Journal of Multivariate Analysis 141 (2015) 217–227

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

A robust test for sphericity of high-dimensional covariance matrices

Xintao Tian [a,b], Yuting Lu [c], Weiming Li [a,*]

[a] School of Science, Beijing University of Posts and Telecommunications, China
[b] School of Statistics, Renmin University of China, China
[c] School of Mathematics, Beijing Normal University, China

ARTICLE INFO

Article history:
Received 11 April 2015
Available online 5 August 2015

AMS 2010 subject classifications:
62H15
62H10

Keywords:
Covariance matrix
High-dimensional data
Sphericity test

ABSTRACT

This paper discusses the problem of testing the sphericity of a covariance matrix in high-dimensional frameworks. A new test procedure is put forward by taking the maximum of two existing statistics which are proved weakly independent in our settings. Asymptotic distribution of the new statistic is derived for generally distributed population with a finite fourth moment. Extensive simulations demonstrate that the proposed test has a great improvement in robustness of power against various models under the alternative hypothesis.

1. Introduction

Large-scale statistical inferences involving covariance matrices are increasingly encountered in many scientific research fields, such as signal processing, image processing, genetics, and stock marketing. A basic problem among such inferences is the sphericity test for covariance matrices when the number of observations is not negligible with respect to the sample size.

Generally, let $\mathbf{x}_1, \ldots, \mathbf{x}_n,\ \mathbf{x}_i \in \mathbb{R}^p$, be a sequence of independent and identically distributed zero-mean random vectors with a common population covariance matrix Σ_p. The sample covariance matrix takes the form $S_n = \sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'/n$. Hypotheses regarding to the sphericity test are

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

where σ^2 is the unknown scalar proportion. Our interest is to study this test based on S_n for general population, in an asymptotic framework where both p and n tend to infinity with $p/n \to c \in (0, \infty)$.

There are a number of works in the literature addressing the sphericity test in high-dimensions. Ledoit and Wolf [12] generalized the locally best invariant (LBI) test which was set up by [9,10] in a fixed p context. This result was later refined in [18] by applying an unbiased estimator of $\mathrm{tr}(\Sigma_p^2)/p$. Fisher et al. [8] studied a homogeneous test constructed from unbiased estimators of $\mathrm{tr}(\Sigma_p^k)/p,\ k = 2, 4$. However, these tests heavily rely on an assumption that the sample is normally distributed. For non-normal cases, Chen et al. [6] developed a new method where the statistic was constituted by some well selected U-statistics, but this technique carries a burden of doing extensive computations. Srivastava et al. [20] proved that the test in [18] is still valid when the kurtosis of the underlying distribution is close to 3, the Gaussian case. Recently, Wang and

* Corresponding author.
E-mail addresses: xttian@ruc.edu.cn (X. Tian), luyut99@sina.com (Y. Lu), liwm@bupt.edu.cn (W. Li).

http://dx.doi.org/10.1016/j.jmva.2015.07.010
0047-259X/

Yao [21] corrected the LBI test and the Gaussian likelihood ratio test [1] in high-dimensions for arbitrary distribution, and provided a finite fourth moment. For more references, one is referred to [11,19,5,15,4], etc.

Among all these tests, we are particularly interested in those from [18,8]. Asymptotic joint distribution of these two test statistics is derived for general populations under both the null and alternative hypotheses. With this joint distribution, we are surprised to find that the two tests are almost independent when the limiting ratio c is large, so that one test statistic has significant changes while the other may not, and vice versa. This inspires us to propose a new test procedure that can reject the sphericity hypothesis when any of the two statistics is large. It turns out that the new test achieves excellent performance on robustness of power compared with the original ones.

The rest of the paper is organized as follows. Section 2 reviews the unbiased estimators of $\mathrm{tr}(\Sigma_p^k)/p$ built on normality assumption and investigates their asymptotic behaviors for general population under moment conditions. In Section 3, we discuss the relationship between Srivastava's test and Fisher et al.'s test, and then formulate our new test procedure. Section 4 reports simulation results and Section 5 presents conclusions and remarks. Technical proofs are deferred to the last section.

2. Estimators of $\mathrm{tr}(\Sigma_p^k)/p$ and their asymptotic properties

Let H_p and F_n be spectral distributions of Σ_p and S_n, respectively. Then integer moments of H_p and F_n are defined by

$$\alpha_k := \int t^k dH_p(t) = \frac{1}{p}\mathrm{tr}(\Sigma_p^k) \quad \text{and} \quad \hat{\beta}_k := \int x^k dF_n(x) = \frac{1}{p}\mathrm{tr}(S_n^k),$$

$k = 0, 1, 2, \ldots$. Assuming the sample data are normally distributed, estimators of $\alpha_i,\ i = 1, 2, 3, 4$, employed in succession in [18,8,7] were proved to be unbiased, consistent, and asymptotically normal. Moreover, these estimators can be expressed as polynomials of $\hat{\beta}_k$'s, i.e.,

$$\begin{aligned}
\hat{\alpha}_1 &= \hat{\beta}_1, \\
\hat{\alpha}_2 &= \tau_2\left(\hat{\beta}_2 - c_n\hat{\beta}_1^2\right), \\
\hat{\alpha}_3 &= \tau_3\left(\hat{\beta}_3 - 3c_n\hat{\beta}_2\hat{\beta}_1 + 2c_n^2\hat{\beta}_1^3\right), \\
\hat{\alpha}_4 &= \tau_4\left(\hat{\beta}_4 - 4c_n\hat{\beta}_3\hat{\beta}_1 - \frac{2n^2+3n-6}{n^2+n+2}c_n\hat{\beta}_2^2 + \frac{10n^2+12n}{n^2+n+2}c_n^2\hat{\beta}_2\hat{\beta}_1^2 - \frac{5n^2+6n}{n^2+n+2}c_n^3\hat{\beta}_1^4\right),
\end{aligned}$$

where $c_n = p/n,\ \tau_2 = n^2/[(n-1)(n+2)],\ \tau_3 = n^4/[(n-1)(n-2)(n+2)(n+4)]$, and $\tau_4 = n^5(n^2+n+2)/[(n+1)(n+2)(n+4)(n+6)(n-1)(n-2)(n-3)]$. When the underlying distribution is not normal, we show that the unbiasedness does not hold any more for $\hat{\alpha}_2, \hat{\alpha}_3$, and $\hat{\alpha}_4$, but the consistency and asymptotic normality can be retained under suitable assumptions.

Assumption (a). The sample and population sizes n, p both tend to infinity, in such a way that $c_n = p/n \to c \in (0, \infty)$.

Assumption (b). There is a doubly infinite array of i.i.d. random variables $(w_{ij}),\ i,\ j \geq 1$, satisfying

$$E(w_{11}) = 0, \qquad E(w_{11}^2) = 1, \qquad E(w_{11}^4) < \infty,$$

such that for each p, n, letting $W_n = (w_{ij})_{1\leq i\leq p, 1\leq j\leq n}$, the observation vectors can be represented as $\mathbf{x}_j = \Sigma_p^{1/2} w_{\cdot j}$ where $w_{\cdot j} = (w_{ij})_{1\leq i\leq p}$ denotes the jth column of W_n.

Assumption (c). The population spectral distribution H_p of Σ_p weakly converges to a probability distribution H, as $p \to \infty$, and the sequence of spectral norms $(\|\Sigma_p\|)$ is bounded.

Assumptions (a)–(c) are classical conditions of the central limit theorem for linear spectral statistics of sample covariance matrices, see [2,3]. From the third assumption, moments of H_p converge to the corresponding moments of H, that is,

$$\alpha_k \to \tilde{\alpha}_k := \int t^k dH(t),$$

as $p \to \infty$, for any fixed $k \in \mathbb{N}$.

Lemma 1. *Suppose that Assumptions (a)–(c) hold, then*

(i) *the estimator $\hat{\alpha}_k$ is strongly consistent, i.e.,*

$$\hat{\alpha}_k - \alpha_k \xrightarrow{a.s.} 0, \quad k = 1, 2, 3, 4.$$