Journal of Multivariate Analysis 107 (2012) 119–140

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

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

Model selection and estimation in the matrix normal graphical model

Jianxin Yin, Hongzhe Li *

School of Statistics and Center for Applied Statistics, Renmin University of China, No. 59 Zhongguancun Street, Haidian District, Beijing 100872, China
Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104-6021, USA

ARTICLE INFO

Article history:
Received 1 June 2011
Available online 10 January 2012

AMS subject classifications:
62
92

Keywords:
Gaussian graphical model
Gene networks
High dimensional data
l_1 penalized likelihood
Matrix normal distribution
Sparsistency

ABSTRACT

Motivated by analysis of gene expression data measured over different tissues or over time, we consider matrix-valued random variable and matrix-normal distribution, where the precision matrices have a graphical interpretation for genes and tissues, respectively. We present a l_1 penalized likelihood method and an efficient coordinate descent-based computational algorithm for model selection and estimation in such matrix normal graphical models (MNGMs). We provide theoretical results on the asymptotic distributions, the rates of convergence of the estimates and the sparsistency, allowing both the numbers of genes and tissues to diverge as the sample size goes to infinity. Simulation results demonstrate that the MNGMs can lead to a better estimate of the precision matrices and better identifications of the graph structures than the standard Gaussian graphical models. We illustrate the methods with an analysis of mouse gene expression data measured over ten different tissues.

1. Introduction

Gaussian graphical models (GGMs) provide natural tools for modeling the conditional independence relationships among a set of random variables [23,37]. Many methods of estimating the standard GGMs have been developed in recent years, especially in high-dimensional settings. Meinshausen and Bühlmann [27] took a neighborhood selection approach to this problem by fitting a l_1 penalized regression or Lasso [33] to each variable using the other variables as predictors. They show that this neighborhood selection procedure consistently estimates the set of non-zero elements of the precision matrix. Other authors have proposed algorithms for the exact maximization of the l_1-penalized log-likelihood. Yuan and Lin [39], Banerjee et al. [4] and Dahl et al. [9] adapted an interior point optimization method for the solution to this problem. Based on the work of Banerjee et al. [4] and a block-wise coordinate descent algorithm, Friedman et al. [16] developed the graphical Lasso (glasso) for sparse inverse covariance estimation, which is computationally very efficient even when the dimension is greater than the sample size. Yuan [38] developed a linear programming procedure for high dimensional inverse covariance matrix estimation and obtained oracle inequalities for the estimation error in terms of several matrix norms. Some theoretical properties of this type of methods have also been developed by Yuan and Lin [39], Ravikumar et al. [30], Rothman et al. [31] and Lam and Fan [22]. Cai et al. [6] developed a constrained l_1 minimization approach to sparse precision matrix estimation, extending the idea of the Dantzig selector [7] developed for sparse high dimensional regressions.

The standard likelihood framework for building Gaussian graphical models assumes that samples are independent and identically distributed from a multivariate Gaussian distribution. This assumption is often limited in certain applications. For example, in genomics, gene expression data of p genes collected over q different tissues from the same subject are

* Corresponding author at: Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104-6021, USA.
E-mail address: hongzhe@upenn.edu (H. Li).

0047-259X/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2012.01.005

often correlated. For a given sample, let $\mathbf{Y}$ be the $p \times q$ matrix of the expression data, where the jth column corresponds to the expression data of p genes measured in the jth tissue, and the ith row corresponds to gene expressions of the ith gene over q different tissues. Instead of assuming that the columns or rows are independent, we assume that the matrix variate random variable $\mathbf{Y}$ follows a matrix normal distribution [10,23,18], where both row and column precision matrices can be specified. The matrix-variate normal distribution has been studied in analysis of multivariate linear model under the assumption of independence and homoscedasticity for the structure of the among-row and among-column covariance matrices of the observation matrix [15,34]. Such a model has also be applied to spatio-temporal data [26,21]. In genomics, Teng and Huang [32] proposed to use the Kronecker product matrix to model gene-experiment interactions, which leads to a gene expression matrix following a matrix-normal distribution. The gene expression matrix measured over multiple tissues is transposable, meaning that potentially both the rows and/or columns are correlated. Such matrix-valued normal distribution was also used in [2,12] for modeling gene expression data in order to account for gene expression dependency across different experiments. Dutilleul [11] developed the maximum likelihood estimation (MLE) algorithm for the matrix normal distribution. Mitchell et al. [28] developed a likelihood ratio test for separability of the covariances. Muralidharan [29] used a matrix normal framework for detecting column dependence when rows are correlated and estimating the strength of the row correlation.

The precision matrices of the matrix normal distribution provide the conditional independence structures of the row and column variables [23], where the non-zero off-diagonal elements of the precision matrices correspond to conditional dependencies among the elements in row or column of the matrix normal distribution. The matrix normal models with specified non-zero elements of the precision matrices define the matrix normal graphical models (MNGMs). This is analogous to the relationship between the Gaussian graphical model and the precision matrix of a multivariate normal distribution. Despite the flexibility of the matrix normal distribution and the MNGMs in modeling the transposable data, methods for model selection and estimation of such models have not been developed fully, especially in high dimensional settings. Wang and West [36] developed a Bayesian approach for the MNGMs using Markov Chain Monte Carlo sampling scheme that employs an efficient method for simulating hyper-inverse Wishart variates for both decomposable and nondecomposable graphs. Allen and Tibshirani [2,3] proposed penalized likelihood approaches for such matrix normal models, where both l_1-norm and l_2-norm penalty functions are used on the precision matrices.

The focus of this paper is to develop a model selection and estimation method for the MNGMs based on a l_1 penalized likelihood approach under the assumption of both row and column precision matrices being sparse. Our penalized estimation method is the same as that proposed in [2,1,3] when l_1 penalty is used. Allen and Tibshirani [2,3] only considered the setting when there is one observed matrix-variate normal data and used the estimated covariance matrices for imputing the missing data and for de-correlating the noise in the underlying data. We focus on evaluating how well such a l_1 penalized estimation method recovers the underlying graphical structures that correspond to the row and column precision matrices when we have n i.i.d. samples from a matrix normal distribution. In addition, we provide asymptotic justification of the estimates and show that the estimates enjoy similar asymptotic and oracle properties as the penalized estimates for the standard GGMs [13,22,39] even when the dimensions $p = p_n$ and $q = q_n$ diverge as the number of observations $n \to \infty$. In addition, if consistent estimates of the precision matrices are available and are used in the adaptive l_1 penalty functions, the resulting estimates have the property of sparsistency.

The rest of the paper is organized as follows. We introduce the MNGMs as motivated by analysis of gene expression data across multiple tissues in Section 2. In Section 3 we present a l_1 penalized likelihood estimate of such a MNGM and an iterative coordinate descent procedure for the optimization. We present the asymptotic properties of the estimates in Section 4 in both the classic setting when the dimensions are fixed and the setting allowing the dimensions to diverge as the sample size goes to infinity. In Section 5 we present simulation results and comparisons with the standard Gaussian graphical model. We present an application of the MNGM in Section 6 to an analysis of mouse gene expression data measured over 10 different tissues. Finally, in Section 7 we give a brief discussion. The proofs of all the theorems are given in the Appendix.

2. Matrix normal graphical model for multi-tissue gene expression data

We consider the gene expression data measured over different tissues. Let $\mathbf{Y}$ be the random $p \times q$ matrix of the gene expression levels of p genes over q tissues. Let vec($\mathbf{A}$) be the vectorization of a matrix $\mathbf{A}$ obtained by stacking the columns of the matrix $\mathbf{A}$ on top of one another. Instead of assuming that the expression levels are independent over different tissues, following [32], we can model this gene expression matrix as

$$\mathbf{Y} = \mathbf{G} + \mathbf{T} + \mathbf{I}_{GT} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{G}$ and $\mathbf{T}$ are expected (constant) effects from the genes and tissues respectively, $\mathbf{I}_{GT}$ are the interaction effects that are assumed to be random with vec($\mathbf{I}_{GT}$) following a multivariate normal distribution with zero means and a covariance matrix $\mathbf{V} \otimes \mathbf{U}$, where the covariance matrices $\mathbf{U}$ and $\mathbf{V}$ respectively represent the gene and tissue dependencies, and $\boldsymbol{\epsilon}$ represents small random normal noises with zero means arising from all nuisance sources. With negligible nuisance effects, vec($\mathbf{Y}$) follows a multivariate normal distribution with means vec($\mathbf{M}$) = vec($\mathbf{G} + \mathbf{T}$) and a covariance matrix $\mathbf{V} \otimes \mathbf{U}$ [32].

Treating the data $\mathbf{Y}$ as a matrix-valued random variable, we say $\mathbf{Y}$ follows a matrix normal distribution, if $\mathbf{Y}$ has a density function

$$p(\mathbf{Y}|\mathbf{M}, \mathbf{U}, \mathbf{V}) = k(\mathbf{U}, \mathbf{V}) \exp(-\mathrm{tr}\{(\mathbf{Y} - \mathbf{M})^T \mathbf{U}^{-1} (\mathbf{Y} - \mathbf{M}) \mathbf{V}^{-1} / 2\}), \tag{2}$$

where $k(\mathbf{U}, \mathbf{V}) = (2\pi)^{-pq/2}|\mathbf{U}|^{-q/2}|\mathbf{V}|^{-p/2}$ is the normalizing constant, $\mathbf{M}$ is the mean matrix, $\mathbf{U}$ is the row covariance matrix and $\mathbf{V}$ is the column covariance matrix. This definition is equivalent to the definition via the Kronecker product [17, Section 8.8 and 9.2]. Specifically,

$$\mathbf{Y} \sim MN_{p,q}(\mathbf{M}; \mathbf{U}, \mathbf{V}) \quad \text{if and only if } \mathrm{vec}(\mathbf{Y}) \sim N_{pq}(\mathrm{vec}(\mathbf{M}), \mathbf{V} \otimes \mathbf{U}). \tag{3}$$

We denote the corresponding precision matrices as $\mathbf{A} = \mathbf{U}^{-1}$, $\mathbf{B} = \mathbf{V}^{-1}$ for $\mathbf{U}$ and $\mathbf{V}$, respectively. This model assumes a particular decomposable covariance matrix for $\mathrm{vec}(\mathbf{Y})$ that is separable in the geostatistics context [8]. The parameters $\mathbf{U}$ and $\mathbf{V}$ are defined up to a positive multiplicative constant. We can set $\mathbf{B}_{11}$ to any positive constant to make the parameters identifiable.

The following proposition shows that there is a graphical model interpretation for the two precision matrices $\mathbf{A}$ and $\mathbf{B}$ in the matrix normal model (2).

Proposition 2.1. *Assume that $\mathbf{Y} \sim MN_{p,q}(\mathbf{M}; \mathbf{U}, \mathbf{V})$. If we partition the columns of $\mathbf{Y}$ as $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_q)$, then it holds for $\gamma, \mu \in \Gamma = \{1, \ldots, q\}$ with $\gamma \neq \mu$ that*

$$\mathbf{y}_\gamma \perp\!\!\!\perp \mathbf{y}_\mu \mid \mathbf{y}_{\Gamma\setminus\{\gamma,\mu\}} \quad \textit{if and only if } b_{\gamma\mu} = 0, \tag{4}$$

where $\mathbf{B} = \{b_{\alpha\beta}\}_{\alpha,\beta\in\Gamma} = \mathbf{V}^{-1}$ is the column precision matrix of the distribution; similarly, if we partition the rows of $\mathbf{Y}$ as $\mathbf{Y} = (\mathbf{y}^1, \ldots, \mathbf{y}^p)^T$, then it holds for $\delta, \eta \in \Xi = \{1, \ldots, p\}$ with $\delta \neq \eta$ that

$$\mathbf{y}^\delta \perp\!\!\!\perp \mathbf{y}^\eta \mid \mathbf{y}^{\Delta\setminus\{\delta,\eta\}} \quad \textit{if and only if } a_{\delta\eta} = 0 \tag{5}$$

where $\mathbf{A} = \{a_{\delta\eta}\}_{\delta,\eta\in\Xi} = \mathbf{U}^{-1}$ is the row precision matrix of the distribution.

This proposition is based on a proposition in [23]. A detailed proof can be found in the Appendix. Without loss of generality, we assume $M = 0$ in this paper since it can be easily estimated.

3. l_1-penalized maximum likelihood estimation of the precision matrices

We propose to estimate the precision matrices $\mathbf{A} = \mathbf{U}^{-1}$, $\mathbf{B} = \mathbf{V}^{-1}$ in model (2) by maximizing a penalized likelihood function. Since for any $c > 0$, $p(\mathbf{Y} \mid \mathbf{A}, \mathbf{B}) = p(\mathbf{Y} \mid c\mathbf{A}, \mathbf{B}/c)$, $\mathbf{A}$ and $\mathbf{B}$ are not uniquely identified. We set $b_{11} = 1$ for the purpose of parameter identification. We propose to estimate $\mathbf{A}$ and $\mathbf{B}$ by minimizing the following penalized negative log-likelihood function

$$\phi(\mathbf{A}, \mathbf{B}) = -q\log(|\mathbf{A}|) - p\log(|\mathbf{B}|) + \frac{1}{n}\sum_{k=1}^{n}\mathrm{tr}\{\mathbf{A}\mathbf{y}_k\mathbf{B}\mathbf{y}_k^T\} + \sum_{i\neq j} p_{\lambda_{ij}}(a_{ij}) + \sum_{i\neq j} p_{\rho_{ij}}(b_{ij}), \tag{6}$$

where $p_{\lambda_{ij}}(\cdot)$ is the penalty function for the element a_{ij} of $\mathbf{A}$ with tuning parameter λ_{ij}, while $p_{\rho_{ij}}(\cdot)$ is the corresponding penalty function for b_{ij} with tuning parameter ρ_{ij}. We consider both l_1-penalty with $p_{\lambda_{ij}}(a_{ij}) = \lambda|a_{ij}|$ and $p_{\rho_{ij}}(b_{ij}) = \rho|b_{ij}|$ and adaptive l_1 penalty with $p_{\lambda_{ij}}(a_{ij}) = \lambda|\tilde{a}_{ij}|^{-\gamma_1}|a_{ij}|$ and $p_{\rho_{ij}}(b_{ij}) = \rho|\tilde{b}_{ij}|^{-\gamma_2}|b_{ij}|$, where $\tilde{\mathbf{A}} = \{\tilde{a}_{ij}\}$ and $\tilde{\mathbf{B}} = \{\tilde{b}_{ij}\}$ are some consistent estimates of $\mathbf{A}$ and $\mathbf{B}$ and $\gamma_1 > 0$ and $\gamma_2 > 0$ are two constants.

It is easy to check that the objective function (6) is a bi-convex function in $\mathbf{A}$ and $\mathbf{B}$. We propose the following iterative procedure to minimize this function:

1. Initialization: $\hat{\mathbf{B}}^{(0)} = \mathbf{I}_q$.
2. In ith step, given the current estimate of $\mathbf{B}$, $\hat{\mathbf{B}}^{(i)}$, we update $\mathbf{A}$ by
$$\hat{\mathbf{A}}^{(i+1)} = \arg\min_{\mathbf{A}} \left\{ -\log(|\mathbf{A}|) + \mathrm{tr}(\hat{\mathbf{S}}_{\mathbf{A}}^{(i)}\mathbf{A}) + \sum_{i\neq j} p^*_{\lambda_{ij}}(a_{ij}) \right\}, \tag{7}$$
where $\hat{\mathbf{S}}_{\mathbf{A}}^{(i)} = 1/(nq)\sum_{k=1}^{n}\mathbf{y}_k\hat{B}^{(i)}\mathbf{y}_k^T$, $\lambda^*_{ij} = \lambda_{ij}/q$.
3. In $(i+1)$th step, given the current estimate of $\mathbf{A}$, $\hat{\mathbf{A}}^{(i+1)}$, we update $\mathbf{B}$ by
$$\hat{\mathbf{B}}^{(i+1)} = \arg\min_{\mathbf{B}} \left\{ -\log(|\mathbf{B}|) + \mathrm{tr}(\hat{\mathbf{S}}_{\mathbf{B}}^{(i+1)}\mathbf{B}) + \sum_{i\neq j} p^*_{\rho_{ij}}(b_{ij}) \right\}, \tag{8}$$
when $\hat{\mathbf{S}}_{\mathbf{B}}^{(i+1)} = 1/(np)\sum_{k=1}^{n}\mathbf{y}_k^T\hat{\mathbf{A}}^{(i+1)}\mathbf{y}_k$, $\rho^*_{ij} = \rho_{ij}/p$.
4. Iterate Steps 2 and 3 until convergence.
5. Scale $(\hat{\mathbf{A}}, \hat{\mathbf{B}}) = (\hat{\mathbf{A}}/c, c\hat{\mathbf{B}})$ such that $\hat{b}_{11} = 1$.

Optimizations (7) and (8) can be solved using the block coordinate descent algorithm in the same way as that developed for estimating the precision matrix in standard Gaussian graphical models [16]. We use the program glasso [16] in this paper

for these optimizations when the l_1 or the adaptive l_1 penalty functions are used. The glasso algorithm guarantees that the estimates $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ are positive definite.

Note that in Step 5 of the algorithm, we rescale the $\mathbf{A}$ and $\mathbf{B}$ matrices to ensure that $\hat{b}_{11} = 1$. However, when l_1 or the adaptive l_1 penalty functions are used, the solution to (6) is always unique in the sense that for a given λ and ρ, there is a unique scaling factor c^*,

$$c^* = \sqrt{\rho\|\mathbf{B}_0\|_1/\left(\lambda\|\mathbf{A}_0\|_1\right)}$$

in the equivalent class $\mathcal{C}_{\mathbf{A}_0,\mathbf{B}_0} := \left\{(\mathbf{A},\mathbf{B})|\mathbf{A} = c\mathbf{A}_0, \mathbf{B} = c^{-1}\mathbf{B}_0, \text{ for some } 0 < c < \infty\right\}$ that minimizes $\Phi(\mathbf{A},\mathbf{B})$, where $\|\mathbf{A}_0\|_1 = \sum_{i=1}^p \sum_{j=1}^p |\mathbf{A}_0(i,j)|$ is the matrix l_1 norm. This can be seen by

$$\lambda_1\|\mathbf{A}_0\|_1 c + \lambda_2\|\mathbf{B}_0\|_1\frac{1}{c} \geq 2\sqrt{\lambda_1\lambda_2\|\mathbf{A}_0\|_1\|\mathbf{B}_0\|_1},$$

based on the algebra–geometry inequality. Equality holds and hence the minimum is attained when $c^* = \sqrt{\lambda_2\|\mathbf{B}_0\|_1/\left(\lambda_1\|\mathbf{A}_0\|_1\right)}$. Hence $\mathbf{A} = c^*\mathbf{A}_0$, $\mathbf{B} = \mathbf{B}_0/c^*$ are uniquely determined.

Finally, the tuning parameters λ and ρ in the l_1 penalty functions are chosen using the cross-validated likelihood function.

4. Asymptotic theorems

Throughout this paper, for a given $p \times q$ matrix $\mathbf{A} = (a_{ij})$, we denote $\|\mathbf{A}\| = \max\{\|\mathbf{A}\mathbf{x}\|/\|\mathbf{x}\|, \mathbf{x} \in \Re^q, \mathbf{x} \neq 0\}$ as the operator or spectral norm of $\mathbf{A}$, $\|A\|_\infty = \max|a_{ij}|$ as the element-wise l_∞ norm of $\mathbf{A}$, and $|||\mathbf{A}|||_\infty = \max_{1\leq i\leq p}\sum_{j=1}^q |a_{ij}|$ as the matrix l_∞ norm of $\mathbf{A}$. Furthermore, we use $\|\mathbf{A}\|_F = \sqrt{\sum_{ij} a_{ij}^2}$ as the Frobenius norm of $\mathbf{A}$. Denote $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ the smallest and largest eigenvalues of the matrix $\mathbf{A}$.

4.1. Asymptotic theorems when p and q are fixed

We first consider the asymptotic distributions of the penalized maximum likelihood estimates in the setting when p and q are fixed as $n \to \infty$. The following theorem provides the asymptotic distribution of the estimate $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$.

Theorem 1. *For n independent identically distributed observations $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ from a matrix normal distribution $MN(0; \mathbf{A}^{-1}, \mathbf{B}^{-1})$, the optimizer $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ of the penalized negative log-likelihood function (6) with the l_1 penalty functions has the following property:*

If $n^{1/2}\lambda \to \lambda_0 \geq 0$, $n^{1/2}\rho \to \rho_0 \geq 0$, as $n \to \infty$, then

$$n^{1/2}\{(\hat{\mathbf{A}}, \hat{\mathbf{B}}) - (\mathbf{A},\mathbf{B})\} \to \operatorname{argmin}_{\mathbf{M}=\mathbf{M}^T,\mathbf{N}=\mathbf{N}^T} f(\mathbf{M},\mathbf{N})$$

in distribution, where

$$f(\mathbf{M},\mathbf{N}) = q\operatorname{tr}(\mathbf{MUMU}) + p\operatorname{tr}(\mathbf{NVNV}) + \operatorname{tr}(\mathbf{MU})\operatorname{tr}(\mathbf{NV}) + W + \lambda_0\sum_{i\neq j}\{m_{ij}\operatorname{sgn}(a_{ij})I(a_{ij}\neq 0) + |m_{ij}|I(a_{ij}=0)\}$$
$$+ \rho_0\sum_{i\neq j}\{n_{ij}\operatorname{sgn}(b_{ij})I(b_{ij}\neq 0) + |n_{ij}|I(b_{ij}=0)\},$$

in which W is a random variable such that $W \sim N(0,\sigma^2)$, where

$$\sigma^2 = 2\{q\operatorname{tr}(\mathbf{MUMU}) + p\operatorname{tr}(\mathbf{NVNV}) + 2\operatorname{tr}(\mathbf{MU})\operatorname{tr}(\mathbf{NV})\}.$$

This result parallels to that of Yuan and Lin [39] for the l_1 penalized likelihood estimate of the precision matrix in the standard Gaussian graphical model.

Suppose that we have c_n-consistent estimators of $\mathbf{A}$ and $\mathbf{B}$, denoted by $\tilde{\mathbf{A}} = (\tilde{a}_{ij})_{1\leq i,j\leq p}$ and $\tilde{\mathbf{B}} = (\tilde{b}_{ij})_{1\leq i,j\leq q}$, that is $c_n(\tilde{\mathbf{A}} - \mathbf{A}) = O_p(1)$, $c_n(\tilde{\mathbf{B}} - \mathbf{B}) = O_p(1)$, we consider the penalized likelihood estimates using the adaptive l_1 penalty function

$$\sum_{i\neq j} p_{\lambda_{ij}}(a_{ij}) = \lambda\sum_{i\neq j}|\tilde{a}_{ij}|^{-\gamma_1}|a_{ij}|, \sum_{i\neq j} p_{\rho_{ij}}(b_{ij}) = \rho\sum_{i\neq j}|\tilde{b}_{ij}|^{-\gamma_2}|b_{ij}|$$

in the objective function (6), where γ_1 and γ_2 are two constants. The following theorem shows that the resulting estimates of the precision matrices have the oracle property that parallels to that of Fan et al. [13] for the standard Gaussian graphical model.

Theorem 2. *For n independent identically distributed observations $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ from a matrix normal distribution $MN(0; \mathbf{A}^{-1}, \mathbf{B}^{-1})$, the optimizer $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ of the object function (6) with adaptive l_1 penalty functions has the oracle property in the sense of*

Fan and Li [14]. *That is, when* $n^{1/2}\lambda = O_p(1)$, $n^{1/2}\rho = O_p(1)$, $n^{1/2}\lambda c_n^{\gamma} \to \infty$ *and* $n^{1/2}\rho c_n^{\gamma} \to \infty$ *as* $n \to \infty$ *for some* $\gamma_1 > 0$ *and* $\gamma_2 > 0$, *then*

(1) *asymptotically, the estimates* $\hat{\mathbf{A}}$ *and* $\hat{\mathbf{B}}$ *have the same sparsity pattern as the true precision matrix* $\mathbf{A}$ *and* $\mathbf{B}$,
(2) *the non-zero entries of* $\hat{\mathbf{A}}$ *and* $\hat{\mathbf{B}}$ *are* c_n*-consistent and asymptotically normal.*

4.2. Asymptotic theorems when $p = p_n$ and $q = q_n$ diverge

The next two theorems provide the convergence rates and sparsistency properties of the estimates allowing $p = p_n$, $q = q_n$ to diverge as $n \to \infty$. We use $\mathbf{A}_0 = (a_{ij}^{(0)})$ and $\mathbf{B}_0 = (b_{kl}^{(0)})$ to denote the true precision matrices and $S_{\mathbf{A}} = \{(i,j) : a_{ij}^{(0)} \neq 0\}$ and $S_{\mathbf{B}} = \{(k,l) : b_{kl}^{(0)} \neq 0\}$ to denote the support of the true matrices, respectively. Let $s_{n1} = \mathrm{card}(S_{\mathbf{A}}) - p_n$ and $s_{n2} = \mathrm{card}(S_{\mathbf{B}}) - q_n$ be the number of nonzero elements in the off-diagonal entries of $\mathbf{A}_0$ and $\mathbf{B}_0$, respectively. We assume the following regularity conditions:

(A) There exist constants ε_1 and ε_2 such that

$$0 < \varepsilon_1 \le \lambda_{\min}(\mathbf{A}_0) \le \lambda_{\max}(\mathbf{A}_0) \le \varepsilon_2 < \infty, \quad \text{for all } n.$$

(B) There exist constants ε_3 and ε_4 such that

$$0 < \varepsilon_3 \le \lambda_{\min}(\mathbf{B}_0) \le \lambda_{\max}(\mathbf{B}_0) \le \varepsilon_4 < \infty, \quad \text{for all } n.$$

(C) The tuning parameter λ_n satisfies

$$\lambda_n = O\left\{\left(1 + \frac{\sqrt{p_n}}{\sqrt{s_{n1}} + 1}\right) q_n \sqrt{\frac{q_n(\log p_n + \log q_n)}{n}}\right\}.$$

(D) The tuning parameter ρ_n satisfies

$$\rho_n = O\left\{\left(1 + \frac{\sqrt{q_n}}{\sqrt{s_{n2}} + 1}\right) p_n \sqrt{\frac{p_n(\log p_n + \log q_n)}{n}}\right\}.$$

Conditions (A) and (B) bound uniformly the eigenvalues of $\mathbf{A}_0$ and $\mathbf{B}_0$, which facilitates the proof of consistency. These conditions are also assumed for the penalized likelihood estimation for the standard Gaussian graphical models [5,22]. The upper bounds on λ_n and ρ_n in condition (C) and (D) are related to the control of bias due to the l_1 penalty terms in the objective function [14,42,22].

Denote $\mathbf{S}_n = 1/n \sum_{k=1}^{n} \mathbf{Y}_k \otimes \mathbf{Y}_k$. It is easy to check that $\mathbf{\Sigma}_0 = (\mathrm{vec}(U))(\mathrm{vec}(V))^T = E\mathbf{S}_n$. We use the double indices (i,j) and (k,l) to refer to a row or a column in $\mathbf{S}_n$ or $\mathbf{\Sigma}_0$. The following lemma provides the tail probability bound of $(\mathbf{S}_n - \mathbf{\Sigma}_0)$.

Lemma 4.1. *Suppose the matrix observations* $\mathbf{Y}_k$*'s are i.i.d. from a matrix normal distribution,* $\mathbf{Y}_k \sim MN(0; \mathbf{U}, \mathbf{V})$, *and* $\lambda_{\max}(\mathbf{U}) \le \varepsilon_1^{-1}$, $\lambda_{\max}(\mathbf{V}) \le \varepsilon_3^{-1}$. *Then we have the tail bound:*

$$\Pr\Big(\max_{1\le i,j\le p_n,\ 1\le k,l\le q_n} |(\mathbf{S}_n - \mathbf{\Sigma}_0)_{(i,j)(k,l)}| \ge t\Big) \le C_1 p_n^2 q_n^2 \exp(-C_2 n t^2), \quad \text{for } |t| \le \delta, \tag{9}$$

for some constants C_1, C_2 *and* δ *that depend on* ε_1, ε_3 *only.*

In this lemma, if we choose $t = \sqrt{\log(p_n^2 q_n^2)/(nC_2)}M$ for some M such that $|t| \le \delta$, then

$$\|\mathbf{S}_n - \mathbf{\Sigma}_0\|_{\infty} = O_P\Big(\sqrt{(\log p_n + \log q_n)/n}\Big).$$

The next theorem provides the rates of convergence of the penalized likelihood estimates $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ in terms of the Frobenius norms.

Theorem 3 (*Rate of Convergence*)**.** *Under the regularity conditions* (A)–(D), *if* $q_n^3(\log p_n + \log q_n)/n = O(\lambda_n^2)$ *and* $q_n\big(p_n + s_{n1}\big)\big(\log p_n + \log q_n\big)^k/n = O(1)$ *for some* $k > 1$; $p_n^3(\log p_n + \log q_n)/n = O(\rho_n^2)$ *and* $p_n\big(q_n + s_{n2}\big)(\log p_n + \log q_n)^l/n = O(1)$ *for some* $l > 1$. *Then when the* l_1 *penalty functions are used, there exists a local minimizer* $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ *of* (6) *such that* $\|\hat{\mathbf{A}} - \mathbf{A}_0\|_F^2 = O_P\{q_n\big(p_n + s_{n1}\big)\big(\log p_n + \log q_n\big)/n\}$ *and* $\|\hat{\mathbf{B}} - \mathbf{B}_0\|_F^2 = O_P\{p_n\big(q_n + s_{n2}\big)\big(\log p_n + \log q_n\big)/n\}$.

Theorem 3 states explicitly how the number of nonzero elements and dimensionality of both precision matrices affect the rates of convergence of the estimates. Since there are $(q_n + s_{n2})(p_n + s_{n1})$ nonzero elements in the Kronecker product $\mathbf{A} \otimes \mathbf{B}$ and $\mathbf{B} \otimes \mathbf{A}$ and each of them can be estimated at best with rate $n^{-1/2}$, the total square errors are at least of rate $q_n(p_n + s_{n1})/n$ for estimating $\mathbf{A}$ and $p_n(q_n + s_{n2})$ for estimating $\mathbf{B}$. The price that we pay for high dimensionality is a logarithmic factor

$(\log p_n + \log q_n)$. The estimates $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ converge to their true values in Frobenius norm as long as $q_n(p_n + s_{n1})$ and $p_n(q_n + s_{n2})$ are at a rate $O((\log p_n + \log q_n)^{-l})$ for some $l > 1$, which decays to zero slowly. This means that in practice $p_n q_n$ can be comparable to n without violating the results. Compared to the rates of convergence of the l_1 penalized likelihood estimates of the precision matrix in the standard GGM [22], the convergence rates for $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ are increased by a factor q_n and p_n. If q_n (or p_n) is fixed as $n \to \infty$, then the rate for $\hat{\mathbf{A}}$ (or $\hat{\mathbf{B}}$) is exactly the same as that given in [22] for the standard Gaussian graphical models.

When an adaptive l_1 penalty function is used, we have the following sparsistency of the penalized estimates. Here sparsistency refers to the property that all parameters in $\mathbf{A}_0$ and $\mathbf{B}_0$ that are zero are actually estimated as zero with probability tending to one. We use S^c to denote the complement of a set S.

Theorem 4 (*Sparsistency*). *Under the conditions given in Theorem 3, when the penalty functions in (6) are adaptive l_1 penalty, $p_{\lambda_{ij}}(a_{ij}) = |a_{ij}|/|\tilde{a}_{ij}|^{\gamma_1}$, $p_{\rho_{kl}}(b_{kl}) = |b_{kl}|/|\tilde{b}_{kl}|^{\gamma_2}$ for some $\gamma_1 > 0, \gamma_2 > 0$, where $\tilde{\mathbf{A}} = (\tilde{a}_{ij})$ and $\tilde{\mathbf{B}} = (\tilde{b}_{kl})$ are any two e_n- and f_n-consistent estimator, i.e. $e_n\|\tilde{\mathbf{A}} - \mathbf{A}_0\|_\infty = O_P(1), f_n\|\tilde{\mathbf{B}} - \mathbf{B}_0\|_\infty = O_P(1)$. For any local minimizer $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ of (6) satisfying*

$$\|\hat{\mathbf{A}} - \mathbf{A}_0\|_F^2 = O_P\Big(q_n(p_n + s_{n1})(\log p_n + \log q_n)/n\Big),$$
$$\|\hat{\mathbf{B}} - \mathbf{B}_0\|_F^2 = O_P\Big(p_n(q_n + s_{n2})(\log p_n + \log q_n)/n\Big),$$

and $\|\hat{\mathbf{A}} - \mathbf{A}_0\|^2 = O_P(c_n)$, $\|\hat{\mathbf{B}} - \mathbf{B}_0\|^2 = O_P(d_n)$ for sequences $c_n \to 0$ and $d_n \to 0$, if

$$e_n^{-2\gamma_1} q_n \left(\frac{p_n(q_n + s_{n2})(\log p_n + \log q_n)}{n} + c_n q_n \right) = O(\lambda_n^2), \tag{10}$$

and

$$f_n^{-2\gamma_2} p_n \left(\frac{q_n(p_n + s_{n2})(\log p_n + \log q_n)}{n} + d_n p_n \right) = O(\rho_n^2), \tag{11}$$

then with probability tending to 1, $\hat{a}_{ij} = 0$ for all $(i, j) \in S_{\mathbf{A}}^c$ and $\hat{b}_{kl} = 0$ for all $(k, l) \in S_{\mathbf{B}}^c$.

The sparsistency results requires a lower bound on the rates of the regularization parameters λ_n and ρ_n. On the other hand, the regularity conditions (C) and (D) impose an upper bound on λ_n and ρ_n in order to control the estimation biases. These requirements on the tuning parameters are similar to those for the GGMs. However, in the case of the matrix normal estimation, the conditions for λ_n depend not only on the dimension p_n of $\mathbf{A}$, the rate of the consistent estimator $\tilde{\mathbf{A}}$ and the rate of error for $\hat{\mathbf{A}}$ in l_2 norm, but also depend on the dimension q_n and its sparsity s_{n2} of the matrix $\mathbf{B}_0$. Similarly, the conditions for ρ_n depend not only on the rate of $\tilde{\mathbf{B}}$ and rate of error for $\hat{\mathbf{B}}$ in l_2 norm, but also on the dimension and sparsity of $\mathbf{A}_0$. In addition, the condition (10) in the theorem, combined with the regularity condition (C), implies that

$$e_n^{-\gamma_1} \le \left(1 + \frac{\sqrt{p_n}}{\sqrt{s_{n1}} + 1}\right) \frac{q_n}{\sqrt{p_n(q_n + s_{n2})}} < \sqrt{q_n}\left(\frac{1}{\sqrt{p_n}} + \frac{1}{\sqrt{s_{n1}} + 1}\right),$$

and

$$e_n^{-\gamma_1}\sqrt{c_n} \le \sqrt{\frac{p_n q_n(\log p_n + \log q_n)}{n}}\left(\frac{1}{\sqrt{p_n}} + \frac{1}{\sqrt{s_{n1}} + 1}\right).$$

These are the requirements for both the rate of the consistent estimator $\tilde{\mathbf{A}}$ in its element-wise l_∞ norm and rate of the operator norm of $\hat{\mathbf{A}}$. Similarly, condition (11) and regularity condition (D) imply that

$$f_n^{-\gamma_2} < \sqrt{p_n}\left(\frac{1}{\sqrt{q_n}} + \frac{1}{\sqrt{s_{n2}} + 1}\right),$$

and

$$f_n^{-\gamma_2}\sqrt{d_n} \le \sqrt{\frac{p_n q_n(\log p_n + \log q_n)}{n}}\left(\frac{1}{\sqrt{q_n}} + \frac{1}{\sqrt{s_{n2}} + 1}\right).$$

5. Monte Carlo simulations

5.1. Comparison candidates and measurements

In this section we present results form Monte Carlo simulations to examine the performances of the penalized likelihood method and to compare them to several naive methods for estimating the two precision matrices. The first method uses only data from one row or column in order to ensure that the observations are independent. Specifically, to estimate the precision matrix $\mathbf{A}$, we choose the ith column from every observation matrix $\mathbf{Y}_k$ ($k = 1, \ldots, n$), denoted by $\mathbf{y}_{k\cdot i}$ the ith column of $\mathbf{Y}_k$, to estimate the row precision matrix $\mathbf{A}$. Since $\mathbf{y}_{k\cdot i} \sim N(\mathbf{M}_{\cdot i}, v_{ii}\mathbf{U})$, we can estimate the precision matrix $\mathbf{A}$ up to a multiplier by fitting a standard GGM. Without loss of generality, we choose the first column $\mathbf{y}_{k\cdot 1}$ in our simulations. We call this procedure the Gaussian graphical model using the column data (GGM-C). Similarly, we can estimate the precision matrix $\mathbf{B}$ by choosing the first row $\mathbf{y}_{k1\cdot}$ from every $\mathbf{Y}_k$ ($k = 1, \ldots, n$). We call this procedure the Gaussian graphical model using the row data (GGM-R). The second approach simply ignores the dependency of the data across the columns or rows and estimates $\mathbf{A}$ by treating the q columns as independent observations and estimates $\mathbf{B}$ by treating the p rows as independent observations using the Gaussian graphical model. We call this procedure the Gaussian graphical model assuming independence of row variables or column variables (GGM-I). For all three procedures (GGM-C, GGM-C and GGM-I), we use the glasso algorithm to estimate these two matrices. When $p, q < n$, we also consider the adaptive version of the glasso, where the maximum likelihood estimates are used as the initial consistent estimates of the precision matrices.

We compare the performance of different estimators of the precision matrices $\mathbf{A}$ and $\mathbf{B}$ by calculating different matrix norms of the estimation errors. Let $\mathbf{\Delta}_A = \mathbf{A} - \hat{\mathbf{A}}$ and $\mathbf{\Delta}_B = \mathbf{B} - \hat{\mathbf{B}}$ be the estimation errors of the estimators $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$, respectively. We compare $\|\mathbf{\Delta}_A\|_\infty$, $|||\mathbf{\Delta}_A|||_\infty$, $\|\mathbf{\Delta}_A\|$ and $\|\mathbf{\Delta}_A\|_F$ for $\hat{A}$, and $\|\mathbf{\Delta}_B\|_\infty$, $|||\mathbf{\Delta}_B|||_\infty$, $\|\mathbf{\Delta}_B\|$ and $\|\mathbf{\Delta}_B\|_F$ for $\hat{B}$.

In order to evaluate how well different procedures recover the graphical structures defined by the precision matrices, we define the non-zero entry in a sparse precision matrix as "positive" and define the specificity (SPE), sensitivity (SEN) and Matthews correlation coefficient (MCC) scores as following:

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

where TP, TN, FP and FN are the numbers of true positives, true negatives, false positives and false negatives.

5.2. Models and data generation

We generate sparse precision matrices $\mathbf{A}$ and $\mathbf{B}$ using a similar scheme as in [25,13]. To be specific, our generating procedure can be described as:

$$a_{ii} \equiv 1$$
$$a_{ij} \mid (\delta_{ij} = 0) \equiv 0$$
$$a_{ij} \mid (\delta_{ij} = 1) \sim \mathrm{Unif}([-1, -0.5] \cup [0.5, 1]),$$

where $i \neq j$ and δ_{ij} is a Bernoulli random variable with a success probability of p_+. Then the off-diagonal elements of each row a_{ij} ($j = 1, \ldots, p$ and $j \neq i$) are divided by $1.5 \sum |a_{i\cdot}|_1$ (off-diagonal elements only). $\mathbf{A}$ is then symmetrized and $\mathbf{U} = \mathbf{A}^{-1}$ is obtained. Note that the diagonal elements in $\mathbf{U}$ generated in this way are heterozygous. We further modify $\mathbf{A}$ by $\mathbf{WA}$ where $\mathbf{W}$ is a diagonal matrix. Since $\mathbf{A}$ generated as above is diagonal dominant, $\mathbf{W} = \mathrm{diag}(w_1, \ldots, w_p)$ is generated as the follows: first we choose the upper bound $w_{\max}$ for w_i's, here we use $w_{\max} = 1.2$. Then for each j, we generate a uniformly distributed random variable r in the interval $(\sum_{i,i\neq j} |a_{ij}|/|a_{jj}|, 1)$ and let $w_j = r w_{\max}$. Thus we can guarantee the diagonal dominance of the matrix $\mathbf{WA}$ and hence the positive definiteness. We further define $\mathbf{U} = (\mathbf{WA})^{-1}$. Matrices $\mathbf{B}$ and $\mathbf{V}$ are generated in a similar way.

After we generate the parameter $(\mathbf{A}, \mathbf{B})$ and $(\mathbf{U}, \mathbf{V})$, we generate the matrix normal data by first generating a pq-dimensional normal vectors $\mathbf{z}_k$ from $N(0, \mathbf{V} \otimes \mathbf{U})$ and then rearranging them into a matrix Y_k such that $\mathrm{vec}(\mathbf{y}_k) = \mathbf{z}_k$ for $k = 1, \ldots, n$.

In the following, let p_{A+} (or p_{B+}) be the probability that an off-diagonal element of matrix $\mathbf{A}$ (or $\mathbf{B}$) is non-zero, which measures the degree of the sparsity of the matrix. We consider five different models of different dimensions and different degrees of sparsity:

Model 1: $n = 100$, $p = 30$, $q = 30$, $p_{A+} = 1/10$ and $p_{B+} = 1/10$.
Model 2: $n = 100$, $p = 80$, $q = 80$, $p_{A+} = 1/20$ and $p_{B+} = 1/20$.
Model 3: $n = 100$, $p = 150$, $q = 150$, $p_{A+} = 1/40$ and $p_{B+} = 1/40$.
Model 4: $n = 100$, $p = 500$, $q = 500$, $p_{A+} = 1/200$ and $p_{B+} = 1/200$.
Model 5: $n = 20$, $p = 600$, $q = 600$, $p_{A+} = 1/200$ and $p_{B+} = 1/200$.

We use a 5-fold cross validation to tune the regularization parameters for Models 1–4 and 3-fold cross validation for Model 5 due to its small sample size. The simulations are repeated 50 times.

Table 1
Comparison of the performance for simulated data sets of different dimensions when l_1 penalty functions are used. MNGM: the matrix normal graphical model with l_1 penalties; GGM-I: Gaussian graphical model treating rows or columns as independent; GGM-R/GGM-C: Gaussian graphical model that uses only data from the first column or the first row; MLE: maximum likelihood estimates. For each measurement, mean and standard deviation are calculated over 50 replications.

Precision matrix	Measure	MNGM	GGM-I	GGM-C GGM-R	MLE
	Model 1, $n = 100$, $p = 30$, $q = 30$				
A	$\|\mathbf{\Delta}_A\|$	0.17(0.026)	0.27(0.015)	0.62(0.057)	0.25(0.037)
	$\|\mathbf{\Delta}_A\|_\infty$	0.35(0.042)	0.53(0.064)	1.23(0.133)	0.64(0.059)
	$\|\mathbf{\Delta}_A\|_\infty$	0.08(0.019)	0.15(0.013)	0.34(0.066)	0.09(0.021)
	$\|\mathbf{\Delta}_A\|_F$	0.43(0.051)	0.73(0.025)	1.73(0.130)	0.60(0.044)
	SPE_A	0.68(0.025)	0.32(0.156)	0.82(0.147)	
	SEN_A	1.00(0.000)	1.00(0.000)	0.36(0.298)	
	MCC_A	0.54(0.022)	0.28(0.108)	0.23(0.068)	
B	$\|\mathbf{\Delta}_B\|$	0.15(0.026)	0.25(0.013)	0.61(0.052)	0.22(0.027)
	$\|\mathbf{\Delta}_B\|_\infty$	0.32(0.044)	0.48(0.040)	1.28(0.100)	0.60(0.051)
	$\|\mathbf{\Delta}_B\|_\infty$	0.08(0.018)	0.14(0.013)	0.31(0.056)	0.07(0.016)
	$\|\mathbf{\Delta}_B\|_F$	0.38(0.049)	0.68(0.026)	1.64(0.099)	0.57(0.033)
	SPE_B	0.68(0.031)	0.40(0.060)	0.70(0.067)	
	SEN_B	0.99(0.009)	1.00(0.006)	0.63(0.142)	
	MCC_B	0.47(0.027)	0.29(0.037)	0.25(0.055)	
	Model 2, $n = 100$, $p = 80$, $q = 80$				
A	$\|\mathbf{\Delta}_A\|$	0.17(0.022)	0.31(0.020)	0.75(0.089)	0.25(0.021)
	$\|\mathbf{\Delta}_A\|_\infty$	0.37(0.030)	0.58(0.030)	1.26(0.089)	0.93(0.045)
	$\|\mathbf{\Delta}_A\|_\infty$	0.07(0.015)	0.19(0.015)	0.47(0.079)	0.06(0.014)
	$\|\mathbf{\Delta}_A\|_F$	0.67(0.082)	1.56(0.122)	3.30(0.515)	0.96(0.038)
	SPE_A	0.89(0.080)	0.69(0.009)	1.00(0.000)	
	SEN_A	1.00(0.000)	1.00(0.000)	0.00(0.000)	
	MCC_A	0.68(0.100)	0.43(0.008)	–	
B	$\|\mathbf{\Delta}_B\|$	0.14(0.013)	0.13(0.012)	0.72(0.147)	0.22(0.020)
	$\|\mathbf{\Delta}_B\|_\infty$	0.45(0.047)	0.42(0.049)	1.45(0.117)	0.88(0.040)
	$\|\mathbf{\Delta}_B\|_\infty$	0.06(0.008)	0.06(0.010)	0.53(0.180)	0.05(0.011)
	$\|\mathbf{\Delta}_B\|_F$	0.56(0.023)	0.54(0.025)	2.97(0.499)	0.91(0.023)
	SPE_B	0.86(0.102)	0.69(0.010)	1.00(0.000)	
	SEN_B	1.00(0.000)	1.00(0.000)	0.00(0.000)	
	MCC_B	0.64(0.124)	0.42(0.008)	0.06(0.000)	

5.3. Simulation results

We present in Tables 1 and 2 the results of the three different procedures in terms of estimating the precision matrix and recovering the corresponding graphical structures when the l_1 penalty functions are used. For all four models considered, we observe that the MNGM results in smaller estimation errors and better performances in identifying graphical structures defined by the precision matrices than the naive applications of the Gaussian graphical models. This is true both for the settings when $p, q < n$ (Models 1 and 2) and when $p, q > n$ (Models 3 and 4). We observe that when only one row or one column is chosen from each observation and the standard GGM is used (GGM-R or GGM-C), the estimation errors are much higher than the MNGM or the GGM when the rows or columns are treated as independent. Similarly, both sensitivities and specificities are also lower if only data from one row or one columns are used. This can be explained by the relatively small sample sizes. On the other hand, if the dependency of the columns or rows is ignored and the data of the columns or rows are treated as independent, direct application of the Gaussian graphical model (GGM-I) results in smaller specificities and higher false positives. As a benchmark comparison, for Models 1 and 2, we also present in Table 1 the errors of the MLEs of **A** and **B**. It is clear that the MNGM gives better estimates than the MLEs. MLEs for Models 3 and 4 do not exist.

When $p, q < n$, as in Models 1 and 2, we have also implemented the penalized likelihood estimation with adaptive l_1 loss functions and performed simulation comparisons with the standard l_1 loss functions, where the maximum likelihood estimates of **A** and **B** are obtained and used as weights in the adaptive l_1 penalty functions. We present the results in Table 3. Comparing with the results in Table 1, we observe that using the adaptive penalties in the MNGM and the GGM-I can lead to better estimates of the precision matrices and better recovery of the graphical structures defined by these precision matrices. However, if we only select one row or column and estimate the precision matrices using the GGM (GGM-R/GGM-C), the estimates based on the adaptive l_1 penalty functions are in general not as good as those based on the l_1 penalty functions. This is due to the fact that when only one row or one column is used, the sample size is small and the MLEs of the precision matrices may not provide sensible estimates of the weights in the adaptive penalty functions, which can lead to poor performance of the resulting estimates.

As expected, since the precision matrices **A** and **B** are generated similarly and both are of the same dimensions, the estimates of these two precision matrices based on the MNGM are very comparable for all four models considered. Some differences in performances for estimating **A** and **B** in Model 4 are observed when the GGM-I or GGM-R/GGM-C is used. This is largely due to the large variability in selecting the tuning parameters when the dependence of the data is ignored as in GGM-I or when only partial data are used as in GGM-R/GGM-C.

Table 2
Comparison of the performance for simulated data sets of different dimensions when l_1 penalty functions are used. MNGM: the matrix normal graphical model with l_1 penalties; GGM-I: Gaussian graphical model treating rows or columns as independent; GGM-R/GGM-C: Gaussian graphical model that uses only data from the first column or the first row. For each measurement, mean and standard deviation are calculated over 50 replications.

Precision matrix	Measure	MNGM	GGM-I	GGM-C GGM-R
	Model 3, $n = 100$, $p = 150$, $q = 150$			
A	$\|\boldsymbol{\Delta}_A\|$	0.12(0.014)	0.31(0.013)	0.78(0.094)
	$\|\boldsymbol{\Delta}_A\|_\infty$	0.32(0.028)	0.59(0.027)	1.45(0.092)
	$\|\boldsymbol{\Delta}_A\|_\infty$	0.05(0.011)	0.20(0.010)	0.49(0.074)
	$\|\boldsymbol{\Delta}_A\|_F$	0.61(0.069)	2.26(0.120)	4.72(0.802)
	SPE_A	0.84(0.005)	0.80(0.004)	1.00(0.000)
	SEN_A	1.00(0.000)	1.00(0.000)	0.00(0.000)
	MCC_A	0.45(0.006)	0.40(0.004)	0.05(0.022)
B	$\|\boldsymbol{\Delta}_B\|$	0.10(0.009)	0.10(0.009)	0.77(0.186)
	$\|\boldsymbol{\Delta}_B\|$	0.29(0.022)	0.32(0.025)	1.38(0.208)
	$\|\boldsymbol{\Delta}_B\|_\infty$	0.04(0.007)	0.04(0.007)	0.58(0.236)
	$\|\boldsymbol{\Delta}_B\|_F$	0.53(0.025)	0.56(0.024)	4.25(0.856)
	SPE_B	0.83(0.005)	0.80(0.003)	1.00(0.000)
	SEN_B	1.00(0.000)	1.00(0.000)	0.01(0.000)
	MCC_B	0.43(0.007)	0.40(0.004)	0.07(0.021)
	Model 4, $n = 100$, $p = 500$, $q = 500$			
A	$\|\boldsymbol{\Delta}_A\|$	0.10(0.008)	0.22(0.008)	3.69(0.521)
	$\|\boldsymbol{\Delta}_A\|_\infty$	0.27(0.018)	0.45(0.019)	4.23(0.502)
	$\|\boldsymbol{\Delta}_A\|_\infty$	0.04(0.007)	0.14(0.006)	3.63(0.581)
	$\|\boldsymbol{\Delta}_A\|_F$	0.95(0.078)	2.94(0.131)	43.68(6.153)
	SPE_A	0.99(0.001)	0.95(0.001)	1.00(0.002)
	SEN_A	1.00(0.00)	1.00(0.00)	0.01(0.038)
	MCC_A	0.76(0.008)	0.52(0.003)	0.13(0.030)
B	$\|\boldsymbol{\Delta}_B\|$	0.08(0.006)	0.08(0.006)	1.17(0.026)
	$\|\boldsymbol{\Delta}_B\|_\infty$	0.26(0.019)	0.26(0.019)	6.88(0.809)
	$\|\boldsymbol{\Delta}_B\|_\infty$	0.03(0.003)	0.03(0.004)	0.34(0.088)
	$\|\boldsymbol{\Delta}_B\|_F$	0.79(0.028)	0.76(0.031)	13.07(0.773)
	SPE_B	0.98(0.001)	0.97(0.001)	0.64(0.055)
	SEN_B	1.00(0.000)	1.00(0.000)	0.65(0.095)
	MCC_B	0.75(0.007)	0.62(0.003)	0.06(0.015)

Finally, Model 5 with $n = 20$, $p = q = 600$ mimics the scenario when $n \ll \min(p, q)$. The performances of the MNGM as shown in Table 4 are still quite comparable to the previous four models. However, estimates from the GGM-I or GGM-R/RRM-C are significantly worse, resulting much lower sensitivities and larger estimation errors.

6. Real data analysis

We applied the MNGM to an analysis of the mouse gene expression data measured over different tissues from the Atlas of Gene Expression in the Mouse Aging (AGEMAP) database [40]. In this study, the authors profiled the effects of aging on gene expressions in different mouse tissues dissected from C57BL/6 mice. Mice were of ages 1, 6, 16, and 24 months, with ten mice per age cohort and five mice of each sex. Sixteen tissues, the cerebellum, cerebrum, striatum, hippocampus, spinal cord, adrenal glands, heart, lung, liver, kidney, muscle, spleen, thymus, bone marrow, eye, and gonads, were dissected from each mouse. For each issue, mRNA was isolated and hybridized to two filter membranes containing a total of 16,896 cDNA clones corresponding to 8932 genes. In our analysis, we leave out the data from six tissues, including cerebellum, bone-marrows, heart, gonads, striatum and liver, from our analysis due to the fact that some mice did not have data on these tissues. Due to the small sample size $n = 40$, we consider a set of 40 genes that belong to the mouse vascular endothelial growth factor (VEGF) signaling pathway and have measured gene expression levels over all 10 tissues.

Fig. 1 shows the scatter plots of the pair-wise correlations of expression levels of these 40 genes in different tissues, clearly indicating that many gene pairs have similar correlations across different tissues and also the gene expression levels are clearly not independent across multiple tissues. The plots indicate that the assumption of the Kronecker covariance structure for the gene-tissue matrix normal data would be helpful in studying the covariance structure of the genes across different tissues.

Our goal is to study the dependency structure of these 40 genes of the VEGF pathway using the expression data across all 10 tissues. When the standard GGM is used to the data of each of the tissues separately, gene networks are identified from 5 out of 10 tissues, including adrenal, kidney, lung, thymus and eye. However, no gene links are identified for the other five tissues. The corresponding gene network graphs are shown in Fig. 2 for each of the five tissues. The networks identified based on the tissue-specific data only include a few VEGF genes, indicating lack of the power of the recovering biologically meaningful links based on data from single tissue. The differences of the identified networks from difference tissues might also be due to the fact that genes of the VEGF pathways are not perturbed enough in some tissues to make inferences on the conditional independence structures among the genes. On the other hand, if all the data are pooled together and the

Table 3
Comparison of the performance for simulated data sets of different dimensions when adaptive penalty functions are used. MNGM: the matrix normal graphical model with adaptive l_1 penalties; GGM-I: Gaussian graphical model treating rows or columns as independent; GGM-R/GGM-C: Gaussian graphical model that uses only data from the first column or the first row. For each measurement, mean and standard deviation are calculated over 50 replications.

Precision matrix	Measure	MNGM	GGM-I	GGM-C GGM-R
	Model 1, $n = 100$, $p = 30$, $q = 30$			
A	$\|\boldsymbol{\Delta}_A\|$	0.15(0.024)	0.26(0.014)	0.64(0.033)
	$\|\|\boldsymbol{\Delta}_A\|\|_\infty$	0.30(0.037)	0.51(0.037)	1.10(0.079)
	$\|\boldsymbol{\Delta}_A\|_\infty$	0.08(0.021)	0.14(0.010)	0.35(0.061)
	$\|\boldsymbol{\Delta}_A\|_F$	0.37(0.046)	0.69(0.025)	1.74(0.059)
	SPE_A	0.81(0.018)	0.41(0.028)	0.95(0.016)
	SEN_A	1.00(0.000)	1.00(0.000)	0.24(0.069)
	MCC_A	0.67(0.022)	0.34(0.018)	0.27(0.059)
B	$\|\boldsymbol{\Delta}_B\|$	0.14(0.024)	0.24(0.012)	0.66(0.040)
	$\|\|\boldsymbol{\Delta}_B\|\|_\infty$	0.28(0.047)	0.48(0.037)	1.14(0.091)
	$\|\boldsymbol{\Delta}_B\|_\infty$	0.08(0.019)	0.13(0.012)	0.35(0.043)
	$\|\boldsymbol{\Delta}_B\|_F$	0.35(0.052)	0.65(0.025)	1.70(0.078)
	SPE_B	0.81(0.023)	0.42(0.023)	0.94(0.012)
	SEN_B	0.99(0.011)	1.00(0.006)	0.32(0.055)
	MCC_B	0.60(0.029)	0.30(0.015)	0.32(0.056)
	Model 2, $n = 100$, $p = 80$, $q = 80$			
A	$\|\boldsymbol{\Delta}_A\|$	0.12(0.019)	0.28(0.020)	2.36(0.756)
	$\|\|\boldsymbol{\Delta}_A\|\|_\infty$	0.28(0.034)	0.49(0.034)	2.94(0.792)
	$\|\boldsymbol{\Delta}_A\|_\infty$	0.06(0.013)	0.18(0.015)	2.28(0.773)
	$\|\boldsymbol{\Delta}_A\|_F$	0.48(0.064)	1.44(0.123)	12.19(4.586)
	SPE_A	0.92(0.005)	0.85(0.005)	1.00(0.000)
	SEN_A	1.00(0.000)	1.00(0.000)	0.00(0.000)
	MCC_A	0.73(0.012)	0.59(0.008)	–
B	$\|\boldsymbol{\Delta}_B\|$	0.12(0.011)	0.12(0.012)	4.78(1.206)
	$\|\|\boldsymbol{\Delta}_B\|\|_\infty$	0.33(0.045)	0.34(0.049)	12.91(3.251)
	$\|\boldsymbol{\Delta}_B\|_\infty$	0.06(0.009)	0.06(0.011)	0.74(0.320)
	$\|\boldsymbol{\Delta}_B\|_F$	0.44(0.027)	0.47(0.030)	10.48(2.145)
	SPE_B	0.92(0.005)	0.85(0.008)	0.12(0.010)
	SEN_B	1.00(0.000)	1.00(0.000)	0.90(0.019)
	MCC_B	0.73(0.013)	0.59(0.012)	0.02(0.018)

Table 4
Comparison of the performance for simulated data sets when $n \ll \min(p, q)$ and l_1 penalty functions are used (Model 5). MNGM: the matrix normal graphical model with l_1 penalties; GGM-I: Gaussian graphical model treating rows or columns as independent; GGM-R/GGM-C: Gaussian graphical model that uses only data from the first column or the first row. For each measurement, mean and standard deviation are calculated over 50 replications.

Precision matrix	Measure	MNGM	GGM-I	GGM-C GGM-R
	Model 5, $n = 20$, $p = 600$, $q = 600$			
A	$\|\boldsymbol{\Delta}_A\|$	0.14(0.013)	0.85(0.012)	6.85(1.532)
	$\|\|\boldsymbol{\Delta}_A\|\|$	0.67(0.041)	1.52(0.015)	10.9(3.799)
	$\|\boldsymbol{\Delta}_A\|_\infty$	0.05(0.009)	0.43(0.01)	6.78(1.563)
	$\|\boldsymbol{\Delta}_A\|_F$	1.58(0.091)	10.17(0.188)	56.92(14.796)
	SPE_A	0.84(0.005)	1(0)	0.98(0.025)
	SEN_A	1(0)	0.03(0.001)	0.04(0.042)
	MCC_A	0.22(0.004)	0.18(0.004)	0.02(0.003)
B	$\|\boldsymbol{\Delta}_\mathbf{B}\|$	0.15(0.01)	0.54(0.015)	1.12(0.069)
	$\|\|\boldsymbol{\Delta}_B\|\|$	0.75(0.037)	1.32(0.024)	4.06(0.344)
	$\|\boldsymbol{\Delta}_B\|_\infty$	0.05(0.009)	0.23(0.005)	0.74(0.191)
	$\|\boldsymbol{\Delta}_B\|_F$	1.68(0.06)	6.84(0.023)	11.96(0.807)
	SPE_B	0.81(0.006)	1(0)	0.93(0.001)
	SEN_B	1(0)	0.03(0.001)	0.11(0.008)
	MCC_B	0.21(0.004)	0.16(0.004)	0.01(0.003)

dependency of gene expression across issues is ignored, the GGM results in a very dense network with 373 links, which is biologically difficult to interpret given that the biological networks are expected to be sparse.

Fig. 3 shows the gene and the tissue networks estimated by the proposed MNGM, including a gene network of 27 links among 22 VEGF genes and a tissue network with 15 edges among the 10 tissues. Compared to the networks estimated based on data from single tissue (see plots of Fig. 2), we observe that more links are identified among these genes and many links identified by the MNGM appear in one of the graphs identified based on the issue-specific data. The difference between the overall network identified by the MNGM and the tissue-specific networks can also be due to the dependence structures of the VEGF genes being different in different tissues. It is interesting to note that many links identified by the MNGM may reflect the underlying VEGF signaling pathway [20]. For example, the binding of VEGF to VEGFR-2 leads to dimerization of the receptor, followed by intracellular activation of the PLCgamma (Plcg). It is interesting that several forms

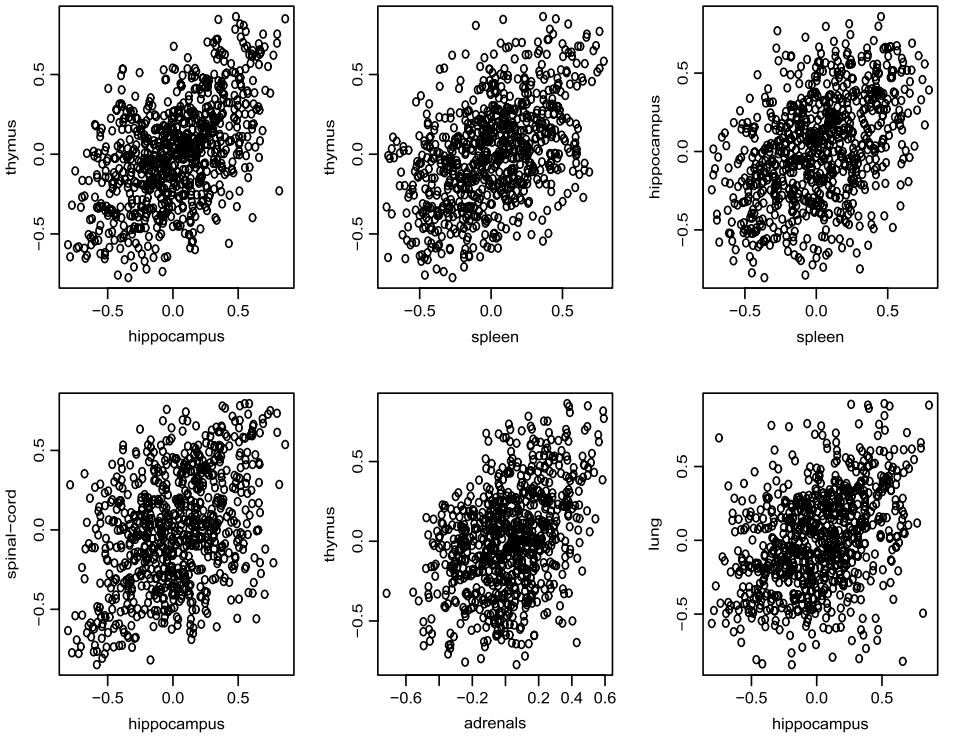
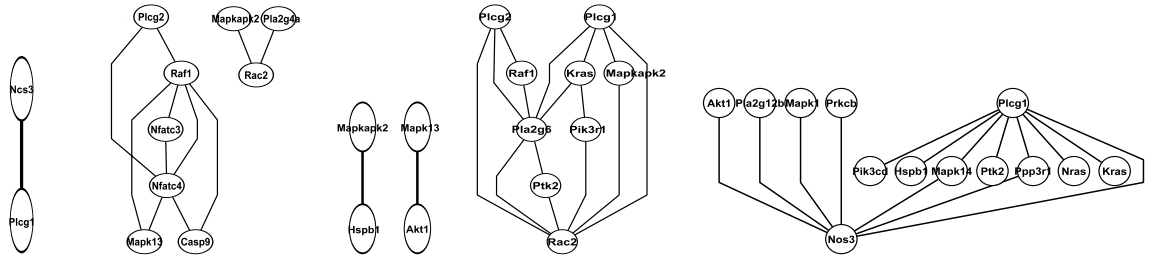


Fig. 1. Mouse gene expression data: scatter plots of pair-wise correlations of 40 genes across different tissues, showing that the expression levels in different issues are dependent.



(a) Adrenal. (b) Lung. (c) Kidney. (d) Thymus. (e) Eye.

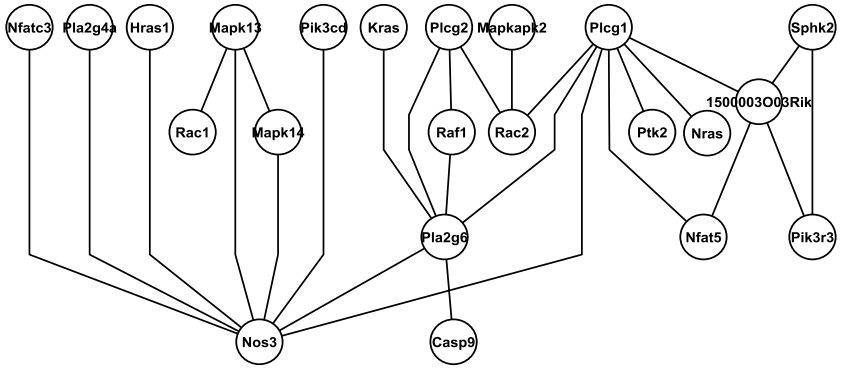
Fig. 2. Analysis of mouse gene expression data: networks identified by the GGM for each of the five tissues, including adrenal, lung, kidney, thymus and eye. The genes that belong to the mouse VEGF pathway are labeled on each of these network graphs. No networks are identified for the other five tissues, including hippocampus, cerebral-cortex, spinal-cord, spleen and skeletal-muscle.

of the PLgamma gene such as Plcg1, Plcg2 and their downstream genes Nfat5 and Pla2g6 are part of network. Several genes on the PKC-Raf kinase-MEK-mitogen-activated protein kinase (MAPK) pathway such as Mapk13, Mapk14 and Mapkapk2 are interconnected.

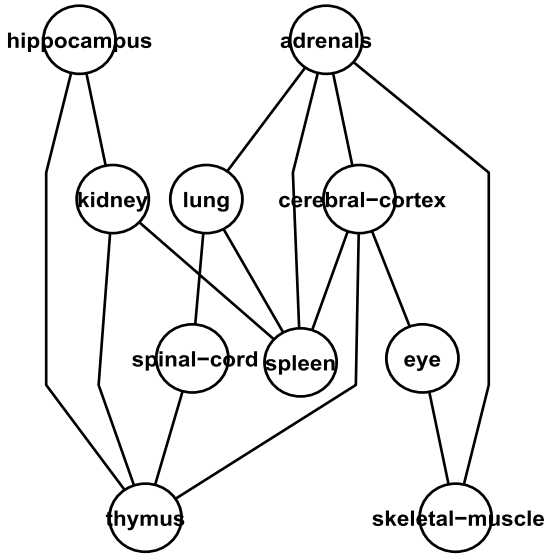
The tissue network as shown in Fig. 3(b) should be interpreted as the conditional dependency structure among the tissues with respect to the gene expression patterns observed among the genes on the VEGF pathway. It is interesting to observe links among lung, spleen and kidney in the vascular tissue group and links between eye and cerebral-cortex and between thymus and hippocampus in the neural tissue group. It is also interesting to observe that the adrenal tissue in the steroid responsive group is linked to both vascular and neural tissue groups. A similar clustering of tissue groups based on their gene expression data is also observed in [40].

7. Discussion

Motivated by analysis of gene expression data measured over different tissues on the same set of samples, we have proposed to apply the matrix normal distribution to model the data jointly and have developed a penalized likelihood method to estimate the row and column precision matrices assuming that both matrices are sparse. Our simulation results have clearly demonstrated such models can result in better estimates of the precision matrices and better identification of the corresponding graphical structures than naive application of the Gaussian graphical models. Our analysis of the



(a) Gene network.



(b) Tissue network.

Fig. 3. Analysis of mouse multi-tissue gene expression data using the MNGM: (a) gene network, where the genes that belong to the mouse VEGF pathway are labeled on the network graph; (b) tissue network based on gene expression data.

mouse gene expression data demonstrated that by effectively combining the expression data from multiple tissues from the same subjects, the matrix normal graphical model can lead to a conditional independence graph with meaningful biological interpretations. We also demonstrated that ignoring the dependency of gene expression across different tissues can lead to higher false positive links and dense graphs, which are difficult to interpret biologically.

The matrix normal distribution provides a natural way of modeling the dependency of data measured over different conditions. 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. Such estimated precision matrices can in turn be applied to the problem of co-expression analysis [32], differential expression analysis [2] and the problem of estimating missing gene expression data. Other applications of the proposed methods include face recognition [41].

The methods proposed in this paper and the related theorems can also be extended to array normal distribution by extending the matrix-variate normal to the tensor array setting using the Tucker product [35]. Such array normal distributions were recently studied by Hoff [19]. Allen [1] proposed an l_1 penalized estimation for such an array normal distribution by regularizing separable tensor precision matrices. Similar techniques can be applied to derive the estimation error bounds and to prove the sparsistency when the adaptive l_1 penalties are used. As multi-dimensional data with possible correlations among the variables of each dimension is becoming more prevalent, further development of estimation methods and relevant theorems are important.

Acknowledgments

This research was supported by NIH grants ES009911 and CA127334. We thank the reviewers for many helpful comments and for pointing out several omitted references.

Appendix

Proof of Proposition 2.1. Before we state the proof of Proposition 1, we need the following lemma [23]:

Lemma A.1. *Using the same notation as in the main text, if we partition the columns of* $\mathbf{Y}$ *as* $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2)$*, where* $\mathbf{Y}_1$ *is a* $p \times r$*,* $\mathbf{Y}_2$ *is* $p \times s$ *random matrix respectively, with* $r + s = q$*. Then the conditional distribution of* $\mathbf{Y}_1$ *given* $\mathbf{Y}_2 = y_2$ *is* $MN_{p\times r}(\mathbf{M}_1 + (y_2 - \mathbf{M}_2)\mathbf{V}_{22}^{-}\mathbf{V}_{21}; \mathbf{U}, \mathbf{V}_{1|2})$*, where* $\mathbf{M} = (\mathbf{M}_1, \mathbf{M}_2)$ *and* $\mathbf{V}_{1|2} = \mathbf{V}_{11} - \mathbf{V}_{12}\mathbf{V}_{22}^{-}\mathbf{V}_{21}$*.* □

Proof of Lemma A.1. See Proposition C.8 in [23, Appendix C]. □

Proof of Proposition 2.1. From Lemma A.1, we know $\mathbf{Y}_{\{\gamma,\mu\}}$ given $\mathbf{Y}_{\Gamma\setminus\{\gamma,\mu\}}$ is distributed as matrix normal $MN(\mathbf{M}_1 + (y_2 - \mathbf{M}_2)\mathbf{V}_{22}^{-}\mathbf{V}_{21}; \mathbf{U}, \mathbf{V}_{1|2})$. From Proposition C.5 of [23, Appendix C], we have

$$\mathbf{B}_{\{\gamma,\mu\}} = \begin{pmatrix} b_{\gamma\gamma} & b_{\gamma\mu} \\ b_{\mu\gamma} & b_{\mu\mu} \end{pmatrix} = \mathbf{V}_{1|2}^{-1}.$$

So

$$\mathbf{V}_{1|2} = \mathbf{V}_{\{\gamma,\mu\}|\Gamma\setminus\{\gamma,\mu\}} = \frac{1}{\det\mathbf{B}_{\{\gamma,\mu\}}} \begin{pmatrix} b_{\mu\mu} & -b_{\gamma\mu} \\ -b_{\mu\gamma} & b_{\gamma\gamma} \end{pmatrix}.$$

From Proposition C.6 of [23, Appendix C] we know $\mathbf{Y}_\gamma \perp\!\!\!\perp \mathbf{Y}_\mu \mid \mathbf{Y}_{\Gamma\setminus\{\gamma,\mu\}}$ if and only if $b_{\gamma\mu} = 0$. A similar argument can be applied to the rows. □

Proof of Theorem 1. Let $\mathbf{M} = \mathbf{M}^T$ be $p \times p$, $\mathbf{N} = \mathbf{N}^T$ be $q \times q$ symmetric random matrices. Denote

$$\begin{aligned} f_n(\mathbf{M}, \mathbf{N}) = {} & -q \log\left|\mathbf{A} + \frac{\mathbf{M}}{n^{1/2}}\right| - p \log\left|\mathbf{B} + \frac{\mathbf{N}}{n^{1/2}}\right| + \frac{1}{n}\sum_{k=1}^{n} \operatorname{tr}\left\{\left(\mathbf{A} + \frac{\mathbf{M}}{n^{1/2}}\right)\mathbf{Y}_k\left(\mathbf{B} + \frac{\mathbf{N}}{n^{1/2}}\right)\mathbf{Y}_k^T\right\} \\ & + \lambda \sum_{i\neq j}\left|a_{ij} + \frac{m_{ij}}{n^{1/2}}\right| + \rho \sum_{i\neq j}\left|b_{ij} + \frac{n_{ij}}{n^{1/2}}\right| + q \log|\mathbf{A}| + p \log|\mathbf{B}| \\ & - \frac{1}{n}\sum_{k=1}^{n} \operatorname{tr}\{\mathbf{A}\mathbf{Y}_k\mathbf{B}\mathbf{Y}_k^T\} - \lambda \sum_{i\neq j}|a_{ij}| - \rho \sum_{i\neq j}|b_{ij}|. \end{aligned}$$

Using the same argument as in [39], we have

$$\begin{aligned} \log\left|\mathbf{A} + \frac{\mathbf{M}}{n^{1/2}}\right| - \log|\mathbf{A}| &= \frac{\operatorname{tr}(\mathbf{MU})}{n^{1/2}} - \frac{\operatorname{tr}(\mathbf{MUMU})}{n} + o(n^{-1}), \\ \log\left|\mathbf{B} + \frac{N}{n^{1/2}}\right| - \log|\mathbf{B}| &= \frac{\operatorname{tr}(\mathbf{NV})}{n^{1/2}} - \frac{\operatorname{tr}(\mathbf{NVNV})}{n} + o(n^{-1}). \end{aligned}$$

Let $T_k = \operatorname{tr}\{(\mathbf{A} + n^{-1/2}\mathbf{M})\mathbf{Y}_k(\mathbf{B} + n^{-1/2}\mathbf{N})\mathbf{Y}_k^T\} - \operatorname{tr}\{\mathbf{A}\mathbf{Y}_k\mathbf{B}\mathbf{Y}_k^T\}$, then

$$\begin{aligned} T_k &= n^{-1/2}\operatorname{tr}(\mathbf{M}\mathbf{Y}_k\mathbf{B}\mathbf{Y}_k^T) + n^{-1/2}\operatorname{tr}(\mathbf{A}\mathbf{Y}_k\mathbf{N}\mathbf{Y}_k^T) + n^{-1}\operatorname{tr}(\mathbf{M}\mathbf{Y}_k\mathbf{N}\mathbf{Y}_k^T) \\ &= n^{-1/2}(\operatorname{vec}\mathbf{Y}_k)^T(\mathbf{B}\otimes\mathbf{M})\operatorname{vec}\mathbf{Y}_k + n^{-1/2}(\operatorname{vec}\mathbf{Y}_k)^T(\mathbf{N}\otimes\mathbf{A})\operatorname{vec}\mathbf{Y}_k + n^{-1}(\operatorname{vec}\mathbf{Y}_k)^T(\mathbf{N}\otimes\mathbf{M})\operatorname{vec}\mathbf{Y}_k. \end{aligned}$$

Denote $Z_k = \operatorname{vec}\mathbf{Y}_k$, then $Z_k \sim N(0, \mathbf{V}\otimes\mathbf{U})$, $T_k = n^{-1/2}Z_k^T(\mathbf{B}\otimes\mathbf{M} + \mathbf{N}\otimes\mathbf{A} + n^{-1/2}\mathbf{N}\otimes\mathbf{M})Z_k$. Next we compute $E(T_k)$ and $\operatorname{var}(T_k)$. First, $E(T_k) = n^{-1/2}\operatorname{tr}\{(\mathbf{B}\otimes M)(\mathbf{V}\otimes\mathbf{U}) + (\mathbf{N}\otimes\mathbf{A})(\mathbf{V}\otimes\mathbf{U}) + n^{-1/2}(\mathbf{N}\otimes\mathbf{M})(V\otimes\mathbf{U})\}$, so $E(T_k) = n^{-1/2}[q\operatorname{tr}(\mathbf{MU}) + p\operatorname{tr}(\mathbf{NV}) + n^{-1/2}\operatorname{tr}(\mathbf{NV})\operatorname{tr}(\mathbf{MU})]$. Next, $\operatorname{var}(T_k) = E(T_k^2) - \{E(T_k)\}^2$ and $E(T_k^2) = n^{-1}E[Z_k^T\mathbf{L}Z_kZ_k^T\mathbf{L}Z_k]$, where $\mathbf{L} = \mathbf{B}\otimes\mathbf{M} + \mathbf{N}\otimes\mathbf{A} + n^{-1/2}\mathbf{N}\otimes\mathbf{M}$. If $Z \sim N(0, \mathbf{\Sigma})$, then

$$E(Z^T\mathbf{A}ZZ^T\mathbf{B}Z) = \operatorname{tr}\{\mathbf{A}\mathbf{\Sigma}(\mathbf{B} + \mathbf{B}^T)\mathbf{\Sigma}\} + \operatorname{tr}(\mathbf{A}\mathbf{\Sigma})\operatorname{tr}(\mathbf{B}\mathbf{\Sigma}). \tag{A.1}$$

Using (A.1), we obtain

$$\begin{aligned} E(T_k^2) = {} & 2n^{-1}\{q\operatorname{tr}(\mathbf{MUMU}) + 2\operatorname{tr}(\mathbf{NV})\operatorname{tr}(\mathbf{MU}) + p\operatorname{tr}(\mathbf{NVNV}) + 2n^{-1/2}\operatorname{tr}(\mathbf{NV})\operatorname{tr}(\mathbf{MUMU}) \\ & + 2n^{-1/2}\operatorname{tr}(\mathbf{MU})\operatorname{tr}(\mathbf{NVNV}) + O(n^{-1})\} + \{E(T_k)\}^2, \end{aligned}$$

and

$$\begin{aligned} \operatorname{var}(T_k) = {} & 2n^{-1}\{q\operatorname{tr}(\mathbf{MUMU}) + p\operatorname{tr}(\mathbf{NVNV}) + 2\operatorname{tr}(\mathbf{MU})\operatorname{tr}(\mathbf{NV}) \\ & + 2n^{-1/2}\operatorname{tr}(\mathbf{NV})\operatorname{tr}(\mathbf{MUMU}) + 2n^{-1/2}\operatorname{tr}(\mathbf{MU})\operatorname{tr}(\mathbf{NVNV}) + O(n^{-1})\}. \end{aligned}$$

Let $\bar{T} = n^{-1}\sum_{k=1}^{n} T_k$, by the central limit theorem, $W_n = n(\bar{T} - ET_k) \to N(0, \sigma^2)$, where $\sigma^2 = 2[q\mathrm{tr}(\mathbf{MUMU}) + p\mathrm{tr}(\mathbf{NVNV}) + 2\mathrm{tr}(\mathbf{MU})\mathrm{tr}(\mathbf{NV})]$. Finally,

$$\begin{aligned} nf_n(\mathbf{M},\mathbf{N}) &= -nq\left[\frac{\mathrm{tr}(\mathbf{MU})}{n^{1/2}} - \frac{\mathrm{tr}(\mathbf{MUMU})}{n} + o(n^{-1})\right] - np\left[\frac{\mathrm{tr}(\mathbf{NV})}{n^{1/2}}\right. \\ &\quad \left. - \frac{\mathrm{tr}(\mathbf{NVNV})}{n} + o(n^{-1})\right] + n[\bar{T} - E(T_k)] + nE(T_k) \\ &\quad + n^{1/2}\lambda\sum_{i\neq j}\{m_{ij}\mathrm{sgn}(a_{ij})I(a_{ij}\neq 0) + |m_{ij}|I(a_{ij}=0)\} \\ &\quad + n^{1/2}\rho\sum_{i\neq j}\{n_{ij}\mathrm{sgn}(b_{ij})I(b_{ij}\neq 0) + |n_{ij}|I(b_{ij}=0)\} \\ &= p\mathrm{tr}(\mathbf{NVNV}) + q\mathrm{tr}(\mathbf{MUMU}) + \mathrm{tr}(\mathbf{MU})\mathrm{tr}(\mathbf{NV}) + W_n \\ &\quad + n^{1/2}\lambda\sum_{i\neq j}\{m_{ij}\mathrm{sgn}(a_{ij})I(a_{ij}\neq 0) + |m_{ij}|I(a_{ij}=0)\} \\ &\quad + n^{1/2}\rho\sum_{i\neq j}\{n_{ij}\mathrm{sgn}(b_{ij})I(b_{ij}\neq 0) + |n_{ij}|I(b_{ij}=0)\}, \end{aligned}$$

so $nf_n(M, N) \to f(\mathbf{M}, \mathbf{N})$. □

Proof of Theorem 2. We prove this theorem by verifying the regularity conditions (A), (B) and (C) of [14] [24, also]. We use $(\mathbf{A})_{ij}$ to denote the (i, j)th element of the matrix, a_{ij}. The log-likelihood function is

$$l = -npq/2\log(2\pi) + nq/2\log(|\mathbf{A}|) + np/2\log(|\mathbf{B}|) - 1/2\sum_{k=1}^{n}\mathrm{tr}(\mathbf{AY}_k\mathbf{BY}_k^T).$$

So

$$\frac{\partial l}{\partial a_{ij}} = \frac{nq}{2}\frac{1}{|\mathbf{A}|}|\mathbf{A}|u_{ij} - \frac{1}{2}\sum_{k=1}^{n}(\mathbf{Y}_k\mathbf{BY}_k^T)_{ij}.$$

On the other hand, $n^{-1}\sum_{k=1}^{n}\mathbf{Y}_k\mathbf{BY}_k^T$ is Wishart-distributed, so

$$E\left\{1/(nq)\sum_{k=1}^{n}(\mathbf{Y}_k\mathbf{BY}_k^T)_{ij}\right\} = u_{ij},$$

and $E(\partial l/\partial a_{ij}) = 0$. Similarly, one can verify $E(\partial l/\partial b_{ij}) = 0$, so the first part of condition (A) is verified. For the second part, we need to check

$$\begin{aligned} &E(\partial l/\partial a_{ij}\partial l/\partial a_{kl}) = E(-\partial^2 l/(\partial a_{ij}\partial a_{kl})), \\ &E(\partial l/\partial b_{ij}\partial l/\partial b_{kl}) = E(-\partial^2 l/(\partial b_{ij}\partial b_{kl})), \\ &E(\partial l/\partial b_{kl}\partial l/\partial a_{ij}) = E(-\partial^2 l/(\partial b_{kl}\partial a_{ij})). \end{aligned}$$

From the property of the Wishart distribution,

$$E\left(\frac{\partial l}{\partial a_{ij}}\frac{\partial l}{\partial a_{kl}}\right) = \frac{1}{4}E\left[\sum_{k=1}^{n}\{(\mathbf{Y}_k\mathbf{BY}_k^T)_{ij} - qu_{ij}\}\{(\mathbf{Y}_k\mathbf{BY}_k^T)_{kl} - qu_{kl}\}\right] = \frac{nq}{2}u_{ij}u_{kl}.$$

On the other hand, $d\mathbf{A}^{-1} = -\mathbf{A}^{-1}d\mathbf{A}\mathbf{A}^{-1}$, so

$$\frac{\partial^2 l}{\partial a_{ij}\partial a_{kl}} = \frac{\partial}{\partial a_{ij}}\left[\frac{nq}{2}\left\{u_{kl} - \frac{1}{nq}\sum_{m=1}^{n}(\mathbf{Y}_m\mathbf{BY}_m^T)_{ij}\right\}\right] = -\frac{nq}{2}\frac{\partial u_{kl}}{\partial a_{ij}} = -\frac{nq}{2}u_{ij}u_{kl}.$$

So $E(\partial l/\partial a_{ij}\partial l/\partial a_{kl}) = E(-\partial^2 l/(\partial a_{ij}\partial a_{kl}))$ holds. We can similarly verify that $E(\partial l/\partial b_{ij}\partial l/\partial b_{kl}) = E(-\partial^2 l/(\partial b_{ij}\partial b_{kl}))$. Denote the orthogonal bases $e_i = (0, \ldots, 1, \ldots, 0)$, which is a vector of all zeros except the ith element. We have

$$
\begin{aligned}
-\frac{\partial^2 l}{\partial b_{kl}\partial a_{ij}} &= \frac{1}{2}\sum_{s=1}^{n}\frac{\partial}{\partial b_{kl}}\{(\mathbf{Y}_s\mathbf{B}\mathbf{Y}_s^T)_{ij} - qu_{ij}\} = \frac{1}{2}\sum_{s=1}^{n}\frac{\partial}{\partial b_{kl}}\mathrm{tr}(\mathbf{B}\mathbf{Y}_s^T e_j e_i^T \mathbf{Y}_s)\\
&= \frac{1}{2}\sum_{s=1}^{n}(\mathbf{Y}_s^T e_i e_j^T \mathbf{Y}_s)_{kl} = \frac{1}{2}\sum_{s=1}^{n}\mathrm{tr}(\mathbf{Y}_s^T e_i e_j^T \mathbf{Y}_s e_l e_k^T)\\
&= \frac{1}{2}\sum_{s=1}^{n}(\mathrm{vec}\mathbf{Y}_s)^T(e_l e_k^T \otimes e_j e_i^T)\mathrm{vec}\mathbf{Y}_s,\\
E\left(-\frac{\partial^2 l}{\partial b_{kl}\partial a_{ij}}\right) &= \frac{1}{2}\sum_{s=1}^{n}\mathrm{tr}\{(e_l e_k^T \otimes e_j e_i^T)(\mathbf{V}\otimes\mathbf{U})\} = \frac{1}{2}\sum_{s=1}^{n}\mathrm{tr}(e_l e_k^T\mathbf{V})\mathrm{tr}(e_j e_i^T\mathbf{U})\\
&= \frac{n}{2}u_{ij}v_{kl}.
\end{aligned}
$$

Using the same notation as in the proof of Theorem 1, let $Z_k = \mathrm{vec}\mathbf{Y}_k$, then $Z_k \sim N(0, \mathbf{V}\otimes\mathbf{U})$. We then have

$$
\begin{aligned}
E\left(\frac{\partial l}{\partial b_{kl}}\frac{\partial l}{\partial a_{ij}}\right) &= \frac{1}{4}\sum_{s=1}^{n}E[\{(\mathbf{Y}_s^T\mathbf{A}\mathbf{Y}_s)_{kl} - pv_{kl}\}\{(\mathbf{Y}_s\mathbf{B}\mathbf{Y}_s^T)_{ij} - qu_{ij}\}]\\
&= \frac{1}{4}\sum_{s=1}^{n}E\{(e_i^T\mathbf{Y}_s\mathbf{B}\mathbf{Y}_s^T e_j)(e_k^T\mathbf{Y}_s^T\mathbf{A}\mathbf{Y}_s e_l) - pqu_{ij}v_{kl}\},
\end{aligned}
$$

and $e_i^T\mathbf{Y}_s\mathbf{B}\mathbf{Y}_s^T e_j = \mathrm{tr}(e_j e_i^T\mathbf{Y}_s\mathbf{B}\mathbf{Y}_s^T) = (\mathrm{vec}\mathbf{Y}_s)^T\{\mathbf{B}\otimes(e_j e_i^T)\}\mathrm{vec}\mathbf{Y}_s = Z_s^T\{\mathbf{B}\otimes(e_j e_i^T)\}Z_s$, and $e_k^T\mathbf{Y}_s^T\mathbf{A}\mathbf{Y}_s e_l = Z_s^T\{(e_k e_l^T)\otimes\mathbf{A}\}Z_s$. Denote $\mathbf{K} = \mathbf{B}\otimes(e_j e_i^T)$, $\mathbf{L} = (e_k e_l^T)\otimes\mathbf{A}$, so $E\{(e_i^T\mathbf{Y}_s\mathbf{B}\mathbf{Y}_s^T e_j)(e_k^T\mathbf{Y}_s^T\mathbf{A}\mathbf{Y}_s e_l)\} = E(Z_s^T\mathbf{K}Z_s Z_s^T\mathbf{L}Z_s)$. Using (A.1), we have

$$
\begin{aligned}
E(\mathbf{Y}_s^T\mathbf{K}Z_s Z_s^T\mathbf{L}Z_s) &= \mathrm{tr}[\{\mathbf{B}\otimes(e_j e_i^T)\}(\mathbf{V}\otimes\mathbf{U})\{(e_k e_l^T)\otimes\mathbf{A} + (e_l e_k^T)\otimes\mathbf{A}\}(\mathbf{V}\otimes\mathbf{U})]\\
&\quad + \mathrm{tr}\{I_q\otimes(e_j e_i^T\mathbf{U})\}\mathrm{tr}\{(e_k e_l^T\mathbf{V})\otimes I_p\}\\
&= v_{lk}u_{ij} + v_{kl}u_{ij} + pqu_{ij}v_{kl} = 2v_{kl}u_{ij} + pqu_{ij}v_{kl}.
\end{aligned}
$$

So $E(\partial l/\partial b_{kl}\partial l/\partial a_{ij}) = n/2u_{ij}v_{kl}$, and $E\{-\partial^2 l/(\partial b_{kl}\partial a_{ij})\} = E(\partial l/\partial b_{kl}\partial l/\partial a_{ij})$. The condition (A) is verified.

Next, we verify condition (B). We have

$$
\begin{aligned}
-2dl &= -nq\mathrm{tr}(\mathbf{U}d\mathbf{A}) - np\mathrm{tr}(\mathbf{V}d\mathbf{B}) + \sum_{k=1}^{n}\{\mathrm{tr}(\mathbf{Y}_k\mathbf{B}\mathbf{Y}_k^T d\mathbf{A}) + \mathrm{tr}(\mathbf{Y}_k^T\mathbf{A}\mathbf{Y}_k d\mathbf{B})\},\\
-2d^2 l &= nq\mathrm{tr}(\mathbf{U}d\mathbf{A}\mathbf{U}d\mathbf{A}) + np\mathrm{tr}(\mathbf{V}d\mathbf{B}\mathbf{V}d\mathbf{B}) + 2\sum_{k=1}^{n}\mathrm{tr}(\mathbf{Y}_k d\mathbf{B}\mathbf{Y}_k^T d\mathbf{A})\\
&= [(d\mathrm{vec}\mathbf{A})^T, (d\mathrm{vec}\mathbf{B})^T]\begin{pmatrix} nq(\mathbf{U}\otimes\mathbf{U}) & \sum_{k=1}^{n}\mathbf{Y}_k\otimes\mathbf{Y}_k\\ \sum_{k=1}^{n}\mathbf{Y}_k^T\otimes\mathbf{Y}_k^T & np(\mathbf{V}\otimes\mathbf{V})\end{pmatrix}\begin{pmatrix} d\mathrm{vec}\mathbf{A}\\ d\mathrm{vec}\mathbf{B}\end{pmatrix}.
\end{aligned}
$$

One can verify that $E(\mathbf{Y}_{ij}^{(k)}\mathbf{Y}_{pq}^{(k)}) = u_{ip}v_{jq}$, where $\mathbf{Y}_{ij}^{(k)}$ is the (i,j)th entry of $\mathbf{Y}_k$'s. So $E(\mathbf{Y}_k\otimes\mathbf{Y}_k) = (\mathrm{vec}\mathbf{U})(\mathrm{vec}\mathbf{V})^T$. Denote $I(\mathbf{A},\mathbf{B})$ as the Fisher information matrix, then

$$
2I(\mathbf{A},\mathbf{B}) = \begin{pmatrix} nq(\mathbf{U}\otimes\mathbf{U}) & n(\mathrm{vec}\mathbf{U})(\mathrm{vec}\mathbf{V})^T\\ n(\mathrm{vec}\mathbf{V})(\mathrm{vec}\mathbf{U})^T & np(\mathbf{V}\otimes\mathbf{V})\end{pmatrix}. \tag{A.2}
$$

To see $I(\mathbf{A},\mathbf{B})$ is non-negative definite, one only needs to check $np(\mathbf{V}\otimes\mathbf{V}) - n/q(\mathrm{vec}\mathbf{V})(\mathrm{vec}\mathbf{U})^T(\mathbf{A}\otimes\mathbf{A})(\mathrm{vec}\mathbf{U})(\mathrm{vec}\mathbf{V})^T = np\{\mathbf{V}\otimes\mathbf{V} - 1/q(\mathrm{vec}\mathbf{V})(\mathrm{vec}\mathbf{V})^T\}$ is so. This is equivalent to checking that for any vector $t \neq 0$ in R^{q^2},

$$
t^T\{\mathbf{V}\otimes\mathbf{V} - 1/q(\mathrm{vec}\mathbf{V})(\mathrm{vec}\mathbf{V})^T\}t \geq 0.
$$

Denote $q\times q$ matrix $\mathbf{D}$ such that $\mathrm{vec}\mathbf{D} = t$, then

$$
t^T\{V\otimes V - 1/q(\mathrm{vec}V)(\mathrm{vec}V)^T\}t = \mathrm{tr}(VD^TVD) - \frac{1}{q}\{\mathrm{tr}(VD)\}^2. \tag{A.3}
$$

Since $\mathbf{V}$ is non-negative definite, $\mathbf{V}^{1/2}$ is well defined, denote $\mathbf{V}^{1/2}\mathbf{D}\mathbf{V}^{1/2} = \mathbf{A}$, then $\mathbf{A}^T = \mathbf{V}^{1/2}\mathbf{D}^T\mathbf{V}^{1/2}$ and (A.3) $= \mathrm{tr}(\mathbf{A}^T\mathbf{A}) - 1/q[\mathrm{tr}(\mathbf{A})]^2$. But in general, one has the inequality $[\mathrm{tr}(\mathbf{A}^T\mathbf{B})]^2 \leq \mathrm{tr}(\mathbf{A}^T\mathbf{A})\mathrm{tr}(\mathbf{B}^T\mathbf{B})$, so $[\mathrm{tr}(\mathbf{A})]^2 = [\mathrm{tr}(\mathbf{A}^T I_q)]^2 \leq \mathrm{tr}(\mathbf{A}^T\mathbf{A})\mathrm{tr}(I_q^2) = q\mathrm{tr}(\mathbf{A}^T\mathbf{A})$, thus we proved the condition (B).

Since the third derivative of the log-likelihood function doesn't involve any random variable, condition (C) is easy to satisfy. Theorem 2 thus holds. □

Proof of Lemma 4.1. We use the notation $\mathbf{Y}^{(k)}$ to refer to $\mathbf{Y}_k$ for convenience. We then have

$$(\mathbf{S}_n - \mathbf{\Sigma}_0)_{(i,j)(k,l)} = 1/n \sum_{s=1}^{n} (\mathbf{Y}_{ik}^{(s)} \mathbf{Y}_{jl}^{(s)} - u_{ij} v_{kl}).$$

Since $\mathrm{vec}(\mathbf{Y}^{(s)})$ is normally distributed, Lemma A.3 of [5] leads to the fact that there exist some constants δ, C_1 and C_2 depending on ε_1 and ε_3 only such that

$$\Pr\left(\left| \sum_{s=1}^{n} (\mathbf{Y}_{ik}^{(s)} \mathbf{Y}_{jl}^{(s)} - u_{ij} u_{kl}) \right| \geq nt \right) \leq C_1 \exp\{-C_2 n t^2\}, \quad \text{for } |t| \leq \delta.$$

Hence we have

$$\Pr\Big(\max_{\substack{(i,j) \\ (k,l)}} |(\mathbf{S}_n - \mathbf{\Sigma}_0)_{(i,j)(k,l)}| \geq t \Big) \leq C_1 p_n^2 q_n^2 \exp\{-C_2 n t^2\}, \quad \text{for } |t| \leq \delta,$$

which proves the lemma. □

Proof of Theorem 3. Let $\mathbf{W}_1$ be a symmetric matrix of dimension p_n and $\mathbf{W}_2$ be a symmetric matrix of dimension q_n. Let $\mathbf{D}_{\mathbf{W}_1}, \mathbf{D}_{\mathbf{W}_2}$ be their diagonal matrices, and $\mathbf{R}_{\mathbf{W}_1} = \mathbf{W}_1 - \mathbf{D}_{\mathbf{W}_1}$, $\mathbf{R}_{\mathbf{W}_2} = \mathbf{W}_2 - \mathbf{D}_{\mathbf{W}_2}$ be their off-diagonal matrices, respectively. Set $\mathbf{\Delta}_{\mathbf{W}_1} = \alpha_n \mathbf{R}_{\mathbf{W}_1} + \beta_n \mathbf{D}_{\mathbf{W}_1}$ and $\mathbf{\Delta}_{\mathbf{W}_2} = \delta_n \mathbf{R}_{\mathbf{W}_2} + \beta_n \mathbf{D}_{\mathbf{W}_2}$. We show that, for $\alpha_n = \sqrt{q_n s_{n1} (\log p_n + \log q_n)/n}$, $\beta_n = \sqrt{q_n p_n (\log p_n + \log q_n)/n}$ and $\delta_n = \sqrt{p_n s_{n2} (\log p_n + \log q_n)/n}$, for a set $\mathcal{A}$ defined as

$$\mathcal{A} = \Big\{ (\mathbf{W}_1, \mathbf{W}_2) : \|\mathbf{\Delta}_{\mathbf{W}_1}\|_F^2 = C_1^2 \alpha_n^2 + C_2^2 \beta_n^2 \text{ and } \|\mathbf{\Delta}_{\mathbf{W}_2}\|_F^2 = C_3^2 \delta_n^2 + C_4^2 \beta_n^2 \Big\},$$
$$\Pr\Big(\inf_{(\mathbf{W}_1, \mathbf{W}_2) \in \mathcal{A}} \phi(\mathbf{A}_0 + \mathbf{\Delta}_{\mathbf{W}_1}, \mathbf{B}_0 + \mathbf{\Delta}_{\mathbf{W}_2}) > \phi(\mathbf{A}_0, \mathbf{B}_0) \Big) \to 1,$$

for sufficiently large constants C_1, C_2, C_3 and C_4. Denote $\mathbf{A}_1 = \mathbf{A}_0 + \mathbf{\Delta}_{\mathbf{W}_1} = (a_{ij}^{(1)})$, $\mathbf{B}_1 = \mathbf{B}_0 + \mathbf{\Delta}_{\mathbf{W}_2} = (b_{ij}^{(1)})$, then

$$\begin{aligned} \phi(\mathbf{A}_1, \mathbf{B}_1) - \phi(\mathbf{A}_0, \mathbf{B}_0) = & -q_n \big[\log(|\mathbf{A}_1|) - \log(|\mathbf{A}_0|) \big] - p_n \big[\log(|\mathbf{B}_1|) - \log(|\mathbf{B}_0|) \big] \\ & + \frac{1}{n} \sum_{s=1}^{n} \mathrm{tr}\big\{ \mathbf{A}_1 \mathbf{Y}^{(s)} \mathbf{B}_1 \mathbf{Y}^{(s)T} \big\} - \frac{1}{n} \sum_{s=1}^{n} \mathrm{tr}\big\{ \mathbf{A}_0 \mathbf{Y}^{(s)} \mathbf{B}_0 \mathbf{Y}^{(s)T} \big\} \\ & + \lambda_n \sum_{i \neq j} \big\{ |a_{ij}^{(1)}| - |a_{ij}^{(0)}| \big\} + \rho_n \sum_{k \neq l} \big\{ |b_{ij}^{(1)}| - |b_{ij}^{(0)}| \big\}. \end{aligned}$$

So $\phi(\mathbf{A}_1, \mathbf{B}_1) - \phi(\mathbf{A}_0, \mathbf{B}_0) = I_1 + I_2 + I_3 + I_4 + I_5$, where

$$\begin{aligned} I_1 = & \frac{1}{n} \sum_{s=1}^{n} \big[\mathrm{tr}\{\mathbf{A}_1 \mathbf{Y}^{(s)} \mathbf{B}_1 \mathbf{Y}^{(s)T}\} - \mathrm{tr}\{\mathbf{A}_0 \mathbf{Y}^{(s)} \mathbf{B}_0 \mathbf{Y}^{(s)T}\} \big] \\ & - q_n \big[\log(|\mathbf{A}_1|) - \log(|\mathbf{A}_0|) \big] - p_n \big[\log(|\mathbf{B}_1|) - \log(|\mathbf{B}_0|) \big], \\ I_2 = & \lambda_n \sum_{(i,j) \in S_A^c} \big\{ |a_{ij}^{(1)}| - |a_{ij}^{(0)}| \big\}, \\ I_3 = & \lambda_n \sum_{\substack{i \neq j \\ (i,j) \in S_A}} \big\{ |a_{ij}^{(1)}| - |a_{ij}^{(0)}| \big\}, \\ I_4 = & \rho_n \sum_{(k,l) \in S_B^c} \big\{ |b_{kl}^{(1)}| - |b_{kl}^{(0)}| \big\}, \\ I_5 = & \rho_n \sum_{\substack{k \neq l \\ (k,l) \in S_B}} \big\{ |b_{kl}^{(1)}| - |b_{kl}^{(0)}| \big\}. \end{aligned} \tag{A.4}$$

Denote $\mathbf{\Delta}_A = \mathbf{\Delta}_{\mathbf{W}_1}$, $\mathbf{\Delta}_B = \mathbf{\Delta}_{\mathbf{W}_2}$ and recall the definitions of $\mathbf{S}_n = 1/n \sum_{s=1}^{n} \mathbf{Y}^{(s)} \otimes \mathbf{Y}^{(s)}$ and $\mathbf{\Sigma}_0 = (\mathrm{vec}\mathbf{U}_0)(\mathrm{vec}\mathbf{V}_0)^T = E\mathbf{S}_n$. Using Taylor's expansion with integral residues, we have

$$\begin{aligned} I_1 = & -q_n \mathrm{tr}\{\mathbf{A}_0^{-1} \mathbf{\Delta}_A\} - p_n \mathrm{tr}\{\mathbf{B}_0^{-1} \mathbf{\Delta}_B\} + (\mathrm{vec}\mathbf{\Delta}_A)^T \big(\mathbf{S}_n - \mathbf{\Sigma}_0 + \mathbf{\Sigma}_0 \big) (\mathrm{vec}\mathbf{B}_0) + (\mathrm{vec}\mathbf{A}_0)^T \big(\mathbf{S}_n - \mathbf{\Sigma}_0 + \mathbf{\Sigma}_0 \big) (\mathrm{vec}\mathbf{\Delta}_B) \\ & + q_n \int_0^1 (1 - v) dv (\mathrm{vec}\mathbf{\Delta}_A)^T \big(\mathbf{A}_v^{-1} \otimes \mathbf{A}_v^{-1} \big) (\mathrm{vec}\mathbf{\Delta}_A) + p_n \int_0^1 (1 - v) dv (\mathrm{vec}\mathbf{\Delta}_B)^T \big(\mathbf{B}_v^{-1} \otimes \mathbf{B}_v^{-1} \big) (\mathrm{vec}\mathbf{\Delta}_B) \\ & + (\mathrm{vec}\mathbf{\Delta}_A)^T \big(\mathbf{S}_n - \mathbf{\Sigma}_0 + \mathbf{\Sigma}_0 \big) (\mathrm{vec}\mathbf{\Delta}_B), \end{aligned} \tag{A.5}$$

where $\mathbf{A}_v = \mathbf{A}_0 + v\boldsymbol{\Delta}_A$, $\mathbf{B}_v = \mathbf{B}_0 + v\boldsymbol{\Delta}_B$. One can easily check that

$$\begin{aligned}(\mathrm{vec}\boldsymbol{\Delta}_A)^T\boldsymbol{\Sigma}_0(\mathrm{vec}\mathbf{B}_0) &= (\mathrm{vec}\boldsymbol{\Delta}_A)^T(\mathrm{vec}\mathbf{U}_0)(\mathrm{vec}\mathbf{V}_0)^T(\mathrm{vec}\mathbf{B}_0) = \mathrm{tr}(\mathbf{U}_0^T\boldsymbol{\Delta}_A)\mathrm{tr}(\mathbf{V}_0^T\mathbf{B}_0)\\ &= q_n\mathrm{tr}(\mathbf{U}_0\boldsymbol{\Delta}_A),\end{aligned}$$

and similarly, $(\mathrm{vec}\mathbf{A}_0)^T\boldsymbol{\Sigma}_0(\mathrm{vec}\boldsymbol{\Delta}_B) = p_n\mathrm{tr}(\mathbf{V}_0\boldsymbol{\Delta}_B)$. Then I_1 can be further simplified as $I_1 = K_1 + K_2 + K_3 + K_4 + K_5 + K_6$, where

$$\begin{aligned}K_1 &= q_n\int_0^1(1-v)dv(\mathrm{vec}\boldsymbol{\Delta}_A)^T\left(\mathbf{A}_v^{-1}\otimes\mathbf{A}_v^{-1}\right)(\mathrm{vec}\boldsymbol{\Delta}_A),\\ K_2 &= p_n\int_0^1(1-v)dv(\mathrm{vec}\boldsymbol{\Delta}_B)^T\left(\mathbf{B}_v^{-1}\otimes\mathbf{B}_v^{-1}\right)(\mathrm{vec}\boldsymbol{\Delta}_B),\\ K_3 &= (\mathrm{vec}\boldsymbol{\Delta}_A)^T\left(\mathbf{S}_n-\boldsymbol{\Sigma}_0\right)(\mathrm{vec}\mathbf{B}_0),\\ K_4 &= (\mathrm{vec}\mathbf{A}_0)^T\left(\mathbf{S}_n-\boldsymbol{\Sigma}_0\right)(\mathrm{vec}\boldsymbol{\Delta}_B),\\ K_5 &= \mathrm{tr}(\mathbf{U}_0\boldsymbol{\Delta}_A)\mathrm{tr}(\mathbf{V}_0\boldsymbol{\Delta}_B),\\ K_6 &= (\mathrm{vec}\boldsymbol{\Delta}_A)^T\left(\mathbf{S}_n-\boldsymbol{\Sigma}_0\right)(\mathrm{vec}\boldsymbol{\Delta}_B).\end{aligned}$$

Note that

$$\begin{aligned}K_1 &\geq q_n\|\boldsymbol{\Delta}_A\|_F^2/2\min_{0\leq v\leq 1}\lambda_{\max}^{-2}(\mathbf{A}_v)\\ &\geq q_n\|\boldsymbol{\Delta}_A\|_F^2/2(\|\mathbf{A}_0\|+\|\boldsymbol{\Delta}_A\|)^{-2}\\ &\geq q_n/2\|\boldsymbol{\Delta}_A\|_F^2(\varepsilon_2+o(1))^{-2} = q_n/2(C_1^2\alpha_n^2+C_2^2\beta_n^2)(\varepsilon_2+o(1))^{-2},\end{aligned}\tag{A.6}$$

$$K_2 \geq p_n/2\|\boldsymbol{\Delta}_B\|_F^2(\varepsilon_4+o(1))^{-2} = p_n/2(C_3^2\delta_n^2+C_4^2\beta_n^2)(\varepsilon_4+o(1))^{-2}.\tag{A.7}$$

Since

$$\mathbf{A}_v^{-1} = (I_p - v\mathbf{U}_0\boldsymbol{\Delta}_A + v^2\mathbf{U}_0\boldsymbol{\Delta}_A\mathbf{U}_0\boldsymbol{\Delta}_A + \cdots)\mathbf{A}_0^{-1},$$

hence

$$\mathrm{tr}\left(\mathbf{A}_v^{-1}\boldsymbol{\Delta}_A^T\mathbf{A}_v^{-1}\boldsymbol{\Delta}_A\right) = \mathrm{tr}(\mathbf{A}_0^{-1}\boldsymbol{\Delta}_A^T\mathbf{A}_0^{-1}\boldsymbol{\Delta}_A)(1+o(1)),$$

then

$$K_1 \geq q_n/2\min_{0\leq v\leq 1}\mathrm{tr}\left(\mathbf{A}_v^{-1}\boldsymbol{\Delta}_A^T\mathbf{A}_v^{-1}\boldsymbol{\Delta}_A\right),$$

so

$$K_1 \geq q_n/2\mathrm{tr}\left(\mathbf{U}_0\boldsymbol{\Delta}_A^T\mathbf{U}_0\boldsymbol{\Delta}_A\right)(1+o(1)),\tag{A.8}$$

and similarly

$$K_2 \geq p_n/2\mathrm{tr}\left(\mathbf{V}_0\boldsymbol{\Delta}_B^T\mathbf{V}_0\boldsymbol{\Delta}_B\right)(1+o(1)).\tag{A.9}$$

Generally, for two squared $p\times p$ matrix $\mathbf{M}$ and $q\times q$ matrix $\mathbf{N}$, we have $\mathrm{tr}(\mathbf{M}) \leq \sqrt{p\mathrm{tr}(\mathbf{M}^T\mathbf{M})}$, $\mathrm{tr}(\mathbf{N}) \leq \sqrt{q\mathrm{tr}(\mathbf{N}^T\mathbf{N})}$ and then

$$|\mathrm{tr}(\mathbf{M})\mathrm{tr}(\mathbf{N})| \leq \sqrt{pq\mathrm{tr}(\mathbf{M}^T\mathbf{M})\mathrm{tr}(\mathbf{N}^T\mathbf{N})} \leq \frac{1}{2}q\mathrm{tr}(\mathbf{M}^T\mathbf{M}) + \frac{1}{2}p\mathrm{tr}(\mathbf{N}^T\mathbf{N}).$$

Let $\mathbf{M} = \mathbf{U}_0^{1/2}\boldsymbol{\Delta}_\mathbf{A}\mathbf{U}_0^{1/2}$, $\mathbf{N} = \mathbf{V}_0^{1/2}\boldsymbol{\Delta}_\mathbf{B}\mathbf{V}_0^{1/2}$, we have

$$\begin{aligned}|K_5| &= |\mathrm{tr}(\mathbf{U}_0\boldsymbol{\Delta}_\mathbf{A})\mathrm{tr}(\mathbf{V}_0\boldsymbol{\Delta}_\mathbf{B})| = |\mathrm{tr}(\mathbf{M})\mathrm{tr}(\mathbf{N})|\\ &\leq \frac{1}{2}q_n\mathrm{tr}(\mathbf{M}^T\mathbf{M}) + \frac{1}{2}p_n\mathrm{tr}(N^T\mathbf{N}) = \frac{1}{2}q_n\mathrm{tr}\left(\mathbf{U}_0\boldsymbol{\Delta}_\mathbf{A}^T\mathbf{U}_0\boldsymbol{\Delta}_\mathbf{A}\right) + \frac{1}{2}p_n\mathrm{tr}\left(\mathbf{V}_0\boldsymbol{\Delta}_\mathbf{B}^T\mathbf{V}_0\boldsymbol{\Delta}_\mathbf{B}\right).\end{aligned}$$

Combining with (A.8) and (A.9) we know that $|K_5|$ is dominated by $K_1 + K_2$ with a large probability.

Next we bound $|K_3|$ and $|K_4|$. We have

$$\begin{aligned}|K_3| &= |(\mathrm{vec}\boldsymbol{\Delta}_A)^T(\mathbf{S}_n-\boldsymbol{\Sigma}_0)(\mathrm{vec}\mathbf{B}_0)| \leq L_1 + L_2,\\ |K_4| &= |(\mathrm{vec}\mathbf{A}_0)^T(\mathbf{S}_n-\boldsymbol{\Sigma}_0)(\mathrm{vec}\boldsymbol{\Delta}_B)| \leq L_3 + L_4,\end{aligned}$$

where if we use double index to indicate a row or column in $\mathbf{S}_n$, $\mathbf{\Sigma}_0$ or a position in $\mathrm{vec}\mathbf{\Delta}_A$, $\mathrm{vec}\mathbf{B}_0$, $\mathrm{vec}\mathbf{A}_0$ and $\mathrm{vec}\mathbf{\Delta}_B$, we have

$$L_1 = \sum_{\substack{i\neq j,\ (i,j)\in S_A \\ (k,l)}} |(\mathbf{\Delta}_A)_{(i,j)}\big(\mathbf{S}_n - \mathbf{\Sigma}_0\big)_{(i,j)(k,l)}(\mathbf{B}_0)_{(k,l)}|,$$

$$L_2 = \sum_{\substack{(i,j)\in S_A^c \\ (k,l)}} |(\mathbf{\Delta}_A)_{(i,j)}\big(\mathbf{S}_n - \mathbf{\Sigma}_0\big)_{(i,j)(k,l)}(\mathbf{B}_0)_{(k,l)}|,$$

$$L_3 = \sum_{\substack{k\neq l,\ (k,l)\in S_B \\ (i,j)}} |(\mathbf{A}_0)_{(i,j)}\big(\mathbf{S}_n - \mathbf{\Sigma}_0\big)_{(i,j)(k,l)}(\mathbf{\Delta}_B)_{(k,l)}|,$$

and

$$L_4 = \sum_{\substack{(k,l)\in S_B^c \\ (i,j)}} |(\mathbf{A}_0)_{(i,j)}\big(\mathbf{S}_n - \mathbf{\Sigma}_0\big)_{(i,j)(k,l)}(\mathbf{\Delta}_B)_{(k,l)}|.$$

From Lemma 4.1 we know that

$$\max_{(i,j)(k,l)} (\mathbf{S}_n - \mathbf{\Sigma}_0)_{(i,j)(k,l)} = O_P\big(\sqrt{(\log p_n + \log q_n)/n}\big).$$

Then

$$\begin{aligned} L_1 &\le \sqrt{s_{n1}}\|\mathbf{\Delta}_A\|_F O_P\left(\sqrt{\frac{\log p_n + \log q_n}{n}}\right) q_n\sqrt{\mathrm{tr}(\mathbf{B}_0^T\mathbf{B}_0)} \\ &\le q_n\sqrt{q_n s_{n1}}\lambda_{\max}(\mathbf{B}_0)\|\mathbf{\Delta}_A\|_F O_P\left(\sqrt{\frac{\log p_n + \log q_n}{n}}\right) \\ &\le q_n O_P\big(\alpha_n\big)\|\mathbf{\Delta}_A\|_F \\ &\le q_n O_P\big(C_1\alpha_n^2 + C_2\beta_n^2\big). \end{aligned} \tag{A.10}$$

This together with (A.6) shows that L_1 is dominated by K_1 by choosing sufficiently large C_1 and C_2. Symmetrically,

$$\begin{aligned} L_3 &\le p_n\sqrt{\mathrm{tr}(\mathbf{A}_0^T\mathbf{A}_0)}O_P\left(\sqrt{\frac{\log p_n + \log q_n}{n}}\right)\sqrt{s_{n2}}\|\mathbf{\Delta}_B\|_F \\ &\le p_n\sqrt{p_n s_{n2}}\lambda_{\max}(\mathbf{A}_0)O_P\left(\sqrt{\frac{\log p_n + \log q_n}{n}}\right)\|\mathbf{\Delta}_B\|_F \\ &\le p_n O_P(C_3\delta_n^2 + C_4\beta_n^2). \end{aligned}$$

By choosing sufficiently large C_3 and C_4, this together with (A.7) shows L_3 can be dominated by K_2. Also

$$\begin{aligned} L_2 &\le \sum_{(i,j)\in S_A^c} |a_{ij}^{(1)}| O_P\left(\sqrt{\frac{\log p_n + \log q_n}{n}}\right)\sum_{k,l}|(\mathbf{B}_0)_{k,l}| \\ &\le \sum_{(i,j)\in S_A^c} |a_{ij}^{(1)}| O_P\left(\sqrt{\frac{\log p_n + \log q_n}{n}}\right) q_n\sqrt{q_n}\lambda_{\max}(\mathbf{B}_0) \\ &\le \sum_{(i,j)\in S_A^c} |a_{ij}^{(1)}| O_P\left(q_n\sqrt{\frac{q_n\big(\log p_n + \log q_n\big)}{n}}\right) \end{aligned} \tag{A.11}$$

from the condition of λ_n in theorem, and using the similar technique in [22], it can be shown that L_2 is dominated by I_2. Similarly, L_4 is dominated by I_4. Thus we proved $|K_3| + |K_4|$ can be dominated by $K_1 + K_2 + I_2 + I_4$. It is easy to show that $|K_6|$ is of smaller order of K_3 and K_4, hence is also dominated by $K_1 + K_2 + I_2 + I_4$. We next show that

$$\begin{aligned} |I_3| &= \lambda_n \sum_{\substack{i\neq j, \\ (i,j)\in S_A}} \big|\{|a_{ij}^{(1)}| - |a_{ij}^{(0)}|\}\big| \\ &\le \lambda_n \sum_{\substack{i\neq j, \\ (i,j)\in S_A}} |a_{ij}^{(1)} - a_{ij}^{(0)}| \le \lambda_n\sqrt{s_{n1}}\|\mathbf{\Delta}_A\|_F \end{aligned}$$

$$\leq \sqrt{s_{n1}}O\left(\left(1+\frac{\sqrt{p_n}}{\sqrt{s_{n1}}+1}\right)q_n\sqrt{\frac{q_n(\log p_n+\log q_n)}{n}}\right)O\big(C_1\alpha_n+C_2\beta_n\big)$$
$$= q_nO\big(C_1\alpha_n^2+C_2\beta_n^2\big), \tag{A.12}$$

where the middle term in (A.12) is from regularity condition (C), thus I_3 is dominated by K_1 if we choose sufficiently large constants C_1 and C_2. Similarly, we get $|I_5| \leq p_nO\big(C_3\delta_n^2+C_4\beta_n^2\big)$ and is dominated by K_2 when C_3 and C_4 are large. Hence the proof. □

Proof of Theorem 4. For $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$, a minimizer of (6), where $\hat{\mathbf{A}} = (a_{ij})$, $\hat{\mathbf{B}} = (b_{kl})$, the derivative of $\phi(\mathbf{A}, \mathbf{B})$ with respect to a_{ij} for $(i, j) \in S_A^c$ evaluated at $\hat{\mathbf{A}}$ is

$$\begin{aligned}\frac{\partial\phi(\hat{\mathbf{A}}, \hat{\mathbf{B}})}{\partial a_{ij}} &= q_nu_{ij}+\frac{1}{n}\sum_{s=1}^{n}\{\mathbf{Y}^{(s)}\hat{\mathbf{B}}\mathbf{Y}^{(s)T}\}_{ij}+\frac{\lambda_n}{|\tilde{a}_{ij}|^{\gamma_1}}\operatorname{sgn}(a_{ij})\\ &= -q_nu_{ij}^{(0)}-q_n(u_{ij}-u_{ij}^{(0)})+\frac{1}{n}\sum_{s=1}^{n}\{\mathbf{Y}^{(s)}\mathbf{B}_0\mathbf{Y}^{(s)T}\}_{ij}\\ &\quad+\frac{1}{n}\sum_{s=1}^{n}\{\mathbf{Y}^{(s)}(\hat{B}-\mathbf{B}_0)\mathbf{Y}^{(s)T}\}_{ij}+\frac{\lambda_n}{|\tilde{a}_{ij}|^{\gamma_1}}\operatorname{sgn}(a_{ij}),\end{aligned}$$

where $\hat{\mathbf{U}} = (u_{ij}) = \hat{\mathbf{A}}^{-1}$ and $\mathbf{U}_0 = (u_{ij}^{(0)})$ and $\mathbf{B}_0$ are the true parameters. If we can show that the sign of $\partial\phi(\hat{\mathbf{A}}, \hat{\mathbf{B}})/\partial a_{ij}$ depends on $\operatorname{sgn}(a_{ij})$ only with probability tending to 1, the optimum is then at 0, so that $a_{ij} = 0$ for $(i, j) \in S_A^c$ with probability tending to 1. Let

$$\begin{aligned}I_1 &= \frac{1}{n}\sum_{s=1}^{n}\{\mathbf{Y}^{(s)}\mathbf{B}_0\mathbf{Y}^{(s)T}\}_{ij}-q_nu_{ij}^{(0)},\\ I_2 &= -q_n(u_{ij}-u_{ij}^{(0)}),\\ I_3 &= \frac{1}{n}\sum_{s=1}^{n}\{\mathbf{Y}^{(s)}(\hat{\mathbf{B}}-\mathbf{B}_0)\mathbf{Y}^{(s)T}\}_{ij},\end{aligned} \tag{A.13}$$

we then have

$$\frac{\partial\phi(\hat{\mathbf{A}}, \hat{\mathbf{B}})}{\partial a_{ij}} = I_1+I_2+I_3+\frac{\lambda_n}{|\tilde{a}_{ij}|^{\gamma_1}}\operatorname{sgn}(a_{ij}).$$

Using the same argument as in [22],

$$\max_{ij}|u_{ij}-u_{ij}^{(0)}| \leq \|\hat{\mathbf{U}}-\mathbf{U}_0\| = \|\hat{\mathbf{U}}(\hat{\mathbf{A}}-\mathbf{A}_0)\mathbf{U}_0\| \leq \|\hat{\mathbf{U}}\|\,\|\hat{\mathbf{A}}-\mathbf{A}_0\|\,\|\mathbf{U}_0\| = O(\|\hat{\mathbf{A}}-\mathbf{A}_0\|),$$

and then $\max_{ij}|I_2| = O_P(q_n\sqrt{c_n})$.

Since $\mathbf{Y}^{(s)} \sim MN(0; \mathbf{U}_0, \mathbf{V}_0)$, then $\mathbf{Y}^{(s)T} \sim MN(0; \mathbf{V}_0, \mathbf{U}_0)$ and $\operatorname{vec}(\mathbf{Y}^{(s)T}) \sim N(0, \mathbf{U}_0 \otimes \mathbf{V}_0)$. Let $\mathbf{Y}^{(s)T} = (\mathbf{Y}_1^{(s)}, \ldots, \mathbf{Y}_{p_n}^{(s)})$, where $\mathbf{Y}_i^{(s)}$ is a $q_n \times 1$ vector, for $i = 1, \ldots, p_n$. We have

$$\{\mathbf{Y}^{(s)}(\hat{\mathbf{B}}-\mathbf{B}_0)\mathbf{Y}^{(s)T}\}_{ij} = \mathbf{Y}_i^{(s)T}(\hat{B}-\mathbf{B}_0)\mathbf{Y}_j^{(s)}, \tag{A.14}$$

$$\{\mathbf{Y}^{(s)}\mathbf{B}_0\mathbf{Y}^{(s)T}\}_{ij} = \mathbf{Y}_i^{(s)T}\mathbf{B}_0\mathbf{Y}_j^{(s)}, \tag{A.15}$$

and

$$\begin{pmatrix}\mathbf{Y}_i^{(s)}\\ \mathbf{Y}_j^{(s)}\end{pmatrix} \sim N\left(0, \begin{pmatrix}u_{ii}^*\mathbf{V}_0, & u_{ij}^*\mathbf{V}_0\\ u_{ij}^*\mathbf{V}_0, & u_{jj}^*\mathbf{V}_0\end{pmatrix}\right), \tag{A.16}$$

$$\begin{pmatrix}\mathbf{B}_0^{1/2}\mathbf{Y}_i^{(s)}\\ \mathbf{B}_0^{1/2}\mathbf{Y}_j^{(s)}\end{pmatrix} \sim N\left(0, \begin{pmatrix}u_{ii}^*I_{q_n}, & u_{ij}^*I_{q_n}\\ u_{ij}^*I_{q_n}, & u_{jj}^*I_{q_n}\end{pmatrix}\right). \tag{A.17}$$

I_1 can be simplified as $I_1 = 1/n\sum_{s=1}^{n}\mathbf{Y}_i^{(s)T}\mathbf{B}_0\mathbf{Y}_j^{(s)} - q_nu_{ij}^*$. We have the following proposition:

Proposition A.1. *Under the notations above, we have*

$$\max_{i,j\in\{1,\ldots,p_n\}}\left|\frac{1}{n}\sum_{s=1}^{n}\mathbf{Y}_i^{(s)T}\mathbf{B}_0\mathbf{Y}_j^{(s)}-q_nu_{ij}^*\right|=O_P\left(q_n\sqrt{\frac{\log p_n}{nq_n}}\right).$$

Proof of Proposition A.1. To save notation, we use q for q_n here or there. Denote $\mathbf{B}_0^{1/2}\mathbf{Y}_i^{(s)}=(z_1,\ldots,z_q)^T$ and $\mathbf{B}_0^{1/2}\mathbf{Y}_j^{(s)}=(w_1,\ldots,w_q)^T$. From (A.17) we have

$$\begin{aligned}&(z_k,w_k)\sim_{i.i.d.}N\left(0,\begin{pmatrix}u_{ii}^* & u_{ij}^*\\ u_{ij}^* & u_{jj}^*\end{pmatrix}\right),\\ &\mathbf{Y}_i^{(s)T}\mathbf{B}_0\mathbf{Y}_j^{(s)}=\left(\mathbf{B}_0^{1/2}\mathbf{Y}_i^{(s)}\right)^T\left(\mathbf{B}_0^{1/2}\mathbf{Y}_j^{(s)}\right)=z_1w_1+\cdots+z_qw_q.\end{aligned}\tag{A.18}$$

Note that (A.18) does not depends on the sample index s, and the sum in I_1 is equivalent to nq_n normal observations. By Lemma A.3 of [5], we have

$$\max_{i,j\in\{1,\ldots,p_n\}}\left|\frac{q_n}{nq_n}\sum_{s=1}^{n}\mathbf{Y}_i^{(s)T}\mathbf{B}_0\mathbf{Y}_j^{(s)}-q_nu_{ij}^*\right|=q_nO_P\left(\sqrt{\frac{\log p_n}{nq_n}}\right)=O_P\left(\sqrt{\frac{q_n\log p_n}{n}}\right).$$

Hence the Proposition A.1 is proved. □

From Proposition A.1, we know $\max_{i,j}|I_1|=O_P(\sqrt{q_n\log p_n/n})$. Next we bound $|I_3|$. From (A.14) we know that $I_3=1/n\sum_{s=1}^n\mathbf{Y}_i^{(s)T}(\hat{\mathbf{B}}-\mathbf{B}_0)\mathbf{Y}_j^{(s)}$, and since

$$\begin{aligned}\mathbf{Y}_i^{(s)T}(\hat{\mathbf{B}}-\mathbf{B}_0)\mathbf{Y}_j^{(s)}&=\mathrm{tr}\big((\hat{B}-\mathbf{B}_0)\mathbf{Y}_j^{(s)}\mathbf{Y}_i^{(s)T}\big)\\&=\mathrm{tr}\Big[(\hat{B}-\mathbf{B}_0)\big(\mathbf{Y}_j^{(s)}\mathbf{Y}_i^{(s)T}-u_{ij}^*\mathbf{V}_0+u_{ij}^*\mathbf{V}_0\big)\Big],\end{aligned}$$

we have

$$I_3=\mathrm{tr}\big[u_{ij}^*\mathbf{V}_0(\hat{\mathbf{B}}-\mathbf{B}_0)\big]+\mathrm{tr}\Big[(\hat{B}-\mathbf{B}_0)\frac{1}{n}\sum_{s=1}^{n}\big(\mathbf{Y}_j^{(s)}\mathbf{Y}_i^{(s)T}-u_{ij}^*\mathbf{V}_0\big)\Big].$$

Then we have $|I_3|\le L_1+L_2$, where

$$\begin{aligned}L_1&=|\mathrm{tr}\big(u_{ij}^*\mathbf{V}_0(\hat{\mathbf{B}}-\mathbf{B}_0)\big)|,\\L_2&=\left|\mathrm{tr}\left[(\hat{\mathbf{B}}-\mathbf{B}_0)\frac{1}{n}\sum_{s=1}^{n}\big(\mathbf{Y}_j^{(s)}\mathbf{Y}_i^{(s)T}-u_{ij}^*\mathbf{V}_0\big)\right]\right|.\end{aligned}$$

Since $\max_{i,j}|u_{ij}^*|\le\|\mathbf{U}_0\|\le\varepsilon_1^{-1}$ and $\|\mathbf{V}_0\|\le\varepsilon_3^{-1}$, then

$$\begin{aligned}L_1&\le\max_{i,j}|u_{ij}^*|\sqrt{\mathrm{tr}(\mathbf{V}_0^T\mathbf{V}_0)}\sqrt{\mathrm{tr}\big((\hat{B}-\mathbf{B}_0)^T(\hat{\mathbf{B}}-\mathbf{B}_0)\big)}\\&\le\sqrt{q_n}\|\mathbf{U}_0\|\,\|\mathbf{V}_0\|\,\|\hat{\mathbf{B}}-\mathbf{B}_0\|_F\le\varepsilon_1^{-1}\varepsilon_3^{-1}\sqrt{q_n}\|\hat{B}-\mathbf{B}_0\|_F.\end{aligned}$$

On the other hand, by (A.16) and Lemma A.3 of [5],

$$\max_{k,l\in\{1,\ldots,q_n\}}\left|\left(\frac{1}{n}\sum_{s=1}^{n}\mathbf{Y}_j^{(s)}\mathbf{Y}_i^{(s)T}-u_{ij}^*\mathbf{V}_0\right)_{(k,l)}\right|=O_P\left(\sqrt{\frac{\log q_n}{n}}\right).$$

Denoting $\mathbb{1}_q$ a $q_n\times 1$ vector of 1's, then

$$\begin{aligned}&\left\|\frac{1}{n}\sum_{s=1}^{n}\mathbf{Y}_j^{(s)}\mathbf{Y}_i^{(s)T}-u_{ij}^*\mathbf{V}_0\right\|_F^2=\mathrm{tr}\left[\left(\frac{1}{n}\sum_{s=1}^{n}\mathbf{Y}_j^{(s)}\mathbf{Y}_i^{(s)T}-u_{ij}^*\mathbf{V}_0\right)^T,\right.\\&\left.\left(\frac{1}{n}\sum_{s=1}^{n}\mathbf{Y}_j^{(s)}\mathbf{Y}_i^{(s)T}-u_{ij}^*\mathbf{V}_0\right)\right]=O_P\left(\frac{\log q_n}{n}\mathrm{tr}(\mathbb{1}_q\mathbb{1}_q^T\mathbb{1}_q\mathbb{1}_q^T)\right)=O_P\left(q_n^2\frac{\log q_n}{n}\right).\end{aligned}$$

Therefore,

$$L_2\le\|\hat{\mathbf{B}}-\mathbf{B}_0\|_F\Big\|\frac{1}{n}\sum_{s=1}^{n}\mathbf{Y}_j^{(s)}\mathbf{Y}_i^{(s)T}-u_{ij}^*\mathbf{V}_0\Big\|_F=O_P\left(\sqrt{\frac{q_n\log q_n}{n}}\sqrt{q_n}\|\hat{\mathbf{B}}-\mathbf{B}_0\|_F\right).$$

Then $L_1 + L_2 = O_P\big(\sqrt{q_n}\|\hat{\mathbf{B}} - \mathbf{B}_0\|_F(1 + \sqrt{q_n \log q_n/n})\big)$. Since the theorem requires the condition $q_n(p_n + s_{n1})(\log p_n + \log q_n)^k/n = O(1)$ for some $k > 1$, we know that $\sqrt{q_n \log q_n/n} = o(1)$. So

$$|I_3| \le L_1 + L_2 = O_P(\sqrt{q_n}\|\hat{\mathbf{B}} - \mathbf{B}_0\|_F) = O_P\left(\sqrt{\frac{p_n q_n(q_n + s_{n2})(\log p_n + \log q_n)}{n}}\right).$$

Concluding from above, we have

$$\begin{aligned}|I_1| + |I_2| + |I_3| &\le O_P\left(\sqrt{\frac{q_n \log p_n}{n}}\right) + O_P\big(q_n\sqrt{c_n}\big) + O_P\left(\sqrt{\frac{p_n q_n(q_n + s_{n2})(\log p_n + \log q_n)}{n}}\right)\\ &= \sqrt{q_n}O_P\left(\sqrt{\frac{p_n(q_n + s_{n2})(\log p_n + \log q_n)}{n}} + \sqrt{q_n c_n}\right).\end{aligned}$$

On the other hand, for $(i, j) \in S_A^c$, $\lambda_n/|\tilde{a}_{ij}|^{\gamma_1} \ge c\lambda_n e_n^{\gamma_1}$ for some constant c. From the condition in the theorem, we have

$$e_n^{-\gamma_1}\sqrt{q_n}\left\{\sqrt{\frac{p_n(q_n + s_{n2})(\log p_n + \log q_n)}{n}} + \sqrt{c_n q_n}\right\} = O(\lambda_n).$$

So the sign of $\partial\phi(\hat{\mathbf{A}}, \hat{\mathbf{B}})/\partial a_{ij}$ is dominated by $\mathrm{sgn}(a_{ij})$, and thus we proved the sparsistency for $\hat{\mathbf{A}}$. A similar proof can be applied to $\hat{\mathbf{B}}$. □

References

[1] G. Allen, Comment on article by Hoff, Bayesian Analysis 6 (2) (2011) 197–202.
[2] G. Allen, R. Tibshirani, Transposable regularized covariance models with an application to missing data imputation, The Annals of Applied Statistics 4 (2) (2010) 764–790.
[3] G. Allen, R. Tibshirani, Inference with transposable data: modeling the effects of row and column correlations, Journal of the Royal Statistical Society, Series B (Theory & Methods) (2011) (in press).
[4] O. Banerjee, L.E. Ghaoui, A. d'Aspremont, Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data, Journal of Machine Learning Research 9 (2008) 485–516.
[5] P. Bickel, E. Levina, Regularized estimation of large covariance matrices, Annals of Statistics 36 (1) (2008) 199–227.
[6] T. Cai, W. Liu, X. Luo, A constrained l_1 minimization approach to sparse precision matrix estimation, Journal of the American Statistical Association 106 (2011) 594–607.
[7] E. Candes, T. Tao, The Dantzig selector: statistical estimation when p is much larger than n, Annals of Statistics 35 (2007) 2313–2351.
[8] N. Cressie, Statistics for Spatial Data, Wiley, New York, 1993.
[9] J. Dahl, L. Vandenberghe, V. Roychowdhury, Covariance selection for non-chordal graphs via chordal embedding, Optimization Methods and Software 23 (2008) 501–520.
[10] A. Dawid, Some matrix-variate distribution theory: notational considerations and a Bayesian application, Biometrika 68 (1981) 265–274.
[11] P. Dutilleul, The mle algorithm for the matrix normal distribution, Journal of Statistical Computation and Simulation 64 (1999) 105–123.
[12] B. Efron, Are a set of microarrays independent of each other? The Annals of Applied Statistics 13 (3) (2009) 922–942.
[13] J. Fan, Y. Feng, Y. Wu, Network exploration via the adaptive lasso and scad penalties, The Annals of Applied Statistics 3 (2009) 521–541.
[14] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, Journal of the American Statistical Association 96 (2001) 1348–1360.
[15] J.D. Finn, A General Model for Multivariate Analysis, Holt, Rinehart and Winston, New York, 1974.
[16] J. Friedman, T. Hastie, R. Tibshirani, Sparse inverse covariance estimation with the graphical lasso, Biostatistics 9 (2008) 432–441.
[17] F.A. Graybill, Matrices with Applications in Statistics, second ed., Wadsworth, Belmont, 1983.
[18] A. Gupta, D. Nagar, Matrix Variate Distributions, in: Monographs and Surveys in Pure and Applied Mathematics, vol. 104, Chapman & Hall, CRC Press, Boca Raton, FL, 1999.
[19] P. Hoff, Separable covariance arrays via the Tucker product, with applications to multivariate relational data, Bayesian Analysis 6 (2) (2011) 179–196.
[20] K. Holmes, O. Roberts, A. Thomas, M. Cross, Vascular endothelial growth factor receptor—2: structure, function, intracellular signalling and therapeutic inhibition, Cellular Signalling 19 (2007) 2003–2012.
[21] H.M. Huizenga, J.C. De Munck, L.J. Waldorp, R. Grasman, Spatiotemporal EEG/MEG source analysis based on a parametric noisecovariance model, IEEE Transactions on Biomedical Engineering 49 (2002) 533–539.
[22] C. Lam, J. Fan, Sparsistency and rates of convergence in large covariance matrices estimation, The Annals of Statistics 37 (2009) 4254–4278.
[23] S.L. Lauritzen, Graphical Models, Clarendon Press, Oxford, 1996.
[24] E. Lehmann, Theory of Point Estimation, Wadsworth and Brooks/Cole, Pacific Grove, CA, 1983.
[25] H. Li, J. Gui, Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks, Biostatistics 7 (2006) 302–317.
[26] K.V. Mardia, C. Goodall, Spatial-temporal analysis of multivariate environmental monitoring data, Multivariate Environmental Statistics 6 (1993) 347–385.
[27] N. Meinshausen, P. Bühlmann, High-dimensional graphs and variable selection with the lasso, Annals of Statistics 34 (2006) 1436–1462.
[28] M. Mitchell, M. Genton, M. Gumpertz, A likelihood ratio test for separability of covariances, Journal of Multivariate Analysis 97 (5) (2006) 1025–1043.
[29] O. Muralidharan, Detecting column dependence when rows are correlated and estimating the strength of the row correlation, Electronic Journal of Statistics 4 (2010) 1527–1546.
[30] P. Ravikumar, M. Wainwright, G. Raskutti, B. Yu, High-dimensional covariance estimation by minimizing l_1-penalized log-determinant divergence, Electronic Journal of Statistics 5 (2011) 935–980.
[31] A. Rothman, E. Levina, J. Zhu, Generalized thresholding of large covariance matrices, Journal of the American Statistical Association 104 (2009) 177–186.
[32] S. Teng, H. Huang, A statistical framework to infer functional gene associations from multiple biologically interrelated microarray experiments, Journal of the American Statistical Association 104 (2009) 465–473.
[33] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B 58 (1996) 267–288.

[34] N.H. Timm, Multivariate analysis of variance of repeated measurements, Handbook of Statistics 1 (1980) 41–87.
[35] L.R. Tucker, The extension of factor analysis to three-dimensional matrices, in: H. Gulliksen, N. Frederiksen (Eds.), Contributions to Mathematical Psychology, Holt, Rinehart and Winston, New York, 1964.
[36] H. Wang, M. West, Bayesian analysis of matrix normal graphical models, Biometrika 96 (2009) 821–834.
[37] J. Whittaker, Graphical Models in Applied Multivariate Analysis, Wiley, 1990.
[38] M. Yuan, Sparse inverse covariance matrix estimation via linear programming, Journal of Machine Learning Research 11 (2010) 2261–2286.
[39] M. Yuan, Y. Lin, Model selection and estimation in the Gaussian graphical model, Biometrika 94 (2007) 19–35.
[40] J.M. Zahn, S. Poosala, A. Owen, D.K. Ingram, A. Lustig, et al., Agemap: a gene expression database for aging in mice, PLoS Genetics 3 (11) (2007) 2326–2337.
[41] Y. Zhang, J. Schneider, Learning multiple tasks with a sparse matrix-normal penalty, in: J. Lafferty, C.K.I. Williams, J. Shawe-Taylor, R. Zemel, A. Culotta (Eds.), in: Advances in Neural Information Processing Systems, vol. 23, 2010, pp. 2550–2558.
[42] H. Zou, The adaptive lasso and its oracle properties, Journal of the American Statistical Association 101 (2006) 1418–1429.