Statistics and Probability Letters 156 (2020) 108613

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Some sphericity tests for high dimensional data based on ratio of the traces of sample covariance matrices

Xue Ding

School of Mathematics, Jilin University, Changchun, 130012, China

ARTICLE INFO

Article history:
Received 3 August 2018
Received in revised form 31 August 2019
Accepted 1 September 2019
Available online 12 September 2019

Keywords:
High dimensional data
Sphericity test
Random matrix theory
Spiked population model

ABSTRACT

In this paper, we investigate the sphericity test for high dimensional data. The test statistic is proposed based on the ratio of traces of sample covariance matrices. The asymptotic distributions of the test statistics under the null hypothesis are established by linear spectral statistics of large sample covariance matrices. We also obtain the asymptotic power functions in the case of the spiked population model as a specific alternative. Compared with some existing tests, the proposed test can handle data having unknown means and non-Gaussian population with general fourth moment. The numerical performance demonstrates that the proposed tests have satisfactory properties in terms of size and power.

1. Introduction

Population covariance matrices tests are an important topic in multivariate statistical analysis. In classical statistical analysis described by Anderson (2003), the test statistics are proposed under the assumption that the sample size n tends to infinity while the dimension p remains fixed. However, this assumption is not suitable for high dimensional data, where the dimension p is proportional to the sample size or even large than the sample size. High-dimensional data analysis has attracted much more attention recently due to its extensive application in many scientific research fields, such as signal processing, image processing, genetics, and stock market analysis. In this work, we consider the sphericity test for covariance matrices when the dimension of observations is as large as or even large than the sample size. Let $X_1, \ldots, X_n, X_i \in \mathbb{R}^p$, be a sequence of independent and identically distributed random vectors with a common population covariance matrix Σ_p. Hypotheses related to the sphericity test are

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

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

The traditional test tools such as those described in Anderson (2003) cannot be used or perform poorly for covariance matrices with high-dimension, since these procedures are justified under a framework where the sample size n tends to infinity while the dimension p remains fixed. Bai and Saranadasa (1996) demonstrate that classical multivariate statistical procedures need to be reexamined when the dimension is high. When $y < 1$, Bai et al. (2009) provide a necessary correction to the classic likelihood ratio test and obtain its central limit theory based on random matrix theories (RMTs). Ledoit and Wolf (2002) (LW test) developed a correction to the existing test statistic for the normally distributed samples that makes it robust against high dimensions. Later, Birke and Dette (2005) extended the LW test to cases where $y = 0$ and

E-mail address: dingxue83@jlu.edu.cn.

https://doi.org/10.1016/j.spl.2019.108613
0167-7152/

∞. For non-normal cases, Chen et al. (2010) proposed a nonparametric test statistic which can accommodate situations where the data dimension is much larger than the sample size. Wang and Yao (2013) propose corrections to the likelihood ratio test and John's test for sphericity in large-dimensions and drive the asymptotic distribution of the two proposed test statistics under the null hypothesis. These asymptotic distributions are valid for a general population having a finite fourth-moment. For other related references, refer to Birke and Dette (2005), Forzani et al. (2017), Onatski et al. (2013), Tian et al. (2015), Wang et al. (2013), Zou et al. (2014) etc.

Srivastava (2005) proposed a test statistic based on the consistent estimator of $tr\,\Sigma_p/p$, $tr\,\Sigma_p^2/p$ for normally distributed samples under more general condition $n = O(p^\delta)$, $0 < \delta < 1$. Later, Fisher et al. (2010) developed a test statistic based on the unbiased estimators of $tr\,\Sigma_p^k/p$, $k = 2, 4$ for normally distributed samples under the condition that $p/n \to y \in (0, \infty)$. Tian et al. (2015) consider the maximum of two existing statistics which are derived for generally a distributed population with a finite fourth moment. In this paper, inspired by Fisher et al. (2010), Srivastava (2005), we use the ratio of trace for sample covariance matrices to construct two test statistics. Instead of constructing unbiased estimators for traces of population covariance matrices such as the above three paper have done, we directly study the asymptotic distribution of the traces of sample covariance matrices by using random matrix theory. The proposed test statistics have some advantages. First, our test can accommodate data with non-Gaussian distributions with unknown mean. The tests in Wang and Yao (2013) and Tian et al. (2015) are also suitable for non-Gaussian distribution, but they assume that sample mean is zero. Compared with the test described in Chen et al. (2010), the proposed test statistics have simple expression and we can obtain their joint distribution. Thus, we can use the best of them in order to improve the test power. The simulation results reveal that the proposed test is more powerful than that of Chen et al. (2010). In addition, due to the simple expression of the proposed test statistics, the study of the power is easier than the existing test statistics and we can obtain the explicit power in the case of the spiked population model as a specific alternative.

The outline of this paper is as follows. In Section 2, we introduce the multivariate model, give the test statistic and obtain the asymptotic distribution under null hypothesis, and finally investigate the asymptotic power under the alternative hypothesis. In Section 3, we report the simulation results. In Section 4, we provide our conclusions. All proofs are described in Section 5.

2. Main results

We consider the following multivariate model which was introduced by Bai and Saranadasa (1996). Let

$$X_i = \Sigma_p^{1/2} Y_i + \mu, \quad \text{for} \quad i = 1, \ldots, n, \tag{2.1}$$

where μ is a p-dimensional constant vector and $Y_i = \{y_{ri}, r = 1, \ldots, p\}$ with $\{y_{ij}, i, j = 1, 2, 3, \ldots\}$ is a double array of independent and identically distributed real random variables satisfying $Ey_{ij} = 0, Var(y_{ij}) = 1, Ey_{ij}^4 = 3 + \Delta$ and $\Sigma_p = (\Sigma_p^{1/2})^2$ is a non-random symmetric nonnegative definite matrix which is the population covariance matrix of X_i. The sample covariance is defined by

$$S_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T,$$

where $\bar{X} = (1/n)\sum_{i=1}^n X_i$. Using this multivariate model, we will show that out test statistics are only dependent on the traces of S_n. Thus, the proposed tests can accommodate data which has unknown means.

2.1. The proposed test statistics

Since the test statistics is invariant, we may assume that $\Sigma_p = diag(\lambda_1, \ldots, \lambda_p)$ without loss of generality. From the Cauchy–Schwarz inequality, it follows that for any positive integer r

$$\psi_r = \frac{1}{p} tr\,\Sigma_p^{2r} \Big/ \left(\frac{1}{p} tr\,\Sigma_p^r\right)^2 \geq 1$$

with equality holding if and only if $\lambda_1 = \cdots = \lambda_p = \lambda$. Thus, we may consider testing $H_0 : \psi_r = 1$ vs. $H_A : \psi_r > 1$ instead of testing $H_0 : \Sigma_p = \sigma^2 I_p$ $vs. H_1 : \Sigma_p \neq \sigma^2 I_p$. For a normal distribution, Srivastava (2005) and Fisher et al. (2010) provided consistent estimators of $(1/p)tr\,\Sigma_p^m$, $m = 1, 2$ and $m = 3, 4$ respectively. Based on these estimators they established corresponding test statistics. In this paper, instead of estimating the $tr\,\Sigma_p^k$, we directly investigate joint distribution of arithmetic means of the eigenvalues of the sample covariance matrix $(1/p)trS_n^m$ for general distribution. For any positive integer r, define

$$\varphi_r = \frac{1}{p} trS_n^{2r} \Big/ \left(\frac{1}{p} trS_n^r\right)^2.$$

We can now provide necessary corrections to the asymptotic distribution of φ_r which can be used as a test statistic to address the high-dimensional effects as Bai et al. (2009) have done for the likelihood ratio test statistic.

Theorem 1. *For the above multivariate statistical model (2.1), as $y_n = p/n \to y \in (0, \infty)$,*

$$(trS_n - p\omega_1, trS_n^2 - p\omega_2, trS_n^4 - p\omega_4)^T \xrightarrow{L} N(\theta, \Xi),$$

where

$$\omega_1 = 1, \quad \omega_2 = 1 + y_n, \quad \omega_4 = (1 + y_n)^3 + 3y_n(1 + y_n),$$

$\theta = (\theta_1, \theta_2, \theta_3) = (-y, -y(y+2), -y^3 - 6y^3 - 34y^2 + 2y)^T$, $\Xi = (\xi_{ij})$ *and* $\xi_{ij} = Cov(f_i, f_j)$ *with* $f_1 = x, f_2 = x^2, f_3 = x^4$ *and*

$$\begin{aligned} Cov(x^r, x^m) = {} & 2y^{m+r} \sum_{k_1=0}^{r-1} \sum_{k_2=0}^{m} \binom{r}{k_1}\binom{m}{k_2}\left(\frac{1-y}{y}\right)^{k_1+k_2} \sum_{l=1}^{r-k_1} l\binom{2r-1-(k_1+l)}{r-1}\binom{2m-1-k_2+l}{m-1} \\ & + \Delta y^{r+m+1} \sum_{k_1=0}^{r} \sum_{k_2=0}^{m} \binom{r}{k_1}\binom{m}{k_2}\left(\frac{1-y}{y}\right)^{k_1+k_2} \binom{2r-k_1}{r-1}\binom{2m-k_2}{m-1}. \end{aligned} \tag{2.2}$$

Theorem 2. *For the above multivariate statistical model (2.1), under the null hypothesis, as $y_n = p/n \to y \in (0, \infty)$,*

$$p\left(\varphi_1 - (1 + y_n)\right) \xrightarrow{L} N(y^2 + (\Delta + 1)y, 4y^2), \quad p\left(\varphi_2 - \left(1 + y_n + \frac{3y_n}{(1+y_n)}\right)\right) \xrightarrow{L} N(\mu_2, \sigma_2^2),$$

where

$$\begin{aligned} \mu_2 &= -2y(\Delta - y - 1)\left(1 + \frac{3y}{(1+y)^2}\right) + \frac{y}{(1+y)^2}(-y^3 + 6y^2(\Delta - 1) + (16\Delta - 1)y + 2 + 6\Delta), \\ \sigma_2^2 &= \frac{1}{\omega_2^4}\xi_{33} + \frac{4\omega_4^2}{\omega_2^6}\xi_{22} - \frac{4\omega_4}{\omega_2^5}\xi_{23}. \end{aligned} \tag{2.3}$$

We notice that the asymptotic distributions of statistics φ_1 and φ_2 are dependent on Δ. This is a common phenomena in central limit theorem of linear spectral statistics of high-dimensional sample covariance matrix and some tests about the high-dimensional covariance matrix, for more details, one can see Bai and Silverstein (2010), Chen et al. (2010), Pan (2014) and Tian et al. (2015). In applications, this constant is same as the mean and variance of samples and can be estimated by classical moment estimate method. After normalization the asymptotic distributions of statistics φ_1 and φ_2 are standard normal distribution and the asymptotic distribution is same if we replaced Δ by its estimator because that Δ's estimator is consistent.

In fact, for any integer r, we can obtain the asymptotic distribution of φ_r by using random matrices theory. However, the power function has a relatively simple form only for $r = 1, 2$ as discussed in next section. Therefore, in this paper, we only investigate the statistics φ_1, φ_2 for the sphere test of covariance matrices.

From the simulation study, we learn that the test statistics φ_1, φ_2 are both satisfactory in terms of the size, but for some population covariance matrices φ_1 is more powerful and for other population covariance matrices φ_2 is more powerful. Considering that we can obtain the joint distribution of φ_1, φ_2, we take the maximum of standardized φ_1, φ_2 in order to improve the power. Let

$$\Psi_1 = \max\left(\frac{p(\varphi_1 - (1 + y_n)) - (y_n^2 + (\Delta + 1)y_n)}{2y_n}, \frac{p\left(\varphi_2 - (1 + y_n + 3y_n/(1+y_n)) - \hat{\mu}_2\right)}{\hat{\sigma}_2}\right).$$

Here $\hat{\mu}_2, \hat{\sigma}_2$ are μ_2, σ_2 replaced y with y_n.

Theorem 3. *Assume that the conditions in Theorem 1 hold, we have under null hypothesis*

$$P(\Psi_1 \le x) \longrightarrow \int_{-\infty}^{x}\int_{-\infty}^{x} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{t^2 - 2\rho ts + s^2}{2(1-\rho^2)}\right\} dtds,$$

where

$$\rho = \frac{1}{2y\sigma_2} u^t \Xi v, \ u = (-2(1+y), 1, 0)^T, \ v = (0, -\frac{2\delta}{\omega_2}, \frac{1}{\omega_2^2})^T, \ \delta = 1 + y + \frac{3y}{1+y}.$$

2.2. Asymptotic power under spiked model

For the high-dimensional test for the covariance matrix in multivariate statistical analysis, it is difficult to investigate the power function under alternative hypothesis. In this subsection , we provide an asymptotic power function for the new statistic φ_1, φ_2 for testing the presence of spike eigenvalues in the population covariance matrix.

For spiked population model, the population covariance matrix is as following

$$\Sigma_p = \begin{pmatrix} \Sigma_{M\times M} & 0 \\ 0 & I_{p-M} \end{pmatrix}$$

where the matrix

$$\Sigma_{M\times M} = \{\underbrace{\alpha_1, \ldots, \alpha_1}_{n_1}, \ldots, \underbrace{\alpha_k, \ldots, \alpha_k}_{n_k}\}. \tag{2.4}$$

Here $n_1, \ldots, n_k$ are fixed and $\sum_{i=1}^{k} n_i = M$. We assume that there are k_1 distinct spikes such that $\alpha_i > 1+\sqrt{y}$ and $M-k_1$ closed spikes such that $1 < \alpha_i \le 1+\sqrt{y}$.

The power function of this test under the general alternative hypothesis is ill-defined. Let us instead consider the null hypothesis $H_0 = I_p$ against an alternative hypothesis of the form:

H_1 : Σ_p has the spiked structure (2.4).

Theorem 4. *If Σ_p has the spiked structure (2.4), then we have*

$$p\left(\varphi_1 - \kappa_2/\kappa_1^2\right) \xrightarrow{L} N(y^2 + (\Delta+1)y, 4y^2), \quad p\left(\varphi_2 - \kappa_4/\kappa_2^2\right) \xrightarrow{L} N(\mu_2, \sigma_2^2)$$

where $\kappa_i, i = 1, 2, 4$ are given by (5.11), μ_2, σ_2^2 are same as in Theorem 1.

Let

$$\Psi_2 = \max\left(\frac{p(\varphi_1 - \kappa_2/\kappa_1^2) - (y_n^2 + (\Delta+1)y_n)}{2y_n}, \frac{p\left(\varphi_2 - \kappa_4/\kappa_2^2\right) - \hat{\mu}_2}{\hat{\sigma}_2}\right).$$

Here $\hat{\mu}_2, \hat{\sigma}_2$ are μ_2, σ_2 replaced y with y_n.

Theorem 5. *Assuming that the conditions in Theorem 4 hold, we have*

$$P(\Psi_2 \le x) \longrightarrow \int_{-\infty}^{x}\int_{-\infty}^{x} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{t^2 - 2\rho ts + s^2}{2(1-\rho^2)}\right\} dtds,$$

where

$$\rho = \frac{1}{2y\sigma_2} u^t \Xi v, \ u = (-2(1+y), 1, 0)^T, \ v = (0, -\frac{2\delta}{\omega_2}, \frac{1}{\omega_2^2})^T, \ \delta = 1 + y + \frac{3y}{1+y}.$$

Based on the conclusions in Theorem 4, we can obtain the asymptotic power function under the spiked population model.

Theorem 6. *For spiked population model as the alternative hypothesis, the asymptotic power function of the two test statistics are*

$$\beta_1(\alpha) = 1 - \Phi\left(\Phi^{-1}(1-\alpha) - \frac{\sum_{i=1}^{k}(1-\alpha_i)^2 n_i}{\sigma_1}\right),$$
$$\beta_2(\alpha) = 1 - \Phi\left(\Phi^{-1}(1-\alpha) - \frac{\pi/(1+y^2)}{\sigma_2}\right).$$

where

$$\pi = \sum_{i=1}^{k} n_i\alpha_i^4 + 4y\sum_{i=1}^{k} n_i\alpha_i^3 + (4y+6y^2)\sum_{i=1}^{k} n_i\alpha_i^2 + (4y+12y^2+4y^3)\sum_{i=1}^{k} n_i\alpha_i - 3My(2+4y+y^2)$$
$$- M(1+y)(1+5y+y^2) - 2(1+5y+y^2)(2y\sum_{i=1}^{k} n_i\alpha_i - 2yM + \sum_{i=1}^{k} n_i\alpha_i^2 - M).$$

3. Simulation results

In this section, we will demonstrate the performance of the proposed test statistic Ψ_1 compared with other existing tests by using a simulation. All simulation results are based on 1000 repetitions. First, we report the empirical size of Ψ_1 and the CZZ test in Chen et al. (2010) and FSG test in Fisher et al. (2010) at a significance level $\alpha = 0.05$ with sample size $n = 50, 100, 150, 200$ and dimension $p = 35, 75, 185, 325$ respectively. In Tables 1 and 2, $\Sigma_p = I_p$ and $(y_{ij})_{p\times n}$ are i.i.d. normal distribution $N(\mu, 1)$ with $\mu = 0$ and 1 respectively. Since the FSG test can only handle normally distributed samples, we compared Ψ_1 with CZZ test statistic in Table 3. Here, $\Sigma_p = I_p$ and $(y_{ij})_{p\times n}$ are i.i.d. Gamma(4,0.5) distribution. The three tables show that the empirical size of the proposed test is close to 0.05 for nearly all combinations of dimension p and sample size n. In contrast, the other tests are not so effective. The empirical size of FSG test is deviated from the nominal size for the combination of large dimension p and small sample size n under the normal distribution.

Table 1
Normal distribution N(0,1).

p\ n	Ψ_1 test				FSG test				CZZ test			
	50	100	150	200	50	100	150	200	50	100	150	200
35	0.049	0.05	0.051	0.056	0.043	0.058	0.057	0.049	0.051	0.06	0.055	0.058
75	0.054	0.044	0.042	0.046	0.061	0.05	0.052	0.06	0.051	0.058	0.064	0.065
185	0.05	0.045	0.048	0.049	0.055	0.062	0.051	0.046	0.052	0.044	0.056	0.056
325	0.049	0.047	0.05	0.052	0.036	0.058	0.062	0.063	0.053	0.06	0.043	0.061

Table 2
Normal distribution N(1,1).

p\ n	Ψ_1 test				FSG test				CZZ test			
	50	100	150	200	50	100	150	200	50	100	150	200
35	0.046	0.05	0.048	0.046	0.046	0.039	0.048	0.061	0.052	0.064	0.061	0.057
75	0.05	0.048	0.051	0.052	0.043	0.054	0.049	0.048	0.053	0.054	0.049	0.05
185	0.048	0.05	0.053	0.053	0.06	0.037	0.058	0.054	0.048	0.065	0.06	0.049
325	0.046	0.045	0.053	0.052	0.054	0.051	0.051	0.056	0.058	0.065	0.057	0.046

Table 3
Gamma distribution Gamma(4,0.5).

p\ n	Ψ_1 test				CZZ test			
	50	100	150	200	50	100	150	200
35	0.051	0.054	0.045	0.053	0.055	0.06	0.066	0.055
75	0.053	0.055	0.055	0.048	0.06	0.045	0.062	0.059
185	0.042	0.053	0.049	0.047	0.048	0.057	0.065	0.067
325	0.054	0.045	0.049	0.052	0.049	0.065	0.053	0.061

The empirical size of CZZ test is not as satisfactory as the empirical size of the proposed test for the combination of large dimension p and large sample size n

To investigate the power of the two statistics, we consider the alternative hypothesis: $\Sigma^{(1)} = diag(2I_{[0.05p]}, I_{p-[0.05p]})$. We report the empirical power at a significance level $\alpha = 0.05$ with dimension $p = 40, 80, 120, 160, 200,240$ and sample size $n = 35, 75, 115, 155, 195, 235$ respectively. In Fig. 1, $(y_{ij})_{p\times n}$ are i.i.d. normal distribution $N(0, 1)$. Since the FSG test can only handle normally distributed samples, we compare the empirical power of Ψ_1 with CZZ test statistic in Fig. 2. Here, $(y_{ij})_{p\times n}$ are i.i.d. standardized Gamma(4,0.5) distribution. From Fig. 1, we can see that the empirical power of FSG test is not good as the empirical power of other tests for large dimension p and small sample size n under normal assumption. Fig. 2 shows that the proposed test is more powerful than the CZZ test for nearly all combinations of dimension p and sample size n.

4. Conclusion

In this article, we study the sphericity test for high dimensional covariance matrix when the dimension is the same as or even larger than the sample size. The proposed test statistics are based on the Cauchy–Schwarz inequality for the population eigenvalues and they can accommodate the data with unknown means and non-Gaussian distributions. Under the null hypothesis, we obtain the asymptotic distributions of these statistics using CLT of LSS of sample covariance matrices. Compared with existing tests, out test statistics have more simple expression. The proposed test statistics make it easy to study the power under alternative hypothesis. In addition, as the alternative hypothesis is from spiked model, we established the asymptotic power function of the test statistics.

5. Proof of main conclusion

We first review key items from random matrices theory. For any $p \times p$ square matrix M with real eigenvalue λ_i^M, F^M denotes the empirical spectral distribution (ESD) of M, that is,

$$F^M(x) = \frac{1}{p}\sum_{i=1}^{p} I_{\{\lambda_i^M \le x\}}, \quad x \in \mathbb{R}.$$

The Stieltjes transform of F^M is defined by

$$s_{F^M}(z) = \int \frac{1}{\lambda - z} dF^M(x) = \frac{1}{p} tr(M - zI_p)^{-1}.$$

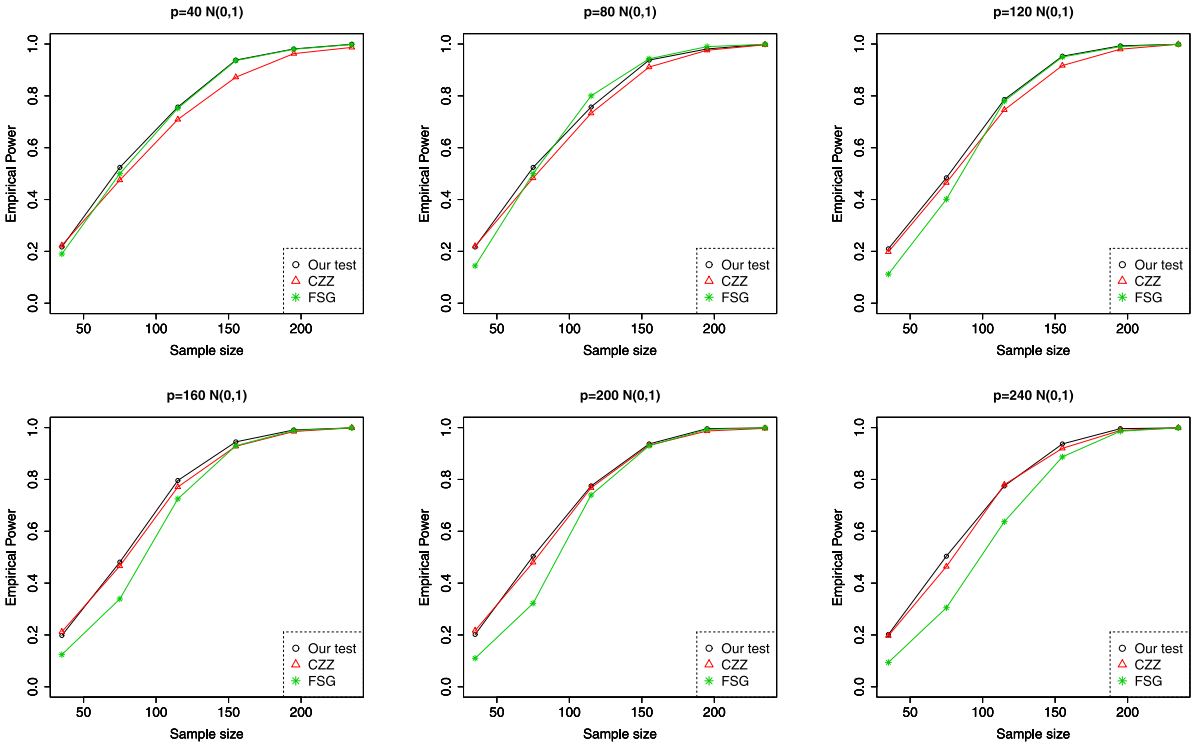

Fig. 1. The empirical power for $\Sigma^{(1)} = diag(2I_{[0.05p]}, I_{p-[0.05p]})$ at $\alpha = 0.05$ significance level under normal assumption.

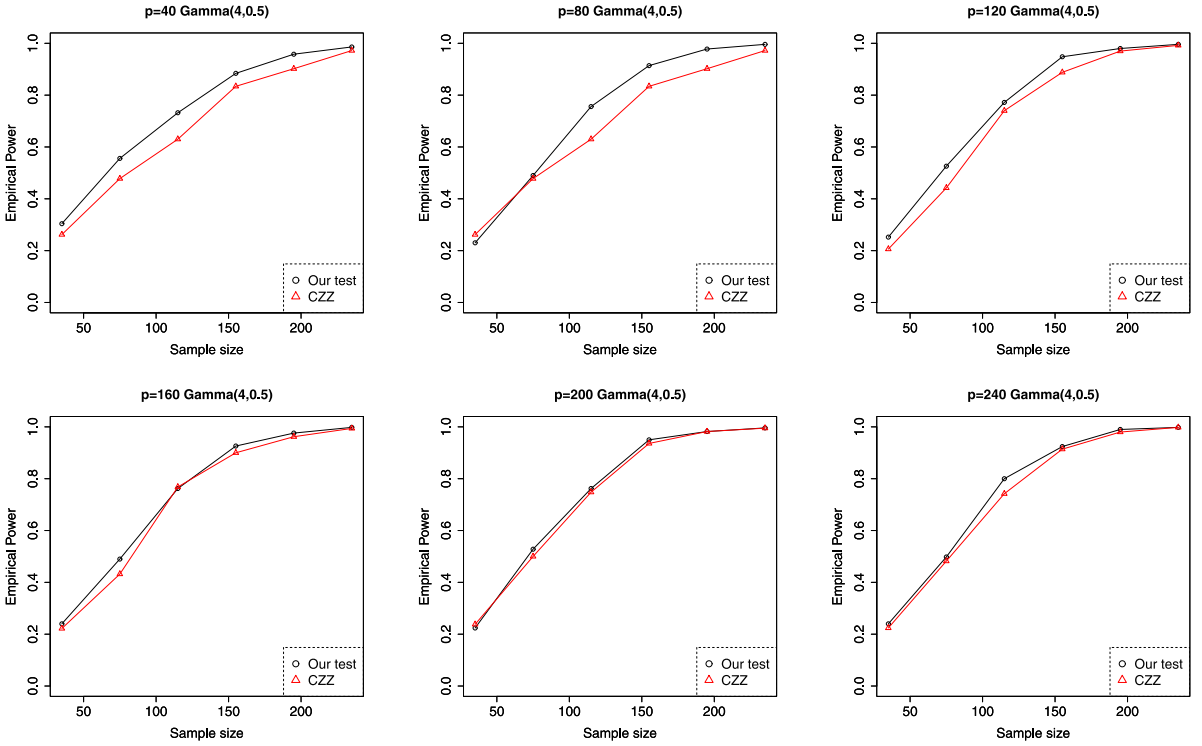

Fig. 2. The empirical power for $\Sigma^{(1)} = diag(2I_{[0.05p]}, I_{p-[0.05p]})$ at $\alpha = 0.05$ significance level under Gamma distribution.

From the random matrix theory in Bai and Silverstein (2010), if the mean vector $\mu = 0$ in (2.1) and the ESD of Σ_p converges to a nonrandom probability distribution H as $p/n \to y \in (0, \infty)$, then the ESD of $\mathcal{S}_n$ converges in distribution to a nonrandom distribution function $F^{y,H}(x)$ whose Stieltjes transform $s(z) = s_{F^{y,H}}(z)$ is, for each $z \in \mathbb{C}^+ = \{z \in \mathbb{C} : \Im z > 0\}$, the unique solution to the equation

$$s = \int \frac{1}{t(1-y-yzs)-z} dH(t). \tag{5.5}$$

If $\Sigma_p = I_p$, the ESD of $\mathcal{S}_n$ converges in distribution to the famous Marčenko–Pastur distribution of index y, denoted as F^y, which is the distribution on $[a_y, b_y]$ with the following density function

$$g_y(x) = \begin{cases} \frac{1}{2\pi yx}\sqrt{[b_y - x][x - a_y]}, & \text{if } a_y \le x \le b_y; \\ 0, & \text{otherwise.} \end{cases}$$

where $a_y = (1-\sqrt{y})^2$, $b_y = (1+\sqrt{y})^2$. Let $\mathcal{U}$ be an open set of the complex plane, including $[\liminf \lambda_{\min}(\Sigma_p) I_{(0,1)}(y) a_y, \limsup \lambda_{\max}(\Sigma_p) b_y]$, and $\mathcal{A}$ be the set of analytic functions $f : \mathcal{U} \mapsto \mathbb{R}$. We consider the empirical process $G_n := \{G_n(f)\}$ indexed by $\mathcal{A}$,

$$G_n(f) = p \int_{-\infty}^{+\infty} f(x)[F^{\mathcal{S}_n} - F^{y_n, H_n}](dx), \quad f \in \mathcal{A},$$

where $F^{\mathcal{S}_n}$ is empirical distribution of $\mathcal{S}_n$ and F^{y_n,H_n} denotes the probability distribution function by substituting $y_n = p/n$ for y, H_n for H in $F^{y,H}$. We then have following conclusion which is useful in computing the asymptotic distribution of the test statistic under the null and alternative hypothesis.

Proposition 1. *For the multivariate statistical model (2.1), assume that $Ey_{11} = 0, Var(y_{11}) = 1, E(y_{11})^4 = 3 + \Delta$ and the condition*

$$\frac{1}{np}\sum_{ij} E(y_{ij})^4 I(|Y_{ij}| > \eta\sqrt{n}) \to 0$$

holds for any fixed $\eta > 0$. Then, for $\Sigma_p = I_p$ and the spiked model in (2.4), the random vector $(G_n(f_1), \ldots, G_n(f_k))$ weakly converges to a k-dimensional Gaussian vector with mean vector,

$$\begin{aligned} m(f_j) = &-\frac{1}{2\pi i}\oint \frac{f_j(z) y \underline{s}^3(z)}{\left(1 - \frac{y\underline{s}^2(z)}{(1+\underline{s}(z))^2}\right)^2 (1+\underline{s}(z))^3} dz + \frac{1}{2\pi i}\oint \frac{f_j(z) y \underline{s}(z)}{z\left(1 - \frac{y\underline{s}^2(z)}{(1+\underline{s}(z))^2}\right)(1+\underline{s}(z))^2} dz \\ &- \frac{\Delta}{2\pi i}\oint \frac{f_j(z) y \underline{s}^3(z)}{\left(1 - \frac{y\underline{s}^2(z)}{(1+\underline{s}(z))^2}\right)(1+\underline{s}(z))^3} dz, \quad j = 1, \ldots, k, \end{aligned}$$

and covariance function

$$\begin{aligned} Cov(f_i, f_j) = &-\frac{1}{2\pi^2}\oint\oint \frac{f_i(z_1) f_j(z_2)}{(\underline{s}(z_1) - \underline{s}(z_2))^2} d\underline{s}(z_1) d\underline{s}(z_2) \\ &- \frac{y\Delta}{4\pi^2}\oint\oint \frac{f_i(z_1) f_j(z_2)}{(1+\underline{s}(z_1))^2 (1+\underline{s}(z_2))^2} d\underline{s}(z_1) d\underline{s}(z_2), \quad \{i, j \in 1, \ldots, k\}. \end{aligned}$$

where $f_1, \ldots, f_k \in \mathcal{A}$, $\underline{s}(z)$ is the Stieltjes transform of $\underline{F}^y = (1-y) I_{(0,+\infty)} + yF^y$. The contours in the above are nonoverlapping and both contain the support of F^y taken in the positive direction in the complex plane.

Proof. These conclusions can be obtained by Theorem 1 in Pan (2014). ■

Proof of Theorem 1. Under the null hypothesis, that is $\Sigma_p = I_p$, the probability distribution F^{y_n,H_n} can easily be shown to be the Marčenko–Pastur law of index y_n denoted by F^{y_n}, by using the variable change $x = 1 + y_n - 2\sqrt{y_n}\cos\theta, 0 < \theta < \pi$, we get

$$\begin{aligned} \omega_m = \int_{-\infty}^{+\infty} x^m dF^{y_n} &= \int_{a_{y_n}}^{b_{y_n}} \frac{x^m}{2\pi x y_n}\sqrt{[b_{y_n} - x][x - a_{y_n}]} \\ &= \frac{1}{2\pi y_n}\int_0^{\pi} (1 + y_n - 2\sqrt{y_n}\cos\theta)^{m-1} 4y_n \sin^2\theta d\theta \\ &= \frac{2}{\pi}\sum_{k=0}^{m-1} C_{m-1}^k (1+y_n)^{m-1-k}(-2\sqrt{y_n})^k \int_0^{\pi} \cos^k\theta \sin^2\theta d\theta. \end{aligned}$$

Then we obtain

$$\omega_1 = 1, \quad \omega_2 = 1 + y_n, \quad \omega_4 = (1 + y_n)^3 + 3y_n(1 + y_n). \tag{5.6}$$

Then, by Proposition 1, we conclude that

$$(trS_n - p\omega_1, trS_n^2 - p\omega_2, trS_n^4 - p\omega_4)^T \xrightarrow{L} N(\theta, \Xi), \tag{5.7}$$

where $\theta = (\theta_1, \theta_2, \theta_3) = (m(f_1), m(f_2), m(f_3))^T$, $\Xi = (\xi_{ij})$ and $\xi_{ij} = Cov(f_i, f_j)$ with $f_1 = x, f_2 = x^2, f_3 = x^4$. Now we compute θ, Ξ. Replacing H with δ_1 in (5.5), and noting the relationship between s and $\underline{s}$, we know that

$$z = -\frac{1}{\underline{s}} + \frac{y}{1 + \underline{s}}.$$

Using Proposition A.1, we know that

$$\begin{aligned} m(f) = &-\frac{1}{2\pi i}\oint \frac{y\underline{s} f\left(-\frac{1}{\underline{s}} + \frac{y}{1+\underline{s}}\right)}{(1 + \underline{s})((1 + \underline{s})^2 - c\underline{s}^2)} d\underline{s} - \frac{\Delta}{2\pi i}\oint \frac{y\underline{s} f\left(-\frac{1}{\underline{s}} + \frac{y}{1+\underline{s}}\right)}{(1 + \underline{s})^3} d\underline{s} \\ &+ \frac{y}{2\pi i}\oint \frac{f\left(-\frac{1}{\underline{s}} + \frac{y}{1+\underline{s}}\right)}{(1 + \underline{s})(y\underline{s} - 1 - \underline{s})} d\underline{s} := I_1 + I_2 + I_3, \end{aligned}$$

where the integral is taken on a contour counterclockwise, when restricted to the real axes, encloses the interval $[\frac{-1}{1-\sqrt{y}}, \frac{-1}{1+\sqrt{y}}]$ when $0 < y < 1$, and encloses the interval $[-1, \frac{-1}{1+\sqrt{y}}]$ when $y \geq 1$. The first and second terms in above equation for $f(x) = x^r$ can be obtained by Bai and Silverstein (2010) and Pan and Zhou (2008). They are

$$\begin{aligned} I_1 &= \frac{1}{4}((1 - \sqrt{y})^{2r} + (1 + \sqrt{y})^{2r}) - \frac{1}{2}\sum_{j=0}^{r}\binom{r}{j}^2 y^j \\ I_2 &= y^{1+r}\sum_{j=0}^{r}\binom{r}{j}\left(\frac{1-y}{y}\right)^j\binom{2r-j}{r-1} - y^{1+r}\sum_{j=0}^{r}\binom{r}{j}\left(\frac{1-y}{y}\right)^j\binom{2r+1-j}{r-1} \end{aligned}$$

For the third term I_3, $\underline{s} = -1$ is the only pole, the residue at $\underline{s} = -1$ for $f(x) = x^r$, $r = 1, 2, 4$ can be calculated respectively as

$$-y, \quad -y(y + 2), \quad -y[(y - 1)^3 + 35 + 45(y - 1) + 15(y - 1)^2].$$

Also, from Bai and Silverstein (2010) and Pan and Zhou (2008), we see that

$$\begin{aligned} Cov(x^r, x^m) = &\ 2y^{m+r}\sum_{k_1=0}^{r-1}\sum_{k_2=0}^{m}\binom{r}{k_1}\binom{m}{k_2}\left(\frac{1-y}{y}\right)^{k_1+k_2}\sum_{l=1}^{r-k_1} l\binom{2r-1-(k_1+l)}{r-1}\binom{2m-1-k_2+l}{m-1} \\ &+ \Delta y^{r+m+1}\sum_{k_1=0}^{r}\sum_{k_2=0}^{m}\binom{r}{k_1}\binom{m}{k_2}\left(\frac{1-y}{y}\right)^{k_1+k_2}\binom{2r-k_1}{r-1}\binom{2m-k_2}{m-1}. \end{aligned} \tag{5.8}$$

The conclusion of this theorem then follows.

Proof of Theorem 2. By the classical spectral theory in Bai and Silverstein (2010), we know that

$$\frac{trS_n^r}{p} \xrightarrow{P} \omega_r, \quad \frac{trS_n^{2r}}{p} \xrightarrow{P} \omega_{2r}.$$

Then combining with the convergence (5.7), and using the Slutsky lemma, we obtain the conclusion that

$$p\left(\varphi_1 - (1 + y_n)\right) \xrightarrow{L} N(\mu_1, \sigma_1^2), \quad p\left(\varphi_2 - \left(1 + y_n + \frac{3y_n}{(1 + y_n)}\right)\right) \xrightarrow{L} N(\mu_2, \sigma_2^2).$$

where

$$\begin{aligned} &\mu_1 = y^2 + (\Delta + 1)y, \quad \sigma_1^2 = 4y^2, \\ &\mu_2 = -2y(\Delta - y - 1)\left(1 + \frac{3y}{(1 + y)^2}\right) + \frac{y}{(1 + y)^2}(-y^3 + 6y^2(\Delta - 1) + (16\Delta - 1)y + 2 + 6\Delta), \\ &\sigma_2^2 = \frac{1}{\omega_2^4}\xi_{33} + \frac{4\omega_4^2}{\omega_2^6}\xi_{22} - \frac{4\omega_4}{\omega_2^5}\xi_{23}. \end{aligned} \tag{5.9}$$

Proof of Theorem 4. According to Theorem 2 in Wang et al. (2014), when the mean vector $\mu = 0$, for the sample covariance matrices $\mathcal{S}_n = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T$, we know that under the alternative hypothesis H_1 given by (2.4), the random vector $(G_n(f_1), G_n(f_2), G_n(f_3))$ converges weakly to 3-dimensional Gaussian vector with $f_1 = x, f_2 = x^2, f_3 = x^4$. The centering terms of the random vector $(G_n(f_1), G_n(f_2), G_n(f_3))$ are

$$\int_{-\infty}^{+\infty} x^r dF^{y_n, H_n} = F^{y_n, H_n}(x^r) = \kappa_r + O\left(\frac{1}{n^2}\right). \tag{5.10}$$

Here

$$\begin{aligned}
\kappa_1 &= 1 + \frac{1}{p}\sum_{i=1}^{k} n_i\alpha_i - \frac{M}{p}, \quad \kappa_2 = \frac{2}{n}\sum_{i=1}^{k} n_i\alpha_i - \frac{2}{n}M + 1 + y_n - \frac{M}{p} + \frac{1}{p}\sum_{i=1}^{k} n_i\alpha_i^2, \\
\kappa_4 &= \frac{1}{p}\sum_{i=1}^{k} n_i\alpha_i^4 + \frac{4}{n}\sum_{i=1}^{k} n_i\alpha_i^3 + \left(\frac{4}{n} + \frac{6y_n}{n}\right)\sum_{i=1}^{k} n_i\alpha_i^2 + \left(\frac{4}{n} + \frac{12y_n}{n}\frac{4y_n^2}{n}\right)\frac{1}{p}\sum_{i=1}^{k} n_i\alpha_i \\
&\quad - \frac{3M}{n}(2 + 4y_n + y_n^2) + \left(1 - \frac{M}{p}\right)(1 + y_n)(1 + 5y_n + y_n^2).
\end{aligned} \tag{5.11}$$

From Pan (2014), we know that for S_n and $\mathcal{S}_n$, $(G_n(f_1), G_n(f_2), G_n(f_3))$ have different central limit theorems and the difference is reflected in mean of the limiting normal distribution. Therefore, we can quote centering term from Wang et al. (2014) in this theorem. Therefore, we conclude that under H_1, for $r = 1, 2$,

$$p\left(\varphi_r - \frac{\kappa_{2r}}{\kappa_r^2}\right) \xrightarrow{L} N(\mu_r, \sigma_r^2),$$

with μ_r, σ_r^2 defined in (5.9). ■

Proof of Theorem 6. For $r = 1, 2$, let

$$W_{n,r} = p\left(\varphi_r - \omega_{2r}/\omega_r^2\right).$$

Then, the power function of the test is $\beta_{n,r} = P(W_{n,r} \geq z_\alpha | H_1 \text{ holds})$. From the conclusions in Theorems 1 and 2, it follows that the asymptotic power function of the test is

$$\beta(\alpha) = 1 - \Phi\left(\Phi^{-1}(1-\alpha) - \frac{\kappa_{2r}/\kappa_r^2 - \omega_{2r}/\omega_r^2}{\sigma_r}\right).$$

With the definition of $\kappa_i, \omega_i, i = 1, 2, 4$, we can complete the proof. ■

Acknowledgments

This work was supported by the NSF of China Grant [No. 11771178], Science and Technology Development Program of Jilin Province, China [No. 20170101152JC].

References

Anderson, T.W., 2003. An Introduction to Multivariate Statistical Analysis, second ed. John Wiley Sons.
Bai, Z.D., Jiang, D.D., Yao, J.F., Zheng, S.R., 2009. Corrections to LRT on large-dimensional covariance matrix by RMT. Ann. Statist. 37, 3822–3840.
Bai, Z.D., Saranadasa, H., 1996. Effect of high dimension: by an example of a two sample problem. Statist. Sinica 6, 311–330.
Bai, Z.D., Silverstein, J.W., 2010. Spectral Analysis of Large Dimensional Random Matrices, second ed. Springer.
Birke, M., Dette, H., 2005. A note on testing the covariance matrix for large dimension. Statist. Probab. Lett. 74 (3), 281–289.
Chen, S.X., Zhang, L.X., Zhong, P.S., 2010. Tests for high- dimensional covariance matrice. J. Amer. Statist. Assoc. 105, 810–819.
Fisher, T., Sun, X., Gallagher, C., 2010. A new test for sphericity of the covariance matrix for high dimensional data. J. Multivariate Anal. 101 (10), 2554–2570.
Forzani, L., Gieco, A., Tolmasky, C., 2017. Likelihood ratio test for partial sphericity in high and ultra-high dimensions. J. Multivariate Anal. 159, 18–38.
Ledoit, O., Wolf, M., 2002. Some hypothesis tests for the covari- ance matrix when the dimension is large compared to the sample size. Ann. Statist. 30, 1081–1102.
Onatski, A., Moreira, M.J., Hallin, M., 2013. Asymptotic power of sphericity tests for high-dimensional data. Ann. Statist. 41 (3), 1204–1231.
Pan, G.M., 2014. Comparision between two types of large sample covari- ance matrices. Ann. Inst. Henri Poincaré Probab. Stat. 50 (2), 655–677.
Pan, G.M., Zhou, W., 2008. Central limit theorem for signal-to-interference ratio of reduced rank linear receiver. Ann. Appl. Probab. 18 (3), 1232–1270.
Srivastava, M.S., 2005. Some tests concerning the covariance matrix in high dimensional data. J. Japan Statist. Soc. 35, 251–272.
Tian, X.T., Lu, Y.T., Li, W.M., 2015. A robust test for sphericity of high-dimensional covariance matrices. J. Multivariate Anal. 141, 217–227.
Wang, Q.W., Silverstein, J.W., Yao, J.F., 2014. A note on the CLT of the LSS for sample covariance matrix from a spiked population model. J. Multivariate Anal. 130, 194–207.
Wang, C., Yang, J., Miao, B.Q., Cao, L.B., 2013. Identity tests for high dimensional data using RMT. J. Multivariate Anal. 128, 128–137.
Wang, Q.W., Yao, J.F., 2013. On the sphericity test with large-dimensional observations. Electron. J. Stat. 7, 2164–2192.
Zou, C.L., Peng, L.H., Feng, L., Wang, Z.J., 2014. Multivariate sign-based high-dimensional tests for sphericity. Biometrika 101 (1), 229–236.