Journal of Multivariate Analysis 128 (2014) 165–185

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Graphical model selection and estimation for high dimensional tensor data

Shiyuan He[a], Jianxin Yin[a,*], Hongzhe Li[b], Xing Wang[a]

[a] *Center for Applied Statistics and School of Statistics, Renmin University of China, No. 59 Zhongguancun Street, Haidian District, Beijing 100872, China*

[b] *Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104-6021, USA*

ARTICLE INFO

Article history:
Received 12 December 2012
Available online 31 March 2014

AMS subject classifications:
62F10
62F12

Keywords:
Gaussian graphical model
Gene networks
l_1 penalized likelihood
Oracle property
Tensor normal distribution

ABSTRACT

Multi-way tensor data are prevalent in many scientific areas such as genomics and biomedical imaging. We consider a K-way tensor-normal distribution, where the precision matrix for each way has a graphical interpretation. We develop an l_1 penalized maximum likelihood estimation and an efficient coordinate descent-based algorithm for model selection and estimation in such tensor normal graphical models. When the dimensions of the tensor are fixed, we drive the asymptotic distributions and oracle property for the proposed estimates of the precision matrices. When the dimensions diverge as the sample size goes to infinity, we present the rates of convergence of the estimates and sparsistency results. Simulation results demonstrate that the proposed estimation procedure can lead to better estimates of the precision matrices and better identifications of the graph structures defined by the precision matrices than the standard Gaussian graphical models. We illustrate the methods with an analysis of yeast gene expression data measured over different time points and under different experimental conditions.

1. Introduction

An increasing number of statistical and data mining problems involves analysis of data that are indexed by more than one way. This type of data is often called the multidimensional matrix, multi-way array or tensor [2]. Recently high-dimensional tensor data have become prevalent in many scientific areas such as genomics, biomedical imaging, remote sensing, bibliometrics, chemometrics and internet. Take a two-way $n \times p$ data matrix as an example, if n samples are not independent, their correlations should be taken into consideration in statistical modeling, which leads to a transposable matrix [1]. In genomic experiments, gene expression data are often collected at different time points during the cell cycle process and under varying experimental conditions. This gives rise to a 3-way tensor data [8]. In social-economics studies, export of commodity k from country i to country j at year t [4] defines a three-way tensor data.

Statistical methods for tensor data analysis are limited. Omberg et al. [8] developed tensor higher-order singular value decomposition for integrative analysis of DNA microarray data from different studies. Tucker and parallel factor analysis (PARAFAC) are useful methods for tensor decomposition [5]. When modeling high dimensional tensor data, a separable covariance matrix structure is often assumed. Such a separable structure on the covariance matrix can dramatically reduce the dimension of the parameter space. Consider a four-way tensor data, suppose that the dimensions are $m_1 = m_2 = m_3 = 100$

[*] Corresponding author.
E-mail addresses: jyin@ruc.edu.cn, jianxinyin@gmail.com (J. Yin).

http://dx.doi.org/10.1016/j.jmva.2014.03.007
0047-259X/© 2014 Elsevier Inc. All rights reserved.

and $m_4 = 10$. The nonseparable model requires a joint covariance matrix of $10^7 \times 10^7$ entries, while the separable model requires only three 100×100 matrices and one 10×10 matrix for each way. The joint covariance matrix is simply the Kronecker product of the matrices over all dimensions. The ratio of dimension between two models is almost of the order of 10^{10}.

In this paper, we consider sparse modeling of the precision matrices of K-way tensor data, assuming a separable covariance matrix structure. The corresponding precision matrices define graphical models for tensor data. In many applications, sparsity in each of the corresponding precision matrices can be assumed to facilitate the interpretation. In addition, tensor normality is a natural assumption for the data distribution when the data are continuous [4]. With the separability assumption on the covariance matrix, the joint covariance matrix of the vectorization of the tensor can be obtained by a Kronecker product of K covariance matrices.

When $K = 2$, the 2-way normal tensor data are also called matrix normal data. Yin and Li [13] discussed the sparse model selection and estimation for the matrix normal distribution using a penalized likelihood approach with Lasso and adaptive Lasso penalties. In their work, the dimensions for row and column can diverge to infinity when the sample size goes to infinity. Other related works in modeling matrix-normal data include [1,16,15,12].

In this paper, we generalize the work by Yin and Li [13] to K-way tensor data and focus our work on graphical model selection and estimation. We develop a penalized maximum likelihood estimation with an adaptive Lasso penalty. The consistency and oracle property are obtained when the tensor dimensions hold fixed. In addition, we derive the rate of convergence and prove sparsistency of the estimates when the dimensions diverge with sample size going to infinity. We further show that the effective sample size for estimating the covariance matrix in each way of the tensor is the product of the number of independent samples and the dimensions of the other $K - 1$ matrices. It is worth noting that this effective sample size is usually very large, hence the convergence is quite fast and the high dimension is actually a bless. Our simulation study demonstrates the high accuracy in estimating the precision matrices with small sample size N.

The rest of the paper is organized as follows. A brief summary of multi-way tensor data is presented in Section 2. Section 3 introduces the definition of the array normal distribution of [4] and its estimation in high dimensional settings. The convexity and optimization of the objective function is discussed in Section 4. In Section 5, the asymptotic properties are derived both for the case of fixed dimensions and the case of diverging dimensions when the sample size goes to infinity. A Monte Carlo simulation study is presented in Section 6. Finally, a 3-way tensor data set on gene expressions [8] is analyzed in Section 7.

2. Multi-way tensor data structure and operations

This section presents a brief summary of multi-way array data or high order tensor data [4,2]. Tensor data are higher order parallels of vector and matrix. Entries in a vector can be indexed by a single index set, while a matrix is indexed by two sets (row and column). In the following presentation, we use non-bold italic letters for scalars, bold-faced lower case letters for vectors, and bold-faced capitals for matrices and the multi-way tensor. For a matrix $\mathbf{A}$, we use $\mathbf{a}(j)$ to denote its j-th column, $\mathbf{a}[i]$ its i-th row, and $A(i, j)$ its (i, j)-th element. Standard matrix identities and inequalities used in this paper can be found in [9].

A K-way tensor is an arrangement of elements, which is indexed by K sets. Suppose $\mathbf{Y}$ is a K-way tensor with dimensions $\{m_1, m_2, \ldots, m_K\}$, then the total number of elements of $\mathbf{Y}$ is $m = m_1 \times m_2 \times \cdots \times m_K$. All the elements in Y are

$$\{y_{(i_1,\ldots,i_K)} : i_k = 1, 2, \ldots, m_k;\ k = 1, 2, \ldots, K\}.$$

Clearly, $\mathbf{Y}$ is a vector when $K = 1$ and a matrix when $K = 2$. We further introduce the notation $\mathbf{Y}_{(\cdots,i_k^0,\ldots)}$, which is a $(K-1)$-subarray of $\mathbf{Y}$. Specifically, $\mathbf{Y}_{(\cdots,i_k^0,\ldots)}$ has the same elements as $\mathbf{Y}$, except that its k-th sub-index is fixed at i_k^0. In other words, all the elements in $\mathbf{Y}_{(\cdots,i_k^0,\ldots)}$ are

$$\{y_{(i_1,\ldots,i_k^0,\ldots,i_K)} : i_h = 1, 2, \ldots, m_h;\ h = 1, 2, \ldots, k-1, k+1, \ldots, K\}.$$

To analyze the properties of the K-way tensor, it is helpful to relate the tensor with vector or matrix. The vectorization of $\mathbf{Y}$ is a vector of dimension m,

$$\begin{aligned}\mathrm{vec}(\mathbf{Y}) = \big(&y_{(1,1,1,\ldots,1)}, y_{(2,1,1,\ldots,1)}, \ldots, y_{(m_1,1,1,\ldots,1)},\\ &y_{(1,2,1,\ldots,1)}, y_{(2,2,1,\ldots,1)}, \ldots, y_{(m_1,2,1,\ldots,1)},\\ &\ldots,\\ &y_{(1,m_2,1,\ldots,1)}, y_{(2,m_2,1,\ldots,1)}, \ldots, y_{(m_1,m_2,1,\ldots,1)},\\ &\ldots,\\ &y_{(1,m_2,m_3,\ldots,m_K)}, y_{(2,m_2,m_3,\ldots,m_K)}, \ldots, y_{(m_1,m_2,\ldots,m_K)}\big)^T.\end{aligned}$$

To be explicit, $y_{(i_1,\ldots,i_K)}$ is the j-th element of $\mathrm{vec}(\mathbf{Y})$ with

$$j = \sum_{k=2}^{K}\Big[(i_k - 1)\Big(\prod_{l=1}^{k-1} m_l\Big)\Big] + i_1.$$

On the other hand, k-mode matrix unfolding results in a $m_k \times (m/m_k)$ matrix, $\mathbf{Y}_{(k)}$, whose i_k^0-th row is $[\mathrm{vec}(\mathbf{Y}_{(\cdots,i_k^0,\ldots)})]^T$ for $i_k^0 = 1, 2, \ldots, m_k$.

The k-mode product of a $m_1 \times \cdots \times m_K$ K-array $\mathbf{Y}$ and a $n \times m_k$ matrix $\mathbf{A}$ is a K-array $\mathbf{Z}$ with dimensions $\{m_1, \ldots, m_{k-1}, n, m_{k+1}, \ldots, m_K\}$. The product is denoted by $\mathbf{Y} \times_k \mathbf{A}$, and the $(i_1, \ldots, i_K)$-th element of $\mathbf{Z}$ is

$$z_{(i_1,\ldots,i_K)} = \sum_{l=1}^{m_k} a_{(i_k,\, l)} y_{(i_1,\ldots,\, i_{k-1},\, l,\, i_{k+1},\ldots,\, i_K)}.$$

The Tucker product is defined based on the k-mode product and is useful for the definition of the tensor normal distribution. For a list of matrices $\mathbf{A} = \{\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_K\}$ with $\mathbf{A}_k$ being of dimension $n_k \times m_k$, the Tucker product of a $m_1 \times \cdots \times m_K$ K-way tensor $\mathbf{Y}$ and $\mathbf{A}$ is

$$\mathbf{Y} \times \mathbf{A} = \mathbf{Y} \times_1 \mathbf{A}_1 \times_2 \mathbf{A}_2 \cdots \times_K \mathbf{A}_K.$$

Let $\mathbf{Z} = \mathbf{Y} \times \mathbf{A}$, then we have the following formula that connects the k-mode unfolding and the Tucker product,

$$\mathbf{Z}_{(k)} = \mathbf{A}_k \mathbf{Y}_{(k)} \big(\mathbf{A}_K \otimes \cdots \otimes \mathbf{A}_{k+1} \otimes \mathbf{A}_{k-1} \otimes \cdots \otimes \mathbf{A}_1\big)^T. \tag{1}$$

3. Tensor normal distribution and penalized likelihood estimation

Our main method builds on the tensor normal distribution introduced by Hoff [4]. Without loss of generality, we assume the mean is zero, and our focus is the estimation of covariance and precision matrices. The probability density function of a tensor normal distribution with zero mean and covariances $\boldsymbol{\Sigma} = \{\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K\}$ is

$$p(\mathbf{Y}|\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K) = (2\pi)^{-m/2}\Big(\prod_{k=1}^{K} |\boldsymbol{\Sigma}_k|^{-m/(2m_k)}\Big) \times \exp(-\|\mathbf{Y} \times \boldsymbol{\Sigma}^{-1/2}\|^2/2),$$

where $\|\mathbf{Y}\|^2 = \sum_{i_1,\ldots,i_K} y^2_{(i_1,\ldots,i_K)}$ and $\boldsymbol{\Sigma}^{-1/2} = \{\boldsymbol{\Sigma}_1^{-1/2}, \ldots, \boldsymbol{\Sigma}_K^{-1/2}\}$. The tensor normal distribution is denoted by $\mathbf{Y} \sim$ anorm $(\mathbf{0}, \boldsymbol{\Sigma}_1 \circ \boldsymbol{\Sigma}_2 \circ \cdots \circ \boldsymbol{\Sigma}_K)$. This definition includes vector normal distribution ($K = 1$) and matrix normal distribution ($K = 2$) as special cases. For $k = 1, 2, \ldots, K$, the inverse of $\boldsymbol{\Sigma}_k$ is called the precision matrix or concentration matrix, denoted by $\boldsymbol{\Omega}_k$. For the purpose of identifiability, we assume

$$\Omega_2(1, 1) = \Omega_3(1, 1) = \cdots = \Omega_K(1, 1) = 1, \tag{2}$$

which requires the $(1, 1)$ entries of $\boldsymbol{\Omega}_2, \boldsymbol{\Omega}_3, \ldots, \boldsymbol{\Omega}_K$ to be one.

Derived from (1), some properties for the tensor normal distribution are given below.

Lemma 1. *Let* $\mathbf{Z} = \mathbf{Y} \times \boldsymbol{\Sigma}^{-1/2}$, $\mathbf{V} = \mathbf{Y}_{(k)}(\boldsymbol{\Omega}_K^{1/2} \otimes \cdots \otimes \boldsymbol{\Omega}_{k+1}^{1/2} \otimes \boldsymbol{\Omega}_{k-1}^{1/2} \otimes \cdots \otimes \boldsymbol{\Omega}_1^{1/2})^T$, *and let* $\mathbf{v}(j)$ *be the* j-th *column of* $\mathbf{V}$, *then we have*

(i) $$\|\mathbf{Y} \times \boldsymbol{\Sigma}^{-1/2}\|^2 = \mathrm{tr}\Big(\mathbf{V}^T \boldsymbol{\Omega}_k \mathbf{V}\Big) = \sum_{j=1}^{m/m_k} \mathbf{v}(j)^T \boldsymbol{\Omega}_k \mathbf{v}(j) = \mathrm{vec}(\mathbf{Y})^T(\boldsymbol{\Omega}_K \otimes \cdots \otimes \boldsymbol{\Omega}_1)\mathrm{vec}(\mathbf{Y});$$

(ii) $\mathrm{vec}(\mathbf{Y}) \sim N(\mathbf{0}, \boldsymbol{\Sigma}_K \otimes \boldsymbol{\Sigma}_{K-1} \otimes \cdots \otimes \boldsymbol{\Sigma}_1)$;

(iii) $\mathbf{Y}$ *can be expressed as*

$$\mathbf{Y} = \mathbf{Z} \times \boldsymbol{\Sigma}^{1/2}$$

with $\boldsymbol{\Sigma}^{1/2} = \{\boldsymbol{\Sigma}_1^{1/2}, \ldots, \boldsymbol{\Sigma}_K^{1/2}\}$ *and* $\mathbf{Z} \sim$ anorm$(\mathbf{0}, \mathbf{I}_1 \circ \mathbf{I}_2 \circ \cdots \circ \mathbf{I}_K)$.

Assuming that we have N i.i.d. observations $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_N$ from a tensor normal distribution with zero mean, we are interested in estimating the true covariance matrices $\{\boldsymbol{\Sigma}_1^0, \ldots, \boldsymbol{\Sigma}_K^0\}$ and their corresponding true precision matrices $\{\boldsymbol{\Omega}_1^0, \ldots, \boldsymbol{\Omega}_K^0\}$. In high dimensional settings, under the sparsity assumption of the precision matrices, we propose to estimate these K precision matrices by maximizing the following penalized likelihood function,

$$\frac{1}{N}\sum_{n=1}^{N} \log(p(\mathbf{Y}_n|\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_K)) - \sum_{k=1}^{K} \lambda_k \sum_{i \neq j} p\big(\Omega_k(i,j)\big) = -\frac{m}{2}\log(2\pi) + \sum_{k=1}^{K} \frac{m}{2m_k}\log|\boldsymbol{\Omega}_k| - \frac{1}{2N}\sum_{n=1}^{N} \mathrm{vec}(\mathbf{Y}_n)^T(\boldsymbol{\Omega}_K \otimes \cdots \otimes \boldsymbol{\Omega}_1)\mathrm{vec}(\mathbf{Y}_n) - \sum_{k=1}^{K} \lambda_k \sum_{i \neq j} p\big(\Omega_k(i,j)\big), \tag{3}$$

where $\Omega_k(i,j)$ is the (i,j)-th element of $\boldsymbol{\Omega}_k$ and λ_k's are the tuning parameters. We focus on the ℓ_1 or Lasso penalty $p(\cdot) = |\cdot|$ and the adaptive Lasso penalty $p(\cdot) = |\cdot|/|\widetilde{\Omega}_k(i,j)|^{\gamma}$ where $\widetilde{\Omega}_k(i,j)$ is a consistent estimator of $\Omega_k(i,j)$.

Maximizing (3) is equivalent to minimizing

$$q(\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_K) = -\sum_{k=1}^{K} \frac{m}{m_k}\log|\boldsymbol{\Omega}_k| + \mathrm{tr}\big[\mathbf{S}(\boldsymbol{\Omega}_K \otimes \cdots \otimes \boldsymbol{\Omega}_1)\big] + \sum_{k=1}^{K} \lambda_k \sum_{i \neq j} p\big(\Omega_k(i,j)\big), \tag{4}$$

where $\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}\mathrm{vec}(\mathbf{Y}_n)\mathrm{vec}(\mathbf{Y}_n)^T$. The optimization can now be expressed as

$$\min_{\boldsymbol{\Omega}_1 \succ 0,\ldots,\boldsymbol{\Omega}_K \succ 0} q(\boldsymbol{\Omega}_1,\ldots,\boldsymbol{\Omega}_K).$$

Denote its solution by $\{\widehat{\boldsymbol{\Omega}}_1,\ldots,\widehat{\boldsymbol{\Omega}}_K\}$.

4. Optimization

The block coordinate descent algorithm can be applied to minimize $q(\boldsymbol{\Omega}_1,\ldots,\boldsymbol{\Omega}_K)$, which leads to local optimal solutions. For $k = 1,\ldots,K$, we iteratively minimize the objective function with respect to one $\boldsymbol{\Omega}_k$, while keeping the other matrices $\boldsymbol{\Omega}_j (j \neq k)$ fixed at current values. As a result of Lemma 1(i) and identity (1), minimizing (4) on a specific $\boldsymbol{\Omega}_k$ is equivalent to minimizing

$$q_3(\boldsymbol{\Omega}_k) = -\log|\boldsymbol{\Omega}_k| + \mathrm{tr}[\mathbf{S}_k\boldsymbol{\Omega}_k] + \lambda_k \cdot \frac{m_k}{m}\sum_{i\neq j} p\big(\Omega_k(i,j)\big) \tag{5}$$

with $\mathbf{S}_k = \frac{m_k}{N\cdot m}\sum_{n=1}^{N}\mathbf{V}_{n(k)}[\mathbf{V}_{n(k)}]^T$ and $\mathbf{V}_{n(k)}$ is the k-mode matrix unfolding of the tensor

$$\mathbf{V}_n = \mathbf{Y}_n \times \{\boldsymbol{\Omega}_1^{1/2},\ldots,\ \boldsymbol{\Omega}_{k-1}^{1/2},\ \mathbf{I},\ \boldsymbol{\Omega}_{k+1}^{1/2},\ldots,\ \boldsymbol{\Omega}_K^{1/2}\}.$$

The optimization problem (5) can be solved by the *glasso* algorithm of Friedman et al. [3]. Through minimizing on $\boldsymbol{\Omega}_k$ iteratively, this procedure decreases the objective function after each iteration and eventually converges to a stationary point [11].

The algorithm is summarized below. Let $\{\boldsymbol{\Omega}_1^{(s)}, \boldsymbol{\Omega}_2^{(s)},\ldots,\boldsymbol{\Omega}_K^{(s)}\}$ be the current estimate at the beginning of the s-th iteration.

Algorithm 1.

```
1. $s = 0$, and $\Omega_k^{(0)} = \mathbf{I}$ for $k = 1, 2, \ldots, K$
2. Repeat
3.     s:=s+1
4.     For $k = 1, 2, \ldots, K$
5.         Compute $\mathbf{V}_n := \mathbf{Y}_n \times \Omega^{(s)k}$, where $\Omega^{(s)k}$ is the matrix list
               $\{[\Omega_1^{(s)}]^{1/2},\ldots[\Omega_{k-1}^{(s)}]^{1/2},\mathbf{I},[\Omega_{k+1}^{(s-1)}]^{1/2},\ldots,[\Omega_K^{(s-1)}]^{1/2}\}$
6.         Compute $\mathbf{S}_k^{(s)} = \frac{m_k}{N\cdot m}\sum_{n=1}^{N}\mathbf{V}_{n(k)}[\mathbf{V}_{n(k)}]^T$.
7.         Update $\Omega_k^{(s-1)}$ to $\Omega_k^{(s)}$ by solving the objective function (5).
8.     End For
9. Until Convergence
10. Let $\omega_k = \Omega_k(1,1)$ and $\omega = \prod_{j>1}\omega_j$, and output
        $\{\omega\cdot\Omega_1^{(s)},\ \Omega_2^{(s)}/\omega_2,\ldots,\ \Omega_K^{(s)}/\omega_K\}$
```

Although the objective function $q(\boldsymbol{\Omega}_1,\ldots,\boldsymbol{\Omega}_K)$ is not convex, we show that as $N \to \infty$, the function is strictly quasi-convex with probability 1. To see this, as $N \to \infty$, the limit of the negative log-likelihood function in $q(\boldsymbol{\Omega}_1,\ldots,\boldsymbol{\Omega}_K)$, is

$$\begin{aligned} l(\mathbf{z}) &= -\sum_{k=1}^{K}\frac{m}{m_k}\log|\boldsymbol{\Omega}_k| + \mathrm{tr}\Big((\boldsymbol{\Sigma}_K^0\otimes\cdots\otimes\boldsymbol{\Sigma}_1^0)(\boldsymbol{\Omega}_K\otimes\cdots\otimes\boldsymbol{\Omega}_1)\Big) \\ &= -\sum_{k=1}^{K}\frac{m}{m_k}\log|\boldsymbol{\Omega}_k| + \mathrm{tr}\big(\boldsymbol{\Sigma}_K^0\boldsymbol{\Omega}_K\big)\cdots\mathrm{tr}\big(\boldsymbol{\Sigma}_1^0\boldsymbol{\Omega}_1\big). \end{aligned}$$

With parameters $\mathbf{z} = (\mathrm{vec}(\boldsymbol{\Omega}_1)^T,\ldots,\mathrm{vec}(\boldsymbol{\Omega}_K)^T)^T$, we find its Hessian matrix $\mathbf{L} = \frac{\partial l(\mathbf{z})}{\partial\mathbf{z}\partial\mathbf{z}^T}$. We then treat $\mathbf{L}$ as a block matrix. For $1 \le i,j \le K$, the (i,j)-th block matrix of this Hessian matrix is

$$\mathbf{L}_{(i,j)} = \frac{\partial l(\mathbf{z})}{\partial\mathbf{z}_i\partial\mathbf{z}_j^T} = \begin{cases} \big(m/m_i\big)\times\boldsymbol{\Omega}_i^{-1}\otimes\boldsymbol{\Omega}_i^{-1}, & i = j \\ \Big[\prod_{k\neq i,j}\mathrm{tr}(\boldsymbol{\Sigma}_k^0\boldsymbol{\Omega}_k)\Big]\times\mathrm{vec}(\boldsymbol{\Sigma}_i^0)\mathrm{vec}(\boldsymbol{\Sigma}_j^0)^T, & i\neq j \end{cases}$$

where $\mathbf{z}_i = \mathrm{vec}(\boldsymbol{\Omega}_i)$. Except at $\mathbf{z}^0 = (\mathrm{vec}(\boldsymbol{\Omega}_1^0)^T,\ldots,\mathrm{vec}(\boldsymbol{\Omega}_K^0)^T)^T$, this Hessian matrix cannot be guaranteed to be nonnegative definite. We linearly transform this matrix without changing its eigenvalues. Due to the fact that the diagonal blocks of the Hessian matrix at $\mathbf{z}^0$ are positive definite and the following result in matrix operation [7]

$$\mathrm{vec}(\boldsymbol{\Omega}_i^0)^T(\boldsymbol{\Sigma}_i^0\otimes\boldsymbol{\Sigma}_i^0)\mathrm{vec}(\boldsymbol{\Omega}_i^0) = \mathrm{tr}(\boldsymbol{\Sigma}_i^0\boldsymbol{\Omega}_i^0\boldsymbol{\Sigma}_i^0\boldsymbol{\Omega}_i^0) = m_i$$

the Hessian matrix $\frac{\partial l(\mathbf{z}_0)}{\partial \mathbf{z}\partial \mathbf{z}^T}$ at $\mathbf{z}^0$ can be linearly transformed into a diagonal block matrix $\mathbf{L}' = \mathrm{diag}\{\mathbf{L}'_{(1,1)}, \ldots, \mathbf{L}'_{(K,K)}\}$, and

$$\mathbf{L}'_{(k,k)} = \begin{cases} (m/m_1)\boldsymbol{\Sigma}_1^0 \otimes \boldsymbol{\Sigma}_1^0, & k = 1 \\ (m/m_k)\boldsymbol{\Sigma}_k^0 \otimes \boldsymbol{\Sigma}_k^0 - (m/m_k^2)\mathrm{vec}(\boldsymbol{\Sigma}_k^0)\mathrm{vec}(\boldsymbol{\Sigma}_k^0)^T, & k = 2, \ldots, K. \end{cases}$$

Clearly, its first diagonal block $\mathbf{L}'_{(1,1)}$ is positive definite. For $k = 2, 3, \ldots, K$, its first diagonal block $\mathbf{L}'_{(k,k)}$ has eigenvalues with the following properties:

(E1) One equals 0, with eigenvector $\mathrm{vec}(\boldsymbol{\Omega}_k^0)$;

(E2) The others are positive, with eigenvectors $\mathbf{v}$ satisfying $\mathrm{vec}(\boldsymbol{\Omega}_k^0)^T\mathbf{v} = 0$.

Property (E1) follows from the fact that

$$\begin{aligned}&(m/m_k)\boldsymbol{\Sigma}_k^0 \otimes \boldsymbol{\Sigma}_k^0\mathrm{vec}(\boldsymbol{\Omega}_k^0) - (m/m_k^2)\mathrm{vec}(\boldsymbol{\Sigma}_k^0)\mathrm{vec}(\boldsymbol{\Sigma}_k^0)^T\mathrm{vec}(\boldsymbol{\Omega}_k^0) \\ &\quad = (m/m_k)\mathrm{vec}(\boldsymbol{\Sigma}_k^0\boldsymbol{\Omega}_k^0\boldsymbol{\Sigma}_k^0) - (m/m_k^2)\mathrm{vec}(\boldsymbol{\Sigma}_k^0)\mathrm{tr}(\boldsymbol{\Sigma}_k^0\boldsymbol{\Omega}_k^0) = (m/m_k)\mathrm{vec}(\boldsymbol{\Sigma}_k^0) - (m/m_k^2)\mathrm{vec}(\boldsymbol{\Sigma}_k^0)\cdot m_k = 0.\end{aligned}$$

Property (E2) can be justified as follows. Suppose $\mathbf{v} \neq \mathbf{0}$ is an eigenvector of $\mathbf{L}'_{(k,k)}$ $(2 \le k \le K)$ satisfying $\mathrm{vec}(\boldsymbol{\Omega}_k^0)^T\mathbf{v} = 0$, and suppose ν is its eigenvalue, then

$$(m/m_k)\boldsymbol{\Sigma}_k^0 \otimes \boldsymbol{\Sigma}_k^0\mathbf{v} - (m/m_k^2)\mathrm{vec}(\boldsymbol{\Sigma}_k^0)\mathrm{vec}(\boldsymbol{\Sigma}_k^0)^T\mathbf{v} = \nu \cdot \mathbf{v}.$$

Multiplying both sides from the left by $\mathbf{v}^T\boldsymbol{\Omega}_k^0 \otimes \boldsymbol{\Omega}_k^0$ we get

$$(m/m_k)\mathbf{v}^T\mathbf{v} - (m/m_k^2)\mathbf{v}^T\boldsymbol{\Omega}_k^0 \otimes \boldsymbol{\Omega}_k^0\mathrm{vec}(\boldsymbol{\Sigma}_k^0)\mathrm{vec}(\boldsymbol{\Sigma}_k^0)^T\mathbf{v} = \nu \cdot \mathbf{v}^T\boldsymbol{\Omega}_k^0 \otimes \boldsymbol{\Omega}_k^0\mathbf{v},$$

which implies

$$(m/m_k)\mathbf{v}^T\mathbf{v} - (m/m_k^2)\mathbf{v}^T\mathrm{vec}(\boldsymbol{\Omega}_k^0)\mathrm{vec}(\boldsymbol{\Sigma}_k^0)^T\mathbf{v} = \nu \cdot \mathbf{v}^T\boldsymbol{\Omega}_k^0 \otimes \boldsymbol{\Omega}_k^0\mathbf{v}.$$

Because $\mathrm{vec}(\boldsymbol{\Omega}_k^0)^T\mathbf{v} = 0$ and $\boldsymbol{\Omega}_k^0 \otimes \boldsymbol{\Omega}_k^0$ is positive definite, we have

$$\nu = \frac{m}{m_k} \times \frac{\mathbf{v}^T\mathbf{v}}{\mathbf{v}^T\boldsymbol{\Omega}_k^0 \otimes \boldsymbol{\Omega}_k^0\mathbf{v}} > 0.$$

Thus, (E2) is established.

From this, we know that $\frac{\partial l(\mathbf{z}_0)}{\partial \mathbf{z}\partial \mathbf{z}^T}$ is non-negative definite. As a result, the negative likelihood function is a convex function although not strictly convex. Since the Lasso penalty function is strictly quasi-convex, we have the following lemma.

Lemma 2. *As $N \to \infty$, the limit of the objective function (4) with parameters $\{\boldsymbol{\Omega}_1, \boldsymbol{\Omega}_2, \ldots, \boldsymbol{\Omega}_K\}$ is strictly quasi-convex with probability one at global optimizer $\{\boldsymbol{\Sigma}_1^0, \ldots, \boldsymbol{\Sigma}_K^0\}$.*

5. Asymptotic results

This section discusses the asymptotic behavior for the optimizer of (4). Theorems 1 and 2 assume that the dimensions $(m_1, m_2, \ldots, m_K)$ are fixed, while Theorems 3 and 4 allow the dimensions $(m_1, m_2, \ldots, m_K)$ to diverge with sample size N. For both scenarios, a fast rate of convergence can be guaranteed and the true sparsity pattern of each precision matrix can be recovered by using the adaptive Lasso penalty with probability tending to 1.

For the multi-way tensor normal distribution, the effective sample size for estimating $\boldsymbol{\Omega}_k^0$ is asymptotically $m/m_k \cdot N$, which is larger than N. In fact, if $\boldsymbol{\Omega}_l^0(l \neq k)$'s are known, the correlation on the $l(\neq k)$-th mode can be removed, the columns of the k-mode matrix unfolding can be treated as the i.i.d. samples from the corresponding vector normal distribution, and these column vectors can be pooled together to estimate $\boldsymbol{\Omega}_k^0$. This can be stated precisely in the following lemma. It helps to explain the fast convergence rate in Theorems 2 and 3 and is used in the proofs.

Lemma 3. *Let $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_N$ be N i.i.d. observations from tensor normal distribution* anorm$(\mathbf{0}, \boldsymbol{\Sigma}_1^0 \circ \boldsymbol{\Sigma}_2^0 \circ \cdots \circ \boldsymbol{\Sigma}_K^0)$, *and suppose $\{\boldsymbol{\Sigma}_1^0, \ldots, \boldsymbol{\Sigma}_{k-1}^0, \boldsymbol{\Sigma}_{k+1}^0, \ldots, \boldsymbol{\Sigma}_K^0\}$ are known and $\boldsymbol{\Sigma}_k^0$ is unknown, then the columns $\mathbf{v}_n^{0k}(j)$ of*

$$\mathbf{V}_n^{0k} = \mathbf{Y}_{n(k)}\left[(\boldsymbol{\Omega}_K^0)^{1/2} \otimes \cdots \otimes (\boldsymbol{\Omega}_{k+1}^0)^{1/2} \otimes (\boldsymbol{\Omega}_{k-1}^0)^{1/2} \otimes \cdots \otimes (\boldsymbol{\Omega}_1^0)^{1/2}\right]$$

are i.i.d samples from m_k-vector normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma}_k^0)$, and

$$\mathbf{S}_k = \frac{m_k}{N \cdot m}\sum_{n=1}^{N}\mathbf{V}_n^{0k}\left[\mathbf{V}_n^{0k}\right]^T = \frac{m_k}{N \cdot m}\sum_{n=1}^{N}\sum_{j=1}^{m/m_k}\mathbf{v}_n^{0k}(j)\left[\mathbf{v}_n^{0k}(j)\right]^T \tag{6}$$

estimates $\boldsymbol{\Sigma}_k^0$ with sample size $(m \cdot N)/m_k$. Furthermore, it holds that for some matrix $\mathbf{R}_k$

$$\sqrt{\frac{N \cdot m}{m_k}}\left(vec(\mathbf{S}_k) - vec(\boldsymbol{\Sigma}_k^0)\right) \to N(\mathbf{0}, \mathbf{R}_k)$$

for fixed m, fixed m_k and $N \to \infty$.

Next, Theorem 1 shows the consistency of estimators from (4) with Lasso penalty when the dimensions $(m_1, m_2, \ldots, m_K)$ are fixed. The tuning parameters may change with sample size N, but we omit the subscript N for simplicity.

Theorem 1 (*Consistency*)**.** *For $k = 1, 2, \ldots, K$, assume $\sqrt{N}\lambda_k \to \lambda_{0k}$ for some constants $\lambda_{0k} \geq 0$, and $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_N$ are N i.i.d. observations from tensor normal distribution* anorm$(\mathbf{0}, \boldsymbol{\Sigma}_1^0 \circ \boldsymbol{\Sigma}_2^0 \circ \cdots \circ \boldsymbol{\Sigma}_K^0)$*, then there exists local optimizer $\{\widehat{\boldsymbol{\Omega}}_1, \ldots, \widehat{\boldsymbol{\Omega}}_K\}$ of (4) with the ℓ_1 norm penalty such that:*

$$\sqrt{N}\{(\widehat{\boldsymbol{\Omega}}_1, \ldots, \widehat{\boldsymbol{\Omega}}_K) - (\boldsymbol{\Omega}_1^0, \ldots, \boldsymbol{\Omega}_K^0)\} \to_d \operatorname{argmin}_{(\mathbf{U}_1, \ldots, \mathbf{U}_K)} g(\mathbf{U}_1, \ldots, \mathbf{U}_K)$$

where

$$\begin{aligned} g(\mathbf{U}_1, \ldots, \mathbf{U}_K) = {} & \frac{1}{2}\sum_{k=1}^{K} \frac{m}{m_k} \mathrm{tr}(\mathbf{U}_k \boldsymbol{\Sigma}_k^0 \mathbf{U}_k \boldsymbol{\Sigma}_k^0) + \sum_{i<j} \frac{m}{m_i m_j} \mathrm{tr}(\mathbf{U}_i \boldsymbol{\Sigma}_i^0) \mathrm{tr}(\mathbf{U}_j \boldsymbol{\Sigma}_j^0) \\ & + \sigma \cdot W + \sum_{k=1}^{K} \lambda_{0k} \sum_{i \neq j} \Big(U_k(i,j) \mathit{sign}(\Omega_k^0(i,j)) I\{\Omega_k^0(i,j) \neq 0\} + |U_k(i,j)| I\{\Omega_k^0(i,j) = 0\} \Big), \end{aligned}$$

W is subject to standard normal distribution $N(0, 1)$ and

$$\sigma^2 = \sum_{k=1}^{K} \frac{2m}{m_k} \mathrm{tr}(\mathbf{U}_k \boldsymbol{\Sigma}_k^0 \mathbf{U}_k \boldsymbol{\Sigma}_k^0) + \sum_{i \neq j} \frac{2m}{m_i m_j} \mathrm{tr}(\mathbf{U}_i \boldsymbol{\Sigma}_i^0) \mathrm{tr}(\mathbf{U}_j \boldsymbol{\Sigma}_j^0).$$

With a slight modification of the proof of Theorem 1, we can show that the consistency also holds for the solutions of (4) with adaptive Lasso penalty. The adaptive penalty is introduced for selecting the non-zero entries in the precision matrix and achieving optimal efficiency for them. For $k = 1, 2, \ldots, K$, define the active sets $\mathscr{A}_k = \{(i, j) : \Omega_k^0(i, j) \neq 0\}$ as the set of indices corresponding to non-zero entries in $\boldsymbol{\Omega}_k^0$.

Theorem 2 (*Oracle Property*)**.** *Consider (4) with adaptive Lasso penalty, and let $\gamma > 0$ be a constant and $\widetilde{\boldsymbol{\Omega}}_k$ be $N^{1/2}$-consistent estimators. When $\sqrt{N}\lambda_k \to 0$ and $N^{(\gamma+1)/2}\lambda_k \to \infty$ for $k = 1, 2, \ldots, K$, there exist local solutions of (4) satisfying the oracle property:*

(1) For $k = 1, 2, \ldots, K$ and all $(i, j) \in \mathscr{A}_k^c$, $\widehat{\Omega}_k(i, j) = 0$ with probability tending to 1.
(2) For $k = 1, 2, \ldots, K$ and elements indexed by $(i, j) \in \mathscr{A}_k$,

$$\mathrm{vec}(\widehat{\boldsymbol{\Omega}}_k - \boldsymbol{\Omega}_k^0)_{\mathscr{A}_k} \to_d \mathrm{N}\Big(\mathbf{0}, \frac{m_k}{m}\Big[\big(\boldsymbol{\Omega}_k^0 \otimes \boldsymbol{\Omega}_k^0\big)_{(\mathscr{A}_k, \cdot)}\Big] \mathbf{R}_k \Big[\big(\boldsymbol{\Omega}_k^0 \otimes \boldsymbol{\Omega}_k^0\big)_{(\mathscr{A}_k, \cdot)}\Big]^T\Big)$$

where $\mathbf{R}_k$ is defined in Lemma 3, $\mathrm{vec}(\widehat{\boldsymbol{\Omega}}_k - \boldsymbol{\Omega}_k^0)_{\mathscr{A}_k}$ *is a sub-vector of* $\mathrm{vec}(\widehat{\boldsymbol{\Omega}}_k - \boldsymbol{\Omega}_k^0)$ *with only elements indexed by $\mathscr{A}_k$ preserved; and $\big(\boldsymbol{\Omega}_k^0 \otimes \boldsymbol{\Omega}_k^0\big)_{(\mathscr{A}_k, \cdot)}$ is a sub-matrix of $\boldsymbol{\Omega}_k^0 \otimes \boldsymbol{\Omega}_k^0$ with all rows corresponding to $\mathscr{A}_k^c$ removed. That is, the l-th row of $\boldsymbol{\Omega}_k^0 \otimes \boldsymbol{\Omega}_k^0$ can be preserved if and only if $l = m_k(j-1) + i$ for some $(i, j) \in \mathscr{A}_k$.*

For the tensor normal distribution, the estimators of precision matrices converge much faster than the vector normal case ($K = 1$). For the tensor case, the limiting covariance matrix for the active entry estimator is

$$\frac{m_k}{m}\Big[\big(\boldsymbol{\Omega}_k^0 \otimes \boldsymbol{\Omega}_k^0\big)_{(\mathscr{A}_k, \cdot)}\Big] \mathbf{R}_k \Big[\big(\boldsymbol{\Omega}_k^0 \otimes \boldsymbol{\Omega}_k^0\big)_{(\mathscr{A}_k, \cdot)}\Big]^T$$

while for the vector normal distribution($K = 1$), the limiting covariance matrix is

$$\Big[\big(\boldsymbol{\Omega}_k^0 \otimes \boldsymbol{\Omega}_k^0\big)_{(\mathscr{A}_k, \cdot)}\Big] \mathbf{R}_k \Big[\big(\boldsymbol{\Omega}_k^0 \otimes \boldsymbol{\Omega}_k^0\big)_{(\mathscr{A}_k, \cdot)}\Big]^T.$$

In the former case, the additional factor m_k/m can be quite small when $\prod_{i \neq k} m_i$ is large. This explains the fast rate of convergence in our simulation studies.

This fast rate of convergence is also observed when the dimensions $(m_1, m_2, \ldots, m_K)$ increase with the sample size N. Results similar to [6] hold with much faster rates, as shown in Theorem 3. Again, the results are stated and proven for the ℓ_1 penalty. Similar results hold for the adaptive Lasso penalty. Let $s_k = |\mathscr{A}_k| - m_k$ be the number of non-zero off-diagonal entries in Ω_k, which also varies with sample size N. Under some conditions, convergence in terms of the Frobenius norm can be guaranteed for (4).

Theorem 3 (*Rate of Convergence*)**.** *Assume $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_N$ are N i.i.d. observations from a tensor normal distribution* anorm $(\mathbf{0}, \boldsymbol{\Sigma}_1^0 \circ \boldsymbol{\Sigma}_2^0 \circ \cdots \circ \boldsymbol{\Sigma}_K^0)$*, with dimensions $(m_1, m_2, \ldots, m_K)$ diverging when sample size N goes to infinity.*

In addition, for $k = 1, 2, \ldots, K$ and some constants τ_{k1}, τ_{k2}, assume the eigenvalues are bounded,

$$0 < \tau_{k1} < \lambda_{\min}(\boldsymbol{\Sigma}_k^0) \leq \lambda_{\max}(\boldsymbol{\Sigma}_k^0) < \tau_{k2} < \infty. \tag{7}$$

If the following conditions on tuning parameters λ_k's ($k = 1, \ldots, K$) are satisfied:

$$\frac{m \log m_k}{m_k N} = O(\lambda_k^2) \quad \text{and} \quad \lambda_k^2 = O\Big(\Big(1 + \frac{m_k}{s_k + 1}\Big)\frac{m \log m_k}{m_k N}\Big) \tag{8}$$

then when the Lasso penalty functions are used, there exists a local minimizer $(\widehat{\boldsymbol{\Omega}}_1, \widehat{\boldsymbol{\Omega}}_2, \ldots, \widehat{\boldsymbol{\Omega}}_K)$ *of* (4) *such that*

$$\|\widehat{\boldsymbol{\Omega}}_k - \boldsymbol{\Omega}_k^0\|_F^2 = O_p\Big(m_k(m_k + s_k)\log m_k/(Nm)\Big).$$

Because $\|\mathbf{A}\| \le \|\mathbf{A}\|_F$ for any matrix $\mathbf{A}$, this rate of convergence also holds for the spectral norm. The rate of convergence for the tensor normal distribution is $(m_k/m)(m_k + s_k)\log m_k/N$, which is much faster than the multivariate normal case where $K = 1$. The rate in the latter case is $(m_1 + s_1)\log m_1/N$, as shown in [6]. Clearly, the results also hold for the adaptive Lasso penalty. Furthermore, with adaptive Lasso penalty, we can recover the true sparsity patterns of the precision matrices with probability tending to one, as shown in the following theorem.

Theorem 4 (*Sparsistency*). *Given the conditions in Theorem 3, for $k = 1, \ldots, K$, suppose $\widetilde{\boldsymbol{\Omega}}_k$ is the f_k-consistent estimator for $\boldsymbol{\Omega}_k^0$ in the sense that*

$$f_k\|\widetilde{\boldsymbol{\Omega}}_k - \boldsymbol{\Omega}_k^0\|_\infty = O_p(1).$$

If $\{\widehat{\boldsymbol{\Omega}}_1, \widehat{\boldsymbol{\Omega}}_2, \ldots, \widehat{\boldsymbol{\Omega}}_K\}$ is a local minimizer of (4) *with adaptive Lasso penalty satisfying*

1. $\|\widehat{\boldsymbol{\Omega}}_k - \boldsymbol{\Omega}_k^0\|_F^2 = O_p\{m_k(m_k + s_k)\log m_k/(Nm)\}$*; and*
2. $\|\widehat{\boldsymbol{\Omega}}_k - \boldsymbol{\Omega}_k^0\|^2 = O_p(\eta_n)$ *for a sequence $\eta_n \to 0$*

and if the tuning parameters satisfy

$$f_k^{-2\gamma}\frac{m^2}{m_k^2}\Big(\frac{m_k \log m_k}{mN} + \eta_n + \sum_{l \ne k}\frac{\tau_{l,2}^2}{mN}(m_l + s_l)\log m_l\Big) = O(\lambda_k^2)$$

then with probability tending to one, we have $\widehat{\Omega}_k(i,j) = 0$ for all $(i,j) \in \mathscr{A}_k^c$ and $k = 1, 2, \ldots, K$.

Similar to [13], the sparsistency results require condition (8) to impose both a lower and a upper bound on the rates of the regularization parameters λ_k's in order to control the model sparsity and estimation biases.

6. Monte Carlo simulation studies

6.1. Comparison candidates and measurements

We evaluate the performances of the proposed penalized likelihood estimation for tensor normal data and compare this to two naive methods using simulations.

The first naive method is an approximate maximum likelihood estimation, which is the MLE without a penalty when the effective sample size is larger than the dimensions m_k's and the ℓ_1 penalized estimate otherwise. Statistical tests are used to select edges when the effective sample size is large. Specifically, for $k = 1, 2, \ldots, K$, the effective sample size for estimating $\boldsymbol{\Omega}_k^0$ is approximately $N_k = Nm/m_k$, where N is the true sample size. In Algorithm 1 of Section 4, if $N_k > m_k$, the inverse of $\mathbf{S}_k$ is directly used to update the estimation of $\boldsymbol{\Omega}_k$ in Step 7, which corresponds to the MLE procedure. However, when $N_k \le m_k$, we update the estimation of $\boldsymbol{\Omega}_k$ through (5) with an ℓ_1 Lasso penalty. When $N_k > m_k$, hypothesis tests are also performed to select edges after estimation. Let ρ_{ij} denote the partial correlation between X_i and X_j adjusting for the remaining elements and $\hat{\rho}_{ij}$ denote its MLE estimator, then

$$\frac{1}{2}\log\Big[\frac{1 + \hat{\rho}_{ij}}{1 - \hat{\rho}_{ij}}\Big] \to \mathrm{N}(0, 1/(n - p - 5)).$$

Based on this result, for $k = 1, 2, \ldots, N$ and $i < j$, let

$$\hat{\rho}_{ij}^k = -\frac{\widehat{\Omega}_k(i,j)}{\sqrt{\widehat{\Omega}_k(i,i)\widehat{\Omega}_k(j,j)}}$$

and we set $\widehat{\Omega}_k(i,j) = \widehat{\Omega}_k(j,i) = 0$ whenever $N_k > m_k$ and

$$\left|\frac{1}{2}\log\Big[\frac{1 + \hat{\rho}_{ij}^k}{1 - \hat{\rho}_{ij}^k}\Big]\right| < \frac{z_{\alpha/2}}{\sqrt{N_k - m_k - 5}}$$

where z_β is the upper $\beta \times 100\%$ quantile of the standard normal distribution. We choose $\alpha = 0.1$.

The second naive method estimates each $\boldsymbol{\Omega}_k$ separately with the adaptive Lasso penalty. It treats the other modes as independent, i.e., assuming $\boldsymbol{\Omega}_j = \mathbf{I}_j$ $(j \neq k)$ in the estimation procedure. In this case, Step 5 of Algorithm 1 in Section 4 is not used and $\mathbf{S}_k$ in Step 6 is computed as

$$\mathbf{S}_k = \frac{m_k}{N \cdot m} \sum_{n=1}^{N} Y_{n(k)} [Y_{n(k)}]^T.$$

For the penalized maximum likelihood estimation, we use the adaptive Lasso penalty with an approximate MLE as the initial estimator $\widetilde{\boldsymbol{\Omega}}_k$. The accuracy of the estimated precision matrix is measured by various matrix norms of $\boldsymbol{\Delta}_k = \boldsymbol{\Omega}_k^0 - \widehat{\boldsymbol{\Omega}}_k$, where $\boldsymbol{\Omega}_k^0$ is the true matrix and $\widehat{\boldsymbol{\Omega}}_k$ is the estimated matrix. We consider the following norms: the Frobenius norm $\| \cdot \|_F$, the operator norm $\| \cdot \|_p$, and the entry-wise max norm $||| \cdot |||_\infty$. In addition, the accuracy of recovering the Gaussian graph structure is also measured. Let TP, TN, FP and FN be the numbers of true positives, true negatives, false positives and false negatives, respectively, where the true positives are the true links on the tensor normal graphs. We define specificity (SPE), sensitivity (SEN), and Matthew's correlation coefficient (MCC) as

$$\text{SPE} = \frac{\text{TN}}{\text{TN}+\text{FP}}, \qquad \text{SEN} = \frac{\text{TP}}{\text{TP}+\text{FN}} \qquad \text{MCC} = \frac{\text{TP}\times\text{TN} - \text{FP}\times\text{FN}}{\left\{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})\right\}^{1/2}}.$$

6.2. Models and data generation

The sparse precision matrix $\boldsymbol{\Omega}_k^0$'s are generated as follows. The non-zero off-diagonal elements for the upper triangle of $\boldsymbol{\Omega}_k^0$ are selected independently with probability p_k. For non-zero elements, their values are generated from

$$\Omega_k^0(i,j) = \Omega_k^0(j,i) \sim \text{Uniform}\big([-0.8, -0.2] \cup [0.2, 0.8]\big).$$

We then make $\boldsymbol{\Omega}_k^0$ diagonally dominant by dividing the i-th row by $1.2 \times \sum_{j \neq i} \left|\Omega_k^0(i,j)\right|$ for $i = 1, 2, \dots, m_k$ and then setting all diagonal entries to be 1. We symmetrize the matrix by letting $\boldsymbol{\Omega}_k^0 := \left[\boldsymbol{\Omega}_k^0 + (\boldsymbol{\Omega}_k^0)^T\right]/2$.

The following four models are considered with sample size $N = 10$. These models have different dimensions and different degrees of sparsity as indicated by p_k. The simulations are repeated 50 times.

1. Model 1: three-way tensor data with dimensions (30, 30, 30) and sparsity $p_1 = p_2 = p_3 = 0.1$.
2. Model 2: three-way tensor data with dimensions (6, 6, 500) and sparsity $p_1 = 0.3,\ p_2 = 0.2,\ p_3 = 0.005$.
3. Model 3: four-way tensor data with dimensions (30, 30, 30, 30) and sparsity $p_1 = 0.05,\ p_2 = 0.075,\ p_3 = 0.1, p_4 = 0.2$.
4. Model 4: four-way tensor data with dimensions (30, 40, 50, 100) and sparsity $p_1 = 0.2,\ p_2 = 0.125,\ p_3 = 0.1,\ p_4 = 0.075$.

6.3. Simulation results

For all simulations, the tuning parameters are chosen based on a validation set of sample size of 10. The results are presented in Tables 1–4. In almost all scenarios, the dimensions of the models are larger than the real sample size $N = 10$. However, we observed that the estimates of the precision matrices are still very accurate. This can be explained by the effective sample size, which is very large for each dimension of the tensor data.

For all four models considered, the proposed penalized likelihood procedure results in better estimation of the precision matrices than the two naive methods in terms of estimation errors. For model selection, the penalized likelihood estimation also gives better results, although the performance of the naive method that assumes independency is comparable in certain circumstances. The effect of the effective sample sizes on precision matrix estimation is also clearly demonstrated in these tables. For Model 1, the effective sample size is $10 \times 30 \times 30 = 9000$ for each way of the tensor data. For Model 3, however, the effective sample size for each way of the data is $10 \times 30^3 = 270{,}000$, which is 30 times larger than Model 1. It is clear from Tables 1 and 3 that the estimates for Model 3 are more accurate than these for Model 1. For Model 2, the effective sample size for estimating $\boldsymbol{\Omega}_3^0$ is $6 \times 6 \times 10 = 360$, which is smaller than its dimension of 500, which leads to larger estimation errors.

7. Real data analysis

Omberg et al. [8] considered the expression levels of 4270 genes of *Saccharomyces cerevisiae* during a time course of cell cycle under two different experimental conditions. Each time course was measured at 12 time points with cell cycles synchronized by α-factor pheromone. Under the depleted condition of Cdc6 or Cdc45 (Cdc6-/Cdc45-), the DNA replication initialization is prevented without delaying cell cycle progression. The gene expressions were also measured in the presence of Cdc6 or Cdc45 (Cdc6+/Cdc45+-) without preventing DNA replication. In our analysis, 4720 genes are averaged on observed values of different probes. After averaging and removing the genes with missing values, a total of 404 genes are used in our analysis. Among these genes, 141, 97, 62, 37 and 67 genes are regulated during the G1, G2/M, M/G1, S and S/G2 phases,

Table 1
Model 1: three-way tensor data with dimensions (30, 30, 30), sample size 10 and sparsity $p_1 = p_2 = p_3 = 0.1$. For each measurement, mean and standard error over 50 replications are shown. P-MLE: penalized maximum likelihood estimates; A-MLE: approximate maximum likelihood estimates; I-MLE: penalized maximum likelihood estimates under independency assumption. Δ_k is the difference between the true and the estimated precision matrix for $k = 1, 2, 3$.

		P-MLE	A-MLE	I-MLE
$\boldsymbol{\Omega}_1$	$\|\Delta_1\|_F$	0.15(0.034)	0.26(0.023)	0.23(0.061)
	$\|\Delta_1\|_\infty$	0.09(0.018)	0.20(0.036)	0.14(0.035)
	$\|\Delta_1\|_2$	0.06(0.013)	0.11(0.017)	0.10(0.025)
	$\|\|\|\Delta_1\|\|\|_\infty$	0.04(0.010)	0.05(0.010)	0.07(0.018)
	SPE	0.95(0.010)	0.90(0.014)	0.98(0.007)
	SEN	1.00(0.000)	1.00(0.000)	1.00(0.000)
	MCC	0.79(0.028)	0.67(0.029)	0.92(0.030)
$\boldsymbol{\Omega}_2$	$\|\Delta_2\|_F$	0.15(0.051)	0.25(0.037)	0.27(0.068)
	$\|\Delta_2\|_\infty$	0.10(0.034)	0.20(0.036)	0.18(0.043)
	$\|\Delta_2\|_2$	0.07(0.022)	0.11(0.020)	0.12(0.030)
	$\|\|\|\Delta_2\|\|\|_\infty$	0.04(0.011)	0.05(0.011)	0.07(0.021)
	SPE	0.99(0.002)	0.90(0.016)	0.96(0.009)
	SEN	1.00(0.000)	1.00(0.000)	1.00(0.000)
	MCC	0.99(0.007)	0.70(0.031)	0.87(0.031)
$\boldsymbol{\Omega}_3$	$\|\Delta_3\|_F$	0.18(0.052)	0.27(0.049)	0.28(0.067)
	$\|\Delta_3\|_\infty$	0.11(0.027)	0.20(0.042)	0.18(0.035)
	$\|\Delta_3\|_2$	0.07(0.020)	0.11(0.026)	0.12(0.027)
	$\|\|\|\Delta_3\|\|\|_\infty$	0.05(0.013)	0.05(0.015)	0.08(0.022)
	SPE	1.00(0.002)	0.90(0.016)	0.96(0.010)
	SEN	1.00(0.000)	1.00(0.000)	1.00(0.000)
	MCC	1.00(0.008)	0.71(0.033)	0.86(0.032)

Table 2
Model 2: three-way tensor data with dimensions (6, 6, 500), sample size 10 and sparsity $p_1 = 0.3,\ p_2 = 0.2,\ p_3 = 0.005$. For each measurement, mean and standard error over 50 replications are shown. P-MLE: penalized maximum likelihood estimates; A-MLE: approximate maximum likelihood estimates; I-MLE: penalized maximum likelihood estimates under independency assumption. Δ_k is the difference between the true and the estimated precision matrix for $k = 1, 2, 3$.

		P-MLE	A-MLE	I-MLE
$\boldsymbol{\Omega}_1$	$\|\Delta_1\|_F$	0.03(0.012)	0.04(0.010)	0.05(0.018)
	$\|\Delta_1\|_\infty$	0.04(0.014)	0.04(0.012)	0.05(0.021)
	$\|\Delta_1\|_2$	0.03(0.011)	0.03(0.010)	0.04(0.017)
	$\|\|\|\Delta_1\|\|\|_\infty$	0.02(0.007)	0.02(0.006)	0.03(0.010)
	SPE	0.99(0.034)	0.91(0.115)	0.99(0.030)
	SEN	1.00(0.000)	1.00(0.000)	1.00(0.000)
	MCC	0.99(0.034)	0.91(0.106)	0.99(0.030)
$\boldsymbol{\Omega}_2$	$\|\Delta_2\|_F$	0.03(0.008)	0.03(0.011)	0.03(0.012)
	$\|\Delta_2\|_\infty$	0.02(0.009)	0.03(0.015)	0.03(0.013)
	$\|\Delta_2\|_2$	0.02(0.008)	0.03(0.009)	0.03(0.011)
	$\|\|\|\Delta_2\|\|\|_\infty$	0.02(0.005)	0.02(0.005)	0.02(0.008)
	SPE	1.00(0.000)	0.92(0.082)	0.99(0.013)
	SEN	1.00(0.000)	1.00(0.000)	1.00(0.000)
	MCC	1.00(0.000)	0.88(0.114)	0.99(0.021)
$\boldsymbol{\Omega}_3$	$\|\Delta_3\|_F$	3.64(0.070)	4.82(0.114)	7.11(0.497)
	$\|\Delta_3\|_\infty$	0.95(0.100)	1.36(0.130)	1.82(0.158)
	$\|\Delta_3\|_2$	0.46(0.023)	0.55(0.022)	0.80(0.063)
	$\|\|\|\Delta_3\|\|\|_\infty$	0.27(0.0378)	0.29(0.026)	0.44(0.074)
	SPE	1.00(0.000)	0.98(0.001)	0.99(0.001)
	SEN	0.84(0.015)	0.92(0.010)	0.58(0.023)
	MCC	0.63(0.010)	0.36(0.006)	0.37(0.019)

respectively [10]. We treat this data set as a 3-way tensor data, where the first way is the gene with $m_1 = 404$, the second way is the time point with $m_2 = 12$ and the third way is the condition with $m_3 = 2$. In addition, each sample batch of [8] is treated as an independent sample for a total of $N = 4$ samples. The original expression data are log-transformed. The expression levels of each gene are scaled to zero mean and unit variance across the four samples.

We apply our penalized estimation using the adaptive Lasso penalty to estimate the precision matrices, where the initial estimates are obtained using the ℓ_1 norm penalty. The tuning parameters are selected based on a 4-fold cross-validation. The conditional independency graph for genes that are linked is shown in Fig. 1. The genes that are regulated at the same cell-cycle phases are colored with the same colors. It is interesting to note that genes that are regulated by the same cell cycle phases tend to link together.

Table 3
Model 3: four-way tensor data with dimensions (30, 30, 30, 30) and sample size 10, $p_1 = 0.05,\ p_2 = 0.075,\ p_3 = 0.1,\ p_4 = 0.2$. For each measurement, mean and standard error over 50 replications are shown. P-MLE: penalized maximum likelihood estimates; A-MLE: approximate maximum likelihood estimates; I-MLE: penalized maximum likelihood estimates under independency assumption. Δ_k is the difference between the true and the estimated precision matrix for $k = 1, 2, 3, 4$.

		P-MLE	A-MLE	I-MLE
$\boldsymbol{\Omega}_1$	$\\|\Delta_1\\|_F$	0.02(0.006)	0.04(0.004)	0.04(0.009)
	$\\|\Delta_1\\|_\infty$	0.01(0.003)	0.03(0.005)	0.02(0.005)
	$\\|\Delta_1\\|_2$	0.01(0.003)	0.02(0.002)	0.02(0.005)
	$\vert\vert\vert\Delta_1\vert\vert\vert_\infty$	0.01(0.002)	0.01(0.002)	0.01(0.003)
	SPE	1.00(0.001)	0.90(0.017)	1.00(0.001)
	SEN	1.00(0.000)	1.00(0.000)	1.00(0.000)
	MCC	1.00(0.010)	0.50(0.035)	1.00(0.007)
$\boldsymbol{\Omega}_2$	$\\|\Delta_2\\|_F$	0.02(0.005)	0.04(0.005)	0.05(0.012)
	$\\|\Delta_2\\|_\infty$	0.01(0.004)	0.03(0.007)	0.03(0.007)
	$\\|\Delta_2\\|_2$	0.01(0.003)	0.02(0.003)	0.02(0.006)
	$\vert\vert\vert\Delta_2\vert\vert\vert_\infty$	0.01(0.002)	0.01(0.002)	0.01(0.004)
	SPE	1.00(0.00)	0.90(0.016)	1.00(0.002)
	SEN	1.00(0.00)	1.00(0.000)	1.00(0.000)
	MCC	1.00(0.00)	0.63(0.033)	0.99(0.012)
$\boldsymbol{\Omega}_3$	$\\|\Delta_3\\|_F$	0.03(0.004)	0.04(0.004)	0.05(0.010)
	$\\|\Delta_3\\|_\infty$	0.02(0.003)	0.04(0.006)	0.03(0.007)
	$\\|\Delta_3\\|_2$	0.01(0.002)	0.02(0.003)	0.02(0.004)
	$\vert\vert\vert\Delta_3\vert\vert\vert_\infty$	0.01(0.002)	0.01(0.002)	0.02(0.005)
	SPE	1.00(0.001)	0.90(0.016)	1.00(0.002)
	SEN	1.00(0.000)	1.00(0.000)	1.00(0.000)
	MCC	1.00(0.002)	0.72(0.032)	0.99(0.008)
$\boldsymbol{\Omega}_4$	$\\|\Delta_4\\|_F$	0.03(0.010)	0.05(0.007)	0.07(0.010)
	$\\|\Delta_4\\|_\infty$	0.02(0.006)	0.04(0.006)	0.05(0.008)
	$\\|\Delta_4\\|_2$	0.01(0.004)	0.02(0.003)	0.03(0.005)
	$\vert\vert\vert\Delta_4\vert\vert\vert_\infty$	0.01(0.003)	0.01(0.002)	0.02(0.004)
	SPE	1.00(0.001)	0.90(0.020)	1.00(0.001)
	SEN	1.00(0.000)	1.00(0.000)	1.00(0.000)
	MCC	1.00(0.003)	0.79(0.033)	1.00(0.002)

Table 4
Model 4: four-way tensor data with dimensions (30, 40, 50, 100) and sample size 10, $p_1 = 0.2,\ p_2 = 0.125,\ p_3 = 0.1,\ p_4 = 0.075$. For each measurement, mean and standard error over 50 replications are shown. P-MLE: penalized maximum likelihood estimates; A-MLE: approximate maximum likelihood estimates; I-MLE: penalized maximum likelihood estimates under independency assumption. Δ_k is the difference between the true and the estimated precision matrix for $k = 1, 2, 3, 4$.

		P-MLE	A-MLE	I-MLE
$\boldsymbol{\Omega}_1$	$\\|\Delta_1\\|_F$	0.01(0.002)	0.02(0.001)	0.01(0.002)
	$\\|\Delta_1\\|_\infty$	0.01(0.001)	0.01(0.002)	0.01(0.002)
	$\\|\Delta_1\\|_2$	0.00(0.001)	0.01(0.001)	0.01(0.001)
	$\vert\vert\vert\Delta_1\vert\vert\vert_\infty$	0.00(0.001)	0.00(0.001)	0.00(0.001)
	SPE	1.00(0.000)	0.90(0.018)	1.00(0.001)
	SEN	1.00(0.000)	1.00(0.000)	1.00(0.000)
	MCC	1.00(0.000)	0.80(0.030)	1.00(0.003)
$\boldsymbol{\Omega}_2$	$\\|\Delta_2\\|_F$	0.01(0.002)	0.03(0.002)	0.02(0.004)
	$\\|\Delta_2\\|_\infty$	0.01(0.001)	0.02(0.002)	0.01(0.002)
	$\\|\Delta_2\\|_2$	0.01(0.001)	0.01(0.001)	0.01(0.002)
	$\vert\vert\vert\Delta_2\vert\vert\vert_\infty$	0.00(0.001)	0.00(0.001)	0.00(0.001)
	SPE	1.00(0.000)	0.90(0.010)	1.00(0.002)
	SEN	1.00(0.000)	1.00(0.000)	1.00(0.000)
	MCC	1.00(0.000)	0.73(0.018)	0.99(0.007)
$\boldsymbol{\Omega}_3$	$\\|\Delta_3\\|_F$	0.02(0.004)	0.03(0.002)	0.02(0.004)
	$\\|\Delta_3\\|_\infty$	0.01(0.002)	0.02(0.003)	0.01(0.002)
	$\\|\Delta_3\\|_2$	0.01(0.001)	0.01(0.001)	0.01(0.001)
	$\vert\vert\vert\Delta_3\vert\vert\vert_\infty$	0.00(0.001)	0.00(0.001)	0.01(0.001)
	SPE	1.00(0.000)	0.90(0.010)	0.99(0.002)
	SEN	1.00(0.000)	1.00(0.000)	1.00(0.000)
	MCC	1.00(0.000)	0.69(0.019)	0.97(0.011)
$\boldsymbol{\Omega}_4$	$\\|\Delta_4\\|_F$	0.04(0.007)	0.09(0.004)	0.06(0.008)
	$\\|\Delta_4\\|_\infty$	0.02(0.002)	0.06(0.004)	0.03(0.004)
	$\\|\Delta_4\\|_2$	0.01(0.002)	0.02(0.002)	0.01(0.002)
	$\vert\vert\vert\Delta_4\vert\vert\vert_\infty$	0.01(0.001)	0.01(0.001)	0.01(0.002)
	SPE	1.00(0.000)	0.90(0.004)	1.00(0.000)
	SEN	1.00(0.000)	1.00(0.000)	1.00(0.000)
	MCC	1.00(0.002)	0.63(0.008)	1.00(0.000)

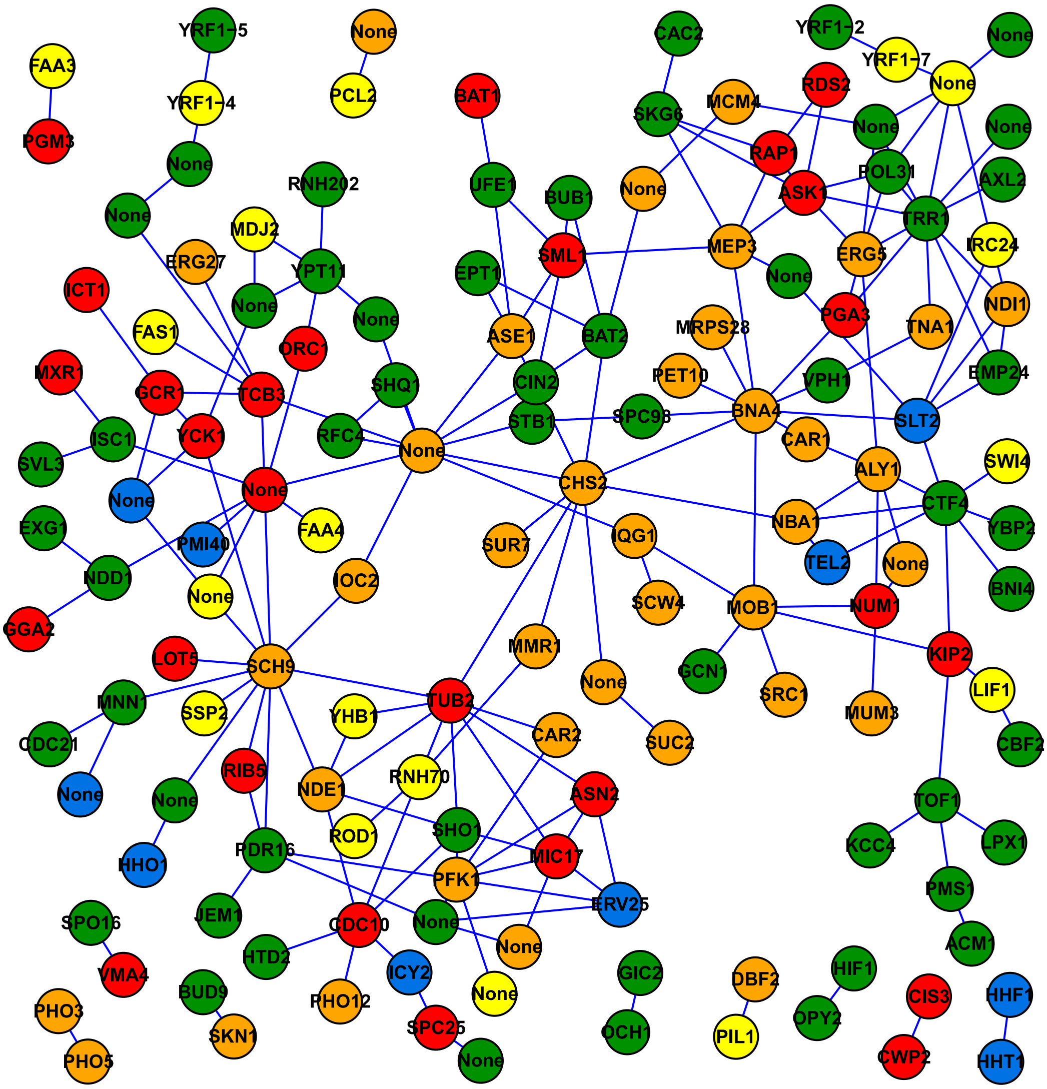
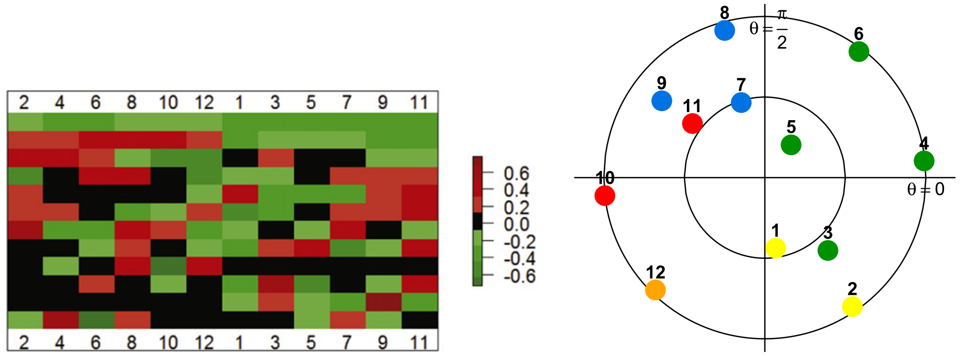


Fig. 1. Gaussian graph of 150 yeast cell cycle associated genes. The colors indicate the cell-cycle phases that the genes are regulated. Green: G1 phase; orange: G2/M; yellow: M/G1; blue: S; Red: S/G2. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 2(a) shows the Raster plot of the eigenvectors of the correlation matrix derived from $\widehat{\boldsymbol{\Sigma}}_2 = \widehat{\boldsymbol{\Omega}}_2^{-1}$. This matrix describes the correlation among the 12 time points during the cell cycle process. Each row of Fig. 2(a) corresponds to an eigenvector sorted by descending eigenvalues. These eigenvectors are the x-eigengenes of [8]. The first x-eigengene represents a constant expression level. The second eigengene represents the contrast in gene expression between the odd and even time points. The third and fourth x-eigengenes reflect the gene expression changes during the cell cycle process. In Fig. 2(b), points are drawn on a plane with the third x-eigengene on the $\theta = 0$-axis and the fourth on the $\theta = \pi/2$-axis, normalized together with the fifth x-eigengene, clearly showing the periodic expression patterns of genes during the cell cycle process.

8. Conclusions and discussion

Motivated by analysis of gene expression data measured at different time points and under different experimental conditions on the same set of samples, we have proposed to apply the tensor normal distribution to model the data jointly and have developed a penalized likelihood method to estimate each way's precision matrix assuming that these matrices are sparse. Our simulation results have clearly demonstrated the proposed penalized estimation method results in better estimates of the precision matrices and better identification of the corresponding graphical structures than naive alternatives. Our theoretical and numerical results show that for the tensor data, the effective sample size for estimating each precision matrix can be quite large although the independent observations are only a few. The tensor normal distribution provides a



(a) Raster plot of the eigenvectors of the correlation matrix of the time points. Each row represents an eigenvector.

(b) Plot of the third and fourth eigenvectors on the $\theta = 0$ and $\theta = \pi/2$ axes.

Fig. 2. Plot of the eigenvectors of the time points correlation matrix based on the estimated time point covariance matrix $\widehat{\boldsymbol{\Sigma}}_2 = \widehat{\boldsymbol{\Omega}}_2^{-1}$.

natural way of modeling the dependency of data indexed by different sets. If the underlying precision matrices are sparse, the proposed penalized likelihood estimation can lead to identification of the non-zero elements in these precision matrices. We observe that the proposed l_1 regularized estimation can lead to better estimates of these sparse precision matrices than the MLEs. How to extend the proposed method to non-normal data is a future research direction.

Acknowledgments

This research was supported by National Natural Science Foundation of China (No. 11201479), Research Fund for the Doctoral Program of Higher Education of China (20120004120007), the Basic Research funds in Renmin University of China from the central government (No. 13XNLJ03, No. 12XNQ049) and NIH grants CA127334 and GM097505.

Appendix

Proof of Theorem 1

Proof. Let $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_K$ be K square matrices of order $m_1, m_2, \ldots, m_K$ respectively. Define a function f_N of them to be

$$f_N(\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_K) = -\sum_{k=1}^{K} \frac{m}{m_k} \log\left|\boldsymbol{\Omega}_k^0 + \frac{\mathbf{U}_k}{\sqrt{N}}\right| + \frac{1}{N}\sum_{n=1}^{N} \mathrm{vec}(\mathbf{Y}_n)^T \left[\left(\boldsymbol{\Omega}_K^0 + \frac{\mathbf{U}_K}{\sqrt{N}}\right) \otimes \cdots \otimes \left(\boldsymbol{\Omega}_1^0 + \frac{\mathbf{U}_1}{\sqrt{N}}\right)\right] \mathrm{vec}(\mathbf{Y}_n)$$
$$+ \sum_{k=1}^{K} \lambda_k \sum_{i \neq j} \left|\Omega_k^0(i,j) + \frac{U_k(i,j)}{\sqrt{N}}\right|.$$

We consider the asymptotic behavior of $N[f_N(\mathbf{U}_1, \ldots, \mathbf{U}_K) - f_N(\mathbf{0}, \ldots, \mathbf{0})]$.

Firstly, the following result is needed for analyzing $N[f_N(\mathbf{U}_1, \ldots, \mathbf{U}_K) - f_N(\mathbf{0}, \ldots, \mathbf{0})]$. Expand the Kronecker product in its first summation to get

$$\frac{1}{N}\sum_{n=1}^{N} \mathrm{vec}(\mathbf{Y}_n)^T \left[\left(\boldsymbol{\Omega}_K^0 + \frac{\mathbf{U}_K}{\sqrt{N}}\right) \otimes \cdots \otimes \left(\boldsymbol{\Omega}_1^0 + \frac{\mathbf{U}_1}{\sqrt{N}}\right)\right] \mathrm{vec}(\mathbf{Y}_n)$$
$$- \frac{1}{N}\sum_{n=1}^{N} \mathrm{vec}(\mathbf{Y}_n)^T (\boldsymbol{\Omega}_K^0 \otimes \cdots \otimes \boldsymbol{\Omega}_1^0) \mathrm{vec}(\mathbf{Y}_n) = \frac{1}{N}\sum_{q>0} \frac{1}{N^{(q/2)}} \sum_{n=1}^{N} \mathrm{vec}(\mathbf{Y}_n)^T (\mathbf{X}_K \otimes \cdots \otimes \mathbf{X}_1) \mathrm{vec}(\mathbf{Y}_n). \tag{A.1}$$

For $k = 1, 2, \ldots, K$, the $\mathbf{X}_k$ takes the value of either $\boldsymbol{\Omega}_k^0$ or $\mathbf{U}_k$. For each combination, denote q to be the number of $\mathbf{X}_k$'s taking $\mathbf{U}_k$. The summation in the third line in (A.1) sums all possible combinations of $\mathbf{X}_k$'s value. For example, the following corresponds to a term with $q = 1$, that is, only one $\mathbf{X}_k$ takes $\mathbf{U}_k$

$$\hat{\mu}_k = \frac{1}{\sqrt{N}}\sum_{n=1}^{N} \mathrm{vec}(\mathbf{Y}_n)^T (\boldsymbol{\Omega}_K^0 \otimes \cdots \otimes \boldsymbol{\Omega}_{k+1}^0 \otimes \mathbf{U}_k \otimes \boldsymbol{\Omega}_{k-1}^0 \otimes \cdots \otimes \boldsymbol{\Omega}_1^0) \mathrm{vec}(\mathbf{Y}_n). \tag{A.2}$$

Now together with similar techniques of Yuan and Lin [14] andYin and Li [13], for N large enough we have

$$f_N(\mathbf{U}_1,\ldots,\mathbf{U}_K)-f_N(\mathbf{0},\ldots,\mathbf{0})=-\sum_{k=1}^{K}\frac{m}{m_k}\Big(\frac{\operatorname{tr}(\mathbf{U}_k\boldsymbol{\Sigma}_k^0)}{\sqrt{N}}-\frac{1}{2}\frac{\operatorname{tr}(\mathbf{U}_k\boldsymbol{\Sigma}_k^0\mathbf{U}_k\boldsymbol{\Sigma}_k^0)}{N}+o\Big(\frac{1}{N}\Big)\Big)$$
$$+\frac{1}{N}\sum_{q>0}\sum_{n=1}^{N}\frac{1}{N^{(q/2)}}\operatorname{vec}(\mathbf{Y}_n)^T(\mathbf{X}_K\otimes\cdots\otimes\mathbf{X}_1)\operatorname{vec}(\mathbf{Y}_n)$$
$$+\sum_{k=1}^{K}\frac{\lambda_k}{\sqrt{N}}\sum_{i\neq j}\Big(U_k(i,j)\cdot\operatorname{sign}(\mathit{\Omega}_k^0(i,j))\cdot I\{\mathit{\Omega}_k^0(i,j)\neq 0\}+|U_k(i,j)|I\{\mathit{\Omega}_k^0(i,j)=0\}\Big),$$

where the Taylor expansion of $\log|A|$ can be found in [7]. Then it follows with $\hat{\mu}_k$ defined in (A.2) that

$$\begin{aligned}N(f_N(\mathbf{U}_1,\ldots,\mathbf{U}_K)-f_N(\mathbf{0},\ldots,\mathbf{0}))=&\sum_{k=1}^{K}\frac{m}{m_k}\Big(\frac{1}{2}\operatorname{tr}(\mathbf{U}_k\boldsymbol{\Sigma}_k^0\mathbf{U}_k\boldsymbol{\Sigma}_k^0)+o(1)\Big)\\&+\sum_{k=1}^{K}\sqrt{N}\Big(\hat{\mu}_k/\sqrt{N}-\frac{m}{m_k}\operatorname{tr}(\mathbf{U}_k\boldsymbol{\Sigma}_k^0)\Big)\\&+\sum_{q>1}\sum_{n=1}^{N}\frac{1}{N^{(q/2)}}\operatorname{vec}(\mathbf{Y}_n)^T(\mathbf{X}_K\otimes\cdots\otimes\mathbf{X}_1)\operatorname{vec}(\mathbf{Y}_n)\\&+\sum_{k=1}^{K}\sqrt{N}\lambda_k\sum_{i\neq j}\Big(U_k(i,j)\cdot\operatorname{sign}(\mathit{\Omega}_k^0(i,j))\cdot I\{\mathit{\Omega}_k^0(i,j)\neq 0\}\\&+|U_k(i,j)|I\{\mathit{\Omega}_k^0(i,j)=0\}\Big).\end{aligned}\tag{A.3}$$

The asymptotic property of the third and the fourth line in the above displayed equation should be addressed. For the third line, define

$$y_{nk}=\operatorname{vec}(\mathbf{Y}_n)^T(\boldsymbol{\Omega}_K^0\otimes\cdots\otimes\boldsymbol{\Omega}_{k+1}^0\otimes\mathbf{U}_k\otimes\boldsymbol{\Omega}_{k-1}^0\otimes\cdots\otimes\boldsymbol{\Omega}_1^0)\operatorname{vec}(\mathbf{Y}_n)$$

and

$$\begin{aligned}\mathbf{W}_k&=(\boldsymbol{\Omega}_K^0\otimes\cdots\otimes\boldsymbol{\Omega}_{k+1}^0\otimes\mathbf{U}_k\otimes\boldsymbol{\Omega}_{k-1}^0\otimes\cdots\otimes\boldsymbol{\Omega}_1^0)(\boldsymbol{\Sigma}_K^0\otimes\cdots\otimes\boldsymbol{\Sigma}_1^0)\\&=\mathbf{I}_K\otimes\cdots\otimes\mathbf{I}_{k+1}\otimes(\mathbf{U}_k\boldsymbol{\Sigma}_k^0)\otimes\mathbf{I}_{k-1}\otimes\cdots\otimes\mathbf{I}_1\end{aligned}$$

then by the results of quadratic form, we have

$$\mu_k=\mathrm{E}(y_{nk})=\operatorname{tr}(\mathbf{W}_k)=\frac{m}{m_k}\operatorname{tr}(\mathbf{U}_k\boldsymbol{\Sigma}_k^0)$$

and

$$\sigma_{kk}=\operatorname{Var}(y_{nk})=2\operatorname{tr}(\mathbf{W}_k\mathbf{W}_k)=\frac{2m}{m_k}\operatorname{tr}(\mathbf{U}_k\boldsymbol{\Sigma}_k^0\mathbf{U}_k\boldsymbol{\Sigma}_k^0),$$

by the central limit theorem, it follows that

$$\sqrt{N}\big(\hat{\mu}_k/\sqrt{N}-\mu_k\big)\to\mathrm{N}(0,\sigma_{kk}).$$

Besides, for $i\neq j$, due to the fact that

$$\sigma_{ij}=\operatorname{Cov}(\hat{\mu}_i,\hat{\mu}_j)=2\operatorname{tr}\big[\mathbf{W}_i\mathbf{W}_j\big]=\frac{2m}{m_im_j}\operatorname{tr}(\mathbf{U}_i\boldsymbol{\Sigma}_i^0)\operatorname{tr}(\mathbf{U}_j\boldsymbol{\Sigma}_j^0)$$

we have

$$\sum_{k=1}^{K}\sqrt{N}\big(\hat{\mu}_k/\sqrt{N}-\mu_k\big)\to\mathrm{N}(0,\sigma^2)$$

where $\sigma^2=\sum_{i,j=1}^{K}\sigma_{ij}$.

Now, we turn to address the fourth line of (A.3). The summands with $q>2$ will ultimately vanish as $N\to\infty$. For the summands with $q=2$, for $i<j$, assume $\mathbf{X}_i$ and $\mathbf{X}_j$ take the value of $\mathbf{U}_i$ and $\mathbf{U}_j$ respectively. Then by weak law of large

numbers,

$$\frac{1}{N}\sum_{n=1}^{N}\mathrm{vec}(\mathbf{Y}_n)^T(\boldsymbol{\Omega}_K^0\otimes\cdots\otimes\boldsymbol{\Omega}_{j+1}^0\otimes\mathbf{U}_j\otimes\cdots\otimes\boldsymbol{\Omega}_{i+1}^0\otimes\mathbf{U}_i\otimes\cdots\otimes\boldsymbol{\Omega}_1^0)\mathrm{vec}(\mathbf{Y}_n)$$

converges to $\frac{m}{m_i m_j}\mathrm{tr}(\mathbf{U}_i\boldsymbol{\Sigma}_i^0)\mathrm{tr}(\mathbf{U}_j\boldsymbol{\Sigma}_j^0)$.

Based on all the results above and the fact that $\sqrt{N}\lambda_k\to\lambda_{0k}$, we can conclude

$$N[f_N(\mathbf{U}_1,\ldots,\mathbf{U}_K)-f_N(\mathbf{0},\ldots,\mathbf{0})]\to_d g(\mathbf{U}_1,\ldots,\mathbf{U}_K)$$

where the function $g(\mathbf{U}_1,\ldots,\mathbf{U}_K)$ is stated in Theorem 1.

The Hessian matrix of

$$\frac{1}{2}\sum_{k=1}^{K}\frac{m}{m_k}\mathrm{tr}(\mathbf{U}_k\boldsymbol{\Sigma}_k^0\mathbf{U}_k\boldsymbol{\Sigma}_k^0)+\sum_{i<j}\frac{m}{m_i m_j}\mathrm{tr}(\mathbf{U}_i\boldsymbol{\Sigma}_i^0)\mathrm{tr}(\mathbf{U}_j\boldsymbol{\Sigma}_j^0)$$

equals to $\frac{\partial l(\mathbf{z}_0)}{\partial\mathbf{z}\partial\mathbf{z}^T}$ of Section 4. By Lemma 2, $g(\mathbf{U}_1,\mathbf{U}_2,\ldots,\mathbf{U}_K)$ is strictly quasi-convex for parameters $\mathbf{z}=(\mathrm{vec}(\mathbf{U}_1)^T,\ldots,\mathrm{vec}(\mathbf{U}_K)^T)^T$. Furthermore, the convergence for $N[f_N(\mathbf{U}_1,\ldots,\mathbf{U}_K)-f_N(\mathbf{0},\ldots,\mathbf{0})]$ is uniform for $\mathbf{z}$ in a compact neighborhood of the origin $\mathbf{0}$. Thus, there is a local minimizer of

$$N[f_N(\mathbf{U}_1,\ldots,\mathbf{U}_K)-f_N(\mathbf{0},\ldots,\mathbf{0})]$$

converging to

$$\mathrm{argmin}_{(\mathbf{U}_1,\ldots,\mathbf{U}_K)}\left\{g(\mathbf{U}_1,\ldots,\mathbf{U}_K)\right\}.$$

The results follows. □

Lemma 4 and its proof

Lemma 4. *Let $\mathbf{Y}_1,\mathbf{Y}_2,\ldots,\mathbf{Y}_N$ be N i.i.d. observations from tensor normal distribution* anorm$(\mathbf{0},\boldsymbol{\Sigma}_1^0\circ\boldsymbol{\Sigma}_2^0\circ\cdots\circ\boldsymbol{\Sigma}_K^0)$, *and $\mathbf{S}_k$ be defined as (6) in Lemma 3. Now, define*

$$\widetilde{\mathbf{S}}_k=\frac{m_k}{N\cdot m}\sum_{n=1}^{N}\widetilde{\mathbf{V}}_n^k\left[\widetilde{\mathbf{V}}_n^k\right]^T=\frac{m_k}{N\cdot m}\sum_{n=1}^{N}\sum_{j=1}^{m/m_k}\widetilde{\mathbf{v}}_n^k(j)\left[\widetilde{\mathbf{v}}_n^k(j)\right]^T$$

where $\widetilde{\mathbf{v}}_n^k(j)$ is the j-th column of

$$\widetilde{\mathbf{V}}_n^k=\mathbf{Y}_{n(k)}(\widehat{\boldsymbol{\Omega}}_K^{1/2}\otimes\cdots\otimes\widehat{\boldsymbol{\Omega}}_{k+1}^{1/2}\otimes\widehat{\boldsymbol{\Omega}}_{k-1}^{1/2}\otimes\cdots\otimes\widehat{\boldsymbol{\Omega}}_1^{1/2})$$

with $\mathbf{Y}_{n(k)}$ being the k-mode matrix unfolding of $\mathbf{Y}_n$, and $\widehat{\boldsymbol{\Omega}}_j(j\neq k)$'s are consistent estimates of $\boldsymbol{\Omega}_j^0(j\neq k)$, which is $\left\|\widehat{\boldsymbol{\Omega}}_j-\boldsymbol{\Omega}_j\right\|_F=o_p(1)$. Then for fixed m, fixed m_k and $N\to\infty$, it holds that

$$\|\widetilde{\mathbf{S}}_k-\mathbf{S}_k\|_F=o_p(1/\sqrt{N}).\tag{A.4}$$

Proof. Comparing $\widetilde{\mathbf{S}}_k$ with $\mathbf{S}_k$, we note that $\boldsymbol{\Omega}_j^0(j\neq k)$'s in $\mathbf{S}_k$ are replaced by $\widehat{\boldsymbol{\Omega}}_j(j\neq k)$'s in $\widetilde{\mathbf{S}}_k$. Let $\mathbf{y}_{n(k)}[j]$ be the jth row of $\mathbf{Y}_{n(k)}$, and consider the (i,j)-th entry of the difference $\widetilde{\mathbf{S}}_k-\mathbf{S}_k$

$$\begin{aligned}\left|S_k(i,j)-\widetilde{S}_k(i,j)\right|&=\left|\frac{m_k}{N\cdot m}\sum_{n=1}^{N}\mathbf{y}_{n(k)}[i]\mathbf{E}_k\mathbf{y}_{n(k)}[j]^T\right|\\&=\left|\frac{m_k}{N\cdot m}\sum_{n=1}^{N}\mathrm{tr}\left(\mathbf{y}_{n(k)}[j]^T\mathbf{y}_{n(k)}[i]\mathbf{E}_k\right)\right|=\left|\mathrm{tr}\left(\mathbf{E}_k\mathbf{F}_{k,i,j}\right)\right|\le\|\mathbf{E}_k\|_F\times\|\mathbf{F}_{k,i,j}\|_F\end{aligned}$$

where

$$\begin{aligned}\mathbf{E}_k&=\widehat{\boldsymbol{\Omega}}_K\otimes\cdots\otimes\widehat{\boldsymbol{\Omega}}_{k+1}\otimes\widehat{\boldsymbol{\Omega}}_{k-1}\otimes\cdots\otimes\widehat{\boldsymbol{\Omega}}_1\\&\quad-\boldsymbol{\Omega}_K^0\otimes\cdots\otimes\boldsymbol{\Omega}_{k+1}^0\otimes\boldsymbol{\Omega}_{k-1}^0\otimes\cdots\otimes\boldsymbol{\Omega}_1^0\\\mathbf{F}_{k,i,j}&=\frac{m_k}{N\cdot m}\sum_{n=1}^{N}\mathbf{y}_{n(k)}[j]^T\mathbf{y}_{n(k)}[i].\end{aligned}$$

Furthermore we have $\|\mathbf{E}_k\|_F = o_p(1)$ by consistency of $\widehat{\boldsymbol{\Omega}}_j (j \neq k)$'s, and $\|\mathbf{F}_{k,i,j}\|_F = O_p(1/\sqrt{N})$ by the central limit theorem. From above, we can conclude that

$$\left|S_k(i,j) - \widetilde{S}_k(i,j)\right| = o_p(1/\sqrt{N})$$

for each i, j and

$$\|\widetilde{\mathbf{S}}_k - \mathbf{S}_k\|_F = o_p(1/\sqrt{N})$$

for fixed m, fixed m_k and $N \to \infty$. □

Proof of Theorem 2

Proof. For $k = 1, 2, \ldots, K$, define $\mathbf{Z}_k$ to be a $m_k \times m_k$ matrix, whose entries satisfy

$$Z_k(i,j) = \begin{cases} 0, & i = j \\ \operatorname{sign}\big(\widehat{\Omega}_k(i,j)\big)/|\widetilde{\Omega}_k(i,j)|^\gamma, & i \neq j \end{cases} \tag{A.5}$$

where $\operatorname{sign}\big(\widehat{\Omega}_k(i,j)\big)$ is equal to 1 if $\widehat{\Omega}_k(i,j) > 0$, equal to -1 if $\widehat{\Omega}_k(i,j) < 0$, or takes the value in the interval $[-1, 1]$ otherwise.

Based on Lemma 1(i), $\widehat{\boldsymbol{\Omega}}_k = \boldsymbol{\Omega}_k^0 + \mathbf{U}_k/\sqrt{N}$ is a local optimizer of (4) with adaptive penalty only if the sub-gradient for $\widehat{\boldsymbol{\Omega}}_k$ equals $\mathbf{0}$, that is

$$-\frac{m}{m_k}\widehat{\boldsymbol{\Omega}}_k^{-1} + \frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{m/m_k}\widetilde{\mathbf{v}}_n^k(j)\big[\widetilde{\mathbf{v}}_n^k(j)\big]^T + \lambda_k\mathbf{Z}_k = \mathbf{0} \tag{A.6}$$

where $\widetilde{\mathbf{v}}_n^k(j)$ is the j-th column of

$$\widetilde{\mathbf{V}}_n^k = \mathbf{Y}_{n(k)}(\widehat{\boldsymbol{\Omega}}_K^{1/2} \otimes \cdots \otimes \widehat{\boldsymbol{\Omega}}_{k+1}^{1/2} \otimes \widehat{\boldsymbol{\Omega}}_{k-1}^{1/2} \otimes \cdots \otimes \widehat{\boldsymbol{\Omega}}_1^{1/2})$$

with $\mathbf{Y}_{n(k)}$ being the k-mode matrix unfolding of $\mathbf{Y}_n$.

Now, define

$$\widetilde{\mathbf{S}}_k = \frac{m_k}{N \cdot m}\sum_{n=1}^{N}\sum_{j=1}^{m/m_k}\widetilde{\mathbf{v}}_n^k(j)\big[\widetilde{\mathbf{v}}_n^k(j)\big]^T.$$

Multiply both sides of (A.6) by $\sqrt{N}$ and take vectorization

$$-\frac{m \cdot \sqrt{N}}{m_k}\operatorname{vec}\big(\widehat{\boldsymbol{\Omega}}_k^{-1}\big) + \frac{m \cdot \sqrt{N}}{m_k}\operatorname{vec}\big(\widetilde{\mathbf{S}}_k\big) + \sqrt{N}\lambda_k\operatorname{vec}\big(\mathbf{Z}_k\big) = \mathbf{0}.$$

From the fact that

$$\operatorname{vec}(\widehat{\boldsymbol{\Omega}}_k^{-1}) = \operatorname{vec}(\boldsymbol{\Sigma}_k^0) - (\boldsymbol{\Sigma}_k^0 \otimes \boldsymbol{\Sigma}_k^0)\frac{\operatorname{vec}(\mathbf{U}_k)}{\sqrt{N}} + o\Big(\frac{1}{\sqrt{N}}\Big)$$

it follows

$$(\boldsymbol{\Sigma}_k^0 \otimes \boldsymbol{\Sigma}_k^0)\operatorname{vec}(\mathbf{U}_k) + \sqrt{N}\Big(\operatorname{vec}(\widetilde{\mathbf{S}}_k) - \operatorname{vec}(\boldsymbol{\Sigma}_k^0)\Big) + \frac{m_k\sqrt{N}}{m}\lambda_k\operatorname{vec}\big(\mathbf{Z}_k\big) + o(1) = \mathbf{0}. \tag{A.7}$$

By Theorem 1, we have already known $\widehat{\boldsymbol{\Omega}}_k$'s are consistent estimates of $\boldsymbol{\Omega}^0$. For the second term on the left side of (A.7), applying Lemma 3 and the conclusion (A.4) of Lemma 4, it follows that

$$\sqrt{N}\Big(\operatorname{vec}(\widetilde{\mathbf{S}}_k) - \operatorname{vec}(\boldsymbol{\Sigma}_k^0)\Big) = \sqrt{N}\Big(\operatorname{vec}(\mathbf{S}_k) - \operatorname{vec}(\boldsymbol{\Sigma}_k^0)\Big) + o_p(1) \to_d \mathrm{N}\Big(0, \frac{m_k}{m}\mathbf{R}_k\Big), \tag{A.8}$$

where $\mathbf{R}_k$ is defined in Lemma 3. Thus, the first two terms on the left side of equation (A.7) is $O_p(1)$. From the assumption of Theorem 2, we also have

$$\sqrt{N}\lambda_k/|\widetilde{\Omega}_k(i,j)|^\gamma \to_p \begin{cases} 0 & \text{if } (i,j) \in \mathscr{A}_k \\ \infty & \text{if } (i,j) \in \mathscr{A}_k^c. \end{cases} \tag{A.9}$$

As a result, for $(i,j) \in \mathscr{A}_k^c$, the probability that $U_k(i,j) = 0$ increases to 1 as $N \to \infty$. Otherwise, the necessary condition (A.7) of local optimality would not hold.

On the other hand, for the entries in $\mathbf{U}_k$ indexed by $(i,j) \in \mathscr{A}_k$,

$$\begin{aligned}\mathrm{vec}(\mathbf{U}_k)_{\mathscr{A}_k} &= \sqrt{N}\Big[(\boldsymbol{\Omega}_k^0 \otimes \boldsymbol{\Omega}_k^0)\big(\mathrm{vec}(\widetilde{\mathbf{S}}_k) - \mathrm{vec}(\boldsymbol{\Sigma}_k^0)\big)\Big]_{\mathscr{A}_k} + o_p(1)\\ &= \big(\boldsymbol{\Omega}_k^0 \otimes \boldsymbol{\Omega}_k^0\big)_{(\mathscr{A}_k,\cdot)} \times \sqrt{N}\big(\mathrm{vec}(\widetilde{\mathbf{S}}_k) - \mathrm{vec}(\boldsymbol{\Sigma}_k^0)\big) + o_p(1)\\ &\to_d \mathrm{N}\Big(0, \frac{m_k}{m}\Big[\big(\boldsymbol{\Omega}_k^0 \otimes \boldsymbol{\Omega}_k^0\big)_{(\mathscr{A}_k,\cdot)}\Big]\mathbf{R}_k\Big[\big(\boldsymbol{\Omega}_k^0 \otimes \boldsymbol{\Omega}_k^0\big)_{(\mathscr{A}_k,\cdot)}\Big]^T\Big)\end{aligned}$$

as a result of (A.7)–(A.9).

Proof of Theorem 3

Proof. The idea is similar to that of Theorem 1, [6]. For $k = 1, 2, \ldots, K$, let $\mathbf{U}_k$ be a symmetric matrix of order m_k, $\mathbf{D}_{U_k}$ be its diagonal matrix, and $\mathbf{R}_{U_k} = \mathbf{U}_k - \mathbf{D}_{U_k}$ be its off-diagonal matrix. Set $\boldsymbol{\Delta}_k = \alpha_{Nk}\mathbf{R}_{U_k} + \beta_{Nk}\mathbf{D}_{U_k}$. The goal is to prove that, for $\alpha_{(N,k)} = (\frac{m_k}{m}s_k \log m_k/N)^{1/2}$ and $\beta_{(N,k)} = (\frac{m_k}{m}m_k \log m_k/N)^{1/2}$, for sets $\mathscr{U}_k = \{\mathbf{U}_k : \|\Delta_k\|_F^2 = A_{(N,k)}^2\alpha_{(N,k)}^2 + B_{(N,k)}^2\beta_{(N,k)}^2\}$ with bounded sequences $\{A_{(N,k)}\}_{N=1}^{\infty}$ and $\{B_{(N,k)}\}_{N=1}^{\infty}$, it holds

$$P\Big(\inf_{\mathbf{U}_1\in\mathscr{U}_1,\ldots,\mathbf{U}_K\in\mathscr{U}_K} q(\boldsymbol{\Omega}_1^0 + \boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Omega}_K^0 + \boldsymbol{\Delta}_K) > q(\boldsymbol{\Omega}_1^0, \ldots, \boldsymbol{\Omega}_K^0)\Big) \to 1. \tag{A.10}$$

As argued in [6], for $k = 1, 2, \ldots, K$, it follows that $\boldsymbol{\Omega}_k^0 + \boldsymbol{\Delta}_k$ is positive definite and there exists local minimizer $(\widehat{\boldsymbol{\Omega}}_1, \widehat{\boldsymbol{\Omega}}_2, \ldots, \widehat{\boldsymbol{\Omega}}_K)$ such that $\|\widehat{\boldsymbol{\Omega}}_k - \boldsymbol{\Omega}_k^0\|_F = O_p(\alpha_{(N,k)} + \beta_{(N,k)})$, which is the desired result.

Instead of being constants as in the proof of Theorem 1 in Lam and Fan, the $\{A_{(N,k)}\}$ and $\{B_{(N,k)}\}$ are modified to sequences bounded by a sufficiently large constant C, that is, for all N and $k = 1, 2, \ldots, K$,

$$\begin{aligned} C &< |A_{(N,k)}| < \tau^{K-1}(K+1)^{K-1}C\\ C &< |B_{(N,k)}| < \tau^{K-1}(K+1)^{K-1}C \end{aligned} \tag{A.11}$$

where $\tau = \max\{\sum_{k=1}^{K}\tau_{k,2}^2, 1\}$. This modification is necessary for the proof of consistency here, as will be shown in **Part b** of this proof. The constant C here will be defined by (A.23) in **Part a** of this proof.

Now, for $\mathbf{U}_k \in \mathscr{U}_k$, $k = 1, 2, \ldots, K$, consider the difference,

$$q(\boldsymbol{\Omega}_1^0 + \boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Omega}_K^0 + \boldsymbol{\Delta}_K) - q(\boldsymbol{\Omega}_1^0, \ldots, \boldsymbol{\Omega}_K^0) = J_1 + J_2 + J_3$$

where

$$\begin{aligned} J_1 &= \mathrm{tr}\big(S\big[(\boldsymbol{\Omega}_K^0 + \boldsymbol{\Delta}_K) \otimes \cdots \otimes (\boldsymbol{\Omega}_1^0 + \boldsymbol{\Delta}_1)\big]\big) - \mathrm{tr}\big(S\big(\boldsymbol{\Omega}_K^0 \otimes \cdots \otimes \boldsymbol{\Omega}_1^0\big)\big)\\ &\quad - \sum_{k=1}^{K}\frac{m}{m_k}(\log|\boldsymbol{\Omega}_k^0 + \boldsymbol{\Delta}_k| - \log|\boldsymbol{\Omega}_k^0|)\\ J_2 &= \sum_{k=1}^{K}\lambda_k\sum_{(i,j)\in\mathscr{A}_k}(|\Omega_k^0(i,j) + \Delta_k(i,j)| - |\Omega_k^0(i,j)|)\\ J_3 &= \sum_{k=1}^{K}\lambda_k\sum_{(i,j)\notin\mathscr{A}_k, i\neq j}(|\Omega_k^0(i,j) + \Delta_k(i,j)| - |\Omega_k^0(i,j)|) \end{aligned}$$

and we can split J_1 as $J_1 = K_1 + K_2$, where

$$K_1 = \mathrm{tr}\big[\mathbf{S}\big((\boldsymbol{\Omega}_K^0 + \boldsymbol{\Delta}_K) \otimes \cdots \otimes (\boldsymbol{\Omega}_1^0 + \boldsymbol{\Delta}_1)\big)\big] - \mathrm{tr}\big(\mathbf{S}\big(\boldsymbol{\Omega}_K^0 \otimes \cdots \otimes \boldsymbol{\Omega}_1^0\big)\big) - \sum_{k=1}^{K}\frac{m}{m_k}\mathrm{tr}(\boldsymbol{\Sigma}_k^0\boldsymbol{\Delta}_k) \tag{A.12}$$

$$K_2 = \sum_{k=1}^{K}\frac{m}{m_k}\mathrm{vec}(\boldsymbol{\Delta}_k)^T\Big\{\int_0^1 g(\nu, \boldsymbol{\Delta}_{k,\nu})(1-\nu)d\nu\Big\}\mathrm{vec}(\boldsymbol{\Delta}_k) \tag{A.13}$$

with $\boldsymbol{\Omega}_{k,\nu} = \boldsymbol{\Omega}_k^0 + \nu\boldsymbol{\Delta}_k$, and $g(\nu, \boldsymbol{\Delta}_{k,\nu}) = \boldsymbol{\Omega}_{k,\nu}^{-1} \otimes \boldsymbol{\Omega}_{k,\nu}^{-1}$. As shown by [6] in the proof of their Theorem 1, we have

$$K_2 \geq \sum_{k=1}^{K}\frac{m}{m_k}(A_{(N,k)}^2\alpha_{(N,k)}^2 + B_{(N,k)}^2\beta_{(N,k)}^2)/2 \cdot (\tau_{k2}^{-1} + o(1))^{-2} = \sum_{k=1}^{K} O\big(\log m_k/N \cdot (A_{(N,k)}^2 s_k + B_{(N,k)}^2 m_k)/2\big). \tag{A.14}$$

Also as argued in [6] in the proof of their Theorem 1, J_2 is dominated by K_2. Besides, notice J_3 is positive. The proof is complete if we can show K_1 is dominated by $K_2 + J_3$. Now, K_1 can be expressed as $H_1 + H_2$, where

$$H_1 = \sum_{k=1}^{K}\Big(tr\big[\mathbf{S}\big(\boldsymbol{\Omega}_K^0 \otimes \cdots \otimes \boldsymbol{\Omega}_{k+1}^0 \otimes \boldsymbol{\Delta}_k \otimes \boldsymbol{\Omega}_{k-1}^0 \otimes \cdots \otimes \boldsymbol{\Omega}_1^0\big)\big] - \frac{m}{m_k}\mathrm{tr}(\boldsymbol{\Sigma}_k^0\boldsymbol{\Delta}_k)\Big)$$
$$H_2 = \sum_{q>1}\mathrm{tr}\big[\mathbf{S}\big(\mathbf{X}_K \otimes \cdots \otimes \mathbf{X}_1\big)\big]$$

where $\mathbf{X}_k$ takes the value of either $\boldsymbol{\Omega}_k^0$ or $\boldsymbol{\Delta}_k$, q is the number of $\mathbf{X}_k$'s taking $\boldsymbol{\Delta}_k$, and the summands in H_2 enumerate all possible combinations for $\mathbf{X}_k$'s values with $q > 1$. For a clearer understanding of the notation used here, please refer to the details in the proof of Theorem 1.

Now, we are about to show that (A.12) is dominated by $K_2 + J_3$, which is positive. The following proof is divided into two parts. **Part a** is devoted to prove H_1 is dominated by $K_2 + J_3$, while **Part b** shows that H_2 is dominated by K_2.

Part a. For H_1, each of the K summands is dominated by a corresponding term in $K_2 + J_3$. We only need to prove this by showing the summand

$$H_1(k) = tr\big[\mathbf{S}(\boldsymbol{\Omega}_K^0 \otimes \cdots \otimes \boldsymbol{\Omega}_{k+1}^0 \otimes \boldsymbol{\Delta}_k \otimes \boldsymbol{\Omega}_{k-1}^0 \otimes \cdots \otimes \boldsymbol{\Omega}_1^0)\big] - \frac{m}{m_k}\mathrm{tr}(\boldsymbol{\Sigma}_k^0\boldsymbol{\Delta}_k) \tag{A.15}$$

in H_1, is dominated by both the term

$$K_2(k) = \frac{m}{m_k}\mathrm{vec}(\boldsymbol{\Delta}_k)^T\Big\{\int_0^1 g(\nu, \boldsymbol{\Delta}_{k,\nu})(1-\nu)d\nu\Big\}\mathrm{vec}(\boldsymbol{\Delta}_k) \geq O\big(\log m_k/N \cdot (A^2_{(N,k)}s_k + B^2_{(N,k)}m_k)/2\big) \tag{A.16}$$

in K_2 and the term

$$J_3(k) = \lambda_k \sum_{(i,j)\notin\mathscr{A}_k, i\neq j}(|\Omega_k^0(i,j) + \Delta_k(i,j)| - |\Omega_k^0(i,j)|) = \lambda_k \sum_{(i,j)\notin\mathscr{A}_k, i\neq j}|\Delta_k(i,j)| \tag{A.17}$$

in J_3. It is worth noticing that $H_1 = \sum_{k=1}^K H_1(k)$, $K_2 = \sum_{k=1}^K K_2(k)$ and $J_3 = \sum_{k=1}^K J_3(k)$.

Let $\mathbf{v}_n^{0k}(j)$ be the j-th column of

$$\mathbf{V}_n^{0k} = \mathbf{Y}_{n(k)}\Big[\big(\boldsymbol{\Omega}_K^0\big)^{1/2} \otimes \cdots \otimes \big(\boldsymbol{\Omega}_{k+1}^0\big)^{1/2} \otimes \big(\boldsymbol{\Omega}_{k-1}^0\big)^{1/2} \otimes \cdots \otimes \big(\boldsymbol{\Omega}_1^0\big)^{1/2}\Big]$$

where $\mathbf{Y}_{n(k)}$ is the k-mode matrix unfolding of $\mathbf{Y}_n$. By Lemma 3,

$$\mathbf{S}_k = \frac{m_k}{N\cdot m}\sum_{n=1}^{N}\sum_{j=1}^{m/m_k}\mathbf{v}_n^{0k}(j)\big[\mathbf{v}_n^{0k}(j)\big]^T$$

is a sample covariance matrix estimating $\boldsymbol{\Sigma}_k^0$ with sample size $N \cdot m/m_K$. Denote the matrix $(\mathbf{S}_k - \boldsymbol{\Sigma}_k^0)$ by $\mathbf{A}_k$, and by Lemma 2 of [6], we have the maximum of elements of $\mathbf{A}_k$ satisfying

$$\max_{ij}|A_k(i,j)| = O_p(\big\{m_k \log m_k/(N\cdot m)\big\}^{1/2}). \tag{A.18}$$

Now, applying Lemma 1 (i) to (A.15), we get that

$$H_1(k) = \frac{m}{m_k}tr\big[(\mathbf{S}_k - \boldsymbol{\Sigma}_k^0)\boldsymbol{\Delta}_k\big]. \tag{A.19}$$

Recalling that we have defined $\mathbf{A}_k = \mathbf{S}_k - \boldsymbol{\Sigma}_k^0$, we can further split (A.19) into $L_1 + L_2$,

$$\begin{aligned} L_1(k) &= \frac{m}{m_k}\sum_{(i,j)\in\mathscr{A}_k}A_k(i,j)\Delta_k(i,j) \\ L_2(k) &= \frac{m}{m_k}\sum_{(i,j)\notin\mathscr{A}_k}A_k(i,j)\Delta_k(i,j). \end{aligned} \tag{A.20}$$

With the result of (A.18) we have,

$$\begin{aligned} |L_1(k)| &= \frac{m}{m_k}\Big|\sum_{(i,j)\in\mathscr{A}_k}A_k(i,j)\Delta_k(i,j)\Big| \leq \frac{m}{m_k}\Big(\sum_{(i,j)\in\mathscr{A}_k}\big[A_k(i,j)\big]^2\Big)^{1/2}\Big(\sum_{(i,j)\in\mathscr{A}_k}\big[\Delta_k(i,j)\big]^2\Big)^{1/2} \\ &\leq \frac{m}{m_k}(s_k + m_k)^{1/2}\max_{i,j}|A_k(i,j)| \cdot \|\boldsymbol{\Delta}_k\|_F \end{aligned}$$

$$
\begin{aligned}
&= \frac{m}{m_k}(s_k+m_k)^{1/2}\cdot O_p\big(\{m_k\log m_k/(N\cdot m)\}^{1/2}\big)\cdot\big(A^2_{(N,k)}\alpha^2_{(N,k)}+B^2_{(N,k)}\beta^2_{(N,k)}\big)^{1/2}\\
&= O_p\Big(\{s_k\log m_k/N+m_k\log m_k/N\}^{1/2}\Big)\times\Big(A^2_{(N,k)}s_k\log m_k/N+B^2_{(N,k)}m_k\log m_k/N\Big)^{1/2}\\
&= O_p\big(A_{(N,k)}\cdot s_k\cdot\log m_k/N+B_{(N,k)}\cdot m_k\cdot\log m_k/N\big) \qquad (A.21)
\end{aligned}
$$

where the last equality is due to the fact that, there exists some constant $\epsilon>0$ such that

$$\epsilon<\frac{[a+bz]^2}{(1+z)(a^2+b^2z)}\le 1$$

for $z\ge 0$, and positive a,b satisfying (A.11).

Besides, there exists a sufficiently large constant C_k, such that when $A_{(N,k)}>C_k$ and $B_{(N,k)}>C_k$,

$$A_{(N,k)}\cdot s_k+B_{(N,k)}\cdot m_k\le\frac{1}{2\cdot(K+1)}(A^2_{(N,k)}s_k+B^2_{(N,k)}m_k). \qquad (A.22)$$

This is because the left side is linear in $A_{(N,k)}$ and $B_{(N,k)}$, the right side is squared in $A_{(N,k)}$ and $B_{(N,k)}$, and the ratio of coefficients on two sides is a constant $1/(2K+2)$. Then it follows that $L_1(k)$ is dominated by $K_2(k)$ due to (A.16) and (A.21).

The fact that $L_2(k)$ is dominated by $J_3(k)$ follows from

$$
\begin{aligned}
J_3(k)-L_2(k)&\ge\sum_{(i,j)\notin\mathscr{A}_K,i\ne j}\Big(\lambda_k|\Delta_k(i,j)|-\frac{m}{m_k}|\Delta_k(i,j)|\cdot O_p\big(\big\{m_k\log m_k/(N\cdot m)\big\}^{1/2}\big)\Big)\\
&=\sum_{(i,j)\notin\mathscr{A}_K,i\ne j}\Big(\lambda_k|\Delta_k(i,j)|-|\Delta_k(i,j)|\cdot O_p\Big(\Big\{\frac{m}{m_k}\cdot\frac{\log m_k}{N}\Big\}^{1/2}\Big)\Big)\ge 0
\end{aligned}
$$

by the assumption on λ_k of Theorem 3. Combined with the result that $L_1(k)$ is dominated by $K_2(k)$, it has been shown that $H_1(k)$ is dominated by $K_2(k)+J_3(k)$ for $k=1,2,\ldots,K$.

Now define a constant C to be

$$C=\max\{C_1,C_2,\ldots,C_K\}. \qquad (A.23)$$

This C is used to control the sequences $\{A_{(N,k)}\}$ and $\{B_{(N,k)}\}$ as mentioned at the beginning of the proof, and would be applied in **Part b** of this proof. Now, it follows that when $A_{(N,k)}>C$ and $B_{(N,k)}>C$ for $k=1,2,\ldots,K$, we have H_1 is dominated by K_2+J_3.

Part b. Define $Q(i,j)$ $(i>j)$ to be the summand in H_2 with exactly two $\mathbf{X}_i$ and $\mathbf{X}_j$ taking $\boldsymbol{\Delta}_i$ and $\boldsymbol{\Delta}_j$, that is

$$Q(i,j)=\mathrm{tr}\big(\mathbf{S}(\boldsymbol{\Omega}^0_K\otimes\cdots\otimes\boldsymbol{\Omega}^0_{i+1}\otimes\boldsymbol{\Delta}_i\otimes\boldsymbol{\Omega}^0_{i-1}\otimes\cdots\otimes\boldsymbol{\Omega}^0_{j+1}\otimes\boldsymbol{\Delta}_j\otimes\boldsymbol{\Omega}^0_{j-1}\otimes\cdots\otimes\boldsymbol{\Omega}^0_1)\big).$$

To prove H_2 is also dominated by K_2, we only need to consider the summands $Q(i,j)(i>j)$ in H_2 with exactly two $\mathbf{X}_k$'s taking $\boldsymbol{\Delta}_k$'s $(q=2)$. For the summands in H_2 with more than two $\mathbf{X}_k$'s taking $\boldsymbol{\Delta}_k$'s $(q>2)$, they are dominated by corresponding summands like $Q(i,j)$.

The method to bound $|H_2|$ is illustrated specifically through the following term

$$Q(K,K-1)=\mathrm{tr}(\mathbf{S}(\boldsymbol{\Delta}_K\otimes\boldsymbol{\Delta}_{K-1}\otimes\boldsymbol{\Omega}^0_{K-2}\otimes\cdots\otimes\boldsymbol{\Omega}^0_1))=Q_1+Q_2 \qquad (A.24)$$

where

$$Q_1=tr\Big(\big(\mathbf{S}-\boldsymbol{\Sigma}^0_K\otimes\cdots\otimes\boldsymbol{\Sigma}^0_1\big)\big(\boldsymbol{\Delta}_K\otimes\boldsymbol{\Delta}_{K-1}\otimes\boldsymbol{\Omega}^0_{K-2}\otimes\cdots\otimes\boldsymbol{\Omega}^0_1\big)\Big) \qquad (A.25)$$

and

$$Q_2=\mathrm{tr}((\boldsymbol{\Sigma}^0_K\otimes\cdots\otimes\boldsymbol{\Sigma}^0_1)(\boldsymbol{\Delta}_K\otimes\boldsymbol{\Delta}_{K-1}\otimes\boldsymbol{\Omega}^0_{K-2}\otimes\cdots\otimes\boldsymbol{\Omega}^0_1))=\frac{m}{m_Km_{K-1}}\mathrm{tr}(\boldsymbol{\Sigma}^0_K\boldsymbol{\Delta}_K)\mathrm{tr}(\boldsymbol{\Sigma}^0_{K-1}\boldsymbol{\Delta}_{K-1}). \qquad (A.26)$$

Notice $H_1(K)$ in **Part a** can be expressed as

$$tr\Big(\big(\mathbf{S}-\boldsymbol{\Sigma}^0_K\otimes\cdots\otimes\boldsymbol{\Sigma}^0_1\big)\big(\boldsymbol{\Delta}_K\otimes\boldsymbol{\Omega}^0_{K-1}\otimes\boldsymbol{\Omega}^0_{K-2}\otimes\cdots\otimes\boldsymbol{\Omega}^0_1\big)\Big). \qquad (A.27)$$

Comparing this with Q_1, we can find that $\boldsymbol{\Omega}^0_{K-1}$ in (A.27) (or $H_1(K)$) is replaced by Δ_{K-1} to get Q_1. Besides, it holds that

$$\max_{i,j}\big|\Delta_{K-1}(i,j)\big|\le\|\boldsymbol{\Delta}_{K-1}\|_F\to 0,\ as\,N\to\infty.$$

Q_1 is of smaller order than $H_1(K)$. At the end of **Part a**, it has been shown that $H_1(K)$ is bounded by $K_2(K)+J_3(K)$. With a similar argument to bound $H_1(K)$, it can be shown that

$$Q_1=o_p\big(K_2(K)+J_3(K)\big). \qquad (A.28)$$

Now, a bound is remained to be derived for Q_2. From $\mathrm{tr}(B) \le \sqrt{p\mathrm{tr}(B^TB)}$ for any matrix B of order p, it follows that

$$\mathrm{tr}(\boldsymbol{\Sigma}_K^0 \boldsymbol{\Delta}_K) \le \sqrt{m_K \mathrm{tr}\big[\boldsymbol{\Delta}_K \big(\boldsymbol{\Sigma}_K^0\big)^2 \boldsymbol{\Delta}_K\big]} \le \sqrt{m_K \|\boldsymbol{\Delta}_K \boldsymbol{\Sigma}_K^0\|_F^2} \le \sqrt{m_K \tau_{K2}^2} \|\boldsymbol{\Delta}_K\|_F.$$

The last inequality is due to Lemma 1 of [6] and assumption (7). Similarly,

$$\mathrm{tr}(\boldsymbol{\Sigma}_{K-1}^0 \boldsymbol{\Delta}_{K-1}) \le \sqrt{m_{K-1}\tau_{K-1,2}^2} \|\boldsymbol{\Delta}_{K-1}\|_F. \tag{A.29}$$

With these, it follows that

$$\begin{aligned} |Q_2| &\le \frac{m}{\sqrt{m_K m_{K-1}}} \cdot \tau_{K,2} \cdot \tau_{K-1,2} \cdot \|\boldsymbol{\Delta}_K\|_F \cdot \|\boldsymbol{\Delta}_{K-1}\|_F \\ &= \tau_{K,2} \cdot \tau_{K-1,2} \cdot (A_{(N,K)}^2 \cdot s_K + B_{(N,K)}^2 \cdot m_K)^{1/2} \cdot (A_{(N,K-1)}^2 \cdot s_{K-1} + B_{(N,K-1)}^2 \cdot m_{K-1})^{1/2} \sqrt{\log m_K \log m_{K-1}}/N. \end{aligned}$$

Without loss of generality, we assume

$$(s_K + m_K)\log m_K \ge (s_{K-1} + m_{K-1})\log m_{K-1} \ge \cdots \ge (s_1 + m_1)\log m_1. \tag{A.30}$$

If the order of (A.30) is violated, we simply need to modify the order of (A.31) accordingly, and then adjust the subsequent proof, the result would still hold. Under this assumption, let

$$\begin{aligned} &A_{(N,K)} = B_{(N,K)} = \tau^{K-1}(K+1)^{K-1}C \\ &A_{(N,K-1)} = B_{(N,K-1)} = \tau^{K-2}(K+1)^{K-2}C \\ &\cdots\cdots \\ &A_{(N,1)} = B_{(N,1)} = C \end{aligned} \tag{A.31}$$

where $\tau = \max\{\sum_{k=1}^K \tau_{k,2}^2, 1\}$. As a result of (A.31),

$$(A_{(N,K-1)}^2 \cdot s_{K-1} + B_{(N,K-1)}^2 \cdot m_{K-1})^{1/2}(\log m_{K-1})^{1/2} \le \tfrac{1}{\tau\cdot(K+1)}(A_{(N,K)}^2 \cdot s_K + B_{(N,K)}^2 \cdot m_K)^{1/2}(\log m_K)^{1/2}.$$

Now it holds that

$$\begin{aligned} |Q_2| &\le \tau_{K,2} \cdot \tau_{K-1,2} \cdot \frac{1}{\tau \cdot (K+1)} \cdot (A_{(N,K)}^2 s_K + B_{(N,K)}^2 m_K) \cdot \log m_K/N \\ &\le \frac{1}{2(K+1)} \cdot (A_{(N,K)}^2 s_K + B_{(N,K)}^2 m_K) \cdot \log m_K/N. \end{aligned}$$

Combining with (A.28), it holds that

$$|Q(K, K-1)| \le \frac{1}{2(K+1)} \cdot \big(A_{(N,K)}^2 \cdot s_K + B_{(N,K)}^2 \cdot m_K\big) \cdot \log m_K/N + o_p\big(K_2(K) + J_3(K)\big). \tag{A.32}$$

Based on these results, we now come to show H_2 is dominated by K_2. Recall that we only need to consider the summands in H_2 with exactly two $\mathbf{X}_k$'s taking $\boldsymbol{\Delta}_k$'s,

$$R = \sum_{k=1}^{K-1} Q(K,k) + \sum_{k=1}^{K-2} Q(K-1,k) + \cdots + \sum_{k=1}^{2} Q(3,k) + Q(2,1).$$

With similar techniques leading to (A.32), it can be proved that

$$\begin{aligned} |Q(K,k)| &\le \frac{\tau_{K,2} \cdot \tau_{K-1,2}}{(K+1)^{K-k} \cdot \tau^{K-k}} \cdot \big(A_{(N,K)}^2 \cdot s_K + B_{(N,K)}^2 \cdot m_K\big) \cdot \log m_K/N + o_p\big(K_2(K) + J_3(K)\big) \\ &\le \frac{1}{2(K+1)} \cdot \big(A_{(N,K)}^2 \cdot s_K + B_{(N,K)}^2 \cdot m_K\big) \cdot \log m_K/N + o_p\big(K_2(K) + J_3(K)\big) \end{aligned}$$

where in the last inequality, we have used $K+1 > 1$, $\tau = \max\{\sum_{k=1}^K \tau_{k,2}^2, 1\} \ge 1$ and $2 \cdot \tau_{K,2} \cdot \tau_{K-1,2} \le \tau$. As a result,

$$\Big|\sum_{k=1}^{K-1} Q(K,k)\Big| \le \tfrac{K-1}{2(K+1)} \cdot (A_{(N,K)}^2 \cdot s_K + B_{(N,K)}^2 \cdot m_K) \cdot \log m_K/N + o_p\big(K_2(K) + J_3(K)\big).$$

Similarly,

$$\Big|\sum_{k=1}^{K-2} Q(K-1,k)\Big| \le \frac{K-2}{2(K+1)} \cdot (A_{(N,K-1)}^2 \cdot s_{K-1} + B_{(N,K-1)}^2 \cdot m_{K-1}) \cdot \log m_{K-1}/N + o_p\big(K_2(K-1) + J_3(K-1)\big).$$

Thus,

$$\begin{aligned}\left|R\right| &\le \sum_{k=2}^{K}\frac{k-1}{2(K+1)}\cdot(A^2_{(N,k)}\cdot s_k+B^2_{(N,k)}\cdot m_k)\cdot\log m_k/N+\sum_{k=2}^{K}o_p\big(K_2(k)+J_3(k)\big)\\ &< \sum_{k=2}^{K}\frac{1}{2}\cdot(A^2_{(N,k)}\cdot s_k+B^2_{(N,k)}\cdot m_k)\cdot\log m_k/N+\sum_{k=2}^{K}o_p\big(K_2(k)+J_3(k)\big).\end{aligned}$$

In this way, we can conclude H_2 is also dominated by K_2. Combining with (A.22), H_1 and H_2 together can be dominated by K_2+J_3. □

Proof of Theorem 4

Similar to the proof of Theorem 2, $\widehat{\boldsymbol{\Omega}}_k$ is a local optimizer of (4) with adaptive penalty only if the sub-gradient for $\widehat{\boldsymbol{\Omega}}_k$ equals $\mathbf{0}$,

$$-\frac{m}{m_k}\widehat{\boldsymbol{\Omega}}_k^{-1}+\frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{m/m_k}\mathbf{v}_n^k(j)\big[\mathbf{v}_n^k(j)\big]^T+\lambda_k\mathbf{Z}_k=\mathbf{0} \tag{A.33}$$

where $\mathbf{Z}_k$ is defined at the beginning of the proof of Theorem 2, and $\mathbf{v}_n^k(j)$ is the j-th column of

$$\mathbf{V}_n^k=\mathbf{Y}_{n(k)}(\widehat{\boldsymbol{\Omega}}_K^{1/2}\otimes\cdots\otimes\widehat{\boldsymbol{\Omega}}_{k+1}^{1/2}\otimes\widehat{\boldsymbol{\Omega}}_{k-1}^{1/2}\otimes\cdots\otimes\widehat{\boldsymbol{\Omega}}_1^{1/2})$$

with $\mathbf{Y}_{n(k)}$ being the k-mode matrix unfolding of $\mathbf{Y}_n$.

Furthermore, define $\mathbf{v}_n^{0k}(j)$ to be the j-th column of

$$\mathbf{V}_n^{0k}=\mathbf{Y}_{n(k)}\Big[\big(\boldsymbol{\Omega}_K^0\big)^{1/2}\otimes\cdots\otimes\big(\boldsymbol{\Omega}_{k+1}^0\big)^{1/2}\otimes\big(\boldsymbol{\Omega}_{k-1}^0\big)^{1/2}\otimes\cdots\otimes\big(\boldsymbol{\Omega}_1^0\big)^{1/2}\Big].$$

Then, (A.33) can be written as

$$\begin{aligned}&-\Big(\widehat{\boldsymbol{\Omega}}_k^{-1}-\boldsymbol{\Sigma}_k^0\Big)+\Big(\frac{m_k}{m\cdot N}\sum_{n=1}^{N}\sum_{j=1}^{m/m_k}\mathbf{v}_n^{0k}(j)\big[\mathbf{v}_n^{0k}(j)\big]^T-\boldsymbol{\Sigma}_k^0\Big)+\Big(\frac{m_k}{m\cdot N}\sum_{n=1}^{N}\sum_{j=1}^{m/m_k}\mathbf{v}_n^{k}(j)\big[\mathbf{v}_n^{k}(j)\big]^T\\ &\quad-\frac{m_k}{m\cdot N}\sum_{n=1}^{N}\sum_{j=1}^{m/m_k}\mathbf{v}_n^{0k}(j)\big[\mathbf{v}_n^{0k}(j)\big]^T\Big)+\lambda_k\frac{m_k}{m}\mathbf{Z}_k=\mathbf{0}.\end{aligned}$$

Define

$$\begin{aligned}\mathbf{A}_k&=\widehat{\boldsymbol{\Omega}}_k^{-1}-\boldsymbol{\Sigma}_k^0\\ \mathbf{B}_k&=\Big(\frac{m_k}{m\cdot N}\sum_{n=1}^{N}\sum_{j=1}^{m/m_k}\mathbf{v}_n^{0k}(j)\big[\mathbf{v}_n^{0k}(j)\big]^T-\boldsymbol{\Sigma}_k^0\Big)\\ \mathbf{C}_k&=\Big(\frac{m_k}{m\cdot N}\sum_{n=1}^{N}\sum_{j=1}^{m/m_k}\mathbf{v}_n^{k}(j)\big[\mathbf{v}_n^{k}(j)\big]^T-\frac{m_k}{m\cdot N}\sum_{n=1}^{N}\sum_{j=1}^{m/m_k}\mathbf{v}_n^{0k}(j)\big[\mathbf{v}_n^{0k}(j)\big]^T\Big).\end{aligned}$$

As shown by [6] in the proof of their Theorem 2

$$\max_{i,j}|A_k(i,j)|=O_p(\eta_n^{1/2}). \tag{A.34}$$

With similar arguments leading to (A.18) in the proof of Theorem 3, we have

$$\max_{i,j}|B_k(i,j)|=O_p\Big(\big\{m_k\log m_k/(m\cdot N)\big\}^{1/2}\Big). \tag{A.35}$$

As for $\mathbf{C}_k$, define $\mathbf{U}_i=\widehat{\boldsymbol{\Omega}}_i-\boldsymbol{\Omega}_i^0$, then we have

$$\begin{aligned}\mathbf{C}_k&=\frac{m_k}{m\cdot N}\sum_{n=1}^{N}\mathbf{Y}_{n(k)}\Big[\Big(\widehat{\boldsymbol{\Omega}}_K\otimes\cdots\otimes\widehat{\boldsymbol{\Omega}}_{k+1}\otimes\widehat{\boldsymbol{\Omega}}_{k-1}\otimes\cdots\otimes\widehat{\boldsymbol{\Omega}}_1\Big)-\Big(\boldsymbol{\Omega}_K^0\otimes\cdots\otimes\boldsymbol{\Omega}_{k+1}^0\otimes\boldsymbol{\Omega}_{k-1}^0\otimes\cdots\otimes\boldsymbol{\Omega}_1^0\Big)\Big]\big(\mathbf{Y}_{n(k)}\big)^T\\ &=\frac{m_k}{m\cdot N}\sum_{n=1}^{N}\sum_{q>0}\mathbf{Y}_{n(k)}\Big(\mathbf{X}_K\otimes\cdots\otimes\mathbf{X}_{k+1}\otimes\mathbf{X}_{k-1}\otimes\cdots\otimes\mathbf{X}_1\Big)\big(\mathbf{Y}_{n(k)}\big)^T\end{aligned}$$

where $\mathbf{X}_l(l\neq k)$'s take the value of either $\mathbf{U}_l$ or $\boldsymbol{\Omega}_l^0$, and q is the number of $\mathbf{X}_l$'s taking $\mathbf{U}_l$. The terms with $q>1$ are dominated by the terms with $q=1$, as a result of consistency of Theorem 3. So we only need to consider the term with $q=1$. For a term with $q = 1$, define for $l \neq k$ that

$$\mathbf{D}_k^l = \frac{1}{N}\sum_{n=1}^{N} \mathbf{Y}_{n(k)}\left(\boldsymbol{\Omega}_K^0 \otimes \cdots \otimes \boldsymbol{\Omega}_{k+1}^0 \otimes \boldsymbol{\Omega}_{k-1}^0 \otimes \cdots \otimes \boldsymbol{\Omega}_{l+1}^0 \otimes \mathbf{U}_l \otimes \boldsymbol{\Omega}_{l-1}^0 \otimes \cdots \otimes \boldsymbol{\Omega}_1^0\right)\left(\mathbf{Y}_{n(k)}\right)^T.$$

Let $D_k^l(i,j)$ be the (i,j)-th element of $\mathbf{D}_k^l$, and $\mathbf{y}_n^k[i]$ be the i-th row of $\mathbf{Y}_{n(k)}$, then

$$\begin{aligned} D_k^l(i,j) &= \frac{m_k}{m \cdot N}\sum_{n=1}^{N} \mathbf{y}_n[i]\left(\boldsymbol{\Omega}_K^0 \otimes \cdots \otimes \boldsymbol{\Omega}_{k+1}^0 \otimes \boldsymbol{\Omega}_{k-1}^0 \otimes \cdots \otimes \boldsymbol{\Omega}_{l+1}^0 \otimes \mathbf{U}_l \otimes \boldsymbol{\Omega}_{l-1}^0 \otimes \cdots \otimes \boldsymbol{\Omega}_1^0\right)\left(\mathbf{y}_n[j]\right)^T \\ &= \frac{m_k}{m \cdot N}\sum_{n=1}^{N} \mathbf{y}_n[i]\left(\boldsymbol{\Omega}_K^0 \otimes \cdots \otimes \boldsymbol{\Omega}_{k+1}^0 \otimes \boldsymbol{\Omega}_{k-1}^0 \otimes \cdots \otimes \boldsymbol{\Omega}_{l+1}^0 \otimes \mathbf{U}_l \otimes \boldsymbol{\Omega}_{l-1}^0 \otimes \cdots \otimes \boldsymbol{\Omega}_1^0\right)\left(\mathbf{y}_n[j]\right)^T \\ &\quad - \frac{1}{m_l} tr\left(\boldsymbol{\Sigma}_l^0 \mathbf{U}_l\right) + \frac{1}{m_l} tr\left(\boldsymbol{\Sigma}_l^0 \mathbf{U}_l\right). \end{aligned}$$

With similar techniques applied to (A.24) in the proof of Theorem 3, we can show that $\left|D_k^l(i,j)\right|$ is dominated by

$$\left|\frac{1}{m_l} tr\left(\boldsymbol{\Sigma}_l^0 \mathbf{U}_l\right)\right| \leq \frac{\tau_{l,2}}{\sqrt{m_l}} \cdot \|U_l\|_F$$

the inequality is a result of (A.29). Thus

$$\max_{i,j}\left|C_k(i,j)\right|$$

is dominated by

$$\sum_{l \neq k} \frac{\tau_{l,2}}{\sqrt{m_l}} \cdot \|U_l\|_F. \tag{A.36}$$

Combining (A.34), (A.35) and (A.36), and by the conditions on λ_k, we have $\widehat{\Omega}_k(i,j) = 0$ with probability increasing to one for $(i,j) \in \mathscr{A}_k^c$. Otherwise, the necessary optimality condition (A.33) would not hold. □

References

[1] G. Allen, R. Tibshirani, Transposable regularized covariance models with an application to missing data imputation, Ann. Appl. Statist. 4 (2) (2010) 764–790.

[2] L. De Lathauwer, B. De Moor, J. Vandewalle, A Multilinear singular value decomposition, SIAM J. Matrix Anal. Appl. 21 (4) (2000) 1253–1278.

[3] J. Friedman, T. Hastie, R. Tibshirani, Sparse inverse covariance estimation with the graphical Lasso, Biostatistics 9 (3) (2008) 432–441.

[4] P.D. Hoff, Separable covariance arrays via the Tucker product, with application to multivariate relational data, Bayesian Anal. 6 (2) (2011) 179–196.

[5] T.G. Kolda, Multilinear operators for higher-order decompositions, Sandia Report, Sand (2006–2081).

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

[7] J.R. Magnus, H. Neudecker, Matrix Differential Calculus with Applications in Statistics and Econometrics, in: Wiley Series in Probability and Statistics: Texts and References Section, 1999.

[8] L. Omberg, et al., Global effects of DNA replication and DNA replication origin activity on eukaryotic gene expression, Mol. Syst. Biol. 5 (1) (2009).

[9] G.A. Seber, A Matrix Handbook for Statisticians, vol. 15, Wiley.com, 2008.

[10] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, B. Futcher, Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization, Mol. Biol. Cell 9 (12) (1998) 3273–3297.

[11] P. Tseng, Convergence of a block coordinate descent method for nondifferentiable minimization, J. Optim. Theory Appl. 109 (3) (2001) 474–494.

[12] T. Tsiligkaridis, A. Hero, S. Zhou, Convergence properties of kronecker graphical Lasso algorithms. 2012. http://arxiv.org/1204.0585v1.pdf.

[13] J. Yin, H. Li, Model selection and estimation in the matrix Normal graphical model, J. Multivar. Anal. 107 (2012) 1190–140.

[14] M. Yuan, Y. Lin, Model selection and estimation in the Gaussian graphical model, Biometrika 94 (1) (2007) 19–35.

[15] Y. Zhang, J. Schneider, Learning multiple tasks with a sparse matrix-normal penalty, in: In Advances in Neural Information Processing Systems, 23, 2010, pp. 2550–2558.

[16] S. Zhou, 2012. Gemini: Graph estimation with matrix variate normal instances, Technical Report. http://arxiv.org/abs/1209.5075.