Sparse Matrix Graphical Models

Chenlei Leng and Cheng Yong Tang *

June 1, 2012

Abstract

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 paper, we propose a novel sparse matrix graphical model for this type 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 rate 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: Graphical models; Matrix graphical models; Matrix-variate normal distribution; Penalized likelihood; Sparsistency; Sparsity.

*Leng and Tang are with Department of Statistics and Applied Probability, National University of Singapore. Research supports from National University of Singapore research grants are gratefully acknowledged. Corresponding author: Chenlei Leng (Email: stalc@nus.edu.sg). We thank co-editor Prof. Xuming He, an associate Editor and two anonymous referees for their constructive comments that have led to a much improved paper.
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, macro-economics 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 100 index for 8 respective expiration dates, such as 30, 60 days. In this example, each observation of the dataset can be denoted by a $89 \times 8$ 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 5.3, for three selected companies Abbott Laboratories, ConocoPhillips and Microsoft is presented in Figure 1.

- US Department of Agriculture reports itemized annual export to major trading partners. A dataset with 40 years US export is collected for 13 trading partners and 36 items. Each observation in the dataset can be denoted by a $13 \times 36$ 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.

![Graphs showing volatility data for Abbott Laboratories (ABT), ConocoPhillips (COP), and Microsoft (MSFT).](image)

**Figure 1:** Volatility data for 142 weeks of trading.

We refer to this type of structured data $X \in \mathbb{R}^{p \times q}$ as matrix-variate data or simply matrix data if no confusion arises. It may appear attempting to stack $X$ as a column vector $\text{vec}(X)$ and model $X$ 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. 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 $\text{vec}(\mathbf{X})$, there exists abundant literature that makes use of penalized
likelihood (Meishausen and Bühlmann, 2006; Yuan and Lin, 2006; Rothman, et al., 2008;
Banejee et al., 2008; Lam and Fan, 2009; Fan et al., 2009; Peng et al., 2009; Guo et al., 2011;
Guo et al., 2010). 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 $\text{vec}(\mathbf{X})$ 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 data sets 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

$$p(\mathbf{X}|\mathbf{M}, \mathbf{\Sigma}, \mathbf{\Psi}) = (2\pi)^{-\frac{qp}{2}}|\mathbf{\Sigma}^{-1}|^{-\frac{q^2}{4}}|\mathbf{\Psi}^{-1}|^{-\frac{p^2}{4}}\text{etr}\{-\frac{1}{2}(\mathbf{X} - \mathbf{M})\mathbf{\Psi}^{-1}(\mathbf{X} - \mathbf{M})^T\mathbf{\Sigma}^{-1}/2\},$$

(1)

where $\mathbf{M} \in \mathbb{R}^{p \times q}$ is the mean matrix; $\mathbf{\Sigma} \in \mathbb{R}^{p \times p}$ and $\mathbf{\Psi} \in \mathbb{R}^{q \times q}$ are the row and col-
umn variance matrices; etr is the exponential of the trace operator. The matrix normal
distribution (1) implies that $\text{vec}(\mathbf{X})$ follows a vector multivariate normal distribution
$N_{pq}(\text{vec}(\mathbf{M}), \mathbf{\Psi} \otimes \mathbf{\Sigma})$, where $\otimes$ is the Kronecker product. Models using matrix instead of
vector normal distributions effectively reduce the dimensionality of the covariance or pre-
cision 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
$\hat{\mathbf{M}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i$ from $\mathbf{X}_i$, where $\mathbf{X}_i$, $i = 1, ..., n$ are assumed to be independent and identi-
cally distributed following the matrix normal distribution (1). Dawid (1981) provided some
theory for matrix-variate distributions and Dutilleul (1999) derived the maximum likeli-
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 \( \Omega = \Sigma^{-1} = (\omega_{ij}) \) and \( \Gamma = \Psi^{-1} = (\gamma_{ij}) \) 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 \( \Gamma \otimes \Omega \) define pairwise conditional independence of corresponding entries (Lauritzen, 1996), given all the other variables. More importantly, zeros in \( \Omega \) and \( \Gamma \) encode conditional independence of the row variables and the column variables in \( X \) 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 \( \omega_{ij} \) and \( \gamma_{kl} \) when \( i \neq k, \ j \neq l \); (2) \( \omega_{ik} = 0 \) when \( i \neq k, \ j = l \); (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

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

\( \rho_{ij,kl} \) is not zero only if both \( \omega_{ik} \) and \( \gamma_{jl} \) are nonzero. If either the \( i \)th and the \( k \)th rows are conditionally independent give other rows, or the \( j \)th and the \( l \)th 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 parametrization of the matrix normal distribution is its identifiability. In later development, we fix \( \omega_{11} = 1 \).

In this paper, we propose sparse matrix graphical models by considering regularized maximum likelihood estimate for matrix data that follow the distribution in (1). By applying appropriate penalty functions on precision matrices \( \Omega \) and \( \Gamma \), we obtain sparse matrix graphical models for the row and the 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 $\Omega$ and $\Gamma$ given iid data $X_1, ..., X_n$ from (1), the standard approach is the maximum likelihood estimation. It is easily seen that the minus log-likelihood up to a constant is

$$
\ell(\Omega, \Gamma) = -\frac{nq}{2} \log |\Omega| - \frac{np}{2} \log |\Gamma| + \frac{1}{2} \sum_{i=1}^{n} \text{tr}(X_i \Gamma X_i^T \Omega).
$$

(2)

The solution minimizing (2) is the familiar maximum likelihood estimate. In order to build sparse models for $\Omega$ and $\Gamma$, we propose the penalized likelihood estimator as the minimizer of

$$
g(\Omega, \Gamma) = \frac{1}{npq} \sum_{i=1}^{n} \text{tr}(X_i \Gamma X_i^T \Omega) - \frac{1}{p} \log |\Omega| - \frac{1}{q} \log |\Gamma| + p_\lambda(\Omega) + p_\lambda(\Gamma),
$$

(3)

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

$$
p_\lambda'(s) = \lambda \{I(s \leq \lambda) + \frac{(3.7\lambda - s)_+}{2.7\lambda} I(s > \lambda)\}
$$
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 (SMGM). When either $\Sigma$ or $\Psi$ 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 $X$ are independently normal distributed with covariance matrix $\Psi$ ($\Sigma$) when $\Sigma = I_p$ ($\Psi = I_q$). 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 the 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 which optimizes one matrix with the other matrix fixed. When $\Omega$ is fixed, we minimize with respect to $\Gamma$

$$
\frac{1}{q} \text{tr}(\Gamma \hat{\Sigma}) - \frac{1}{q} \log |\Gamma| + p\lambda_2(\Gamma),
$$

(4)

where $\hat{\Sigma} = \sum_{i=1}^n X_i^T \Omega X_i / q$. When the LASSO penalty is used, we use the coordinate descent graphical LASSO (gLASSO) algorithm in Friedman et al. (2008) to obtain the solution. When used with warm start for different tuning parameters $\lambda_2$, it can solve large problems very efficiently even when $q \gg n$. For the SCAD penalty, we follow Zou and Li
(2008) to linearized 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
\[
\frac{1}{q} \text{tr}(\Gamma \tilde{\Sigma}) - \frac{1}{q} \log |\Gamma| + \sum_{i \neq j} p'_{\lambda_2}(|\gamma_{ij,k}|)|\gamma_{ij}| \]
which can be solved by using the graphical LASSO algorithm. Here \( \gamma_{ij,k} \) is the \( k \)th step estimate of \( \gamma_{ij} \). 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 graphical LASSO estimates by taking \( p'_{\lambda_2}(|\gamma_{ij,k}|) \) as \( \lambda_2 \). On the other hand, when \( \Gamma \) is fixed, we minimize with respect to \( \Omega \)
\[
\frac{1}{p} \text{tr}(\tilde{\Psi} \Omega) - \frac{1}{p} \log |\Omega| + p_{\lambda_1}(\Omega), \tag{5}
\]
where \( \tilde{\Psi} = \sum_{i=1}^{n} X_i \Gamma X_i^T/p. \)

Synthetically, the computational algorithm is summarized as follows.

1. Start with \( \Gamma^{(0)} = I_q \), and minimize (4) to get \( \Omega^{(0)} \). Normalize \( \Omega^{(0)} \) such that \( \omega_{11}^{(0)} = 1 \). Let \( m = 1; \)
2. Fix \( \Omega^{(m-1)} \) and minimize (5) to get \( \Gamma^{(m)}; \)
3. Fix \( \Gamma^{(m)} \) and minimize (4) to get \( \Omega^{(m)} \). Normalize such that \( \omega_{11}^{(m)} = 1 \). Let \( m \leftarrow m+1; \)
4. Repeat Step 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 \( g(\Omega, \Gamma) \). An argument has been outlined in 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 et al. (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 $\Omega$, the effective sample size is essentially $nq$. More recently, Srivastava et al. (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 $\lambda_1$ and $\lambda_2$ 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 $\rho_L > \rho_{L-1} > \ldots > \rho_1$, where $\rho_L$ is chosen to yield very sparse models. The solution at $\rho_i$ is then used as the initial value for the solution at $\rho_{i-1}$ (Friedman et al., 2008). Note that if $\rho_L$ is large, the resulting estimated $\Omega$ and $\Gamma$ would be very sparse. Thus, the algorithm is not very sensitive to the initial value for the solution at $\rho_L$ because in this case, a much fewer number of parameters in the two precision matrices need to be estimated. Subsequently, for other values of $\rho$, 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 $\Omega_0$ and $\Gamma_0$ 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 $\Omega_0$ and $\Gamma_0$ respectively. We make the following standard regularity assumptions for theoretical analysis of the penalized likelihood.
A1. There exist constant $\tau_1$ such that for all $n$, 
\[ 0 < \tau_1 < \lambda_1(\Sigma_0) \leq \lambda_p(\Sigma_0) < 1/\tau_1 < \infty, \quad 0 < \tau_1 < \lambda_1(\Psi_0) \leq \lambda_q(\Psi_0) < 1/\tau_1 < \infty, \]
where $\lambda_1(A) \leq \lambda_2(A) \leq \cdots \leq \lambda_m(A)$ denote the eigenvalues of an $m$-dimensional symmetric matrix $A$.

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 $\Omega_0$ and $\Gamma_0$,
\[
\max_{(i,j) \in S_1} p'_{\lambda_1}(|\omega_{0ij}|) = O[p^{-1}(1 + p/(s_1 + 1))\sqrt{\log p/(nq)}], \quad \max_{(i,j) \in S_1} p''_{\lambda_1}(|\omega_{0ij}|) = o(p^{-1}),
\]
\[
\max_{(i,j) \in S_2} p'_{\lambda_2}(|\gamma_{0ij}|) = O[q^{-1}(1 + q/(s_2 + 1))\sqrt{\log q/(np)}], \quad \max_{(i,j) \in S_2} p''_{\lambda_2}(|\gamma_{0ij}|) = o(q^{-1}),
\]
\[
\min_{(i,j) \in S_1} |\omega_{0ij}|/\lambda_1 \rightarrow \infty \quad \text{and} \quad \min_{(i,j) \in S_2} |\gamma_{0ij}|/\lambda_2 \rightarrow \infty \quad \text{as} \quad n \rightarrow \infty.
\]

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

All conditions imposed here are mild and comparable to 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 $\lambda_1$ and $\lambda_2$ are allowed to converge to 0 at faster rates than those in Lam and Fan (2009). Thus smaller bias is induced due to this penalization.

We use the notations $\| \cdot \|_F$ 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 $n \rightarrow \infty$, $(p+s_1)\log p/(nq) \rightarrow 0$ and $(q+s_2)\log q/(np) \rightarrow 0$, there exists a local minimizer of (3) such that
\[
\|\hat{\Omega} - \Omega_0\|_F^2/p = O_p\{(1 + s_1/p)\log p/(nq)\} \quad \text{and} \quad \|\hat{\Gamma} - \Gamma_0\|_F^2/q = O_p\{(1 + s_2/q)\log q/(np)\}.
\]
Theorem 1 shows that the rate of convergence is determined by $n$ and $p$ for $\Omega$ and by $n$ and $q$ for $\Gamma$. 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 Appendix that Theorem 1 is also valid for the LASSO penalty if Condition of A4 is changed to 

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

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 $\text{vec}(X)$ by SMGM, the rate of convergence for estimating $\Omega_0 \otimes \Gamma_0$ using our approach is easily seen as

$$
\frac{|\hat{\Omega} \otimes \hat{\Gamma} - \Omega_0 \otimes \Gamma_0|^2}{pq} = O_p[\max\{(1 + \frac{s_1}{p}) \log p/(nq^2), (1 + \frac{s_2}{q}) \log q/(np^2)\}].
$$

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

$$
\frac{|\hat{\Omega} \otimes \hat{\Gamma} - \Omega_0 \otimes \Gamma_0|^2}{pq} = O_p[(1 + \frac{s_1s_2}{pq}) \log(pq)/n].
$$

We immediately see that when $p \to \infty$ and $q \to \infty$, SMGM gives estimates with 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 utilizing a more parsimonious model that naturally incorporates the data structure.

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

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

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 $p$. 

11
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 $\hat{\Omega}$ and $\hat{\Gamma}$ satisfying 
\[ \| \hat{\Omega} - \Omega_0 \|_F^2 = O_p\{ (p + s_1) \log p/(nq) \}, \| \hat{\Gamma} - \Gamma_0 \|_F^2 = O_p\{ (q + s_2) \log q/(np) \}, \]
\[ \| \Omega - \Omega_0 \| = O_p(\eta_1 n), \| \Gamma - \Gamma_0 \| = O_p(\eta_2 n) \] for sequences $\eta_{1n}$ and $\eta_{2n}$ 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 can not 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 seen from Theorem 2. Therefore to simultaneously achieve consistency and sparsistency when applying the LASSO penalty, the numbers of nonzero components in $\Omega_0$ and $\Gamma_0$ 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 $\Omega$ and $\Gamma$ 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 $\Psi_0$ is a
correlation matrix, one can apply SMGM with $\Gamma = I$ and obtain consistent and sparse estimate of $\Omega$ following the proof in Appendix. We conclude that by incorporating the structure information from matrix data, more efficient estimates of the graphic models can be obtained.

We now discuss the asymptotic properties of the estimates. Let $A \otimes 2 = A \otimes A$ and $K_{pq}$ be an $pq \times pq$ commutation matrix that transforms $\text{vec}(A)$ to $\text{vec}(A^T)$ for a $p \times q$ matrix $A$. A rigorous definition and its properties can be found for example in Gupta and Nagar (2000). Let $\Lambda_1^n = \text{diag}\{p''_{\lambda_1}(\Omega_0)\}$, $\Lambda_2^n = \text{diag}\{p''_{\lambda_2}(\Omega)\}$, $b_1^n = p'_{\lambda_1}(|\Omega_0|)\text{sgn}(\Omega_0)$, and $b_2^n = p'_{\lambda_2}(|\Omega_0|)\text{sgn}(\Omega_0)$. Denote $S_1$ as the set of all the indices of nonzero components in $\text{vec}(\Omega_0)$ except $\omega_{11}$, $S_2$ as the set for all the indices of nonzero components in $\text{vec}(\Gamma_0)$. We use the subscript $S$ and $S \times S$ for the corresponding sub-vector and sub-matrix respectively.

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

$$\sqrt{nq} \alpha_p^T \{\Sigma_0^{\otimes 2} (I + K_{pp})\}^{-1/2}_{S_1 \times S_1} (A_{1n} + \Sigma_0^{\otimes 2})_{S_1 \times S_1} \{\text{vec}(\hat{\Omega}) - \text{vec}(\Omega_0) + b_{1n}\}_{S_1} \overset{d}{\rightarrow} N(0, 1),$$

$$\sqrt{np} \alpha_q^T \{\Psi_0^{\otimes 2} (I + K_{qq})\}^{-1/2}_{S_2 \times S_2} (A_{2n} + \Psi_0^{\otimes 2})_{S_2 \times S_2} \{\text{vec}(\hat{\Gamma}) - \text{vec}(\Gamma_0) + b_{2n}\}_{S_2} \overset{d}{\rightarrow} N(0, 1),$$

where $\alpha_d$ denotes a $d$-dimensional unit vector and $\overset{d}{\rightarrow}$ 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 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 Theorem 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 maximum likelihood estimate (MLE) of $\Omega$ and $\Gamma$ using
the algorithm in Dutilleul (1999). Note that this algorithm is similar to the algorithm for computing the SMGM estimate, when the penalty function is absent. We also compare SMGM with the graphical LASSO method for estimating $\Omega_0 \otimes \Gamma_0$ (Friedman, et al. 2008). In particular, this approach ignores the matrix structure of the observations and vectorizes them as $x = \text{vec}(X)$. The graphical LASSO method then optimizes
\[
\text{tr}(\Theta S) - \log |\Theta| + \lambda \sum_{i \neq j} |\theta_{ij}|,
\]
where $S$ is the sample covariance matrix of $x_i$. In addition, we implement a ridge type regularized method by using squared matrix Frobenius norm $\|A\|_F^2$ 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 to $p \times q$. 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 $\Omega$ and $\Gamma$ (Rothman et al., 2008).

1. $A_1$: Inverse AR(1) such that $A_1 = B^{-1}$ with $b_{ij} = 0.7^{|i-j|}$.

2. $A_2$: 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. $A_3 = B + \delta I$: each off-diagonal upper triangle entry in $B$ is generated independently and equals to 0.5 with probability $a = 0.1$ and 0 with probability $1 - a = 0.9$. The diagonals of $B$ are zero and $\delta$ is chosen such that the condition number of $A_3$ is $d$.

All matrices are sparse and are numbered in order of decreasing sparsity. We assess the estimation accuracy for a precision matrix $A \in \mathbb{R}^{d \times d}$ using the Kullback-Leibler loss, defined as
\[
l(A, \hat{A}) = \text{tr}(A^{-1}\hat{A}) - \log |A^{-1}\hat{A}| - d.
\]
Here we use $A = \Omega_0 \otimes \Gamma_0$, the main parameter of interest in the Kullback-Leibler loss, and $\hat{A}$ is an estimate of it. We also summarize the performance in terms of true positive rate.
(TPR) and true negative rate (TNR), defined as

\[
\text{TPR} = \frac{\# \{ \hat{A}_{ij} \neq 0 & A_{ij} \neq 0 \}}{\# \{ A_{ij} \neq 0 \}}, \quad \text{TNR} = \frac{\# \{ \hat{A}_{ij} = 0 & A_{ij} = 0 \}}{\# \{ A_{ij} = 0 \}}.
\]

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), as well as the graphical LASSO vectorizing the matrix data (gLASSO). The results based on 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 \( p = q = 20 \). 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 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 improves 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. In order 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 graphical LASSO 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 and SCAD penalty are the two best approaches in terms of estimation accuracy. In addition, the SCAD perform much better than the LASSO penalty using the
<table>
<thead>
<tr>
<th>$n$</th>
<th>$p$</th>
<th>$q$</th>
<th>$\Omega$</th>
<th>$\Gamma$</th>
<th>MLE</th>
<th>Ridge</th>
<th>LASSO</th>
<th>SCAD</th>
<th>gLASSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>10</td>
<td>10</td>
<td>$A_1$</td>
<td>$A_1$</td>
<td>22.6(5.0)</td>
<td>11.8(2.3)</td>
<td>12.5(4.1)</td>
<td>6.9(2.4)</td>
<td>105.4(9.9)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$A_2$</td>
<td>$A_1$</td>
<td>22.9(4.2)</td>
<td>11.2(1.7)</td>
<td>11.7(2.3)</td>
<td>11.3(2.2)</td>
<td>77.8(5.6)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$A_2$</td>
<td>$A_2$</td>
<td>23.9(4.8)</td>
<td>10.4(1.4)</td>
<td>15.8(3.1)</td>
<td>17.0(3.5)</td>
<td>44.4(3.7)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$A_3$</td>
<td>$A_1$</td>
<td>22.5(3.9)</td>
<td>10.7(1.7)</td>
<td>8.1(2.0)</td>
<td>7.4(2.0)</td>
<td>84.7(6.4)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$A_3$</td>
<td>$A_2$</td>
<td>22.5(4.7)</td>
<td>10.0(1.5)</td>
<td>11.2(2.2)</td>
<td>10.6(2.2)</td>
<td>54.6(4.4)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$A_3$</td>
<td>$A_3$</td>
<td>23.7(4.5)</td>
<td>9.7(1.3)</td>
<td>7.6(1.8)</td>
<td>7.4(2.0)</td>
<td>57.4(5.2)</td>
</tr>
<tr>
<td>100</td>
<td>10</td>
<td>10</td>
<td>$A_1$</td>
<td>$A_1$</td>
<td>1.2(0.2)</td>
<td>1.4(0.2)</td>
<td>0.71(0.12)</td>
<td>0.38(0.09)</td>
<td>30.7(0.8)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$A_2$</td>
<td>$A_1$</td>
<td>1.2(0.2)</td>
<td>1.2(0.1)</td>
<td>1.0(0.1)</td>
<td>0.8(0.2)</td>
<td>17.2(0.5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$A_2$</td>
<td>$A_2$</td>
<td>1.2(0.1)</td>
<td>1.1(0.1)</td>
<td>1.1(0.1)</td>
<td>1.1(0.1)</td>
<td>15.9(0.5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$A_3$</td>
<td>$A_1$</td>
<td>1.2(0.2)</td>
<td>1.1(0.2)</td>
<td>0.68(0.11)</td>
<td>0.38(0.09)</td>
<td>18.0(0.5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$A_3$</td>
<td>$A_2$</td>
<td>1.2(0.2)</td>
<td>1.1(0.1)</td>
<td>1.0(0.1)</td>
<td>0.82(0.13)</td>
<td>13.9(0.5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$A_3$</td>
<td>$A_3$</td>
<td>1.2(0.1)</td>
<td>1.1(0.1)</td>
<td>0.66(0.11)</td>
<td>0.35(0.08)</td>
<td>8.4(0.4)</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>20</td>
<td>$A_1$</td>
<td>$A_1$</td>
<td>89.6(7.9)</td>
<td>45.6(3.5)</td>
<td>28.9(5.8)</td>
<td>13.9(2.6)</td>
<td>509(19)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$A_2$</td>
<td>$A_1$</td>
<td>87.6(8.1)</td>
<td>42.8(2.8)</td>
<td>33.5(3.1)</td>
<td>33.2(3.4)</td>
<td>356(12)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$A_2$</td>
<td>$A_2$</td>
<td>88.6(7.5)</td>
<td>40.3(2.5)</td>
<td>50.2(4.1)</td>
<td>50.6(4.3)</td>
<td>185(8)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$A_3$</td>
<td>$A_1$</td>
<td>88.0(8.0)</td>
<td>42.1(3.2)</td>
<td>21.4(3.3)</td>
<td>16.9(2.9)</td>
<td>410(17)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$A_3$</td>
<td>$A_2$</td>
<td>90.3(9.7)</td>
<td>39.0(2.8)</td>
<td>42.0(5.9)</td>
<td>37.0(5.8)</td>
<td>228(10)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$A_3$</td>
<td>$A_3$</td>
<td>88.5(9.9)</td>
<td>37.2(2.9)</td>
<td>20.8(3.3)</td>
<td>17.6(3.2)</td>
<td>273(11)</td>
</tr>
<tr>
<td>100</td>
<td>20</td>
<td>20</td>
<td>$A_1$</td>
<td>$A_1$</td>
<td>4.5(0.3)</td>
<td>5.5(0.3)</td>
<td>1.9(0.2)</td>
<td>0.9(0.2)</td>
<td>169(1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$A_2$</td>
<td>$A_1$</td>
<td>4.6(0.4)</td>
<td>4.5(0.3)</td>
<td>3.1(0.3)</td>
<td>2.0(0.3)</td>
<td>94.2(0.9)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$A_2$</td>
<td>$A_2$</td>
<td>4.4(0.3)</td>
<td>4.2(0.2)</td>
<td>4.0(0.3)</td>
<td>2.8(0.3)</td>
<td>83.6(1.3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$A_3$</td>
<td>$A_1$</td>
<td>4.6(0.3)</td>
<td>4.6(0.3)</td>
<td>2.1(0.2)</td>
<td>0.81(0.13)</td>
<td>110(1.3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$A_3$</td>
<td>$A_2$</td>
<td>4.5(0.3)</td>
<td>4.2(0.3)</td>
<td>3.3(0.3)</td>
<td>2.0(0.2)</td>
<td>72.1(1.2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$A_3$</td>
<td>$A_3$</td>
<td>4.6(0.4)</td>
<td>4.2(0.3)</td>
<td>1.5(0.2)</td>
<td>0.77(0.13)</td>
<td>67.1(1.2)</td>
</tr>
</tbody>
</table>

Table 1: Simulation results on the Kullback-Leibler loss. The sample standard errors are in parentheses.
Table 2: Simulation results on model selection. The sample standard errors are in parentheses.

<table>
<thead>
<tr>
<th>n</th>
<th>p</th>
<th>q</th>
<th>Ω</th>
<th>Γ</th>
<th>LASSO TPR</th>
<th>LASSO TNR</th>
<th>SCAD TPR</th>
<th>SCAD TNR</th>
<th>gLASSO TPR</th>
<th>gLASSO TNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>10</td>
<td>10</td>
<td>A₁</td>
<td>A₁</td>
<td>1(0) 0.82(0.03)</td>
<td>1(0) 0.95(0.06)</td>
<td>0.50(0.05)</td>
<td>0.92(0.01)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A₂</td>
<td>A₁</td>
<td>0.70(0.11) 0.76(0.07) 0.70(0.10)</td>
<td>0.90(0.11) 0.80(0.11)</td>
<td>0.18(0.02) 0.93(0.01)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A₂</td>
<td>A₂</td>
<td>0.53(0.16) 0.69(0.16) 0.43(0.16)</td>
<td>0.79(0.16) 0.50(0.02)</td>
<td>0.98(0.02)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A₃</td>
<td>A₁</td>
<td>1(0) 0.78(0.05) 1.0(0.01)</td>
<td>0.90(0.07) 0.44(0.04)</td>
<td>0.93(0.01)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A₃</td>
<td>A₂</td>
<td>0.75(0.11) 0.77(0.10) 0.72(0.11)</td>
<td>0.84(0.14) 0.14(0.02)</td>
<td>0.97(0.01)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A₃</td>
<td>A₃</td>
<td>1(0) 0.88(0.04) 1.0(0.01)</td>
<td>0.92(0.05) 0.52(0.04)</td>
<td>0.96(0.01)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>10</td>
<td>10</td>
<td>A₁</td>
<td>A₁</td>
<td>1(0) 0.66(0.08)</td>
<td>1(0) 0.99(0.02)</td>
<td>0.81(0.05)</td>
<td>0.66(0.03)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A₂</td>
<td>A₁</td>
<td>1.0(0.01) 0.38(0.11) 1.0(0.01)</td>
<td>0.74(0.07) 0.35(0.01)</td>
<td>0.83(0.01)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A₂</td>
<td>A₂</td>
<td>1.0(0.00) 0.16(0.06) 1.0(0.01)</td>
<td>0.46(0.11) 0.23(0.01)</td>
<td>0.87(0.01)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A₃</td>
<td>A₁</td>
<td>1(0) 0.77(0.08) 1(0)</td>
<td>0.99(0.01) 0.75(0.02)</td>
<td>0.81(0.02)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A₃</td>
<td>A₂</td>
<td>1(0) 0.38(0.09) 1(0)</td>
<td>0.70(0.07) 0.37(0.01)</td>
<td>0.87(0.01)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A₃</td>
<td>A₃</td>
<td>1(0) 0.78(0.07) 1(0)</td>
<td>0.95(0.01) 0.93(0.01)</td>
<td>0.91(0.01)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>20</td>
<td>A₁</td>
<td>A₁</td>
<td>1(0) 0.93(0.01)</td>
<td>1(0) 0.99(0.01)</td>
<td>0.38(0.03)</td>
<td>0.97(0.00)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A₂</td>
<td>A₁</td>
<td>0.72(0.07) 0.86(0.03) 0.67(0.11)</td>
<td>0.91(0.05) 0.10(0.01)</td>
<td>0.98(0.00)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A₂</td>
<td>A₂</td>
<td>0.51(0.11) 0.79(0.07) 0.47(0.11)</td>
<td>0.84(0.06) 0.02(0.00)</td>
<td>1.0(0.00)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A₃</td>
<td>A₁</td>
<td>1(0) 0.91(0.02) 1(0)</td>
<td>0.97(0.02) 0.31(0.02)</td>
<td>0.98(0.00)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A₃</td>
<td>A₂</td>
<td>0.90(0.06) 0.80(0.04) 0.87(0.06)</td>
<td>0.90(0.03) 0.07(0.01)</td>
<td>0.99(0.00)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A₃</td>
<td>A₃</td>
<td>1(0) 0.92(0.02) 1(0)</td>
<td>0.96(0.02) 0.26(0.02)</td>
<td>0.99(0.00)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>20</td>
<td>20</td>
<td>A₁</td>
<td>A₁</td>
<td>1(0) 0.83(0.02)</td>
<td>1(0) 1.00(0.00)</td>
<td>0.58(0.00)</td>
<td>0.93(0.00)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A₂</td>
<td>A₁</td>
<td>1.00(0.00) 0.60(0.04) 1(0)</td>
<td>0.89(0.02) 0.25(0.01)</td>
<td>0.94(0.00)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A₂</td>
<td>A₂</td>
<td>1(0) 0.36(0.03) 1(0)</td>
<td>0.71(0.03) 0.12(0.01)</td>
<td>0.95(0.00)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A₃</td>
<td>A₁</td>
<td>1(0) 0.78(0.03) 1(0)</td>
<td>1.00(0.00) 0.59(0.00)</td>
<td>0.93(0.00)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A₃</td>
<td>A₂</td>
<td>1.00(0.00) 0.49(0.05) 1.00(0.00)</td>
<td>0.84(0.02) 0.26(0.01)</td>
<td>0.96(0.00)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A₃</td>
<td>A₃</td>
<td>1(0) 0.90(0.03) 1(0)</td>
<td>1.00(0.00) 0.63(0.01)</td>
<td>0.95(0.00)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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 true positive rate and the true negative rate 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 \( (p, q) = (20, 20) \).

<table>
<thead>
<tr>
<th>( n )</th>
<th>( p )</th>
<th>( q )</th>
<th>( \Omega )</th>
<th>( \Gamma )</th>
<th>MLE</th>
<th>Ridge</th>
<th>LASSO</th>
<th>SCAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>100</td>
<td>20</td>
<td>( A_1 )</td>
<td>( A_1 )</td>
<td>504(15)</td>
<td>273(7)</td>
<td>67.7(6.1)</td>
<td>17.1(2.0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( A_2 )</td>
<td>( A_1 )</td>
<td>504(14)</td>
<td>268(6)</td>
<td>131(5)</td>
<td>114.6(3.7)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( A_2 )</td>
<td>( A_2 )</td>
<td>502(16)</td>
<td>245(4)</td>
<td>183(11)</td>
<td>158(12)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( A_3 )</td>
<td>( A_1 )</td>
<td>504(13)</td>
<td>258(6)</td>
<td>133(10)</td>
<td>100(6)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( A_3 )</td>
<td>( A_2 )</td>
<td>506(15)</td>
<td>239(5)</td>
<td>150(9)</td>
<td>118(9)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( A_3 )</td>
<td>( A_3 )</td>
<td>507(16)</td>
<td>222(5)</td>
<td>117(6)</td>
<td>91(7)</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>20</td>
<td>( A_1 )</td>
<td>( A_1 )</td>
<td>59.6(1.4)</td>
<td>54.3(1.1)</td>
<td>12.3(0.7)</td>
<td>2.60(0.27)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( A_2 )</td>
<td>( A_1 )</td>
<td>59.1(1.4)</td>
<td>52.1(1.1)</td>
<td>28.7(0.8)</td>
<td>14.2(0.8)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( A_2 )</td>
<td>( A_2 )</td>
<td>59.5(1.4)</td>
<td>52.3(1.1)</td>
<td>32.9(0.9)</td>
<td>15.3(0.8)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( A_3 )</td>
<td>( A_1 )</td>
<td>59.0(1.1)</td>
<td>53.0(0.8)</td>
<td>24.7(0.8)</td>
<td>8.06(0.47)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( A_3 )</td>
<td>( A_2 )</td>
<td>59.4(1.3)</td>
<td>51.4(0.9)</td>
<td>30.2(1.1)</td>
<td>11.7(0.7)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( A_3 )</td>
<td>( A_3 )</td>
<td>59.7(1.3)</td>
<td>50.7(0.9)</td>
<td>19.9(0.8)</td>
<td>7.2(0.4)</td>
</tr>
<tr>
<td>20</td>
<td>100</td>
<td>50</td>
<td>( A_1 )</td>
<td>( A_1 )</td>
<td>471(11)</td>
<td>325(6)</td>
<td>68.5(3.5)</td>
<td>23.9(2.1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( A_2 )</td>
<td>( A_1 )</td>
<td>471(11)</td>
<td>323(6)</td>
<td>160(4)</td>
<td>123(5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( A_2 )</td>
<td>( A_2 )</td>
<td>472(12)</td>
<td>318(6)</td>
<td>216(6)</td>
<td>148(6)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( A_3 )</td>
<td>( A_1 )</td>
<td>473(11)</td>
<td>314(6)</td>
<td>126(6)</td>
<td>68.4(4.3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( A_3 )</td>
<td>( A_2 )</td>
<td>473(10)</td>
<td>310(6)</td>
<td>176(8)</td>
<td>97(5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( A_3 )</td>
<td>( A_3 )</td>
<td>472(10)</td>
<td>292(5)</td>
<td>165(17)</td>
<td>69.2(3.7)</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>50</td>
<td>( A_1 )</td>
<td>( A_1 )</td>
<td>68.5(1.3)</td>
<td>79.4(1.2)</td>
<td>17.3(0.8)</td>
<td>3.41(0.28)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( A_2 )</td>
<td>( A_1 )</td>
<td>68.4(1.3)</td>
<td>65.9(1.2)</td>
<td>41.8(1.1)</td>
<td>10.9(0.5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( A_2 )</td>
<td>( A_2 )</td>
<td>68.4(1.1)</td>
<td>63.7(1.1)</td>
<td>52.5(1.0)</td>
<td>13.3(0.5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( A_3 )</td>
<td>( A_1 )</td>
<td>68.3(1.3)</td>
<td>76.9(1.1)</td>
<td>45.6(1.3)</td>
<td>8.21(0.32)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( A_3 )</td>
<td>( A_2 )</td>
<td>68.1(1.3)</td>
<td>64.4(1.2)</td>
<td>66.2(2.0)</td>
<td>11.3(0.7)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( A_3 )</td>
<td>( A_3 )</td>
<td>68.4(1.4)</td>
<td>68.9(1.0)</td>
<td>49.2(1.1)</td>
<td>8.21(0.48)</td>
</tr>
</tbody>
</table>

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

Based on these simulation results, we see clearly that, by exploiting the covariance matrix of \( \text{vec}(X) \) as a Kronecker product of two matrices, the proposed method outperforms the graphical LASSO 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.

In order to investigate the sensitivity of the algorithm to the initial values, we conduct further numerical simulations. We take $\Omega = \Gamma = A_3$ with $n = 10$ and $p = q = 20$. We then take three possible initial values. The first is the maximum likelihood estimate computed using our algorithm with identity matrices as initial values when there is no penalty. The second uses identity matrices as 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 $B_j, j = 1, 2, 3$ for estimating $\Omega \otimes \Gamma$ with these three initial values. We then compute the Frobenius norm of $D_1 = B_1 - B_3$ and $D_2 = B_2 - B_3$. 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 $A_3$ to 0.5, we find that the Frobenius norms of 999 $D_1$’s are effectively zero, while those of 1000 $D_2$’s are zero. The final estimates seem to depend on the initial values when the less sparse matrices $\Omega$ and $\Gamma$ 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 US agricultural export data

The US agricultural export is an important part of US export. In March 2011 alone, the US agricultural export exceeds 13.3 billion US dollars and gives an agricultural trade surplus of about 4.5 billion US dollars. Understanding the export pattern can be useful to understand the current and predict future export trends. We extract the annual US agricultural export data between 1970 and 2009 from the United States Department of
Agriculture website. We look at the annual US export data to thirteen 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 etc., 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 data set comprises of 40 matrices with dimensionality $14 \times 36$. A selected 40 years data plot for North America, East Asia and Southeast Asia for these 36 items can be found in Figure 2.

Figure 2: USDA export data for three regions over 40 years for 36 items: Each connected line represents one item.

The original data are denoted in thousands US 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 data set, 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 ten 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 the US 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 eleven largest edges corresponds to either North Africa or Sub-Saharan Africa.

To compare the performance of the proposed decomposition to that of the usual graphical LASSO estimate, we use cross validation. Specifically, each time, one matrix is left out for comparison in terms of its log likelihood. The rest matrices are used to build a model, using either our method or the graphical LASSO 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 graphical LASSO 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.
Figure 3: The fitted graphical models of the regions (left) and of the items (right) for the USDA export data from 1970 to 2009.

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 Standard and Poor (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-Scholes’s model (Black and Scholes, 1973), the implied volatility data are as informative as option pricing data. We consider a dataset of the implied volatilities for standardized call
Figure 4: The estimates and the corresponding 95% confidence intervals for the 43 edges between the regions, where the largest eleven 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, "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.

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 followed 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 applied 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 fifty-nine companies are present in 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), Qualcomm (QCOM),
AT&T (T), Intel (INTC), 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 ConocoPhillips (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 91 and 183
days have six edges each, and options to expire 122 days 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 to the graphical lasso method, we use the
same strategy outlined for analyzing the US agricultural export data. The cross validation
procedure shows that our approach with either LASSO or SCAD penalty outperforms the
graphical LASSO method significantly.
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).

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 modelling such datasets via matrix-variate normal distribution, not only for rate of convergence consideration, but also for illuminating the relationships of the row and the column variables.

In this paper, 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, 2008b). 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 by fully satisfactory. It is desirable to develop more elaborate time series models such as vector 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 done by Rothman et al. (2010), where sparse regression models are employed in addition to sparse covariance estimation.

Appendix

Proof of Theorem 1:

To decompose (3), we define \( \Delta_1 = \Omega - \Omega_0, \Delta_2 = \Gamma - \Gamma_0 \) and note that

\[
\text{tr}(X_i \Gamma X_i^T \Omega) - \text{tr}(X_i \Gamma_0 X_i^T \Omega_0) = \text{tr}(X_i \Gamma_0 X_i^T \Delta_1) + \text{tr}(X_i \Delta_2 X_i^T \Omega_0) + \text{tr}(X_i \Delta_2 X_i^T \Delta_1).
\]

Further, by Taylor’s expansion

\[
\log |\Omega| - \log |\Omega_0| = \text{tr}((\Sigma_0 \Delta_1) - \text{vec}^T(\Delta_1) \left\{ \int_0^1 h(v, \Omega_v^{(1)})(1 - v)dv \right\} \text{vec}(\Delta_1))
\]
where $\Omega^{(1)}_v = \Omega_0 + v\Delta_1$ and $h(v, \Omega) = \Omega^{-1} \otimes \Omega^{-1}$. We define

$$T_1 = \frac{1}{npq} \left\{ \sum_{i=1}^{n} \text{tr}(X_i \Gamma_0 X_i^T \Delta_1) \right\} - \frac{1}{p} \text{tr}(\Sigma_0 \Delta_1),$$

$$T_2 = \frac{1}{npq} \left\{ \sum_{i=1}^{n} \text{tr}(X_i^T \Omega_0 X_i \Delta_2) \right\} - \frac{1}{q} \text{tr}(\Psi_0 \Delta_2),$$

$$T_3 = \frac{1}{npq} \sum_{i=1}^{n} \text{tr}(X_i \Delta_2 X_i^T \Delta_1),$$

$$T_4 = p^{-1} \text{vec}^T(\Delta_1) \left\{ \int_0^1 h(v, \Omega_v^{(1)})(1-v) dv \right\} \text{vec}(\Delta_1),$$

$$T_5 = q^{-1} \text{vec}^T(\Delta_2) \left\{ \int_0^1 h(v, \Omega_v^{(2)})(1-v) dv \right\} \text{vec}(\Delta_2),$$

$$T_6 = \sum_{(i,j) \in S_1'} \{ p_{\lambda_1}(|\omega_{ij}|) - p_{\lambda_1}(|\omega_{0ij}|) \} + \sum_{(i,j) \in S_2'} \{ p_{\lambda_2}(|\gamma_{ij}|) - p_{\lambda_1}(|\gamma_{0ij}|) \} \quad \text{and} \quad T_7 = \sum_{(i,j) \in S_1, i \neq j} \{ p_{\lambda_1}(|\omega_{ij}|) - p_{\lambda_1}(|\omega_{0ij}|) \} + \sum_{(i,j) \in S_2, i \neq j} \{ p_{\lambda_2}(|\gamma_{ij}|) - p_{\lambda_1}(|\gamma_{0ij}|) \}. $$

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

$$g(\Omega, \Gamma) - g(\Omega_0, \Gamma_0) = T_1 + T_2 + T_3 + T_4 + T_5 + T_6 + T_7. \quad (6)$$

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

$$P \left[ \inf_{\Delta_1 \in \mathcal{A}, \Delta_2 \in \mathcal{B}} \{ g(\Omega_0 + \Delta_1, \Gamma_0 + \Delta_2) - g(\Omega_0, \Gamma_0) \} > 0 \right] \rightarrow 1.$$

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

$$T_4 \geq p^{-1} \|\text{vec}(\Delta_1)\|^2 \int_0^1 (1-v) \text{min}_{0 < e \leq 1} \lambda_1(\Omega^{-1} \otimes \Omega^{-1}) dv$$

$$\geq (2p)^{-1} \{ \tau_1^{-1} + o(1) \}^{-2} (C_1^2 \alpha_1^2 + C_2^2 \beta_1^2)$$

and similarly,

$$T_5 \geq (2q)^{-1} \{ \tau_1^{-1} + o(1) \}^{-2} (C_1^2 \alpha_2^2 + C_2^2 \beta_2^2).$$

Then, we consider $T_1$ and $T_2$,

$$(npq)^{-1} \sum_{i=1}^{n} \text{tr}(X_i \Gamma_0 X_i^T \Delta_1) = (npq)^{-1} \sum_{i=1}^{n} \text{tr}(Y_i Y_i^T \Delta_1) = p^{-1} \text{tr}(Q_1 \Delta_1)$$
where \( Q_1 = (nq)^{-1} \sum_{i=1}^{n} Y_i Y_i^T \), and \( Y_i = X_i \Gamma_0^{1/2} \) follows matrix-variate normal distribution with parameters \( \Sigma_0 \) and \( I \) by the properties of matrix-variate normal distribution (Gupta and Nagar, 2000). In other words, the columns of \( Y_i \) are independent and identically distributed normal random vectors with covariance matrix \( \Sigma_0 \). Therefore

\[
T_1 = p^{-1} \text{tr}\{(Q_1 - \Sigma_0)\Delta_1\} = p^{-1} \left\{ \sum_{(i,j) \in S_1} + \sum_{(i,j) \in S_1^c} \right\} (Q_1 - \Sigma_0)_{ij}(\Delta_1)_{ij} = T_{11} + T_{12}
\]

where \( T_{11} \) and \( T_{12} \) are respectively the two sums over \( S_1 \) and \( S_1^c \). Since \( \max_{i,j} |(Q_1 - \Sigma_0)_{ij}| = O_p[\{\log p/(nq)\}^{1/2}] \) by Bickel and Levina (2008a),

\[
|T_{11}| \leq p^{-1}(s_1 + p)^{1/2}||\Delta_1||_F \max_{i,j} |(Q_1 - \Sigma_0)_{ij}| = p^{-1}O_p(\alpha_1 + \beta_1)||\Delta_1||_F.
\]

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

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

\[
\sum_{(i,j) \in S_1^c} \{p_{\lambda_i}(|\Delta_{ij}|) - p^{-1}(Q_1 - \Sigma_0)_{ij}(\Delta_1)_{ij}\} \geq \sum_{(i,j) \in S_1^c} \{\lambda_1 k_1 - p^{-1} \max_{i,j} |(Q_1 - \Sigma_0)_{ij}|\} |\Delta_{1ij}|
\]

\[
= \lambda_1 \sum_{(i,j) \in S_1^c} \{k_1 - o_p(1)\} |\Delta_{1ij}|,
\]

which is greater than 0 with probability tending to 1, and the last equality is due to Condition A4 so that \( \lambda_1^{-1} p^{-1} \max_{i,j} |(Q_1 - \Sigma_0)_{ij}| = O_p\{\lambda_1^{-1} p^{-1} \sqrt{\log p/(nq)}\} = o_p(1) \).

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 \( \lambda_1 \) and \( \lambda_2 \) 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,

\[
p_{\lambda_i}(|\omega_{ij}|) = p_{\lambda_i}(|\omega_{0ij}|) + p'_{\lambda_i}(|\omega_{0ij}|)\text{sgn}(\omega_{0ij})(\omega_{ij} - \omega_{0ij}) + 2^{-1}p''_{\lambda_i}(|\omega_{0ij}|)(\omega_{ij} - \omega_{0ij})^2\{1 + o(1)\}.
\]
Noting Condition A3, we have
\[
\sum_{(i,j) \in S_1, i \neq j} \left\{ p\lambda_1(|\omega_{ij}|) - p\lambda_1(|\omega_{0ij}|) \right\}
\leq \sqrt{s_1}C_1 \alpha_1 \max_{(i,j) \in S_1} p\lambda_1'(|\omega_{0ij}|) + C_1 \max_{(i,j) \in S_1} p\lambda_1''(|\omega_{0ij}|) \alpha_1^2 \left\{ 1 + \omega(1) \right\}
\]
\[
= O\{ p^{-1}(\alpha_1^2 + \beta_1^2) \}.
\]
Hence, it is dominated by the positive term \( T_4 \). Same arguments apply for the penalty on \( \Gamma \). Therefore \( T_7 \) is dominated by \( T_4 + T_5 \).

Finally, note that \( E\{ \text{tr}(X_i \Delta_2 X_i^T \Delta_1) \} = \text{tr}(\Delta_2 \Psi_0) \text{tr}(\Sigma_0 \Delta_1) \) by the properties of matrix-variate normal distribution (Gupta and Nagar, 2000). Hence by law of large numbers
\[
T_3 = (npq)^{-1} \sum_{i=1}^{n} \text{tr}(X_i \Delta_2 X_i^T \Delta_1) = (pq)^{-1} \text{tr}(\Delta_2 \Psi_0) \text{tr}(\Sigma_0 \Delta_2) \{ 1 + \omega_p(1) \}.
\]
Let \( \xi_{1j} \) and \( \xi_{2k}, j = 1, \ldots, p, k = 1, \ldots, q \) be the eigen values of \( \Delta_1 \) and \( \Delta_2 \). Then by the von Neumann inequality
\[
|\text{tr}(\Sigma_0 \Delta_1)| \leq \sum_{j=1}^{p} \lambda_j(\Sigma_0) |\xi_{1j}| \leq (\tau_1)^{-1} \sum_{j=1}^{p} |\xi_{1j}| \leq \tau_1^{-1} p^{1/2} \| \Delta_1 \|_F \leq \tau_1^{-1} p^{1/2} (C_1^2 \alpha_1^2 + C_2^2 \beta_1^2)^{1/2}.
\]
By applying the argument on \( |\text{tr}(\Psi_0)\Delta_2| \), we have as \( n \to \infty \),
\[
|T_3| \leq 1/\sqrt{pq} \tau_1^{-2} (C_1^2 \alpha_1^2 + C_2^2 \beta_1^2)(C_1^2 \alpha_2^2 + C_2^2 \beta_2^2) \leq T_4 + T_5.
\]
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 \( \Omega \) and \( \Gamma \), we have
\[
\frac{\partial g(\Omega, \Gamma)}{\partial \omega_{ij}} = 2(t_{ij} - p^{-1} \sigma_{ij} + p\lambda_1'(|\omega_{ij}|)) \text{sgn}(\omega_{ij})
\]
where \( t_{ij} \) is the \((i, j)\)th element of
\[
(npq)^{-1} \sum_{i=1}^{n} (X_i \Gamma X_i^T) = p^{-1} Q_1 + (npq)^{-1} \sum_{i=1}^{n} X_i(\Gamma - \Gamma_0)X_i^T.
\]
It is seen that the \((i, j)\)th element in \((npq)^{-1} \sum_{i=1}^{n} X_i (\Gamma - \Gamma_0) X_i^\top\) is \(O_p\{p^{-1/2} \eta_{2n}\}\) following Lam and Fan (2009). In addition, \(p^{-1} \max_{ij} |q_{ij} - \sigma_{0ij}| = O_p\{p^{-1/2} \log p/(npq)\}\) by Bickel and Levina (2008a). Further, following Lam and Fan (2009), \(p^{-1}|\sigma_{0ij} - \sigma_{ij}| = O(p^{-1/2} \eta_{1n}^2)\). Therefore we need \(\log p/(nq) + \eta_{1n} + \eta_{2n} = O(\lambda_1^2 p^2)\) to ensure that \(\text{sgn}(\omega_{ij})\) dominates the first derivative of \(g(\Omega, \Gamma)\), and by the same arguments for the minimizer \(\Gamma\). Theorem 2 then follows.

**Proof of Theorem 3:**

Since \(\hat{\Omega}\) and \(\hat{\Gamma}\) solves \(0 = \{\partial g(\Omega, \Gamma)/\partial \vec(\Omega), \partial g(\Omega, \Gamma)/\partial \vec(\Gamma)\}^\top\), for the local minimizer \(\hat{\Omega}\) and \(\hat{\Gamma}\) in Theorem 1, we use Taylor’s expansion at the truth. Let \(A \otimes^2 = A \otimes A\). Standard calculations imply

\[
\left\{ \begin{array}{c}
\Sigma_0 \otimes^2 \\
\Psi_0 \otimes^2
\end{array} \right\} + \Lambda_n + \left( \begin{array}{c}
R_{1n} \\
R_{2n}
\end{array} \right)_{S \times S} \left( \begin{array}{c}
\vec(\hat{\Omega}) - \vec(\Omega_0) \\
\vec(\hat{\Gamma}) - \vec(\Gamma_0)
\end{array} \right)_S
\]

where the remainder terms satisfy \(\|R_{1n}\| = O_p(\alpha_1 + \beta_1)\) and \(\|R_{2n}\| = O_p(\alpha_2 + \beta_2)\) following the proof of Theorem 1, the subscript \(S\) and \(S \times S\) indicate the sub-vector and sub-matrix corresponding to those nonzero components of \(\Omega_0\) and \(\Gamma_0\) excluding \(\omega_{11}\), \(\Lambda_n = \text{diag}\{p_{n1}^\top(\hat{\Omega}), p_{n2}^\top(\hat{\Omega})\}\), and \(b_n = \{p_{n1}^\top(\Omega_0)\text{sgn}(\Omega_0), p_{n2}^\top(\Omega_0)\text{sgn}(\Omega_0)\}^\top\). The follow facts by results in Gupta and Nagar (2000) imply the moments of the right hand side in (7).

First note that \(E\{\text{vec}(X_i \Gamma_0 X_i^\top)\} = q \Sigma_0\) and \(E\{\text{vec}(X_i^\top \Omega_0 X_i)\} = p \Psi_0\). Let \(S_1 = X_i \Gamma_0 X_i^\top\), \(S_2 = X_i^\top \Omega_0 X_i\), then

\[
\text{cov}\{\text{vec}(S_1), \text{vec}(S_1)\} = q \Sigma_0 \otimes^2 (I + K_{pp}), \text{cov}\{\text{vec}(S_2), \text{vec}(S_2)\} = p \Psi_0 \otimes^2 (I + K_{qq})
\]

where \(K_{ab}\) 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 \(C\), \(E(S_1 C S_2) = \Sigma_0 C \Psi_0 (2 + pq)\)
and $E(S_1)CE(S_2) = pq = \Sigma_0 C \Psi_0$. Then for any $q \times p$ matrix $D$, we have

$$\text{tr}[(C^T \otimes D)\text{cov}\{\text{vec}(S_1), \text{vec}(S_2)\}]$$

$$= \text{tr}(C^T \otimes D)[E\{\text{vec}(S_1)\text{vec}^T(S_2)\} - E\{\text{vec}(S_1)\}E\{\text{vec}^T(S_2)\}]$$

$$= \text{tr}[E\{\text{vec}(DS_1C)\text{vec}^T(S_2)\} - E\{\text{vec}(DS_1C)\}E\{\text{vec}^T(S_2)\}]$$

$$= \text{tr}\{E(S_1CS_2D) - E(S_1)CE(S_2)D\} = 2\Sigma_0 C \Psi_0 D$$

$$= \text{vec}^T(\Psi_0)(C^T \otimes D)\text{vec}(\Sigma_0) = \text{tr}\{(C^T \otimes D)\text{vec}(\Psi_0)\text{vec}^T(\Sigma_0)\}.$$


