Bernoulli **29**(2), 2023, 984–1006
https://doi.org/10.3150/22-BEJ1487

Central limit theorem of linear spectral statistics of high-dimensional sample correlation matrices

YANQING YIN[1,a], SHURONG ZHENG[2,b] and TINGTING ZOU[3,c]

[1]*Chongqing University, Chongqing, China,* [a]*yinyq799@nenu.edu.cn*
[2]*Northeast Normal University, Changchun, China,* [b]*zhengsr@nenu.edu.cn*
[3]*Jilin University, Changchun, China,* [c]*zoutt260@nenu.edu.cn*

A high-dimensional sample correlation matrix is an important random matrix in multivariate statistical analysis. Its central limit theory is one of the main theoretical bases for making statistical inferences on high-dimensional correlation matrices. Under the high-dimensional framework in which the data dimension tends to infinity proportionally with the sample size, we establish the central limit theorems (CLT) for the linear spectral statistics (LSS) of sample correlation matrices in two settings: (1) the population follows an independent component structure; (2) the population follows an elliptical structure, including some heavy-tailed distributions. The results show that the CLTs of the LSS of the sample correlation matrices are very different in the two settings. In particular, even if the population correlation matrix is an identity matrix, the CLTs are different in the two settings. An application of our two established CLTs is provided.

Keywords: Sample correlation matrix; high-dimensional; independent component structure; elliptical distribution; central limit theorem; random matrix theory

1. Introduction

A correlation matrix is an important matrix in multivariate statistical analysis. Suppose that $\mathbf{y}_1,\ldots,\mathbf{y}_n$ are *independent and identically distributed* (i.i.d.) from a p-dimensional population $\mathbf{y}$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The population correlation matrix is defined as

$$\mathbf{R} = [\operatorname{diag}(\boldsymbol{\Sigma})]^{-1/2}\boldsymbol{\Sigma}[\operatorname{diag}(\boldsymbol{\Sigma})]^{-1/2},$$

where $\operatorname{diag}(\boldsymbol{\Sigma})$ is a diagonal matrix formed by the diagonal elements of $\boldsymbol{\Sigma}$. The sample covariance matrix $\mathbf{S}_n$ and sample correlation matrix $\widehat{\mathbf{R}}_n$ are defined as

$$\mathbf{S}_n = (n-1)^{-1}\sum_{j=1}^{n}(\mathbf{y}_j-\bar{\mathbf{y}})(\mathbf{y}_j-\bar{\mathbf{y}})^T,\ \ \widehat{\mathbf{R}}_n = [\operatorname{diag}(\mathbf{S}_n)]^{-1/2}\mathbf{S}_n[\operatorname{diag}(\mathbf{S}_n)]^{-1/2}, \tag{1.1}$$

where $\bar{\mathbf{y}} = n^{-1}\sum_{j=1}^{n}\mathbf{y}_j$ is the sample mean.

A large body of literature has studied the spectral properties of the sample covariance matrix. (1). Regarding the *limiting spectral distribution* (LSD) of a high-dimensional sample covariance matrix, [13] conducted pioneering work and established the M-P law. Later, [22] and [20] further extended the results of the Marčenko-Pastur law and established the famous Stieltjes equations of the LSD; (2). For the *central limit theorem* (CLT) of *linear spectral statistics* (LSS) of a high-dimensional sample covariance matrix, [12] proved that the normalized power sums of sample eigenvalues were asymptotically distributed as the standardized Gaussian distribution under $\boldsymbol{\Sigma} = \mathbf{I}_p$ by using the moment method. [3] established the CLT for the LSS of high-dimensional sample covariance matrices under the Gaussian-like moment conditions. The CLT was further extended to a situation with an unknown mean vector and any

fourth moment by [16,27]. [8] further extended the CLT to elliptical distributions. (3). Regarding the limiting distribution of the extreme eigenvalues of a high-dimensional covariance matrix, [11] derived the Tracy-Widom law of the largest eigenvalue. [23] extended the result to elliptical distributions. For more results of high-dimensional covariance matrices, readers may refer to the book by [2].

Regarding the spectral properties of high-dimensional sample correlation matrices $\widehat{\mathbf{R}}_n$, some studies have been conducted, but the literature is relatively limited. (1). For the LSD of a high-dimensional sample correlation matrix, [9] proved that the *empirical spectral distribution* (ESD) of $\widehat{\mathbf{R}}_n$ weakly converged to the Marčenko-Pastur law under $\mathbf{R} = \mathbf{I}_p$. [5] derived the Marčenko-Pastur-type system of equations of the LSD of $\widehat{\mathbf{R}}_n$ for elliptical distributions, allowing $\mathbf{R} \neq \mathbf{I}_p$. (2). For the extreme eigenvalues of $\widehat{\mathbf{R}}_n$, the almost sure limit of the smallest eigenvalue of $\widehat{\mathbf{R}}_n$ was derived by [24] under $\mathbf{R} = \mathbf{I}_p$. [4] and [18] derived the Tracy-Widom law of extreme eigenvalues of $\widehat{\mathbf{R}}_n$. (3). For the CLT of the LSS of $\widehat{\mathbf{R}}_n$, [7] established the CLT for the LSS of $\widehat{\mathbf{R}}_n$ under $\mathbf{R} = \mathbf{I}_p$. [14] relaxed the condition $\mathbf{R} = \mathbf{I}_p$ but required a Gaussian assumption since they used Gaussian tools based on the integration by parts formula and the Poincaré–Nash inequality. [10] studied the CLT of special LSS of sample correlation matrices under Gaussian assumptions. [25] established the CLT for LSS of rescaled sample correlation matrix $\widehat{\mathbf{R}}_n\mathbf{R}^{-1}$ under the independent component structure. (4). The spiked model of correlation matrix was investigated in [15] as an extension of the spiked covariance model in [17].

The existing results for the CLT of the LSS of sample correlation matrices are for certain special conditions, namely, the Gaussian assumption, $\mathbf{R} = \mathbf{I}_p$, a rescaled sample correlation matrix, or some special LSS. To provide a general CLT of the LSS of sample correlation matrices, this study establishes the CLT of the LSS of sample correlation matrices in two settings:

- The population $\mathbf{y}$ follows an independent component structure

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\Gamma}\mathbf{x}, \tag{1.2}$$

  where $\boldsymbol{\Gamma}$ is a $p \times p$ non-random matrix, $\mathrm{E}\mathbf{y} = \boldsymbol{\mu}$, and $\mathbf{x}$ is a p-dimensional random vector with i.i.d. entries.
- The population $\mathbf{y}$ follows an elliptical structure

$$\mathbf{y} = \boldsymbol{\mu} + \rho\boldsymbol{\Gamma}\mathbf{x}, \tag{1.3}$$

  where $\boldsymbol{\Gamma}$ is a $p \times p$ non-random matrix with $\mathrm{rank}(\boldsymbol{\Gamma}) = p$, ρ is a non-negative random radius of $\mathbf{y}$, and $\mathbf{x}$ is a p-dimensional random direction independent of ρ, and uniformly distributed on the unit sphere S^{p-1} in $\mathbb{R}^p$ (denoted by $\mathbf{x} \sim U(S^{p-1})$).

The elliptical structure and the independent component structure both include Gaussian distributions. However, the elliptical structure includes some heavy-tailed distributions (e.g., multivariate t distribution), but the independent component structure does not. Three examples of elliptical distributions are given as follows:

- **Multivariate Gaussian distribution**: When $\rho^2 \sim \chi^2_m$, where χ^2_m denotes the chi-square distribution with the degree of freedom m, then $\mathbf{y}$ follows $N(\boldsymbol{\mu}, \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)$.
- **Double exponential distribution**: When $\rho \sim Gamma(p,1)$, where $Gamma(p,1)$ denotes the Gamma distribution with shape and scale parameters p and 1, $\mathbf{y}$ follows a double exponential distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T$.
- **Multivariate t distribution**: When $\rho^2/p \sim F(p,\nu)$ with $\nu > 4$, where $F(p,\nu)$ denotes the F distribution with degrees of freedom p and ν, then $\mathbf{y} \sim t(\nu, \boldsymbol{\mu}, \nu(\nu-2)^{-1}\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)$, which is a multivariate t distribution with mean vector $\boldsymbol{\mu}$, covariance matrix $\nu(\nu-2)^{-1}\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T$, and degree of freedom ν.

For other elliptical distributions, such as Kotz-type distributions and Pearson type II distributions, see [6]. Multivariate t and Gaussian distributions are also mentioned in [8].

Our contributions are as follows:

- Deriving the CLT of the LSS of a high-dimensional sample correlation matrix $\widehat{\mathbf{R}}_n$ under the independent component structure (1.2);
- Deriving the CLT of the LSS of a high-dimensional sample correlation matrix $\widehat{\mathbf{R}}_n$ under the elliptical structure (1.3).

It shows that the CLTs of the LSS of sample correlation matrices are very different under the independent component structure and the elliptical structure. Even if the population correlation matrix is an identity matrix, that is, $\mathbf{R} = \mathbf{I}_p$, the two CLTs differ under the two structures.

This paper is organized as follows: Section 2 derives the LSD of high-dimensional $\widehat{\mathbf{R}}_n$ under both independent component and elliptical structures. The CLT of the LSS of $\widehat{\mathbf{R}}_n$ under the elliptical and independent component structures is established in Sections 3 and 4, respectively. Section 5 presents the application as well as some simulation results. The sketches of proofs of theorems can be found in the Appendix. The detailed proofs of some theorems are included in the supplement.

2. Limiting spectral distribution of $\widehat{\mathbf{R}}_n$

Assumption $\mathbf{A}_E$. Assume that the i.i.d. samples $\mathbf{y}_1,\ldots,\mathbf{y}_n$ satisfy the following elliptical structure:

$$\mathbf{y}_j = \rho_j \mathbf{\Gamma} \mathbf{x}_j + \boldsymbol{\mu},\ j = 1,\ldots,n, \tag{2.1}$$

where $\mathbf{\Gamma}$ is a $p \times p$ non-random matrix with $\operatorname{rank}(\mathbf{\Gamma}) = p$, $\mathbf{x}_j$ is a p-dimensional random direction independent of ρ_j and uniformly distributed on the unit sphere S^{p-1} in $\mathbb{R}^p$, and ρ_j is a non-negative random radius satisfying

$$\mathrm{E}\rho_j^2 = p, \quad \mathrm{E}\rho_j^4 = p^2 + \tau p + o(p), \quad \mathrm{E}\left|p^{-1/2}(\rho_j^2 - p)\right|^{2+\varepsilon} < \infty, \tag{2.2}$$

for constants $\tau \geq 0$ and $\varepsilon > 0$.

Assumption $\mathbf{A}_L$. Assume that the i.i.d. samples $\mathbf{y}_1,\ldots,\mathbf{y}_n$ satisfy the following linear independent component structure:

$$\mathbf{y}_j = \mathbf{\Gamma} \mathbf{x}_j + \boldsymbol{\mu},\ j = 1,\ldots,n,$$

where $\mathbf{\Gamma}$ is a $p \times p$ non-random matrix, $\mathrm{E}\mathbf{y}_j = \boldsymbol{\mu}$, and $\mathbf{x}_j = (x_{1j},\ldots,x_{pj})^T$ is a p-dimensional random vector with i.i.d. entries satisfying

$$\mathrm{E}x_{ij} = 0,\ \mathrm{E}x_{ij}^2 = 1,\ \mathrm{E}x_{ij}^4 = \beta_x + 3 + o(1),\ \mathrm{E}(|x_{ij}|^4(\log(|x_{ij}|))^{2+2\epsilon}) < \infty.$$

Assumption B. Let $\mathbf{G} = [\operatorname{diag}(\mathbf{\Sigma})]^{-1/2}\mathbf{\Gamma}$. Assume that the ESD H_n of $\mathbf{R} = \mathbf{G}\mathbf{G}^T$ weakly converges to a proper distribution H as $p \to \infty$. Moreover, the spectral norm of $\mathbf{R}$ is uniformly bounded in p.

Assumption C. A convergence regime is required as $y_n = p/n \to y \in (0, +\infty)$.

Assumptions A_E and A_L are for two model structures, where A_E applies to elliptical structures, including some heavy-tailed distributions, and A_L applies to independent component structures. For the aforementioned elliptical distributions, we have

- When $\mathbf{y}$ follows a multivariate normal distribution, $\tau = 2$ and $\rho^2 \sim \chi^2_p$;
- When $\mathbf{y}$ follows a double exponential distribution, $\tau = 4$ and $\rho \sim Gamma(p, 1)$;
- When $\mathbf{y}$ follows a multivariate t distribution, τ does not exist and $\rho^2 \sim pF(p, \nu)$.

Assumption B is for the spectral norm of the population correlation matrix $\mathbf{R}$. Assumption C states that the population dimension p and sample size n tend to infinity proportionally.

For simplicity, let $\{\hat{\lambda}_i, i = 1, \ldots, p\}$ be the sample eigenvalues of $\widehat{\mathbf{R}}_n$ defined in (1.1). The ESD of $\widehat{\mathbf{R}}_n$ is defined as

$$F_n(x) = p^{-1} \sum_{i=1}^{p} \delta_{\{\hat{\lambda}_i \le x\}},$$

where $\delta_{\{\cdot\}}$ is an indicator function and x is any real number. The following theorem provides the LSD $F^{y,H}(x)$ of $F_n(x)$.

Theorem 2.1. *Under Assumptions A_L-B-C or A_E-B-C, the ESD $F_n(x)$ of $\widehat{\mathbf{R}}_n$ converges almost surely to the LSD $F^{y,H}$, whose Stieltjes transform $s(z)$ is the only solution to the equation*

$$s(z) = \int \frac{1}{t[1 - y - yzs(z)] - z} dH(t), \quad z \in \mathbb{C}^+, \tag{2.3}$$

in the set $\{s(z) : -(1-y)/z + ys(z) \in \mathbb{C}^+\}$, where $\mathbb{C}^+ = \{z \in \mathbb{C} : \Im(z) > 0\}$ with $\Im(z)$ being the imaginary part of z. Letting

$$\underline{s}(z) = -\frac{1 - y}{z} + ys(z), \quad z \in \mathbb{C}^+,$$

then (2.3) can be re-expressed as

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

Let $[a, b]$ be the support set of the LSD $F^{y,H}(x)$; then, $F^{y,H}(x) = 0$, $x < 0$ and

$$F^{y,H}(x) = \begin{cases} \int_0^x f^{y,H}(t)dt, & y \le 1, x \ge 0, \\ \int_0^x f^{y,H}(t)dt + (1 - 1/y)\delta_{\{x \ge 0\}}, & y > 1, x \ge 0, \end{cases}$$

with the limiting spectral density being

$$f^{y,H}(x) = (y\pi)^{-1} \lim_{z \to x} \Im(\underline{s}(z))\delta_{\{0 \le a \le x \le b\}}. \tag{2.4}$$

Note that the Stieltjes transform $s(z)$ of the LSD of the high-dimensional sample correlation matrix $\widehat{\mathbf{R}}_n$ in (2.3) has the same form as that of the high-dimensional sample covariance matrix in [20]. Moreover, the Stieltjes transforms of the LSD of $\widehat{\mathbf{R}}_n$ are the same for elliptical and independent component structures.

3. Central limit theorem of linear spectral statistics of $\widehat{\mathbf{R}}_n$ under elliptical structure

Let $\{\hat{\lambda}_i, i = 1,\ldots,p\}$ be the sample eigenvalues of $\widehat{\mathbf{R}}_n$ defined in (1.1), where $\mathbf{y}_1,\ldots,\mathbf{y}_n$ satisfy Assumption A_E. Define the LSS of $\widehat{\mathbf{R}}_n$ as

$$L_{g_\ell} = \sum_{i=1}^{p} g_\ell(\hat{\lambda}_i), \quad \ell = 1,\ldots,K, \tag{3.1}$$

where $g_\ell(\cdot), \ell = 1,\ldots,K$ are some known analytic functions. To establish the CLT of the LSS of $\widehat{\mathbf{R}}_n$, for a fixed K and known functions $g_1,\ldots,g_K$, we consider the K-dimensional random vector $(W(g_1), \ldots, W(g_K))$, where

$$W(g_\ell) = \sum_{i=1}^{p} g_\ell(\hat{\lambda}_i)\delta_{\{\hat{\lambda}_i>0\}} - p\int_a^b g_\ell(x) f^{y_{n-1},H}(x)dx, \ \ell = 1,\ldots,K, \tag{3.2}$$

and $f^{y_{n-1},H}$ is defined in (2.4) with $y_{n-1} = p/(n-1)$.

We need an additional assumption as follows:

Assumption D. The functions $g_1,\ldots,g_K$ are known analytic functions in a domain containing

$$\left[\liminf_p \lambda_{\min}^{\mathbf{R}} \cdot \delta_{\{0\le y\le 1\}}(1-\sqrt{y})^2, \limsup_p \lambda_{\max}^{\mathbf{R}} \cdot (1+\sqrt{y})^2\right]$$

where $\lambda_{\min}^{\mathbf{R}}$ and $\lambda_{\max}^{\mathbf{R}}$ are the minimum and maximum eigenvalues of $\mathbf{R}$, respectively.

3.1. Central limit theorem under elliptical structure

This section establishes the CLT of the random vector $(W(g_1),\ldots,W(g_K))$ under the elliptical structure assumption A_E.

Theorem 3.1. *Under Assumptions A_E-B-C-D, the random vector $(W(g_1),\ldots,W(g_K))$ weakly converges to a multivariate Gaussian random vector $(X_{g_1},\ldots,X_{g_K})$ with the mean and covariance functions as follows:*

$$\mathrm{E}X_{g_\ell} = -\frac{1}{2\pi\mathbf{i}}\oint_C g_\ell(z)\mathrm{E}M(z)\,dz,$$

$$\mathrm{Cov}(X_{g_{\ell_1}}, X_{g_{\ell_2}}) = -\frac{1}{4\pi^2}\oint_{C_1}\oint_{C_2} g_{\ell_1}(z_1)g_{\ell_2}(z_2)\mathrm{Cov}(M(z_1),M(z_2))\,dz_2\,dz_1,$$

where

$$\begin{aligned}\mathrm{E}M(z) =& y\int \frac{[t\underline{s}'(z)]^2}{\underline{s}(z)[1+t\underline{s}(z)]^3}dH(t) + (\tau-2)[1+z\underline{s}(z)]\int \frac{t\underline{s}'(z)}{[1+t\underline{s}(z)]^2}dH(t)\\ &+ \lim_{n\to\infty}\frac{2}{n}\sum_{k=1}^{p}\frac{\partial}{\partial z}\left[\underline{s}(z)\left(\mathbf{e}_k^T\mathbb{R}(z)\mathbf{R}\mathbf{e}_k\right)^2\right]\\ &+ \lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{p}\sum_{\ell=1}^{p}r_{k\ell}^2\cdot\frac{\partial}{\partial z}\left[\left(\mathbf{e}_k^T\mathbb{R}(z)\mathbf{e}_\ell\right)^2\right]\end{aligned} \tag{3.3}$$

and

$$\begin{aligned}\mathrm{Cov}(M(z_1),M(z_2)) =& 2\left\{\frac{\underline{s}'(z_1)\underline{s}'(z_2)}{[\underline{s}(z_2)-\underline{s}(z_1)]^2}-\frac{1}{(z_1-z_2)^2}\right\}\\ &+\lim_{n\to+\infty}\frac{2}{n}\sum_{k=1}^{p}\sum_{\ell=1}^{p}r_{k\ell}^2\cdot\frac{\partial}{\partial z_1}\mathbf{e}_k^T\mathbb{R}(z_1)\mathbf{e}_k\cdot\frac{\partial}{\partial z_2}\mathbf{e}_\ell^T\mathbb{R}(z_2)\mathbf{e}_\ell\\ &+\lim_{n\to+\infty}\frac{2}{n}\sum_{k=1}^{p}\frac{\partial}{\partial z_1}\mathbf{e}_k^T\mathbb{R}(z_1)\mathbf{e}_k\cdot\frac{\partial}{\partial z_2}\left[\frac{\mathbf{e}_k^T\mathbb{R}(z_2)\mathbf{e}_k-1}{\underline{s}(z_2)}\right]\\ &+\lim_{n\to+\infty}\frac{2}{n}\sum_{k=1}^{p}\frac{\partial}{\partial z_1}\left[\frac{\mathbf{e}_k^T\mathbb{R}(z_1)\mathbf{e}_k-1}{\underline{s}(z_1)}\right]\cdot\frac{\partial}{\partial z_2}\mathbf{e}_k^T\mathbb{R}(z_2)\mathbf{e}_k,\end{aligned} \tag{3.4}$$

for $\ell,\ell_1,\ell_2\in\{1,\dots,K\}$, C, C_1, and C_2 are three contours enclosing the support $[a,b]$ of $F^{y,H}(x)$, C_1, C_2 are non-overlapping, the contour integral $\oint$ is anticlockwise, $r_{k\ell}$ is the (k,ℓ)th element of the population correlation matrix $\mathbf{R}$, $\mathbf{e}_k$ is the kth column of $p\times p$ identity matrix $\mathbf{I}_p$, $\underline{s}'(z)$ is the derivative of $\underline{s}(z)$ at z, and

$$\mathbb{R}(z)=\left(\mathbf{I}_p+\underline{s}(z)\mathbf{R}\right)^{-1}.$$

Remark 3.1. When $\mathbf{R}=\mathbf{I}_p$, (3.3) and (3.4) can be simplified as

$$\mathrm{E}M(z)=\frac{y[\underline{s}'(z)]^2}{\underline{s}(z)[1+\underline{s}(z)]^3}+(\tau-4)[1+z\underline{s}(z)]\frac{\underline{s}'(z)}{[1+\underline{s}(z)]^2},$$

$$\mathrm{Cov}(M(z_1),M(z_2))=2\left\{\frac{\underline{s}'(z_1)\underline{s}'(z_2)}{[\underline{s}(z_2)-\underline{s}(z_1)]^2}-\frac{1}{(z_1-z_2)^2}\right\}-2y\frac{\underline{s}'(z_1)\underline{s}'(z_2)}{[1+\underline{s}(z_1)]^2[1+\underline{s}(z_2)]^2}.$$

Theorem 3.2. *Under Assumptions A_E-B-C-D and $\mathbf{R}=\mathbf{I}_p$, the random vector*

$$(W(g_1),\dots,W(g_K))$$

converges weakly to a multivariate Gaussian random vector $(X_{g_1},\ldots,X_{g_K})$ *with*

$$\begin{aligned}\mathrm{E}X_{g_\ell} &= \lim_{r\to 1^+}\frac{1}{2\pi\mathbf{i}}\oint_{|\xi|=1} g_\ell(|1+\sqrt{y}\xi|^2)\left(\frac{\xi}{\xi^2-r^{-2}}-\frac{1}{\xi}\right)d\xi\\ &\quad+\frac{\tau-4}{2\pi\mathbf{i}}\oint_{|\xi|=1}\frac{g_\ell(|1+\sqrt{y}\xi|^2)}{\xi^3}\,d\xi\end{aligned} \tag{3.5}$$

and

$$\begin{aligned}\mathrm{Cov}(X_{g_{\ell_1}},X_{g_{\ell_2}}) &= \lim_{r\to 1^+}\frac{-1}{2\pi^2}\oint_{|\xi_1|=1}\oint_{|\xi_2|=1}\frac{g_{\ell_1}(|1+\sqrt{y}\xi|^2)g_{\ell_2}(|1+\sqrt{y}\xi|^2)}{(\xi_1-r\xi_2)^2}\,d\xi_2 d\xi_1\\ &\quad+\frac{1}{2\pi^2}\oint_{|\xi_1|=1}\frac{g_{\ell_1}(|1+\sqrt{y}\xi_1|^2)}{\xi_1^2}d\xi_1\oint_{|\xi_2|=1}\frac{g_{\ell_2}(|1+\sqrt{y}\xi_2|^2)}{\xi_2^2}d\xi_2,\end{aligned} \tag{3.6}$$

where $\ell,\ell_1,\ell_2\in\{1,\ldots,K\}$*, the contour* $\oint$ *is anticlockwise, and*

$$|1+\sqrt{y}\xi|^2 = 1+\sqrt{y}\xi+\sqrt{y}\xi^{-1}+y$$

satisfies $|\xi|=1$.

Note that the CLTs of the LSS of $\widehat{\mathbf{R}}_n$ in Theorems 3.1 and 3.2 differ from those of $\widehat{\mathbf{R}}_n\mathbf{R}^{-1}$ in [25].

3.2. Two examples

The following two examples show two applications of the CLT of the LSS of sample correlation matrices under elliptical structure: one is for spatial-sign samples, and the other is for general samples from elliptical structure.

Example 3.1. Letting $g_\ell(x)=x^\ell$ for $\ell=1,2,3,4$ and $g_5(x)=\log x$, under Assumptions A_E-B-C-D and $\mathbf{R}=\mathbf{I}_p$, we have

- Centering terms:

$$\begin{aligned}&\textstyle\int g_1(x)f^{y_{n-1}}(x)dx=1,\quad \int g_2(x)f^{y_{n-1}}(x)dx=1+y_{n-1},\\ &\textstyle\int g_3(x)f^{y_{n-1}}(x)dx=1+3y_{n-1}+y_{n-1}^2,\\ &\textstyle\int g_4(x)f^{y_{n-1}}(x)dx=1+6y_{n-1}+6y_{n-1}^2+y_{n-1}^3,\\ &\textstyle\int g_5(x)f^{y_{n-1}}(x)dx=\frac{y_{n-1}-1}{y_{n-1}}\log(1-y_{n-1})-1,\ \ y_{n-1}<1,\end{aligned} \tag{3.7}$$

- Mean terms:

$$\begin{aligned}&\mathrm{E}X_{g_1}=0,\ \mathrm{E}X_{g_2}=-3y+\tau y,\\ &\mathrm{E}X_{g_3}=-9y(1+y)+3\tau y(1+y),\\ &\mathrm{E}X_{g_4}=6(\tau-3)y(1+y)^2+(4\tau-10)y^2,\\ &\mathrm{E}X_{g_5}=0.5\log(1-y)-0.5y(\tau-4).\end{aligned} \tag{3.8}$$

- Variance and covariance terms:

$$\begin{aligned}
&\mathrm{Var}(X_{g_1}) = 0, \quad \mathrm{Var}(X_{g_2}) = 4y^2,\\
&\mathrm{Var}(X_{g_3}) = 6y^3 + 36y^2(1+y)^2,\\
&\mathrm{Var}(X_{g_4}) = 8y^4 + 96y^3(1+y)^2 + 16y^2[2y + 3(1+y)^2]^2,\\
&\mathrm{Var}(X_{g_5}) = -2\log(1-y) - 2y, \quad y < 1,\\
&\mathrm{Cov}(X_{g_1}, X_{g_\ell}) = 0, \quad \ell = 2,3,4,5,\\
&\mathrm{Cov}(X_{g_2}, X_{g_3}) = 12y^2(1+y),\\
&\mathrm{Cov}(X_{g_2}, X_{g_4}) = 8y^2[2y + 3(1+y)^2],\\
&\mathrm{Cov}(X_{g_2}, X_{g_5}) = -2y^2,\\
&\mathrm{Cov}(X_{g_3}, X_{g_4}) = 24(1+y)y^3 + 24y^2(1+y)[2y + 3(1+y)^2],\\
&\mathrm{Cov}(X_{g_3}, X_{g_5}) = 2y^3 - 6y^2(1+y),\\
&\mathrm{Cov}(X_{g_4}, X_{g_5}) = -2y^4 + 8(1+y)^2y^3 - 4y^2\left(2y + 3(1+y)^2\right).
\end{aligned} \tag{3.9}$$

Example 3.2. Suppose that the samples $\mathbf{y}_1^0, \mathbf{y}_2^0, \ldots, \mathbf{y}_n^0$ are i.i.d. from a p-dimensional elliptical structure

$$\mathbf{y}_j^0 = \rho_j^0 \mathbf{\Gamma} \mathbf{x}_j, \quad j = 1, \ldots, n,$$

in (2.1) and (2.2). Let the spatial-sign samples be

$$\mathbf{y}_j = \sqrt{p}\frac{\mathbf{y}_j^0}{\|\mathbf{y}_j^0\|}, \quad j = 1, \ldots, n.$$

Let the sample correlation matrix $\widehat{\mathbf{R}}_n$ be from the spatial-sign samples $\mathbf{y}_1, \ldots, \mathbf{y}_n$,

$$\mathbf{S}_n = (n-1)^{-1}\sum_{j=1}^{n}(\mathbf{y}_j - \bar{\mathbf{y}})(\mathbf{y}_j - \bar{\mathbf{y}})^T, \quad \widehat{\mathbf{R}}_n = [\mathrm{diag}(\mathbf{S}_n)]^{-1/2}\mathbf{S}_n[\mathrm{diag}(\mathbf{S}_n)]^{-1/2},$$

and $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$ be the sample eigenvalues of $\widehat{\mathbf{R}}_n$. Then, when $\mathbf{y}_1^0, \mathbf{y}_2^0, \ldots, \mathbf{y}_n^0$ satisfy Assumptions A_E-B-C-D and $\mathrm{Cov}(\mathbf{y}_j^0) = \mathbf{\Sigma} = \sigma^2\mathbf{I}_p$, where σ^2 is an unknown scalar parameter, the random vector

$$(W(g_1), \ldots, W(g_K))$$

converges weakly to a multivariate Gaussian random vector $(X_{g_1}, \ldots, X_{g_K})$ with the mean function $\mathrm{E}X_{g_\ell}$ and covariance function $\mathrm{Cov}(X_{g_{\ell_1}}, X_{g_{\ell_2}})$, which are the same as (3.5) and (3.6) with $\tau = 0$; that is,

$$\mathrm{E}X_{g_\ell} = \lim_{r\to 1^+}\frac{1}{2\pi\mathbf{i}}\oint_{|\xi|=1} g_\ell(|1+\sqrt{y}\xi|^2)\left(\frac{\xi}{\xi^2 - r^{-2}} - \frac{1}{\xi}\right) - \frac{4}{2\pi\mathbf{i}}\oint_{|\xi|=1}\frac{g_\ell(|1+\sqrt{y}\xi|^2)}{\xi^3}\,d\xi, \tag{3.10}$$

$$\begin{aligned}
\mathrm{Cov}(X_{g_{\ell_1}}, X_{g_{\ell_2}}) = &\lim_{r\to 1^+}\frac{-1}{2\pi^2}\oint_{|\xi_1|=1}\oint_{|\xi_2|=1}\frac{g_{\ell_1}(|1+\sqrt{y}\xi|^2)g_{\ell_2}(|1+\sqrt{y}\xi|^2)}{(\xi_1 - r\xi_2)^2}\,d\xi_2 d\xi_1\\
&+ \frac{1}{2\pi^2}\oint_{|\xi_1|=1}\frac{g_{\ell_1}(|1+\sqrt{y}\xi_1|^2)}{\xi_1^2}d\xi_1\oint_{|\xi_2|=1}\frac{g_{\ell_2}(|1+\sqrt{y}\xi_2|^2)}{\xi_2^2}d\xi_2.
\end{aligned} \tag{3.11}$$

In particular, for $g_1(x)=x$, $g_2(x)=x^2$, $g_3(x)=x^3$, $g_4(x)=x^4$, $g_5(x)=\log x$, $\int g_\ell(x)f^{y_{n-1}}(x)dx$, $\mathrm{E}X_{g_\ell}$ and $\mathrm{Cov}(X_{g_{\ell_1}},X_{g_{\ell_2}})$ are given in (3.7), (3.8) and (3.9) with $\tau=0$.

Proof. Since $\mathbf{x}_j$ is uniformly distributed on the unit sphere S^{p-1} in $\mathbb{R}^p$, then $\mathbf{x}_j$ can be equivalently written as

$$\mathbf{x}_j=\mathbf{w}_j/\|\mathbf{w}_j\|,\ \ \mathbf{w}_j\sim N(0,\mathbf{I}_p).$$

Thus, we have

$$\mathbf{y}_j^0=\rho_j^0\mathbf{\Gamma}\frac{\mathbf{w}_j}{\|\mathbf{w}_j\|},\ \ \|\mathbf{y}_j^0\|=\rho_j^0\frac{\|\mathbf{\Gamma}\mathbf{w}_j\|}{\|\mathbf{w}_j\|}.$$

Then, by $\mathbf{\Gamma}\mathbf{\Gamma}^T=\mathbf{\Sigma}=\sigma^2\mathbf{I}_p$, we obtain

$$\mathbf{y}_j=\sqrt{p}\frac{\mathbf{\Gamma}\mathbf{w}_j}{\|\mathbf{\Gamma}\mathbf{w}_j\|}\overset{d}{=}\sqrt{p}\frac{\mathbf{w}_j}{\|\mathbf{w}_j\|}=\rho_j\mathbf{x}_j,$$

where $\mathbf{\Gamma}\mathbf{w}_j\sim N(0,\sigma^2\mathbf{I}_p)$, $\overset{d}{=}$ denotes having the same distributions, $\rho_j=\sqrt{p}$, and $\mathrm{E}\rho_j^4=p^2$ with $\tau=0$. □

4. Central limit theorem of linear spectral statistics of $\widehat{\mathbf{R}}_n$ under linear independent component structure

This section establishes the CLT of the random vector $(W(g_1),\ldots,W(g_K))$ under the linear independent component structure assumption A_L. Let $\{\hat{\lambda}_i,i=1,\ldots,p\}$ be the sample eigenvalues of $\widehat{\mathbf{R}}_n$ defined in (1.1), where $\mathbf{y}_1,\ldots,\mathbf{y}_n$ satisfy Assumption A_L.

Theorem 4.1. *Under Assumptions A_L-B-C-D and $\mathbf{G}^T\mathbf{G}$ being a diagonal matrix or $\beta_x=0$, the random vector $(W(g_1),\cdots,W(g_K))$ converges to a multivariate Gaussian random vector $(X_{g_1},\cdots,X_{g_K})$ with the mean and covariance functions as follows:*

$$\mathrm{E}X_{g_\ell}=-\frac{1}{2\pi\mathbf{i}}\oint_C g_\ell(z)\mathrm{E}M(z)\,dz,$$

$$\mathrm{Cov}(X_{g_{\ell_1}},X_{g_{\ell_2}})=-\frac{1}{4\pi^2}\oint_{C_1}\oint_{C_2}g_{\ell_1}(z_1)g_{\ell_2}(z_2)\mathrm{Cov}(M(z_1),M(z_2))\,dz_2\,dz_1,$$

where

$$\begin{aligned}\mathrm{E}M(z) =& y\int \frac{[t\underline{s}'(z)]^2}{\underline{s}(z)[1+t\underline{s}(z)]^3}dH(t)+\beta_x y\int \frac{t^2\underline{s}'(z)\underline{s}(z)}{[1+t\underline{s}(z)]^3}dH(t)\\ &+\lim_{n\to\infty}\frac{2}{n}\sum_{k=1}^{p}\frac{\partial}{\partial z}[\underline{s}(z)(\mathbf{e}_k^T\mathbb{R}(z)\mathbf{R}\mathbf{e}_k)^2]\\ &+\lim_{n\to\infty}\frac{\beta_x}{n}\sum_{k=1}^{p}\sum_{\ell=1}^{p}g_{k\ell}^2\cdot\frac{\partial}{\partial z}\left[\underline{s}(z)(\mathbf{e}_\ell^T\mathbf{G}^T\mathbb{R}(z)\mathbf{e}_k)^2\right]\\ &+\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{p}\sum_{\ell=1}^{p}r_{k\ell}^2\cdot\frac{\partial}{\partial z}\left[\left(\mathbf{e}_k^T\mathbb{R}(z)\mathbf{e}_\ell\right)^2\right]\\ &+\frac{\beta_x}{2}\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{p}\sum_{\ell=1}^{p}\frac{\partial}{\partial z}\left[\left(\mathbf{e}_k^T\mathbb{R}(z)\mathbf{e}_\ell\right)^2\right]\sum_{j=1}^{p}g_{\ell j}^2g_{kj}^2,\end{aligned}\tag{4.1}$$

and

$$\begin{aligned}&\mathrm{Cov}(M(z_1),M(z_2))\\ &=2\left\{\frac{\underline{s}'(z_1)\underline{s}'(z_2)}{[\underline{s}(z_2)-\underline{s}(z_1)]^2}-\frac{1}{(z_1-z_2)^2}\right\}+\beta_x y\int\frac{t^2\underline{s}'(z_1)\underline{s}'(z_2)}{[1+t\underline{s}(z_1)]^2[1+t\underline{s}(z_2)]^2}dH(t)\\ &+\lim_{n\to+\infty}\frac{1}{n}\sum_{k=1}^{p}\sum_{\ell=1}^{p}\left(\beta_x\sum_{j=1}^{p}g_{kj}^2g_{\ell j}^2+2r_{k\ell}^2\right)\cdot\left[\frac{\partial}{\partial z_1}\mathbf{e}_k^T\mathbb{R}(z_1)\mathbf{e}_k\frac{\partial}{\partial z_2}\mathbf{e}_\ell^T\mathbb{R}(z_2)\mathbf{e}_\ell\right]\\ &+\lim_{n\to+\infty}\frac{2}{n}\sum_{k=1}^{p}\frac{\partial}{\partial z_1}\mathbf{e}_k^T\mathbb{R}(z_1)\mathbf{e}_k\cdot\frac{\partial}{\partial z_2}\left[\frac{\mathbf{e}_k^T\mathbb{R}(z_2)\mathbf{e}_k-1}{\underline{s}(z_2)}\right]\\ &+\lim_{n\to+\infty}\frac{2}{n}\sum_{k=1}^{p}\frac{\partial}{\partial z_1}\left[\frac{\mathbf{e}_k^T\mathbb{R}(z_1)\mathbf{e}_k-1}{\underline{s}(z_1)}\right]\cdot\frac{\partial}{\partial z_2}\mathbf{e}_k^T\mathbb{R}(z_2)\mathbf{e}_k\\ &+\lim_{n\to+\infty}\frac{\beta_x}{n}\sum_{k=1}^{p}\sum_{\ell=1}^{p}g_{k\ell}^2\cdot\frac{\partial}{\partial z_1}\mathbf{e}_k^T\mathbb{R}(z_1)\mathbf{e}_k\cdot\frac{\partial}{\partial z_2}\left[\underline{s}(z_2)\mathbf{e}_\ell^T\mathbf{G}^T\mathbb{R}(z_2)\mathbf{G}\mathbf{e}_\ell\right]\\ &+\lim_{n\to+\infty}\frac{\beta_x}{n}\sum_{k=1}^{p}\sum_{\ell=1}^{p}g_{k\ell}^2\cdot\frac{\partial}{\partial z_2}\mathbf{e}_k^T\mathbb{R}(z_2)\mathbf{e}_k\cdot\frac{\partial}{\partial z_1}\left[\underline{s}(z_1)\mathbf{e}_\ell^T\mathbf{G}^T\mathbb{R}(z_1)\mathbf{G}\mathbf{e}_\ell\right],\end{aligned}\tag{4.2}$$

for $\ell,\ell_1,\ell_2\in\{1,\ldots,K\}$, C, C_1, and C_2 are three contours enclosing the support $[a,b]$ of $F^{y,H}(x)$, C_1, C_2 are non-overlapping, the contour integral $\oint$ is anticlockwise, $g_{k\ell}$ is the (k,ℓ)th element of $\mathbf{G}$, and

$$\mathbb{R}(z)=\left(\mathbf{I}_p+\underline{s}(z)\mathbf{R}\right)^{-1}.$$

Note that the CLT of the LSS of $\widehat{\mathbf{R}}_n$ in Theorem 4.1 differ from that of $\widehat{\mathbf{R}}_n\mathbf{R}^{-1}$ in [25] for the case $\mathbf{R}\neq\mathbf{I}_p$. However, for $\mathbf{R}=\mathbf{I}_p$, the CLT of the LSS of $\widehat{\mathbf{R}}_n$ in Theorem 4.1 is the same as that of $\widehat{\mathbf{R}}_n\mathbf{R}^{-1}$ in [25].

Remark 4.1. When $\tau = 2$ and $\beta_x = 0$, the CLT under elliptical assumption A_E, is the same as the CLT under linear independent component assumption A_L. That is, the mean functions (3.3) and (4.1) are the same, and the covariance functions (3.4) and (4.2) are also the same.

- **Influence of τ:** The mean function in the CLT under the elliptical assumption A_E depends on the parameter τ. It is interesting that the covariance function in the CLT under the elliptical assumption A_E is **independent of** the parameter τ.
- **Influence of β_x:** The mean and covariance functions in the CLT under the independent component structure assumption A_L both depend on the parameter β_x.

Corollary 4.1. *Under Assumptions A_L-B-C-D and $\mathbf{R} = \mathbf{I}_p$, the random vector*

$$(W(g_1), \ldots, W(g_K))$$

converges weakly to a multivariate Gaussian random vector $(X_{g_1}, \ldots, X_{g_K})$ with

$$\mathrm{E}X_{g_\ell}, \ \mathrm{Cov}(X_{g_{\ell_1}}, X_{g_{\ell_2}})$$

being the same as those in Example 3.1 of [25].

Example 4.1. Letting $g_\ell(x) = x^\ell$ for $\ell = 1,2,3,4$ and $g_5(x) = \log x$, under Assumptions A_L-B-C-D and $\mathbf{R} = \mathbf{I}_p$, we have

- Centering terms: For $\ell = 1,2,3,4,5$, $\int g_\ell(x) f^{y_{n-1}}(x)dx$ is the same as (3.7) in Example 3.1.
- Mean and Covariance terms: For $\ell, \ell_1, \ell_2 = 1,2,3,4,5$, $\mathrm{E}X_{g_\ell}$ and $\mathrm{Cov}(X_{g_{\ell_1}}, X_{g_{\ell_2}})$ are the same as Example 3.3 in [25].

5. An application

In this section, we test the uncorrelation of a group of random variables $Y_1, \ldots, Y_p$; that is,

$$H_0 : \mathrm{Corr}(Y_k, Y_\ell) = 0 \ \text{ for } \ 1 \le k < \ell \le p \quad \text{v.s.} \quad H_1 : \text{not } \ H_0, \tag{5.1}$$

where $\mathrm{Corr}(\cdot)$ denotes the Pearson correlation coefficient. Since the elliptical distributions include heavy-tailed distributions, for example, multivariate t distribution, we will test (5.1) under the elliptical distributions.

5.1. Tests

Suppose that the samples $\mathbf{y}_1^0, \mathbf{y}_2^0, \ldots, \mathbf{y}_n^0$ are i.i.d. from a p-dimensional elliptical structure

$$\mathbf{y}_j^0 = \rho_j^0 \mathbf{\Gamma} \mathbf{x}_j, \ \ j = 1, \ldots, n,$$

in (2.1). Let the spatial-sign samples be

$$\mathbf{y}_j = \sqrt{p} \frac{\mathbf{y}_j^0}{\|\mathbf{y}_j^0\|}, \ \ j = 1, \ldots, n.$$

Let $\hat{\lambda}_1,\ldots,\hat{\lambda}_p$ be the sample eigenvalues of the sample correlation matrix $\widehat{\mathbf{R}}_n$ from the samples $\mathbf{y}_1,\ldots,\mathbf{y}_n$ in (1.1). We consider two test statistics as follows:

$$T_L = \log\det(\widehat{\mathbf{R}}_n) = \sum_{i=1}^{p}\log\hat{\lambda}_i,$$

$$T_{S'} = \mathrm{tr}(\widehat{\mathbf{R}}_n - \mathbf{I}_p)^2 = \sum_{i=1}^{p}(\hat{\lambda}_i - 1)^2 = \sum_{i=1}^{p}\hat{\lambda}_i^2 - p,$$

where T_L is similar to the *generalized likelihood ratio test* (GLRT) statistic in [1] and $T_{S'}$ is similar to the *Schott test* (SCT) statistic in [19]. Without loss of generality, we let

$$T_S = \sum_{i=1}^{p}\hat{\lambda}_i^2$$

replace $T_{S'}$. Motivated by [21], to enhance the powers, we also consider the following statistic:

$$T_F = \mathrm{tr}[(\widehat{\mathbf{R}}_n - \mathbf{I}_p)^4] + 3p = \sum_{i=1}^{p}\hat{\lambda}_i^4 - 4\sum_{i=1}^{p}\hat{\lambda}_i^3 + 6\sum_{i=1}^{p}\hat{\lambda}_i^2.$$

By Examples 3.1 and 3.2, and setting

$$y_{n-1} = p/(n-1),\ \ y_n = p/n,$$

under H_0 and the assumptions of Example 3.2, we have

$$\sigma_L^{-1}(T_L - \mu_L) \to N(0,1),\ \ 0 < y_{n-1} < 1,$$
$$\sigma_S^{-1}(T_S - \mu_S) \to N(0,1),\ \ 0 < y_{n-1},$$
$$\sigma_F^{-1}(T_F - \mu_F) \to N(0,1),\ \ 0 < y_{n-1},$$

where

$$\mu_L = p(y_{n-1}-1)(y_{n-1})^{-1}\log(1-y_{n-1}) - p + 0.5\log(1-y_n) + 2y_n,$$
$$\mu_S = py_{n-1} + p - 3y_n,$$
$$\mu_F = py_{n-1}^3 + 2py_{n-1}^2 + 3p + e_4 - 4e_3 + 6e_2,$$
$$\sigma_L^2 = -2\log(1-y_n) - 2y_n,\ \ \sigma_S^2 = 4y_n^2,$$
$$\sigma_F^2 = v_4 + 16v_3 + 36v_2 - 8v_{3,4} + 12v_{2,4} - 48v_{2,3},$$

with

$$e_2 = -3y_n,\quad e_3 = -9y_n(1+y_n),\quad e_4 = -18y_n(1+y_n)^2 - 10y_n^2,$$
$$v_2 = 4y_n^2,\quad v_3 = 6y_n^3 + 36y_n^2(1+y_n)^2,$$
$$v_4 = 8y_n^4 + 96y_n^3(1+y_n)^2 + 16y_n^2[2y_n + 3(1+y_n)^2]^2,$$
$$v_{2,3} = 12y_n^2(1+y_n),\quad v_{2,4} = 8y_n^2[2y_n + 3(1+y_n)^2],$$

$$v_{3,4} = 24(1+y_n)y_n^3 + 24y_n^2(1+y_n)[2y_n + 3(1+y_n)^2].$$

The rejection regions of the three tests based on the statistics T_L, T_S, and T_F for testing (5.1) at the test level 5% are as follows:

$$\{\mathbf{y}_1,\ldots,\mathbf{y}_n:\ \sigma_L^{-1}|T_L-\mu_L|>q_{0.975}\}, \tag{5.2}$$

$$\{\mathbf{y}_1,\ldots,\mathbf{y}_n:\ \sigma_S^{-1}|T_S-\mu_S|>q_{0.975}\}, \tag{5.3}$$

$$\{\mathbf{y}_1,\ldots,\mathbf{y}_n:\ \sigma_F^{-1}|T_F-\mu_F|>q_{0.975}\}, \tag{5.4}$$

where $q_{0.975}$ is the 97.5% quantile of $N(0,1)$.

5.2. Simulation study

To test (5.1), this section compares the performances of our proposed three tests: "CLRT" from (5.2), "SCT" from (5.3), and "FT" from (5.4). The matrix $\mathbf{\Gamma}$ is set as

- Scenario 1: $\mathbf{\Gamma}=\mathbf{\Sigma}^{1/2}$, $\mathbf{\Sigma}=(s_{i,j,\theta})_{p\times p}$, $s_{i,j,\theta}=\delta_{\{i=j\}}+\delta_{\{i\neq j\}}\theta^{|i-j|}$,
- Scenario 2: $\mathbf{\Gamma}=\mathbf{\Sigma}^{1/2}$, $\mathbf{\Sigma}=(s_{i,j,\eta})_{p\times p}$, $s_{i,j,\eta}=\delta_{\{i=j\}}+\delta_{\{i\neq j\}}\eta$, $i,j=1,\ldots,p$,

where $\mathbf{1}_p$ is a p-dimensional vector with p elements being one and θ, η are two parameters satisfying $|\theta|<1$, $0<\eta<1$. In Scenario 1, all eigenvalues of $\mathbf{\Sigma}$ deviate from 1 with faint signals. And in Scenario 2, the maximum eigenvalue of $\mathbf{\Sigma}$ is $1+(p-1)\eta$ with the order $O(p)$, and other $p-1$ eigenvalues of $\mathbf{\Sigma}$ are all $1-\eta$.

The samples $\mathbf{y}_1^0,\mathbf{y}_2^0,\ldots,\mathbf{y}_n^0$ are i.i.d. from a p-dimensional elliptical structure

$$\mathbf{y}_j^0=\rho_j^0\mathbf{\Gamma}\mathbf{x}_j,\quad j=1,\ldots,n$$

in (2.1), where $\mathbf{x}_j$ is uniformly distributed on the unit sphere S^{p-1} in $\mathbb{R}^p$. Then, the spatial-sign samples are $\mathbf{y}_j=\sqrt{p}\mathbf{y}_j^0/\|\mathbf{y}_j^0\|$ for $j=1,\ldots,n$. We consider multivariate t distribution of $\mathbf{y}_j^0$ as follows:

- **Multivariate t distribution**: Let $(\rho_j^0)^2/p\sim F(p,\nu)$, where ρ_j^0 is independent of $\mathbf{x}_j$. Then, $\mathbf{y}_j^0$ is distributed as the multivariate t distribution with mean vector $\mathbf{0}_p$, covariance matrix $\nu(\nu-2)^{-1}\mathbf{\Gamma}\mathbf{\Gamma}^T$, and the degree of freedom ν.

The parameter setting is as follows:

- Dimension: $p=100,250,500,1000$;
- Ratio of dimension and sample size: $p/n=0.1,0.5,0.8,1.5,5$;
- $\theta=0,0.03,0.05$ with $\theta=0$ to evaluate empirical sizes and $\theta\neq 0$ to evaluate empirical powers in Scenario 1;
- $\eta=0,0.001,0.002,\cdots,0.012$ with $\eta=0$ to evaluate empirical sizes and $\eta\neq 0$ to evaluate empirical powers in Scenario 2;
- Degree of freedom: $\nu=5$.

The test level was set as $\alpha=5\%$, the simulation times were 10000, and the rejection regions of our proposed three tests CLRT, SCT, and FT were (5.2)-(5.3)-(5.4). Table 1 presents the empirical sizes and powers of our three tests in Scenario 1. It seems that the SCT performs better than the CLRT and FT. Fig. 1 shows the performances of the three tests for $p=500$ and $n=600$ in Scenario 2. In this setting, the FT performs better than the CLRT and SCT. The simulation results show that the test

Table 1. Percentages for empirical sizes and powers of the three tests CLRT, SCT, and FT in Scenario 1.

		Empirical sizes			Empirical powers					
		$\theta = 0$			$\theta = 0.03$			$\theta = 0.05$		
p	p/n	CLRT	SCT	FT	CLRT	SCT	FT	CLRT	SCT	FT
100	0.1	4.65	4.72	3.81	14.23	14.99	11.42	67.26	70.36	57.98
	0.5	4.75	4.42	3.32	5.52	4.67	3.45	7.05	8.04	6.16
	0.8	5.76	4.30	3.06	5.17	4.61	3.68	6.01	6.54	5.12
	1.5	/	4.14	3.04	/	4.56	3.25	/	5.00	4.09
	5	/	3.89	3.80	/	4.29	4.08	/	4.03	3.68
250	0.1	4.99	4.78	4.77	58.6	61.8	49.6	100	100	100
	0.5	5.02	4.65	4.25	6.36	7.02	6.05	18.1	24.9	18.9
	0.8	5.15	4.53	3.61	5.61	6.18	4.73	8.33	12.8	10.0
	1.5	/	4.73	3.84	/	4.98	4.24	/	7.03	6.07
	5	/	3.90	3.77	/	4.32	4.15	/	4.73	4.44
500	0.1	5.39	5.12	4.91	99.0	99.4	97.6	100	100	100
	0.5	4.62	4.66	4.53	11.7	14.9	11.3	51.8	70.4	56.8
	0.8	4.96	4.75	4.45	6.71	9.31	7.63	16.6	34.4	26.6
	1.5	/	4.58	4.21	/	6.02	5.17	/	13.8	11.6
	5	/	4.82	4.34	/	4.99	4.93	/	5.45	5.58
1000	0.1	5.28	4.86	4.99	100	100	100	100	100	100
	0.5	5.39	4.69	4.44	30.8	44.6	33.8	98.0	100	99.1
	0.8	5.42	5.03	4.69	10.8	19.7	15.3	50.9	87.8	77.1
	1.5	/	4.84	4.76	/	9.06	7.58	/	37.7	30.4
	5	/	4.90	4.82	/	5.35	5.16	/	7.99	7.84

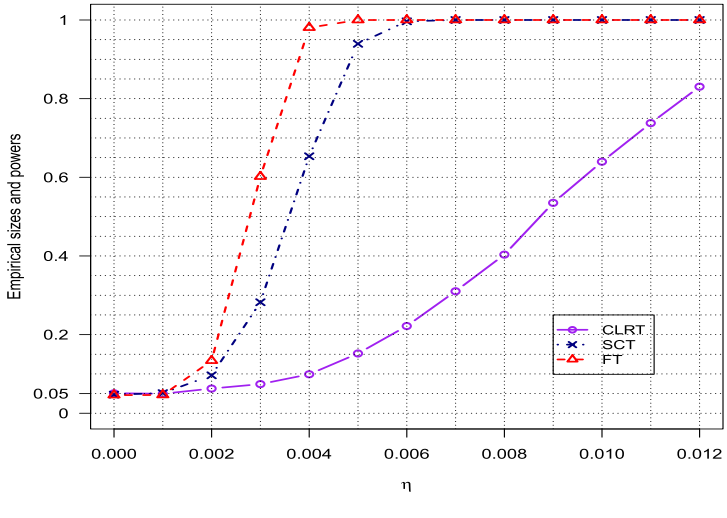

Figure 1. Empirical sizes and powers of the three tests CLRT, SCT, and FT in Scenario 2. The dimension and sample size is $p = 500$ and $n = 600$, respectively.

"SCT" is powerful for the alternative that all its eigenvalues $\{\lambda_j, j = 1, \cdots, p\}$ deviate from 1 with faint and dense signals because $\sum_{j=1}^{p}(\lambda_j - 1)^2$ will be very large. For example, if $\lambda_1 = \cdots = \lambda_{p/2} = 1.2$ and $\lambda_{p/2+1} = \cdots = \lambda_p = 0.8$, then $\sum_{j=1}^{p}(\lambda_j - 1)^2 = 40$ for $p = 1000$. The test "FT" is powerful for the alternative with its spike eigenvalues far exceeding 1 because the fourth power of the spike eigenvalues will be very large.

Appendix A: Proofs of Theorem 2.1, Theorem 3.1, and Theorem 4.1

This section provides the skeleton proof of some theorems in Sections 3 and 4. The technical details can be found in the supplementary file: Supplement on "Central Limit Theorem of Linear Spectral Statistics of High-dimensional Sample Correlation Matrices."

A.1. Remove the sample mean

First, notice that

$$\mathbf{S}_n = (n-1)^{-1}\sum_{j=1}^{n}(\mathbf{y}_j - \bar{\mathbf{y}})(\mathbf{y}_j - \bar{\mathbf{y}})^T = (n-1)^{-1}\sum_{j=1}^{n}(\mathbf{y}_j^0 - \bar{\mathbf{y}}^0)(\mathbf{y}_j^0 - \bar{\mathbf{y}}^0)^T,$$

where

$$\mathbf{y}_j^0 = \begin{cases} \rho_j \mathbf{\Gamma}\mathbf{x}_j, & \text{under elliptical case,} \\ \mathbf{\Gamma}\mathbf{x}_j, & \text{under linear case,} \end{cases}$$

and $\bar{\mathbf{y}}^0 = n^{-1}\sum_{j=1}^{n}\mathbf{y}_j^0$. Denote $\mathbf{S}_n^0 = n^{-1}\sum_{j=1}^{n}\mathbf{y}_j^0\mathbf{y}_j^{0T}$. Note that the rank of $\bar{\mathbf{y}}^0\bar{\mathbf{y}}^{0T}$ is at most one, and combined with Theorem A.43 in [2], we know that for a large n, the difference between the ESD of $\mathbf{S}_n^0$ and the ESD of $\mathbf{S}_n$ is negligible. Thus, we consider $\mathbf{S}_n^0$ instead of $\mathbf{S}_n$ in the proof of Theorem 2.1 in the next section. Then, by the proof of substitution principle for the CLT of the LSS of the sample covariance matrix in [27], we also study the matrix $\mathbf{S}_n^0$ instead of $\mathbf{S}_n$ in the proof of Theorems 3.1 and 4.1. The difference between these two CLTs is indicated by the change in the ratio of dimension to sample size, from y_n to y_{n-1}. By checking the proofs, we find that the main task in proving the substitution principle is the proof of (5.7) in [27]. By the decomposition equation (1.18) and Lemma 2.3 given in the supplement [26], combined with (5.6) and Lemmas 5.1 and 5.6 in [27], the substitution principle for the CLT of the LSS of the sample correlation matrix also holds. Section 5.3.2 in [27] contains further details.

In the following sections, we still use $\mathbf{S}_n$ instead of $\mathbf{S}_n^0$, and the definition of the sample correlation matrix is redefined accordingly. Before presenting the proof, we offer some notations. Let

$$\mathbf{G} = (g_{kh}) = [\mathrm{diag}(\mathbf{\Sigma})]^{-1/2}\mathbf{\Gamma}, \tag{A.1}$$

$$\mathbf{\Xi}_j = \begin{cases} \rho_j^2 \mathbf{G}\mathbf{x}_j\mathbf{x}_j^T\mathbf{G}^T, & \text{under elliptical case,} \\ \mathbf{G}\mathbf{x}_j\mathbf{x}_j^T\mathbf{G}^T, & \text{under linear case,} \end{cases} \tag{A.2}$$

$$\mathbf{\Xi} = \frac{1}{n}\sum_{j=1}^{n}\mathbf{\Xi}_j. \tag{A.3}$$

Observe that

$$\mathrm{E}(\boldsymbol{\Xi}_j) = \mathrm{E}(\boldsymbol{\Xi}) = \mathbf{R}, \quad \text{hence } \operatorname{diag}(\mathrm{E}(\boldsymbol{\Xi}_j)) = \operatorname{diag}(\mathrm{E}(\boldsymbol{\Xi})) = \mathbf{I}_p. \tag{A.4}$$

It is easy to see that $\widehat{\mathbf{R}}_n$ can also be written as

$$\widehat{\mathbf{R}}_n = [\operatorname{diag}(\boldsymbol{\Xi})]^{-1/2}\, \boldsymbol{\Xi}\, [\operatorname{diag}(\boldsymbol{\Xi})]^{-1/2}.$$

Throughout this paper, we also note that C and $C_{(\cdot)}$ denote constants that may take different values from one appearance to another.

A.2. Proof of Theorem 2.1

The result under the linear model was proven in [5], and we now consider the elliptical case. Note that $\|\boldsymbol{\Xi}\| \le C, a.s.$ as $n \to \infty$ (see the supplement for details); it can be verified from Lemma 2.3 in the supplement that

$$\begin{aligned}
\left\|\widehat{\mathbf{R}}_n - \boldsymbol{\Xi}\right\| =& \left\| \left([\operatorname{diag}(\boldsymbol{\Xi})]^{-1/2} - \mathbf{I}_p + \mathbf{I}_p\right) \boldsymbol{\Xi} \left([\operatorname{diag}(\boldsymbol{\Xi})]^{-1/2} - \mathbf{I}_p + \mathbf{I}_p\right) - \boldsymbol{\Xi} \right\| \\
\le& \left\| \left([\operatorname{diag}(\boldsymbol{\Xi})]^{-1/2} - \mathbf{I}_p\right) \boldsymbol{\Xi} \left([\operatorname{diag}(\boldsymbol{\Xi})]^{-1/2} - \mathbf{I}_p\right) \right\| \\
&+ \left\| \boldsymbol{\Xi} \left([\operatorname{diag}(\boldsymbol{\Xi})]^{-1/2} - \mathbf{I}_p\right) \right\| + \left\| \left([\operatorname{diag}(\boldsymbol{\Xi})]^{-1/2} - \mathbf{I}_p\right) \boldsymbol{\Xi} \right\| \\
\le& \left\| \left([\operatorname{diag}(\boldsymbol{\Xi})]^{-1/2} - \mathbf{I}_p\right) \right\|^2 \|\boldsymbol{\Xi}\| + 2 \left\| \left([\operatorname{diag}(\boldsymbol{\Xi})]^{-1/2} - \mathbf{I}_p\right) \right\| \|\boldsymbol{\Xi}\| \to 0 \quad a.s..
\end{aligned} \tag{A.5}$$

Then, the result of this theorem in the elliptical case follows from Theorem 2.1 in [8] and Weyl's inequality.

A.3. Sketch of proofs of Theorem 3.1 and Theorem 4.1

This section provides the main sketch of the proof of Theorems 3.1 and 4.1. The details are included in the Supplementary Material. Recall that, for any analytic function g in a domain containing the support interval $F^{y,H}$,

$$W(g) = \sum_{i=1}^{p} g(\hat{\lambda}_i) - p \int g(x)\, dF^{y_n, H_n}(x),$$

where $\{\hat{\lambda}_i, i = 1, \ldots, p\}$ are the eigenvalues of $\widehat{\mathbf{R}}_n$.

First, we need to conduct truncation, centralization, and rescaling. Specifically, in the elliptical case, we perform truncation and rescaling on $\{\rho_j, j = 1, \cdots, n\}$ and provide the following notation:

- *Truncation:* Denote $\check{\rho}_j = \rho_j I_{\{\left|\rho_j^2 - p\right| < \eta_n p\}}$, $\check{\boldsymbol{\Xi}} = n^{-1} \sum_{j=1}^{n} \check{\rho}_j^2 \mathbf{G}\mathbf{x}_j \mathbf{x}_j^T \mathbf{G}^T$, and

$$\check{\mathbf{R}}_n = \left[\operatorname{diag}(\check{\boldsymbol{\Xi}})\right]^{-1/2} \check{\boldsymbol{\Xi}} \left[\operatorname{diag}(\check{\boldsymbol{\Xi}})\right]^{-1/2},$$

  where the sequence $\eta_n \downarrow 0$ as $n \to \infty$ and satisfies

$$\eta_n \sqrt{n} \to \infty, \quad \eta_n^{-2} p^{-1} \mathrm{E}\left[\left(\rho_1^2 - p\right)^2 I_{\{|\rho_1^2 - p| \ge \eta_n p\}} \right] \to 0.$$

- *Rescaling:* Denote $\tilde{\rho}_j = \check{\rho}_j/\sigma_n$, $\widetilde{\mathbf{\Xi}} = n^{-1}\sum_{j=1}^n \tilde{\rho}_j^2 \mathbf{G}\mathbf{x}_j\mathbf{x}_j^T\mathbf{G}^T$, and

$$\widetilde{\mathbf{R}}_n = \left[\operatorname{diag}(\widetilde{\mathbf{\Xi}})\right]^{-1/2}\widetilde{\mathbf{\Xi}}\left[\operatorname{diag}(\widetilde{\mathbf{\Xi}})\right]^{-1/2},$$

  where $\sigma_n^2 = \mathrm{E}\left(\check{\rho}_1^2\right)/p$.

In the linear case, we perform truncation, centralization, and rescaling on $\{x_{ij}, i = 1,\cdots,p, j = 1,\cdots,n\}$ and provide the following notation:

- *Truncation:* Denote $\check{x}_{ij} = x_{ij}I_{\{|x_{ij}|<\eta_n\sqrt{n}\}}$, $\check{\mathbf{X}} = \left(\check{x}_{ij}\right)$, $\check{\mathbf{\Xi}} = n^{-1}\mathbf{G}\check{\mathbf{X}}\check{\mathbf{X}}^T\mathbf{G}$, and

$$\check{\mathbf{R}}_n = n^{-1}\left[\operatorname{diag}(\check{\mathbf{\Xi}})\right]^{-1/2}\check{\mathbf{\Xi}}\left[\operatorname{diag}(\check{\mathbf{\Xi}})\right]^{-1/2},$$

  where the sequence of $\eta_n = (\log n)^{-(1+\varepsilon)/2} \downarrow 0$ as $n \to \infty$.
- *Centralization and Rescaling:* Denote $\tilde{x}_{ij} = \left(\check{x}_{ij} - \mathrm{E}\,\check{x}_{ij}\right)/\sqrt{\mathrm{E}\left(\check{x}_{ij} - \mathrm{E}\,\check{x}_{ij}\right)^2}$, $\widetilde{\mathbf{X}} = \left(\tilde{x}_{ij}\right)$, $\widetilde{\mathbf{\Xi}} = n^{-1}\mathbf{G}\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T\mathbf{G}$, and

$$\widetilde{\mathbf{R}}_n = \left[\operatorname{diag}(\widetilde{\mathbf{\Xi}})\right]^{-1/2}\widetilde{\mathbf{\Xi}}\left[\operatorname{diag}(\widetilde{\mathbf{\Xi}})\right]^{-1/2}.$$

We define $\widetilde{W}(g)$ as the analogs of $W(g)$ with $\widehat{\mathbf{R}}_n$ replaced by $\widetilde{\mathbf{R}}_n$. We can prove that

$$W(g) = \widetilde{W}(g) + o_{\mathrm{a.s.}}(1)$$

in the elliptical and linear cases, respectively. Therefore, we only need to derive the limiting distribution of $\widetilde{W}(g)$. For simplicity, we still use $W(g)$, $\widehat{\mathbf{R}}_n$, $\mathbf{\Xi}$, ρ_j, and x_{ij} instead of $\widetilde{W}(g)$, $\widetilde{\mathbf{R}}_n$, $\widetilde{\mathbf{\Xi}}$, $\tilde{\rho}_j$, and $\tilde{x}_{ij}$, respectively. Moreover, we assume that

$$\forall j, \quad |\rho_j^2 - p| < \eta_n p, \quad \mathrm{E}(\rho_j^2) = p, \quad \mathrm{E}(\rho_j^4) = p^2 + \tau p + o(p)$$

in the elliptical case and

$$\forall i,j, \quad \left|x_{ij}\right| < \eta_n\sqrt{n}, \quad \mathrm{E}(x_{ij}) = 0, \quad \mathrm{E}(x_{ij}^2) = 1$$

in the linear case.

The proofs of Theorems 3.1 and 4.1 rely on analyzing the Stieltjes transform $s_n(z)$ of the ESD of $\widehat{\mathbf{R}}_n$. We denote $M_n(z) := p(s_n(z) - s_{y_n}(z))$, where $s_{y_n}(z)$ is the Stieltjes transform of the distribution F^{y_n,H_n}. Notice that by using the Cauchy integral formula, we have

$$W(g) = \sum_{i=1}^p g(\hat{\lambda}_i) - p\int g(x)\,dF^{y_n,H_n}(x) = -\frac{1}{2\pi i}\oint_C g(z)M_n(z)\,dz, \tag{A.6}$$

where C is any contour inside the domain and surrounding the support interval of $F^{y,H}$. This suggests that we change our target to analyze the random process $M_n(z)$. Following the ideas of the arguments on pages 562–563 in [3] and pages 542–543 in [8], we investigate a truncated version $\widehat{M}_n(z)$ of $M_n(z)$. Let x_r be any number greater than $\limsup_n \lambda_{\max}^{\mathbf{R}}(1+\sqrt{y})^2$, and x_ℓ be any negative number if $\liminf_n \lambda_{\min}^{\mathbf{R}} I_{(0,1)}(y)(1-\sqrt{y})^2 = 0$. Otherwise choose $x_\ell \in (0, \liminf_n \lambda_{\min}^{\mathbf{R}} I_{(0,1)}(y)(1-\sqrt{y})^2)$. Then, define a contour C as $C = C_\ell \cup C_u \cup C_b \cup C_r$, where

$$\begin{aligned} C_u &= \{x + i\nu_0 : x \in [x_\ell, x_r]\}, \quad C_\ell = \{x_\ell + i\nu : |\nu| \le \nu_0\} \\ C_b &= \{x - i\nu_0 : x \in [x_\ell, x_r]\}, \quad C_r = \{x_r + i\nu : |\nu| \le \nu_0\} \end{aligned}$$

and $v_0 > 0$ is to be determined. Let $C_n = C \cap \{z : |\Im z| > n^{-2}\}$. Moreover, for a sufficiently small $\varepsilon > 0$, such that

$$x_\ell + \varepsilon \le \liminf_n \lambda_{\min}^{\mathbf{R}} I_{(0,1)}(y)(1-\sqrt{y})^2 - \varepsilon \le \limsup_n \lambda_{\max}^{\mathbf{R}}(1+\sqrt{y})^2 + \varepsilon \le x_r - \varepsilon.$$

Define

$$\mathcal{B}_n = \{\liminf_n \lambda_{\min}^{\mathbf{R}} I_{(0,1)}(y)(1-\sqrt{y})^2 - \varepsilon \le \lambda_{\min}(\Xi) < \lambda_{\max}(\Xi) < \limsup_n \lambda_{\max}^{\mathbf{R}}(1+\sqrt{y})^2 + \varepsilon\}.$$

According to equations (1.9a) and (1.9b) in [3] and Lemma A.4 in [8], we have $P(\mathcal{B}_n) = o(n^{-t})$ in elliptical and linear cases for any given $t > 0$. Let

$$\widehat{M}_n(z) = \begin{cases} M_n(z), & \text{if } z \in C_n \\ M_n(x_\ell + in^{-2}), & \text{if } \Re z = x_\ell, \Im z \in [0, n^{-2}] \\ M_n(x_\ell - in^{-2}), & \text{if } \Re z = x_\ell, \Im z \in [-n^{-2}, 0) \\ M_n(x_r + in^{-2}), & \text{if } \Re z = x_r, \Im z \in [0, n^{-2}] \\ M_n(x_r - in^{-2}), & \text{if } \Re z = x_r, \Im z \in [-n^{-2}, 0). \end{cases}$$

Observe that in event $\mathcal{B}_n$, when $\Re z$ equals either x_ℓ or x_r, $|M_n(z)| \le 1/\varepsilon$, then we have

$$\left| p \oint_C g(z)(M_n(z) - \widehat{M}_n(z))\, dz \right| = \left| p \oint_{C \setminus C_n} g(z)(M_n(z) - \widehat{M}_n(z))\, dz \right| \le K \frac{p}{n^2} \cdot 1/\varepsilon = o(1).$$

Therefore, to establish the central limit theorem for $p \oint_C g(z) M_n(z)\, dz$, it suffices to study $p \oint_C g(z) \widehat{M}_n(z)\, dz$. Furthermore, since $\Im(z)$ can be chosen to be arbitrarily small, the contributions from segments C_ℓ and C_r can also be small. This allows us to focus only on $z \in C_u \cup C_b$ when analyzing $\widehat{M}_n(z)$. We still use the notation $M_n(z)$ instead of $\widehat{M}_n(z)$ in the following for notation convenience.

Next, to investigate the random process $M_n(z)$, we split $M_n(z)$ into several parts. Specifically, we have

$$M_n(z) = V + M_0 + zM_1 + z^2 M_2 + z^2 O_1 + z^3 O_2, \tag{A.7}$$

where

$$\begin{aligned} V &= \mathrm{tr}\mathbf{A}^{-1}(z) - \mathrm{E}\left(\mathrm{tr}\mathbf{A}^{-1}(z)\right) + \mathrm{tr}\left(\mathbf{A}^{-1}(z)\mathbf{D}\right) - \mathrm{Etr}\left(\mathbf{A}^{-1}(z)\mathbf{D}\right) \\ &\quad + z\mathrm{tr}\left(\mathbf{A}^{-2}(z)\mathbf{D}\right) - z\mathrm{Etr}\left(\mathbf{A}^{-2}(z)\mathbf{D}\right), \\ M_0 &= \mathrm{E}\left(\mathrm{tr}\mathbf{A}^{-1}(z)\right) - p s_{y_n}(z) + \mathrm{Etr}\left(\mathbf{A}^{-1}(z)\mathbf{D}\right) + z\mathrm{Etr}\left(\mathbf{A}^{-2}(z)\mathbf{D}\right), \\ M_1 &= \mathrm{tr}\left(\mathbf{A}^{-1}(z)\mathbf{D}\right)^2, \quad M_2 = \mathrm{tr}\left(\left(\mathbf{A}^{-1}(z)\mathbf{D}\right)^2 \mathbf{A}^{-1}(z)\right), \\ O_1 &= \mathrm{tr}\left(\left(\mathbf{A}^{-1}(z)\mathbf{D}\right)^2 \mathbf{H}^{-1}(z)\mathbf{D}\right), \quad O_2 = \mathrm{tr}\left(\left(\mathbf{A}^{-1}(z)\mathbf{D}\right)^3 \mathbf{H}^{-1}(z)\right), \\ \mathbf{A}(z) &= \Xi - z\mathbf{I}_p, \quad \mathbf{D} = \mathrm{diag}(\Xi) - \mathbf{I}_p, \quad \mathbf{H}(z) = \Xi - z\,\mathrm{diag}(\Xi). \end{aligned}$$

Consequently

(1): In both elliptical and linear cases, the terms O_1 and O_2 converge in probability to zero; thus, they have no contribution to the limit properties of $M_n(z)$.

(2): In both elliptical and linear cases, the terms M_1 and M_2 converge in probability to their means $\mathrm{E}\,M_1$ and $\mathrm{E}\,M_2$; thus, they have contributions to the limit of the mean of $M_n(z)$ but do not contribute to the limit of the variance–covariance function of $M_n(z)$. Moreover, we have in the elliptical case

$$z\,\mathrm{E}\,M_1 + z^2\,\mathrm{E}\,M_2 \to E_1(z) = \lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{p}\sum_{\ell=1}^{p} r_{k\ell}^2\frac{\partial}{\partial z}\left(\left[\mathbf{e}_k^T\left(\mathbf{I}+\underline{s}(z)\mathbf{R}\right)^{-1}\mathbf{e}_\ell\right]^2\right).$$

While in the linear case,

$$\begin{aligned}E_1(z) =& \frac{\beta_x}{2}\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{p}\sum_{\ell=1}^{p}\frac{\partial}{\partial z}\left(\left[\mathbf{e}_k^T\left(\mathbf{I}+\underline{s}(z)\mathbf{R}\right)^{-1}\mathbf{e}_\ell\right]^2\right)\sum_{j=1}^{p}\left(\mathbf{e}_\ell^T\mathbf{G}\mathbf{e}_j\mathbf{e}_j^T\mathbf{G}^T\mathbf{e}_k\right)^2\\ &+\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{p}\sum_{\ell=1}^{p}\frac{\partial}{\partial z}\left(\left[\mathbf{e}_k^T\left(\mathbf{I}+\underline{s}(z)\mathbf{R}\right)^{-1}\mathbf{e}_\ell\right]^2\right)\left(\mathbf{e}_k^T\mathbf{R}\mathbf{e}_\ell\right)^2.\end{aligned}$$

(3): The term M_0 converges to certain limits in both cases. The limit of M_0 in the elliptical case is

$$\begin{aligned}E_0(z) =& y\int\frac{\left(\underline{s}'(z)t\right)^2 dH(t)}{\underline{s}(z)(1+t\underline{s}(z))^3}+(\tau-2)\left(1+z\underline{s}(z)\right)\int\frac{\underline{s}'(z)t dH(t)}{(1+t\underline{s}(z))^2}\\ &+\lim_{n\to\infty}\frac{2}{n}\sum_{k=1}^{p}\frac{\partial}{\partial z}\left(\underline{s}(z)(\mathbf{e}_k^T\left(\mathbf{I}+\underline{s}(z)\mathbf{R}\right)^{-1}\mathbf{R}\mathbf{e}_k)^2\right).\end{aligned}$$

The limit of M_0 in the linear case is

$$\begin{aligned}E_0(z) =& y\int\frac{\left(\underline{s}'(z)t\right)^2 dH(t)}{\underline{s}(z)(1+t\underline{s}(z))^3}+\beta_x y\int\frac{\underline{s}'(z)\underline{s}(z)t^2 dH(t)}{(1+t\underline{s}(z))^3}\\ &+\lim_{n\to\infty}\frac{2}{n}\sum_{k=1}^{p}\frac{\partial}{\partial z}\left(\underline{s}(z)(\mathbf{e}_k^T\left(\mathbf{I}+\underline{s}(z)\mathbf{R}\right)^{-1}\mathbf{R}\mathbf{e}_k)^2\right)\\ &+\lim_{n\to\infty}\frac{\beta_x}{n}\sum_{k,h=1}^{p}g_{kh}^2\frac{\partial}{\partial z}\left(\underline{s}(z)(\mathbf{e}_h^T\mathbf{G}^T\left(\mathbf{I}+\underline{s}(z)\mathbf{R}\right)^{-1}\mathbf{e}_k)^2\right).\end{aligned}$$

(4): The term V converges weakly to a zero mean Gaussian process in both cases. The process is tight in both cases. The variance–covariance function is

$$\begin{aligned}&v(z_1,z_2)\\ =&2\left[\frac{\underline{s}'(z_1)\underline{s}'(z_2)}{(\underline{s}(z_2)-\underline{s}(z_1))^2}-\frac{1}{(z_1-z_2)^2}\right]\\ &+\lim_{n\to+\infty}\frac{2}{n}\sum_{k,\ell=1}^{p}r_{k\ell}^2\frac{\partial}{\partial z_1}\mathbf{e}_k^T\left(\mathbf{I}+\underline{s}(z_1)\mathbf{R}\right)^{-1}\mathbf{e}_k\frac{\partial}{\partial z_2}\mathbf{e}_\ell^T\left(\mathbf{I}+\underline{s}(z_2)\mathbf{R}\right)^{-1}\mathbf{e}_\ell\end{aligned}$$

$$+\lim_{n\to+\infty}\frac{2}{n}\sum_{k=1}^{p}\frac{\partial}{\partial z_2}\left\{\frac{\mathbf{e}_k^T\left(\mathbf{I}+\underline{s}(z_2)\mathbf{R}\right)^{-1}\mathbf{e}_k-1}{\underline{s}(z_2)}\right\}\frac{\partial}{\partial z_1}\mathbf{e}_k^T\left(\mathbf{I}+\underline{s}(z_1)\mathbf{R}\right)^{-1}\mathbf{e}_k$$

$$+\lim_{n\to+\infty}\frac{2}{n}\sum_{k=1}^{p}\frac{\partial}{\partial z_1}\left\{\frac{\mathbf{e}_k^T\left(\mathbf{I}+\underline{s}(z_1)\mathbf{R}\right)^{-1}\mathbf{e}_k-1}{\underline{s}(z_1)}\right\}\frac{\partial}{\partial z_2}\mathbf{e}_k^T\left(\mathbf{I}+\underline{s}(z_2)\mathbf{R}\right)^{-1}\mathbf{e}_k$$

in the elliptical case. While the variance–covariance function in the linear case becomes

$$\begin{aligned}
&v(z_1,z_2)\\
=&2\left[\frac{\underline{s}'(z_1)\underline{s}'(z_2)}{(\underline{s}(z_2)-\underline{s}(z_1))^2}-\frac{1}{(z_1-z_2)^2}\right]+\beta_x y\underline{s}'(z_1)\underline{s}'(z_2)\int\frac{t^2dH(t)}{(1+\underline{s}(z_1))^2(1+\underline{s}(z_2))^2}\\
&+\lim_{n\to+\infty}\frac{1}{n}\sum_{k,\ell=1}^{p}\left(\beta_x\sum_{j=1}^{p}g_{kj}^2g_{\ell j}^2+2r_{k\ell}^2\right)\left[\frac{\partial}{\partial z_1}\mathbf{e}_k^T(\mathbf{I}+\underline{s}(z_1)\mathbf{R})^{-1}\mathbf{e}_k\frac{\partial}{\partial z_2}\mathbf{e}_\ell^T(\mathbf{I}+\underline{s}(z_2)\mathbf{R})^{-1}\mathbf{e}_\ell\right]\\
&+\lim_{n\to+\infty}\frac{2}{n}\sum_{k=1}^{p}\frac{\partial}{\partial z_1}\mathbf{e}_k^T(\mathbf{I}+\underline{s}(z_1)\mathbf{R})^{-1}\mathbf{e}_k\frac{\partial}{\partial z_2}\left\{\frac{\mathbf{e}_k^T\left(\mathbf{I}+\underline{s}(z_2)\mathbf{R}\right)^{-1}\mathbf{e}_k-1}{\underline{s}(z_2)}\right\}\\
&+\lim_{n\to+\infty}\frac{2}{n}\sum_{k=1}^{p}\frac{\partial}{\partial z_2}\mathbf{e}_k^T(\mathbf{I}+\underline{s}(z_2)\mathbf{R})^{-1}\mathbf{e}_k\frac{\partial}{\partial z_1}\left\{\frac{\mathbf{e}_k^T\left(\mathbf{I}+\underline{s}(z_1)\mathbf{R}\right)^{-1}\mathbf{e}_k-1}{\underline{s}(z_1)}\right\}\\
&+\lim_{n\to+\infty}\frac{\beta_x}{n}\sum_{k,\ell=1}^{p}g_{k\ell}^2\frac{\partial}{\partial z_1}\mathbf{e}_k^T(\mathbf{I}+\underline{s}(z_1)\mathbf{R})^{-1}\mathbf{e}_k\frac{\partial}{\partial z_2}\left\{\underline{s}(z_2)\mathbf{e}_\ell^T\mathbf{G}^T(\mathbf{I}+\underline{s}(z_2)\mathbf{R})^{-1}\mathbf{G}\mathbf{e}_\ell\right\}\\
&+\lim_{n\to+\infty}\frac{\beta_x}{n}\sum_{k,\ell=1}^{p}g_{k\ell}^2\frac{\partial}{\partial z_2}\mathbf{e}_k^T(\mathbf{I}+\underline{s}(z_2)\mathbf{R})^{-1}\mathbf{e}_k\frac{\partial}{\partial z_1}\left\{\underline{s}(z_1)\mathbf{e}_\ell^T\mathbf{G}^T(\mathbf{I}+\underline{s}(z_1)\mathbf{R})^{-1}\mathbf{G}\mathbf{e}_\ell\right\}.
\end{aligned}$$

Combining (1)–(4) and (A.6), we conclude that the random vector

$$\left(\sum_{i=1}^{p}g_1(\hat{\lambda}_i)-p\int g_1(x)\,dF^{y_n,H_n}(x),\ldots,\sum_{i=1}^{p}g_K(\hat{\lambda}_i)-p\int g_K(x)\,dF^{y_n,H_n}(x)\right)\tag{A.8}$$

converges to a K-dimensional normal random vector $\left(X_{g_1},\ldots,X_{g_K}\right)$ in both cases. The mean function is

$$\mathrm{E}X_{g_\ell}=-\frac{1}{2\pi\mathbf{i}}\oint g_\ell(z)(E_0(z)+E_1(z))dz,\quad 1\le\ell\le K.$$

The variance–covariance function is

$$\mathrm{Cov}(X_{g_{\ell_1}},X_{g_{\ell_2}})=-\frac{1}{4\pi^2}\oint\oint g_{\ell_1}(z_1)g_{\ell_2}(z_2)v(z_1,z_2)dz_1dz_2,\quad 1\le\ell_1,\ell_2\le K.$$

This completes the proof of Theorems 3.1 and 4.1.

Appendix B: Proof of Theorem 3.2 and Example 3.1

B.1. Proof of Theorem 3.2

Let $\underline{s}(z) = -\frac{1}{1+\sqrt{y}\xi}$; then, we have $dz = h(1-\xi^{-2})d\xi$ and $z = 1 + h\xi + h\xi^{-1} + h^2 = |1 + h\xi|^2$, where $|\xi| = 1$ and $h = \sqrt{y}$. When ξ runs counter clockwise on the unit circle, z runs counter clockwise on a contour that encloses the support interval $[a,b] = [(1-h)^2,(1+h)^2]$. Thus, we regard z as a function of ξ. Theorem 3.2 follows.

B.2. Proof of Example 3.1

First, notice that when $\mathbf{R} = \mathbf{I}$, we have

$$\underline{s}'(z) = \frac{\underline{s}^2(z)[1+\underline{s}(z)]^2}{[1+\underline{s}(z)]^2 - y\underline{s}^2(z)},$$

since

$$z = -\frac{1}{\underline{s}(z)} + \frac{y}{1+\underline{s}(z)} = -\frac{-[1+(1-y)\underline{s}(z)]}{\underline{s}(z)[1+\underline{s}(z)]} = -(1-y)\frac{[\underline{s}(z)+(1-y)^{-1}]}{\underline{s}(z)[1+\underline{s}(z)]}.$$

Let $g_1(x) = x$, $g_2(x) = x^2$, $g_3(x) = x^3$, $g_4(x) = x^4$, and $g_5(x) = \log(x)$. We know that the moments of the standard M-P distribution with index y take the values

$$m_k(y) = \sum_{r=0}^{k-1} \frac{1}{r+1}\binom{r}{k}\binom{k-1}{r} y^k,$$

see Lemma 3.1 in [2]. From this, we can easily calculate the centering terms

$$\int g_1(x) f^{y_{n-1}}(x)dx = m_1(y_{n-1}) = 1, \quad \int g_2(x) f^{y_{n-1}}(x)dx = m_2(y_{n-1}) = 1 + y_{n-1},$$

$$\int g_3(x) f^{y_{n-1}}(x)dx = m_3(y_{n-1}) = 1 + 3y_{n-1} + y_{n-1}^2,$$

$$\int g_4(x) f^{y_{n-1}}(x)dx = m_4(y_{n-1}) = 1 + 6y_{n-1} + 6y_{n-1}^2 + y_{n-1}^3.$$

We also have the centering term

$$\int g_5(x) f^{y_{n-1}}(x)dx = \frac{y_{n-1}-1}{y_{n-1}}\log(1-y_{n-1}) - 1, \quad y_{n-1} < 1,$$

see Section 9.12.3 in [2]. Denote

$$A(g) = -\frac{1}{2\pi i}\oint g(z)\frac{y\left(\underline{s}'(z)\right)^2}{\underline{s}(z)(1+\underline{s}(z))^3}\,dz,$$

$$B(g) = -\frac{1}{2\pi i}\oint g(z)(\tau-4)\left(1+z\underline{s}(z)\right)\frac{\underline{s}'(z)}{(1+\underline{s}(z))^2}\,dz,$$

One can also derive from Theorem 3.2 and some calculations that

$$
\begin{aligned}
A(g_1) &= 0, \quad B(g_1) = 0,\\
A(g_2) &= y, \quad B(g_2) = y,\\
A(g_3) &= 3y(1+y),\\
B(g_3) &= 3y(1+y),\\
A(g_4) &= 6y^2 + 6(1+y)^2 y,\\
B(g_4) &= 4y^2 + 6(1+y)^2 y,\\
A(g_5) &= \frac{\log(1-y)}{2},\\
B(g_5) &= -\frac{y}{2}.
\end{aligned}
$$

The results of this example follows.

Acknowledgements

We are grateful to the Editor, the Associate Editor and two referees for their constructive comments, which helped us to improve the manuscript. Yanqing Yin is partially supported by NSFC 11801234. Shurong Zheng and Tingting Zou are the corresponding authors who were partially supported by NSFC grant 12071066 and KLAS.

Supplementary Material

Supplement on "Central limit theorem of linear spectral statistics of high-dimensional sample correlation matrices" (DOI: 10.3150/22-BEJ1487SUPP; .pdf). This supplementary material consists of detailed proofs of Theorems 3.1 and 4.1.

References

[1] Anderson, T.W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. *Wiley Series in Probability and Statistics*. Hoboken, NJ: Wiley. MR1990662

[2] Bai, Z. and Silverstein, J.W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*, 2nd ed. *Springer Series in Statistics*. New York: Springer. MR2567175 https://doi.org/10.1007/978-1-4419-0661-8

[3] Bai, Z.D. and Silverstein, J.W. (2004). CLT for linear spectral statistics of large-dimensional sample covariance matrices. *Ann. Probab.* **32** 553–605. MR2040792 https://doi.org/10.1214/aop/1078415845

[4] Bao, Z., Pan, G. and Zhou, W. (2012). Tracy-Widom law for the extreme eigenvalues of sample correlation matrices. *Electron. J. Probab.* **17** 1-32. MR2988403 https://doi.org/10.1214/EJP.v17-1962

[5] El Karoui, N. (2009). Concentration of measure and spectra of random matrices: Applications to correlation matrices, elliptical distributions and beyond. *Ann. Appl. Probab.* **19** 2362–2405. MR2588248 https://doi.org/10.1214/08-AAP548

[6] Fang, K.T. and Zhang, Y.T. (1990). *Generalized Multivariate Analysis*. Berlin: Springer. MR1079542

[7] Gao, J., Han, X., Pan, G. and Yang, Y. (2017). High dimensional correlation matrices: The central limit theorem and its applications. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 677–693. MR3641402 https://doi.org/10.1111/rssb.12189

[8] Hu, J., Li, W., Liu, Z. and Zhou, W. (2019). High-dimensional covariance matrices in elliptical distributions with application to spherical test. *Ann. Statist.* **47** 527–555. MR3909941 https://doi.org/10.1214/18-AOS1699
[9] Jiang, T. (2004). The limiting distributions of eigenvalues of sample correlation matrices. *Sankhyā* **66** 35–48. MR2082906
[10] Jiang, T. (2019). Determinant of sample correlation matrix with application. *Ann. Appl. Probab.* **29** 1356–1397. MR3914547 https://doi.org/10.1214/17-AAP1362
[11] Johnstone, I.M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. MR1863961 https://doi.org/10.1214/aos/1009210544
[12] Jonsson, D. (1982). Some limit theorems for the eigenvalues of a sample covariance matrix. *J. Multivariate Anal.* **12** 1–38. MR0650926 https://doi.org/10.1016/0047-259X(82)90080-X
[13] Marčenko, V.A. and Pastur, L.A. (1967). Distribution of eigenvalues for some sets of random matrices. *Math. USSR, Sb.* **1** 457–483.
[14] Mestre, X. and Vallet, P. (2017). Correlation tests and linear spectral statistics of the sample correlation matrix. *IEEE Trans. Inf. Theory* **63** 4585–4618. MR3666978 https://doi.org/10.1109/TIT.2017.2689780
[15] Morales-Jimenez, D., Johnstone, I.M., McKay, M.R. and Yang, J. (2021). Asymptotics of eigenstructure of sample correlation matrices for high-dimensional spiked models. *Statist. Sinica* **31** 571–601. MR4286186 https://doi.org/10.5705/ss.20
[16] Pan, G.M. and Zhou, W. (2008). Central limit theorem for signal-to-interference ratio of reduced rank linear receiver. *Ann. Appl. Probab.* **18** 1232–1270. MR2418244 https://doi.org/10.1214/07-AAP477
[17] Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17** 1617–1642. MR2399865
[18] Pillai, N.S. and Yin, J. (2012). Edge universality of correlation matrices. *Ann. Statist.* **40** 1737–1763. MR3015042 https://doi.org/10.1214/12-AOS1022
[19] Schott, J.R. (2005). Testing for complete independence in high dimensions. *Biometrika* **92** 951–956. MR2234197 https://doi.org/10.1093/biomet/92.4.951
[20] Silverstein, J.W. (1995). Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices. *J. Multivariate Anal.* **55** 331–339. MR1370408 https://doi.org/10.1006/jmva.1995.1083
[21] Tian, X., Lu, Y. and Li, W. (2015). A robust test for sphericity of high-dimensional covariance matrices. *J. Multivariate Anal.* **141** 217–227. MR3390068 https://doi.org/10.1016/j.jmva.2015.07.010
[22] Wachter, K.W. (1978). The strong limits of random matrix spectra for sample matrices of independent elements. *Ann. Probab.* **6** 1–18. MR0467894 https://doi.org/10.1214/aop/1176995607
[23] Wen, J. and Zhou, W. (2019). Tracy-Widom limit for the largest eigenvalue of high-dimensional covariance matrices in elliptical distributions. arXiv preprint. Available at arXiv:1901.05166.
[24] Xiao, H. and Zhou, W. (2010). Almost sure limit of the smallest eigenvalue of some sample correlation matrices. *J. Theoret. Probab.* **23** 1–20. MR2591901 https://doi.org/10.1007/s10959-009-0270-2
[25] Yin, Y.Q., Li, C.C., Tian, G.L. and Zheng, S.R. (2021). Spectral properties of rescaled sample correlation matrix. Accepted by Statistica Sinica.
[26] Yin, Y., Zheng, S. Zou, T. (2023). Supplement to "Central Limit Theorem of Linear Spectral Statistics of High-dimensional Sample Correlation Matrices." https://doi.org/10.3150/22-BEJ1487SUPP
[27] Zheng, S., Bai, Z. and Yao, J. (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. MR3316190 https://doi.org/10.1214/14-AOS1292

Received November 2021 and revised February 2022