Submitted to Bernoulli

Covariance Estimation via Sparse Kronecker Structures

CHENLEI LENG[1,*] and GUANGMING PAN[2,**]

[1] *Department of Statistics, University of Warwick, Coventry CV4 7AL, UK.*
E-mail: [*] C.Leng@warwick.ac.uk

[2] *School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371, Republic of Singapore. E-mail:* [**] gmpan@ntu.edu.sg

The problem of estimating covariance matrices is central to statistical analysis and is extensively addressed when data are vectors. This paper studies a novel Kronecker-structured approach for estimating such matrices when data are matrices and arrays. Focusing on matrix-variate data, we present simple approaches to estimate the row and the column correlation matrices, formulated separately via convex optimization. We also discuss simple thresholding estimators motivated by the recent development in the literature. Non-asymptotic results show that the proposed method greatly outperforms methods that ignore the matrix structure of the data. In particular, our framework allows the dimensionality of data to be arbitrary order even for fixed sample size, and works for flexible distributions beyond normality. Simulations and data analysis further confirm the competitiveness of the method. An extension to general array-data is also outlined.

Keywords: Covariance matrix, Kronecker structure, Matrix data, Non-asymptotic bound.

1. Introduction

Matrix and array observations are becoming increasingly available in the big data era thanks to the rapid advance in the information technology and the need to store data in structured forms; see, for example, Li et al. (2010), Hoff (2011), Leng and Tang (2012), Zhou et al. (2013), and Zhou et al. (2014). Consider independent and identically distributed matrix-variates $X_1, ..., X_n \in \mathbb{R}^{p\times q}$ that are realizations of a matrix random variable X following a matrix-variate distribution (Gupta and Nagar, 2000). Writing vec as the vector operator that stacks the columns of a matrix into a vector, we denote

$$\mathrm{var}(\mathrm{vec}(X_k)) = \Gamma$$

as the $pq \times pq$ dimensional covariance matrix. Without loss of generality, we assume that $\mathrm{E}(X_k)$ is known or a consistent estimator of $\mathrm{E}(X_k)$ such as the sample mean is available. For the latter case, we require $n > 1$. In the sequel, we work with $X_k - \mathrm{E}(X_k)$.

The covariance matrix Γ plays an indispensable role in multivariate data analysis and is a central quantity for estimation and inference. To begin with, a simple estimate of Γ is the familiar sample covariance matrix after these observations are vectorized. However,

whenever the data dimension pq is larger than the sample size n, this estimator can be of little use due to its singularity. Based on this observation, a plethora of approaches, built upon various sparsity assumptions on Γ, have attracted increasing attention. See, for example, Bickel and Levina (2008a,b), Rothman et al. (2009), Cai and Liu (2011), Bien and Tibshirani (2011), Rothman (2012), Xue et al. (2012), and Cui et al. (2016).

Stacking matrices into vectors incurs a loss of information in the matrix form of the data. An attractive alternative is to assume (Hoff, 2011; Leng and Tang, 2012; Tsiligkaridis and Hero, 2013)

$$\Gamma = \Psi \otimes \Sigma,$$

where, loosely speaking, $\Psi = (\psi_{ij}) \in \mathbb{R}^{q\times q}$ depicts the covariance of the columns of X_i and $\Sigma = (\sigma_{ij}) \in \mathbb{R}^{p\times p}$ that of the rows. Using a Kronecker product for the overall covariance matrix Γ retains the matrix structure of the data. Another immediate advantage is that the number of the unknown parameters in Γ reduces from an order of p^2q^2 to an order of $p^2 + q^2$, making the problem more tractable. As will become clear, with appropriate sparsity assumptions on Ψ and Σ, this decomposition enables one to estimate Γ at a higher rate of convergence, and allows substantially larger dimensional covariances to be estimated, even with a fixed sample size. Without considering sparsity, Srivastava et al. (2008) estimated the Kronecker structure when p and q are fixed. There are also a growing number of papers on estimating the concentration matrix Γ^{-1} via a Kronecker product representation by estimating sparse concentration matrices Ψ^{-1} and Σ^{-1} (Allen and Tibshirani, 2010; Yin and Li, 2012; Leng and Tang, 2012; Zhou, 2014). These papers assume matrix normality for the data distribution. None of them addresses the issue of estimating sparse Ψ or Σ.

This paper is motivated by the neuroimaging data in Section 4.1. When we applied existing approaches for estimating a sparse Gaussian graphical model in Γ^{-1} (Yuan and Lin, 2007), or for estimating two sparse Gaussian graphical models in Ψ^{-1} and Σ^{-1} (Leng and Tang, 2012), or for estimating a sparse covariance matrix in Γ (Cui et al., 2016), they all give a final estimated Γ which is diagonal. A formal test of the null hypothesis that the covariance matrix is diagonal, however, is rejected (Chen et al., 2010). On the other hand, the proposed class of estimators, collectively named sparse Kronecker-structured estimators for huge dimensional Ψ and Σ under sparsity assumptions, is found to be useful for depicting the correlation structures in Ψ and Σ. See Figure 6. At the core of these estimators is to estimate non-iteratively two correlation matrices by convex optimization, one for Ψ and the other for Σ. The resulting estimates are guaranteed positive definite. The technical tools used for the non-asymptotic analysis are totally different from those in Leng and Tang (2012) and Zhou (2014) and can be of independent interest. By "non-asymptotic analysis" we here mean that the sample size n does not need to go to infinity. Apart from this, there are two major innovations in our non-asymptotic analysis. First, the non-asymptotic results cover not only the usual Gaussian distribution, but also distributions such as the exponential tail type distributions (Cai and Liu, 2011) and the Bernoulli distribution, substantially enhancing the usefulness of the method. Second, our model allows the dimensionality to be arbitrary order even when the sample size is fixed, thanks to the Kronecker structure assumption that greatly

reduces the number of parameters needed. For modelling covariance of random vectors, the dimensionality is allowed at most to be of sub-exponential order of the sample size (Bickel and Levina, 2008a,b). Methodologically, the proposed method for matrix data can be easily extended to study array data, which is straightforward operationally and theoretically, and is discussed in the paper. Our non-asymptotic analysis indicates that the proposed method gives fast rate of convergence for estimating Γ. As a result, simple estimates by soft thresholding usually suffice to guarantee its positive definiteness in contrast to the thresholding estimator for the covariance matrix of vector data where special care is recommended (Rothman, 2012; Xue et al., 2012). This is in sharp contrast to the more difficult problem of estimation concentration matrices where loss function based approaches have to be employed (Leng and Tang, 2012).

The following notations are used throughout the paper. For a square matrix $A = (a_{ij}) \in \mathbb{R}^{m\times m}$, $\mathrm{diag}(A)$ denotes a matrix consisting of the diagonal terms of A as $\mathrm{diag}\{a_{11}, ..., a_{mm}\}$. The Frobenius norm of A is denoted as $\|A\|_F$ and its element ℓ_1 norm is denoted as $|A|_1$. We write $|A|_{\max} = \max_{i,j} |a_{ij}|$ as the maximum entry of A. The spectral norm of A is $\|A\|_2$ denoting the largest singular value of A. The trace of a square matrix A is denoted as $\mathrm{tr}(A)$ and the Kronecker product between matrices is denoted by $\otimes$. A positive semi-definite matrix is denoted by $A \succeq 0$. The (i, j)th entry of A is denoted either as a_{ij} or A_{ij}. If A denotes a covariance matrix, the corresponding correlation matrix is denoted as R^A such that $\{\mathrm{diag}(A)\}^{1/2} R^A \{\mathrm{diag}(A)^{1/2}\} = A$. Finally, I_k denotes the $k \times k$ identity matrix.

The rest of the paper is organised as follows. We present the proposed Kronecker-structured estimation method in Section 2, where several variants of the approach are discussed. In Section 3, we provide the main theory, and outline the generalisation of the proposed method for estimating the covariance matrix of array data. Simulation studies and data analysis are presented in Section 4 with a short conclusion in Section 5. All the proofs are relegated to the Appendix.

2. Kronecker-structured estimation

We make the following assumption on the matrix variate data X that includes multivariate normality as a special case.

Condition (A). The matrix variate data X have the structure

$$X = BSA^{\mathsf{T}}, \tag{1}$$

where A, B are square matrices such that $AA^{\mathsf{T}} = \Psi$, $BB^{\mathsf{T}} = \Sigma$ and the entries of $S = (s_{ij})$ are independent and identically distributed with mean 0 and variance 1.

First we derive a simple sample estimate for $\Gamma = \Psi \otimes \Sigma$. Write

$$\Psi_n = \frac{1}{n}\sum_{k=1}^{n} X_k^{\mathsf{T}} X_k, \ \Sigma_n = \frac{1}{n}\sum_{k=1}^{n} X_k X_k^{\mathsf{T}}.$$

Clearly, the expectations of these two matrices satisfy

$$E(\Psi_n) = \mathrm{tr}(\Sigma)\Psi,\ E(\Sigma_n) = \mathrm{tr}(\Psi)\Sigma,$$

respectively, giving rise to a reasonable estimate of $\Gamma = \Psi \otimes \Sigma$ as $\Psi_n \otimes \Sigma_n/(\mathrm{tr}(\Sigma)\mathrm{tr}(\Psi))$. Noting that $\mathrm{tr}(\Gamma) = \mathrm{tr}(\Sigma)\mathrm{tr}(\Psi)$, we can estimate $\mathrm{tr}(\Sigma)\mathrm{tr}(\Psi)$ consistently by $\sum_{k=1}^n \|X_k\|_F^2/n$. Therefore, a simple sample estimate of Γ admitting the Kronecker structure is

$$\Gamma_n = \Psi_n \otimes \Sigma_n/(\frac{1}{n}\sum_{k=1}^n \|X_k\|_F^2). \tag{2}$$

If the sample size goes to infinity and the dimensionality of the covariance matrix is fixed, it is not difficult to see that Γ_n is a consistent estimator of Γ.

Next we discuss how to estimate Γ when p and q are much larger than n. Define the marginal variance matrices of Ψ_n and Σ_n as

$$(W_1^\Psi)^2 = \mathrm{diag}(\Psi_n),\ (W_1^\Sigma)^2 = \mathrm{diag}(\Sigma_n),$$

respectively. The sample row and column sample correlation matrices can be written respectively as R_n^Ψ and R_n^Σ, where

$$\Psi_n = W_1^\Psi R_n^\Psi W_1^\Psi,\ \Sigma_n = W_1^\Sigma R_n^\Sigma W_1^\Sigma.$$

These two sample correlation matrices can be seen as estimates of the population correlation matrices R^Ψ and R^Σ respectively. The proposed Kronecker estimator replaces R_n^Ψ and R_n^Σ by their penalized estimates in

$$\widehat{R}^\Psi = \arg\min_{R\in\mathbb{R}^{q\times q}} \frac{1}{2}\|R - R_n^\Psi\|_F^2 + \lambda_\Psi |R|_1,\quad \text{s. t. } R \succeq \epsilon I_q,\ R_{jj} = 1,\ j = 1, ..., q, \tag{3}$$

$$\widehat{R}^\Sigma = \arg\min_{R\in\mathbb{R}^{p\times p}} \frac{1}{2}\|R - R_n^\Sigma\|_F^2 + \lambda_\Sigma |R|_1,\quad \text{s. t. } R \succeq \epsilon I_p,\ R_{jj} = 1,\ j = 1, ..., p, \tag{4}$$

where λ_Ψ and λ_Σ are penalty parameters, and ϵ is an arbitrary small positive constant that guarantees positive definiteness of the estimates. Here the simplified notation $\widehat{R}^\Psi$ ($\widehat{R}^\Sigma$) suppresses its dependence on the sample size n and the penalty parameter λ_Ψ (λ_Σ). The optimization problem in (3) and (4) is convex and thus convex optimization techniques can be applied. In this paper, we use the efficient accelerated proximal gradient algorithm in Cui et al. (2016). After obtaining the penalized estimates of Ψ and Σ, the final Kronecker estimator assembles them as

$$\widehat{\Psi} = W_1^\Psi \widehat{R}^\Psi W_1^\Psi,\quad \widehat{\Sigma} = W_1^\Sigma \widehat{R}^\Sigma W_1^\Sigma,\quad \widehat{\Gamma} = \widehat{\Psi} \otimes \widehat{\Sigma}/\left(\frac{1}{n}\sum_{k=1}^n \|X_k\|_F^2\right). \tag{5}$$

Motivated by the adaptive Lasso (Zou, 2006), we can replace the penalty $|R|_1$ in (3) by $\sum_{j<k} |R_{jk}|/|(R_n^\Psi)_{jk}|$, and the penalty in (4) by $\sum_{j<k} |R_{jk}|/|(R_n^\Sigma)_{jk}|$. After the correlation matrices are obtained, we estimate Σ, Ψ and Γ as in (5). The estimate is denoted as $\hat{\Gamma}^A$ and referred to as the adaptive Kronecker estimator.

It turns out numerically that the positive definiteness constraints $R \succeq \epsilon I$ in (3) and (4) are often redundant in the sense that the resulting estimates without these constraints are positive definite. Thanks to the Kronecker product structure of Γ, the non-asymptotic analysis in the next session reveals that in estimating the $p \times p$ matrix Σ, the effective sample size becomes nq, and in estimating the $q \times q$ matrix Ψ, the effective sample size becomes np. Thus, the curse of dimensionality is greatly alleviated. As a consequence, we can directly use thresholding as in in Bickel and Levina (2008b) and Rothman et al. (2009) without the positive definite constraints, giving rise to simple and fast estimators. Although thresholding estimators for covariance of vector data are also found to be positive definite with high probability (Bickel and Levina, 2008b; Rothman et al., 2009), in practice, these estimators are based on an effective sample size n for estimating an $(pq)\times(pq)$ matrix in our setup, and thus are more prone to the violation of the constraints. Finally, if our estimate after thresholding is not positive definite, we then evoke the algorithm in Cui et al. (2016).

In particular, we define soft-thresholding Kronecker estimator by replacing $\widehat{R}^{\Psi}$ by $S_{\lambda_\Psi}(R_n^{\Psi})$ and $\widehat{R}^{\Sigma}$ by $S_{\lambda_\Sigma}(R_n^{\Sigma})$ in (5) respectively, where S_λ is the soft-thresholding operator such that the (i,j)th entry $(i \neq j)$ of $S_\lambda(A)$ is $\text{sign}(a_{ij}) \cdot \max\{|a_{ij}| - \lambda, 0\}$ for some $\lambda \geq 0$, and the diagonals of $S_\lambda(A)$ are the same as those in A. These matrices can be seen as the solutions to (3) and (4) with a tuning parameter λ but without the positive definiteness constraint $R \succeq \epsilon I$ respectively.

Similarly, define the hard-thresholding operator $H_\lambda(\cdot)$ such that the (i,j)th entry $(i \neq j)$ of $H_\lambda(A)$ is $a_{ij} \cdot I(|a_{ij}| \geq \lambda)$ for some $\lambda \geq 0$ and an indicator function $I(\cdot)$. The hard-thresholding estimator $\hat{\Gamma}^H$ is defined by replacing $\widehat{R}^{\Psi}$ by $H_{\lambda_\Psi}(R_n^{\Psi})$ and $\widehat{R}^{\Sigma}$ by $H_{\lambda_\Sigma}(R_n^{\Sigma})$ in (5) respectively.

Finally, we have the covariance estimate by vectorizing X_k as $\text{vec}(X_k)$. Writing

$$\widetilde{\Gamma}_n = \frac{1}{n}\sum_{k=1}^{n}\{\text{vec}(X_k)\}\{\text{vec}(X_k)\}^{\mathsf{T}},$$

we can estimate a sparse correlation matrix as

$$\widetilde{R}^{\widetilde{\Gamma}} = \arg\min_{R\in\mathbb{R}^{pq\times pq}} \frac{1}{2}\|R - R^{\widetilde{\Gamma}_n}\|_F^2 + \lambda|R|_1, \quad \text{s. t. } R \succeq \epsilon I_{pq},\ R_{jj} = 1,\ j = 1, ..., pq.$$

We note again that the constraint $R \succeq \epsilon I$ is enforced to guarantee the positive definiteness of the resulting estimate. Without this constraint, the resulting estimate is just the soft-thresholding estimate as in Bickel and Levina (2008a). However, Rothman (2012) and Xue et al. (2012) observed that when the dimension pq is high relative to the sample size n, the soft-thresholding estimator can be seriously non positive definite, giving rise to invalid covariance matrices. This phenomenon is especially true when both p and q are very large, a scenario that is appropriate for our study. Write $\widetilde{W}^2 = \text{diag}(\widetilde{\Gamma}_n)$. The vectorized estimator is formally defined as $\widetilde{\Gamma}^V = \widetilde{W}\widetilde{R}^{\widetilde{\Gamma}}\widetilde{W}$. More details of this approach can be found in Cui et al. (2016).

3. Theory

We now present non-asymptotic bounds for the proposed Kronecker estimate in terms of the spectral and Frobenius norms when p and q diverge to infinity. One novelty about our results is that the dimensionality p and q can diverge to infinity at any rate under suitable sparsity conditions, even when the sample size n is fixed. This is in marked contrast to other approaches where the dimensionality is only allowed to be of sub-exponential order of the sample size, and the matrix is sufficiently sparse (Bickel and Levina, 2008a,b). Another novelty of the results is that we study distributions well beyond matrix normality, providing theoretical guarantees for studying a variety of problems much broader than those in Leng and Tang (2012) and Zhou (2014). For $\Psi = (\psi_{ij})$, $\Sigma = (\sigma_{ij})$, let $s_\Psi = \sum_{i\neq j} I(\psi_{ij} \neq 0)$ and $s_\Sigma = \sum_{i\neq j} I(\sigma_{ij} \neq 0)$ denote the numbers of nonzero off-diagonal parameters in Ψ and Σ, respectively. The uniform upper bounds of ψ_{ij} and σ_{ij} are denoted as $\psi_{\max}$ and $\sigma_{\max}$ respectively. We impose the following conditions.

Condition (B). There exists constants $0 < c_1 < c_2$ such that $c_1 \leq \lambda_j(\Psi) \leq c_2, c_1 \leq \lambda_i(\Sigma) \leq c_2$ for $j \leq q, i \leq p$. Here $\lambda_j(\Psi), \lambda_i(\Sigma)$ are the eigenvalues of Ψ, Σ, respectively, in decreasing order.

Condition (C). Assume $p,\ q \to \infty$, $c_4 < \frac{\log p}{\log q} \leq c_5$ with $c_4 > 0$.

Condition (D). We assume $Es_{ij}^{48} < \infty$.

Basically, Condition (B) states that the eigenvalues of Ψ and Σ are bounded away from zero and infinity. Hence the diagonal elements of Ψ and Σ are also bounded from above. The model in (1) means that matrix variate X is a linear transformation of some $p \times q$ variate random matrix S with independent components. It generates a rich collection of X from S with given row and column covariance matrices $\Psi = AA^\mathsf{T}$ and $\Sigma = BB^\mathsf{T}$ such that $\mathrm{var}(\mathrm{vec}(X)) = \Gamma$. In particular, if X follows a matrix normal distribution, the data structure (1) is satisfied. Condition (D) is satisfied by many commonly used distributions such as the normal distribution, Bernoulli distribution and the exponential tail type distributions in Cai and Liu (2011). The moment condition can be further weakened by truncation. But we do not pursue this direction because otherwise a much lengthier proof is needed. The condition $c_4 < \frac{\log p}{\log q} \leq c_5$ in (C) means that p and q are not necessarily in the same order. For example, we can allow $p = O(q^k)$ for any finite k.

For simplicity, we focus on the non-asymptotic bounds of the Kronecker estimate. The properties of the other estimates can be derived likewise. We first present the accuracy of the estimated correlation matrices defined in (3) and (4). Then we spell out the non-asymptotic bounds for estimating Ψ and Σ that eventually give rise to the bounds of estimating Γ.

Theorem 1. *Assume that the true correlation matrix R^Ψ and R^Σ are both positive definite. Under (A)-(D) if we set the thresholding parameters as $\lambda_\Psi = O\Big(\sqrt{\frac{\log q}{np}}\Big)$ and $\lambda_\Sigma = O\Big(\sqrt{\frac{\log p}{nq}}\Big)$, then we have*

$$\|\widehat{R}^\Psi - R^\Psi\|_F \leq C\sqrt{(s_\Psi + 1)\frac{\log q}{np}},\ \text{and}\ \|\widehat{R}^\Sigma - R^\Sigma\|_F \leq C\sqrt{(s_\Sigma + 1)\frac{\log p}{nq}},$$

with probability $1-q^{-0.9}$ *and* $1-p^{-0.9}$*, respectively. Here (and in what follows)* C *is a positive constant independent of* n, p, q *but may take different values in different places.*

Note that Theorem 1 does not require the sample size n to go to ∞. For fixed n, Theorem 1 continues to hold. Loosely speaking, the theorem states that in estimating R^{Σ}, the sample size becomes from n to nq, and that it becomes np in estimating R^{Ψ}. Therefore, for the non-asymptotic results to take effect, we only need to let the effective sample sizes np and nq diverge to infinity, as compared to the usual non-asymptotic arguments for which the sample size n must diverge to infinity in estimating variance of vectorized data (Bickel and Levina, 2008b). We note that, if the vectorized estimate is used, the bound in terms of the Frobenius norm becomes $O(\sqrt{s_\Gamma \log(pq)/n})$ where s_Γ is the number of the nonzero off-diagonal terms of Γ (Rothman, 2012), for which a sufficient condition for the convergence is $\log(pq) = o(n)$ when Γ is sparse. Obviously, the convergence rate of the vectorized estimate is much slower.

We have the following corollary regarding the convergence rates of the Kronecker estimates $\widehat{\Psi}$ and $\widehat{\Sigma}$ of the two covariance matrices.

Corollary 1. *Assume that Conditions (A)-(D) are satisfied. We have*

$$\left\|\frac{1}{tr\Sigma}\widehat{\Psi}-\Psi\right\|_2 \le C\sqrt{(s_\Psi+1)\frac{\log q}{np}}, \qquad \left\|\frac{1}{tr\Sigma}\widehat{\Psi}-\Psi\right\|_F \le C\sqrt{(s_\Psi+q)\frac{\log q}{np}},$$

and

$$\left\|\frac{1}{tr\Psi}\widehat{\Sigma}-\Sigma\right\|_2 \le C\sqrt{(s_\Sigma+1)\frac{\log p}{nq}}, \qquad \left\|\frac{1}{tr\Psi}\widehat{\Sigma}-\Sigma\right\|_F \le C\sqrt{(s_\Sigma+p)\frac{\log p}{nq}},$$

with probability $1-q^{-0.9}$ *and* $1-p^{-0.9}$*, respectively.*

Now, we have the following corollary regarding the rate of convergence of the Kronecker estimate $\widehat{\Gamma}$ for estimating Γ.

Corollary 2. *Assume that Conditions (A)-(D) are satisfied,* $(s_\Psi+q)\frac{\log q}{np} \to 0$ *and* $(s_\Sigma+p)\frac{\log p}{nq} \to 0$*. We have*

$$\|\widehat{\Gamma}-\Gamma\|_2 \le C\left(\sqrt{(s_\Psi+1)\frac{\log q}{np}}+\sqrt{(s_\Sigma+1)\frac{\log p}{nq}}\right),$$
$$\|\widehat{\Gamma}-\Gamma\|_F \le C\left(\sqrt{(s_\Psi+q)\frac{\log q}{n}}+\sqrt{(s_\Sigma+p)\frac{\log p}{n}}\right),$$

with probability $1-p^{-0.9}-q^{-0.9}-(pq)^{-0.9}$.

As this paper is the first for studying the estimation of a sparse covariance matrix when it admits a Kronecker structure, we can only compare its rate of convergence to a few methods for estimating a sparse precision matrix. Assuming matrix normality, Leng and Tang (2012) and Zhou (2014) both show that their estimators have similar rates of convergence, when additional assumptions are posed on the sparsity of Ψ^{-1} and Σ^{-1}. Our results improve those in Leng and Tang (2012) that required $\max(p\log p/q, q\log q/p)/n \to 0$ ruling out problems with fixed n. In addition, the method in Leng and Tang (2012) involves non-convex optimization, while Zhou (2014) employs graphical lasso. Our method usually involves thresholding, and thus is computationally more attractive. When p or q is 1, the results are consistent with those in Xue et al. (2012).

We now compare the Kronecker estimate to the vectorized estimate. If we vectorize the matrix observations, the standard arguments in Xue et al. (2012) and Cui et al. (2016) can show that estimation errors satisfy

$$\|\widehat{\Gamma}^V - \Gamma\|_2 = O_p\left(\sqrt{(s_\Psi s_\Sigma + 1)\frac{\log pq}{n}}\right),\ \|\widehat{\Gamma}^V - \Gamma\|_F = O_p\left(\sqrt{(s_\Psi + q)(s_\Sigma + p)\frac{\log pq}{n}}\right).$$

The Kronecker structure assumption on Γ effectively increases the sample sizes to np and nq for estimating Ψ and Σ respectively.

Denote the non-diagonal support of Ψ as $A_{0\Psi} = \{(i,j) : i \neq j, (\Psi)_{ij} \neq 0\}$ and similarly $A_{0\Sigma}$ for Σ. We have the following consistency results for covariance selection.

Corollary 3. *Assume that Conditions (A)-(D) are satisfied.*
If $\lambda_q(R^\Psi) \gg C\sqrt{(s_\Psi + 1)\frac{\log q}{np}}$ *and* $\min_{(i,j)\in A_{0\Psi}}(R^\Psi)_{ij} \gg \sqrt{\frac{\log q}{np}}$ *hold then* $(\hat{R}^\Psi)_{ij} = 0$ *for* $(i,j) \in A_{0\Psi}^C$, *and* $(\hat{R}^\Psi)_{ij} \neq 0$ *for* $(i,j) \in A_{0\Psi}$ *with probability tending to one;*
If $\lambda_p(R^\Sigma) \gg C\sqrt{(s_\Psi + 1)\frac{\log p}{nq}}$ *and and* $\min_{(i,j)\in A_{0\Sigma}}(R^\Sigma)_{ij} \gg \sqrt{\frac{\log p}{nq}}$ *then* $(\hat{R}^\Sigma)_{ij} = 0$ *for* $(i,j) \in A_{0\Sigma}^C$, *and* $(\hat{R}^\Sigma)_{ij} \neq 0$ *for* $(i,j) \in A_{0\Sigma}$ *with probability tending to one.*

The condition $\lambda_q(R^\Psi) \gg C\sqrt{(s_\Psi + 1)\frac{\log q}{np}}$ ensures that the solution to (2) without the constraint is still positive definite with probability tending to one. It is clear that the solution to (2) without this constraint becomes the soft thresholding Kronecker estimator. Therefore with probability tending to one, $(\hat{R}^\Psi)_{ij} = \text{sign}(R_n^\Psi)_{ij}(|(R_n^\Psi)_{ij}| - \lambda_\Psi)_+$ where $(b)_+ > 0$ for $b > 0$, and $(b)_+ = 0$ otherwise. Corollary 3 establishes the consistency of covariance selection and is attractive from a model selection perspective.

We below discuss selection of λ_Σ and λ_Ψ via cross validation that is done by spliting the sample randomly into a training and a test set randomly N times. As in Bickel and Levina (2008b), it suffices to prove the result when $N = 1$. We consider choosing λ_Σ and λ_Ψ by a grid search on $\{j\sqrt{\frac{\log p}{nq}}\}, 0 < j \leq J$ and $\{j\sqrt{\frac{\log p}{nq}}\}, 0 < j \leq J_1$ respectively. For convenience write the observations as

$$X_1, \cdots, X_m, X_{m+1}, \cdots, X_{m+B}$$

with $n = m + B$, where $\{X_1, \cdots, X_m\}$ is the training set and the $\{X_{m+1}, \cdots, X_{m+B}\}$ is the test set. The two tuning parameters are chosen separately, such that for choosing λ_Σ for example, the estimated covariance matrix of the training dataset for Σ is the closest to the sample covariance matrix of the test set for Σ in terms of the F norm. The selection of λ_Ψ is done similarly. Denote the estimators obtained from cross validation by $\widehat{R}^{\Psi}_{\lambda_\Psi}$ and $\widehat{R}^{\Sigma}_{\lambda_\Sigma}$ respectively. Inspired by Bickel and Levina (2008b) we have the following theory.

Theorem 2. *Suppose that S in (1) consists of i.i.d standard normal variables. If $B = n\varepsilon_n$, $(\log J)^3 = o(\sqrt{nq}p^{-1}\varepsilon_n\sqrt{\log p(s_\Sigma + p)})$ and $J_1 = o(\sqrt{np}q^{-1}\varepsilon_n\sqrt{\log q(s_\Psi + q)})$, then with probability tending to one*

$$\|\widehat{R}^{\Psi}_{\lambda_\Psi} - R^{\Psi}\|_F \leq C\sqrt{(s_\Psi + 1)\frac{\log q}{np}}, \text{ and } \|\widehat{R}^{\Sigma}_{\lambda_\Sigma} - R^{\Sigma}\|_F \leq C\sqrt{(s_\Sigma + 1)\frac{\log p}{nq}}.$$

Corollary 4. *Under the condition of Theorem 2, the following two inequalities hold with probability tending to one*

$$\|\widehat{\Psi}^{\Psi}_{\lambda_\Psi} - \Psi\|_F \leq C\sqrt{(s_\Psi + q)\frac{\log p}{nq}}, \text{ and } \|\frac{1}{tr\Psi}\hat{\Sigma_{\lambda_\Sigma}} - \Sigma\|_F \leq C\sqrt{(s_\Sigma + p)\frac{\log p}{nq}}.$$

In theory, a convenient choice of ε_n is $\frac{C}{\log n}$ for some constant C.

3.1. Array data

The proposed framework can be easily extended to array-type data in a straightforward manner as we discuss below. Consider independent and identically distributed array variables $X_k \in \mathbb{R}^{p_1 \times \cdots \times p_L}$, $k = 1, \cdots, n$. We assume that they are properly centred such that $\mathrm{E}(X_k)$ is a zero array and $\mathrm{var}(\mathrm{vec}(X_k)) = \Gamma = \Sigma_1 \otimes \cdots \otimes \Sigma_L$.

As the higher-order analog of matrix rows and columns, a fiber is defined by fixing every index but one (Kolda and Bader, 2009). For example, the mode-ℓ fibers of an array X are all vectors $x_{i_1 \cdots i_{\ell-1} : i_{\ell+1} \cdots i_L}$ that are obtained by fixing the values of $\{i_1, ..., i_L\} \backslash i_\ell$. The mode-ℓ unfolding (also known as matricization or flattening) of a tensor X, denoted as $X^{(\ell)}$, is an $p_\ell \times q_\ell$ matrix with $q_\ell = \prod_{k=1(\neq \ell)}^{L} p_k$ by replacing the model-ℓ fibers in its columns. More specifically, the (i_ℓ, j)th element is the $(i_1, \cdots, i_L)$th element of X where

$$j = 1 + \prod_{k=1(\neq \ell)}^{L} (i_k - 1)J_k, \text{ with } J_k = \prod_{m=1(\neq \ell)}^{k-1} p_m.$$

With this definition, we see that $\Sigma_n^{(\ell)} = \frac{1}{n}\sum_{k=1}^{n}\left\{X_k^{(\ell)}\right\}\left\{X_k^{(\ell)}\right\}^{\mathsf{T}}$ is an unbiased estimator of $a_\ell\Sigma_\ell$ where $a_\ell = \prod_{k=1(\neq \ell)}^{L}\mathrm{tr}(\Sigma_k)$. Thus, we have that $\otimes_{\ell=1}^{L}\Sigma_n^{(\ell)}$ is an unbiased estimate of $\{\prod_{\ell=1}^{L}\mathrm{tr}(\Sigma_\ell)\}^{L-1}\Gamma$. Replacing $\prod_{\ell=1}^{L}\mathrm{tr}(\Sigma_\ell)$ by its consistent estimate

$b_n = \sum_{k=1}^n \|\mathrm{vec}(X_k)\|^2/n$, we obtain a moment estimate of Γ as

$$\Gamma_n = \otimes_{\ell=1}^L \Sigma_n^{(\ell)}/b_n^{L-1}.$$

The Kronecker estimate for array data is a straightforward extension of that for matrix data. Following previous ideas, we can define various estimates similar to those in Section 2. For instance, we minimize (3) by replacing R_n^Ψ by $R^{\Sigma_n^{(l)}}$ to get the Kronecker estimate of $\widehat{R}^{(\ell)}$, where $R^{\Sigma_n^{(l)}}$ is the correlation of $\Sigma_n^{(\ell)}$ and $\lambda^{(\ell)}$ is the penalty parameter. Let $\widehat{\Sigma}^{(\ell)} = W^{(\ell)}\widehat{R}^{(\ell)}W^{(\ell)}$, where $(W^{(\ell)})^2 = \mathrm{diag}(\Sigma_n^{(\ell)})$. The final Kronecker estimate of Γ is

$$\widehat{\Gamma}_{array} = \otimes_{\ell=1}^L \widehat{\Sigma}^{(\ell)}/b_n^{L-1}.$$

To establish the asymptotic results, we can impose similar conditions as those in Section 3, under which one can show that if $(s_\ell+1)\frac{\log p_\ell}{n\prod_{k=1(\neq\ell)}^L p_k} \to 0, \ell \le L$, we have

$$\|\widehat{\Gamma}_{array}-\Gamma\|_2 \le C\sum_{\ell=1}^L \sqrt{(s_\ell+1)\frac{\log p_\ell}{n\prod_{k=1(\neq\ell)}^L p_k}},\ \|\widehat{\Gamma}_{array}-\Gamma\|_F \le C\sum_{\ell=1}^L \sqrt{(s_\ell+p_\ell)\frac{\log p_\ell}{n}},$$

with probability $1-2\sum_{k=1}^L p_k^{-0.9}$. The detail of these results will be pursued elsewhere.

4. Numerical Study

Extensive simulation studies are conducted to assess the finite-sample performance of the proposed estimators. In particular, we consider the following matrices as the building blocks for generating the covariance matrices Ψ and Σ throughout the simulation study. For all the simulation studies, we choose $\epsilon = 10^{-6}$.

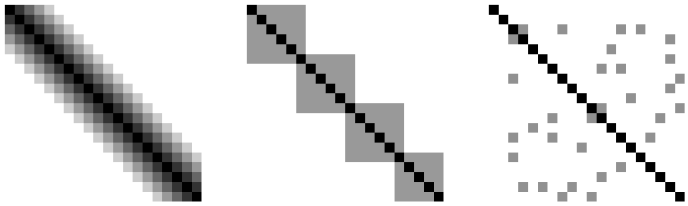
Case 1 (Banded matrix). The (ij)th entry of the matrix is $a_{ij} = (1-\frac{|i-j|}{5})_+$.

Case 2 (Block diagonal matrix). Partition the indices $\{1, 2, \cdots, p\}$ into $K = p/5$ non-overlapping subsets I_k of equal size. Let i_k denote the maximum index in I_k. We set

$$a_{ij} = 0.6I_{\{i=j\}} + 0.4\sum_{i=1}^K I_{\{i\in I_k, j\in J_k\}} + 0.4\sum_{k=1}^{K-1}(I_{\{i=i_k, j\in i_{k+1}\}} + I_{\{i\in I_{k+1}, j=i_k\}}).$$

Case 3 (Random sparse matrix). Let $A = B + \delta I$: each off-diagonal upper triangle entry in B is generated independently and equals to 0.5 with probability 0.1 and 0 with probability 0.9. The diagonals of B are zero and δ is chosen such that the conditional number of A is p.

The first two cases are similar to those in Xue et al. (2012). Case 3, a random sparse covariance matrix, is adopted from Rothman et al. (2008) and Leng and Tang (2012). The patterns of the sparsity of these three matrices can be seen from Figure 1 for $p = 20$, where a random realization of the matrix in Case 3 is illustrated. We denote these three matrices as $A_j,\ j = 1, 2, 3$.



(a) Case 1 (b) Case 2 (c) Case 3

Figure 1. Heat maps of the three matrices used for simulation when $p = 20$. The figure for Case 3 is random realization of the matrix generating process. Black color denotes 1 and white color denotes 0.

We generate 50 data sets for each simulation setup, each data matrix taking the form $X = BSA^{\mathsf{T}}$ where A, B are square matrices such that $AA^{\mathsf{T}} = A_j$, $BB^{\mathsf{T}} = A_k$, $3 \geq j \geq k \geq 1$ and entries of S are independent t_{10} random variates normalized to have variance one. We consider a sample size $n = 20$ and various dimensions as $(p, q) = (20, 20), (320, 320)$ or $(640, 20)$. For each setup, we generate a test data with the same sample size and choose the tuning parameter that minimises the Frobenius norm of the difference between the estimate and the empirical covariance matrix of the test data.

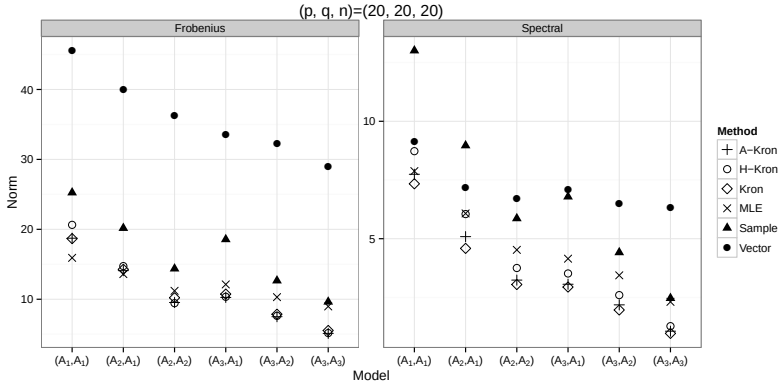


Figure 2. The accuracy when $(p, q, n) = (20, 20, 20)$. A-Kron: the adaptive Kronecker estimator; H-Kron: the hard-thresholding estimator; Kron: the Kronecker estimator in (5); MLE: the maximum likelihood estimator; Sample: the sample estimator in (2); Vector: the vectorized estimator

We first examine the estimation accuracy in the Frobenius and spectral norm respectively between the truth and an estimator. The performance of various estimators of Γ when $(p, q) = (20, 20)$ is presented in Figure 2. When pq is large, computing the sparse vectorized estimator or a spectral norm is very slow. Thus, we only present the accu-

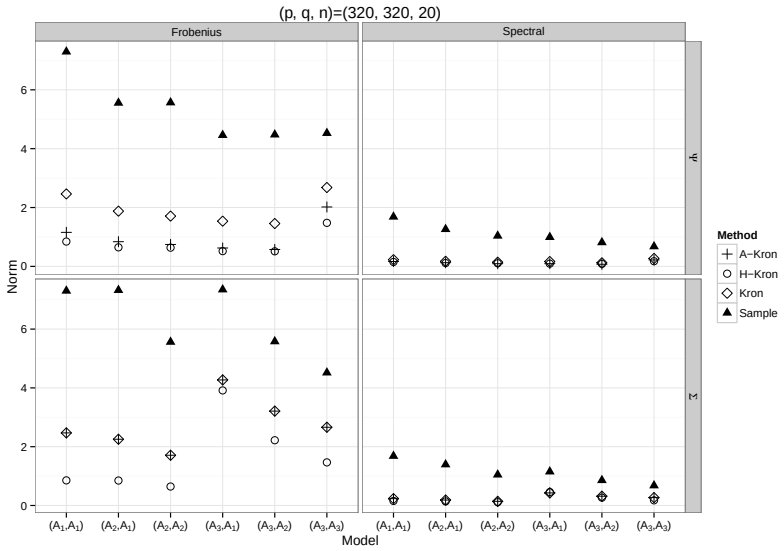


Figure 3. The accuracy when $(p, q, n) = (320, 320, 20)$. Short notations can be found in the caption of Figure 2

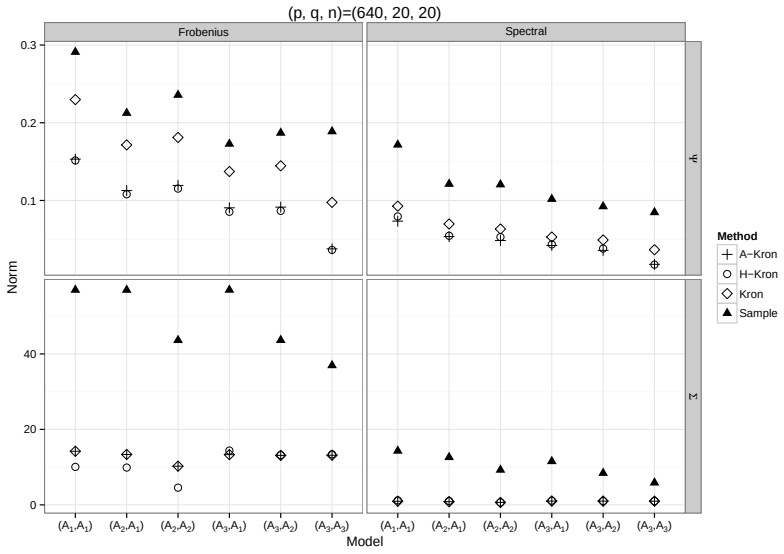


Figure 4. The accuracy when $(p, q, n) = (640, 20, 20)$. Short notations can be found in the caption of Figure 2

racy of estimating Ψ and Σ for $(p, q) = (320, 320)$ in Figure 3 and for $(p, q) = (640, 20)$ in Figure 4 respectively. Here the maximum likelihood estimator is obtained via the flip-flop algorithm in Srivastava et al. (2008) which fails to give convergent solutions if $n < \min\{p/q, q/p\} + 1$ when $(p, q) = (640, 20)$. This phenomenon happens because during the iteration, some of the estimated intermediate matrices are degenerate and thus cannot be used for updating. Finally we also note that the naive sample estimator $\tilde{\Gamma}_n$ is outperformed by all the estimators in Figure 2 by a large magnitude (not shown) and thus is omitted for better visualization.

We can draw the following conclusions from these figures. The sample estimator and the sparse vectorized estimator are outperformed by the Kronecker estimators. The maximum likelihood estimator works well when $(p, q) = (20, 20)$, but loses out when dimensionality becomes large. Among the Kronecker estimators, the hard-thresholding estimator and the adaptive Kronecker estimator perform among the best, followed by the Kronecker estimator overall.

For variable selection, we follow Leng and Tang (2012) to record the true positive rate defined as $\#\{\hat{A}_{ij} \neq 0 \ \& \ A_{ij} \neq 0\}/\#\{A_{ij} \neq 0\}$ and the true negative rate defined as $\#\{\hat{A}_{ij} = 0 \ \& \ A_{ij} = 0\}/\#\{A_{ij} = 0\}$. From Table 1, we can see that the Kronecker estimators perform satisfactorily in general, especially so for the hard-thresholding and adaptive estimators.

An interesting question arises regarding how robust the method is if the assumed Kronecker structure is not true. Towards this end, we generate data by assuming

$$\Gamma = \alpha\Psi \otimes \Sigma + (1 - \alpha)I,$$

where Ψ and Σ are specified as previously, I is an $pq \times pq$ dimensional identity matrix and $\alpha \in [0, 1]$ is a constant. Apparently, whenever $\alpha \in (0, 1)$, the assumed Kronecker structure does not hold. We find that under this perturbation scheme, the proposed method continues to perform better than the sparse vectorized estimator, an example of which with $\alpha = 0.5$ is illustrated in Figure 5. This is most likely due to the reduction of the large number of parameters, the simple structure of Γ and the small sample size $n = 20$. We have also conducted additional simulations by assuming $\Gamma = A_i, i = 1, 2, 3$ which does not admit the Kronecker structure. We have observed that the proposed methods continue to outperform the sparse vectorized estimator.

4.1. A data analysis

As an illustration, we apply the proposed covariance matrix estimation method to analyze the Neuro Bureau ADHD-200 preprocessed data (http://www.nitrc.org/frs/?group_id=383). We examine the resting state functional magnetic resonance imaging (fMRI) data collected by Oregon Health Sciences University by focusing on the 42 typically developing children. These children served as the baseline for comparison to those diagonized with attention deficit hyperactivity disorder (ADHD). For this dataset, we examine the so-called automated anatomical labeling (AAL) atlas with 116 regions of interest (ROI). The labels in the atlas indicate macroscopic brain structures which were

Table 1. Model selection result in percentages. TPR, true positive rate; TNR, true negative rate; Kron, Kronecker estimator; A-Kron, adaptive Kronecker estimator; H-Kron, Kronecker estimator via hard-thresholding.

			$(p,q,n)=(320,320,20)$			$(p,q,n)=(640,20,20)$		
(Ψ,Σ)	Matrix		Kron	A-Kron	H-Kron	Kron	A-Kron	H-Kron
(A_1,A_1)	Ψ	TPR	100	100	100	100	100	100
		TNR	92	99	100	42	93	98
	Σ	TPR	100	100	100	92	88	78
		TNR	92	99	100	94	99	100
(A_2,A_1)	Ψ	TPR	100	100	100	100	100	100
		TNR	91	99	100	39	95	97
	Σ	TPR	100	100	100	100	99	92
		TNR	92	99	100	95	99	100
(A_2,A_2)	Ψ	TPR	100	100	100	100	100	100
		TNR	91	100	100	51	99	100
	Σ	TPR	100	100	100	100	100	99
		TNR	91	100	100	95	99	100
(A_3,A_1)	Ψ	TPR	100	100	100	100	100	100
		TNR	85	100	100	45	96	99
	Σ	TPR	100	99	91	2	2	2
		TNR	81	90	81	99	100	100
(A_3,A_2)	Ψ	TPR	100	100	100	100	100	100
		TNR	91	100	100	50	99	100
	Σ	TPR	100	100	99	7	2	2
		TNR	75	93	100	99	100	100
(A_3,A_3)	Ψ	TPR	100	100	100	100	100	100
		TNR	84	97	84	75	100	100
	Σ	TPR	100	100	100	17	5	2
		TNR	84	97	100	97	99	100

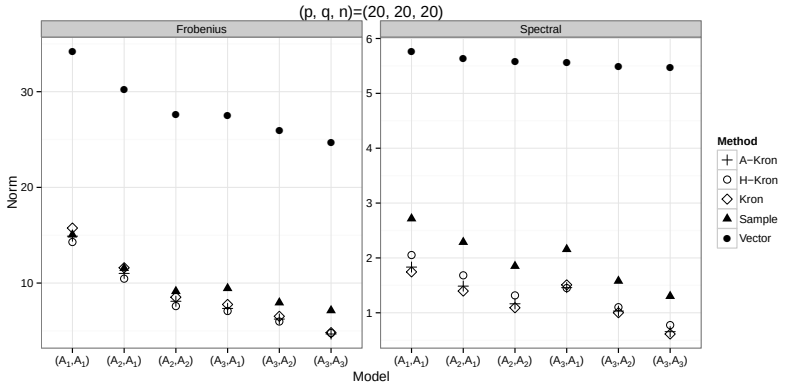
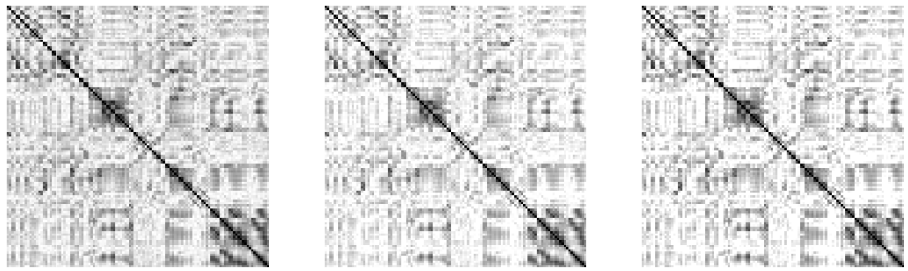
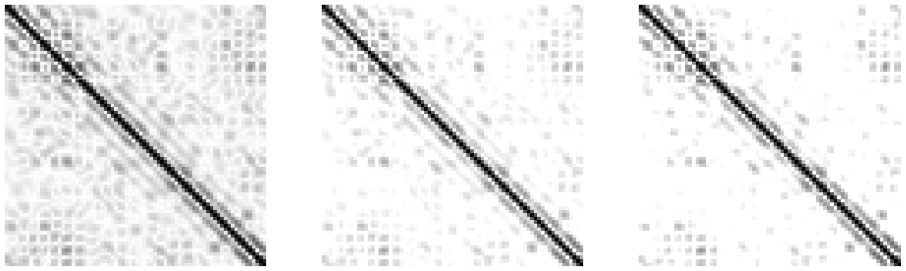


Figure 5. The accuracy when $(p,q,n)=(20,20,20)$ when the true covariate matrix does not admit the Kronecker structure. Short notations can be found in the caption of Figure 2

obtained by fractionating the brain into functional space using nearest-neighbor interpolation (Tzourio-Mazoyer et al., 2002). In an fMRI study, brain activities are measured by detecting associated changes in blood flow through low frequency blood oxygenation level dependent (BOLD) signal in the brain. For our data, the signals of these children were recorded over 74 scans equally spaced in time. Thus, this dataset consists of $n = 42$ observations, each of which can be seen as a $p \times q$ matrix with a temporal dimension $p = 74$ and a spatial dimension $q = 116$.



(a) Sample R^{Ψ}, sparsity=0% (b) Kronecker R^{Ψ}, sparsity=18.6% (c) Adaptive Kronecker R^{Ψ}, sparsity=39.4%



(d) Sample R^{Σ}, sparsity=0% (e) Kronecker R^{Σ}, sparsity=53.7% (f) Adaptive Kronecker R^{Σ}, sparsity=74.4%

Figure 6. The sample Ψ and Σ and their estimators based on the method in this paper. The sparsity is the percentage of the zeros in the corresponding matrix.

To explore the covariance structure of the data, we start by vectorizing the matrices as vectors each with a dimension $pq = 8584$. We fit the penalized Gaussian graphical model in Yuan and Lin (2007) for estimating a sparse Γ^{-1} and the sparse correlation matrix estimation method in Cui et al. (2016) for estimating a sparse Γ. Using 10-fold cross validation for choosing their respective tuning parameter, both method estimate Γ as a diagonal matrix. A formal test that the $(pq) \times (pq)$ dimensional correlation matrix is an identity matrix is rejected (Chen et al., 2010). We proceed to use the sparse matrix variate graphical model in Leng and Tang (2012) to estimate Γ whose inverse is represented as

the Kronecker product of a sparse Ψ^{-1} and Σ^{-1}. The tuning parameter is again chosen by 10-fold cross validation and we found that both matrices are estimated as diagonal.

We then apply the Kronecker structured estimators in this paper for which 10-fold cross validation is again used to choose the tuning parameters. We first plot the sample estimators of R^{Ψ} in panel (a) and R^{Σ} in panel (d) of Figure 4.1. Interestingly, the correlation matrix among the scans at different times clearly exhibit a banded structure, indicating that adjacent observations along the temporal dimension are strongly correlated, while distant observations are not as correlated. This agrees with the intuition for fMRI datasets as BOLD signals at consecutive times can be related to each other. We plot the Kronecker estimator of R^{Ψ} in panel (b) and R^{Σ} in panel (e) of Figure 6. The corresponding adaptive Kronecker estimators are plotted in panel (c) and (f). We note that the adaptive Kronecker estimators are much more sparse than their non-adaptive versions. For example, the sparsity of R^{Σ} is 74.4% when the adaptive Kronecker estimator is used, while it is 53.7% for the Kronecker estimator. Comparing the 10-fold cross validation out-of-sample errors, we further find that the errors of the Kronecker and the adaptive Kronecker estimators are much smaller than the sample estimators in panel (a) and (d) of the figure, and much better than the sample estimator when data are treated as vectors. Finally, our Kronecker estimators from panel (c) and (f) confirm that a banded structure for the covariance matrix of the observed BOLD signals over the temporal domain might be appropriate.

5. Conclusion

We have proposed a novel sparse Kronecker structured method for estimating huge dimensional covariance matrices for matrix and array data. Our approach is simple, requiring no-iteration as opposed to iterative procedures for example in Leng and Tang (2012), easy to compute, and enjoy superior non-asymptotic results under flexible distributions as in Section 3. We impose no constraints on the dimensionality of the data compared to the sample size, as opposed to the usual practice in the literature, for example in Bickel and Levina (2008b). This highlights the significant gain in analysing high dimensional structured data by assuming a Kronecker structure on the covariance matrix.

Appendix A

We state a roadmap of the proof to Theorem 1, consisting of two steps.

1. We first establish the following two bounds

$$\Big|\frac{1}{\mathrm{tr}\Sigma}\Psi_n - \Psi\Big|_{\max} \leq C\sqrt{\frac{\log q}{np}} \quad \text{and} \quad \Big|\frac{1}{\mathrm{tr}\Psi}\Sigma_n - \Sigma\Big|_{\max} \leq C\sqrt{\frac{\log p}{nq}}, \tag{6}$$

with certain probabilities. Moderate deviation for martingale is employed to establish (6) under the moment condition. To this end, a key step is to construct a

martingale by appropriately rewriting the entry of the random matrix of interest. Once this is done, the next difficulty is to characterize the difference between the conditional variance and the variance of the martingale, which is accomplished by evaluating its higher moment.

2. We further derive the convergence rate of the correlation matrices by putting the correlation matrices into (3) and (4) in the main paper. Specifically, we obtain the convergence rate of $(\widehat{R}^{\Psi} - R^{\Psi})$ and $(\widehat{R}^{\Sigma} - R^{\Sigma})$ under the spectral norm and the Frobenius norm, respectively.

To prove (6) we first write the expression of the elements of the matrices in (6) as

$$\begin{aligned}\Big|\frac{\Psi_n}{p} - \frac{\mathrm{tr}(\Sigma)\Psi}{p}\Big|_{\max} &= \max_{u,v\le q}\Big|\frac{1}{n}\sum_{k=1}^{n}\Big[\frac{1}{p}\sum_{i=1}^{p}(x_{k,iu}x_{k,iv} - \psi_{uv}\sigma_{ii})\Big]\Big|,\\ \Big|\frac{\Sigma_n}{q} - \frac{\mathrm{tr}(\Psi)\Sigma}{q}\Big|_{\max} &= \max_{u,v\le p}\Big|\frac{1}{n}\sum_{k=1}^{n}\Big[\frac{1}{q}\sum_{j=1}^{q}(x_{k,uj}x_{k,vj} - \psi_{jj}\sigma_{uv})\Big]\Big|.\end{aligned} \tag{7}$$

We then prove (6) by considering the (u,v)th element of the above matrices as independent sums or martingales.

For ease reference we cite a moderate deviation result for martingales in Grama (1997).

Lemma 1. *Let z_n be a martingale difference sequence with respect to the increasing σ-field $\mathfrak{F}_n$. Suppose that for some $\delta > 0$*

$$L_{2\delta}^n = E\sum_{i=1}^{n}|z_i|^{2+2\delta} \to 0, \quad N_{2\delta}^n = E\Big|\sum_{i=1}^{n}E(z_i^2|\mathfrak{F}_{i-1}) - 1\Big|^{1+\delta} \to 0.$$

Suppose that x is such that $1 \le x \le \alpha(L_{2\delta}^n + N_{2\delta}^n)^{-1}$ with $\alpha > 0$. Then

$$P(|\sum_{i=1}^{n} z_i| \ge r) = 2(1-\Phi(r))\Big[1 + \theta C(\alpha,\delta)x^{\frac{1}{3+2\delta}}(L_{2\delta}^n + N_{2\delta}^n)^{\frac{1}{3+2\delta}}\Big],$$

where $|\theta| \le 1$, $C(\alpha,\delta)$ is a constant depending only on α and δ and

$$r^2 = 2\log x - \theta_1 2c(\delta)\log(1+\sqrt{2\log x}),$$

with $0 \le \theta_1 \le 1$ and $c(\delta) = 3 + 6\delta$.

Applying this lemma, we establish the following result whose proof can be found in supplementary materials.

Lemma 2. *Assume that Conditions (A), (B) and (C) are satisfied. Then for some $M > 0$,*

$$P\Big(\Big|\frac{\Psi_n}{tr\Sigma} - \Psi\Big|_{\max} \le \frac{p}{tr\Sigma}\sqrt{M\frac{\log q}{np}}\Big) \ge 1 - q^{-0.94},\ P\Big(\Big|\frac{\Sigma_n}{tr\Psi} - \Sigma\Big|_{\max} \le \frac{q}{tr\Psi}\sqrt{M\frac{\log p}{nq}}\Big) \ge 1 - p^{-0.94}.$$

Proof of Lemma 2. Throughout the paper we use C and C_j to denote constants which may change from line to line.

We only prove the second inequality and the first one can be proved similarly. Define

$$Q_{quv} = \sqrt{nq}\Big[\frac{\Sigma_n}{q} - \frac{\mathrm{tr}(\Psi)\Sigma}{q}\Big]_{uv}.$$

By (7) write

$$Q_{quv} = \frac{1}{\sqrt{nq}}\sum_{k=1}^{n}\Big[\sum_{j=1}^{q}(x_{k,uj}x_{k,vj} - \psi_{jj}\sigma_{uv})\Big] = \frac{1}{\sqrt{nq}}\sum_{j=1}^{q}\Big[\sum_{k=1}^{n}(x_{k,uj}x_{k,vj} - \psi_{jj}\sigma_{uv})\Big]. \quad (8)$$

In order to decompose Q_{nuv} into a sum of some manageable terms, we introduce the following notation. Let

$$\begin{aligned}
A &= (a_{ij})_{q\times q} = (a_1, \ldots, a_q)^{\mathsf{T}}, \quad a_i = (a_{i1}, \ldots, a_{iq})^{\mathsf{T}}, \quad i = 1, \ldots, q,\\
B &= (b_{ij})_{p\times p} = (b_1, \ldots, b_p)^{\mathsf{T}}, \quad b_i = (b_{i1}, \ldots, b_{ip})^{\mathsf{T}}, \quad i = 1, \ldots, p,\\
S_k &= (s_{k,ij})_{p\times q} = (s_{k,1}, \ldots, s_{k,q}), \quad s_{k,\ell} = (s_{k,1\ell}, \ldots, s_{k,p\ell})^{\mathsf{T}}, \quad \ell = 1, \ldots, q,
\end{aligned}$$

Recalling (1) in the main paper and the covariance matrices $\Phi = AA^{\mathsf{T}}$, $\Sigma = BB^{\mathsf{T}}$, we have

$$\phi_{ij} = a_i^{\mathsf{T}}a_j, \quad i,j \le q, \qquad \sigma_{ij} = b_i^{\mathsf{T}}b_j, \quad i,j \le p,\ x_{k,ij} = b_i^{\mathsf{T}}S_k a_j = \sum_{\ell=1}^{q} a_{j\ell}b_i^{\mathsf{T}}s_{k,\ell}. \quad (9)$$

Denote the (i,j) entry of $A^{\mathsf{T}}A$ by $\phi_{ij} = \sum\limits_{k=1}^{q} a_{ki}a_{kj}$.

Using (8) and (9) and the fact that $trA^{\mathsf{T}}A = trAA^{\mathsf{T}}$, we have $Q_{quv} = \sum_{\ell=1}^{q} J_\ell$, where

$$J_\ell = \frac{1}{\sqrt{nq}}\sum_{\alpha<\ell}\phi_{\ell\alpha}(J_{1\alpha\ell} + J_{2\alpha\ell}) + \frac{1}{\sqrt{nq}}\phi_{\ell\ell}J_{3\ell}$$

with $J_{1\alpha\ell} = \sum\limits_{k=1}^{n} b_u^{\mathsf{T}}s_{k,\ell}s_{k,\alpha}^{\mathsf{T}}b_v$, $J_{2\alpha\ell} = \sum\limits_{k=1}^{n} b_u^{\mathsf{T}}s_{k,\alpha}s_{k,\ell}^{\mathsf{T}}b_v$ and $J_{3\ell} = \sum\limits_{k=1}^{n}\Big(b_u^{\mathsf{T}}s_{k,\ell}s_{k,\ell}^{\mathsf{T}}b_v - b_u^{T}b_v\Big)$. Define the σ fields $\mathfrak{F}_\ell = \sigma(s_{km}, k = 1, \cdots, n, m = 1, \cdots, \ell)$. Then one may verify that $E\Big(J_\ell|\mathfrak{F}_{\ell-1}\Big) = 0$. Furthermore a direct calculation indicates that $E|\sum\limits_{\ell}^{q} J_\ell| < \Big(Var(\sum\limits_{\ell}^{q} J_\ell)\Big)^{1/2} < \infty$, as (10) below shows. Therefore $\{J_\ell, \mathfrak{F}_\ell\}$ is a sequence of martingale differences.

We next calculate the variance of Q_{quv}. One may verify that

$$E\Big(\sum_{\alpha<\ell}\phi_{\ell\alpha}J_{1\alpha\ell}\Big)^2 = E\Big(\sum_{\alpha<\ell}\phi_{\ell\alpha}J_{2\alpha\ell}\Big)^2 = nb_u^{T}b_ub_v^{T}b_v\sum_{\alpha<\ell}\phi_{\ell\alpha}^2$$

and

$$E\Big(\sum_{\alpha_1<\ell}\phi_{\ell\alpha_1}J_{1\alpha_1\ell}\Big)\Big(\sum_{\alpha_2<\ell}\phi_{\ell\alpha_2}J_{2\alpha_2\ell}\Big)=n(b_u^Tb_v)^2\sum_{\alpha<\ell}\phi_{\ell\alpha}^2.$$

It follows that the variance of Q_{quv} is

$$Var(Q_{quv})=\sum_{\ell=1}^{q}Var(J_\ell)=\frac{1}{nq}\sum_{\ell=1}^{q}E\Big(\sum_{\alpha<\ell}\phi_{\ell\alpha}(J_{1\alpha\ell}+J_{2\alpha\ell})\Big)^2+\frac{1}{nq}\sum_{\ell=1}^{q}\phi_{\ell\ell}^2E\Big(J_{3\ell}\Big)^2$$

$$=\frac{b_u^Tb_ub_v^Tb_v+(b_u^Tb_v)^2}{q}\sum_{\ell\neq\alpha}^{q}\phi_{\ell\alpha}^2+\frac{1}{q}\sum_{\ell=1}^{q}\phi_{\ell\ell}^2\Big[(Es_{1,11}^4-3)\sum_{i=1}^{p}b_{ui}^2b_{vi}^2+b_u^Tb_ub_v^Tb_v+(b_u^Tb_v)^2\Big].$$

Therefore

$$Var(Q_{quv})\asymp\frac{\|\Phi\|_F^2}{q},\tag{10}$$

where $a_n\asymp b_n$ means that there exist constants c_1 and c_2 such that $c_1a_n\le b_n\le c_2a_n$ as $n\to\infty$.

We now investigate $N_{2\delta}^n$ in Lemma 1. To this end, we first evaluate the terms involved in $E\Big(J_\ell^2|\mathfrak{F}_{\ell-1}\Big)$. Note that

$$E\Big[\Big(\phi_{\ell\ell}J_{3\ell}\Big)^2|\mathfrak{F}_{\ell-1}\Big]=E\Big(\phi_{\ell\ell}J_{3\ell}\Big)^2,$$

$$E\Big[\Big(\sum_{\alpha<\ell}\phi_{\ell\alpha}J_{1\alpha\ell}\times\phi_{\ell\ell}J_{3\ell}\Big)|\mathfrak{F}_{\ell-1}\Big]=\phi_{\ell\ell}\sum_{\alpha<\ell}\Big(\phi_{\ell\alpha}\sum_{k=1}^{n}s_{k,\alpha}^{\mathsf{T}}b_v\Big)\sum_{i=1}^{p}b_{ui}^2b_{vi}Es_{1,11}^3,$$

$$E\Big[\Big(\sum_{\alpha<\ell}\phi_{\ell\alpha}J_{2\alpha\ell}\times\phi_{\ell\ell}J_{3\ell}\Big)|\mathfrak{F}_{\ell-1}\Big]=\phi_{\ell\ell}\sum_{\alpha<\ell}\Big(\phi_{\ell\alpha}\sum_{k=1}^{n}b_u^Ts_{k,\alpha}\Big)\sum_{i=1}^{p}b_{vi}^2b_{ui}Es_{1,11}^3,$$

$$E\Big[\Big(\sum_{\alpha_1<\ell}\phi_{\ell\alpha_1}J_{1\alpha_1\ell}\sum_{\alpha_2<\ell}\phi_{\ell\alpha_2}J_{2\alpha_2\ell}\Big)|\mathfrak{F}_{\ell-1}\Big]=\sum_{\alpha_1<\ell,\alpha_2<\ell}\Big(\phi_{\ell\alpha_1}\phi_{\ell\alpha_2}\sum_{k=1}^{n}s_{k,\alpha_1}^Tb_vb_u^Ts_{k,\alpha_2}\Big)b_u^Tb_v,$$

$$E\Big[\Big(\sum_{\alpha<\ell}\phi_{\ell\alpha}J_{1\alpha\ell}\Big)^2|\mathfrak{F}_{\ell-1}\Big]=b_u^{\mathsf{T}}b_u\sum_{k=1}^{n}\Big(\sum_{\alpha<\ell}\phi_{\ell\alpha}s_{k,\alpha}^{\mathsf{T}}b_v\Big)^2,$$

$$E\Big[\Big(\sum_{\alpha<\ell}\phi_{\ell\alpha}J_{2\alpha\ell}\Big)^2|\mathfrak{F}_{\ell-1}\Big]=b_v^{\mathsf{T}}b_v\sum_{k=1}^{n}\Big(\sum_{\alpha<\ell}\phi_{\ell\alpha}b_u^Ts_{k,\alpha}\Big)^2.$$

It follows that

$$\sum_{\ell}^{q} E\Big(J_\ell^2|\mathfrak{F}_{\ell-1}\Big) = \tfrac{1}{nq}\sum_{\ell}^{q} E\Big[\Big(\sum_{\alpha<\ell}\phi_{\ell\alpha}(J_{1\alpha\ell}+J_{2\alpha\ell})+\phi_{\ell\ell}J_{3\ell}\Big)^2|\mathfrak{F}_{\ell-1}\Big] =$$
$$\tfrac{1}{nq}\sum_{\ell}^{q} E\Big[\Big(\sum_{\alpha<\ell}\phi_{\ell\alpha}(J_{1\alpha\ell}+J_{2\alpha\ell}\Big)^2|\mathfrak{F}_{\ell-1}\Big] + E\Big[\Big(\phi_{\ell\ell}J_{3\ell}\Big)^2|\mathfrak{F}_{\ell-1}\Big]$$
$$+2E\Big[\Big(\sum_{\alpha<\ell}\phi_{\ell\alpha}(J_{1\alpha\ell}+J_{2\alpha\ell}\Big)\phi_{\ell\ell}J_{3\ell}|\mathfrak{F}_{\ell-1}\Big]$$
$$= \tfrac{1}{nq}\sum_{\ell}^{q}\Big[b_u^{\mathsf{T}}b_u\sum_{k=1}^{n}\Big(\sum_{\alpha<\ell}\phi_{\ell\alpha}s_{k,\alpha}^{\mathsf{T}}b_v\Big)^2 + b_v^{\mathsf{T}}b_v\sum_{k=1}^{n}\Big(\sum_{\alpha<\ell}\phi_{\ell\alpha}b_u^T s_{k,\alpha}\Big)^2$$
$$+2\sum_{\alpha_1<\ell,\alpha_2<\ell}\Big(\phi_{\ell\alpha_1}\phi_{\ell\alpha_2}\sum_{k=1}^{n}s_{k,\alpha_1}^T b_v b_u^T s_{k,\alpha_2}\Big)b_u^T b_v + E\Big(\phi_{\ell\ell}J_{3\ell}\Big)^2\Big] + Q_{n8} + Q_{n9},$$

where

$$Q_{n8} = 2Es_{1,11}^3\frac{1}{nq}\sum_{\ell}^{q}\phi_{\ell\ell}\Big[\sum_{\alpha<\ell}\Big(\phi_{\ell\alpha}\sum_{k=1}^{n}s_{k,\alpha}^{\mathsf{T}}b_v\Big)\sum_{i=1}^{p}b_{ui}^2b_{vi}\Big], \tag{11}$$

and

$$Q_{n9} = 2Es_{1,11}^3\frac{1}{nq}\sum_{\ell}^{q}\phi_{\ell\ell}\Big[\sum_{\alpha<\ell}\Big(\phi_{\ell\alpha}\sum_{k=1}^{n}b_u^T s_{k,\alpha}\Big)\sum_{i=1}^{p}b_{vi}^2b_{ui}\Big]. \tag{12}$$

We conclude from (10) and (11) that

$$\sum_{\ell}^{q} E\Big(J_\ell^2|\mathfrak{F}_{\ell-1}\Big) - Var(\sum_{\ell}^{q}J_\ell) = \Big(Q_{n1}+Q_{n2}+Q_{n3}+Q_{n4}+Q_{n5}+Q_{n6}+Q_{n7}+Q_{n8}+Q_{n9}\Big),$$

where

$$Q_{n1} = b_u^Tb_u\Big[\frac{1}{nq}\sum_{\ell}^{q}\sum_{k=1}^{n}\sum_{\alpha<\ell}\phi_{\ell\alpha}^2\Big(s_{k,\alpha}^{\mathsf{T}}b_v\Big)^2 - \frac{b_v^{\mathsf{T}}b_v}{q}\sum_{\ell}^{q}\sum_{\alpha<\ell}\phi_{\ell\alpha}^2\Big], \tag{13}$$

$$Q_{n2} = \frac{2b_u^Tb_u}{nq}\sum_{\ell}^{q}\sum_{k=1}^{n}\sum_{\alpha_1<\alpha_2<\ell}\phi_{\ell\alpha_1}\phi_{\ell\alpha_2}s_{k,\alpha_1}^{\mathsf{T}}b_vs_{k,\alpha_2}^{\mathsf{T}}b_v,$$

$$Q_{n3} = b_v^Tb_v\Big[\frac{1}{nq}\sum_{\ell}^{q}\sum_{k=1}^{n}\sum_{\alpha<\ell}\phi_{\ell\alpha}^2\Big(b_u^Ts_{k,\alpha}\Big)^2 - \frac{b_u^{\mathsf{T}}b_u}{q}\sum_{\ell}^{q}\sum_{\alpha<\ell}\phi_{\ell\alpha}^2\Big],$$

$$Q_{n4} = \frac{2b_v^Tb_v}{nq}\sum_{\ell}^{q}\sum_{k=1}^{n}\sum_{\alpha_1<\alpha_2<\ell}\phi_{\ell\alpha_1}\phi_{\ell\alpha_2}b_u^{\mathsf{T}}s_{k,\alpha_1}b_u^{\mathsf{T}}s_{k,\alpha_2},$$

$$Q_{n5} = 2b_u^Tb_v\Big[\frac{1}{nq}\sum_{\ell}^{q}\sum_{k=1}^{n}\sum_{\alpha<\ell}\Big(\phi_{\ell\alpha}^2s_{k,\alpha}^Tb_vb_u^Ts_{k,\alpha}\Big) - b_u^Tb_v\frac{1}{q}\sum_{\ell}^{q}\sum_{\alpha<\ell}^{q}\phi_{\ell\alpha}^2\Big],$$

$$Q_{n6} = 2b_u^T b_v \Big[\frac{1}{nq}\sum_{\ell}^{q}\sum_{\alpha_1<\alpha_2<\ell}\Big(\phi_{\ell\alpha_1}\phi_{\ell\alpha_2}\sum_{k=1}^{n} s_{k,\alpha_1}^T b_v b_u^T s_{k,\alpha_2}\Big)\Big]$$

and

$$Q_{n7} = 2b_u^T b_v \Big[\frac{1}{nq}\sum_{\ell}^{q}\sum_{\alpha_2<\alpha_1<\ell}\Big(\phi_{\ell\alpha_1}\phi_{\ell\alpha_2}\sum_{k=1}^{n} s_{k,\alpha_1}^T b_v b_u^T s_{k,\alpha_2}\Big)\Big]. \tag{14}$$

In order to offset p^2 caused by max in the inequality (17), below we evaluate the higher moments of $Q_{nj}, j = 1, \cdots, 9$ in Lemma 3 below. By (10) and Lemma 3 we obtain

$$E\Big|\frac{\sum_{\ell}^{q} E\Big(J_\ell^2|\mathfrak{F}_{\ell-1}\Big)}{Var(\sum_{\ell}^{q} J_\ell)} - 1\Big|^{12} \le \frac{Cq^2\|\Phi_1^0(\Phi_1^0)^T\|_F^{12}}{n^6\|\|\Phi\|_F^{24}} + \frac{Cq^6}{n^6\|\|\Phi\|_F^{24}} + \frac{C\|\Phi_1^0(\Phi_1^0)^T\|_F^6}{n^6\|\|\Phi\|_F^{12}} \le \frac{C}{n^6q^3}, \tag{15}$$

where we use Lemma 2.1 of Bhansali et al. (2007).

We next consider $L_{2\delta}^n$ with $\delta = 11$ in condition (1). By Rosenthal's inequality

$$E(J_\ell)^{24} \le \frac{C}{n^{12}q^{12}}E\Big(\sum_{\alpha<\ell}\phi_{\ell\alpha}(J_{1\alpha\ell} + J_{2\alpha\ell})\Big)^{24} + \frac{C}{n^{12}q^{12}}\phi_{\ell\ell}^{24}EJ_{3\ell}^{24} \le \frac{C}{q^{12}}\Big(\sum_{\alpha<\ell}\phi_{\ell\alpha}^2\Big)^{12}.$$

This, together with (10), yields that

$$\frac{1}{\Big(Var(\sum_{\ell}^{q} J_\ell)\Big)^{12}}\sum_{\ell=1}^{q} E(J_\ell)^{24} \le \frac{C}{q^{11}}, \tag{16}$$

where we also use the fact that $\sum_{\alpha<\ell}\phi_{\ell\alpha}^2$ and $\phi_{\ell\alpha}$ are both bounded for any ℓ.

From (15) and (16) we see that $Cq^3 \le \alpha(L_{2\delta}^n + N_{2\delta}^n)^{-1}$ with some appropriate C, independent of q. Therefore choose x to be Cq^3 in Lemma 1. Note that $Var(\sum_{\ell=1}^{q} J_\ell) \le C$ by (10). It follows from Lemma 1 that

$$\begin{aligned} P(\max_{u,v\le p}|\textstyle\sum_{\ell=1}^{q} J_\ell| \ge M\sqrt{\log p}) &\le p^2 P(|\frac{\sum_{\ell=1}^{q} J_\ell}{\sqrt{Var(\sum_{\ell=1}^{q} J_\ell)}}| \ge CM\sqrt{\log p}) \\ &\le p^2 P(|\frac{\sum_{\ell=1}^{q} J_\ell}{\sqrt{Var(\sum_{\ell=1}^{q} J_\ell)}}| \ge \sqrt{2t\log q^3}) \le \tfrac{C}{p^{3t-2}}, \end{aligned} \tag{17}$$

where M is chosen so that $CM\sqrt{\log p} > \sqrt{2t\log q^3}$ with $2 < 3t < 3$. Selecting t so that $3t - 2 = 0.95$ and then summarizing the above, the proof is complete.

Lemma 3. *Recall the definitions of $Q_{nj}, j = 1, \cdots, 9$ in (11)-(14). Then*

$$E(Q_{n1})^{12} + \cdots + E(Q_{n9})^{12} \le \frac{C\|\Phi_1^0(\Phi_1^0)^T\|_F^{12}}{n^6q^{10}} + \frac{C}{n^6q^6} + \frac{C\,\|\,\Phi\,\|_F^{12}\,\|\Phi_1^0(\Phi_1^0)^T\|_F^6}{n^6q^{12}},$$

where $\Phi_1^0 = (\phi_{\ell\alpha}^0)$ stands for the matrix obtained from $A^TA = (\phi_{\ell\alpha})$ with $\phi_{\ell\alpha}^0 = \phi_{\ell\alpha}$ if $\alpha < \ell$ and zero otherwise.

Proof of Lemma 3. Note that the terms Q_{n2}, Q_{n4}, Q_{n6} and Q_{n7} are similar (their upper bounds are the same up to the constants involving $b_u^T b_v, b_u^T b_u, b_v^T b_v$). Therefore we only estimate Q_{n2} below. Define

$$u_{\alpha_1\alpha_2} = \sum_{\ell > \alpha_2} \phi_{\ell\alpha_1}\phi_{\ell\alpha_2}, \qquad v_{\alpha_1\alpha_3} = \sum_{\alpha_2 > max(\alpha_1,\alpha_3)} u_{\alpha_1\alpha_2}u_{\alpha_3\alpha_2}.$$

Write

$$Q_{n2} = \frac{2}{nq}\sum_{k=1}^{n}\sum_{\alpha_1<\alpha_2} u_{\alpha_1\alpha_2}s_{k,\alpha_1}^{\mathsf{T}}b_v s_{k,\alpha_2}^{\mathsf{T}}b_v.$$

By Rosenthal's inequality,

$$\begin{aligned}
E\Big(Q_{n2}\Big)^{12} &\le \frac{C}{n^{12}q^{12}}\Big|\sum_{k=1}^{n}E\Big(\sum_{\alpha_2}\sum_{\alpha_1<\alpha_2} u_{\alpha_1\alpha_2}s_{k,\alpha_1}^{\mathsf{T}}b_v s_{k,\alpha_2}^{\mathsf{T}}b_v\Big)^2\Big|^6 \\
&\quad + \frac{C}{n^{12}q^{12}}\sum_{k=1}^{n}E\Big(\sum_{\alpha_2}\sum_{\alpha_1<\alpha_2} u_{\alpha_1\alpha_2}s_{k,\alpha_1}^{\mathsf{T}}b_v s_{k,\alpha_2}^{\mathsf{T}}b_v\Big)^{12} \\
&\le \frac{C}{n^6q^{12}}\Big|\sum_{\alpha_2}\sum_{\alpha_1<\alpha_2} u_{\alpha_1\alpha_2}^2\Big|^6 + \frac{C}{n^{12}q^{12}}\sum_{k=1}^{n}E\Big(\sum_{\alpha_2}\Big(\sum_{\alpha_1<\alpha_2} u_{\alpha_1\alpha_2}s_{k,\alpha_1}^{\mathsf{T}}b_v\Big)^2\Big)^6 \\
&\quad + \frac{C}{n^{12}q^{12}}\sum_{k=1}^{n}\sum_{\alpha_2}E\Big(\sum_{\alpha_1<\alpha_2} u_{\alpha_1\alpha_2}s_{k,\alpha_1}^{\mathsf{T}}b_v\Big)^{12} \\
&\le \frac{C}{n^6q^{12}}\|\Phi_1^0(\Phi_1^0)^T\|_F^{12} + \frac{C}{n^{12}q^{12}}\sum_{k=1}^{n}E\Big(\sum_{\alpha_1,\alpha_3} v_{\alpha_1\alpha_3}s_{k,\alpha_1}^{\mathsf{T}}b_v s_{k,\alpha_3}^{\mathsf{T}}b_v\Big)^6 \\
&\quad + \frac{C}{n^{12}q^{12}}\sum_{k=1}^{n}\sum_{\alpha_2}\Big(\sum_{\alpha_1<\alpha_2}E\Big(u_{\alpha_1\alpha_2}s_{k,\alpha_1}^{\mathsf{T}}b_v\Big)^2\Big)^6 + \frac{C}{n^{12}q^{12}}\sum_{k=1}^{n}\sum_{\alpha_2}\sum_{\alpha_1<\alpha_2}E\Big(u_{\alpha_1\alpha_2}s_{k,\alpha_1}^{\mathsf{T}}b_v\Big)^{12} \\
&\le \frac{C}{n^6q^{12}}\|\Phi_1^0(\Phi_1^0)^T\|_F^{12} + \frac{C}{n^{12}q^{12}}\sum_{k=1}^{n}E\Big(\sum_{\alpha_3}\Big(\sum_{\alpha_1<\alpha_3} v_{\alpha_1\alpha_3}s_{k,\alpha_1}^{\mathsf{T}}b_v\Big)^2\Big)^3 \\
&\quad + \frac{C}{n^{12}q^{12}}\sum_{k=1}^{n}\sum_{\alpha_3}E\Big(\sum_{\alpha_1<\alpha_3} v_{\alpha_1\alpha_3}s_{k,\alpha_1}^{\mathsf{T}}b_v\Big)^6 + \frac{C}{n^{11}q^{12}}\sum_{\alpha_2}\Big(\sum_{\alpha_1<\alpha_2} u_{\alpha_1\alpha_2}^2\Big)^6 \\
&\le \frac{C}{n^6q^{12}}\|\Phi_1^0(\Phi_1^0)^T\|_F^{12} + \frac{C}{n^{12}q^{10}}\sum_{k=1}^{n}\sum_{\alpha_3}E\Big(\sum_{\alpha_1<\alpha_3} v_{\alpha_1\alpha_3}s_{k,\alpha_1}^{\mathsf{T}}b_v\Big)^6 + \frac{C}{n^{11}q^{12}}\Big(\sum_{\alpha_2}\sum_{\alpha_1<\alpha_2} u_{\alpha_1\alpha_2}^2\Big)^6 \\
&\le \frac{C}{n^6q^{12}}\|\Phi_1^0(\Phi_1^0)^T\|_F^{12} + \frac{C}{n^{11}q^{10}}\sum_{\alpha_3}\Big(\sum_{\alpha_1<\alpha_3} v_{\alpha_1\alpha_3}^2\Big)^3 + \frac{C}{n^{11}q^{10}}\sum_{\alpha_3}\sum_{\alpha_1<\alpha_3} v_{\alpha_1\alpha_3}^6
\end{aligned}$$

$$\leq \frac{C}{n^6q^{12}}\|\Phi_1^0(\Phi_1^0)^T\|_F^{12} + \frac{C}{n^{11}q^{10}}\Big(\sum_{\alpha_3}\sum_{\alpha_1<\alpha_3} v_{\alpha_1\alpha_3}^2\Big)^3$$

$$\leq \frac{C}{n^6q^{12}}\|\Phi_1^0(\Phi_1^0)^T\|_F^{12} + \frac{C}{n^{11}q^{10}}\Big(\sum_{\alpha_3}\sum_{\alpha_1<\alpha_3}\sum_{\alpha_2>\alpha_1} u_{\alpha_1\alpha_2}^2 \sum_{\alpha_2>\alpha_3} u_{\alpha_3\alpha_2}^2\Big)^3$$

$$\leq \frac{C}{n^6q^{10}}\|\Phi_1^0(\Phi_1^0)^T\|_F^{12},$$

where the third step uses the fact that

$$\sum_{\alpha_2}\Big(\sum_{\alpha_1<\alpha_2} u_{\alpha_1\alpha_2}s_{k,\alpha_1}^{\mathsf{T}}b_v\Big)^2 = \sum_{\alpha_1,\alpha_3} v_{\alpha_1\alpha_3}s_{k,\alpha_1}^{\mathsf{T}}b_v s_{k,\alpha_3}^{\mathsf{T}}b_v,$$

and the fact that

$$\sum_{\alpha_2}\sum_{\alpha_1<\alpha_2} u_{\alpha_1\alpha_2}^2 = \sum_{\ell_1,\ell_2}\sum_{\alpha_1<\alpha_2<min(\ell_1,\ell_2)} \phi_{\ell_1\alpha_1}\phi_{\ell_1\alpha_2}\phi_{\ell_2\alpha_1}\phi_{\ell_2\alpha_2} \leq \|\Phi_1^0(\Phi_1^0)^T\|_F^2,$$

and the step next to the last one uses Cauchy's inequality.

Since the terms Q_{n1}, Q_{n3} and Q_{n5} are similar (their upper bounds are the same up to the constants involving $b_u^Tb_v, b_u^Tb_u, b_v^Tb_v$) we only consider Q_{n1} next. Let $v_\alpha = \sum_{\ell>\alpha}\phi_{\ell\alpha}^2$. Write

$$Q_{n1} = \frac{1}{nq}\sum_{k=1}^n\sum_\alpha v_\alpha\Big(s_{k,\alpha}^{\mathsf{T}}b_v\Big)^2 - \frac{b_v^{\mathsf{T}}b_v}{q}\sum_\alpha v_\alpha,$$

It follows that

$$E|Q_{n1}|^{12} \leq \frac{C}{n^{12}q^{12}}\Big(\sum_{k=1}^n E\Big(\sum_\alpha v_\alpha\big(s_{k,\alpha}^{\mathsf{T}}b_vb_v^Ts_{k,\alpha} - b_v^Tb_v\big)\Big)^2\Big)^6$$

$$+\frac{C}{n^{12}q^{12}}\sum_{k=1}^n E\Big(\sum_\alpha v_\alpha\big(s_{k,\alpha}^{\mathsf{T}}b_vb_v^Ts_{k,\alpha} - b_v^Tb_v\big)\Big)^{12}$$

$$\leq \frac{C}{n^{12}q^{12}}\Big(\sum_{k=1}^n\sum_\alpha v_\alpha^2\Big)^6 + \frac{C}{n^{12}q^{12}}\sum_{k=1}^n\sum_\alpha v_\alpha^{12} \leq \frac{C}{n^6q^{12}}\Big(\sum_\alpha v_\alpha^2\Big)^6 \leq \frac{C}{n^6q^6},$$

where we use the fact that v_α is bounded.

We next consider Q_{n8} only since Q_{n8} and Q_{n9} are similar (see (11) and (12)). Note that $|\sum_{i=1}^p b_{ui}^2b_{vi}|$ is bounded. Define $u_\alpha = \sum_{\ell>\alpha}\big(\phi_{\ell\ell}\phi_{\ell\alpha}\big)$. Rewrite Q_{n8} as

$$Q_{n8} = \frac{2Es_{1,11}^3}{nq}\sum_{k=1}^n\sum_\alpha s_{k,\alpha}^{\mathsf{T}}b_vu_\alpha\sum_{i=1}^p b_{ui}^2b_{vi},$$

By Rosenthal's inequality

$$E|Q_{n8}|^{12} \le \frac{C}{n^{12}q^{12}}\Big|\sum_{k=1}^{n}\sum_{\alpha} b_v^T b_v u_\alpha^2\Big|^6 + \frac{C}{n^{12}q^{12}}\sum_{k=1}^{n} E\Big|\sum_{\alpha} s_{k,\alpha}^{\mathsf{T}} b_v u_\alpha\Big|^{12}$$

$$\le \frac{C}{n^6q^{12}}\Big|\sum_{\alpha} u_\alpha^2\Big|^6 + \frac{C}{n^{12}q^{12}}\sum_{k=1}^{n}\Big|\sum_{\alpha} v_v^{\mathsf{T}} b_v u_\alpha^2\Big|^6 + \frac{C}{n^{12}q^{12}}\sum_{k=1}^{n}\sum_{\alpha} E|s_{k,\alpha}^{\mathsf{T}} b_v|^{12} u_\alpha^{12}$$

$$\le \frac{C}{n^6q^{12}}\Big|\sum_{\alpha} u_\alpha^2\Big|^6 + \frac{C}{n^{11}q^{12}}\Big|\sum_{\alpha} u_\alpha^2\Big|^6 + \frac{C}{n^{11}q^{12}}\sum_{\alpha} u_\alpha^{12}$$

$$\le \frac{C}{n^6q^{12}}\Big|\sum_{\alpha} u_\alpha^2\Big|^6 \le \frac{C \parallel \Phi \parallel_F^{12} \parallel \Phi_1^0(\Phi_1^0)^T\parallel_F^6}{n^6q^{12}},$$

because

$$\sum_{\alpha} u_\alpha^2 = \sum_{\ell_1,\ell_2}\sum_{\alpha<min(\ell_1,\ell_2)} \phi_{\ell_1\ell_1}\phi_{\ell_1\alpha_1}\phi_{\ell_2\ell_2}\phi_{\ell_2\alpha_2}$$

$$\le \sum_{\ell}\phi_{\ell\ell}^2\Big(\sum_{\ell_1,\ell_2}\Big|\sum_{\alpha<min(\ell_1,\ell_2)}\phi_{\ell_1\alpha}\phi_{\ell_2\alpha}\Big|^2\Big)^{1/2} \le C \parallel \Phi \parallel_F^2 \parallel \Phi_1^0(\Phi_1^0)^T\parallel_F.$$

Appendix B

The aim of this section is to prove Theorem 1, Corollary 1 and Corollary 2. To simplify the presentation, we define the following two events,

$$\mathcal{X}_\Psi = \Big\{\Big|\frac{1}{\mathrm{tr}\Sigma}\Psi_n - \Psi\Big|_{\max} \le C\sqrt{\frac{\log q}{np}}\Big\},\ \mathcal{X}_\Sigma = \Big\{\Big|\frac{1}{\mathrm{tr}\Psi}\Sigma_n - \Sigma\Big|_{\max} \le C\sqrt{\frac{\log p}{nq}}\Big\},$$

where C is a constant which may have different values in different places. By Lemma 2 we have the following Lemma.

Lemma 4. *Assume that Conditions (A), (B) and (C) are satisfied. We have*

$$P(\mathcal{X}_\Psi) = 1 - q^{-0.94}, \quad P(\mathcal{X}_\Sigma) = 1 - p^{-0.94}.$$

In the following, to ease the presentation, we assume that the two events $\mathcal{X}_\Psi$ and $\mathcal{X}_\Sigma$ hold.

We next derive the convergence rate regarding the correlation matrices R_n^Ψ and R_n^Σ.

Lemma 5. *Assume that the events $\mathcal{X}_\Psi$ and $\mathcal{X}_\Sigma$ happen. We have*

$$\big|R_n^\Psi - R^\Psi\big|_{\max} \le C\sqrt{\frac{\log q}{np}}, \quad \big|R_n^\Sigma - R^\Sigma\big|_{\max} \le C\sqrt{\frac{\log p}{nq}},$$

where C is a positive constant.

Proof of Lemma 5. We only prove the first inequality while the second one can be similarly proved. Recall $R_n^\Psi = (W_1^\Psi)^{-1}\Psi_n(W_1^\Psi)^{-1}$ and $R^\Psi = (W^\Psi)^{-1}\Psi(W^\Psi)^{-1}$, where $(W^\Psi)^2 = \mathrm{diag}(\Psi)$. When the event $\mathcal{X}_\Psi$ happens, we have $\left|\frac{1}{\mathrm{tr}\Sigma}(W_1^\Psi)_{ii}^2 - \psi_{ii}\right| \le C\sqrt{\frac{\log q}{np}}$ for all $i \le q$. Then, there exist positive constants c and C such that

$$c \le |(W_1^\Psi)_{ii}/\sqrt{\mathrm{tr}\Sigma}| \le C, \qquad \left|\frac{1}{\sqrt{\mathrm{tr}\Sigma}}(W_1^\Psi)_{ii} - \sqrt{\psi_{ii}}\right| \le C\sqrt{\frac{\log q}{np}}.$$

It follows that

$$\left|(R_n^\Psi)_{ij} - (R^\Psi)_{ij}\right| = \left|\frac{\frac{1}{\mathrm{tr}\Sigma}(\Psi_n)_{ij}}{(W_1^\Psi)_{ii}(W_1^\Psi)_{jj}/\mathrm{tr}\Sigma} - \frac{\psi_{ij}}{\sqrt{\psi_{ii}\psi_{jj}}}\right| \le C\sqrt{\frac{\log q}{np}}.$$

Proof of Theorem 1. Suppose that the events $\mathcal{X}_\Psi$ and $\mathcal{X}_\Sigma$ happen. Let

$$\widehat{\Delta} = \widehat{R}^\Psi - R^\Psi = \arg\min_\Delta F(\Delta) \ \text{ s.t. } R^\Psi + \Delta \succeq \epsilon I_q \text{ and } \Delta_{jj} = 0,$$

where $F(\Delta) = \frac{1}{2}\|R^\Psi + \Delta - R_n^\Psi\|_F^2 + \lambda_\Psi|R^\Psi + \Delta|_1$.

Let $\lambda_\Psi = C\sqrt{\frac{\log q}{np}}$. Then by Lemma 5, we have $\left|R_n^\Psi - R^\Psi\right|_{\max} \le \lambda_\Psi$. Let A_0 be the matrix constructed from R^Ψ by replacing its nonzero entries with 1 and Δ_{A_0} be the Hadamard product $\Delta \circ A_0 = (\Delta_{ij} \cdot A_{o,ij})$. Consider for any $\Delta \in \{\Delta : \Delta = \Delta^T, R^\Psi + \Delta \succeq \epsilon I_q, \Delta_{jj} = 0, \|\Delta\|_F \ge 4(s_\Psi + 1)^{1/2}\lambda_\Psi\}$. Then one can see that

$$\begin{aligned}
F(\Delta) - F(0) &= \frac{1}{2}\|R^\Psi + \Delta - R_n^\Psi\|_F^2 - \frac{1}{2}\|R^\Psi - R_n^\Psi\|_F^2 + \lambda_\Psi(|R^\Psi + \Delta|_1 - |R^\Psi|_1) \\
&= \frac{1}{2}\|\Delta\|_F^2 + \langle \Delta, R^\Psi - R_n^\Psi\rangle + \lambda_\Psi|\Delta_{A_0^c}|_1 + \lambda_\Psi(|\Delta_{A_0} + R_{A_0}^\Psi|_1 - |R_{A_0}^\Psi|_1) \\
&\ge \frac{1}{2}\|\Delta\|_F^2 - \lambda_\Psi|\Delta|_1 + \lambda_\Psi|\Delta_{A_0^c}|_1 - \lambda_\Psi|\Delta_{A_0}|_1 = \frac{1}{2}\|\Delta\|_F^2 - 2\lambda_\Psi|\Delta_{A_0}|_1 \\
&\ge \frac{1}{2}\|\Delta\|_F^2 - 2\lambda_\Psi\sqrt{s_\Psi}\|\Delta\|_F > 0.
\end{aligned}$$

Here in the second inequality, we use the fact that $\Delta_{jj} = 0$ and

$$|\Delta_{A_0}|_1 = \sum_{i\ne j}\Delta_{ij} \cdot A_{0,ij} \le (s_\Psi \sum_{i,j}\Delta_{ij}^2)^{1/2} \le \sqrt{s_\Psi}\|\Delta\|_F.$$

By the convexity of the objective function $F(\Delta)$, we immediately see that the global optimizer must satisfy

$$\|\widehat{R}^\Psi - R^\Psi\|_F^2 \le 16(s_\Psi + 1)\lambda_\Psi^2 \le C(s_\Psi + 1)\frac{\log q}{np}.$$

Appealing to the same method, we also have

$$\|\widehat{R}^\Sigma - R^\Sigma\|_F^2 \le C(s_\Psi + 1)\frac{\log p}{nq}.$$

Hence, via Lemma 4, we have proved Theorem 1.

Proof of Corollary 1. Assume that the two events $\mathcal{X}_\Psi$ and $\mathcal{X}_\Sigma$ happen. Note that

$$\widehat{\Psi} = \widehat{W}_1^\Psi \widehat{R}^\Psi \widehat{W}_1^\Psi, \qquad \Psi = W^\Psi R^\Psi W^\Psi.$$

We have

$$\begin{aligned} \frac{1}{\mathrm{tr}\Sigma}\widehat{\Psi} - \Psi &= \frac{1}{\mathrm{tr}\Sigma}\widehat{W}_1^\Psi(\widehat{R}^\Psi - R^\Psi)\widehat{W}_1^\Psi \\ &+ (\frac{1}{\sqrt{\mathrm{tr}\Sigma}}\widehat{W}_1^\Psi - W^\Psi)R^\Psi\frac{1}{\sqrt{\mathrm{tr}\Sigma}}\widehat{W}_1^\Psi + W^\Psi R^\Psi(\frac{1}{\sqrt{\mathrm{tr}\Sigma}}\widehat{W}_1^\Psi - W^\Psi). \end{aligned} \tag{18}$$

Consider the spectral norm of $\frac{1}{\mathrm{tr}\Sigma}\widehat{\Psi} - \Psi$ first. Since $\widehat{W}_1^\Psi, W^\Psi$ are diagonal matrices, we obtain

$$\Big\|\frac{1}{\sqrt{\mathrm{tr}\Sigma}}\widehat{W}_1^\Psi\Big\|_2 \le C, \quad \Big\|\frac{1}{\sqrt{\mathrm{tr}\Sigma}}\widehat{W}_1^\Psi - W^\Psi\Big\|_2 \le \Big|\frac{1}{\sqrt{\mathrm{tr}\Sigma}}\widehat{W}_1^\Psi - W^\Psi\Big|_{\max} \le C\sqrt{\frac{\log q}{np}}.$$

Hence, by (18), we have

$$\begin{aligned} \Big\|\frac{1}{\mathrm{tr}\Sigma}\widehat{\Psi} - \Psi\Big\|_2 &\le \Big\|\frac{1}{\sqrt{\mathrm{tr}\Sigma}}\widehat{W}_1^\Psi\Big\|_2^2 \cdot \Big\|\widehat{R}^\Psi - R^\Psi\Big\|_F + \Big\|\frac{1}{\sqrt{\mathrm{tr}\Sigma}}\widehat{W}_1^\Psi - W^\Psi\Big\|_2 \cdot \Big\|R^\Psi\Big\|_2 \cdot \Big\|\frac{1}{\sqrt{\mathrm{tr}\Sigma}}\widehat{W}_1^\Psi\Big\|_2 \\ &+ \Big\|W^\Psi\Big\|_2 \cdot \Big\|R^\Psi\Big\|_2 \cdot \Big\|\frac{1}{\sqrt{\mathrm{tr}\Sigma}}\widehat{W}_1^\Psi - W^\Psi\Big\|_2 \le C\sqrt{(s_\Psi+1)\frac{\log q}{np}}. \end{aligned}$$

Consider the Frobenius norm now. Note that

$$\Big\|\frac{1}{\sqrt{\mathrm{tr}\Sigma}}\widehat{W}_1^\Psi\Big\|_F \le C\sqrt{q}, \quad \Big\|\frac{1}{\sqrt{\mathrm{tr}\Sigma}}\widehat{W}_1^\Psi - W^\Psi\Big\|_F \le \sqrt{q}\Big|\frac{1}{\sqrt{\mathrm{tr}\Sigma}}\widehat{W}_1^\Psi - W^\Psi\Big|_{\max} \le C\sqrt{\frac{q\log q}{np}}.$$

The above result, together with the formula $\|AB\|_F \le \|A\|_2 \cdot \|B\|_F$, implies

$$\begin{aligned} \Big\|\frac{1}{\mathrm{tr}\Sigma}\widehat{\Psi} - \Psi\Big\|_F &\le \Big\|\frac{1}{\sqrt{\mathrm{tr}\Sigma}}\widehat{W}_1^\Psi\Big\|_2^2 \cdot \Big\|\widehat{R}^\Psi - R^\Psi\Big\|_F \\ &+ \Big\|\frac{1}{\sqrt{\mathrm{tr}\Sigma}}\widehat{W}_1^\Psi - W^\Psi\Big\|_F \cdot \Big\|R^\Psi\Big\|_2 \cdot \Big\|\frac{1}{\sqrt{\mathrm{tr}\Sigma}}\widehat{W}_1^\Psi\Big\|_2 \\ &+ \Big\|W^\Psi\Big\|_2 \cdot \Big\|R^\Psi\Big\|_2 \cdot \Big\|\frac{1}{\sqrt{\mathrm{tr}\Sigma}}\widehat{W}_1^\Psi - W^\Psi\Big\|_F \\ &\le C(\sqrt{s_\Psi+1}+\sqrt{q})\sqrt{\frac{\log q}{np}} \le C\sqrt{(s_\Psi+q)\frac{\log q}{np}}. \end{aligned}$$

Similarly, these arguments also work for the estimate $\widehat{\Sigma}$. Hence when $\mathcal{X}_\Psi$ and $\mathcal{X}_\Sigma$ happen, we have

$$\Big\|\frac{1}{\mathrm{tr}\Psi}\widehat{\Sigma} - \Sigma\Big\|_2 \le C\sqrt{(s_\Sigma+1)\frac{\log p}{nq}}, \ \Big\|\frac{1}{\mathrm{tr}\Psi}\widehat{\Sigma} - \Sigma\Big\|_F \le C\sqrt{(s_\Sigma+p)\frac{\log p}{nq}}.$$

Therefore, Corollary 1 follows from Lemma 4.

Proof of Corollary 2. Define the following event

$$\mathcal{X}_0 = \Big\{ \frac{1}{npq}\sum_{k=1}^{n} \|X_k\|_F^2 - \Big(\frac{1}{q}\mathrm{tr}\Psi\Big)\Big(\frac{1}{p}\mathrm{tr}\Sigma\Big) \le C\sqrt{\frac{\log(pq)}{npq}} \Big\}.$$

By a tedious proof, as in Lemma 2, we have

$$P\Big(\mathcal{X}_0 | \text{under condition (C)}\Big) = 1 - (pq)^{-0.95}.$$

Suppose that the events $\mathcal{X}_\Psi, \mathcal{X}_\Sigma$, and $\mathcal{X}_0$ happen. Applying the formulas $\|A+B\|_F \le \|A\|_F + \|B\|_F$, $\|AB\|_F \le \|A\|_F\|B\|_2$, $\|A\otimes B\|_F = \|A\|_F\|B\|_F$ and Corollary 1, we write

$$\begin{aligned}
\|\widehat{\Gamma} - \Gamma\|_F &= \left\| \frac{\widehat{\Psi}\otimes\widehat{\Sigma}}{\big(\frac{1}{n}\sum_{k=1}^{n}\|X_k\|_F^2\big)} - \Psi\otimes\Sigma \right\|_F \\
&\le \left\| \frac{pq\widehat{\Psi}\otimes\widehat{\Sigma}}{\big(\frac{1}{n}\sum_{k=1}^{n}\|X_k\|_F^2\big)\cdot(\mathrm{tr}\Psi)(\mathrm{tr}\Sigma)} \right\|_F \cdot \left| \frac{1}{npq}\sum_{k=1}^{n}\|X_k\|_F^2 - (\frac{1}{p}\mathrm{tr}\Sigma)(\frac{1}{q}\mathrm{tr}\Psi) \right| \\
&\quad + \left\| \left(\frac{\widehat{\Psi}}{\mathrm{tr}(\Sigma)}\right)\otimes\left(\frac{\widehat{\Sigma}}{\mathrm{tr}(\Psi)}\right) - \Psi\otimes\Sigma \right\|_F \\
&\le \frac{pq\cdot 2\|\Psi\|_F\cdot 2\|\Sigma\|_F}{\mathrm{tr}\Psi\mathrm{tr}\Sigma} C\sqrt{\frac{\log(pq)}{npq}} + \left\| \left(\frac{\widehat{\Psi}}{\mathrm{tr}(\Sigma)} - \Psi\right)\otimes\left(\frac{\widehat{\Sigma}}{\mathrm{tr}(\Psi)} - \Sigma\right) \right\|_F
\end{aligned}$$

$$\begin{aligned}
&+ \left\| \Psi\otimes\left(\frac{\widehat{\Sigma}}{\mathrm{tr}(\Psi)} - \Sigma\right) \right\|_F + \left\| \left(\frac{\widehat{\Psi}}{\mathrm{tr}(\Sigma)} - \Psi\right)\otimes\Sigma \right\|_F \\
&\le C\left(\sqrt{\frac{\log(pq)}{n}} + \sqrt{(s_\Psi+q)(s_\Sigma+p)\frac{\log p\log q}{n^2pq}} + \sqrt{(s_\Psi+q)\frac{\log q}{n}} + \sqrt{(s_\Sigma+p)\frac{\log p}{n}} \right), \\
&\le C\left(\sqrt{(s_\Psi+q)\frac{\log q}{n}} + \sqrt{(s_\Sigma+p)\frac{\log p}{n}} \right),
\end{aligned}$$

where in the last inequality, by condition (B) we use the fact that $\|\Psi\|_F \le \lambda_1(\Psi)\sqrt{q}$ and $\|\Sigma\|_F \le \lambda_1(\Sigma)\sqrt{p}$. Note that $\|A\otimes B\|_2 \le \|A\|_2\cdot\|B\|_2$. From Corollary 1, we have

$$\|\widehat{\Gamma} - \Gamma\|_2 = O_p\left(\sqrt{(s_\Psi+1)\frac{\log q}{np}} + \sqrt{(s_\Sigma+1)\frac{\log p}{nq}} \right).$$

Therefore, Corollary 2 follows from the above inequalities and Lemma 4.

Proof of Corollary 3.. By the Weyl inequality and Theorem 1, we have

$$\lambda_q(\hat{R}^\Psi) \geq \lambda_q(R^\Psi) - \lambda_1(\hat{R}^\Psi - R^\Psi) > 0,$$

because $\lambda_q(R^\Psi) \gg C\sqrt{(s_\Psi+1)\frac{\log q}{np}}$ and $\lambda_1(\hat{R}^\Psi - R^\Psi) \leq \|\hat{R}^\Psi - R^\Psi\|_F$. This implies that $\hat{R}^\Psi$ is positive definite with probability tending to one.

Note that the above argument still holds if the constraint $R^\Psi \geq \varepsilon I_q$ is removed (one may also refer to the proof of Theorem 1). It is clear that the minimizing solution to (2) in the main paper without this constraint becomes the soft thresholding covariance estimator. Therefore with probability tending to one, $(\hat{R}^\Psi)_{ij} = sgn(R_n^\Psi)_{ij}(|(R_n^\Psi)_{ij}| - \lambda_\Psi)_+$ where $(b)_+ > 0$ for $b > 0$, and $(b)_+ = 0$ otherwise. In view of the assumption $\min_{(i,j)\in A_{0\Psi}}(R^\Psi)_{ij} \gg \sqrt{\frac{\log q}{np}}$, $\lambda_\Psi = O(\sqrt{\frac{\log q}{np}})$ and Lemma 5, we have $(\hat{R}^\Psi)_{ij} = 0$ for $(i,j) \in A_{0\Psi}^C$ and $i \neq j$, and $(\hat{R}^\Psi)_{ij} \neq 0$ for $(i,j) \in A_{0\Psi}$ and $i \neq j$ with probability tending to one. One can prove a similar result for $\hat{R}^\Sigma$ and details are omitted here.

Appendix C: Cross Validation

We below consider $\hat{\lambda}_\Sigma$ only and $\hat{\lambda}_\Psi$ can be handled similarly where $\hat{\lambda}_\Sigma$ and $\hat{\lambda}_\Psi$ are obtained from cross validation. The proof of Theorem 2 is straightforward by following that of Theorem 1. Indeed, one should notice that Lemmas 4 and 5 have nothing to do with $\hat{\lambda}_\Sigma$ (which is different from the proof in Bickel and Levina (2008b)). Therefore the argument for Theorem 1 is also applicable to Theorem 2 as long as $\hat{\lambda}_\Sigma = O_p(\sqrt{\frac{\log p}{np}})$. Theorem 2 immediately implies Corollary 4 as in the proof of Corollary 1.

Below we also provide an alternative proof for Corollary 4 (which is enough for the resulting estimator) since our cross validation is based on sample covariance matrices. Recall that the matrix data are generated from normal random matrices $X = BSA^T$, where $\{S\} = (s_{ij})$ consists of i.i.d standard normal random variables. Denote the spectral decomposition of A^TA by $U\Lambda U^T$ with $\Lambda = diag(\lambda_1, \cdots, \lambda_q)$. Define $\Sigma_1^0 = \frac{1}{tr\Psi}BSA^TAS^TB$. Following Lemma A.2 in Bickel and Levina (2008b) one can prove that for symmetric matrix V with $\|V\|_F = 1$

$$P\Big(\frac{\sqrt{q}}{\sqrt{p}}\Big|trV\Sigma_1^0 - \sum\gamma_j\Big| \geq t\Big) \leq Ke^{-\delta t(1+o(1))}, \tag{19}$$

where $t \to \infty$, $\gamma_1, \cdots, \gamma_p$ are eigenvalues of B^TVB and $K > 0, \delta > 0$. Indeed, recalling the notations above (9) by the normality of s_{ij} one can see that $trV\Sigma_1^0$ has the same distribution as

$$\frac{1}{tr\Psi}\sum_{j=1}^{q}\sum_{i=1}^{p}\gamma_i\lambda_j s_{ij}^2.$$

Hence one may repeat the argument for Lemma A.2 in Bickel and Levina (2008b) to obtain (19). Applying (19) and repeating the arguments for Lemma A.3 in Bickel and

Levina (2008b) we have $\rho(J) \leq \frac{C(\log J)^2 p}{q}$, where $\rho(J)$ is the upper bound involved in condition A2 of Theorem 3 in Bickel and Levina (2008b).

Moreover if we change all $\Omega_p(r_n)$ in Theorem 3 of Bickel and Levina (2008b) to all $O_P(r_n)$, the conclusion still holds. Indeed, solving the quadratic inequality in terms of $a_n^{1/2}$ (proved similarly to Theorem 3 of Bickel and Levina (2008b))

$$a_n \leq a_n^{1/2} o_P(r_n^{1/2}) + r_n(1 + o_P(1))$$

yields $a_n = O_p(r_n)$, so that Theorem 3 of Bickel and Levina (2008b) holds for $O_P(r_n)$. Then the argument for proving Theorem 4 of Bickel and Levina (2008b) is also applicable here. Hence Corollary 4 follows.

Acknowledgements

We are grateful to Prof. Holger Dette, an associate editor and two anonymous referees for their constructive comments that have led to a much improved paper.

References

Allen, G. I. and Tibshirani, R. (2010). Transposable regularized covariance models with an application to missing data imputation. *Ann. Appl. Statist.* **4**, 764–790.

Bellec, P., Chu, C., Chouinard-Decorte, F., Benhajali, Y., Margulies, D. S., and Craddock, R. C. (2017). The Neuro Bureau ADHD-200 Preprocessed Repository. *Neuroimage* **144**, 275–286.

Bhansali, R., Giraitis, L. and Kokoszka, P. (2007). Convergence of quadratic forms with nonvanishing diagonal. *Statist. Prob. Letters* **77**, 726–734.

Bickel, P. and Levina, E. (2008a). Regularized estimation of large covariance matrices. *Ann. Statist.* **36**, 199–227.

Bickel, P. and Levina, E. (2008b). Covariance regularization by thresholding. *Ann. Statist.* **36**, 2577–2604.

Bien, J. and Tibshirani, R. (2011). Sparse estimation of a covariance matrix. *Biometrika* **98**, 807–820.

Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Am. Statist. Assoc.* **106**, 1–13.

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

Cui, Y., Leng, C. and Sun, D. (2016). Sparse estimation of high-dimensional correlation matrices. *Comp. Statist. Data Anal.* **93**, 390–403.

Grama, I. G. (1997). On moderate deviations for margingales. *Ann. Prob.* **25**, 152–183.

Gupta, A. K. and Nagar, D. K. (2000). *Matrix Variate Distributions*. Chapman and Hall/CRC, London.

Hitczenko, P. (1990). Best constants in martingale version in Rosenthal's inequality. *Ann. Prob.* **18**, 1656–1668.

Hoff, P. (2011). Separable covariance arrays via the Tucker product, with applications to multivariate relational data. *Bayes. Anal.* **6**, 179–196.

Johnson, W., Schechtman, G. and Zinn, J. (1985). Best constants in moment inequalities for linear combinations of independent and exchangeable random variables. *Ann. Prob.* **13**, 234–253.

Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review* **51**, 455–500.

Leng, C. and Tang, C. Y. (2012). Sparse matrix graphical models. *J. Am. Statist. Assoc.* **107**, 1287–1300.

Li, B., Kim, M. K. and Altman, N. (2010). On dimension folding of matrix- or array-valued statistical objects. *Ann. Statist.* **38**, 1094–1121.

Petrov, V. V. (1975). *Sums of Independent Random Variables*. Springer, New York.

Rothman, A. (2012). Positive definite estimators of large covariance matrices. *Biometrika* **99**, 733–740.

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electro. J. Statist.* **2**, 494–515.

Rothman, A., Levina, L., and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *J. Am. Statist. Assoc,* **104**, 177–186.

Srivastava, M. S., von Rosen, T. and von Rosen, D. (2008). Models with a Kronecker product covariance structure: Estimation and testing. *Math. Meth. Statist.* **17**, 357–370.

Tsiligkaridis, T. and Hero, A. O. (2013). Covariance estimation in high dimensions via Kronecker product expansions. *IEEE Tran. Sig. Proc.* **61**, 5347–5360.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage, **15**, 273–289.

Xue, L., Ma, S. and Zou, H. (2012). Positive definite L_1 penalized estimation of large covariance matrices. *J. Am. Statist. Assoc.* **107**, 1480–1491.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94**, 19–35.

Yin, J. and Li, H. (2012). Model selection and estimation in the matrix normal graphical model. *J. Multiv. Anal.* **107**, 119–140.

Zahn, J. M., Poosala, S., Owen, A., Ingram, K., Lustig, A., et al. (2007). Agemap: a gene expression database for aging in mice. *PLoS Genet.* **3**, 2326–2337.

Zhou, H., Li, L. and Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. *J. Am. Statist. Assoc.* **108**, 540–552.

Zhou, J., Bhattacharya, A., Herring, A. H., and Dunson, D. B. (2014). Bayesian factorizations of big sparse tensors. *J. Am. Statist. Assoc.*, to appear.

Zhou, S. (2014). Gemini: Graph estimation with matrix variate normal instances. *Ann. Statist.* **42**, 532–562.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *J. Am. Statist. Assoc.* **101**, 1418–1429.