Statistica Sinica **27** (2017), 1299-1317
doi:https://doi.org/10.5705/ss.202015.0011

THE EFFECT OF L_1 PENALIZATION ON CONDITION NUMBER CONSTRAINED ESTIMATION OF PRECISION MATRIX

Xiao Guo[1] and Chunming Zhang[2,3]

[1] *University of Science and Technology of China,* [2] *Nankai Unversity and* [3] *University of Wisconsin-Madison*

Abstract: Estimation of large precision matrices is fundamental to high-dimensional inference. An important issue is to deal with ill-conditioning of the precision matrix estimate, typically encountered in finite-samples, but rarely studied in the literature. In this paper, we focus on estimating the precision matrix by imposing a bound on the condition number of the estimate, which effectively ensures well-conditioning. Specifically, we propose a correlation-based estimator, constrained with both the condition number and the L_1 penalty, yielding a precision matrix estimator with theoretically guaranteed rate of convergence. This result further enables us to demonstrate that incorporating the L_1 penalty is necessary for achieving consistency of the resulting estimator in typical high-dimensional settings, while inconsistency will occur when the L_1 penalty is absent. An algorithm based on the alternating direction method of multipliers is developed to implement the proposed method, which reveals the satisfactory performance in simulation studies. An application of the method to a call center data is illustrated.

Key words and phrases: Condition number, covariance matrix, penalization, precision matrix, sparsity.

1. Introduction

Estimation of a large precision matrix has been an important and challenging problem with applications in many scientific fields. For example, in linear discriminant analysis, optimal portfolio selection, recovery of the structure of undirected Gaussian graphical model, and detection of activated brain regions for neuroimaging data estimation of the precision matrix is needed. Given n i.i.d. p_n-variate random vectors, the inverse of the sample covariance matrix, $\mathbf{S}_n^{-1}$, is commonly used for estimating the precision matrix $\mathbf{\Sigma}_0^{-1}$, where $\mathbf{\Sigma}_0$ is the true covariance matrix. When the dimension p_n is fixed, $\mathbf{S}_n$ is consistent for $\mathbf{\Sigma}_0$, but when $p_n > n$, the singularity of $\mathbf{S}_n$ makes its inverse unavailable. Even if $p_n \le n$, $\mathbf{S}_n^{-1}$ may be inconsistent when $\lim_{n\to\infty} p_n/n = c$ for a constant $c \in (0, 1]$

(Marčenko and Pastur (1967)). Besides the potential issue of inconsistency associated with precision matrix estimator, another issue is that its condition number is large, or inflated in practice.

In the literature, several approaches for estimating the covariance matrix have been developed. See Bickel and Levina (2008), Cai and Liu (2011), Cai and Zhou (2012), Liu, Wang and Zhao (2014), Rothman (2012), Rothman, Levina and Zhu (2009), and Xue, Ma and Zou (2012), among others. The inverses of these estimators are consistent for the precision matrix. At the same time, regularization schemes have been proposed to estimate the precision matrix directly. For example, Meinshausen and Bühlmann (2006) proposed an L_1 penalized regression approach, which was extended by Peng et al. (2009). Other works utilizing the penalized log-likelihood approach include Banerjee, El Ghaoui and d'Aspremont (2008), Friedman, Hastie and Tibshirani (2008), and Lam and Fan (2009). These methods simultaneously recover the sparsity structure of the precision matrix. Although the aforementioned estimators are consistent, the problem of ill-conditioning was not taken into consideration. To protect the condition number of the precision matrix estimate from being inflated, a natural way is to shrink the eigenvalues of the estimator. For example, Won et al. (2013) developed an estimator of $\mathbf{\Sigma}_0$ by imposing a bound on the condition number of the estimator, but did not address the issue of inconsistency.

In this work, we focus on estimating the precision matrix with a condition number constraint. We consider a correlation-based estimator of the precision matrix with the condition number constraint, and study its asymptotic properties. We incorporate the L_1 penalty with the proposed estimator and examine its effect on the consistency of the condition number constrained estimator. We show that if the L_1 penalty is absent, the estimator is consistent only in restrictive cases and inconsistent in many circumstances. Under regularity conditions, we find the convergence rate of our estimator with the L_1 penalty incorporated in high-dimensional cases, allowing $\lim_{n\to\infty} p_n/n > 0$.

Our estimator with the L_1 penalty is asymptotically equivalent to the correlation-based SPICE estimator developed in Rothman et al. (2008), but has an advantage in that it possesses a constrained condition number and enjoys better finite-sample performance. To implement our estimation method, we develop an algorithm based on the alternating direction method of multipliers (Boyd et al. (2010)). Simulations and data analyses reveal the satisfactory performance of our proposed estimator with the L_1 penalty included.

The rest of this paper is organized as follows. Section 2 proposes the condition number constrained estimator of the precision matrix, details situations in which consistency and inconsistency take place and derives the convergence rate of the estimator with the L_1 penalty incorporated. Section 3 develops the algorithm for the estimator $\widehat{\mathbf{\Theta}}_{\text{prop}-2}$ defined at (2.5) and (2.6). Section 4 discusses data-driven choices of the tuning parameters. Section 5 presents simulations and Section 6 analyzes data. The supplementary material includes the proofs of the results.

We introduce some notation here. For any set G, denote by $|G|$ the cardinality of G. For matrices A and B of size $m \times m$, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the smallest and largest eigenvalues of A, respectively, $A \succ 0$ means that A is positive definite, and $A \otimes B$ is the Kronecker matrix product. We write $A(i,j)$ for the element of A in the ith row and jth column. The trace of A is denoted by $\text{tr}(A)$ and $\det(A)$ is the determinant of A. The off-diagonal elementwise L_1 norm of A is

$$|A|_1 = \sum_{1 \le i \neq j \le m} \sum |A(i,j)|. \tag{1.1}$$

The L_2, L_∞, and Frobenius norms of A are $\|A\|_2 = \{\lambda_{\max}(A^T A)\}^{1/2}$, $\|A\|_\infty = \max_{1 \le i \le m} \sum_{j=1}^m |A(i,j)|$, and $\|A\|_F = \{\text{tr}(A^T A)\}^{1/2}$, respectively. Denote by $\mathbf{I}_m$ the $m \times m$ identity matrix and by $\boldsymbol{e}_{q,m}$ the qth column of $\mathbf{I}_m$. The empirical spectral distribution of A is $\mathbb{F}^A(x) = m^{-1}|\{j \le m : \lambda_j \le x\}|$, where $\{\lambda_j\}_{j=1}^m$ are the eigenvalues of A. For a sequence of random distribution functions $\{F_n\}_{n\ge1}$ and a deterministic distribution function F_0, we write $F_n(x) \xrightarrow{\text{P}} F_0(x)$ as $n \to \infty$ at any continuous point x of F_0 to denote that F_n converges weakly to F_0 in probability. For a vector $\boldsymbol{v} = (v_1, \ldots, v_m)^T$, the L_1 norm is $\|\boldsymbol{v}\|_1 = \sum_{i=1}^m |v_i|$. For two sequences of positive real numbers $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$, $a_n \asymp b_n$ denotes that $a_n = O(b_n)$ and $b_n = O(a_n)$. In the following, C and c are generic finite constants that may vary from place to place and do not depend on n.

2. Condition Number Constrained Estimator of $\mathbf{\Sigma}_0^{-1}$

Throughout, $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are i.i.d. p_n-variate random vectors, with $\boldsymbol{X}_i = (X_{i,1}, \ldots, X_{i,p_n})^T$. We write $\boldsymbol{\mu}_0 = E(\boldsymbol{X}_i) \in \mathbb{R}^{p_n}$ and $\mathbf{\Sigma}_0 = \text{cov}(\boldsymbol{X}_i, \boldsymbol{X}_i)$ and assume $\mathbf{\Sigma}_0$ is positive definite. If $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are normally distributed, the log-likelihood is

$$\ell_n(\boldsymbol{\mu}_0, \mathbf{\Sigma}_0) = -\frac{np_n}{2}\log(2\pi)$$

$$-\frac{1}{2}\Big[-n\log\{\det(\boldsymbol{\Sigma}_0^{-1})\}+\sum_{i=1}^{n}(\boldsymbol{X}_i-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{X}_i-\boldsymbol{\mu}_0)\Big].$$

Write the maximum likelihood estimators (MLEs) of $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ as $\overline{\boldsymbol{X}}$ and $\mathbf{S}_n$. If we replace $\boldsymbol{\mu}_0$ with $\overline{\boldsymbol{X}}$, the Gaussian log-likelihood function is

$$\ell_n(\overline{\boldsymbol{X}},\boldsymbol{\Sigma}_0)=-\frac{np_n}{2}\log(2\pi)-\frac{n}{2}[-\log\{\det(\boldsymbol{\Sigma}_0^{-1})\}+\mathrm{tr}(\mathbf{S}_n\boldsymbol{\Sigma}_0^{-1})].$$

Let $\boldsymbol{\Theta}_0=\boldsymbol{\Sigma}_0^{-1}$ be the precision matrix of $\boldsymbol{X}_i$. We focus on estimating $\boldsymbol{\Theta}_0$. It is known that, when p_n is fixed, $\mathbf{S}_n^{-1}$ is a well-behaved estimator of $\boldsymbol{\Theta}_0$. However, when $\lim_{n\to\infty}p_n/n=c$ with a constant $c\in(0,1]$, $\mathbf{S}_n^{-1}$ may be ill-conditioned (Marčenko and Pastur (1967)), i.e. the condition number is inflated where condition number of a positive-definite matrix $\boldsymbol{\Sigma}$ is defined as $\mathrm{cond}(\boldsymbol{\Sigma})=\lambda_{\max}(\boldsymbol{\Sigma})/\lambda_{\min}(\boldsymbol{\Sigma})$. Particularly, if $p_n>n$, then $\mathbf{S}_n$ is not invertible. When the dimension is large, $\mathbf{S}_n^{-1}$ is either numerically unavailable or ill-conditioned.

We consider estimating $\boldsymbol{\Theta}_0$ by imposing a condition number constraint while minimizing the negative Gaussian log-likelihood function. Although the underlying distribution of $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ may not be Gaussian, its log-likelihood still performs well as a loss function in both theoretical and practical aspects. Sections 2.1 and 2.2 study the properties of the condition number constrained estimators of $\boldsymbol{\Theta}_0$, in the absence and presence of an L_1 penalty, respectively. Proofs of the results are given in the supplementary material.

2.1. Condition number constrained estimator: without L_1 penalty

Let $\boldsymbol{\Sigma}_0=W_0\Gamma_0W_0$, where Γ_0 is the true correlation matrix and W_0 is the diagonal matrix of the true standard deviations. If $\Omega_0=\Gamma_0^{-1}$, then $\boldsymbol{\Theta}_0=W_0^{-1}\Omega_0W_0^{-1}$. Let $\mathbf{S}_n=\widehat{W}\mathbf{R}_n\widehat{W}$, where $\widehat{W}^2$ is the diagonal matrix with the same diagonal as $\mathbf{S}_n$ and $\mathbf{R}_n$ is the sample correlation matrix.

We propose an estimator of the precision matrix

$$\widehat{\boldsymbol{\Theta}}_{\mathrm{prop}-1}=\widehat{W}^{-1}\widetilde{\Omega}_{\kappa_n}\widehat{W}^{-1}, \tag{2.1}$$

where $\widetilde{\Omega}_{\kappa_n}$ is the solution to

$$\begin{cases}\underset{\Omega\succ 0}{\text{minimize}} & -\log\{\det(\Omega)\}+\mathrm{tr}(\mathbf{R}_n\Omega),\\ \text{subject to} & \mathrm{cond}(\Omega)\le\kappa_n,\end{cases} \tag{2.2}$$

with a tuning parameter $\kappa_n\ge 1$.

Won et al. (2013) developed a well-conditioned estimator $\widehat{\boldsymbol{\Sigma}}_{\mathrm{WLKR}}$ of $\boldsymbol{\Sigma}_0$. The inverse of their estimator, $\widehat{\boldsymbol{\Theta}}_{\mathrm{WLKR}}=\widehat{\boldsymbol{\Sigma}}_{\mathrm{WLKR}}^{-1}$, is also well-conditioned and can be obtained by solving

$$\begin{cases} \underset{\boldsymbol{\Theta}\succ 0}{\text{minimize}} & -\log\{\det(\boldsymbol{\Theta})\} + \mathrm{tr}(\mathbf{S}_n\boldsymbol{\Theta}), \\ \text{subject to} & \mathrm{cond}(\boldsymbol{\Theta}) \le \kappa_n, \end{cases} \tag{2.3}$$

where $\kappa_n \ge 1$ is a tuning parameter. The difference between $\widehat{\boldsymbol{\Theta}}_{\mathrm{prop}-1}$ in (2.1) and $\widehat{\boldsymbol{\Theta}}_{\mathrm{WLKR}}$ in (2.3) is that $\widehat{\boldsymbol{\Theta}}_{\mathrm{prop}-1}$ is a correlation-based estimator while $\widehat{\boldsymbol{\Theta}}_{\mathrm{WLKR}}$ regularizes the precision matrix directly.

From Won et al. (2013), $\widetilde{\Omega}_{\kappa_n}^{-1}$ in (2.2) truncates the eigenvalues of $\mathbf{R}_n$. However, because of the condition number constraint, $\widetilde{\Omega}_{\kappa_n}^{-1}$ and $\mathbf{R}_n$ have different asymptotic behaviors. We examine the asymptotic properties of $\widehat{\boldsymbol{\Theta}}_{\mathrm{prop}-1}$. The following conditions will be involved.

A1. $\lim_{n\to\infty}\log(p_n)/n = 0$, $\max_{1\le j\le p_n} E[e^{t\{X_{1,j}-E(X_{1,j})\}^2}] < C$ for $|t| < c$, $\boldsymbol{\Sigma}_0$ is diagonal, and $\kappa_n = 1$, where $C \in (0,\infty)$ and $c \in (0,\infty)$.

A2. $\lim_{n\to\infty} p_n^{4/\beta}/n = 0$, $\max_{1\le j\le p_n} E\{|X_{1,j} - E(X_{1,j})|^\beta\} < C$, $\boldsymbol{\Sigma}_0$ is diagonal, and $\kappa_n = 1$, where $\beta \in [4,\infty)$ and $C \in (0,\infty)$.

A3. $\lim_{n\to\infty} p_n/n = 0$, $\max_{1\le j\le p_n} E\{|X_{1,j} - E(X_{1,j})|^4\} < C$, $\|\boldsymbol{\Sigma}_0^{-1/2}\|_\infty < C$ with constant $C \in (0,\infty)$, and $\liminf_{n\to\infty}\{\kappa_n - \mathrm{cond}(\Omega_0)\} > 0$. For any $i = 1,\ldots,n$, $\{\boldsymbol{e}_{j,p_n}^T\boldsymbol{\Sigma}_0^{-1/2}(\boldsymbol{X}_i - \boldsymbol{\mu}_0) : j = 1,\ldots,p_n\}$ are i.i.d. random variables.

Theorem 1 (consistency of $\widehat{\boldsymbol{\Theta}}_{\mathrm{prop}-1}$)**.** *Suppose $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n \in \mathbb{R}^{p_n}$ are* i.i.d. *with mean vector $\boldsymbol{\mu}_0$ and covariance matrix $\boldsymbol{\Sigma}_0$, and $0 < c \le \lambda_{\min}(\boldsymbol{\Sigma}_0) \le \lambda_{\max}(\boldsymbol{\Sigma}_0) \le C < \infty$. For $\widehat{\boldsymbol{\Theta}}_{\mathrm{prop}-1}$ in* (2.1)*, under Condition* A1 *or* A2 *or* A3*, we have $\|\widehat{\boldsymbol{\Theta}}_{\mathrm{prop}-1} - \boldsymbol{\Theta}_0\|_2 \xrightarrow{\mathrm{P}} 0$ and $\|\widehat{\boldsymbol{\Theta}}_{\mathrm{prop}-1}^{-1} - \boldsymbol{\Sigma}_0\|_2 \xrightarrow{\mathrm{P}} 0$ as $n \to \infty$.*

The results of Theorem 1 also hold for $\widehat{\boldsymbol{\Theta}}_{\mathrm{WLKR}}$ if we replace Conditions A1, A2, and A3 respectively by the following.

A1*. $\lim_{n\to\infty}\log(p_n)/n = 0$, $\max_{1\le j\le p_n} E[e^{t\{X_{1,j}-E(X_{1,j})\}^2}] < C$ for $|t| < c$, $\boldsymbol{\Sigma}_0 = \mathbf{I}_{p_n}$, and $\kappa_n = 1$, where $C \in (0,\infty)$ and $c \in (0,\infty)$.

A2*. $\lim_{n\to\infty} p_n^{4/\beta}/n = 0$, $\max_{1\le j\le p_n} E\{|X_{1,j} - E(X_{1,j})|^\beta\} < C$, $\boldsymbol{\Sigma}_0 = \mathbf{I}_{p_n}$, and $\kappa_n = 1$, where $\beta \in [4,\infty)$ and $C \in (0,\infty)$.

A3*. $\lim_{n\to\infty} p_n/n = 0$, $\max_{1\le j\le p_n} E\{|X_{1,j} - E(X_{1,j})|^4\} < C$, $\|\boldsymbol{\Sigma}_0^{-1/2}\|_\infty < C$ with constant $C \in (0,\infty)$, and $\liminf_{n\to\infty}\{\kappa_n - \mathrm{cond}(\boldsymbol{\Theta}_0)\} > 0$. For any $i = 1,\ldots,n$, $\{\boldsymbol{e}_{j,p_n}^T\boldsymbol{\Sigma}_0^{-1/2}(\boldsymbol{X}_i - \boldsymbol{\mu}_0) : j = 1,\ldots,p_n\}$ are i.i.d. random variables.

Condition A3 has been considered by Bai and Yin (1993) and El Karoui (2009), and holds when $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ are, for example, normal. Comparing Conditions A1 and A2 with A1* and A2*, in high-dimensional settings, $\widehat{\mathbf{\Theta}}_{\text{prop}-1}$ is consistent for $\mathbf{\Theta}_0$ when $\mathbf{\Sigma}_0$ is diagonal, while the consistency of $\widehat{\mathbf{\Theta}}_{\text{WLKR}}$ requires $\mathbf{\Sigma}_0 = \mathbf{I}_{p_n}$. Conditions A1–A3 are restrictive. Theorem 2 details situations in which $\widehat{\mathbf{\Theta}}_{\text{prop}-1}$ is inconsistent.

Denote by $\mathbb{F}^{\Gamma_0}$ the empirical spectral distribution of Γ_0. Suppose $\mathbb{F}^{\Gamma_0}$ converges to a probability distribution function F_0 weakly as $n \to \infty$, and let

$$\begin{aligned} l_{\min} &= \inf\{x : F_0(x) > 0\}, & l_{\max} &= \sup\{x : F_0(x) < 1\}, \\ c_{\min} &= \inf\{F_0(x) : F_0(x) > 0\}, & c_{\max} &= \sup\{F_0(x) : F_0(x) < 1\}. \end{aligned} \tag{2.4}$$

Denote by $\mathbb{F}^{\mathbf{R}_n}$ the empirical spectral distribution of $\mathbf{R}_n$. From Theorem 1 of El Karoui (2009), if $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are i.i.d. random vectors, $\{\boldsymbol{e}_{j,p_n}^T \mathbf{\Sigma}_0^{-1/2}(\boldsymbol{X}_i - \boldsymbol{\mu}_0) : j = 1, \ldots, p_n\}$ are i.i.d. random variables for any $i = 1, \ldots, n$, $\mathbb{F}^{\Gamma_0}$ converges to F_0 weakly, $\max_{1 \le j \le p_n} E\{|X_{1,j} - E(X_{1,j})|^\beta\} < C$, $\|\mathbf{\Sigma}_0^{-1/2}\|_\infty < C$, $\|\Gamma_0\|_2 < C$, and $\lim_{n\to\infty} p_n/n = y$ with constants $\beta \in (4, \infty)$, $C \in (0, \infty)$, and $y \in (0, \infty)$, then $\mathbb{F}^{\mathbf{R}_n}$ converges weakly to a distribution function F in probability. We provide regularity conditions that are needed in Theorem 2.

- B1. $\|\widehat{W}^2 - W_0^2\|_2 = o_{\text{P}}(1)$, $\lim_{n\to\infty} p_n/n = \infty$, and $F_0 \neq \mathrm{I}_{[C,\infty)}$ for any $C \in [0, \infty)$.
- B2. $\|\widehat{W}^2 - W_0^2\|_2 = o_{\text{P}}(1)$ and $|\min\{\kappa_n, \text{cond}(\mathbf{R}_n)\} - \text{cond}(\Gamma_0)| \nrightarrow 0$ in probability as $n \to \infty$.
- B3. $\{\boldsymbol{e}_{j,p_n}^T \mathbf{\Sigma}_0^{-1/2}(\boldsymbol{X}_i - \boldsymbol{\mu}_0) : j = 1, \ldots, p_n\}$ are i.i.d. random variables, for any $i = 1, \ldots, n$. $0 < l_{\min} < l_{\max} < \infty$, $\max_{1 \le j \le p_n} E\{|X_{1,j} - E(X_{1,j})|^\beta\} < C$, $\|\mathbf{\Sigma}_0^{-1/2}\|_\infty < C$, $\lim_{n\to\infty} p_n/n = y$, $\lim_{n\to\infty} \kappa_n = l_{\max}/l_{\min}$, and $F_0 \neq F\, \mathrm{I}_{[l_{\min}, l_{\max})} + \mathrm{I}_{[l_{\max},\infty)}$, where $l_{\min}$ and $l_{\max}$ are defined in (2.4), $\beta \in (4, \infty)$, $C \in (0, \infty)$, and $y \in (0, \infty)$ are constants and F is the limit that $\mathbb{F}^{\mathbf{R}_n}$ converges weakly to in probability.

Theorem 2 (inconsistency of $\widehat{\mathbf{\Theta}}_{\text{prop}-1}$)**.** *Suppose* $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^{p_n}$ *are* i.i.d. *with mean vector* $\boldsymbol{\mu}_0$ *and covariance matrix* $\mathbf{\Sigma}_0$, $0 < c \le \lambda_{\min}(\mathbf{\Sigma}_0) \le \lambda_{\max}(\mathbf{\Sigma}_0) \le C < \infty$. *If* $\mathbb{F}^{\Gamma_0}$ *converges to a probability distribution function* F_0 *weakly, for* $\widehat{\mathbf{\Theta}}_{\text{prop}-1}$ *in* (2.1)*, under Condition* B1 *or* B2 *or* B3*, we have* $\|\widehat{\mathbf{\Theta}}_{\text{prop}-1} - \mathbf{\Theta}_0\|_2 \nrightarrow 0$ *and* $\|\widehat{\mathbf{\Theta}}_{\text{prop}-1}^{-1} - \mathbf{\Sigma}_0\|_2 \nrightarrow 0$ *in probability.*

From Condition B1, $F_0 \neq \mathrm{I}_{[C,\infty)}$ for any $C \in [0, \infty)$ excludes the case of $\mathbf{\Sigma}_0$ being diagonal. In this situation, if p_n is much larger than n, then $\widehat{\mathbf{\Theta}}_{\text{prop}-1}$ is not a consistent estimator. Condition B2 implies that a well-selected tuning parameter κ_n is very important for the consistency of $\widehat{\mathbf{\Theta}}_{\text{prop}-1}$. For Condition

B3, $F_0 \neq F\,\mathrm{I}_{[l_{\min},l_{\max})} + \mathrm{I}_{[l_{\max},\infty)}$ is satisfied in many situations. For example, from Silverstein and Choi (1995), if $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{\text{i.i.d.}}{\sim} N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, $\mathbb{F}^{\Gamma_0}$ converges to F_0 weakly, $\lim_{n\to\infty} p_n/n = y \in (0,1)$ and some other regularity conditions hold, then F has a continuous density function on $(0,\infty)$. If F_0 does not have a continuous density, for example, F_0 is discrete, and $c_{\min} + c_{\max} < 1$ with $c_{\min}$ and $c_{\max}$ defined in (2.4), then condition $F_0 \neq F\,\mathrm{I}_{[l_{\min},l_{\max})} + \mathrm{I}_{[l_{\max},\infty)}$ holds. In this situation, $\widehat{\boldsymbol{\Theta}}_{\mathrm{prop}-1}$ is not consistent for $\boldsymbol{\Theta}_0$ in high-dimensional cases.

The results of Theorem 2 also hold for $\widehat{\boldsymbol{\Theta}}_{\mathrm{WLKR}}$ if we replace Ω_0, Γ_0, $\mathbf{R}_n$, and $\widehat{\boldsymbol{\Theta}}_{\mathrm{prop}-1}$ with $\boldsymbol{\Theta}_0$, $\boldsymbol{\Sigma}_0$, $\mathbf{S}_n$, and $\widehat{\boldsymbol{\Theta}}_{\mathrm{WLKR}}$, respectively, in the conditions of Theorem 2.

2.2. Condition number constrained estimator: with L_1 penalty

To obtain consistent estimators of the precision matrix in high-dimensional cases, several estimation methods have been developed (see, e.g., Lam and Fan (2009) and Rothman et al. (2008)). The commonly used approach is the penalized log-likelihood method. In this section, we develop an estimator of $\boldsymbol{\Theta}_0$ with the condition number constraint and L_1 penalty, and study its asymptotic property.

By adding the L_1 penalty of Ω to the objective function in (2.2), we propose the estimator,

$$\widehat{\boldsymbol{\Theta}}_{\mathrm{prop}-2} = \widehat{W}^{-1}\widehat{\Omega}_{\mu_n,\kappa_n}\widehat{W}^{-1}, \tag{2.5}$$

where $\widehat{\Omega}_{\mu_n,\kappa_n}$ solves the optimization problem,

$$\begin{cases} \underset{\Omega\succ 0}{\text{minimize}} & -\log\{\det(\Omega)\} + \mathrm{tr}(\mathbf{R}_n\Omega) + \mu_n|\Omega|_1, \\ \text{subject to} & \mathrm{cond}(\Omega) \leq \kappa_n, \end{cases} \tag{2.6}$$

with $\mu_n > 0$ and $\kappa_n \geq 1$ the tuning parameters and $|\cdot|_1$ the matrix off-diagonal elementwise L_1 norm defined in (1.1). From the proof of Lemma 3 in Ravikumar et al. (2011), since the objective function in (2.6) is strictly convex for $\Omega \succ 0$ (Ravikumar et al. (2011)) and $\{\Omega \succ 0, \mathrm{cond}(\Omega) \leq \kappa_n\}$ is a convex constraint, there exists a unique solution to (2.6).

A similar estimator without the condition number constraint developed in Rothman et al. (2008) is given by

$$\widehat{\boldsymbol{\Theta}}_{\mathrm{RBLZ}} = \widehat{W}^{-1}\widehat{\Omega}_{\mathrm{RBLZ}}\widehat{W}^{-1}, \tag{2.7}$$

where

$$\widehat{\Omega}_{\mathrm{RBLZ}} = \underset{\Omega\succ 0}{\arg\min}\,[-\log\{\det(\Omega)\} + \mathrm{tr}(\mathbf{R}_n\Omega) + \mu_n|\Omega|_1],$$

with a tuning parameter $\mu_n > 0$. The convergence rate of $\widehat{\mathbf{\Theta}}_{\rm RBLZ}$ has been demonstrated in Rothman et al. (2008) under normal $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. In Theorem 3, we re-examine the consistency of $\widehat{\mathbf{\Theta}}_{\rm RBLZ}$ under the exponential tail assumption. The convergence rate of $\widehat{\mathbf{\Theta}}_{\rm RBLZ}$ is also established under the polynomial tail condition. In the following, for sets T, $T' \subseteq \{1, \ldots, p_n\} \times \{1, \ldots, p_n\}$, let $(\Gamma_0 \otimes \Gamma_0)_{TT'}$ denote the $|T| \times |T'|$ submatrix of $\Gamma_0 \otimes \Gamma_0$ with rows and columns indexed by T and T', respectively (see Section 3.1 in Ravikumar et al. (2011)). Specifically, if $T = \{(i_u, j_u) : u = 1, \ldots, h\}$ and $T' = \{(i'_v, j'_v) : v = 1, \ldots, h'\}$, then $\boldsymbol{e}_{u,h}^T (\Gamma_0 \otimes \Gamma_0)_{TT'} \boldsymbol{e}_{v,h'} = \Gamma_0(i_u, i'_v)\Gamma_0(j_u, j'_v)$ for $u = 1, \ldots, h$ and $v = 1, \ldots, h'$.

For the next result, let $s_n = |\{(i, j) : i \neq j \text{ and } \mathbf{\Theta}_0(i, j) \neq 0\}|$ and $t_n = \max_{i=1,\ldots,p_n} |\{j = 1, \ldots, p_n : \mathbf{\Theta}_0(i, j) \neq 0\}|$ and consider the following conditions

C1. (exponential tail condition) $\max_{1 \le j \le p_n} E[e^{t\{X_{1,j} - E(X_{1,j})\}^2}] < C$ for $|t| < c$ with certain constants $C \in (0, \infty)$ and $c \in (0, \infty)$, $\mu_n \asymp \{\log(p_n)/n\}^{1/2}$ and $r_n = o(1)$ where $r_n = \min(1 + s_n, t_n^2) \log(p_n)/n$.

C2. (polynomial tail condition) $\max_{1 \le j \le p_n} E\{|X_{1,j} - E(X_{1,j})|^\beta\} < C$ for certain constants $\beta \in [4, \infty)$ and $C \in (0, \infty)$, $\mu_n \asymp p_n^{2\tau/\beta}/n^{1/2}$ and $r_n = o(1)$ where $r_n = \min(1 + s_n, t_n^2) p_n^{4\tau/\beta}/n$ and $\tau \in (2, \infty)$ is a constant.

Theorem 3. *Suppose $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^{p_n}$ are* i.i.d. *random vectors with covariance matrix $\mathbf{\Sigma}_0$ such that $\lambda_{\min}(\mathbf{\Sigma}_0) \ge c > 0$ and $\|\mathbf{\Sigma}_0\|_\infty \le C < \infty$. Let $S = \{(i, j) : \mathbf{\Theta}_0(i, j) \neq 0\}$ and assume $\|\{(\Gamma_0 \otimes \Gamma_0)_{SS}\}^{-1}\|_\infty \le C < \infty$ and $\max_{e \in \{\{(i,j)\} : \mathbf{\Theta}_0(i,j)=0\}} \|(\Gamma_0 \otimes \Gamma_0)_{eS}\{(\Gamma_0 \otimes \Gamma_0)_{SS}\}^{-1}\|_1 \le c$ with constant $c \in (0, 1)$. For $\widehat{\mathbf{\Theta}}_{\rm RBLZ}$ in* (2.7)*, if either* C1 *or* C2 *holds, we have $\|\widehat{\mathbf{\Theta}}_{\rm RBLZ} - \mathbf{\Theta}_0\|_2^2 = O_{\rm P}(r_n) = \|\widehat{\mathbf{\Theta}}_{\rm RBLZ}^{-1} - \mathbf{\Sigma}_0\|_2^2$.*

Rothman et al. (2008) demonstrated that the convergence rate of $\widehat{\mathbf{\Theta}}_{\rm RBLZ}$ under the L_2 norm is $\{(1 + s_n) \log(p_n)/n\}^{1/2}$ by assuming that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are multivariate normal and the eigenvalues of $\mathbf{\Sigma}_0$ are bounded away from 0 and ∞. If, in $\mathbf{\Theta}_0$, the maximum number of non-zeros per row is large relative to the total number of non-zero off-diagonal elements, $s_n = O(t_n^2)$, then the convergence rate of $\widehat{\mathbf{\Theta}}_{\rm RBLZ}$ in Theorem 3 under the exponential tail condition is equivalent to that in Rothman et al. (2008). However, when $t_n^2 = o(s_n)$, our result provides a faster convergence rate while requiring stronger conditions than that in Rothman et al. (2008). Particularly, conditions $\|\{(\Gamma_0 \otimes \Gamma_0)_{SS}\}^{-1}\|_\infty \le C < \infty$ and $\max_{e \in \{\{(i,j)\} : \mathbf{\Theta}_0(i,j)=0\}} \|(\Gamma_0 \otimes \Gamma_0)_{eS}\{(\Gamma_0 \otimes \Gamma_0)_{SS}\}^{-1}\|_1 \le c$ in Theorem 3 are adopted from Ravikumar et al. (2011), the latter of which is regarded as the mutual incoherence or irrepresentability condition.

We now study the asymptotic properties of $\widehat{\boldsymbol{\Theta}}_{\text{prop}-2}$.

Theorem 4 (consistency of $\widehat{\boldsymbol{\Theta}}_{\text{prop}-2}$)**.** *For* $\widehat{\boldsymbol{\Theta}}_{\text{prop}-2}$ *in* (2.5)*, under the conditions in Theorem 3 and* $\liminf_{n\to\infty}\{\kappa_n - \text{cond}(\Omega_0)\} > 0$*,*

$$\lim_{n\to\infty} \mathrm{P}(\widehat{\boldsymbol{\Theta}}_{\text{prop}-2} = \widehat{\boldsymbol{\Theta}}_{\text{RBLZ}}) = 1,$$
$$\|\widehat{\boldsymbol{\Theta}}_{\text{prop}-2} - \boldsymbol{\Theta}_0\|_2^2 = O_{\mathrm{P}}(r_n) = \|\widehat{\boldsymbol{\Theta}}_{\text{prop}-2}^{-1} - \boldsymbol{\Sigma}_0\|_2^2.$$

Comparing the results of Theorems 1, 2, and 4, we can see the effect of the L_1 penalty. Under the condition that $\max(r_n, p_n/n) = o(1)$, $\widehat{\boldsymbol{\Theta}}_{\text{prop}-1}$ and $\widehat{\boldsymbol{\Theta}}_{\text{prop}-2}$ are consistent. In high-dimensional cases with $\lim_{n\to\infty} p_n/n > 0$ and $r_n = o(1)$, if $\boldsymbol{\Sigma}_0$ is not diagonal, then, under some regularity conditions, $\widehat{\boldsymbol{\Theta}}_{\text{prop}-2}$ converges to $\boldsymbol{\Theta}_0$ but $\widehat{\boldsymbol{\Theta}}_{\text{prop}-1}$ may not. The L_1 penalty of Ω in the objective function of (2.6) is necessary for the convergence of $\widehat{\boldsymbol{\Theta}}_{\text{prop}-2}$ in high-dimensional settings.

Although $\widehat{\boldsymbol{\Theta}}_{\text{prop}-2}$ is asymptotically equivalent to $\widehat{\boldsymbol{\Theta}}_{\text{RBLZ}}$, we find in numerical studies in Sections 5 and 6 that our proposed method performs better in finite-sample situations while effectively controlling the condition number of the estimate.

Another competitive estimator is the graphical lasso estimator (Friedman, Hastie and Tibshirani (2008)) defined as

$$\widehat{\boldsymbol{\Theta}}_{\text{GLasso}} = \underset{\boldsymbol{\Theta}\succ 0}{\arg\min}\,[-\log\{\det(\boldsymbol{\Theta})\} + \text{tr}(\mathbf{S}_n\boldsymbol{\Theta}) + \mu_n|\boldsymbol{\Theta}|_1], \tag{2.8}$$

with a tuning parameter $\mu_n > 0$. Consistency of $\widehat{\boldsymbol{\Theta}}_{\text{GLasso}}$ has been demonstrated by Rothman et al. (2008) and Ravikumar et al. (2011) under different conditions. The convergence rate of $\widehat{\boldsymbol{\Theta}}_{\text{GLasso}}$ under the L_2 norm derived in Rothman et al. (2008) is $\{(p_n+s_n)\log(p_n)/n\}^{1/2}$ under normality. Comparing the rate of $\widehat{\boldsymbol{\Theta}}_{\text{GLasso}}$ in Rothman et al. (2008) with that of $\widehat{\boldsymbol{\Theta}}_{\text{prop}-2}$ in Theorem 4 under Condition C1, we find that the two rates are equivalent if $p_n = O(s_n)$ and $s_n = O(t_n^2)$, while $\widehat{\boldsymbol{\Theta}}_{\text{prop}-2}$ converges faster if $t_n^2 = o(p_n + s_n)$ or $s_n = o(p_n)$. In Ravikumar et al. (2011), the derived convergence rate of $\widehat{\boldsymbol{\Theta}}_{\text{GLasso}}$ is $\{\min(p_n+s_n, t_n^2)\log(p_n)/n\}^{1/2}$ under the exponential tail assumption and $\{\min(p_n+s_n, t_n^2)p_n^{4\tau/\beta}/n\}^{1/2}$ under the polynomial tail assumption. Under either the exponential tail or polynomial tail assumption, the convergence rate of $\widehat{\boldsymbol{\Theta}}_{\text{prop}-2}$ is equivalent to that of $\widehat{\boldsymbol{\Theta}}_{\text{GLasso}}$ in Ravikumar et al. (2011) if $p_n = O(s_n)$ or $t_n^2 = O(s_n)$, while the rate of $\widehat{\boldsymbol{\Theta}}_{\text{prop}-2}$ is sharper if $s_n = o(t_n^2)$ and $s_n = o(p_n)$. To sum up, compared with the convergence rates of $\widehat{\boldsymbol{\Theta}}_{\text{GLasso}}$ obtained in past work (Rothman et al. (2008); Ravikumar et al. (2011)), the rate of convergence for $\widehat{\boldsymbol{\Theta}}_{\text{prop}-2}$ is faster under certain situations, for example, in cases where $\boldsymbol{\Theta}_0$ is sparse and the total number of non-zero off-diagonal elements in $\mathbf{\Theta}_0$ is small relative to the maximum number of non-zeros per row. Under some other situations, where s_n dominates p_n or t_n^2, $\widehat{\mathbf{\Theta}}_{\text{prop}-2}$ and $\widehat{\mathbf{\Theta}}_{\text{GLasso}}$ converge to $\mathbf{\Theta}_0$ at the same rate.

3. Algorithm for Solving (2.6)

Numerically solving the optimization problem (2.6) is a non-trivial task. Different algorithms for penalized sparse precision matrix estimation have been developed (Boyd et al. (2010); Friedman, Hastie and Tibshirani (2008)). For example, Boyd et al. (2010) proposed an ADMM (alternating direction method of multipliers) algorithm, which can be used to solve the optimization problems for $\widehat{\mathbf{\Theta}}_{\text{RBLZ}}$ and $\widehat{\mathbf{\Theta}}_{\text{GLasso}}$. Based on the ADMM algorithm, we develop an algorithm for our estimator. The problem (2.6) is equivalent to the optimization problem

$$\begin{cases} \underset{\Omega\succ 0,\, Z}{\text{minimize}} & -\log\{\det(\Omega)\} + \operatorname{tr}(\mathbf{R}_n\Omega) + \mu_n|Z|_1, \\ \text{subject to} & \operatorname{cond}(\Omega) \le \kappa_n,\ \Omega = Z. \end{cases} \tag{3.1}$$

To deal with (3.1), we minimize the corresponding scaled augmented Lagrangian,

$$\begin{cases} \underset{\Omega\succ 0,\, Z}{\text{minimize}} & L_\rho(\Omega, Z; U), \\ \text{subject to} & \operatorname{cond}(\Omega) \le \kappa_n,\ \Omega = Z, \end{cases} \tag{3.2}$$

where

$$\begin{aligned} L_\rho(\Omega, Z; U) = &-\log\{\det(\Omega)\} + \operatorname{tr}(\mathbf{R}_n\Omega) + \mu_n|Z|_1 \\ &+ \frac{\rho}{2}\|\Omega - Z + U\|_F^2 - \frac{\rho}{2}\|U\|_F^2, \end{aligned} \tag{3.3}$$

is the scaled augmented Lagrange function, $\rho \in (0,\infty)$ is an arbitrary constant, and U is the Lagrange multiplier. The objective functions and constraints in both (3.1) and (3.2) are convex, and therefore there exist unique solutions to the two optimization problems (see, e.g., the proof of Lemma 3 in Ravikumar et al. (2011)). The problems (2.6), (3.1), and (3.2) are equivalent, so we use (3.2).

Motivated by the ADMM algorithm (Boyd et al. (2010)), we calculate the limit of $Z^{(i)}$ as the solution to (3.2) by iterations (with $i = 1, 2, \ldots$) of three steps until convergence

$$\begin{aligned} &\text{Step 1}:\ \Omega^{(i)} \leftarrow \underset{\Omega\succ 0,\, \operatorname{cond}(\Omega)\le\kappa_n}{\arg\min} L_\rho(\Omega, Z^{(i-1)}; U^{(i-1)}), \\ &\text{Step 2}:\ Z^{(i)} \leftarrow \underset{Z}{\arg\min}\, L_\rho(\Omega^{(i)}, Z; U^{(i-1)}), \\ &\text{Step 3}:\ U^{(i)} \leftarrow U^{(i-1)} + \Omega^{(i)} - Z^{(i)}. \end{aligned}$$

The criterion for declaring algorithmic convergence is

$$\frac{\sum_{j=1}^{p_n}\sum_{k=1}^{p_n}|\Omega^{(i+1)}(j,k)-\Omega^{(i)}(j,k)|}{\sum_{j=1}^{p_n}\sum_{k=1}^{p_n}|\Omega^{(i)}(j,k)|}\le 10^{-4}.$$

Issues on the global convergence of the ADMM algorithm can be found in Boyd et al. (2010) (see Section 3.2 therein for details). In practice, we use the zero matrix as the initial values for $Z^{(0)}$ and $U^{(0)}$, and set $\rho = 1$ for each iteration to control the step size.

From Boyd et al. (2010), the optimization problem for $\widehat{\mathbf{\Theta}}_{\mathrm{RBLZ}}$ can be solved using the ADMM algorithm by iterations of three steps with an explicit solution for each step. The ADMM algorithm for $\widehat{\mathbf{\Theta}}_{\mathrm{prop}-2}$ here differs from that for $\widehat{\mathbf{\Theta}}_{\mathrm{RBLZ}}$ in the sense that we consider the constraint $\{\Omega \succ 0,\, \mathrm{cond}(\Omega) \le \kappa_n\}$ in Step 1 while they use $\Omega \succ 0$.

Next, we calculate the solutions for Steps 1 and 2 in our algorithm. For Step 2, from (3.3),

$$\begin{aligned} Z^{(i)} &= \arg\min_Z L_\rho(\Omega^{(i)}, Z; U^{(i-1)}) \\ &= \arg\min_Z \left\{\mu_n|Z|_1 + \frac{\rho}{2}\|\Omega^{(i)} - Z + U^{(i-1)}\|_F^2\right\} \\ &= \arg\min_Z \left[\frac{1}{2}\|Z - \{\Omega^{(i)} + U^{(i-1)}\}\|_F^2 + \frac{\mu_n}{\rho}|Z|_1\right]. \end{aligned} \tag{3.4}$$

The last optimization problem in (3.4) is similar to problem (1) in Xue, Ma and Zou (2012), which has a closed-form solution by soft-thresholding (see Paragraph 2 of Section 1 in Xue, Ma and Zou (2012) for details). By arguments similar to those in Xue, Ma and Zou (2012), we can show that there also exists a closed-form solution to (3.4) based on soft-thresholding: for $j, k = 1, \ldots, p_n$,

$$Z^{(i)}(j,k) = \begin{cases} A^{(i)}(j,k), & \text{if } j = k, \\ \mathrm{sign}\{A^{(i)}(j,k)\}\max\{|A^{(i)}(j,k)| - \mu_n/\rho, 0\}, & \text{otherwise}, \end{cases}$$

where $A^{(i)} = \Omega^{(i)} + U^{(i-1)}$. For Step 1, the solution is not straightforward due to the constraint $\{\Omega \succ 0,\, \mathrm{cond}(\Omega) \le \kappa_n\}$. To obtain $\Omega^{(i)}$, we now propose a method, the proof of which is available in the online supplementary material.

Proposition 1. *For the optimization problem in Step* 1*, let* VDV^T *be the eigen-decomposition of* $\mathbf{R}_n/\rho - Z^{(i-1)} + U^{(i-1)}$ *with* $D = \mathrm{diag}(d_1, \ldots, d_{p_n})$ *and* $d_1 \ge \cdots \ge d_{p_n}$*. Let* $\delta_j = -d_j/2 + \sqrt{d_j^2/4 + 1/\rho}$ *for* $j = 1, \ldots, p_n$*,*

$$\widetilde{D} = \begin{cases} \mathrm{diag}(\delta_1, \ldots, \delta_{p_n}), & \text{if } \delta_{p_n}/\delta_1 \le \kappa_n, \\ \mathrm{diag}(\widetilde{d}_1, \ldots, \widetilde{d}_{p_n}), & \text{if } \delta_{p_n}/\delta_1 > \kappa_n, \end{cases}$$

where $\widetilde{d}_j = \min\{\max(\tau_0, \delta_j), \kappa_n \tau_0\}$ *and*

$$\tau_0 = \frac{\begin{array}{c}\Big[-\rho\Big(\sum_{j=1}^{\alpha_0} d_j + \kappa_n \sum_{j=\beta_0}^{p_n} d_j \Big) + \Big\{ \rho^2 \Big(\sum_{j=1}^{\alpha_0} d_j + \kappa_n \sum_{j=\beta_0}^{p_n} d_j \Big)^2 \\ +4\rho(\alpha_0 + {\kappa_n}^2 p_n - {\kappa_n}^2 \beta_0 + {\kappa_n}^2)(\alpha_0 + p_n - \beta_0 + 1) \Big\}^{1/2} \Big]\end{array}}{2\rho(\alpha_0 + {\kappa_n}^2 p_n - {\kappa_n}^2 \beta_0 + {\kappa_n}^2)},$$

with α_0 *the largest index in* $\{1, \ldots, p_n\}$ *such that* $\tau_0 > \delta_{\alpha_0}$ *and* β_0 *the smallest index in* $\{1, \ldots, p_n\}$ *such that* $\kappa_n \tau_0 < \delta_{\beta_0}$. *Then, the solution to Step* 1 *is* $\Omega^{(i)} = V\widetilde{D}V^T$. *The quantities* α_0 *and* β_0 *can be found in* $O(p_n)$ *operations.*

Since Step 1 also requires the eigendecomposition of $\mathbf{R}_n/\rho - Z^{(i-1)} + U^{(i-1)}$ which takes $O(p_n^3)$ operations, the number of operations for each iteration of the algorithm is $O(p_n^3)$. To calculate $\widehat{\mathbf{\Theta}}_{\mathrm{RBLZ}}$ or $\widehat{\mathbf{\Theta}}_{\mathrm{GLasso}}$, the ADMM algorithm also needs $O(p_n^3)$ operations.

4. Tuning Parameter Selection

This section illustrates a data-driven method to select the tuning parameters for $\widehat{\mathbf{\Theta}}_{\mathrm{prop}-2}$. In practice, we choose κ_n and μ_n in an iterative way. At each step, one of them is fixed and the other one is updated, with cross validation used for choosing κ_n and BIC for μ_n. Specifically, for a fixed μ_n, we divide the data into k folds and choose κ_n by minimizing

$$\mathrm{CV}(\kappa_n; \mu_n) = \sum_{i=1}^{k} \frac{n}{2k} [-\log\{\det(\widehat{\Omega}^{[-i]}_{\mu_n,\kappa_n})\} + \mathrm{tr}(\mathbf{R}_n^{[i]} \widehat{\Omega}^{[-i]}_{\mu_n,\kappa_n})],$$

where $\mathbf{R}_n^{[i]}$ is the sample correlation matrix based on the ith fold and $\widehat{\Omega}^{[-i]}_{\mu_n,\kappa_n}$ is the estimate of Ω_0 calculated with all observations except those in the ith fold. Given κ_n, we choose μ_n that minimizes the BIC function

$$\begin{aligned} \mathrm{BIC}(\mu_n; \kappa_n) = & -n \log\{\det(\widehat{\Omega}_{\mu_n,\kappa_n})\} + n\mathrm{tr}(\mathbf{R}_n \widehat{\Omega}_{\mu_n,\kappa_n}) \\ & + \log(n) \sum\sum_{1 \le i \le j \le p_n} \mathrm{I}\{\widehat{\Omega}_{\mu_n,\kappa_n}(i,j) \ne 0\}. \end{aligned}$$

The details for selecting κ_n and μ_n are as follows.

Step I : Initialize κ_n^0.

Step II : Repeat the following steps (with $i = 1, 2, \ldots$) until convergence:

$$\mu_n^i = \arg\min_{\mu_n} \mathrm{BIC}(\mu_n; \kappa_n^{i-1}),$$
$$\kappa_n^i = \arg\min_{\kappa_n} \mathrm{CV}(\kappa_n; \mu_n^i),$$

where the optimization problems are solved by grid search.

In the numerical studies in Sections 5 and 6, the initial value in Step I is $\kappa_n^0 = \infty$. We also observe that the averaged number of iterations for the numerical convergence of the algorithm in Step II is moderate in each situation, and does not increase as p_n increases.

5. Simulation Evaluation

Simulation studies were conducted to compare estimators $\widehat{\mathbf{\Theta}}_{\text{prop}-1}$ in (2.1), $\widehat{\mathbf{\Theta}}_{\text{prop}-2}$ in (2.5), $\widehat{\mathbf{\Theta}}_{\text{WLKR}}$ in (2.3), $\widehat{\mathbf{\Theta}}_{\text{RBLZ}}$ in (2.7), $\widehat{\mathbf{\Theta}}_{\text{GLasso}}$ in (2.8), and $\widehat{\mathbf{\Theta}}_{\text{banded}}$, with $n = 300$ and $p_n \in \{100, 200, 400\}$, where $\widehat{\mathbf{\Theta}}_{\text{banded}}$ is the banded estimate of $\mathbf{\Theta}_0$ by Cholesky decomposition as defined in Bickel and Levina (2008) (see Section 2.2 therein). To generate data, we considered the following schemes

I. $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{\text{i.i.d.}}{\sim} N(\mathbf{0}, \mathbf{\Sigma}_0)$, where $\mathbf{\Sigma}_0 = \mathrm{diag}(10, \ldots, 10, 0.01, \ldots, 0.01)$. The proportion of the "high" eigenvalues is 80% among the p_n eigenvalues. Similar schemes were used in Won et al. (2013).

II. $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{\text{i.i.d.}}{\sim} t_3(\mathbf{0}, \mathbf{\Sigma}^*)$ with $(\mathbf{\Sigma}^*)^{-1} = \mathrm{diag}(3K, K, \ldots, K)$ where K is a 50×50 matrix such that $K(i,j) = \mathrm{I}(i=j) + 0.1\,\mathrm{I}(|i-j|=1) + 0.4\,\mathrm{I}(|i-j|=3)$ for $i,\, j = 1, \ldots, 50$.

III. $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{\text{i.i.d.}}{\sim} t_3(\mathbf{0}, \mathbf{\Sigma}^*)$ with $\boldsymbol{e}_{i,p_n}^T (\mathbf{\Sigma}^*)^{-1} \boldsymbol{e}_{j,p_n} = \{3\,\mathrm{I}(i=j) + 1.49\,\mathrm{I}(|i-j|=1)\}\{1 - 1/2\,\mathrm{I}(i \le p_n/2)\}\{1 - 1/2\,\mathrm{I}(j \le p_n/2)\}$ for $i,\, j = 1, \ldots, p_n$.

Here, $t_\nu(\mathbf{0}, \mathbf{\Sigma}^*)$ denotes the multivariate t distribution with degree of freedom ν, location vector $\mathbf{0} \in \mathbb{R}^{p_n}$, and scale matrix $\mathbf{\Sigma}^*$. Specifically, if random quantities $\boldsymbol{Y} \sim N(\mathbf{0}, \mathbf{\Sigma}^*)$ and $w \sim \chi^2_\nu$ are independent, then $\boldsymbol{Y}/(w/\nu)^{1/2} \sim t_\nu(\mathbf{0}, \mathbf{\Sigma}^*)$ (see, for example, Section 5.6 in DeGroot (2004)). Thus $\mathbf{\Sigma}_0 = 3\mathbf{\Sigma}^*$ in schemes II and III. For schemes I–III, the structures of $\mathbf{\Theta}_0$ are diagonal, block diagonal, and banded, respectively.

Monte Carlo simulations were replicated 400 times in each setting. For a generic estimator $\widehat{\Theta}$ of $\mathbf{\Theta}_0$, we calculated the averaged losses $\|\widehat{\Theta} - \mathbf{\Theta}_0\|_F$ and $\|\widehat{\Theta} - \mathbf{\Theta}_0\|_2$. The selection performance was measured by the false positive rate (FPR) and false negative rate (FNR):

Table 1. Comparison of $\widehat{\mathbf{\Theta}}_{\text{prop}-1}$, $\widehat{\mathbf{\Theta}}_{\text{prop}-2}$, $\widehat{\mathbf{\Theta}}_{\text{WLKR}}$, $\widehat{\mathbf{\Theta}}_{\text{RBLZ}}$, $\widehat{\mathbf{\Theta}}_{\text{GLasso}}$, and $\widehat{\mathbf{\Theta}}_{\text{banded}}$ with data generated by scheme I. Each metric is averaged over 400 replications with the standard error shown in the bracket.

p_n	$\widehat{\Theta}$	$\Vert\widehat{\Theta}-\mathbf{\Theta}_0\Vert_F$	$\Vert\widehat{\Theta}-\mathbf{\Theta}_0\Vert_2$	FPR	FNR
100	$\widehat{\mathbf{\Theta}}_{\text{prop}-1}$	36.96 (0.31)	18.81 (0.23)	0.00 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\text{prop}-2}$	36.96 (0.31)	18.81 (0.23)	0.00 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\text{WLKR}}$	140.22 (0.15)	31.46 (0.03)	1.00 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\text{RBLZ}}$	36.97 (0.31)	18.82 (0.23)	0.00 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\text{GLasso}}$	36.96 (0.31)	18.81 (0.23)	0.00 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\text{banded}}$	49.92 (1.07)	24.66 (0.56)	0.02 (0.00)	0.00 (0.00)
200	$\widehat{\mathbf{\Theta}}_{\text{prop}-1}$	52.64 (0.31)	21.42 (0.25)	0.00 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\text{prop}-2}$	52.64 (0.31)	21.42 (0.25)	0.00 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\text{WLKR}}$	631.40 (0.00)	99.83 (0.00)	1.00 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\text{RBLZ}}$	52.65 (0.31)	21.42 (0.25)	0.00 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\text{GLasso}}$	52.64 (0.31)	21.42 (0.25)	0.00 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\text{banded}}$	72.62 (1.75)	28.26 (0.72)	0.01 (0.00)	0.00 (0.00)
400	$\widehat{\mathbf{\Theta}}_{\text{prop}-1}$	74.40 (0.31)	23.48 (0.21)	0.00 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\text{prop}-2}$	74.40 (0.31)	23.47 (0.21)	0.00 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\text{WLKR}}$	892.90 (0.00)	99.83 (0.00)	1.00 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\text{RBLZ}}$	74.41 (0.31)	23.48 (0.21)	0.00 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\text{GLasso}}$	74.40 (0.31)	23.48 (0.21)	0.00 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\text{banded}}$	113.92 (2.85)	33.39 (0.81)	0.01 (0.00)	0.00 (0.00)

$$\text{FPR} = \frac{|\{(i,j) : \mathbf{\Theta}_0(i,j)=0,\ \widehat{\Theta}(i,j)\neq 0\}|}{|\{(i,j) : \mathbf{\Theta}_0(i,j)=0\}|},$$

$$\text{FNR} = \frac{|\{(i,j) : \mathbf{\Theta}_0(i,j)\neq 0,\ \widehat{\Theta}(i,j)= 0\}|}{|\{(i,j) : \mathbf{\Theta}_0(i,j)\neq 0\}|}.$$

For $\widehat{\mathbf{\Theta}}_{\text{prop}-2}$, the tuning parameters were selected as described in Section 4 with $k=5$. To calculate $\widehat{\mathbf{\Theta}}_{\text{prop}-1}$ and $\widehat{\mathbf{\Theta}}_{\text{WLKR}}$, we solved (2.2) and (2.3) using the algorithm in Won et al. (2013) (see (8), Lemma 1, and Theorem 1). The tuning parameter κ_n of $\widehat{\mathbf{\Theta}}_{\text{prop}-1}$ and $\widehat{\mathbf{\Theta}}_{\text{WLKR}}$ were selected by the 5-fold CV as illustrated in Won et al. (2013). The graphical lasso algorithm in Friedman, Hastie and Tibshirani (2008) was applied to calculate $\widehat{\mathbf{\Theta}}_{\text{RBLZ}}$ and $\widehat{\mathbf{\Theta}}_{\text{GLasso}}$, where μ_n was selected by the 5-fold cross validation illustrated in Rothman et al. (2008). For these methods, κ_n was chosen from $\{1.4226^i : i=0,1,\ldots,29\}$ and μ_n from $\{0.02\times 1.2390^i : i=0,1,\ldots,29\}$. The banding parameter for $\widehat{\mathbf{\Theta}}_{\text{banded}}$ was selected by the random splitting method in Bickel and Levina (2008).

The averaged losses, FPR, and FNR of different precision matrix estimators are presented in Tables 1–3 corresponding to schemes I–III, respectively. For

Table 2. Comparison of $\widehat{\mathbf{\Theta}}_{\mathrm{prop}-1}$, $\widehat{\mathbf{\Theta}}_{\mathrm{prop}-2}$, $\widehat{\mathbf{\Theta}}_{\mathrm{WLKR}}$, $\widehat{\mathbf{\Theta}}_{\mathrm{RBLZ}}$, $\widehat{\mathbf{\Theta}}_{\mathrm{GLasso}}$, and $\widehat{\mathbf{\Theta}}_{\mathrm{banded}}$ with data generated by scheme II. Each metric is averaged over 400 replications with the standard error shown in the bracket.

p_n	$\widehat{\Theta}$	$\Vert\widehat{\Theta}-\mathbf{\Theta}_0\Vert_F$	$\Vert\widehat{\Theta}-\mathbf{\Theta}_0\Vert_2$	FPR	FNR
100	$\widehat{\mathbf{\Theta}}_{\mathrm{prop}-1}$	4.49 (0.02)	1.40 (0.01)	1.00 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\mathrm{prop}-2}$	2.59 (0.03)	0.84 (0.01)	0.20 (0.00)	0.09 (0.00)
	$\widehat{\mathbf{\Theta}}_{\mathrm{WLKR}}$	4.76 (0.02)	1.25 (0.01)	1.00 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\mathrm{RBLZ}}$	2.93 (0.04)	0.94 (0.01)	0.13 (0.00)	0.10 (0.00)
	$\widehat{\mathbf{\Theta}}_{\mathrm{GLasso}}$	2.94 (0.04)	0.89 (0.01)	0.15 (0.00)	0.11 (0.00)
	$\widehat{\mathbf{\Theta}}_{\mathrm{banded}}$	5.06 (0.09)	1.62 (0.04)	0.07 (0.00)	0.15 (0.01)
200	$\widehat{\mathbf{\Theta}}_{\mathrm{prop}-1}$	5.10 (0.02)	1.37 (0.01)	1.00 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\mathrm{prop}-2}$	2.91 (0.03)	0.87 (0.01)	0.11 (0.00)	0.08 (0.00)
	$\widehat{\mathbf{\Theta}}_{\mathrm{WLKR}}$	6.11 (0.03)	1.55 (0.01)	1.00 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\mathrm{RBLZ}}$	3.31 (0.05)	0.97 (0.01)	0.08 (0.00)	0.10 (0.00)
	$\widehat{\mathbf{\Theta}}_{\mathrm{GLasso}}$	3.84 (0.04)	1.07 (0.01)	0.09 (0.00)	0.11 (0.00)
	$\widehat{\mathbf{\Theta}}_{\mathrm{banded}}$	5.76 (0.14)	1.77 (0.05)	0.05 (0.00)	0.10 (0.01)
400	$\widehat{\mathbf{\Theta}}_{\mathrm{prop}-1}$	6.14 (0.02)	1.38 (0.01)	1.00 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\mathrm{prop}-2}$	3.57 (0.03)	0.91 (0.01)	0.06 (0.00)	0.08 (0.00)
	$\widehat{\mathbf{\Theta}}_{\mathrm{WLKR}}$	7.33 (0.03)	1.69 (0.00)	1.00 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\mathrm{RBLZ}}$	3.91 (0.05)	0.97 (0.01)	0.05 (0.00)	0.09 (0.00)
	$\widehat{\mathbf{\Theta}}_{\mathrm{GLasso}}$	4.65 (0.04)	1.20 (0.01)	0.06 (0.00)	0.11 (0.00)
	$\widehat{\mathbf{\Theta}}_{\mathrm{banded}}$	7.17 (0.22)	1.94 (0.06)	0.03 (0.00)	0.06 (0.01)

scheme I, where $\mathbf{\Sigma}_0$ is diagonal, $\widehat{\mathbf{\Theta}}_{\mathrm{prop}-1}$, $\widehat{\mathbf{\Theta}}_{\mathrm{prop}-2}$, $\widehat{\mathbf{\Theta}}_{\mathrm{RBLZ}}$, and $\widehat{\mathbf{\Theta}}_{\mathrm{GLasso}}$ perform comparably well and outperform $\widehat{\mathbf{\Theta}}_{\mathrm{WLKR}}$ and $\widehat{\mathbf{\Theta}}_{\mathrm{banded}}$. The results for $\widehat{\mathbf{\Theta}}_{\mathrm{prop}-1}$, $\widehat{\mathbf{\Theta}}_{\mathrm{prop}-2}$, $\widehat{\mathbf{\Theta}}_{\mathrm{RBLZ}}$, and $\widehat{\mathbf{\Theta}}_{\mathrm{GLasso}}$ are similar in Table 1, since the calculated estimates are all close to $\widehat{W}^2$. From Table 2, by comparing the losses of the precision matrix estimators, we see that $\widehat{\mathbf{\Theta}}_{\mathrm{prop}-2}$ outperforms the other estimators. From Table 3, $\widehat{\mathbf{\Theta}}_{\mathrm{prop}-2}$ has smaller averaged losses than $\widehat{\mathbf{\Theta}}_{\mathrm{prop}-1}$, $\widehat{\mathbf{\Theta}}_{\mathrm{WLKR}}$, $\widehat{\mathbf{\Theta}}_{\mathrm{GLasso}}$, and $\widehat{\mathbf{\Theta}}_{\mathrm{banded}}$. Compared with $\widehat{\mathbf{\Theta}}_{\mathrm{RBLZ}}$, when $p_n = 400$, $\widehat{\mathbf{\Theta}}_{\mathrm{prop}-2}$ has a slightly larger loss under the L_2 norm. In the other cases, $\widehat{\mathbf{\Theta}}_{\mathrm{prop}-2}$ performs better than $\widehat{\mathbf{\Theta}}_{\mathrm{RBLZ}}$.

As suggested by one referee, the averaged condition numbers of $\widehat{\mathbf{\Theta}}_{\mathrm{prop}-2}$ and $\widehat{\mathbf{\Theta}}_{\mathrm{RBLZ}}$ were also compared under the same amount of L_1 regularization. After calculating $\widehat{\mathbf{\Theta}}_{\mathrm{prop}-2}$ with the tuning parameters $\kappa_n = \widehat{\kappa}_n$ and $\mu_n = \widehat{\mu}_n$ selected by the data-driven method in Section 4, we calculated $\widehat{\mathbf{\Theta}}_{\mathrm{RBLZ}}$ with μ_n equal to $\widehat{\mu}_n$ instead of selected by CV, and denote the resulting estimator by $\widehat{\mathbf{\Theta}}^*_{\mathrm{RBLZ}}$. Hence, $\widehat{\mathbf{\Theta}}_{\mathrm{prop}-2}$ and $\widehat{\mathbf{\Theta}}^*_{\mathrm{RBLZ}}$ have the same amount of L_1 regularization. However, for scheme I where $\Omega_0 = \mathbf{I}_{p_n}$, the data-driven choice of κ_n for $\widehat{\mathbf{\Theta}}_{\mathrm{prop}-2}$ is exactly 1

Table 3. Comparison of $\widehat{\mathbf{\Theta}}_{\text{prop}-1}$, $\widehat{\mathbf{\Theta}}_{\text{prop}-2}$, $\widehat{\mathbf{\Theta}}_{\text{WLKR}}$, $\widehat{\mathbf{\Theta}}_{\text{RBLZ}}$, $\widehat{\mathbf{\Theta}}_{\text{GLasso}}$, and $\widehat{\mathbf{\Theta}}_{\text{banded}}$ with data generated by scheme III. Each metric is averaged over 400 replications with the standard error shown in the bracket.

p_n	$\widehat{\Theta}$	$\Vert\widehat{\Theta}-\mathbf{\Theta}_0\Vert_F$	$\Vert\widehat{\Theta}-\mathbf{\Theta}_0\Vert_2$	FPR	FNR
100	$\widehat{\mathbf{\Theta}}_{\text{prop}-1}$	4.22 (0.03)	1.48 (0.02)	1.00 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\text{prop}-2}$	2.34 (0.04)	0.84 (0.01)	0.22 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\text{WLKR}}$	5.44 (0.03)	1.32 (0.01)	1.00 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\text{RBLZ}}$	3.02 (0.05)	0.95 (0.01)	0.14 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\text{GLasso}}$	3.63 (0.05)	1.01 (0.01)	0.18 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\text{banded}}$	4.28 (0.12)	1.46 (0.04)	0.04 (0.00)	0.09 (0.01)
200	$\widehat{\mathbf{\Theta}}_{\text{prop}-1}$	6.67 (0.05)	1.60 (0.01)	1.00 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\text{prop}-2}$	3.86 (0.06)	0.97 (0.01)	0.13 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\text{WLKR}}$	8.91 (0.04)	1.54 (0.01)	1.00 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\text{RBLZ}}$	4.58 (0.08)	1.03 (0.01)	0.10 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\text{GLasso}}$	5.93 (0.07)	1.14 (0.01)	0.12 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\text{banded}}$	6.90 (0.26)	1.91 (0.08)	0.04 (0.00)	0.06 (0.01)
400	$\widehat{\mathbf{\Theta}}_{\text{prop}-1}$	11.34 (0.05)	1.58 (0.01)	1.00 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\text{prop}-2}$	6.69 (0.06)	1.14 (0.01)	0.07 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\text{WLKR}}$	13.92 (0.04)	1.67 (0.00)	1.00 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\text{RBLZ}}$	7.00 (0.10)	1.11 (0.01)	0.06 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\text{GLasso}}$	9.43 (0.08)	1.25 (0.01)	0.07 (0.00)	0.00 (0.00)
	$\widehat{\mathbf{\Theta}}_{\text{banded}}$	9.81 (0.39)	2.08 (0.09)	0.02 (0.00)	0.03 (0.01)

in many replications, and hence $\widehat{\Omega}_{\mu_n,\kappa_n} = \mathbf{I}_{p_n}$ for any $\mu_n > 0$. In this situation, $\widehat{\mathbf{\Theta}}_{\text{prop}-2}$ is not sensitive to μ_n at all. The performance of $\widehat{\mathbf{\Theta}}_{\text{RBLZ}}$ depends on the choice of μ_n. Therefore, it is difficult to make a fair comparison of $\widehat{\mathbf{\Theta}}_{\text{prop}-2}$ and $\widehat{\mathbf{\Theta}}_{\text{RBLZ}}$ under the same amount of L_1 regularization for scheme I. In Table 4, the condition number of the true precision matrix and the averaged condition numbers of $\widehat{\mathbf{\Theta}}_{\text{prop}-2}$ and $\widehat{\mathbf{\Theta}}^*_{\text{RBLZ}}$ are presented for schemes II and III only. Table 4 reveals that the averaged condition numbers of $\widehat{\mathbf{\Theta}}^*_{\text{RBLZ}}$ are larger than the true values and have relatively large standard errors.

6. Data Application

To illustrate the applicability of the proposed method, we applied the precision matrix estimator to call center data (available at `http://iew3.technion.ac.il/serveng2012S/callcenterdata/index.html`). The data recorded the time of the phone calls entering the call center of "Anonymous Bank" in Israel every day in 1999. Because of the difference of the arrival patterns between the weekdays and weekends, we discarded the data for the weekends (Friday and

Table 4. Comparison of the averaged condition numbers of $\widehat{\mathbf{\Theta}}_{\text{prop}-2}$ and $\widehat{\mathbf{\Theta}}^*_{\text{RBLZ}}$ for schemes II and III. Results are averaged over 400 replications with the standard errors shown in the brackets.

Scheme	p_n	cond($\mathbf{\Theta}_0$)	cond($\widehat{\mathbf{\Theta}}_{\text{prop}-2}$)	cond($\widehat{\mathbf{\Theta}}^*_{\text{RBLZ}}$)
II	100	453.34	309.51 (4.59)	491.00 (29.70)
	200	453.34	344.18 (4.43)	645.07 (22.85)
	400	453.34	383.53 (4.42)	914.96 (47.99)
III	100	992.11	741.21 (9.50)	1,247.41 (41.03)
	200	1,127.59	862.59 (9.17)	1,710.53 (71.18)
	400	1,176.39	982.94 (9.23)	2,360.38 (132.83)

Saturday), and only used those on the weekdays (258 days). Since there are relatively fewer calls before 7:00am, we only considered the time period from 7:00am to midnight. On each day, we divided the 17-hour period into 3-minute intervals and counted the number of calls $X_{i,j}$ for the ith day and jth time period with $i \in \{1,\dots,258\}$ and $j \in \{1,\dots,340\}$.

We aimed to use the arrival counts in the first half of the day to predict those in the second half of the day. Take $\boldsymbol{X}_i = (X_{i,1},\dots,X_{i,340})^T$ to be the vector of observations on the ith day, for $i = 1,\dots,258$, and let $\boldsymbol{X}_i^{(1)} = (X_{i,1},\dots,X_{i,170})^T$ and $\boldsymbol{X}_i^{(2)} = (X_{i,171},\dots,X_{i,340})^T$ be the observations in the first and second halves of the day, respectively. We partitioned the mean vector and covariance matrix of $\boldsymbol{X}_i$ correspondingly by

$$\boldsymbol{\mu}_0 = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \mathbf{\Sigma}_0 = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix}.$$

The best linear predictor of $\boldsymbol{X}_i^{(2)}$ is expressed as

$$\widehat{\boldsymbol{X}}_i^{(2)} = \boldsymbol{\mu}_2 + \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}(\boldsymbol{X}_i^{(1)} - \boldsymbol{\mu}_1). \tag{6.1}$$

We used $\{\boldsymbol{X}_1,\dots,\boldsymbol{X}_{100}\}$ as the training set and $\{\boldsymbol{X}_{101},\dots,\boldsymbol{X}_{258}\}$ as the testing set. The estimates of $\boldsymbol{\mu}_0$ and $\mathbf{\Sigma}_0$, calculated based on the training data, were plugged into (6.1) for prediction. The sample mean $100^{-1}\sum_{i=1}^{100}\boldsymbol{X}_i$ was used to estimate $\boldsymbol{\mu}_0$ while the inverses of different precision matrix estimates were applied to estimate $\mathbf{\Sigma}_0$. Specifically, we calculated $\widehat{\mathbf{\Theta}}_{\text{prop}-1}$, $\widehat{\mathbf{\Theta}}_{\text{prop}-2}$, $\widehat{\mathbf{\Theta}}_{\text{WLKR}}$, $\widehat{\mathbf{\Theta}}_{\text{RBLZ}}$, $\widehat{\mathbf{\Theta}}_{\text{GLasso}}$, and $\widehat{\mathbf{\Theta}}_{\text{banded}}$ as the estimates of $\mathbf{\Theta}_0$ first, and took their inverses to estimate $\mathbf{\Sigma}_0$. For the precision matrix estimation problem, $(n,\, p_n) = (100,\, 340)$ and the tuning parameters were selected in the same way as that in Section 5 with grid points $\kappa_n \in \{1.2143^i : i = 0,1,\dots,19\}$ and $\mu_n \in \{0.1 \times 1.1708^i : i = 0,1,\dots,19\}$. The performance of different methods for estimating $\mathbf{\Theta}_0$ was compared by the averaged absolute forecast error (AFE) based on the testing

Table 5. Comparison of the AFE based on $\widehat{\mathbf{\Theta}}_{\text{prop}-1}$, $\widehat{\mathbf{\Theta}}_{\text{prop}-2}$, $\widehat{\mathbf{\Theta}}_{\text{WLKR}}$, $\widehat{\mathbf{\Theta}}_{\text{RBLZ}}$, $\widehat{\mathbf{\Theta}}_{\text{GLasso}}$, and $\widehat{\mathbf{\Theta}}_{\text{banded}}$ for the call center data, with SE denoting the standard error.

$\widehat{\Theta}$	$\widehat{\mathbf{\Theta}}_{\text{prop}-1}$	$\widehat{\mathbf{\Theta}}_{\text{prop}-2}$	$\widehat{\mathbf{\Theta}}_{\text{WLKR}}$	$\widehat{\mathbf{\Theta}}_{\text{RBLZ}}$	$\widehat{\mathbf{\Theta}}_{\text{GLasso}}$	$\widehat{\mathbf{\Theta}}_{\text{banded}}$
AFE	1.83	1.76	1.83	1.81	1.84	1.82
SE	0.01	0.01	0.01	0.01	0.01	0.01

data. Specifically,

$$\text{AFE} = \frac{1}{158 \times 170} \sum_{i=101}^{258} \sum_{j=1}^{170} |\boldsymbol{e}_{j,170}^T (\widehat{\boldsymbol{X}}_i^{(2)} - \boldsymbol{X}_i^{(2)})|.$$

Table 5 presents the AFE of the best linear predictor calculated with $\mathbf{\Theta}_0$ estimated by different methods. From Table 5, it is clear that $\widehat{\mathbf{\Theta}}_{\text{prop}-2}$ corresponds to a smaller averaged absolute forecast error than the other estimators.

Supplementary Materials

The detailed proofs of Theorems 1–4 and Proposition 1 are relegated to the Supplementary Material.

Acknowledgment

C. Zhang's research is supported by the NSF grants DMS-1308872 and DMS-1521761, Wisconsin Alumni Research Foundation, and National Natural Science Foundation of China, grants 11690014. X. Guo's research is supported by National Natural Science Foundation of China, grants 11601500, and the Fundamental Research Funds for the Central Universities.

References

Bai, Z. D. and Yin, Y. Q. (1993). Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *Ann. Prob.* **21**, 1275–1294.

Banerjee, O., El Ghaoui, L. and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.* **9**, 485–516.

Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36**, 199–227.

Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. (2010). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundat. Trends Mach. Learn.* **3**, 1–122.

Cai, T. T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.* **106**, 672–684.

Cai, T. T. and Zhou, H. H. (2012). Minimax estimation of large covariance matrices under ℓ_1-norm. *Statistica Sinica* **22**, 1319–1378.

DeGroot, M. H. (2004). *Optimal Statistical Decisions*. Wiley-Interscience.

El Karoui, N. (2009). Concentration of measure and spectra of random matrices: applications to correlation matrices, elliptical distributions and beyond. *Ann. Appl. Prob.* **19**, 2362–2405.

Farrell, R. H. (1985). *Multivariate Calculation*. Springer, New York.

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–441.

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

Liu, H., Wang, L. and Zhao, T. (2014). Sparse covariance matrix estimation with eigenvalue constraints. *J. Comput. Graph. Statist.* **23**, 439–459.

Marčenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Math. USSR-Sb.* **1**, 507–536.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34**, 1436–1462.

Peng, J., Wang, P., Zhou, N. and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *J. Amer. Statist. Assoc.* **104**, 735–746.

Ravikumar, P., Wainwright, M. J., Raskutti, G. and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Statist.* **5**, 935–980.

Rothman, A. J. (2012). Positive definite estimators of large covariance matrices. *Biometrika* **99**, 733–740.

Rothman, A. J., Bickel, P. J., Levina, E. and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Statist.* **2**, 494–515.

Rothman, A. J., Levina, E. and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc.* **104**, 177–186.

Silverstein, J. W. and Choi, S. I. (1995). Analysis of the limiting spectral distribution of large dimensional random matrices. *J. Mult. Anal.* **54**, 295–309.

Stewart, G. W. and Sun, J. (1990). *Matrix Perturbation Theory*. Academic Press, Boston.

Won, J., Lim, J., Kim, S. and Rajaratnam, B. (2013). Condition-number-regularized covariance estimation. *J. R. Statist. Soc.* B **75**, 427–450.

Xue, L., Ma, S. and Zou, H. (2012). Positive-definite ℓ_1-penalized estimation of large covariance matrices. *J. Amer. Statist. Assoc.* **107**, 1480–1491.

Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei, Anhui 230026, China.

E-mail: xiaoguo@ustc.edu.cn

School of Mathematical Sciences, Nankai University, Tianjin 300071, China;

Department of Statistics, University of Wisconsin, Madison, WI 53706, USA.

E-mail: cmzhang@stat.wisc.edu

(Received January 2015; accepted June 2016)