Computational Statistics and Data Analysis 104 (2016) 10–23

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

# Power computation for hypothesis testing with high-dimensional covariance matrices

Ruitao Lin [a], Zhongying Liu [b], Shurong Zheng [b,*], Guosheng Yin [a]

[a] Department of Statistics and Actuarial Science, The University of Hong Kong, China
[b] School of Mathematics & Statistics and KLAS, Northeast Normal University, China

ARTICLE INFO

*Article history:*
Received 4 June 2015
Received in revised form 23 February 2016
Accepted 11 May 2016
Available online 6 June 2016

*Keywords:*
Central limit theorem
Confidence interval
High-dimensional covariance matrix
Hypothesis testing
Power calculation
Stieltjes transform

ABSTRACT

Based on the random matrix theory, a unified numerical approach is developed for power calculation in the general framework of hypothesis testing with high-dimensional covariance matrices. In the central limit theorem of linear spectral statistics for sample covariance matrices, the theoretical mean and covariance are computed numerically. Based on these numerical values, the power of the hypothesis test can be evaluated, and furthermore the confidence interval for the unknown parameters in the high-dimensional covariance matrix can be constructed. The validity of the proposed algorithms is well supported by a convergence theorem. Our numerical method is assessed by extensive simulation studies, and a real data example of the S&P 100 index data is analyzed to illustrate the proposed algorithms.

© 2016 Elsevier B.V. All rights reserved.

## 1. Introduction

High-dimensional data become commonplace in the modern big data era, particularly in the fields of finance, genomics, informatics, image analysis, and so on. Estimation and inference based on the sample covariance matrix are fundamentally important in multivariate statistical analysis. However, when the dimensionality grows with the sample size, the conventional estimate of the covariance matrix is inconsistent (Bai and Yin, 1993), which thus leads to poor inference procedures in the high-dimensional setting. To address the inconsistency issue, extensive research has been conducted on estimation of high-dimensional covariance matrices. In particular, the random matrix theory is able to attenuate the randomness of the sample covariance matrix, and thus is widely used as a popular tool to achieve efficient statistical inference on the high-dimensional covariance matrices. Bai and Silverstein (2004) and Zheng et al. (2015) established the central limit theorem (CLT) of linear spectral statistics (LSS) for large dimensional sample covariance matrices when the dimension $p$ and the sample size $n$ increase proportionally; that is, $p/n \to c$ where $c$ is some constant in $(0, \infty)$.

Based on the CLT of LSS of sample covariance matrices, extensive research has been carried out on testing the high-dimensional covariance matrix structures. For example, Bai et al. (2009) proposed to test whether the high-dimensional covariance matrix is equal to a specified matrix. Nevertheless, there is no explicit power function, because the asymptotic mean and covariance in the CLT of LSS of large dimensional covariance matrices are expressed by the contour integrals that cannot be derived in closed forms. In hypothesis testing, power functions are important for evaluating the effectiveness of the testing methods. For example, Chen et al. (2012), Onatski et al. (2013) and Wang (2014) studied the power functions for

* Corresponding author.
*E-mail address:* zhengsr@nenu.edu.cn (S. Zheng).

http://dx.doi.org/10.1016/j.csda.2016.05.008
0167-9473/© 2016 Elsevier B.V. All rights reserved.

different covariance testing problems. However, all these power functions are derived for some specific high-dimensional covariance structure, and may not be generally applicable.

Our method is motivated by a real example in financial studies. As a preliminary step, the equicorrelation structure is commonly adopted for analyzing the correlation structures among financial assets (Ledoit and Wolf, 2004; Kyj et al., 2009; Engle and Kelly, 2012). Given the return data on a set of stocks, it is of great importance to examine whether the stock returns are equicorrelated. Therefore, the hypothesis testing procedure can be used to reduce the risk of inappropriate modeling for the covariance matrix. In addition to testing the covariance structure, the confidence intervals of the equicorrelation coefficients can provide useful information on the movement of stock prices. Unfortunately, it is often challenging to compute the power and confidence intervals as the number of assets is typically large relative to the number of observations.

Our goal is to provide a general numerical algorithm for computing the asymptotic mean and covariance in the CLT of LSS of large-dimensional sample covariance matrices in Zheng et al. (2015). The numerical algorithm can be used to obtain the power functions of high-dimensional covariance tests. Moreover, when the covariance matrix $\mathbf{\Sigma}$ is characterized by some low dimensional parameters $\boldsymbol{\rho}$, say $\mathbf{\Sigma} = \mathbf{\Sigma}(\boldsymbol{\rho})$, the confidence region of $\boldsymbol{\rho}$ can be numerically constructed as well.

The rest of the paper is arranged as follows. Section 2 establishes a general lemma based on Zheng et al. (2015) and provides a numerical algorithm for computation of the theoretical mean and variance in the CLT of LSS of sample covariance matrices. The power function for hypothesis testing and confidence intervals for the parameters in the covariance matrix are also derived. Section 3 exhibits some simulation studies, and Section 4 conducts a real data analysis on the S&P 100 index data for illustration of our proposal. Section 5 concludes with a discussion and technical details are delineated to Appendix.

## 2. Power and confidence interval

### 2.1. Motivating example: Hypothesis testing of a covariance matrix

Let $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ be a sample from the population with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$, where $\mathbf{y}_i$ is a $p$-dimensional random vector, $i = 1, \ldots, n$. It is of interest in multivariate statistical analysis to test

$$H_0 : \mathbf{\Sigma} = \mathbf{I}_p,$$

where $\mathbf{I}_p$ is the identity matrix of dimension $p$ (Anderson, 2003). For a more general framework, the identity matrix $\mathbf{I}_p$ in the null hypothesis $H_0$ can be replaced by any positive-definite covariance matrix $\mathbf{\Sigma}_0$. Many well-known test statistics are constructed as functionals of the eigenvalues of the sample covariance matrix, which is defined as

$$\mathbf{S}_n = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^T,$$

where $\bar{\mathbf{y}} = n^{-1}\sum_{i=1}^{n}\mathbf{y}_i$ is the sample mean and the superindex "$^T$" denotes the transpose. Hereafter, such statistics are referred as linear spectral statistics (LSS). For example, the likelihood ratio test (LRT) statistic for the Gaussian case is given by

$$T_n = \sum_{j=1}^{p} \mathrm{tr}\mathbf{S}_n - \log|\mathbf{S}_n| - p,$$

where "tr" denotes the trace of a matrix. Bai et al. (2009) established the CLT for $T_n$ under the null hypothesis $H_0$ for high-dimensional cases when $c_n = p/n \to c \in (0, \infty)$, while the theoretical power function was not derived despite its importance in hypothesis testing. Moreover, when $\mathbf{\Sigma}$ is characterized by some parameters $\boldsymbol{\rho}$, i.e., $\mathbf{\Sigma} = \mathbf{\Sigma}(\boldsymbol{\rho})$, it is also of interest to construct the confidence region for $\boldsymbol{\rho}$.

### 2.2. CLT of LSS for sample covariance matrices

Let $\{\lambda_1 \le \cdots \le \lambda_p\}$ be the eigenvalues of the population covariance matrix $\mathbf{\Sigma}$. The empirical spectral distribution (ESD) of $\mathbf{\Sigma}$ is given by

$$H_p(x) = \frac{1}{p}\sum_{j=1}^{p}\delta\{\lambda_j \le x\}, \quad \text{for any } x \in R$$

where $\delta\{\cdot\}$ is the indicator function and $R$ is the real line. Let $\{\hat{\lambda}_1 \le \cdots \le \hat{\lambda}_p\}$ be the eigenvalues of the sample covariance matrix $\mathbf{S}_n$, and correspondingly the ESD of $\mathbf{S}_n$ is

$$F_n(x) = \frac{1}{p}\sum_{j=1}^{p}\delta\{\hat{\lambda}_j \le x\}, \quad \text{for any } x \in R.$$

The Marčenko–Pastur scheme is adopted; that is,

$$c_n = p/n \to c \in (0, \infty), \quad \text{as } p, n \to \infty.$$

In addition, it is assumed that $H_p$ converges weakly to $H$, the limiting spectral distribution (LSD) of $\mathbf{\Sigma}$, and that $F_n$ converges weakly to $F^{c,H}$, the LSD of $\mathbf{S}_n$, as $p, n \to \infty$. Typically, when $\mathbf{\Sigma}$ is an identity matrix, $F^{c,H}$ is the Marčenko–Pastur law with the support set $[(1-\sqrt{c})^2\delta_{\{c<1\}}, (1+\sqrt{c})^2]$ (Marčenko and Pastur, 1967). When $\mathbf{\Sigma}$ is not an identity matrix, the support set of $F^{c,H}$ may include $[a, b]$ with

$$a = \liminf_p \lambda_1(1-\sqrt{c})^2\delta_{\{c<1\}}, \qquad b = \limsup_p \lambda_p(1+\sqrt{c})^2,$$

where "lim inf" and "lim sup" denote the infimum and supremum limits, respectively (Bai and Silverstein, 2004).

The LSS of $\mathbf{S}_n$ is $p\int f(x)dF_n(x) = \sum_{j=1}^p f(\hat{\lambda}_j)$, where $f$ is an analytic function in the support set $[a, b]$. Zheng et al. (2015) established the CLT of $G_p(f)$,

$$G_p(f) = \sum_{j=1}^p f(\hat{\lambda}_j) - p\int_a^b f(x)dF^{c_{n-1},H_p}(x),$$

where $c_{n-1} = p/(n-1)$ and $F^{c_{n-1},H_p}(x)$ is obtained by plugging $(c_{n-1}, H_p)$ in $F^{c,H}$. Bai and Silverstein (2004) indicated that

$$\frac{d}{dx}F^{c_{n-1},H_p}(x) = \frac{1}{c_{n-1}\pi}\lim_{z\to x}\Im(\underline{m}_{H_p}(z)),$$

where $\Im(\cdot)$ denotes the imaginary part and $\underline{m}_{H_p}(z)$ is the unique solution of the Stieltjes equation

$$z = -\frac{1}{\underline{m}_{H_p}(z)} + c_{n-1}\int \frac{t}{1+t\underline{m}_{H_p}(z)}dH_p(t). \tag{1}$$

We impose three assumptions as those in Zheng et al. (2015). Let $\mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Gamma}^T$ where $\mathbf{\Gamma} = \mathbf{A}\text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_p})$ with the eigenvector matrix $\mathbf{A}$.

Assumption (a). The sample $\mathbf{y}_i$ satisfies $\mathbf{y}_i = \boldsymbol{\mu} + \mathbf{\Gamma}\mathbf{x}_i$ where $\mathbf{x}_i = (x_{1i}, \ldots, x_{pi})^T$ for $i = 1, \ldots, n$, and $\{x_{ji}, j \le p, i \le n\}$ are independent random variables with common moments,

$$\mathrm{E}x_{ji} = 0, \qquad \mathrm{E}x_{ji}^2 = 1, \qquad \beta_x = \mathrm{E}x_{ji}^4 - 3 < \infty,$$

and the Lindeberg condition holds,

$$\frac{1}{np}\sum_{j=1}^p\sum_{i=1}^n \mathrm{E}\{x_{ji}^4\delta(|x_{ji}| \ge \eta\sqrt{n})\} \to 0 \quad \text{as } p/n \to c \in (0, \infty),$$

for any $\eta > 0$.

Assumption (b) The asymptotic regime $c_n = p/n \to c \in (0, \infty)$ is satisfied.

Assumption (c) The sequence of $\{\mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Gamma}^T\}_{p\ge 1}$ is bounded in the spectral norm. That is, there exists a constant $M$ satisfying $0 \le \lambda_1 \le \cdots \le \lambda_p \le M$. The ESD $H_p$ of $\mathbf{\Sigma}$ converges weakly to the LSD $H$ as $p \to \infty$.

**Lemma 1.** *Under Assumptions* (a)–(c)*, if* $f_1, \ldots, f_L$ *are analytic in the support set* $[a, b]$*, then the random vector* $(G_p(f_1), \ldots, G_p(f_L))$ *with*

$$G_p(f_\ell) = \sum_{j=1}^p f_\ell(\hat{\lambda}_j) - p\int_a^b f_\ell(x)dF^{c_{n-1},H_p}(x), \quad \ell = 1, \ldots, L$$

*converges weakly to an L-dimensional Gaussian vector* $(X_{f_1}, \ldots, X_{f_L})$ *with mean functions*

$$\mathrm{E}X_{f_\ell} = \frac{1}{2\pi}\int_a^b f_\ell'(x)\arg\left(1 - c\int\frac{t^2\underline{m}^2(x)}{(1+t\underline{m}(x))^2}dH(t)\right)dx + \frac{\beta_x}{2\pi}\int_a^b f_\ell'(x)\Im\left(1 - c\int\frac{\underline{m}^2(z)t^2dH(t)}{(1+t\underline{m}(z))^2}\right)\Big|_{z=x+0i}dx,$$

*and covariance functions*

$$\begin{aligned}\mathrm{Cov}(X_{f_{\ell_1}}, X_{f_{\ell_2}}) = {} & \frac{1}{2\pi^2}\int_{a-\epsilon}^{b+\epsilon}\int_a^b f_{\ell_1}'(x)f_{\ell_2}'(y)\log\left(1 + 4\frac{\underline{m}_\Im(x)\underline{m}_\Im(y)}{|\underline{m}(x) - \underline{m}(y)|^2}\right)dxdy \\ & + \frac{c\beta_x}{\pi^2}\int\left(\int_a^b\Im\left(\frac{f_{\ell_1}'(x)}{1+t(\underline{m}(x))}\right)dx\int_a^b\Im\left(\frac{f_{\ell_2}'(y)}{1+t(\underline{m}(y))}\right)dy\right)dH(t),\end{aligned}$$

*where* $f_\ell'$ *denotes the first derivative of* $f_\ell$*,* $\epsilon \to 0$*, and* $\Re(\cdot)$ *denotes the real part, and* $\underline{m}(x) = \lim_{z\to x}\underline{m}(z)$*,* $z \in \mathbb{C}^+ = \{z : z = x + \mathbf{i}y, y > 0\}$ *with* $\underline{m}_\Im(x) = \lim_{z\to x}\Im(\underline{m}(z))$ *indicating the limiting imaginary part of* $\underline{m}(x)$*, and*

$$z = -\frac{1}{\underline{m}(z)} + c\int\frac{t}{1+t\underline{m}(z)}dH(t).$$

The proof of the lemma is provided in the Appendix.

### 2.3. Power function and confidence interval

Consider a two-sided test, the rejection region is given by

$$\left\{ \mathbf{y}_1, \ldots, \mathbf{y}_n : \sum_{j=1}^{p} f(\hat{\lambda}_j) \geq r_1 \text{ or } \sum_{j=1}^{p} f(\hat{\lambda}_j) \leq r_2 \right\}, \tag{2}$$

where $r_1$ and $r_2$ are critical constants. For an arbitrary positive-definite matrix $\mathbf{\Sigma}_A$ under $H_A$, the theoretical power is given by

$$\begin{aligned} 1 - \beta(\mathbf{\Sigma}_A) &= 1 - \Pr\left( r_2 < \sum_{j=1}^{p} f(\hat{\lambda}_j) < r_1 \mid \mathbf{\Sigma}_A \right) \\ &= 1 - \Phi\left( \frac{r_1 - p\int_a^b f(x) dF^{c_{n-1}, H_p}(x) - \mathrm{E}X_f}{\sqrt{\mathrm{Cov}(X_f, X_f)}} \right) + \Phi\left( \frac{r_2 - p\int_a^b f(x) dF^{c_{n-1}, H_p}(x) - \mathrm{E}X_f}{\sqrt{\mathrm{Cov}(X_f, X_f)}} \right), \end{aligned} \tag{3}$$

where $\mathrm{E}X_f$, $p\int_a^b f(x) dF^{c_{n-1}, H_p}(x)$ and $\mathrm{Cov}(X_f, X_f)$ are evaluated under $H_A$. With an arbitrary covariance matrix $\mathbf{\Sigma}_A$, however, the power function cannot be explicitly expressed, since there are no closed forms for all quantities in (3).

Suppose that the population covariance matrix $\mathbf{\Sigma}(\boldsymbol{\rho})$ depends on some low dimensional parameters $\boldsymbol{\rho}$, where $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_k)^T$ is a $k$-dimensional parameter vector. The $100(1-\alpha)\%$ level confidence interval of the functional $g(\boldsymbol{\rho})$ can be easily constructed based on the relationship between hypothesis testing and interval estimation.

**Algorithm 1** (*Computation of the Theoretical Confidence Interval*)**.**

Step 1. Assume that $\rho_j$ lies inside $[\rho_j^{(1)}, \rho_j^{(2)}]$ for $j = 1, \ldots, k$. Divide $[\rho_j^{(1)}, \rho_j^{(2)}]$ as

$$\rho_j^{(1)} \equiv \rho_{j,0} < \rho_{j,1} < \cdots < \rho_{j,(N-1)} < \rho_{j,N} \equiv \rho_j^{(2)},$$

where

$$\rho_{j,l} = \rho_j^{(1)} + \frac{l}{N}(\rho_j^{(2)} - \rho_j^{(1)}),$$

for $j = 1, \ldots, k$. This partition results in a total of $kN$ grid points $\{\boldsymbol{\rho}_m, m = 1, \ldots, kN\}$, where $\boldsymbol{\rho}_m = (\rho_{1,l_1}, \ldots, \rho_{k,l_k})^T$, $1 \leq l_1, \ldots, l_k \leq N$.

Step 2. Based on the sample $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$, we obtain the observed LSS, $\sum_{j=1}^{p} f(\hat{\lambda}_j)$.

Step 3. Given $\boldsymbol{\rho}_m$, $m = 1, \ldots, kN$, then $[\widehat{g(\boldsymbol{\rho})}_L, \widehat{g(\boldsymbol{\rho})}_U]$ can be constructed as the $100(1-\alpha)\%$ confidence interval for $g(\boldsymbol{\rho})$, where

$$\widehat{g(\boldsymbol{\rho})}_U = \max_{\boldsymbol{\rho}_m} \left\{ g(\boldsymbol{\rho}_m) : \left| \frac{\sum_{j=1}^{p} f(\hat{\lambda}_j) - p\int_a^b f(x) dF^{c_{n-1}, H_p}(x) - \mathrm{E}_{\boldsymbol{\rho}_m} X_f}{\sqrt{\mathrm{Cov}_{\boldsymbol{\rho}_m}(X_f, X_f)}} \right| \leq z_{1-\alpha/2} \right\},$$

$$\widehat{g(\boldsymbol{\rho})}_L = \min_{\boldsymbol{\rho}_m} \left\{ g(\boldsymbol{\rho}_m) : \left| \frac{\sum_{j=1}^{p} f_i(\hat{\lambda}_j) - p\int_a^b f(x) dF^{c_{n-1}, H_p}(x) - \mathrm{E}_{\boldsymbol{\rho}_m} X_f}{\sqrt{\mathrm{Cov}_{\boldsymbol{\rho}_m}(X_f, X_f)}} \right| \leq z_{1-\alpha/2} \right\},$$

where $z_{1-\alpha/2}$ is the $100(1-\alpha/2)\%$ quantile of the standard normal distribution, $H_p$ is the ESD of $\mathbf{\Sigma}(\boldsymbol{\rho}_m)$, and the theoretical mean and variance terms are computed under $\mathbf{\Sigma}(\boldsymbol{\rho}_m)$.

All of the aforementioned formulae and algorithm involve the theoretical mean and variance terms of the LSS. In most scenarios, the covariance matrix is arbitrary and not ideally structured. As a result, there are usually no explicit expressions for these terms under a general covariance matrix setting, which makes the computations of power and confidence interval challenging. Such a problem would become more prominent when the dimension $p$ is relatively large in comparison to the sample size $n$.

### 2.4. Numerical algorithm for theoretical means and covariances

Based on Lemma 1, we can obtain the CLT of the LSS for sample covariance matrices by the Riemann integration instead of the contour integration, which enables us to compute the theoretical mean $p\int_a^b f_\ell(x)dF^{c_{n-1},H_p}(x) + \mathrm{E}X_{f_\ell}$ and theoretical covariance $\mathrm{Cov}(X_{f_{\ell_1}}, X_{f_{\ell_2}})$ numerically. As indicated by the lemma, the terms $p\int_a^b f_\ell(x)dF^{c_{n-1},H_p}(x)$, $\mathrm{E}X_{f_\ell}$ and $\mathrm{Cov}(X_{f_{\ell_1}}, X_{f_{\ell_2}})$ depend on the Stieltjes transform $\underline{m}(x)$. We first focus on the algorithm to compute $\underline{m}(x)$. Let $\underline{m}_\Re(z)$ and $\underline{m}_\Im(z)$ be the real and imaginary parts of $\underline{m}(z)$, respectively; that is, $\underline{m}(z) = \underline{m}_\Re(z) + \mathbf{i}\underline{m}_\Im(z)$. Based on Eq. (1), the Stieltjes transform $\underline{m}(x)$ is closely related to $H$, the limiting distribution of $H_p$. For any $z$, replacing $H$ by $H_p$ in (1), we have

$$\begin{aligned} z = x + \mathbf{i}y &\approx -\frac{1}{\underline{m}(z)} + \frac{c}{p}\sum_{j=1}^{p}\frac{\lambda_j}{1+\lambda_j\underline{m}(z)} \\ &= \frac{-\underline{m}_\Re(z) + \mathbf{i}\underline{m}_i(z)}{\underline{m}_\Re^2(z)+\underline{m}_\Im^2(z)} + \frac{1}{n}\sum_{j=1}^{p}\frac{\lambda_j + \lambda_j^2\underline{m}_\Re(z) - \mathbf{i}\lambda_j^2\underline{m}_\Im(z)}{(1+\lambda_j\underline{m}_\Re(z))^2+\lambda_j^2\underline{m}_\Im^2(z)} \\ &= \frac{-\underline{m}_\Re(z)}{\underline{m}_\Re^2(z)+\underline{m}_\Im^2(z)} + \frac{1}{n}\sum_{j=1}^{p}\frac{\lambda_j + \lambda_j^2\underline{m}_\Re(z)}{(1+\lambda_j\underline{m}_\Re(z))^2+\lambda_j^2\underline{m}_\Im^2(z)} \\ &\quad + \mathbf{i}\left(\frac{\underline{m}_\Im(z)}{\underline{m}_\Re^2(z)+\underline{m}_\Im^2(z)} + \frac{1}{n}\sum_{j=1}^{p}\frac{-\lambda_j^2\underline{m}_\Im(z)}{(1+\lambda_j\underline{m}_\Re(z))^2+\lambda_j^2\underline{m}_\Im^2(z)}\right). \end{aligned}$$

Accordingly, we obtain two equations as follows

$$\begin{cases} x \approx \dfrac{-\underline{m}_\Re(z)}{\underline{m}_\Re^2(z)+\underline{m}_\Im^2(z)} + \dfrac{1}{n}\displaystyle\sum_{j=1}^{p}\frac{\lambda_j + \lambda_j^2\underline{m}_\Re(z)}{(1+\lambda_j\underline{m}_\Re(z))^2+\lambda_j^2\underline{m}_\Im^2(z)}, \\ y \approx \dfrac{\underline{m}_\Im(z)}{\underline{m}_\Re^2(z)+\underline{m}_\Im^2(z)} + \dfrac{1}{n}\displaystyle\sum_{j=1}^{p}\frac{-\lambda_j^2\underline{m}_i(z)}{(1+\lambda_j\underline{m}_\Re(z))^2+\lambda_j^2\underline{m}_\Im^2(z)}, \end{cases} \tag{4}$$

which form the basis of numerical algorithms to compute the theoretical mean and covariance functions. When $z = x + \mathbf{i}y$ is given, the system of Eq. (4) can be treated as two nonlinear equations of $(\underline{m}_\Re(z), \underline{m}_\Im(z))$, which can be solved via some numerical procedure, for example, the Newton–Raphson method.

After obtaining the numerical values of $(\underline{m}_\Re(z), \underline{m}_\Im(z))$, the theoretical mean and covariance functions can be computed according to the following algorithm.

**Algorithm 2** (*Computation of the Theoretical Mean and Covariance Functions*)**.**

Step 1. Divide the support set $[a, b]$ as

$$a \equiv x_0 < x_1 = a + \frac{b-a}{N} < \cdots < x_{N-1} = a + \frac{(N-1)(b-a)}{N} < x_N \equiv b,$$

where $N$ is a large integer, e.g., $N = 10\,000$. The finer the partition, the more accurate the calculation of the theoretical mean and covariance functions.

Step 2. Given $z_s = x_s + i\epsilon_1$, the solution $(\underline{m}_\Re(x_s + i\epsilon_1), \underline{m}_\Im(x_s + i\epsilon_1))$ is obtained by solving the nonlinear equations (4), where $\epsilon_1$ is a small positive number, e.g., $\epsilon_1 = 10^{-3}$ for $s = 1, \ldots, N$. Similarly, $(\underline{m}_\Re(z_t), \underline{m}_\Im(z_t))$ can be computed for $z_t = x_t + i\epsilon_2$ for $t = 1, \ldots, N$, where $\epsilon_2 \neq \epsilon_1$.

Step 3. Based on the Riemann sum of the integral, we can approximate $\mathrm{E}X_{f_{\ell_1}}$ and $\mathrm{Cov}(X_{f_{\ell_1}}, X_{f_{\ell_2}})$ through

$$\begin{aligned} \mathrm{E}X_{f_\ell} &\approx \frac{b-a}{2\pi N}\sum_{s=1}^{N} f_\ell'(x_s)\arg\left(1 - \frac{c_n}{p}\sum_{j=1}^{p}\frac{\lambda_j^2\underline{m}_\Im^2(z_s)}{(1+\lambda_j\underline{m}_\Im(z_s))^2}\right) \\ &\quad - \frac{\beta_x(b-a)}{2\pi N}\sum_{s=1}^{N} f_\ell'(x_s)\Im\left(1 - \frac{c_n}{p}\sum_{j=1}^{p}\frac{\underline{m}^2(z_s)\lambda_j^2}{(1+\lambda_j\underline{m}(z_s))^2}\right) \\ &\triangleq \widehat{M}_\ell \end{aligned}$$

and

$$\begin{aligned}\mathrm{Cov}(X_{f_{\ell_1}}, X_{f_{\ell_2}}) &\approx \frac{(b-a)^2}{2\pi^2 N^2}\sum_{s,t=1}^{N} f'_{\ell_1}(x_s) f'_{\ell_2}(x_t)\log\left(1+4\frac{\underline{m}_{\mathfrak{I}}(z_s)\underline{m}_{\mathfrak{I}}(z_t)}{|\underline{m}(z_s)-\underline{m}(z_t)|^2}\right)\\ &\quad -\frac{c_n\beta_x(b-a)^2}{p\pi^2 N^2}\sum_{j=1}^{p}\Re\left(\sum_{s,t=1}^{N}\frac{f'_{\ell_1}(x_j)f'_{\ell_2}(x_t)}{\{1+\lambda_j\underline{m}(z_s)\}\{1+\lambda_j\underline{m}(z_t)\}}\right)\\ &\triangleq \widehat{\mathrm{Cov}}_{\ell_1\ell_2}\end{aligned}$$

where $\Re$ denotes the real part.

Step 4. The limiting spectral density $f^{c_{n-1},H_p}(x)$ of the LSD $F^{c_{n-1},H_p}(x)$ is approximated by

$$f^{c_{n-1},H_p}(x_s) \approx \frac{1}{c_{n-1}\pi}\underline{m}_{\mathfrak{I}}(z_s), \quad s=1,\ldots,N,$$

which leads to

$$\begin{aligned}\int_a^b f_\ell(x)dF^{c_{n-1},H_p}(x) &\approx \frac{b-a}{c_{n-1}\pi N}\sum_{s=1}^{N} f_\ell(x_s)\underline{m}_{\mathfrak{I}}(z_s)\\ &\triangleq \widehat{C}_\ell.\end{aligned}$$

Algorithm 2 facilitates the computations of power and confidence interval by substituting the numerical values of $\widehat{M}_l$, $\widehat{\mathrm{Cov}}_{l_1l_2}$, and $\widehat{C}_l$ into the power formula and confidence interval algorithm in Section 2.3. The following theorem shows that the approximated mean and covariance functions in Algorithm 2 converge respectively to their theoretical counterparts.

**Theorem 2.1.** *As $\epsilon_1, \epsilon_2 \to 0$, $N \to \infty$, $p \to \infty$ and $c_n \to c$, we have*

$$\widehat{M}_\ell \to \mathrm{E}X_{f_\ell}, \qquad \widehat{\mathrm{Cov}}_{\ell_1\ell_2} \to \mathrm{Cov}(X_{f_{\ell_1}}, X_{f_{\ell_2}}),$$

*and*

$$\widehat{C}_\ell - \int_a^b f_\ell(x)dF^{c_{n-1},H_p}(x) \to 0.$$

The proof of Theorem 2.1 is provided in the Appendix. This theorem guarantees the validity of the proposed algorithms for calculation of the power and confidence interval.

## 3. Numerical study

### 3.1. Likelihood ratio test for identity matrix

In the numerical study, we consider testing $H_0 : \mathbf{\Sigma} = \mathbf{I}_p$, the LRT statistic is

$$T_n = \sum_{j=1}^{p}\mathrm{tr}\mathbf{S}_n - \log|\mathbf{S}_n| - p = \sum_{j=1}^{p} f(\hat{\lambda}_j),$$

where $f(x) = x - \log x - 1$. Bai et al. (2009) established the explicit rejection region of $T_n$ under $H_0$, which is one-sided with

$$r_1 = p\left[1-\frac{c_{n-1}-1}{c_{n-1}}\log(1-c_{n-1})\right] - \frac{1}{2}\log(1-c_{n-1}) + \frac{1}{2}c_{n-1}\beta_x - z_\alpha\sqrt{-2c_{n-1}-2\log(1-c_{n-1})},$$

and $r_2 = -\infty$ in (2). For an arbitrary positive-definite matrix $\mathbf{\Sigma}_A$, the theoretical power is

$$1-\beta(\mathbf{\Sigma}_A) = 1-\Phi\left(\frac{r_1 - p\int_a^b f(x)dF^{c_{n-1},H_p}(x) - \mathrm{E}X_f}{\sqrt{\mathrm{Cov}(X_f, X_f)}}\right).$$

In addition, the theoretical local power function under the local alternative $\mathbf{\Sigma}_A = \mathbf{I}_p + \mathbf{I}_p/\sqrt{n}$ is

$$1-\beta(\mathbf{\Sigma}_A) = 1-\Phi\left(\frac{r_1-\mu_A}{\sigma_A}\right),$$

where

$$\mu_A = p\left[\sqrt{n}-1-\log(1+\sqrt{n}) - \left(1-\frac{1}{c_{n-1}}\right)\log(1-c_{n-1})\right] + \frac{\log(1-c_{n-1}) - c_{n-1}\beta_x}{2},$$

and $\sigma_A = \sqrt{-2\log(1-c_{n-1}) - 2c_{n-1} + (\beta_x+2)n^{-1}c_{n-1}}$. The detailed derivation is provided in the Appendix.

**Table 1**
Comparison between theoretical and empirical means and variances under the equicorrelation matrix $\mathbf{\Sigma} = ((1-\rho)\delta_{\{i=j\}} + \rho)_{i,j=1}^{p}$.

| $\rho$ | Theoretical | | Empirical | |
|---|---|---|---|---|
| | Mean | Variance | Mean | Variance |
| $(p, n) = (50, 500)$ | | | | |
| 0.050 | 3.943 | 0.035 | 3.916 | 0.036 |
| 0.075 | 4.958 | 0.066 | 4.920 | 0.065 |
| 0.100 | 6.078 | 0.101 | 6.028 | 0.111 |
| 0.125 | 7.278 | 0.164 | 7.221 | 0.159 |
| 0.150 | 8.548 | 0.231 | 8.489 | 0.231 |
| 0.175 | 9.878 | 0.311 | 9.813 | 0.320 |

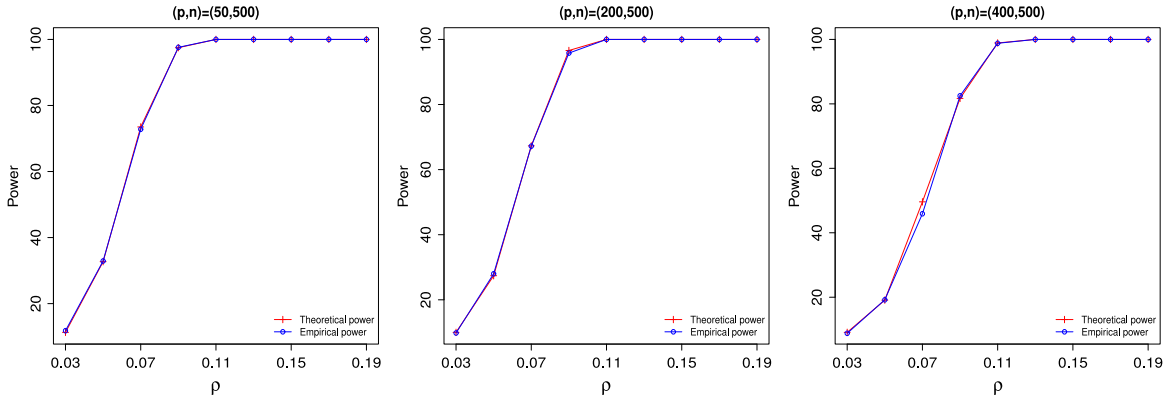


**Fig. 1.** Theoretical and empirical power curves for hypothesis testing on the banded matrix $\mathbf{\Sigma}_A = ((1-\rho)\delta_{\{i=j\}} + \rho\delta\{|i-j| \le 1\})_{i,j=1}^{p}$ with sample size $n = 500$.

For a more general case, if the local alternative is $\mathbf{\Sigma}_A = \mathbf{I}_p + \mathbf{\Omega}/\sqrt{n}$, where $\mathbf{\Omega}$ is an arbitrary matrix that makes $\mathbf{\Sigma}_A$ positive-definite, then the local power formula does not have a closed form. Instead, it can be computed by the proposed numerical method. Under this setting, the theoretical mean and variance terms in (3) should be computed based on $\mathbf{\Sigma}_A = \mathbf{I}_p + \mathbf{\Omega}/\sqrt{n}$.

### 3.2. Simulation study

In the simulation study, we consider the LRT as described in Section 3.1. We observe the sample $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$, which are independent and identically distributed from $N(\mathbf{0}_p, \mathbf{\Sigma})$. We consider three cases for $\mathbf{\Sigma}$: equicorrelation matrix $((1-\rho)\delta_{\{i=j\}} + \rho)_{i,j=1}^{p}$, banded matrix $((1-\rho)\delta_{\{i=j\}} + \rho\delta\{|i-j| \le d\})_{i,j=1}^{p}$ and AR(1) covariance matrix $(\rho^{|i-j|})_{i,j=1}^{p}$, where $\rho$ is the prespecified parameter that leads to a positive-definite matrix. We are interested in exploring the finite-sample performance of the theoretical mean and covariance in Lemma 1, the power function for hypothesis testing of $H_0$ and the confidence interval of the parameter $\boldsymbol{\rho} = \rho$ in $\mathbf{\Sigma}(\boldsymbol{\rho})$.

We first compare the theoretical mean and variance with their empirical counterparts, for which the theoretical values are computed by Algorithm 2, and the empirical ones are the sample mean and variance of $T_n$ from 10 000 replications. Table 1 presents the theoretical mean and variance, and their empirical counterparts under the equicorrelation matrix, for $n = 500$, $p = 50$, and $\rho$ ranging from 0.05 to 0.175. Table 2 provides the results under the AR(1) covariance matrix, for $n = 500$, $p = 50, 100, 300$, and $\rho$ ranging from 0.05 to 0.5. Table 3 corresponds to the banded matrix, for $n = 500$, $p = 50, 200, 400$, and $\rho$ ranging from 0.03 to 0.35. Tables 1–3 show that both the theoretical mean and variance by Algorithm 2 coincide with the empirical counterparts, even when the dimension $p$ is large compared to the sample size $n$.

To further evaluate the power function for hypothesis testing of $H_0$, Fig. 1 shows the theoretical and empirical power functions under the alternative banded matrix with the test size 10%. The theoretical power (3) is computed based on the theoretical mean and variance by Algorithm 2, while the empirical power is obtained from 10 000 simulations. We consider the dimension $p = 50, 200, 400$ and the sample size $n = 500$. Fig. 1 shows that for fixed $(n, p)$, the power increases with the value of $\rho$. On the other hand, there is a general trend that the power decreases with an increasing dimension $p$ for a fixed sample size. When $\rho > 0.1$, it is easy for the LRT to detect a banded matrix as the power is greater than 80% for all combinations of $(n, p)$ considered in the simulation study. Overall, Fig. 1 shows that the theoretical and the empirical power function are almost indistinguishable, which demonstrates the satisfactory approximation of the proposed numerical method.

Last, we examine the performance of the proposed algorithm for confidence interval, and only the result with an AR(1) covariance matrix is presented. In particular, we take the autoregressive parameter $\rho$ as an unknown parameter and compute the theoretical 90% confidence interval. For this evaluation, the dimension and the sample size are $(p, n) = (50, 500)$, $(100, 500)$, $(300, 500)$, $(100, 1000)$, $(250, 2500)$, and the correlation parameter $\rho$ varies from 0.05 to 0.50.

**Table 2**
Comparison between theoretical and empirical means and variances under the AR(1) covariance matrix $\mathbf{\Sigma} = (\rho^{|i-j|})_{i,j=1}^{p}$.

| $\rho$ | Theoretical | | Empirical | |
|---|---|---|---|---|
| | Mean | Variance | Mean | Variance |
| $(p, n) = (50, 500)$ | | | | |
| 0.05 | 2.773 | 0.012 | 2.762 | 0.012 |
| 0.10 | 3.153 | 0.015 | 3.132 | 0.015 |
| 0.15 | 3.783 | 0.020 | 3.756 | 0.020 |
| 0.20 | 4.688 | 0.027 | 4.641 | 0.027 |
| 0.25 | 5.853 | 0.037 | 5.803 | 0.037 |
| 0.30 | 7.263 | 0.049 | 7.226 | 0.050 |
| 0.35 | 9.043 | 0.065 | 9.051 | 0.064 |
| 0.40 | 11.188 | 0.085 | 11.183 | 0.086 |
| 0.45 | 13.730 | 0.109 | 13.724 | 0.108 |
| 0.50 | 16.739 | 0.140 | 16.734 | 0.140 |
| $(p, n) = (100, 500)$ | | | | |
| 0.05 | 11.127 | 0.048 | 11.100 | 0.049 |
| 0.10 | 11.872 | 0.054 | 11.850 | 0.055 |
| 0.15 | 13.139 | 0.064 | 13.108 | 0.067 |
| 0.20 | 14.940 | 0.079 | 14.896 | 0.080 |
| 0.25 | 17.270 | 0.099 | 17.244 | 0.100 |
| 0.30 | 20.193 | 0.124 | 20.193 | 0.126 |
| 0.35 | 23.793 | 0.156 | 23.780 | 0.160 |
| 0.40 | 28.123 | 0.196 | 28.121 | 0.196 |
| 0.45 | 33.253 | 0.246 | 33.247 | 0.252 |
| 0.50 | 39.397 | 0.308 | 39.340 | 0.318 |
| $(p, n) = (300, 500)$ | | | | |
| 0.05 | 117.949 | 0.632 | 117.949 | 0.627 |
| 0.10 | 120.163 | 0.651 | 120.192 | 0.657 |
| 0.15 | 123.974 | 0.683 | 124.005 | 0.696 |
| 0.20 | 129.403 | 0.729 | 129.389 | 0.741 |
| 0.25 | 136.483 | 0.789 | 136.493 | 0.794 |
| 0.30 | 145.395 | 0.867 | 145.396 | 0.874 |
| 0.35 | 156.253 | 0.963 | 156.279 | 0.963 |
| 0.40 | 169.332 | 1.083 | 169.344 | 1.083 |
| 0.45 | 184.852 | 1.237 | 184.876 | 1.241 |
| 0.50 | 203.202 | 1.424 | 203.225 | 1.437 |

**Table 3**
Comparison between theoretical and empirical means and variances under the banded matrix $\mathbf{\Sigma} = ((1-\rho)\delta_{\{i=j\}} + \rho\delta\{|i-j| \le 1\})_{i,j=1}^{p}$.

| $\rho$ | Theoretical | | Empirical | |
|---|---|---|---|---|
| | Mean | Variance | Mean | Variance |
| $(p, n) = (50, 500)$ | | | | |
| 0.05 | 2.762 | 0.012 | 2.762 | 0.012 |
| 0.10 | 3.136 | 0.015 | 3.137 | 0.015 |
| 0.15 | 3.781 | 0.019 | 3.782 | 0.019 |
| 0.20 | 4.727 | 0.026 | 4.728 | 0.027 |
| 0.25 | 6.032 | 0.035 | 6.033 | 0.036 |
| 0.30 | 7.788 | 0.046 | 7.794 | 0.045 |
| 0.35 | 10.166 | 0.058 | 10.174 | 0.060 |
| $(p, n) = (200, 500)$ | | | | |
| 0.05 | 47.498 | 0.236 | 47.506 | 0.230 |
| 0.10 | 49.054 | 0.244 | 49.028 | 0.239 |
| 0.15 | 51.690 | 0.257 | 51.640 | 0.258 |
| 0.20 | 55.541 | 0.276 | 55.486 | 0.289 |
| 0.25 | 60.807 | 0.303 | 60.801 | 0.326 |
| 0.30 | 67.952 | 0.339 | 67.961 | 0.362 |
| 0.35 | 77.659 | 0.387 | 77.664 | 0.425 |
| $(p, n) = (400, 500)$ | | | | |
| 0.05 | 240.856 | 1.573 | 240.865 | 1.656 |
| 0.10 | 243.891 | 1.610 | 243.924 | 1.686 |
| 0.15 | 249.147 | 1.648 | 249.170 | 1.732 |
| 0.20 | 256.845 | 1.715 | 256.871 | 1.794 |
| 0.25 | 267.498 | 1.785 | 267.515 | 1.829 |
| 0.30 | 281.958 | 1.895 | 281.893 | 1.939 |
| 0.35 | 301.667 | 2.071 | 301.376 | 2.025 |

**Table 4**
Average width and coverage rate of 90% confidence intervals for parameter $\rho$ of the AR(1) covariance matrix $\mathbf{\Sigma}(\rho) = (\rho^{|i-j|})_{i,j=1}^{p}$.

| $\rho$ | Confidence interval | Width | Coverage rate (%) |
|---|---|---|---|
| $(p, n) = (50, 500)$ | | | |
| 0.05 | [0.011, 0.076] | 0.065 | 89.79 |
| 0.10 | [0.076, 0.117] | 0.041 | 89.66 |
| 0.15 | [0.133, 0.163] | 0.030 | 88.27 |
| 0.20 | [0.186, 0.211] | 0.025 | 89.44 |
| 0.25 | [0.237, 0.260] | 0.023 | 89.62 |
| 0.30 | [0.289, 0.311] | 0.022 | 89.73 |
| 0.35 | [0.340, 0.360] | 0.020 | 89.76 |
| 0.40 | [0.390, 0.410] | 0.020 | 87.88 |
| 0.45 | [0.441, 0.459] | 0.018 | 87.44 |
| 0.50 | [0.491, 0.509] | 0.018 | 90.25 |
| $(p, n) = (100, 500)$ | | | |
| 0.05 | [0.010, 0.075] | 0.065 | 89.49 |
| 0.10 | [0.077, 0.116] | 0.039 | 89.03 |
| 0.15 | [0.135, 0.162] | 0.027 | 89.30 |
| 0.20 | [0.188, 0.210] | 0.022 | 89.38 |
| 0.25 | [0.240, 0.259] | 0.019 | 89.72 |
| 0.30 | [0.292, 0.308] | 0.017 | 87.75 |
| 0.35 | [0.342, 0.358] | 0.016 | 90.20 |
| 0.40 | [0.393, 0.407] | 0.016 | 87.60 |
| 0.45 | [0.443, 0.457] | 0.014 | 87.95 |
| 0.50 | [0.493, 0.506] | 0.013 | 89.66 |
| $(p, n) = (300, 500)$ | | | |
| 0.05 | [0.009, 0.082] | 0.073 | 90.49 |
| 0.10 | [0.073, 0.119] | 0.046 | 89.67 |
| 0.15 | [0.135, 0.164] | 0.029 | 88.14 |
| 0.20 | [0.189, 0.210] | 0.021 | 89.44 |
| 0.25 | [0.241, 0.260] | 0.019 | 90.05 |
| 0.30 | [0.293, 0.307] | 0.014 | 87.99 |
| 0.35 | [0.344, 0.356] | 0.012 | 87.13 |
| 0.40 | [0.395, 0.406] | 0.011 | 87.74 |
| 0.45 | [0.445, 0.455] | 0.010 | 88.20 |
| 0.50 | [0.496, 0.504] | 0.008 | 89.71 |

**Table 5**
Average width and coverage rate of 90% confidence intervals for parameter $\rho$ of the AR(1) covariance matrix $\mathbf{\Sigma}(\rho) = (\rho^{|i-j|})_{i,j=1}^{p}$ with $p/n = 0.1$.

| $\rho$ | Confidence interval | Width | Coverage rate (%) |
|---|---|---|---|
| $(p, n) = (100, 1000)$ | | | |
| 0.05 | [0.025, 0.065] | 0.040 | 89.72 |
| 0.10 | [0.090, 0.109] | 0.019 | 90.04 |
| 0.15 | [0.143, 0.157] | 0.014 | 89.97 |
| 0.20 | [0.194, 0.206] | 0.012 | 90.07 |
| 0.25 | [0.244, 0.255] | 0.011 | 90.13 |
| 0.30 | [0.295, 0.305] | 0.010 | 89.95 |
| 0.35 | [0.345, 0.355] | 0.010 | 89.76 |
| 0.40 | [0.395, 0.405] | 0.010 | 90.15 |
| 0.45 | [0.446, 0.455] | 0.009 | 89.75 |
| 0.50 | [0.495, 0.503] | 0.008 | 89.65 |
| $(p, n) = (250, 2500)$ | | | |
| 0.05 | [0.044, 0.056] | 0.012 | 89.22 |
| 0.10 | [0.097, 0.104] | 0.007 | 89.52 |
| 0.15 | [0.147, 0.153] | 0.006 | 90.30 |
| 0.20 | [0.198, 0.203] | 0.005 | 89.80 |
| 0.25 | [0.248, 0.252] | 0.004 | 90.12 |
| 0.30 | [0.298, 0.302] | 0.004 | 89.75 |
| 0.35 | [0.348, 0.352] | 0.004 | 90.20 |
| 0.40 | [0.398, 0.402] | 0.004 | 90.13 |
| 0.45 | [0.448, 0.452] | 0.004 | 89.95 |
| 0.50 | [0.499, 0.502] | 0.003 | 90.01 |

Tables 4 and 5 present the average lower and upper boundaries, width and empirical coverage rate of the confidence intervals from 10 000 replications. It can be seen that the average width of the 90% confidence interval decreases with the increasing $\rho$ for fixed $(n, p)$. On the other hand, if the sample size $n$ is fixed, the average width increases with $p$. In addition, when the ratio of dimension-to-sample size $p/n$ is fixed, the width of confidence interval shrinks as the sample size increases. For example, the widths of confidence intervals when $(p, n) = (100, 1000)$ are approximately half of those for $(p, n) = (50, 500)$. Despite

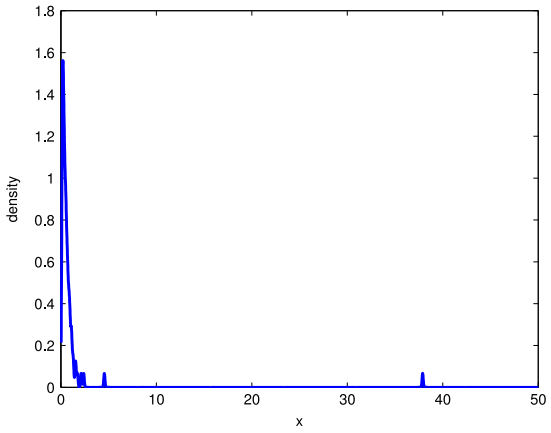


**Fig. 2.** Kernel estimation of LSD based on the sample coefficient correlation matrix of the S&P 100 index data.

different values of $p$, $n$ and $\rho$, the coverage rates under various configurations are very close to the nominal level 90%, which shows the validity of the proposed method.

## 4. Real data example

To further demonstrate the practical utility of the proposed method, we apply our algorithm to analyze the correlation coefficient matrix of stock returns. The correlation coefficient matrix of stock returns plays an important role in financial applications. For example, the plug-in approach to the high-dimensional Markowitz optimization problem based on the sample mean and covariance estimators performs poorly in practice (Michaud, 1989). This is mainly due to the fact that the sample covariance matrix is a poor estimator of the population covariance matrix in the high-dimensional setting. Other than the common approaches to deriving a point estimate, our method can make inference about the population matrix based on some interval estimates.

Specifically, we investigate a set of US stocks included in the S&P 100 index (OEX index) from November 21, 2011 to November 18, 2013 (501 trading days). Let $P_{ji}$ denote the stock closing price of the $j$th stock at day $i$. We compute the log-return of each stock, which is the first-lag difference of the logarithm of the stock prices

$$y_{ji} = \log(P_{ji}) - \log(P_{j,i-1}), \quad i = 1, \ldots, 500,$$

where $P_{j0}$ is the stock closing price on November 21, 2011. After eliminating the stocks with missing values, we obtain a total of $p = 95$ stocks with the sample size $n = 500$. In this example, we assume that the log-returns of each stock are independent, and we are interested in estimating the correlation among the stocks of the S&P 100 index. Although the independence assumption may not be realistic, it may serve as a reasonable first-cut approximation (Ledoit and Wolf, 2003).

Let $\mathbf{S}_n$ be the sample correlation matrix of $\{\mathbf{y}_1, \ldots, \mathbf{y}_{500}\}$ where $\mathbf{y}_i = (y_{1i}, \ldots, y_{95i})^T$. Fig. 2 exhibited the estimated LSD based on the kernel smoothing method (Jing et al., 2010). It is obvious that there exists a spike eigenvalue, which is around the point of 38.5, in the LSD of the S&P 100 index correlation matrix. Among the existing standard matrix structures, the LSD of an equicorrelation matrix, as defined by all the off-diagonal elements of the correlation matrix having the same correlation, also has one spike. Therefore, we assume that the stocks from S&P 100 index are equicorrelated, and let $\rho$ represent the correlation coefficient. Such equicorrelation structures have been commonly utilized in estimating the covariance matrix of financial data (Engle and Kelly, 2012).

We are interested in obtaining the 90% confidence interval of $\rho$. To illustrate the proposed Algorithms 1 and 2, we take $\rho = 0.45$ as an example, and examine whether $\rho = 0.45$ lies inside the 90% confidence interval. For this purpose, we need to consider testing

$$H_0 : \rho = 0.45, \quad \text{versus} \quad H_A : \rho \neq 0.45.$$

Based on $\mathbf{S}_n$, the observed LRT statistic is

$$T_n = \sum_{j=1}^{p} \text{tr}\mathbf{S}_n - \log|\mathbf{S}_n| - p = 72.45,$$

where $p = 95$ and $n = 500$. Under the null hypothesis $H_0 : \rho = 0.45$, the theoretical mean and variance of $T_n$ computed by Algorithm 2 are 62.2 and 7.45, respectively. As a result, the observed LRT statistic is outside of the acceptance region of $H_0$.

The aforementioned procedure is repeated for a grid of points in (0, 1). Finally, the 90% confidence interval for $\rho$ is [0.482, 0.534], which demonstrates that the stocks in the S&P 100 index have a large correlation, and the correlation cannot be ignored.

## 5. Concluding remarks

The central limit theorem has recently been developed for the linear spectral statistics of high-dimensional sample covariance matrices. To evaluate the power of hypothesis testing, we have developed the numerical algorithms to compute the theoretical means and variances, and the power function and confidence interval can be derived as well for high-dimensional data. Simulation studies show that the numerical algorithms behave well. As the proposed method can deal with general covariance matrices, it can be generalized to more complicated high-dimensional problems.

## Acknowledgments

We thank the Editor, the Associate Editor and the referee for their helpful comments that have led to substantial improvements of the paper. The research was supported in part by NSFC (11522015) and Fundamental Research Funds for the Central Universities (Shurong Zheng); The research was supported in part by a grant (17125814) from the Research Grants Council of Hong Kong (Guosheng Yin).

## Appendix

### A.1. Proof of Lemma 1

Based on the work of Zheng et al. (2015), it is demonstrated that $\{G_p(f_\ell), \ell = 1, \ldots, k\}$ converges weakly to a Gaussian vector $(X_{f_1}, \ldots, X_{f_L})$ with mean functions

$$\mathrm{EX}_{f_\ell} = -\frac{1}{2\pi \mathbf{i}} \oint_{\mathcal{C}} f_\ell(z) \frac{c \int \underline{m}^3(z) t^2 (1 + t\underline{m}(z))^{-3} dH(t)}{\left[1 - c \int \underline{m}^2(z) t^2 (1 + t\underline{m}(z))^{-2} dH(t)\right]^2} dz - \frac{\beta_x}{2\pi \mathbf{i}} \oint_{\mathcal{C}} f_\ell(z) \frac{c \int \frac{t^2 \underline{m}^3(z)}{(\underline{m}(z)t+1)^3} dH(t)}{1 - c \int \frac{\underline{m}^2(z) t^2 dH(t)}{(1+t\underline{m}(z))^2}} dz,$$

and covariance functions

$$\begin{aligned}\mathrm{Cov}(X_{f_{\ell_1}}, X_{f_{\ell_2}}) = & -\frac{1}{2\pi^2} \oint_{\mathcal{C}_1} \oint_{\mathcal{C}_2} \frac{f_{\ell_1}(z_1) f_{\ell_2}(z_2)}{(\underline{m}(z_1) - \underline{m}(z_2))^2} d\underline{m}(z_1) d\underline{m}(z_2) \\ & - \frac{c\beta_x}{4\pi^2} \oint_{\mathcal{C}} \oint_{\mathcal{C}} f_{\ell_1}(z_1) f_{\ell_2}(z_2) \left[ \int \frac{t}{(\underline{m}(z_1)t + 1)^2} \frac{t}{(\underline{m}(z_2)t + 1)^2} dH(t) \right] d\underline{m}(z_1) d\underline{m}(z_2).\end{aligned}$$

Bai and Silverstein (2004) showed that

$$-\frac{1}{2\pi \mathbf{i}} \oint_{\mathcal{C}} f_\ell(z) \frac{c \int \underline{m}^3(z) t^2 (1 + t\underline{m}(z))^{-3} dH(t)}{\left[1 - c \int \underline{m}^2(z) t^2 (1 + t\underline{m}(z))^{-2} dH(t)\right]^2} dz = \frac{1}{2\pi} \int_a^b f'_\ell(x) \arg\left(1 - c \int \frac{t^2 \underline{m}^2(x)}{(1 + t\underline{m}(x))^2} dH(t)\right) dx$$

and

$$-\frac{1}{2\pi^2} \oint_{\mathcal{C}_1} \oint_{\mathcal{C}_2} \frac{f_{\ell_1}(z_1) f_{\ell_2}(z_2)}{(\underline{m}(z_1) - \underline{m}(z_2))^2} d\underline{m}(z_1) d\underline{m}(z_2) = \frac{1}{2\pi^2} \int_{a-\epsilon}^{b+\epsilon} \int_a^b f'_{\ell_1}(x) f_{\ell_2}(y) \ln\left(1 + 4 \frac{\underline{m}_\Im(x) \underline{m}_\Im(y)}{|\underline{m}(x) - \underline{m}(y)|^2}\right) dx dy.$$

Moreover, we have

$$\begin{aligned}-\frac{\beta_x}{2\pi \mathbf{i}} \oint_{\mathcal{C}} f_\ell(z) \frac{c \int \frac{t^2 \underline{m}^3(z)}{(\underline{m}(z)t+1)^3} dH(t)}{1 - c \int \frac{\underline{m}^2(z) t^2 dH(t)}{(1+t\underline{m}(z))^2}} dz &= \frac{\beta_x}{4\pi \mathbf{i}} \oint_{\mathcal{C}} f_\ell(z) d\left(1 - c \int \frac{\underline{m}^2(z) t^2 dH(t)}{(1 + t\underline{m}(z))^2}\right) \\ &= -\frac{\beta_x}{4\pi \mathbf{i}} \oint_{\mathcal{C}} f'_\ell(z) \left(1 - c \int \frac{\underline{m}^2(z) t^2 dH(t)}{(1 + t\underline{m}(z))^2}\right) dz \\ &= \frac{\beta_x}{2\pi} \int_a^b f'_\ell(x) \Im\left(1 - c \int \frac{\underline{m}^2(z) t^2 dH(t)}{(1 + t\underline{m}(z))^2}\right)\Big|_{z=x+0i} dx\end{aligned}$$

and

$$\begin{aligned}&-\frac{c\beta_x}{4\pi^2} \oint_{\mathcal{C}} \oint_{\mathcal{C}} f_{\ell_1}(z_1) f_{\ell_2}(z_2) \left( \int \frac{t}{(\underline{m}(z_1)t + 1)^2} \frac{t}{(\underline{m}(z_2)t + 1)^2} dH(t) \right) d\underline{m}(z_1) d\underline{m}(z_2) \\ &= -\frac{c\beta_x}{4\pi^2} \int \left( \oint \frac{f'_{\ell_1}(z_1)}{1 + t\underline{m}(z_1)} dz_1 \int \oint \frac{f'_{\ell_2}(z_2)}{1 + t\underline{m}(z_2)} dz_2 \right) dH(t) \\ &= \frac{c\beta_x}{\pi^2} \int \left( \int_a^b \Im\left(\frac{f'_{\ell_1}(x)}{1 + t(\underline{m}(x))}\right) dx \int_a^b \Im\left(\frac{f'_{\ell_2}(y)}{1 + t(\underline{m}(y))}\right) dy \right) dH(t).\end{aligned}$$

Therefore, the proof of Lemma 1 is completed.

### A.2. Convergence of Algorithm 1

We only consider the first term in the mean function, i.e., to show that the numerical estimate converges to

$$(2\pi)^{-1}\int_a^b f_\ell'(x)\arg\left(1-c\int\frac{t^2\underline{m}^2(x)}{(1+t\underline{m}(x))^2}dH(t)\right)dx.$$

The remaining part can be proved with similar argument. Because $H$ is the limiting distribution of $H_p$, we have

$$\begin{aligned}&-\frac{1}{2\pi\mathbf{i}}\oint_{\mathcal{C}} f_\ell(z)\frac{c\int\underline{m}^3(z)t^2(1+t\underline{m}(z))^{-3}dH_p(t)}{\left[1-c\int\underline{m}^2(z)t^2(1+t\underline{m}(z))^{-2}dH_p(t)\right]^2}dz\\&\to-\frac{1}{2\pi\mathbf{i}}\oint_{\mathcal{C}} f_\ell(z)\frac{c\int\underline{m}^3(z)t^2(1+t\underline{m}(z))^{-3}dH(t)}{\left[1-c\int\underline{m}^2(z)t^2(1+t\underline{m}(z))^{-2}dH(t)\right]^2}dz.\end{aligned}$$

Moreover, we have

$$-\frac{1}{2\pi\mathbf{i}}\oint_{\mathcal{C}} f_\ell(z)\frac{c\int\underline{m}^3(z)t^2(1+t\underline{m}(z))^{-3}dH_p(t)}{\left[1-c\int\underline{m}^2(z)t^2(1+t\underline{m}(z))^{-2}dH_p(t)\right]^2}dz=\frac{1}{2\pi}\int_a^b f_\ell'(x)\arg\left(1-c\int\frac{t^2\underline{m}^2(x)}{(1+t\underline{m}(x))^2}dH_p(t)\right)dx,$$

and then

$$\begin{aligned}&\int_a^b f_\ell'(x)\arg\left(1-c\int\frac{t^2\underline{m}^2(x+\mathbf{i}\epsilon_1)}{(1+t\underline{m}(x+\mathbf{i}\epsilon_1))^2}dH_p(t)\right)dx\\&=\sum_{s=1}^N\int_{x_{s-1}}^{x_s} f_\ell'(x)\arg\left(1-c\int\frac{t^2\underline{m}^2(z_s)}{(1+t\underline{m}(z_s))^2}dH_p(t)\right)dx\\&=\frac{b-a}{N}\sum_{s=1}^N f_\ell'(\tilde{x}_s)\arg\left(1-c\int\frac{t^2\underline{m}^2(\tilde{z}_s)}{(1+t\underline{m}(\tilde{z}_s))^2}dH_p(t)\right)\\&=\frac{b-a}{N}\sum_{s=1}^N f_\ell'(x_s)\arg\left(1-\frac{c}{p}\sum_{j=1}^p\frac{\lambda_j^2\underline{m}_{\underline{\mathcal{I}}}^2(z_s)}{(1+\lambda_j\underline{m}_{\underline{\mathcal{I}}}(z_s))^2}\right)\\&\quad+\frac{b-a}{N}\sum_{s=1}^N(f_\ell'(\tilde{x}_s)-f_\ell'(x_s))\arg\left(1-\frac{c}{p}\sum_{j=1}^p\frac{\lambda_j^2\underline{m}_{\underline{\mathcal{I}}}^2(\tilde{z}_s)}{(1+\lambda_j\underline{m}_{\underline{\mathcal{I}}}(\tilde{z}_s))^2}\right)\\&\quad+\frac{b-a}{N}\sum_{s=1}^N f_\ell'(x_s)\left[\arg\left(1-\frac{c}{p}\sum_{j=1}^p\frac{\lambda_j^2\underline{m}_{\underline{\mathcal{I}}}^2(\tilde{z}_s)}{(1+\lambda_j\underline{m}_{\underline{\mathcal{I}}}(\tilde{z}_s))^2}\right)-\arg\left(1-\frac{c}{p}\sum_{j=1}^p\frac{\lambda_j^2\underline{m}_{\underline{\mathcal{I}}}^2(z_s)}{(1+\lambda_j\underline{m}_{\underline{\mathcal{I}}}(z_s))^2}\right)\right]\end{aligned}$$

where $z_s=x_s+\mathbf{i}\epsilon_1$ and $\tilde{z}_s=\tilde{x}_s+\mathbf{i}\epsilon_1$, and $\tilde{x}_s$ lies between $x_{s-1}$ and $x_s$. Since $f_\ell'$ is analytic in $[a,b]$, we have

$$|f_\ell'(\tilde{x}_s)-f_\ell'(x_s)|\le K|\tilde{x}_s-x_s|\le K\frac{b-a}{N},$$

where $K$ is a constant. Consequently, we obtain

$$\left|\frac{b-a}{N}\sum_{s=1}^N(f_\ell'(\tilde{x}_s)-f_\ell'(x_s))\arg\left(1-\frac{c}{p}\sum_{j=1}^p\frac{\lambda_j^2\underline{m}_{\underline{\mathcal{I}}}^2(\tilde{z}_s)}{(1+\lambda_j\underline{m}_{\underline{\mathcal{I}}}(\tilde{z}_s))^2}\right)\right|\le\frac{2\pi K(b-a)^2}{N},$$

which converges to 0 as $N\to\infty$. Moreover, by the definition of the Stieltjes transform, we have

$$|\underline{m}_{\underline{\mathcal{I}}}(\tilde{z}_s)-\underline{m}_{\underline{\mathcal{I}}}(z_s)|\le K\frac{b-a}{N}\quad\text{and}\quad|\underline{m}_{\underline{\mathcal{I}}}^2(\tilde{z}_s)-\underline{m}_{\underline{\mathcal{I}}}^2(z_s)|\le K\frac{b-a}{N}.$$

Thus we have

$$\begin{aligned}&\left|\arg\left(1-\frac{c}{p}\sum_{j=1}^p\frac{\lambda_j^2\underline{m}_{\underline{\mathcal{I}}}^2(\tilde{z}_s)}{(1+\lambda_j\underline{m}_{\underline{\mathcal{I}}}(\tilde{z}_s))^2}\right)-\arg\left(1-\frac{c}{p}\sum_{j=1}^p\frac{\lambda_j^2\underline{m}_{\underline{\mathcal{I}}}^2(z_s)}{(1+\lambda_j\underline{m}_{\underline{\mathcal{I}}}(z_s))^2}\right)\right|\\&\le c\left|\frac{1}{p}\sum_{j=1}^p\frac{\lambda_j^2\underline{m}_{\underline{\mathcal{I}}}^2(\tilde{z}_s)}{(1+\lambda_j\underline{m}_{\underline{\mathcal{I}}}(\tilde{z}_s))^2}-\frac{1}{p}\sum_{j=1}^p\frac{\lambda_j^2\underline{m}_{\underline{\mathcal{I}}}^2(z_s)}{(1+\lambda_j\underline{m}_{\underline{\mathcal{I}}}(z_s))^2}\right|\\&\le\frac{c}{p}\sum_{j=1}^p\left|\frac{\lambda_j^2\underline{m}_{\underline{\mathcal{I}}}^2(\tilde{z}_s)}{(1+\lambda_j\underline{m}_{\underline{\mathcal{I}}}(\tilde{z}_s))^2}-\frac{\lambda_j^2\underline{m}_{\underline{\mathcal{I}}}^2(z_s)}{(1+\lambda_j\underline{m}_{\underline{\mathcal{I}}}(z_s))^2}\right|\end{aligned}$$

$$\leq \frac{c}{p}\sum_{j=1}^{p}\lambda_j^2\left|\frac{\underline{m}_{\Im}^2(\tilde{z}_s)-\underline{m}_{\Im}^2(z_s)}{(1+\lambda_j\underline{m}_{\Im}(\tilde{z}_s))^2}\right|+\frac{c}{p}\sum_{j=1}^{p}\lambda_j^2\underline{m}_{\Im}^2(z_s)\left|\frac{(1+\lambda_j\underline{m}_{\Im}(\tilde{z}_s))^2-(1+\lambda_j\underline{m}_{\Im}(z_s))^2}{(1+\lambda_j\underline{m}_{\Im}(\tilde{z}_s))^2(1+\lambda_j\underline{m}_{\Im}(z_s))^2}\right|$$

$$\leq \frac{(b-a)K}{pN}\sum_{j=1}^{p}\lambda_j^2\to 0,\quad \text{as } N\to\infty,$$

where $\underline{m}_{\Im}(z)>0$ in $\mathcal{C}^+=\{z:\Im(z)>0\}$, $\mathbf{\Sigma}$ is bounded in the spectral norm. Then, it can be shown that

$$\left|\frac{1}{N}\sum_{s=1}^{N}f_\ell'(x_s)\left[\arg\left(1-\frac{c}{p}\sum_{j=1}^{p}\frac{\lambda_j^2\underline{m}_{\Im}^2(\tilde{z}_s)}{(1+\lambda_j\underline{m}_{\Im}(\tilde{z}_s))^2}\right)-\arg\left(1-\frac{c}{p}\sum_{j=1}^{p}\frac{\lambda_j^2\underline{m}_{\Im}^2(z_s)}{(1+\lambda_j\underline{m}_{\Im}(z_s))^2}\right)\right]\right|\to 0,\quad \text{as } N\to\infty.$$

Therefore, we obtain

$$\left|\int_a^b f_\ell'(x)\arg\left(1-c\int\frac{t^2\underline{m}^2(x+\mathbf{i}\epsilon_1)}{(1+t\underline{m}(x+\mathbf{i}\epsilon_1))^2}dH(t)\right)dx\right.$$
$$\left.-\frac{b-a}{N}\sum_{s=1}^{N}f_\ell'(x_s)\arg\left(1-\frac{c}{p}\sum_{j=1}^{p}\frac{\lambda_j^2\underline{m}_{\Im}^2(z_s)}{(1+\lambda_j\underline{m}_{\Im}(z_s))^2}\right)\right|\to 0,$$

as $N\to\infty$. Furthermore, we have

$$\left|\int_a^b f_\ell'(x)\arg\left(1-c\int\frac{t^2\underline{m}^2(x+\mathbf{i}\epsilon_1)}{(1+t\underline{m}(x+\mathbf{i}\epsilon_1))^2}dH(t)\right)dx-\int_a^b f_\ell'(x)\arg\left(1-c\int\frac{t^2\underline{m}^2(x)}{(1+t\underline{m}(x))^2}dH(t)\right)dx\right|\to 0$$

as $\epsilon_1\to 0$. Thus, we have

$$\frac{b-a}{N}\sum_{s=1}^{N}f_\ell'(x_s)\arg\left(1-\frac{c}{p}\sum_{j=1}^{p}\frac{\lambda_j^2\underline{m}_{\Im}^2(z_s)}{(1+\lambda_j\underline{m}_{\Im}(z_s))^2}\right)\to\int_a^b f_\ell'(x)\arg\left(1-c\int\frac{t^2\underline{m}^2(x)}{(1+t\underline{m}(x))^2}dH(t)\right)dx$$

as $N\to\infty$, $\epsilon_1\to 0$ and $p\to\infty$.

### *A.3. Theoretical local power for $H_0:\mathbf{\Sigma}=\mathbf{I}_p$*

If $\mathbf{S}_n$ is the sample covariance matrix under the alternative hypothesis where $\mathbf{\Sigma}_A=\mathbf{I}_p+\mathbf{I}_p/\sqrt{n}$, and $\{\hat{\lambda}_1\leq\cdots\leq\hat{\lambda}_p\}$ are the eigenvalues of $\mathbf{S}_n$, then

$$\mathrm{tr}\mathbf{S}_n=\sum_{j=1}^{p}\hat{\lambda}_j=(1+n^{-1/2})\sum_{j=1}^{p}\hat{\lambda}_j/(1+n^{-1/2}),$$

where $\{\hat{\lambda}_j/(1+n^{-1/2}),j=1,\ldots,p\}$ can be treated as the sample eigenvalues with the population covariance matrix $\mathbf{I}_p$. Moreover,

$$\log|\mathbf{S}_n|=\sum_{j=1}^{p}\log\hat{\lambda}_j=p\log(1+n^{-1/2})+\sum_{j=1}^{p}\log\hat{\lambda}_j/(1+n^{-1/2}).$$

Then, the LRT statistic is given by

$$T_n=(1+n^{-1/2})\sum_{j=1}^{p}\hat{\lambda}_j/(1+n^{-1/2})-\sum_{j=1}^{p}\log\hat{\lambda}_j/(1+n^{-1/2})-p-p\log(1+n^{-1/2}).$$

Under the alternative hypothesis $\mathbf{\Sigma}_A=\mathbf{I}_p+\mathbf{I}_p/\sqrt{n}$, it can be shown that

$$\left(\sum_{j=1}^{p}\frac{\hat{\lambda}_j}{1+n^{-1/2}}-p,\sum_{j=1}^{p}\log\frac{\hat{\lambda}_j}{1+n^{-1/2}}-\frac{p(c_{n-1}-1)}{c_{n-1}}\log(1-c_{n-1})+p\right)^T$$
$$\to N\left(\begin{pmatrix}0\\ \frac{1}{2}\log(1-c)-\frac{1}{2}c\beta_x\end{pmatrix},\begin{pmatrix}(\beta_x+2)c & (\beta_x+2)c\\ (\beta_x+2)c & -2\log(1-c)+\beta_x c\end{pmatrix}\right).$$

Based on the delta method,

$$\frac{T_n-\mu_A}{\sigma_A}\to N(0,1)\text{ under }H_A,$$

where

$$\mu_A = p\left[\sqrt{n} - 1 - \log(1+\sqrt{n}) - \left(1 - \frac{1}{c_{n-1}}\right)\log(1-c_{n-1})\right] + \frac{\log(1-c_{n-1}) - c_{n-1}\beta_x}{2},$$

and $\sigma_A = \sqrt{-2\log(1-c_{n-1}) - 2c_{n-1} + (\beta_x+2)n^{-1}c_{n-1}}$. Consequently, the theoretical local power function is

$$1 - \beta(\mathbf{\Sigma}_A) = 1 - \Phi\left(\frac{r_1 - \mu_A}{\sigma_A}\right).$$

## References


Anderson, T.W., 2003. An Introduction to Multivariate Statistical Analysis, third ed. Wiley, New York.
Bai, Z.D., Jiang, D.D., Yao, J.F., Zheng, S.R., 2009. Corrections to LRT on large-dimensional covariance matrix by RMT. Ann. Statist. 37, 3822–3840.
Bai, Z.D., Silverstein, J.W., 2004. CLT for linear spectral statistics of large-dimensional sample covariance matrices. Ann. Probab. 32, 553–605.
Bai, Z.D., Yin, Y.Q., 1993. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. Ann. Probab. 21, 1275–1294.
Chen, S.X., Zhang, L.X., Zhong, P.S., 2012. Tests for high-dimensional covariance matrices. J. Amer. Statist. Assoc. 105, 810–819.
Engle, R., Kelly, B., 2012. Dynamic equicorrelation. J. Bus. Econom. Statist. 30, 212–228.
Jing, B.Y., Pan, G.M., Shao, Q.M., Zhou, W., 2010. Nonparametric estimate of spectral density functions of sample covariance matrices: A first step. Ann. Statist. 38, 3724–3750.
Kyj, L., Ostdiek, B., Ensor, K., 2009. Realized Covariance Estimation in Dynamic Portfolio Optimization. Tech. Rep., AFA 2010 Atlanta Meetings Paper.
Ledoit, O., Wolf, M., 2003. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. J. Empir. Finance 10, 603–621.
Ledoit, O., Wolf, M., 2004. A well-conditioned estimator for large-dimensional covariance matrices. J. Multivariate Anal. 88, 365–411.
Marčenko, V.A., Pastur, L.A., 1967. Distribution of eigenvalues for some sets of random matrices. Math. USSR-Sb. 1, 457–483.
Michaud, R.O., 1989. The markowitz optimization enigma: is 'optimized' optimal? Financ. Anal. J. 45, 31–42.
Onatski, A., Moreira, M.J., Hallin, M., 2013. Asymptotic power of sphericity tests for high-dimensional data. Ann. Statist. 41, 1204–1231.
Wang, C., 2014. Asymptotic power of likelihood ratio tests for high dimensional data. Statist. Probab. Lett. 88, 184–189.
Zheng, S.R., Bai, Z.D., Yao, J.F., 2015. Substitution principle for CLT of linear spectral statistics of high-dimensional sample covariance matrices with applications to hypothesis testing. Ann. Statist. 43, 546–591.