Journal of Multivariate Analysis 128 (2014) 108–119

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Hypothesis testing for high-dimensional covariance matrices[☆]

Weiming Li[a], Yingli Qin[b,*]

[a] Beijing University of Posts and Telecommunications, Beijing, China

[b] Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada

ARTICLE INFO

Article history:
Received 23 March 2013
Available online 31 March 2014

AMS subject classifications:
62H15
62H10

Keywords:
Covariance matrix
Empirical spectral distribution
High-dimensional
Hypothesis testing
Stieltjes transform

ABSTRACT

This paper discusses the problem of testing for high-dimensional covariance matrices. Tests for an identity matrix and for the equality of two covariance matrices are considered when the data dimension and the sample size are both large. Most importantly, the dimension can be much larger than the sample size. The proposed test statistics are built upon the Stieltjes transform of the spectral distribution of the sample covariance matrix. We prove that the proposed statistics are asymptotically chi-square distributed under the null hypotheses, and normally distributed under the alternative hypotheses. Simulation results show that for finite dimension and sample size the proposed tests outperform some existing methods in various cases.

1. Introduction

Modern statistical analysis often encounters high-dimensional data. For instance, in DNA microarray analysis, the data dimension p (e.g. the number of genes of interest) is typically in the thousands, while the sample size n (e.g. the number of biological samples) is normally in the dozens or at most a couple hundreds. The majority of multivariate statistical procedures are no longer valid since their asymptotic properties are built within a framework where p is fixed and n approaches infinity. Therefore, novel statistical procedures which can handle "large p, large n" or even "large p, small n" need to be developed.

In this paper, we are interested in testing for population covariance matrices when the data dimension p can be much larger than the sample size n. Two tests will be discussed: a one-sample test, testing the identity of a $p \times p$ covariance matrix Σ_p,

$$H_0 : \Sigma_p = I_p \quad vs.\ H_1 : \Sigma_p \neq I_p, \tag{1.1}$$

and a two-sample test for the equality of two $p \times p$ covariance matrices Σ_{1p} and Σ_{2p},

$$H_0 : \Sigma_{1p} = \Sigma_{2p} \quad vs.\ H_1 : \Sigma_{1p} \neq \Sigma_{2p}. \tag{1.2}$$

☆ Weiming Li's research was supported by the Fundamental Research Funds for the Central Universities, No. 2014RC0905. Yingli Qin's research was supported by the University of Waterloo's start-up grant No. 203123.

* Corresponding author.
E-mail addresses: liwm601@gmail.com (W. Li), yingli.qin@uwaterloo.ca (Y. Qin).

http://dx.doi.org/10.1016/j.jmva.2014.03.013

Classical tests based on the likelihood ratio [1] suffer poor performance when p is not negligible with respect to n, since the sample covariance matrix does not converge to its population counterpart in high-dimensional situations. An important improvement in [2] corrects these likelihood ratio tests to accommodate situations where both p and n can be large but $p < n$. However, this correction cannot be easily extended to situations where $p > n$ as the corrected statistics involve the logarithm of the determinant of the sample covariance matrix, which is singular when $p > n$.

There are a number of works in the literature addressing the testing problems in situations where $p > n$. Ledoit and Wolf [13] modified certain tests, originally proposed by John [11] and Nagao [17], to adapt situations where p and n increase at the same rate. The results were later extended in [5] to the case when p/n tends to infinity or zero. In [23], a likelihood ratio type test was put forward for the identity (and sphericity) test, where only the non-zero sample eigenvalues were included in the likelihood ratio statistic. Schott [20] proposed a test for the equality of several covariance matrices based on the sum of squared differences between elements of the sample covariance matrices. In [24,26,25,9], unbiased estimators of $\mathrm{tr}\,\Sigma^k/p$ $(k = 1, 2, 3, 4)$ were constructed using functions of the sample covariance matrices, from which some tests were developed. In [8,14,7], new estimators of $\mathrm{tr}\,\Sigma^k/p$ were introduced, and some tests were then investigated based on these estimators.

A common feature among these methods is that they put emphasis on the eigenvalues of sample or population covariance matrices. In particular, the differences in the first few moments of the eigenvalues between the null and the alternative hypotheses. Most recently, [6] presented a two sample test in sparse settings based on the maximum of standardized differences between the entries of the sample covariances. For the study of the asymptotic properties of relevant tests, one is referred to [24,8,14,18].

In this paper, we investigate the testing problems (1.1) and (1.2) from random matrix theory point of view. We focus on the empirical spectral distributions rather than the moments of population (or sample) eigenvalues. By looking at the empirical spectral distributions, we are able to utilize the complete distribution of all eigenvalues. One may see that the new tests can detect a small shift of the bulk eigenvalues. Moreover, the power of the tests increases as the dimension p increases.

The success of this strategy relies on the convergence theorem of empirical spectral distributions in random matrix theory. The *spectral distribution* (SD) G^A of an $m \times m$ Hermitian matrix (or symmetric in real case) A is the measure generated by the eigenvalues $\{\lambda_j\}$ of A,

$$G^A = \frac{1}{m}\sum_{j=1}^{m}\delta_{\lambda_j},$$

where δ_b denotes the Dirac point measure at b. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a sequence of i.i.d. zero-mean random vectors in $\mathbb{R}^p$ or $\mathbb{C}^p$, with a common population covariance matrix Σ_p. The sample covariance matrix takes the form of $S_n = \sum_{k=1}^{n}\mathbf{x}_k\mathbf{x}_k^*/n$, where $\mathbf{x}_k^*$ stands for the conjugate transpose of $\mathbf{x}_k$. We are interested in the limiting relationship between the following two SDs as both $p \to \infty$ and $n \to \infty$:

$$H_p := G^{\Sigma_p} \quad \text{and} \quad F_n := G^{S_n},$$

which are referred as *population spectral distribution* (PSD) and *empirical spectral distribution* (ESD), respectively.

Following the conventional assumption in random matrix theory, we assume that H_p weakly converges to a limiting distribution H, as $p \to \infty$. Then under some regularity conditions, as p, n both tend to infinity with $p/n \to c > 0$, the ESD F_n converges to a non-random distribution F which relates H via the Marčenko–Pastur (MP) equation through Stieltjes transform, see (2.1). Particularly, if $H = \delta_1$ then the distribution F is the MP law.

The properties of the limiting distribution F offer us a new way to test for population covariances matrices. To test the hypothesis (1.1), we propose to measure the difference between the ESD F_n and the MP law; to test the hypothesis (1.2), we propose to measure the difference between the two ESDs.

The rest of this paper is organized as follows. In the next section, we discuss the test for the identity covariance matrix (1.1). The proposed test statistic is extended to testing for the equality of two covariance matrices (1.2) in Section 3. Section 4 reports simulation results. Conclusions and remarks are presented in Section 5, and proofs are postponed to Appendix.

2. Test for the identity of Σ_p

2.1. Main assumptions and the Marčenko–Pastur equation

Our model assumptions are as follows.

Assumption (a). Both $n, p \to \infty$ with $c_n = p/n \to c \in (0, \infty)$.

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

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

such that for each pair of (p, n), let $W = (w_{jk})_{1\leq j\leq p, 1\leq k\leq n}$, the observation vectors can be represented as $\mathbf{x}_k = \Sigma_p^{1/2} w_{.k}$ where $w_{.k} = (w_{jk})_{1\leq j\leq p}$ denotes the k-th column of W and $\Sigma_p^{1/2}$ stands for any Hermitian square root of Σ_p.

Assumption (c). The PSD H_p of Σ_p weakly converges to a probability distribution H on $[0, \infty)$ as $p \to \infty$. Moreover, the sequence of spectral norms $\{\|\Sigma_p\|\}$ is uniformly bounded for any p.

Assumption (a) defines our asymptotic regime which is commonly seen in high-dimensional data analysis. Assumption (b) states that the observation vector $\mathbf{x}_k$ can be expressed as a linear transformation of a vector whose elements are i.i.d. random variables with zero mean, unit variance, and finite fourth moment. This assumption is obviously satisfied for the normal distribution. Assumption (c) is natural in our asymptotic regime since the dimension p tends to infinity.

These assumptions are also required by the celebrated MP theorem, which states that the ESD F_n weakly converges to a determinate distribution F, called the *limiting spectral distribution* (LSD) [16,21,22,3]. More precisely, let us denote $\underline{s}_n(z)$ the Stieltjes transform of $c_nF_n + (1 - c_n)\delta_0$, i.e.

$$\underline{s}_n(z) = -\frac{1 - c_n}{z} + c_n \int \frac{1}{x - z} dF_n(x).$$

Then, under the above assumptions, for any $z \in \mathbb{C}^+ := \{z \in \mathbb{C} : \Im(z) > 0\}$, almost surely, $\underline{s}_n(z)$ converges to $\underline{s}(z)$, the Stieltjes transform of $cF + (1 - c)\delta_0$, which solves the following fundamental MP equation:

$$z = -\frac{1}{\underline{s}(z)} + c \int \frac{t}{1 + t\underline{s}(z)} dH(t), \quad z \in \mathbb{C}^+, \tag{2.1}$$

where $\Im(z)$ stands for the imaginary part of z. Recently, the convergence of $\underline{s}_n(z)$ and the MP equation were both extended to a subset of the real line [15].

We also introduce the finite dimensional version of the MP equation for fixed p and n,

$$z = -\frac{1}{\underline{s}_{c_n,H_p}(z)} + c_n \int \frac{t}{1 + t\underline{s}_{c_n,H_p}(z)} dH_p(t), \quad z \in \mathbb{C}^+, \tag{2.2}$$

where the Stieltjes transform $\underline{s}_{c_n,H_p}(z)$ is used as the centralization term when establishing the central limit theorem for $\underline{s}_n(z)$ in [4]. Motivated by this, we employ $\underline{s}_{c_n,H_p}(z)$ to construct test statistics and to investigate their asymptotic properties in the remaining of Sections 2 and 3. It is easy to see that $\underline{s}_{c_n,H_p}(z)$ converges to $\underline{s}(z)$ as $p, n \to \infty$.

2.2. Test statistic and its asymptotic distribution

Consider the testing problem (1.1), under the null hypothesis, we may assume that $H_p = H = \delta_1$, and hence the ESD F_n converges to the MP law. On the other hand, the Stieltjes transform solved from (2.2) is

$$\underline{s}_{c_n,\delta_1}(z) = \frac{c_n - 1 - z + \sqrt{(1 - c_n + z)^2 - 4z}}{2z}, \quad z \in \mathbb{C}^+, \tag{2.3}$$

where the square root is specified as the one with positive imaginary part. Therefore, we can measure the difference between the ESD F_n and the MP law by computing a distance between $\underline{s}_n(z)$ and $\underline{s}_{c_n,\delta_1}(z)$. Following this idea, we present our test statistic T_1 for testing the identity covariance matrix:

$$T_1 = \sum_{l=1}^{m} \|1/\underline{s}_n(z_l) - 1/\underline{s}_{c_n,\delta_1}(z_l)\|^2,$$

where a *testing-net* $\{z_1, \dots, z_m\} \subset \mathbb{C}^+$ is a set of mutually non-identical complex numbers with positive imaginary parts.

Remark 1. An important step in constructing the test statistic is to choose the testing-net. In Section 4, we propose a simple method to select $\{z_1, \dots, z_m\}$.

Remark 2. We define T_1 as a function $\|1/\underline{s}_n(z) - 1/\underline{s}_{c_n,\delta_1}(z)\|$ rather than the naive one $\|\underline{s}_n(z) - \underline{s}_{c_n,\delta_1}(z)\|$. This is because that the former represents a relative distance and thus is more stable than the latter in which such distances are summed.

Next, we present two functions which will be used to describe the limiting distribution of T_1. Define for any two complex numbers z and $\tilde{z}$:

$$m(z, f) := \frac{-c(f'(z))^2}{f^3(z)(1 + f(z))^3}, \tag{2.4}$$

$$v(z, \tilde{z}, f) := \begin{cases} \left(\dfrac{f'(z)f'(\tilde{z})}{(f(z) - f(\tilde{z}))^2} - \dfrac{1}{(z - \tilde{z})^2} \right) \dfrac{2}{f^2(z)f^2(\tilde{z})} & (z \neq \tilde{z}), \\ \dfrac{2f'(z)f'''(z) - 3(f''(z))^2}{6\,(f'(z))^2 f^4(z)} & (z = \tilde{z}). \end{cases} \tag{2.5}$$

Theorem 1. *Suppose that the assumptions* (a)–(c) *are fulfilled.*

(i) *If* w_{11} *in Assumption* (b) *is real and* $\mathrm{E}(w_{11}^4) = 3$, *then under the null hypothesis in* (1.1),

$$n^2 T_1 \xrightarrow{\mathcal{D}} \|W_1\|^2, \qquad W_1 \sim N_{2m}(\mu, \Omega), \tag{2.6}$$

where $N_{2m}(\mu, \Omega)$ *denotes a* $2m$ *dimensional multivariate normal distribution with mean vector* μ *and covariance matrix* Ω *defined as follows:*

$$\mu = [\Re(M), \Im(M)]',$$
$$\Omega = \frac{1}{2}\begin{pmatrix} \Re(\Gamma_1 + \Gamma_2) & \Im(\Gamma_1 - \Gamma_2) \\ \Im(\Gamma_1 + \Gamma_2) & \Re(\Gamma_2 - \Gamma_1) \end{pmatrix}, \tag{2.7}$$

where $M = [m(z_l, \underline{s})]_{l=1}^m$, $\Gamma_1 = [\upsilon(z_s, z_t, \underline{s})]_{s,t=1}^m$, $\Gamma_2 = [\upsilon(z_s, z_t^*, \underline{s})]_{s,t=1}^m$, *and* $\underline{s}$ *is the limit of* $\underline{s}_{c_n,\delta_1}(z)$ *defined in* (2.3). *The notations* $\Re(\cdot)$ *and* $\Im(\cdot)$ *respectively stand for the real and imaginary parts of a vector or a matrix.*

(ii) *If* w_{11} *is complex satisfying* $\mathrm{E}(w_{11}^2) = 0$ *and* $\mathrm{E}(|w_{11}|^4) = 2$, *then* (2.6) *also holds, except the mean vector is zero and the covariance is* $1/2$ *of that given in* (2.7).

(iii) *Under the alternative hypothesis, let* $e(H_p) = \sum_{l=1}^m \|1/\underline{s}_{c_n,H_p}(z_l) - 1/\underline{s}_{c_n,\delta_1}(z_l)\|^2$. *If the conditions in* (i) *or* (ii) *hold, then* $n[T_1 - e(H_p)]$ *converges in distribution to a normal variable, whose asymptotic mean and variance are given in the Appendix.*

To carry out the test procedure, we need to find quantiles of $\|W_1\|^2$ which is generalized chi-square distributed according to (2.6). To avoid this technical difficulty, one may employ a centralized test statistic $\widetilde{T}_1$,

$$\widetilde{T}_1 = \sum_{l=1}^m \|1/\underline{s}_n(z_l) - 1/\underline{s}_{c_n,\delta_1}(z_l) - \frac{1}{n} m(z_l, \underline{s})\|^2. \tag{2.8}$$

The conclusion (i) in Theorem 1 becomes

$$n^2 \widetilde{T}_1 \xrightarrow{\mathcal{D}} \|\widetilde{W}_1\|^2 \sim \sum_{i=1}^{2m} \sigma_i^2 \chi_i^2,$$

where $\sigma_1^2, \ldots, \sigma_{2m}^2$ are the eigenvalues of Ω, and $\chi_1^2, \ldots, \chi_{2m}^2$ are i.i.d $\chi^2(1)$. The matrix Ω is known if the limiting ratio c is known. In finite dimensional and finite sample situations, we replace c with c_n and plug $\underline{s}_{c_n,\delta_1}(z)$ into (2.7) to calculate Ω and its eigenvalues. Given these eigenvalues, any quantiles of $\sum_{i=1}^{2m} \sigma_i^2 \chi_i^2$ can be approximated by Monte Carlo simulations.

3. Tests for the equality of two population covariances

Let $\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i}$ be a sequence of i.i.d. zero-mean random vectors in $\mathbb{R}^p$ or $\mathbb{C}^p$, with population covariance matrix Σ_{ip}, for $i = 1, 2$ respectively. Denote by F_{in_i} the ESD of the sample covariance matrix $S_{in_i} = \sum_{j=1}^{n_i} \mathbf{x}_{ij}\mathbf{x}_{ij}^*/n_i$, and write $\underline{s}_{in_i}(z)$ for the Stieltjes transform of $(p/n_i)F_{in_i} + (1 - p/n_i)\delta_0$, $i = 1, 2$.

We list some assumptions for the two-sample test. These assumptions are parallel to the ones for the one-sample test in Section 2.

Assumption (a′). The dimensions $n_1, n_2, p \to \infty$ with $c_{in_i} = p/n_i \to c_i \in (0, \infty)$, $i = 1, 2$.

Assumption (b′). The Assumption (b) holds for each individual sample.

Assumption (c′). The Assumption (c) holds with the SD H_{ip} of Σ_{ip} weakly converges to H_i as $p \to \infty$, $i = 1, 2$.

3.1. Test when the two sample sizes are equal

Suppose that $n_1 = n_2 := n$ and let $c = \lim_{p,n\to\infty} p/n$. Under the null hypothesis in (1.2), we have $H_{1p} = H_{2p} \to H_1 = H_2 := H$, and thus the Stieltjes transforms $\underline{s}_{1n}(z)$ and $\underline{s}_{2n}(z)$ are independent and identically distributed with the same limit $\underline{s}(z)$.

Motivated by the definition of T_1, we propose our statistic for testing the hypothesis in (1.2) when $n_1 = n_2$ as

$$T_2 = \sum_{l=1}^m \|1/\underline{s}_{1n}(z_l) - 1/\underline{s}_{2n}(z_l)\|^2,$$

where $\{z_1, \ldots, z_m\}$ is a testing-net in $\mathbb{C}^+$.

Theorem 2. *Suppose that the Assumptions* (a′)–(c′) *are fulfilled.*

(i) *If w_{11} in Assumption* (b) *is real and* $\mathrm{E}(w_{11}^4) = 3$, *then under the null hypothesis in* (1.2),

$$n^2 T_2 \to \|W_2\|^2, \quad W_2 \sim N_{2m}(0, \Psi), \tag{3.1}$$

where $N_{2m}(0, \Psi)$ *denotes a* $2m$ *dimensional multivariate normal distribution with zero mean and covariance matrix* Ψ *where*

$$\Psi = \begin{pmatrix} \Re(\Gamma_1 + \Gamma_2) & \Im(\Gamma_1 - \Gamma_2) \\ \Im(\Gamma_1 + \Gamma_2) & \Re(\Gamma_2 - \Gamma_1) \end{pmatrix}, \tag{3.2}$$

where $\Gamma_1 = [\upsilon(z_s, z_t, \underline{s})]_{s,t=1}^m$ *and* $\Gamma_2 = [\upsilon(z_s, z_t^*, \underline{s})]_{s,t=1}^m$ *with* $\underline{s} = \underline{s}(z)$, *the common limit of* $\underline{s}_{in}(z)$*'s.*

(ii) *If w_{11} is complex satisfying* $\mathrm{E}(w_{11}^2) = 0$ *and* $\mathrm{E}(|w_{11}|^4) = 2$, *then* (3.1) *also holds, except the covariance is* $1/2$ *of that given in* (3.2).

(iii) *Under the alternative hypothesis, let* $e(H_{1p}, H_{2p}) = \sum_{l=1}^m \|1/\underline{s}_{c_n,H_{1p}}(z_l) - 1/\underline{s}_{c_n,H_{2p}}(z_l)\|^2$. *If the conditions in* (i) *or* (ii) *hold, then* $n[T_2 - e(H_{1p}, H_{2p})]$ *converges in distribution to a normal variable whose asymptotic mean and variance are given in the Appendix.*

When carrying out the test procedure, we may use $\widehat{\Psi}_w$ to estimate the limiting covariance Ψ by replacing $\underline{s}(z)$ in Γ_1 and Γ_2 with

$$\underline{s}_w(z) = \frac{1}{2}\underline{s}_{1n}(z) + \frac{1}{2}\underline{s}_{2n}(z).$$

From the strong convergence of $\underline{s}_{in}(z)$, we immediately get $\underline{s}_w(z) \xrightarrow{a.s.} \underline{s}(z)$, for every $z \in \mathbb{C}^+$. Hence $\widehat{\Psi}_w \xrightarrow{a.s.} \Psi$, as $n \to \infty$. Similar to the discussion in Section 2, one may estimate any quantiles of $\|W_2\|^2$ from its approximated representation by linear combination of i.i.d. chi-square distributed variables.

3.2. Test when the two sample sizes are not equal

Let $\underline{s}_i(z)$ be the limit of the Stieltjes transform $\underline{s}_{in_i}(z)$, $i = 1, 2$. Without loss of generality, we suppose $n_1 < n_2$, and thus $c_1 = \lim p/n_1 > c_2 = \lim p/n_2$. The statistic T_2 is no longer suitable as $\underline{s}_1(z) \neq \underline{s}_2(z)$ for every $z \in \mathbb{C}^+$ under the null hypothesis. In fact, these two limiting Stieltjes transforms are related in the following way.

Proposition 1. *In addition to the Assumptions* (a′)–(c′), *suppose that* $c_1 > c_2$, *then*

$$H_1 = H_2 \Leftrightarrow \underline{s}_1(z_1) = \underline{s}_2(z_2), \quad \forall z_1 \in \mathbb{C}^+,$$

where

$$z_2 = \frac{c_2}{c_1} z_1 + \left(\frac{c_2}{c_1} - 1\right) \frac{1}{\underline{s}_1(z_1)}.$$

Therefore, we present our statistic for testing the null hypothesis in (1.2) when $n_1 < n_2$ as

$$T_3 = \sum_{l=1}^m \|1/\underline{s}_{1n_1}(z_{1l}) - 1/\underline{s}_{2n_2}(\widehat{z}_{2l})\|^2,$$

where

$$\widehat{z}_{2l} = \frac{c_{2n_2}}{c_{1n_1}} z_{1l} + \left(\frac{c_{2n_2}}{c_{1n_1}} - 1\right) \frac{1}{\underline{s}_{1n_1}(z_{1l})},$$

and $\{z_{11}, \ldots, z_{1m}\}$ is a testing-net chosen from $\mathbb{C}^+$.

It is clear that the test statistic T_3 reduces to T_2 if the two sample sizes are equal (i.e. $c_{1n_1} = c_{2n_2}$). We will investigate the limiting distribution of T_3 in a future paper. When conducting the testing procedure, we employ the renown permutation test method evolved from the works of Fisher [10] and Pitman [19], which can provide satisfactory results as shown in our simulation section.

4. Simulation

We report on simulations carried out to compare the performance of the proposed tests with several congeneric tests in the literature. For the identity test, we consider four tests given respectively in [23] (referred as T_{S06}), [8] (referred as T_{C10}), [25] (referred as T_{S11}), and the test $\widetilde{T}_1$ defined in (2.8). The test in [23] is a modified version of the likelihood ratio test; the tests in [8] and [25] are based on different unbiased estimators of $\mathrm{tr}(\Sigma - I)^2$. For the equality test, our tests are compared with three tests based on different measures of the distance between two covariance matrices $\Sigma_1 = (\sigma_{ij1})$ and $\Sigma_2 = (\sigma_{ij2})$, the test in [14] (referred as T_{L12}) based on $\mathrm{tr}(\Sigma_1 - \Sigma_2)^2$, the test in [26] (referred as T_{S10}) based on

Table 1
Empirical sizes in percentiles of four identity tests with $\alpha = 0.05$.

n	$p = 100$				$p = 200$				$p = 500$			
	T_{C10}	T_{S06}	T_{S11}	$\widetilde{T}_1$	T_{C10}	T_{S06}	T_{S11}	$\widetilde{T}_1$	T_{C10}	T_{S06}	T_{S11}	$\widetilde{T}_1$
20	6.28	4.96	5.58	5.03	6.01	4.85	5.25	5.02	6.00	5.10	4.99	4.92
50	5.39	7.61	5.22	4.70	5.74	4.88	4.69	5.13	5.14	4.95	5.26	4.97
80	5.63	87.8	5.46	5.08	5.72	7.26	4.78	5.01	5.34	4.89	5.17	5.06

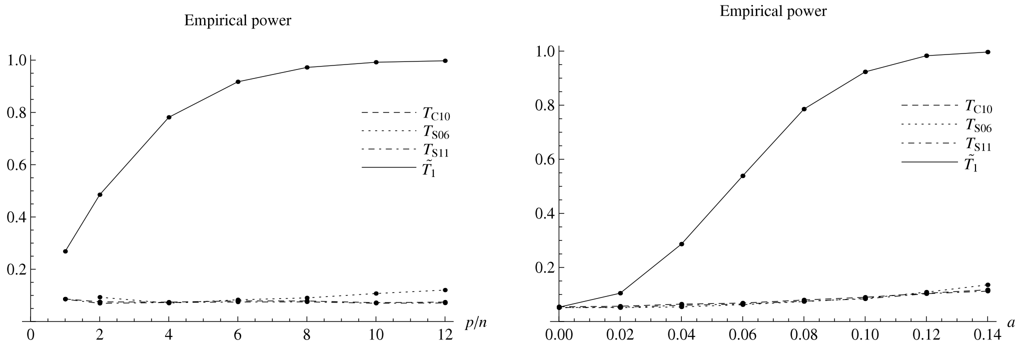

Fig. 1. Empirical powers of the identity tests for Model 1. The parameter settings are $(n, a) = (50, 0.08)$, $1 \le p/n \le 12$ in the left panel, and $(n, p) = (50, 200)$, $0 \le a \le 0.14$ in the right panel.

$\mathrm{tr}(\Sigma_1^2)/\mathrm{tr}^2(\Sigma_1) - \mathrm{tr}(\Sigma_2^2)/\mathrm{tr}^2(\Sigma_2)$, and the test in [6] (referred as T_{C13}) based on $\max_{1\le i\le j\le p}(\sigma_{ij1} - \sigma_{ij2})^2$. The nominal significant level for all the tests is set at $\alpha = 0.05$, and all statistics are computed from 10 000 independent replications.

For our proposed test procedures, we need to choose a testing-net from $\mathbb{C}^+$. The principle is to pick those points $\{z_l\}$ such that $\underline{s}(z_l)$ is sensitive to the changes of the LSD F. From the inverse formula of the Stieltjes transform, the density function of F (if exists) has the form

$$f(x) = \frac{1}{c} \lim_{\varepsilon\to 0^+} \Im \underline{s}(x + i\varepsilon), \quad x \in S_F \setminus \{0\},$$

where S_F stands for the support of F, we observe that the sensitive points are roughly those points whose real part belongs to S_F and imaginary part is not large. Notice that if the imaginary parts are extremely small, these points may introduce large errors to $\underline{s}_n(z)$ due to its unboundedness at each point of the sample eigenvalues. We then choose to construct a *testing-set* as follows.

$$\mathcal{Z} = \{u + vI : \lambda_{\min} \le u \le \lambda_{\max},\ v = u\},$$

where $\lambda_{\min}$ and $\lambda_{\max}$ are the smallest and largest nonzero sample eigenvalues, respectively. (Note that the proposed testing-set may not be optimal from the test efficiency point of view, but it is considerably efficient as attested by our preliminary numerical studies.) Then, we choose m equally spaced points from $\mathcal{Z}$ as our testing-net. We name this process as adaptive choice of the testing-net. It will be shown in the end of this section that the size m of the testing-net is not a key factor for test efficiency. Hence we simply set $m = 5$ for our tests in all simulations when comparing with other tests.

For the identity test $H_0 : \Sigma_p = I_p$, the samples $\{\mathbf{x}_i\}$ are generated from mean-zero real normal population. The α-quantile of $\|\widetilde{W}_1\|^2$ is approximated as the corresponding quantile of $\{\tilde{w}_j^2 := \sum_{i=1}^{2m} \sigma_i^2 u_{ij}^2,\ j = 1, \ldots, 100,000\}$ with (u_{ij}) i.i.d. normal. Table 1 illustrates the empirical sizes of the four tests with different combinations of sample sizes and dimensions. One can see that all the sizes satisfactorily match the nominal level except that of T_{S06} when p and n are close. This size distortion is due to the assumption of the test that one dimension is fixed and the other tends to infinity.

We study two models under the alternative hypothesis, which are listed below. The empirical powers are calculated with a fixed sample size $n = 50$. The power of T_{S06} for $p/n = 1$ is excluded because of its serious size distortion.

Model 1: $\Sigma_p = I_p + \mathrm{diag}(0, \ldots, 0, a, \ldots, a)$, where the number of as is $p/2$ (p even).
Model 2: $\Sigma_p = bI_p + \mathrm{diag}(u_1, \ldots, u_p)$, where $u_1, \ldots u_p$ are i.i.d. Unif$(-0.2, 0.2)$.

Fig. 1 illustrates the powers for Model 1. In the left panel, $a = 0.08$ is fixed and p/n increases from 1 to 12; while on the right, $p/n = 4$ is fixed and a increases from 0 to 0.14. The curves in the figure show that the proposed test has much higher power than the others, and is sensitive to the changes of p/n and a. The power of $\widetilde{T}_1$ rapidly increases to 1 as p/n increases, which convinces us that the increased dimensionality may become a blessing. The power curves of T_{C10} and T_{S11} are almost overlapping, which can be understood from the fact that they are both based on unbiased estimators of $\mathrm{tr}(\Sigma - I)^2$. Fig. 2 exhibits the powers for Model 2 with similar parameter settings. The only difference in the results is that shown in the right panel, no test performs particularly well when the factor b is close to 1. The crash of $\widetilde{T}_1$ at this point may be attribute to the

Table 2
Empirical sizes in percentiles of four equality tests with $\alpha = 0.05$ for Model 3 and Model 4.

Model 3	$p = 100$				$p = 200$				$p = 500$			
(n_1, n_2)	T_{L12}	T_{S10}	T_{C13}	T_2	T_{L12}	T_{S10}	T_{C13}	T_2	T_{L12}	T_{S10}	T_{C13}	T_2
(20, 20)	4.05	3.07	10.54	4.45	4.45	0.95	11.89	4.32	3.70	0.04	16.17	4.42
(50, 50)	4.72	4.33	5.08	5.03	4.28	3.47	4.94	4.83	4.60	1.68	5.51	4.73
Model 4	$p = 100$				$p = 200$				$p = 500$			
(n_1, n_2)	T_{L12}	T_{S10}	T_{C13}	T_3	T_{L12}	T_{S10}	T_{C13}	T_3	T_{L12}	T_{S10}	T_{C13}	T_3
(20, 40)	4.98	4.93	5.34	5.29	4.76	4.97	4.82	5.17	4.84	4.89	4.73	5.06
(60, 90)	4.87	5.20	5.26	4.98	5.19	5.13	5.21	5.14	4.73	5.15	5.13	4.99

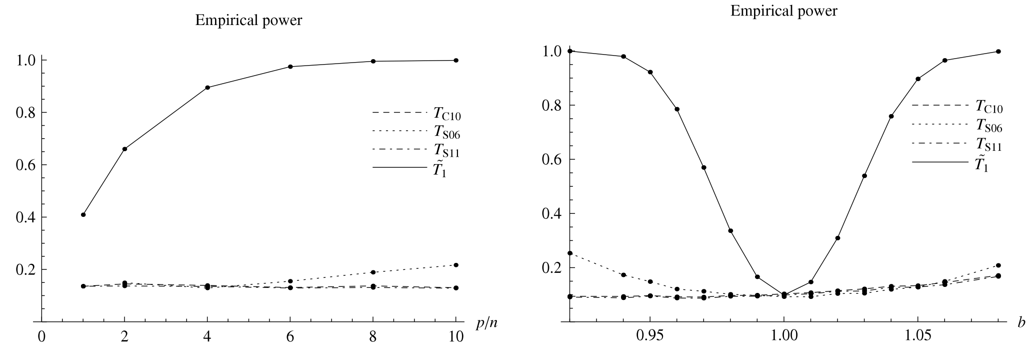

Fig. 2. Empirical powers of the identity tests for Model 2. The parameter settings are $(n, b) = (50, 1.05)$, $1 \le p/n \le 10$ in the left panel, and $(n, p) = (50, 200)$, $0.92 \le b \le 1.08$ in the right panel.

symmetry (around 1) of the PSD under the alternative hypothesis, which offsets the difference between the ESD and the MP law. However, the empirical power of $\widetilde{T}_1$ increases rapidly to 1 when b runs away from 1.

For the equality test $H_0 : \Sigma_{1p} = \Sigma_{2p}$, we study two models listed below, one is from [14] and the other is from [26]. Some parameters in the models are re-adjusted for better comparison.

Model 3: Let (z_{ijk}) be i.i.d. real standard normal. Define a moving average model $x_{ijk} = z_{ijk} + (1/4)z_{ijk+1} + az_{ijk+2}$, $i = 1, 2,\ j = 1, \dots, n,\ k = 1, \dots, p$. The two samples are independently generated from this model with the first taking $a = 0$.

Model 4: Let $O = \mathrm{diag}(\omega_1, \dots, \omega_p)$ where $\omega_1, \dots, \omega_p$ are i.i.d. Unif(1, 5), $\Delta_b = (\delta_{ijb})$ be a $p \times p$ matrix whose element are defined by $\delta_{ijb} = (-1)^{i+j}(0.4)^{|i-j|^{1/10}+b/10}$. The two samples are independently generated from $N(0, O\Delta_b O)$ with the first taking $b = 0$.

In the study of Model 3, we use the test T_2 with equal sample sizes. The quantile of $\|W_2\|^2$ is approximated in the same way as previous. In the study of Model 4, we use the test T_3 with unequal sample sizes. The permutation test procedure is employed across all the tests under Model 4 for the seek of fairness. Take T_3 as an example: With two samples $\mathbf{X}$ and $\mathbf{Y}$ of sizes n_1 and n_2, respectively, denote by $\mathbf{Z}$ their pooled sample. Randomly choose a sample $\widetilde{\mathbf{X}}$ of size n_1 from $\mathbf{Z}$ without replacement, denote the remains $\widetilde{\mathbf{Y}}$, and calculate T_3 based on $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$. Repeating the procedure r times yields r realizations of T_3. The ranking of original T_3 gives a p-value of the test. We take $r = 1000$ in our simulations.

One can see from Table 2 that, under Model 3, the empirical sizes of the tests T_2 and T_{L12} are close to the nominal level 0.05 in all cases, while the tests T_{S10} and T_{C13} (with small sample sizes) both have different levels of size distortion. This shows that the null distributions of the tests T_2 and T_{L12} can be well approximated by their asymptotic distributions. The empirical sizes of the tests under Model 4 are all close to the nominal one.

Fig. 3 shows that the proposed test is more powerful than the other tests in all the settings. It can be seen in the left panel that, as p grows, the power curves of T_{L12} and T_{C13} remain flat, and the power of T_{S10} decreases. Similar phenomena are shown in Fig. 4. Overall, the new tests significantly outperform all the other three tests for the models under investigation.

At last, we present some results on the relationship between the size m of the testing-net and the powers of the proposed tests. In Fig. 5, empirical powers with respect to different m values are plotted for Models 1–4. The results show that the empirical powers have little fluctuations as m changes. Hence, the proposed tests are robust with respect to the adaptive choice of the testing-net.

5. Conclusions and remarks

This paper discusses the problems of testing hypotheses in (1.1) and (1.2) for high-dimensional data. Unlike most existing methods in the literature, we adopt the ESD of the sample covariance matrix in random matrix theory. By measuring the

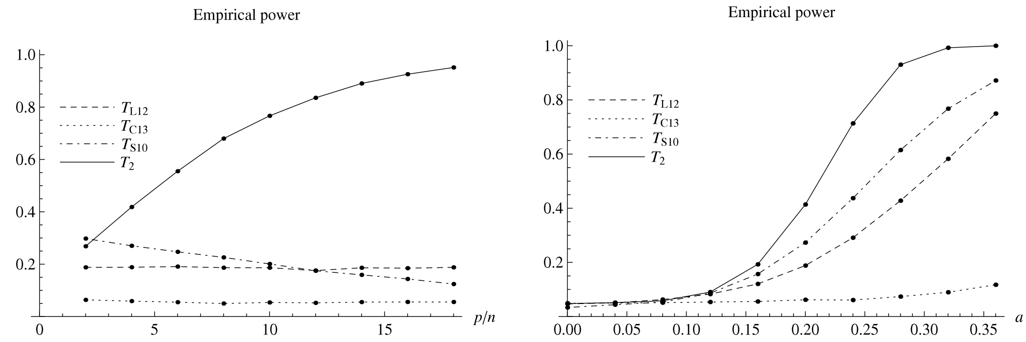

Fig. 3. Empirical powers of the equality tests for Model 3. The parameter settings are $(n, a) = (50, 0.2)$, $2 \le p/n \le 18$ in the left panel, and $(n, p) = (50, 200)$, $0 \le a \le 0.36$ in the right panel.

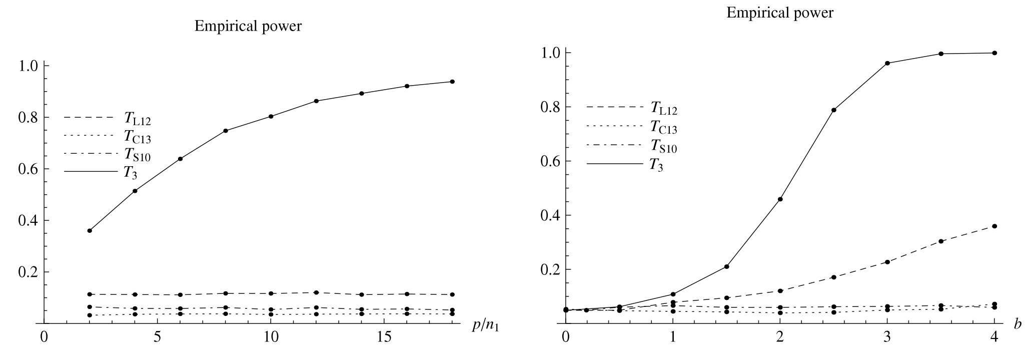

Fig. 4. Empirical powers of the equality tests for Model 4. The parameter settings are $(n_1, n_2, b) = (60, 90, 2)$, $2 \le p/n_1 \le 18$ in the left panel, and $(n_1, n_2, p) = (60, 90, 200)$, $0 \le b \le 4$ in the right panel.

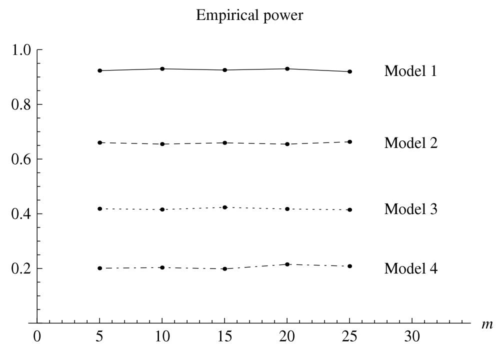

Fig. 5. Empirical powers of the proposed tests for Models 1–4 with respect to the size m of the testing-net (m ranging from 5 to 25). The parameter settings are $(n, p, a) = (50, 200, 0.1)$ for Model 1, $(n, p, b) = (50, 100, 1.05)$ for Model 2, $(n, p, a) = (50, 200, 0.2)$ for Model 3, and $(n_1, n_2, p, b) = (60, 90, 200, 1.5)$ for Model 4.

distance between the Stieltjes transform of the ESD F_n and that of the MP law, we propose the statistic T_1 to test for the identity covariance matrix in one-sample case. By measuring the distance of two Stieltjes transforms of ESDs, we propose T_2 and T_3 to test for the equality of two covariance matrices, where T_2 can be viewed as a special case of T_3 when the two sample sizes are equal. Asymptotic investigations reveal that T_1 and T_2 each has a generalized chi-square limiting distribution under null hypothesis, and a normal limiting distribution under alternative hypothesis.

Our tests are designed to detect the variation of the bulk of sample eigenvalues from the MP law in one-sample case and from the other set of sample eigenvalues in two-sample case. As our simulation studies demonstrate, when the bulk eigenvalues have some (even small) shift or asymmetric expansion/contraction, our tests are sensitive to these signals and thus outperform the competing tests; our tests lose some power but still perform as well as the competing tests when the expansion or contraction is symmetric. This is because the Stieltjes transform may attenuate such kind of signals to certain extent. In a spiked covariance model (see Johnstone [12]), our tests may compare unfavourably with some tests which focus on extreme sample eigenvalues.

Acknowledgments

The authors would like to thank the two referees for their helpful suggestions which have helped improve the presentation of this paper.

Appendix

A.1. Lemmas

Denote by $\underline{s}_n(z)$ the Stieltjes transform of $c_nF_n + (1-c_n)\delta_0$, $\underline{s}_{c_n,H_p}(z)$ the solution to Eq. (2.2) on $\mathbb{C}^+$, and $\underline{s}(z)$ the common limit of $\underline{s}_n(z)$ and $\underline{s}_{c_n,H_p}(z)$.

Lemma A.1. *Let $z_1, \ldots, z_m$ be m complex numbers on $\mathbb{C}^+$, and suppose that the Assumptions* (a)–(c) *are fulfilled. Then*

(i) *the random vector*

$$n\left[\underline{s}_n(z_1) - \underline{s}_{c_n,H_p}(z_1), \ldots, \underline{s}_n(z_m) - \underline{s}_{c_n,H_p}(z_m)\right] \tag{A.1}$$

forms a tight sequence in n.

(ii) *If w_{11} and Σ_p are real and $\mathrm{E}(w_{11}^4) = 3$, then (A.1) converges weakly to a normal vector $(V_1, \ldots, V_m)$, with means*

$$\mathrm{E}V_l = \frac{c\int \underline{s}^3(z_l)t^2(1+t\underline{s}(z_l))^{-3}dH(t)}{(1-c\int \underline{s}^2(z_l)t^2(1+t\underline{s}(z_l))^{-2}dH(t))^2}, \quad l = 1, \ldots, m, \tag{A.2}$$

and covariance function

$$\begin{aligned}\mathrm{Cov}(V_s, V_t) &\equiv \mathrm{E}(V_s - \mathrm{E}V_s)(V_t - \mathrm{E}V_t)\\ &= 2\left(\frac{\underline{s}'(z_s)\underline{s}'(z_t)}{(\underline{s}(z_s)-\underline{s}(z_t))^2} - \frac{1}{(z_s - z_t)^2}\right), \quad z_s \neq z_t.\end{aligned} \tag{A.3}$$

(iii) *If w_{11} is complex with $\mathrm{E}(w_{11}^2) = 0$ and $\mathrm{E}(|w_{11}|^4) = 2$, then* (ii) *also holds, except the means are zero and the covariance function is $1/2$ the function given in (A.3).*

Proof. See Lemma 1.1 in [4]. □

Lemma A.1 establishes the central limit theorem for the vector (A.1). However, the covariance function (A.3) excludes the case $z_s = z_t$, and thus cannot be applied to calculate the variance of V_l $(l = 1, \ldots, m)$ which is needed in our testing procedures.

Lemma A.2. *Under the assumptions of Lemma A.1. The mean and variance of V_l $(l = 1, \ldots, m)$ in Lemma A.1 are*

$$\mathrm{E}V_l = \frac{c(\underline{s}'(z_l))^2\int t^2(1+t\underline{s}(z_l))^{-3}dH(t)}{\underline{s}(z_l)},$$

$$\mathrm{Cov}(V_l, V_l) = \frac{2\underline{s}'(z_l)\underline{s}'''(z_l) - 3(\underline{s}''(z_l))^2}{6\left(\underline{s}'(z_l)\right)^2}.$$

Proof. The mean function can be derived from the Eq. (A.2) and the derivative of the MP equation. More precisely, taking the derivative of z on both sides of the MP equation (2.1), we get

$$\frac{\underline{s}^2(z)}{\underline{s}'(z)} = 1 - c\int \frac{t^2\underline{s}^2(z)}{(1+t\underline{s}(z))^2}dH(t), \quad z \in \mathbb{C}^+.$$

Substitute this to (A.2), we then get the formula.

For the variance of V_l, we first show that it is the limit of the covariance function $\mathrm{Cov}(V_l, V_t)$ as $z_t \to z_l$. Let $z_t = z_l + 1/t$, $f(x) = 1/(x - z_l)$ and $f_t(x) = 1/(x - z_t)$, from Theorem 1.1 in [4],

$$\mathrm{Var}(V_l) = -\frac{1}{2\pi^2}\oint_{C_1}\oint_{C_2} f(z)f(\tilde{z})\frac{\underline{s}'(z)\underline{s}'(\tilde{z})}{(\underline{s}(z)-\underline{s}(\tilde{z}))^2}dzd\tilde{z},$$

$$\mathrm{Cov}(V_l, V_t) = -\frac{1}{2\pi^2}\oint_{C_1}\oint_{C_2} f(z)f_t(\tilde{z})\frac{\underline{s}'(z)\underline{s}'(\tilde{z})}{(\underline{s}(z)-\underline{s}(\tilde{z}))^2}dzd\tilde{z} \tag{A.4}$$

where C_1 and C_2 are two positive oriented non-overlapping contours, each enclosing the support of the LSD F, and excluding the points z_l and $\{z_t : t \geq 1\}$. By the continuity of $\underline{s}'(z)$, the integrand in (A.4) is bounded on the contours. From this and Lebesgue dominated convergence theorem, as $t \to \infty$, the limit of $\mathrm{Cov}(V_l, V_t)$ exists and is equal to $\mathrm{Var}(V_l)$.

Therefore, from (A.3) we have

$$\begin{aligned}
\mathrm{Var}(V_l) &= \lim_{t\to\infty} \mathrm{Cov}(V_l, V_t) \\
&= 2\lim_{z\to z_l}\left(\frac{\underline{s}'(z)\underline{s}'(z_l)}{(\underline{s}(z)-\underline{s}(z_l))^2} - \frac{1}{(z-z_l)^2}\right) \\
&= 2\lim_{z\to z_l}\frac{(z-z_l)^2\underline{s}'(z)\underline{s}'(z_l) - (\underline{s}(z)-\underline{s}(z_l))^2}{(\underline{s}(z)-\underline{s}(z_l))^2(z-z_l)^2} \\
&= 2\lim_{z\to z_l}\frac{\underline{s}'(z_l)(\underline{s}'(z)-\underline{s}'(z_l)) - \underline{s}'(z_l)\underline{s}''(z_l)(z-z_l)}{(\underline{s}(z)-\underline{s}(z_l))^2} - \frac{3(\underline{s}''(z_l))^2 + 4\underline{s}'(z_l)\underline{s}'''(z_l)}{12(\underline{s}'(z_l))^2} \\
&= \frac{2\underline{s}'(z_l)\underline{s}'''(z_l) - 3(\underline{s}''(z_l))^2}{6(\underline{s}'(z_l))^2},
\end{aligned}$$

where the last two equations are respectively obtained from

$$\begin{aligned}
&\underline{s}(z)-\underline{s}(z_l) = \underline{s}'(z_l)(z-z_l) + \frac{1}{2}\underline{s}''(z_l)(z-z_l)^2 + \frac{1}{6}\underline{s}'''(z_l)(z-z_l)^3 + o(z-z_l)^3, \\
&\underline{s}'(z)-\underline{s}'(z_l) = \underline{s}''(z_l)(z-z_l) + \frac{1}{2}\underline{s}'''(z_l)(z-z_l)^2 + o(z-z_l)^2. \quad \square
\end{aligned}$$

Lemma A.3. *Suppose the Assumptions* (a)–(c) *hold. Let $z_1, \ldots, z_m$ be m complex numbers on $\mathbb{C}^+$, define $U_n = n[1/\underline{s}_n(z_1) - 1/\underline{s}_{c_n,H_p}(z_1), \ldots, 1/\underline{s}_n(z_m) - 1/\underline{s}_{c_n,H_p}(z_m)]$.*

(i) *If w_{11} and Σ_p are real and $\mathrm{E}(w_{11}^4) = 3$, then the random vector $W_n = [\Re(U_n), \Im(U_n)]'$ converges to a $2m$ dimensional normal vector $W_{c,H}$. The asymptotic mean is*

$$\mu_{c,H} = [\Re(m_H(z_1,\underline{s})), \ldots, \Re(m_H(z_m,\underline{s})), \Im(m_H(z_1,\underline{s})), \ldots, \Im(m_H(z_m,\underline{s}))]' \tag{A.5}$$

with $m_H(z,\underline{s}) = -c(\underline{s}'(z))^2 \int t^2[(1+t\underline{s}(z))\underline{s}(z)]^{-3}dH(t)$, and the asymptotic covariance matrix is

$$\Sigma_{c,H} = \frac{1}{2}\begin{pmatrix} \Re(\Gamma_1+\Gamma_2) & \Im(\Gamma_1-\Gamma_2) \\ \Im(\Gamma_1+\Gamma_2) & \Re(\Gamma_2-\Gamma_1)\end{pmatrix}, \tag{A.6}$$

where $\Gamma_1 = [v(z_s, z_t, \underline{s}(z))]_{s,t=1}^m$ and $\Gamma_2 = [v(z_s, z_t^*, \underline{s}(z))]_{s,t=1}^m$ with $v(z,\tilde{z},f)$ defined in (2.5).

(ii) *If w_{11} is complex with $\mathrm{E}(w_{11}^2) = 0$ and $\mathrm{E}(|w_{11}|^4) = 2$, then* (i) *also holds, except the mean vector is zero and the covariance matrix is $\Sigma_{c,H}/2$.*

Proof. From Lemma A.1 and a standard application of the Delta method, we have

$$\begin{aligned}
U_n &\xrightarrow{\mathcal{D}} -[V_1/\underline{s}^2(z_1), \ldots, V_m/\underline{s}^2(z_m)] \\
&:= U = (U_1, \ldots, U_m).
\end{aligned}$$

For the real case, U is complex normal with $\mathrm{E}(U_l) = -\mathrm{E}(V_l)/\underline{s}^2(z_l) = m_H(z_l,\underline{s})$ and $\mathrm{Cov}(U_s, U_t) = \mathrm{Cov}(V_s, V_t)/[\underline{s}(z_s)\underline{s}(z_t)]^2 = v(z_s, z_t, \underline{s})$. Vectorizing the real and imaginary parts of U, then $W_{c,H} = [\Re(U), \Im(U)]'$ is a $2m$ dimensional normal vector with the mean vector (A.5) and the covariance matrix

$$\begin{aligned}
\Sigma_{c,H} &= \begin{pmatrix} \mathrm{E}[(\Re(U)-\mathrm{E}\Re(U))'(\Re(U)-\mathrm{E}\Re(U))] & \mathrm{E}[(\Re(U)-\mathrm{E}\Re(U))'(\Im(U)-\mathrm{E}\Im(U))] \\ \mathrm{E}[(\Im(U)-\mathrm{E}\Im(U))'(\Re(U)-\mathrm{E}\Re(U))] & \mathrm{E}[(\Im(U)-\mathrm{E}\Im(U))'(\Im(U)-\mathrm{E}\Im(U))]\end{pmatrix} \\
&= \frac{1}{2}\begin{pmatrix} \Re(\Gamma_1+\Gamma_2) & \Im(\Gamma_1-\Gamma_2) \\ \Im(\Gamma_1+\Gamma_2) & \Re(\Gamma_2-\Gamma_1)\end{pmatrix}.
\end{aligned}$$

For the complex case, U is also complex normal with mean zero and covariance $\Sigma_{c,H}/2$. $\square$

A.2. Proofs of Theorems 1 and 2, and Proposition 1

Proof of Theorem 1. We only consider the real case. The complex case follows from the real case and the second conclusion of Lemma A.3. Under the null hypothesis, we assume $H = H_p = \delta_1$. From Lemma A.3,

$$n^2 T_1 = n^2 \sum_{l=1}^m \|1/\underline{s}_n(z_l) - 1/\underline{s}_{c_n,\delta_1}(z_l)\|^2 = U_n U_n^* \xrightarrow{\mathcal{D}} \|W_{c,\delta_1}\|^2,$$

where W_{c,δ_1} (denoted as W_1 in the theorem) is normal with the mean vector and covariance matrix given by (A.5) and (A.6), respectively. Moreover, the mean function $m_H(z, f)$ can be simplified from $H = \delta_1$ as in (2.4).

Under the alternative hypothesis, let $E_n = [1/\underline{s}_n(z_l) + 1/\underline{s}_{c_n,H_p}(z_l) - 2/\underline{s}_{c_n,\delta_1}(z_l)]_{l=1}^m$. Almost surely, E_n converges to $E = 2[1/\underline{s}_{c,H}(z_l) - 1/\underline{s}_{c,\delta_1}(z_l)]_{l=1}^m$. Write $\alpha = [\Re(E), \Im(E)]'$, then

$$\begin{aligned} n[T_1 - e(H_p)] &= n\sum_{l=1}^m \left(\|1/\underline{s}_n(z_l) - 1/\underline{s}_{c_n,\delta_1}(z_l)\|^2 - \|1/\underline{s}_{c_n,H_p}(z_l) - 1/\underline{s}_{c_n,\delta_1}(z_l)\|^2 \right) \\ &= [\Re(E_n), \Im(E_n)][\Re(U_n), \Im(U_n)]' \\ &\xrightarrow{\mathcal{D}} \alpha' W_{c,H} \sim N(\alpha'\mu_{c,H}, \alpha'\Sigma_{c,H}\alpha). \end{aligned}$$

Proof of Theorem 2. Similar to the proof of Theorem 1, we only consider the real case. Under the Assumptions (a′)–(c′), the conditions of Lemma A.3 hold for both the two samples. Write $U_{in} = n[1/\underline{s}_{in}(z_1) - 1/\underline{s}_{c_n,H_{ip}}(z_1), \ldots, 1/\underline{s}_{in}(z_m) - 1/\underline{s}_{c_n,H_{ip}}(z_m)]$. Then $W_{in} = [\Re(U_{in}), \Im(U_{in})]' \xrightarrow{\mathcal{D}} W_{c,H_i}$, $i = 1, 2$, where W_{c,H_1} and W_{c,H_2} are independent normal vectors with mean and covariance functions given in (A.5) and (A.6), respectively.

Under the null hypothesis, we assume $H_{1p} = H_{2p} \to H$ and thus W_{c,H_1} and W_{c,H_2} are identically distributed. We get

$$n^2 T_2 = \sum_{l=1}^m \|1/\underline{s}_{1n}(z_l) - 1/\underline{s}_{2n}(z_l)\|^2 = (U_{1n} - U_{2n})(U_{1n} - U_{2n})^* \xrightarrow{\mathcal{D}} \|W_{c,H_1} - W_{c,H_2}\|^2,$$

where $W_{c,H_1} - W_{c,H_2}$ (denoted by W_2 in the theorem) is a $2m$-dimensional normal vector with mean zero and covariance matrix

$$\Psi = \begin{pmatrix} \Re(\Gamma_1 + \Gamma_2) & \Im(\Gamma_1 - \Gamma_2) \\ \Im(\Gamma_1 + \Gamma_2) & \Re(\Gamma_2 - \Gamma_1) \end{pmatrix},$$

where $\Gamma_1 = [v(z_s, z_t, \underline{s})]_{s,t=1}^m$, $\Gamma_2 = [v(z_s, z_t^*, \underline{s})]_{s,t=1}^m$, and $\underline{s} = \underline{s}(z)$ is the common limit of $\underline{s}_{1n}(z)$ and $\underline{s}_{2n}(z)$.

Under the alternative hypothesis, let $E_{in} = [1/\underline{s}_{in}(z_l) + 1/\underline{s}_{c_n,H_{ip}}(z_l)]_{l=1}^m$, $i = 1, 2$. Almost surely, E_{in} converges to $E_i = 2[1/\underline{s}_{c,H_i}(z_l)]_{l=1}^m$, $i = 1, 2$. Write $\alpha = [\Re(E_1 - E_2), \Im(E_1 - E_2)]'$. From Lemma A.3,

$$\begin{aligned} n[T_2 - e(H_{1p}, H_{2p})] &= \sum_{l=1}^m \left(\|1/\underline{s}_{1n}(z_l) - 1/\underline{s}_{2n}(z_l)\|^2 - \|1/\underline{s}_{c_n,H_{1p}}(z_l) - 1/\underline{s}_{c_n,H_{2p}}(z_l)\|^2 \right) \\ &= [\Re(E_{1n} - E_{2n}), \Im(E_{1n} - E_{2n})][\Re(U_{1n} - U_{2n}), \Im(U_{1n} - U_{2n})]' \\ &\xrightarrow{\mathcal{D}} \alpha'(W_{c,H_1} - W_{c,H_2}), \end{aligned}$$

which is a normal variable with mean $\alpha'(\mu_{c,H_1} - \mu_{c,H_2})$ and variance $\alpha'(\Sigma_{c,H_1} + \Sigma_{c,H_2})\alpha$.

Proof of Proposition 1. For any $z_1 \in \mathbb{C}^+$, define

$$z_2 = \frac{c_2}{c_1} z_1 + \left(\frac{c_2}{c_1} - 1 \right) \frac{1}{\underline{s}_1(z_1)}, \tag{A.7}$$

which is also a complex number in $\mathbb{C}^+$ since $\underline{s}_1(z_1) \in \mathbb{C}^+$ and $c_2 < c_1$.

Under the Assumptions (a′)–(c′), we have

$$z_1 = -\frac{1}{\underline{s}_1(z_1)} + c_1 \int \frac{t}{1 + t\underline{s}_1(z_1)} dH_1(t), \tag{A.8}$$

$$z_2 = -\frac{1}{\underline{s}_2(z_2)} + c_2 \int \frac{t}{1 + t\underline{s}_2(z_2)} dH_2(t). \tag{A.9}$$

Replace z_1 in (A.7) by the right hand side of (A.8), we get

$$z_2 = -\frac{1}{\underline{s}_1(z_1)} + c_2 \int \frac{t}{1 + t\underline{s}_1(z_1)} dH_1(t). \tag{A.10}$$

From the uniqueness of the solution to the MP equation on $\mathbb{C}^+$, and by the comparison of the equations between (A.9) and (A.10), we conclude that

$$H_1 = H_2 \Leftrightarrow \underline{s}_1(z_1) = \underline{s}_2(z_2), \quad \forall z_1 \in \mathbb{C}^+.$$

References

[1] T.W. Anderson, An Introduction to Multivariate Statistical Analysis, third ed., Wiley & Sons, Hoboken, NJ, 2003.
[2] Z.D. Bai, D.D. Jiang, J.F. Yao, S.R. Zheng, Corrections to LRT on largedimensional covariance matrix by RMT, Ann. Statist. 37 (2009) 3822–3840.
[3] Z.D. Bai, J.W. Silverstein, Spectral Analysis of Large Dimensional Random Matrices, second ed., Springer, New York, 2010.
[4] Z.D. Bai, J.W. Silverstein, CLT for linear spectral statistics of large-dimensional sample covariance matrices, Ann. Probab. 32 (2004) 553–605.
[5] M. Birke, H. Dette, A note on testing the covariance matrix for large dimension, Statist. Probab. Lett. 74 (2005) 281–289.
[6] T. Cai, W.D. Liu, Y. Xia, Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings, J. Amer. Statist. Assoc. 108 (2013) 265–277.
[7] T. Cai, Z. Ma, Optimal hypothesis testing for high dimensional covariance matrices, Bernoulli 19 (2013) 2153–2779.
[8] S.X. Chen, L.X. Zhang, P.S. Zhong, Testing for high dimensional covariance matrices, J. Amer. Statist. Assoc. 109 (2010) 810–819.
[9] T.J. Fisher, On testing for an identity covariance matrix when the dimensionality equals or exceeds the sample size, J. Statist. Plann. Inference 142 (2012) 312–326.
[10] R.A. Fisher, Design of Experiments, Oliver and Boyd, Edinburgh, 1935.
[11] S. John, Some optimal multivariate tests, Biometrika 58 (1971) 123–127.
[12] I. Johnstone, On the distribution of the largest principal component, Ann. Statist. 29 (2001) 295–327.
[13] O. Ledoit, M. Wolf, Some hypothesis tests for the covariance matrix when the dimension is large compare to the sample size, Ann. Statist. 30 (2002) 1081–1102.
[14] J. Li, S.X. Chen, Two sample tests for high dimensional covariance matrices, Ann. Statist. 40 (2012) 908–940.
[15] W. Li, J. Chen, Y. Qin, J. Yao, Z. Bai, Estimation of the population spectral distribution from a large dimensional sample covariance matrix, J. Statist. Plann. Inference 143 (2013) 1887–1897.
[16] V.A. Marčenko, L.A. Pastur, Distribution of eigenvalues in certain sets of random matrices, Mat. Sb. (N.S.) 72 (1967) 507–536.
[17] H. Nagao, On some test criteria for covariance matrix, Ann. Statist. 1 (1973) 700–709.
[18] A. Onatski, M.J. Moreira, M. Hallin, Asymptotic power of sphericity tests for high-dimensional data, Ann. Statist. 41 (2013) 1204–1231.
[19] E.J.G. Pitman, Significance tests which may be applied to samples from any populations. Part III. The analysis of variance test, Biometrika 29 (1938) 322–335.
[20] J.R. Schott, A test for the equality of covariance matrices when the dimension is large relative to the sample size, Comput. Statist. Data Anal. 51 (2007) 6535–6542.
[21] J.W. Silverstein, Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices, J. Multivariate Anal. 55 (1995) 331–339.
[22] J.W. Silverstein, Z.D. Bai, On the empirical distribution of eigenvalues of a class of large-dimensional random matrices, J. Multivariate Anal. 54 (1995) 175–192.
[23] M.S. Srivastava, Some tests criteria for the covariance matrix with fewer observations than the dimension, Acta Comment. Univ. Tartu. Math. 10 (2006) 77–93.
[24] M.S. Srivastava, Some tests concerning the covariance matrix in high dimensional data, J. Japan Statist. Soc. 35 (2005) 251–272.
[25] M.S. Srivastava, T. Kollo, D. von Rosen, Some tests for the covariance matrix with fewer observations than the dimension under non-normality, J. Multivariate Anal. 102 (2011) 1090–1103.
[26] M.S. Srivastava, H. Yanagihara, Testing the equality of several covariance matrices with fewer observations than the dimension, J. Multivariate Anal. 101 (2010) 1319–1329.