Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

Fast and Separable Estimation in High-Dimensional Tensor Gaussian Graphical Models

Keqian Min, Qing Mai & Xin Zhang

To cite this article: Keqian Min, Qing Mai & Xin Zhang (2022) Fast and Separable Estimation in High-Dimensional Tensor Gaussian Graphical Models, Journal of Computational and Graphical Statistics, 31:1, 294-300, DOI: 10.1080/10618600.2021.1938086

To link to this article: https://doi.org/10.1080/10618600.2021.1938086

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2022, VOL. 31, NO. 1, 294–300
https://doi.org/10.1080/10618600.2021.1938086

Taylor & Francis
Taylor & Francis Group

SHORT TECHNICAL NOTE

Fast and Separable Estimation in High-Dimensional Tensor Gaussian Graphical Models

Keqian Min, Qing Mai, and Xin Zhang

Department of Statistics, Florida State University, Tallahassee, FL

ABSTRACT

In the tensor data analysis, the Kronecker covariance structure plays a vital role in unsupervised learning and regression. Under the Kronecker covariance model assumption, the covariance of an M-way tensor is parameterized as the Kronecker product of M individual covariance matrices. With normally distributed tensors, the key to high-dimensional tensor graphical models becomes the sparse estimation of the M inverse covariance matrices. Unable to maximize the tensor normal likelihood analytically, existing approaches often require cyclic updates of the M sparse matrices. For the high-dimensional tensor graphical models, each update step solves a regularized inverse covariance estimation problem that is computationally nontrivial. This computational challenge motivates our study of whether a noncyclic approach can be as good as the cyclic algorithms in theory and practice. To handle the potentially very high-dimensional and high-order tensors, we propose a separable and parallel estimation scheme. We show that the new estimator achieves the same minimax optimal convergence rate as the cyclic estimation approaches. Numerically, the new estimator is much faster and often more accurate than the cyclic approach. Moreover, another advantage of the separable estimation scheme is its flexibility in modeling, where we can easily incorporate user-specified or specially structured covariances on any modes of the tensor. We demonstrate the efficiency of the proposed method through both simulations and a neuroimaging application. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received August 2020
Revised April 2021

KEYWORDS
Graphical models; Kronecker covariance; Sparse precision matrix; Tensor

1. Introduction

Modeling high-dimensional tensor data, that is, multi-way array-valued variable $\mathbf{X}_i \in \mathbb{R}^{p_1 \times \cdots \times p_M}$ for each sample $i = 1, \ldots, n$, are frequently involved in many areas of research such as neuroimaging, computational biology, and signal processing. Statistical methods for tensor data analysis are becoming increasingly popular in the recent years (see Chi and Kolda 2012; Zhou, Li and Zhu 2013; Hoff 2015; Lock 2018; Pan, Mai and Zhang 2019; Bi et al. 2021; Pfeiffer, Kapla and Bura 2021, among others). In this article, we study the tensor Gaussian graphical models, which are valuable tools to reveal the dependence structure among the large number of random variables in a tensor. The tensor Gaussian graphical models are generalizations of the Gaussian graphical model for multivariate data. For background on graphical models and the conditional independence interpretation of the Gaussian graphical models, see, for example, Lauritzen (1996).

A straightforward way to study the dependence of variables in tensor is to vectorize the tensor into a vector and then estimate the inverse of the covariance matrix (i.e., the precision matrix). Under the sparsity assumption, many penalized methods have been proposed to obtain a sparse precision matrix (e.g., Yuan and Lin 2007; Friedman, Hastie and Tibshirani 2008; Witten, Friedman and Simon 2011; Danaher, Wang and Witten 2014; Zhang and Zou 2014; Cai, Liu and Zhou 2016; Molstad and Rothman 2018). However, the high dimensionality of the vectorized data is a great challenge to the existing methods. For example, a typical 3D MRI scan of the brain has dimension $p_1 \times p_2 \times p_3 = 256 \times 256 \times 256$. The vectorized data has over 16 million variables, making it unrealistic to estimate the precision matrix. Therefore, to exploit the tensor structural information, the parsimonious and interpretable Kronecker covariance structure is widely used in modeling tensor data. Under this assumption, the problem is then simplified to the estimation of three much smaller precision matrices of size $p_m \times p_m$, $m = 1, 2, 3$, in the MRI example.

Sparse precision matrix estimation under the Kronecker covariance structure has been gaining increasing attention in the recent years: From matrix data (Leng and Tang 2012; Yin and Li 2012; Zhou 2014; Zhu and Li 2018) to higher order tensors (Tsiligkaridis, Hero III and Zhou 2013; He et al. 2014; Xu, Zhang, and Gu 2017; Lyu et al. 2019). Assuming the precision matrices are sparse, the standard way is to use a penalized negative log-likelihood function. Then the estimators are obtained through iterative algorithms to cyclic update one precision matrix while fixing the other $M - 1$ precision matrices at the estimated value from the previous iteration. Such cyclic update algorithms can be computationally costly, especially when the tensor dimensions or the tensor order

CONTACT Xin Zhang henry@stat.fsu.edu Department of Statistics, Florida State University, 214 OSB, 117 N. Woodward Ave., P.O. Box 3064330, Tallahassee, FL 32306-4330.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

are not small. Even for matrix Gaussian graphical models, as the dimensions increase, the computational cost increases drastically, especially when cross-validation is used for tuning parameter selection. In addition, when using the cyclic iterative methods, all precision matrices are assumed to be sparse and estimated in a nested sequence. If we are only interested in the dependence of one particular mode, then the estimation accuracy would be a concern when other irrelevant precision matrices violate the sparsity assumption. Because all precision matrices are estimated through regularized optimization, the inaccurate estimates of nonsparse precision matrices can affect the estimation of the sparse precision matrices.

To tackle this challenge, we propose a noncyclic and parallel approach for estimating the sparse tensor Gaussian graphical model with a relaxed assumption. Specifically, only the precision matrices of interest are assumed to be sparse. Each precision matrix is estimated by solving an independent ℓ_1 constrained minimization problem. For example, in some cases, such as modeling spatio-temporal data, some particular non-sparse correlation patterns can be adapted into our framework. The proposed method can also be incorporated into jointly estimating regression coefficients and covariances in tensor response regression. The proposed method improves the computation efficiency and the estimation accuracy in numerical studies. In theory, the proposed estimator enjoys the same optimal convergence rate in the graphical model literature.

The rest of the paper is organized as follows. In Section 2, we present some preliminaries on tensor algebra and introduce the tensor graphical model. In Section 3, we develop our separable non-iterative method and its theoretical properties. Simulation studies and a real data illustration are presented in Sections 4 and 5, respectively. Online supplementary materials contains proofs, R code, and applications to partially sparse models and tensor regression.

2. Background and Problem Set-up

2.1. Notation

We review some basic tensor notation and operations that are commonly used (e.g., Kolda and Bader 2009). Multidimensional array $\mathbf{A} \in \mathbb{R}^{p_1 \times \cdots \times p_M}$ is called a tensor of order M. The vectorization of a tensor $\mathbf{A}$ is denoted by $\text{vec}(\mathbf{A}) \in \mathbb{R}^{(\prod_m p_m)\times 1}$, with $A_{i_1,\ldots,i_M}$ being its jth element, where $j = 1 + \sum_{m=1}^{M}(i_m - 1)\prod_{m'=1}^{m-1} p_{m'}$. The Frobenius norm of $\mathbf{A}$ is defined as $\|\mathbf{A}\|_F = (\sum_{i_1,\ldots,i_M} A^2_{i_1,\ldots,i_M})^{1/2}$. We define $p = \prod_{m=1}^{M} p_m$ and $p_{-m} = \prod_{m'=1,m'\neq m}^{M} p_{m'}$. The mode-$m$ matricization of $\mathbf{A}$ is denoted by $\mathbf{A}_{(m)} \in \mathbb{R}^{p_m \times p_{-m}}$, which is obtained by combining the $(M-1)$ modes of the tensor similar to vectorization. The mode-m product of tensor $\mathbf{A}$ with a matrix $\boldsymbol{\alpha} \in \mathbb{R}^{d\times p_m}$ is defined as $\mathbf{A} \times_m \boldsymbol{\alpha}$ and it yields a tensor of size $p_1 \times \cdots \times p_{m-1} \times d \times p_{m+1} \times \cdots \times p_M$. Elementwise, we have $(\mathbf{A} \times_m \boldsymbol{\alpha})_{i_1,\ldots,i_{m-1},j,i_{m+1},\ldots,i_M} = \sum_{i_m=1}^{p_m} A_{i_1,\ldots,i_M}\boldsymbol{\alpha}_{j,i_m}$. For a list of matrices $\{\boldsymbol{\alpha}\} = \{\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_M\}$ with $\boldsymbol{\alpha}_m \in \mathbb{R}^{d_m\times p_m}$, we define $\mathbf{A} \times \{\boldsymbol{\alpha}\} = \mathbf{A}\times_1\boldsymbol{\alpha}_1 \cdots \times_M\boldsymbol{\alpha}_M$. Let $\mathbf{Y} = \mathbf{A}\times\{\boldsymbol{\alpha}\}$, then $\mathbf{Y}_{(m)} = \boldsymbol{\alpha}_m\mathbf{A}_{(m)}(\boldsymbol{\alpha}_M \otimes \cdots \otimes \boldsymbol{\alpha}_{m+1} \otimes \boldsymbol{\alpha}_{m-1} \otimes \cdots \otimes \boldsymbol{\alpha}_1)^\top$ where $\otimes$ is the Kronecker product. Let $\{\boldsymbol{\alpha}\}_{-m}$ be the subset of $\{\boldsymbol{\alpha}\}$ without the m-th matrix.

2.2. Tensor Gaussian Graphical Model

The tensor random variable $\mathbf{Z} \in \mathbb{R}^{p_1\times\cdots\times p_M}$ follows a standard tensor normal distribution if all elements of $\mathbf{Z}$ are independent standard normal random variables. Let $\mathbf{X} = \boldsymbol{\mu} + \mathbf{Z} \times \{\boldsymbol{\Sigma}^{1/2}\}$, where $\{\boldsymbol{\Sigma}^{1/2}\} = \{\boldsymbol{\Sigma}_1^{1/2},\ldots,\boldsymbol{\Sigma}_M^{1/2}\}$ is a list of symmetric positive definite matrices, then $\mathbf{X}$ follows a tensor normal (TN) distribution with mean $\boldsymbol{\mu}$ and Kronecker separable covariance structure, denoted as $\mathbf{X} \sim \text{TN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_M)$. If $M = 2$, the tensor normal distribution reduces to the matrix normal (MN) distribution (Gupta and Nagar 2018). Suppose that a mode-M tensor $\mathbf{X} \in \mathbb{R}^{p_1\times\cdots\times p_M}$ follows a tensor normal distribution TN $(\mathbf{0}; \boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_M)$, then its probability density function is defined as

$$p(\mathbf{X} \mid \boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_M) = (2\pi)^{-p/2}\left\{\prod_{m=1}^{M}|\boldsymbol{\Sigma}_m|^{-p_{-m}/2}\right\} \times \exp\left(-\frac{1}{2}\left\|\mathbf{X}\times\{\boldsymbol{\Sigma}^{-1/2}\}\right\|_F^2\right), \quad (1)$$

where $\{\boldsymbol{\Sigma}^{-1/2}\} = \{\boldsymbol{\Sigma}_1^{-1/2},\ldots,\boldsymbol{\Sigma}_M^{-1/2}\}$. For $m = 1,\ldots,M$, we call $\boldsymbol{\Omega}_m \equiv \boldsymbol{\Sigma}_m^{-1}$ the mode-m precision matrix, which is our target parameter.

To distinguish the population true parameter and the argument in optimization, we consider the model that $\mathbf{X}_1,\ldots,\mathbf{X}_n$ are iid samples from TN $(\mathbf{0}; \boldsymbol{\Sigma}_1^*,\ldots,\boldsymbol{\Sigma}_M^*)$. We study the problem of estimating the precision matrices $\{\boldsymbol{\Omega}_1^*,\ldots,\boldsymbol{\Omega}_M^*\}$, where $\boldsymbol{\Omega}_m^* = (\boldsymbol{\Sigma}_m^*)^{-1}$ for $m = 1,\ldots,M$. In particular, we focus on high-dimensional settings that $\boldsymbol{\Omega}_m^*$ is sparse for $m = 1,\ldots,M$. This sparse precision matrix assumption has the conditional independence interpretation that is analogous to the classical multivariate (vector) settings (Lauritzen 1996). Let $\mathbf{X}_{(m)}$ denote the mode-m matricization of $\mathbf{X}$ and $\mathbf{X}_{(m),i}$ denote the i-th row of $\mathbf{X}_{(m)}$. Then $[\boldsymbol{\Omega}_m^*]_{i,j} = 0$ if and only if $\mathbf{X}_{(m),i}$ is independent of $\mathbf{X}_{(m),j}$ given $\mathbf{X}_{(m),k}$, $k \neq i, j$.

The precision matrices $\{\boldsymbol{\Omega}_1^*,\ldots,\boldsymbol{\Omega}_M^*\}$ are identifiable up to $(M-1)$ scaling constants. For example, it is easy to verify that TN $(\mathbf{0}; \boldsymbol{\Sigma}_1^*, \boldsymbol{\Sigma}_2^*,\ldots,\boldsymbol{\Sigma}_M^*)$ and TN $(\mathbf{0}; \boldsymbol{\Sigma}_1^*/c, c\boldsymbol{\Sigma}_2^*,\ldots,\boldsymbol{\Sigma}_M^*)$ have the same probability density function for any $c > 0$. To address the identifiability problem in the tensor normal model, we assume that $\|\boldsymbol{\Omega}_m^*\|_F = 1$ for all $m = 1,\ldots,M$, which can be easily satisfied by standardization. Alternatively, one can impose $(M-1)$ constraints that $\|\boldsymbol{\Omega}_m^*\|_F = 1$ for all $m = 2,\ldots,M$, and absorb the scaling constant into $\boldsymbol{\Omega}_1^*$. Clearly, the scaling does not affect the sparsity pattern of the precision matrices and is a nuisance parameter in graphical model.

A standard approach to the tensor graphical model estimation is based on minimizing the penalized negative log-likelihood function,

$$\text{L}_n(\boldsymbol{\Omega}_1,\ldots,\boldsymbol{\Omega}_M) = \frac{1}{p}\text{tr}\{\mathbf{S}(\boldsymbol{\Omega}_M \otimes \cdots \otimes \boldsymbol{\Omega}_1)\} - \sum_{m=1}^{M}\frac{1}{p_m}\log|\boldsymbol{\Omega}_m| + \sum_{m=1}^{M}P_{\lambda_m}(\boldsymbol{\Omega}_m), \quad (2)$$

where $\mathbf{S} = n^{-1}\sum_{i=1}^{n}\text{vec}(\mathbf{X}_i)\,\text{vec}(\mathbf{X}_i)^\top$ and $P_{\lambda_m}(\cdot)$ is a penalized function indexed by the tuning parameter $\lambda_m > 0$, $m = 1,\ldots,M$. In this paper, we focus on the ℓ_1 penalty on the off-diagonal elements: $P_{\lambda_m}(\boldsymbol{\Omega}_m) = \lambda_m\|\boldsymbol{\Omega}_m\|_{1,\text{off}}$, where

$\|\mathbf{\Omega}_m\|_{1,\text{off}} = \sum_{i\neq j}|[\mathbf{\Omega}_m]_{i,j}|$. The penalized model in (2) is referred to as the sparse tensor graphical model.

3. Proposed Method

3.1. The Separable Estimation Approach

In this section, we propose a scalable and parallelizable approach for solving the tensor graphical lasso problem that is computationally much more efficient than the widely used cyclic updating scheme. The idea is to estimate each $\mathbf{\Omega}_m^*$ separately so that the method is robust and flexible in dealing with different modes of the tensors.

The proposed estimator is obtained through a penalized pseudo-likelihood approach. When estimating $\mathbf{\Omega}_m^*$, suppose that the other $M-1$ precision matrices are fixed at a list of symmetric and positive definite matrices $\{\widetilde{\mathbf{\Omega}}\}_{-m}$. Then we can obtain the estimator $\widehat{\mathbf{\Omega}}_m$ by minimizing the following pseudo-log-likelihood function with respect to $\mathbf{\Omega}_m$:

$$\mathrm{L}_{nm}(\mathbf{\Omega}_m;\{\widetilde{\mathbf{\Omega}}\}_{-m}) = \frac{1}{p_m}\operatorname{tr}\left(\widetilde{\mathbf{S}}_m\mathbf{\Omega}_m\right) - \frac{1}{p_m}\log|\mathbf{\Omega}_m| + \lambda_m||\mathbf{\Omega}_m||_{1,\text{off}}, \quad (3)$$

where $\widetilde{\mathbf{S}}_m = \frac{1}{np_{-m}}\sum_{i=1}^n \widetilde{\mathbf{V}}_i^m(\widetilde{\mathbf{V}}_i^m)^T$ and $\widetilde{\mathbf{V}}_i^m = \mathbf{X}_{i(m)}(\widetilde{\mathbf{\Omega}}_M^{1/2}\otimes\cdots\otimes\widetilde{\mathbf{\Omega}}_{m+1}^{1/2}\otimes\widetilde{\mathbf{\Omega}}_{m-1}^{1/2}\cdots\otimes\widetilde{\mathbf{\Omega}}_1^{1/2})$. Recall that we use normalization to achieve $\|\widetilde{\mathbf{\Omega}}_m\|_F = \|\widehat{\mathbf{\Omega}}_m\|_F = 1$ for all m. For our separable estimation, we use

$$\widetilde{\mathbf{\Omega}}_m = \begin{cases} \left\{\frac{1}{np_{-m}}\sum_{i=1}^n \mathbf{X}_{i(m)}(\mathbf{X}_{i(m)})^T\right\}^{-1}, & np_{-m} > p_m(p_m-1)/2; \\ \mathbf{I}_{p_m}, & np_{-m} \le p_m(p_m-1)/2. \end{cases} \quad (4)$$

Clearly, the key advantage of our separable estimation is that the optimization problem (3) for $\mathbf{\Omega}_m$ does not depend on the sparse estimators for other precision matrices $\{\widehat{\mathbf{\Omega}}\}_{-m}$. Moreover, the precision matrices $\{\widetilde{\mathbf{\Omega}}\}_{-m}$ used in optimization (4) are adaptive to the tensor dimensions and sample size. The number $n_m = np_{-m}$ is the effective sample size for estimating $\mathbf{\Omega}_m$, which has $p_m(p_m-1)/2$ unique parameters.

We present the following results for the Kronecker separable covariance model and matrix/tensor normal distribution to gain more insight into our estimator. Similar results have been obtained in recent studies (Lyu et al. 2019; Pan, Mai and Zhang 2019; Drton, Kuriki and Hoff 2020).

Lemma 1. Suppose that $\mathbf{X}_1,\ldots,\mathbf{X}_n$ are iid from $\mathrm{TN}(\mathbf{0};\mathbf{\Sigma}_1^*,\ldots,\mathbf{\Sigma}_M^*)$. When $np_{-j} > p_j$, $\sum_{i=1}^n \mathbf{X}_{i(j)}(\mathbf{X}_{i(j)})^T$ is positive definite with probability 1. When $\{\widetilde{\mathbf{\Omega}}\}_{-m} = \{\boldsymbol{\alpha}\}_{-m}$ is non-stochastic, the sample covariance matrix $\widetilde{\mathbf{S}}_m$ in Equation (3) has expected value $\mathrm{E}(\widetilde{\mathbf{S}}_m \mid \{\widetilde{\mathbf{\Omega}}\}_{-m} = \{\boldsymbol{\alpha}\}_{-m}) = \frac{1}{p_{-m}}\left\{\prod_{j\neq m}\operatorname{tr}\left(\mathbf{\Sigma}_j^*\boldsymbol{\alpha}_j\right)\right\}\mathbf{\Sigma}_m^*$.

By Lemma 1, when $n_j > p_j(p_j-1)/2$, the sample covariance matrix $\widetilde{\mathbf{S}}_m$ in Equation (3) is positive-definite with probability 1. Moreover, it is an unbiased estimator for $\mathbf{\Sigma}_m^*$ after scaling when we simply use any fixed $\{\widetilde{\mathbf{\Omega}}\}_{-m}$, for example, the identity matrix in (4). Of course, when the sample size increases, we expect the well-conditioned sample estimator $\widetilde{\mathbf{\Omega}}_m = \left\{\frac{1}{np_{-m}}\sum_{i=1}^n \mathbf{X}_{i(m)}(\mathbf{X}_{i(m)})^T\right\}^{-1}$ improves estimation accuracy over the identity matrix.

Algorithm 1 Parallel algorithm for tensor Gaussian graphical models

1. **Input:** Tensor samples $\mathbf{X}_1,\ldots,\mathbf{X}_n$, tuning parameters $\lambda_1,\ldots,\lambda_M$.
2. **Initialization:** For each $m\in\{1,\ldots,M\}$, compute $\widetilde{\mathbf{\Omega}}_m$ using (4).
3. **Sparse Estimation:** For each $m\in\{1,\ldots,M\}$, solve the optimization problem (3) for $\widehat{\mathbf{\Omega}}_m$ using the glasso algorithm (Friedman, Hastie and Tibshirani 2008).

After obtaining $\{\widetilde{\mathbf{\Omega}}\}_{-m}$ from Equation (4), solving the optimization problem in Equation (3) is the same as solving a classical graphical lasso problem (Friedman, Hastie and Tibshirani 2008). When estimating a list of precision matrices, minimizing L_{nm} is independent of minimizing $\mathrm{L}_{nm'}$. Thus, the set of optimization problems are separable and can be solved simultaneously by implementing parallel computing. We summarize our scalable and parallelizable algorithm in Algorithm 1. Even when comparing it with the one-iteration estimator from cyclic algorithms, we still see a decrease in computation cost in the numerical studies. The M sparse estimation problems in Algorithm 1 are parallel, making the proposed method differs from the one-iteration estimator. It has better performance, especially when some $\mathbf{\Omega}_m^*$ is not very sparse and can not be estimated well by sparse solutions.

3.2. Theoretical Properties

We show that the proposed estimator converges to the true precision matrix at an optimal rate. Since each optimization problem is independent of the others and can be solved separately or simultaneously by parallel computing, we present the theoretical results for an arbitrary mode, that is, analysis only on $\widehat{\mathbf{\Omega}}_m$.

We explicitly express the sample-based minimization to Equation (3) and the proposed estimator $\widehat{\mathbf{\Omega}}_m$, which requires normalization, as follows,

$$\widehat{M}_m\left(\{\widetilde{\mathbf{\Omega}}\}_{-m}\right) = \arg\min_{\mathbf{\Omega}_m}\mathrm{L}_{nm}\left(\mathbf{\Omega}_m;\{\widetilde{\mathbf{\Omega}}\}_{-m}\right),$$
$$\widehat{\mathbf{\Omega}}_m = \frac{\widehat{M}_m\left(\{\widetilde{\mathbf{\Omega}}\}_{-m}\right)}{\left\|\widehat{M}_m\left(\{\widetilde{\mathbf{\Omega}}\}_{-m}\right)\right\|_F}. \quad (5)$$

To characterize the sparsity of the true precision matrix $\mathbf{\Omega}_m^*$, we define a sparsity parameter $s_m = |\mathbb{S}_m| - p_m$, where $\mathbb{S}_m = \left\{(i,j):\left[\mathbf{\Omega}_m^*\right]_{i,j}\neq 0\right\}$. Then s_m is the number of nonzero off-diagonal elements in $\mathbf{\Omega}_m^*$. We construct the convergence theory for the proposed estimator in the following theorem, under two mild technical conditions.

Condition 1. Bounded Eigenvalues. For any $m = 1,\ldots,M$, there is a constant $C_1 > 0$ such that $0 < C_1 \le \lambda_{\min}\left(\mathbf{\Sigma}_m^*\right) \le \lambda_{\max}\left(\mathbf{\Sigma}_m^*\right) \le 1/C_1 < \infty$, where $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the minimal and maximal eigenvalues.

Condition 2. Tuning. There is a constant $C_2 > 0$ such that the tuning parameter λ_m satisfies $1/C_2\sqrt{\log p_m/\left(npp_m\right)} \le \lambda_m \le C_2\sqrt{\log p_m/\left(npp_m\right)}$.

Theorem 1. Suppose that Conditions (C1) and (C2) hold. The proposed estimator $\widehat{\mathbf{\Omega}}_m$ in (5) satisfies $\left\|\widehat{\mathbf{\Omega}}_m - \mathbf{\Omega}_m^*\right\|_F = O_P\left(\sqrt{\frac{(p_m+s_m)\log p_m}{np_{-m}}}\right)$.

Condition (C1) requires that the eigenvalues of the true covariance matrices are bounded uniformly. It is a commonly used assumption to study the estimation consistency in graphical models (He et al. 2014; Xu, Zhang, and Gu 2017; Lyu et al. 2019). Condition (C2) gives the growth rate of the tuning parameters in terms of n and p. This type of assumption is also widely used in penalized methods to control the estimation bias and sparsity (Zhou 2014; Lyu et al. 2019).

Theorem 1 shows that the proposed estimator $\widehat{\mathbf{\Omega}}_m$ converges to the true precision matrix $\mathbf{\Omega}_m^*$ at a rate of $\sqrt{\left(p_m + s_m\right)\log p_m/(np_{-m})}$ in the Frobenius norm. The same convergence rate is achieved by the Tlasso estimator after one iteration and it is minimax-optimal (Cai, Liu and Zhou 2016; Lyu et al. 2019). When the true precision matrices are known and let $\{\widetilde{\mathbf{\Omega}}\}_{-m} = \{\mathbf{\Omega}^*\}_{-m}$, the effective sample size for estimating $\widetilde{\mathbf{S}}_m$ is np_{-m} since $\widetilde{\mathbf{V}}_i^m$ has independent columns. From the analysis in Cai, Liu and Zhou (2016), the best rate one can obtain is $\sqrt{\left(p_m + s_m\right)\log p_m/(np_{-m})}$ which matches ours. Therefore, the convergence rate in Theorem 1 is also minimax-optimal.

We compare the theoretical results with other existing works. Leng and Tang (2012) showed the existence of a local minimizer that enjoyed the same convergence rate as ours at $M = 2$. However, they did not show that their algorithm could guarantee the local minimizer. Yin and Li (2012) established the convergence rates at $\sqrt{p_2(p_1 + s_1)(\log p_1 + \log p_2)/n}$ which was slightly slower. To achieve the optimal estimation error, Xu, Zhang, and Gu (2017) required the number of iteration to be no less than $C\log(np/(p_m s_m))$ where C is a positive constant. In contrast, our proposed method does not have convergence concerns and enjoys the same optimal convergence property.

3.3. Connection With Existing Approaches

Clearly when $M = 1$, our method reduces to the sparse Gaussian graphical model which has been extensively studied (e.g., Yuan and Lin 2007; Banerjee, Ghaoui and d'Aspremont 2008; Friedman, Hastie and Tibshirani 2008; Rothman et al. 2008; Ravikumar et al. 2011). When $M = 2$, our method is then applicable to the sparse matrix graphical model. Various iterative algorithm can be viewed as the matrix version of our problem (Leng and Tang see 2012; Yin and Li see 2012; Tsiligkaridis, Hero III and Zhou see 2013). The only exception is Zhou (2014), which proposed a noncyclic algorithm for inverse correlation matrix estimation. Our method extends Zhou (2014) from matrix to tensor, and from inverse correlation to inverse covariance (which is easier to interpret in graphical models), and the plug-in estimator $\widetilde{\mathbf{\Omega}}_j$ is not fixed as $\mathbf{I}_{p_j}$ for $j \ne m$ for more efficient estimation in larger sample scenarios. Since the effective sample size of tensor data is generally larger than the actual sample size, we can get an improved estimation.

For the sparse tensor graphical model with a general $M \ge 2$, He et al. (2014) and Lyu et al. (2019) used a ℓ_1 penalty and provided a coordinate descent-based algorithm; Xu, Zhang, and Gu (2017) used a l_0 penalty and proposed an alternating gradient descent algorithm. To the best of our knowledge, all existing methods propose iterative cyclic-updating algorithms.

The proposed separable and parallelizable algorithm circumvents many limitations in the cyclic algorithms. First, the efficiency of iterative algorithms is a concern due to the high dimensionality of tensor data. The new algorithm significantly reduces the computational cost through parallelism. Even comparing with the first iteration in the cyclic algorithms, we still see a decrease in computation cost in numerical studies. Second, the choice of initial value and the convergence speed of the algorithm should be considered for iterative approaches. The initialization step in Algorithm 1 offers a simple initial value with theoretical guarantees. Thirdly, with a well-chosen initial value for $\widetilde{\mathbf{\Omega}}_m$, we can finely tune λ_m at low computational cost for each mode. In practice, the more affordable and efficient tuning process can further improve performance. Finally, the cyclically iterative methods assume that all the precision matrices are sparse, which can be restrictive if we are only interested in estimating a subset of the precision matrices. In contrast, thanks to the independence of the optimization problems in our algorithm, we can be more flexible at scenarios with different sparsity levels. In Section B of the supplementary materials, we will discuss the applications to partially sparse graphical models, prespecified covariance structured (e.g. longitudinal, spatial–temporal models), and joint model of mean and covariance in regression.

4. Simulations

We compare the proposed estimator (Separate) with three alternative solutions: (Oracle) The oracle estimator using the true parameter $\mathbf{\Omega}_{-m}^*$ when estimates $\mathbf{\Omega}_m^*$ in (3); (Cyclic) The cyclic-updates algorithm to the penalized MLE problem (2), where each sparse precision matrix is updated using current estimates of other sparse precision matrices and and convergence is often reached after 5–100 iterations; and ("Sequential") The cyclic algorithm with only one iteration, which means $\widehat{\mathbf{\Omega}}_m$ is obtained using the sparse estimators $\widehat{\mathbf{\Omega}}_j, j < m$.

For fair comparison, we use the same R package `glasso` to compute the ℓ_1 penalized optimizations for all four estimators. We expect similar results if different penalties or packages are used consistently for the four estimators. Additional implementation details are given in the supplementary materials.

We consider the following three types of covariance/precision matrices $\mathbf{\Sigma}_m^*, \mathbf{\Omega}_m^* \in \mathbb{R}^{p_m \times p_m}$.

- Triangle (TR) covariance. We set $\left[\mathbf{\Sigma}_m^*\right]_{i,j} = \exp\left(-\left|h_i - h_j\right|/2\right)$ with $h_1 < h_2 < \cdots < h_{p_m}$. The difference $h_i - h_{i-1}, i = 2, \ldots, p_m$, is generated iid from Unif(0.5, 1). This generated covariance matrix mimics autoregressive process of order one, that is, AR(1).

- Autoregressive (AR) precision. We set $\left[\mathbf{\Omega}_m^*\right]_{i,j} = 0.8^{|i-j|}$.
- Compound symmetry (CS) precision. We set $\left[\mathbf{\Omega}_m^*\right]_{i,j} = 0.6$ for $i \neq j$ and $\left[\mathbf{\Omega}_m^*\right]_{i,i} = 1$.

Note that the first model (TR) has a sparse precision matrix while the other two (AR and CS) have a nonsparse precision matrix.

We consider six models as follows. Models 1–3 are fully sparse models, and Models 4–6 are partially sparse models. In Models 3 and 6, we consider the unbalanced setting in dimension. In all models, we normalize the precision matrices such that $\|\mathbf{\Omega}_m^*\|_F = 1$ for $m = 1, \ldots, M$. For all the following models, we set $M = 3$ and $n = 100$.

- Model 1: $\mathbf{\Omega}_1^*, \mathbf{\Omega}_2^*, \mathbf{\Omega}_3^*$ are all from the TR covariance model, $(p_1, p_2, p_3) = (30, 36, 30)$.
- Model 2: $\mathbf{\Omega}_1^*, \mathbf{\Omega}_2^*, \mathbf{\Omega}_3^*$ are all from the TR covariance model, $(p_1, p_2, p_3) = (100, 100, 100)$.
- Model 3: $\mathbf{\Omega}_1^*, \mathbf{\Omega}_2^*, \mathbf{\Omega}_3^*$ are all from the TR covariance model, $(p_1, p_2, p_3) = (5, 5, 500)$.
- Model 4: Same as Model 1, except for $\mathbf{\Omega}_1^* = \mathrm{AR}(0.8)$.
- Model 5: Same as Model 1, except for $\mathbf{\Omega}_1^* = \mathbf{\Omega}_2^* = \mathrm{CS}(0.6)$.
- Model 6: Same as Model 3, except for $\mathbf{\Omega}_1^* = \mathbf{\Omega}_2^* = \mathrm{CS}(0.6)$.

We also consider the six models but with sample size $n = 20$ and another two fully sparse models with $M = 4$ and $M = 5$. The results show similar findings and are included in the supplementary materials. To measure the estimation performance, we report the errors in Frobenius norm $\|\widehat{\mathbf{\Omega}}_m - \mathbf{\Omega}_m^*\|_F$ and the errors in max norm $\|\widehat{\mathbf{\Omega}}_m - \mathbf{\Omega}_m^*\|_{\max}$. We also report the true positive rate and true negative rate to measure the performance of support recovery. For fully sparse models, we further report the averages of these criteria across all modes.

The computation time for Models 1–3 is summarized in Table 1. The proposed separable estimation approach is much faster than the cyclic-updating approach, especially as the dimension increases. The computational costs of Separable, Oracle and Sequential methods are roughly the same, as we expected. For Models 4–6, the computation time is not included because the sequential and cyclic methods failed in these partially sparse models.

The results of estimation accuracy are summarized in Table 2. For both fully sparse models and partially sparse models, we can observe that the separable estimator outperforms the sequential one-iteration estimator and cyclic estimator on every mode and is comparable to the oracle estimator. In particular, when the dimension is unbalanced in Models 3 and 6, the cyclic estimator performs much worse than the proposed method. Moreover, comparing Models 1 and 2 with Models 4 and 5, we notice that

Table 1. Averaged computation cost (in sec) and standard errors (in parentheses) for Models 1–3 from 20 replicates.

	(p_1, p_2, p_3)	Separate	Oracle	Sequential	Cyclic
Model 1	(30,36,30)	1.17 (0.01)	0.84 (0.01)	3.10 (0.01)	13.24 (0.92)
Model 2	(100,100,100)	44.13 (0.07)	34.44 (0.11)	93.22 (0.05)	379.13 (1.88)
Model 3	(5,5,500)	14.21 (0.04)	13.90 (0.03)	13.94 (0.02)	120.91 (0.07)

Table 2. Comparison of means and the maximum standard errors (in parentheses) of different performance measures for Models 1–3 from 100 replicates.

(%)	Model 1				Model 2				Model 3				
	Sep.	Ora.	Seq.	Cyc.	Sep.	Ora.	Seq.	Cyc.	Sep.	Ora.	Seq.	Cyc.	S.E.≤
						Frobenius norm loss							
AVG.	0.62	0.62	2.34	0.80	0.21	0.21	1.28	0.27	1.97	1.83	3.33	13.74	(0.01)
$m = 1$	0.59	0.59	3.74	0.73	0.21	0.21	2.29	0.27	0.50	0.30	0.79	1.87	(0.02)
$m = 2$	0.66	0.65	2.58	0.92	0.21	0.21	1.29	0.27	0.52	0.31	2.43	2.05	(0.02)
$m = 3$	0.62	0.62	0.71	0.75	0.22	0.21	0.26	0.27	4.88	4.88	6.77	37.29	(0.01)
						Max norm loss							
AVG.	0.15	0.15	0.29	0.15	0.03	0.03	0.06	0.03	0.27	0.21	0.61	0.78	(0.01)
$m = 1$	0.15	0.14	0.44	0.15	0.03	0.03	0.10	0.03	0.22	0.14	0.37	0.81	(0.01)
$m = 2$	0.16	0.15	0.27	0.16	0.03	0.03	0.06	0.03	0.23	0.14	1.08	0.89	(0.01)
$m = 3$	0.16	0.16	0.17	0.16	0.04	0.03	0.04	0.04	0.37	0.36	0.38	0.64	(0.01)
						True negative rate							
AVG.	66.21	66.19	19.28	41.08	81.88	81.42	26.09	62.18	45.96	46.02	54.34	27.32	(1.21)
$m = 1$	65.77	65.75	0.07	43.55	82.04	81.50	0.04	62.19	20.25	21.00	40.83	37.83	(2.2)
$m = 2$	67.95	68.06	2.29	36.46	81.58	81.10	2.76	62.04	23.83	24.33	42.17	42.83	(2.48)
$m = 3$	64.9	64.75	55.50	43.23	82.01	81.67	75.48	62.31	93.79	92.72	80.02	1.30	(0.52)
(%)	Model 4				Model 5				Model 6				
	Sep.	Ora.	Seq.	Cyc.	Sep.	Ora.	Seq.	Cyc.	Sep.	Ora.	Seq.	Cyc.	S.E.≤
						Frobenius norm loss							
$m = 2$	0.67	0.66	2.78	1.64	–	–	–	–	–	–	–	–	(0.01)
$m = 3$	0.60	0.59	0.71	1.37	0.61	0.61	1.40	1.39	4.91	4.88	13.41	38.64	(0.02)
						Max norm loss							
$m = 2$	0.16	0.16	0.29	0.17	–	–	–	–	–	–	–	–	(0.01)
$m = 3$	0.16	0.15	0.16	0.17	0.16	0.16	0.17	0.17	0.36	0.36	0.36	0.66	(0.01)
						True negative rate							
$m = 2$	67.59	67.28	1.47	0.68	–	–	–	–	–	–	–	–	(0.35)
$m = 3$	64.48	64.56	48.63	0.81	65.77	65.63	0	0	91.34	92.17	47.34	0.18	(0.83)

NOTES: All methods have achieved 100% true positive rate (and hence not shown in the table). Comparison of means and the maximum standard errors (in parentheses) of different performance measures for Models 4–6 from 100 replicates. Results for $m = 2$ in Models 5 and 6 are not reported because $\mathbf{\Omega}_2^*$ is not sparse.

the cyclically iterative algorithm cannot improve the estimation accuracy in the iteration for partially sparse models. This confirms our concern that violation of sparsity assumption hurts the performance of iterative methods. It also indicates that the performance of the proposed method is more stable and consistent. The proposed method also shows superior performance in terms of support recovery.

5. Real Data Analysis

As an illustration, we apply the proposed method on the electroencephalography (EEG) data from a study to examine EEG correlates of genetic predisposition to alcoholism. The dataset is available at *http://kdd.ics.uci.edu/databases/eeg/*. In the study, 64 channels of electrodes were placed on the scalp at standard sites. There were 122 subjects among which 77 were alcohol individuals and 45 were nonalcoholic individuals. Each subject had 120 trials under exposure to different picture stimuli and measurements were collected from the electrodes which were sampled at 256 Hz (3.9-msec epoch) for 1 second. More collection details could be found in Zhang et al. (1995). We used the same part of the data that was analyzed in Li, Kim and Altman (2010). The data focused on single stimulus condition and were averages from all the corresponding trials. Therefore, each EEG sample was a 256×64 matrix. We further downsized it to 64×64. We divided the dataset into an alcoholic group and a nonalcoholic group. For each group, we standardized the data and applied the proposed method to estimate the precision matrix for channels. Tuning parameter was chosen by five-fold cross-validation.

The correlation networks identified by the proposed method are displayed in Figure 1. To have a clear insight of the graph pattern, we only present the top 100 correlations. We can observe that channels located close to each other tend to be correlated. For example, channels F4, F2, FZ, F1, F3, and F5 that are physically placed in line on the scalp are strongly correlated. The centered channels such as AFZ and OZ also have strong connections with the adjacent nodes AF1, AF2 and O1, O2. We also notice that the correlations among POZ and PO1, PO2 are stronger in nonalcoholic group than in alcoholic group. As POZ is located near the parietal-occipital junction, there is a large probability for it to be most sensitive to the visual stimuli (Lei and Liao 2017). A possible explanation for the weaker correlations in alcoholic group is that alcohol can slow down the brain's information processing (Tzambazis and Stough 2000).

The results from the cyclic algorithm and the colored version of the network graphs are given in the supplementary materials. While the cyclic algorithm can identify most of the strong correlation pairs, the total number of correlation pairs is about two or three times of the number identified by the proposed method. For nonalcoholic group, the proposed method identifies 387 and 425 correlation pairs respectively for the two modes, while the cyclic algorithm identifies 888 and 1598 pairs. For alcoholic group, the proposed method identifies 416 and 441 pairs and the cyclic algorithm identifies 947 and 1536 pairs. The separable method offers a more sparse solution that is easier to interpret.

Supplementary Materials

Supp.pdf. This file contains proofs, additional numerical results, and application to partially sparse model. (pdf)

Rcode.zip. The file includes R code files for simulation and real data analysis and a readme file for demonstration. (zip)

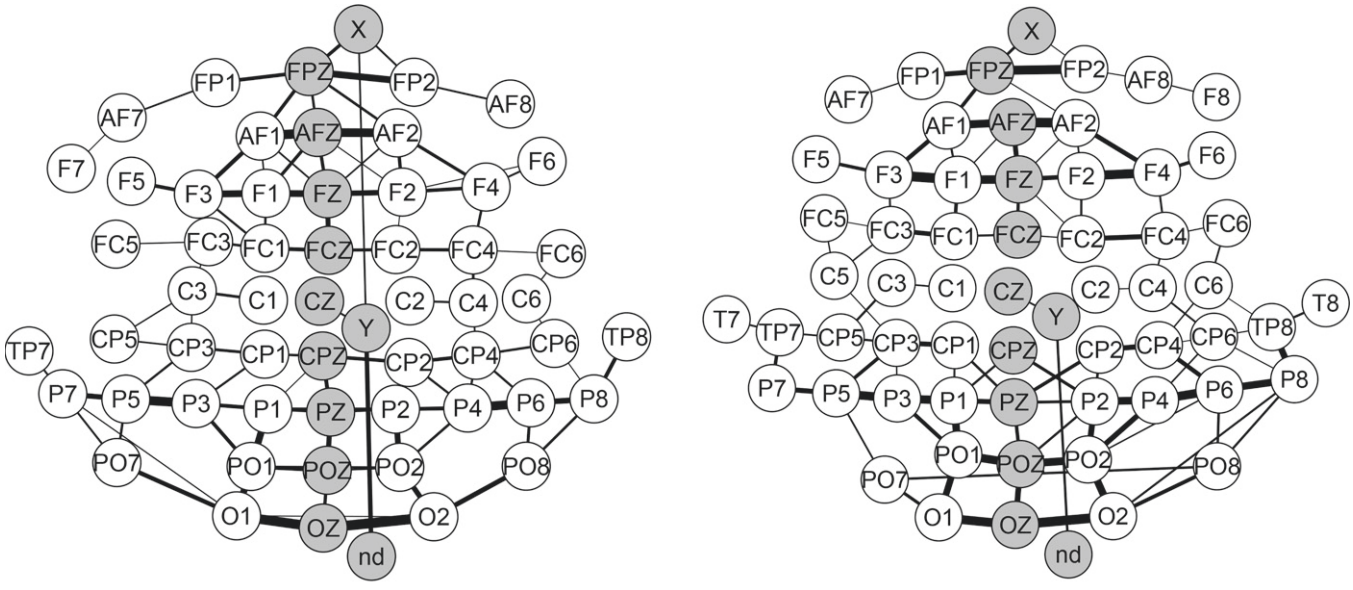

Figure 1. Correlation networks constructed by the proposed method using the estimated precision matrices among channels for alcoholic group and nonalcoholic group. Only the first 100 strongest correlations are displayed. Nodes are labeled with EEG electrode identifiers. The gray nodes are placed on the middle of the scalp whose left and right nodes are placed on the left and right of the scalp respectively. Nodes that are not correlated with the others are not included. The thickness of edges represents the magnitude of correlations.

Acknowledgments

The authors thank to the editor, the associate editor and two reviewers for helpful comments. Research for this paper was supported in part by grants CCF-1908969 from the National Science Foundation.

References

Banerjee, O., Ghaoui, L. E., and d'Aspremont, A. (2008), "Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data," *Journal of Machine Learning Research*, 9, 485–516. [297]

Bi, X., Tang, X., Yuan, Y., Zhang, Y., and Qu, A. (2021), "Tensors in Statistics," *Annual Review of Statistics and Its Application*, 8, 345–336. [294]

Cai, T. T., Liu, W., and Zhou, H. H. (2016), "Estimating Sparse Precision Matrix: Optimal Rates of Convergence and Adaptive Estimation," *The Annals of Statistics*, 44, 455–488. [294,297]

Chi, E. C., and Kolda, T. G. (2012), "On Tensors, Sparsity, and Nonnegative Factorizations," *SIAM Journal on Matrix Analysis and Applications*, 33, 1272–1299. [294]

Danaher, P., Wang, P., and Witten, D. M. (2014), "The Joint Graphical Lasso for Inverse Covariance Estimation Across Multiple Classes," *Journal of the Royal Statistical Society*, Series B, 76, 373. [294]

Drton, M., Kuriki, S., and Hoff, P. (2020), "Existence and Uniqueness of the Kronecker Covariance MLE," arXiv no. 2003.06024 . [296]

Friedman, J., Hastie, T., and Tibshirani, R. (2008), "Sparse Inverse Covariance Estimation With the Graphical Lasso," *Biostatistics*, 9, 432–441. [294,296,297]

Gupta, A. K., and Nagar, D. K. (2018), *Matrix Variate Distributions*, Boca Ration: Chapman and Hall/CRC. [295]

He, S., Yin, J., Li, H. and Wang, X. (2014), "Graphical Model Selection and Estimation for High Dimensional Tensor Data," *Journal of Multivariate Analysis*, 128, 165–185. [294,297]

Hoff, P. D. (2015), "Multilinear Tensor Regression for Longitudinal Relational Data," *The Annals of Applied Statistics*, 9, 1169–1193. [294]

Kolda, T. G., and Bader, B. W. (2009), "Tensor Decompositions and Applications," *SIAM Review*, 51, 455–500. [295]

Lauritzen, S. L. (1996), *Graphical Models* (Vol. 17), Oxford: Clarendon Press. [294,295]

Lei, X., and Liao, K. (2017), "Understanding the Influences of Eeg Reference: A Large-Scale Brain Network Perspective," *Frontiers in Neuroscience*, 11, 205. [299]

Leng, C., and Tang, C. Y. (2012), "Sparse Matrix Graphical Models," *Journal of the American Statistical Association*, 107, 1187–1200. [294,297]

Li, B., Kim, M. K., and Altman, N. (2010), "On Dimension Folding of Matrix-or Array-Valued Statistical Objects," *The Annals of Statistics*, 38, 1094–1121. [299]

Lock, E. F. (2018), "Tensor-on-Tensor Regression," *Journal of Computational and Graphical Statistics*, 27, 638–647. [294]

Lyu, X., Sun, W. W., Wang, Z., Liu, H., Yang, J. and Cheng, G. (2019), "Tensor Graphical Model: Non-Convex Optimization and Statistical Inference," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42. [294,296,297]

Molstad, A. J., and Rothman, A. J. (2018), "Shrinking Characteristics of Precision Matrix Estimators," *Biometrika*, 105, 563–574. [294]

Pan, Y., Mai, Q., and Zhang, X. (2019), "Covariate-Adjusted Tensor Classification in High Dimensions," *Journal of the American Statistical Association*, 114, 1305–1319. [294,296]

Pfeiffer, R. M., Kapla, D. B., and Bura, E. (2021), "Least Squares and Maximum Likelihood Estimation of Sufficient Reductions in Regressions With Matrix-Valued Predictors," *International Journal of Data Science and Analytics*, 11, 11–26. [294]

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011), "High-Dimensional Covariance Estimation by Minimizing l1-Penalized Log-Determinant Divergence," *Electronic Journal of Statistics*, 5, 935–980. [297]

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008), "Sparse Permutation Invariant Covariance Estimation," *Electronic Journal of Statistics*, 2, 494–515. [297]

Tsiligkaridis, T., Hero III, A. O., and Zhou, S. (2013), "On Convergence of Kronecker Graphical Lasso Algorithms," *IEEE Transactions on Signal Processing*, 61, 1743–1755. [294,297]

Tzambazis, K., and Stough, C. (2000), "Alcohol Impairs Speed of Information Processing and Simple and Choice Reaction Time and Differentially Impairs Higher-Order Cognitive Abilities," *Alcohol and Alcoholism*, 35, 197–201. [299]

Witten, D. M., Friedman, J. H., and Simon, N. (2011), "New Insights and Faster Computations for the Graphical Lasso," *Journal of Computational and Graphical Statistics*, 20, 892–900. [294]

Xu, P., Zhang, T., and Gu, Q. (2017), "Efficient Algorithm for Sparse Tensor-Variate Gaussian Graphical Models Via Gradient Descent," in Aarti Singh and Jerry Zhu (eds.), *Artificial Intelligence and Statistics*, pp. 923–932. [294,297]

Yin, J., and Li, H. (2012), "Model Selection and Estimation in the Matrix Normal Graphical Model," *Journal of Multivariate Analysis*, 107, 119–140. [294,297]

Yuan, M., and Lin, Y. (2007), "Model Selection and Estimation in the Gaussian Graphical Model," *Biometrika*, 94, 19–35. [294,297]

Zhang, T., and Zou, H. (2014), "Sparse Precision Matrix Estimation Via Lasso Penalized d-Trace Loss," *Biometrika*, 101, 103–120. [294]

Zhang, X. L., Begleiter, H., Porjesz, B., Wang, W. and Litke, A. (1995), "Event Related Potentials During Object Recognition Tasks," *Brain Research Bulletin*, 38, 531–538. [299]

Zhou, H., Li, L., and Zhu, H. (2013), "Tensor Regression With Applications in Neuroimaging Data Analysis," *Journal of the American Statistical Association*, 108, 540–552. [294]

Zhou, S. (2014), "Gemini: Graph Estimation With Matrix Variate Normal Instances," *The Annals of Statistics*, 42, 532–562. [294,297]

Zhu, Y., and Li, L. (2018), "Multiple Matrix Gaussian Graphs Estimation," *Journal of the Royal Statistical Society*, Series B, 80, 927–850. [294]