Computational Statistics and Data Analysis 78 (2014) 43–55

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Testing proportionality of two large-dimensional covariance matrices

Lin Xu [a], Baisen Liu [b], Shurong Zheng [a,*], Shaokun Bao [c]

[a] KLAS and School of Mathematics and Statistics, Northeast Normal University, Changchun 130024, China
[b] School of Statistics, Dongbei University of Finance and Economics, Dalian 116025, China
[c] Technology Center, Qinhuangdao Port Co.,LTD, Qinhuangdao 066002, China

ARTICLE INFO

Article history:
Received 21 November 2012
Received in revised form 9 September 2013
Accepted 26 March 2014
Available online 13 April 2014

Keywords:
Proportionality
Large-dimensional data
Likelihood ratio test
Limiting spectral distribution
Random F-matrices

ABSTRACT

Testing the proportionality of two large-dimensional covariance matrices is studied. Based on modern random matrix theory, a pseudo-likelihood ratio statistic is proposed and its asymptotic normality is proved as the dimension and sample sizes tend to infinity proportionally.

1. Introduction

The structure of covariance matrices is an important issue in statistical analysis. In multi-group setting, it is important to know what relations there are between covariance matrices. The simplest situation is that these covariance matrices are identical. However, in many cases, these covariance matrices are not identical but share some common features, such as proportionality to a common matrix. The equality of covariance matrices describes the homoscedasticity of the population covariance matrices, while the proportionality of covariance matrices implies the simplest heteroscedasticity of the population covariance matrices. Thus it is important to check whether the population covariance matrices from different groups are equal or proportional. In fact, testing equality or proportionality of covariance matrices is often used as an initial test in analyzing experimental data.

In this article, we consider the proportionality test of two population covariance matrices

$$H_0 : \mathbf{\Sigma}_1 = c\mathbf{\Sigma}_2, \tag{1}$$

where $c > 0$ is an unknown scalar and $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are two population covariance matrices. This hypothesis has extensive applications in many fields, such as discriminant analysis (Dargahi-Noubary, 1981; Hawkins and Raath, 1982), genetics (Jensen and Madsen, 2004), and principal components analysis (Flury and Riedwyl, 1988; Krzanowski, 1979; Schott, 1991). For example, it is related to a quantitative genetic experiment, called *paternal half-sib design*. The aim of the experiment is to model measurements of some quantitative traits in two independent populations of animal offsprings. This experiment is conducted under the hypothesis of equal heritabilities in the two populations, and it corresponds to the hypothesis of

* Corresponding author.
E-mail address: zhengsr@nenu.edu.cn (S. Zheng).

http://dx.doi.org/10.1016/j.csda.2014.03.014
0167-9473/

proportionality between population covariance matrices which we discuss here. A detailed description of this design refers to Jensen and Madsen (2004).

There has been a long history in studying proportionality of covariance matrices. Some researches have been done on the proportionality test of covariance matrices, and most of them are based on the classical limit theorems where the dimension p is fixed. To our knowledge, Federer (1951) was the first to research on the proportionality of covariance matrices and developed a maximum likelihood method for two groups of normal populations with dimension $p \le 3$. Khatri (1967) and Pillai et al. (1969) studied the distribution of ratios of the characteristic roots of the product of one sample covariance matrix and the inverse matrix of another sample covariance matrix. However, they neither explicitly constructed tests for proportionality nor indicated how to estimate proportional covariance matrices.

Kim (1971) extensively studied the problem of proportionality between covariance matrices. Kim (1971) showed that the solution of the likelihood equations was unique, which was later published by Guttman et al. (1983). Independently, Rao (1983) also considered the likelihood ratio test for proportionality of covariance matrices from two normal populations. However, they did not obtain satisfactory solutions for $k > 2$ populations. As an extension, Flury (1986) constructed a likelihood ratio statistic for $k \ge 2$ populations, and proved that the test statistic had an asymptotic Chi-square distribution with the fixed dimension p and sample sizes n_1, n_2 both go to infinity. For $k = 2$ and under H_0, the statement is

$$(n_1 + n_2) \sum_{j=1}^{p} \log(\hat{\lambda}_j) - n_1 \log(|\hat{\boldsymbol{\Sigma}}_1|) + n_2(p \log(\hat{c}) - \log(|\hat{\boldsymbol{\Sigma}}_2|)) \Rightarrow \chi^2_{(p^2+p-2)/2}, \tag{2}$$

where $(\hat{\lambda}_j, \hat{c})$ are obtained by an iterative algorithm and unbiased sample covariance matrices $(\hat{\boldsymbol{\Sigma}}_1, \hat{\boldsymbol{\Sigma}}_2)$ are defined as

$$\hat{\boldsymbol{\Sigma}}_1 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T, \qquad \hat{\boldsymbol{\Sigma}}_2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (\mathbf{y}_j - \bar{\mathbf{y}})(\mathbf{y}_j - \bar{\mathbf{y}})^T, \tag{3}$$

with samples $(\mathbf{x}_i, i = 1, \ldots, n_1)$ and $(\mathbf{y}_j, j = 1, \ldots, n_2)$ from two multivariate normal populations $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, respectively. Let $\mathbf{A}^T$ denote the transpose of $\mathbf{A}$. Independently, Eriksen (1987) obtained a Bartlett adjusted likelihood ratio test for $k \ge 2$ samples from multivariate normal populations. To consider the non-normal case, Schott (1999) gave a Wald test for samples from any k populations with finite fourth moments and fixed dimensions.

With the rapid development and wide applications of computer techniques, the data we collect and store become more and more huge [see Bai and Silverstein (2006)]. Some classical asymptotic theories are often no more valid or perform badly for large-dimensional data, because they often assume that the dimension p is fixed and the sample sizes tend to infinity, that is, the dimension must be very small with respect to the sample sizes. For example, as shown by Bai et al. (2009), Type I errors are dramatically enlarged for the classical Chi-square approximation of the likelihood ratio statistic for testing (1) taking $c = 1$. Simulation results in Bai et al. (2009) showed that the Type I errors were 21.2% and 49.5% for $(p, n_1, n_2) = (40, 800, 400)$ and $(80, 1600, 800)$, respectively. Thus when the dimension p is large compared with the sample sizes (n_1, n_2), the classical limit theorems for (1) taking $c = 1$ become invalid. All such problems and limitations bring up the need for further study on proportionality test of covariance matrices.

In this paper, we extend the work of Bai et al. (2009) to the proportionality test of two population covariance matrices. In fact, the equality test of two covariance matrices by Bai et al. (2009) is just a special case of (1) when $c = 1$. Moreover, in this paper, a pseudo-likelihood ratio test (*PLRT*) for (1) is proposed and its asymptotic normality is established. The proposed PLRT can be applied to Gaussian and non-Gaussian distributions with finite fourth moments for small and large dimension p.

The rest of the paper is organized as follows. In Section 2, some preliminary and random matrix theory (*RMT*) results are briefly reviewed. In Section 3, a pseudo-likelihood ratio test is proposed and its asymptotic normality is established. In Section 4, some simulations are performed to illustrate the proposed pseudo-likelihood test. Finally, this paper concludes in Section 5. Some proofs and technical derivations are postponed to the Appendix.

2. Review of some random matrix theories

Several results from RMT are reviewed for the proposed test procedure. For any $p \times p$ matrix $\mathbf{M}$ with real eigenvalues $\lambda_i^{\mathbf{M}}, i = 1, \ldots, p$, the *empirical spectral distribution* (ESD) is defined as

$$F_n^{\mathbf{M}}(x) = \frac{1}{p} \sum_{i=1}^{p} I(\lambda_i^{\mathbf{M}} \le x),$$

where $I(\cdot)$ is an indicator function. Consider a random matrix $\mathbf{M}$ whose ESD $F_n^{\mathbf{M}}$ converges (in a sense to be precised) to a *limiting spectral distribution* (LSD) $F^{\mathbf{M}}$. To make statistical inference about a parameter $\theta = \int f(x) dF^{\mathbf{M}}(x)$, it is natural to use an estimator

$$\hat{\theta} = \int f(x) dF_n^{\mathbf{M}}(x) = \frac{1}{p} \sum_{i=1}^{p} f(\lambda_i^{\mathbf{M}}),$$

which is a so-called *linear spectral statistic* (LSS) of $\mathbf{M}$.