pa_1}\left(\frac{1}{n_1} + \frac{a_2}{n_2 a_1}\right)$. It follows Theorem 4.1, then the proof is completed.

Appendix

Lemma A.1 (The delta method)

Suppose $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ are random vectors in the $\Re^k$ Euclidean space and assume that $\tau_n(\mathbf{X}_n - \boldsymbol{\mu}) \xrightarrow{D} N_k(\mathbf{0}, \Sigma)$, where $\boldsymbol{\mu}$ is a constant vector and $\{\tau_n\}$ is a sequence of constants $\tau_n \to \infty$. In addition, presume that $g(.)$ is a function from $\Re^k$ to $\Re$ which is differentiable at $\boldsymbol{\mu}$ with a gradient (vector of first partial derivatives) of dimension $1 \times k$ at $\boldsymbol{\mu}$ equal to $g'(\boldsymbol{\mu})$, then

$$\tau_n[g(\mathbf{X}_n) - g(\boldsymbol{\mu})] \xrightarrow{D} N\left(0, g'(\boldsymbol{\mu})\Sigma g'(\boldsymbol{\mu})^T\right).$$

Proof see [8]

5. References

[1] T.W. Anderson, *An Introduction to Multivariate Statistical Analysis*, 2nd ed., New York, John Wiley & Sons, 1984.

[2] S. Chaipitak and S. Chongcharoen, A test for testing the equality of two covariance matrices for high-dimensional data, *J. Appl. Sci*. 13 (2013), 270-277.

[3] S. Chaipitak, *Tests for covariance matrices with high-dimensional data*, Ph.D. dissertation, National Institute of Development Administration, 2012.

[4] Y. Fujikoshi, V. Ulyanov and R. Shimizu, *Multivariate Statistics : High-Dimensional and Large-Sample Approximations*, New Jersey, John Wiley & Sons, 2010.

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

[6] J.G. Ibrahim, M. Chen and R.J. Gray, Baysian models for gene expression with DNA microarray data. *J. Am. Stat. Assoc*. 97 (2002): 88-99.

[7] D.E. Johnson, *Applied Multivariate Methods for Data Analysts***,** California, Duxbury, 1998.

[8] E.L. Lehmann and J.P. Romano, *Testing Statistical Hypotheses,* 3rd ed., New York, Springer, 2005.

[9] J.R. Schott, A test for the equality of covariance matrices when the dimension is large relative to the sample sizes, *Comput. Stat. Data. An.*, 51 (2007), 6535-6542.

[10] M.S. Srivastava, *Methods of multivariate statistics*, New York, John Wiley & Sons, 2002.

[11] M.S. Srivastava, Multivatiate theory for analyzing high-dimensional data. *J. Japan. Statist. Soc***.**, 37 (2007), 53-86.

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