Statistica Sinica Preprint No: SS-2024-0395	
Title	Large Dimensional Spearman's Rank Correlation Matrices: The Central Limit Theorem and Its Applications
Manuscript ID	SS-2024-0395
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0395
Complete List of Authors	Hantao Chen and Cheng Wang
Corresponding Authors	Cheng Wang
E-mails	chengwang@sjtu.edu.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

Large dimensional Spearman's rank correlation matrices: The central limit theorem and its applications

Hantao Chen and Cheng Wang

School of Mathematical Sciences, MOE-LSC, Shanghai Jiao Tong University

Abstract: This paper is concerned with Spearman's correlation matrices under large dimensional regime, in which the data dimension diverges to infinity proportionally with the sample size. We establish the central limit theorem for the linear spectral statistics of Spearman's correlation matrices, which extends the results of Bao et al. (2015). We also study the improved Spearman's correlation matrices of Hoeffding (1948) which is a standard U-statistic of order 3. As applications, we propose three new test statistics for large dimensional independent test and numerical studies demonstrate the applicability of our proposed methods.

Key words and phrases: Central limit theorem, large-dimensional independence test, linear spectral statistics, Spearman's rank correlation matrices, U-statistics.

1. Introduction

In multivariate statistical analyses, the covariance matrix is a fundamental tool used to describe the relationships among features. Its theoretical property is crucial for understanding many statistical methods. For the classical setting where the data dimension p is fixed and the sample size n tends to infinity, these properties and their applications in

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

various methods are summarized in textbooks, e.g., Anderson (2003).

In the last few decades, large amounts of work are focused on the large dimensional regime,

$$n \to \infty, \quad p = p(n) \to \infty, \quad p/n = y_n \to y \in (0, \infty). \tag{1.1}$$

Random matrix theory, as a powerful tool, provides insights into the behavior of large dimensional sample covariance matrices, extending the famous Wishart distribution theory. The pioneering work Marčenko and Pastur (1967) derived the limiting spectral distribution (LSD) which is called Marčenko-Pastur (MP) law. With the LSD, we can describe the limits of linear spectral statistics (LSS). Furthermore, Bai and Silverstein (2004) firstly derived the central limit theorem (CLT) of LSS. The following works include Pan and Zhou (2008); Anderson and Zeitouni (2008); Lytova and Pastur (2009); Pan (2014); Zheng et al. (2015) and so on. As applications of RMT on the sample covariance matrix, Dobriban and Wager (2018) and Wang and Jiang (2018) studied the prediction errors of ridge regression and regularized linear discriminant analysis; Hastie et al. (2022) demonstrated the double descent phenomenon in the simple linear regression; Bai et al. (2009) proposed a bias correction to the likelihood ratio test; Wang et al. (2013) and Wang and Yao (2013) considered the identify test and the sphericity test of covariance matrices, respectively. For more results on large dimensional covariance matrices, it is referred to Paul and Aue (2014) and Yao et al. (2015) for a comprehensive review.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

Normalization is a common procedure in data analysis. By standardizing the sample covariance matrix, we obtain Pearson's correlation matrix, a scale-invariant measure. Recent research has extensively studied Pearson's correlation matrices. Jiang (2004b) first derived the limiting spectral distribution. Bao et al. (2012) and Pillai and Yin (2012) established limiting distributions for the extreme eigenvalues. Mestre and Vallet (2017) and Gao et al. (2017) developed the CLT of LSS of Pearson's correlation matrices. Zheng et al. (2019) extended the CLT to general covariance structures. See also Jiang (2019) and Parolya et al. (2024). For a large class of population distributions, El Karoui (2009) demonstrated that the spectral properties of Pearson's correlation matrices resemble those of sample covariance matrices. Typically, these studies assume finite fourth moments for the features. However, for distributions with infinite fourth moments, such as heavy-tailed populations, the applicability of these results may require additional verification or may no longer hold. For instance, Heiny and Parolya (2024) justified the CLT of log-determinant statistics of Pearson's correlation matrices and Heiny and Yao (2022) discovered a new LSD result for heavy-tailed distributions.

To address the challenges posed by heavy-tailed data, non-parametric statistics offer robust correlation measures. Among these, Spearman's rank correlation matrix and Kendall's rank correlation matrix are particularly popular due to their distribution-free nature, making them suitable for heavy-tailed data. Recent research has explored the properties of these rank-based correlation matrices. For example, Leung and Drton (2018) and Wang et al. (2024) studied a class of rank-based U-statistics for independence test.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

In the realm of random matrix theory, Bai and Zhou (2008) and Wu and Wang (2022) investigated the LSD of Spearman's correlation matrices. Bandeira et al. (2017) and Li et al. (2023) studied the LSD of Kendall's correlation matrices. As far as the CLT, Bao et al. (2015) considered asymptotic distributions of polynomial functions of Spearman's correlation matrices and Li et al. (2021) studied the CLT of LSS of Kendall's correlation matrices.

In this paper, we focus on Spearman's correlation matrices and aim to establish a central limit theorem for general linear spectral statistics. Due to introducing ranking, the independence among samples are violated and thus, we turn to consider Gram matrices. The rescaled Gram matrix is a sample covariance matrix related to the distribution which are independent and uniformly distributed on the permutations of $\{1, \cdots, n\}$. In Bao et al. (2015), they adopted the celebrated moment method and derived the CLT for polynomial functions. In this work, we follow the classical technique developed by Bai and Silverstein (2004) and consider the asymptotic distribution of Stieltjes transforms. Key challenges arise in computing the covariance of quadratic forms and establishing concentration inequalities for these forms. For uniform distribution on $\{1, \cdots, n\}$, it is challenging to derive the explicit covariance of quadratic forms. We derive the three leading terms which all contribute to the final CLT. More details can be found in our Lemma S2.1 and Lemma S2.2 of Supplement Materials. The obtained results are consistent with Bao et al. (2015) for polynomial functions and are also applicable to more general LSS such as log-determinant functions. The resulting CLT of Stieltjes transform can connect to many

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

other covariance or correlation matrices.

In non-parametric statistics, Hoeffding (1948) theoretical analyzed the Spearman's correlation from the perspective of U-statistics and proposed an improved version. Specifically, Spearman's correlation can be expressed as a U-statistics of order 3 with an additional term. To address this, Hoeffding (1948) introduced an improved Spearman's correlation which is a standard U-statistic of order 3. Sample covariance matrices and Kendall's correlation matrices are well-known examples of U-statistics of order 2, and their CLTs have been extensively studied in Pan (2014) and Li et al. (2021), respectively. To the best of our knowledge, there are no CLTs for general LSS of U-statistics of order higher than 2. While the improved Spearman's correlation matrix is challenging to analyze directly, we can evaluate the difference between it and the classical Spearman's correlation matrix. This approach enables us to establish a CLT for standard U-statistics of order 3. This result is of interest for covariance/correlation matrices of U-statistic types and may contribute to the development of CLTs for LSS of general U-statistics of higher order.

As applications of such CLTs, we revisit hypothesis testing for independence. Numerous studies have proposed various test statistics based on different correlation matrices, including Jiang (2004a), Zhou (2007), Gao et al. (2017), Bao et al. (2015), Leung and Drton (2018), Bao (2019), Li et al. (2021). Our proposed test statistics fall into two categories: those based on Euclidean distance and those based on Stein's loss. Through extensive numerical experiments, we demonstrate the competitive performance of our proposed methods compared to well-established approaches.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

Our contributions are summarized as follows:

1. For Gram matrices, we study a novel population distribution which is uniformly distributed on the permutations of $\{1, \cdots, n\}$. Unlike the independent component model or elliptical distributions, the quadratic forms associated with this distribution exhibit a complex covariance structure. By carefully analyzing three leading terms, we derive a new central limit theorem.

2. For Spearman's correlation matrices, we establish a CLT of general linear spectral statistics, extending the work of Bao et al. (2015) which focused on polynomial functions. Our approach, based on classical random matrix techniques and the Stieltjes transform, provides a more direct connection to other classical results, shedding light on the underlying structure of Spearman's correlation.

3. From a U-statistic perspective, Spearman's correlation is not a standard U-statistic. Hoeffding (1948) proposed an improved version which is a U-statistic of order 3. By carefully evaluating the difference between the classical and improved Spearman's correlation matrices, we derive the explicit impact on the asymptotic mean and establish a CLT for the improved Spearman's correlation matrix. As we know, this is the first CLT for standard U-statistic of order 3 in random matrix theory.

4. Spearman's correlation matrices, derived from ranking and standardizing the original data matrix, can be viewed as both sample covariance and Pearson-type correlation matrices. From a U-statistic perspective, Spearman's correlation matrices are of

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

order 3, while Kendall's correlation matrices are of order 2. Thus, Spearman's correlation matrices can be connected with many existing random matrix models and the corresponding CLT results can also be connected with well-established CLT results. The obtained results allow us to gain deeper insights into the asymptotic distribution of linear spectral statistics for various sample covariance and correlation matrices.

The remainder of the paper is structured as follows: Section 2 introduces the necessary background knowledge and tools from random matrix theory. Section 3 presents our main results, including the CLTs for Gram matrices, Spearman's correlation matrices, and improved Spearman's correlation matrices. Section 4 applies our theoretical results to hypothesis testing for independence and conducts numerical experiments to demonstrate the effectiveness of our proposed methods. In Section 5, we summarize our CLTs with discussions and the Appendix provides detailed proofs of our theoretical results.

2. Preliminary result in RMT

Let $\mathbf{H}_n$ be any $n \times n$ Hermitian matrix with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$. The empirical spectral distribution (ESD) is defined as

$$F^{\mathbf{H}_n}(x) = \frac{1}{n}\sum_{i=1}^{n} I(\lambda_i \leq x), \tag{2.1}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

where $I(\cdot)$ is the indicator function. If $F^{\mathbf{H}_n}$ converges weakly to some limiting distribution F, then we call F the limiting spectral distribution of $\mathbf{H}_n$.

With the LSD, we can study the linear spectral statistic which is defined as

$$\frac{1}{n}\sum_{i=1}^{n} f(\lambda_i) = \int f(x) dF^{\mathbf{H}_n}(x).$$

Here $f(\cdot)$ is any bounded and continuous function. By the property of weak convergence, we can conclude

$$\int f(x) dF^{\mathbf{H}_n}(x) \to \int f(x) dF(x).$$

Some common functions in statistics include

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \lambda_i^k =& \frac{1}{n}\mathrm{tr}(\mathbf{H}_n^k), \ k = 1, 2, \cdots, \\
\frac{1}{n}\sum_{i=1}^{n} (\lambda_i - 1)^2 =& \frac{1}{n}\|\mathbf{H}_n - \mathbf{I}_n\|_F^2, \\
\frac{1}{n}\sum_{i=1}^{n} \lambda_i - \log(\lambda_i) - 1 =& \frac{1}{n}\mathrm{tr}(\mathbf{H}_n) - \frac{1}{n}\log\det(\mathbf{H}_n) - 1,
\end{aligned}$$

and so on. If $\mathbf{H}_n$ is a random matrix, we can further consider the central limit theorem of linear spectral statistics.

In random matrix theory, one of the most powerful tools is Stieltjes transform, which

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

is defined as

$$m_F(z) = \int \frac{1}{x - z} dF(x), \quad z \in \mathbb{C}^+, \tag{2.2}$$

with respect to any distribution function F. Here $\mathbb{C}^+$ is the upper half space of the complex plane. Similar to the characteristic function in probability, there is a one-to-one correspondence between the probability distribution and its Stieltjes transform. With the Stieltjes transform, by the residue theorem of complex analysis,

$$\frac{1}{n}\sum_{i=1}^{n} f(\lambda_i) = \int f(x) dF^{\mathbf{H}_n}(x) = -\frac{1}{2\pi i} \oint_{\mathcal{C}} f(z) m_{F^{\mathbf{H}_n}}(z) dz,$$

where $\oint_{\mathcal{C}}$ is closed and taken in the positive direction, enclosing the support of $F^{\mathbf{H}_n}$. Furthermore, we can study the asymptotic distribution, e.g.,

$$\int f(x) dF^{\mathbf{H}_n}(x) - \int f(x) dF(x) = \frac{1}{2\pi i} \oint_{\mathcal{C}} f(z) \left(m_F(z) - m_{F^{\mathbf{H}_n}}(z) \right) dz.$$

In summary, to find the LSD of a random matrix $\mathbf{H}_n$, we can study its Stieltjes transform

$$m_{F^{\mathbf{H}_n}}(z) = \frac{1}{n} \mathrm{tr} \left(\mathbf{H}_n - z\mathbf{I}_n \right)^{-1}.$$

To explore the asymptotic distribution of the LSS, we need to find the asymptotic distri-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

bution of

$$m_F(z) - m_{F^{\mathbf{H}_n}}(z),$$

which is usually a Gaussian process. The Gaussian process further yields the asymptotically normal distribution of the LSS. It is referred to Bai and Silverstein (2010) for a comprehensive survey on random matrix theory.

3. Main result

For independent and identically distributed (i.i.d.) samples $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^p$, we denote their rank statistics as $\mathbf{r}_i = (r_{i1}, \cdots, r_{ip})^\top,\ i = 1, \ldots, n$. For each feature $j \in \{1, \cdots, p\}$, $(r_{1j}, \ldots, r_{nj})$ are uniformly distributed on the permutations of $\{1, \cdots, n\}$. Then,

$$\mathbb{E}r_{ij} = \frac{n+1}{2},\ \mathrm{var}(r_{ij}) = \frac{n^2-1}{12}.$$

With the rank statistics, the Spearman's rank correlation matrix is defined by

$$\boldsymbol{\rho}_n = \frac{12}{n(n^2-1)} \sum_{k=1}^{n} (\mathbf{r}_k - \frac{n+1}{2}\mathbf{1}_p)(\mathbf{r}_k - \frac{n+1}{2}\mathbf{1}_p)^\top, \tag{3.1}$$

which is the Pearson's correlation matrix based on $\mathbf{r}_1, \ldots, \mathbf{r}_n \in \mathbb{R}^p$. Due to ranking, $\mathbf{r}_1, \ldots, \mathbf{r}_n \in \mathbb{R}^p$ are not independent anymore and it is hard to tackle the Spearman's rank correlation matrix directly. Here we turn to its Gram matrix.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

3.1 Gram matrix

Standardizing rank statistics, we denote

$$\sqrt{\frac{12}{n^2-1}}\begin{pmatrix} r_{11}-\frac{n+1}{2} & \cdots & r_{1p}-\frac{n+1}{2} \\ \vdots & \cdots & \vdots \\ r_{n1}-\frac{n+1}{2} & \cdots & r_{np}-\frac{n+1}{2} \end{pmatrix} = \begin{pmatrix} \mathbf{s}_1, & \cdots, & \mathbf{s}_p \end{pmatrix}.$$

If the features are completely independent, $\mathbf{s}_1, \ldots, \mathbf{s}_p \in \mathbb{R}^n$ are i.i.d. and have been centered, e.g.,

$$\mathbb{E}\mathbf{s}_i = \mathbf{0}_n,\ \boldsymbol{\Sigma} = \mathrm{cov}(\mathbf{s}_1) = \frac{n}{n-1}\left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\right).$$

Then, we can study the sample covariance matrix of $\mathbf{s}_1, \ldots, \mathbf{s}_p$,

$$\mathbf{g}_n = \frac{1}{p}\sum_{i=1}^{p}\mathbf{s}_i\mathbf{s}_i^\top. \tag{3.2}$$

This sample covariance matrix can also be regarded as the Gram matrix of the original rank statistics, that is,

$$\mathbf{g}_n = \frac{12}{p(n^2-1)}\left((\mathbf{r}_i - \frac{n+1}{2}\mathbf{1}_p)^\top(\mathbf{r}_j - \frac{n+1}{2}\mathbf{1}_p)\right)_{n\times n}.$$

Thus, $\mathbf{g}_n$ and $\boldsymbol{\rho}_n/y_n$ share the same non-zero eigenvalues.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

Denoting $m_n(z)$ as the Stieltjes transforms of $\boldsymbol{\rho}_n$ i.e.,

$$m_n(z) = \frac{1}{p}\mathrm{tr}(\boldsymbol{\rho}_n - z\mathbf{I}_p)^{-1},$$

it is proven in Bai and Zhou (2008) that $m_n(z) \to m(z)$ almost surely and

$$m(z) = \frac{1 - y - z + \sqrt{(1+y-z)^2 - 4y}}{2yz}. \tag{3.3}$$

This result shows that the LSD of $\boldsymbol{\rho}_n$ converges weakly to the M-P law F_y almost surely, whose density function is

$$p_y(x) = \frac{\sqrt{(x-(1-\sqrt{y})^2)((1+\sqrt{y})^2 - x)}}{2\pi xy} I\left((1-\sqrt{y})^2 < x < (1+\sqrt{y})^2\right),$$

for $y \leq 1$ and has a point mass $1 - 1/y$ at origin for $y > 1$.

We further denote $s_n(z)$ and $\underline{s}_n(z)$ as the Stieltjes transform of $\mathbf{g}_n$ and $\boldsymbol{\rho}_n/y_n$, respectively

$$s_n(z) = \frac{1}{n}\mathrm{tr}(\mathbf{g}_n - z\mathbf{I}_n)^{-1} = y_n^2 m_n(y_n z) - \frac{1-y_n}{z},$$
$$\underline{s}_n(z) = \frac{1}{p}\mathrm{tr}(\frac{\boldsymbol{\rho}_n}{y_n} - z\mathbf{I}_p)^{-1} = \frac{1}{y_n}(s_n(z) + \frac{1}{z}) - \frac{1}{z},$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

and almost surely

$$s_n(z) \to s(z) = \frac{1 - y_0 - z + \sqrt{(1 + y_0 - z)^2 - 4y_0}}{2y_0 z},$$
$$\underline{s}_n(z) \to \underline{s}(z) = \frac{-(1 - y_0 + z) + \sqrt{(1 + y_0 - z)^2 - 4y_0}}{2z},$$

where $y_0 = 1/y$. Then, the LSD of $\mathbf{g}_n$ converges weakly to the M-P law F_{y_0} almost surely.

For the LSS of $\mathbf{g}_n$,

$$\int f(x) dF^{\mathbf{g}_n}(x) = \frac{1}{n} \sum_{i=1}^{n} f\left(\lambda_i(\mathbf{g}_n)\right),$$

where f is an analytic function and $\lambda_1(\mathbf{g}_n) \geq \cdots \geq \lambda_n(\mathbf{g}_n)$ are eigenvalues of $\mathbf{g}_n$, we have almost surely

$$\int f(x) dF^{\mathbf{g}_n}(x) \to \int f(x) dF_{y_0}(x).$$

Further, we study the asymptotic distribution of the LSS. Let

$$G_n(x) = n\left(F^{\mathbf{g}_n}(x) - F_{n/p}(x)\right),$$

and we focus on

$$\int f(x) dG_n(x) = n\left(\int f(x) dF^{\mathbf{g}_n}(x) - \int f(x) dF_{n/p}(x)\right). \tag{3.4}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

Our central limit theorem is presented as follows.

Theorem 1. Assume that $\{X_{ij} : i = 1, \ldots, n; j = 1, \ldots, p\}$ are doubly independent and absolutely continuous with respect to the Lebesgue measure. Let $f_1, \ldots, f_k$ be functions on $\mathbb{R}$ and analytic on an open interval containing

$$[I(y_0 < 1)(1 - \sqrt{y_0})^2, (1 + \sqrt{y_0})^2]. \tag{3.5}$$

Then, as $n/p \to y_0 \in (0, \infty)$, the random vector

$$\left(\int f_1(x) dG_n(x), \cdots, \int f_k(x) dG_n(x) \right)$$

converges weakly to a Gaussian vector $(G_{f_1}, \cdots, G_{f_k})$ with the asymptotic mean

$$\mathbb{E}G_f = -\frac{1}{2\pi i} \oint_{\mathcal{C}} f(z)\mu(z)dz,$$

and the asymptotic covariance function

$$\operatorname{cov}(G_f, G_g) = -\frac{1}{4\pi^2} \oint_{\mathcal{C}_1} \oint_{\mathcal{C}_2} f(z_1)g(z_2)\sigma(z_1, z_2)dz_1 dz_2,$$

where

$$\mu(z) = \frac{y_0 \underline{s}^3(z) \left(1 + \underline{s}(z)\right)}{\left((1 + \underline{s}(z))^2 - y_0 \underline{s}^2(z)\right)^2} - \frac{2y_0 \underline{s}^3(z)}{\left((1 + \underline{s}(z))^2 - y_0 \underline{s}^2(z)\right)\left(1 + \underline{s}(z)\right)}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

$$+\frac{\underline{s}^3(z)}{(1+\underline{s}(z))^2-y_0\underline{s}^2(z)},$$
$$\sigma(z_1,z_2)=\frac{2\underline{s}'(z_1)\underline{s}'(z_2)}{(\underline{s}(z_1)-\underline{s}(z_2))^2}-\frac{2}{(z_1-z_2)^2}-\frac{2y_0\underline{s}'(z_1)\underline{s}'(z_2)}{(1+\underline{s}(z_1))^2(1+\underline{s}(z_2))^2}.$$

The contour $\oint_{\mathcal{C}}$ is closed and taken in the positive direction, each enclosing the support (3.5).

For concrete functions such as logarithms and polynomials, we will derive CLTs in next section. In details, the integral involving $\underline{s}(z)$ can be calculated explicitly for most cases.

Remark 1. Interestingly, our CLT is quite related to the existing results for sample covariance matrices based on independent components model. For instance, we consider a spike model with population covariance matrix being $\mathrm{diag}(0,\frac{n}{n-1},\cdots,\frac{n}{n-1})$ whose elements are the eigenvalues of $\mathbf{\Sigma}$. The CLT for LSS of the sample covariance matrix is derived by Pan and Zhou (2008), where the centering term is $\int f(x)dF^{y_n,H_n}(x)$, the asymptotic mean is

$$\mathbb{E}G_f=-\frac{1}{2\pi i}\int f(z)\frac{y\underline{m}^3(z)\,(1+\underline{m}(z))}{\left((1+\underline{m}(z))^2-y\underline{m}^2(z)\right)^2}dz$$
$$-\frac{\mathbb{E}Z_{11}^4-3}{2\pi i}\int f(z)\frac{2y\underline{m}^3(z)}{\left((1+\underline{m}(z))^2-y\underline{m}^2(z)\right)(1+\underline{m}(z))}dz,$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

and the asymptotic covariance is

$$\begin{aligned}\operatorname{cov}(G_f, G_g) =& -\frac{1}{4\pi^2}\iint f(z_1)g(z_2)\left(\frac{2\underline{m}'(z_1)\underline{m}'(z_2)}{(\underline{m}(z_1)-\underline{m}(z_2))^2}-\frac{2}{(z_1-z_2)^2}\right)dz_1dz_2\\ &-\frac{\mathbb{E}Z_{11}^4-3}{4\pi^2}\iint f(z_1)g(z_2)\frac{y\underline{m}'(z_1)\underline{m}'(z_2)}{(1+\underline{m}(z_1))^2(1+\underline{m}(z_2))^2}dz_1dz_2.\end{aligned}$$

As can be seen, the asymptotic covariance and the first two terms of asymptotic mean are essentially the same, with only the coefficients differing. For the new term in the asymptotic mean, it comes from the discrepancy between $\mathbf{\Sigma}$ and $\mathbf{I}$. More specifically, the Marčenko-Pastur equation of $F_{n/p}$ is

$$s_n^{(0)}(z)=\frac{1}{(1-\frac{n}{p}-\frac{n}{p}zs_n^{(0)}(z))-z}.$$

For $\mathbf{\Sigma}$ which has $n-1$ eigenvalues equal to $n/(n-1)$ and one zero eigenvalue, the corresponding Marčenko-Pastur equation is

$$\begin{aligned}s_n^{(1)}(z)=&\int\frac{1}{t(1-\frac{n}{p}-\frac{n}{p}zs_n^{(1)}(z))-z}dF^{\mathbf{\Sigma}}(t)\\ =&\frac{n-1}{n}\left[\frac{n}{n-1}\left(1-\frac{n}{p}-\frac{n}{p}zs\right)-z\right]^{-1}-\frac{1}{nz}.\end{aligned}$$

Through careful calculation, we can conclude

$$n\left(s_n^{(1)}(z)-s_n^{(0)}(z)\right)\to\frac{\underline{s}^3(z)}{(1+\underline{s}(z))^2-y_0\underline{s}^2(z)}.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

More details can be found in the proof, e.g., the equation (S1.24) of Supplement Materials.

3.2 Spearman's rank correlation matrix

For Spearman's rank correlation matrix $\boldsymbol{\rho}_n \in \mathbb{R}^{p\times p}$ which has the same non-zero eigenvalues as the ones of $y_n\mathbf{g}_n \in \mathbb{R}^{n\times n}$, we have

$$\begin{aligned}\int f(x)dF^{\boldsymbol{\rho}_n}(x) =&\frac{1}{p}\sum_{i=1}^{p} f(\lambda_i(\boldsymbol{\rho}_n)) = \frac{1}{p}\sum_{i=1}^{n} f(y_n\lambda_i(\mathbf{g}_n)) + \frac{p-n}{p}f(0)\\ =&\frac{1}{y_n}\int f(y_nx)dF^{\mathbf{g}_n}(x) + \frac{p-n}{p}f(0).\end{aligned}$$

In addition, by the property of M-P law,

$$\int f(y_nx)dF_{n/p}(x) = y_n\int f(x)dF_{y_n}(x) + (1-y_n)f(0),$$

which yields

$$\int f(x)dF_{y_n}(x) = \frac{1}{y_n}\int f(y_nx)dF_{n/p}(x) + \frac{p-n}{p}f(0).$$

Therefore, we can study the asymptotic distribution of

$$T(f) = p\left(\int f(x)dF^{\boldsymbol{\rho}_n}(x) - \int f(x)dF_{y_n}(x)\right) = \int f(y_nx)dG_n(x).$$

By Theorem 1, we have proven the CLT of $\int f(yx)dG_n(x)$ and the remaining is to

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

show

$$\int f(y_n x) dG_n(x) - \int f(yx) dG_n(x) = o_p(1),$$

whose details can be found in the proof. Based on these observations, we state the CLT for $\boldsymbol{\rho}_n$ in the following theorem.

Theorem 2. Assume that $\{X_{ij} : i = 1, \ldots, n; j = 1, \ldots, p\}$ are doubly independent and absolutely continuous with respect to the Lebesgue measure. Let $f_1, \cdots, f_k$ be functions analytic on an open interval containing

$$\left[I(y < 1)(1 - \sqrt{y})^2, (1 + \sqrt{y})^2\right]. \tag{3.6}$$

Then, as $p/n \to y \in (0, \infty)$, the random vector $(T(f_1), \cdots, T(f_k))$ converges weakly to a Gaussian vector $(Z_{f_1}, \cdots, Z_{f_k})$ with mean function

$$\mathbb{E}Z_f = -\frac{1}{2\pi i} \oint_{\mathcal{C}'} f(yz)\mu(z)dz = -\frac{1}{2\pi i} \oint_{\mathcal{C}} f(z)\frac{\mu(z/y)}{y}dz, \tag{3.7}$$

and covariance function

$$\begin{aligned} \operatorname{cov}(Z_f, Z_g) &= -\frac{1}{4\pi^2} \oint_{\mathcal{C}_1'} \oint_{\mathcal{C}_2'} f(yz_1)g(yz_2)\sigma(z_1, z_2)dz_1 dz_2, \\ &= -\frac{1}{4\pi^2} \oint_{\mathcal{C}_1} \oint_{\mathcal{C}_2} f(z_1)g(z_2)\frac{\sigma(z_1/y, z_2/y)}{y^2}dz_1 dz_2, \end{aligned} \tag{3.8}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

where $\mu(z)$ and $\sigma(z_1, z_2)$ are defined in Theorem 1, and the contour $\oint_{\mathcal{C}}$ is closed and taken in the positive direction, each enclosing the support (3.6).

Remark 2. In Bao et al. (2015), they derived the asymptotic distribution of $\mathrm{tr}(\boldsymbol{\rho}_n^k)$ for any positive integer $k \geq 2$. Technically, they utilized Anderson and Zeitouni's cumulant method (Anderson and Zeitouni, 2008) and proposed a two-step comparison approach to obtain the explicit mean and covariance of CLTs. Here we adopt the classical proof technique from the seminal work of Bai and Silverstein (2004). Taking $f(x) = x^k$, we refer to results (4.3) in Bao et al. (2015).

Remark 3. Note that $\boldsymbol{\rho}_n$ is a correlation matrix which means $\mathrm{tr}(\boldsymbol{\rho}_n) = p$. Thus, $f(x) = x$ is a degenerate case. In Theorem 4, we derive the asymptotic mean and the asymptotic variance for $f(x) = x^k$ with $k \geq 1$. Taking $k = 1$, we can obtain $\mu_x = 0$ and $\sigma_x^2 = 0$.

Remark 4. For sample covariance matrices, using the sample mean or the true population mean has impacts on the final CLT and Pan (2014) compared the two types of sample covariance matrices. More specifically, Zheng et al. (2015) proposed a substitution principle which adjusted the sample size from n to $n-1$ for the sample covariance matrix based on the sample mean. For Spearman's rank correlation matrices, although the sample mean of the rank statistics is constant, e.g.,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{r}_i = \frac{n+1}{2}\mathbf{1}_p,$$

it still contributes to the CLT. In details, the population covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

related the Gram matrix $\mathbf{g}_n$ has one zero eigenvalue. Following Zheng et al. (2015), we can also use $n-1$ and consider the ratio $p/(n-1)$. Then, the additional term such as $\mu_3(z)$ can be removed and more details can be found in Remark 1.

3.3 Improved Spearman's correlation matrix

The Spearman's rank correlation is a classical method in non-parametric statistics. We note that the rank statistics can be transformed into the sum of indicators, which implies

$$r_{ij} - \frac{n+1}{2} = \frac{1}{2}\sum_{k\neq i} \text{sign}(X_{ij} - X_{kj}).$$

Here $\text{sign}(\cdot)$ is the sign function. Denoting the sign vector

$$\mathbf{A}_{ij} = \text{sign}(\mathbf{X}_i - \mathbf{X}_j) = \begin{pmatrix} \text{sign}(X_{i1} - X_{j1}) \\ \vdots \\ \text{sign}(X_{ip} - X_{jp}) \end{pmatrix},$$

we have

$$\boldsymbol{\rho}_n = \frac{3}{n(n^2-1)} \sum_{i=1}^{n} \sum_{j,k\neq i} \mathbf{A}_{ij}\mathbf{A}_{ik}^{\top}.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

In the form of non-parametric U-statistics, it can be decomposed into two U-statistics,

$$\boldsymbol{\rho}_n = \frac{3}{n(n^2-1)} \sum_{i,j}^{*} \mathbf{A}_{ij}\mathbf{A}_{ij}^{\top} + \frac{3}{n(n^2-1)} \sum_{i,j,k}^{*} \mathbf{A}_{ij}\mathbf{A}_{ik}^{\top} = \frac{3}{n+1}\mathbf{K}_n + \frac{n-2}{n+1}\widetilde{\boldsymbol{\rho}}_n,$$

where

$$\mathbf{K}_n = \frac{1}{n(n-1)} \sum_{i,j}^{*} \mathbf{A}_{ij}\mathbf{A}_{ij}^{\top}$$

is Kendall's rank correlation matrix and

$$\widetilde{\boldsymbol{\rho}}_n = \frac{3}{n(n-1)(n-2)} \sum_{i,j,k}^{*} \mathbf{A}_{ij}\mathbf{A}_{ik}^{\top}$$

is the improved Spearman's rank correlation matrix proposed by Hoeffding (1948). Here $\sum^{*}$ denotes summation over mutually different indices and more details can be found in Wu and Wang (2022).

As can be seen, the Kendall's correlation matrix $\mathbf{K}_n$ is a U-statistic of order 2 and Li et al. (2021) studied the asymptotic distribution of its linear spectral statistics. The improved Spearman's rank correlation matrix $\widetilde{\boldsymbol{\rho}}_n$ is a U-statistic of order 3, that is difficult to analyze it directly. Based on the CLT of $\boldsymbol{\rho}_n$, we can study the difference between $\widetilde{\boldsymbol{\rho}}_n$ and $\boldsymbol{\rho}_n$ which is $3(\mathbf{K}_n - \widetilde{\boldsymbol{\rho}}_n)/(n+1)$. For LSD (Wu and Wang, 2022), this difference can be ignored. However, for the CLT of LSS, this deviation does contribute a non-trivial term

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

to the asymptotic distribution. Considering the centered statistic

$$\widetilde{T}(f) = p\left(\int f(x)dF^{\widetilde{\rho}_n}(x) - \int f(x)dF_{y_n}(x)\right),$$

we present the main result in the following theorem.

Theorem 3. Under the conditions of Theorem 2, as $y_n \to y$, the random vector $(\widetilde{T}(f_1), \cdots, \widetilde{T}(f_k))$ converges weakly to a Gaussian vector $(\widetilde{Z}_{f_1}, \cdots, \widetilde{Z}_{f_k})$ with mean function

$$\mathbb{E}\widetilde{Z}_f = -\frac{1}{2\pi i}\oint_{\mathcal{C}} f(yz)\left(\mu(z) + \widetilde{\mu}(z)\right)dz, \tag{3.9}$$

and covariance function

$$\mathrm{cov}(\widetilde{Z}_f, \widetilde{Z}_g) = -\frac{1}{4\pi^2}\oint_{\mathcal{C}_1}\oint_{\mathcal{C}_2} f(yz_1)g(yz_2)\sigma(z_1, z_2)dz_1dz_2, \tag{3.10}$$

where $\mu(z)$, $\sigma(z_1, z_2)$ are defined in Theorem 1, and

$$\widetilde{\mu}(z) = \frac{\underline{s}^3(z)\left(2 + \underline{s}(z)\right)}{\left(1 + \underline{s}(z)\right)^2 - \underline{s}^2(z)/y}.$$

Compared with Theorem 2 for $\boldsymbol{\rho}_n$, the asymptotic variance is the same and there is a new additional term to the asymptotic mean which is due to the difference $\widetilde{\boldsymbol{\rho}}_n - \boldsymbol{\rho}_n$. Specifically, this difference can be depicted by the discrepancy of their Stieltjes transforms, i.e., $p(m_{F^{\widetilde{\rho}_n}}(z) - m_{F^{\rho_n}}(z))$. Through rigorous derivation, we find that its limit is a non-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

random term. Consequently, it only introduces a drift to the asymptotic mean and does not affect the asymptotic variance. For more details, please refer to the proof in the Supplementary Material.

4. Application

For sample correlation matrices, one important application is to test mutual independence among features. More specifically, we consider the hypotheses testing problem

$$H_0 : \mathbf{R} = \mathbf{I},$$

where $\mathbf{R}$ is a population correlation matrix such as the classical Pearson's correlation matirx, Spearman's correlation matrix, Kendall's correlation matrix and so on.

To evaluate $\mathbf{R} = \mathbf{I}$, we have the following equivalent definitions

$$\begin{aligned} \ell_2 \text{ loss:} \quad & \|\mathbf{R} - \mathbf{I}\|_2^2 = \mathrm{tr}(\mathbf{R}^2) - 2\mathrm{tr}(\mathbf{R}) + p = 0; \\ \text{Stein's loss:} \quad & \mathrm{tr}(\mathbf{R}) - \log(\mathbf{R}) - p = 0; \\ \ell_\infty \text{ loss:} \quad & \|\mathbf{R} - \mathbf{I}\|_\infty = 0. \end{aligned}$$

Based on an estimator $\widehat{\mathbf{R}}$, intuitively one can conduct test statistics

$$T_1(\widehat{\mathbf{R}}) = \mathrm{tr}(\widehat{\mathbf{R}}^2) - 2\mathrm{tr}(\widehat{\mathbf{R}}) + p,$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

$$T_2(\widehat{\mathbf{R}}) = \mathrm{tr}(\widehat{\mathbf{R}}) - \log(\widehat{\mathbf{R}}) - p,$$

$$T_3(\widehat{\mathbf{R}}) = \|\widehat{\mathbf{R}} - \mathbf{I}\|_\infty.$$

For large dimensional data, it is challenging to derive the asymptotic distribution of these statistics. In the past 20 years, significant progress has been made in this field and many important methods were proposed in literature. In special, ℓ_2 loss and Stein's loss can be expressed by linear spectral statistics and in RMT, they motivate the study on the LSS of these correlation matrices. We summarize the developments of testing correlation matrices in Table 1.

Table 1: Developments of testing correlation matrices in RMT

	Sample correlation	Kendall's τ	Spearman's ρ
ℓ_2 loss	Gao et al. (2017); Zheng et al. (2019)	Li et al. (2021)	Bao et al. (2015)
Stein's loss	Gao et al. (2017)	Li et al. (2021)	
ℓ_∞ loss:	Zhou (2007)	Han et al. (2017)	Han et al. (2017)

In special, Chen et al. (2010) proposed a U-statistic's trick to estimate the ℓ_2 loss directly which can obtain a better convergence rate. Using this trick, several statistics based on $T_1(\widehat{\mathbf{R}})$ can be extended to high-dimensional data case where $p \gg n$. See Leung and Drton (2018) for more details.

As application of Theorem 2 for Spearman's correlation matrix $\boldsymbol{\rho}_n$, we can fill the gap for Stein's loss, i.e., obtaining the asymptotic distribution of $T_2(\boldsymbol{\rho}_n)$. Similarly, based on the improved Spearman's correlation matrix $\widetilde{\boldsymbol{\rho}}_n$, we can also conduct three test statistics. Han et al. (2017) has studied $T_3(\widetilde{\boldsymbol{\rho}}_n)$ and here we can derive the distributions of $T_1(\widetilde{\boldsymbol{\rho}}_n)$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

and $T_2(\widetilde{\boldsymbol{\rho}}_n)$.

Taking $f(x)$ being logarithm $\log(x)$ or polynomial x^k for $k \geq 2$, by direct calculations of Theorem 2 and Theorem 3, we obtain the following asymptotic distributions.

Theorem 4. Under the conditions of Theorem 2, we have

$$\log|\boldsymbol{\rho}_n| + (n-p)\log(1-y_n) + p \xrightarrow{d} N(\mu_{\log}, \sigma^2_{\log}), \tag{4.1}$$

$$\log|\widetilde{\boldsymbol{\rho}}_n| + (n-p)\log(1-y_n) + p \xrightarrow{d} N(\widetilde{\mu}_{\log}, \sigma^2_{\log}), \tag{4.2}$$

$$\mathrm{tr}(\boldsymbol{\rho}_n^k) - \sum_{j=0}^{k-1} \frac{p y_n^j}{(j+1)} \binom{k}{j} \binom{k-1}{j} \xrightarrow{d} N(\mu_{x^k}, \sigma^2_{x^k}), \tag{4.3}$$

$$\mathrm{tr}(\widetilde{\boldsymbol{\rho}}_n^k) - \sum_{j=0}^{k-1} \frac{p y_n^j}{(j+1)} \binom{k}{j} \binom{k-1}{j} \xrightarrow{d} N(\widetilde{\mu}_{x^k}, \sigma^2_{x^k}), \tag{4.4}$$

where the asymptotic means are

$$\begin{aligned}
\mu_{\log} =& \frac{3}{2}\log(1-y) + 2y, \\
\widetilde{\mu}_{\log} =& \mu_{\log} - \frac{y^2}{1-y}, \\
\mu_{x^k} =& \frac{1}{4}\left[(1-\sqrt{y})^{2k} + (1+\sqrt{y})^{2k}\right] - \frac{1}{2}\sum_{j=0}^{k}\binom{k}{j}^2 y^{k-j} - \frac{2}{y}\sum_{j=0}^{k}\binom{k}{j}(y-1)^j\binom{2k-j}{k-2} \\
& + \sum_{j=0}^{k}\binom{k}{j}(y-1)^j\binom{2k-j-1}{k-2}, \\
\widetilde{\mu}_{x^k} =& \mu_{x^k} - \sum_{j=0}^{k-1}\binom{k}{j}(y-1)^j\binom{2k-j-2}{k-1} + \sum_{j=0}^{k}\binom{k}{j}(y-1)^j\binom{2k-j}{k-1},
\end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

and the asymptotic variances are

$$\begin{aligned}
\sigma^2_{\log} =& -2\log(1-y) - 2y,\\
\sigma^2_{x^k} =& 2\sum_{j_1=0}^{k-1}\sum_{j_2=0}^{k}\binom{k}{j_1}\binom{k}{j_2}(y-1)^{j_1+j_2}\sum_{l=1}^{k-j_1} l\binom{2k-1-(j_1+l)}{k-1}\binom{2k-1-j_2+l}{k-1}\\
&-\frac{2}{y}\sum_{j_1=0}^{k}\sum_{j_2=0}^{k}\binom{k}{j_1}\binom{k}{j_2}(y-1)^{j_1+j_2}\binom{2k-j_1}{k-1}\binom{2k-j_2}{k-1}.
\end{aligned}$$

For polynomials of $\boldsymbol{\rho}_n$, our results (4.3) are consistent with ones of Bao et al. (2015) and the other three asymptotic distributions are new which can be used to derive the asymptotic distribution of test statistics.

Noting $\mathrm{tr}(\boldsymbol{\rho}_n) = p$ and $\mathrm{tr}(\widetilde{\boldsymbol{\rho}}_n) = p$, we can simplify the test statistics of $T_1(\boldsymbol{\rho}_n),\ T_1(\widetilde{\boldsymbol{\rho}}_n)$, $T_2(\boldsymbol{\rho}_n),\ T_2(\widetilde{\boldsymbol{\rho}}_n)$ and consider

$$L_{\boldsymbol{\rho},2} = \mathrm{tr}(\boldsymbol{\rho}_n^2), \quad L_{\boldsymbol{\rho},\log} = \log(|\boldsymbol{\rho}_n|), \quad L_{\widetilde{\boldsymbol{\rho}},2} = \mathrm{tr}(\widetilde{\boldsymbol{\rho}}_n^2), \quad L_{\widetilde{\boldsymbol{\rho}},\log} = \log(|\widetilde{\boldsymbol{\rho}}_n|).$$

With Theorem 4, we can get four rejection regions for testing the null distribution

$$\begin{aligned}
R_1 =& \{L_{\boldsymbol{\rho},2} - \frac{p^2}{n} - p > y_n^2 - y_n + 2y_n Z_\alpha\},\\
R_2 =& \{L_{\boldsymbol{\rho},\log} + (n-p)\log(1-y_n) + p < \frac{3}{2}\log(1-y_n) + 2y_n\\
&-\sqrt{-2\log(1-y_n) - 2y_n}Z_\alpha\},\\
R_3 =& \{L_{\widetilde{\boldsymbol{\rho}},2} - \frac{p^2}{n} - p > 3y_n^2 - y_n + 2y_n Z_\alpha\},
\end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

$$R_4 = \{L_{\widetilde{\boldsymbol{\rho}},\log} + (n-p)\log(1-y_n) + p < \frac{3}{2}\log(1-y_n) + \frac{2y_n - 3y_n^2}{1-y_n} - \sqrt{-2\log(1-y_n) - 2y_n}Z_\alpha\},$$

where Z_α is the upper-α quantile of $N(0,1)$.

To examine the finite sample performance of these test statistics, we conduct the following null hypotheses with data $\mathbf{X}_n = (X_{ij})_{n\times p}$ generated from different models. Specifically, we consider three types of null distributions:

- Normal distribution: X_{ij} are i.i.d. $N(0,1)$ for $1 \le i \le n$ and $1 \le j \le p$.
- Cauchy distribution: X_{ij} are i.i.d. Cauchy distribution with location 0 and scale 1 (Cauchy$(0,1)$) for $1 \le i \le n$ and $1 \le j \le p$.
- Mixed distribution: X_{ij_1} are i.i.d. Cauchy$(0,1)$ for $1 \le i \le n$, $1 \le j_1 \le \lfloor p/4 \rfloor$; X_{ij_2} are i.i.d. $N(0,1)$ for $1 \le i \le n$, $\lfloor p/4 \rfloor + 1 \le j_2 \le \lfloor p/2 \rfloor$; X_{ij_3} are i.i.d. $\chi^2(2)$ for $1 \le i \le n$, $\lfloor p/2 \rfloor + 1 \le j_2 \le p$.

It is noted that Cauchy$(0,1)$ is a well known heavy-tailed distribution without expectation, and the mixed distribution is from Li et al. (2021).

As for comparison, we consider other 8 test statistics based on Spearman, Kendall and Pearson's correlation matrices:

1. $L_{\boldsymbol{\rho},\max}$: maximum test based on Spearman's correlations (Han et al., 2017);
2. $L_{\widetilde{\boldsymbol{\rho}},\max}$: maximum test based on improved Spearman's correlations (Han et al., 2017);

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

3. $L_{\mathbf{K},2}$: ℓ_2 test based on Kendall's correlations (Li et al., 2021);

4. $L_{\mathbf{K},\log}$: Stein's test based on Kendall's correlations (Li et al., 2021);

5. $L_{\mathbf{K},\max}$: maximum test based on Kendall's correlations (Han et al., 2017);

6. $L_{\mathbf{R},2}$: ℓ_2 test based on Pearson's correlations (Han et al., 2017);

7. $L_{\mathbf{R},\log}$: Stein's test based on Pearson's correlations (Han et al., 2017);

8. $L_{\mathbf{R},\max}$: maximum test based on Pearson's correlations (Zhou, 2007).

We conduct numerical experiments to evaluate the performance of our proposed test statistics. We consider various combinations of sample size n, dimension p, and underlying distributions, and compare our methods with existing approaches. Table 2 presents the empirical sizes of the tests at a nominal significance level of 5% based on 1000 replications. Our results demonstrate that Pearson's correlation-based tests are sensitive to distributional assumptions and may not perform well under heavy-tailed distributions. In contrast, rank-based test statistics, including our proposed $L_{\boldsymbol{\rho},\log}$, $L_{\widetilde{\boldsymbol{\rho}},2}$, and $L_{\widetilde{\boldsymbol{\rho}},\log}$, exhibit robust performance across different distributions. The empirical sizes of our proposed tests are close to the nominal 5% level, confirming the validity of our theoretical results.

To evaluate the power of our proposed test statistics, we generate data under various alternative hypotheses. We start with data generated from the above three null distributions and then generate the correlated data $\mathbf{X}_n$ as follows:

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

Table 2: Empirical sizes of independence test statistics based on Pearson, Spearman and Kendall's correlations.

n	100	200	400	100	200	400	100	200	400
p	50	100	200	70	140	280	200	400	800
y	0.5	0.5	0.5	0.7	0.7	0.7	2	2	2
Normal distribution									
$L_{\boldsymbol{\rho},2}$	0.046	0.05	0.046	0.044	0.041	0.045	0.051	0.059	0.058
$L_{\boldsymbol{\rho},\log}$	0.037	0.049	0.041	0.052	0.043	0.048	-	-	-
$L_{\boldsymbol{\rho},\max}$	0.023	0.028	0.037	0.025	0.033	0.041	0.022	0.023	0.049
$L_{\widetilde{\boldsymbol{\rho}},2}$	0.047	0.052	0.047	0.047	0.042	0.046	0.063	0.062	0.059
$L_{\widetilde{\boldsymbol{\rho}},\log}$	0.068	0.075	0.05	0.069	0.049	0.052	-	-	-
$L_{\widetilde{\boldsymbol{\rho}},\max}$	0.023	0.027	0.037	0.025	0.033	0.042	0.022	0.023	0.049
$L_{\mathbf{K},2}$	0.043	0.05	0.047	0.046	0.045	0.047	0.059	0.067	0.06
$L_{\mathbf{K},\log}$	0.047	0.054	0.045	0.055	0.045	0.046	0.119	0.09	0.064
$L_{\mathbf{K},\max}$	0.039	0.039	0.048	0.035	0.042	0.048	0.034	0.029	0.058
$L_{\mathbf{R},2}$	0.053	0.046	0.051	0.041	0.044	0.04	0.065	0.055	0.051
$L_{\mathbf{R},\log}$	0.053	0.06	0.051	0.045	0.039	0.041	-	-	-
$L_{\mathbf{R},\max}$	0.027	0.023	0.04	0.027	0.028	0.037	0.012	0.029	0.039
Cauchy distribution									
$L_{\boldsymbol{\rho},2}$	0.056	0.055	0.039	0.042	0.058	0.041	0.048	0.051	0.044
$L_{\boldsymbol{\rho},\log}$	0.065	0.05	0.055	0.056	0.052	0.045	-	-	-
$L_{\boldsymbol{\rho},\max}$	0.023	0.029	0.044	0.03	0.031	0.039	0.016	0.02	0.038
$L_{\widetilde{\boldsymbol{\rho}},2}$	0.058	0.058	0.04	0.047	0.058	0.041	0.057	0.057	0.05
$L_{\widetilde{\boldsymbol{\rho}},\log}$	0.072	0.064	0.064	0.072	0.061	0.049	-	-	-
$L_{\widetilde{\boldsymbol{\rho}},\max}$	0.023	0.027	0.044	0.03	0.031	0.039	0.016	0.02	0.038
$L_{\mathbf{K},2}$	0.06	0.058	0.041	0.043	0.06	0.045	0.056	0.058	0.052
$L_{\mathbf{K},\log}$	0.062	0.063	0.053	0.057	0.051	0.05	0.125	0.074	0.072
$L_{\mathbf{K},\max}$	0.031	0.041	0.051	0.037	0.037	0.042	0.029	0.032	0.038
$L_{\mathbf{R},2}$	0.303	0.329	0.394	0.306	0.331	0.396	0.305	0.343	0.385
$L_{\mathbf{R},\log}$	0.766	0.924	0.989	0.896	0.984	1	-	-	-
$L_{\mathbf{R},\max}$	1	1	1	1	1	1	1	1	1
Mixed distribution									
$L_{\boldsymbol{\rho},2}$	0.053	0.061	0.051	0.038	0.054	0.049	0.043	0.041	0.044
$L_{\boldsymbol{\rho},\log}$	0.057	0.052	0.055	0.056	0.053	0.05	-	-	-
$L_{\boldsymbol{\rho},\max}$	0.02	0.033	0.036	0.021	0.034	0.031	0.017	0.028	0.032
$L_{\widetilde{\boldsymbol{\rho}},2}$	0.056	0.063	0.052	0.042	0.054	0.049	0.049	0.043	0.045
$L_{\widetilde{\boldsymbol{\rho}},\log}$	0.071	0.066	0.065	0.077	0.063	0.056	-	-	-
$L_{\widetilde{\boldsymbol{\rho}},\max}$	0.02	0.033	0.036	0.021	0.034	0.031	0.017	0.028	0.032
$L_{\mathbf{K},2}$	0.06	0.064	0.054	0.047	0.061	0.048	0.049	0.042	0.044
$L_{\mathbf{K},\log}$	0.071	0.062	0.059	0.052	0.065	0.049	0.115	0.066	0.061
$L_{\mathbf{K},\max}$	0.036	0.047	0.043	0.03	0.043	0.036	0.033	0.044	0.04
$L_{\mathbf{R},2}$	1	1	1	1	1	1	1	1	1
$L_{\mathbf{R},\log}$	1	1	1	1	1	1	-	-	-
$L_{\mathbf{R},\max}$	0.91	1	1	0.712	0.988	1	1	1	1

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

- Global correlation: $\mathbf{X}_n = \mathbf{Z}_n\mathbf{\Sigma}$, where $\mathbf{\Sigma} = (\sigma_{ij})_{p\times p}$ is the Toeplitz matrix with $\sigma_{ii} = 1$, $\sigma_{ij} = \rho^{|i-j|}$;

- Sparse correlation: $\mathbf{X}_n = \mathbf{Z}_n\mathbf{\Sigma}$, where $\mathbf{\Sigma} = (\sigma_{ij})_{p\times p}$ is the Toeplitz matrix with $\sigma_{ii} = 1$, $\sigma_{i-1,i} = \sigma_{i,i+1} = \rho$, $\sigma_{ij} = 0$ for $|i-j| > 1$;

By combining these data generation methods, we obtain six different alternative hypotheses. For each alternative hypothesis, we set $(n, p) = (100, 200)$ or $(n, p) = (200, 100)$ to assess the power of the tests for $y < 1$ or $y > 1$. Since Pearson's correlation-based tests are sensitive to distributional assumptions and maximum-norm-type tests are always undersized when (n, p) are not large enough, we focus on rank-based tests (Spearman and Kendall) with ℓ_2 loss and Stein's loss in our power analysis. Simulations are presented in Figure 1 and Figure 2.

From Figure 1 and Figure 2, we observe that the power of all test statistics increases as the absolute value of correlation strength ρ increases. This demonstrates the effectiveness of rank-based tests, especially under heavy-tailed distributions. When $(n, p) = (100, 200)$, although $L_{\mathbf{K},\log}$ always exhibits much higher power, it fails to control the size under the null hypothesis. Therefore, for both correlation structures, tests based on the ℓ_2 loss demonstrate superior performance. Interestingly, for all the alternative hypotheses, $L_{\mathbf{K},2}$ always performs the best and $L_{\widetilde{\boldsymbol{\rho}},2}$ always performs better than $L_{\boldsymbol{\rho},2}$ since it is a combination of $\boldsymbol{\rho}_n$ and $\mathbf{K}_n$. We leave the theoretical analysis of these powers as a future work. Overall, our proposed test statistics $L_{\boldsymbol{\rho},\log}$, $L_{\widetilde{\boldsymbol{\rho}},2}$, and $L_{\widetilde{\boldsymbol{\rho}},\log}$ demonstrate comparable performance

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

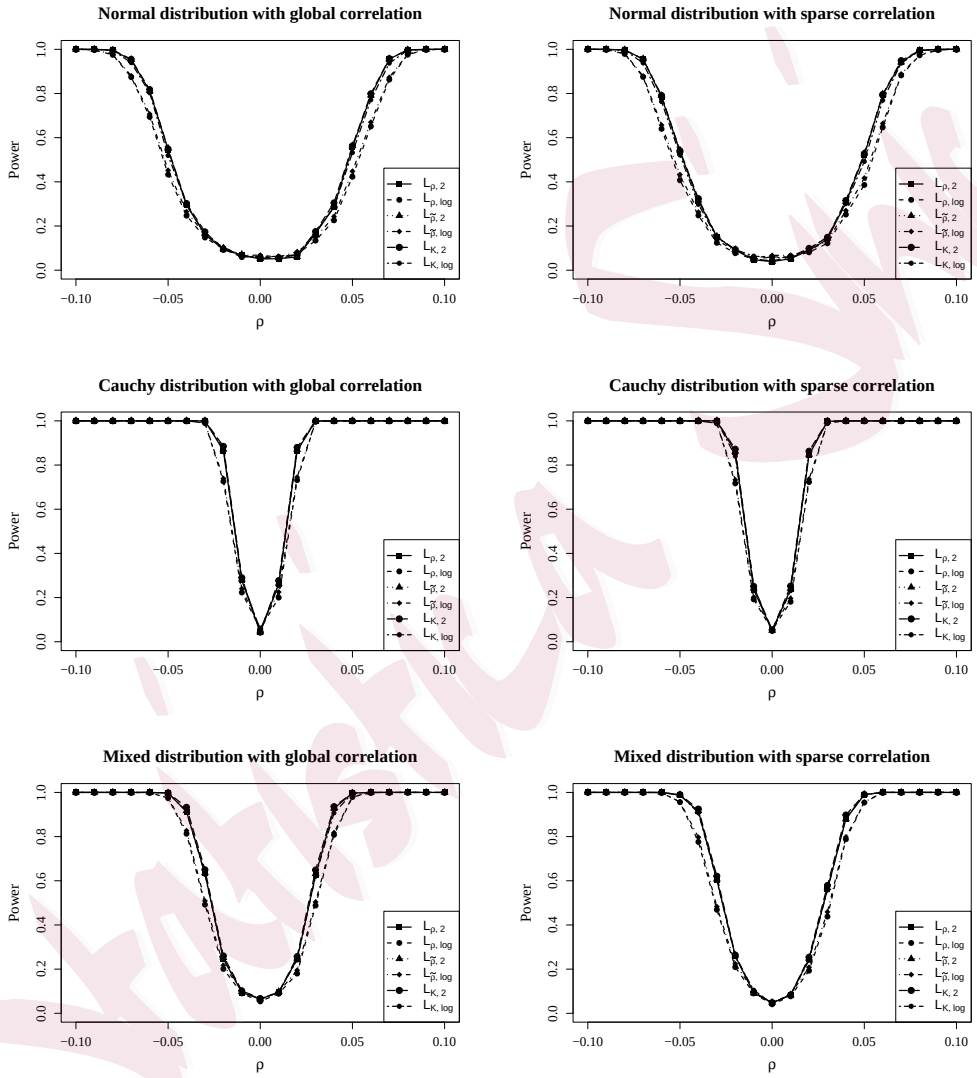

Figure 1: The empirical powers of the tests with the variation of correlation strength ρ when $(n, p) = (200, 100)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

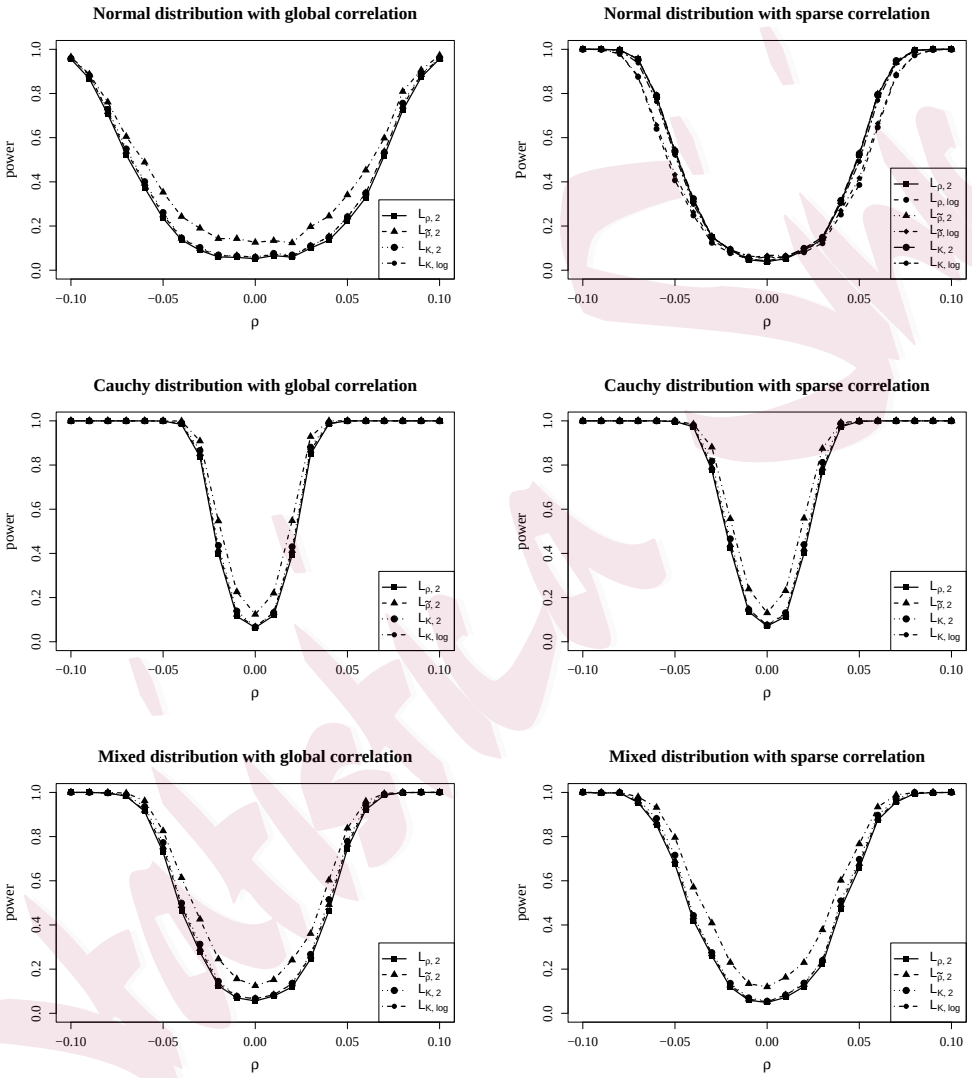

Figure 2: The empirical powers of the tests with the variation of correlation strength ρ when $(n, p) = (100, 200)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

across various scenarios.

5. Discussion

For the Stieltjes transform of the considered matrix, we can get study its limit which yields the limit of LSS and the CLT which yields the CLT of LSS. We summarize the results as following for $\Im(z) > c$.

- For the Gram matrix $\mathbf{g}_n \in \mathbb{R}^{n\times n}$, we have

$$\begin{aligned}
&\text{LSD}: \ F^{\mathbf{g}_n} \xrightarrow{d} F_{1/y}, \ a.s.;\\
&\text{Stieltjes transform: } s_n(z) = \frac{1}{n}\text{tr}(\mathbf{g}_n - z\mathbf{I}_n)^{-1} \xrightarrow{a.s.} s(z) = m(1/y, z);\\
&\text{CLT: } n\left(s_n(z) - \mathbb{E}s_n(z)\right) \xrightarrow{d} \text{Gaussian Processes}\left(0, \sigma(z_1, z_2)\right),
\end{aligned}$$

where

$$n\mathbb{E}s_n(z) = n \cdot m(n/p, z) + \mu(z) + o(1).$$

- For re-scaled Spearman's rank correlation matrix $\boldsymbol{\rho}_n/\mathbf{y}_n \in \mathbb{R}^{p\times p}$,

$$\begin{aligned}
&\text{LSD}: \ F^{\boldsymbol{\rho}_n/\mathbf{y}_n} \xrightarrow{d} \underline{F}_{1/y}, \ a.s.;\\
&\text{Stieltjes transform: } \underline{s}_n(z) = \frac{1}{p}\text{tr}(\boldsymbol{\rho}_n/\mathbf{y}_n - z\mathbf{I}_n)^{-1} = \frac{n}{p}\left(s_n(z) + \frac{1}{z}\right) - \frac{1}{z}
\end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

$$\xrightarrow{a.s.} \underline{s}(z) = \frac{1}{y}\left(s(z) + \frac{1}{z}\right) - \frac{1}{z};$$

$$\text{CLT: } p\left(\underline{s}_n(z) - \mathbb{E}\underline{s}_n(z)\right) \xrightarrow{d} \text{Gaussian Processes}\left(0, \sigma(z_1, z_2)\right),$$

where

$$\begin{aligned} p\mathbb{E}\underline{s}_n(z) =& n \cdot m(n/p, z) + \mu(z) + \frac{n-p}{z} + o(1) \\ =& n\left(m(n/p, z) + (1 - \frac{p}{n})\frac{1}{z}\right) + \mu(z) + o(1). \end{aligned}$$

- For Spearman's rank correlation matrix $\boldsymbol{\rho}_n \in \mathbb{R}^{p\times p}$

$$\text{LSD}: \ F^{\boldsymbol{\rho}_n} \xrightarrow{d} F_y, \ a.s.;$$

$$\text{Stieltjes transform: } m_n(z) = \frac{1}{p}\text{tr}(\boldsymbol{\rho}_n - z\mathbf{I}_n)^{-1} = \frac{1}{y_n}\underline{s}_n(z/y_n)$$

$$\xrightarrow{a.s.} \frac{1}{y}\underline{s}(z/y) = m(z) = m(y, z);$$

$$\text{CLT: } p\left(m_n(z) - \mathbb{E}m_n(z)\right) \xrightarrow{d} \text{Gaussian Processes}\left(0, \frac{\sigma(z_1/y, z_2/y)}{y^2}\right),$$

where

$$p\mathbb{E}m_n(z) = p \cdot m(p/n, z) + \frac{\mu(z/y)}{y} + o(1).$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

- For improved Spearman's rank correlation matrix $\widetilde{\boldsymbol{\rho}}_n \in \mathbb{R}^{p\times p}$

$$\begin{aligned}
&\text{LSD}: \ F^{\widetilde{\boldsymbol{\rho}}_n} \xrightarrow{d} F_y, \ a.s.;\\
&\text{Stieltjes transform: } \widetilde{m}_n(z) = \frac{1}{p}\text{tr}(\widetilde{\boldsymbol{\rho}}_n - z\mathbf{I}_n)^{-1} \xrightarrow{a.s.} m(z) = m(y,z);\\
&\text{CLT: } p\left(\widetilde{m}_n(z) - \mathbb{E}\widetilde{m}_n(z)\right) \xrightarrow{d} \text{Gaussian Processes}\left(0, \frac{\sigma(z_1/y, z_2/y)}{y^2}\right),
\end{aligned}$$

where

$$p\mathbb{E}\widetilde{m}_n(z) = p\cdot m(p/n, z) + \frac{\mu(z/y)}{y} + \frac{\widetilde{\mu}(z/y)}{y} + o(1).$$

With these CLTs, we can construct hypothesis tests based on Spearman's and improved Spearman's correlation matrices. Our simulation studies demonstrate the practical applicability of these new test statistics.

In this work, we study the improved Pearson's correlation which is a standard U-statistic of order 3. Studying general U-statistic typed correlation matrices could be a topic of future work. Moreover, we compare the test statistics through simulations. Investigating the asymptotic distribution of test statistics under local alternatives could be another interesting future work.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

Supplementary Material

The online Supplementary Material includes the detailed proofs of the main theorems and additional lemmas.

Acknowledgments

The authors thank the Editor, the Associate Editor, and anonymous reviewers for their insightful comments on earlier versions of this paper. Cheng Wang's research is partially supported by NSFC 12031005, NSFC 72495121 and the fundamental research funds for the central universities.

References

Anderson, G. W. and O. Zeitouni (2008). A CLT for regularized sample covariance matrices. *Annals of Statistics 36*(6), 2553–2576.

Anderson, T. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley.

Bai, Z., D. Jiang, J. Yao, and S. Zheng (2009). Corrections to lrt on large dimensional covariance matrix by rmt. *Annals of Statistics 37*(6B), 3822–3840.

Bai, Z. and J. W. Silverstein (2004). CLT for linear spectral statistics of large-dimensional sample covariance matrices. *Annals of Probability 32*(1A), 553–605.

Bai, Z. and J. W. Silverstein (2010). *Spectral analysis of large dimensional random matrices*. New York: Springer.

Bai, Z. and W. Zhou (2008). Large sample covariance matrices without independence structures in columns.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

Statistica Sinica 18, 425–442.

Bandeira, A. S., A. Lodhia, and P. Rigollet (2017). Marčenko-Pastur law for Kendall's tau. *Electronic Communications in Probability 22*(32), 1–7.

Bao, Z. (2019). Tracy–widom limit for Kendall's tau. *Annals of Statistics 47*(6), 3504–3532.

Bao, Z., L.-C. Lin, G. Pan, and W. Zhou (2015). Spectral statistics of large dimensional Spearman's rank correlation matrix and its application. *Annals of Statistics 43*(6), 2588–2623.

Bao, Z., G. Pan, and W. Zhou (2012). Tracy-Widom law for the extreme eigenvalues of sample correlation matrices. *Electronic Journal of Probability 17*(88), 1–32.

Chen, S. X., L.-X. Zhang, and P.-S. Zhong (2010). Tests for high-dimensional covariance matrices. *Journal of the American Statistical Association 105*(490), 810–819.

Dobriban, E. and S. Wager (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *Annals of Statistics 46*(1), 247–279.

El Karoui, N. (2009). Concentration of measure and spectra of random matrices: Applications to correlation matrices, elliptical distributions and beyond. *Annals of Applied Probability 19*(6), 2362–2405.

Gao, J., X. Han, G. Pan, and Y. Yang (2017). High dimensional correlation matrices: The central limit theorem and its applications. *Journal of the Royal Statistical Society, Series B 79*(3), 677–693.

Han, F., S. Chen, and H. Liu (2017). Distribution-free tests of independence in high dimensions. *Biometrika 104*(4), 813–828.

Hastie, T., A. Montanari, S. Rosset, and R. J. Tibshirani (2022). Surprises in high-dimensional ridgeless least squares interpolation. *Annals of Statistics 50*(2), 949–986.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

REFERENCES

Heiny, J. and N. Parolya (2024). Log determinant of large correlation matrices under infinite fourth moment. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques 60*(2), 1048–1076.

Heiny, J. and J. Yao (2022). Limiting distributions for eigenvalues of sample correlation matrices from heavy-tailed populations. *Annals of Statistics 50*(6), 3249–3280.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics 19*(3), 293–325.

Jiang, T. (2004a). The asymptotic distributions of the largest entries of sample correlation matrices. *Annals of Applied Probability 14*(2), 865–880.

Jiang, T. (2004b). The limiting distributions of eigenvalues of sample correlation matrices. *Sankhyā: The Indian Journal of Statistics 66*(1), 35–48.

Jiang, T. (2019). Determinant of sample correlation matrix with application. *Annals of Applied Probability 29*(3), 1356–1397.

Leung, D. and M. Drton (2018). Testing independence in high dimensions with sums of rank correlations. *Annals of Statistics 46*(1), 280–307.

Li, Z., C. Wang, and Q. Wang (2023). On eigenvalues of a high-dimensional Kendall's rank correlation matrix with dependence. *Science China Mathematics 66*(11), 2615–2640.

Li, Z., Q. Wang, and R. Li (2021). Central limit theorem for linear spectral statistics of large dimensional Kendall's rank correlation matrices and its applications. *Annals of Statistics 49*(3), 1569–1593.

Lytova, A. and L. Pastur (2009). Central limit theorem for linear eigenvalue statistics of random matrices with independent entries. *Annals of Probability 37*(5), 1778–1840.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

Marčenko, V. and L. Pastur (1967). Distribution of eigenvalues for some sets of random matrices. *Sbornik: Mathematics 1*(4), 457–483.

Mestre, X. and P. Vallet (2017). Correlation tests and linear spectral statistics of the sample correlation matrix. *IEEE Transactions on Information Theory 63*(7), 4585–4618.

Pan, G. (2014). Comparison between two types of large sample covariance matrices. *Annales de l'IHP Probabilités et statistiques 50*(2), 655–677.

Pan, G. and W. Zhou (2008). Central limit theorem for signal-to-interference ratio of reduced rank linear receiver. *Annals of Applied Probability 18*(3), 1232–1270.

Parolya, N., J. Heiny, and D. Kurowicka (2024). Logarithmic law of large random correlation matrices. *Bernoulli 30*(1), 346–370.

Paul, D. and A. Aue (2014). Random matrix theory in statistics: A review. *Journal of Statistical Planning and Inference 150*, 1–29.

Pillai, N. S. and J. Yin (2012). Edge universality of correlation matrices. *Annals of Statistics 40*(3), 1737–1763.

Wang, C. and B. Jiang (2018). On the dimension effect of regularized linear discriminant analysis. *Electronic Journal of Statistics 12*(2), 2709–2742.

Wang, C., J. Yang, B. Miao, and L. Cao (2013). Identity tests for high dimensional data using RMT. *Journal of Multivariate Analysis 118*, 128–137.

Wang, H., B. Liu, L. Feng, and Y. Ma (2024). Rank-based max-sum tests for mutual independence of high-dimensional random vectors. *Journal of Econometrics 238*(1), 105578.

Wang, Q. and J. Yao (2013). On the sphericity test with large-dimensional observations. *Electronic Journal of*

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0395

Statistics 7, 2164–2192.

Wu, Z. and C. Wang (2022). Limiting spectral distribution of large dimensional Spearman's rank correlation matrices. *Journal of Multivariate Analysis 191*, 105011.

Yao, J., S. Zheng, and Z. Bai (2015). *Large sample covariance matrices and high-dimensional data analysis*. Cambridge: Cambridge University Press.

Zheng, S., Z. Bai, and J. Yao (2015). Substitution principle for CLT of linear spectral statistics of high-dimensional sample covariance matrices with applications to hypothesis testing. *Annals of Statistics 43*(2), 546–591.

Zheng, S., G. Cheng, J. Guo, and H. Zhu (2019). Test for high dimensional correlation matrices. *Annals of Statistics 47*(5), 2887–2921.

Zhou, W. (2007). Asymptotic distribution of the largest off-diagonal entry of correlation matrices. *Transactions of the American Mathematical Society 359*(11), 5345–5363.

Hantao Chen

School of Mathematical Sciences, MOE-LSC, Shanghai Jiao Tong University

E-mail: htchen2000@sjtu.edu.cn

Cheng Wang(corresponding author)

School of Mathematical Sciences, MOE-LSC, Shanghai Jiao Tong University

E-mail: chengwang@sjtu.edu.cn