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: Preprint
doi:10.5705/ss.202016.0364

Statistica Sinica

High-Dimensional Semiparametric Estimate of Latent Covariance Matrix for Matrix-variate

Lu Niu[1] and Junlong Zhao[2]

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

2. School of Statistics, Beijing Normal University, China

Abstract: Estimating the covariance matrix of a high-dimensional matrix-variate is an important issue. As such, many methods have been developed, typically based on the sample covariance matrix under a Gaussian or sub-Gaussian assumption. However, the sub-Gaussian assumption is restrictive and the estimate based on the sample covariance matrix is not robust. In this study, we estimate the covariance matrix of a high-dimensional matrix-variate using a transelliptical distribution and Kendall's τ correlation. Because the covariance matrix of a matrix-variate is commonly assumed to have a low-dimensional structure, we consider the structure of the Kronecker expansion. The asymptotic results of the estimator are established. Simulation results and a real-data analysis confirm the effectiveness of our method.

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

Corresponding author: Junlong Zhao. Email: zhaojunlong928@126.com

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

1. Introduction

Covariance matrices are widely used in statistical inferences, such as principal components analysis (PCA), as well as in test statistics in multivariate analyses. Thus, estimations of covariance matrices have attracted attention in diverse fields, including bioinformatics (Jones et al., 2012) and economics and financial time series analyses, such as portfolio selection (Ledoit and Wolf, 2001), risk management (Karceski and Lakonishok, 1999), and asset pricing (Engle et al., 2010), among others. Because the sample covariance matrix is singular when the dimension is larger than the sample size, the estimation problem is generally challenging, especially when the dimension is high. To estimate the covariance matrix efficiently, some low-dimensional structures are often assumed, such as sparsity or low rank. For vector-valued variates, many works have estimated sparse or low-rank covariance matrices (Johnson et al., 2011; Bickel and Levina, 2008, 2009; Lam and Fan, 2009; Rigollet and Tsybakov, 2012, etc.). For a detailed review on this topic, refer to Fan et al. (2015).

With the rapid development of new technology, in many applications, researchers often collect data for a matrix-variate $\{X_k \in \mathbb{R}^{p\times q}, 1 \leq k \leq n\}$, with $X_k = (X_{ij,k})_{1\leq i\leq p, 1\leq j\leq q} \in \mathbb{R}^{p\times q}$, such as nuclear magnetic res-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

onance(NMR) data (Wallbacks and Norden, 2006) and electroencephalograph (EEG) data (Sejnowski et al., 2007). Covariance matrix estimations for such data are important in applications. Most works focus on the case in which both p and q are fixed (Dutilleul, 1999; Gupta and Nagar, 1999). In recent years, researchers have begun focusing on the case in which p and q diverge under an additional low-dimensional structure, such as sparsity and the 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 denotes a $q \times q$ linearly independent matrix, B_i denotes a $p \times p$ linearly independent matrix, and $r \leq \min\{p^2, q^2\}$. Here, linear independence means that vectors $\{\mathrm{vec}(A_i), i = 1, \cdots, r\}$ are linearly independent, as are $\{\mathrm{vec}(B_i), i = 1, \cdots, r\}$. Because $\mathbf{\Sigma}$ is symmetric and positive semidefinite, equation (1.1) imposes certain restrictions on A_i and B_i. For example, when $r = 1$, A_i and B_i should be symmetric and positive semidefinite.

Model (1.1) with $r \geq 1$ has applications in various fields, including video

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

modeling and classification, network anomaly detection, and Magnetoencephalography(MEG)/EEG covariance modeling (Greenewald and Hero, 2014a) (Greenewald and Hero, 2014c; Tsiligkaridis and Hero, 2013). For example, Greenewald and Hero (2014b) analyzed a yeast metabolic cell cycle data set, where 9335 gene probes are sampled approximately every 24 minutes for a total of 36 time points. The data include about three different cell cycles. According to this study, the matrix B_i serves as a spatial factor describing the dependencies among the genes. Matricx A_i with dimension 36×36 serves as a temporal factor, describing the dependencies among different time points. Because spatial and temporal dependency patterns may vary between cell cycles, r represents the number of different dependence patterns. The value of r estimated by Greenewald and Hero (2014b) is three, which matches the number of 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. The 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 of model (1.1) is given by $\mathbf{\Sigma} = A \otimes B$ (i.e. $r = 1$). This has been widely considered in low-dimensional cases with a nor-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

mal matrix-variate (Dutilleul, 1999), and in high-dimensional cases with a Gaussian assumption on X_k and a 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. As argued by
Han and Liu (2014), this approach has several disadvantages. (i) These estimates are not robust to outliers or a heavy-tailed distribution. (ii) 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 a weak assumption on the distribution.

In the traditional case of the 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) have relaxed the sub-Gaussian assumption, proposing a transelliptical family of distributions. $\mathbf{Y}$ follows a transelliptical distribution if there exists a set of nonspecific strictly increasing functions $(f_1, \cdots, f_p)$, such that $(f_1(Y_1), \cdots, f_p(Y_p))$ follows an elliptical distribution with the location parameter zero and the scale parameter Γ^0, the diagonal elements of which are one. Γ^0 is called a *latent generalized correlation matrix* (Han and Liu, 2014). Moreover,

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

Liu et al. (2009, 2012) and Han and Liu (2014) introduced a *latent covariance matrix*, denoted as Γ, in a margin-preserved nonparanormal distribution. Note that the inverse function f_j^{-1} exists because 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 a 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 a robust estimate of Γ^0.

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

Our study makes two major contributions to the literature. 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 non-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

trivial. For vector-valued variables, some works estimate the correlation matrix based on a 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 proposed method, 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) Our theoretical analysis is quite different. Our proposed method involves a linear operator $\mathcal{T}$ (see Section 2.3 for details), and we need to examine the error of $\mathcal{T}(\hat{\mathbf{R}}^{\tau})$ rather than that of $\hat{\mathbf{R}}^{\tau}$, where $\hat{\mathbf{R}}^{\tau}$ denotes the estimate of the correlation matrix based on Kendall's τ correlation. The main challenge is that $\mathcal{T}(\hat{\mathbf{R}}^{\tau})$ is asymmetric, but the matrix concentration inequalities used to examine $\hat{\mathbf{R}}^{\tau}$ are not applicable here (Han and Liu, 2017; Wegkamp et al., 2016). Thus, we use a different approach to establish the convergence rate.

Second, we study the statistical properties of our estimator when r is known beforehand and $r = 1$. This case differs from that considered by Tsiligkaridis and Hero (2013). In particular, for fixed $r = 1$, we obtain estimates of A and B. The asymptotic results show that the estimator is effective, even when the dimensions p and q have an exponential order of sample size n when the matrices A and B are dense.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

Notation. 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 in 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 $|S|$ as the cardinality of S. In addition, 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. In addition, 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.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

2. High-dimensional latent covariance matrix estimation for matrix-variates

2.1 Brief review of concepts

We first review several concepts related to transelliptical distributions (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: $\mathbf{Y} \overset{d}{=} \mu + \xi \mathbf{A}\mathbf{U}$. 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 $\Gamma = \mathbf{A}\mathbf{A}^\top$, 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 exist 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 a latent generalized correlation matrix. In particular, if

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

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

In Definitions 2–3, f is unspecified and is unknown, in practice. Similarly to the latent correlation matrix, if we view $\{f_j\}_{j=1}^p$ 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. Kendall's τ correlation coefficient between Y_i and Y_j is defined as

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

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

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

2.2 Estimate of latent correlation matrix for matrix-variates

In many applications, $\{X_k \in \mathbb{R}^{p\times q}, 1 \le k \le n\}$ are contaminated or not Gaussian. We examine a transelliptical distribution. Assume that $\mathrm{vec}(X_k)$ 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 the Pearson correlation matrix $\mathbf{R}$; that is,

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

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 $D = (\mathrm{diag}(\mathbf{\Sigma}))^{1/2}$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

is the diagonal matrix of the standard deviation and $\mathbf{R}$ is the correlation matrix. Naturally, a robust estimate of $\mathbf{\Sigma}$ can be constructed by combining the robust estimate of $\mathbf{R}$ and D.

To estimate the correlation matrix $\mathbf{R}$, we consider Kendall's τ correlation, which is a robust measure for the relation between two variables. Recall Definition 4 on Kendall's τ correlation matrix. We denote 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 $(i_s, j_s) \in [p] \times [q], s = 1, 2$, $\mathbf{T}_{(j_1-1)p+i_1,(j_2-1)p+i_2}$ denotes the Kendall's τ correlation coefficient between variables $f_{i_1 j_1}(X_{i_1 j_1, k})$ and $f_{i_2 j_2}(X_{i_2 j_2, k})$. 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 given 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 the estimate of $\mathbf{T}_{(j_1-1)p+i_1,(j_2-1)p+i_2}$. Similarly to Han and Liu

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

(2014), we estimate $\mathbf{T}_{(j_1-1)p+i_1,(j_2-1)p+i_2}$ by

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

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 these, we estimate $\mathbf{R}$ by

$$\hat{\mathbf{R}}^{\tau} = \left(\sin(\frac{\pi}{2}\hat{\mathbf{T}}_{i,j})\right) = \sin\left(\frac{\pi}{2}\hat{\mathbf{T}}\right). \tag{2.2}$$

2.3 Estimate of latent covariance matrix of matrix-variates

Similarly to Tsiligkaridis and Hero (2013), we assume that the latent covariance matrix $\mathbf{\Sigma}$ has the Kronecker structure given in (1.1). That is, $\mathbf{\Sigma} = \sum_{i=1}^{r} A_i \otimes B_i$, where A_i is a $q \times q$ linearly independent matrix, B_i is a $p \times p$ linearly independent matrix, and $r \leq \min\{p^2, q^2\}$. We estimate the 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 the transelliptical distribution. The main reason for the stronger assumption is that the standard

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

deviation is, in general, 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. Because D is a diagonal matrix, we denote the diagonal elements as a 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

$$\hat{\sigma}_{ij} = c_{ij} \cdot \text{median}\{|X_{ij,k} - X_{ij}^{med}|, k = 1, \cdots, n\}, \tag{2.3}$$

where $X_{ij}^{med} = \text{median}\{X_{ij,k}, k = 1, \cdots, n\}$ and c_{ij}^{-1} is equal to the 0.5 quantile of the distribution of $|X_{ij,k} - \xi_{ij,0.5}|/\sigma_{ij}$, which can be written as a function of the 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 a χ^2 distribution with degree of freedom one. We show later that the estimate $\hat{\sigma}_{ij}$ is uniformly consistent over $(i,j) \in [p] \times [q]$ under a mild assumption on the densities of the marginal distributions.

Remark 1. Here, we aim to give a robust estimate of the variance of

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

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 the MAD-type estimate cannot be used. In this case, many other robust estimators can be used. Catoni (2012) proposed a robust estimator of the variance that allows for heavy-tailed distributions with a bounded kurtosis. Suppose that $\{Z_k, 1 \leq k \leq n\}$ are independent and identically distributed (i.i.d) copies 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 the 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 a good convergence rate. Specifically, $P(\max\limits_{1\leq i,j\leq p} |\hat{\check{\sigma}}_{1\leq i,j\leq p}^{RA} - \check{\sigma}_{ij}| \geq 4v\sqrt{a(\log p)/n}) \leq 2p^{2-a}$, where v is a constant and $a > 2$ (Fan et al. (2017)). Obviously, by assuming $\max\limits_{1\leq i,j\leq p} E(X_{ij}^4)$ and replacing p with 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 $\hat{D}^{(d)} = (\hat{\sigma}_{ij}, 1 \leq i \leq p, 1 \leq j \leq q)$. Moreover, $\mathbf{R}$ can be estimated as shown in the previous section. Combining these, we have the following robust estimate of $\mathbf{\Sigma}$:

$$\hat{\mathbf{\Sigma}}^{\tau} = \hat{D}\hat{\mathbf{R}}^{\tau}\hat{D}. \tag{2.4}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

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 in each row and q^2 blocks in total. Denote $\mathbf{N} = (N(i,j))_{i,j=1}^{q}$, where $N(i,j) \in \mathbb{R}^{p\times p}$ is a block in the ith row and the 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 $\mathrm{vec}(N(i,j))^\top \in \mathbb{R}^{p^2}$. For further details on this transformation, refer to Tsiligkaridis and Hero (2013). Moreover, we define $\mathcal{T}^{-1} : \mathbb{R}^{q^2\times p^2} \mapsto \mathbb{R}^{pq\times pq}$ as the inverse operator of $\mathcal{T}(\cdot)$. Based on the definition of $\mathcal{T}(\cdot)$, we have

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

Because A_i and B_i are linearly independent, the above equation implies that the matrix $\mathcal{T}(\mathbf{\Sigma})$ has rank r. When r is small, $\mathcal{T}(\mathbf{\Sigma})$ has a 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 p^2-dimensional vectors with $\|u\| = \|v\| = 1$ and $c > 0$ is a constant, we can show that $\tilde{M}$ is positive semidefinite if and only if $u = v$.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

In fact, when $u \neq v$, for any vector $w \in \mathbb{R}^{p^2}$ such that $w^\top(u+v)/2 = 0$, we have 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 positive semidefinite. For the matrix $\mathcal{T}(\mathbf{\Sigma})$ considered here, $A_i^\top$ is, in general, not equal to B_i. Therefore, $\mathcal{T}(\mathbf{\Sigma})$ may not be positive semidefinite, in general.

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

$$\hat{\mathbf{\Sigma}}_{\mathcal{T}}^{\tau} = \arg \min_{S \in \mathbb{R}^{q^2 \times p^2}} \|\mathcal{T}(\hat{\mathbf{\Sigma}}^{\tau}) - S\|_F^2 + \lambda \|S\|_*, \tag{2.5}$$

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

$$\hat{\mathbf{\Sigma}}_{\mathcal{T}}^{\tau} = \sum_{i=1}^{\min\{p^2,q^2\}} \left(\hat{\varphi}_i(\mathcal{T}(\hat{\mathbf{\Sigma}}^{\tau})) - \frac{\lambda}{2} \right)_+ \mathbf{u}_i \mathbf{v}_i^\top, \tag{2.6}$$

where $\hat{\varphi}_i(\mathcal{T}(\hat{\mathbf{\Sigma}}^{\tau}))$ is 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

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

where LR in the subscript of the estimator indicates that the low-rank

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

Kronecker structure in (1.1) has been taken into account. The tuning parameter λ can be selected using the cross-validation (CV) method of Bickel and Levina (2009).

Finally, note that $\mathbf{R}$ itself may be of interest in many applications. In such cases, the above procedure can be used to estimate $\mathbf{R}$ when the latent correlation matrix $\mathbf{R}$ is assumed to have the Kronecker form given in (1.1). By replacing $\hat{\mathbf{\Sigma}}^{\tau}$ in (2.5) and (2.6) with $\hat{\mathbf{R}}^{\tau}$ and denoting the optimal solution of (2.6) as $\hat{\mathbf{R}}^{\tau}_{\mathcal{T}}$, we obtain 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 $\mathbf{R}$. First, we consider $\mathbf{\Sigma}$. 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 do not impose the sparsity assumption further. Clearly, this can be extended to include the sparsity assumption in our setting. For identification, we rewrite the model as

$$\mathbf{\Sigma} = \gamma \cdot A \otimes B, \tag{2.7}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

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 $V_A = \text{vec}(A^\top), V_B = \text{vec}(B)$. 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 (2.4). We estimate (γ, V_A, V_B) by minimizing the following objective function:

$$(\hat{\gamma}, \hat{V}_A, \hat{V}_B) = \arg\min_{\Theta} \|\mathcal{T}(\hat{\mathbf{\Sigma}}^\tau) - dv_1 v_2^\top\|_F^2,$$

where $\Theta = \{(d, v_1, v_2)\colon\ d \in \mathbb{R}, v_1 \in \mathbb{R}^{q^2}, v_2 \in \mathbb{R}^{p^2}, d > 0, \|v_1\| = \|v_2\| = 1\}$. Obviously, the estimator $\hat{\gamma}$ is the largest singular value of the SVD decomposition of $\mathcal{T}(\hat{\mathbf{\Sigma}}^\tau)$, and $(\hat{V}_A, \hat{V}_B)$ represents the associated left and right eigenvectors, respectively. Consequently, $\mathbf{\Sigma}$ in (2.7) can be estimated using

$$\hat{\mathbf{\Sigma}}^\tau_{(rk=1)} = \mathcal{T}^{-1}(\hat{\gamma}\hat{V}_A\hat{V}_B^\top). \tag{2.8}$$

Let $\hat{A} \in \mathbb{R}^{q\times q}$ and $\hat{B} \in \mathbb{R}^{p\times p}$ be 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 $\hat{\mathbf{\Sigma}}^\tau_{(rk=1)}$.

Similarly to $\mathbf{\Sigma}$, when $\mathbf{R}$ has the Kronecker structure with $r = 1$, it can

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

be estimated in the same way. Similarly to (2.7), denote

$$\mathbf{R} = \tilde{\gamma} \cdot \tilde{A} \otimes \tilde{B}, \tag{2.9}$$

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 (2.2). By replacing $\hat{\mathbf{\Sigma}}^\tau$ with $\hat{\mathbf{R}}^\tau$ in the above procedure, we can estimate $\mathbf{R}$ in (2.9). Define $V_{\tilde{A}}$ and $V_{\tilde{B}}$ in the same way as V_A and V_B, respectively. Then, $\mathbf{R}$ can be estimated using

$$\hat{\mathbf{R}}^\tau_{(rk=1)} = \mathcal{T}^{-1}(\hat{\tilde{\gamma}}\hat{V}_{\tilde{A}}\hat{V}_{\tilde{B}}^\top), \tag{2.10}$$

where $\hat{\tilde{\gamma}}$ is the largest singular value of the SVD decomposition of $\mathcal{T}(\hat{\mathbf{R}}^\tau)$, and $(\hat{V}_{\tilde{A}}, \hat{V}_{\tilde{B}})$ denotes 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 estimates 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

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

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 $\|\hat{\mathbf{R}}^{\tau} - \mathbf{R}\|_{op}$, because the former is no longer a symmetric matrix, and thus, 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 a 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 in deriving the matrix concentration inequality.

Proposition 3.1 (Tropp, 2012) *Let W be a random self-adjoint matrix.*

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

For all $t \in \mathbb{R}$,

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

For further details, refer to Tropp (2012). This inequality is the key step in the proof of Han and Liu (2017).

To simplify the argument, we first introduce some notation. 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 Kendall's τ correlation matrix for the population 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 the U-statistic (Keener et al., 1998; Borovskikh and Weber, 2003), because n is large, the tail probability

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

of $\sqrt{n}\mathbf{a}^\top E_n$ is similar to that of the normal distribution $N(0, \mathbf{a}^\top \mathbf{W}\mathbf{a})$ under certain 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 high-level assumption that $\sqrt{n}\mathbf{a}^\top E_n$ has a tail probability similar to that of a sub-Gausssian variable.

(A1) (Tail probability) Assume that $\|\mathbf{W}\|_{op} < C < \infty$ for any positive integers p and q. In addition, assume that, when 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 $\mathbf{a} \in \mathbb{R}^{p^2 q^2}$, 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 a 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) cannot be applied directly, we expect that a similar inequality still holds under certain conditions.

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

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

Theorem 1. *Assume* $\mathrm{vec}(X_k)$ *follows a transelliptical distribution, denoted by* $\mathrm{vec}(X_k) \sim TE_p(\mathbf{R}, \xi;\, f_{11}, \cdots, f_{pq})$. *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 $X^{(0)}_{ij,k} = c_{ij}|X_{ij,k} - \xi_{ij,0.5}|$. Note that $X^{(0)}_{ij,k}$ is different to $X^{(sv)}_{ij,k}$ in Section 2. Recall from (2.3) that c^{-1}_{ij} is the 0.5 quantile of the distribution of $|X_{ij,k} - \xi_{ij,0.5}|/\sigma_{ij}$. Clearly, $c_{ij} > 0$ and

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

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

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

Suppose that the variables $X_{ij,k}$ and $X^{(0)}_{ij,k}$ have densities denoted by $f_{ij}(x)$ and $g_{ij}(x)$, respectively. Recall that $\xi_{ij,0.5}$ and σ_{ij} are the 0.5 quantiles

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

of the distributions of $X_{ij,k}$ and $X^{(0)}_{ij,k}$, respectively, for $(i,j) \in [p] \times [q]$. We make the following assumption.

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

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

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

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 $\|\mathbf{R}\|_F/\min\{p,q\} \le \|\mathcal{T}(\mathbf{R})\|_{op} \le \|\mathbf{R}\|_F$.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

In addition, because $\sqrt{pq} \leq \|\mathbf{R}\|_F \leq pq$, $\|\mathcal{T}(\mathbf{R})\|_{op}$ can be as small as $\sqrt{pq}/\min\{p,q\}$, which is $O(1)$ if $p \asymp q$. If $\|\mathcal{T}(\mathbf{R})\|_{op}$ is small, such that $\omega_n^{(2)}\|\mathcal{T}(\mathbf{R})\|_{op} \leq \omega_n^{(1)}$, then from Theorem 2, we have $r(\omega_n^{(0)})^2 = O\left(r \cdot [n^{-1}(p^2+q^2+\log n) + (n^{-1}pq\log(pq))^2]\right)$. Then, the inequality in Theorem 2 can be written as

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

for some constant $C > 0$. When $\mathbf{\Sigma}$ has a low-rank structure, as in (1.1), then the first term is zero and we get the convergence rate

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

On the other hand, when the normal distribution is assumed, Tsiligkaridis and Hero (2013) obtained a convergence rate of 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 a normal distribution, although the semi-parametric estimate relaxes the assumption of normal distribution greatly.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

Finally, when the 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. Thus, 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 one,*

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

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

Now, we consider the asymptotic behavior of the estimators $\hat{\mathbf{\Sigma}}^{\tau}_{(rk=1)}$ 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 $\gamma = \|\mathbf{\Sigma}\|_F = \|\mathcal{T}(\mathbf{\Sigma})\|_{op}$. Furthermore, recall that $\text{vec}(\hat{A}^{\top}) = \hat{V}_A$, $\text{vec}(A^{\top}) = V_A$, $\text{vec}(\hat{B}) = \hat{V}_B$, and $\text{vec}(B) = V_B$. We have the following conclusions.

Theorem 3. *Recall $\mathbf{\Sigma}$ in (2.7) and the estimate $\hat{\mathbf{\Sigma}}^{\tau}_{(rk=1)}$ in (2.8). Under the assumptions of Theorem 2, we have*

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

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

where c *and* c' *take values 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*

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

Similarly to Theorem 3, we have the following conclusion for $\hat{\mathbf{R}}^\tau_{(rk=1)}$, which is the estimator of $\mathbf{R}$ in (2.9). The proof is similar to that of Theorem 3. Thus, we omit it here.

Proposition 3. *Recall* $\mathbf{R}$ *in (2.9) and the estimate* $\hat{\mathbf{R}}^\tau_{(rk=1)}$ *in (2.10). Under the assumption 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 *and* c' *take values 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 $X = H^\top Y L \in \mathbb{R}^{p\times q}$, where $Y \in \mathbb{R}^{s_1\times s_2}$, with $\mathrm{cov}(\mathrm{vec}(Y)) = \mathbf{I}_{s_1 s_2}$, and $H \in \mathbb{R}^{s_1\times p}$ and $L \in \mathbb{R}^{s_2\times q}$ are

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

such that each column of H and L has a 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 sufficient large constant C. 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).

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

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

Therefore, we can handle the case of p and q being an exponential order of

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

n. In addition, for Case (ii), using a 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 and 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), because p and q are large. The main reason is that we do not impose the sparsity assumption here. Note that the error term $\omega_n^{(2)}$ is the result of estimating D, and $\omega_n^{(1)}$ is the result of $\mathbf{R}$. 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 $\mathbf{R}$ to be larger than that of D. Moreover, we have the following relative error for Case (i):

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

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

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

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

which is better than Case (i) of $\mathbf{\Sigma}$. In addition, 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 following 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):

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

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

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 further by using the penalized method with the ℓ_1 penalty in order 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 a low-rank approximation of a 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 $H_i \in \mathbb{R}^{s_1\times p}$ and $L_i \in \mathbb{R}^{s_2\times q}$, for $1 \le i \le r$, are constant matrices, and $Y_{ki} \in \mathbb{R}^{s_1\times s_2}$, for $i = 1, \cdots, r$, 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} is specified later. For this model, it is easy to see that

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

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

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

Example 1. Set $r = 1$ and $Y_{k1} \in \mathbb{R}^{s_1 \times s_2}$, with $\text{vec}(Y_{k1}) \sim N(0, I_{s_1 s_2})$, 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 $l_{ij} = 0.2^{|i-j|}$. Replace the first observation X_1 with the contaminated observation $\tilde{X}_1 = \delta_0 \mathbf{I} + X_1$, where $\mathbf{I} \in \mathbb{R}^{p \times q} = (\mathbf{I}_p, \mathbf{0}_{p \times (q-p)})$. 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 $t(3)$, a t-distribution with 3 degrees of freedom. Then, H_1 and L_1 are the same as in Example 1. Obviously, we have $\text{cov}(\text{vec}(Y_{k1})) = 3I_{s_1 s_2}$.

Example 3. Set $r = 3$. Y_{k1}, H_1, and L_1 are the same as in Example 1. Y_{k2} and Y_{k3} are i.i.d. copies of Y_{k1}, for $1 \leq k \leq n$. $H_2 = (h'_{ij}) \in \mathbb{R}^{s_1 \times p}$ 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 $l''_{ij} = 0.1^{|i-j|}$. Similarly to Example 1, we replace the first observation X_1 by the contaminated observation $\tilde{X}_1$ defined in Example 1.

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

We consider two cases: (s_1, s_2) is equal to (p, q) and $(\lceil p/4 \rceil, \lceil q/4 \rceil)$ where

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

$\lceil a \rceil$ denotes the largest integer no more than a for any constant $a \in \mathbb{R}$. In Examples 1 and 2, we assume that $r = 1$ is known. The robust estimator $\hat{\mathbf{\Sigma}}^{\tau}_{(rk=1)}$ is obtained as in Section 2.4. In addition, the nonrobust estimate, denoted as $\hat{\mathbf{\Sigma}}^{sam}_{(rk=1)}$, is the rank-one Kronecker approximation of the sample covariance matrix. For Examples 3 and 4, the robust estimator $\hat{\mathbf{\Sigma}}^{\tau}_{LR}$ is obtained using the approach in Section 2.3, and the nonrobust estimator is the PRLS estimator of Tsiligkaridis and Hero (2013), denoted as $\hat{\mathbf{\Sigma}}^{prls}$. Then, $\hat{r}$ is determined by the tuning parameter λ, which is selected using the CV method of Bickel and Levina (2009).

Because the nonrobust estimators, $\hat{\mathbf{\Sigma}}^{prls}$ and $\hat{\mathbf{\Sigma}}^{sam}_{(rk=1)}$, are derived from the sample covariance matrix, to simplify the description, we denote $\hat{\mathbf{\Sigma}}^{prls}$ and $\hat{\mathbf{\Sigma}}^{sam}_{(rk=1)}$ as $\hat{\mathbf{\Sigma}}^{sam}$. Furthermore, the robust estimators ($\hat{\mathbf{\Sigma}}^{\tau}_{(rk=1)}$ and $\hat{\mathbf{\Sigma}}^{\tau}_{LR}$) are 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 replications, denoted as $Err^{(rob)}_F, Err^{(rob)}_2$, and $Err^{(rob)}_\infty$, respectively. Similarly, we compute those of $Err^{(sam)}$ and define $Err^{(sam)}_F, Err^{(sam)}_2$, and $Err^{(sam)}_\infty$ in the same way.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

4.2 Simulation results

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

Moreover, we compare the following two settings: $(s_1, s_2) = (p, q)$ and $(s_1, s_2) = (\lceil p/4 \rceil, \lceil q/4 \rceil)$. For Examples 1 and 2, comparing Table 1 and Table S1 in the Supplementary Material, we note that the two estimators show similar performance under the two different values of (s_1, s_2). However, for Examples 3 and 4, comparing Table 2 and Table S2 in the Supplementary Material, we find significant differences in the performance of the estimators under the two different values of (s_1, s_2).

(2). *Simulation results on the selection of rank.* In our simulations,

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

Table 1: Simulation results for Examples 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 Examples 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

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

when r is unknown, the tuning parameter λ in our method is selected following Bickel and Levina (2009). To check the effectiveness of this method, we report in Tables 3–4 the simulation results on the rank selection for Examples 3 and 4, where the true rank is three.

Set $(n, p, q, s_1, s_2) = (100, 15, 15, 15, 15)$ in Example 3 and Example 4. Table 3 reports the results for the event $\{\hat{r} = i\}$, with $i = 1, 2, 3$, and the event $\{\hat{r} > 3\}$ over 200 replications. From Table 3, we see that the 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% over 200 replications, in most of the cases. In addition, Table 4 reports $Err_F^{(rob)}, Err_2^{(rob)}$, and $Err_\infty^{(rob)}$ with fixed rank=1,2,3, respectively. From Table 4, we see that rank=3 leads to the best results. In addition, we see that $Err_F^{(rob)}$ is affected most by the selection of the rank, whereas $Err_\infty^{(rob)}$ is least affected by the selection of the rank.

4.3 Real-Data analysis

We apply our method to the Atlas of Gene Expression in the Mouse Aging (AGEMAP) database, which is a resource of gene expressions 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 four age states:

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

Table 3: Rank estimation for Examples 3 and 4, where the true rank is three and $(n, p, q, s_1, s_2) = (100, 15, 15, 15, 15)$. We set the numbers of the estimated rank $\hat{r}$ equal to $1, 2, 3$ and $\hat{r} > 3$ over 200 replications. The rank is selected correctly 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 Examples 3 and 4, where the true rank is three 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

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

1, 6, 16, and 24 months. For each age state, researchers chose ten mice, with five for each gender, yielding 40 observations in total. Similarly 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 the genes related to the mitogen-activated protein kinase signaling pathway, long-term potentiation, insulin signaling pathway, and vascular endothelial growth factor signaling pathway, as documented at $http://rgd.mcw.edu/rgdweb/pathway/pathwayRecord.html?accid = PW:0000243\&species = Mouse\sharp Pathway$. We apply our method to males and females 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 aging. Because 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: (i) the robust estimate of our proposal; (ii) the PRLS estimator of Tsiligkaridis and Hero (2013), which is nonrobust; and (iii) the estimator of Leng and Pan (2017), who assume that $\Sigma = A \otimes B$, that is, $r = 1$ in (1.1). For our proposed method and the PRLS estimator, r is estimated using the data rather than fixed, with the estimator denoted as $\hat{r}$. Heat maps of the covariance matrices for both

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

males and females obtained by three estimators are presented, respectively, in Figures 1–3, which are available in the Supplementary Material. In all of these figures, the diagonal blocks from the lower, left corner to the upper, right corner are associated with seven tissues.

According to Yin and Li (2012), genes associated with aging have dependencies not only inside the same tissue, but also across different tissues. From Figure 1, we observe a weak dependency between Hippocampus and Thymus in males, and a clear dependency between Cerebrum and Thymus in females. These observations coincide with those of Yin and Li (2012), where the authors found that gene expressions in Thymus are related to 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 and female mice, and that the patterns of dependency between the tissues are different for males and females. These coincide with our results in plot (a) and (b) in Figure 1.

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

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

PLRS estimator (Tsiligkaridis and Hero, 2013) show almost no, or very weak correlations between different tissues for females, which is inconsistent with the findings of Yin and Li (2012).

5. Discussion

We have proposed a method for covariance matrix estimation for a high-dimensional matrix-variate in the framework of a transelliptical distribution, taking into account the Kronecker structure of the covariance matrix. Recall that $\mathbf{T} = (\mathbf{T}_{i,j})$ is Kendall's correlation matrix with estimate $\hat{\mathbf{T}} = (\hat{\mathbf{T}}_{i,j})$, and $\mathbf{R} = (\mathbf{R}_{i,j})$ is 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 $\hat{\mathbf{R}}^{sam} = (\hat{\mathbf{R}}^{sam}_{i,j})$ as the sample correlation matrix. When X_k follows a normal distribution, the sample Pearson's correlation is asymptotically unbiased and reaches the Cramér-Rao lower bound as the sample size tends to infinity (Xu et al., 2013). Hence, $\hat{\mathbf{R}}^{sam}_{i,j}$ is, in general, more efficient than $\hat{\mathbf{R}}^{\tau}_{i,j}$ when X_k is normal.

Consider a bivariate normal distribution with a correlation coefficient ρ. Let $\hat{\mathbf{T}}_{\rho}$ be 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, that is, $\hat{\rho}_{\mathcal{K}} = \sin(\frac{\pi}{2}\hat{\mathbf{T}}_{\rho})$, and let $\hat{\rho}_{\mathcal{P}}$ denote the sample Pearson correlation. According to Xu et al. (2013), the estimator $\hat{\rho}_{\mathcal{K}}$ has variance $\mathrm{Var}(\hat{\rho}_{\mathcal{K}}) \approx$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

$[\pi^2(4-\rho^2)/36]\mathrm{Var}(\hat{\mathbf{T}}_\rho)$,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

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

In particular, 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 the price paid for the use of a robust estimate.

Supplementary Material

The Supplementary Material consist of two parts. The first part contains the proofs of Proposition 1, Theorem 1, Lemma 1, and Theorem 3 in the main text. The second part contains the simulation results for the case of $(s_1, s_2) = (\lceil p/4 \rceil, \lceil q/4 \rceil)$ for Examples 1–4 in Section 4, and Figures 1–3 for the real-data analysis in Section 5.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

Acknowledgements

The work of Dr. Zhao is supported by the National Natural Science Foundation of China (No. 11471030, 11871104) and the Fundamental Research Funds for the Central Universities. Lu Niu is supported by the National Natural Science Foundation of China (No. 11571267, 91538112).

References

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

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

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

Borovskikh, Y. V. and N. C. Weber (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., V. Ng, and M. Rothschild (2010). Asset pricing with a factor arch covariance

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

structure: empirical estimates for treasury bills. *National Bureau of Economic Research*, 213–237.

Fan, J., Q. Li, and Y. Wang (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., Y. Liao, and H. Liu (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., K. T. Fang, and S. Kotz (2002). The meta-elliptical distributions with given marginals. *Journal of Multivariate Analysis 94*(1), 1–16.

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

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

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

Gupta, A. K. and D. K. Nagar (1999). *Matrix variate distributions*, Volume 104. CRC Press.

Han, F. and H. Liu (2014). Scale-invariant sparse PCA on high dimensional meta-elliptical

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

data. *Journal of the American Statistical Association 109*(505), 275–287.

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

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

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

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

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

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

Ledoit, O. and M. Wolf (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. and G. Pan (2017). Covariance estimation via sparse kronecker structure. *Bernoulli In press*.

Leng, C. and C. Y. Tang (2012). Sparse matrix graphical models. *Journal of the American*

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

Statistical Association 107(499), 1187–1200.

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

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

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

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

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

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

Sejnowski, T., S. Makeig, and A. Delorme (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*

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

Computational Mathematics 12(4), 389–434.

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

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

Wallbacks, L. and U. E. B. Norden (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., Y. Zhao, et al. (2016). Adaptive estimation of the copula correlation matrix for semiparametric elliptical copulas. *Bernoulli 22*(2), 1184–1226.

Xu, W., Y. Hou, Y. S. Hung, and Y. Zou (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. and H. Li (2012). Model selection and estimation in the matrix normal graphical model. *Journal of Multivariate Analysis 107*(3), 119–140.

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

School of Mathematics and System Science, Beihang University, China

E-mail: niulu2010@yeah.net

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0364

School of Statistics, Beijing Normal University, China

E-mail: zhaojunlong928@126.com