Robust estimation for number of factors in high dimensional factor modeling via Spearman correlation matrix

Jiaxin Qiu
Department of Statistics and Actuarial Science
The University of Hong Kong, Hong Kong, China
qiujx@connect.hku.hk
and
Zeng Li*
Department of Statistics and Data Science
Southern University of Science and Technology, Shenzhen, China
liz9@sustech.edu.cn
and
Jianfeng Yao
School of Data Science
The Chinese University of Hong Kong (Shenzhen), Shenzhen, China
jeffyao@cuhk.edu.cn

August 29, 2024

Abstract

Determining the number of factors in high-dimensional factor modeling is essential but challenging, especially when the data are heavy-tailed. In this paper, we introduce a new estimator based on the spectral properties of Spearman sample correlation matrix under the high-dimensional setting, where both dimension and sample size tend to infinity proportionally. Our estimator is robust against heavy tails in either the common factors or idiosyncratic errors. The consistency of our estimator is established under mild conditions. Numerical experiments demonstrate the superiority of our estimator compared to existing methods.

Keywords: Factor model; Heavy tails; High dimensionality; Spearman correlation matrix; Spiked eigenvalue.

*Zeng Li (liz9@sustech.edu.cn) is the corresponding author.

1 Introduction

Factor models are helpful tools for understanding the common dependence among high-dimensional outputs. They are widely used in data analysis in various areas like finance, genomics, and economics. Estimating the total number of factors is one of the most fundamental challenges when applying factor models in practice. This paper focuses on the following factor model:

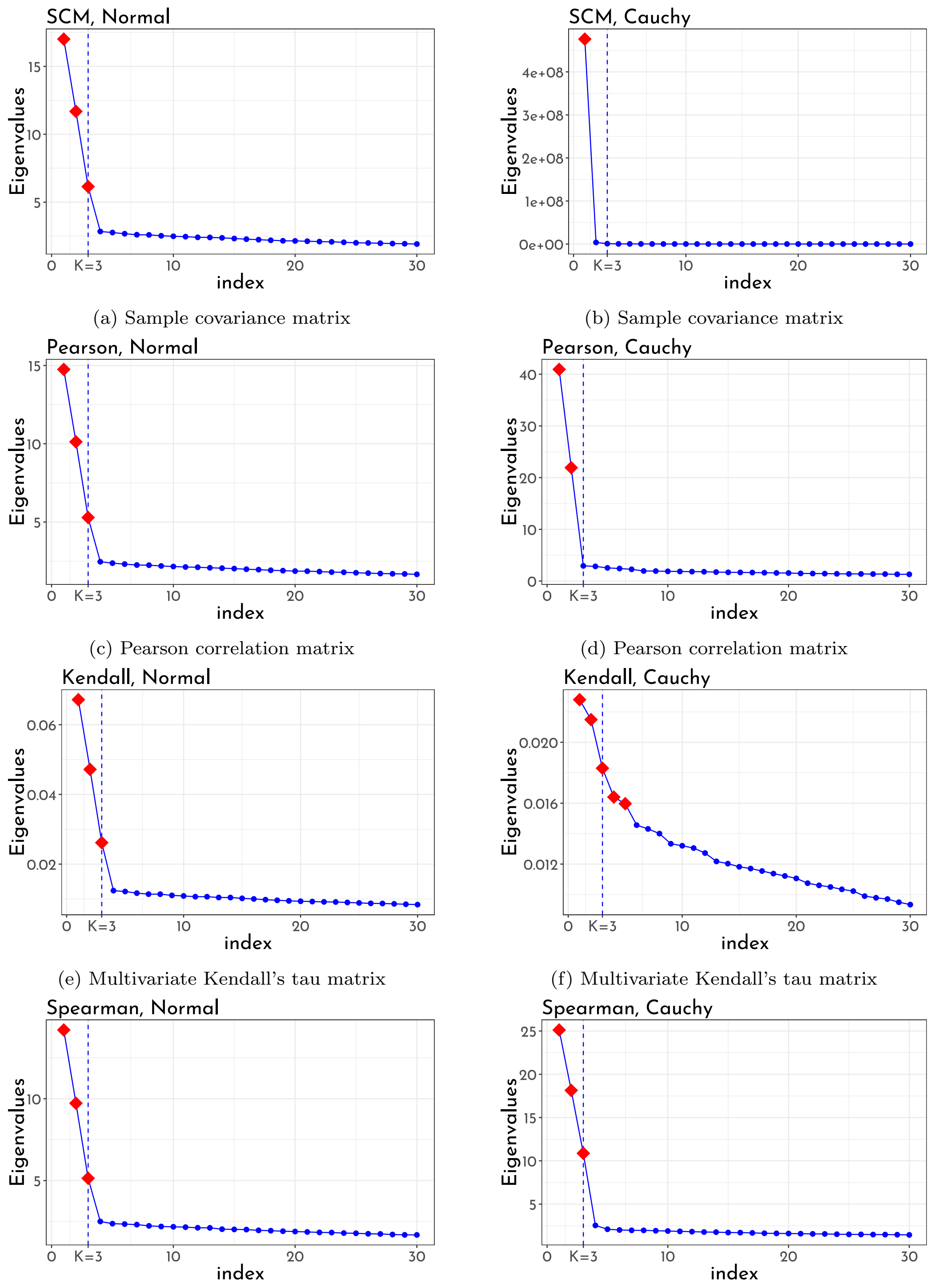
$$\mathbf{y}_i = \mathbf{B}\mathbf{f}_i + \mathbf{\Psi}\mathbf{e}_i, \qquad i \in [n] := \{1, 2, \ldots, n\}, \tag{1}$$

where $\{\mathbf{y}_i\}_{i=1}^n$ are the p-dimensional observation vectors, $\{\mathbf{f}_i\}_{i=1}^n$ the K-dimensional latent common factor vectors, $\{\mathbf{e}_i\}_{i=1}^n$ the p-dimensional idiosyncratic error vectors, $\mathbf{B}$ the $p \times K$ factor loading matrix, and $\mathbf{\Psi}$ a $p \times p$ diagonal matrix. The objective of this paper is to estimate the number of common factors when the observed data are heavy-tailed.

There is a large literature on this estimation problem which can generally be categorized into two types of approaches. The first type is based on information criteria. The seminal work Bai and Ng (2002) proposed several information criteria, which were formulated in many different forms, through modifications of the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Hallin and Liška (2007) proposed an information criterion that utilized spectral density matrix estimation. Alessi et al. (2010) modified Bai and Ng (2002)'s criteria by tuning the penalty function to enhance their performance. Kong (2017) employed similar ideas and put forth a local principal component analysis (PCA) approach to study a continuous-time factor model with time-varying factor loadings using high-frequency data. Li et al. (2017a) used information criteria akin to those proposed by Bai and Ng (2002) for factor models when the number of factors increases with the cross-section size and time period. The first type of approach usually requires strong signals. The second type of approach is based on the eigenvalue behavior of various types of covariance/correlation matrices. As for sample covariance matrices, Nadakuditi and Edelman (2008) proposed an estimator by exploiting the distribution properties of the moments of eigenvalues. Ahn and Horenstein (2013) proposed two estimators by utilizing the ratios of adjacent eigenvalues, namely the eigenvalue ratio (ER) estimator and the growth ratio (GR) estimator. Onatski (2010, 2012) proposed an alternative edge distribution (ED) estimator based on the maximum differences between consecutive eigenvalues instead of their ratios. Owen and Wang (2016) introduced an estimator that utilizes the bi-cross-validation (BCV) technique from Owen and Perry (2009). This estimator is based on the theoretical results concerning the spiked sample covariance matrix. For lagged sample autocovariance matrices, Lam and Yao (2012) developed a ratio-based estimator for factor modeling of multivariate time series. This estimator was further extended by Li et al. (2017b) to accommodate weak factors. As for correlation matrices, Fan et al. (2020) proposed a tuning-free and scale-invariant adjusted correlation thresholding method. This approach has been further extended to a time series tensor factor model in Lam (2021) and Chen and Lam (2024).

The aforementioned methods have been proved to be inadequate when dealing with heavy-tailed data, and would mostly result in biased or inconsistent estimators. Heavy-tailed data are common in various real-world applications. For instance, prices of stock returns often exhibit heavy tails due to the occurrence of extreme events in the market. However, little literature has focused on estimating the number of factors in the context of heavy-tailed data. Assuming a jointly elliptical distribution for both common factors and idiosyncratic errors (as discussed in Fan et al. (2018)), Yu et al. (2019) proposed two estimators utilizing the sample multivariate Kendall's tau matrix. He et al. (2022b) further extended it to the matrix factor model. He et al. (2022a) recovered factor loadings and scores by performing PCA on the multivariate Kendall's tau matrix. It is worth mentioning that Yu et al. (2019)'s method requires that $\|\mathbf{B}^\intercal\mathbf{B}/p - \mathbf{\Sigma}_B\|_2 \to 0$, where $\mathbf{\Sigma}_B$ is a $K \times K$ positive definite matrix with bounded and distinct eigenvalues (see their Assumptions 2.3). The factor model is considered to have a strong factor structure (Bai and Ng, 2002) when both $\mathbf{B}^\intercal\mathbf{B}/p$ and $\sum_{i=1}^n \mathbf{f}_i\mathbf{f}_i^\intercal/n$ converge to positive definite matrices. In this paper, we consider the weak loading scenario by assuming $\mathbf{B}^\intercal\mathbf{B} = \mathbf{I}_K$, which is a commonly used identifiability condition in the literature on factor models (see, for example, Bai and Li (2012)). Moreover, we address a more challenging scenario where both the factors and idiosyncratic errors may be heavy-tailed, potentially leading to the non-existence of the limit of $\sum_{i=1}^n \mathbf{f}_i\mathbf{f}_i^\intercal/n$. We propose an estimator based on *Spearman correlation matrix* (Spearman, 1961) which shows significant improvements over existing methods. Here, we use a toy example to demonstrate the robustness of Spearman correlation matrix. Data are generated following factor model (1) with $K = 3$. The factors and idiosyncratic errors follow either standard normal distribution or standard Cauchy distribution. As shown in Figure 1, when the common factors and the idiosyncratic noise are light-tailed, all four sample covariance/correlation matrices have three spiked eigenvalues, and all factors can be detected. When the data distribution is heavy-tailed, only our method can clearly identify all three factors.

The Spearman correlation matrix is defined as the Pearson correlation matrix of the ranks of the data. It is a valuable tool when dealing with heavy-tailed data. However, the nonlinear structure of rank-based correlation brings significant difficulties when analyzing its eigenvalue behavior. To address this, we need to resort to tools in random matrix theory



(a) Sample covariance matrix (b) Sample covariance matrix

(c) Pearson correlation matrix (d) Pearson correlation matrix

(e) Multivariate Kendall's tau matrix (f) Multivariate Kendall's tau matrix

(g) Spearman correlation matrix (h) Spearman correlation matrix

Figure 1: Scatter plots of the first 30 eigenvalues of the sample covariance matrix (SCM), Pearson correlation matrix, Spearman correlation matrix, and multivariate Kendall's tau matrix. Data are generated following the factor model (1) with $K = 3$. The factors and idiosyncratic errors are drawn independently from standard **Normal** distribution (**Left panel**: Figures (a), (c), (e), (g)) or standard **Cauchy** distribution (**Right panel**: Figures (b), (d), (f), (h)). The symbol "♦" represents the spiked eigenvalues. Further details regarding the matrices $\mathbf{B}$ and $\boldsymbol{\Psi}$ can be found in the case (C1) in Section 3.

(RMT). Unfortunately, most existing work in RMT focuses on very restrictive settings where data has independent components. Bai and Zhou (2008) showed that its limiting spectral distribution (LSD) is the well-known Marčenko-Pastur law. Bao et al. (2015) established the central limiting theorem (CLT) for its linear spectral statistics (LSS). Bao (2019) showed that the Tracy-Widom law holds for its largest eigenvalues. To the best of our knowledge, the first investigation of the Spearman sample correlation matrix for general dependent data was conducted very recently by Wu and Wang (2022), which derived its LSD under the non-paranormal distribution proposed by Liu et al. (2009). Many other spectral properties, including the extreme eigenvalues, CLT for LSS, and spiked eigenvalues for dependent data, still remain open. Our work is the first to investigate the eigenvalue behavior of the Spearman sample correlation matrix under spike models, and successfully apply the theories to identify the number of factors in high-dimensional factor modeling for heavy-tailed data.

To summarize, the main contributions of this paper are two-fold. First, we propose a new estimator based on the Spearman sample correlation matrix for the number of common factors in the high-dimensional factor model (1). This estimator is distribution-free and capable of estimating the number of factors even when the data is heavy-tailed. Second, we provide a theoretical explanation of the phase-transition phenomenon for the top eigenvalues of the Spearman sample correlation matrix under the spike model. From a technical point of view, we investigate this phase-transition theory by establishing the universality of the asymptotic law of a low-dimensional random matrix (see Lemma 6.2 and Remark 2.6 for more details), and our method does not require the commonly used independent component structure.

Before moving forward, let us introduce some notations that will be used throughout this paper. We use $[n]$ to denote the set $\{1, 2, \ldots, n\}$. We adopt the convention of using regular letters for scalars, and bold-face letters for vectors or matrices. For any matrix $\mathbf{A}$, we denote its (i, j)-th entry by A_{ij}, its transpose by $\mathbf{A}^\intercal$, its trace by $\mathrm{tr}(\mathbf{A})$, its j-th largest eigenvalue by $\lambda_j(\mathbf{A})$ (when the eigenvalues of $\mathbf{A}$ are real), its spectral norm by $\|\mathbf{A}\|_2 = \sqrt{\lambda_1(\mathbf{A}\mathbf{A}^\intercal)}$, and its element-wise maximum norm by $\|\mathbf{A}\|_{\max} = \max_{i,j} |A_{ij}|$. We use $\mathrm{diag}(\mathbf{A})$ to denote the diagonal matrix of $\mathbf{A}$ (replacing all off-diagonal entries with zero). For a sequence of random variables $\{X_n\}_{n=1}^\infty$ and a corresponding set of nonnegative real numbers $\{a_n\}_{n=1}^\infty$, we write $X_n = O_P(a_n)$ if $X_n/a_n = O_P(1)$ (bounded in probability), and we write $X_n = o_P(a_n)$ if $X_n/a_n \to 0$ in probability. For any univariate function f, we denote $f(\mathbf{A}) = [f(A_{ij})]$ as a matrix with f applied on each entry of $\mathbf{A}$. Throughout this paper, C stands for some positive constant whose value is not important and may change from line to line. The notation "$i_1 \neq i_2 \neq \cdots \neq i_m$" indicates that the m indices $\{i_\ell\}_{\ell=1}^m$ are pairwise different. All limits are for $n \to \infty$, unless explicitly stated otherwise.

The rest of this article is organized as follows. Section 2 proposes a new estimator for the number of common factors in the factor model (1). The consistency of our estimator is proven based on the spectral properties of the Spearman sample correlation matrix. Section 3 offers comprehensive simulation experiments, comparing our estimator with others. In Section 4, we evaluate the performance of the proposed estimator on a real dataset. A brief discussion is given in Section 5. Section 6 presents some important technical lemmas and proofs. Auxiliary lemmas and technical proofs are relegated to the supplementary material.

2 Main results

2.1 Spearman correlation matrix

For p-dimensional i.i.d. data sample $\{\mathbf{y}_i\}_{i=1}^n$, we denote the ranks of the data as follows:

$$\mathbf{Y}_n = \begin{pmatrix} \mathbf{y}_1^\intercal \\ \vdots \\ \mathbf{y}_n^\intercal \end{pmatrix} = \underbrace{\begin{pmatrix} y_{11} & \cdots & y_{1p} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{np} \end{pmatrix}}_{\text{raw data matrix}} \Longrightarrow \underbrace{\begin{pmatrix} r_{11} & \cdots & r_{1p} \\ \vdots & \ddots & \vdots \\ r_{n1} & \cdots & r_{np} \end{pmatrix}}_{\text{ranks matrix}},$$

where $r_{ij} = \sum_{\ell=1}^n \mathbb{1}\{y_{\ell j} \leqslant y_{ij}\}$ is the rank of y_{ij} among $\{y_{\ell j}\}_{\ell=1}^n$, and $\mathbb{1}\{\cdot\}$ denotes the indicator function. The Spearman correlation matrix of the raw data matrix $\mathbf{Y}_n$ is the Pearson correlation matrix of the ranks matrix. Define the normalized ranks matrix

$$\mathbf{R} = \sqrt{\frac{12}{n^2-1}}\left(r_{ij} - \frac{n+1}{2}\right)_{n\times p}, \tag{2}$$

and let $\mathbf{R}_i^\intercal$ be the i-th row of the matrix $\mathbf{R}$. The Spearman sample correlation matrix of $\mathbf{Y}_n$ is

$$\boldsymbol{\rho}_n = \frac{1}{n}\mathbf{R}^\intercal\mathbf{R} = \frac{1}{n}\sum_{i=1}^n \mathbf{R}_i\mathbf{R}_i^\intercal. \tag{3}$$

The empirical spectral distribution (ESD) of $\boldsymbol{\rho}_n$ is referred to as a random measure $F^{\boldsymbol{\rho}_n} = p^{-1}\sum_{j=1}^p \delta_{\lambda_j(\boldsymbol{\rho}_n)}$, where $\delta_{\lambda_j(\boldsymbol{\rho}_n)}$ is the Dirac mass at the point $\lambda_j(\boldsymbol{\rho}_n)$. The limit of $F^{\boldsymbol{\rho}_n}$ is called limiting spectral distribution (LSD). Under the assumption that the components of $\mathbf{y}_i$ are i.i.d., Bai and Zhou (2008) proved that the LSD of $\boldsymbol{\rho}_n$ is the well-known Marčenko-Pastur distribution. Recently, Wu and Wang (2022) extended this result to the non-paranormal distribution. In this study, we further extend their findings to encompass the scale mixture of normal distributions (see Definition 2.1), as stated in Lemma 6.1.

Throughout this paper, we assume that both common factors and idiosyncratic errors follow continuous distributions. Therefore, with probability one, there are no ties among $\{y_{ij}, i \in [n]\}$ for each j. For any $j \in [p]$ and $i, \ell \in [n]$ with $i \neq \ell$, we have $\mathbb{1}\{y_{\ell j} \leqslant y_{ij}\} = \frac{1}{2} + \frac{1}{2}\mathrm{sign}(y_{ij} - y_{\ell j})$, where $\mathrm{sign}(\cdot)$ denotes the sign function. Hence, we have

$$r_{ij} - \frac{n+1}{2} = 1 + \frac{1}{2}\sum_{i \neq \ell}\big\{1 + \mathrm{sign}(y_{ij} - y_{\ell j})\big\} - \frac{n+1}{2} = \frac{1}{2}\sum_{i \neq \ell}\mathrm{sign}(y_{ij} - y_{\ell j}). \tag{4}$$

For two sample vectors $\mathbf{y}_i$ and $\mathbf{y}_\ell$, we define the sign vector

$$\mathbf{A}_{i\ell} = \mathrm{sign}(\mathbf{y}_i - \mathbf{y}_\ell) = \big(\mathrm{sign}(y_{i1} - y_{\ell 1}), \ldots, \mathrm{sign}(y_{ip} - y_{\ell p})\big)^\intercal.$$

Then, from (2) and (4), we can rewrite the Spearman sample correlation matrix (3) as

$$\boldsymbol{\rho}_n = \frac{3}{n(n^2-1)}\sum_{i=1}^{n}\sum_{\ell_1, \ell_2 \neq i}\mathbf{A}_{i\ell_1}\mathbf{A}_{i\ell_2}^\intercal.$$

The application of sign transformations to the data introduces an intractable nonlinear correlation structure. To address this challenge, we utilize *Hoeffding's decomposition* (Hoeffding, 1948) to handle the nonlinear correlation within $\mathbf{A}_{i\ell}$. By employing this decomposition, we can identify the dominant term of $\boldsymbol{\rho}_n$. Let $\mathbf{A}_i = \mathbb{E}(\mathbf{A}_{i\ell} \mid \mathbf{y}_i)$ with $i \neq \ell$, Hoeffding's decomposition of $\mathbf{A}_{i\ell}$ can be expressed as follows:

$$\mathbf{A}_{i\ell} = \mathbf{A}_i - \mathbf{A}_\ell + \boldsymbol{\varepsilon}_{i\ell}, \tag{5}$$

where $\boldsymbol{\varepsilon}_{i\ell} := \mathbf{A}_{i\ell} - \mathbf{A}_i + \mathbf{A}_\ell$. Note that $\mathbb{E}\mathbf{A}_{i\ell} = \mathbb{E}\mathbf{A}_i = \mathbf{0}$, and the covariance matrix of $\mathbf{A}_i$ is $\mathbb{E}(\mathbf{A}_i\mathbf{A}_i^\intercal)$. With Hoeffding's decomposition defined in (5), we have

$$\boldsymbol{\rho}_n = \frac{3}{n(n^2-1)}\Bigg\{\sum_{i=1}^{n}\sum_{\ell_1, \ell_2 \neq i}(\mathbf{A}_i - \mathbf{A}_{\ell_1})(\mathbf{A}_i - \mathbf{A}_{\ell_2})^\intercal + \sum_{i=1}^{n}\sum_{\ell_1, \ell_2 \neq i}(\mathbf{A}_i - \mathbf{A}_{\ell_1})\boldsymbol{\varepsilon}_{i\ell_2}^\intercal$$

$$+\sum_{i=1}^{n}\sum_{\ell_1,\ell_2\neq i}\boldsymbol{\varepsilon}_{i\ell_1}(\mathbf{A}_i-\mathbf{A}_{\ell_2})^\intercal+\sum_{i=1}^{n}\sum_{\ell_1,\ell_2\neq i}\boldsymbol{\varepsilon}_{i\ell_1}\boldsymbol{\varepsilon}_{i\ell_2}^\intercal\Bigg\}.$$

It will be shown that the cross-terms in the above identity are negligible (see Lemma 2.2 and its proof in Section S2.1 in the supplementary material). We can then focus on the first term,

$$\begin{aligned}&\frac{3}{n(n^2-1)}\sum_{i=1}^{n}\sum_{\ell_1,\ell_2\neq i}(\mathbf{A}_i-\mathbf{A}_{\ell_1})(\mathbf{A}_i-\mathbf{A}_{\ell_2})^\intercal\\ =&\frac{n-2}{n+1}\cdot\frac{3}{n}\sum_{i=1}^{n}\left\{\frac{1}{(n-1)(n-2)}\sum_{\ell_1\neq\ell_2\neq i}(\mathbf{A}_i-\mathbf{A}_{\ell_1})(\mathbf{A}_i-\mathbf{A}_{\ell_2})^\intercal\right\}+\frac{3}{n+1}\boldsymbol{\tau}_n,\end{aligned}\tag{6}$$

where $\boldsymbol{\tau}_n := \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\ell\neq i}(\mathbf{A}_i-\mathbf{A}_\ell)(\mathbf{A}_i-\mathbf{A}_\ell)^\intercal$ is the sample marginal Kendall's tau correlation matrix (Bandeira et al., 2017; Li et al., 2023). The second term, $3\boldsymbol{\tau}_n/(n+1)$, and the cross-terms in (6) are negligible (see Lemma 2.2 and its proof in Section S2.1 in the supplementary material). Hence, the leading order term of $\boldsymbol{\rho}_n$ is

$$\mathbf{W}_n=\frac{3}{n}\sum_{i=1}^{n}\mathbf{A}_i\mathbf{A}_i^\intercal.\tag{7}$$

Through direct calculations, it can be demonstrated that the difference between the expected values of $\mathbf{W}_n$ and $\boldsymbol{\rho}_n$ is of the order $O_P(n^{-1})$. To be more precise, we have $\mathbb{E}\boldsymbol{\rho}_n=\frac{n}{n+1}\mathbb{E}\mathbf{W}_n+\frac{3}{n+1}\mathbb{E}\boldsymbol{\varepsilon}_{12}\boldsymbol{\varepsilon}_{12}^\intercal$.

Under certain assumptions, we can show that the spectrum of $\boldsymbol{\rho}_n$ can be approximated by that of $\mathbf{W}_n$ (see Lemma 2.2). This allows us to study the spectral properties of $\boldsymbol{\rho}_n$ via those of $\mathbf{W}_n$. Under the scale mixture of normals framework defined in Section 2.2, we find that the population covariance matrix $\boldsymbol{\Sigma}_\rho=\mathbb{E}\mathbf{W}_n$ has a finite-rank perturbation structure with K spiked eigenvalues (see Lemma 2.3). Thus naturally $\mathbf{W}_n$ has K relatively large eigenvalues too. Subsequently, we can estimate the number of factors based on the top eigenvalues of $\mathbf{W}_n$ or $\boldsymbol{\rho}_n$. By establishing the phase-transition theory of the spiked eigenvalues of $\mathbf{W}_n$, the consistency of the new estimator follows.

2.2 Phase transition theory

From the perspective of RMT, the spectrum of $\mathbf{W}_n$ relies on the structure of $\mathbf{\Sigma}_\rho = \mathbb{E}\mathbf{W}_n$. Although from factor model (1), it is clear that $\mathbf{\Sigma}_y := \mathrm{Cov}(\mathbf{y}_i) = \mathbf{B}\mathbf{B}^\intercal + \mathbf{\Psi}$ has a finite-rank-K perturbation structure, the relationship between $\mathbf{\Sigma}_\rho$ and $\mathbf{\Sigma}_y$ is unclear. The structure of $\mathbf{\Sigma}_\rho$ changes for different distributions of $\mathbf{y}_i$. Therefore, extra distribution assumption of $\mathbf{y}_i$ is needed to maintain the finite-rank perturbation structure of $\mathbf{\Sigma}_\rho$. Specifically, we assume that both the common factors and the idiosyncratic errors follow a *scale mixture of normal distributions*, defined as follows:

Definition 2.1 (Scale mixture of normals, Andrews and Mallows (1974))**.** A p-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_p)^\intercal$ follows a *scale mixture of normal distributions* (SMN) if $\mathbf{X}$ has the stochastic representation $\mathbf{X} \overset{\mathsf{d}}{=} \sqrt{W}\mathbf{Z}$, where W is a scalar-valued random variable with positive support, and $\mathbf{Z}$ follows p-dimensional normal distribution $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$ independent of W, $\mathbf{\Sigma}$ is a positive semi-definite matrix. The notation "$X \overset{\mathsf{d}}{=} Y$" means X and Y have the same distribution.

Our motivation for using this scale mixture of normals is two-fold. First, the scale mixture of normals contains heavy-tailed distributions, such as Student's t distribution. If W follows the inverse Gamma distribution $\mathsf{invGamma}(\nu/2, \nu/2)$ with probability density function $g_W(w) = \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} w^{-(\nu/2+1)} \exp\{-\nu/(2w)\}$, then $\mathbf{X}$ follows the p-dimentional Student's t distribution $t_\nu(\mathbf{0}, \mathbf{\Sigma})$ with location parameter $\mathbf{0}$, scale matrix $\mathbf{\Sigma}$, and degrees of freedom ν, the probability density function of which is $f_{\mathbf{X}}(\mathbf{x}) = \frac{\Gamma((\nu+p)/2)}{\Gamma(\nu/2)\nu^{p/2}\pi^{p/2}|\mathbf{\Sigma}|^{1/2}}(1 + \frac{1}{\nu}\mathbf{x}^\intercal\mathbf{\Sigma}^{-1}\mathbf{x})^{-(\nu+p)/2}$. A number of well-known distributions can be written as scale mixtures of normals. We refer the readers to Section 2 of Heinen and Valdesogo (2020) for more examples. The second motivation is for technical advantage. From the fact that $\mathbf{X} \mid (W = w) \sim \mathcal{N}_p(\mathbf{0}, w\mathbf{\Sigma})$, we can relate the Spearman sample correlation matrix of $\mathbf{X}$ to the scale matrix $\mathbf{\Sigma}$ using Grothendieck's identity (see Lemma S1.5 in the supplementary material). Heinen and Valdesogo (2020) derived an explicit expression for Spearman correlation of bivariate scale mixture of normals. We extend this result to a more complicated bivariate population (see Lemma S1.6 in the supplementary material) and utilize it to examine the structure of $\mathbf{\Sigma}_\rho$ (see Lemma 2.3). This direct connection to the scale matrix $\mathbf{\Sigma}_\rho$ is a fundamental step in the analysis of our proposed estimator.

Furthermore, we need the following assumptions:

Assumption (A1). *As $n \to \infty$, $p = p(n) \to \infty$ and $p/n = c_n \to c \in (0, \infty)$.*

Assumption (A2). *All pairs $\{(\mathbf{f}_i^\intercal, \mathbf{e}_i^\intercal)^\intercal\}_{i=1}^n$ are i.i.d., and $\mathbf{f}_i$ is independent of $\mathbf{e}_i$, and both of them follow the scale mixture of normal distributions. Suppose that w_f and w_e are two independent random variables with positive support. The common factor $\mathbf{f}_i$ has a stochastic representation $\mathbf{f}_i \overset{\mathsf{d}}{=} \sqrt{w_i^f}\mathbf{x}_i$, where w_i^f is an independent copy of w_f, and $\mathbf{x}_i$ follows K-dimensional standard normal distribution. The idiosyncratic error $\mathbf{e}_i$ has a stochastic representation $\mathbf{e}_i \overset{\mathsf{d}}{=} (\sqrt{w_{i1}^e}z_{i1}, \dots, \sqrt{w_{ip}^e}z_{ip})^\intercal$, where $\{w_{ij}^e\}_{j=1}^p$ are i.i.d. copies of w_e, and $(z_{i1}, \dots, z_{ip})^\intercal$ follows p-dimensional standard normal distribution.*

Assumption (A3). *The loading matrix $\mathbf{B}$ is normalized by the constraint $\mathbf{B}^\intercal\mathbf{B} = \mathbf{I}_K$. All the entries of $\mathbf{B}$ are of order $O(p^{-1/2})$.*

Assumption (A4). *The matrix $\mathbf{\Psi}$ is diagonal with entries of order $O(1)$.*

Assumption (A1) is common in the RMT literature. Assumption (A2) pertains to the distribution of the common factors and the idiosyncratic errors, and allows them to be heavy-tailed. This assumption is crucial for examining the structure of the population covariance matrix $\mathbf{\Sigma}_\rho$. The constraint $\mathbf{B}^\intercal\mathbf{B} = \mathbf{I}_K$ in (A3) is a commonly used identifiability condition, see Bai and Li (2012). It follows that by singular value decomposition, we can represent $\mathbf{B}$ as $\mathbf{U}\mathbf{V}^\intercal$, where $\mathbf{U} \in \mathbb{R}^{p\times K}$ and $\mathbf{V} \in \mathbb{R}^{K\times K}$, and both have orthonormal columns. The column vectors of $\mathbf{U}$ and $\mathbf{V}$ are unit vectors in $\mathbb{R}^p$ and $\mathbb{R}^K$, respectively. Thus, the condition $B_{ij} = O(p^{-1/2})$ in (A3), for any $i \in [p]$ and $j \in [K]$, is not overly restrictive. This condition facilitates our technical proofs. Assumption (A4) is standard in the factor models literature.

In what follows, we develop some important spectral properties of the Spearman sample correlation matrix. First, we show that the spectrum of $\boldsymbol{\rho}_n$ can be approximated by that of $\mathbf{W}_n$, as stated in the following lemma.

Lemma 2.2. *Under Assumptions (A1) – (A4), for any $j \in [p]$, $|\lambda_j(\boldsymbol{\rho}_n) - \lambda_j(\mathbf{W}_n)| = O_P(n^{-1/2})$ as $n \to \infty$.*

The proof of this lemma is provided in the supplementary material. From this Lemma, we can investigate the properties of $\boldsymbol{\rho}_n$ through its surrogate $\mathbf{W}_n$. As $\boldsymbol{\Sigma}_\rho$ represents the expectation of $\mathbf{W}_n$, examining the structure of $\boldsymbol{\Sigma}_\rho$ provides us with valuable insights of $\mathbf{W}_n$. The spike structure of $\boldsymbol{\Sigma}_\rho$ is illustrated in the following lemma.

Lemma 2.3 (Finite-rank perturbation)**.** *Under Assumptions (A1) – (A4), we have*

$$\left\|\boldsymbol{\Sigma}_\rho - \left\{\mathrm{diag}\big(\mathbf{I}_p - \gamma\boldsymbol{\Psi}^{-1}\mathbf{B}\mathbf{B}^\intercal\boldsymbol{\Psi}^{-1}\big) + \gamma\boldsymbol{\Psi}^{-1}\mathbf{B}\mathbf{B}^\intercal\boldsymbol{\Psi}^{-1}\right\}\right\|_2 = o(1) \tag{8}$$

as $n \to \infty$, where $\gamma := (6/\pi)\mathbb{E}\big[w^f/\{(w_1^e + w_2^e)(w_3^e + w_4^e)\}^{1/2}\big]$, and w_j^e, $j = 1, 2, 3, 4$, are independent copies of w_e.

Remark 2.4. *Note that both $\boldsymbol{\Sigma}_\rho$ and its approximation in (8) are correlation-type matrices, and all the diagonal entries equal to one. Consequently, the average of their eigenvalues are both one, and their bulk eigenvalues are clustered around one.*

The proof of this lemma is provided in the supplementary material. In this lemma, we derive a "consistent" approximation of the population covariance matrix $\boldsymbol{\Sigma}_\rho$. From Weyl's lemma (Lemma S1.1 in the supplementary material), the spectrum of the matrix $\mathbf{\Sigma}_\rho$ can be approximated by that of a rank-K perturbation of a diagonal matrix. Intuitively, $\mathbf{\Sigma}_\rho$ would have at most K relatively larger eigenvalues. As for the sample counterpart, at most K spiked sample eigenvalues of $\boldsymbol{\rho}_n$ would lay outside the support of its LSD. Naturally, by counting the number of spiked eigenvalues of $\boldsymbol{\rho}_n$, we can obtain a promising estimator of total number of factors. However, a very important yet intuitive observation here is that, for $j \in [K]$, $\lambda_j(\boldsymbol{\rho}_n)$ is not always far away from the bulk eigenvalues $\{\lambda_j(\boldsymbol{\rho}_n)\}_{j=K+1}^p$. It depends on whether the signal $\lambda_j(\mathbf{\Sigma}_\rho)$ is strong enough. If $\lambda_j(\mathbf{\Sigma}_\rho)$ is too weak, $\lambda_j(\boldsymbol{\rho}_n)$ would lie on the boundary of the support of bulk eigenvalues. This phenomenon is commonly referred to as the *phase-transition phenomenon*, which is described in the following theorem.

Theorem 2.5 (Phase transition)**.** *For the high-dimensional factor model* (1)*, assume that Assumptions (A1) – (A4) hold, and the ESD of $\mathbf{\Sigma}_\rho$ tends to a proper probability measure H as $n \to \infty$. Denoting $\psi(\alpha) = \alpha + c\int \frac{t\alpha}{\alpha - t}\,\mathrm{d}H(t)$, we have*

(a) *For $j \in [K]$ satisfying $\psi'\big(\lambda_j(\mathbf{\Sigma}_\rho)\big) > 0$, the j-th sample eigenvalue of $\boldsymbol{\rho}_n$ converges almost surely to $\psi\big(\lambda_j(\mathbf{\Sigma}_\rho)\big)$, which is outside the support of the LSD of $\boldsymbol{\rho}_n$.*

(b) *For $j \in [K]$ satisfying $\psi'\big(\lambda_j(\mathbf{\Sigma}_\rho)\big) \leqslant 0$, the j-th sample eigenvalue of $\boldsymbol{\rho}_n$ converges almost surely to the right endpoint of the support of the LSD of $\boldsymbol{\rho}_n$.*

The proof of Theorem 2.5 can be found in Section 6.

Remark 2.6. *From Lemma 2.2, we can investigate the asymptotic behavior of spiked eigenvalues of $\boldsymbol{\rho}_n$ via those of $\mathbf{W}_n$. Although $\mathbf{W}_n = (3/n)\sum_{i=1}^n \mathbf{A}_i\mathbf{A}_i^\intercal$ is a Wishart-type random matrix, the nonlinear correlation structure of $\mathbf{A}_i$ makes it difficult to directly apply the current phase-transition analysis techniques. The reason is that the vectors $\{\mathbf{A}_i\}_{i=1}^n$ do not follow the commonly used* independent component structure *as in Bai and Yao (2008, 2012) and Jiang and Bai (2021). Specifically, the vector $\mathbf{A}_i$ cannot be written as $\mathbf{A}_i = \mathbf{\Sigma}^{1/2}\mathbf{x}_i$,* *where $\mathbf{\Sigma}$ is non-negative definite and all elements of $\mathbf{x}_i \in \mathbb{R}^p$ are i.i.d. with zero mean and unit variance. To remove the independent component structure assumption, we first show that replacing $\{\sqrt{3}\mathbf{A}_i\}_{i=1}^n$ in $\mathbf{W}_n$ with i.i.d. $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma}_\rho)$ random vectors does not change the asymptotic behavior of spiked eigenvalues of $\mathbf{W}_n$ (see Lemma 6.2 and Section 6.2 for more details). The substitution of $\{\sqrt{3}\mathbf{A}_i\}_{i=1}^n$ with i.i.d. $\mathcal{N}_p(0, \mathbf{\Sigma}_\rho)$ is feasible because each entry of $\mathbf{A}_i$ follows a $\mathsf{Uniform}(-1, 1)$ distribution and the universality phenomenon holds for light-tailed distributions. The universality phenomenon reveals that as long as $\mathbf{A}_i$ has light-tailed entries, the first-order asymptotic behavior of the eigenvalues of $\frac{3}{n}\sum_{i=1}^n \mathbf{A}_i\mathbf{A}_i^\intercal$ remains the same when replacing $\{\sqrt{3}\mathbf{A}_i\}_{i=1}^n$ with Gaussian vectors. To guarantee the feasibility of this replacement, we establish concentration properties related to certain quadratic forms and their higher-order moments under the nonlinear correlation structure (see Lemma S1.7 in the supplementary material). Then since Gaussian random vectors naturally follows the independent component structure, we can directly apply the phase transition theory in Bai and Yao (2008, 2012) and Jiang and Bai (2021) to complete the proof of Theorem 2.5.*

To summarize, we first establish in Lemma 2.2 that $|\lambda_j(\rho_n) - \lambda_j(\mathbf{W}_n)| = o_P(1)$, and subsequently turn to analyze the eigenvalues of $\mathbf{W}_n$. Secondly, we prove that $\mathbf{\Sigma}_\rho = \mathbb{E}\mathbf{W}_n$ exhibits a rank-K perturbation structure as in Lemma 2.3. Thirdly, we confirm the phase-transition phenomenon for $\lambda_j(\mathbf{W}_n)$, where $j \in [K]$. Therefore, the corresponding result of $\lambda_j(\boldsymbol{\rho}_n)$ follows naturally, as demonstrated in Theorem 2.5.

2.3 Estimation of the number of factors

With the phase-transition theory in Theorem 2.5, we now propose our new estimator for the number of factors. As stated in Theorem 2.5, if $\psi'\big(\lambda_j(\mathbf{\Sigma}_\rho)\big) \leqslant 0$ for some $j \in [K]$, the corresponding sample eigenvalue $\lambda_j(\boldsymbol{\rho}_n)$ will converge to the right endpoint of the support of the LSD of $\boldsymbol{\rho}_n$, which is also the limit of the largest noise eigenvalue $\lambda_{K+1}(\boldsymbol{\rho}_n)$. Hence, such weak factors will be merged into the noise component, making their signal undetectable.

By taking this into account, we define the *number of significant factors* as

$$K_0 = \#\{j \in [K] : \psi'\big(\lambda_j(\boldsymbol{\Sigma}_\rho)\big) > 0\}, \tag{9}$$

where the notation $\# \mathcal{S}$ denotes the cardinality number of the set $\mathcal{S}$. By Theorem 2.5, the leading K_0 eigenvalues of $\boldsymbol{\rho}_n$ will lay outside the support of its LSD .

The LSD of $\boldsymbol{\rho}_n$, denoted by $F_{c,H}$, is the generalized Marčenko-Pastur law as stated in Lemma 6.1. Let $\mathsf{supp}(F_{c,H})$ denote the support of $F_{c,H}$. The Stieltjes transform of $F_{c,H}$ is defined as $m(x) = \int \frac{1}{t-x}\,\mathrm{d}F_{c,H}(t)$ for $x \in \mathbb{R} \setminus \mathsf{supp}(F_{c,H})$. Its first-order derivative $m'(x)$ is also only defined outside $\mathsf{supp}(F_{c,H})$, and can be extended as a function mapping the entire real line $\mathbb{R}$ to $\mathbb{R} \cup \{+\infty\}$ as follows:

$$m'(x) = \begin{cases} \int \frac{1}{(t-x)^2}\,\mathrm{d}F_{c,H}(t), & \text{if } x \in \mathbb{R} \setminus \mathsf{supp}(F_{c,H}), \\ +\infty, & \text{if } x \in \mathsf{supp}(F_{c,H}). \end{cases} \tag{10}$$

This implies that $m'(\lambda_j(\boldsymbol{\rho}_n))$ takes either finite or infinite values, depending on whether $\lambda_j(\boldsymbol{\rho}_n)$ is a spiked eigenvalue or a bulk eigenvalue. Based on this observation, we utilize the derivative of the Stieltjes transform defined in (10) to identify all the spiked eigenvalues and estimate the total number of significant factors. Let $K_{\max}$ be a predetermined upper bound on the true number of significant factors, K_0. As the LSD H of $\boldsymbol{\Sigma}_\rho$ is unknown, we cannot obtain the explicit expression of $m'(x)$. Therefore, we utilize

$$\widehat{m}'_{n,j}(x) = \frac{1}{p-j} \sum_{\ell=j+1}^{p} \frac{1}{\big\{x - \lambda_\ell(\boldsymbol{\rho}_n)\big\}^2} \tag{11}$$

to estimate $m'(x)$ for $1 \leqslant j \leqslant K_{\max}$. Intuitively, if $\lambda_j(\boldsymbol{\rho}_n)$ lies within the bulk spectrum of $\boldsymbol{\rho}_n$, we would expect $\widehat{m}'_{n,j}\big(\lambda_j(\boldsymbol{\rho}_n)\big)$ to be very large. On the contrary, if $\lambda_j(\boldsymbol{\rho}_n)$ is a spiked eigenvalue, $\widehat{m}'_{n,j}\big(\lambda_j(\boldsymbol{\rho}_n)\big)$ should be relatively small. This phenomenon bears similarities to the behavior of $m'(x)$ described in the equation (10). Actually, it will be shown that

$$\widehat{m}'_{n,j}\big(\lambda_j(\boldsymbol{\rho}_n)\big) = \begin{cases} O_P(1), & \text{for } 1 \leqslant j \leqslant K_0, \\ O_P(p^{1/3}), & \text{for } K_0 + 1 \leqslant j \leqslant K_{\max}, \end{cases} \tag{12}$$

as $n \to \infty$ (see the proof of Theorem 2.8 in Section 6.3). Hence, a natural estimator of the number of significant factors is

$$\widehat{K}_{\mathtt{SR}} = \mathop{\arg\max}_{1 \leqslant j \leqslant K_{\max}} \frac{\widehat{m}'_{n,j+1}(\lambda_{j+1}(\boldsymbol{\rho}_n))}{\widehat{m}'_{n,j}(\lambda_j(\boldsymbol{\rho}_n))}, \tag{13}$$

where the "S" in subscript stands for Stieltjes transform, and the "R" stands for Ratio.

Remark 2.7. *Here, we explain our preference for choosing $\widehat{m}'_{n,j}(\cdot)$ over directly employing eigen-ratio type estimators. Under our current framework, both eigenvalues $\lambda_j(\boldsymbol{\rho}_n)$ and the ratios $\frac{\lambda_{j+1}(\boldsymbol{\rho}_n)}{\lambda_j(\boldsymbol{\rho}_n)}$ are of constant order for all $1 \leqslant j \leqslant K_{\max}$. The ratios of eigenvalues exhibit the following behavior:*

$$\frac{\lambda_{j+1}(\boldsymbol{\rho}_n)}{\lambda_j(\boldsymbol{\rho}_n)} \begin{cases} < 1, & 1 \leqslant j \leqslant K_0, \\ = 1 - \varepsilon_p, & K_0 + 1 \leqslant j \leqslant K_{\max}, \end{cases}$$

where $\varepsilon_p = o_P(1)$ and $\varepsilon_p > 0$. When factor signals are weak, the values of $\lambda_{j+1}(\boldsymbol{\rho}_n)/\lambda_j(\boldsymbol{\rho}_n)$ may be close to 1 for both $j = K_0$ and $j = K_0 + 1$, making it difficult to distinguish between them. This can lead to inaccurate estimation. Performing the derivative of Stieltjes transform $\widehat{m}'_{n,j}(\cdot)$ on the corresponding eigenvalues $\lambda_j(\boldsymbol{\rho}_n)$ significantly amplifies the ratio *at $j = K_0$. As shown in Section 6.3, we found out that*

$$\frac{\widehat{m}'_{n,j+1}(\lambda_{j+1}(\boldsymbol{\rho}_n))}{\widehat{m}'_{n,j}(\lambda_j(\boldsymbol{\rho}_n))} \begin{cases} = O_P(1), & 1 \leqslant j \leqslant K_0 - 1, \\ \to \infty, & j = K_0, \\ = O_P(1), & K_0 + 1 \leqslant j \leqslant K_{\max}. \end{cases}$$

As a result, the sequence of ratios $\{\widehat{m}'_{n,j+1}(\lambda_{j+1}(\boldsymbol{\rho}_n))/\widehat{m}'_{n,j}(\lambda_j(\boldsymbol{\rho}_n))\}$ blows up at $j = K_0$. Therefore, estimating K_0 using $\widehat{m}'_{n,j}(\lambda_j(\boldsymbol{\rho}_n))$ is more efficient compared to using ratios of $\lambda_j(\boldsymbol{\rho}_n)$. Based on this observation, we propose the `SR` *estimator.*

The consistency of this estimator is established in the following theorem, and its proof is postponed to Section 6.

Theorem 2.8 (Consistency of $\widehat{K}_{\mathtt{SR}}$)**.** *For the high-dimensional factor model* (1)*, assume that Assumptions (A1) – (A4) hold. Let K_0 be the number of significant factors defined in* (9) *and $\widehat{K}_{\mathtt{SR}}$ be the proposed estimator defined in* (11) – (13)*. Then, we have*

$$\lim_{n\to\infty} \mathbb{P}(\widehat{K}_{\mathtt{SR}} = K_0) = 1.$$

3 Simulation studies

In this section, we conduct some simulations to examine the finite sample performance of the proposed estimator. We compare several estimators in the current literature, including our `SR` estimator; the `NE` estimator (Nadakuditi and Edelman, 2008); the `ED` estimator (Onatski, 2010, 2012); the `BCV` estimator (Owen and Wang, 2016); the `MKTCR` estimator (Yu et al., 2019), as well as the `ACT` estimator (Fan et al., 2020). The `MKTCR` estimator is designed to handle heavy-tailed data and can only detect strong factors. Conversely, other estimators are also capable of identifying weak factors. Specifically, these competing estimators are defined as follows:

1. Our $\mathtt{SR}$ estimator: Here we use a slightly modified version of $\widehat{K}_{\mathtt{SR}}$ to circumvent numerical instability:
$$\widetilde{K}_{\mathtt{SR}} = \mathop{\arg\max}_{1\leqslant j\leqslant K_{\max}} \frac{\widetilde{m}'_{n,j+1}(\lambda_{j+1}(\boldsymbol{\rho}_n))}{\widetilde{m}'_{n,j}(\lambda_j(\boldsymbol{\rho}_n))}, \tag{14}$$
where $\widetilde{m}'_{n,j}(\lambda_j(\boldsymbol{\rho}_n)) = \frac{1}{p-j}\sum_{\ell=j+1}^{p} \frac{1}{\{\lambda_j(\boldsymbol{\rho}_n)-\lambda_\ell(\boldsymbol{\rho}_n)\}^2+p^{-4/3}}$. Here we add $p^{-4/3}$ to the denominator in case that $\lambda_j(\boldsymbol{\rho}_n) - \lambda_\ell(\boldsymbol{\rho}_n)$ is too small. Similar to $\widehat{m}'_{n,j}(\lambda_j(\boldsymbol{\rho}_n))$, $\widetilde{m}'_{n,j}(\lambda_j(\boldsymbol{\rho}_n))$ still satisfies (12). $\widetilde{K}_{\mathtt{SR}}$ also retains the same consistency properties as $\widehat{K}_{\mathtt{SR}}$. Detailed proof is provided in Section 6.3.

2. $\mathtt{NE}$ estimator: Let $\{\mathbf{y}_i\}_{i=1}^n$ be an i.i.d. sample from the factor model (1). The sample covariance matrix of $\{\mathbf{y}_i\}_{i=1}^n$ is defined as $\mathbf{S}_n = n^{-1}\sum_{i=1}^n(\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{y}_i - \overline{\mathbf{y}})^\intercal$, where $\overline{\mathbf{y}} = n^{-1}\sum_{i=1}^n \mathbf{y}_i$ is the sample mean. Based on the eigenvalues of $\mathbf{S}_n$, Nadakuditi and Edelman (2008) introduced the $\mathtt{NE}$ estimator as follows:
$$\widehat{K}_{\mathtt{NE}} = \mathop{\arg\min}_{0\leqslant j<\min(p,n)} \left\{\frac{1}{4}\left(\frac{n}{p}\right)^2 t_j^2 + 2(j+1)\right\},$$
where $t_j = p\big[(p-j)\{\sum_{i=j+1}^p \lambda_i(\mathbf{S}_n)\}^{-2}\sum_{i=j+1}^p \lambda_i^2(\mathbf{S}_n) - 1 - p/n\big] - p/n$.

3. $\mathtt{ED}$ estimator: Based on the eigenvalues of $\mathbf{S}_n$, Onatski (2010, 2012) proposed an eigenvalue difference criterion, defined as
$$\widehat{K}_{\mathtt{ED}} = \max\{1 \leqslant j \leqslant K_{\max}, \lambda_j(\mathbf{S}_n) - \lambda_{j+1}(\mathbf{S}_n) \geqslant \delta\},$$
where δ is a predetermined threshold calculated using a calibration method described in Onatski (2010, Section IV).

4. $\mathtt{BCV}$ estimator: Owen and Wang (2016) introduced an algorithm to determine the number of factors based on $\mathbf{S}_n$ and the bi-cross-validation (BCV) technique from

Owen and Perry (2009). This method involves randomly holding out some rows and some columns of the observed data, fitting a factor model to the held-in data, and comparing held-out data to corresponding fitted values. We utilize Owen and Wang's R package "esaBcv" to implement the BCV method in our simulation studies.

5. MKTCR estimators: The sample multivariate Kendall's tau matrix is defined as $\mathbf{K}_n = \frac{2}{n(n-1)}\sum_{1\leqslant i<\ell\leqslant n}\frac{(\mathbf{y}_i-\mathbf{y}_\ell)(\mathbf{y}_i-\mathbf{y}_\ell)^\intercal}{\|\mathbf{y}_i-\mathbf{y}_\ell\|^2}$. Based on the eigenvalues of $\mathbf{K}_n$, Yu et al. (2019) constructed the MKTCR estimator as follows:

$$\widehat{K}_{\mathtt{MKTCR}} = \mathop{\arg\max}_{1\leqslant j\leqslant K_{\max}} \frac{\ln\{1+\lambda_j(\mathbf{K}_n)/V_{j-1}\}}{\ln\{1+\lambda_{j+1}(\mathbf{K}_n)/V_j\}},$$

where $V_j = \sum_{i=j+1}^{\min(p,n)}\lambda_i(\mathbf{K}_n)$, $0\leqslant j\leqslant \min(p,n)-1$.

6. ACT estimator: The sample Pearson correlation matrix of $\{\mathbf{y}_i\}_{i=1}^n$ is defined as $\mathbf{P}_n = \left[\operatorname{diag}(\mathbf{S}_n)\right]^{-1/2}\mathbf{S}_n\left[\operatorname{diag}(\mathbf{S}_n)\right]^{-1/2}$. Based on the spectral properties of $\mathbf{P}_n$, Fan et al. (2020) proposed an estimator to estimate the factor number as follows:

$$\widehat{K}_{\mathtt{ACT}} = \max\left\{1\leqslant j\leqslant K_{\max} : \widehat{\alpha}_j(\mathbf{P}_n) > 1+\sqrt{p/(n-1)}\right\},$$

where $\{\widehat{\alpha}_j(\mathbf{P}_n)\}_{j=1}^p$ are bias correction of sample eigenvalues of $\mathbf{P}_n$, defined as $\widehat{\alpha}_j(\mathbf{P}_n) = -1/\underline{m}_{n,j}(\lambda_j(\mathbf{P}_n))$ with $\underline{m}_{n,j}(x) = -(1-c_j)/x + c_j m_{n,j}(x)$, $c_j = (p-j)/(n-1)$, and

$$m_{n,j}(x) = \frac{1}{p-j}\left[\sum_{\ell=j+1}^{p}\frac{1}{\lambda_\ell(\mathbf{P}_n)-x} + \frac{1}{\{3\lambda_j(\mathbf{P}_n)+\lambda_{j+1}(\mathbf{P}_n)\}/4-x}\right].$$

Our simulation studies consider various combinations of dimension and sample size, namely $(p,n) = (50,100)$, $(100,200)$, $(150,300)$, and $(200,400)$, which all have the ratio $p/n = 1/2$. We take the true number of common factors $K=3$ and set the possible maximum value of the number of common factors $K_{\max} = 20$. Recalling our distribution assumption (A2) for both common factors $\mathbf{f}_i$ and idiosyncratic errors $\mathbf{e}_i$, we generate $\mathbf{f}_i$ and $\mathbf{e}_i$ by $\mathbf{f}_i = (w_i^f)^{1/2}\mathbf{x}_i$ and $e_{ij} = (w_{ij}^e)^{1/2} z_{ij}$, where e_{ij} denotes the j-th component of $\mathbf{e}_i$, $\{\mathbf{x}_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{N}_K(\mathbf{0}, \mathbf{I}_K)$, and $\{z_{ij}, i \in [n], j \in [p]\} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. We employ four different scenarios to generate sample data for w_i^f and w_{ij}^e:

1. (Normal population, see Table 1) Let $w_i^f = w_{ij}^e = 1$ for all $i \in [n]$ and $j \in [p]$;

2. (Uniform and Chi-squared population, see Table 1) Let $\{w_i^f\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathsf{Uniform}(0,1)$ and $\{w_{ij}^e, i \in [n], j \in [p]\} \overset{\text{i.i.d.}}{\sim} \chi^2(1)$;

3. (Student's $t(2)$ population, see Table 2) Let $\{w_i^f\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathsf{invGamma}(1,1)$ and $\{w_{ij}^e, i \in [n], j \in [p]\} \overset{\text{i.i.d.}}{\sim} \mathsf{invGamma}(1,1)$, where $\mathsf{invGamma}(\alpha,\beta)$ denotes the inverse Gamma distribution with shape parameter α and scale parameter β. In this scenario, both $\mathbf{f}_i$ and e_{ij} follow (multivariate) Student's $t(2)$ distributions;

4. (Cauchy population, see Table 2) Let $\{w_i^f\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathsf{invGamma}(1/2,1/2)$ and $\{w_{ij}^e, i \in [n], j \in [p]\} \overset{\text{i.i.d.}}{\sim} \mathsf{invGamma}(1/2,1/2)$. In this scenario, both $\mathbf{f}_i$ and e_{ij} follow (multivariate) Cauchy distribution.

Furthermore, we consider three cases for the loading matrix $\mathbf{B} = (B_{ij})_{p\times K}$ and the matrix $\mathbf{\Psi}$ as follows. (C1) is from (Harding, 2013; Fan et al., 2020). (C2) and (C3) are both from Onatski (2012).

(C1) For any $j \in [K]$, let $B_{ij} = \sqrt{5j/p}$ for $i \in [K]$, and let $B_{ij} = a_{ij}\sqrt{5j/(p-j)}$ for $i \in \{K+1,\dots,p\}$, where $a_{ij} = -1$ if $i = rj$ or $a_{ij} = 1$ if $i \neq rj$, $r \in \mathbb{N}^+$. Let $\mathbf{\Psi} = \mathbf{I}_p$.

(C2) For any $i \in [p]$ and $j \in [K]$, let $\sqrt{p}B_{ij}/\sqrt{10j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Let $\mathbf{\Psi} = \mathbf{I}_p$.

(C3) For any $i \in [p]$ and $j \in [K]$, let $\sqrt{p}B_{ij}/\sqrt{10j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Let $\mathbf{\Psi} = \mathbf{T}^{1/2}$, where $\mathbf{T}$ is a Toeplitz matrix with its (i,j)-th entry equal to $0.45^{|i-j|}$.

The simulation results are reported in Tables 1 - 2 and Figures 2 - 5. In the case of light-tailed data (see Tables 1 and Figures 2 - 3), our SR estimator performs comparably to other estimators. However, when handling heavy-tailed data (see Tables 2 and Figures 4 - 5), the NE, ED, and BCV estimators, which are based on sample covariance matrices, prove ineffective, whereas our SR estimator outperforms the MKTCR and ACT estimators.

Table 1: Percentages (%) of estimated number of common factors in 1000 simulations. Entries of common factors and idiosyncratic errors are generated from **light-tailed** distributions. The results are reported in the form $a(b|c)$, in which a, b, c are the percentages of true estimates, overestimates, and underestimates, respectively. The notation "ave($\widehat{K}$)" denotes mean estimators for the case $(p, n) = (200, 400)$.

Case	p	NE	ED	BCV	MKTCR	ACT	SR
				Normal population			
(C1)	50	97.7(2.3|0)	98.9(1.1|0)	97.5(0.5|2)	86.1(0|13.9)	100(0|0)	**97.4**(0.5|2.1)
	100	97.2(2.8|0)	99(1|0)	100(0|0)	84.4(0|15.6)	100(0|0)	**99.8**(0|0.2)
	150	96.7(3.3|0)	99.5(0.5|0)	100(0|0)	80.2(0|19.8)	100(0|0)	**100**(0|0)
	200	96.3(3.7|0)	99.6(0.4|0)	100(0|0)	78.3(0|21.7)	99.9(0.1|0)	**100**(0|0)
	ave($\widehat{K}$)	3.038	3.004	3	2.783	3.001	**3**
(C2)	50	97.1(2.9|0)	98.8(1.2|0)	99.1(0.8|0.1)	91.7(0|8.3)	100(0|0)	**93.8**(0|6.2)
	100	97.9(2.1|0)	99.3(0.7|0)	100(0|0)	98(0|2)	100(0|0)	**100**(0|0)
	150	96.7(3.3|0)	99.6(0.4|0)	100(0|0)	100(0|0)	100(0|0)	**100**(0|0)
	200	97.1(2.9|0)	99.6(0.4|0)	100(0|0)	100(0|0)	100(0|0)	**100**(0|0)
	ave($\widehat{K}$)	3.03	3.007	3	3	3	**3**
(C3)	50	0(100|0)	97(3|0)	80.3(19.7|0)	99.6(0|0.4)	99.9(0.1|0)	**96.3**(0.3|3.4)
	100	0(100|0)	99(1|0)	82.9(17.1|0)	100(0|0)	72.2(27.8|0)	**99.9**(0.1|0)
	150	0(100|0)	99.4(0.6|0)	82.3(17.7|0)	100(0|0)	39.9(60.1|0)	**100**(0|0)
	200	0(100|0)	99.3(0.7|0)	77.7(22.3|0)	100(0|0)	4.4(95.6|0)	**100**(0|0)
	ave($\widehat{K}$)	55.225	3.008	3.28	3	5.448	**3**
				Uniform and Chi-squared population			
(C1)	50	37.4(62.3|0.3)	58.7(2.7|38.6)	23.2(0.4|76.4)	22.4(0|77.6)	89.5(0|10.5)	**88.3**(0.6|11.1)
	100	26.7(73.3|0)	91(1.1|7.9)	34.4(0|65.6)	12.9(0|87.1)	99.8(0.1|0.1)	**98.9**(0.1|1)
	150	22.7(77.2|0.1)	97.7(0.3|2)	36.4(0.1|63.5)	5.1(0|94.9)	99.6(0.4|0)	**99.9**(0|0.1)
	200	23.3(76.7|0)	99.1(0.8|0.1)	35.6(0|64.4)	3.4(0|96.6)	99.4(0.6|0)	**100**(0|0)
	ave($\widehat{K}$)	4.204	3.009	2.356	1.919	3.006	**3**
(C2)	50	34.6(65.3|0.1)	91.5(2.4|6.1)	63.6(1.7|34.7)	27.2(0|72.8)	96.7(0|3.3)	**87.1**(0.2|12.7)
	100	23.9(76.1|0)	98.7(1.3|0)	99.8(0.2|0)	99.3(0|0.7)	100(0|0)	**100**(0|0)
	150	22.1(77.9|0)	99.3(0.7|0)	100(0|0)	82.8(0|17.2)	99.8(0.2|0)	**100**(0|0)
	200	16.6(83.4|0)	99(1|0)	100(0|0)	98.5(0|1.5)	100(0|0)	**100**(0|0)
	ave($\widehat{K}$)	4.343	3.01	3	2.985	3	**3**
(C3)	50	0(100|0)	96.5(2.6|0.9)	58.3(41.6|0.1)	57.7(0|42.3)	94.8(5.2|0)	**75.8**(2.2|22)
	100	0(100|0)	98.2(1.8|0)	79.7(20.3|0)	99.9(0|0.1)	27.7(72.3|0)	**99.5**(0.5|0)
	150	0(100|0)	99.3(0.7|0)	82(18|0)	97.9(0|2.1)	5.3(94.7|0)	**100**(0|0)
	200	0(100|0)	99.2(0.8|0)	84(16|0)	100(0|0)	0.1(99.9|0)	**100**(0|0)
	ave($\widehat{K}$)	56.198	3.01	3.193	3	7.554	**3**

Table 2: Percentages (%) of estimated number of common factors in 1000 simulations. Entries of common factors and idiosyncratic errors are generated from **heavy-tailed** distributions. The results are reported in the form $a(b|c)$, in which a, b, c are the percentages of true estimates, overestimates, and underestimates, respectively. The notation "ave($\widehat{K}$)" denotes mean estimators for the case $(p, n) = (200, 400)$.

Case	p	NE	ED	BCV	MKTCR	ACT	SR
				$t(2)$ **population**			
(C1)	50	0(100|0)	16.9(29.4|53.7)	26.6(6.3|67.1)	22.6(0.2|77.2)	83.3(0.3|16.4)	**93.2**(0|6.8)
	100	0(100|0)	16.1(33.6|50.3)	31.3(4|64.7)	15.2(0|84.8)	93.5(4.3|2.2)	**99.6**(0|0.4)
	150	0(100|0)	13.8(33.2|53)	34(4.2|61.8)	12.9(0|87.1)	91.9(7.4|0.7)	**100**(0|0)
	200	0(100|0)	14.1(38|47.9)	37.9(2.9|59.2)	11(0|89)	87.6(11.6|0.8)	**100**(0|0)
	ave($\widehat{K}$)	37.884	3.316	2.228	1.941	3.12	**3**
(C2)	50	0(100|0)	24.1(44.2|31.7)	47.6(10.6|41.8)	26.1(0.2|73.7)	90.4(0.3|9.3)	**92.2**(0|7.8)
	100	0(100|0)	19.1(44.9|36)	59.3(8.3|32.4)	31.2(0|68.8)	96(2.9|1.1)	**100**(0|0)
	150	0(100|0)	14.4(49.7|35.9)	80(4|16)	77.5(0|22.5)	93.3(6.6|0.1)	**100**(0|0)
	200	0(100|0)	14.6(53.1|32.3)	85(2.1|12.9)	94.8(0|5.2)	89.4(10.5|0.1)	**100**(0|0)
	ave($\widehat{K}$)	38.064	4.501	2.836	2.944	3.109	**3**
(C3)	50	0(100|0)	28.8(50.1|21.1)	36.8(49.5|13.7)	38.4(0.1|61.5)	67.5(28.8|3.7)	**78.1**(2.3|19.6)
	100	0(100|0)	19.9(57.1|23)	47.5(43.6|8.9)	57.8(0|42.2)	18(81.4|0.6)	**94.4**(2.8|2.8)
	150	0(100|0)	14.8(64.5|20.7)	66.8(28.6|4.6)	92.6(0|7.4)	6.3(93.5|0.2)	**99.9**(0.1|0)
	200	0(100|0)	13.1(66.3|20.6)	72.2(24.8|3)	99.3(0|0.7)	2.6(97.3|0.1)	**100**(0|0)
	ave($\widehat{K}$)	70.59	4.911	3.47	2.993	7.959	**3**
				Cauchy population			
(C1)	50	0(100|0)	11.2(77.6|11.2)	0.8(0.3|98.9)	7.7(5.9|86.4)	24.9(2.1|73)	**90.1**(0.2|9.7)
	100	0(100|0)	7.2(85|7.8)	0.1(0.1|99.8)	6.8(5|88.2)	38(12.4|49.6)	**98.9**(0.2|0.9)
	150	0(100|0)	8(84.4|7.6)	0.1(0|99.9)	6.3(4.2|89.5)	36.7(28.4|34.9)	**99.2**(0.7|0.1)
	200	0(100|0)	7.5(86.5|6)	0.2(0.1|99.7)	4.9(4.8|90.3)	35.7(41.7|22.6)	**99.1**(0.9|0)
	ave($\widehat{K}$)	118.068	8.138	0.038	1.507	3.439	**3.009**
(C2)	50	0(100|0)	12.4(76.9|10.7)	1(0.7|98.3)	6.8(2.6|90.6)	29.1(2.3|68.6)	**92**(0|8)
	100	0(100|0)	8.5(83.3|8.2)	0.1(0.1|99.8)	6.3(3.2|90.5)	49.9(17|33.1)	**100**(0|0)
	150	0(100|0)	8(86.8|5.2)	0.4(0|99.6)	6.2(2.1|91.7)	46.7(35.6|17.7)	**100**(0|0)
	200	0(100|0)	5.6(88|6.4)	0.1(0.2|99.7)	6.4(2.8|90.8)	37(49.5|13.5)	**100**(0|0)
	ave($\widehat{K}$)	118.511	8.236	0.069	1.486	3.596	**3**
(C3)	50	0(100|0)	12.6(77.5|9.9)	2.5(10.1|87.4)	5.6(3.8|90.6)	13.9(74.6|11.5)	**15.5**(17.8|66.7)
	100	0(100|0)	10.2(81|8.8)	1.9(4.7|93.4)	5.6(1.4|93)	3.3(93.4|3.3)	**46**(27.3|26.7)
	150	0(100|0)	8.8(83.5|7.7)	1.1(4.8|94.1)	3.7(1.3|95)	0.8(97.3|1.9)	**92.3**(7.5|0.2)
	200	0(100|0)	5.7(87.2|7.1)	1.5(3.5|95)	4.6(1.2|94.2)	0.8(98.5|0.7)	**93.3**(6.5|0.2)
	ave($\widehat{K}$)	126.004	7.919	0.421	1.416	16.854	**3.268**

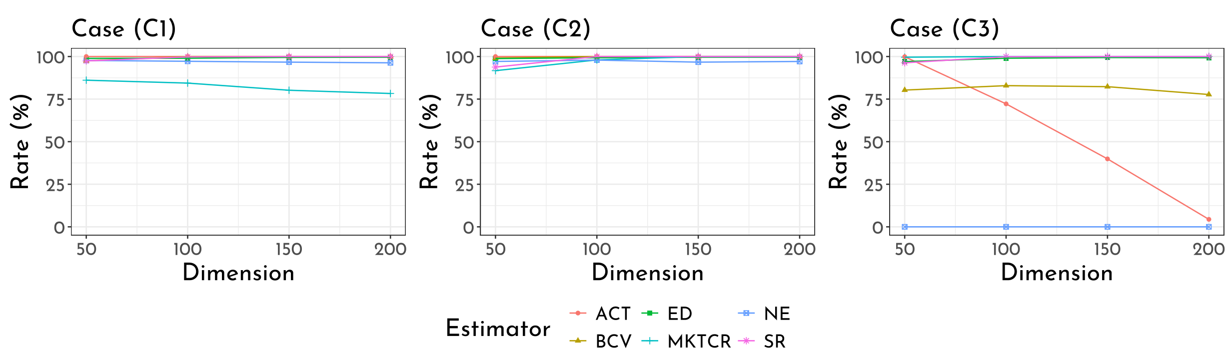


Figure 2: Correct identification rate of six estimators. Entries of common factors and idiosyncratic errors are generated from the standard normal distribution.

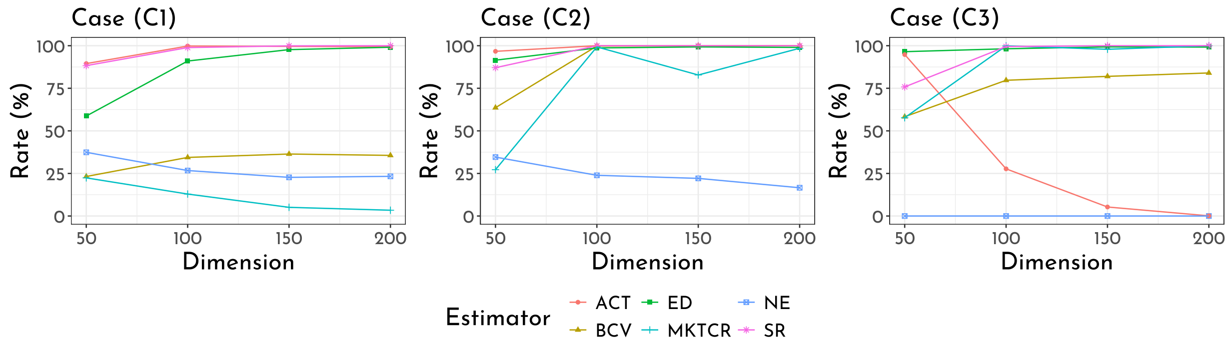


Figure 3: Correct identification rate of six estimators. Entries of common factors are generated from a scale mixture of normals with $\{w_i^f\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathsf{Uniform}(0,1)$, and entries of idiosyncratic errors are generated from a scale mixture of normals with $\{w_{ij}^e, i \in [n], j \in [p]\} \overset{\text{i.i.d.}}{\sim} \chi^2(1)$.

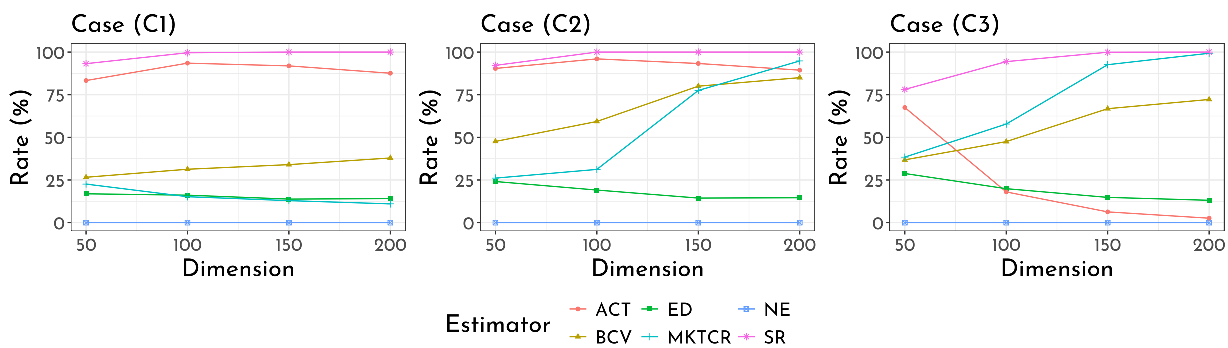


Figure 4: Correct identification rate of six estimators. Entries of common factors and idiosyncratic errors are generated from Student's $t(2)$ distribution.

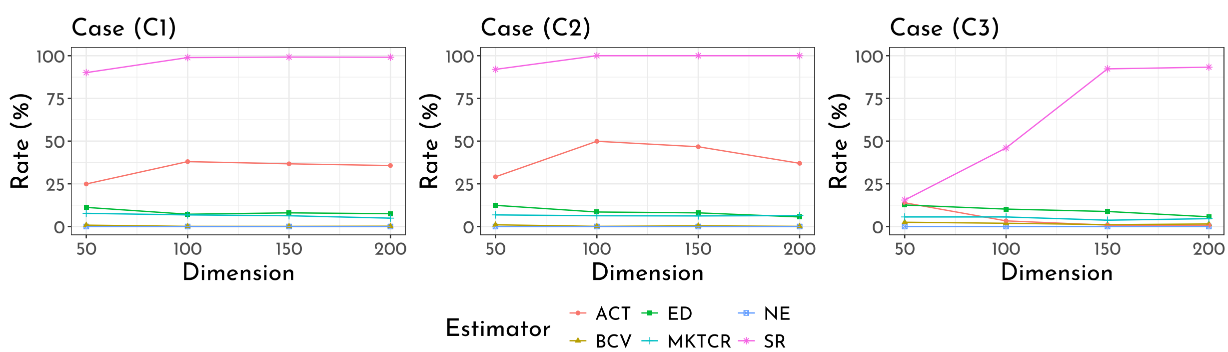
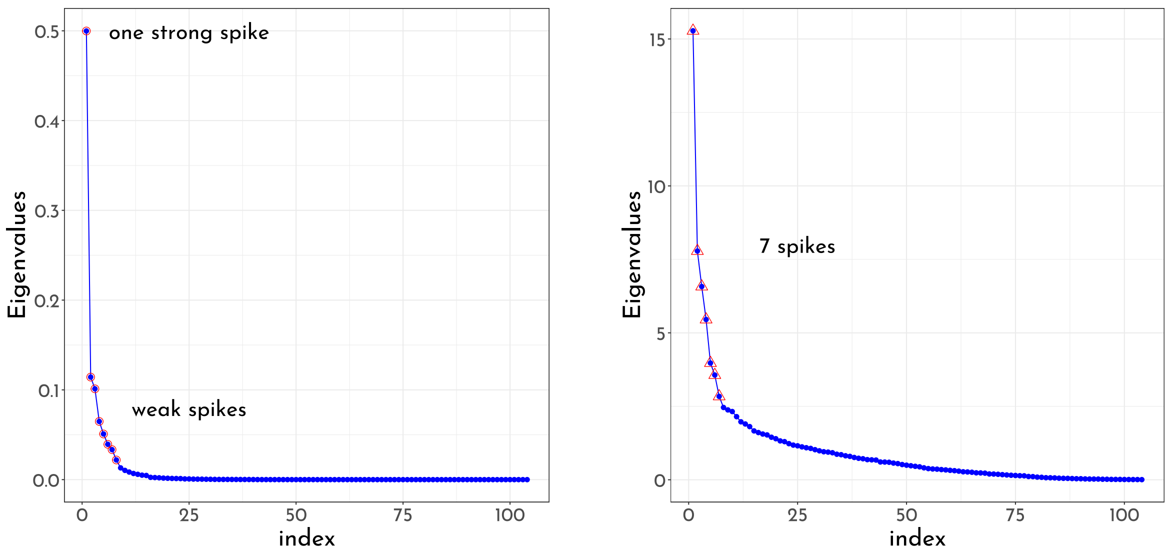


Figure 5: Correct identification rate of six estimators. Entries of common factors and idiosyncratic errors are generated from the standard Cauchy distribution.

4 Real data analysis

In this section, we analyze the monthly macroeconomic dataset (FRED-MD, McCracken and Ng (2016)) from March 1959 to January 2023. The data can be downloaded from the website `http://research.stlouisfed.org/econ/mccracken/fred-md/`, and includes the monthly series of 128 macroeconomic variables. Following McCracken and Ng (2016), the series with missing values are removed and the remaining dataset is transformed to a stationary form. After this preprocessing procedure, the data dimension is $p = 105$ and the sample size is $n = 767$. McCracken and Ng (2016)'s recommendation to remove outliers has not been implemented in our data analysis, as we believe that data with heavy-tailed distributions will inevitably contain extreme observations that cannot be circumvented. Since our estimator is tailored to heavy-tailed observations, we directly use it to identify the number of factors.

The dataset reveals that more than 67% of the macroeconomic variables exhibit a sample kurtosis that exceeds 9, which is the theoretical kurtosis of the Student's $t(5)$ distribution. This indicates that the dataset is probably heavy-tailed. Compared to other estimators, `MKTCR`, `ACT`, and `SR` have slightly higher accuracy under heavy-tailed conditions, so we employ these three methods for estimation. The results are as follows: $\widehat{K}_{\texttt{ACT}} = 13$, $\widehat{K}_{\texttt{MKTCR}} = 1$, and $\widehat{K}_{\texttt{SR}} = 7$. As shown in the simulation studies in Section 3, `ACT` has similar performance to our estimator when data is light-tailed, while it tends to overestimate when data is heavy-tailed. The same story happens for this real dataset. Both our `SR` estimator and Yu et al. (2019)'s `MKTCR` estimator are based on eigenvalues of certain type of sample correlation matrices as plotted in Figure 6. From Figure 6(a), it is evident that the multivariate Kendall's tau matrix exhibits one "strong" spike and several "weak" spikes. However, the `MKTCR` estimator only detects the strong spike while ignoring the weaker spikes. It potentially underestimates the total number of factors, similarly as shown in the simulation studies in Section 3. On the other hand, Figure 6(b) illustrates that our SR estimator has successfully detected all seven spikes of the Spearman sample correlation matrix. Therefore, $\widehat{K}_{\texttt{SR}} = 7$ is a more persuasive estimation for this dataset.



(a) Multivariate Kendall's tau matrix (b) Spearman correlation matrix

Figure 6: Scatter plots of all the eigenvalues of multivariate Kendall's tau matrix and Spearman correlation matrix generated from the real dataset. The MKTCR estimator only detects one "strong" spiked eigenvalue of the multivariate Kendall's tau matrix, and neglects several "weak" spikes. Our SR estimator detects all seven spikes of Spearman correlation matrix.

5 Discussions

In summary, we propose a novel estimator to identify the number of common factors in high-dimensional factor models when the data is heavy-tailed. We demonstrate that, under certain assumptions, the number of spiked eigenvalues of the Spearman sample correlation matrix is consistent with the total number of significant factors. Our estimator is constructed based on this observation, and its consistency is proved under mild assumptions. From the perspective of RMT, we investigate the eigenstructure of the Spearman sample correlation matrix under spike models and establish the phase-transition theory of its spiked eigenvalues. Simulation results demonstrate that our proposed estimator outperforms competing methods in various scenarios, especially with heavy-tailed observations. However, our SR estimator does not perform well when the sample size is not large enough, such as when $(p, n) = (50, 100)$. The possible reason is that the estimation of $m'(x)$ is inaccurate when the sample size is small. A more accurate estimator for $m'(x)$ would improve the accuracy of our SR estimator. Furthermore, it is worth extending our method for factor modeling in high-dimensional time series (Lam and Yao, 2012; Li et al., 2017b) and tensor data (Lam, 2021; Chen and Lam, 2024). These extensions are beyond the scope of the current paper, and we leave them for future work.

6 Proof of Theorems 2.5 and 2.8

6.1 Some technical lemmas

In this section, we propose two technical lemmas in preparation for proving Theorems 2.5 and 2.8. The proofs of these lemmas are relegated to supplementary material.

Lemma 6.1 provides the LSD of $\boldsymbol{\rho}_n$ and $\mathbf{W}_n$, extending the result of Wu and Wang (2022). Their result is restricted to the non-paranormal distribution, and our study considers the case where the data follows a scale mixture of normal distributions, as indicated in Assumption (A2).

Lemma 6.1 (Limiting spectral distribution)**.** *For the high-dimensional factor model* (1)*, assume that Assumptions (A1) – (A4) hold, and the ESD of $\boldsymbol{\Sigma}_\rho = \mathbb{E}\mathbf{W}_n$ tends to a proper probability measure H as $n \to \infty$. Then, with probability one, both $F^{\boldsymbol{\rho}_n}$ and $F^{\mathbf{W}_n}$ tend to a non-random probability distribution $F_{c,H}$, the Stieltjes transform $m = m(z)$ $(z \in \mathbb{C}^+)$ of which is the unique solution to the equation*

$$m = \int \frac{1}{t(1 - c - czm) - z} \, \mathrm{d}H(t). \tag{15}$$

The following Lemma 6.2 concerns the limiting behavior of $\boldsymbol{\Omega}_K(\cdot, \cdot)$ defined in (18), which plays a crucial role in the proof of Theorem 2.5.

Lemma 6.2. *Let* $\mathbf{X} = (X_{ij})_{n\times p} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\intercal$ *be an* $n \times p$ *random matrix. Assume that* $\mathbf{X}$ *satisfies Assumption (A1) and the following assumptions:*

(B1) The vectors $\{\mathbf{x}_\ell\}_{\ell=1}^n$ *are i.i.d., but the entries of each* $\mathbf{x}_\ell$ *are not necessarily i.i.d.*

(B2) (Moment condition) For any $i, j, s, t \in [p]$ *with* $i \neq j \neq s \neq t$*, we have* $\mathbb{E}X_{1i} = 0$, $\mathbb{E}X_{1i}^2 = 1$, $\mathbb{E}X_{1i}X_{1j} = 0$, $\mathbb{E}X_{1i}^4 = O(1)$, $\mathbb{E}X_{1i}^2X_{1j}X_{1s} = O(p^{-1})$, $\mathbb{E}X_{1i}X_{1j}X_{1s}X_{1t} = O(p^{-2})$.

(B3) (Weak dependency) For any $p \times p$ *symmetric matrix* $\mathbf{T}$ *with bounded spectral norm, we have* $\mathrm{Var}(\mathbf{x}_1^\intercal \mathbf{T}\mathbf{x}_1) = o(p^2)$ *as* $p \to \infty$.

(B4) (Concentration) For any convex 1*-Lipschitz (with respect to the Euclidean norm) function* F *from* $\mathbb{R}^p$ *to* $\mathbb{R}$*, let* m_F *denote a median of* F*, we have* $\mathbb{P}\big(|F(\mathbf{x}_1) - m_F| > t\big) \leqslant C\exp\big\{-c(p)t^2\big\}$*, where* C *and* $c(p)$ *are independent of* F*, and* C *is independent of* p*. We allow* $c(p)$ *to be a constant or to go to zero with* p *like* $p^{-\alpha}$, $0 \leqslant \alpha < 1$.

Moreover, let $\mathbf{Y} = (Y_{ij})_{n\times p} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^\intercal$ *be a random matrix independent of* $\mathbf{X}$*, satisfying Assumptions (A1) and (B1) – (B4) with* X_{ij} *and* $\mathbf{x}_\ell$ *replaced by* Y_{ij} *and* $\mathbf{y}_\ell$*, respectively. Then,* $\mathbf{\Omega}_K(\lambda, \mathbf{X})$ *and* $\mathbf{\Omega}_K(\lambda, \mathbf{Y})$ *have the same limiting distribution, where* $\mathbf{\Omega}_K(\cdot, \cdot)$ *is defined in* (18).

Remark 6.3. *Assumption (B4) is from El Karoui (2009). In (El Karoui, 2009, p. 2386), the author gave some examples of distributions satisfying Assumption (B4), such as:*

- *Gaussian random vectors with covariance matrix* $\mathbf{\Sigma}_p$ *and* $c(p) = 1/\|\mathbf{\Sigma}_p\|_2$ *(according to Theorem 2.7 in Ledoux (2001)).*
- *Random vectors with i.i.d. entries bounded by* $1/\sqrt{c(p)}$ *(according to Lemma S1.3).*

6.2 Proof of Theorem 2.5

From Lemma 2.2, we investigate the phase-transition theory of spiked eigenvalues of $\boldsymbol{\rho}_n$ by those of $\mathbf{W}_n = (3/n)\sum_{i=1}^n \mathbf{A}_i\mathbf{A}_i^\intercal$. Recall that $\boldsymbol{\Sigma}_\rho = \mathbb{E}\mathbf{W}_n$. Define the spectral decomposition of $\boldsymbol{\Sigma}_\rho$ as $\boldsymbol{\Sigma}_\rho = \mathbf{U}\left(\begin{smallmatrix}\mathbf{D}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{D}_2\end{smallmatrix}\right)\mathbf{U}^\intercal$, where $\mathbf{U}$ is a $p\times p$ orthogonal matrix, $\mathbf{D}_1$ is the diagonal matrix consisting of the K spiked population eigenvalues, and $\mathbf{D}_2$ is the diagonal matrix consisting of the remaining $p-K$ non-spiked eigenvalues. Let $\widetilde{\mathbf{A}}_i := \sqrt{3}\boldsymbol{\Sigma}_\rho^{-1/2}\mathbf{A}_i$ denote a transformed version of $\sqrt{3}\mathbf{A}_i$. It is obvious that $\widetilde{\mathbf{A}}_i$ is isotropic, that is, $\mathrm{Cov}(\widetilde{\mathbf{A}}_i) = \mathbf{I}_p$. By using these notations and the spectral decomposition of $\boldsymbol{\Sigma}_\rho$, we have

$$0 = |\lambda\mathbf{I}_p - \mathbf{W}_n| = \left|\lambda\mathbf{I}_p - \mathbf{U}\begin{pmatrix}\mathbf{D}_1^{1/2} & \\ & \mathbf{D}_2^{1/2}\end{pmatrix}\mathbf{U}^\intercal\widetilde{\mathbf{W}}_n\mathbf{U}\begin{pmatrix}\mathbf{D}_1^{1/2} & \\ & \mathbf{D}_2^{1/2}\end{pmatrix}\mathbf{U}^\intercal\right|, \tag{16}$$

where $\widetilde{\mathbf{W}}_n := n^{-1}\boldsymbol{\mathcal{A}}^\intercal\boldsymbol{\mathcal{A}}$ with $\boldsymbol{\mathcal{A}} := (\widetilde{\mathbf{A}}_1,\ldots,\widetilde{\mathbf{A}}_n)^\intercal$. Let $\mathbf{Q} = \mathbf{U}^\intercal\widetilde{\mathbf{W}}_n\mathbf{U}$ and partition it as $\mathbf{Q} = \left(\begin{smallmatrix}\mathbf{Q}_{11} & \mathbf{Q}_{12}\\ \mathbf{Q}_{21} & \mathbf{Q}_{22}\end{smallmatrix}\right) = \left(\begin{smallmatrix}\mathbf{U}_1^\intercal\widetilde{\mathbf{W}}_n\mathbf{U}_1 & \mathbf{U}_1^\intercal\widetilde{\mathbf{W}}_n\mathbf{U}_2\\ \mathbf{U}_2^\intercal\widetilde{\mathbf{W}}_n\mathbf{U}_1 & \mathbf{U}_2^\intercal\widetilde{\mathbf{W}}_n\mathbf{U}_2\end{smallmatrix}\right)$, where $\mathbf{U}_1$ is the submatrix formed by the first K columns of $\mathbf{U}$, and $\mathbf{U}_2$ is the remaining submatrix. Plugging this identity into (16) yields that

$$\begin{aligned}
0 &= \left|\lambda\mathbf{I}_p - \begin{pmatrix}\mathbf{D}_1^{1/2}\mathbf{Q}_{11}\mathbf{D}_1^{1/2} & \mathbf{D}_1^{1/2}\mathbf{Q}_{12}\mathbf{D}_2^{1/2}\\ \mathbf{D}_2^{1/2}\mathbf{Q}_{21}\mathbf{D}_1^{1/2} & \mathbf{D}_2^{1/2}\mathbf{Q}_{22}\mathbf{D}_2^{1/2}\end{pmatrix}\right|\\
&= \left|\lambda\mathbf{I}_{p-K} - \mathbf{D}_2^{1/2}\mathbf{Q}_{22}\mathbf{D}_2^{1/2}\right|\\
&\quad\times\left|\lambda\mathbf{I}_K - \mathbf{D}_1^{1/2}\mathbf{Q}_{11}\mathbf{D}_1^{1/2} - \mathbf{D}_1^{1/2}\mathbf{Q}_{12}\mathbf{D}_2^{1/2}(\lambda\mathbf{I}_{p-K} - \mathbf{D}_2^{1/2}\mathbf{Q}_{22}\mathbf{D}_2^{1/2})^{-1}\mathbf{D}_2^{1/2}\mathbf{Q}_{21}\mathbf{D}_1^{1/2}\right|,
\end{aligned}$$

where the last equality follows from the formula $\det\left(\begin{smallmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{smallmatrix}\right) = \det(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})\cdot\det(\mathbf{D})$. Suppose that λ is a spiked eigenvalue, then we have $|\lambda\mathbf{I}_{p-K} - \mathbf{D}_2^{1/2}\mathbf{Q}_{22}\mathbf{D}_2^{1/2}| \neq 0$, and

$$0 = \left|\lambda\mathbf{D}_1^{-1} - \mathbf{Q}_{11} - \mathbf{Q}_{12}\mathbf{D}_2^{1/2}(\lambda\mathbf{I}_{p-K} - \mathbf{D}_2^{1/2}\mathbf{Q}_{22}\mathbf{D}_2^{1/2})^{-1}\mathbf{D}_2^{1/2}\mathbf{Q}_{21}\right|$$

$$= \left|\lambda \mathbf{D}_1^{-1} - \frac{1}{n}\mathbf{U}_1^\intercal \boldsymbol{\mathcal{A}}^\intercal \left[\mathbf{I}_n + \frac{1}{n}\boldsymbol{\mathcal{A}}\mathbf{U}_2\mathbf{D}_2^{1/2}\left(\lambda \mathbf{I}_{p-K} - \frac{1}{n}\mathbf{D}_2^{1/2}\mathbf{U}_2^\intercal \boldsymbol{\mathcal{A}}^\intercal \boldsymbol{\mathcal{A}}\mathbf{U}_2\mathbf{D}_2^{1/2}\right)^{-1}\mathbf{D}_2^{1/2}\mathbf{U}_2^\intercal \boldsymbol{\mathcal{A}}^\intercal\right]\boldsymbol{\mathcal{A}}\mathbf{U}_1\right|$$

$$= \left|\lambda \mathbf{D}_1^{-1} - \frac{\lambda}{n}\mathrm{tr}\left\{\left(\lambda \mathbf{I}_n - \frac{1}{n}\boldsymbol{\mathcal{A}}\boldsymbol{\Gamma}\boldsymbol{\mathcal{A}}^\intercal\right)^{-1}\right\}\mathbf{I}_K + n^{-1/2}\boldsymbol{\Omega}_K(\lambda, \boldsymbol{\mathcal{A}})\right|, \tag{17}$$

where $\boldsymbol{\Gamma} = \mathbf{U}_2\mathbf{D}_2\mathbf{U}_2^\intercal$ and

$$\boldsymbol{\Omega}_K(\lambda, \boldsymbol{\mathcal{A}}) = \frac{\lambda}{\sqrt{n}}\left[\mathrm{tr}\left\{\left(\lambda \mathbf{I}_n - \frac{1}{n}\boldsymbol{\mathcal{A}}\boldsymbol{\Gamma}\boldsymbol{\mathcal{A}}^\intercal\right)^{-1}\right\}\mathbf{I}_K - \mathbf{U}_1^\intercal \boldsymbol{\mathcal{A}}^\intercal\left(\lambda \mathbf{I}_n - \frac{1}{n}\boldsymbol{\mathcal{A}}\boldsymbol{\Gamma}\boldsymbol{\mathcal{A}}^\intercal\right)^{-1}\boldsymbol{\mathcal{A}}\mathbf{U}_1\right]. \tag{18}$$

From Lemma 6.2 and Remark 6.3, if $\widetilde{\mathbf{A}}_i$ satisfies Assumptions (B1) – (B4), we can replace entries of $\boldsymbol{\mathcal{A}}$ by the standard Gaussian entries without changing the phase-transition theory of the spiked eigenvalues. Then, our Theorem 2.5 follows from Lemma 3.1 and Theorems 4.1 – 4.2 in Bai and Yao (2012).

It remains to prove that random vector $\widetilde{\mathbf{A}}_i$ satisfies Assumptions (B1) – (B4), which shows that our Lemma 6.2 applies. It is obvious that $\widetilde{\mathbf{A}}_i$ satisfies Assumption (B1). By using Lemma 2.3, we conclude that $\liminf_{p\to\infty}\lambda_p(\boldsymbol{\Sigma}_\rho^{1/2}) > 0$, and thus $\boldsymbol{\Sigma}_\rho^{-1/2}$ is bounded in spectral norm. Combining this information and the fact that each component of the random vector $\mathbf{A}_i$ follows $\mathsf{Uniform}(-1, 1)$ distribution, we conclude that each component of the random vector $\widetilde{\mathbf{A}}_i$ has bounded fourth moment. This, together with Lemma 2.3 and similar calculations in Section S2.3.1 of the supplementary material, implies that $\widetilde{\mathbf{A}}_i$ satisfies the moment condition (B2). From (S2.16) in the supplementary material and the fact that $\|\boldsymbol{\Sigma}_\rho^{-1/2}\|_2 = O(1)$, we have, for any $p \times p$ symmetry matrix $\mathbf{T}$ with bounded spectral norm, $\mathrm{Var}(\widetilde{\mathbf{A}}_i^\intercal \mathbf{T}\widetilde{\mathbf{A}}_i) = \mathrm{Var}(3\mathbf{A}_i^\intercal \boldsymbol{\Sigma}_\rho^{-1/2}\mathbf{T}\boldsymbol{\Sigma}_\rho^{-1/2}\mathbf{A}_i) = o(p^2)$. Hence, $\widetilde{\mathbf{A}}_i$ satisfies Assumption (B3). By Assumptions (A3) and (A4), Lemma 2.3, and the fact that each component of $\mathbf{A}_i$ follows $\mathsf{Uniform}(-1, 1)$, we conclude that each component of $\widetilde{\mathbf{A}}_i$ is bounded, and thus satisfies the concentration assumption (B4), according to Lemma S1.3 in the supplementary material. Therefore, the random vector $\widetilde{\mathbf{A}}_i$ satisfies Assumptions (B1) – (B4), which shows that our Lemma 6.2 applies. This completes the proof of Theorem 2.5.

6.3 Proof of Theorem 2.8

From Theorem 2.5, we have, for any $j \in [K_0]$, $j < \ell \leqslant p$, $\frac{1}{\{\lambda_j(\boldsymbol{\rho}_n)-\lambda_\ell(\boldsymbol{\rho}_n)\}^2} \asymp 1$, here $x \asymp 1$ means that there exist constants a and b such that $a < x < b$ a.s.. Thus,

$$\widehat{m}'_{n,j}\big(\lambda_j(\boldsymbol{\rho}_n)\big) = \frac{1}{p-j}\sum_{\ell=j+1}^{p}\frac{1}{\big\{\lambda_j(\boldsymbol{\rho}_n)-\lambda_\ell(\boldsymbol{\rho}_n)\big\}^2} \asymp 1,$$

and

$$\frac{\widehat{m}'_{n,j+1}\big(\lambda_{j+1}(\boldsymbol{\rho}_n)\big)}{\widehat{m}'_{n,j}\big(\lambda_j(\boldsymbol{\rho}_n)\big)} = O_P(1), \qquad \text{for } j \in [K_0-1]. \tag{19}$$

From the fluctuation of edge eigenvalues (Péché, 2009; Bao, 2019), we have $\lambda_j(\boldsymbol{\rho}_n) - \lambda_{j+1}(\boldsymbol{\rho}_n) = O_P(p^{-2/3})$ for $K_0+1 \leqslant j \leqslant K_{\max}$. Thus, for $K_0+1 \leqslant j \leqslant K_{\max}$, we get

$$\widehat{m}'_{n,j}\big(\lambda_j(\boldsymbol{\rho}_n)\big) = \frac{1}{p-j}\sum_{\ell=j+1}^{p}\frac{1}{\big\{\lambda_j(\boldsymbol{\rho}_n)-\lambda_\ell(\boldsymbol{\rho}_n)\big\}^2} = O_P(p^{1/3}).$$

Therefore,

$$\frac{\widehat{m}'_{n,K_0+1}\big(\lambda_{K_0+1}(\boldsymbol{\rho}_n)\big)}{\widehat{m}'_{n,K_0}\big(\lambda_{K_0}(\boldsymbol{\rho}_n)\big)} \to \infty. \tag{20}$$

Moreover, for $K_0+1 \leqslant j \leqslant K_{\max}$, we have

$$\frac{\widehat{m}'_{n,j+1}\big(\lambda_{j+1}(\boldsymbol{\rho}_n)\big)}{\widehat{m}'_{n,j}\big(\lambda_j(\boldsymbol{\rho}_n)\big)} = \frac{p-j}{p-j-1}\,\frac{\sum_{\ell=j+2}^{p}\big\{\lambda_{j+1}(\boldsymbol{\rho}_n)-\lambda_\ell(\boldsymbol{\rho}_n)\big\}^{-2}}{\sum_{\ell=j+1}^{p}\big\{\lambda_j(\boldsymbol{\rho}_n)-\lambda_\ell(\boldsymbol{\rho}_n)\big\}^{-2}} \asymp 1. \tag{21}$$

By combining (19) – (21), we obtain that $\widehat{K}_{\mathsf{SR}}$ is consistent. Moreover, if $\widehat{m}'_{n,j}(x)$ is replaced by $\widetilde{m}'_{n,j}(x)$, the derivations above remain valid, thereby ensuring the consistency of $\widetilde{K}_{\mathsf{SR}}$ in (14) as well.

SUPPLEMENTARY MATERIAL

This supplementary material contains some auxiliary lemmas and the technical proofs of Lemmas 2.2, 2.3, 6.1, 6.2, S1.6, and S1.7.

References

Ahn, S. C. and Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3):1203–1227.

Alessi, L., Barigozzi, M., and Capasso, M. (2010). Improved penalization for determining the number of factors in approximate factor models. *Statistics & Probability Letters*, 80(23-24):1806–1813.

Andrews, D. F. and Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(1):99–102.

Bai, J. and Li, K. (2012). Statistical analysis of factor models of high dimension. *The Annals of Statistics*, 40(1):436–465.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Bai, Z. and Yao, J. (2008). Central limit theorems for eigenvalues in a spiked population model. *Annales de l'IHP Probabilités et statistiques*, 44(3):447–474.

Bai, Z. and Yao, J. (2012). On sample eigenvalues in a generalized spiked population model. *Journal of Multivariate Analysis*, 106:167–177.

Bai, Z. and Zhou, W. (2008). Large sample covariance matrices without independence structures in columns. *Statistica Sinica*, 18(2):425–442.

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

Bao, Z. (2019). Tracy-Widom limit for Spearman's rho. Technical report, The Hong Kong University of Science and Technology.

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

Chen, W. and Lam, C. (2024). Rank and factor loadings estimation in time series tensor factor model by pre-averaging. *The Annals of Statistics*, 52(1).

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

Fan, J., Guo, J., and Zheng, S. (2020). Estimating number of factors by adjusted eigenvalues thresholding. *Journal of the American Statistical Association*, 117(538):852–861.

Fan, J., Liu, H., and Wang, W. (2018). Large covariance estimation through elliptical factor models. *The Annals of statistics*, 46(4):1383–1414.

Hallin, M. and Liška, R. (2007). Determining the number of factors in the general dynamic factor model. *Journal of the American Statistical Association*, 102(478):603–617.

Harding, M. (2013). Estimating the number of factors in large dimensional factor models. Technical report, University of California-Irvine.

He, Y., Kong, X., Yu, L., and Zhang, X. (2022a). Large-dimensional factor analysis without moment constraints. *Journal of Business & Economic Statistics*, 40(1):302–312.

He, Y., Wang, Y., Yu, L., Zhou, W., and Zhou, W.-X. (2022b). Matrix Kendall's tau in high-dimensions: A robust statistic for matrix factor model. *arXiv preprint arXiv:2207.09633*.

Heinen, A. and Valdesogo, A. (2020). Spearman rank correlation of the bivariate Student t and scale mixtures of normal distributions. *Journal of Multivariate Analysis*, 179.

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

Jiang, D. and Bai, Z. (2021). Generalized four moment theorem and an application to CLT for spiked eigenvalues of high-dimensional covariance matrices. *Bernoulli*, 27(1):274–294.

Kong, X.-B. (2017). On the number of common factors with high-frequency data. *Biometrika*, 104(2):397–410.

Lam, C. (2021). Rank determination for time series tensor factor model using correlation thresholding. Technical report, Working paper LSE.

Lam, C. and Yao, Q. (2012). Factor modeling for high-dimensional time series: Inference for the number of factors. *The Annals of Statistics*, 40(2).

Ledoux, M. (2001). *The concentration of measure phenomenon*, volume 89. American Mathematical Society.

Li, H., Li, Q., and Shi, Y. (2017a). Determining the number of factors when the number of factors can increase with sample size. *Journal of Econometrics*, 197(1):76–86.

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

Li, Z., Wang, Q., and Yao, J. (2017b). Identifying the number of factors from singular values of a large sample auto-covariance matrix. *The Annals of Statistics*, 45(1):257 – 288.

Liu, H., Lafferty, J., and Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10(80):2295–2328.

McCracken, M. W. and Ng, S. (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589.

Nadakuditi, R. and Edelman, A. (2008). Sample eigenvalue based detection of high-dimensional signals in white noise using relatively few samples. *IEEE Transactions on Signal Processing*, 56(7):2625 – 2638.

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *Review of Economics and Statistics*, 92(4):1004–1016.

Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics*, 168(2):244–258.

Owen, A. B. and Perry, P. O. (2009). Bi-cross-validation of the svd and the nonnegative matrix factorization. *The Annals of Applied Statistics*, 3(2).

Owen, A. B. and Wang, J. (2016). Bi-cross-validation for factor analysis. *Statistical Science*, 31(1).

Péché, S. (2009). Universality results for the largest eigenvalues of some sample covariance matrix ensembles. *Probability Theory and Related Fields*, 143(3–4):481–516.

Spearman, C. (1961). The proof and measurement of association between two things. *The American Journal of Psychology*, 15:72–101.

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

Yu, L., He, Y., and Zhang, X. (2019). Robust factor number specification for large-dimensional elliptical factor model. *Journal of Multivariate analysis*, 174.

Supplementary material for “Robust estimation for number of factors in high dimensional factor modeling via Spearman correlation matrix”

Jiaxin Qiu
Department of Statistics and Actuarial Science
The University of Hong Kong, Hong Kong, China
qiujx@connect.hku.hk
and
Zeng Li
Department of Statistics and Data Science
Southern University of Science and Technology, Shenzhen, China
liz9@sustech.edu.cn
and
Jianfeng Yao
School of Data Science
The Chinese University of Hong Kong (Shenzhen), Shenzhen, China
jeffyao@cuhk.edu.cn

September 7, 2024

Contents

This supplementary document contains some auxiliary lemmas and the technical proofs of Lemmas 2.2, 2.3, 6.1, 6.2, S1.6 and S1.7.

S1 Auxiliary lemmas

This section introduces several auxiliary lemmas used in the technical proofs of our theoretical results. Lemmas S1.1 – S1.5 are from existing literature, while Lemmas S1.6 and S1.7 are our original contributions. We provide the proofs of these two new lemmas in Sections S2.5 and S2.6, respectively.

Lemma S1.1 (Weyl's inequality, Corollary 6.3.4 in Horn and Johnson (2012))**.** *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be two* $n \times n$ *normal matrices, and let* $\lambda_1(\mathbf{A}) \geqslant \cdots \geqslant \lambda_n(\mathbf{A})$ *and* $\lambda_1(\mathbf{B}) \geqslant \cdots \geqslant \lambda_n(\mathbf{B})$ *be the nonincreasingly ordered eigenvalues of* $\mathbf{A}$ *and* $\mathbf{B}$*, respectively. Then*

$$\max_{1 \leqslant i \leqslant n} |\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B})| \leqslant \|\mathbf{A} - \mathbf{B}\|_2,$$

where $\|\mathbf{A} - \mathbf{B}\|_2$ *denotes the spectral norm of* $\mathbf{A} - \mathbf{B}$*.*

Lemma S1.2 (Lemma 7 in El Karoui (2009))**.** *Suppose that the random vector* $\mathbf{r} \in \mathbb{R}^p$ *has the property that for any convex* 1*-Lipschitz (with respect to the Euclidean norm) function* F *from* $\mathbb{R}^p$ *to* $\mathbb{R}$*, we have*

$$\mathbb{P}\big(|F(\mathbf{r}) - m_F| > t\big) \leqslant C \exp\big\{-c(p)t^2\big\},$$

where m_F denotes a median of F, and C and $c(p)$ are independent of F, and C is independent of p. We allow $c(p)$ to be a constant or to go to zero with p like $p^{-\alpha}$, $0 \leqslant \alpha < 1$. Suppose, further, that $\mathbb{E}(\mathbf{r}) = \mathbf{0}$, $\mathbb{E}(\mathbf{r}\mathbf{r}^\intercal) = \mathbf{\Sigma}$, with $\|\mathbf{\Sigma}\|_2 \leqslant \log(p)$. If $\mathbf{M}$ is a complex deterministic matrix such that $\|\mathbf{M}\|_2 \leqslant \xi$, where ξ is independent of p, then $p^{-1}\mathbf{r}^\intercal\mathbf{M}\mathbf{r}$ is strongly concentrated around its mean, $p^{-1}\mathrm{tr}(\mathbf{M}\mathbf{\Sigma})$. In particular, if, for $\varepsilon > 0$, $t_p(\varepsilon) = (\log p)^{1+\varepsilon}/\sqrt{pc(p)}$, then

$$\log\left\{\mathbb{P}\left(\left|\frac{1}{p}\mathbf{r}^\intercal\mathbf{M}\mathbf{r} - \frac{1}{p}\mathrm{tr}(\mathbf{M}\mathbf{\Sigma})\right| > t_p(\varepsilon)\right)\right\} \asymp -(\log p)^{1+2\varepsilon}.$$

Lemma S1.3 (Corollary 4.10 in Ledoux (2001))**.** *For every product probability $\mathbb{P}$ on $[0,1]^n$, every convex 1-Lipschitz function F on $\mathbb{R}^n$, and every $r \geqslant 0$, we have*

$$\mathbb{P}\big(|F - m_F| \geqslant r\big) \leqslant 4e^{-r^2/4},$$

where m_F is a median of F for $\mathbb{P}$.

Lemma S1.4 ((3.3.41) in Horn and Johnson (1991))**.** *For any $n \times n$ Hermitian $\mathbf{A} = (A_{ij})$ with eigenvalues $\lambda_1, \ldots, \lambda_n$, and convex f, we have*

$$\sum_{i=1}^{n} f(A_{ii}) \leqslant \sum_{i=1}^{n} f(\lambda_i).$$

Lemma S1.5 (Grothendieck's identity, Lemma 3.6.6 in Vershynin (2018), Lemma 4.1 in Li et al. (2023))**.** *Consider a bivariate normal vector:*

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \gamma \\ \gamma & \sigma_2^2 \end{pmatrix}\right),$$

we have

$$\mathbb{E}[\mathrm{sign}(Z_1)\mathrm{sign}(Z_2)] = \frac{2}{\pi}\arcsin\Big(\frac{\gamma}{\sigma_1\sigma_2}\Big).$$

Lemma S1.6 is utilized to analyze the structure of $\mathbf{\Sigma}_\rho$, as stated in Lemma 2.3.

Lemma S1.6. *Consider a bivariate random vector:*

$$\mathbf{Q} := \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix} \overset{\mathsf{d}}{=} \begin{pmatrix} \sqrt{W_X}X_1 + \sqrt{W_{Y_1}}Y_1 \\ \sqrt{W_X}X_2 + \sqrt{W_{Y_2}}Y_2 \end{pmatrix}, \tag{S1.1}$$

where W_X*,* W_{Y_1}*, and* W_{Y_2} *are three independent scalar-valued random variables with positive support, and* $(X_1, X_2, Y_1, Y_2)^\intercal \sim \mathcal{N}_4(\mathbf{0}, \mathbf{\Omega}_0)$ *with*

$$\mathbf{\Omega}_0 = \begin{pmatrix} \mathbf{\Sigma}_X & O \\ O & \mathbf{\Sigma}_Y \end{pmatrix}, \qquad \mathbf{\Sigma}_X = \begin{pmatrix} \sigma_{X_1}^2 & \gamma \\ \gamma & \sigma_{X_2}^2 \end{pmatrix}, \qquad \mathbf{\Sigma}_Y = \begin{pmatrix} \sigma_{Y_1}^2 & 0 \\ 0 & \sigma_{Y_2}^2 \end{pmatrix}.$$

Suppose that $\widetilde{Q}_1$ *and* $\widetilde{Q}_2$ *(they are independent) are independent copies of* Q_1 *and* Q_2*, respectively, then, we have*

$$\mathbb{E}[\mathrm{sign}(Q_1 - \widetilde{Q}_1)\,\mathrm{sign}(Q_2 - \widetilde{Q}_2)] = \frac{2}{\pi}\mathbb{E}\big\{\arcsin(r)\big\},$$

where

$$r = \frac{\gamma W_X}{\sqrt{(W_X + \widetilde{W}_{X_1})\sigma_{X_1}^2 + (W_{Y_1} + \widetilde{W}_{Y_1})\sigma_{Y_1}^2}\sqrt{(W_X + \widetilde{W}_{X_2})\sigma_{X_2}^2 + (W_{Y_2} + \widetilde{W}_{Y_2})\sigma_{Y_2}^2}},$$

and $(\widetilde{W}_{X_1}, \widetilde{W}_{Y_1})$ *and* $(\widetilde{W}_{X_2}, \widetilde{W}_{Y_2})$ *are independent copies of* (W_X, W_{Y_1}) *and* (W_X, W_{Y_2})*, respectively.*

Lemma S1.7 is used in the proof of Lemma 6.2.

Lemma S1.7. *Let* $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\intercal$ *and* $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^\intercal$ *be two* $n \times p$ *independent random matrices satisfying the same assumptions as those in Lemma 6.2. For any* $i \in [n]$, *we denote*

$$
\begin{aligned}
\mathbf{X}_{i0} &= (\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{y}_{i+1}, \ldots, \mathbf{y}_n)^\intercal, \\
\beta_{i0} &= 1 - n^{-1}\mathrm{tr}\big\{\boldsymbol{\Gamma}(\lambda\mathbf{I} - n^{-1}\boldsymbol{\Gamma}^{1/2}\mathbf{X}_{i0}^\intercal\mathbf{X}_{i0}\boldsymbol{\Gamma}^{1/2})^{-1}\big\}, \\
\beta_i &= 1 - n^{-1}\mathbf{x}_i^\intercal\boldsymbol{\Gamma}^{1/2}(\lambda\mathbf{I} - n^{-1}\boldsymbol{\Gamma}^{1/2}\mathbf{X}_{i0}^\intercal\mathbf{X}_{i0}\boldsymbol{\Gamma}^{1/2})^{-1}\boldsymbol{\Gamma}^{1/2}\mathbf{x}_i, \\
\mathbf{H}_1 &= n^{-1}\mathbf{X}_{i0}^\intercal(\lambda\mathbf{I}_{n-1} - n^{-1}\mathbf{X}_{i0}\boldsymbol{\Gamma}\mathbf{X}_{i0}^\intercal)^{-1}\mathbf{X}_{i0}\boldsymbol{\Gamma}, \\
\mathbf{H}_2 &= n^{-1}\boldsymbol{\Gamma}\mathbf{X}_{i0}^\intercal(\lambda\mathbf{I}_{n-1} - n^{-1}\mathbf{X}_{i0}\boldsymbol{\Gamma}\mathbf{X}_{i0}^\intercal)^{-2}\mathbf{X}_{i0}\boldsymbol{\Gamma}, \\
\boldsymbol{\tau}_{i0} &= (1 + n^{-1}\mathrm{tr}\mathbf{H}_2)\mathbf{I}_K - \mathbf{U}_1^\intercal(\mathbf{I}_p + \mathbf{H}_1)(\mathbf{I}_p + \mathbf{H}_1)^\intercal\mathbf{U}_1, \\
\boldsymbol{\tau}_i &= n^{-1}(\mathbf{x}_i^\intercal\mathbf{H}_2\mathbf{x}_i - \mathrm{tr}\mathbf{H}_2)\mathbf{I}_K - \mathbf{U}_1^\intercal(\mathbf{I}_p + \mathbf{H}_1)(\mathbf{x}_i\mathbf{x}_i^\intercal - \mathbf{I}_p)(\mathbf{I}_p + \mathbf{H}_1)^\intercal\mathbf{U}_1,
\end{aligned}
$$

with $\boldsymbol{\Gamma} = \mathbf{U}_2\mathbf{D}_2\mathbf{U}_2^\intercal$. *Then, for any* $K \times K$ *symmetric matrix* $\mathbf{W}$, *it holds that, as* $n \to \infty$,

$$\beta_{i0} \to -\frac{1}{\lambda\underline{m}(\lambda)}, \qquad \varepsilon_i := \beta_i - \beta_{i0} \to 0, \tag{S1.2}$$

$$\mathbb{E}\,\varepsilon_i^2 = o(1), \qquad \mathbb{E}\,\varepsilon_i^4 = o(n^{-1}), \tag{S1.3}$$

$$\mathbb{E}\{\mathrm{tr}(\mathbf{W}\boldsymbol{\tau}_i)\}^2 = O(1), \qquad \mathbb{E}\{\mathrm{tr}(\mathbf{W}\boldsymbol{\tau}_{i0})\}^2 = O(1), \tag{S1.4}$$

where $\underline{m}(\cdot)$ *denotes the Stieltjes transform of the LSD of the matrix* $n^{-1}\mathbf{X}_{i0}\boldsymbol{\Gamma}\mathbf{X}_{i0}^\intercal$.

Remark S1.8. *The conclusions* (S1.2) – (S1.4) *presented in Lemma S1.7 are identical to those in Lemma C.3 and Lemma D.1 of Jiang and Bai (2021). However, the proof provided by Jiang and Bai (2021) is not applicable for our scenario as we do not assume that* $\mathbf{X}$ *(or* $\mathbf{Y}$*) possesses i.i.d. entries, as stated in Assumption (B1). Specifically, we need to establish the concentration properties of certain quadratic forms without the i.i.d. assumption. We* *address this challenge in Section S2.6 by leveraging concentration inequalities from Ledoux (2001) and El Karoui (2009). See the proof in Section S2.6 for more details.*

S2 Proofs of lemmas

S2.1 Proof of Lemma 2.2

Let $\mathbf{\Delta}_n$ denote the difference between $\boldsymbol{\rho}_n$ and $\mathbf{W}_n$, that is,

$$\mathbf{\Delta}_n := \boldsymbol{\rho}_n - \mathbf{W}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{R}_i\mathbf{R}_i^\intercal - \frac{3}{n}\sum_{i=1}^{n}\mathbf{A}_i\mathbf{A}_i^\intercal, \tag{S2.1}$$

where $\mathbf{R}_i^\intercal$ is the i-th row of $\mathbf{R} = \sqrt{\frac{12}{n^2-1}\big(r_{ij} - \frac{n+1}{2}\big)_{n\times p}}$, and $\mathbf{A}_i = \mathbb{E}\{\text{sign}(\mathbf{y}_i - \mathbf{y}_\ell) \mid \mathbf{y}_i\}$ with $i \neq \ell$. From Lemma S1.1, it is sufficient to show that $\|\sqrt{n}\mathbf{\Delta}_n\|_2 = O_P(1)$. To this end, we show that $\sqrt{n}\mathbf{\Delta}_n$ is a Wigner-type matrix by the following moment estimations:

$$\mathbb{E}\{(\sqrt{n}\mathbf{\Delta}_n)_{j\ell}\} = O(n^{-1/2}), \quad \text{Var}\{(\sqrt{n}\mathbf{\Delta}_n)_{j\ell}\} = O(n^{-1}), \quad j \neq \ell, \tag{S2.2}$$

$$\text{Corr}\big((\sqrt{n}\mathbf{\Delta}_n)_{j\ell}, (\sqrt{n}\mathbf{\Delta}_n)_{j\ell'}\big) = O(n^{-1/2}), \qquad j \neq \ell \neq \ell', \tag{S2.3}$$

$$\text{Corr}\big((\sqrt{n}\mathbf{\Delta}_n)_{jj}, (\sqrt{n}\mathbf{\Delta}_n)_{\ell r}\big) = O(n^{-1/2}), \qquad \ell \neq r. \tag{S2.4}$$

Let y_{ij} be the j-th component of $\mathbf{y}_i$. For each $j \in [p]$, we define the *empirical cumulative distribution function* (ECDF) as

$$\widehat{F}_j(y) = \frac{1}{n}\sum_{\ell=1}^{n}\mathbb{1}\{y_{\ell j} \leqslant y\}, \tag{S2.5}$$

and let $r_{ij} = n\widehat{F}_j(y_{ij})$ be the rank of y_{ij} among $\{y_{\ell j}\}_{\ell=1}^{n}$, where and $i \in [n]$ and $j \in [p]$. For notational simplicity, we denote $F_{ij} = F_j(y_{ij})$ and $\widehat{F}_{ij} = \widehat{F}_j(y_{ij})$, where $F_j(\cdot)$ is the CDF of

y_{1j}. Using these notations, we write the (i,j)-th entry of the matrix $\mathbf{R}$ as

$$R_{ij} = \sqrt{\frac{12}{n^2-1}}\Big(r_{ij} - \frac{n+1}{2}\Big) = \sqrt{\frac{12n^2}{n^2-1}}\Big(\widehat{F}_{ij} - \frac{n+1}{2n}\Big), \tag{S2.6}$$

and write the j-th entry of the vector $\mathbf{A}_i$ as

$$A_{ij} = \mathbb{E}\big\{\operatorname{sign}\big(y_{ij} - y_{\ell j}\big) \mid y_{ij}\big\} = 2F_{ij} - 1. \tag{S2.7}$$

First, we prove (S2.2). From (S2.1) and (S2.6) – (S2.7), the (j,ℓ)-th entry of $\mathbf{\Delta}_n$, where $j \neq \ell$, can be written as

$$\begin{aligned}
&(\mathbf{\Delta}_n)_{j\ell} \\
&= \frac{1}{n}\left\{\sum_{i=1}^{n}\Big(\frac{12n^2}{n^2-1}\Big)\Big(\widehat{F}_{ij} - \frac{n+1}{2n}\Big)\Big(\widehat{F}_{i\ell} - \frac{n+1}{2n}\Big) - \sum_{i=1}^{n} 12\Big(F_{ij} - \frac{1}{2}\Big)\Big(F_{i\ell} - \frac{1}{2}\Big)\right\} \\
&= \frac{6}{n}\sum_{i=1}^{n}\left\{\Big(\frac{2n^2}{n^2-1}\widehat{F}_{ij}\widehat{F}_{i\ell} - 2F_{ij}F_{i\ell}\Big) - \Big(\frac{n}{n-1}\widehat{F}_{ij} - F_{ij}\Big) - \Big(\frac{n}{n-1}\widehat{F}_{i\ell} - F_{i\ell}\Big) + \frac{1}{n-1}\right\} \\
&=: \frac{6}{n}\sum_{i=1}^{n}(\mathbf{\Delta}_n)_{j\ell}^{(i)}.
\end{aligned}$$

From the fact that $\{F_{ij}\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} \mathsf{Uniform}(0,1)$ for fixed j and the definition (S2.5) of ECDF, we obtain

$$\mathbb{E}F_{ij} = \frac{1}{2}, \qquad \mathbb{E}F_{ij}^2 = \frac{1}{3}, \qquad \mathbb{E}\widehat{F}_{ij} = \frac{1}{n}\sum_{\ell=1}^{n}\mathbb{E}\mathbb{1}\{y_{\ell j} \leqslant y_{ij}\} = \frac{n+1}{2n},$$

and

$$\mathbb{E}\widehat{F}_{ij}\widehat{F}_{i\ell} = \frac{(n-1)(n-2)}{n^2}\mathbb{E}F_{ij}F_{i\ell} + \frac{n-1}{n^2}\mathbb{E}\mathbb{1}\{y_{sj} \leqslant y_{ij}\}\mathbb{1}\{y_{s\ell} \leqslant y_{i\ell}\} + \frac{1}{n}, \quad (s \neq i).$$

These identities imply that

$$\mathbb{E}\big\{(\boldsymbol{\Delta}_n)_{j\ell}^{(1)}\big\} = -\frac{6}{n+1}\mathbb{E}F_{1j}F_{1\ell} + \frac{2}{n+1}\mathbb{E}\mathbb{1}\{y_{2j} \leqslant y_{1j}\}\mathbb{1}\{y_{2\ell} \leqslant y_{1\ell}\} + \frac{1}{n+1} = O(n^{-1}).$$

As $\{(\boldsymbol{\Delta}_n)_{j\ell}^{(i)}\}_{i=1}^n$ are i.i.d. for fixed j and ℓ, we have

$$\mathbb{E}\big\{(\boldsymbol{\Delta}_n)_{j\ell}\big\} = 6\mathbb{E}\big\{(\boldsymbol{\Delta}_n)_{j\ell}^{(1)}\big\} = O(n^{-1}). \tag{S2.8}$$

Moreover, a similar calculation gives

$$\begin{aligned}
&\operatorname{Var}\big\{(\boldsymbol{\Delta}_n)_{j\ell}^{(1)}\big\} \\
&= \frac{n^3}{(n^2-1)^2}\mathbb{E}\bigg[\frac{1}{3} + 4F_{1j}F_{1\ell}\big(3F_{1j} + 3F_{1\ell} - 4F_{1j}F_{1\ell}\big) - 10F_{1j}F_{1\ell} \\
&\qquad\qquad - 4\Big(F_{1j} + F_{1\ell} - 2F_{1j}F_{1\ell} - \frac{1}{2}\Big)\mathbb{1}\{y_{2j} \leqslant y_{1j}\}\mathbb{1}\{y_{2\ell} \leqslant y_{1\ell}\}\bigg] + O(n^{-2}) \\
&= O(n^{-1}).
\end{aligned}$$

Thus, we have

$$\operatorname{Var}\big\{(\boldsymbol{\Delta}_n)_{j\ell}\big\} = \frac{36}{n}\operatorname{Var}\big\{(\boldsymbol{\Delta}_n)_{j\ell}^{(1)}\big\} = O(n^{-2}). \tag{S2.9}$$

This, together with (S2.8), implies (S2.2).

Now, we prove (S2.3). For any $j, \ell, \ell' \in [p]$ with $j \neq \ell \neq \ell'$, direct calculation gives us

$$\begin{aligned}
&\operatorname{Cov}\big((\boldsymbol{\Delta}_n)_{j\ell}, (\boldsymbol{\Delta}_n)_{j\ell'}\big) \\
&= \frac{36n^2}{(n^2-1)^2}\mathbb{E}\Big\{-16F_{1j}^2F_{1\ell}F_{1\ell'} + 8F_{1j}^2F_{1\ell} + 4F_{1j}^2\mathbb{1}\{y_{2\ell} \leqslant y_{1\ell}\}\mathbb{1}\{y_{2\ell'} \leqslant y_{1\ell'}\} \\
&\qquad + 12F_{1j}F_{1\ell}F_{1\ell'} + 4F_{1j}\mathbb{1}\{y_{2j} \leqslant y_{1j}\}\big[F_{1\ell}\mathbb{1}\{y_{2\ell'} \leqslant y_{1\ell'}\} + F_{1\ell'}\mathbb{1}\{y_{2\ell} \leqslant y_{1\ell}\}\big] \\
&\qquad - 3F_{1j}\big(F_{1\ell} + F_{1\ell'}\big) - 2F_{1j}\mathbb{1}\{y_{2j} \leqslant y_{1j}\}\big[\mathbb{1}\{y_{2\ell} \leqslant y_{1\ell}\} + \mathbb{1}\{y_{2\ell'} \leqslant y_{1\ell'}\}\big]
\end{aligned}$$

$$
\begin{aligned}
&- 4F_{1j}\mathbb{1}\{y_{2\ell} \leqslant y_{1\ell}\}\mathbb{1}\{y_{2\ell'} \leqslant y_{1\ell'}\} - F_{1\ell}F_{1\ell'}\\
&- 2\mathbb{1}\{y_{2j} \leqslant y_{1j}\}\big[F_{1\ell}\mathbb{1}\{y_{2\ell'} \leqslant y_{1\ell'}\} + F_{1\ell'}\mathbb{1}\{y_{2\ell} \leqslant y_{1\ell}\}\big]\\
&+ \mathbb{1}\{y_{2j} \leqslant y_{1j}\}\mathbb{1}\{y_{2\ell} \leqslant y_{1\ell}\} + \mathbb{1}\{y_{2j} \leqslant y_{1j}\}\mathbb{1}\{y_{2\ell'} \leqslant y_{1\ell'}\}\\
&+ \mathbb{1}\{y_{2\ell} \leqslant y_{1\ell}\}\mathbb{1}\{y_{2\ell'} \leqslant y_{1\ell'}\} + \frac{1}{6}\bigg\} + O(n^{-3}).
\end{aligned} \tag{S2.10}
$$

From Assumptions (A1), (A3), and (A4), the constant term in the curly brackets of (S2.10) is of order $O(n^{-1/2})$, then we have $\mathrm{Cov}\big((\mathbf{\Delta}_n)_{j\ell}, (\mathbf{\Delta}_n)_{j\ell'}\big) = O(n^{-5/2})$. This, together with (S2.9), implies that

$$
\mathrm{Corr}\big((\mathbf{\Delta}_n)_{j\ell}, (\mathbf{\Delta}_n)_{j\ell'}\big) = \frac{\mathrm{Cov}\big((\mathbf{\Delta}_n)_{j\ell}, (\mathbf{\Delta}_n)_{j\ell'}\big)}{\sqrt{\mathrm{Var}\big\{(\mathbf{\Delta}_n)_{j\ell}\big\} \cdot \mathrm{Var}\big\{(\mathbf{\Delta}_n)_{j\ell'}\big\}}} = O(n^{-1/2}),
$$

and thus (S2.3) holds true.

Finally, we prove (S2.4). From (S2.1) and (S2.6) – (S2.7), the j-th diagonal entries of $\mathbf{\Delta}_n$ can be written as $(\mathbf{\Delta}_n)_{jj} = 1 - (12/n)\sum_{i=1}^n (F_{ij} - 1/2)^2$. By using the fact that $\{F_{ij}\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathsf{Uniform}(0,1)$ for fixed j, we obtain

$$
\mathrm{Var}\{(\mathbf{\Delta}_n)_{jj}\} = \frac{4}{5n}. \tag{S2.11}
$$

For any $j, \ell, r \in [p]$ with $\ell \neq r$, by using similar argument above, we have

$$
\begin{aligned}
&\mathrm{Cov}\big((\mathbf{\Delta}_n)_{jj}, (\mathbf{\Delta}_n)_{\ell r}\big)\\
&= \frac{72}{n}\frac{1}{n+1}\bigg\{\Big(2\mathbb{E}\big[F_{1j}\mathbb{1}\{y_{2\ell} \leqslant y_{1\ell}\}\mathbb{1}\{y_{2r} \leqslant y_{1r}\}\big] - 6\mathbb{E}F_{1j}F_{1\ell}F_{1r} + 2\mathbb{E}(F_{1j}F_{1\ell} + F_{1j}F_{1r}) - \frac{1}{2}\Big)\\
&\qquad - \Big(2\mathbb{E}\big[F_{1j}^2\mathbb{1}\{y_{2\ell} \leqslant y_{1\ell}\}\mathbb{1}\{y_{2r} \leqslant y_{1r}\}\big] - 6\mathbb{E}F_{1j}^2F_{1\ell}F_{1r} + 2\mathbb{E}(F_{1j}^2F_{1\ell} + F_{1j}^2F_{1r}) - \frac{1}{3}\Big)\\
&\qquad - \frac{1}{6}\Big[-6\mathbb{E}F_{1\ell}F_{1r} + 2\mathbb{E}\mathbb{1}\{y_{2\ell} \leqslant y_{1\ell}\}\mathbb{1}\{y_{2r} \leqslant y_{1r}\} + 1\Big]\bigg\}
\end{aligned}
$$

$$= O(n^{-5/2}).$$

This, together with (S2.9) and (S2.11), implies (S2.4) and completes the proof.

S2.2 Proof of Lemma 2.3

Recall that $\mathbf{A}_{i\ell} = \operatorname{sign}(\mathbf{y}_i - \mathbf{y}_\ell)$ and $\mathbf{A}_i = \mathbb{E}(\mathbf{A}_{i\ell} \mid \mathbf{y}_i)$ for any $i, \ell \in [n]$ with $i \neq \ell$. By using the law of iterated expectations repeatedly, we obtain

$$\begin{aligned}
\boldsymbol{\Sigma}_\rho = 3\mathbb{E}(\mathbf{A}_1\mathbf{A}_1^\intercal) &= 3\mathbb{E}\big\{\mathbb{E}(\mathbf{A}_{12} \mid \mathbf{y}_1)\mathbf{A}_1^\intercal\big\} = 3\mathbb{E}(\mathbf{A}_{12}\mathbf{A}_1^\intercal) \\
&= 3\mathbb{E}\big\{\mathbf{A}_{12}\mathbb{E}(\mathbf{A}_{13}^\intercal \mid \mathbf{y}_1, \mathbf{y}_2)\big\} = 3\mathbb{E}\big\{\mathbb{E}(\mathbf{A}_{12}\mathbf{A}_{13}^\intercal \mid \mathbf{y}_1, \mathbf{y}_2)\big\} \\
&= 3\mathbb{E}(\mathbf{A}_{12}\mathbf{A}_{13}^\intercal) = 3\mathbb{E}\big\{\operatorname{sign}(\mathbf{y}_1 - \mathbf{y}_2)\operatorname{sign}(\mathbf{y}_1 - \mathbf{y}_3)^\intercal\big\}.
\end{aligned}$$

Let $\mathbf{b}_i^\intercal$ denote the i-th row of $\mathbf{B}$, and Ψ_{ii} denote the i-th diagonal element of $\boldsymbol{\Psi}$. From Assumption (A2) and Definition of SMN, the i-th component of $\mathbf{y}_1 = \mathbf{B}\mathbf{f}_1 + \boldsymbol{\Psi}\mathbf{e}_1$ has a stochastic representation (below, we omit "1" in subscripts to simplify notations)

$$y_i \overset{\mathsf{d}}{=} \sqrt{w^f}\mathbf{b}_i^\intercal\mathbf{x} + \sqrt{w_i^e}\Psi_{ii}z_i, \quad i \in [p], \tag{S2.12}$$

where $\mathbf{x} \sim \mathcal{N}_K(\mathbf{0}, \mathbf{I}_K)$, $\{z_i\}_{i=1}^p \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, and they are independent. For any $i \in [p]$, let

$$\tilde{\mathbf{v}}_i := \frac{\sqrt{w^f}\mathbf{b}_i}{\sqrt{(w^f + \widetilde{w}^f)\mathbf{b}_i^\intercal\mathbf{b}_i + (w_i^e + \widetilde{w}_i^e)\Psi_{ii}^2}}, \qquad \hat{\mathbf{v}}_i := \frac{\sqrt{w^f}\mathbf{b}_i}{\sqrt{(w^f + \widehat{w}^f)\mathbf{b}_i^\intercal\mathbf{b}_i + (w_i^e + \widehat{w}_i^e)\Psi_{ii}^2}},$$

where $(\widetilde{w}^f, \widetilde{w}_i^e)$ and $(\widehat{w}^f, \widehat{w}_i^e)$ are two independent copies of (w^f, w_i^e). It follows from Lemma S1.6 that, the (i,j)-th element of $\mathbf{\Sigma}_\rho$ can be written as

$$(\mathbf{\Sigma}_\rho)_{ij} = \begin{cases} (6/\pi)\mathbb{E}\{\arcsin(\tilde{\mathbf{v}}_i^\intercal \hat{\mathbf{v}}_j)\}, & i \neq j, \\ 1, & i = j. \end{cases}$$

By using Taylor's expansion, we derive that $(\mathbf{\Sigma}_\rho)_{ij} = \frac{6}{\pi}\mathbb{E}(\tilde{\mathbf{v}}_i^\intercal \hat{\mathbf{v}}_j) + O(p^{-2})$ for $i \neq j$. By definitions of $\tilde{\mathbf{v}}_i$ and $\hat{\mathbf{v}}_j$ and Assumption (A3), we have

$$\begin{aligned} &\mathbb{E}(\tilde{\mathbf{v}}_i^\intercal \hat{\mathbf{v}}_j) \\ &= \mathbb{E}\left\{ \frac{\sqrt{w^f}}{\sqrt{(w^f + \widetilde{w}^f)\mathbf{b}_i^\intercal \mathbf{b}_i + (w_i^e + \widetilde{w}_i^e)\Psi_{ii}^2}} \cdot \frac{\sqrt{w^f}}{\sqrt{(w^f + \widehat{w}^f)\mathbf{b}_j^\intercal \mathbf{b}_j + (w_j^e + \widehat{w}_j^e)\Psi_{jj}^2}} \right\} \mathbf{b}_i^\intercal \mathbf{b}_j \\ &= \mathbb{E}\left\{ \left(\frac{\sqrt{w^f}}{\sqrt{(w_i^e + \widetilde{w}_i^e)\Psi_{ii}^2}} + o(1) \right) \left(\frac{\sqrt{w^f}}{\sqrt{(w_j^e + \widehat{w}_j^e)\Psi_{jj}^2}} + o(1) \right) \right\} \mathbf{b}_i^\intercal \mathbf{b}_j \\ &= \mathbb{E}\left\{ \frac{w^f}{\sqrt{w_i^e + \widetilde{w}_i^e}\sqrt{w_j^e + \widehat{w}_j^e}} \right\} \frac{\mathbf{b}_i^\intercal \mathbf{b}_j}{\Psi_{ii}\Psi_{jj}} + o(p^{-1}). \end{aligned}$$

By using the above estimations, we obtain

$$\mathbf{\Sigma}_\rho = \operatorname{diag}(\mathbf{I}_p - \gamma \mathbf{\Psi}^{-1}\mathbf{B}\mathbf{B}^\intercal \mathbf{\Psi}^{-1}) + \gamma \mathbf{\Psi}^{-1}\mathbf{B}\mathbf{B}^\intercal \mathbf{\Psi}^{-1} + \mathbf{E}_n, \tag{S2.13}$$

where $\mathbf{E}_n := \mathbf{\Sigma}_\rho - \operatorname{diag}(\mathbf{I}_p - \gamma \mathbf{\Psi}^{-1}\mathbf{B}\mathbf{B}^\intercal \mathbf{\Psi}^{-1}) - \gamma \mathbf{\Psi}^{-1}\mathbf{B}\mathbf{B}^\intercal \mathbf{\Psi}^{-1}$ with $\|\mathbf{E}_n\|_{\max} = o(p^{-1})$ and

$$\gamma := \frac{6}{\pi}\mathbb{E}\big[w^f / \{(w_i^e + \widetilde{w}_i^e)(w_j^e + \widehat{w}_j^e)\}^{1/2}\big]. \tag{S2.14}$$

By the basic norm inequality $\|\mathbf{A}\|_2 \leqslant p\|\mathbf{A}\|_{\max}$ for any $p \times p$ matrix $\mathbf{A}$, we have

$$\|\mathbf{E}_n\|_2 \leqslant p \cdot o(p^{-1}) = o(1). \tag{S2.15}$$

Since $(\widetilde{w}_i^e, \widehat{w}_j^e)$ is an independent copy of (w_i^e, w_j^e) for $i \neq j$ and random variables $\{w_i^e\}_{i=1}^p$ are i.i.d., we can replace $w_i^e, w_j^e, \widetilde{w}_i^e, \widehat{w}_j^e$ in (S2.14) by $w_1^e, w_2^e, w_3^e, w_4^e$, respectively. This, together with (S2.13) and (S2.15), completes the proof of Lemma 2.3.

S2.3 Proof of Lemma 6.1

From Lemma 2.2, $\boldsymbol{\rho}_n$ and $\mathbf{W}_n$ share the same LSD, thus we only need to derive the LSD of $\mathbf{W}_n$. Since the vectors $\{\mathbf{A}_i\}_{i=1}^n$ are i.i.d., by Theorem 1.1 of Bai and Zhou (2008), we can prove Lemma 6.1 by verifying that the elements of $\mathbf{A}_1$ are *weakly dependent* in the following sense: for any non-random $p \times p$ matrix $\mathbf{D}$ with bounded spectral norm,

$$\mathrm{Var}(\mathbf{A}_1^\intercal \mathbf{D} \mathbf{A}_1) = o(p^2). \tag{S2.16}$$

From the Corollary 1.1 in Bai and Zhou (2008), (S2.16) holds true if

$$\sum_{\Lambda} \Big\{ \mathbb{E}\big(A_{1i}A_{1j} - \sigma_{ij}\big)\big(A_{1i'}A_{1j'} - \sigma_{i'j'}\big) \Big\}^2 = o(p^2), \tag{S2.17}$$

$$\max_{i \neq j} \mathbb{E}\big|A_{1i}A_{1j} - \sigma_{ij}\big|^2 = o(p), \tag{S2.18}$$

where A_{1i} is the i-th component of $\mathbf{A}_1$, $\sigma_{ij} := \mathbb{E}\, A_{1i}A_{1j}$, and

$$\Lambda = \{(i,j,i',j') : i,j,i',j' \in [p]\} \setminus \{(i,j,i',j') : i = i' \neq j = j' \text{ or } i = j' \neq j = i'\}.$$

Now, we prove (S2.17) and (S2.18). From (S2.7), we have $A_{1i} \sim \mathsf{Uniform}(-1,1)$ for any $i \in [p]$, and thus

$$\mathbb{E}\, A_{1i}^4 = \frac{1}{5}, \qquad \sigma_{ii} = \mathbb{E}\, A_{1i}^2 = \frac{1}{3}. \tag{S2.19}$$

Two moment conditions (S2.17) and (S2.18) follow from (S2.19) and the following estimations:

$$\sigma_{ij} = O(p^{-1}), \qquad \mathbb{E}\, A_{1i}A_{1j}A_{1i'}A_{1j'} = O(p^{-2}), \qquad \mathbb{E}\, A_{1i}^2 A_{1i'}A_{1j'} = O(p^{-1}), \tag{S2.20}$$

$$\mathbb{E}\, A_{1i}^3 A_{1j} = O(p^{-1}), \qquad \mathbb{E}\big(A_{1i}^2 - \sigma_{ii}\big)\big(A_{1j}^2 - \sigma_{jj}\big) = O(p^{-2}), \tag{S2.21}$$

where $i, j, i', j' \in [p]$ and $i \neq j \neq i' \neq j'$. Proofs of (S2.20) and (S2.21) are provided in Section S2.3.1. From (S2.19) – (S2.21), we obtain

$$\max_{i \neq j} \mathbb{E}\left|A_{1i}A_{1j} - \sigma_{ij}\right|^2 = O(1),$$

and

$$\begin{aligned}
&\sum_{\Lambda}\Big\{\mathbb{E}\big(A_{1i}A_{1j} - \sigma_{ij}\big)\big(A_{1i'}A_{1j'} - \sigma_{i'j'}\big)\Big\}^2 \\
&= \sum_{i\neq j\neq i'\neq j'} \big(\mathbb{E}\, A_{1i}A_{1j}A_{1i'}A_{1j'} - \sigma_{ij}\sigma_{i'j'}\big)^2 + 2\sum_{i\neq i'\neq j'} \big(\mathbb{E}\, A_{1i}^2A_{1i'}A_{1j'} - \sigma_{ii}\sigma_{i'j'}\big)^2 \\
&\qquad + 2\sum_{i\neq j\neq j'} \big(\mathbb{E}\, A_{1i}^2A_{1j}A_{1j'} - \sigma_{ij}\sigma_{ij'}\big)^2 + 2\sum_{i\neq j}\big(\mathbb{E}\, A_{1i}^3A_{1j} - \sigma_{ij}\sigma_{ii}\big)^2 \\
&\qquad + \sum_{i\neq j}\big(\mathbb{E}\, A_{1i}^2A_{1j}^2 - \sigma_{ii}\sigma_{jj}\big)^2 + \sum_{i}\big(\mathbb{E}\, A_{1i}^4 - \sigma_{ii}^2\big)^2 \\
&= O(p^4)\cdot O(p^{-4}) + O(p^3)\cdot O(p^{-2}) + O(p^3)\cdot O(p^{-2}) + O(p^2)\cdot O(p^{-2}) \\
&\qquad + O(p^2)\cdot O(p^{-4}) + O(p)\cdot O(1) \\
&= O(p).
\end{aligned}$$

This implies (S2.17) and (S2.18), completing the proof of Lemma 6.1.

S2.3.1 Proofs of (S2.20) and (S2.21)

In this section, we provide the proofs of some moment estimations (S2.20) and (S2.21), which are used in the proof of Lemma 6.1.

From equations (S2.7), (S2.12), and Assumption (A3), we have

$$
\begin{aligned}
\mathbb{E}A_{1i} &= 2\mathbb{E}\big\{F_i\big(\mathbf{b}_i^\intercal \mathbf{f}_1 + \Psi_{ii}e_{1i}\big)\big\} - 1 \\
&= \mathbb{E}\left\{2F_i\big(\Psi_{ii}e_{1i}\big) - 1\right\} + O(p^{-1/2}) =: \mathbb{E}\tilde{A}_{1i} + O(p^{-1/2}).
\end{aligned}
$$

Since $\mathbb{E}A_{1i} = 0$, we obtain that $\mathbb{E}\tilde{A}_{1i} = O(p^{-1/2})$. By using the same argument, we obtain

$$
\mathbb{E}\,\tilde{A}_{1i}^2 = \mathbb{E}\,A_{1i}^2 + O(p^{-1}), \qquad \mathbb{E}\,\tilde{A}_{1i}^3 = O(p^{-1/2}).
$$

Since $\tilde{A}_{1i}$ and $\tilde{A}_{1j}$ are independent for $i \neq j$, we have

$$
\sigma_{ij} = \mathbb{E}\,A_{1i}A_{1j} = \mathbb{E}\,\tilde{A}_{1i} \cdot \mathbb{E}\,\tilde{A}_{1j} + O(p^{-1}) = O(p^{-1}).
$$

By using the above estimations, for $i \neq j \neq i' \neq j'$, we have

$$
\mathbb{E}\,A_{1i}A_{1j}A_{1i'}A_{1j'} = \mathbb{E}\,\tilde{A}_{1i} \cdot \mathbb{E}\,\tilde{A}_{1j} \cdot \mathbb{E}\,\tilde{A}_{1i'} \cdot \mathbb{E}\,\tilde{A}_{1j'} + O(p^{-2}) = O(p^{-2}),
$$

$$
\begin{aligned}
\mathbb{E}\,A_{1i}^2A_{1i'}A_{1j'} &= \mathbb{E}\Big[A_{1i}^2\big\{\tilde{A}_{1i'} + O(p^{-1/2})\big\}\big\{\tilde{A}_{1j'} + O(p^{-1/2})\big\}\Big] \\
&= \mathbb{E}A_{1i}^2 \cdot \mathbb{E}\tilde{A}_{1i'} \cdot \mathbb{E}\tilde{A}_{1j'} + O(p^{-1}) = O(p^{-1}),
\end{aligned}
$$

$$
\mathbb{E}\,A_{1i}^3A_{1j} = \mathbb{E}\Big[\big\{\tilde{A}_{1i} + O(p^{-1/2})\big\}^3\big\{\tilde{A}_{1j} + O(p^{-1/2})\big\}\Big]
$$

$$
\begin{aligned}
&= \mathbb{E}\Big[\Big\{\tilde{A}_{1i}^3 + O(p^{-1/2}) \cdot \tilde{A}_{1i}^2 + O(p^{-1}) \cdot \tilde{A}_{1i} + O(p^{-3/2})\Big\}\Big\{\tilde{A}_{1j} + O(p^{-1/2})\Big\}\Big] \\
&= O(p^{-1}),
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}\big(A_{1i}^2 - \sigma_{ii}\big)\big(A_{1j}^2 - \sigma_{jj}\big) \\
&= \mathbb{E}\Big[\big\{\tilde{A}_{1i}^2 - \sigma_{ii} + O(p^{-1/2}) \cdot \tilde{A}_{1i} + O(p^{-1})\big\}\big\{\tilde{A}_{1j}^2 - \sigma_{jj} + O(p^{-1/2}) \cdot \tilde{A}_{1j} + O(p^{-1})\big\}\Big] \\
&= O(p^{-2}).
\end{aligned}
$$

These equations complete the proofs of (S2.20) and (S2.21).

S2.4 Proof of Lemma 6.2

Throughout this proof, $\mathbf{\Omega}_K(\lambda, \mathbf{X})$ is simply denoted as $\mathbf{\Omega}_K(\mathbf{X})$ if no confusion. Recall that $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\intercal$ and $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^\intercal$. We denote

$$
\mathbf{X}_i = (\mathbf{x}_1, \ldots, \mathbf{x}_i, \mathbf{y}_{i+1}, \ldots, \mathbf{y}_n)^\intercal, \qquad \mathbf{X}_{i0} = (\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{y}_{i+1}, \ldots, \mathbf{y}_n)^\intercal,
$$

with convention $\mathbf{X} = \mathbf{X}_n$ and $\mathbf{Y} = \mathbf{X}_0$. Recall that $\mathbf{\Gamma} = \mathbf{U}_2\mathbf{D}_2\mathbf{U}_2^\intercal$, and for any $i \in [n]$,

$$
\begin{aligned}
\beta_{i0} &= 1 - n^{-1}\mathrm{tr}\big\{\mathbf{\Gamma}(\lambda\mathbf{I} - n^{-1}\mathbf{\Gamma}^{1/2}\mathbf{X}_{i0}^\intercal\mathbf{X}_{i0}\mathbf{\Gamma}^{1/2})^{-1}\big\}, \\
\beta_i &= 1 - n^{-1}\mathbf{x}_i^\intercal\mathbf{\Gamma}^{1/2}(\lambda\mathbf{I} - n^{-1}\mathbf{\Gamma}^{1/2}\mathbf{X}_{i0}^\intercal\mathbf{X}_{i0}\mathbf{\Gamma}^{1/2})^{-1}\mathbf{\Gamma}^{1/2}\mathbf{x}_i, \\
\mathbf{H}_1 &= n^{-1}\mathbf{X}_{i0}^\intercal(\lambda\mathbf{I}_{n-1} - n^{-1}\mathbf{X}_{i0}\mathbf{\Gamma}\mathbf{X}_{i0}^\intercal)^{-1}\mathbf{X}_{i0}\mathbf{\Gamma}, \\
\mathbf{H}_2 &= n^{-1}\mathbf{\Gamma}\mathbf{X}_{i0}^\intercal(\lambda\mathbf{I}_{n-1} - n^{-1}\mathbf{X}_{i0}\mathbf{\Gamma}\mathbf{X}_{i0}^\intercal)^{-2}\mathbf{X}_{i0}\mathbf{\Gamma}, \\
\boldsymbol{\tau}_{i0} &= (1 + n^{-1}\mathrm{tr}\mathbf{H}_2)\mathbf{I}_K - \mathbf{U}_1^\intercal(\mathbf{I}_p + \mathbf{H}_1)(\mathbf{I}_p + \mathbf{H}_1)^\intercal\mathbf{U}_1, \\
\boldsymbol{\tau}_i &= n^{-1}(\mathbf{x}_i^\intercal\mathbf{H}_2\mathbf{x}_i - \mathrm{tr}\mathbf{H}_2)\mathbf{I}_K - \mathbf{U}_1^\intercal(\mathbf{I}_p + \mathbf{H}_1)(\mathbf{x}_i\mathbf{x}_i^\intercal - \mathbf{I}_p)(\mathbf{I}_p + \mathbf{H}_1)^\intercal\mathbf{U}_1.
\end{aligned}
$$

Using these notations and the Sherman-Morrison formula, we obtain

$$
\begin{aligned}
&\boldsymbol{\Omega}_K(\mathbf{X}_i) - \boldsymbol{\Omega}_K(\mathbf{X}_{i0}) \\
&= \frac{1}{\beta_i\sqrt{n}}\Big\{(1 + n^{-1}\mathbf{x}_i^{\mathsf{T}}\mathbf{H}_2\mathbf{x}_i)\mathbf{I}_K - \mathbf{U}_1^{\mathsf{T}}(\mathbf{I}_p + \mathbf{H}_1)\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}(\mathbf{I}_p + \mathbf{H}_1)^{\mathsf{T}}\mathbf{U}_1\Big\} \\
&= \frac{1}{\beta_{i0}\sqrt{n}}(\boldsymbol{\tau}_{i0} + \boldsymbol{\tau}_i) - \frac{\varepsilon_i}{\beta_{i0}^2\sqrt{n}}(\boldsymbol{\tau}_{i0} + \boldsymbol{\tau}_i) + \frac{\varepsilon_i^2}{\beta_{i0}^2\beta_i\sqrt{n}}(\boldsymbol{\tau}_{i0} + \boldsymbol{\tau}_i).
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
&\boldsymbol{\Omega}_K(\mathbf{X}_{i-1}) - \boldsymbol{\Omega}_K(\mathbf{X}_{i0}) \\
&= \frac{1}{\beta_{i0}\sqrt{n}}(\boldsymbol{\tau}_{i0} + \boldsymbol{\tau}_{iy}) - \frac{\varepsilon_{iy}}{\beta_{i0}^2\sqrt{n}}(\boldsymbol{\tau}_{i0} + \boldsymbol{\tau}_{iy}) + \frac{\varepsilon_{iy}^2}{\beta_{i0}^2\beta_{iy}\sqrt{n}}(\boldsymbol{\tau}_{i0} + \boldsymbol{\tau}_{iy}),
\end{aligned}
$$

where $\beta_{iy}, \boldsymbol{\tau}_{iy}$, and ε_{iy} are similarly defined as $\beta_i, \boldsymbol{\tau}_i$, and ε_i with $\mathbf{x}_i$ replaced by $\mathbf{y}_i$.

Now, we begin to prove that $\boldsymbol{\Omega}_K(\mathbf{X})$ and $\boldsymbol{\Omega}_K(\mathbf{Y})$ have the same limiting distribution. To this end, we show that the difference between the characteristic functions tends to zero, that is, for any $K \times K$ deterministic symmetric matrix $\mathbf{W}$,

$$
\mathbb{E}\Big(\exp\big[\mathrm{i}\,\mathrm{tr}\big\{\mathbf{W}\boldsymbol{\Omega}_K(\mathbf{X})\big\}\big]\Big) - \mathbb{E}\Big(\exp\big[\mathrm{i}\,\mathrm{tr}\big\{\mathbf{W}\boldsymbol{\Omega}_K(\mathbf{Y})\big\}\big]\Big) \to 0,
$$

where $\mathrm{i} = \sqrt{-1}$. Using the notations we introduced above, we can write

$$
\begin{aligned}
&\mathbb{E}\Big(\exp\big[\mathrm{i}\,\mathrm{tr}\big\{\mathbf{W}\boldsymbol{\Omega}_K(\mathbf{X})\big\}\big]\Big) - \mathbb{E}\Big(\exp\big[\mathrm{i}\,\mathrm{tr}\big\{\mathbf{W}\boldsymbol{\Omega}_K(\mathbf{Y})\big\}\big]\Big) \\
&= \sum_{i=1}^{n} \mathbb{E}\exp\Big(\mathrm{i}\,\mathrm{tr}\Big[\mathbf{W}\Big\{\boldsymbol{\Omega}_K(\mathbf{X}_{i0}) + \frac{\boldsymbol{\tau}_{i0}}{\beta_{i0}\sqrt{n}}\Big\}\Big]\Big)\Big\{\mathbb{E}_i\exp\Big(\mathrm{i}\,\mathrm{tr}\Big[\mathbf{W}\Big\{\frac{\boldsymbol{\tau}_i}{\beta_{i0}\sqrt{n}} - \frac{(\boldsymbol{\tau}_{i0} + \boldsymbol{\tau}_i)\varepsilon_i}{\beta_{i0}^2\sqrt{n}}\Big\}\Big]\Big) \\
&\qquad - \mathbb{E}_i\exp\Big(\mathrm{i}\,\mathrm{tr}\Big[\mathbf{W}\Big\{\frac{\boldsymbol{\tau}_{iy}}{\beta_{i0}\sqrt{n}} - \frac{(\boldsymbol{\tau}_{i0} + \boldsymbol{\tau}_{iy})\varepsilon_{iy}}{\beta_{i0}^2\sqrt{n}}\Big\}\Big]\Big)\Big\} \\
&\quad + \sum_{i=1}^{n} \mathbb{E}\exp\Big(\mathrm{i}\,\mathrm{tr}\Big[\mathbf{W}\Big\{\boldsymbol{\Omega}_K(\mathbf{X}_{i0}) + \frac{\boldsymbol{\tau}_{i0}}{\beta_{i0}\sqrt{n}}\Big\}\Big]\Big)
\end{aligned}
$$

$$\Bigg\{\mathbb{E}_i \exp\bigg(\mathrm{i}\,\mathrm{tr}\Big[\mathbf{W}\Big\{\frac{\boldsymbol{\tau}_i}{\beta_{i0}\sqrt{n}} - \frac{(\boldsymbol{\tau}_{i0}+\boldsymbol{\tau}_i)\varepsilon_i}{\beta_{i0}^2\sqrt{n}}\Big\}\Big]\bigg)\bigg(\exp\Big[\mathrm{i}\,\mathrm{tr}\Big\{\frac{\mathbf{W}(\boldsymbol{\tau}_{i0}+\boldsymbol{\tau}_i)\varepsilon_i^2}{\beta_{i0}^2\beta_i\sqrt{n}}\Big\}\Big]-1\bigg)$$
$$-\mathbb{E}_i \exp\bigg(\mathrm{i}\,\mathrm{tr}\Big[\mathbf{W}\Big\{\frac{\boldsymbol{\tau}_{iy}}{\beta_{i0}\sqrt{n}} - \frac{(\boldsymbol{\tau}_{i0}+\boldsymbol{\tau}_{iy})\varepsilon_{iy}}{\beta_{i0}^2\sqrt{n}}\Big\}\Big]\bigg)\bigg(\exp\Big[\mathrm{i}\,\mathrm{tr}\Big\{\frac{\mathbf{W}(\boldsymbol{\tau}_{i0}+\boldsymbol{\tau}_{iy})\varepsilon_{iy}^2}{\beta_{i0}^2\beta_{iy}\sqrt{n}}\Big\}\Big]-1\bigg)\Bigg\},$$

where $\mathbb{E}_i(\cdot) := \mathbb{E}(\cdot \mid \mathbf{X}_{i0})$ denotes the conditional expectation given $\mathbf{X}_{i0}$. The second term on the RHS of the above identity can be shown to be negligible by using the same argument as in (Jiang and Bai, 2021, p. 16) and our Lemma S1.7. Hence, we have

$$\begin{aligned}
&\mathbb{E}\Big(\exp\big[\mathrm{i}\,\mathrm{tr}\{\mathbf{W}\boldsymbol{\Omega}_K(\mathbf{X})\}\big]\Big) - \mathbb{E}\Big(\exp\big[\mathrm{i}\,\mathrm{tr}\{\mathbf{W}\boldsymbol{\Omega}_K(\mathbf{Y})\}\big]\Big)\\
&= \sum_{i=1}^n \mathbb{E}\exp\bigg(\mathrm{i}\,\mathrm{tr}\Big[\mathbf{W}\Big\{\boldsymbol{\Omega}_K(\mathbf{X}_{i0}) + \frac{\boldsymbol{\tau}_{i0}}{\beta_{i0}\sqrt{n}}\Big\}\Big]\bigg)\cdot\Bigg\{\mathbb{E}_i\exp\bigg(\mathrm{i}\,\mathrm{tr}\Big[\mathbf{W}\Big\{\frac{\boldsymbol{\tau}_i}{\beta_{i0}\sqrt{n}} - \frac{(\boldsymbol{\tau}_{i0}+\boldsymbol{\tau}_i)\varepsilon_i}{\beta_{i0}^2\sqrt{n}}\Big\}\Big]\bigg)\\
&\qquad - \mathbb{E}_i\exp\bigg(\mathrm{i}\,\mathrm{tr}\Big[\mathbf{W}\Big\{\frac{\boldsymbol{\tau}_{iy}}{\beta_{i0}\sqrt{n}} - \frac{(\boldsymbol{\tau}_{i0}+\boldsymbol{\tau}_{iy})\varepsilon_{iy}}{\beta_{i0}^2\sqrt{n}}\Big\}\Big]\bigg)\Bigg\} + o(1).
\end{aligned}$$

Moreover, by using the same argument, we obtain

$$\begin{aligned}
&\mathbb{E}\Big(\exp\big[\mathrm{i}\,\mathrm{tr}\{\mathbf{W}\boldsymbol{\Omega}_K(\mathbf{X})\}\big]\Big) - \mathbb{E}\Big(\exp\big[\mathrm{i}\,\mathrm{tr}\{\mathbf{W}\boldsymbol{\Omega}_K(\mathbf{Y})\}\big]\Big)\\
&= \sum_{i=1}^n \mathbb{E}\exp\bigg(\mathrm{i}\,\mathrm{tr}\Big[\mathbf{W}\Big\{\boldsymbol{\Omega}_K(\mathbf{X}_{i0}) + \frac{\boldsymbol{\tau}_{i0}}{\beta_{i0}\sqrt{n}}\Big\}\Big]\bigg)\cdot\Bigg(\mathbb{E}_i\exp\Big[\mathrm{i}\,\mathrm{tr}\Big\{\mathbf{W}\Big(\frac{\boldsymbol{\tau}_i}{\beta_{i0}\sqrt{n}} - \frac{\boldsymbol{\tau}_{i0}\varepsilon_i}{\beta_{i0}^2\sqrt{n}}\Big)\Big\}\Big]\\
&\qquad - \mathbb{E}_i\exp\Big[\mathrm{i}\,\mathrm{tr}\Big\{\mathbf{W}\Big(\frac{\boldsymbol{\tau}_{iy}}{\beta_{i0}\sqrt{n}} - \frac{\boldsymbol{\tau}_{i0}\varepsilon_{iy}}{\beta_{i0}^2\sqrt{n}}\Big)\Big\}\Big]\Bigg) + o(1).
\end{aligned}$$

This, together with Taylor's expansion, gives us

$$\begin{aligned}
&\mathbb{E}\Big(\exp\big[\mathrm{i}\,\mathrm{tr}\{\mathbf{W}\boldsymbol{\Omega}_K(\mathbf{X})\}\big]\Big) - \mathbb{E}\Big(\exp\big[\mathrm{i}\,\mathrm{tr}\{\mathbf{W}\boldsymbol{\Omega}_K(\mathbf{Y})\}\big]\Big)\\
&\leqslant \Bigg|\sum_{i=1}^n \mathbb{E}\exp\bigg(\mathrm{i}\,\mathrm{tr}\Big[\mathbf{W}\Big\{\boldsymbol{\Omega}_K(\mathbf{X}_{i0}) + \frac{\boldsymbol{\tau}_{i0}}{\beta_{i0}\sqrt{n}}\Big\}\Big]\bigg)\times\\
&\qquad C\Bigg\{\mathbb{E}_i\bigg(1 + \mathrm{i}\,\mathrm{tr}\Big\{\mathbf{W}\Big(\frac{\boldsymbol{\tau}_i}{\beta_{i0}\sqrt{n}} - \frac{\boldsymbol{\tau}_{i0}\varepsilon_i}{\beta_{i0}^2\sqrt{n}}\Big)\Big\} - \frac{1}{2}\Big[\mathrm{tr}\Big\{\mathbf{W}\Big(\frac{\boldsymbol{\tau}_i}{\beta_{i0}\sqrt{n}} - \frac{\boldsymbol{\tau}_{i0}\varepsilon_i}{\beta_{i0}^2\sqrt{n}}\Big)\Big\}\Big]^2 + o(n^{-1})\bigg)
\end{aligned}$$

$$
-\,\mathbb{E}_i\bigg(1+\mathrm{i}\,\mathrm{tr}\Big\{\mathbf{W}\Big(\frac{\boldsymbol{\tau}_{iy}}{\beta_{i0}\sqrt{n}}-\frac{\boldsymbol{\tau}_{i0}\varepsilon_{iy}}{\beta_{i0}^2\sqrt{n}}\Big)\Big\}-\frac{1}{2}\Big[\mathrm{tr}\Big\{\mathbf{W}\Big(\frac{\boldsymbol{\tau}_{iy}}{\beta_{i0}\sqrt{n}}-\frac{\boldsymbol{\tau}_{i0}\varepsilon_{iy}}{\beta_{i0}^2\sqrt{n}}\Big)\Big\}\Big]^2+o(n^{-1})\bigg)\bigg\}\bigg|
$$

$$
=o(1),
$$

where we use some facts as follows:

$$
\begin{aligned}
&\mathbb{E}_i\mathrm{tr}\Big\{\mathbf{W}\Big(\frac{\boldsymbol{\tau}_{i}}{\beta_{i0}\sqrt{n}}-\frac{\boldsymbol{\tau}_{i0}\varepsilon_{i}}{\beta_{i0}^2\sqrt{n}}\Big)\Big\}=\mathbb{E}_i\mathrm{tr}\Big\{\mathbf{W}\Big(\frac{\boldsymbol{\tau}_{iy}}{\beta_{i0}\sqrt{n}}-\frac{\boldsymbol{\tau}_{i0}\varepsilon_{iy}}{\beta_{i0}^2\sqrt{n}}\Big)\Big\}=0,\\
&\mathbb{E}_i\Big[\mathrm{tr}\Big\{\mathbf{W}\Big(\frac{\boldsymbol{\tau}_{i}}{\beta_{i0}\sqrt{n}}-\frac{\boldsymbol{\tau}_{i0}\varepsilon_{i}}{\beta_{i0}^2\sqrt{n}}\Big)\Big\}\Big]^2-\mathbb{E}_i\Big[\mathrm{tr}\Big\{\mathbf{W}\Big(\frac{\boldsymbol{\tau}_{iy}}{\beta_{i0}\sqrt{n}}-\frac{\boldsymbol{\tau}_{i0}\varepsilon_{iy}}{\beta_{i0}^2\sqrt{n}}\Big)\Big\}\Big]^2=o_P(n^{-1}). \quad \text{(S2.22)}
\end{aligned}
$$

Finally, it remains to prove (S2.22). Note that

$$
\begin{aligned}
&\Big[\mathrm{tr}\Big\{\mathbf{W}\Big(\frac{\boldsymbol{\tau}_{i}}{\beta_{i0}\sqrt{n}}-\frac{\boldsymbol{\tau}_{i0}\varepsilon_{i}}{\beta_{i0}^2\sqrt{n}}\Big)\Big\}\Big]^2\\
&=\frac{1}{n}\Big[\frac{1}{\beta_{i0}^2}\{\mathrm{tr}(\mathbf{W}\boldsymbol{\tau}_i)\}^2-\frac{2\varepsilon_i}{\beta_{i0}^3}\mathrm{tr}(\mathbf{W}\boldsymbol{\tau}_i)\mathrm{tr}(\mathbf{W}\boldsymbol{\tau}_{i0})+\frac{\varepsilon_i^2}{\beta_{i0}^4}\{\mathrm{tr}(\mathbf{W}\boldsymbol{\tau}_{i0})\}^2\Big],
\end{aligned}
$$

and from Lemma S1.7, we have

$$
\begin{aligned}
&\mathbb{E}\Big\{\frac{\varepsilon_i}{\beta_{i0}^3}\mathrm{tr}(\mathbf{W}\boldsymbol{\tau}_i)\mathrm{tr}(\mathbf{W}\boldsymbol{\tau}_{i0})\Big\}\asymp\mathbb{E}\,\varepsilon_i=o(1),\\
&\mathbb{E}\Big[\frac{\varepsilon_i^2}{\beta_{i0}^4}\big\{\mathrm{tr}(\mathbf{W}\boldsymbol{\tau}_{i0})\big\}^2\Big]\asymp\mathbb{E}\,\varepsilon_i^2=o(1).
\end{aligned}
$$

Since $\mathbf{X}$ and $\mathbf{Y}$ satisfy (B1) – (B4), by using above estimates and Lemma S1.7, we obtain

$$
\begin{aligned}
&\mathbb{E}_i\Big[\mathrm{tr}\Big\{\mathbf{W}\Big(\frac{\boldsymbol{\tau}_{i}}{\beta_{i0}\sqrt{n}}-\frac{\boldsymbol{\tau}_{i0}\varepsilon_{i}}{\beta_{i0}^2\sqrt{n}}\Big)\Big\}\Big]^2-\mathbb{E}_i\Big[\mathrm{tr}\Big\{\mathbf{W}\Big(\frac{\boldsymbol{\tau}_{iy}}{\beta_{i0}\sqrt{n}}-\frac{\boldsymbol{\tau}_{i0}\varepsilon_{iy}}{\beta_{i0}^2\sqrt{n}}\Big)\Big\}\Big]^2\\
&=\frac{1}{n}\mathbb{E}_i\bigg(\frac{1}{\beta_{i0}^2}\Big[\big\{\mathrm{tr}(\mathbf{W}\boldsymbol{\tau}_i)\big\}^2-\big\{\mathrm{tr}(\mathbf{W}\boldsymbol{\tau}_{iy})\big\}^2\Big]+o(1)\bigg)=o_P(n^{-1}),
\end{aligned}
$$

which is (S2.22), completing the proof of Lemma 6.2.

S2.5 Proof of Lemma S1.6

From the definition of $\mathbf{Q}$ in (S1.1), we obtain $\mathbf{Q}|(W_X, W_{Y_1}, W_{Y_2}) \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{pmatrix} W_X\sigma_{X_1}^2 + W_{Y_1}\sigma_{Y_1}^2 & \gamma W_X \\ \gamma W_X & W_X\sigma_{X_2}^2 + W_{Y_2}\sigma_{Y_2}^2 \end{pmatrix}.$$

For $\widetilde{Q}_1$ and $\widetilde{Q}_2$, they have the following stochastic representations:

$$\widetilde{Q}_1 \overset{\mathsf{d}}{=} \sqrt{\widetilde{W}_{X_1}}\widetilde{X}_1 + \sqrt{\widetilde{W}_{Y_1}}\widetilde{Y}_1, \qquad \widetilde{Q}_2 \overset{\mathsf{d}}{=} \sqrt{\widetilde{W}_{X_2}}\widetilde{X}_2 + \sqrt{\widetilde{W}_{Y_2}}\widetilde{Y}_2,$$

where $(\widetilde{W}_{X_1}, \widetilde{W}_{Y_1})$, $(\widetilde{W}_{X_2}, \widetilde{W}_{Y_2})$, $(\widetilde{X}_1, \widetilde{Y}_1)$, and $(\widetilde{X}_2, \widetilde{Y}_2)$ are independent copies of (W_X, W_{Y_1}), (W_X, W_{Y_2}), (X_1, Y_1), and (X_2, Y_2), respectively. Note that $\widetilde{X}_1$ and $\widetilde{X}_2$ are independent. Two random vectors $\mathbf{Q}$ and $\widetilde{\mathbf{Q}} = (\widetilde{Q}_1, \widetilde{Q}_2)^\intercal$ are independent, and $\widetilde{\mathbf{Q}}|(\widetilde{W}_{X_1}, \widetilde{W}_{X_2}, \widetilde{W}_{Y_1}, \widetilde{W}_{Y_2}) \sim \mathcal{N}_2(\mathbf{0}, \widetilde{\boldsymbol{\Sigma}})$, where

$$\widetilde{\boldsymbol{\Sigma}} = \begin{pmatrix} \widetilde{W}_{X_1}\sigma_{X_1}^2 + \widetilde{W}_{Y_1}\sigma_{Y_1}^2 & 0 \\ 0 & \widetilde{W}_{X_2}\sigma_{X_2}^2 + \widetilde{W}_{Y_2}\sigma_{Y_2}^2 \end{pmatrix}.$$

Comnbining $\mathbf{Q}$ and $\widetilde{\mathbf{Q}}$ into a vector $\boldsymbol{\mathcal{Q}} := \left(\begin{smallmatrix} \mathbf{Q} \\ \widetilde{\mathbf{Q}} \end{smallmatrix}\right)$, then we have $\boldsymbol{\mathcal{Q}}|\mathbf{W} \sim \mathcal{N}_4(\mathbf{0}, \boldsymbol{\Omega})$, where $\mathbf{W} := (W_X, W_{Y_1}, W_{Y_2}, \widetilde{W}_{X_1}, \widetilde{W}_{X_2}, \widetilde{W}_{Y_1}, \widetilde{W}_{Y_2})$, and $\boldsymbol{\Omega} = \left(\begin{smallmatrix} \boldsymbol{\Sigma} & O \\ O & \widetilde{\boldsymbol{\Sigma}} \end{smallmatrix}\right)$. From

$$\left(\begin{smallmatrix} Q_1 - \widetilde{Q}_1 \\ Q_2 - \widetilde{Q}_2 \end{smallmatrix}\right) = (\mathbf{I}_2, -\mathbf{I}_2)\boldsymbol{\mathcal{Q}} =: \mathbf{B}_0\boldsymbol{\mathcal{Q}},$$

we obtain $\left(\begin{smallmatrix} Q_1 - \widetilde{Q}_1 \\ Q_2 - \widetilde{Q}_2 \end{smallmatrix}\right) \Big| \mathbf{W} \sim \mathcal{N}_2(\mathbf{0}, \mathbf{B}_0\boldsymbol{\Omega}\mathbf{B}_0^\intercal)$, where

$$\mathbf{B}_0\boldsymbol{\Omega}\mathbf{B}_0^\intercal = \begin{pmatrix} (W_X + \widetilde{W}_{X_1})\sigma_{X_1}^2 + (W_{Y_1} + \widetilde{W}_{Y_1})\sigma_{Y_1}^2 & \gamma W_X \\ \gamma W_X & (W_X + \widetilde{W}_{X_2})\sigma_{X_2}^2 + (W_{Y_2} + \widetilde{W}_{Y_2})\sigma_{Y_2}^2 \end{pmatrix}.$$

Given $\mathbf{W}$, the Pearson correlation coefficient between $Q_1 - \widetilde{Q}_1$ and $Q_2 - \widetilde{Q}_2$ is

$$r = \frac{\gamma W_X}{\sqrt{(W_X + \widetilde{W}_{X_1})\sigma_{X_1}^2 + (W_{Y_1} + \widetilde{W}_{Y_1})\sigma_{Y_1}^2}\sqrt{(W_X + \widetilde{W}_{X_2})\sigma_{X_2}^2 + (W_{Y_2} + \widetilde{W}_{Y_2})\sigma_{Y_2}^2}}.$$

By Lemma S1.5, we conclude that

$$\mathbb{E}\{\text{sign}(Q_1 - \widetilde{Q}_1)\text{sign}(Q_2 - \widetilde{Q}_2)\} = \mathbb{E}\Big[\mathbb{E}\Big\{\text{sign}(Q_1 - \widetilde{Q}_1)\text{sign}(Q_2 - \widetilde{Q}_2)\Big|\mathbf{W}\Big\}\Big] = \frac{2}{\pi}\mathbb{E}\big\{\arcsin(r)\big\}.$$

This completes the proof of Lemma S1.6.

S2.6 Proof of Lemma S1.7

The convergence of β_{i0} can be proven by using the same argument in the proof of Lemma C.3 of Jiang and Bai (2021). The conclusion $\varepsilon_i \to 0$ follows from (S1.3). First, we prove that $\mathbb{E}\,\varepsilon_i^2 = o(1)$. Since $\|\boldsymbol{\Gamma}\|_2$ is bounded and the variable λ (spiked eigenvalue) stays away from the bulk eigenvalues of $n^{-1}\boldsymbol{\Gamma}^{1/2}\mathbf{X}_{i0}^\intercal\mathbf{X}_{i0}\boldsymbol{\Gamma}^{1/2}$, the spectral norm of $\mathbf{B} := \boldsymbol{\Gamma}^{1/2}(\lambda\mathbf{I} - n^{-1}\boldsymbol{\Gamma}^{1/2}\mathbf{X}_{i0}^\intercal\mathbf{X}_{i0}\boldsymbol{\Gamma}^{1/2})^{-1}\boldsymbol{\Gamma}^{1/2}$ is bounded. By Assumptions (A1) and (B3), we have

$$\mathbb{E}\,\varepsilon_i^2 = \frac{1}{n^2}\mathbb{E}\big|\mathbf{x}_i^\intercal\mathbf{B}\mathbf{x}_i - \text{tr}\mathbf{B}\big|^2 = o(1).$$

Second, we prove $\mathbb{E}\varepsilon_i^4 = o(n^{-1})$. From Assumption (B4), we can apply Lemma S1.2 to estimating moments of ε_i. For $\delta > 0$, let $t_n(\delta) = (p/n)(\log p)^{1+\delta}/\sqrt{pc(p)}$, where $c(p)$ is defined in Assumption (B4). Define an event

$$\mathcal{F}_n(\delta) = \{\mathbf{x}_i : n^{-1}(\mathbf{x}_i^\intercal\mathbf{B}\mathbf{x}_i - \text{tr}\mathbf{B}) < t_n(\delta)\}.$$

From Lemma S1.2 and the fact $\mathbb{E}(\mathbf{x}_1\mathbf{x}_1^\intercal) = \mathbf{I}_p$ (see Assumption (B2)), we have

$$\mathbb{P}\big(\mathcal{F}_n^c(\delta)\big) = o(n^{-1}). \tag{S2.23}$$

By the elementary inequality $|a+b|^r \leqslant 2^{r-1}(|a|^r + |b|^r)$ for $r > 1$, we have

$$\mathbb{E}\,\varepsilon_i^4 = \mathbb{E}\big|\varepsilon_i \mathbb{1}_{\{\mathcal{F}_n(\delta)\}} + \varepsilon_i \mathbb{1}_{\{\mathcal{F}_n^c(\delta)\}}\big|^4 \leqslant 8\big[\mathbb{E}\varepsilon_i^4 \mathbb{1}_{\{\mathcal{F}_n(\delta)\}} + \mathbb{E}\varepsilon_i^4 \mathbb{1}_{\{\mathcal{F}_n^c(\delta)\}}\big]. \tag{S2.24}$$

For the first term, we have

$$\mathbb{E}\,\varepsilon_i^4 \mathbb{1}_{\{\mathcal{F}_n(\delta)\}} = \int_0^\infty 4t^3 \mathbb{P}\big(\big|\varepsilon_i \mathbb{1}_{\{\mathcal{F}_n(\delta)\}}\big| > t\big)\,\mathrm{d}t \leqslant \int_0^{t_n(\delta)} 4t^3\,\mathrm{d}t = o(n^{-1}). \tag{S2.25}$$

For the second term, by (S2.23) and the fact that $|\varepsilon_i| = |n^{-1}\big(\mathbf{x}_i^\intercal \mathbf{B}\mathbf{x}_i - \mathrm{tr}\mathbf{B}\big)| \leqslant C\|\mathbf{B}\|_2$ (using Assumption (B2) and the Courant–Fischer principle) and $\|\mathbf{B}\|_2 = O(1)$, we obtain

$$\mathbb{E}\,\varepsilon_i^4 \mathbb{1}_{\{\mathcal{F}_n^c(\delta)\}} \leqslant C^4 \|\mathbf{B}\|_2^4 \mathbb{P}\big(\mathcal{F}_n^c(\delta)\big) = o(n^{-1}).$$

This, together with (S2.24) and (S2.25), implies that $\mathbb{E}\,\varepsilon_i^4 = o(n^{-1})$.

Finally, we prove (S1.4). The proofs of $\mathbb{E}\{\mathrm{tr}(\mathbf{W}\boldsymbol{\tau}_i)\}^2 = O(1)$ and $\mathbb{E}\{\mathrm{tr}(\mathbf{W}\boldsymbol{\tau}_{i0})\}^2 = O(1)$ are similar, and thus we only prove the first conclusion. Let

$$\begin{aligned}
\boldsymbol{\tau}_{i1} &= n^{-1}(\mathbf{x}_i^\intercal \mathbf{H}_2 \mathbf{x}_i - \mathrm{tr}\mathbf{H}_2)\mathbf{I}_K, \\
\boldsymbol{\tau}_{i2} &= \mathbf{U}_1^\intercal(\mathbf{I}_p + \mathbf{H}_1)(\mathbf{x}_i\mathbf{x}_i^\intercal - \mathbf{I}_p)(\mathbf{I}_p + \mathbf{H}_1)^\intercal \mathbf{U}_1,
\end{aligned}$$

then we can write $\boldsymbol{\tau}_i = \boldsymbol{\tau}_{i1} - \boldsymbol{\tau}_{i2}$. To prove $\mathbb{E}\{\mathrm{tr}(\mathbf{W}\boldsymbol{\tau}_i)\}^2 = O(1)$, it suffices to show that

$\mathbb{E}\{\mathrm{tr}(\mathbf{W}\boldsymbol{\tau}_{i2})\}^2 = O(1)$, since

$$\mathbb{E}\{\mathrm{tr}(\mathbf{W}\boldsymbol{\tau}_{i})\}^2 = \mathbb{E}\{\mathrm{tr}\mathbf{W}(\boldsymbol{\tau}_{i1} - \boldsymbol{\tau}_{i2})\}^2 \leqslant 2\mathbb{E}\{\mathrm{tr}(\mathbf{W}\boldsymbol{\tau}_{i1})\}^2 + 2\mathbb{E}\{\mathrm{tr}(\mathbf{W}\boldsymbol{\tau}_{i2})\}^2,$$

and

$$\mathbb{E}\{\mathrm{tr}(\mathbf{W}\boldsymbol{\tau}_{i1})\}^2 = \frac{1}{n^2}(\mathrm{tr}\mathbf{W})^2\mathbb{E}(\mathbf{x}_i^\intercal\mathbf{H}_2\mathbf{x}_i - \mathrm{tr}\mathbf{H}_2)^2 = o(1),$$

which follows from Assumption (B3) and the fact $\|\mathbf{H}_2\|_2 = O(1)$. Let

$$\boldsymbol{\Delta} = (\mathbf{I}_p + \mathbf{H}_1^\intercal)\mathbf{U}_1\mathbf{W}\mathbf{U}_1^\intercal(\mathbf{I}_p + \mathbf{H}_1),$$

then we have

$$\mathbb{E}[\mathrm{tr}(\mathbf{W}\boldsymbol{\tau}_{i2})]^2 = \mathbb{E}|\mathbf{x}_i^\intercal\boldsymbol{\Delta}\mathbf{x}_i - \mathrm{tr}\boldsymbol{\Delta}|^2 \leqslant 2\mathbb{E}|\mathbf{x}_i^\intercal\boldsymbol{\Delta}\mathbf{x}_i|^2 + 2\mathbb{E}|\mathrm{tr}\boldsymbol{\Delta}|^2.$$

Note that

$$\begin{aligned}\mathrm{tr}(\boldsymbol{\Delta}) &= \mathrm{tr}\big\{\mathbf{W}\mathbf{U}_1^\intercal(\mathbf{I}_p + \mathbf{H}_1)(\mathbf{I}_p + \mathbf{H}_1^\intercal)\mathbf{U}_1\big\} \\ &\leqslant K\|\mathbf{W}\mathbf{U}_1^\intercal(\mathbf{I}_p + \mathbf{H}_1)(\mathbf{I}_p + \mathbf{H}_1^\intercal)\mathbf{U}_1\|_2 = O(1),\end{aligned}$$

and similarly, $(\mathrm{tr}\boldsymbol{\Delta})^2 = O(1)$. Thus, we need only to show $\mathbb{E}|\mathbf{x}_i^\intercal\boldsymbol{\Delta}\mathbf{x}_i|^2 = O(1)$. Let $\Delta_{k\ell}$ denote the (k, ℓ)-th entry of $\boldsymbol{\Delta}$. From Lemma S1.4, we obtain $\sum_{k=1}^p \Delta_{kk}^2 \leqslant \mathrm{tr}(\boldsymbol{\Delta}^2) = O(1)$. By using Cauchy-Schwarz inequality, we obtain

$$\sum_{k\neq\ell}\Delta_{kk}\Delta_{\ell\ell} \leqslant \Big(\sum_k \Delta_{kk}\Big)\Big(\sum_\ell \Delta_{\ell\ell}\Big) = (\mathrm{tr}\boldsymbol{\Delta})^2 = O(1),$$

$$\sum_{k\neq\ell\neq s}\Delta_{k\ell}\Delta_{ks}\asymp\sum_{k\neq\ell\neq s}\Delta_{kk}\Delta_{\ell s}\leqslant\Big(\sum_{k}\Delta_{kk}\Big)\cdot\Big(\sum_{\ell\neq s}\Delta_{\ell s}\Big)\leqslant(\mathrm{tr}\boldsymbol{\Delta})\cdot p(\mathrm{tr}\boldsymbol{\Delta}^2)^{1/2}=O(p),$$

$$\sum_{k\neq\ell\neq s\neq t}\Delta_{k\ell}\Delta_{st}\leqslant\Big(\sum_{k\neq\ell}\Delta_{k\ell}\Big)\Big(\sum_{s\neq t}\Delta_{st}\Big)\leqslant p^2\mathrm{tr}(\boldsymbol{\Delta}^2)=O(p^2).$$

It follows from the above inequalities and Assumption (B2) that

$$\begin{aligned}\mathbb{E}|\mathbf{x}_i^\intercal\boldsymbol{\Delta}\mathbf{x}_i|^2&=\sum_{k=1}^{p}\mathbb{E}\Delta_{kk}^2\cdot\mathbb{E}X_{ik}^4+\sum_{k\neq\ell}\mathbb{E}(\Delta_{kk}\Delta_{\ell\ell}+2\Delta_{k\ell}^2)\cdot\mathbb{E}X_{ik}^2X_{i\ell}^2\\&\quad+\sum_{k\neq\ell\neq s}\mathbb{E}(2\Delta_{kk}\Delta_{\ell s}+4\Delta_{k\ell}\Delta_{ks})\cdot\mathbb{E}X_{ik}^2X_{i\ell}X_{is}\\&\quad+\sum_{k\neq\ell\neq s\neq t}\mathbb{E}\Delta_{k\ell}\Delta_{st}\cdot\mathbb{E}X_{ik}X_{i\ell}X_{is}X_{it}\\&=O(1).\end{aligned}$$

This completes the proof of Lemma S1.7.

References

Bai, Z. and Zhou, W. (2008). Large sample covariance matrices without independence structures in columns. *Statistica Sinica*, 18(2):425–442.

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

Horn, R. A. and Johnson, C. R. (1991). *Topic in matrix analysis*. Cambridge University Press.

Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis*. Cambridge University Press.

Jiang, D. and Bai, Z. (2021). Supplement to “Generalized four moment theorem and an application to CLT for spiked eigenvalues of high-dimensional covariance matrices”. *Bernoulli*.

Ledoux, M. (2001). *The concentration of measure phenomenon*, volume 89. American Mathematical Society.

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

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press.