Jurnal Teknologi

Full paper

A Statistical Test for the Stability of Covariance Structure

Nur Syahidah Yusoff[a*], Maman Abdurachman Djauhari[b]

[a]*Department of Sciences, Faculty of Industrial Sciences & Technology, Universiti Malaysia Pahang, 26300, Gambang, Pahang, Malaysia*
[b]*Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia*

*Corresponding author: wnsyahidah@ump.edu.my

Article history

Received :21 January 2013
Received in revised form :
7 May 2013
Accepted :25 June 2013

Graphical abstract

$$J = \sum_{i=1}^{m} \left(\frac{1}{2} tr(Z_i^2) \right)$$

Abstract

The stability of covariance matrix is a major issue in multivariate analysis. As can be seen in the literature, the most popular and widely used tests are Box M-test and Jennrich J-test introduced by Box in 1949 and Jennrich in 1970, respectively. These tests involve determinant of sample covariance matrix as multivariate dispersion measure. Since it is only a scalar representation of a complex structure, it cannot represent the whole structure. On the other hand, they are quite cumbersome to compute when the data sets are of high dimension since they do not only involve the computation of determinant of covariance matrix but also the inversion of a matrix. This motivates us to propose a new statistical test which is computationally more efficient and, if it is used simultaneously with M-test or J-test, we will have a better understanding about the stability of covariance structure. An example will be presented to illustrate its advantage

Keywords: Covariance determinant; covariance matrix; vector variance

Abstrak

Kestabilan matrik kovarians adalah isu utama dalam analisis multivariat. Di dalam kajian literatur, terdapat ujian statistik yang paling popular dan digunakan secara meluas ialah ujian Box-M yang diperkenalkan oleh Box pada tahun 1949 dan ujian Jennrich-J oleh Jennrich pada tahun 1970. Ujian-ujian tersebut menggunakan penentu sampel matrik kovarians sebagai ukuran serakan multivariat. Oleh kerana ia hanya mewakili skalar bagi struktur kompleks, jadi ia tidak boleh mewakili keseluruhan struktur. Selain itu, pengiraan ujian-ujian tersebut agak rumit dan sukar apabila berhadapan dengan data yang mempunyai dimensi yang tiggi kerana mereka melibatkan pengiraan penentu sampel matrik kovarians dan juga *inversion* matrik. Ini mendorong kami untuk membina satu ujian statistik yang baru di mana pengiraannya lebih efisien dan jika is digunakan serentak dengan ujian M dan ujian J,pemahaman kita terhadap kestabilan struktur kovarians menjadi lebih baik. Satu contoh akan dibentangkan untuk menggambarkan kelebihan ujian tersebut.

Kata kunci: Matrik kovarians; penentu kovarians; *vector variance*

© 2013 Penerbit UTM Press. All rights reserved.

1.0 INTRODUCTION

Nowadays, testing the stability of process variability in multivariate setting is a serious problem. For that purpose, there are many different methods available in the literature. All those methods are constructed based on the notion of multivariate variability measure. According to Djauhari *et al*. [1], a multivariate variability measure is defined as a non-negative, real valued function of a covariance matrix such that the more scattered the population, the larger the value of that function and, conversely, the less scattered the population, the smaller the value of that function. In multivariate setting, variability is numerically represented by covariance matrix. The importance of covariance structure stability has been shown in many areas. For example, in medical research [2], in genetic research [3], [4], [5] and [6], in personality research [7], in financial industry [8], [9] and [10], in real estate industry [11] and service industry [12] and [13].

In medical research, Emil [2] mentioned that covariance structure stability has been used to model the error structure of both observed and latent variables. While in genetic research, covariance structure stability is used to reconstruct historical patterns of selection and to test genetic drift as a null model for differentiation [3]. In financial [9] and real estate [11] industries, the covariance structure stability is needed in portfolio optimization and to determine the allocation of international real estate securities investments, respectively. Besides that, covariance structure stability is also used to improve the quality and performance throughout the entire chain of marketing, development, production and sales processes which are aimed at the delivery a very high quality of product to the customers [12].

Last but not least in personality research [7], it has been used to avoid inappropriate restandardization of the variables which can easily produce seriously misleading results.

Tang [9] mentioned that the stability of covariance matrices can only be fully examined by testing directly the equality of covariance matrices across time periods. For that purpose, the most popular and widely used tests are Box-M test [10], [14], [15] and [16] and Jennrich-J test [8], [11] and [17]. These tests involve determinant of sample covariance matrix, i.e., generalized variance (GV) as multivariate dispersion measurement. Due to that, these tests are quite cumbersome to compute when the data sets are of high dimension since they do not only involve the computation of determinant of covariance matrix but also the inversion of a matrix. This motivates us to propose a new statistical test which is computationally more efficient. Furthermore, instead of GV, we use vector variance (VV) as the measure of multivariate dispersion. An example will be presented to illustrate its advantage.

The rest of the paper is organized as follows. In the next section, we will discuss the limitation of the existing tests, i.e., Box-M test and Jennrich-J test and then, a new statistical test will be proposed. In Section 3, an example will be presented and discussed. This paper will be closed with the conclusion in the fourth section.

2.0 METHODOLOGY

Let m independent samples from p-variate normal distribution $N_p(\mu_1,\Sigma_1),...,N_p(\mu_m,\Sigma_m)$. Therefore, the hypothesis for testing the stability of covariance matrices is

$$H_0:\Sigma_1=\Sigma_2=...=\Sigma_m \text{ versus } H_1:\Sigma_i\neq\Sigma_j$$

for a pair (i,j); $i,j = 1, 2, \ldots, m$. In the next two sub sections, we present the statistic of M-test and J-test.

2.1 Box M-test

The statistic for M-test [14] is

$$M = N\ln\left|\bar{S}\right| - \sum_{i=1}^{m} n_i \ln\left|S_i\right| \qquad (1)$$

where

(i) $\bar{S}=\frac{1}{N}\sum_{i=1}^{m} n_i S_i$ is the pooled of sample covariance matrix,

(ii) $N=n_1+n_2+...+n_m$; $n_i=$ sample size of covariance matrix.

Box [14] shows that the statistical test (1) can be approximated by χ^2 distribution if p and m are not greater than five and $n_i>20$ and F distribution for p and m are greater than five.

$$\frac{M}{c}\xrightarrow{d}\chi_a^2 \text{ and } \frac{M}{b}\xrightarrow{d}F_{f_1,f_2}$$

where

(i) $a=f_1=\frac{1}{2}(m-1)p(p+1)$; p is dimension of the covariance matrix.

(ii) $b=\dfrac{f_1}{1-A_1-\dfrac{f_1}{f_2}}$

(iii) $c=\dfrac{1}{(1-A_1)}$; $A_1=\dfrac{(2p^2+3p-1)}{6(p+1)(m-1)}\left(\sum_{i=1}^{m}\dfrac{1}{n_i}-\dfrac{1}{n}\right)$

(iv) $f_2=\dfrac{f_1+2}{A_2-A_1^2}$; $A_2=\dfrac{(p-1)(p+2)}{6(m-1)}\left(\sum_{i=1}^{m}\dfrac{1}{n_i^2}-\dfrac{1}{n^2}\right)$

Pearson [18] compares the accuracy of these two approximations and concluded that χ^2 approximation will be sufficient to use in any practical purposes. However, Pearson [18] also shows that F approximation, clearly, more accurate than χ^2 approximation. Hence, in this study, we employ the approximation of F distribution. Therefore, in this case, H_0 is rejected at level of significance α if $\frac{M}{b}$ exceeds F_{f_1,f_2}, the $(1-\alpha)-$th quantile of F distribution.

2.2 Jennrich J-test

According to Jennrich [17], the statistic for J-test is

$$J=\sum_{i=1}^{m}\left(\frac{1}{2}tr(Z_i^2)\right) \qquad (2)$$

where

(i) $Z_i=\sqrt{n_i}\bar{S}^{-1}(S_i-\bar{S})$; $S=$ sample covariance matrix,

(ii) $\bar{S}=(\bar{s}_{ij})=\dfrac{n_1S_1+n_2S_2+...+n_mS_m}{N}$ is the pooled of sample covariance matrix,

(iii) $N=n_1+n_2+...+n_m$; $n_i=$ sample size of covariance matrix.

Jennrich [17] shows that the statistical test (2) is asymptotically χ^2 distributed with degree of freedom k = $(m-1)p(p-1)/2$ where p is dimension of the covariance matrix. Therefore, H_0 is rejected at level of significance α if J exceeds $\chi^2_{\alpha;k}$, the $(1-\alpha)-$th quantile of χ^2 distribution.

2.3 Proposed Statistical Test

It is important to note that M-test and J-test involve the determinant of sample covariance matrix, i.e., GV as a measure multivariate dispersion. Due to that application of GV, these tests are quite cumbersome to compute when the data sets are of high dimension since they do not only involve the computation of determinant of covariance matrix but also the inversion of a matrix. This limitation motivates us to propose a new statistical test which is computationally more efficient. It is because the proposed statistical test used VV as a measure multivariate dispersion. VV is a sum squared of the elements of sample covariance matrix, i.e., $Tr(S^2)$. Djauhari et.al [1] mentioned that by using Cholesky decomposition, the computational complexity of VV is far less than GV, i.e., VV is of order $O(p^2)$ and GV is of order $O(p^3)$. Consequently, the computation of the proposed statistical test is more efficient than the existing tests, i.e., M-test and J-test.

According to Montgomery [19], by using Multivariate Statistical Process Control (MSPC), the hypothesis testing the stability of covariance structure, i.e., $H_0 : \Sigma_1 = \Sigma_2 = ... = \Sigma_m$ versus $H_1 : \Sigma_i \neq \Sigma_j$ for at least one pair (i,j) is equivalently to testing repeatedly $H_0 : \Sigma_i = \Sigma_0$ versus $H_1 : \Sigma_i \neq \Sigma_0$ where $i = 1, 2, ..., m$. Σ_0 is a reference sample, i.e., covariance matrix of the whole studied data. Therefore, the statistic of the proposed statistical test is

$$Z_i = \frac{\sqrt{n_i - 1}\left(Tr\left(S_i^2\right) - Tr\left(\Sigma_0^2\right)\right)}{\sqrt{8Tr\left(\Sigma_0^4\right)}} \tag{3}$$

distributed to $N(0,1)$ where

(i) $Tr\left(\Sigma_0^4\right) = \sum_{i=1}^{p}\sum_{j=1}^{p}\left(\sigma_{ij}^*\right)^2$; $\sigma_{ij}^* =$ elements (i,j) of Σ_0^2

(ii) $Tr\left(\Sigma_0^2\right) = \sum_{i=1}^{p}\sum_{j=1}^{p}\sigma_{ij}^2$; $\sigma_{ij} =$ elements (i,j) of Σ_0

(iii) $Tr\left(S_i^2\right) = \sum_{i=1}^{p}\sum_{j=1}^{p}s_{ij}^2$; $s_{ij} =$ elements (i,j) of S

(iv) $n_i =$ sample size of covariance matrix.

In this case, H_0 is rejected if $|Z| > z_{\frac{\alpha}{2}}$.

3.0 EXAMPLE

Covariance matrices of foreign exchange rate time series are analyzed for 78 world currencies, retrieved from Pacific Exchange Rate Service (http://fx.sauder.ubc.ca/EUR/analysis.html). Let S_1 and S_2 are the covariance matrices of first quarter in year 2000 and third quarter in 2000, respectively. The first quarter consists of the data from January 2000 until April 2000 and third quarter consists of the data from August 2000 until October 2000. The sample size for both covariance matrices, n_1 and n_2 is equal to 64. In this case, (1) and (2) cannot be calculated since the determinant of S_1 and S_2 are equal 0.

To test the stability of covariance matrices, i.e., equality of covariance matrices by using proposed statistical test (3), the hypotheses involve are $H_0 : \Sigma_1 = \Sigma_0$ versus $H_1 : \Sigma_1 \neq \Sigma_0$ and $H_0 : \Sigma_2 = \Sigma_0$ versus $H_1 : \Sigma_2 \neq \Sigma_0$. For $\alpha = 0.05$, we obtain Z_1 and Z_2 are as follows.

(i) $H_0 : \Sigma_1 = \Sigma_0$ versus $H_1 : \Sigma_1 \neq \Sigma_0$

$|Z_1| = 5.4574 > z_{0.025} = 1.96$; reject H_0

(ii) $H_0 : \Sigma_2 = \Sigma_0$ versus $H_1 : \Sigma_2 \neq \Sigma_0$

$|Z_2| = 2.5732 > z_{0.025} = 1.96$; reject H_0

Therefore, by using proposed statistical test, we conclude that S_1 and S_2 are not stable.

4.0 CONCLUSION

According to that example, it shows that M-test and J-test cannot be calculated because of the determinant of S_1 and S_2 are equal 0. To avoid this singularity problem, the sample size of the data, n must be greater than number of dimension, p. This signifies that the both tests are not apt to use when data sets are of high dimension.

That finding illustrates us the advantages of the proposed statistical test. Those advantages are as follows.

(i) Do not necessities the condition $n > p$
(ii) Can be used efficiently for the high dimension data set since proposed statistical test is a quadratic form.

Acknowledgement

We acknowledge financial support from the Ministry of Higher Education of Malaysia, via Research University Grant (RUG) vote number Q.J13000.7126.02H18. The authors would like to thank the Ministry of Higher Education and Universiti Teknologi Malaysia for the sponsorship and the opportunity to do this research. The first author is in debt to Universiti Malaysia Pahang for this opportunity. Special thanks go to the anonymous referees for their constructive comments and suggestions.

References

[1] M., A. Djauhari, M. Mashuri and D., E. Herwindiati. 2008. Multivariate Process Variability Monitoring. *Communications in Statistics-Theory and Method*. 37: 1742–1754.
[2] K. Emil. 2002. Bias and Heteroscedastic Memory Error in Self-reported Health Behavior: An Investigation Using Covariance Structure Analysis. *BMC Medical Research Methodology*. 2(14): 1–12.
[3] S., J. Arnold, R. Burger, P., A. Holenhole, C., A. Beverley, and A., G. Jones. 2008. Understanding the Evolution and Stability of the G-matrix. *Evolution*. 62: 2451–2461.
[4] F. Eroukhmanoff. 2009. Just How Much is the G-matrix Actually Constraining Adaptation? *Evol Biol*. 36: 323–326.
[5] J. Schafer and K. Strimmer. 2005. A Shrinkage Approach to Large Scale Covariance Matrix Estimation and Implications for Functional Genomics. *Statistical Applications in Genetics and Molecular Biology*. 4(1): 1–29.
[6] S., J. Steppan, P., C. Phillips and D. Houle. 2002. Comparative Quantitative Genetics: Evolution of the G-matrix. *Trends in Ecology and Evolution*. 17: 320–327.
[7] F., F. John and G., W. Stephen. 1997. The Investigation of Personality Structure: Statistical Models. *Journal of Research in Personality*. 31: 439–485.
[8] F. Schindler. 1998. Correlation Structure of Real Estate Markets Over Time. *Journal of Property Investment and Finance*. 27(6): 579–592.
[9] G., Y., N. Tang. 1998. The Intertemporal Stability of the Covariance and Correlation Matrices of Hong Kong Stock Returns. *Applied Financial Economics*. 8: 359–365.
[10] L. Stephen. 2006. The Stability of the co-movements between Real Estate Returns in the UK. *Journal of Property Investment & Finance*. 24(5): 434–442.
[11] P., M., A. Eichholtz. 1996. The stability of the covariances of international property share returns. *Journal of Real Estate Research*. 11(2): 149-158.
[12] K., C., B. Roes and D. Dorr. 1997. Implementing Statistical Process Control in Service Processes. *International Journal of Quality Science*. 2(3): 149–166.
[13] M. Wood. 1994. Statistical Methods for Monitoring Service Processes. *International Journal of Service Industry Management*. 5(4): 53–68.
[14] G., E., P. Box. 1949. A General Distribution Theory for A Class of Likelihood Criteria. *Biometrika*. 36 (3/4): 317–346.
[15] K., V. Mardia, J., T. Kent and J., M. Bibby. 1979. *Multivariate Analysis*. London: Academic Press.
[16] M., R. David. 2000. Time-varying Stock Market Correlations and Correlation Breakdown. Risk Management Team.
[17] R., I. Jennrich. 1970. Test for the Equality of Two Correlation Matrices. *Journal of the American Statistical Association*. 65(330): 904–912.
[18] E., S. Pearson. 1969. Some Comments on the Accuracy of Box's Approximations to the Distribution of M. *Biometrika*. 56(1): 219–220.
[19] D., C. Montgomery. 2005. *Introduction to Statistical Quality Control*. Fifth Edition. New York: John Wiley and Son, Inc.