Statistics and Probability Letters 117 (2016) 89–92

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

A note on tests for high-dimensional covariance matrices

Guangyu Mao *

School of Economics and Management, Beijing Jiaotong University, China

ARTICLE INFO

Article history:
Received 23 April 2016
Received in revised form 3 May 2016
Accepted 3 May 2016
Available online 22 May 2016

Keywords:
High dimension
Identity test
Sphericity test

ABSTRACT

This paper notes that two test statistics proposed by Chen et al. (2010) and another two recently developed by Srivastava et al. (2014) for sphericity and identity of covariance matrices respectively under non-normality are essentially the same except for a scale factor.

1. Introduction

Testing for the sphericity and identity of covariance matrices are two old topics in statistics. Suppose $\{x_i\}_{i=1}^N$ is a random sample from a population of random vectors of dimension p with positive-definite covariance matrix Σ_p. Based on the data, the two types of tests are designed to check

$$H_0 : \Sigma_p = \sigma^2 I_p \quad \text{against} \quad H_1 : \Sigma_p \neq \sigma^2 I_p,$$

and

$$\tilde{H}_0 : \Sigma_p = I_p \quad \text{against} \quad \tilde{H}_1 : \Sigma_p \neq I_p,$$

respectively, where σ^2 is an unknown but finite positive constant, and I_p is the identity matrix of dimension p. For simplicity, below we refer to the tests for H_0 and $\tilde{H}_0$ as sphericity test and identity test respectively.

When p is assumed to be fixed, the sphericity test and identity test have been well studied as documented, e.g., by Muirhead (1982). In recent years, growing attention has been paid to the two tests in the high-dimensional settings due to the increasing availability of big data sets, which typically have the feature that the sample size N is not much larger, or even far less, than the dimension p. Therefore, to test H_0 or $\tilde{H}_0$ in high dimensions, it is natural to postulate that p diverges as N approaches infinity, denoted by $(N, p) \to \infty$ in this paper. In the literature, related contributions include, but are not limited to, Ledoit and Wolf (2002), Srivastava (2005), Chen et al. (2010), Fisher et al. (2012), Fisher (2012) and Srivastava et al. (2014).

Of the existing papers, Chen et al. (2010) is the earliest one that investigates the high-dimensional sphericity test and identity test under non-normal population. Since $ptr(\Sigma_p^2)/[tr(\Sigma_p)]^2 - 1 \geq 0$ with equality holding if and only if H_0 is true, and $tr(\Sigma_p^2)/p - 2tr(\Sigma_p)/p + 1 \geq 0$ with equality holding if and only if $\tilde{H}_0$ is true, to formulate effective test statistics, it is

* Correspondence to: Siyuan East Building #807, Beijing Jiaotong University, Shang Yuan Cun #3, Beijing, 100044, China.
E-mail address: gymao@bjtu.edu.cn.

http://dx.doi.org/10.1016/j.spl.2016.05.002
0167-7152/

helpful to employ estimators of the unknown $tr(\Sigma_p)$ and $tr(\Sigma_p^2)$. Having this in mind, Chen, Zhang and Zhong (henceforth CZZ) proposed the following two estimators:

$$T_{1,CZZ} = Y_{1,N} - Y_{3,N}, \tag{1}$$

$$T_{2,CZZ} = Y_{2,N} - 2Y_{4,N} + Y_{5,N}, \tag{2}$$

where

$$Y_{1,N} = \frac{1}{N}\sum_{i=1}^{N} x_i'x_i, \quad Y_{3,N} = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i}^{N} x_i'x_j,$$

$$Y_{2,N} = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i}^{N}(x_i'x_j)^2, \qquad Y_{4,N} = \frac{1}{N(N-1)(N-2)}\sum_{i,j,k}^{*} x_i'x_jx_j'x_k,$$

$$Y_{5,T} = \frac{1}{N(N-1)(N-2)(N-3)}\sum_{i,j,k,l}^{*} x_i'x_jx_k'x_l,$$

in which $\sum^*$ denotes summation over mutually different indices. CZZ proved that $E(T_{1,CZZ}) = tr(\Sigma_p)$ and $E(T_{2,CZZ}) = tr(\Sigma_p^2)$ under non-normality. By virtue of the two unbiased estimators, CZZ formulated two statistics $\frac{N}{2}U_{CZZ}$ and $\frac{N}{2}V_{CZZ}$ to test H_0 and $\tilde{H}_0$ respectively, where $U_{CZZ} = p(\frac{T_{2,CZZ}}{T_{1,CZZ}^2}) - 1$ and $V_{CZZ} = \frac{1}{p}T_{2,CZZ} - \frac{2}{p}T_{1,CZZ} + 1$, and showed that $\frac{N}{2}U_{CZZ} \xrightarrow{D} N(0,1)$ under H_0 and $\frac{N}{2}V_{CZZ} \xrightarrow{D} N(0,1)$ under $\tilde{H}_0$ as $(N,p) \to \infty$, where $\xrightarrow{D}$ signifies convergence in distribution.

Based on the same idea, Srivastava et al. (2014) recently constructed another two estimators for $tr(\Sigma_p)$ and $tr(\Sigma_p^2)$. Let $\tilde{x}_i = x_i - \frac{1}{N}\sum_{k=1}^{N} x_k$, and $\tilde{X} = [\tilde{x}_1, \ldots, \tilde{x}_N]$. Then, the sample covariance matrix of x_i is $\hat{\Sigma}_p = \frac{1}{n}\tilde{X}\tilde{X}'$, where $n = N-1$. Using the traces of $\hat{\Sigma}_p$ and $\hat{\Sigma}_p^2$, the estimators proposed by Srivastava, Yanagihara and Kubokawa (henceforth SYK) have the following forms:

$$T_{1,SYK} = tr(\hat{\Sigma}_p), \tag{3}$$

$$T_{2,SYK} = \frac{1}{N(N-1)(N-2)(N-3)}\left\{(N-2)n^3 tr(\hat{\Sigma}_p^2) + n^2[tr(\hat{\Sigma}_p)]^2 - Nn\sum_{i=1}^{N}(\tilde{x}_i'\tilde{x}_i)^2\right\}. \tag{4}$$

SYK showed that $T_{1,SYK}$ and $T_{2,SYK}$ are also unbiased under non-normality for $tr(\Sigma_p)$ and $tr(\Sigma_p^2)$ respectively. As $(N,p) \to \infty$ with a restriction $N = O(p^\delta)$, where δ is a constant satisfying $1/2 < \delta < 1$, SYK further proved that $\frac{n}{2}U_{SYK} \xrightarrow{D} N(0,1)$ under H_0 and $\frac{n}{2}V_{SYK} \xrightarrow{D} N(0,1)$ under $\tilde{H}_0$, where $U_{SYK} = p(\frac{T_{2,SYK}}{T_{1,SYK}^2}) - 1$ and $V_{SYK} = \frac{1}{p}T_{2,SYK} - \frac{2}{p}T_{1,SYK} + 1$.

SYK's test statistics $\frac{n}{2}U_{SYK}$ and $\frac{n}{2}V_{SYK}$ look much like CZZ's $\frac{N}{2}U_{CZZ}$ and $\frac{N}{2}V_{CZZ}$ in form. The main difference between SYK's tests and CZZ's lies in the way they estimate $tr(\Sigma_p)$ and $tr(\Sigma_p^2)$. At first glance, it seems that SYK's $T_{1,SYK}$ and $T_{2,SYK}$ are much simpler than CZZ's $T_{1,CZZ}$ and $T_{2,CZZ}$, and SYK's estimators are more easily computed. In fact, SYK commented in their paper that CZZ's tests require "computing time of the order $O(N^4)$" since $T_{2,CZZ}$ in (2) "has summation over four indices", and theirs only require "computing time of the order $O(N^2)$". However, our finding below shows that SYK's estimators and CZZ's are exactly the same.

Claim 1. *SYK's $T_{1,SYK}$ in (3) and $T_{2,SYK}$ in (4) are identical to CZZ's $T_{1,CZZ}$ in (1) and $T_{2,CZZ}$ in (2) respectively. That is, $T_{1,SYK} = T_{1,CZZ}$ and $T_{2,SYK} = T_{2,CZZ}$.*

The claim suggests that CZZ's test statistics can be efficiently computed using the traces of $\hat{\Sigma}_p$ and $\hat{\Sigma}_p^2$ even though they have more complicated forms. Besides, it is also helpful to explain some phenomena in simulations. To illustrate this, we employ a simple simulation experiment about sphericity test here. Let $\{u_{ji}\}$ be a double array of *i.i.d. Gamma*$(4, \sqrt{2}/2)$ random variables, where 4 is the specified shape parameter and $\sqrt{2}/2$ is the specified scale parameter. To study the empirical size of the above two sphericity tests under 5% significance level, we sample x_i by

$$x_i = (x_{1i}, x_{2i}, \ldots, x_{pi})' = (u_{1i}, u_{2i}, \ldots, u_{pi})'.$$

To investigate the empirical power, we employ the following sampling scheme:

$$x_i = (x_{1i}, x_{2i}, \ldots, x_{pi})' = \frac{84}{85}(u_{1i}, \ldots, u_{pi})' + \frac{13}{85}(u_{2i}, \ldots, u_{p+1,i})'.$$

The test results based on 2000 replications are reported in Table 1. As we can find, the empirical size and power of SYK's sphericity test are always not larger than those of CZZ's test when (N,p) is given. In fact, this is a result caused by