Stockholms universitet

A shrinkage test for large-dimensional covariance matrix

Martin Nilsson

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2021:17
Matematisk statistik
September 2021

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematiska institutionen

Mathematical Statistics
Stockholm University
Master Thesis **2021:17**
http://www.math.su.se

A shrinkage test for large-dimensional covariance matrix

Martin Nilsson*

September 2021

Abstract

In this thesis, we use an optimal linear shrinkage estimator for the covariance matrix along with modern results on linear spectral statistics to establish a new test for sphericity under the large-dimensional asymptotics, namely when both the number of variables p and the sample size n tend to infinity such that $p/n \to c > 0$. Using similar techniques, we also show that a previously established test based on the Cauchy-Schwarz inequality remains usable under weaker assumptions than originally stated. We perform a Monte Carlo simulation study to verify our results, to assess the quality of our new test, and to see how well it performs compared to other tests.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: martin.nilsson@math.su.se. Supervisor: Taras Bodnar.

Acknowledgements

I am very grateful to my supervisor Taras Bodnar, professor at the Stockholm University Department of Mathematics (incl. Math. Statistics), for recommending to me the topic of random matrix theory and for his helpful suggestions and feedback during the writing process. I would also like to thank my family for their patience and support.

Contents

1 Introduction

With advances in data collection and storage, modern multivariate statistics must increasingly often handle situations where the number of variables p is comparable to the sample size n or possibly even exceeds it, as is the case e.g. for DNA microarray data. This entails a large $p \times p$ covariance matrix Σ along with a large inverse that in many applications nevertheless must be accurately estimated, as when using the Markowitz mean-variance approach to select an efficient portfolio from a large number of stocks (see [17]). There is also a need for statistical hypothesis tests that perform well in this large-dimensional setting, e.g. tests for *sphericity* of the covariance matrix; that is, when $\Sigma = \sigma^2 I_p$ is a scalar matrix for some unknown and unspecified constant $\sigma^2 > 0$.

In statistics at large, estimators and tests are often derived and have their performance appraised under the assumption that the sample size n tends to infinity, since this opens up for asymptotic techniques (see e.g. [10]). While the usual estimator for Σ, the sample covariance matrix S_n, is unbiased and consistent under the "standard asymptotics" where p is fixed and $n \to \infty$, it is known to behave very differently under the "large-dimensional asymptotics" (a.k.a. "general asymptotics" or "Kolmogorov asymptotics" or "Marchenko-Pastur scheme") where $p = p(n)$ depends on n and $p/n \to c \in (0, +\infty)$ as $n \to \infty$. Consequently, in the large-dimensional regime many classical tests such as Hotelling's T^2 test lose power (see [3]) and portfolio selection strategies that depends on "plug-in" estimators may underperform when these depend on S_n or its inverse (see [7]). Moreover, S_n is singular and some tests degenerate when $p > n$, e.g. the original likelihood ratio test (LRT) for sphericity as originally established by Mauchly [19].

On the one hand, efforts have been made to develop new estimators and tests under the large-dimensional asymptotics. In regard to testing for sphericity, these efforts include the tests of Srivastava [23] and Fisher *et al.* [9], both of which are based on the Cauchy-Schwarz inequality and derived under the assumptions of Gaussian data and convergence of $\operatorname{tr}(\Sigma^i)/p$ as $n \to \infty$ up to order $i = 8$ and $i = 16$. In search of a new well-conditioned estimator for Σ, Ledoit and Wolf [11] studied linear shrinkage estimators—in this context defined as linear combinations of S_n and I_p with coefficients estimated from data—and found one that is optimal in the sense that it asymptotically minimizes a Frobenius loss in quadratic mean; most importantly, this optimality is retained under the large-dimensional asymptotics.

To arrive at this result, Ledoit and Wolf [11] had to assume the existence of the eighth moment of the data.

On the other hand, efforts have been made to introduce large-dimensional "corrections" to already existing estimators and tests. Research in this direction includes the paper by Bai *et al.* [1] who modified the sphericity LRT to account for $p/n \to c \in (0,1)$ while simultaneously dropping the assumption of normality (later amendments by Wang and Yao [24] removed a restrictive assumption on the fourth moment of the data-generating model). Notably, the approach used by these authors differs from those of the previous paragraph in that they make heavy use of recent advances in random matrix theory (RMT).

RMT rose to prominence with a series of papers in the 1950's by famous physicist E. Wigner in which he showed that certain classes of symmetric matrices with random elements have, in the limit of increasing matrix size, a non-random distribution of eigenvalues (see e.g. [25]). Later, in 1967, a seminal paper by Marčenko and Pastur [18] established an integral equation for the Stieltjes transformation of eigenvalue distributions arising in this manner; importantly for statistical applications, the class of matrices under consideration is wide enough to include sample covariance matrices. This result was investigated further by several authors and subsequently re-established under very general conditions in a 1995 paper by Silverstein [22].

Recently, Bodnar *et al.* [8] extended the optimal linear shrinkage estimator (OLSE) of Ledoit and Wolf [11] to the case when the shrinkage target is an arbitrary covariance matrix. Unlike Ledoit and Wolf [11], Bodnar *et al.* [8] used the RMT to (1) get by with laxer assumptions on the moments, and (2) to show that the new estimator minimizes the Frobenius loss, not just in quadratic mean, but almost surely. The new OLSE is written $\widehat{\Sigma}_{\text{OLSE}} = \hat{\alpha}^* S_n + \hat{\beta}^* \Sigma_0$ where Σ_0 is the shrinkage target and $\hat{\alpha}^*$, $\hat{\beta}^*$ are *bona fide* estimators of the optimal coefficients. As we will later see, if one chooses $\Sigma_0 = I_p/p$ then $\hat{\alpha}^*$, $\hat{\beta}^*$ will be functions of $\text{tr}(S_n)/p$, $\text{tr}(S_n^2)/p$ where the latter pair of statistics can be regarded as the sample equivalents of $\text{tr}(\Sigma)/p$, $\text{tr}(\Sigma^2)/p$.

The aforementioned statistics have the useful property of being expressible in terms of the eigenvalues of S_n; indeed and more generally, $\text{tr}(S_n^k)/p = p^{-1}\sum_{j=1}^p \ell_j^k$ where $\ell_1, \ldots, \ell_p$ are the eigenvalues. This makes them special cases of statistics of the form $p^{-1}\sum_{j=1}^p f(\ell_j)$ where f is a suitable function, a.k.a. linear spectral statistics (LSS), a term coined in a 2004 paper by Bai

and Silverstein [5]. In the same paper the authors prove that under the large-dimensional asymptotics and some general conditions, a vector of LSSs will converge in distribution to a multivariate Gaussian whose mean and covariance functions are given by contour integrals of expressions containing the Stieltjes transforms of certain limiting eigenvalue distributions. Although this central limit theorem (CLT) does not assume Gaussian data, the moments still need to match those of a standard Gaussian up to order four, rendering it very difficult to use anything but a Gaussian. This restriction on the fourth moment was later removed by Pan and Zhou [20]. Moreover, Wang and Yao [24] recently derived new, more explicit formulas for the mean and covariance functions.

The last few years has seen the RMT taking on an increasingly important role in the mathematical toolbox of the multivariate and/or large-dimensional statistician, and the topic is very popular. Recent applications include: a new method for deriving covariance matrix estimators that are optimal under Stein's loss [12], a CLT for LSSs of so-called *separable* sample covariance matrices [2], new tests for the independence of two large-dimensional vectors [6], a bootstrap procedure for LSSs [16], a family of rotation-invariant tests for general linear hypotheses in a large-dimensional multivariate linear regression model [14], an analytical (i.e. non-numerical) nonlinear shrinkage estimator for large-dimensional covariance matrices [13], and a CLT for the joint distribution of the m largest eigenvalues and trace of large-dimensional covariance matrices in a generalized spiked population model [15].

The rest of this thesis is organized as follows. Section 2 presents the necessary background material from the RMT including the important central limit theorem for linear spectral statistics. In Section 3, we use the RMT to prove our main results: (1) a new test for sphericity based on the OLSE of Bodnar *et al.* [8], and (2) a re-establishment of the test of Fisher *et al.* [9] under more general assumptions. In Section 4, we perform a Monte Carlo simulation study to validate our theoretical findings and to investigate the performance of our new test. Section 5 discusses these results and summarizes the thesis. All proofs and calculations are relegated to the Appendix.

2 Preliminaries

2.1 Notation and assumptions

The following notation and assumptions will be in effect throughout the entire thesis. Let

- $X_{n;ij}$, $n \geq 1$, $i \geq 1$, $j \geq 1$ be identically distributed real random variables such that $X_{n;ij}$, $i \geq 1$, $j \geq 1$ are independent for each fixed n and $\mathbb{E}(X_{n;ij}) = 0$, $\mathbb{E}(X_{n;ij}^2) = 1$, $\mathbb{E}(X_{n;ij}^4) < \infty$;
- $\beta = \mathbb{E}(X_{1;11}^4) - 3$ be the excess kurtosis;
- $c \in (0, +\infty)$ be a nonrandom number called the *concentration*;
- $p = p(n)$ be positive integers such that $c_n := p/n \to c$ as $n \to \infty$;
- X_n be the $p \times n$ matrix having $X_{n;ij}$, $1 \leq i \leq p$, $1 \leq j \leq n$ as elements;
- Σ_n be a nonrandom positive definite $p \times p$ matrix;
- $Y_n = \Sigma_n^{1/2} X_n$;
- $S_n = n^{-1} Y_n Y_n^T$;
- unless otherwise stated, all limits (in probability, distribution, etc.) shall be taken as $n \to \infty$.

These choices are motivated in part by the random matrix theory and in part by the intended statistical applications. For example, our assumptions on the $X_{n;ij}$ imply that the columns $X_{n;\cdot 1}, \ldots, X_{n;\cdot n}$ of X_n are i.i.d. with zero mean vector and identity covariance matrix, and so the columns $Y_{n;\cdot 1}, \ldots, Y_{n;\cdot n}$ of Y_n are i.i.d. with mean vector

$$\mathbb{E}(Y_{n;\cdot 1}) = \Sigma_n^{1/2} \mathbb{E}(X_{n;\cdot 1}) = \Sigma_n^{1/2} 0 = 0$$

and covariance matrix

$$\mathrm{Cov}(Y_{n;\cdot 1}) = \Sigma_n^{1/2} \mathrm{Cov}(X_{n;\cdot 1}) (\Sigma_n^{1/2})^T = \Sigma_n^{1/n} I_p \Sigma_n^{1/2} = \Sigma_n.$$

The intended interpretation is therefore that the columns of Y_n constitute a random sample from some multivariate distribution $D(0, \Sigma_n)$ and that S_n is

the associated sample covariance matrix. Hence, in applications we assume that Y_n is observed and S_n computed, while X_n and Σ_n are unknown. In contrast, the existence of the fourth moment is just a technical requirement for the CLT on linear spectral statistics to hold.

Remark. Although the sample covariance matrix associated to an i.i.d. Gaussian sample $x_1, \dots, x_N \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ is usually defined as

$$S = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T,$$

there always exists an orthogonal transformation of vectors that will transform $x_1, \dots, x_N$ into a new i.i.d. Gaussian sample $y_1, \dots, y_N \sim \mathcal{N}(0, \Sigma)$ such that $y_N = \sqrt{N}\bar{x}$ and

$$S = \frac{1}{N-1} \sum_{i=1}^{N-1} y_i y_i^T$$

(see Lemma 1.1 of [21]). Hence, by discarding y_N and letting $n = N - 1$ we see that in the case of Gaussian data there is no loss of generality in assuming that the covariance matrix can be computed using the formula $S_n = n^{-1} Y_n Y_n^T$. If instead one does not assume Gaussian data but rather only that $y_1, \dots, y_n$ is an i.i.d. sample from some multivariate distribution $D(\boldsymbol{\mu}, \Sigma_n)$ then this argument does not work. Fortunately, if one additionally knows that $\boldsymbol{\mu} = \mathbf{0}$ (as is the case for our "data" $Y_{n;\cdot 1}, \dots, Y_{n;\cdot n}$), then there is no need to estimate the mean and therefore no need to decrement n by one to account for what would otherwise be a smaller number of degrees of freedom. Thus, in our setting the formula

$$S = \frac{1}{n} \sum_{i=1}^{n} y_i y_i^T$$

is valid nonetheless, and $S_n = n^{-1} Y_n Y_n^T$ retains its interpretation as the sample covariance matrix.

Remark. One easy way to satisfy the moment requirements is to let $X_{n;ij}$ be standard Gaussians, in which case $\beta = 0$. It is however very difficult to find a standardized non-Gaussian distribution that simultaneously satisfies $\beta = 0$.

We shall on occasion need an additional assumption on the rate of convergence to the concentration, namely

(A) $c_n = c + o(1/n)$ or equivalently $n(c_n - c) \to 0$ as $n \to \infty$.

This is a comparably tame and common assumption that will be used in conjunction with the delta method in some of the proofs.

2.2 Random matrix theory

In this section and this section only we allow the fourth moment of $X_{n;ij}$ to be infinite. Recall that the empirical spectral distribution (e.s.d.) of a real symmetric $p \times p$ matrix A with eigenvalues $\lambda_1, \ldots, \lambda_p$ is the c.d.f.

$$F^A(x) = \frac{1}{p} \sum_{i=1}^{p} \mathbb{1}(\lambda_i \leq x), \tag{1}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function. One of the central concerns of RMT is to understand the asymptotic behaviour of the (random) spectrum of S_n; this includes computing the limit of F^{S_n} if it exists, and for this limit to exist we need the limit of F^{Σ_n} to exist. Let for this purpose H be a nonrandom c.d.f. with support on $[0, +\infty)$. It is then known that for each $z \in \mathbb{C}^+ = \{z \in \mathbb{C} : \operatorname{Im} z > 0\}$ the equation

$$m = \int \frac{1}{\tau(1 - c - czm) - z} dH(\tau)$$

has a unique solution $m = m_{c,H}(z)$ in the set $\{m \in \mathbb{C} : -(1-c)/z + cm \in \mathbb{C}^+\}$ (see [22]) and there exists a unique c.d.f. $F^{c,H}$ such that $z \mapsto m_{c,H}(z)$ is the Stieltjes transform of $F^{c,H}$; that is,

$$m_{c,H}(z) = \int \frac{1}{\lambda - z} \, dF^{c,H}(\lambda) \quad \text{for all} \quad z \in \mathbb{C}^+.$$

The fundamental result on the asymptotic behaviour of the spectrum of S_n is

Theorem 2.1 ([22]). *Assume that* $F^{\Sigma_n} \xrightarrow{d} H$. *Then* $\mathbb{P}(F^{S_n} \xrightarrow{d} F^{c,H}) = 1$.

Remark. In its full generality, the result as stated in [22] holds under weaker assumptions and allows for complex-valued $X_{n;ij}$, nonzero mean, and random Σ_n, but we shall not need this stronger version.

Recall that the Marchenko-Pastur law with parameter $c > 0$ is the probability distribution F^c on $\mathbb{R}$ given by

$$dF^c(x) = \left(1 - \frac{1}{c}\right)_+ \mathbb{1}(x = 0) + \frac{1}{2\pi cx}\sqrt{(b-x)(x-a)}\mathbb{1}(a \leq x \leq b)\,dx$$

where $x_+ = \max\{0, x\}$, $a = (1-\sqrt{c})^2$ and $b = (1+\sqrt{c})^2$. (Note that F^c has an atom at $x = 0$ if and only if $c > 1$.) A special case of Theorem 2.1 is the celebrated Marchenko–Pastur theorem, which is one of the few cases where the limiting distribution $F^{c,H}$ has a known analytic expression:

Theorem 2.2 ([26])**.** *Assume that $\Sigma_n = I_p$ for all n. Then $\mathbb{P}(F^{S_n} \xrightarrow{d} F^c) = 1$.*

Remark. Since $\Sigma_n = I_p$ implies $H_n = \chi_{[1,+\infty)} = H$ where $\chi_{[1,+\infty)}$ is the indicator function (a.k.a. characteristic function) of the set $[1, +\infty)$, the Marchenko-Pastur theorem is more or less equivalent to the statement that $F^{c,\chi_{[1,+\infty)}} = F^c$.

For later convenience we here introduce some notation. Any probability measure F on $\mathbb{R}$ induces a functional $L^1(\mathbb{R}, F) \to \mathbb{R}$ by

$$f \mapsto F(f) = \int f(x)\,dF.$$

This definition is equivalent to $F(f) = \mathbb{E}_F[f(X)]$ where X is any random variable with probability distribution F. In particular, if $x^r \in L^1(\mathbb{R}, F)$ then $F(x^r)$ is the rth raw moment of F. In the case that F is the Marchenko-Pastur law we have the following explicit formula for these moments (see Lemma 3.1 of [4]).

Proposition 2.3 ([4])**.** *Let $c > 0$. Then*

$$F^c(x^r) = \sum_{k=0}^{r-1} \frac{c^k}{k+1}\binom{r}{k}\binom{r-1}{k} \quad \textit{for all} \quad r = 1, 2, 3, \ldots$$

In later proofs we shall need the first four moments; these are given by

Corollary 2.4. *Let $c > 0$. Then*

$$\begin{aligned} F^c(x) &= 1, \\ F^c(x^2) &= 1 + c, \\ F^c(x^3) &= 1 + 3c + c^2, \\ F^c(x^4) &= 1 + 6c + 6c^2 + c^3. \end{aligned}$$

2.3 Linear spectral statistics

As mentioned in the introduction, linear spectral statistics (of S_n) are functionals of the form

$$\frac{1}{p}\sum_{i=1}^{p} f(\ell_i) \tag{2}$$

where f is a suitable function on $[0,\infty)$ and $\ell_1,\ldots,\ell_p$ are the eigenvalues of S_n. Note that for sufficiently regular f one may use (1) to rewrite (2) as a Riemann-Stieltjes integral:

$$\frac{1}{p}\sum_{i=1}^{p} f(\ell_i) = \int f(x)\, dF^{S_n} = F^{S_n}(f).$$

In this thesis we shall only find use for LSSs of the form

$$a_r := \frac{1}{p}\operatorname{tr}(S_n^r) = \frac{1}{p}\sum_{i=1}^{p} \ell_i^r = F^{S_n}(x^r) \tag{3}$$

where r is a positive integer, but for the sake of completeness we cover the central limit theorem for LSSs in its general form.

Setting aside their ubiquity in multivariate statistics, LSSs have assumed increasing importance in the past decade due to the discovery that they possess a very general and powerful CLT. With a mind to formulating this result, let $H_n = F^{\Sigma_n}$, recall $p/n = c_n \to c$, and assume $H_n \xrightarrow{d} H$ for some c.d.f H. Because of Theorem 2.1, one then expects the (random) quantity $F^{S_n}(f) - F^{c_n,H_n}(f)$ to be small for sufficiently large n, and is lead to wonder about the rate at which it approaches zero. A natural guess is that this rate is essentially $1/p$, so that the sequence of random variables

$$X_n(f) := p\left\{F^{S_n}(f) - F^{c_n,H_n}(f)\right\}$$

might converge in distribution; in particular, if $\Sigma_n = I_p$ for all n then the remark after Theorem 2.2 implies that $F^{c_n,H_n} = F^{c_n}$ (the Marchenko-Pastur law with parameter c_n) and we expect

$$X_n(f) = p\left\{F^{S_n}(f) - F^{c_n}(f)\right\} \tag{4}$$

to converge in distribution. Indeed, we have the following

Theorem 2.5 ([24]). *Assume that $\Sigma_n = I_p$ for all n. Let $f_1, \ldots, f_k$ be real-valued analytic functions on $[0, \infty)$. The random vector $\{X_n(f_1), \ldots, X_n(f_k)\}$ converges in distribution to a multivariate Gaussian vector $(X_{f_1}, \ldots, X_{f_k})$ with mean function and covariance function*

$$\begin{aligned}\mathbb{E}(X_f) &= I_1(f) + \beta I_2(f),\\ \mathrm{Cov}(X_f, X_g) &= 2J_1(f,g) + \beta J_2(f,g),\end{aligned}$$

where

$$\begin{aligned}I_1(f) &= \lim_{r\downarrow 1} \frac{1}{2\pi i} \oint_{|\xi|=1} f(|1+h\xi|^2) \left[\frac{\xi}{\xi^2 - r^{-2}} - \frac{1}{\xi}\right] d\xi,\\ I_2(f) &= \frac{1}{2\pi i} \oint_{|\xi|=1} f(|1+h\xi|^2) \frac{1}{\xi^3} d\xi,\\ J_1(f,g) &= \lim_{r\downarrow 1} -\frac{1}{4\pi^2} \oint_{|\xi_1|=1} \oint_{|\xi_2|=1} \frac{f(|1+h\xi_1|^2) g(|1+h\xi_2|^2)}{(\xi_1 - r\xi_2)^2} d\xi_1 d\xi_2,\\ J_2(f,g) &= -\frac{1}{4\pi^2} \oint_{|\xi_1|=1} \frac{f(|1+h\xi_1|^2)}{\xi_1^2} d\xi_1 \oint_{|\xi_2|=1} \frac{g(|1+h\xi_2|^2)}{\xi_2^2} d\xi_2,\end{aligned}$$

and $h = \sqrt{c}$.

Remark. With appropriate modifications, this CLT even allows for complex-valued $X_{n;ij}$ and $f_1, \ldots, f_k$. An even more general version that allows for non-identity covariance matrices can be found in [20].

As an application of this CLT and their (at the time) newly discovered formulas for the mean function and covariance function, Wang and Yao [24] took $f_r(x) = x^r$, $r = 1, 2$ and obtained the following result.

Lemma 2.6 ([24]). *Assume that $\Sigma_n = I_p$ for all n. Then*

$$p\left\{\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 + c_n \end{pmatrix}\right\} \xrightarrow{d} \mathcal{N}(\boldsymbol{\mu}, \Sigma)$$

where

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ (1+\beta)c \end{pmatrix}$$

and

$$\Sigma = \begin{pmatrix} (2+\beta)c & 2(2+\beta)(c+c^2) \\ 2(2+\beta)(c+c^2) & 4c^2 + 4(2+\beta)(c+2c^2+c^3) \end{pmatrix}.$$

We shall use this lemma together with the delta method to derive our main result: a new test for sphericity. Later, we shall also consider a different statistic that depends not only on a_1, a_2 but also a_3, a_4, and as such we need an extended version of the lemma. Though the formulas in Theorem 2.5 for the mean and covariances are analytically tractable, they still demand very long computations (see the proof of Lemma 2.1 in [24]), and as the number of additional parameters is 14 the author does not deem it within the scope of this thesis to carry out these computations. Luckily, in the case $\beta = 0$ a pair of general formulas (see below) was known already to Bai and Silverstein [5], who were the first authors to prove a general CLT of this kind.

Proposition 2.7 ([5])**.** *Assume $\beta = 0$. With notation as in Theorem 2.5,*

$$\mathbb{E}(X_r) = \frac{1}{4}\big((1-\sqrt{c})^{2r} + (1+\sqrt{c})^{2r}\big) - \frac{1}{2}\sum_{j=0}^{r}\binom{r}{j}^2 c^j$$

and

$$\begin{aligned}\operatorname{Cov}(X_{x^{r_1}}, X_{x^{r_2}}) = 2c^{r_1+r_2}\sum_{k_1=0}^{r_1-1}\sum_{k_2=0}^{r_2}\binom{r_1}{k_1}\binom{r_2}{k_2}\left(\frac{1-c}{c}\right)^{k_1+k_2} \\ \times \sum_{\ell=1}^{r_1-k_1}\ell\binom{2r_1-1-(k_1+\ell)}{r_1-1} \\ \times \binom{2r_2-1-k_2+\ell}{r_2-1}\end{aligned}$$

for all positive integers r, r_1, r_2.

Thus we shall be satisfied with the following lemma:

Lemma 2.8. *Assume that $\beta = 0$ and $\Sigma_n = I_p$ for all n. Then*

$$p\left\{\begin{pmatrix}a_1\\a_2\\a_3\\a_4\end{pmatrix} - \begin{pmatrix}1\\1+c_n\\1+3c_n+c_n^2\\1+6c_n+6c_n^2+c_n^3\end{pmatrix}\right\} \xrightarrow{d} \mathcal{N}(\boldsymbol{\mu}, \Sigma) \tag{5}$$

with

$$\boldsymbol{\mu} = \begin{pmatrix}\mu_1\\\mu_2\\\mu_3\\\mu_4\end{pmatrix} \quad \textit{and} \quad \Sigma = \begin{pmatrix}\sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14}\\ \sigma_{12} & \sigma_{22} & \sigma_{23} & \sigma_{24}\\ \sigma_{13} & \sigma_{23} & \sigma_{33} & \sigma_{34}\\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_{44}\end{pmatrix}$$

where

$$\begin{aligned}\mu_1 &= 0,\\ \mu_2 &= c,\\ \mu_3 &= 3c(1+c),\\ \mu_4 &= c(6+17c+6c^2)\end{aligned}$$

and

$$\begin{aligned}\sigma_{11} &= 2c,\\ \sigma_{22} &= 4c(2+5c+2c^2),\\ \sigma_{33} &= 6c(3+24c+46c^2+24c^3+3c^4),\\ \sigma_{44} &= 8c(4+66c+300c^2+485c^3+300c^4+66c^5+4c^6),\\ \sigma_{12} &= 4c(1+c),\\ \sigma_{13} &= 6c(1+3c+c^2),\\ \sigma_{14} &= 8c(1+6c+6c^2+c^3),\\ \sigma_{23} &= 12c(1+5c+5c^2+c^3),\\ \sigma_{24} &= 8c(2+17c+32c^2+17c^3+2c^4),\\ \sigma_{34} &= 24c(1+12c+37c^2+37c^3+12c^4+c^5).\end{aligned}$$

The presence of c_n in the two preceding lemmas will later become an issue when we try to apply the delta method. The purpose of including (A) in our assumptions is to circumvent this issue by allowing c_n to be replaced with c.

Lemma 2.9. *Assume (A) and $\Sigma_n = I_p$ for all n. Then the conclusions of Lemma* 2.6 *and Lemma* 2.8 *hold with c_n replaced by c.*

2.4 Linear shrinkage estimators

In the terminology of Bodnar *et al.* [8], a general linear shrinkage estimator (GLSE) of Σ_n is a linear combination

$$\widehat{\Sigma}_{\mathrm{GLSE}} = \alpha_n S_n + \beta_n \Sigma_0 \tag{6}$$

where α_n, β_n are real numbers (the "shinkage intensities") and Σ_0 is a $p \times p$ nonrandom symmetric positive definite matrix (the "shrinkage target"). Consider now the problem of minimizing $\left\|\widehat{\Sigma}_{\mathrm{GLSE}} - \Sigma_n\right\|_F$ with respect to the

shrinkage intensities, where $\|A\|_F = \sqrt{\mathrm{tr}(AA^T)}$ denotes the Frobenius norm of a square real matrix A. (The idea is that under the large-dimensional asymptotics the unmodified sample covariance matrix S_n can be a poor estimator for Σ_n in the Frobenius norm sense, whereas an appropriate choice of shrinkage target and shrinkage intensities can make $\widehat{\Sigma}_{\mathrm{GLSE}}$ a much better estimate.) Bodnar *et al.* [8] solve this minimization problem to find the *optimal* shrinkage intensities, denoted α_n^*, β_n^*. We are here only interested in the former; it is given by

$$\alpha_n^* = \frac{\mathrm{tr}(S_n\Sigma_n)\,\|\Sigma_0\|_F^2 - \mathrm{tr}(\Sigma_n\Sigma_0)\,\mathrm{tr}(S_n\Sigma_0)}{\|S_n\|_F^2\,\|\Sigma_0\|_F^2 - (\mathrm{tr}(S_n\Sigma_0))^2}. \tag{7}$$

Unfortunately, α_n^* depends on the in applications unknown covariance matrix Σ_n, and so it becomes necessary to find a *bona fide* estimator for this optimal shrinkage intensity. To this end, Bodnar *et al.* [8] use the RMT to show that α_n^* is asymptotically equivalent (in a certain sense to be specified below) to the nonrandom quantity

$$\alpha^* = 1 - \frac{\frac{c}{p}(\mathrm{tr}(\Sigma_n))^2\,\|\Sigma_0\|_F^2}{(\|\Sigma_n\|_F^2 + \frac{c}{p}(\mathrm{tr}(\Sigma_n))^2)\,\|\Sigma_0\|_F^2 - (\mathrm{tr}(\Sigma_n\Sigma_0))^2} \tag{8}$$

and that this quantity is in turn asymptotically equivalent to the *bona fide* shrinkage intensity

$$\widehat{\alpha}^* = 1 - \frac{\frac{1}{n}(\mathrm{tr}(S_n))^2\,\|\Sigma_0\|_F^2}{\|S_n\|_F^2\,\|\Sigma_0\|_F^2 - (\mathrm{tr}(S_n\Sigma_0))^2} \tag{9}$$

which is a consistent estimator for α_n^*. More specifically, α_n^*, α^* and $\widehat{\alpha}_n^*$ are asymptotically equivalent in the sense of the following

Theorem 2.10 ([8])**.** *Assume that*

(i) $\mathbb{E}(|X_{n;ij}|^{4+\epsilon}) < +\infty$ *for some* $\epsilon > 0$*;*

(ii) $F^{\Sigma_n} \xrightarrow{d} H$ *for some c.d.f.* H*;*

(iii) the order of only a finite number of eigenvalues of Σ_n *can depend on* p*, and* $\lambda_{\max}(\Sigma_n)$ *is at most of order* $O(\sqrt{p})$*;*

(iv) $\sup_n \mathrm{tr}(\Sigma_0) < +\infty$.

Then $|\alpha_n^* - \alpha^*|$, $|\widehat{\alpha}^* - \alpha^*|$, *and* $|\alpha_n^* - \widehat{\alpha}^*|$ *all converge to zero a.s.*

Remark. Assumption (i) replaces our previous assumption of $\mathbb{E}(X_{n;ij}^4) < +\infty$. Note also that we in our notation have suppressed the dependence of Σ_0, α^* and $\widehat{\alpha}^*$ on n.

Similar results hold for the optimal and *bona fide* shrinkage intensities β_n and $\widehat{\beta}_n$, and with this Bodnar *et al.* [8] suggest the following "optimal linear shrinkage estimator" (OLSE) for Σ_n:

$$\widehat{\Sigma}_{\text{OLSE}} = \widehat{\alpha}^* S_n + \widehat{\beta}^* \Sigma_0. \tag{10}$$

The shrinkage target Σ_0 may be chosen as desired to take into account prior information or to speculate about the structure of Σ_n.

3 Tests for sphericity

The covariance matrix Σ_n is said to be *spherical* (a.k.a *isotropic*) if it is a scalar matrix; that is, if $\Sigma_n = \sigma^2 I_p$ for some unspecified scalar $\sigma^2 > 0$. A common problem in multivariate statistics is to test for sphericity, i.e. to test

$$H_0 : \Sigma_n = \sigma^2 I_p \quad \text{against} \quad H_A : \Sigma_n \neq \sigma^2 I_p.$$

Note that the seemingly more general problem of testing $H_1 : \Sigma_n = \Sigma_0$ where Σ_0 is a given $p \times p$ covariance matrix can be reduced to a sphericity test by first transforming the data as $\widetilde{Y}_n = \Sigma_0^{-1/2} Y_n$; this is because the covariance matrix of $\widetilde{Y}_n$ is

$$\widetilde{\Sigma}_n = \mathrm{Cov}(\Sigma_0^{-1/2} Y_n) = \Sigma_0^{-1/2} \mathrm{Cov}(Y_n)(\Sigma_0^{-1/2})^T = \Sigma_0^{-1/2} \Sigma_n \Sigma_0^{-1/2}$$

and thus $H_1 : \Sigma_n = \Sigma_0$ holds if and only $\widetilde{H}_0 : \widetilde{\Sigma}_n = I_p$ holds.

3.1 Description of the new shrinkage–based test

We shall now use the OLSE of Bodnar *et al.* [8] to establish our main result: a new test for sphericity. Let us henceforth take $\Sigma_0 = p^{-1} I_p$ in (6), so that under H_0 the problem of finding

$$\underset{\alpha_n, \beta_n}{\arg\min} \left\| \widehat{\Sigma}_{\mathrm{GLSE}} - \Sigma_n \right\|_F = \underset{\alpha_n, \beta_n}{\arg\min} \left\| \alpha_n S_n + \beta_n p^{-1} I_p - \sigma^2 I_p \right\|_F$$

has the obvious solution $\alpha_n^* = 0$, $\beta_n^* = p\sigma^2$. Alternatively, under the aforementioned shrinkage target, the optimal shrinkage intensity (7) reduces to

$$\alpha_n^* = \frac{\mathrm{tr}(S_n \Sigma_n) - \mathrm{tr}(\Sigma_n)\,\mathrm{tr}(S_n)}{\|S_n\|_F^2 - (\mathrm{tr}(S_n))^2}, \tag{11}$$

(we have rescaled the fraction by a factor p^2) which evidently is zero under H_0. (Note, however, that this relationship is an implication and not an equivalence.) The asymptotic quantity in (8), too, vanishes under H_0: First of all, it is invariant under rescaling of Σ_n by a nonzero constant so we may

without loss of generality assume $\Sigma_n = I_p$. Then

$$\begin{aligned}\alpha^* &= 1 - \frac{\frac{c}{p}(\mathrm{tr}(I_p))^2 \left\|p^{-1}I_p\right\|_F^2}{(\left\|I_p\right\|_F^2 + \frac{c}{p}(\mathrm{tr}(I_p))^2) \left\|p^{-1}I_p\right\|_F^2 - (\mathrm{tr}(p^{-1}I_p))^2} \\ &= 1 - \frac{c}{(p+cp)p^{-1} - 1} \\ &= 1 - \frac{c}{1+c-1} \\ &= 0.\end{aligned}$$

By combining one or both of these observations with Theorem 2.10 we immediately obtain

Proposition 3.1. *Under H_0, $\widehat{\alpha}^* \to 0$ a.s.*

Remark. The factor p^{-1} in the shrinkage target is just a necessity for condition (iv) in the theorem and holds no greater importance.

This suggests that one could conduct a test of H_0 by computing the *bona fide* shrinkage intensity $\widehat{\alpha}^*$ with the intention of rejecting H_0 if $\widehat{\alpha}^*$ deviates too much from zero. To determine how large this deviation would have to be in order to be considered statistically significant we shall prove a CLT for $\widehat{\alpha}^*$. The technical tools that enable this result are the previously stated CLT for LSSs along with the delta method. In the formulation of the result, recall that $\beta = \mathbb{E}(X_{n;ij}^4) - 3$ is the excess kurtosis.

Theorem 3.2. *Assume (A). Under H_0,*

$$T = \frac{p\widehat{\alpha}^* - (1+\beta)}{2} \xrightarrow{d} N(0,1).$$

We have therefore established the following approximate test of H_0 at significance level $\gamma \in (0,1)$:

Reject H_0 if and only if $T > z_\gamma$, where z_γ is the upper $100\gamma\%$ critical value of the standard Gaussian distribution, i.e., $\mathbb{P}(N(0,1) > z_\gamma) = \gamma$.

We will call our new test the *shrinkage test* (ST) for sphericity.

3.2 A strengthening of the test of Fisher *et al.*

For our second result we shall use the RMT to strengthen the sphericity test of Fisher *et al.* [9] in the sense that a certain CLT of theirs will be shown to hold under a weaker set of assumptions than that of the original paper. First, we give a brief explanation of how their test is derived: By conjugating both sides of $H_0 : \Sigma_n = \sigma^2 I_p$ with a matrix whose columns form an eigenbasis of Σ_n we may assume without loss of generality that Σ_n is diagonal, say $\Sigma_n = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$. Then H_0 holds if and only if $\lambda_1 = \cdots = \lambda_p = \sigma^2$. The key insight is now that given any positive integer r the Cauchy-Schwarz inequality yields

$$\left(\sum_{i=1}^{p} \lambda_i^r\right)^2 = \left(\sum_{i=1}^{p} 1 \cdot \lambda_i^r\right)^2 \leq \left(\sum_{i=1}^{p} 1^2\right)\left(\sum_{i=1}^{p} \lambda_i^{2r}\right) = p \sum_{i=1}^{p} \lambda_i^{2r}$$

with equality if and only if the vectors $(1, \ldots, 1)^T$ and $(\lambda_1, \ldots, \lambda_p)^T$ are colinear; that is, if and only if H_0 holds. As a result the quantity

$$1 \leq \psi_r = \frac{p \sum \lambda_i^{2r}}{\left(\sum \lambda_i^r\right)^2} = \frac{p^{-1} \operatorname{tr}(\Sigma_n^{2r})}{\left(p^{-1} \operatorname{tr}(\Sigma_n^r)\right)^2}$$

can be said to measure the deviation from sphericity in the sense that H_0 is true if and only $\psi_r = 1$. Thus, if one knows the asymptotic distribution of ψ_r one also obtains a test for H_0. Srivastava [23] and Fisher *et al.* [9] consider the cases $r = 1$ and $r = 2$ respectively.

Although a natural candidate estimator for ψ_2 is its sample counterpart

$$\frac{p^{-1} \operatorname{tr}(S_n^4)}{\left(p^{-1} \operatorname{tr}(S_n^2)\right)^2} = \frac{a_4}{a_2^2},$$

this naive estimator happens to be inconsistent under the large-dimensional asymptotics and so will not do (more on this below). As an alternative, Fisher *et al.* [9] show that if one additionally assumes that

(a) the $X_{n;ij}$ are standard Gaussians, and

(b) $p^{-1} \operatorname{tr}(\Sigma_n^i) \to a_i^0 \in (0, +\infty)$ for $i = 1, 2, \ldots, 16$,

then the estimator

$$\hat{\psi}_2 = \frac{\hat{a}_4}{\hat{a}_2^2}$$

is a consistent estimator of ψ_2, where

$$\hat{a}_2 = \frac{n^2}{(n-1)(n+2)}\frac{1}{p}\left(\operatorname{tr}(S_n^2) - \frac{1}{n}(\operatorname{tr}(S_n))^2\right)$$

is an unbiased and consistent estimator of $p^{-1}\operatorname{tr}(\Sigma_n^2)$, and

$$\hat{a}_4 = \frac{\tau}{p}\left(\operatorname{tr}(S_n^4) + b\operatorname{tr}(S_n^3)\operatorname{tr}(S_n) + c^*(\operatorname{tr}(S_n^2))^2 + d\operatorname{tr}(S_n^2)(\operatorname{tr}(S_n))^2 + e(\operatorname{tr}(S_n))^4\right)$$

is an unbiased and consistent estimator of $p^{-1}\operatorname{tr}(\Sigma_n^4)$, where

$$\tau = \frac{n^5(n^2+n+2)}{(n+1)(n+2)(n+4)(n+6)(n-1)(n-2)(n-3)},$$

$$b = -\frac{4}{n}, \quad c^* = -\frac{2n^2+3n-6}{n(n^2+n+2)}, \quad d = \frac{2(5n+6)}{n(n^2+n+2)}, \quad e = -\frac{5n+6}{n^2(n^2+n+2)}.$$

Remark. If condition (b) of Fisher *et al.* and conditions (i)-(iv) of Theorem 2.10 all hold, then Theorem 3.2 of Bodnar *et al.* [8] implies that $a_2 \xrightarrow{a.s.} a_2^0 + ca_1^0$. On the other hand, $\psi_2 \to a_4^0/(a_2^0)^2$, and thus we expect a_4/a_2^2 to be an inconsistent estimator for ψ_2 since the limit of the denominator of the former depends on c while the limit of the denominator of the latter does not.

Fisher *et al.* [9] then proceed to derive the unconditional (read: regardless of whether H_0 is true or not) asymptotic distribution of their estimator, and as a special case they obtain a CLT for $\hat{\psi}_r$ under H_0, which may then be used as basis for a test of H_0. Therefore, given that one is only interested in the testing aspect, our contribution shall be to show that the assumptions (a)-(b) are unnecessarily strong by using the RMT to re-establish the CLT under an arguably much weaker set of conditions. This more general CLT is stated in the following

Theorem 3.3. *Assume (A) and $\beta = 0$. Under H_0,*

$$T_F = n\frac{\psi_2 - 1}{\sqrt{8(8+12c+c^2)}} \xrightarrow{d} N(0,1).$$

Remark. While we consider this result valuable in that it does not require (b), we do feel obliged to point out that it is very hard to find non-Gaussian $X_{n;ij}$ such that $X_{n;ij}$ is both standardized and have excess kurtosis $\beta = 0$, and therefore one might in practice be forced to assume (a).

3.3 The corrected likelihood ratio test

For the sake of comparison in our simulation study we also include a third test for sphericity, the corrected likelihood ratio test (CLRT), though we shall have nothing new to prove about it. Originally derived in 1940 by Mauchly [19] for multivariate Gaussian samples, the likelihood ratio test statistic for H_0 is

$$V_n = \det(S_n) \cdot \left(p^{-1} \operatorname{tr}(S_n)\right)^{-p}.$$

Note that V_n is degenerate when $p > n$ since S_n is singular in this case, so the likelihood ratio test should not be used when the number of variables exceeds the sample size. A classical result says that under the standard asymptotics (read: p is fixed) and when H_0 holds, $-n \log V_n \xrightarrow{d} \chi^2(f)$, a chi-square distribution with degree of freedom $f = \frac{1}{2}p(p+1)+1$. After observing that

$$\frac{1}{p} \log\left(\det(S_n)\right) = \frac{1}{p} \sum_{j=1}^{p} \log \ell_j$$

is a linear spectral statistic, Wang and Yao [24] use Theorem 2.5 to prove the following CLT which serves as a large-dimensional correction of the likelihood ratio test.

Theorem 3.4 ([24])**.** *Assume that $c \in (0,1)$. Under H_0,*

$$T_L = -\log V_n + (p-n)\log\left(1 - \frac{p}{n}\right) - p \xrightarrow{d} N(\mu, \sigma_1^2)$$

where

$$\mu = -\frac{1}{2}\log(1-c) + \frac{1}{2}\beta c$$

and

$$\sigma_1^2 = -2\log(1-c) - 2c.$$

Remark. With appropriate modifications to the mean and variance this CLT is also valid for complex-valued $X_{n;ij}$ (see [24] for details).

Wang and Yao [24] point out that the asymptotic variance of T_L depends on c through a factor $-\log(1-c)$ and thus blows up when c approaches 1; as such, their prediction was that the power will break down when c is close to 1. This phenomenon was indeed observed in their Monte Carlo simulations, and their conclusion was that in general the CLRT may be preferable to other statistics only when the ratio p/n is much lower than 1.

4 Simulation study

We conduct a Monte Carlo simulation study in RStudio to investigate how the size and power of our new test, the shrinkage test (ST), compares to existing tests; namely, the test of Fisher *et al.* [9]—which henceforth will be referred to as the Cauchy-Schwarz test (CST)—and the corrected likelihood ratio test (CLRT). The following conventions shall be used throughout:

- T, T_F, and T_L denote the test statistics of the ST, the CST, and the CLRT, respectively.
- Since under H_0 all three statistics are invariant under rescaling of σ^2, we assume without loss of generality that $\sigma^2 = 1$.
- For each simulated value, the Monte Carlo sample size (a.k.a. the number of repetitions) is set to $N = 10^4$.
- We use a pre-set significance level equal to $\gamma = 0.05$.

Remark. Given a statistical hypothesis test of a null hypothesis H_0 vs. an alternative hypothesis H_1, the *size* is the probability of falsely rejecting the null hypothesis (a.k.a. the probability of making a type I error), i.e.

$$\text{Size} := \mathbb{P}(\text{test rejects } H_0 \mid H_0),$$

while the *power* is the probability of correctly rejecting the null hypothesis (a.k.a. the probability of *not* making a type II error), i.e.

$$\text{Power} := \mathbb{P}(\text{test rejects } H_0 \mid H_1).$$

In addition, the test is said to be *consistent* if the power approaches 1 as the sample size tends to infinity.

4.1 Size

Since each statistic T, T_F, T_L is asymptotically standard Gaussian under H_0, our criterion for rejection shall naturally be $T > z_\gamma$, where z_γ as before is the upper $100\gamma\%$ critical value of the standard Gaussian distribution. As estimator for size we use the *attained significance level* (ASL), which Fisher *et al.* [9] define as

$$\text{ASL}(T) = \frac{(\#T > z_\gamma)}{N}.$$

Thus, if Theorem 3.2 indeed is true and n is taken sufficiently large, we expect to see the outcome $\mathrm{ASL}(T) \approx \gamma$ under H_0. For the sake of comparison we also simulate the ASL of T_F, and, whenever it is applicable (read: $c < 1$), the ASL of T_L.

Remark. More precisely, if we for the sake of argument assume that the approximation $T \sim N(0,1)$ holds exactly, then $\mathbb{P}(T > z_\gamma) = \gamma$ and consequently $N \cdot \mathrm{ASL}(T) \sim \mathrm{Bin}(N, \gamma)$, so in this case we can construct an exact confidence interval for $\mathrm{ASL}(T)$ at confidence level $1 - \gamma = 0.95$, namely

$$\left(\frac{Q(\gamma/2; N, \gamma)}{N}, \frac{Q(1-\gamma/2; N, \gamma)}{N} \right) \approx (0.0458, 0.0543),$$

where $q \mapsto Q(q; N, \gamma)$ is the quantile function of $\mathrm{Bin}(N, \gamma)$. So, for all sufficiently large n we expect T to fall within the above confidence interval approximately 95% of the time.

Firstly, we simulate the ASL for all three statistics under H_0 and standard Gaussian $X_{n;ij}$ on a grid of values of n and c. Tables 1-3 provide the results, and although these seem to confirm the asymptotic normality of T, we observe that the rate of convergence appears much slower for c close to zero. In comparison, the statistics T_F and T_L appear to converge at a much faster rate across the board, and especially T_F seems to be well-approximated by a standard Gaussian even for as low sample sizes as $n = 50$.

Secondly, to investigate the impact of non-Gaussian data and nonzero excess kurtosis, we simulate the ASL for all three *uncorrected* statistics under H_0 (meaning we substitute $\beta = 0$ in the formulas for these statistics even if $\mathbb{E}(X_{n;ij}^4) - 3 \neq 0$). More specifically, we simulate t-distributed $X_{n;ij}$ with 10 degrees of freedom, so that $\beta = 1$. Tables 4-6 provide the results for the same grid of values of n and c as before, and we observe that even though β is rather small in magnitude, all values are significantly worse than the case with standard Gaussian $X_{n;ij}$.

Thirdly, we simulate the ASL under H_0 for the *corrected* versions of T and T_L under t-distributed $X_{n;ij}$ with 10 degrees of freedom to see if the corrections are indeed working as they should. (Since the author is are not aware of any existing corrected version of T_F we do not include this statistic in our considerations.) Tables 7-8 provide the results and appear to show that the statistics are indeed converging in distribution to a standard Gaussian as the sample size tends to infinity and that the rate is comparable to the situation with standard Gaussian data, if perhaps slightly slower.

Finally, we compute kernel density estimates for T and plot these alongside their corresponding large-dimensional asymptotic approximation: the standard Gaussian density. We use the same values for the concentration c as before, but this time we choose the dimension p adaptively and by hand until the ASL is deemed close enough to γ to proceed. The results are shown in Figure 1, and we observe that although the central parts of the density estimates appear somewhat non-Gaussian in the sense that the peaks are a bit uneven with mass slightly off to the side, it does seem like the general shapes of the densities closely match that of a standard Gaussian.

Table 1: ASL(T) under Gaussian $X_{n;ij}$.

$p = cn$	$c = 0.2$	$c = 0.5$	$c = 0.8$	$c = 1.5$	$c = 2$	$c = 5$
$n = 50$	0.0044	0.0231	0.0335	0.0413	0.0363	0.0464
$n = 100$	0.0171	0.0376	0.0407	0.0435	0.0443	0.0489
$n = 150$	0.0252	0.0402	0.0427	0.0433	0.0466	0.0521
$n = 200$	0.0308	0.0369	0.0440	0.0516	0.0501	0.0499

Table 2: ASL(T_F) under Gaussian $X_{n;ij}$.

$p = cn$	$c = 0.2$	$c = 0.5$	$c = 0.8$	$c = 1.5$	$c = 2$	$c = 5$
$n = 50$	0.0452	0.0479	0.0489	0.0529	0.0507	0.0492
$n = 100$	0.0514	0.0568	0.0547	0.0537	0.0548	0.0486
$n = 150$	0.0522	0.0529	0.0556	0.0488	0.0513	0.0546
$n = 200$	0.0575	0.0501	0.0544	0.0559	0.0508	0.0495

Table 3: ASL(T_L) under Gaussian $X_{n;ij}$.

$p = cn$	$c = 0.2$	$c = 0.5$	$c = 0.8$
$n = 50$	0.0628	0.0564	0.0566
$n = 100$	0.0529	0.0548	0.0523
$n = 150$	0.0529	0.0535	0.0513
$n = 200$	0.0506	0.0475	0.0473

Table 4: Uncorrected ASL(T) under t-distributed $X_{n;ij}$.

$p = cn$	$c = 0.2$	$c = 0.5$	$c = 0.8$	$c = 1.5$	$c = 2$	$c = 5$
$n = 50$	0.0147	0.0720	0.0885	0.1057	0.1107	0.1220
$n = 100$	0.0588	0.0982	0.1067	0.1222	0.1222	0.1227
$n = 150$	0.0783	0.1067	0.1163	0.1211	0.1231	0.1241
$n = 200$	0.0923	0.1122	0.1116	0.1257	0.1242	0.1267

Table 5: Uncorrected ASL(T_F) under t-distributed $X_{n;ij}$.

$p = cn$	$c = 0.2$	$c = 0.5$	$c = 0.8$	$c = 1.5$	$c = 2$	$c = 5$
$n = 50$	0.0926	0.1032	0.1003	0.0892	0.0850	0.0772
$n = 100$	0.1193	0.1157	0.1085	0.0947	0.0929	0.0749
$n = 150$	0.1241	0.1126	0.1082	0.0935	0.0845	0.0750
$n = 200$	0.1255	0.1150	0.1032	0.0913	0.0877	0.0721

Table 6: Uncorrected ASL(T_L) under t-distributed $X_{n;ij}$.

$p = cn$	$c = 0.2$	$c = 0.5$	$c = 0.8$
$n = 50$	0.1140	0.1100	0.0912
$n = 100$	0.1220	0.1049	0.0915
$n = 150$	0.1206	0.1110	0.0938
$n = 200$	0.1248	0.1079	0.0882

Table 7: Corrected ASL(T) under t-distributed $X_{n;ij}$.

$p = cn$	$c = 0.2$	$c = 0.5$	$c = 0.8$	$c = 1.5$	$c = 2$	$c = 5$
$n = 50$	0.0021	0.0202	0.0311	0.0409	0.0450	0.0475
$n = 100$	0.0131	0.0331	0.0400	0.0480	0.0457	0.0494
$n = 150$	0.0197	0.0372	0.0403	0.0490	0.0476	0.0482
$n = 200$	0.0294	0.0405	0.0444	0.0512	0.0496	0.0478

Table 8: Corrected ASL(T_L) under t-distributed $X_{n;ij}$.

$p = cn$	$c = 0.2$	$c = 0.5$	$c = 0.8$
$n = 50$	0.0633	0.0581	0.0543
$n = 100$	0.0597	0.0517	0.0515
$n = 150$	0.0532	0.0537	0.0540
$n = 200$	0.0574	0.0488	0.0495

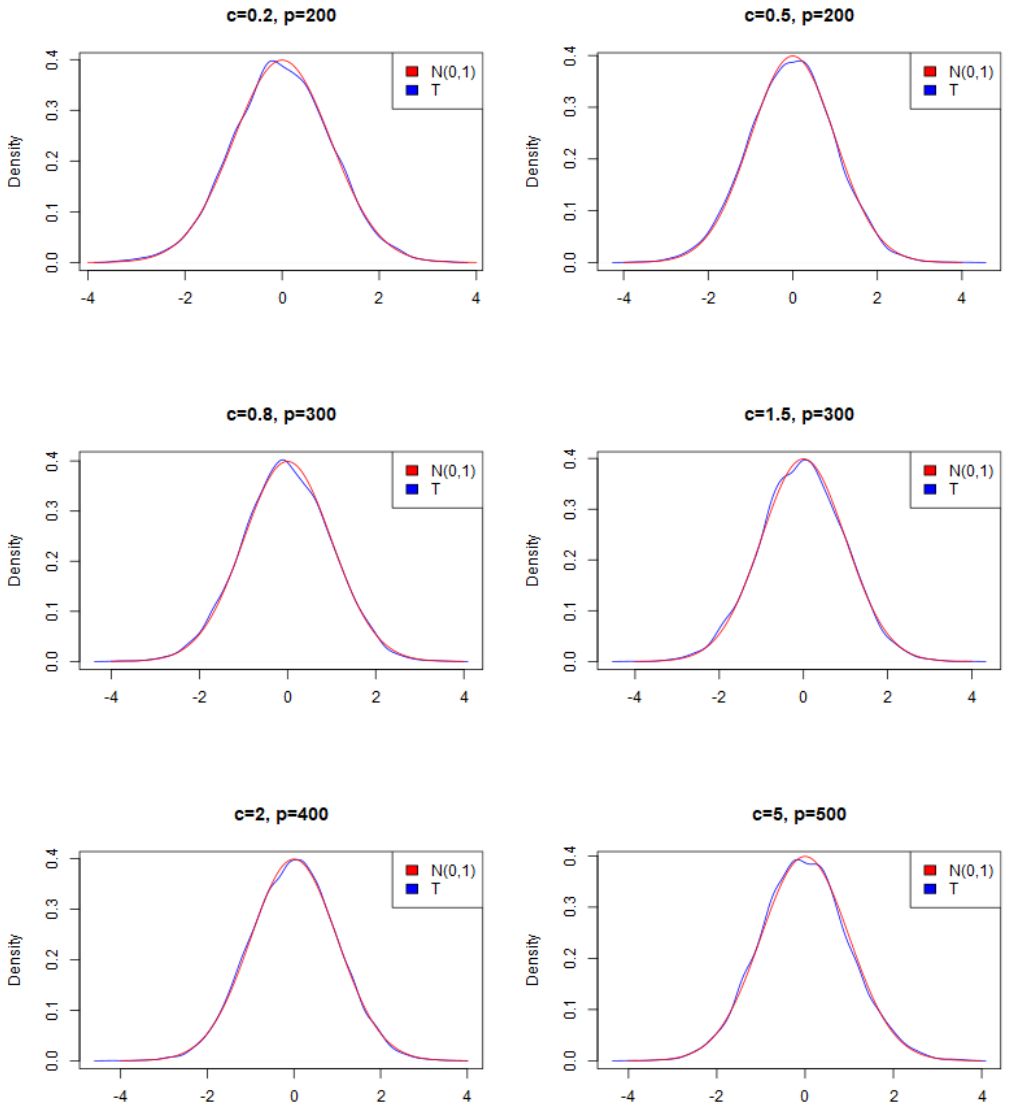

Figure 1: The large-dimensional asymptotic approximation of the density of T together with kernel density estimators for T.

4.2 Power

Next, we simulate the power of all three tests under various realizations of H_1 with a mind to seeing how the ST compares to the other two tests and under what conditions the former might be preferable (if at all). For the sake of simplicity we let $X_{n;ij}$ be standard Gaussian in all our power simulations. This leaves us with the issue of deciding (1) what criterion to use for rejection, and (2) what non-spherical covariance matrix to use for H_1. Concerning (1), we have seen in our simulations of the ASL that the rate at which T converges in distribution to a standard Gaussian (as $n \to \infty$) under H_0 is strongly dependent on the concentration constant c, and it is natural to presume that similar differences in rates of convergence are present also under H_1. Thus, using $T > z_\gamma$ as the criterion for rejection may be problematic, and we have instead opted to use the same two-step procedure as Srivastava [23] and Fisher *et al.* [9]:

1. Under H_0, simulate a Monte Carlo sample from T and use this sample to compute an estimate $\hat{T}_\gamma$ for the critical value of T at significance level γ, i.e. the real number T_γ satisfying $\mathbb{P}(T > T_\gamma \mid H_0) = \gamma$.

2. Under H_1, simulate a new Monte Carlo sample from T and use this sample to compute the frequency

$$\frac{(\#T > \hat{T}_\gamma)}{N},$$

   which is then taken as an estimator of the power.

(We shall use the R function `quantile` with default options to compute the critical value in step 1.)

Remark. The author's understanding of this procedure and why it may be preferable is that it essentially uses an exact test for H_0, since if we ignore the estimation uncertainty present in $\hat{T}_\gamma$ then $\mathbb{P}(T > \hat{T}_\gamma) = \gamma$ and thus $\mathbb{P}(\text{test rejects } H_0 \mid H_0) = \gamma$. This is in contrast to the approximate test based on the limiting distribution of T, since then $\mathbb{P}(T > z_\gamma \mid H_0) = \gamma$ only holds in the limit and thus $\mathbb{P}(\text{test rejects } H_0 \mid H_0) \approx \gamma$. Of course, if n is large enough that T is well-approximated by a standard Gaussian, then $\hat{T}_\gamma \approx z_\gamma$ and so these two methods of computing the power will give approximately the same value.

Concerning (2), a number of different scenarios for H_1 will be investigated. We start by considering the case where the empirical spectral distribution of Σ_n remains approximately the same for all n; more specifically, let Σ_n be diagonal with equal proportions of eigenvalues 0.8, 1, and 1.5. (That is, one third of the diagonal elements are equal to 0.8, one third are equal to 1, and one third are equal to 1.5, and if p is not a multiple of three then we include or exclude one additional element 1.5 compared to the number of 0.8s and 1s.) Table 9 presents the simulated powers under this Σ_n for a grid of values of n and c. We observe that all three tests appear to be consistent, but that the ST has higher power than both other tests for all considered choices of parameters; moreover, when n is fixed and c increases, the power of the ST increases slightly while the powers of the other two tests decrease slightly.

For our next few scenarios for H_1 we will again follow the lead of Fisher *et al.* [9] and consider two types of "near spherical" covariance matrices; more precisely, let

$$\Sigma_\theta = \text{diag}(\theta, 1, 1, \ldots, 1) \quad \text{for} \quad \theta > 0$$

and

$$\Sigma_\Theta = \begin{pmatrix} \Theta & \mathbf{0}^T \\ \mathbf{0} & I \end{pmatrix}$$

where

$$\Theta = \text{diag}(0.75, 1.25, 1.75, 2.25, 2.75, 3.25).$$

Firstly, we simulate the powers under $\Sigma_n = \Sigma_\theta$ for $\theta = 3$ and $\theta = 4$ on the same grid of values of n and c as previously. Tables 10-11 present the results, and we observe that the ST appears to be consistent, at least for $c < 1$, but also that the rate of convergence rapidly drops off with increasing c, and for $c > 1$ the power almost seems to be plateauing. In general, when $c < 1$ the ST performs much better than the CLRT and slightly more worse than the CST. When $c > 1$, the ST performs worse than the CST with the difference being more pronounced for larger c.

Secondly, we simulate the power under $\Sigma_n = \Sigma_\Theta$ on the same grid of values of n and c. The results are found in Table 12, and we observe that the power appears to behave essentially the same as in the previous situation. In general, when c is close to 0, the ST performs comparably to the CST while the CLRT performs comparably or worse. When $c > 1$, the CST once again dominates, but the difference is less pronounced compared to the case with $\Sigma_n = \Sigma_\theta$.

Thirdly, we investigate how the powers of T and T_F under $\Sigma_n = \Sigma_\theta$ behave as functions of θ by simulating these powers on a grid of values of θ. Figure 2 provides the results for $n = 50$, $c = 3$, and $\theta \in (0, 8]$, and it appears that the CST outperforms the ST for all values of θ.

(Note that we have excluded the case with concentration $c = 5$ in our power simulations. This is because the time it took to simulate the ASL suggests that the power under $n = 200$, $c = 5$, $p = cn = 1000$ would take a very long time to simulate.)

Table 9: Simulated power under diagonal Σ_n with spectrum $\{0.8, 1, 1.5\}$.

$p = cn$	$c = 0.2$			$c = 0.5$		
	T	T_F	T_L	T	T_F	T_L
$n = 50$	0.4463	0.3252	0.4233	0.5044	0.3092	0.3671
$n = 100$	0.9055	0.8067	0.8557	0.9393	0.7453	0.8316
$n = 150$	0.9966	0.9814	0.9932	0.9994	0.9614	0.9906
$n = 200$	1.0000	0.9995	1.0000	1.0000	0.9991	0.9995

$p = cn$	$c = 0.8$			$c = 1.5$		$c = 2$	
	T	T_F	T_L	T	T_F	T	T_F
$n = 50$	0.5079	0.2726	0.2745	0.5090	0.2175	0.5431	0.1966
$n = 100$	0.9532	0.6793	0.6612	0.9589	0.5227	0.9624	0.4601
$n = 150$	0.9994	0.9236	0.9374	0.9998	0.8333	0.9998	0.7525
$n = 200$	1.0000	0.9945	0.9951	1.0000	0.9662	1.0000	0.9270

Table 10: Simulated power under $\Sigma_n = \Sigma_\theta$ with $\theta = 3$.

$p = cn$	$c = 0.2$			$c = 0.5$		
	T	T_F	T_L	T	T_F	T_L
$n = 50$	0.9307	0.9566	0.8076	0.7477	0.8409	0.3567
$n = 100$	0.9915	0.9981	0.9119	0.8550	0.9631	0.3724
$n = 150$	0.9977	0.9997	0.9442	0.8980	0.9907	0.3903
$n = 200$	0.9994	1.0000	0.9623	0.9276	0.9962	0.4056

$p = cn$	$c = 0.8$			$c = 1.5$		$c = 2$	
	T	T_F	T_L	T	T_F	T	T_F
$n = 50$	0.5847	0.7332	0.1597	0.3312	0.4818	0.2559	0.3743
$n = 100$	0.6725	0.8834	0.1593	0.3486	0.6126	0.2691	0.4494
$n = 150$	0.6948	0.9298	0.1656	0.3730	0.6588	0.2752	0.4883
$n = 200$	0.7202	0.9620	0.1689	0.3753	0.7048	0.2570	0.5169

Table 11: Simulated power under $\Sigma_n = \Sigma_\theta$ with $\theta = 4$.

$p = cn$	$c = 0.2$			$c = 0.5$		
	T	T_F	T_L	T	T_F	T_L
$n = 50$	0.9939	0.9970	0.9684	0.9661	0.9886	0.6746
$n = 100$	1.0000	1.0000	0.9965	0.9947	0.9996	0.7282
$n = 150$	1.0000	1.0000	0.9998	0.9983	1.0000	0.7544
$n = 200$	1.0000	1.0000	0.9998	0.9999	1.0000	0.7695

$p = cn$	$c = 0.8$			$c = 1.5$		$c = 2$	
	T	T_F	T_L	T	T_F	T	T_F
$n = 50$	0.9007	0.9599	0.3100	0.7042	0.8529	0.5650	0.7613
$n = 100$	0.9619	0.9959	0.3310	0.7810	0.9614	0.6361	0.8930
$n = 150$	0.9846	0.9996	0.3300	0.8269	0.9838	0.6767	0.9453
$n = 200$	0.9907	0.9998	0.3292	0.8435	0.9932	0.6769	0.9656

Table 12: Simulated power under $\Sigma_n = \Sigma_\Theta$.

$p = cn$	$c = 0.2$			$c = 0.5$		
	T	T_F	T_L	T	T_F	T_L
$n = 50$	0.9928	0.9608	0.9899	0.9835	0.9736	0.8424
$n = 100$	1.0000	1.0000	1.0000	0.9995	0.9998	0.9391
$n = 150$	1.0000	1.0000	1.0000	1.0000	1.0000	0.9636
$n = 200$	1.0000	1.0000	1.0000	1.0000	1.0000	0.9667

$p = cn$	$c = 0.8$			$c = 1.5$		$c = 2$	
	T	T_F	T_L	T	T_F	T	T_F
$n = 50$	0.9439	0.9344	0.4675	0.7724	0.7888	0.6597	0.6755
$n = 100$	0.9915	0.9978	0.5312	0.8644	0.9403	0.7410	0.8268
$n = 150$	0.9971	0.9998	0.5658	0.9026	0.9785	0.7552	0.8980
$n = 200$	0.9991	1.0000	0.5786	0.9233	0.9888	0.7550	0.9282

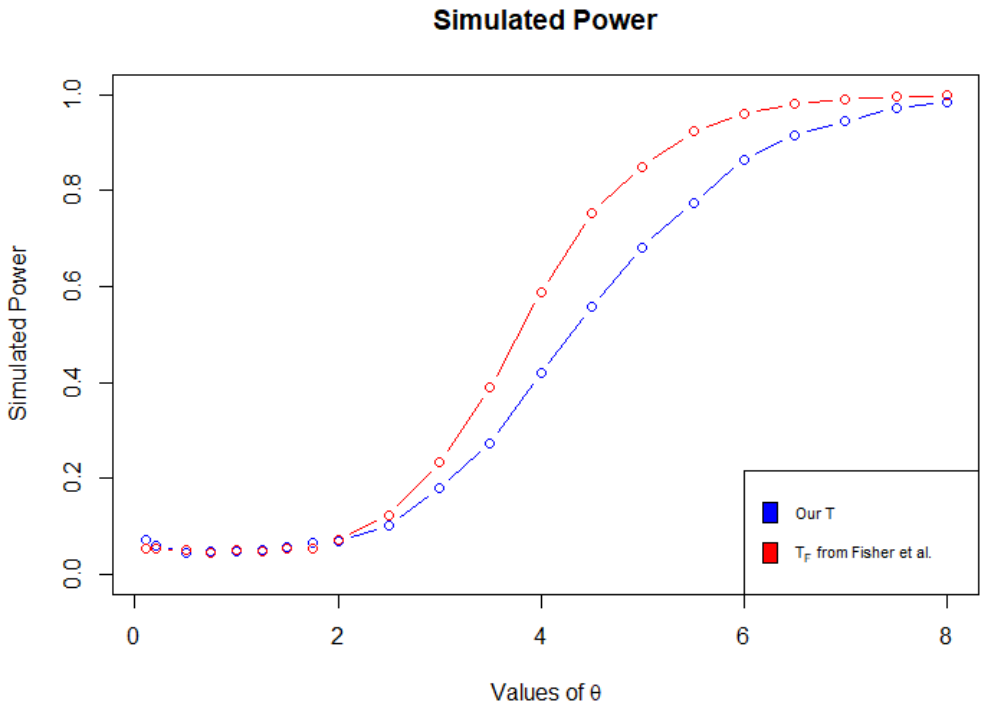

Figure 2: Simulated powers of T and T_F as θ increases.

5 Discussion

5.1 Performance of the new test

From our simulations we conclude the following:

- T is asymptotically standard Gaussian under H_0, thus confirming our Theorem 3.2.
- However, the rate of convergence may be slow in comparison to that of statistics associated with other tests, especially when c is close to zero.
- The exact ST does well at detecting non-sphericity when the number of non-identity elements is proportional to p and outperforms both the CST and the CLRT in this setting.
- Under near-spherical covariance matrix, the exact ST outperforms the CLRT but is outperformed by the CST.

Restricting ourselves to the three tests considered in this thesis, it would therefore seem that the choice lies mostly between our new shrinkage test and the CST. We shall discuss some additional pros and cons of these two tests before giving our verdict.

The ST enjoys an important computational advantage compared with the CST, namely that the former uses fewer linear spectral statistics than the latter (2 vs. 4): Computing S_n is unavoidable and requires $O(p^2n)$ operations, but each additional linear spectral statistic a_r we use (assuming $a_r = p^{-1}\operatorname{tr}(S_n^r)$ is computed as it is written and we store the intermediate matrices S_n^r) requires an additional matrix multiplication and $O(p^3)$ additional operations to compute. When p is very large this leads to the CST taking significantly more time to execute compared to the ST, which could make the latter preferable. (Of course, in principle one could numerically compute the eigenvalues $\ell_1, \ldots, \ell_p$ of S_n and use these to compute each $a_r = p^{-1}\sum_{i=1}^p \ell_i^r$ using only $O(p)$ operations, but this method might be too numerically unstable to rely on for testing purposes.)

On the other hand, a demerit of the ST and a possible explanation for its low power in certain circumstances is its relative "detachment" from the hypothesis it purports to test: To begin with, in the derivation of the test itself we technically only have that the sphericity hypothesis $H_0 : \Sigma_n = \sigma^2 I_p$ *implies* a second auxiliary hypothesis $H_0' : \alpha_n = 0$, where the latter is what

we actually test. (Indeed, (11) shows that $\alpha_n = 0$ if and only if $\operatorname{tr}(S_n\Sigma_n) = \operatorname{tr}(S_n)\operatorname{tr}(\Sigma_n)$ which can happen even if Σ_n is not spherical.) Secondly, since we cannot compute α_n directly, we have to settle for approximating it using the *bona fide* estimate $\widehat{\alpha}^*$, and though we know that the distance between these quantities converges to zero almost surely, this method still introduces a further degree of separation between H_0 and our test statistic. In comparison, the CST takes advantage of an if-and-only-if–relationship between the null hypothesis H_0 and the auxiliary hypothesis $\psi_2 = 1$. Furthermore, both the numerator and denominator of $\hat{\psi}_2$ are computed using estimators that are unbiased under Gaussian data, which may improve the accuracy of the estimation and the quality of the test in certain cases. In theory the CST can also extract (in a vague sense) more information from data than the ST since the former uses more linear spectral statistics than the latter.

With these considerations in mind, we conclude that the (approximate) ST can indeed be preferable to other tests, and we recommend that it be used when (1) c is large and/or n is large, and (2) one suspects that the true covariance matrix is non-spherical but not necessarily near-spherical.

5.2 Future work

What follows are some of our suggestions for future work based on the content and findings of this thesis.

- Include the ST in an empirical study, say applied to financial data.
- Use the RMT to find the asymptotic distribution of $p\hat{\alpha}^*$ under the alternative hypothesis (given that this is possible, that is).
- Derive a generalization of the CST test statistic to the case with non-zero excess kurtosis β.

An additional observation we would like to make is that we have relied on the formulas in Proposition 2.7 to provide us with the asymptotic means and covariances of the random variables a_r when $\beta = 0$, and the existence of these formulas, along with the simple appearance of the expressions for the means and covariances in Lemma 2.6, lead us to believe that it should be possible to use Theorem 2.5 to derive generalizations of these formulas to arbitrary β. Such formulas could prove useful in future applications of this CLT, since the case $f_r(x) = x^r$ is bound to appear often.

5.3 Summary

This thesis is dedicated to the problem of testing the covariance matrix for sphericity in the large-dimensional setting. Using recent central limit theorems on linear spectral statistics of large-dimensional sample covariance matrices, and an optimal linear shrinkage estimator for large-dimensional covariance matrices, we derive the asymptotic distribution of a *bona fide* optimal linear shrinkage intensity and use this to establish a new test for sphericity. Monte Carlo simulations reveal that the new test performs better than an existing test based on the likelihood ratio, while depending on the nature of the alternative hypothesis it may perform better or worse than a different existing test based on the Cauchy-Schwarz inequality. Moreover, we generalize the latter test to non-Gaussian data and weaker assumptions on the covariance matrix.

Appendix

Proof of Corollary 2.4. Using Proposition 2.3 we find that the first four moments of the Marchenko-Pastur law F^c are

$$\begin{aligned} F^c(x) &= \frac{1}{1} \cdot 1 \cdot 1 \\ &= 1, \end{aligned}$$

$$\begin{aligned} F^c(x^2) &= \frac{1}{1} \cdot 1 \cdot 1 + \frac{c}{2} \cdot 2 \cdot 1 \\ &= 1 + c, \end{aligned}$$

$$\begin{aligned} F^c(x^3) &= \frac{1}{1} \cdot 1 \cdot 1 + \frac{c}{2} \cdot 3 \cdot 2 + \frac{c^2}{3} \cdot 3 \cdot 1 \\ &= 1 + 3c + c^2, \end{aligned}$$

$$\begin{aligned} F^c(x^4) &= \frac{1}{1} \cdot 1 \cdot 1 + \frac{c}{2} \cdot 4 \cdot 3 + \frac{c^2}{3} \cdot 6 \cdot 3 + \frac{c^3}{4} \cdot 4 \cdot 1 \\ &= 1 + 6c + 6c^2 + c^3. \end{aligned}$$

□

Proof of Lemma 2.8. Theorem 2.5 gives that $\{X_n(x), X_n(x^2), X_n(x^3), X_n(x^4)\}$ converges in distribution to a multivariate Gaussian vector $(X_x, X_{x^2}, X_{x^3}, X_{x^4})$. For each $r = 1, 2, 3, 4$ we have upon combining (4) with (3) that

$$X_n(x^r) = p\left\{a_r - F^{c_n}(x^r)\right\}$$

where $F^{c_n}(x^r)$ is the rth raw moment of the Marchenko-Pastur law with parameter c_n and is therefore given by Corollary 2.4. It remains to compute the parameters $\mu_i = \mathbb{E}(X_{x^i})$ and $\sigma_{ij} = \mathrm{Cov}(X_{x^i}, X_{x^j})$ for $1 \leq i, j \leq 4$. The parameters $\mu_1, \mu_2, \sigma_{11}, \sigma_{22}, \sigma_{12}$ may immediately be obtained by taking $\beta = 0$ in Lemma 2.6, so it suffices to compute $\mu_3, \mu_4, \sigma_{13}, \sigma_{14}, \sigma_{23}, \sigma_{24}, \sigma_{33}, \sigma_{34}, \sigma_{44}$

using Proposition 2.7. The two means are

$$\begin{aligned}\mu_3 &= \frac{1}{4}\big((1-\sqrt{c})^6 + (1+\sqrt{c})^6\big) \\ &\qquad - \frac{1}{2}\left(\binom{3}{0}^2 + \binom{3}{1}^2 c + \binom{3}{2}^2 c^2 + \binom{3}{3}^2 c^3\right) \\ &= \frac{1}{4}\big(2 + 30c + 30c^2 + 2c^3\big) - \frac{1}{2}\left(1 + 9c + 9c^2 + c^3\right) \\ &= 3c(1+c),\end{aligned}$$

$$\begin{aligned}\mu_4 &= \frac{1}{4}\big((1-\sqrt{c})^8 + (1+\sqrt{c})^8\big) \\ &\qquad - \frac{1}{2}\left(\binom{4}{0}^2 + \binom{4}{1}^2 c + \binom{4}{2}^2 c^2 + \binom{4}{3}^2 c^3 + \binom{4}{4}^2 c^4\right) \\ &= \frac{1}{4}\big(2 + 56c + 140c^2 + 56c^3 + 2c^4\big) - \frac{1}{2}\left(1 + 16c + 36c^2 + 16c^3 + c^4\right) \\ &= c(6 + 17c + 6c^2),\end{aligned}$$

while the first of the covariances is

$$\begin{aligned}
\sigma_{13} &= 2c^4 \sum_{k_1=0}^{0} \sum_{k_2=0}^{3} \binom{1}{k_1} \binom{3}{k_2} \left(\frac{1-c}{c}\right)^{k_1+k_2} \\
&\qquad \times \sum_{\ell=1}^{1-k_1} \ell \binom{1-(k_1+\ell)}{0} \binom{5-k_2+\ell}{2} \\
&= 2c^4 \sum_{k_2=0}^{3} \binom{1}{0} \binom{3}{k_2} \left(\frac{1-c}{c}\right)^{k_2} \\
&\qquad \times \sum_{\ell=1}^{1} \ell \binom{1-\ell}{0} \binom{5-k_2+\ell}{2} \\
&= 2c^4 \sum_{k_2=0}^{3} \binom{3}{k_2} \left(\frac{1-c}{c}\right)^{k_2} \cdot 1 \cdot \binom{0}{0} \binom{6-k_2}{2} \\
&= 2c^4 \sum_{k_2=0}^{3} \binom{3}{k_2} \binom{6-k_2}{2} \left(\frac{1-c}{c}\right)^{k_2} \\
&= 2c^4 \left(1 \cdot 15 + 3 \cdot 10 \cdot \frac{1-c}{c} + 3 \cdot 6 \cdot \left(\frac{1-c}{c}\right)^2 + 1 \cdot 3 \cdot \left(\frac{1-c}{c}\right)^3\right) \\
&= 6c[5c^3 + 10c^2(1-c) + 6c(1-2c+c^2) + (1-3c+3c^2-c^3)] \\
&= 6c[5c^3 + 10c^2 - 10c^3 + 6c - 12c^2 + 6c^3 + 1 - 3c + 3c^2 - c^3] \\
&= 6c(1+3c+c^2).
\end{aligned}$$

The remaining covariances can be obtained in a similar fashion and as such we omit their computations.

□

Proof of Lemma 2.9. Let $r \in \{1,2,3,4\}$. Writing

$$p\left\{a_r - F^c(x^r)\right\} = p\left\{a_r - F^{c_n}(x^r)\right\} + p\left\{F^{c_n}(x^r) - F^c(x^r)\right\},$$

we see by Slutsky's theorem that it suffices to show that the rightmost quantity converges to zero. For each positive integer k the polynomial $x^k - c^k$ has a root at $x = c$ and hence the factor theorem yields a polynomial $p_k(x)$ with coefficients depending only on k and c such that $x^k - c^k = (x-c)p_k(x)$. Let

us also define $p_0(x) \equiv 0$. It follows by Proposition 2.3 that

$$F^{c_n}(x^r) - F^c(x^r) = (c_n - c) \sum_{k=0}^{r-1} \frac{p_k(c_n)}{k+1} \binom{r}{k} \binom{r-1}{k}.$$

Since the above sum is bounded in n (owing to c_n being convergent and therefore bounded) we need only multiply both sides with $p = c_n n$ and use (A). □

Proof of Theorem 3.2. Recall that the shrinkage target is $\Sigma_0 = p^{-1} I_p$. To use the RMT we first need to express $\widehat{\alpha}^*$ in terms of linear spectral statistics:

$$\begin{aligned}
\widehat{\alpha}^* &= 1 - \frac{\frac{1}{n}(\operatorname{tr}(S_n))^2 \left\|p^{-1} I_p\right\|_F^2}{\left\|S_n\right\|_F^2 \left\|p^{-1} I_p\right\|_F^2 - (\operatorname{tr}(S_n p^{-1} I_p))^2} \\
&= 1 - \frac{\frac{1}{n}(\operatorname{tr}(S_n))^2 p^{-1}}{\left\|S_n\right\|_F^2 p^{-1} - (p^{-1} \operatorname{tr}(S_n))^2} \\
&= 1 - \frac{\frac{p}{n}(p^{-1} \operatorname{tr}(S_n))^2}{p^{-1} \operatorname{tr}(S_n^2) - (p^{-1} \operatorname{tr}(S_n))^2} \\
&= 1 - c_n \frac{a_1^2}{a_2 - a_1^2}
\end{aligned}$$

where $a_r = p^{-1} \operatorname{tr}(S_n^r)$, $r = 1, 2$ are LSSs with $f_r(x) = x^r$, respectively. We shall now use the delta method. Let us for this purpose introduce the function $g(x_1, x_2) = x_1^2/(x_2 - x_1^2)$ and write

$$\widehat{\alpha}^* = 1 - c_n g(a_1, a_2).$$

Under H_0, $S_n = n^{-1}\sigma^2 X_n X_n^T$. It is easy to see that $g(a_1, a_2)$ is invariant under rescaling of S_n by a positive constant, so we may without loss of generality assume that $\sigma^2 = 1$, which under H_0 is equivalent to $\Sigma_n = I_p$. Then the conditions of Lemma 2.9 are satisfied and we have

$$p \left\{ \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 + c \end{pmatrix} \right\} \xrightarrow{d} \mathcal{N}(\boldsymbol{\mu}, \Sigma) \tag{12}$$

where

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ (1 + \beta)c \end{pmatrix} \tag{13}$$

and

$$\Sigma = \begin{pmatrix} (2+\beta)c & 2(2+\beta)(c+c^2) \\ 2(2+\beta)(c+c^2) & 4c^2 + 4(2+\beta)(c+2c^2+c^3) \end{pmatrix}. \tag{14}$$

By multiplying both "sides" of (12) with p^{-1} and using Slutsky's theorem we obtain $a_1 \xrightarrow{p} 1$, $a_2 \xrightarrow{p} 1+c$ and hence

$$g(a_1, a_2) \xrightarrow{p} g(1, 1+c) = \frac{1^2}{(1+c)-1^2} = \frac{1}{c}.$$

Together with (A) this implies that the second term in

$$p\widehat{\alpha}^* = p\left\{1 - cg(a_1, a_2)\right\} + c_n n(c - c_n) g(a_1, g_2)$$

converges in probability to zero, and by Slutsky's theorem it therefore suffices to show that the first term converges in distribution to $N(1+\beta, 4)$. As already mentioned we have $g(1, 1+c) = 1/c$, so the delta method yields

$$p\left\{g(a_1, a_2) - 1/c\right\} \xrightarrow{d} N(J\mu, J\Sigma J^T).$$

where

$$\begin{aligned} J &= \nabla g(1, 1+c) \\ &= \left(\frac{2x_1 x_2}{(x_2 - x_1^2)^2}, -\frac{x_1^2}{(x_2 - x_1^2)^2} \right) \Bigg|_{(a_1,a_2)=(1,1+c)} \\ &= \left(\frac{2(1+c)}{c^2}, -\frac{1}{c^2} \right). \end{aligned}$$

From (13) we get

$$J\boldsymbol{\mu} = -\frac{1+\beta}{c}$$

and from (14) we get

$$\begin{aligned} J\Sigma J^T &= J \begin{pmatrix} (2+\beta)c & 2(2+\beta)(c+c^2) \\ 2(2+\beta)(c+c^2) & 4c^2+4(2+\beta)(c+2c^2+c^3) \end{pmatrix} \frac{1}{c^2} \begin{pmatrix} 2(1+c) \\ -1 \end{pmatrix} \\ &= J\frac{1}{c} \begin{pmatrix} (2+\beta) & 2(2+\beta)(1+c) \\ 2(2+\beta)(1+c) & 4c+4(2+\beta)(1+c)^2 \end{pmatrix} \begin{pmatrix} 2(1+c) \\ -1 \end{pmatrix} \\ &= J\frac{1}{c} \begin{pmatrix} 0 \\ -4c \end{pmatrix} \\ &= J \begin{pmatrix} 0 \\ -4 \end{pmatrix} \\ &= \frac{4}{c^2} \end{aligned}$$

which gives

$$p\left\{g(a_1,a_2)-1/c\right\} \xrightarrow{d} N\left(-\frac{1+\beta}{c}, \frac{4}{c^2}\right)$$

We therefore only need to multiply the quantity on the left hand side by $-c$ and apply Slutsky's theorem to obtain

$$p\left\{1-cg(a_1,a_2)\right\} \xrightarrow{d} N(1+\beta,4)$$

and the proof is done. □

Proof of Theorem 3.3. Our strategy shall as before be to use the RMT together with the delta method. As $\hat{\psi}_2$ is easily seen to be invariant under rescaling of S_n by a positive constant we may as in the previous proof assume $\Sigma_n = I_p$ without loss of generality, so Lemma 2.9 is applicable. Multiply both "sides" of (5) (the version that has c instead of c_n) with p^{-1} and apply Slutsky's theorem to obtain

$$\begin{aligned} a_1 &\xrightarrow{p} 1, \\ a_2 &\xrightarrow{p} 1+c, \\ a_3 &\xrightarrow{p} 1+3c+c^2, \\ a_4 &\xrightarrow{p} 1+6c+6c^2+c^3. \end{aligned} \tag{15}$$

We start by considering $\hat{a}_2$ which we rewrite as

$$\hat{a}_2 = A\left(a_2 - c_n a_1^2\right)$$

where

$$A := \frac{n^2}{(n-1)(n+2)}.$$

Then (15) implies

$$\hat{a}_2 \xrightarrow{p} 1 \cdot ((1+c) - c \cdot 1^2) = 1.$$

We next introduce the function

$$g_2(x_1, x_2) = x_2 - cx_1^2$$

along with the random variable $a_2' = g_2(a_1, a_2) = a_2 - ca_1^2$. Similarly to $\hat{a}_2$ we have $a_2' \xrightarrow{p} 1$. Consider now the decomposition

$$\begin{aligned} n(\hat{a}_2 - 1) &= n(AA^{-1}\hat{a}_2 - A^{-1}\hat{a}_2 + A^{-1}\hat{a}_2 - a_2' + a_2' - 1) \\ &= n(A-1)A^{-1}\hat{a}_2 + n(A^{-1}\hat{a}_2 - a_2') + n(a_2' - 1) \\ &= t_1 + t_2 + n(a_2' - 1). \end{aligned}$$

We show that $t_1 = n(A-1)A^{-1}\hat{a}_2$ and $t_2 = n(A^{-1}\hat{a}_2 - a_2')$ each converges in probability to a constant: Firstly,

$$n(A-1) = -\frac{n(n-2)}{(n-1)(n+2)} \to -1$$

and hence $t_1 \xrightarrow{p} -1$ where we have used $A \to 1$ and $\hat{a}_2 \xrightarrow{p} 1$. Secondly,

$$t_2 = n((a_2 - c_n a_1^2) - (a_2 - ca_1^2)) = n(c - c_n)a_1^2 \xrightarrow{p} 0$$

where we have used (A) and $a_1 \xrightarrow{p} 1$. In total we hence have

$$n(\hat{a}_2 - 1) = n(a_2' - 1) - 1 + o_p(1)$$

where $o_p(1)$ denotes a term that converges to zero in probability. We proceed to investigate $\hat{a}_4$ in a similar fashion by first rewriting it as

$$\begin{aligned} \hat{a}_4 &= \frac{\tau}{p}(\operatorname{tr}(S_n^4) + b \operatorname{tr}(S_n^3)\operatorname{tr}(S_n) + c^*(\operatorname{tr}(S_n^2))^2 + d \operatorname{tr}(S_n^2)(\operatorname{tr}(S_n))^2 + e(\operatorname{tr}(S_n))^4) \\ &= \tau(a_4 + pb \cdot a_3 a_1 + pc^* \cdot a_2^2 + p^2 d \cdot a_2 a_1^2 + p^3 e \cdot a_1^4) \\ &= \tau(a_4 + Ba_3 a_1 + Ca_2^2 + Da_2 a_1^2 + Ea_1^4) \end{aligned}$$

where

$$
\begin{aligned}
B &:= pb = -4c_n \\
C &:= pc^* = -\frac{2n^2+3n-6}{n^2+n+2}c_n \\
D &:= p^2 d = \frac{10n^2+12n}{n^2+n+2}c_n^2 \\
E &:= p^3 e = -\frac{5n^2+6n}{n^2+n+2}c_n^3.
\end{aligned}
$$

We have

$$
\tau = \frac{n^7+n^6+O(n^5)}{n^7+7n^6+O(n^5)} \to 1
$$

and

$$
\begin{aligned}
B &\to -4c =: B_0, \\
C &\to -2c =: C_0, \\
D &\to 10c^2 =: D_0, \\
E &\to -5c^3 =: E_0.
\end{aligned}
$$

which together with (15) implies

$$
\begin{aligned}
\hat{a}_4 &\xrightarrow{p} (1+6c+6c^2+c^3) - 4c(1+3c+c^2) - 2c(1+c)^2 + 10c^2(1+c) - 5c^3 \\
&= 1.
\end{aligned}
$$

We next introduce the function

$$
\begin{aligned}
g_4(x_1,x_2,x_3,x_4) &= x_4 + B_0 x_3 x_1 + C_0 x_2^2 + D_0 x_2 x_1^2 + E_0 x_1^4 \\
&= x_4 - 4c x_3 x_1 - 2c x_2^2 + 10c^2 x_2 x_1^2 - 5c^3 x_1^4
\end{aligned}
$$

along with the random variable $a_4' = g(a_1,a_2,a_3,a_4)$. Similarly to $\hat{a}_4$ we have $a_4' \xrightarrow{p} 1$. Consider now the decomposition

$$
\begin{aligned}
n(\hat{a}_4 - 1) &= n(\tau\tau^{-1}\hat{a}_4 - \tau^{-1}\hat{a}_4 + A^{-1}\hat{a}_4 - \tilde{a}_4 + \tilde{a}_4 - 1) \\
&= n(\tau-1)\tau^{-1}\hat{a}_4 + n(\tau^{-1}\hat{a}_4 - \tilde{a}_4) + n(\tilde{a}_4 - 1) \\
&= t_3 + t_4 + n(\tilde{a}_4 - 1).
\end{aligned}
$$

We show that both $t_3 = n(\tau - 1)\tau^{-1}\hat{a}_4$ and $t_4 = n(\tau^{-1}\hat{a}_4 - \tilde{a}_4)$ each converge in probability to a constant. Firstly,

$$n(\tau - 1) = \frac{-6n^7 + O(n^6)}{n^7 + O(n^6)} \to -6$$

and hence $t_3 \xrightarrow{p} -6$. Secondly, we may decompose t_4 further as

$$t_4 = n(B - B_0)a_3a_1 + n(C - C_0)a_2^2 + n(D - D_0)a_2a_1^2 + n(E - E_0)a_1^4$$

where, due to (A),

$$n(B - B_0) = 4n(c - c_n) \to 0$$

and

$$\begin{aligned}
n(C - C_0) &= n\left(2c - \frac{2n^2 + 3n - 6}{n^2 + n + 2}c_n\right) \\
&= n\left(2c - 2c_n + 2c_n - \frac{2n^2 + 3n - 6}{n^2 + n + 2}c_n\right) \\
&= 2n(c - c_n) + n\left(2 - \frac{2n^2 + 3n - 6}{n^2 + n + 2}\right)c_n \\
&= o(1) - \frac{n(n - 10)}{n^2 + n + 2}c_n \\
&\to -c
\end{aligned}$$

and

$$\begin{aligned}
n(D - D_0) &= n\left(\frac{10n^2 + 12n}{n^2 + n + 2}c_n^2 - 10c^2\right) \\
&= n\left(\frac{10n^2 + 12n}{n^2 + n + 2}c_n^2 - 10c_n^2 + 10c_n^2 - 10c^2\right) \\
&= n\left(\frac{10n^2 + 12n}{n^2 + n + 2} - 10\right)c_n^2 + 10n(c_n^2 - c^2) \\
&= \frac{2n(n - 10)}{n^2 + n + 2}c_n^2 + 10n(c_n - c)(c_n + c) \\
&\to 2c^2
\end{aligned}$$

and

$$\begin{aligned} n(E - E_0) &= n\left(5c^3 - \frac{5n^2+6n}{n^2+n+2}c_n^3\right) \\ &= n\left(5c^3 - 5c_n^3 + 5c_n^3 - \frac{5n^2+6n}{n^2+n+2}c_n^3\right) \\ &= 5n(c^3 - c_n^3) + n\left(5 - \frac{5n^2+6n}{n^2+n+2}\right)c_n^3 \\ &= 5n(c - c_n)(c^2 + cc_n + c_n^2) - \frac{n(n-10)}{n^2+n+2}c_n^3 \\ &\to -c^3. \end{aligned}$$

These limits together with (15) imply

$$\begin{aligned} t_4 &\xrightarrow{p} 0 - c(1+c)^2 + 2c^2(1+c)\cdot 1^2 - c^3 \cdot 1^4 \\ &= -c. \end{aligned}$$

We may thus write

$$n(\hat{a}_4 - 1) = n(a_4' - 1) - 6 - c + o_p(1).$$

Summarizing the proof so far, we have introduced two new random variables a_2', a_4' and proved that

$$n\left\{\begin{pmatrix}\hat{a}_2\\ \hat{a}_4\end{pmatrix} - \begin{pmatrix}1\\1\end{pmatrix}\right\} = n\left\{\begin{pmatrix}a_2'\\ a_4'\end{pmatrix} - \begin{pmatrix}1\\1\end{pmatrix}\right\} - \begin{pmatrix}1\\ 6+c\end{pmatrix} + o_p(1). \tag{16}$$

If we define $g : \mathbb{R}^4 \to \mathbb{R}^2$ by $g = (g_2, g_4)^T$, then $(a_2', a_4')^T = g(a_1, a_2, a_3, a_4)$. Let

$$\begin{aligned} J &:= \nabla g(1, 1+c, 1+3c+c^2, 1+6c+6c^2+c^3) \\ &= \begin{pmatrix} -2c & 1 & 0 & 0 \\ -4c(1-c)^2 & 2c(-2+3c) & -4c & 1 \end{pmatrix}. \end{aligned}$$

Then Lemma 2.9 along with the delta method yields

$$p\left\{\begin{pmatrix}a_2'\\ a_4'\end{pmatrix} - \begin{pmatrix}1\\1\end{pmatrix}\right\} \xrightarrow{d} \mathcal{N}(J\boldsymbol{\mu}, J\Sigma J^T) \tag{17}$$

where $\boldsymbol{\mu}$, Σ are as in Lemma 2.8. A straightforward matrix calculation that we omit for brevity gives

$$J\boldsymbol{\mu} = \begin{pmatrix} c \\ c(6+c) \end{pmatrix}.$$

Thus, upon multiplying the left hand side of (17) with $1/c_n = n/p$ and using Slutsky's theorem we obtain

$$n\left\{\begin{pmatrix} a_2' \\ a_4' \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \end{pmatrix}\right\} \xrightarrow{d} \mathcal{N}\left\{\begin{pmatrix} 1 \\ 6+c \end{pmatrix}, \frac{1}{c^2} J\Sigma J^T\right\}.$$

Evidently the asymptotic mean in this CLT is equal to the negative of the constant vector in the right hand side of (16), and hence another application of Slutsky's theorem yields

$$n\left\{\begin{pmatrix} \hat{a}_2 \\ \hat{a}_4 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \end{pmatrix}\right\} \xrightarrow{d} \mathcal{N}(\mathbf{0}, c^{-2} J\Sigma J^T).$$

Let us write $\hat{\psi}_2 = h(\hat{a}_2, \hat{a}_4)$ where $h(y_2, y_4) = y_4/y_2^2$. We have $h(1,1) = 1$ and

$$J_1 := \nabla h(1,1) = (-2, 1)$$

and so another application of the delta method yields

$$n\{\hat{\psi}_2 - 1\} \xrightarrow{d} N(0, c^{-2}(J_1 J)\Sigma(J_1 J)^T).$$

Some final matrix calculations which we again omit for brevity give

$$c^{-2}(J_1 J)\Sigma(J_1 J)^T) = 8(8 + 12c + c^2)$$

and the proof is complete. □

References

[1] Z. Bai, D. Jiang, J.-F. Yao, and S. Zheng, *Corrections to LRT on large-dimensional covariance matrix by RMT*, Ann. Statist., 37 (2009), pp. 3822–3840.

[2] Z. Bai, H. Li, and G. Pan, *Central limit theorem for linear spectral statistics of large dimensional separable sample covariance matrices*, Bernoulli, 25 (2019), pp. 1838–1869.

[3] Z. Bai and H. Saranadasa, *Effect of high dimension: by an example of a two sample problem*, Statist. Sinica, 6 (1996), pp. 311–329.

[4] Z. Bai and J. W. Silverstein, *Spectral analysis of large dimensional random matrices*, Springer Series in Statistics, Springer, New York, second ed., 2010.

[5] Z. D. Bai and J. W. Silverstein, *CLT for linear spectral statistics of large-dimensional sample covariance matrices*, Ann. Probab., 32 (2004), pp. 553–605.

[6] T. Bodnar, H. Dette, and N. Parolya, *Testing for independence of large dimensional vectors*, Ann. Statist., 47 (2019), pp. 2977–3008.

[7] T. Bodnar, S. Dmytriv, Y. Okhrin, N. Parolya, and W. Schmid, *Statistical inference for the expected utility portfolio in high dimensions*, IEEE Trans. Signal Process., 69 (2021), pp. 1–14.

[8] T. Bodnar, A. K. Gupta, and N. Parolya, *On the strong convergence of the optimal linear shrinkage estimator for large dimensional covariance matrix*, J. Multivariate Anal., 132 (2014), pp. 215–228.

[9] T. J. Fisher, X. Sun, and C. M. Gallagher, *A new test for sphericity of the covariance matrix for high dimensional data*, J. Multivariate Anal., 101 (2010), pp. 2554–2570.

[10] L. Le Cam and G. L. Yang, *Asymptotics in statistics*, Springer Series in Statistics, Springer-Verlag, New York, second ed., 2000. Some basic concepts.

[11] O. Ledoit and M. Wolf, *A well-conditioned estimator for large-dimensional covariance matrices*, J. Multivariate Anal., 88 (2004), pp. 365–411.

[12] ——, *Optimal estimation of a large-dimensional covariance matrix under Stein's loss*, Bernoulli, 24 (2018), pp. 3791–3832.

[13] ——, *Analytical nonlinear shrinkage of large-dimensional covariance matrices*, Ann. Statist., 48 (2020), pp. 3043–3065.

[14] H. Li, A. Aue, and D. Paul, *High-dimensional general linear hypothesis tests via non-linear spectral shrinkage*, Bernoulli, 26 (2020), pp. 2541–2571.

[15] Z. Li, F. Han, and J. Yao, *Asymptotic joint distribution of extreme eigenvalues and trace of large sample covariance matrix in a generalized spiked population model*, Ann. Statist., 48 (2020), pp. 3138–3160.

[16] M. E. Lopes, A. Blandino, and A. Aue, *Bootstrapping spectral statistics in high dimensions*, Biometrika, 106 (2019), pp. 781–801.

[17] H. Markowitz, *Portfolio selection [reprint of J. Finance* **7** *(1952), no. 1, 77–91]*, in Financial risk measurement and management, vol. 267 of Internat. Lib. Crit. Writ. Econ., Edward Elgar, Cheltenham, 2012, pp. 197–211.

[18] V. A. Marčenko and L. A. Pastur, *Distribution of eigenvalues in certain sets of random matrices*, Mat. Sb. (N.S.), 72 (114) (1967), pp. 507–536.

[19] J. W. Mauchly, *Significance test for sphericity of a normal n-variate distribution*, Ann. Math. Statistics, 11 (1940), pp. 204–209.

[20] G. M. Pan and W. Zhou, *Central limit theorem for signal-to-interference ratio of reduced rank linear receiver*, Ann. Appl. Probab., 18 (2008), pp. 1232–1270.

[21] V. Serdobolskii, *Multivariate statistical analysis*, vol. 41 of Theory and Decision Library. Series B: Mathematical and Statistical Methods, Kluwer Academic Publishers, Dordrecht, 2000. A high-dimensional approach.

[22] J. W. Silverstein, *Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices*, J. Multivariate Anal., 55 (1995), pp. 331–339.

[23] M. S. Srivastava, *Some tests concerning the covariance matrix in high dimensional data*, J. Japan Statist. Soc., 35 (2005), pp. 251–272.

[24] Q. Wang and J. Yao, *On the sphericity test with large-dimensional observations*, Electron. J. Stat., 7 (2013), pp. 2164–2192.

[25] E. P. Wigner, *On the distribution of the roots of certain symmetric matrices*, Ann. of Math. (2), 67 (1958), pp. 325–327.

[26] P. Yaskov, *A short proof of the Marchenko-Pastur theorem*, C. R. Math. Acad. Sci. Paris, 354 (2016), pp. 319–322.