Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Sparse Matrix Graphical Models

Chenlei Leng & Cheng Yong Tang

Sparse Matrix Graphical Models

Chenlei LENG and Cheng Yong TANG

Matrix-variate observations are frequently encountered in many contemporary statistical problems due to a rising need to organize and analyze data with structured information. In this article, we propose a novel sparse matrix graphical model for these types of statistical problems. By penalizing, respectively, two precision matrices corresponding to the rows and columns, our method yields a sparse matrix graphical model that synthetically characterizes the underlying conditional independence structure. Our model is more parsimonious and is practically more interpretable than the conventional sparse vector-variate graphical models. Asymptotic analysis shows that our penalized likelihood estimates enjoy better convergent rates than that of the vector-variate graphical model. The finite sample performance of the proposed method is illustrated via extensive simulation studies and several real datasets analysis.

KEY WORDS: Conditional independence; Matrix-variate normal distribution; Penalized likelihood; Sparsistency; Sparsity.

1. INTRODUCTION

The rapid advance in information technology has brought an unprecedented array of high-dimensional data. Besides a large number of collected variables, structural information is often available in the data collection process. Matrix-variate variable is an important way of organizing high-dimensional data to incorporate such structural information, and is commonly encountered in multiple applied areas such as brain imaging studies, financial market trading, macroeconomics analysis, and many others. Consider the following concrete examples:

- To meet the demands of investors, options contingent on equities and market indices are frequently traded with multiple combinations of striking prices and expiration dates. A dataset collects weekly implied volatilities, which are equivalent to the prices, of options for 89 component equities in the Standard and Poor (S&P) 100 index for eight respective expiration dates, such as 30, 60 days. In this example, each observation of the dataset can be denoted by an 89 × 8 matrix, whose rows are the companies and whose columns encode information of the expiration dates. A snapshot of the dataset, after processing as discussed in Section 5.2, for three selected companies, Abbott Laboratories, ConocoPhillips, and Microsoft, is presented in Figure 1.
- United States Department of Agriculture (USDA) reports itemized annual export to major trading partners. A dataset with 40 years U.S. export is collected for 13 trading partners and 36 items. Each observation in the dataset can be denoted by a 13 × 36 matrix where the trading partners and items, as the rows and columns of this matrix, are used as structural information for the observations.
- In brain imaging studies, it is routine to apply electroencephalography (EEG) on an individual in an attempt to understand the relationship between brain imaging and the trigger of some events, for example, alcohol consumption. A typical experiment scans each subject from a large number of channels of electrodes at hundreds of time points. Therefore, the observation of each subject is conveniently denoted as a large channel by time matrix.

We refer to this type of structured data $\mathbf{X} \in \mathbb{R}^{p\times q}$ as matrix-variate data or simply matrix data if no confusion arises. It may appear attempting to stack $\mathbf{X}$ as a column vector vec($\mathbf{X}$) and model $\mathbf{X}$ as a $p \times q$-dimensional vector. Gaussian graphical models (Lauritzen 1996), when applied to vector data, are useful for representing conditional independence structure among the variables. A graphical model in this case consists of a vertex set and an edge set. The absence of an edge between two vertices denotes that the corresponding pair of variables are conditionally independent, given all the other variables. To build a sparse graphical model for vec($\mathbf{X}$), there exists abundant literature that makes use of penalized likelihood (Meinshausen and Bühlmann 2006; Yuan and Lin 2007; Banejee, El Ghaoui, and d'Aspremont 2008; Rothman et al. 2008; Fan, Feng, and Wu 2009; Lam and Fan 2009; Peng et al. 2009; Guo et al. 2010, 2011). However, these approaches suffer from at least two obvious shortcomings when applied to matrix data. First, the need to estimate a $p^2 \times q^2$-dimensional covariance (or precision) matrix can be a daunting task due to the extremely high dimensionality. Second, any analysis based on vec($\mathbf{X}$) effectively ignores all row and column structural information, an incoherent part of the data characteristics. In practice, this structural information is useful and sometimes vital for interpretation purposes, and, as discussed later, for convergent rate considerations. New approaches that explore the matrix nature of such datasets are therefore called for to meet the emerging challenges in analyzing such data.

As an extension of the familiar multivariate Gaussian distribution for vector data, we consider the matrix-variate normal distribution for $\mathbf{X}$ with probability density function

$$p(\mathbf{X}|\mathbf{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}) = (2\pi)^{-qp/2}|\boldsymbol{\Sigma}^{-1}|^{q/2}|\boldsymbol{\Psi}^{-1}|^{p/2}\mathrm{etr}\{-(\mathbf{X}-\mathbf{M})\boldsymbol{\Psi}^{-1} \times (\mathbf{X}-\mathbf{M})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}/2\}, \quad (1)$$

where $\mathbf{M} \in \mathbb{R}^{p\times q}$ is the mean matrix, $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ and $\boldsymbol{\Psi} \in \mathbb{R}^{q\times q}$ are the row and column variance matrices, and etr is the exponential of the trace operator. The matrix normal distribution

Chenlei Leng is Associate Professor, Department of Statistics and Applied Probability, National University of Singapore, Singapore (E-mail: *stalc@nus.edu.sg*). Cheng Yong Tang is Assistant Professor, the Business School, University of Colorado Denver, and Assistant Professor, Department of Statistics and Applied Probability, National University of Singapore, Singapore (E-mail: *chengyong.tang@ucdenver.edu*). Research support from National University of Singapore research grants is gratefully acknowledged. We thank coeditor Prof. Xuming He, an associate editor, and two anonymous referees for their constructive comments that have led to a much improved article.

© 2012 American Statistical Association
Journal of the American Statistical Association
September 2012, Vol. 107, No. 499, Theory and Methods
DOI: 10.1080/01621459.2012.706133

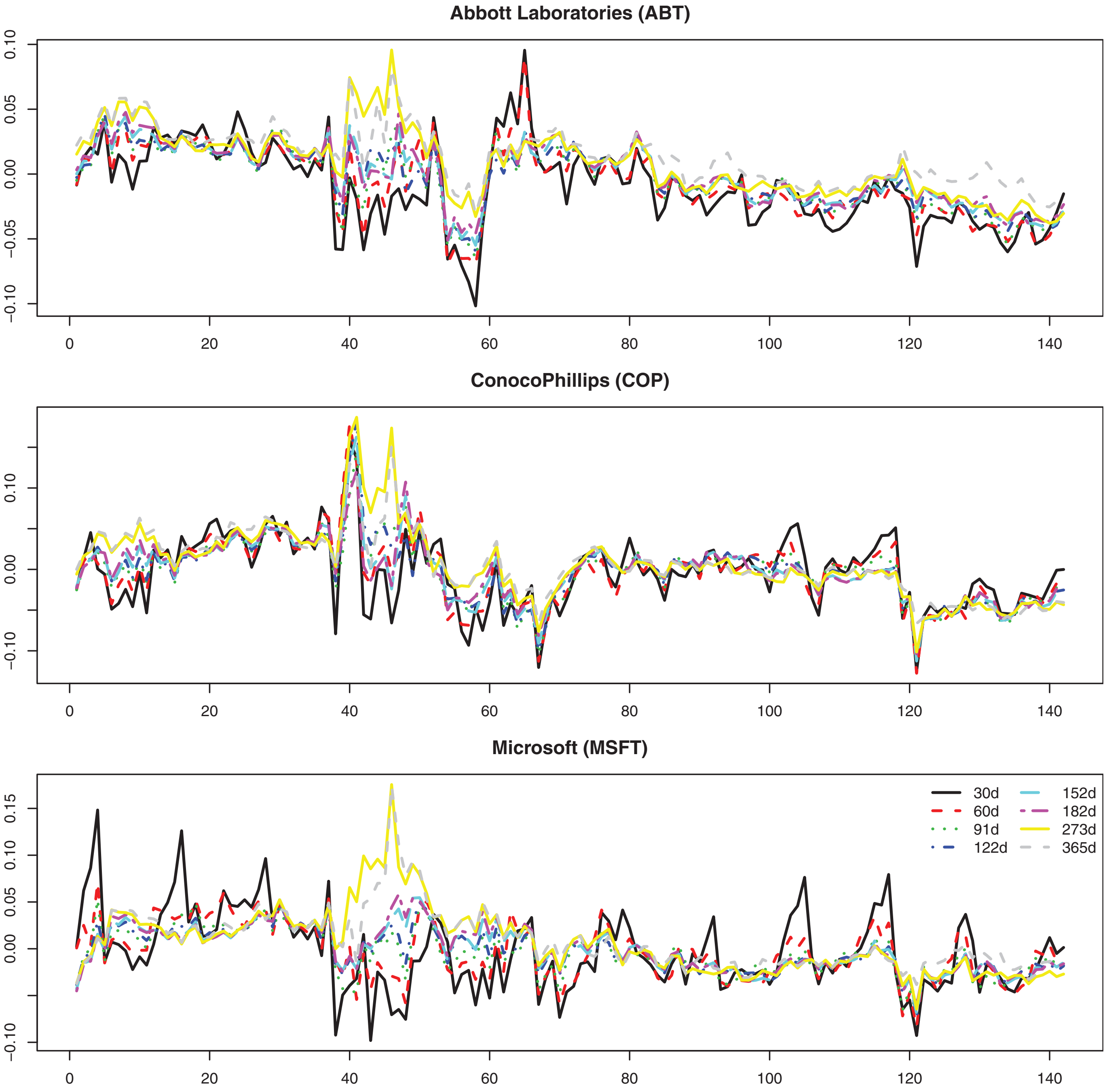

Figure 1. Volatility data for 142 weeks of trading. The online version of this figure is in color.

(1) implies that vec($\mathbf{X}$) follows a vector multivariate normal distribution $N_{pq}(\mathrm{vec}(\mathbf{M}), \mathbf{\Psi} \otimes \mathbf{\Sigma})$, where $\otimes$ is the Kronecker product. Models using matrix instead of vector normal distributions effectively reduce the dimensionality of the covariance or precision matrix from the order of $p^2 \times q^2$ to an order of $p^2 + q^2$. Without loss of generality, we assume $\mathbf{M} = 0$ from now on. Otherwise, we can always center the data by subtracting $\hat{\mathbf{M}} = \sum_{i=1}^{n} \mathbf{X}_i / n$ from $\mathbf{X}_i$, where $\mathbf{X}_i$, $i = 1, \ldots, n$, are assumed to be independent and identically distributed (iid) following the matrix normal distribution (1). Dawid (1981) provided some theory for matrix-variate distributions and Dutilleul (1999) derived the maximum likelihood estimation. Allen and Tibshirani (2010) studied regularized estimation in a model similar to (1) with a single observation when $n = 1$, but did not provide any theoretical result. Wang and West (2009) studied Bayesian inference for such models.

Matrix-variate Gaussian distributions are useful for characterizing conditional independence in the underlying matrix variables; see Dawid (1981) and Gupta and Nagar (2000). Let $\mathbf{\Omega} = \mathbf{\Sigma}^{-1} = (\omega_{ij})$ and $\mathbf{\Gamma} = \mathbf{\Psi}^{-1} = (\gamma_{ij})$ be the precision matrices for row and column vectors, respectively. Similar to that in a vector-variate graphical model, all the conditional independence can be analogously read off from the precision matrices. In particular, zeros in $\mathbf{\Gamma} \otimes \mathbf{\Omega}$ define pairwise conditional independence of corresponding entries (Lauritzen 1996), given all the other variables. More importantly, zeros in $\mathbf{\Omega}$ and $\mathbf{\Gamma}$

encode conditional independence of the row variables and the column variables in $\mathbf{X}$, respectively, often creating a much simpler graphical model for understanding matrix data. Formally, in matrix normal distribution, two arbitrary entries X_{ij} and X_{kl} are conditionally independent, given remaining entries if and only if (1) at least one zero in ω_{ij} and γ_{kl} when $i \neq k$, $j \neq l$; (2) $\omega_{ik} = 0$ when $i \neq k$, $j = l$; and (3) $\gamma_{jl} = 0$ when $i = k$, $j \neq l$. In terms of the partial correlation between variables X_{ij} and X_{kl}, defined as

$$\rho_{ij,kl} = -\frac{\omega_{ik}}{\sqrt{\omega_{ii}\omega_{kk}}} \cdot \frac{\gamma_{jl}}{\sqrt{\gamma_{jj}\gamma_{ll}}},$$

$\rho_{ij,kl}$ is not zero only if both ω_{ik} and γ_{jl} are nonzero. If either the ith and the kth rows are conditionally independent given other rows, or the jth and the lth columns are conditionally independent given other columns, or both, the partial correlation between variables X_{ij} and X_{kl} would be zero. An obvious issue with the parameterization of the matrix normal distribution is its identifiability. In later development, we fix $\omega_{11} = 1$.

In this article, we propose sparse matrix graphical models by considering regularized maximum likelihood estimate (MLE) for matrix data that follow the distribution in (1). By applying appropriate penalty functions on precision matrices $\mathbf{\Omega}$ and $\mathbf{\Gamma}$, we obtain sparse matrix graphical models for the row and column variables. Theoretical results show that the structural information in the rows and columns of the matrix variable can be exploited to yield more efficient estimates of these two precision matrices than those in the vector graphical model. Given appropriate penalty functions, we show that our method gives sparsistent graphical models. Namely, we are able to identify zeros in these two precision matrices correctly with probability tending to one. We demonstrate the performance of the proposed approach in extensive simulation studies. Through several real data analysis, we illustrate the attractiveness of the proposed approach in disentangling structures of complicated high-dimensional data as well as improving interpretability.

The rest of the article is organized as follows. We present the proposed methodology in Section 2, followed by a discussion on an iterative algorithm for fitting the model. The asymptotic properties of the proposed method are provided in Section 3. Simulations are presented in Section 4, and data analyses are given in Section 5. Section 6 summarizes the main findings and outlines future research. All proofs are found in the Appendix.

2. METHOD

To estimate $\mathbf{\Omega}$ and $\mathbf{\Gamma}$, given iid data $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from (1), the standard approach is the maximum likelihood estimation. It is easily observed that the minus log-likelihood up to a constant is

$$\ell(\mathbf{\Omega}, \mathbf{\Gamma}) = -\frac{nq}{2}\log|\mathbf{\Omega}| - \frac{np}{2}\log|\mathbf{\Gamma}| + \frac{1}{2}\sum_{i=1}^{n}\operatorname{tr}\left(\mathbf{X}_i\mathbf{\Gamma}\mathbf{X}_i^{\mathrm{T}}\mathbf{\Omega}\right). \tag{2}$$

The solution minimizing (2) is the familiar MLE. To build sparse models for $\mathbf{\Omega}$ and $\mathbf{\Gamma}$, we propose the penalized likelihood estimator as the minimizer of

$$g(\mathbf{\Omega}, \mathbf{\Gamma}) = \frac{1}{npq}\sum_{i=1}^{n}\operatorname{tr}\left(\mathbf{X}_i\mathbf{\Gamma}\mathbf{X}_i^{\mathrm{T}}\mathbf{\Omega}\right) - \frac{1}{p}\log|\mathbf{\Omega}| - \frac{1}{q}\log|\mathbf{\Gamma}| + p_{\lambda_1}(\mathbf{\Omega}) + p_{\lambda_2}(\mathbf{\Gamma}), \tag{3}$$

where $p_\lambda(\mathbf{A}) = \sum_{i \neq j} p_\lambda(|a_{ij}|)$ for a square matrix $\mathbf{A} = (a_{ij})$, with a suitably defined penalty function p_λ. In this article, we use the LASSO penalty (Tibshirani 1996) defined as $p_\lambda(s) = \lambda|s|$ or the smoothly clipped absolute derivation (SCAD) penalty (Fan and Li 2001), whose first derivative is given by

$$p'_\lambda(s) = \lambda\left\{I(s \leq \lambda) + \frac{(3.7\lambda - s)_+}{2.7\lambda}I(s > \lambda)\right\}$$

for $s > 0$. Here, $(s)_+ = s$ if $s > 0$ and is zero otherwise. The SCAD penalty is introduced to overcome the induced bias when estimating those nonzero parameters (Fan and Li 2001). Loosely speaking, by imposing a small or zero penalty on a nonzero estimate, the SCAD effectively produces unbiased estimates of the corresponding entries. In contrast, the LASSO, imposing a constant penalty on all estimates, inevitably introduces biases that may affect the rate of convergence and consistency in terms of model selection (Lam and Fan 2009). We shall call collectively the resulting model sparse matrix graphical models (SMGMs). When either $\mathbf{\Sigma}$ or $\mathbf{\Psi}$ is an identity matrix, the distribution in (1) is equivalent to a q- or p-dimensional multivariate normal distribution. This is seen by observing that the rows (columns) of $\mathbf{X}$ are independently normal distributed, with covariance matrix $\mathbf{\Psi}$ ($\mathbf{\Sigma}$) when $\mathbf{\Sigma} = \mathbf{I}_p$ ($\mathbf{\Psi} = \mathbf{I}_q$). Therefore, SMGM includes the multivariate Gaussian graphical model as a special case.

The matrix normal distribution is a natural way to encode the structural information in the row and column variables. The SMGM provides a principled framework for studying such information content when conditional independence is of particular interest. In high-dimensional data analysis, imposing sparsity constraints can often stabilize estimation and improve prediction accuracy. More importantly, it is an effective mechanism to overcome the curse of dimensionality for such data analysis.

We now discuss how to implement SMGM. Note that the penalized log-likelihood is not a convex function, but it is conditional convex if the penalty function is convex, for example, in the LASSO case. This naturally suggests an iterative algorithm that optimizes one matrix with the other matrix fixed. When $\mathbf{\Omega}$ is fixed, we minimize with respect to $\mathbf{\Gamma}$

$$\frac{1}{q}\operatorname{tr}(\mathbf{\Gamma}\tilde{\mathbf{\Sigma}}) - \frac{1}{q}\log|\mathbf{\Gamma}| + p_{\lambda_2}(\mathbf{\Gamma}), \tag{4}$$

where $\tilde{\mathbf{\Sigma}} = \sum_{i=1}^{n}\mathbf{X}_i^{\mathrm{T}}\mathbf{\Omega}\mathbf{X}_i/q$. When the LASSO penalty is used, we use the coordinate descent graphical LASSO (gLASSO) algorithm by Friedman, Hastie, and Tibshirani (2008) to obtain the solution. When used with warm start for different tuning parameters λ_2, it can solve large problems very efficiently even when $q \gg n$. For the SCAD penalty, we follow Zou and Li (2008) to linearize the penalty as

$$p_{\lambda_2}(s) = p_{\lambda_2}(s_0) + p'_{\lambda_2}(s_0)(|s| - |s_0|)$$

for the current estimate s_0. Thus, we only need to minimize

$$\frac{1}{q}\operatorname{tr}(\mathbf{\Gamma}\tilde{\mathbf{\Sigma}}) - \frac{1}{q}\log|\mathbf{\Gamma}| + \sum_{i \neq j} p'_{\lambda_2}(|\gamma_{ij,k}|)|\gamma_{ij}|,$$

which can be solved by using the gLASSO algorithm. Here, $\gamma_{ij,k}$ is the kth-step estimate of γ_{ij}. In practice, it is noted that in a different context, one iteration of this linearization procedure is sufficient given a good initial value (Zou and Li 2008). In our implementation, we take the initial value as the gLASSO estimates by taking $p'_{\lambda_2}(|\gamma_{ij,k}|)$ as λ_2. On the other hand, when $\boldsymbol{\Gamma}$ is fixed, we minimize with respect to $\boldsymbol{\Omega}$

$$\frac{1}{p}\mathrm{tr}(\tilde{\boldsymbol{\Psi}}\boldsymbol{\Omega}) - \frac{1}{p}\log|\boldsymbol{\Omega}| + p_{\lambda_1}(\boldsymbol{\Omega}), \qquad (5)$$

where $\tilde{\boldsymbol{\Psi}} = \sum_{i=1}^{n} \mathbf{X}_i\boldsymbol{\Gamma}\mathbf{X}_i^{\mathrm{T}}/p$.

Synthetically, the computational algorithm is summarized as follows:

1. Start with $\boldsymbol{\Gamma}^{(0)} = \mathbf{I}_q$, and minimize (4) to get $\boldsymbol{\Omega}^{(0)}$. Normalize $\boldsymbol{\Omega}^{(0)}$ such that $\omega_{11}^{(0)} = 1$. Let $m = 1$.
2. Fix $\boldsymbol{\Omega}^{(m-1)}$ and minimize (5) to get $\boldsymbol{\Gamma}^{(m)}$.
3. Fix $\boldsymbol{\Gamma}^{(m)}$ and minimize (4) to get $\boldsymbol{\Omega}^{(m)}$. Normalize such that $\omega_{11}^{(m)} = 1$. Let $m \leftarrow m + 1$.
4. Repeat Steps 2 and 3 until convergence.

Although there is no guarantee that the algorithm converges to the global minimum, the algorithm converges to a local stationary point of $g(\boldsymbol{\Omega}, \boldsymbol{\Gamma})$. An argument has been outlined by Allen and Tibshirani (2010) when the LASSO penalty is used in our approach. A detailed discussion of various algorithms and their convergence properties can be found in Gorski, Pfeuffer, and Klamroth (2007).

When there is no penalty, Dutilleul (1999) showed that each step of the iterative algorithm is well defined if $n \geq \max\{p/q, q/p\} + 1$. This is understandable since for estimating, for example, $\boldsymbol{\Omega}$, the effective sample size is essentially nq. More recently, Srivastava, von Rosen, and Von Rosen (2008) showed that the MLE exists and is unique, provided $n > \max\{p, q\}$. Thus, if the sample size is larger than the row and the column dimension, the algorithm guarantees to find the global optimal solution, if a convex penalty such as LASSO or ridge is used in our formulation.

Since the optimal tuning parameters λ_1 and λ_2 are not known in advance, a useful practice is to apply a warm-start method on decreasing sequences of values for these two parameters. For example, if we set these two parameters to be the same as $\lambda_1 = \lambda_2 = \rho$, we compute a sequence of solutions of (3) for $\rho_L > \rho_{L-1} > \cdots > \rho_1$, where ρ_L is chosen to yield very sparse models. The solution at ρ_i is then used as the initial value for the solution at ρ_{i-1} (Friedman, Hastie, and Tibshirani 2008). Note that if ρ_L is large, the resulting estimated $\boldsymbol{\Omega}$ and $\boldsymbol{\Gamma}$ would be very sparse. Thus, the algorithm is not very sensitive to the initial value for the solution at ρ_L because in this case, a much fewer number of parameters in the two precision matrices need to be estimated. Subsequently, for other values of ρ, the initial values would be very close, yielding the resulting estimates very close. Thus, the sensitivity of the algorithm is greatly reduced. Intuitively, the warm-start trick would work well for very sparse models. Further numerical evidence of this observation will be presented in the simulation studies.

3. ASYMPTOTICS

We now study the asymptotic properties of the proposed method in this section. Let $\boldsymbol{\Omega}_0$ and $\boldsymbol{\Gamma}_0$ be the true parameter of the underlying model, $S_1 = \{(i,j) : \omega_{0ij} \neq 0\}$ and $S_2 = \{(i,j) : \gamma_{0ij} \neq 0\}$. Define $s_1 = |S_1| - p$ and $s_2 = |S_2| - q$ as the number of nonzero off-diagonal parameters in $\boldsymbol{\Omega}_0$ and $\boldsymbol{\Gamma}_0$, respectively. We make the following standard regularity assumptions for theoretical analysis of the penalized likelihood.

A1. There exist constant τ_1 such that for all n,

$$0 < \tau_1 < \lambda_1(\boldsymbol{\Sigma}_0) \leq \lambda_p(\boldsymbol{\Sigma}_0) < 1/\tau_1 < \infty,$$
$$0 < \tau_1 < \lambda_1(\boldsymbol{\Psi}_0) \leq \lambda_q(\boldsymbol{\Psi}_0) < 1/\tau_1 < \infty,$$

where $\lambda_1(\mathbf{A}) \leq \lambda_2(\mathbf{A}) \leq \cdots \leq \lambda_m(\mathbf{A})$ denote the eigenvalues of an m-dimensional symmetric matrix $\mathbf{A}$.

A2. The penalty function $p_\lambda(\cdot)$ is singular at the origin, and $\lim_{t\downarrow 0} p_\lambda(t)/(\lambda t) = k > 0$.

A3. For the nonzero components in $\boldsymbol{\Omega}_0$ and $\boldsymbol{\Gamma}_0$,

$$\max_{(i,j)\in S_1} p'_{\lambda_1}(|\omega_{0ij}|) = O[p^{-1}\{1+p/(s_1+1)\}\sqrt{\log p/(nq)}],$$
$$\max_{(i,j)\in S_1} p''_{\lambda_1}(|\omega_{0ij}|) = o(p^{-1}),$$
$$\max_{(i,j)\in S_2} p'_{\lambda_2}(|\gamma_{0ij}|) = O[q^{-1}\{1+q/(s_2+1)\}\sqrt{\log q/(np)}],$$
$$\max_{(i,j)\in S_2} p''_{\lambda_2}(|\gamma_{0ij}|) = o(q^{-1}),$$
$$\min_{(i,j)\in S_1} |\omega_{0ij}|/\lambda_1 \to \infty \text{ and } \min_{(i,j)\in S_2} |\gamma_{0ij}|/\lambda_2 \to \infty \text{ as } n \to \infty.$$

A4. The tuning parameters satisfy $\lambda_1^{-2}p^{-2}(p+s_1)\log p/(nq) \to 0$, $\lambda_2^{-2}q^{-2}(q+s_2)\log q/(np) \to 0$ as $n \to \infty$.

All conditions imposed here are mild and comparable with those in Lam and Fan (2009). Both the LASSO and the SCAD penalty satisfy Condition A2, while Condition A3 is less restrictive for the unbiased SCAD penalty than for the LASSO. The relations among the penalty functions, Condition A3, and the properties of the resulting estimates are discussed in more detail later. We also note that in Condition A4, due to the information from the matrix-variate structure, the tuning parameters λ_1 and λ_2 are allowed to converge to 0 at faster rates than those in Lam and Fan (2009). Thus, a smaller bias is induced due to this penalization.

We use the notations $\|\cdot\|_{\mathrm{F}}$ and $\|\cdot\|$ for the Frobenius and operator norms of a matrix. The following theorem establishes the rate of convergence of SMGM.

Theorem 1. [Rate of convergence] Under Conditions A1–A4, as $n \to \infty$, $(p+s_1)\log p/(nq) \to 0$, and $(q+s_2)\log q/(np) \to 0$, there exists a local minimizer of (3) such that

$$\frac{\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0\|_{\mathrm{F}}^2}{p} = O_p\left\{\left(1+\frac{s_1}{p}\right)\log p/(nq)\right\}$$

and

$$\frac{\|\hat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}_0\|_{\mathrm{F}}^2}{q} = O_p\left\{\left(1+\frac{s_2}{q}\right)\log q/(np)\right\}.$$

Theorem 1 shows that the rate of convergence is determined by n and p for $\boldsymbol{\Omega}$ and by n and q for $\boldsymbol{\Gamma}$. This represents the fact that the column (row) information in the matrix-variate data is incorporated in estimating the concentration matrix of the row (column), which indeed confirms the merit of the proposed SMGM. This result is a generalization of the results for which independent vectors are used in estimating concentration matrices. We show in the Appendix that Theorem 1 is also valid for the LASSO penalty if Condition A4 is changed to $\lambda_1 = O(p^{-1}\sqrt{\log p/(nq)})$ and $\lambda_2 = O(q^{-1}\sqrt{\log q/(np)})$. We comment that the desirable local minimizer may be difficult to identify in practice. If there exist multiple local minimizers, the one computed by the algorithm may not be the one having the asymptotic property, and there does not seem to exist an algorithm that can always give such a local minimizer.

As to the implied graphical model for vec($\mathbf{X}$) by SMGM, the rate of convergence for estimating $\boldsymbol{\Omega}_0 \otimes \boldsymbol{\Gamma}_0$ using our approach is easily seen as

$$\frac{\|\hat{\boldsymbol{\Omega}} \otimes \hat{\boldsymbol{\Gamma}} - \boldsymbol{\Omega}_0 \otimes \boldsymbol{\Gamma}_0\|_{\mathrm{F}}^2}{pq} = O_p\left[\max\left\{\left(1+\frac{s_1}{p}\right)\log p/(nq^2),\right.\right. \left.\left.\left(1+\frac{s_2}{q}\right)\log q/(np^2)\right\}\right].$$

If we apply the SCAD penalty to the vectorized observations, Lam and Fan (2009) gave the following rate of convergence:

$$\frac{\|\hat{\boldsymbol{\Omega}} \otimes \hat{\boldsymbol{\Gamma}} - \boldsymbol{\Omega}_0 \otimes \boldsymbol{\Gamma}_0\|_{\mathrm{F}}^2}{pq} = O_p\left[\left(1+\frac{s_1 s_2}{pq}\right)\log(pq)/n\right].$$

We immediately observe that when $p \to \infty$ and $q \to \infty$, SMGM gives estimates with a much faster rate of convergence. Indeed, the rate of convergence of our model is strictly faster as long as the dimensionality grows with the sample size. The improvement is due to our using a more parsimonious model that naturally incorporates the data structure.

In addition, we also note that the conditions in Theorem 1 inexplicitly put constraints on the growth rates of p and q. More specifically,

$$\max(p\log p/q, q\log q/p)/n \to 0 \text{ and}$$
$$\max(s_1\log p/q, s_2\log q/p)/n \to 0$$

are required. If we, without loss of generality, consider $p > q$, then $p\log p/q = o(n)$, $s_1\log p/q = o(n)$, and $s_2\log q/p = o(n)$ effectively determine the upper bound of p.

We establish the model selection consistency, also known as sparsistency, of the proposed approach in the following theorem.

Theorem 2. [Sparsistency] Under Conditions A1–A4, for local minimizers $\hat{\boldsymbol{\Omega}}$ and $\hat{\boldsymbol{\Gamma}}$ satisfying $\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0\|_{\mathrm{F}}^2 = O_p\{(p + s_1)\log p/(nq)\}$, $\|\hat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}_0\|_{\mathrm{F}}^2 = O_p\{(q + s_2)\log q/(np)\}$, $\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0\| = O_p(\eta_{1n})$, $\|\hat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}_0\| = O_p(\eta_{2n})$ for sequences η_{1n} and η_{2n} converging to 0, $\log p/(nq) + \eta_{1n} + \eta_{2n} = O_p(\lambda_1^2 p^2)$, and $\log q/(np) + \eta_{1n} + \eta_{2n} = O_p(\lambda_2^2 q^2)$, with probability tending to 1, we have $\hat{\omega}_{ij} = 0$ and $\hat{\gamma}_{ij} = 0$ for all $(i, j) \in S_1^c$ and $(i, j) \in S_2^c$.

If the LASSO penalty is used, the consistency result in Theorem 1 requires controlling the incurred bias by Condition A3 so that the tuning parameters cannot be too large. This effectively imposes upper bounds for the tuning parameters in the LASSO penalty as $\lambda_1 = O[p^{-1}\{1 + p/(s_1 + 1)\}\sqrt{\log p/(nq)}]$ and $\lambda_2 = O[q^{-1}\{1 + q/(s_2 + 1)\}\sqrt{\log q/(np)}]$. On the other hand, larger penalization is needed to achieve sparsistency as observed from Theorem 2. Therefore, to simultaneously achieve consistency and sparsistency when applying the LASSO penalty, the numbers of nonzero components in $\boldsymbol{\Omega}_0$ and $\boldsymbol{\Gamma}_0$ are restricted, even under optimal conditions. Similar to the discussions in Lam and Fan (2009), s_1 and s_2 can be at most $O(p)$ and $O(q)$, respectively, to ensure the consistency and sparsistency. While for those unbiased penalty functions such as the SCAD, s_1 and s_2 could be allowed to grow at $O(p^2)$ and $O(q^2)$ in the SMGM, benefited from the fact that Condition A3 does not impose upper bounds on tuning parameters for an unbiased penalty function.

It is remarkable that the restriction on the sample size n is largely relaxed due to incorporating the structural information from the multiple rows and columns. Even if $n < p$ and $n < q$, consistent estimates of $\boldsymbol{\Omega}$ and $\boldsymbol{\Gamma}$ are still achievable. When $p = q$, a sufficient condition for convergence of SMGM with the SCAD penalty is $\log(pq) = o(n)$ if the matrices are sparse enough. On the other hand, for the vector graphical model, we require at least $(pq)\log(pq) = o(n)$. In the extreme case when n is finite, it is still possible to obtain consistent estimates of the precision matrices by applying SMGM. To appreciate this, consider, for example, that $n = 1$, p is fixed, and q is growing. If $\boldsymbol{\Psi}_0$ is a correlation matrix, one can apply SMGM, with $\boldsymbol{\Gamma} = \mathbf{I}$, and obtain consistent and sparse estimate of $\boldsymbol{\Omega}$ following the proof in the Appendix. We conclude that by incorporating the structure information from matrix data, more efficient estimates of the graphical models can be obtained.

We now discuss the asymptotic properties of the estimates. Let $\mathbf{A}^{\otimes 2} = \mathbf{A} \otimes \mathbf{A}$ and $\mathbf{K}_{pq}$ be a $pq \times pq$ commutation matrix that transforms vec($\mathbf{A}$) to vec($\mathbf{A}^{\mathrm{T}}$) for a $p \times q$ matrix $\mathbf{A}$. A rigorous definition and its properties can be found, for example, in Gupta and Nagar (2000). Let $\boldsymbol{\Lambda}_{1n} = \mathrm{diag}\{p''_{\lambda_1}(\hat{\boldsymbol{\Omega}})\}$, $\boldsymbol{\Lambda}_{2n} = \mathrm{diag}\{p''_{\lambda_2}(\hat{\boldsymbol{\Omega}})\}$, $\mathbf{b}_{1n} = p'_{\lambda_1}(|\boldsymbol{\Omega}_0|)\mathrm{sgn}(\boldsymbol{\Omega}_0)$, and $\mathbf{b}_{2n} = p'_{\lambda_1}(|\boldsymbol{\Omega}_0|)\mathrm{sgn}(\boldsymbol{\Omega}_0)$. Denote S_1 as the set of all the indices of nonzero components in vec($\boldsymbol{\Omega}_0$) except ω_{11}, and S_2 as the set for all the indices of nonzero components in vec($\boldsymbol{\Gamma}_0$). We use the subscripts S and $S \times S$ for the corresponding subvector and submatrix, respectively.

Theorem 3. [Asymptotic normality] Under Conditions A1–A4, $(p + s_1)^2/(nq) \to 0$ and $(q + s_2)^2/(np) \to 0$ as $n \to \infty$, for the local minimizer $\hat{\boldsymbol{\Omega}}$ and $\hat{\boldsymbol{\Gamma}}$ in Theorem 1, we have

$$\sqrt{nq}\boldsymbol{\alpha}_p^{\mathrm{T}}\left\{\boldsymbol{\Sigma}_0^{\otimes 2}(\mathbf{I} + \mathbf{K}_{pp})\right\}_{S_1\times S_1}^{-1/2}\left(\boldsymbol{\Lambda}_{1n} + \boldsymbol{\Sigma}_0^{\otimes 2}\right)_{S_1\times S_1} \times \{\mathrm{vec}(\hat{\boldsymbol{\Omega}}) - \mathrm{vec}(\boldsymbol{\Omega}_0) + \mathbf{b}_{1n}\}_{S_1} \xrightarrow{d} N(0, 1),$$

$$\sqrt{np}\boldsymbol{\alpha}_q^{\mathrm{T}}\left\{\boldsymbol{\Psi}_0^{\otimes 2}(\mathbf{I} + \mathbf{K}_{qq})\right\}_{S_2\times S_2}^{-1/2}\left(\boldsymbol{\Lambda}_{2n} + \boldsymbol{\Psi}_0^{\otimes 2}\right)_{S_2\times S_2} \times \{\mathrm{vec}(\hat{\boldsymbol{\Gamma}}) - \mathrm{vec}(\boldsymbol{\Gamma}_0) + \mathbf{b}_{2n}\}_{S_2} \xrightarrow{d} N(0, 1),$$

where $\boldsymbol{\alpha}_d$ denotes a d-dimensional unit vector and $\xrightarrow{d}$ denotes convergence in distribution.

Theorem 3 clearly illustrates the impact of using the matrix-variate data. In comparison with those in Lam and Fan (2009), fast rates of convergence $\sqrt{nq}$ and $\sqrt{np}$ are achieved for estimating the precision matrices of the columns and the rows, respectively. Similar to that in Theorem 1, a caution is that the local minimizer may be different from the one computed by the algorithm. It is of great interest to develop an algorithm that guarantees identifying the local minimizer in Theorems 1 and 3.

4. SIMULATION AND DATA ANALYSIS

We conduct extensive simulation studies in this section. For comparison purposes, we tabulate the performance of the MLE of $\boldsymbol{\Omega}$ and $\boldsymbol{\Gamma}$ using the algorithm by Dutilleul (1999). Note that this algorithm is similar to the algorithm for computing the SMGM estimate, when the penalty function is absent. We

Table 1. Simulation results based on the Kullback–Leibler loss. The sample standard errors are in parentheses

n	p	q	$\mathbf{\Omega}$	$\mathbf{\Gamma}$	MLE	Ridge	LASSO	SCAD	gLASSO
10	10	10	$\mathbf{A}_1$	$\mathbf{A}_1$	22.6 (5.0)	11.8 (2.3)	12.5 (4.1)	6.9 (2.4)	105.4 (9.9)
			$\mathbf{A}_2$	$\mathbf{A}_1$	22.9 (4.2)	11.2 (1.7)	11.7 (2.3)	11.3 (2.2)	77.8 (5.6)
			$\mathbf{A}_2$	$\mathbf{A}_2$	23.9 (4.8)	10.4 (1.4)	15.8 (3.1)	17.0 (3.5)	44.4 (3.7)
			$\mathbf{A}_3$	$\mathbf{A}_1$	22.5 (3.9)	10.7 (1.7)	8.1 (2.0)	7.4 (2.0)	84.7 (6.4)
			$\mathbf{A}_3$	$\mathbf{A}_2$	22.5 (4.7)	10.0 (1.5)	11.2 (2.2)	10.6 (2.2)	54.6 (4.4)
			$\mathbf{A}_3$	$\mathbf{A}_3$	23.7 (4.5)	9.7 (1.3)	7.6 (1.8)	7.4 (2.0)	57.4 (5.2)
100	10	10	$\mathbf{A}_1$	$\mathbf{A}_1$	1.2 (0.2)	1.4 (0.2)	0.71 (0.12)	0.38 (0.09)	30.7 (0.8)
			$\mathbf{A}_2$	$\mathbf{A}_1$	1.2 (0.2)	1.2 (0.1)	1.0 (0.1)	0.8 (0.2)	17.2 (0.5)
			$\mathbf{A}_2$	$\mathbf{A}_2$	1.2 (0.1)	1.1 (0.1)	1.1 (0.1)	1.1 (0.1)	15.9 (0.5)
			$\mathbf{A}_3$	$\mathbf{A}_1$	1.2 (0.2)	1.1 (0.2)	0.68 (0.11)	0.38 (0.09)	18.0 (0.5)
			$\mathbf{A}_3$	$\mathbf{A}_2$	1.2 (0.2)	1.1 (0.1)	1.0 (0.1)	0.82 (0.13)	13.9 (0.5)
			$\mathbf{A}_3$	$\mathbf{A}_3$	1.2 (0.1)	1.1 (0.1)	0.66 (0.11)	0.35 (0.08)	8.4 (0.4)
10	20	20	$\mathbf{A}_1$	$\mathbf{A}_1$	89.6 (7.9)	45.6 (3.5)	28.9 (5.8)	13.9 (2.6)	509 (19)
			$\mathbf{A}_2$	$\mathbf{A}_1$	87.6 (8.1)	42.8 (2.8)	33.5 (3.1)	33.2 (3.4)	356 (12)
			$\mathbf{A}_2$	$\mathbf{A}_2$	88.6 (7.5)	40.3 (2.5)	50.2 (4.1)	50.6 (4.3)	185 (8)
			$\mathbf{A}_3$	$\mathbf{A}_1$	88.0 (8.0)	42.1 (3.2)	21.4 (3.3)	16.9 (2.9)	410 (17)
			$\mathbf{A}_3$	$\mathbf{A}_2$	90.3 (9.7)	39.0 (2.8)	42.0 (5.9)	37.0 (5.8)	228 (10)
			$\mathbf{A}_3$	$\mathbf{A}_3$	88.5 (9)	37.2 (2.9)	20.8 (3.3)	17.6 (3.2)	273 (11)
100	20	20	$\mathbf{A}_1$	$\mathbf{A}_1$	4.5 (0.3)	5.5 (0.3)	1.9 (0.2)	0.9 (0.2)	169 (1)
			$\mathbf{A}_2$	$\mathbf{A}_1$	4.6 (0.4)	4.5 (0.3)	3.1 (0.3)	2.0 (0.3)	94.2 (0.9)
			$\mathbf{A}_2$	$\mathbf{A}_3$	4.4 (0.3)	4.2 (0.2)	4.0 (0.3)	2.8 (0.3)	83.6 (1.3)
			$\mathbf{A}_3$	$\mathbf{A}_1$	4.6 (0.3)	4.6 (0.3)	2.1 (0.2)	0.81 (0.13)	110 (1.3)
			$\mathbf{A}_3$	$\mathbf{A}_2$	4.5 (0.3)	4.2 (0.3)	3.3 (0.3)	2.0 (0.2)	72.1 (1.2)
			$\mathbf{A}_3$	$\mathbf{A}_3$	4.6 (0.4)	4.2 (0.3)	1.5 (0.2)	0.77 (0.13)	67.1 (1.2)

also compare SMGM with the gLASSO method for estimating $\mathbf{\Omega}_0 \otimes \mathbf{\Gamma}_0$ (Friedman, Hastie, and Tibshirani 2008). In particular, this approach ignores the matrix structure of the observations and vectorizes them as $\mathbf{x} = \text{vec}(\mathbf{X})$. The gLASSO method then optimizes

$$\text{tr}(\mathbf{\Theta}\mathbf{S}) - \log|\mathbf{\Theta}| + \lambda \sum_{i \neq j} |\theta_{ij}|,$$

where $\mathbf{S}$ is the sample covariance matrix of $\mathbf{x}_i$. In addition, we implement a ridge-type regularized method by using squared matrix Frobenius norm $\|\mathbf{A}\|_{\text{F}}^2$ in the penalty function in (3). We do not report the sample covariance estimates because for most of the simulations reported here, the sample size is too small comparing with $p \times q$. For each simulation setup, we conduct 50 replications. To choose the tuning parameters, we generate a random test dataset with the sample size equal to the training data.

We use the following $d \times d$ matrices as the building block for generating sparse precision matrices for $\mathbf{\Omega}$ and $\mathbf{\Gamma}$ (Rothman et al. 2008).

1. $\mathbf{A}_1$: Inverse AR(1) such that $\mathbf{A}_1 = \mathbf{B}^{-1}$ with $b_{ij} = 0.7^{|i-j|}$.
2. $\mathbf{A}_2$: AR(4) with $a_{ij} = I(|i-j| = 0) + 0.4I(|i-j| = 1) + 0.2I(|i-j| = 2) + 0.2I(|i-j| = 3) + 0.1I(|i-j| = 4)$.
3. $\mathbf{A}_3 = \mathbf{B} + \delta\mathbf{I}$: each off-diagonal upper triangle entry in $\mathbf{B}$ is generated independently and equals to 0.5, with probability $a = 0.1$, and 0, with probability $1 - a = 0.9$. The diagonals of $\mathbf{B}$ are zero and δ is chosen such that the condition number of $\mathbf{A}_3$ is d.

All matrices are sparse and are numbered in order of decreasing sparsity. We assess the estimation accuracy for a precision matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ using the Kullback–Leibler loss, defined as

$$l(\mathbf{A}, \hat{\mathbf{A}}) = \text{tr}(\mathbf{A}^{-1}\hat{\mathbf{A}}) - \log|\mathbf{A}^{-1}\hat{\mathbf{A}}| - d.$$

Here, we use $\mathbf{A} = \mathbf{\Omega}_0 \otimes \mathbf{\Gamma}_0$, the main parameter of interest in the Kullback–Leibler loss, and $\hat{\mathbf{A}}$ is an estimate of it. We also summarize the performance in terms of true positive rate (TPR) and true negative rate (TNR), defined as

$$\text{TPR} = \frac{\#\{\hat{A}_{ij} \neq 0 \;\&\; A_{ij} \neq 0\}}{\#\{A_{ij} \neq 0\}}, \ \text{TNR} = \frac{\#\{\hat{A}_{ij} = 0 \;\&\; A_{ij} = 0\}}{\#\{A_{ij} = 0\}}.$$

We first generate random datasets with sample sizes $n = 10$ or $n = 100$, with $p = q = 10$ or 20. For these relatively small-dimensional datasets, we compare the MLE, the ridge estimator (Ridge), our proposed method using LASSO penalty (LASSO), our proposed method with the SCAD penalty (SCAD), and the graphical LASSO vectorizing the matrix data (gLASSO). The results based on the Kullback–Leibler loss are summarized in Table 1, with the model selection results summarized in Table 2. From Table 1, we see clearly that the gLASSO method by ignoring the matrix structure performs much worse than all the other approaches by a large margin. The MLE gives similar accuracy as the ridge estimator with $n = 100$, but performs much worse with $n = 10$. This illustrates the benefit of regularization. Among the three regularized estimates, denoted as Ridge, LASSO, and SCAD, the two sparse estimates LASSO and SCAD prevail in general, especially when

Table 2. Simulation results based on model selection. The sample standard errors are in parentheses

					LASSO		SCAD		gLASSO	
n	p	q	$\mathbf{\Omega}$	$\mathbf{\Gamma}$	TPR	TNR	TPR	TNR	TPR	TNR
10	10	10	$\mathbf{A}_1$	$\mathbf{A}_1$	1 (0)	0.82 (0.03)	1 (0)	0.95 (0.06)	0.50 (0.05)	0.92 (0.01)
			$\mathbf{A}_2$	$\mathbf{A}_1$	0.70 (0.11)	0.76 (0.07)	0.70 (0.10)	0.80 (0.11)	0.18 (0.02)	0.93 (0.01)
			$\mathbf{A}_2$	$\mathbf{A}_2$	0.53 (0.16)	0.69 (0.16)	0.43 (0.16)	0.79 (0.16)	0.05 (0.02)	0.98 (0.02)
			$\mathbf{A}_3$	$\mathbf{A}_1$	1 (0)	0.78 (0.05)	1.0 (0.01)	0.90 (0.07)	0.44 (0.04)	0.93 (0.01)
			$\mathbf{A}_3$	$\mathbf{A}_2$	0.75 (0.11)	0.77 (0.10)	0.72 (0.11)	0.84 (0.14)	0.14 (0.02)	0.97 (0.01)
			$\mathbf{A}_3$	$\mathbf{A}_3$	1 (0)	0.88 (0.04)	1.0 (0.01)	0.92 (0.05)	0.52 (0.04)	0.96 (0.01)
100	10	10	$\mathbf{A}_1$	$\mathbf{A}_1$	1 (0)	0.66 (0.08)	1 (0)	0.99 (0.02)	0.81 (0.05)	0.66 (0.03)
			$\mathbf{A}_2$	$\mathbf{A}_1$	1.0 (0.01)	0.38 (0.11)	1.0 (0.01)	0.74 (0.07)	0.35 (0.01)	0.83 (0.01)
			$\mathbf{A}_2$	$\mathbf{A}_2$	1.0 (0.00)	0.16 (0.06)	1.0 (0.01)	0.46 (0.11)	0.23 (0.01)	0.87 (0.01)
			$\mathbf{A}_3$	$\mathbf{A}_1$	1 (0)	0.77 (0.08)	1 (0)	0.99 (0.01)	0.75 (0.02)	0.81 (0.02)
			$\mathbf{A}_3$	$\mathbf{A}_2$	1 (0)	0.38 (0.09)	1 (0)	0.70 (0.07)	0.37 (0.01)	0.87 (0.01)
			$\mathbf{A}_3$	$\mathbf{A}_3$	1 (0)	0.78 (0.07)	1 (0)	0.95 (0.01)	0.93 (0.01)	0.91 (0.01)
10	20	20	$\mathbf{A}_1$	$\mathbf{A}_1$	1 (0)	0.93 (0.01)	1 (0)	0.99 (0.01)	0.38 (0.03)	0.97 (0.00)
			$\mathbf{A}_2$	$\mathbf{A}_1$	0.72 (0.07)	0.86 (0.03)	0.67 (0.11)	0.91 (0.05)	0.10 (0.01)	0.98 (0.00)
			$\mathbf{A}_2$	$\mathbf{A}_2$	0.51 (0.11)	0.79 (0.07)	0.47 (0.11)	0.84 (0.06)	0.02 (0.00)	1.0 (0.00)
			$\mathbf{A}_3$	$\mathbf{A}_1$	1 (0)	0.91 (0.02)	1 (0)	0.97 (0.02)	0.31 (0.02)	0.98 (0.00)
			$\mathbf{A}_3$	$\mathbf{A}_2$	0.90 (0.06)	0.80 (0.04)	0.87 (0.06)	0.90 (0.03)	0.07 (0.01)	0.99 (0.00)
			$\mathbf{A}_3$	$\mathbf{A}_3$	1 (0)	0.92 (0.02)	1 (0)	0.96 (0.02)	0.26 (0.02)	0.99 (0.00)
100	20	20	$\mathbf{A}_1$	$\mathbf{A}_1$	1 (0)	0.83 (0.02)	1 (0)	1.00 (0.00)	0.58 (0.00)	0.93 (0.00)
			$\mathbf{A}_2$	$\mathbf{A}_1$	1.00 (0.00)	0.60 (0.04)	1 (0)	0.89 (0.02)	0.25 (0.01)	0.94 (0.00)
			$\mathbf{A}_2$	$\mathbf{A}_2$	1 (0)	0.36 (0.03)	1 (0)	0.71 (0.03)	0.12 (0.01)	0.95 (0.00)
			$\mathbf{A}_3$	$\mathbf{A}_1$	1 (0)	0.78 (0.03)	1 (0)	1.00 (0.00)	0.59 (0.00)	0.93 (0.00)
			$\mathbf{A}_3$	$\mathbf{A}_2$	1.00 (0.00)	0.49 (0.05)	1.00 (0.00)	0.84 (0.02)	0.26 (0.01)	0.96 (0.00)
			$\mathbf{A}_3$	$\mathbf{A}_3$	1 (0)	0.90 (0.03)	1 (0)	1.00 (0.00)	0.63 (0.01)	0.95 (0.00)

$p = q = 20$. The SCAD consistently outperforms the LASSO in terms of the Kullback–Leibler loss, especially so when n becomes large ($n = 100$). In terms of model selection, Table 2 reveals that the SMGM with either the LASSO or the SCAD penalty outperforms the gLASSO estimates drastically. Note "1" in the table means that model selection is 100% correct, while "1.0" means that the rate is 100% after rounding. The TPR and TNR of the SCAD method improve those of the LASSO, especially when n is large. This difference is due to the bias issue inherent in the LASSO type of penalizations (Fan and Li 2001).

We now investigate the situation when p and q are large. To study the case when $p \neq q$, we let $(p, q) = (100, 20)$ and $(p, q) = (100, 50)$ with $n = 20$ or 100. Since for this setting, the gLASSO takes much longer time to compute, we report only the results for all the other approaches. The results are summarized in Table 3 for estimation. Again, the penalized estimates usually outperform the MLE. The two sparse estimates using either LASSO or SCAD penalty are the two best approaches in terms of estimation accuracy. In addition, the SCAD performs much better than the LASSO penalty using the matrix normal distribution. This is true also for model selection (results not shown). Even for a relatively small sample size ($n = 20$), the SMGM method with the SCAD penalty gives the TPR and the TNR very close to one. Comparing to the results in Table 2 with a comparable sample size, it seems that we are blessed with high dimensionality. For example, when $(p, q) = (100, 50)$ for $n = 100$, the model selection results are consistently better than those in Table 2 when $(p, q) = (20, 20)$.

Based on these simulation results, we observe clearly that by exploiting the covariance matrix of vec($\mathbf{X}$) as a Kronecker product of two matrices, the proposed method outperforms the gLASSO in general, regardless of whether any penalty is used. When the two precision matrices are sparse, the simulation results demonstrate clearly that by imposing appropriate penalties, we can estimate the covariance matrices more accurately with excellent model selection results. In addition, as expected from the asymptotic results, the SCAD penalty gives better performance in terms of Kullback–Leibler loss and model selection than the LASSO.

To investigate the sensitivity of the algorithm to the initial values, we conduct further numerical simulations. We take $\mathbf{\Omega} = \mathbf{\Gamma} = \mathbf{A}_3$ with $n = 10$ and $p = q = 20$. We then take three possible initial values. The first is the MLE computed using our algorithm, with identity matrices as the initial values when there is no penalty. The second uses identity matrices as the initial values. The third uses the true matrices as the initial values. We implement the warm-start approach and compare the final estimated precision matrices with the tuning parameters chosen as outlined before. The three estimates are denoted as $\mathbf{B}_j$, $j = 1, 2, 3$, for estimating $\mathbf{\Omega} \otimes \mathbf{\Gamma}$ with these three initial values. We then compute the Frobenius norm of $\mathbf{D}_1 = \mathbf{B}_1 - \mathbf{B}_3$ and $\mathbf{D}_2 = \mathbf{B}_2 - \mathbf{B}_3$. We find that these norms are effectively all zero across 1000 simulations, suggesting that this algorithm is robust with respect to the initial values. If we change the sparsity parameter a for generating $\mathbf{A}_3$ to 0.5, we find that the Frobenius norms of 999 $\mathbf{D}_1$'s are effectively zero, while those of 1000 $\mathbf{D}_2$'s are zero. The final estimates seem to depend on the initial values

Table 3. Simulation results based on the Kullback–Leibler loss. The sample standard errors are in parentheses

n	p	q	$\boldsymbol{\Omega}$	$\boldsymbol{\Gamma}$	MLE	Ridge	LASSO	SCAD
20	100	20	$\mathbf{A}_1$	$\mathbf{A}_1$	504 (15)	273 (7)	67.7 (6.1)	17.1 (2.0)
			$\mathbf{A}_2$	$\mathbf{A}_1$	504 (14)	268 (6)	131 (5)	114.6 (3.7)
			$\mathbf{A}_2$	$\mathbf{A}_2$	502 (16)	245 (4)	183 (11)	158 (12)
			$\mathbf{A}_3$	$\mathbf{A}_1$	504 (13)	258 (6)	133 (10)	100 (6)
			$\mathbf{A}_3$	$\mathbf{A}_2$	506 (15)	239 (5)	150 (9)	118 (9)
			$\mathbf{A}_3$	$\mathbf{A}_3$	507 (16)	222 (5)	117 (6)	91 (7)
100	100	20	$\mathbf{A}_1$	$\mathbf{A}_1$	59.6 (1.4)	54.3 (1.1)	12.3 (0.7)	2.60 (0.27)
			$\mathbf{A}_2$	$\mathbf{A}_1$	59.1 (1.4)	52.1 (1.1)	28.7 (0.8)	14.2 (0.8)
			$\mathbf{A}_2$	$\mathbf{A}_2$	59.5 (1.4)	52.3 (1.1)	32.9 (0.9)	15.3 (0.8)
			$\mathbf{A}_3$	$\mathbf{A}_1$	59.0 (1.1)	53.0 (0.8)	24.7 (0.8)	8.06 (0.47)
			$\mathbf{A}_3$	$\mathbf{A}_2$	59.4 (1.3)	51.4 (0.9)	30.2 (1.1)	11.7 (0.7)
			$\mathbf{A}_3$	$\mathbf{A}_3$	59.7 (1.3)	50.7 (0.9)	19.9 (0.8)	7.2 (0.4)
20	100	50	$\mathbf{A}_1$	$\mathbf{A}_1$	471 (11)	325 (6)	68.5 (3.5)	23.9 (2.1)
			$\mathbf{A}_2$	$\mathbf{A}_1$	471 (11)	323 (6)	160 (4)	123 (5)
			$\mathbf{A}_2$	$\mathbf{A}_2$	472 (12)	318 (6)	216 (6)	148 (6)
			$\mathbf{A}_3$	$\mathbf{A}_1$	473 (11)	314 (6)	126 (6)	68.4 (4.3)
			$\mathbf{A}_3$	$\mathbf{A}_2$	473 (10)	310 (6)	176 (8)	97 (5)
			$\mathbf{A}_3$	$\mathbf{A}_3$	472 (10)	292 (5)	165 (17)	69.2 (3.7)
100	100	50	$\mathbf{A}_1$	$\mathbf{A}_1$	68.5 (1.3)	79.4 (1.2)	17.3 (0.8)	3.41 (0.28)
			$\mathbf{A}_2$	$\mathbf{A}_1$	68.4 (1.3)	65.9 (1.2)	41.8 (1.1)	10.9 (0.5)
			$\mathbf{A}_2$	$\mathbf{A}_2$	68.4 (1.1)	63.7 (1.1)	52.5 (1.0)	13.3 (0.5)
			$\mathbf{A}_3$	$\mathbf{A}_1$	68.3 (1.3)	76.9 (1.1)	45.6 (1.3)	8.21 (0.32)
			$\mathbf{A}_3$	$\mathbf{A}_2$	68.1 (1.3)	64.4 (1.2)	66.2 (2.0)	11.3 (0.7)
			$\mathbf{A}_3$	$\mathbf{A}_3$	68.4 (1.4)	68.9 (1.0)	49.2 (1.1)	8.21 (0.48)

when the less sparse matrices $\boldsymbol{\Omega}$ and $\boldsymbol{\Gamma}$ are used. However, the effect seems to be minimal, although in this case, convergence cannot be guaranteed due to the need to estimate a large number of parameters.

5. DATA ANALYSIS

5.1 U.S. Agricultural Export Data

The U.S. agricultural export is an important part of U.S. export. In March 2011 alone, the U.S. agricultural export exceeds 13.3 billion U.S. dollars and gives an agricultural trade surplus of about 4.5 billion U.S. dollars. Understanding the export pattern can be useful to understand the current trend and predict future export trends. We extract the annual U.S. agricultural export data between 1970 and 2009 from the USDA website. We look at the annual U.S. export data to 13 regions, including North America, Caribbean, Central America, South America, European Union-27, Other Europe, East Asia, Middle East, North Africa, Sub-Saharan Africa, South Asia, Southeast Asia, and Oceania. The 36 export items are broadly categorized as bulk items, including, for example, wheat, rice, and so on; intermediate items, consisting of, for example, wheat flour, soybean oil, and so on; and consumer-oriented items, including snack food, breakfast cereals, and so on. These product groups are adopted from Foreign Agricultural Service BICO HS-10 codes. Thus, the dataset comprises of 40 matrices with dimensionality 14 × 36. A selected 40 years' data plot for North America, East Asia, and Southeast Asia for these 36 items can be found in Figure 2.

The original data are denoted in thousands U.S. dollars. After imputing the one missing value with zero, we take logarithm of the original data plus one. To reduce the potential serial correlations in this multivariate time series dataset, we first take the lag one difference for each region item combination, and apply our method to the differences as if they were independent for further analysis. Box–Pierce tests of lag 10 for these 36 × 13 time series after the differencing operation suggest that most of them can pass the test for serial correlations. We note that this simple operation can be best seen as a rough preprocessing step, and may not achieve independence among data. Therefore, more elaborate models may be needed if that is desirable. We choose the tuning parameter using leave-out-one cross-validation, which minimizes the average Kullback–Leibler loss on the testing data. Since the SCAD penalty gives sparser models, here we only report the results for this penalty.

The final fitted graphical models for the regions and the items are presented in Figure 3, where there are 43 edges for the regions and 254 edges for the items. It is found that all of the nonzero edges for the regions are negative, indicating that the U.S. export to one region is negatively affected by the export to other regions. We then obtain the standard errors using 1000 bootstrap samples for the nonzero edges and plot in Figure 4 the asymptotic 95% pointwise confidence intervals. Among these edges, the magnitude between Europe Union and Other Europe, and that between East Asia and Southeast Asia are the strongest. Interestingly, none of the 11 largest edges corresponds to either North Africa or Sub-Saharan Africa.

To compare the performance of the proposed decomposition with that of the usual gLASSO estimate, we use cross-validation. Specifically, each time, one matrix is left out for comparison in

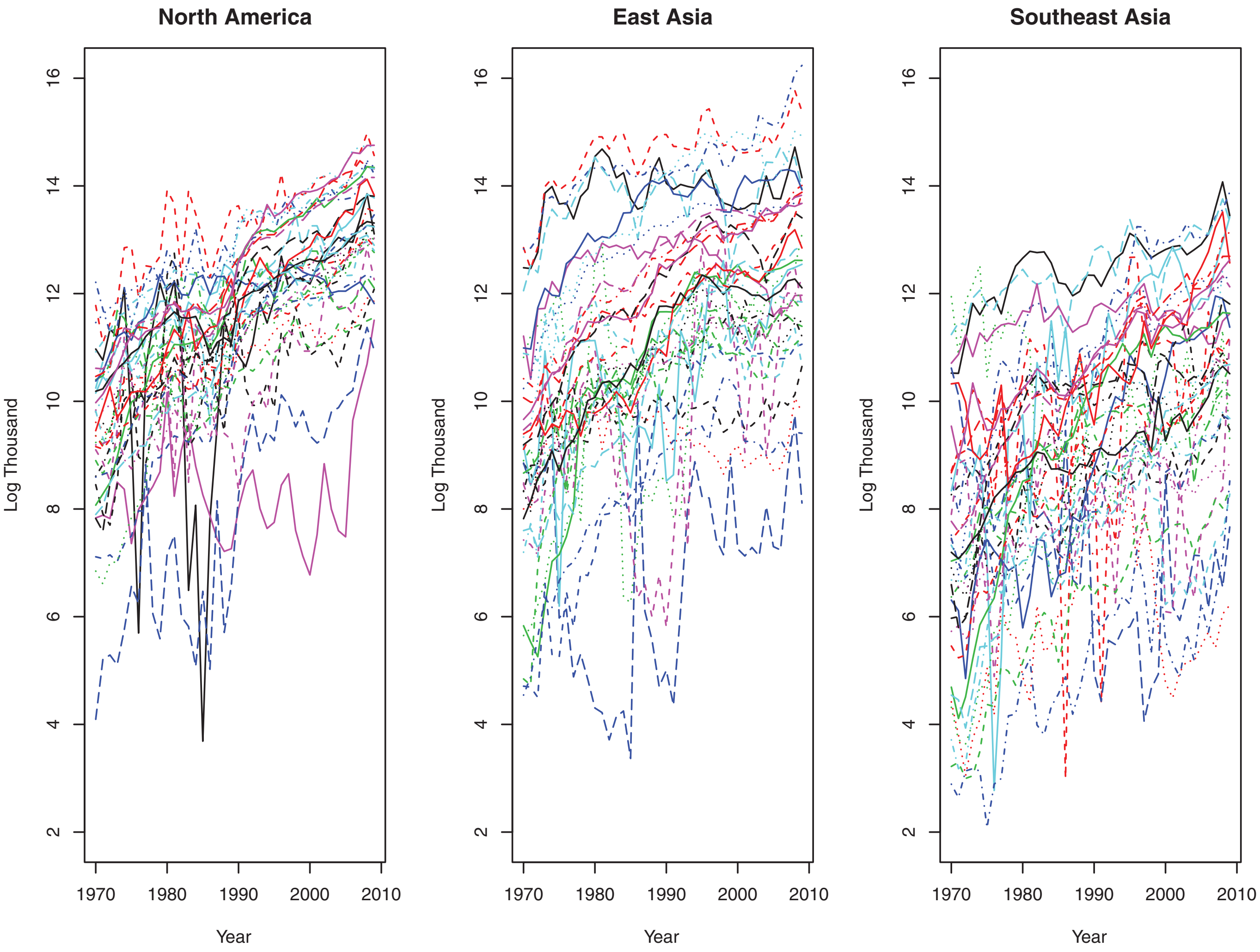

Figure 2. USDA export data for three regions over 40 years for 36 items. Each connected line represents one item. The online version of this figure is in color.

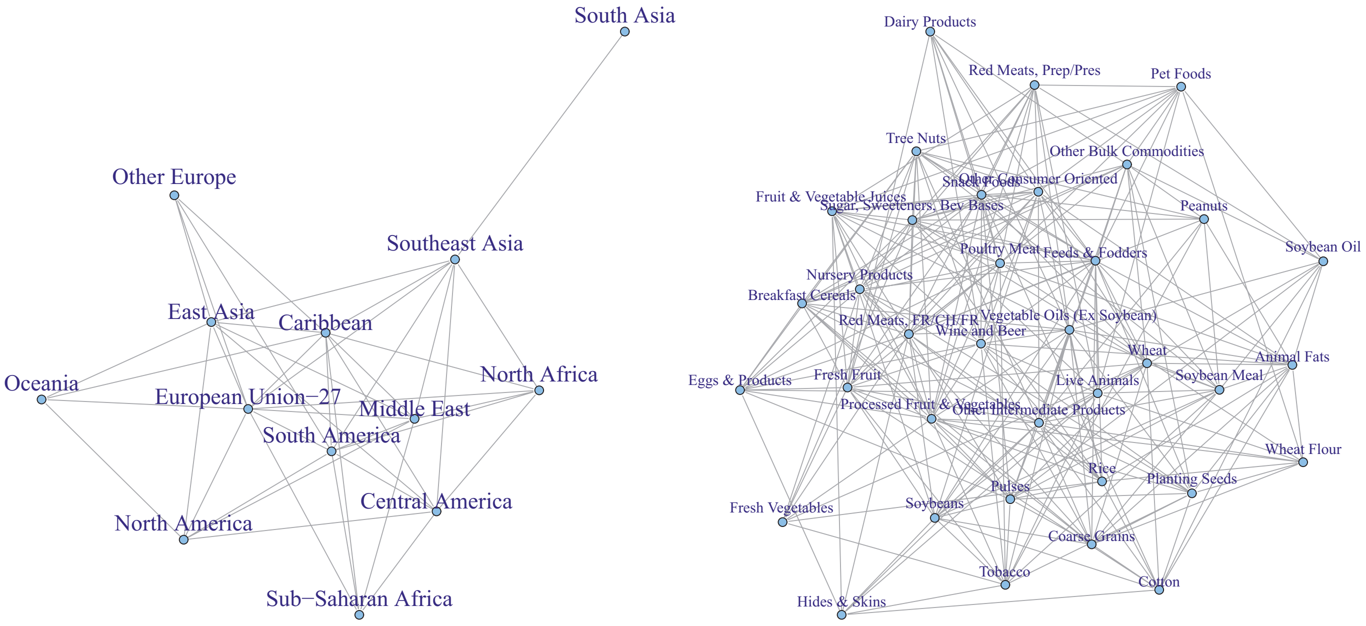

Figure 3. The fitted graphical models of the regions (left) and of the items (right) for the USDA export data from 1970 to 2009. The online version of this figure is in color.

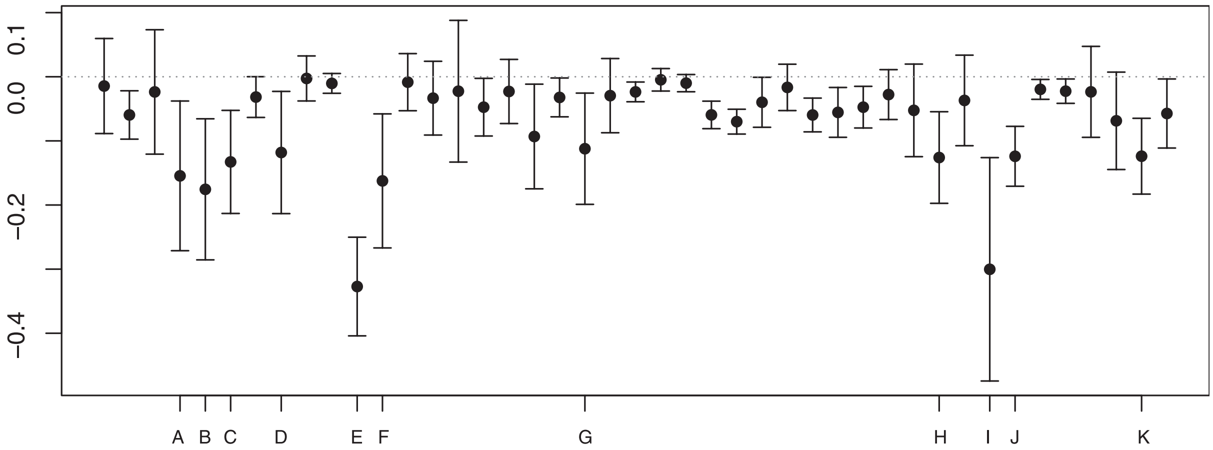

Figure 4. The estimates and the corresponding 95% confidence intervals for the 43 edges between the regions, where the largest 11 edges are marked. “A” is for CAR and SAM, “B” for CAM and SAM, “C” for NAM and EU, “D” for SAM and EU, “E” for EU and OE, “F” for NAM and EA, “G” for SAM and EU, “H” for SAM and SEA, “I” for EA and SEA, “J” for ME and SEA, and “K” for EU and OC, where EU denotes Europe Union, OE denotes Other Europe, EA denotes East Asia, SEA denotes Southeast Asia, CAM denotes Central America, NAM denotes North America, SAM denotes South America, CAR denotes Caribbean, and OC denotes Oceania.

terms of its log-likelihood. The rest matrices are used to build a model, using either our method or the gLASSO method, with the tuning parameters chosen by leave-out-one cross-validation. We then compute the log-likelihood of the matrix that is left out and take average. A simple two-sample t-test shows that our method with SCAD and the LASSO penalty perform similarly (p-value $= 0.9$), while our method with either penalty outperforms the usual gLASSO method with both p-values close to 0.01. This shows that our decomposition of the variance matrix is preferred over the decomposition that ignores the matrix structure of the data.

5.2 Implied Volatilities of Equity Options

The contemporary pricing theory (Merton 1990) for contingent claims is the most prominent development of finance in the past decades. However, understanding the pricing mechanism in actual market trading remains not satisfactorily resolved and attractive for many recent investigations; see, for example, Duan and Wei (2009) and reference therein. How market and economics factors affect the pricing mechanism is a fundamentally important question for both theoretical development and industrial practice. For demonstrating the application of the proposed approach in this area with abundant trading data, we consider the weekly option pricing data of 89 equities in the S&P 100 index, which includes the leading U.S. stocks with exchange-listed options. Constituents of the S&P 100 are selected for sector balance and represent about 57% of the market capitalization of the S&P 500 and almost 45% of the market capitalization of the U.S. equity markets.

Since the price of an option is equivalent to the implied volatility from the Black and Scholes’s (1973) model, the implied volatility data are as informative as option pricing data. We consider a dataset of the implied volatilities for standardized call options of 89 equities in S&P 100 from January 2008 to November 2010. The options are standardized in the sense that they are all at-the-money (striking price being equal to equity price) and expiring on unified dates, say 30, 60, 91, 122, 152, 192, 273, and 365 days in the future. High-level correlations are expected for the implied volatilities of different equities, and so for those expiring on different dates. To eliminate the systematic impact among the data, we first follow the treatment as in Duan and Wei (2009) to regress the implied volatilities using linear regression against the implied volatilities of the standardized options on the S&P 100 index. We then apply the proposed approach to the residuals to explore the unexplained correlation structures. We plot the connected components with more than two companies of the graphical models in Figure 5. The only other connected component consists of Bank of New York (BK) and JP Morgan Chase (JPM), which are independent of all the other companies. Other than these two, few financial firms are found connected in the estimated graphical model. This may imply the fact that the correlations of the implied volatilities among the financial clusters, as well as their impact on other clusters, can be well explained by those induced from the market index. As for the correlations among other firms on the estimated graph, in total, 59 companies are present on the left panel of Figure 5, where 219 out of 3916 possible edges are present. It is remarkable that very clear industrial groups can be identified. For instance, the tech companies, such as Amazon (AMZN), Microsoft (MSFT), Cisco (CSCO), IBM (IBM), QaulComm (QCOM), AT&T (T), Intel (INTC), and EMC Corporation (EMC), are tightly connected and form a community with MasterCard (MA). Home Depot and Lowe’s are connected, and the four oil companies ConocoPhilips (COP), Chevron (CVX), Occidental Petroleum Corporation (OXY), and Exxon Mobile (XOM) are closely connected. As for the graph associated with expiration dates on the right panel of Figure 5, 22 out of 28 possible edges are presented, where intuitive interpretation is quite clear. In particular, it seems that the call options to expire in 30, 60, and 365 days are most loosely connected with four edges each, and the call option to expire in 273 days has five edges, and the options to expire in

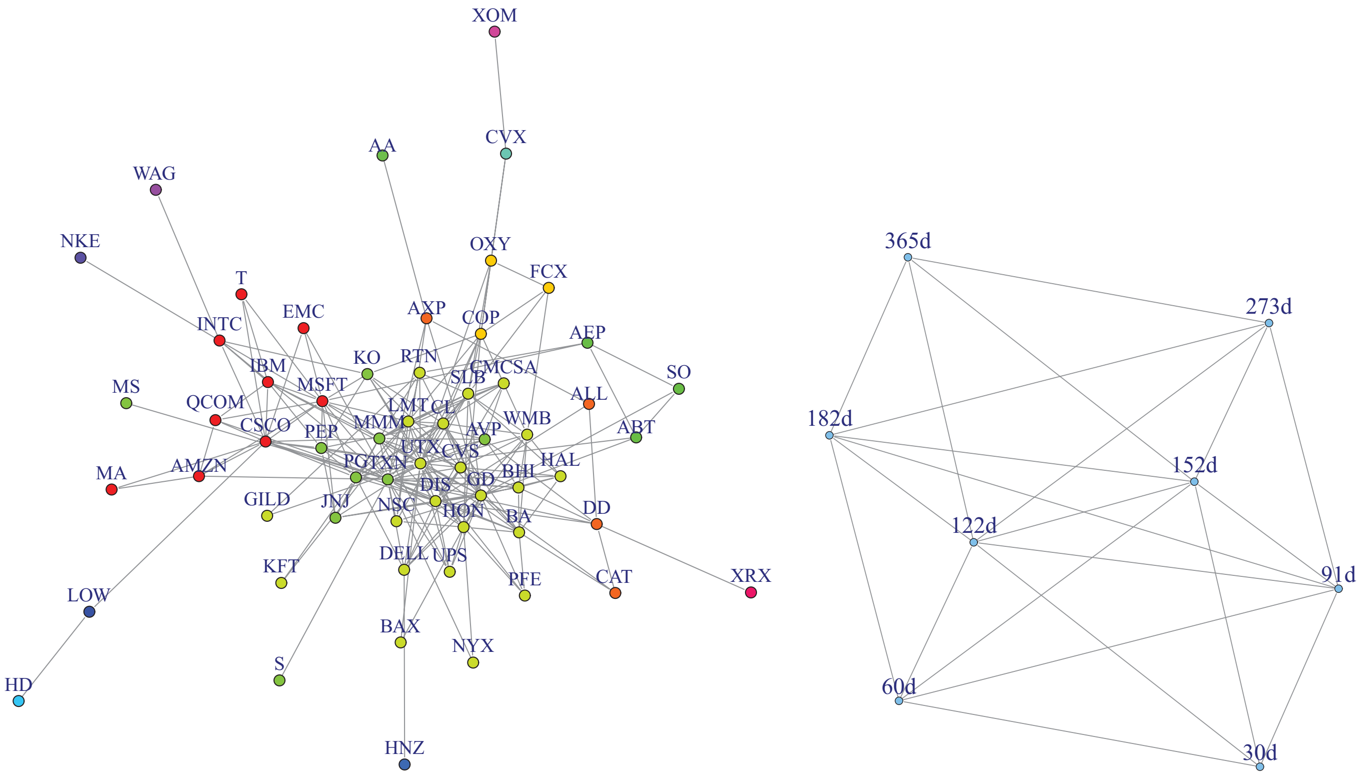

Figure 5. S&P 100: firms are labeled by their tickers with more detail on *http://www.standardandpoors.com/*. The colors indicate the community structure extracted via short random walks by Pons and Latapy (2005). The online version of this figure is in color.

91 and 183 days have six edges each, and the options to expire in 122 and 152 days are connected to all other options.

In summary, we observe clearly industrial and expiration dates pattern from the data analysis even after controlling the level of the implied volatility index of the S&P 100. Such finding echoes those in the constrained factor analysis on excess returns of equity stocks as in Tsai and Tsay (2010), and can be informative in studying the pricing mechanism of options contingent on equities. Again, caution is needed because we assume the data are independent.

To compare the performance of our method with the gLASSO method, we use the same strategy outlined for analyzing the U.S. agricultural export data. The cross-validation procedure shows that our approach with either LASSO or SCAD penalty outperforms the gLASSO method significantly.

6. CONCLUSION

We have proposed a novel framework to model high-dimensional matrix-variate data. We demonstrate via simulation and real data analysis that the structural information of this type of data deserves special treatment. Theoretical analysis shows that it is advantageous in modeling such datasets via matrix-variate normal distribution, not only for the rate of convergence consideration, but also for illuminating the relationships of the row and column variables.

In this article, we only study sparsity in the precision matrices. There are obvious ways to extend our work. For example, our framework can be modified to accommodate a sparse row precision matrix and a sparse column modified Cholesky matrix (Pourahmadi 1999; Huang et al. 2006), or two sparse covariance matrices (Bickel and Levina 2008a, b). The former would be reasonable for the volatility data because the column variables' expiration dates are ordered according to time. The proposed framework can be also applied to array-type data, where independent arrays with more than two indices are observed (Hoff 2011). With the many choices to model multiple matrices and a framework to study multidimensional data, these issues are of great interest and will be studied separately.

In illustrating our method through data analysis, we have simply treated the time series data as independent after some elementary operations. In practice, this is a rough approximation and may not be fully satisfactory. It is desirable to develop more elaborate time series models such as vector autoregressive moving average (ARMA) models to investigate the mean structure. Given the high dimensionality of the matrix observations, it is also of great interest to develop models that can handle the sparsity in the mean structure as well. A promising, related approach for the usual Gaussian graphical model has been found by Rothman, Levina, and Zhu (2010), where sparse regression models are employed in addition to sparse covariance estimation.

APPENDIX

Proof of Theorem 1. To decompose (3), we define $\mathbf{\Delta}_1 = \mathbf{\Omega} - \mathbf{\Omega}_0$, $\mathbf{\Delta}_2 = \mathbf{\Gamma} - \mathbf{\Gamma}_0$ and note that

$$\begin{aligned}&\operatorname{tr}\left(\mathbf{X}_i\mathbf{\Gamma}\mathbf{X}_i^{\mathrm{T}}\mathbf{\Omega}\right) - \operatorname{tr}\left(\mathbf{X}_i\mathbf{\Gamma}_0\mathbf{X}_i^{\mathrm{T}}\mathbf{\Omega}_0\right)\\ &\quad = \operatorname{tr}\left(\mathbf{X}_i\mathbf{\Gamma}_0\mathbf{X}_i^{\mathrm{T}}\mathbf{\Delta}_1\right) + \operatorname{tr}\left(\mathbf{X}_i\mathbf{\Delta}_2\mathbf{X}_i^{\mathrm{T}}\mathbf{\Omega}_0\right) + \operatorname{tr}\left(\mathbf{X}_i\mathbf{\Delta}_2\mathbf{X}_i^{\mathrm{T}}\mathbf{\Delta}_1\right).\end{aligned}$$

Further, by Taylor's expansion

$$\log|\mathbf{\Omega}| - \log|\mathbf{\Omega}_0| = \mathrm{tr}(\Sigma_0\mathbf{\Delta}_1) - \mathrm{vec}^{\mathrm{T}}(\mathbf{\Delta}_1) \times \left\{\int_0^1 h\left(v, \mathbf{\Omega}_v^{(1)}\right)(1-v)dv\right\}\mathrm{vec}(\mathbf{\Delta}_1),$$

where $\mathbf{\Omega}_v^{(1)} = \mathbf{\Omega}_0 + v\mathbf{\Delta}_1$ and $h(v, \mathbf{\Omega}) = \mathbf{\Omega}_v^{-1} \otimes \mathbf{\Omega}_v^{-1}$. We define

$$\begin{aligned}
T_1 &= \frac{1}{npq}\left\{\sum_{i=1}^{n}\mathrm{tr}\left(\mathbf{X}_i\mathbf{\Gamma}_0\mathbf{X}_i^{\mathrm{T}}\mathbf{\Delta}_1\right)\right\} - \frac{\mathrm{tr}(\mathbf{\Sigma}_0\mathbf{\Delta}_1)}{p},\\
T_2 &= \frac{1}{npq}\left\{\sum_{i=1}^{n}\mathrm{tr}\left(\mathbf{X}_i^{\mathrm{T}}\mathbf{\Omega}_0\mathbf{X}_i\mathbf{\Delta}_2\right)\right\} - \frac{\mathrm{tr}(\mathbf{\Psi}_0\mathbf{\Delta}_2)}{q},\\
T_3 &= \frac{1}{npq}\sum_{i=1}^{n}\mathrm{tr}\left(\mathbf{X}_i\mathbf{\Delta}_2\mathbf{X}_i^{\mathrm{T}}\mathbf{\Delta}_1\right),\\
T_4 &= p^{-1}\mathrm{vec}^{\mathrm{T}}(\mathbf{\Delta}_1)\left\{\int_0^1 h\left(v, \mathbf{\Omega}_v^{(1)}\right)(1-v)dv\right\}\mathrm{vec}(\mathbf{\Delta}_1),\\
T_5 &= q^{-1}\mathrm{vec}^{\mathrm{T}}(\mathbf{\Delta}_2)\left\{\int_0^1 h\left(v, \mathbf{\Omega}_v^{(2)}\right)(1-v)dv\right\}\mathrm{vec}(\mathbf{\Delta}_2),\\
T_6 &= \sum_{(i,j)\in S_1^c}\left\{p_{\lambda_1}(|\omega_{ij}|) - p_{\lambda_1}(|\omega_{0ij}|)\right\}\\
&\quad + \sum_{(i,j)\in S_2^c}\left\{p_{\lambda_2}(|\gamma_{ij}|) - p_{\lambda_1}(|\gamma_{0ij}|)\right\}, \text{ and}\\
T_7 &= \sum_{(i,j)\in S_1, i\neq j}\left\{p_{\lambda_1}(|\omega_{ij}|) - p_{\lambda_1}(|\omega_{0ij}|)\right\}\\
&\quad + \sum_{(i,j)\in S_2, i\neq j}\left\{p_{\lambda_2}(|\gamma_{ij}|) - p_{\lambda_1}(|\gamma_{0ij}|)\right\}.
\end{aligned}$$

We have the following decomposition from the definition (3):

$$g(\mathbf{\Omega}, \mathbf{\Gamma}) - g(\mathbf{\Omega}_0, \mathbf{\Gamma}_0) = T_1 + T_2 + T_3 + T_4 + T_5 + T_6 + T_7. \quad \text{(A.1)}$$

Let $\alpha_1 = \{s_1\log p/(nq)\}^{1/2}$, $\beta_1 = \{p\log p/(nq)\}^{1/2}$, $\alpha_2 = \{s_2\log q/(np)\}^{1/2}$, and $\beta_2 = \{q\log q/(np)\}^{1/2}$. For all m-dimensional diagonal matrix $\mathbf{D}_m$ and symmetric matrix $\mathbf{R}_m$ whose diagonal components are zero such that $\|\mathbf{R}_m\|_F = C_1$ and $\|\mathbf{D}_m\|_F = C_2$ for constants C_1 and C_2, we define $\mathcal{A} = \{\mathbf{M} : \mathbf{M} = \alpha_1\mathbf{R}_p + \beta_1\mathbf{D}_p\}$ and $\mathcal{B} = \{\mathbf{M} : \mathbf{M} = \alpha_2\mathbf{R}_q + \beta_2\mathbf{D}_q\}$, and we need to show that

$$P\left[\inf_{\mathbf{\Delta}_1\in\mathcal{A}, \mathbf{\Delta}_2\in\mathcal{B}}\{g(\mathbf{\Omega}_0 + \mathbf{\Delta}_1, \mathbf{\Gamma}_0 + \mathbf{\Delta}_2) - g(\mathbf{\Omega}_0, \mathbf{\Gamma}_0)\} > 0\right] \to 1.$$

First, following Rothman et al. (2008) and Lam and Fan (2009), we have

$$\begin{aligned}
T_4 &\geq p^{-1}\|\mathrm{vec}(\mathbf{\Delta}_1)\|^2\int_0^1(1-v)\min_{0<v<1}\lambda_1\left(\mathbf{\Omega}_v^{-1}\otimes\mathbf{\Omega}_v^{-1}\right)dv\\
&\geq (2p)^{-1}\left\{\tau_1^{-1} + o(1)\right\}^{-2}\left(C_1^2\alpha_1^2 + C_2^2\beta_1^2\right) \text{ and, similarly,}\\
T_5 &\geq (2q)^{-1}\left\{\tau_1^{-1} + o(1)\right\}^{-2}\left(C_1^2\alpha_2^2 + C_2^2\beta_2^2\right).
\end{aligned}$$

Then, we consider T_1 and T_2,

$$\begin{aligned}
(npq)^{-1}\sum_{i=1}^{n}\mathrm{tr}\left(\mathbf{X}_i\mathbf{\Gamma}_0\mathbf{X}_i^{\mathrm{T}}\mathbf{\Delta}_1\right) &= (npq)^{-1}\sum_{i=1}^{n}\mathrm{tr}\left(\mathbf{Y}_i\mathbf{Y}_i^{\mathrm{T}}\mathbf{\Delta}_1\right)\\
&= p^{-1}\mathrm{tr}(\mathbf{Q}_1\mathbf{\Delta}_1),
\end{aligned}$$

where $\mathbf{Q}_1 = (nq)^{-1}\sum_{i=1}^{n}\mathbf{Y}_i\mathbf{Y}_i^{\mathrm{T}}$, and $\mathbf{Y}_i = \mathbf{X}_i\mathbf{\Gamma}_0^{1/2}$ follows matrix-variate normal distribution with parameters $\mathbf{\Sigma}_0$ and $\mathbf{I}$ by the properties of matrix-variate normal distribution (Gupta and Nagar 2000). In other words, the columns of $\mathbf{Y}_i$ are iid normal random vectors with covariance matrix $\mathbf{\Sigma}_0$. Therefore,

$$\begin{aligned}
T_1 &= p^{-1}\mathrm{tr}\{(\mathbf{Q}_1 - \mathbf{\Sigma}_0)\mathbf{\Delta}_1\}\\
&= p^{-1}\left\{\sum_{(i,j)\in S_1} + \sum_{(i,j)\in S_1^c}\right\}(\mathbf{Q}_1 - \mathbf{\Sigma}_0)_{ij}(\mathbf{\Delta}_1)_{ij} = T_{11} + T_{12},
\end{aligned}$$

where T_{11} and T_{12} are, respectively, the two sums over S_1 and S_1^c. Since $\max_{i,j}|(\mathbf{Q}_1 - \mathbf{\Sigma}_0)_{ij}| = O_p[\{\log p/(nq)\}^{1/2}]$ by Bickel and Levina (2008a),

$$\begin{aligned}
|T_{11}| &\leq p^{-1}(s_1 + p)^{1/2}\|\mathbf{\Delta}_1\|_F\max_{i,j}|(\mathbf{Q}_1 - \mathbf{\Sigma}_0)_{ij}|\\
&= p^{-1}O_p(\alpha_1 + \beta_1)\|\mathbf{\Delta}_1\|_F.
\end{aligned}$$

Hence, by choosing C_1 and C_2 sufficiently large, T_{11} is dominated by the positive term T_4.

By Condition A2, $p_{\lambda_1}(|\mathbf{\Delta}_{ij}|) \geq \lambda_1 k_1|\mathbf{\Delta}_{ij}|$ for some constant $k_1 > 0$ when n is sufficiently large. Therefore,

$$\begin{aligned}
&\sum_{(i,j)\in S_1^c}\left\{p_{\lambda_1}(|\mathbf{\Delta}_{ij}|) - p^{-1}(\mathbf{Q}_1 - \mathbf{\Sigma}_0)_{ij}(\mathbf{\Delta}_1)_{ij}\right\}\\
&\geq \sum_{(i,j)\in S_1^c}\left\{\lambda_1 k_1 - p^{-1}\max_{i,j}|(\mathbf{Q}_1 - \mathbf{\Sigma}_0)_{ij}|\right\}|\mathbf{\Delta}_{1ij}|\\
&= \lambda_1\sum_{(i,j)\in S_1^c}\{k_1 - o_p(1)\}|\mathbf{\Delta}_{1ij}|,
\end{aligned}$$

which is greater than 0 with probability tending to 1, and the last equality is due to Condition A4 so that $\lambda_1^{-1}p^{-1}\max_{i,j}|(\mathbf{Q}_1 - \mathbf{\Sigma}_0)_{ij}|\}| = O_p\{\lambda_1^{-1}p^{-1}\sqrt{\log p/(nq)}\} = o_p(1)$. Applying exactly the same arguments on T_2, we have shown that $T_1 + T_2 + T_4 + T_5 + T_6 > 0$ with probability tending to 1. If the LASSO penalty is applied, Condition A4 for λ_1 and λ_2 can be relaxed to

$$\lambda_1 = O(p^{-1}\sqrt{\log p/(nq)}) \text{ and } \lambda_2 = O(q^{-1}\sqrt{\log q/(np)}).$$

As for T_7, applying Taylor's expansion,

$$\begin{aligned}
p_{\lambda_1}(|\omega_{ij}|) &= p_{\lambda_1}(|\omega_{0ij}|) + p'_{\lambda_1}(|\omega_{0ij}|)\mathrm{sgn}(\omega_{0ij})(\omega_{ij} - \omega_{0ij})\\
&\quad + 2^{-1}p''_{\lambda_1}(|\omega_{0ij}|)(\omega_{ij} - \omega_{0ij})^2\{1 + o(1)\}.
\end{aligned}$$

Noting Condition A3, we have

$$\begin{aligned}
&\left|\sum_{(i,j)\in S_1, i\neq j}\left\{p_{\lambda_1}(|\omega_{ij}|) - p_{\lambda_1}(|\omega_{0ij}|)\right\}\right|\\
&\leq \sqrt{s_1}C_1\alpha_1\max_{(i,j)\in S_1}p'_{\lambda_1}(|\omega_{0ij}|)\\
&\quad + C_1\max_{(i,j)\in S_1}p''_{\lambda_1}(|\omega_{0ij}|)\alpha_1^2/2\{1 + o(1)\}\\
&= O\left\{p^{-1}\left(\alpha_1^2 + \beta_1^2\right)\right\}.
\end{aligned}$$

Hence, it is dominated by the positive term T_4. Same arguments apply for the penalty on $\mathbf{\Gamma}$. Therefore, T_7 is dominated by $T_4 + T_5$.

Finally, note that $E\{\mathrm{tr}(\mathbf{X}_i\mathbf{\Delta}_2\mathbf{X}_i^{\mathrm{T}}\mathbf{\Delta}_1)\} = \mathrm{tr}(\mathbf{\Delta}_2\Psi_0)\mathrm{tr}(\mathbf{\Sigma}_0\mathbf{\Delta}_1)$ by the properties of matrix-variate normal distribution (Gupta and Nagar 2000). Hence, by law of large numbers,

$$\begin{aligned}
T_3 &= (npq)^{-1}\sum_{i=1}^{n}\mathrm{tr}\left(\mathbf{X}_i\mathbf{\Delta}_2\mathbf{X}_i^{\mathrm{T}}\mathbf{\Delta}_1\right)\\
&= (pq)^{-1}\mathrm{tr}(\mathbf{\Delta}_2\Psi_0)\mathrm{tr}(\mathbf{\Sigma}_0\mathbf{\Delta}_2)\{1 + o_p(1)\}.
\end{aligned}$$

Let ξ_{1j} and ξ_{2k}, $j = 1, \ldots, p, k = 1, \ldots, q$, be the eigenvalues of $\boldsymbol{\Delta}_1$ and $\boldsymbol{\Delta}_2$. Then by the von Neumann inequality,

$$\begin{aligned}|\mathrm{tr}(\boldsymbol{\Sigma}_0\boldsymbol{\Delta}_1)| &\le \sum_{j=1}^{p}\lambda_j(\boldsymbol{\Sigma}_0)|\xi_{1j}| \\ &\le (\tau_1)^{-1}\sum_{j=1}^{p}|\xi_{1j}| \le \tau_1^{-1}p^{1/2}\|\Delta_1\|_F \\ &\le \tau_1^{-1}p^{1/2}\left(C_1^2\alpha_1^2 + C_2^2\beta_1^2\right)^{1/2}.\end{aligned}$$

By applying the argument on $|\mathrm{tr}(\boldsymbol{\Psi}_0)\boldsymbol{\Delta}_2|$, we have as $n \to \infty$,

$$|T_3| \le 1/\sqrt{pq}\,\tau_1^{-2}\left(C_1^2\alpha_1^2 + C_2^2\beta_1^2\right)\left(C_1^2\alpha_2^2 + C_2^2\beta_2^2\right) \le T_4 + T_5.$$

In summary, we conclude that $T_1 + T_2 + T_3 + T_4 + T_5 + T_6 + T_7 > 0$ with probability tending to 1 and this proves Theorem 1.

Proof of Theorem 2. For the minimizers $\boldsymbol{\Omega}$ and $\boldsymbol{\Gamma}$, we have

$$\frac{\partial g(\boldsymbol{\Omega}, \boldsymbol{\Gamma})}{\partial \omega_{ij}} = 2(t_{ij} - p^{-1}\sigma_{ij} + p'_{\lambda_1}(|\omega_{ij}|))\mathrm{sgn}(\omega_{ij}),$$

where t_{ij} is the (i, j)th element of

$$(npq)^{-1}\sum_{i=1}^{n}\left(\mathbf{X}_i\boldsymbol{\Gamma}\mathbf{X}_i^{\mathrm{T}}\right) = p^{-1}\mathbf{Q}_1 + (npq)^{-1}\sum_{i=1}^{n}\mathbf{X}_i(\boldsymbol{\Gamma} - \boldsymbol{\Gamma}_0)\mathbf{X}_i^{\mathrm{T}}.$$

It is seen that the (i, j)th element in $(npq)^{-1}\sum_{i=1}^{n}\mathbf{X}_i(\boldsymbol{\Gamma} - \boldsymbol{\Gamma}_0)\mathbf{X}_i^{\mathrm{T}}$ is $O_p\{p^{-1}\eta_{2n}^{1/2}\}$ following Lam and Fan (2009). In addition, $p^{-1}\max_{ij}|q_{ij} - \sigma_{0ij}| = O_p\{p^{-1}\sqrt{\log p/(nq)}\}$ by Bickel and Levina (2008a). Further, following Lam and Fan (2009), $p^{-1}|\sigma_{0ij} - \sigma_{ij}| = O(p^{-1}\eta_{1n}^{1/2})$. Therefore, we need $\log p/(nq) + \eta_{1n} + \eta_{2n} = O(\lambda_1^2 p^2)$ to ensure that $\mathrm{sgn}(\omega_{ij})$ dominates the first derivative of $g(\boldsymbol{\Omega}, \boldsymbol{\Gamma})$, and by the same arguments for the minimizer $\boldsymbol{\Gamma}$. Theorem 2 then follows.

Proof of Theorem 3. Since $\hat{\boldsymbol{\Omega}}$ and $\hat{\boldsymbol{\Gamma}}$ solves $\mathbf{0} = \{\partial g(\boldsymbol{\Omega}, \boldsymbol{\Gamma}))/\partial^{\mathrm{T}}\mathrm{vec}(\boldsymbol{\Omega}), \partial g(\boldsymbol{\Omega}, \boldsymbol{\Gamma}))/\partial^{\mathrm{T}}\mathrm{vec}(\boldsymbol{\Gamma})\}^{\mathrm{T}}$, for the local minimizer $\hat{\boldsymbol{\Omega}}$ and $\hat{\boldsymbol{\Gamma}}$ in Theorem 1, we use Taylor's expansion at the truth. Let $\mathbf{A}^{\otimes 2} = \mathbf{A} \otimes \mathbf{A}$. Standard calculations imply

$$\begin{aligned}&\left\{\begin{pmatrix}\boldsymbol{\Sigma}_0^{\otimes 2} & n^{-1}q^{-1}\sum_{i=1}^{n}\mathbf{X}_i^{\otimes 2} \\ n^{-1}p^{-1}\sum_{i=1}^{n}\mathbf{X}_i^{\mathrm{T}\otimes 2} & \boldsymbol{\Psi}_0^{\otimes 2}\end{pmatrix} + \boldsymbol{\Lambda}_n \right. \\ &\left. + \begin{pmatrix}\mathbf{R}_{1n} \\ \mathbf{R}_{2n}\end{pmatrix}\right\}_{S\times S}\begin{pmatrix}\mathrm{vec}(\hat{\boldsymbol{\Omega}}) - \mathrm{vec}(\boldsymbol{\Omega}_0) \\ \mathrm{vec}(\hat{\boldsymbol{\Gamma}}) - \mathrm{vec}(\boldsymbol{\Gamma}_0)\end{pmatrix}_S \\ &= n^{-1}\begin{pmatrix}q^{-1}\sum_{i=1}^{n}\mathrm{vec}(\mathbf{X}_i\boldsymbol{\Gamma}_0\mathbf{X}_i^{\mathrm{T}}) - \mathrm{vec}(\boldsymbol{\Sigma}_0) \\ p^{-1}\sum_{i=1}^{n}\mathrm{vec}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\Omega}_0\mathbf{X}_i) - \mathrm{vec}(\boldsymbol{\Psi}_0)\end{pmatrix}_S + (\mathbf{b}_n)_S,\end{aligned} \tag{A.2}$$

where the remainder terms satisfy $\|\mathbf{R}_{1n}\| = O_p(\alpha_1 + \beta_1)$ and $\|\mathbf{R}_{2n}\| = O_p(\alpha_2 + \beta_2)$ following the proof of Theorem 1, the subscripts S and $S \times S$ indicate the subvector and submatrix corresponding to those nonzero components of $\boldsymbol{\Omega}_0$ and $\boldsymbol{\Gamma}_0$, excluding ω_{11}, $\boldsymbol{\Lambda}_n = \mathrm{diag}\{p''^{\mathrm{T}}_{\lambda_1}(\hat{\boldsymbol{\Omega}}), p''^{\mathrm{T}}_{\lambda_2}(\hat{\boldsymbol{\Omega}})\}$, and $\mathbf{b}_n = \{p'^{\mathrm{T}}_{\lambda_1}(|\boldsymbol{\Omega}_0|)\mathrm{sgn}(\boldsymbol{\Omega}_0), p'^{\mathrm{T}}_{\lambda_1}(|\boldsymbol{\Omega}_0|)\mathrm{sgn}(\boldsymbol{\Omega}_0)\}^{\mathrm{T}}$. The follow facts from the results in Gupta and Nagar (2000) imply the moments of the right-hand side in (A.2). First note that $E\{\mathrm{vec}(\mathbf{X}_i\boldsymbol{\Gamma}_0\mathbf{X}_i^{\mathrm{T}})\} = q\boldsymbol{\Sigma}_0$ and $E\{\mathrm{vec}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\Omega}_0\mathbf{X}_i)\} = p\boldsymbol{\Psi}_0$. Let $\mathbf{S}_1 = \mathbf{X}_i\boldsymbol{\Gamma}_0\mathbf{X}_i^{\mathrm{T}}$ and $\mathbf{S}_2 = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\Omega}_0\mathbf{X}_i$, then

$$\begin{aligned}\mathrm{cov}\{\mathrm{vec}(\mathbf{S}_1), \mathrm{vec}(\mathbf{S}_1)\} &= q\boldsymbol{\Sigma}_0^{\otimes 2}(\mathbf{I} + \mathbf{K}_{pp}), \\ \mathrm{cov}\{\mathrm{vec}(\mathbf{S}_2), \mathrm{vec}(\mathbf{S}_2)\} &= p\Psi_0^{\otimes 2}(\mathbf{I} + \mathbf{K}_{qq}),\end{aligned}$$

where $\mathbf{K}_{ab}$ is a commutation matrix of order $ab \times ab$; see Gupta and Nagar (2000) for its definition and properties. In addition, for any $p \times q$ matrix $\mathbf{C}$, $E(\mathbf{S}_1\mathbf{C}\mathbf{S}_2) = \boldsymbol{\Sigma}_0\mathbf{C}\boldsymbol{\Psi}_0(2 + pq)$ and $E(\mathbf{S}_1)\mathbf{C}E(\mathbf{S}_2) = pq = \boldsymbol{\Sigma}_0\mathbf{C}\boldsymbol{\Psi}_0$. Then, for any $q \times p$ matrix $\mathbf{D}$, we have

$$\begin{aligned}&\mathrm{tr}[(\mathbf{C}^{\mathrm{T}} \otimes \mathbf{D})\mathrm{cov}\{\mathrm{vec}(\mathbf{S}_1), \mathrm{vec}(\mathbf{S}_2)\}] \\ &\quad = \mathrm{tr}((\mathbf{C}^{\mathrm{T}} \otimes \mathbf{D})[E\{\mathrm{vec}(\mathbf{S}_1)\mathrm{vec}^{\mathrm{T}}(\mathbf{S}_2)\} \\ &\qquad - E\{\mathrm{vec}(\mathbf{S}_1)\}E\{\mathrm{vec}^{\mathrm{T}}(\mathbf{S}_2)\}]) \\ &\quad = \mathrm{tr}[E\{\mathrm{vec}(\mathbf{D}\mathbf{S}_1\mathbf{C})\mathrm{vec}^{\mathrm{T}}(\mathbf{S}_2)\} - E\{\mathrm{vec}(\mathbf{D}\mathbf{S}_1\mathbf{C})\}E\{\mathrm{vec}^{\mathrm{T}}(\mathbf{S}_2)\}] \\ &\quad = \mathrm{tr}\{E(\mathbf{S}_1\mathbf{C}\mathbf{S}_2\mathbf{D}) - E(\mathbf{S}_1)\mathbf{C}E(\mathbf{S}_2)\mathbf{D}\} = 2\boldsymbol{\Sigma}_0\mathbf{C}\boldsymbol{\Psi}_0\mathbf{D} \\ &\quad = \mathrm{vec}^{\mathrm{T}}(\boldsymbol{\Psi}_0)(\mathbf{C}^{\mathrm{T}} \otimes \mathbf{D})\mathrm{vec}(\boldsymbol{\Sigma}_0) \\ &\quad = \mathrm{tr}\{(\mathbf{C}^{\mathrm{T}} \otimes \mathbf{D})\mathrm{vec}(\boldsymbol{\Psi}_0)\mathrm{vec}^{\mathrm{T}}(\boldsymbol{\Sigma}_0)\}.\end{aligned}$$

This implies $\mathrm{cov}\{\mathrm{vec}(\mathbf{S}_1), \mathrm{vec}(\mathbf{S}_2)\} = \mathrm{vec}(\boldsymbol{\Sigma}_0)\mathrm{vec}^{\mathrm{T}}(\boldsymbol{\Psi}_0)$. Furthermore,

$$\begin{aligned}E(\mathbf{X}_i \otimes \mathbf{X}_i) &= \mathrm{vec}(\boldsymbol{\Sigma}_0)\mathrm{vec}^{\mathrm{T}}(\boldsymbol{\Psi}_0), \\ E\left(\mathbf{X}_i^{\mathrm{T}} \otimes \mathbf{X}_i^{\mathrm{T}}\right) &= \mathrm{vec}(\boldsymbol{\Psi}_0)\mathrm{vec}^{\mathrm{T}}(\boldsymbol{\Sigma}_0).\end{aligned}$$

Since $p \to \infty$ and $q \to \infty$, $n^{-1}q^{-1}\sum_{i=1}^{n}\mathbf{X}_i^{\otimes 2}$ and $n^{-1}p^{-1}\sum_{i=1}^{n}\mathbf{X}_i^{\mathrm{T}\otimes 2}$ are negligible on the left-hand side of (A.2), so does the covariance between components on the right-hand side of (A.2). Then, following standard arguments in penalized likelihood (Lam and Fan 2009), we establish Theorem 3.

[Received July 2011. Revised May 2012.]

REFERENCES

Allen, G. I., and Tibshirani, R. (2010), "Transposable Regularized Covariance Models With an Application to Missing Data Imputation," *The Annals of Applied Statistics*, 4, 764–790. [1188,1190]

Banejee, O., El Ghaoui, L., and d'Aspremont, A. (2008), "Model Selection Through Sparse Maximum Likelihood Estimation," *Journal of Machine Learning Research*, 9, 485–516. [1187]

Bickel, P., and Levina, E. (2008a), "Regularized Estimation of Large Covariance Matrices," *The Annals of Statistics*, 36, 199–227. [1197,1198,1199]

——— (2008b), "Covariance Regularization by Thresholding," *The Annals of Statistics*, 36, 2577–2604. [1197]

Black, F., and Scholes, M. (1973), "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy*, 81, 673–654. [1196]

Dawid, A. P. (1981), "Some Matrix-Variate Distribution Theory: Notational Considerations and a Bayesian Application," *Biometrika*, 68, 265–274. [1188]

Duan, J. C., and Wei, J. (2009), "Systematic Risk and the Price Structure of Individual Equity Options," *Review of Financial Studies*, 22, 1981–2006. [1196]

Dutilleul, P. (1999), "The MLE Algorithm for the Matrix Normal Distribution," *Journal of Statistical Computation and Simulation*, 64, 105–123. [1188,1190,1191]

Fan, J., Feng, Y., and Wu, Y. (2009), "Network Exploration via the Adaptive LASSO and SCAD Penalties," *The Annals of Applied Statistics*, 3, 521–541. [1187]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [1189,1193]

Friedman, J., Hastie, T., and Tibshirani, R. (2008), "Sparse Inverse Covariance Estimation With the Graphical Lasso," *Biostatistics*, 9, 432–441. [1189,1190,1192]

Gorski, J., Pfeuffer, F., and Klamroth, K. (2007), "Biconvex Sets and Optimization With Biconvex Functions: A Survey and Extensions," *Mathematical Methods of Operations Research*, 66, 373–408. [1190]

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2010), "Estimating Heterogeneous Graphical Models for Discrete Data," Technical Report. [1187]

——— (2011), "Joint Estimation of Multiple Graphical Models," *Biometrika*, 98, 1–15. [1187]

Gupta, A. K., and Nagar, D. K. (2000), *Matrix Variate Distributions*, Boca Raton, FL: Chapman and Hall/CRC Press. [1188,1191,1198,1199]

Hoff, P. D. (2011), "Separable Covariance Arrays via the Tucker Product, With Applications to Multivariate Relational Data," *Bayesian Analysis*, 6, 179–196. [1197]

Huang, J. Z., Liu, N., Pourahmadi, M., and Liu, L. (2006), "Covariance Selection and Estimation via Penalised Normal Likelihood," *Biometrika*, 93, 85–98. [1197]

Lam, C., and Fan, J. (2009), "Sparsistency and Rates of Convergence in Large Covariance Matrices Estimation," *The Annals of Statistics*, 37, 4254–4278. [1187,1189,1190,1191,1198,1199]

Lauritzen, S. L. (1996), *Graphical Models*, Oxford: Clarendon Press. [1187,1188]

Meinshausen, N., and Bühlmann, P. (2006), "High-Dimensional Graphs and Variable Selection With the Lasso," *The Annals of Statistics*, 34, 1436–1462. [1187]

Merton, R. C. (1990), *Continuous-Time Finance*, Cambridge, MA: Basil Blackwell. [1196]

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009), "Partial Correlation Estimation by Joint Sparse Regression Model," *Journal of the American Statistical Association*, 104, 735–746. [1187]

Pons, P., and Latapy, M. (2005), "Computing Communities in Large Networks Using Random Walks," *Journal of Graph Algorithms and Applications* [online], 10, 191–218. Available at *http://arxiv.org/abs/physics/0512106*. [1197]

Pourahmadi, M. (1999), "Joint Mean-Covariance Models With Applications to Longitudinal Data: Unconstrained Parameterisation," *Biometrika*, 86, 677–690. [1197]

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008), "Sparse Permutation Invariant Covariance Estimation," *Electronic Journal of Statistics*, 2, 494–515. [1187,1192,1198]

Rothman, A. J., Levina, E., and Zhu, J. (2010), "Sparse Multivariate Regression With Covariance Estimation," *Journal of Computational and Graphical Statistics*, 19, 947–962. [1197]

Srivastava, M. S., von Rosen, T., and Von Rosen, D. (2008), "Models With a Kronecker Product Covariance Structure: Estimation and Testing," *Mathematical Methods of Statistics*, 17, 357–370. [1190]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society,* Series B, 58, 267–288. [1189]

Tsai, H., and Tsay, R. S. (2010), "Constrained Factor Models," *Journal of the American Statistical Association*, 105, 1593–1605. [1197]

Wang, H., and West, M. (2009), "Bayesian Analysis of Matrix Normal Graphical Models," *Biometrika*, 96, 821–834. [1188]

Yuan, M., and Lin, Y. (2007), "Model Selection and Estimation in the Gaussian Graphical Model," *Biometrika*, 94, 19–35. [1187]

Zou, H., and Li, R. (2008), "One-Step Sparse Estimates in Nonconcave Penalized Likelihood Models" (with discussion), *The Annals of Statistics*, 36, 1509–1533. [1189]