Statistica Sinica Preprint No: SS-2016-0364	
Title	High dimensional semiparametric estimate of latent covariance matrix for matrix-variate
Manuscript ID	SS-2016-0364
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0364
Complete List of Authors	Lu Niu and Junlong Zhao
Corresponding Author	Junlong Zhao
E-mail	zhaojunlong928@126.com
Notice: Accepted version subject to English editing.	

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

High dimensional semiparametric estimate of latent covariance matrix for matrix-variate

Lu Niu[1], Junlong Zhao[2*]

1. School of Mathematics and System Science, Beihang University, China

2. School of Statistics, Beijing Normal University, China

December 26, 2017

Abstract

Estimation of the covariance matrix of high dimensional matrix-variate is an important issue. Many methods have been developed, based on sample covariance matrix under the Gaussian or sub-Gaussian assumption. However, sub-Gaussian assumption is restrictive and the estimate based on the sample covariance matrix is not robust. In this paper, we consider the estimate of covariance matrix for high dimensional matrix-variate in the frame of transelliptical distribution and the Kendall's τ correlation. Since the covariance matrix of matrix-variate is commonly assumed to own some low dimension structure, we consider the structure of Kronecker expansion in this paper. The asymptotic results of the estimator are established. Simulation results and real data analysis confirm the effectiveness of our method.

Key words: matrix-variate, latent covariance (correlation) matrix, robust estimate, Kronecker product

1 Introduction

Covariance matrix has been widely used in the inference in statistical inference, such as PCA and all sorts of testing statistics in multivariate analysis. Estimation of covariance matrix has attracted great attention in diverse fields, including bioinformatics (Jones et al., 2012) and economics and financial time series analysis, e.g., portfolio selection (Ledoit and Wolf, 2001), risk management (Karceski and Lakonishok, 1999) and asset pricing (Engle et al., 2010), etc. Since the sample covariance matrix will be singular when

*Corresponding author. Email:zhaojunlong928@126.com. Junlong Zhao was supported by National Science Foundation of China, No.11471030 and No.11101022, and the Fundamental Research Funds for the Central Universities.

the dimension is larger than the sample size, the estimation problem is generally challenging, especially when the dimension is high. To estimate covariance matrix efficiently, some low dimension structures are often assumed, such as sparsity or low rank. For the vector-valued variate, many works have been developed on the estimation of sparse or low rank covariance matrices (Johnson et al., 2011; Bickel and Levina, 2008, 2009; Lam and Fan, 2009; Rigollet and Tsybakov, 2012, etc.). A detailed review on this topic can be referred to Fan et al. (2015).

With the rapid development of new technology, in many applications, researchers often collect the data of matrix-variate $\{X_k \in \mathbb{R}^{p\times q}, 1 \le k \le n\}$ with $X_k = (X_{ij,k})_{1\le i\le p, 1\le j\le q} \in \mathbb{R}^{p\times q}$, such as Nuclear Magnetic Resonance(NMR) data (Wallbacks and Norden, 2006) and electroencephalograph (EEG) data (Sejnowski et al., 2007). Covariance matrix estimation of this kind of data is important in application. When both p and q are fixed, many works have been developed (Dutilleul, 1999; Gupta and Nagar, 1999). In recent years, some works have been developed for the case of p and q being diverging under additional low dimensional structure, such as sparsity and Kronecker structure (Leng and Pan, 2017; Tsiligkaridis and Hero, 2013, etc.).

When the dimensions p and q are large, to estimate the covariance matrix of X_k efficiently, Tsiligkaridis and Hero (2013) considered the case where the covariance matrix $\mathbf{\Sigma} = \mathrm{cov}(\mathrm{vec}(X_k))$ has the following Kronecker form

$$\mathbf{\Sigma} = \sum_{i=1}^{r} A_i \otimes B_i, \tag{1.1}$$

where A_i's are $q \times q$ linearly independent matrices and B_i's are $p \times p$ linearly independent matrices, and $r \le \min\{p^2, q^2\}$. Here linear independence means that vectors $\{\mathrm{vec}(A_i), i = 1, \cdots, r\}$ are linearly independent, so are $\{\mathrm{vec}(B_i), i = 1, \cdots, r\}$. Since $\mathbf{\Sigma}$ is symmetric and semi-positive definite, the equation (1.1) imposes some restrictions on A_i's and B_i's. For example, when $r = 1$, A_i's and B_i's should be symmetric and semi-positive definite.

Model (1.1) with $r \ge 1$ has applications in various fields, including video modeling and classification, network anomaly detection and Magnetoencephalography(MEG)/EEG covariance modeling (Greenewald and Hero, 2014a,c; Tsiligkaridis and Hero, 2013). For example, Greenewald and Hero (2014b) analyzed the yeast metabolic cell cycle data set,

where 9335 gene probes are sampled approximately every 24 minutes for a total of 36 time points, and there are about three different cell cycles in this data. According to this study, matrices B_i's serve as spatial factors describing the dependencies among the genes, and matrices A_i's with dimension 36×36 serve as temporal factors, describing the dependencies among different time points. Since spatial and temporal dependency pattern may vary in different cell cycles, r stands for the number of different dependence patterns. The estimated value of r by Greenewald and Hero (2014b) is 3, which matches the number of the cell cycles. Moreover, as pointed by Loan and Pitsianis (1992), any $pq \times pq$ matrix $\mathbf{M}$ can be represented by (1.1) with sufficiently large r. Covariance matrix in (1.1) with small r has a low dimension structure. Tsiligkaridis and Hero (2013) proposed a Permuted Rank-penalized Least Squares (PRLS) estimator to estimate the covariance matrix with structure (1.1).

A special case for model (1.1) is that $\mathbf{\Sigma} = A \otimes B$ (i.e. $r = 1$). This special case has been widely considered in low dimensional case with normal matrix-variate (Dutilleul, 1999), and high dimensional case with Gaussian assumption on X_k and sparsity assumption on both A and B (Leng and Tang, 2012; Tsiligkaridis et al., 2012).

However, the PRLS method (Tsiligkaridis and Hero, 2013) and many others mentioned above utilize the sample covariance matrix under the Gaussian or sub-Gaussian assumption. Consequently, as argued by Han and Liu (2014), there are several disadvantages. (1) These estimates are not robust to outliers or heavy tailed distribution. (2) The theory of these methods relies heavily on the Gaussian or sub-Gaussian assumption, which may not be realistic for many real-world applications. Therefore, it is desirable to develop a robust estimate under weak assumption on distribution.

In the traditional case of vector-valued variable $\mathbf{Y} = (Y_1, \cdots, Y_p)^\top \in \mathbb{R}^p$, several works (Liu et al., 2012; Han and Liu, 2014, 2017) relaxed the sub-Gaussian assumption, proposing a transelliptical family of distribution. $\mathbf{Y}$ follows a transelliptical distribution, if there exists unspecific strictly increasing function $(f_1, \cdots, f_p)$ such that $(f_1(Y_1), \cdots, f_p(Y_p))$ follows an elliptical distribution with the location parameter 0 and the scale parameter Γ^0, whose diagonal elements are 1. Γ^0 is called *latent generalized correlation matrix* (Han and Liu, 2014). Moreover, Liu et al. (2009, 2012) and Han and Liu (2014) introduced *latent covariance matrix*, denoted as Γ, in the margin-preserved nonparanormal distribution. Note that inverse function f_j^{-1} exists, since f_j is a strictly increasing function in

the above definitions. Consequently, viewing the nuisance parameter $(f_j, 1 \leq j \leq p)$ as a kind of contamination, $\mathbf{Y}$ can be viewed as contaminated observation of some elliptical or normal variable with correlation matrix Γ^0, which is the parameter of interest. Han and Liu (2014, 2017) developed their scale-invariant PCA based on the robust estimate of Γ^0.

In this paper, we extend some ideas of Han and Liu (2014) to matrix-variate, considering the estimate of latent covariance matrix of matrix-variate $X_k \in \mathbb{R}^{p\times q}$. However, different from Han and Liu (2014), we consider the case where the latent covariance matrix has the structure (1.1). Our method is also an extension of Tsiligkaridis and Hero (2013), relaxing the Gaussian assumption. Both our method and those of Tsiligkaridis and Hero (2013) are two-step estimate but with different initial values.

There are two major contributions. First, when r is unknown in (1.1), we propose an estimator based on Kendall's τ correlation. The study of the statistical properties of the estimator is nontrivial. For vector-valued variables, there are some works estimating the correlation matrix based on nonparanormal distribution and Kendall's τ correlation (e.g. Liu et al., 2012; Han and Liu, 2014, 2017; Wegkamp et al., 2016). Although Kendall's τ correlation is used in both these estimators and our proposal, there are two significant differences. (i) The works on vector-valued data (e.g. Liu et al., 2012; Han and Liu, 2014) do not take into account the structure of (1.1). (ii) The theoretical analysis is quite different. Our proposal involves a linear operator $\mathcal{T}$ (see Section 2.3 for details) and we need to study the error of $\mathcal{T}(\hat{\mathbf{R}}^\tau)$ rather than $\hat{\mathbf{R}}^\tau$, where $\hat{\mathbf{R}}^\tau$ denotes the estimate of correlation matrix based on Kendall's τ correlation. The main challenge is that $\mathcal{T}(\hat{\mathbf{R}}^\tau)$ is asymmetric, while the matrix concentration inequalities used in studying $\hat{\mathbf{R}}^\tau$ are not applicable here (Han and Liu, 2017; Wegkamp et al., 2016). A different approach is used in this paper to establish the convergence rate.

Second, we study the statistical property of our estimator when r is given in advance and $r = 1$. This case is different from that considered by Tsiligkaridis and Hero (2013). Particularly, for fixed $r = 1$, we get the estimate of A and B, respectively. The asymptotic results show that the estimator is effective even in the case of the dimension p, q being exponential order of sample size n, when the matrices A and B are dense.

Notations. For any scalar $a \in \mathbb{R}$, let $a_+ = \max\{a, 0\}$. For any integer m, $[m] = \{1, \cdots, m\}$. $\mathbf{1}_m = (1, 1, \cdots, 1)^\top \in \mathbb{R}^m$. For any vector $\mathbf{v} \in \mathbb{R}^m$, $\|\mathbf{v}\|$ denotes the Euclidean norm of $\mathbf{v}$. For any $m \times m$ matrix $M = (M_{ij})$, $\|M\|_{op}$ denotes the operator norm,

$\|M\|_{\max} = \max_{i,j} |M_{ij}|$ and $\|M\|_F$ is the Frobenius norm of M. $\|M\|_*$ denotes the nuclear norm and $\|M\|_* = \sum_{l=1}^{rk_M} \varphi_l(M)$, where $rk_M = \text{rank}(M)$ and $\varphi_l(M)$ is the l-th largest singular value of M. $\text{diag}(M)$ denotes the vector consisting of the diagonal elements of M and $(\text{diag}(M))$ denotes the diagonal matrix of which the diagonal elements are $\text{diag}(M)$. $\text{tr}(M)$ denotes the trace of M. For any matrices $M_1, M_2 \in \mathbb{R}^{m\times n}$, $M_1 \circ M_2$ denotes the Hadamard product of M_1 and M_2. For any set S, denote $|S|$ the cardinality of S. And for two series $\{a_n\}$ and $\{b_n\}$, $a_n \asymp b_n$ means that $0 < c^{-1} \leq \lim_n a_n/b_n \leq c < \infty$ for some constant c. For clarity, for any random vector $\mathbf{Y} = (Y_1, \cdots, Y_p) \in \mathbb{R}^p$, the Pearson correlation and Kendall's τ correlation between Y_i and Y_j are denoted as $\text{corr}(Y_i, Y_j)$ and $\tau(Y_i, Y_j)$, respectively. And the Pearson correlation matrix and Kendall's τ correlation matrix of $\mathbf{Y}$ are denoted as $\text{corr}(\mathbf{Y}) = (\text{corr}(Y_i, Y_j)) \in \mathbb{R}^{p\times p}$ and $\text{corr}^{\mathcal{K}}(\mathbf{Y}) = (\tau(Y_i, Y_j)) \in \mathbb{R}^{p\times p}$, respectively.

2 High dimensional latent covariance matrix estimation for matrix-variate

2.1 Brief review of some concepts

We first review some concepts on transelliptical distribution (Fang et al., 2002; Liu et al., 2009, 2012; Han and Liu, 2014, 2017).

Definition 1 (Elliptical distribution). A random vector $\mathbf{Y} = (Y_1, ..., Y_p)^\top \in \mathbb{R}^p$ follows an elliptical distribution if and only if $\mathbf{Y}$ has a stochastic representation: $\mathbf{Y} \overset{d}{=} \mu + \xi\mathbf{AU}$. Here $\mu \in \mathbb{R}^p$, $\mathbf{A} \in \mathbb{R}^{p\times q}$ with $q = \text{rank}(\mathbf{A})$, $\xi \geq 0$ is a random variable independent of $\mathbf{U}$ and $\mathbf{U} \in \mathcal{S}^{q-1}$ is uniformly distributed on the unit sphere $\mathcal{S}^{q-1}$ in $\mathbb{R}^q$. Letting $\Gamma = \mathbf{AA}^\top$, we denote $\mathbf{Y} \sim EC_p(\mu, \Gamma, \xi)$. Γ is called the scatter matrix.

Definition 2 (Transelliptical family). A continuous random vector $\mathbf{Y} = (Y_1, \cdots, Y_p)^\top \in \mathbb{R}^p$ follows a transelliptical distribution, denoted by $\mathbf{Y} \sim TE_p(\Gamma^0, \xi; f_1, \cdots, f_p)$, if there exists univariate strictly increasing functions $f_1, \cdots, f_p$ such that

$$(f_1(Y_1), \cdots, f_p(Y_p)) \sim EC_p(0, \Gamma^0, \xi),$$

where $EC_p(0, \Gamma^0, \xi)$ denotes an elliptical distribution with $\text{diag}(\Gamma^0) = \mathbf{1}_p$. Here Γ^0 is called latent generalized correlation matrix. Particularly, if the elliptical distribution is replaced

by the normal distribution $N(0, \Gamma^0)$ with $\text{diag}(\Gamma^0) = \mathbf{1}_p$, this model is also called Gaussian copula model or nonparanormal model, and Γ^0 is called *latent correlation matrix*.

Definition 3 (Margin-preserved Nonparanormal Distribution). A random vector $\mathbf{Y} = (Y_1, \cdots, Y_p)^\top \in \mathbb{R}^p$ with means $\mu = (\mu_1, \cdots, \mu_p)^\top$ and standard deviations $\{\sigma_1^{(Y)}, \cdots, \sigma_d^{(Y)}\}$ is said to follow a margin-preserved nonparanormal distribution $MNPN_p(\mu, \Gamma, f)$ if and only if there exists a set of strictly increasing univariate functions $f = \{f_j\}_{j=1}^p$ such that $f(\mathbf{Y}) = (f_1(Y_1), \cdots, f_p(Y_p))^\top \sim N_p(\mu, \Gamma)$, where $\text{diag}(\Gamma) = ((\sigma_1^{(Y)})^2, \cdots, (\sigma_p^{(Y)})^2)^\top \in \mathbb{R}^p$. We call Γ the latent covariance matrix.

In the above definitions 2-3, f is unspecified and is unknown in practice. Similar to latent correlation matrix, if one views $\{f_j\}_{j=1}^p$ as a kind of contamination, $\mathbf{Y}$ is the contaminated observation of some normal variable with covariance matrix Γ, which is the parameter of interest.

Definition 4 (Kendall's τ correlation) Let $\mathbf{Y} = (Y_1, \cdots, Y_p)^\top$ be a p-dimensional random vector. The Kendall's τ correlation coefficient between Y_i and Y_j is defined as

$$\tau(Y_i, Y_j) := P((Y_i - \tilde{Y}_i)(Y_j - \tilde{Y}_j) > 0) - P((Y_i - \tilde{Y}_i)(Y_j - \tilde{Y}_j) < 0)$$

where $\tilde{Y} = (\tilde{Y}_1, \cdots, \tilde{Y}_p)^\top$ is an independent copy of $\mathbf{Y}$. Denote $\text{corr}^{\mathcal{K}}(\mathbf{Y}) = (\tau(Y_i, Y_j)) \in \mathbb{R}^{p \times p}$ the Kendall's τ correlation matrix.

2.2 Estimate of latent correlation matrix for matrix-variate

In many applications, $\{X_k \in \mathbb{R}^{p \times q}, 1 \le k \le n\}$ are contaminated or not Gaussian. We study in the frame of transelliptical distribution. Assume that $\text{vec}(X_k)$ follows the transelliptical distribution $TE_{pq}(\mathbf{R}, \xi; f)$, where $f = (f_{11}, \cdots, f_{pq})$. Or equivalently, the uncontaminated variables of vector

$$f(X_k) = (f_{11}(X_{11,k}), \cdots, f_{p1}(X_{p1,k}), \cdots, f_{1q}(X_{11,k}), \cdots, f_{pq}(X_{pq,k})) \in \mathbb{R}^{pq}$$

follow an elliptical distribution with Pearson correlation matrix $\mathbf{R}$, i.e.

$$\mathbf{R} = \text{corr}(\text{vec}(f(X_k))) = (\mathbf{R}_{i,j}) \in \mathbb{R}^{pq \times pq}.$$

For $(i_s, j_s) \in [p] \times [q], s = 1, 2$, it is easy to see that

$$\mathbf{R}_{(j_1-1)p+i_1,(j_2-1)p+i_2} = \mathrm{corr}(f_{i_1j_1}(X_{i_1j_1,k}), f_{i_2j_2}(X_{i_2j_2,k})).$$

The main idea of our robust estimate of the latent covariance matrix $\mathbf{\Sigma}$ comes from the observation that $\mathbf{\Sigma} = D\mathbf{R}D$, where $D = (\mathrm{diag}(\mathbf{\Sigma}))^{1/2}$ is the diagonal matrix of standard deviation, and $\mathbf{R}$ is the correlation matrix. Naturally, the robust estimate of $\mathbf{\Sigma}$ can be constructed by combining the robust estimate of $\mathbf{R}$ and D, respectively.

To obtain the estimate of the correlation matrix $\mathbf{R}$, we consider here the Kendall's τ correlation, which is a robust measure for the relation between two variables. Recall the definition 4 on Kendall's τ correlation matrix. We denote the Kendall's τ correlation matrix as

$$\mathbf{T} = \mathrm{corr}^{\mathcal{K}}(\mathrm{vec}(X_k)) = (\mathbf{T}_{i,j}) \in \mathbb{R}^{pq \times pq}.$$

For $(i_s, j_s) \in [p] \times [q], s = 1, 2$, $\mathbf{T}_{(j_1-1)p+i_1,(j_2-1)p+i_2}$ stands for the Kendall's τ correlation coefficient between variables $f_{i_1j_1}(X_{i_1j_1,k})$ and $f_{i_2j_2}(X_{i_2j_2,k})$. The relationship between $\mathbf{T}_{(j_1-1)p+i_1,(j_2-1)p+i_2}$ and $\mathbf{R}_{(j_1-1)p+i_1,(j_2-1)p+i_2}$ is shown as follows (Han and Liu (2014))

$$\mathbf{R}_{(j_1-1)p+i_1,(j_2-1)p+i_2} = \sin\left(\frac{\pi}{2}\mathbf{T}_{(j_1-1)p+i_1,(j_2-1)p+i_2}\right).$$

This motivates us to construct a robust estimate of $\mathbf{R}$, denoted as $\mathbf{R}^\tau \in \mathbb{R}^{pq \times pq}$, based on estimate of $\mathbf{T}_{(j_1-1)p+i_1,(j_2-1)p+i_2}$. Similar to Han and Liu (2014), we estimate $\mathbf{T}_{(j_1-1)p+i_1,(j_2-1)p+i_2}$ by

$$\hat{\mathbf{T}}_{(j_1-1)p+i_1,(j_2-1)p+i_2} = \frac{2}{n(n-1)} \sum_{k_1<k_2} \mathrm{sign}(X_{i_1j_1,k_1} - X_{i_1j_1,k_2})\mathrm{sign}(X_{i_2j_2,k_1} - X_{i_2j_2,k_2}).$$

where $(i_s, j_s) \in [p] \times [q], s = 1, 2$. Then $\hat{\mathbf{T}} = (\hat{\mathbf{T}}_{i,j})$ is the estimate of $\mathbf{T}$. Combining together, we estimate $\mathbf{R}$ by

$$\hat{\mathbf{R}}^\tau = \left(\sin(\frac{\pi}{2}\hat{\mathbf{T}}_{i,j})\right) = \sin\left(\frac{\pi}{2}\hat{\mathbf{T}}\right). \tag{2.1}$$

2.3 Estimate of latent covariance matrix of matrix-variate

Similar to Tsiligkaridis and Hero (2013), we assume that the latent covariance matrix $\mathbf{\Sigma}$ has the Kronecker structure (1.1). That is, $\mathbf{\Sigma} = \sum_{i=1}^{r} A_i \otimes B_i$, where A_i's are $q \times q$

linearly independent matrices and B_i's are $p \times p$ linearly independent matrices, and $r \leq \min\{p^2, q^2\}$. We consider the estimate of covariance matrix under the nonparanormal distribution, that is, $\text{vec}(X_k) \sim MNPN_{pq}(\mu, \mathbf{\Sigma}, f)$, where $f = (f_{11}, \cdots, f_{pq})$. Equivalently, vector $f(X_k) \sim N(\mu, \mathbf{\Sigma})$, where $f(X_k) = (f_{ij}(X_{ij,k}), 1 \leq i \leq p, 1 \leq j \leq q) \in \mathbb{R}^{pq}$ and $\text{var}(f_{ij}(X_{ij,k})) = \text{var}(X_{ij,k})$ for any $1 \leq i \leq p, 1 \leq j \leq q$.

Note that $MNPN_{pq}(\mu, \mathbf{\Sigma}; f)$ is a special case of transelliptical distribution. The main reason for a stronger assumption is that standard deviation is generally not invariant under the increasing function f.

Now we turn to the robust estimate of D. Clearly, matrix D can be estimated by the robust estimate of the standard deviation of each element of X_k. Since D is a diagonal matrix, we denote diagonal elements as vector $D^{(d)} = (\sigma_{11}, \cdots, \sigma_{1q}, \cdots, \sigma_{p1}, \cdots, \sigma_{pq})^\top$. Let $\xi_{ij,0.5}$ denote the 0.5 quantile of the distribution of $X_{ij,k}$, for $(i, j) \in [p] \times [q]$. A natural robust estimate for σ_{ij} is the median absolute deviation (MAD) type estimate $\hat{\sigma}_{ij}$ defined as

$$\hat{\sigma}_{ij} = c_{ij} \cdot \text{median}\{|X_{ij,k} - X_{ij}^{med}|, k = 1, \cdots, n\}, \tag{2.2}$$

where $X_{ij}^{med} = \text{median}\{X_{ij,k}, k = 1, \cdots, n\}$ and c_{ij}^{-1} equals the 0.5 quantile of the distribution of $|X_{ij,k} - \xi_{ij,0.5}|/\sigma_{ij}$, which can be written as the function of standardized variable $X_{ij,k}^{(sv)} = (X_{ij,k} - E(X_{ij,k}))/\sigma_{ij}$. When the distribution of $X_{ij,k}^{(sv)}$ is known, c_{ij} can be calculated directly. For example, when $X_{ij,k}$ is normal, we have $c_{ij} = \sqrt{1/\chi^2_{0.5}(1)}$, where $\chi^2_{0.5}(1)$ is the 0.5 quantile of χ^2 distribution with degree of freedom 1. We show later that the estimate $\hat{\sigma}_{ij}$ is consistent uniformly over $(i, j) \in [p] \times [q]$, under mild assumption on the densities of the marginal distributions.

Remark 1. *Here we aim at giving a robust estimate of variance of σ_{ij}. In practice, when the distribution of $X_{ij,k}^{(sv)}$ is unknown for some index (i, j), c_{ij} will be unknown and MAD type estimate cannot be used. In this case, many other robust estimators can be used. Catoni (2012) proposed a robust estimator of variance that allows for heavy-tailed distributions with a bounded kurtosis. Suppose that $\{Z_k, 1 \leq k \leq n\}$ are the i.i.d. copy of some random vector $Z = (z_1, \cdots, z_p)^\top \in \mathbb{R}^p$ with covariance matrix $\check{\Sigma} = (\check{\sigma}_{ij})$. Assuming that the maximum of fourth moment $\max_{1 \leq i \leq p} E(z_i^4)$ exists, Fan et al (2017) proposed a robust approximate (RA) quadratic loss function and showed that the corresponding estimator $\hat{\check{\sigma}}_{ij}^{RA}$ has good convergence rate. Specifically,*

$P(\max_{1\le i,j\le p}|\hat{\check{\sigma}}^{RA}_{1\le i,j\le p}-\check{\sigma}_{ij}|\ge 4v\sqrt{a(\log p)/n})\le 2p^{2-a}$, *where* v *is a constant and* $a>2$ *(Fan et al. (2017)). Obviously, by assuming* $\max_{1\le i,j\le p}E(X_{ij}^4)$ *and replacing* p *by* p^2*, the estimator of Fan et al. (2017) can be applied to our setting to construct the robust estimator* $\hat{\sigma}_{ij}^{RA}$.

Denote the estimate of $D^{(d)}$ as $\hat{D}^{(d)}=(\hat{\sigma}_{ij},1\le i\le p,1\le j\le q)$. Moreover, $\mathbf{R}$ can be estimated as shown in the previous section. Combining these together, we have the robust estimate of $\mathbf{\Sigma}$ as

$$\hat{\mathbf{\Sigma}}^{\tau}=\hat{D}\hat{\mathbf{R}}^{\tau}\hat{D}. \tag{2.3}$$

To simplify the following argument, we first introduce the transformation operator $\mathcal{T}(\cdot)$. For any $\mathbf{N}\in\mathbb{R}^{pq\times pq}$, split $\mathbf{N}$ into blocks of sub-matrices of size $p\times p$ with q blocks each row and q^2 blocks in total. Denote $\mathbf{N}=(N(i,j))_{i,j=1}^{q}$, with $N(i,j)\in\mathbb{R}^{p\times p}$ being the block in ith row and jth column. Define the permutation operator $\mathcal{T}:\mathbb{R}^{pq\times pq}\to\mathbb{R}^{q^2\times p^2}$ by setting the $((i-1)q+j)$-th row of $\mathcal{T}(\mathbf{N})$ equal to $\mathrm{vec}(N(i,j))^{\top}\in\mathbb{R}^{p^2}$, details on this transformation referred to Tsiligkaridis and Hero (2013). Moreover, we define $\mathcal{T}^{-1}:\mathbb{R}^{q^2\times p^2}\mapsto\mathbb{R}^{pq\times pq}$ being the inverse operator of $\mathcal{T}(\cdot)$. Based on the definition of $\mathcal{T}(\cdot)$, we have

$$\mathcal{T}(\mathbf{\Sigma})=\sum_{i=1}^{r}\mathrm{vec}(A_i^{\top})(\mathrm{vec}(B_i))^{\top}.$$

Since A_i's and B_i's are linearly independent respectively, the above equation implies that the matrix $\mathcal{T}(\mathbf{\Sigma})$ has rank r. When r is small, $\mathcal{T}(\mathbf{\Sigma})$ has low rank structure.

Note that we do not require that $p=q$. Although $\mathbf{\Sigma}$ is positive definite, $\mathcal{T}(\mathbf{\Sigma})$ may not be semi-positive definite even if $p=q$. To see this, we consider a simple case where $r=1$ and $p=q$. For any matrix $\tilde{M}=c\cdot uv^{\top}$, where u and v are the p^2-dimensional vectors with $\|u\|=\|v\|=1$ and $c>0$ being a constant, one can show that $\tilde{M}$ is semi-positive definite if and only if $u=v$. In fact, when $u\neq v$, for any vector $w\in\mathbb{R}^{p^2}$ such that $w^{\top}(u+v)/2=0$, we see that $w^{\top}u=w^{\top}(u-v)/2=-w^{\top}(v-u)/2=-w^{\top}v$. Consequently, $w^T\tilde{M}w<0$. In addition, when $u=v$, it is easy to see that $\tilde{M}$ is semi-positive definite. For matrix $\mathcal{T}(\mathbf{\Sigma})$ considered here, $A_i^{\top}$ is generally not equal to B_i. Therefore, $\mathcal{T}(\mathbf{\Sigma})$ may not be semi-positive definite generally.

To take the Kronecker structure into account, we consider the following optimization problem,

$$\hat{\mathbf{\Sigma}}^{\tau}_{\mathcal{T}}=\arg\min_{S\in\mathbb{R}^{q^2\times p^2}}\|\mathcal{T}(\hat{\mathbf{\Sigma}}^{\tau})-S\|_F^2+\lambda\|S\|_* \tag{2.4}$$

where λ is the tuning parameter, which leads to the optimal solution

$$\hat{\mathbf{\Sigma}}^{\tau}_{\mathcal{T}} = \sum_{i=1}^{\min\{p^2,q^2\}} \left(\hat{\varphi}_i(\mathcal{T}(\hat{\mathbf{\Sigma}}^{\tau})) - \frac{\lambda}{2}\right)_+ \mathbf{u}_i\mathbf{v}_i^{\top}, \tag{2.5}$$

where $\hat{\varphi}_i(\mathcal{T}(\hat{\mathbf{\Sigma}}^{\tau}))$ stands for the i-th largest singular value of $\mathcal{T}(\hat{\mathbf{\Sigma}}^{\tau})$, and $\mathbf{u}_i$ and $\mathbf{v}_i$ are the corresponding left and right eigenvectors, respectively. Then the estimate of $\mathbf{\Sigma}$ can be defined as

$$\hat{\mathbf{\Sigma}}^{\tau}_{LR} = \mathcal{T}^{-1}(\hat{\mathbf{\Sigma}}^{\tau}_{\mathcal{T}}),$$

where LR in the subscript of the estimator indicates that the low rank Kronecker structured in (1.1) has been taken into account. The tuning parameter λ can be selected by cross validation (CV) method of Bickel and Levina (2009).

Finally, we note that $\mathbf{R}$ itself can be of interest in many applications. The above procedure can be used to estimate $\mathbf{R}$, when the latent correlation matrix $\mathbf{R}$ is assumed to have the Kronecker form as (1.1). By replacing $\hat{\mathbf{\Sigma}}^{\tau}$ in (2.4) and (2.5) by $\hat{\mathbf{R}}^{\tau}$ and denoting the optimal solution of (2.5) as $\hat{\mathbf{R}}^{\tau}_{\mathcal{T}}$, we get the estimate of $\mathbf{R}$, denoted as $\hat{\mathbf{R}}^{\tau}_{LR} = \mathcal{T}^{-1}(\hat{\mathbf{R}}^{\tau}_{\mathcal{T}})$.

2.4 The special case of $r = 1$

We consider the special case of $r = 1$ in $\mathbf{\Sigma}$ or $\mathbf{R}$. Consider $\mathbf{\Sigma}$ first. Note that $\mathbf{\Sigma} = A \otimes B$, when $r = 1$. This special case has been studied by Leng and Tang (2012), Leng and Pan (2017) and many others. Leng and Pan (2017) considered the estimate when A and B are sparse. Here, we focus on the semiparametric estimation and will not impose the sparsity assumption further. Clearly, one can extend the idea of sparsity assumption into our setting. For identification, we rewrite the mode as

$$\mathbf{\Sigma} = \gamma \cdot A \otimes B, \tag{2.6}$$

where $\gamma = \|\mathbf{\Sigma}\|_F$, $A = (a_{ij}) \in R^{q\times q}$ and $B = (b_{ij}) \in R^{p\times p}$ with $\|A\|_F = \|B\|_F = 1$. Let $V_A = \mathrm{vec}(A^{\top}), V_B = \mathrm{vec}(B)$. Then $\mathcal{T}(\mathbf{\Sigma}) = \gamma V_A V_B^{\top}$, according to the definition of $\mathcal{T}(\cdot)$. Recall that $\hat{\mathbf{\Sigma}}^{\tau}$ is the robust estimate of $\mathbf{\Sigma}$ obtained in (2.3). We estimate (γ, V_A, V_B) by minimizing the following objective function

$$(\hat{\gamma}, \hat{V}_A, \hat{V}_B) = \arg\min_{\Theta} \|\mathcal{T}(\hat{\mathbf{\Sigma}}^{\tau}) - dv_1v_2^{\top}\|_F^2,$$

where $\Theta = \{(d, v_1, v_2) : d \in \mathbb{R}, v_1 \in \mathbb{R}^{q^2}, v_2 \in \mathbb{R}^{p^2}, d > 0, \|v_1\| = \|v_2\| = 1\}$. Obviously, the estimator $\hat{\gamma}$ is the largest singular value of the SVD decomposition of $\mathcal{T}(\hat{\mathbf{\Sigma}}^{\tau})$, and $(\hat{V}_A, \hat{V}_B)$ are the associated left and right eigenvectors, respectively. Consequently, $\mathbf{\Sigma}$ in (2.6) can be estimated by

$$\hat{\mathbf{\Sigma}}^{\tau}_{(rk=1)} = \mathcal{T}^{-1}(\hat{\gamma}\hat{V}_A\hat{V}_B^{\top}). \tag{2.7}$$

Let $\hat{A} \in \mathbb{R}^{q\times q}$ and $\hat{B} \in \mathbb{R}^{p\times p}$ are the matrices such that $\text{vec}(\hat{A}^{\top}) = \hat{V}_A$ and $\text{vec}(\hat{B}) = \hat{V}_B$. Then $\hat{A}$ and $\hat{B}$ are estimates of A and B, respectively. In the next section, we establish the asymptotic results of $\hat{A}, \hat{B}$ and $\hat{\mathbf{\Sigma}}^{\tau}_{(rk=1)}$.

Similar to $\mathbf{\Sigma}$, when $\mathbf{R}$ has the Kronecker structure with $r = 1$, it can be estimated in the same way. Similar to (2.6), denote

$$\mathbf{R} = \tilde{\gamma} \cdot \tilde{A} \otimes \tilde{B} \tag{2.8}$$

where $\tilde{\gamma} = \|\mathbf{R}\|_F$, $\tilde{A} = (\tilde{a}_{ij}) \in \mathbb{R}^{q\times q}$ and $\tilde{B} = (\tilde{b}_{ij}) \in \mathbb{R}^{p\times p}$ with $\|\tilde{A}\|_F = \|\tilde{B}\|_F = 1$. Recall the definition of $\hat{\mathbf{R}}^{\tau}$ in (2.1). Just replacing $\hat{\mathbf{\Sigma}}^{\tau}$ by $\hat{\mathbf{R}}^{\tau}$ in the above procedure, one can get the estimate of $\mathbf{R}$ in (2.8). Define $V_{\tilde{A}}$ and $V_{\tilde{B}}$ in the same way as those of V_A and V_B. Then $\mathbf{R}$ can be estimated by

$$\hat{\mathbf{R}}^{\tau}_{(rk=1)} = \mathcal{T}^{-1}(\hat{\tilde{\gamma}}\hat{V}_{\tilde{A}}\hat{V}_{\tilde{B}}^{\top}), \tag{2.9}$$

where $\hat{\tilde{\gamma}}$ is the largest singular value of SVD decomposition of $\mathcal{T}(\hat{\mathbf{R}}^{\tau})$, and $(\hat{V}_{\tilde{A}}, \hat{V}_{\tilde{B}})$ are the associated left and right eigenvectors, respectively. In addition, let $\hat{\tilde{A}} \in \mathbb{R}^{q\times q}$ and $\hat{\tilde{B}} \in \mathbb{R}^{p\times p}$ such that $\text{vec}(\hat{\tilde{A}}^{\top}) = \hat{V}_{\tilde{A}}$ and $\text{vec}(\hat{\tilde{B}}) = \hat{V}_{\tilde{B}}$. Then $\hat{\tilde{A}}$ and $\hat{\tilde{B}}$ are the estimate of $\tilde{A}$ and $\tilde{B}$, respectively.

3 The asymptotic properties of the estimates

3.1 Asymptotic properties of $\|\mathcal{T}(\hat{\mathbf{R}}^{\tau}) - \mathcal{T}(\mathbf{R})\|_{op}$

To establish the bound of the estimator, we need to establish the upper bound of the term $\|\mathcal{T}(\hat{\mathbf{R}}^{\tau}) - \mathcal{T}(\mathbf{R})\|_{op}$. A related quantity is $\|\hat{\mathbf{R}}^{\tau} - \mathbf{R}\|_{op}$, of which the associated upper bound has been studied by Han and Liu (2017), Mitra and Zhang (2014) and Wegkamp et al. (2016). However, $\|\mathcal{T}(\hat{\mathbf{R}}^{\tau}) - \mathcal{T}(\mathbf{R})\|_{op}$ is quite different from $\|\hat{\mathbf{R}}^{\tau} - \mathbf{R}\|_{op}$, since the former is not a symmetric matrix any more and the matrix concentration inequality

used by Han and Liu (2017) and Wegkamp et al. (2016) is not applicable. In fact, the theoretical analyses of Han and Liu (2017) and Wegkamp et al. (2016) rely on the matrix concentration inequality of Tropp (2012), where the candidate must be square and symmetric.

Remark 2. *Tropp (2012) considered the finite sequence $\{W_k\}$ of random, self-adjoint matrices with dimension d. Based on the matrix Laplacian transformation, Tropp (2012) derived bound on the probability*

$$P\left(\lambda_{\max}\left(\sum_k W_k\right) \geq t\right),$$

where $\lambda_{\max}(\cdot)$ denotes the algebraically largest eigenvalue of a self-adjoint matrix. The following Proposition 3.1 of Tropp (2012) on the Laplace Transformation Method plays a critical role to derive the matrix concentration inequality.

Proposition 3.1 *(Tropp, 2012) Let W be a random self-adjoint matrix. For all $t \in \mathbb{R}$,*

$$P(\lambda_{\max}(W) \geq t) \leq \inf_{\theta>0}\{e^{-\theta t}E(\mathrm{tr}(e^{\theta W}))\}.$$

Details are referred to Tropp (2012). This inequality is the key step in the proof of Han and Liu (2017).

To simplify the argument, we introduce some notations first. Let $Z_k = \mathrm{vec}(X_k) \in \mathbb{R}^{pq}$, $U_{kk'} = \mathrm{sign}(Z_k - Z_{k'}) \in \mathbb{R}^{pq}, 1 \leq k \neq k' \leq n$, and $V_{kk'} = \mathrm{vec}(U_{kk'}U_{kk'}^\top - E(U_{kk'}U_{kk'}^\top))$. Recall that $\mathbf{T} = E(U_{kk'}U_{kk'}^\top)$ and $\hat{\mathbf{T}} = 2(n(n-1))^{-1}\sum_{1\leq k\neq k'\leq n} U_{kk'}^\top U_{kk'}$ are the population Kendall's τ correlation matrix and its estimate, respectively. According to the definition of $\hat{\mathbf{R}}^\tau$, we have $\hat{\mathbf{R}}^\tau = \sin(\frac{\pi}{2}\hat{\mathbf{T}})$. Let

$$E_n := \mathrm{vec}(\hat{\mathbf{T}} - \mathbf{T}) = 2(n(n-1))^{-1}\sum_{1\leq k\neq k'\leq n} V_{kk'},$$

which is a U-statistic. Note that $\mathbf{a}^\top E_n$ is also a U-statistic. Based on the asymptotic normality of U-statistics, under some conditions, $\sqrt{n}\mathbf{a}^\top E_n$ will converge in distribution to $N(0, \mathbf{a}^\top\mathbf{W}\mathbf{a})$, for any $\mathbf{a} \in \mathbb{R}^{p^2q^2}$ with $\|\mathbf{a}\| = 1$, where $\mathbf{W} = \mathrm{cov}(V_{kk'})$. By the tail probability of U-statistic (Keener et al., 1998; Borovskikh and Weber, 2003), as n is large, the tail probability of $\sqrt{n}\mathbf{a}^\top E_n$ will be similar to that of normal $N(0, \mathbf{a}^\top\mathbf{W}\mathbf{a})$ under some

conditions. Assume that $\|\mathbf{W}\|_{op} < C < \infty$ for some constant $C > 0$ and for any positive integers p and q. Then $\mathbf{a}^\top \mathbf{W}\mathbf{a}$ is upper bounded by a constant. To simplify the proof, we make the following high level assumption that $\sqrt{n}\mathbf{a}^\top E_n$ has the tail probability similar to that of sub-Gausssian variable.

(A1) (Tail probability) Assume that $\|\mathbf{W}\|_{op} < C < \infty$ for any positive integers p and q. And assume that, as n is large, $P(\sqrt{n}\mathbf{a}^\top E_n > t) \le C' \exp(-t^2/K^2)$ for any positive $t > 0$ with $t = o(n^{1/2}\log n)$ and any $\mathbf{a} \in \mathbb{R}^{p^2q^2}$ with $\|\mathbf{a}\| = 1$, where $0 < C, C', K < \infty$ are constants.

Han and Liu (2017) showed that if $Z_k \sim TE_{pq}(I_{pq}, \xi; f_1, \cdots, f_{pq})$, then Z_k satisfies the sign sub-Gaussian condition, that is, for any unit vector v, $E(\exp(t\langle V_{kk'}, vv^\top\rangle)) \le e^{ct^2}$ for any $0 < t < t_0$ for some t_0. Barber and Kolar (2015) proved that if $Z_k \sim N(0, \Sigma)$, then $\text{sign}(Z_k)$ is sub-Gaussian and similar inequality holds. In our setting, we need to consider $\mathbf{a}^T V_{kk'}$. Although the results of Han and Liu (2017) and Barber and Kolar (2015) may not be applied directly, it is expected that similar inequality still holds under some conditions.

Proposition 1. *Suppose that* $E(\exp(t\mathbf{a}^\top V_{kk'})) \le e^{ct^2}$ *for any* $0 < t < t_0$, *where* $t_0 > 0$ *and* $c > 0$ *are constants. Then Assumption (A1) holds.*

Theorem 1. *Assume that* $\text{vec}(X_k)$ *follows transelliptical distribution, denoted by* $\text{vec}(X_k) \sim TE_p(\mathbf{R}, \xi;\ f_{11}, \cdots, f_{pq})$. *Under assumption (A1), we have*

$$\|\mathcal{T}(\hat{\mathbf{R}}^\tau) - \mathcal{T}(\mathbf{R})\|_{op} = O_p\left(\sqrt{\frac{p^2 + q^2 + \log n}{n}} + \frac{pq\log(pq)}{n}\right).$$

3.2 Asymptotic properties of $\hat{\mathbf{\Sigma}}^\tau_{LR}$ and $\hat{\mathbf{R}}^\tau_{LR}$

To show the convergence of $\hat{\mathbf{\Sigma}}^\tau_{LR}$, we establish the rate of

$$\|\hat{D} - D\|_{\max} = \sup_{(i,j)\in[p]\times[q]} |\hat{\sigma}_{ij} - \sigma_{ij}|.$$

Let $X^{(0)}_{ij,k} = c_{ij}|X_{ij,k} - \xi_{ij,0.5}|$. One should be aware that $X^{(0)}_{ij,k}$ is different from $X^{(sv)}_{ij,k}$ in Section 2. Remind the fact in (2.2) that c_{ij}^{-1} is the 0.5 quantile of the distribution of $|X_{ij,k} - \xi_{ij,0.5}|/\sigma_{ij}$. It is obvious that $c_{ij} > 0$ and that

$$P(X^{(0)}_{ij,k} \ge \sigma_{ij}) = P(|X_{ij,k} - \xi_{ij,0.5}|/\sigma_{ij} \ge c_{ij}^{-1}) \ge 1/2,$$

$$P(X_{ij,k}^{(0)} \le \sigma_{ij}) = P(|X_{ij,k} - \xi_{ij,0.5}|/\sigma_{ij} \le c_{ij}^{-1}) \ge 1/2.$$

The above inequalities imply that σ_{ij} is the 0.5 quantile of $X_{ij,k}^{(0)}$.

Suppose that variables $X_{ij,k}$ and $X_{ij,k}^{(0)}$ have the densities, denoted by $f_{ij}(x)$ and $g_{ij}(x)$, respectively. Recall that $\xi_{ij,0.5}$ and σ_{ij} are the 0.5 quantile of the distribution of $X_{ij,k}$ and $X_{ij,k}^{(0)}$, respectively, for $(i,j) \in [p] \times [q]$. We make the following assumptions.

(A2) Assume that $\min_{ij}\{f_{ij}(\xi_{ij,0.5}) \wedge g_{ij}(\sigma_{ij})\} > c_0 > 0$ for some constant c_0, where $a \wedge b = \min\{a, b\}$.

Lemma 1. *Assume that (A2) holds and that* $n^{-1}\log(\max\{p,q\}) \to 0$. *Then we have*

$$\|\hat{D} - D\|_{\max} = O_p\left(\sqrt{\frac{\log(\max(p,q))}{n}}\right).$$

Let $\omega_n^{(1)} = \sqrt{(p^2+q^2+\log n)/n} + n^{-1}pq\log(pq)$. Based on Lemma 1 and Theorem 1, we get the following convergence rate of $\hat{\mathbf{\Sigma}}_{LR}^{\tau}$ in the following Theorem 2. To simplify the notations, we denote $\omega_n^{(2)} = (n^{-1}\log(\max(p,q))^{1/2}$ and $\omega_n^{(0)} = \omega_n^{(1)}\|D\|_{\max}^2 + \omega_n^{(2)}\|\mathcal{T}(\mathbf{R})\|_{op}\|D\|_{\max}$.

Theorem 2. *Suppose that* $\mathrm{vec}(X_k)$ *follows margin-preserved nonparanormal distribution in Definition 3. Under (A1) and (A2), as* $\lambda > C\omega_n^{(0)}$ *for some constant* $C > 0$, *we have with probability tending to 1,*

$$\|\hat{\mathbf{\Sigma}}_{LR}^{\tau} - \mathbf{\Sigma}\|_F^2 \le \inf_{\substack{G \in \mathbb{R}^{q^2 \times p^2} \\ rank(G) \le r}} \|G - \mathcal{T}(\mathbf{\Sigma})\|_F^2 + r \cdot (\omega_n^{(0)})^2.$$

Note that $\|\mathcal{T}(\mathbf{R})\|_F = \|\mathbf{R}\|_F$ and that $\mathcal{T}(\mathbf{R})$ is a matrix of dimension $q^2 \times p^2$. It is easy to see that $\|\mathbf{R}\|_F/\min\{p,q\} \le \|\mathcal{T}(\mathbf{R})\|_{op} \le \|\mathbf{R}\|_F$. In addition, since $\sqrt{pq} \le \|\mathbf{R}\|_F \le pq$, $\|\mathcal{T}(\mathbf{R})\|_{op}$ can be as small as $\sqrt{pq}/\min\{p,q\}$, which is of order $O(1)$, if $p \asymp q$. If $\|\mathcal{T}(\mathbf{R})\|_{op}$ is small such that $\omega_n^{(2)}\|\mathcal{T}(\mathbf{R})\|_{op} \le \omega_n^{(1)}$, then from Theorem 2, we get $r(\omega_n^{(0)})^2 = O\left(r \cdot [n^{-1}(p^2+q^2+\log n) + (n^{-1}pq\log(pq))^2]\right)$. Then the inequality in Theorem 2 can be written as

$$\|\hat{\mathbf{\Sigma}}_{LR}^{\tau} - \mathbf{\Sigma}\|_F^2 \le \inf_{\substack{G \in \mathbb{R}^{q^2 \times p^2} \\ rank(G) \le r}} \|G - \mathcal{T}(\mathbf{\Sigma})\|_F^2 + Cr \cdot \left[\frac{p^2+q^2+\log n}{n} + \left(\frac{pq\log(pq)}{n}\right)^2\right].$$

for some constant $C > 0$. When $\mathbf{\Sigma}$ has the the low rank structure as in (1.1), then the

first term will be zero and we get the convergence rate

$$\|\hat{\mathbf{\Sigma}}^{\tau}_{LR} - \mathbf{\Sigma}\|_F^2 = O_p\left(r \cdot \left[\frac{p^2+q^2+\log n}{n} + \left(\frac{pq\log(pq)}{n}\right)^2\right]\right).$$

On the other hand, when the normal distribution is assumed, Tsiligkaridis and Hero (2013) obtained the convergence rate of the order

$$r \cdot \max\left\{\frac{p^2+q^2+\log M_0}{n}, \left(\frac{p^2+q^2+\log M_0}{n}\right)^2\right\},$$

where $M_0 = \max\{n, p, q\}$. Therefore, the convergence rate of the robust estimator is comparable to that of normal distribution, although the semiparametric estimate relaxes the assumption of normal distribution greatly.

Finally, when latent correlation matrix $\mathbf{R}$ is of interest, we have the following conclusion on the estimator $\hat{\mathbf{R}}^{\tau}_{LR}$ obtained in the last paragraph of Section 2.3. The proof is similar to Step 2 of the proof of Theorem 2, and we omit it here.

Proposition 2. *Under the assumption of Theorem 1, taking $\lambda > C\omega_n^{(1)}$ for some constant $C > 0$, we have, with probability tending to 1,*

$$\|\hat{\mathbf{R}}^{\tau}_{LR} - \mathbf{R}\|_F^2 \leq \inf_{\substack{G\in\mathbb{R}^{q^2\times p^2}\\ rank(G)\leq r}} \|G - \mathcal{T}(\mathbf{R})\|_F^2 + r\cdot(\omega_n^{(1)})^2.$$

3.3 The asymptotic results for the case of $r = 1$

Now we consider the asymptotic behavior of the estimator $\hat{\mathbf{\Sigma}}^{\tau}_{(rk=1)}$ and $\hat{\mathbf{R}}^{\tau}_{(rk=1)}$ in Section 2.4 for the case of $r = 1$. We first consider $\hat{\mathbf{\Sigma}}^{\tau}_{(rk=1)}$. Recall that $\mathbf{\Sigma} = \gamma \cdot A \otimes B$, where $A \in \mathbb{R}^{q\times q}, B \in \mathbb{R}^{p\times p}$ with $\|A\|_F = \|B\|_F = 1$ and $\gamma = \|\mathbf{\Sigma}\|_F = \|\mathcal{T}(\mathbf{\Sigma})\|_{op}$. Recall that $\mathrm{vec}(\hat{A}^{\top}) = \hat{V}_A$, $\mathrm{vec}(A^{\top}) = V_A$, $\mathrm{vec}(\hat{B}) = \hat{V}_B$ and $\mathrm{vec}(B) = V_B$. We have the following conclusions.

Theorem 3. *Recall $\mathbf{\Sigma}$ in (2.6) and the estimate $\hat{\mathbf{\Sigma}}^{\tau}_{(rk=1)}$ in (2.7). Under the assumptions of Theorem 2, we have*

$$\|\hat{A} - cA\|_F = O_p(\omega_n^{(0)}/\|\mathbf{\Sigma}\|_F), \qquad \|\hat{B} - c'B\|_F = O_p(\omega_n^{(0)}/\|\mathbf{\Sigma}\|_F),$$

where c, c' take value of 1 or -1, such that $c\hat{V}_A^{\top}V_A \geq 0$ and $c'\hat{V}_B^{\top}V_B \geq 0$. In addition, we

have

$$\|\hat{\mathbf{\Sigma}}^{\tau}_{(rk=1)} - \mathbf{\Sigma}\|_F^2 = O_p\left((\omega_n^{(0)})^2\right).$$

Similar to Theorem 3, we have the following conclusion for $\hat{\mathbf{R}}^{\tau}_{(rk=1)}$, which is the estimator of $\mathbf{R}$ in (2.8). Its proof is similar to that of Theorem 3, and we omit it here.

Proposition 3. *Recall* $\mathbf{R}$ *in (2.8) and the estimate* $\hat{\mathbf{R}}^{\tau}_{(rk=1)}$ *in (2.9). Under the assumptions of Theorem 1, we have*

$$\|\hat{\tilde{A}} - c\tilde{A}\|_F = O_p(\omega_n^{(1)}/\|\mathbf{R}\|_F), \qquad \|\hat{\tilde{B}} - c'\tilde{B}\|_F = O_p(\omega_n^{(1)}/\|\mathbf{R}\|_F),$$

where c, c' *take value of* 1 *or* -1*, such that* $c\hat{V}_{\tilde{A}}^{\top}V_{\tilde{A}} \geq 0$ *and* $c'\hat{V}_{\tilde{B}}^{\top}V_{\tilde{B}} \geq 0$*. In addition, we have*

$$\|\hat{\mathbf{R}}^{\tau}_{(rk=1)} - \mathbf{R}\|_F^2 = O_p\left((\omega_n^{(1)})^2\right).$$

We discuss the above results briefly. Let $X = H^{\top}YL \in \mathbb{R}^{p\times q}$, where $Y \in \mathbb{R}^{s_1\times s_2}$ with $\mathrm{cov}(\mathrm{vec}(Y)) = \mathbf{I}_{s_1s_2}$, $H \in \mathbb{R}^{s_1\times p}$ and $L \in \mathbb{R}^{s_2\times q}$ are such that each column of H and L has unit ℓ_2 norm. Then

$$\mathrm{cov}(\mathrm{vec}(X)) = \mathrm{corr}(\mathrm{vec}(X)) = L^{\top}L \otimes H^{\top}H := A \otimes B,$$

where $A = L^{\top}L$ and $B = H^{\top}H$. Therefore, $\mathbf{R} = \mathbf{\Sigma}$ and consequently $\|D\|_{\max} = 1$. Let $A = (a_{ij}), B = (b_{ij})$, $N_A = \{(i,j) : 0 < C^{-1} < |a_{ij}| \leq C < \infty, 1 \leq i,j \leq q\}$ and $N_B = \{(i,j) : 0 < C^{-1} < |b_{ij}| \leq C < \infty, 1 \leq i,j \leq p\}$ for some constant C large enough. We consider the following two cases: (i) A, B are dense such that $|N_A| \asymp q^2$ and $|N_B| \asymp p^2$; (ii) A, B are sparse such that $|N_A| \asymp q$ and $|N_B| \asymp p$. Then, $\|\mathbf{R}\|_F = \|\mathbf{\Sigma}\|_F = O(pq)$ for Case (i), and $\|\mathbf{R}\|_F = \|\mathbf{\Sigma}\|_F = O(\sqrt{pq})$ for Case (ii).

Consider $\mathbf{\Sigma}$ first. Remind that $\|\mathbf{\Sigma}\|_F = \|\mathcal{T}(\mathbf{\Sigma})\|_{op}$. Then $\omega_n^{(0)}/\|\mathbf{\Sigma}\|_F = \omega_n^{(1)}/\|\mathbf{\Sigma}\|_F + \omega_n^{(2)}$. Then for Case (i), we have

$$\omega_n^{(1)}/\|\mathbf{\Sigma}\|_F \asymp \omega_n^{(1)}/(pq) = \sqrt{\frac{p^2+q^2+\log n}{np^2q^2}} + \frac{\log(pq)}{n}.$$

By Theorem 3 and the definition of $\omega_n^{(2)}$, we get for Case (i)

$$\|\hat{A} - cA\|_F = \|\hat{B} - c'B\|_F = O_p(\omega_n^{(2)}) = O_p\left(\sqrt{\frac{\log(\max\{p,q\})}{n}}\right).$$

Therefore, we can handle the case of p, q being exponential order of n. And for Case (ii), by similar argument, it holds that

$$\omega_n^{(1)}/\|\mathbf{\Sigma}\|_F \asymp \omega_n^{(1)}/\sqrt{pq} = \sqrt{\frac{p^2+q^2+\log n}{n}} + \frac{\sqrt{pq}\log(pq)}{n}.$$

Suppose that q, p are diverging. By Theorem 3, we get the following bound for Case (ii)

$$\|\hat{A} - cA\|_F = \|\hat{B} - c'B\|_F = O_p\left(\frac{\sqrt{pq}\log(pq)}{n} + \omega_n^{(2)}\right).$$

Note that the convergence rate for Case (ii) can be worse than that of Case (i), as p, q are large. The main reason is that we do not impose the sparsity assumption here. Note that the error term $\omega_n^{(2)}$ comes from estimating D and $\omega_n^{(1)}$ from estimating $\mathbf{R}$. For the sparse case, without the sparsity assumption, the estimate on $\mathbf{R}$ is less efficient, which makes it possible for the estimation error of $\mathbf{R}$ being larger than that of D. Moreover, we have the relative error for Case (i)

$$\frac{1}{\|\mathbf{\Sigma}\|_F^2}\|\hat{\mathbf{\Sigma}}^{\tau}_{(rk=1)} - \mathbf{\Sigma}\|_F^2 = O_p\left(p^{-2}q^{-2}(\omega_n^{(1)})^2 + (\omega_n^{(2)})^2\right) = O_p((\omega_n^{(2)})^2) = O_p\left(\frac{\log(\max\{p,q\})}{n}\right).$$

Similarly, we get the relative error for Case (ii)

$$\frac{1}{\|\mathbf{\Sigma}\|_F^2}\|\hat{\mathbf{\Sigma}}^{\tau}_{(rk=1)} - \mathbf{\Sigma}\|_F^2 = O_p((pq)^{-1}(\omega_n^{(1)})^2 + (\omega_n^{(2)})^2) = O_p\left(\frac{pq\log^2(pq)}{n^2} + (\omega_n^{(2)})^2\right).$$

Now we discuss the $\mathbf{R}$, where the error involves only $\omega_n^{(1)}$. Similar to the discussion above, we have for Case (i)

$$\|\hat{\tilde{A}} - c\tilde{A}\|_F = \|\hat{\tilde{B}} - c'B\|_F = O_p(n^{-1}\log(pq)),$$

which is better than Case (i) of $\mathbf{\Sigma}$. And for Case (ii), we have

$$\|\hat{\tilde{A}} - c\tilde{A}\|_F = \|\hat{\tilde{B}} - c'\tilde{B}\|_F = O_p\left(\frac{\sqrt{pq}\log(pq)}{n}\right),$$

which has the same rate as that of Case (ii) of $\mathbf{\Sigma}$. Similarly, we have the relative error for Case (i)

$$\frac{1}{\|\mathbf{R}\|_F^2}\|\hat{\mathbf{R}}^{\tau}_{(rk=1)} - \mathbf{R}\|_F^2 = O_p\left(p^{-2}q^{-2}(\omega_n^{(1)})^2\right) = O_p\left(\frac{\log^2(pq)}{n^2}\right),$$

and the relative error for Case (ii)

$$\frac{1}{\|\mathbf{R}\|_F^2}\|\hat{\mathbf{R}}^{\tau}_{(rk=1)} - \mathbf{R}\|_F^2 = O_p((pq)^{-1}(\omega_n^{(1)})^2) = O_p\left(\frac{pq\log^2(pq)}{n^2}\right).$$

From the above discussion, we see that the relative error for Case (i) is much better than that of Case (ii). The efficiency may be improved if one utilizes the penalized method with ℓ_1 penalty to encourage the sparsity.

4 Simulation and real data analysis

4.1 Simulation setup

In this section, we compare our method with the PRLS estimator of Tsiligkaridis and Hero (2013) which is the low rank approximation of sample covariance matrix and is non-robust. We generate $X_k \in \mathbb{R}^{p\times q}$ i.i.d. according to the following models. For simplicity, we set $p \le q$. Let

$$X_k = \sum_{i=1}^{r} H_i^{\top} Y_{ki} L_i, \quad k = 1, \cdots, n,$$

where $H_i \in \mathbb{R}^{s_1\times p}, L_i \in \mathbb{R}^{s_2\times q}, 1 \le i \le r$ are constant matrices, and $Y_{ki} \in \mathbb{R}^{s_1\times s_2}, i = 1, \cdots, r$ are independent random matrices with $\mathrm{cov}(\mathrm{vec}(Y_{ki})) = c_i \cdot I_{s_1 s_2}$ for some constant $c_i > 0$. The distribution of Y_{ki} will be specified later. For this model, it is easy to see that

$$\mathbf{\Sigma} = \mathrm{cov}(\mathrm{vec}(X_k)) = \sum_{i=1}^{r} c_i^2 (L_i^{\top} \otimes H_i^{\top})(L_i \otimes H_i) = \sum_{i=1}^{r} c_i^2 L_i^{\top} L_i \otimes H_i H_i^{\top}$$

Therefore, $\mathbf{\Sigma}$ has the structure of (1.1).

Example 1. Set $r = 1$ and $Y_{k1} \in \mathbb{R}^{s_1\times s_2}$ with $\mathrm{vec}(Y_{k1}) \sim N(0, I_{s_1 s_2})$, where $H_1 = (h_{ij}) \in \mathbb{R}^{s_1\times p}$ and $L_1 = (l_{ij}) \in \mathbb{R}^{s_2\times q}$ with $h_{ij} = 0.5^{|i-j|}$ and $l_{ij} = 0.2^{|i-j|}$. We replace the first observation X_1 by the contaminated one $\tilde{X}_1 = \delta_0 \mathbf{I} + X_1$, where $\mathbf{I} \in \mathbb{R}^{p\times q} = (\mathbf{I}_p, \mathbf{0}_{p\times(q-p)})$. We take $\delta_0 = 0, 10, 20, 50$. Clearly, when $\delta_0 = 0$, there are no outliers.

Example 2. Set $r = 1$ and $Y_{k1} \in \mathbb{R}^{s_1\times s_2}$, where the elements of Y_{k1}' are i.i.d. variables with distribution $t(3)$, t distribution with degree of freedom 3. H_1 and L_1 are the same as Example 1. Obviously, we have $\mathrm{cov}(\mathrm{vec}(Y_{k1})) = 3I_{s_1 s_2}$.

Example 3. Set $r = 3$. Y_{k1}, H_1 and L_1 are the same as Example 1. Y_{k2}, Y_{k3} are i.i.d. copies of Y_{k1}, $1 \le k \le n$. $H_2 = (h'_{ij}) \in \mathbb{R}^{s_1\times p}$ with $h'_{ij} = 0.4^{|i-j|}$, $L_2 = (l'_{ij}) \in \mathbb{R}^{s_2\times p}$ with

$l'_{ij} = 0.3^{|i-j|}$, $H_3 = (h''_{ij}) \in \mathbb{R}^{s_1 \times p}$ with $h''_{ij} = 0.1^{|i-j|}$ and $L_3 = (l''_{ij}) \in \mathbb{R}^{s_2 \times p}$ with $l''_{ij} = 0.1^{|i-j|}$. Similar to Example 1, we replace the first observation X_1 by the contaminated one $\tilde{X}_1$ defined in Example 1.

Example 4. Set $r = 3$. $Y_{k1}, 1 \le k \le n$ and H_1, L_1 are the same as Example 2. For $1 \le k \le n$, Y_{k2}, Y_{k3} are i.i.d. copies of Y_{k1}, and H_2, H_3, L_2, L_3 are the same as Example 3.

We consider two cases: (s_1, s_2) equals to (p, q) and $(\lceil p/4 \rceil, \lceil q/4 \rceil)$, respectively, where $\lceil a \rceil$ denotes the largest integer no more than a for any constant $a \in \mathbb{R}$. In Example 1 and 2, we assume that $r = 1$ is known. The robust estimator $\hat{\mathbf{\Sigma}}^{\tau}_{(rk=1)}$ is obtained by Section 2.4. And the non-robust estimate, denoted as $\hat{\mathbf{\Sigma}}^{sam}_{(rk=1)}$, is the rank one Kronecker approximation of the sample covariance matrix. For Example 3 and 4, the robust estimator $\hat{\mathbf{\Sigma}}^{\tau}_{LR}$ is obtained by the approach in Section 2.3 and the non-robust estimator is the PRLS estimator of Tsiligkaridis and Hero (2013), denoted as $\hat{\mathbf{\Sigma}}^{prls}$. The $\hat{r}$ is determined by tuning parameter λ, which is selected by CV method of Bickel and Levina (2009).

Since the non-robust estimators, i.e. $\hat{\mathbf{\Sigma}}^{prls}$ and $\hat{\mathbf{\Sigma}}^{sam}_{(rk=1)}$, are derived from sample covariance matrix, to simplify the description, $\hat{\mathbf{\Sigma}}^{prls}$ and $\hat{\mathbf{\Sigma}}^{sam}_{(rk=1)}$ will be denote as $\hat{\mathbf{\Sigma}}^{sam}$. And the robust estimators ($\hat{\mathbf{\Sigma}}^{\tau}_{(rk=1)}$ and $\hat{\mathbf{\Sigma}}^{\tau}_{LR}$) will be denoted as $\hat{\mathbf{\Sigma}}^{rob}$. Let $Err^{(rob)} = \hat{\mathbf{\Sigma}}^{rob} - \mathbf{\Sigma}$ and $Err^{(sam)} = \hat{\mathbf{\Sigma}}^{sam} - \mathbf{\Sigma}$. We compute the average of $\|Err^{(rob)}\|_F$, $\|Err^{(rob)}\|_{op}$ and $\|Err^{(rob)}\|_\infty$ over 100 replicas, respectively, denoted as $Err^{(rob)}_F, Err^{(rob)}_2, Err^{(rob)}_\infty$. Similarly, we compute those of $Err^{(sam)}$ and define $Err^{(sam)}_F, Err^{(sam)}_2, Err^{(sam)}_\infty$ in the same way.

4.2 Simulation Results

(1). *Simulation results on estimate error.* The simulation results on $(s_1, s_2) = (p, q)$ are presented in Table 1-2, and those on $(s_1, s_2) = (\lceil p/4 \rceil, \lceil q/4 \rceil)$ are presented in the Supplementary material, due to the limited space. For Example 1, we see from Table 1 that for $\delta_0 = 0$, the non-robust estimator has better performance than the robust estimation. However, as δ_0 increases, the performance of non-robust estimator deteriorates and the robust estimate becomes better. And for Example 2, as shown in Table 2, the robust estimate is slightly better than non-robust estimator. For Example 3, it can be inferred from Table 2 that the robust estimate is better than the non-robust estimator when δ_0 is large. And for Example 4, the robust estimate is much better than non-robust estimate.

Table 1: Simulation results for Example 1 and 2 with $s_1 = p, s_2 = q$

n, p, q		Example 1				Example 2
		$\delta_0 = 0$	$\delta_0 = 10$	$\delta_0 = 20$	$\delta_0 = 50$	
(100,15,15)	$Err_F^{(rob)}$	0.0237	0.0265	0.0304	0.0436	0.0026
	$Err_F^{(sam)}$	0.0098	0.0172	0.0324	0.1139	0.0029
	$Err_2^{(rob)}$	0.0194	0.0215	0.0244	0.0345	0.0021
	$Err_2^{(sam)}$	0.0076	0.0140	0.0254	0.0875	0.0023
	$Err_\infty^{(rob)}$	0.0046	0.0050	0.0055	0.0078	0.0005
	$Err_\infty^{(sam)}$	0.0017	0.0032	0.0058	0.0189	0.0008
(100,25,25)	$Err_F^{(rob)}$	0.0156	0.0165	0.0175	0.0221	0.0008
	$Err_F^{(sam)}$	0.0073	0.0094	0.0141	0.0419	0.0008
	$Err_2^{(rob)}$	0.0114	0.0117	0.0123	0.0151	0.0006
	$Err_2^{(sam)}$	0.0048	0.0067	0.0099	0.0276	0.0005
	$Err_\infty^{(rob)}$	0.0018	0.0019	0.0020	0.0025	0.0001
	$Err_\infty^{(sam)}$	0.0007	0.0010	0.0016	0.0045	0.0001

Table 2: Simulation results for Example 3 and 4 with $s_1 = p, s_2 = q$

n, p, q		Example 3				Example 4
		$\delta_0 = 0$	$\delta_0 = 10$	$\delta_0 = 20$	$\delta_0 = 50$	
(100,15,15)	$Err_F^{(rob)}$	0.3143	0.3151	0.3174	0.3210	0.6955
	$Err_F^{(sam)}$	0.2329	0.2516	0.2958	0.7404	2.1451
	$Err_2^{(rob)}$	0.1153	0.1159	0.1198	0.1255	0.2469
	$Err_2^{(sam)}$	0.0991	0.1099	0.1304	0.3306	1.7175
	$Err_\infty^{(rob)}$	0.0116	0.0116	0.0124	0.0223	0.0436
	$Err_\infty^{(sam)}$	0.0190	0.0216	0.0425	0.1753	0.6533
(100,25,25)	$Err_F^{(rob)}$	0.2238	0.2238	0.2254	0.2273	0.5238
	$Err_F^{(sam)}$	0.1235	0.1247	0.1476	0.3444	2.3158
	$Err_2^{(rob)}$	0.0647	0.0648	0.0652	0.0663	0.1465
	$Err_2^{(sam)}$	0.0368	0.0375	0.0446	0.1164	1.0147
	$Err_\infty^{(rob)}$	0.0042	0.0043	0.0047	0.0081	0.0198
	$Err_\infty^{(sam)}$	0.0066	0.0072	0.0142	0.0616	0.4079

Moreover, we compare the two settings: $(s_1, s_2) = (p, q)$ and $(s_1, s_2) = (\lceil p/4 \rceil, \lceil q/4 \rceil)$. For Example 1 and 2, comparing Table 1 with Table S1 in the Supplementary material, we note that each estimator has similar performance under two different values of (s_1, s_2). However, for Example 3 and 4, comparing Table 2 with Table S2 in the Supplementary material, we see that there exist significant differences in the performance of each estimator under two different values of (s_1, s_2).

(2). *Simulation results on the selection of rank.* In our simulations, when r is unknown, the tuning parameter λ in our method is selected according to Bickel and Levina (2009). To check the effectiveness of the method of Bickel and Levina (2009), we report in Table

3–4 the simulation results on the rank selection for Example 3 and 4, where the true rank is 3.

Set $(n, p, q, s_1, s_2) = (100, 15, 15, 15, 15)$ in Example 3 and Example 4. We report in Table 3 the numbers of the event $\{\hat{r} = i\}$ with $i = 1, 2, 3$, and the event $\{\hat{r} > 3\}$ over 200 replicas, respectively. From Table 3, we see that method of Bickel and Levina (2009) works well and the estimated $\hat{r}$ is robust to outliers. For example, the empirical probability of event $\{\hat{r} = 3\}$ is 80% in 200 replicas in most of the cases. In addition, we also report in Table 4 $Err_F^{(rob)}, Err_2^{(rob)}, Err_\infty^{(rob)}$ with fixed rank=1,2,3, respectively. Obviously, from Table 4, we see that rank=3 leads to the best results. In addition, we see that the $Err_F^{(rob)}$ is affected most by the selection of the rank, while the $Err_\infty^{(rob)}$ is less affected by the selection of the rank.

Table 3: Rank estimation for Example 3 and 4, where the true rank is 3 and $(n, p, q, s_1, s_2) = (100, 15, 15, 15, 15)$. We account the numbers of the estimated rank $\hat{r}$ equal to $1, 2, 3$ and $\hat{r} > 3$ over 200 replicas. Rank can be correctly selected in most cases.

	Example 3				Example 4
	$\delta_0 = 0$	$\delta_0 = 10$	$\delta_0 = 20$	$\delta_0 = 50$	
$\hat{r} = 1$	5	6	8	10	3
$\hat{r} = 2$	25	27	33	35	30
$\hat{r} = 3$	168	165	158	155	167
$\hat{r} > 3$	2	2	1	0	0

Table 4: Estimation error with fixed rank for Example 3 and 4, where the true rank is 3 and $(n, p, q, s_1, s_2) = (100, 15, 15, 15, 15)$.

rank		Example 3				Example 4
		$\delta_0 = 0$	$\delta_0 = 10$	$\delta_0 = 20$	$\delta_0 = 50$	
rank=1	$Err_F^{(rob)}$	0.3282	0.3582	0.3672	0.3884	0.7950
	$Err_2^{(rob)}$	0.1302	0.1392	0.1430	0.1461	0.3219
	$Err_\infty^{(rob)}$	0.0212	0.0221	0.0233	0.0379	0.0786
rank=2	$Err_F^{(rob)}$	0.2885	0.2919	0.2990	0.3106	0.7351
	$Err_2^{(rob)}$	0.1211	0.1222	0.1241	0.1286	0.2745
	$Err_\infty^{(rob)}$	0.0114	0.0117	0.0137	0.0273	0.0579
rank=3	$Err_F^{(rob)}$	0.2586	0.2607	0.2667	0.2983	0.6844
	$Err_2^{(rob)}$	0.1112	0.1127	0.1131	0.1256	0.2138
	$Err_\infty^{(rob)}$	0.0109	0.0113	0.0134	0.0235	0.0391

4.3 Real Data Analysis

We apply our method to Atlas of Gene Expression in the Mouse Aging (AGEMAP) database, which is a resource of gene expression as a function of age in mice, including

expression changes for 8,932 genes in 16 tissues as a function of age (Zahn et al., 2007). There are 4 age states: 1, 6, 16 and 24 months. For each age state, researchers chose ten mice with five for each gender, and consequently there are 40 observations in total. Similar to Leng and Pan (2017) and Yin and Li (2012), we select seven tissues: Cerebrum, Hippocampus, Kidney, Lung, Muscle, Thymus and Spinal cord (i.e. $q = 7$), and examine genes related to mitogen activated protein kinase signaling pathway, long term potentiation, insulin signaling pathway and vascular endothelial growth factor signaling pathway, as documented on $http://rgd.mcw.edu/rgdweb/pathway/pathwayRecord.html?accid = PW:0000243\&species = Mouse\sharp Pathway$. We apply our method to male and female, respectively, where the sample size is $n = 20$ for each gender. According to Yin and Li (2012), there are 70 genes that are closely related to ageing. Since the sample size is small, we choose the first 30 genes of the largest variance among these 70 genes for analysis. Therefore, for each gender, we have $(n, p, q) = (20, 30, 7)$.

We compare three different estimators: (1) the robust estimate of our proposal; (2) PRLS estimator of Tsiligkaridis and Hero (2013), which is non-robust; (3) the estimator of Leng and Pan (2017), where it is assume that $\Sigma = A \otimes B$, that is, $r = 1$ in (1.1). For our proposal and PRLS estimator, r is estimated by data rather than fixed, with the estimator denoted as $\hat{r}$. Heat maps of the covariance matrices for both male and female obtained by three estimators are presented respectively in Figure 1–3, which are available in the Supplementary materials. In all these figures, the diagonal blocks from lower left corner to upper right corner are associated with seven tissues, respectively.

According to Yin and Li (2012), genes associated with aging have dependency not only inside same tissue but also across different tissues. From the Figure 1, one observes weak dependency between Hippocampus and Thymus in male, and clear dependency between Cerebrum and Thymus in female. These observations coincide with those of Yin and Li (2012), where the authors found that gene expressions in Thymus are related with those in Hippocampus, Cerebrum, Spinal cord, Lung and Kidney. Moreover, Lustig et al. (2007) indicated that some genes chosen from Thymus express differently between male mice and female ones, and that the patterns on dependency among the tissues are different for male and female. These coincide with our results in plot (a) and (b) in Figure 1.

From the Figure 2, PRLS estimator (Tsiligkaridis and Hero, 2013) also reveals clear dependency between Thymus and Lung. On the other hand, it also shows the dependency

between Thymus and Muscle for male, which is not supported by the results of Yin and Li (2012). In addition, the estimator of Leng and Pan (2017) in Figure 3 and PLRS estimator (Tsiligkaridis and Hero, 2013) show almost no or very weak correlations among different tissues for female, which is inconsistent with the findings of Yin and Li (2012).

5 Discussion

In this paper, we propose a method of covariance matrix estimation for high dimensional matrix-variate in the frame of Transelliptical distribution, taking into account Kronecker structure of the covariance matrix. Recall that $\mathbf{T} = (\mathbf{T}_{i,j})$ stands for Kendall's correlation matrix with estimate $\hat{\mathbf{T}} = (\hat{\mathbf{T}}_{i,j})$, and $\mathbf{R} = (\mathbf{R}_{i,j})$ for the Pearson correlation matrix with robust estimate $\hat{\mathbf{R}}^{\tau} = (\hat{\mathbf{R}}^{\tau}_{i,j}) = (\sin(\frac{\pi}{2}\hat{\mathbf{T}}_{i,j}))$. Denote $\hat{\mathbf{R}}^{sam} = (\hat{\mathbf{R}}^{sam}_{i,j})$ the sample correlation matrix. When X_k follows normal distribution, the sample Pearson's correlation is asymptotically unbiased and reaches the Cramér-Rao lower bound as sample size tends to infinity (Xu et al., 2013). Hence, $\hat{\mathbf{R}}^{sam}_{i,j}$ is generally more efficient than $\hat{\mathbf{R}}^{\tau}_{i,j}$, when X_k is normal.

Consider bivariate normal with correlation coefficient ρ. Let $\hat{\mathbf{T}}_{\rho}$ stand for the sample version of Kendall's τ correlation. Let $\hat{\rho}_{\mathcal{K}}$ be the robust estimator of ρ constructed from Kendall's τ correlation as above, i.e. $\hat{\rho}_{\mathcal{K}} = \sin(\frac{\pi}{2}\hat{\mathbf{T}}_{\rho})$, and let $\hat{\rho}_{\mathcal{P}}$ denote sample Pearson correlation. According to Xu et al. (2013), estimator $\hat{\rho}_{\mathcal{K}}$ has the variance $\mathrm{Var}(\hat{\rho}_{\mathcal{K}}) \approx [\pi^2(4-\rho^2)/36]\mathrm{Var}(\hat{\mathbf{T}}_{\rho})$ with

$$\mathrm{Var}(\hat{\mathbf{T}}_{\rho}) = \frac{2}{n(n-1)}\left[1 - \frac{4S_1^2}{\pi^2} + 2(n-2)\left(\frac{1}{9} - \frac{4S_2^2}{\pi^2}\right)\right],$$

where $S_1 = \sin^{-1}(\rho)$ and $S_2 = \sin^{-1}(\rho/2)$. Moreover, Xu et al. (2013) considered the asymptotic relative efficiency, defined as

$$\mathrm{ARE}^{\mathcal{K}}(\rho) \triangleq \lim_{n\to\infty} \frac{\mathrm{Var}(\hat{\rho}_{\mathcal{P}})}{\mathrm{Var}(\hat{\rho}_{\mathcal{K}})} = \frac{9(1-\rho^2)}{\pi^2 - 36(\sin^{-1}\frac{1}{2}\rho)^2},$$

Particularly, for $\rho = 0$, $\mathrm{ARE}^{\mathcal{K}}(0) = 9/\pi^2$ and for $\rho \to \pm 1$,

$$\mathrm{ARE}^{\mathcal{K}}|_{\rho\to\pm1} = \frac{1}{4}\frac{\rho\sqrt{4-\rho^2}}{\sin^{-1}\frac{1}{2}\rho}\bigg|_{\rho\to\pm1} = \frac{3\sqrt{3}}{2\pi} \approx 0.8270.$$

This can be viewed as a price paid for the use of robust estimate.

6 Appendix

The proof of Theorem 2 is presented as follows. The proofs of other theorems and lemmas are available in the Supplementary materials.

6.1 Proof of Theorem 2

Step 1. We first establish the bound $\|\mathcal{T}(\hat{\mathbf{\Sigma}}^{\tau}) - \mathcal{T}(\mathbf{\Sigma})\|_{op} = O_p(\omega_n^{(0)})$.

Recalling that $\hat{\mathbf{\Sigma}}^{\tau} = \hat{D}\hat{\mathbf{R}}^{\tau}\hat{D}$, we have

$$\begin{aligned}\|\mathcal{T}(\hat{\mathbf{\Sigma}}^{\tau}) - \mathcal{T}(\mathbf{\Sigma})\|_{op} &= \|\mathcal{T}(\hat{D}\hat{\mathbf{R}}^{\tau}\hat{D}) - \mathcal{T}(D\mathbf{R}D)\|_{op} \\ &\leq \|\mathcal{T}(\hat{D}\hat{\mathbf{R}}^{\tau}\hat{D}) - \mathcal{T}(\hat{D}\mathbf{R}\hat{D})\|_{op} + \|\mathcal{T}(\hat{D}\mathbf{R}\hat{D}) - \mathcal{T}(D\mathbf{R}D)\|_{op} \\ &:= \|J_1\|_{op} + \|J_2\|_{op}, \end{aligned} \tag{6.1}$$

where J_1 and J_2 are defined accordingly.

Consider $\|J_1\|_{op}$ first. Letting $\Delta = \{(u,v) : u \in \mathbb{S}^{q^2-1}, v \in \mathbb{S}^{p^2-1}\}$ and $\mathbf{F} = \hat{\mathbf{R}}^{\tau} - \mathbf{R}$, we have

$$\|J_1\|_{op} = \sup_{(u,v)\in\Delta} u^{\top} J_1 v = \sup_{(u,v)\in\Delta} u^{\top}\mathcal{T}(\hat{D}\mathbf{F}\hat{D})v.$$

Recall that $D^{(d)}$ is the vector of diagonal elements of D. Then $\hat{D}\mathbf{F}\hat{D} = \mathbf{F} \circ \hat{D}^{(d)}\hat{D}^{(d)\top}$ and $\mathcal{T}(\hat{D}\mathbf{F}\hat{D}) = \mathcal{T}(\mathbf{F}) \circ \mathcal{T}(\hat{D}^{(d)}\hat{D}^{(d)\top})$. Therefore,

$$u^{\top}\mathcal{T}(\hat{D}\mathbf{F}\hat{D})v = \langle \mathcal{T}(\mathbf{F}), \mathcal{T}(\hat{D}^{(d)}\hat{D}^{(d)\top}) \circ (u \otimes v^{\top})\rangle \leq \|\hat{D}\|_{\max}^2 \langle \mathcal{T}(\mathbf{F}), u \otimes v^{\top}\rangle = \|\hat{D}\|_{\max}^2 u^{\top}\mathcal{T}(\mathbf{F})v.$$

In addition, by Lemma 1, we have $\|\hat{D}\|_{\max}^2 = \|D\|_{\max}^2 + O_p\left(\omega_n^{(2)}\right)$. Therefore, combining with Theorem 1, we have

$$\|J_1\|_{op} \leq \|\hat{D}\|_{\max}^2 \|\mathcal{T}(\mathbf{F})\|_{op} = \|\mathcal{T}(\hat{\mathbf{R}}^{\tau}) - \mathcal{T}(\mathbf{R})\|_{op}(\|D\|_{\max}^2 + O_p(\omega_n^{(2)})) = O_p(\omega_n^{(1)}\|D\|_{\max}^2).$$

Now we consider $\|J_2\|_{op}$. It is easy to see that

$$\begin{aligned} J_2 &= \mathcal{T}(\hat{D}\mathbf{R}\hat{D}) - \mathcal{T}(D\mathbf{R}D) \\ &= \mathcal{T}((\hat{D} - D)\mathbf{R}D) + \mathcal{T}(D\mathbf{R}(\hat{D} - D)) + \mathcal{T}((\hat{D} - D)\mathbf{R}(\hat{D} - D)) \\ &:= J_{21} + J_{22} + J_{23}. \end{aligned}$$

Applying the same argument as that of J_1 on each term in J_2, we get

$$\|J_2\|_{op} \leq 2\|\mathcal{T}(\mathbf{R})\|_{op}\|D\|_{\max}\|\hat{D} - D\|_{\max} + \|\mathcal{T}(\mathbf{R})\|_{op}\|\hat{D} - D\|_{\max}^2.$$

Combining together and noting the definition of $\omega_n^{(0)}$, we have

$$\|\mathcal{T}(\hat{\mathbf{\Sigma}}^\tau) - \mathcal{T}(\mathbf{\Sigma})\|_{op} = O_p(\omega_n^{(1)}\|D\|_{\max}^2 + \omega_n^{(2)}\|\mathcal{T}(\mathbf{R})\|_{op}\|D\|_{\max}) = O_p(\omega_n^{(0)}).$$

Step 2. We prove the final conclusion. By the proof similar to that of the Theorem 2 of Tsiligkaridis and Hero (2013), as $\lambda > 2\|\mathcal{T}(\hat{\mathbf{\Sigma}}^\tau) - \mathcal{T}(\mathbf{\Sigma})\|_{op}$, we have

$$\|\hat{\mathbf{\Sigma}}_{\mathcal{T}}^\tau - \mathcal{T}(\mathbf{\Sigma})\|_F^2 \leq \inf_{G \in \mathbb{R}^{q^2 \times p^2}} \left\{ \|G - \mathcal{T}(\mathbf{\Sigma})\|_F^2 + \frac{(1+\sqrt{2})^2}{4}\lambda^2 \text{rank}(G) \right\}$$

Combining with results in Step 1 on the operator norm of $\|\mathcal{T}(\hat{\mathbf{\Sigma}}^\tau) - \mathcal{T}(\mathbf{\Sigma})\|_{op}$, it follows that, as $\lambda > C\omega_n^{(0)}$ for some constant $C > 0$, with probability tending to 1,

$$\|\hat{\mathbf{\Sigma}}_{\mathcal{T}}^\tau - \mathcal{T}(\mathbf{\Sigma})\|_F^2 \leq \inf_{\substack{G \in \mathbb{R}^{q^2 \times p^2} \\ rank(G) \leq r}} \|G - \mathcal{T}(\mathbf{\Sigma})\|_F^2 + r(\omega_n^{(0)})^2.$$

The finally conclusion is derived by noting that $\|\hat{\mathbf{\Sigma}}_{LR}^\tau - \mathbf{\Sigma}\|_F^2 = \|\hat{\mathbf{\Sigma}}_{\mathcal{T}}^\tau - \mathcal{T}(\mathbf{\Sigma})\|_F^2$. ■

References

Barber, R. F., Kolar, M., 2015. Rocket: Robust confidence intervals via kendall's tau for transelliptical graphical models. arXiv preprint arXiv:1502.07641.

Bickel, P. J., Levina, E., 2008. Regularized estimation of large covariance matrices. Annals of Statistics 36 (1), 199–227.

Bickel, P. J., Levina, E., 2009. Covariance regularization by thresholding. Annals of Statistics 36 (6), 2577–2604.

Borovskikh, Y. V., Weber, N. C., 2003. Large deviations of U-statistics. i. Lithuanian Mathematical Journal 43 (43), 11–33.

Catoni, O., 2012. Challenging the empirical mean and empirical variance: a deviation study. Annales De L Institut Henri Poincar Probabilits Et Statistiques 48 (4), 1148–1185.

Dutilleul, P., 1999. The mle algorithm for the matrix normal distribution. Journal of Statistical Computation & Simulation 64 (2), 105–123.

Engle, R. F., Ng, V., Rothschild, M., 2010. Asset pricing with a factor arch covariance structure: empirical estimates for treasury bills. National Bureau of Economic Research, 213–237.

Fan, J., Li, Q., Wang, Y., 2017. Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. Journal of the Royal Statistical Society B 79 (1), 247–265.

Fan, J., Liao, Y., Liu, H., 2015. An overview on the estimation of large covariance and precision matrices. Proceedings of the IEEE International Conference on Micro Electro Mechanical Systems, 415–418.

Fang, H. B., Fang, K. T., Kotz, S., 2002. The meta-elliptical distributions with given marginals. Journal of Multivariate Analysis 94 (1), 1–16.

Greenewald, K., Hero, A. O., 2014a. Regularized block toeplitz covariance matrix estimation via kronecker product expansions. In: Statistical Signal Processing (SSP), 2014 IEEE Workshop on. IEEE, pp. 9–12.

Greenewald, K., Hero, A. O., 2014b. Robust kronecker product PCA for spatio-temporal covariance estimation. IEEE Transactions on Signal Processing 63 (23), 6368–6378.

Greenewald, K. H., Hero, A. O., 2014c. Kronecker PCA based spatio-temporal modeling of video for dismount classification. arXiv preprint arXiv:1405.4574.

Gupta, A. K., Nagar, D. K., 1999. Matrix variate distributions. Vol. 104. CRC Press.

Han, F., Liu, H., 2014. Scale-invariant sparse PCA on high dimensional meta-elliptical data. Journal of the American Statistical Association 109 (505), 275–287.

Han, F., Liu, H., 2017. Statistical analysis of latent generalized correlation matrix estimation in transelliptical distribution. Bernoulli 23 (1), 23–57.

Johnson, C. C., Jalali, A., Ravikumar, P., 2011. High-dimensional sparse inverse covariance estimation using greedy methods. Computer Science 96 (3), 497–512.

Jones, D. T., Buchan, D. W., Cozzetto, D., Pontil, M., 2012. Psicov: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 28 (2), 184.

Karceski, J., Lakonishok, J., 1999. On portfolio optimization: Forecasting covariances and choosing the risk model. Review of Financial Studies 12 (5), 937–974.

Keener, R. W., Robinson, J., Weber, N. C., 1998. Tail probability approximations for U-statistics. Statistics & Probability Letters 37 (1), 59–65.

Lam, C., Fan, J., 2009. Sparsistency and rates of convergence in large covariance matrix estimation. Annals of Statistics 37 (6B), 4254–4278.

Ledoit, O., Wolf, M., 2001. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. Journal of Empirical Finance 10 (5), 603–621.

Leng, C., Pan, G., 2017. Covariance estimation via sparse kronecker structure. Bernoulli In press.

Leng, C., Tang, C. Y., 2012. Sparse matrix graphical models. Journal of the American Statistical Association 107 (499), 1187–1200.

Liu, H., Han, F., Yuan, M., Lafferty, J., Wasserman, L., 2012. High dimensional semiparametric gaussian copula graphical models. Annals of Statistics 40 (4), 2293–2326.

Liu, H., Lafferty, J., Wasserman, L., 2009. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. Journal of Machine Learning Research 10 (3), 2295–2328.

Loan, C. V., Pitsianis, N., 1992. Approximation with kronecker products. Cornell University, 293–314.

Lustig, A., Weeraratna, A. T., Wood, W. W., Teichberg, D., Bertak, D., Carter, A., Poosala, S., Firman, J., Becker, K. G., Zonderman, A. B., et al., 2007. Transcriptome analysis of age-, gender-and diet-associated changes in murine thymus. Cellular immunology 245 (1), 42–61.

Mitra, R., Zhang, C. H., 2014. Multivariate analysis of nonparametric estimates of large correlation matrices. arXiv preprint arXiv:1403.6195.

Rigollet, P., Tsybakov, A., 2012. Estimation of covariance matrices under sparsity constraints. arXiv preprint arXiv:1205.1210.

Sejnowski, T., Makeig, S., Delorme, A., 2007. Enhanced detection of artifacts in eeg data using higher-order statistics and independent component analysis. Neuroimage 34 (34), 1443–1449.

Tropp, J. A., 2012. User-friendly tail bounds for sums of random matrices. Foundations of Computational Mathematics 12 (4), 389–434.

Tsiligkaridis, T., Hero, A. O., 2013. Covariance estimation in high dimensions via kronecker product expansions. Signal Processing IEEE Transactions on 61 (21), 5347–5360.

Tsiligkaridis, T., Iii, A. O. H., Zhou, S., 2012. Convergence properties of kronecker graphical lasso algorithms. IEEE Transactions on Signal Processing 61 (7), 1743–1755.

Wallbacks, L., Norden, U. E. B., 2006. Multivariate data analysis of in situ pulp kinetics using 13 c cp/mas nmr. Journal of Wood Chemistry & Technology 9 (2), 235–249.

Wegkamp, M., Zhao, Y., et al., 2016. Adaptive estimation of the copula correlation matrix for semiparametric elliptical copulas. Bernoulli 22 (2), 1184–1226.

Xu, W., Hou, Y., Hung, Y. S., Zou, Y., 2013. A comparative analysis of spearman's rho and kendall's tau in normal and contaminated normal models. Signal Processing 93 (1), 261–276.

Yin, J., Li, H., 2012. Model selection and estimation in the matrix normal graphical model. Journal of Multivariate Analysis 107 (3), 119–140.

Zahn, J. M., Poosala, S., Owen, A. B., Ingram, D. K., Lustig, A., Carter, A., Weeraratna, A. T., Taub, D. D., Gorospe, M., Mazanmamczarz, K., 2007. Agemap: A gene expression database for aging in mice. Plos Genetics 3 (11), e201.