The Canadian Journal of Statistics
Vol. 00, No. 00, 2020, Pages 000–000
La revue canadienne de statistique

On variable ordination of Cholesky-based estimation for a sparse covariance matrix

Xiaoning KANG[1*] and Xinwei DENG[2]

[1]*Institute of Supply Chain Analytics and International Business College, Dongbei University of Finance and Economics, Dalian, China*
[2]*Department of Statistics, Virginia Tech, Blacksburg, VA, U.S.A.*

Key words and phrases: Covariance matrix; high dimensionality; order of variables; sparsity.

MSC 2010: Primary 62H12, 62H20.

Abstract: Estimation of a large sparse covariance matrix is of great importance for statistical analysis, especially in high-dimensional settings. The traditional approach such as the sample covariance matrix performs poorly due to the high dimensionality. The modified Cholesky decomposition (MCD) is a commonly used method for sparse covariance matrix estimation. However, the MCD method relies on the order of variables, which often is not available or cannot be pre-determined in practice. In this work, we solve this order issue by obtaining a set of covariance matrix estimates based on assuming different orders of variables used in the MCD. Then we consider an ensemble estimator as the "centre" of such a set of covariance matrix estimates with respect to the Frobenius norm. Our proposed method not only ensures that the estimator is positive definite, but also captures the underlying sparse structure of the covariance matrix. Under some regularity conditions, we establish both algorithmic and asymptotic convergence of the proposed method. Its merits are illustrated via simulation studies and a practical example using data from a prostate cancer study. *The Canadian Journal of Statistics* 00: 000–000; 2020 © 2020 Statistical Society of Canada

Résumé: L'estimation de grandes matrices de covariance éparses revêt une grande importance pour l'analyse statistique, notamment en haute dimension. Les estimés traditionnels tels que la matrice de covariance empirique offrent de piètres performances en haute dimension. La décomposition de Cholesky modifiée (DCM) est couramment utilisée pour l'estimation de matrices de covariances éparses. Elle dépend toutefois de l'ordre des variables qui est souvent indisponible ou inconnu d'avance en pratique. Les auteurs résolvent ce problème d'ordre en obtenant un ensemble d'estimés de la DCM avec des matrices de covariance dont les entrées sont ordonnées selon différents arrangements. Ils considèrent un estimateur ensembliste correspondant au centre selon la norme de Frobenius d'un tel ensemble d'estimés générés par des agencements différents des variables. Leur méthode garantit un estimateur positif défini en plus de capturer la structure éparse sous-jacente de la matrice de covariance. Les auteurs établissent la convergence algorithmique et asymptotique de la méthode proposée sous des conditions de régularité. Ils illustrent ses mérites par des études de simulation et l'analyse d'un exemple pratique avec les données d'une étude sur le cancer de la prostate. *La revue canadienne de statistique* 00: 000–000; 2020 © 2020 Société statistique du Canada

Additional Supporting Information may be found in the online version of this article at the publisher's website.

**Author to whom correspondence may be addressed.*
E-mail: xiaoningmike@126.com

1. INTRODUCTION

Estimation of a large covariance matrix from high-dimensional data is an important and challenging problem in multivariate data analysis. For example, dimension reduction using principal component analysis usually relies on accurate estimation of covariance matrix. In the context of graphical models, the estimate of a covariance matrix or its inverse is often used to infer the network structure of the graph. However, conventional covariance estimation is known to perform poorly due to the dimensionality of the estimation problem when the number of variables is close to or larger than the sample size (Johnstone, 2001). To overcome this curse of dimensionality, various methods proposed in literature often assume certain patterns of sparsity in the covariance matrices.

In this work we focus on the problem of estimating a sparse covariance matrix for high-dimensional data. Early work on covariance estimation includes shrinking eigenvalues of the sample covariance matrix (Dey & Srinivasan, 1985; Haff, 1991), a linear combination of the sample covariance matrix and a proper diagonal matrix (Ledoit & Wolf, 2004), improving the estimation using the matrix condition number (Aubry et al., 2012; Won et al., 2013), and regularizing the eigenvectors of the matrix logarithm (Deng & Tsui, 2013; Yu, Wang & Zhu, 2017). However, the above-mentioned methods do not exploit the sparse structure of the covariance matrix. A sparse covariance estimate can be useful in subsequent inference, such as inferring the correlation pattern among the variables. Bickel & Levina (2009) proposed to threshold the small entries of the sample covariance matrix to zeroes and studied the theoretical behaviour of their method when the number of variables is large. Rothman, Levina & Zhu (2009) suggested thresholding the sample covariance matrix with more general thresholding functions. Wagaman & Levina (2009) introduced a method of sparse estimation for a covariance matrix with banded structure based on the correlations between variables using the Isomap. Cai & Yuan (2012) suggested estimating a covariance matrix using block thresholding. Their estimator is constructed by dividing the sample covariance matrix into blocks and then simultaneously estimating the entries in a block using thresholding. However, the threshold-based estimator is not guaranteed to be positive definite. To make the resulting estimate both sparse and positive definite simultaneously, Bien & Tibshirani (2011) proposed using a penalized likelihood method with a Lasso penalty (Tibshirani, 1996) on the entries in the covariance matrix. Their idea is similar to the graphical Lasso for inverse covariance matrix estimation found in the literature (Yuan & Lin, 2007; Friedman, Hastie & Tibshirani, 2008; Rocha, Zhao & Yu, 2008; Rothman et al., 2008; Yuan, 2008, 2010; Deng & Yuan, 2009), but the computation is much more complicated due to the non-convexity of the objective function. Xue, Ma & Zou (2012) developed a sparse covariance matrix estimator for high-dimensional data based on a convex objective function with positive definite constraint and L_1 penalty. They also derived a fast algorithm to solve the constraint optimization problem. Additional research on the problem of estimating a high-dimensional covariance matrix can be found in Fan, Liao & Mincheva (2013), Liu, Wang & Zhao (2014), Xiao et al. (2016), Cai, Ren & Zhou (2016), Huang, Farewell & Pan (2017) and Kang, Xie & Wang (2020). A comprehensive review of the development of covariance matrix estimation can be found in Pourahmadi (2013) and Fan, Liao & Liu (2016).

Another direction of sparse covariance estimation is to take advantage of matrix decomposition. One popular and effective tool is the modified Cholesky decomposition (MCD) (Pourahmadi, 1999; Wu & Pourahmadi, 2003; Pourahmadi, Daniels & Park, 2007; Rothman, Levina & Zhu, 2009; Dellaportas & Pourahmadi, 2012; Xue, Ma & Zou, 2012; Rajaratnam & Salzman, 2013). This method assumes that the variables have a natural order which enables them to be sequentially orthogonalized to re-parameterize the covariance matrix. By imposing a certain sparse pattern on the Cholesky factor, we obtain the sparse structure in the estimated covariance matrix. For example, Huang et al. (2006) considered imposing an L_1 (Lasso) penalty on the entries in the Cholesky factor. Rothman, Levina & Zhu (2010) suggested using a banded

estimator of the Cholesky factor, which can be obtained by regressing each variable on only its closest k predecessors. However, the MCD-based approach for estimating a covariance matrix depends on the order of the variables, and such a pre-specification may not always be appropriate. A natural order of variables is often not available or cannot be pre-determined in many applications such as those involving gene expression or stock market pricing data.

In this article, we adopt the MCD approach for estimating a large covariance matrix, but resolve the drawback of order dependency in this method using the permutation idea suggested by Zheng et al. (2017). By considering a set of covariance estimates under different orders of variables in the MCD, Zheng et al. (2017) introduced an order-averaged estimator for a large covariance matrix which is positive definite. However, their approach cannot ensure that the resulting estimate is sparse. In addition, Zheng et al. (2017) did not provide any theoretical results. To overcome these drawbacks, we address the order issue and ensure that the resulting estimate is suitably sparse. We also show that the estimator of Zheng et al. (2017) represents a special case of our proposed estimator when the penalty tuning parameter in the objective function is set equal to zero. It is worth remarking that it is not straightforward to simultaneously address both the order issue and the sparsity of the covariance matrix within the framework of MCD. To achieve these two goals we first obtain a collection of estimates of a covariance matrix from different orders of variables using the permutation idea. With such estimates, our proposed estimator is obtained as the "centre" of this collection under the Frobenius norm through an L_1 penalized objective function, where the L_1 regularization is imposed to achieve the sparsity of the estimate. We also develop an efficient algorithm that makes the computation of our estimator attractive. Furthermore, under certain regularity conditions we establish the consistency of our proposed estimator with respect to the Frobenius norm.

The remainder of this article is organized as follows. Section 2 briefly reviews the MCD method of estimating a covariance matrix. Section 3 introduces our proposed method for addressing the order issue. We also outline an efficient algorithm to solve the objective function. In Section 4, we identify the theoretical properties of our proposed method of estimation. A simulation study and a practical example are reported in Sections 5 and 6, respectively. Some summary comments are provided in Section 7.

2. A REVIEW OF MCD

Suppose that $\boldsymbol{X} = (X_1, \ldots, X_p)'$ is a p-dimensional random vector with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be n independent and identically distributed observations following $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Pourahmadi (1999) proposed the MCD to estimate a covariance matrix, which is statistically meaningful and guarantees the estimate is positive definite. This decomposition arises from regressing each variable X_j on its predecessors $X_1, \ldots, X_{j-1}$ for $2 \leq j \leq p$. Specifically, consider fitting a series of regression models

$$X_j = \sum_{k=1}^{j-1} (-t_{jk}) X_k + \epsilon_j = \hat{X}_j + \epsilon_j,$$

where ϵ_j is the error term for the jth regression model with $E\epsilon_j = 0$ and $Var(\epsilon_j) = d_j^2$. Let $\epsilon_1 = X_1$ and $\boldsymbol{D} = diag(d_1^2, \ldots, d_p^2)$ be the diagonal covariance matrix of $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_p)'$. Construct the unit lower triangular matrix $\boldsymbol{T} = (t_{jk})_{p \times p}$ with ones on its diagonal and regression coefficients $(t_{j1}, \ldots, t_{j,j-1})'$ as its jth row. It follows that

$$\boldsymbol{D} = Var(\boldsymbol{\epsilon}) = Var(\boldsymbol{X} - \hat{\boldsymbol{X}}) = Var(\boldsymbol{TX}) = \boldsymbol{T\Sigma T}',$$

and thus

$$\boldsymbol{\Sigma} = \boldsymbol{T}^{-1}\boldsymbol{D}\boldsymbol{T}'^{-1}. \tag{1}$$

Consequently, the MCD method reduces the challenge of estimating a covariance matrix to the task of fitting $(p-1)$ linear regressions, and is applicable in high-dimensional settings. However, directly imposing a sparse structure on the Cholesky factor matrix $\boldsymbol{T}$ in Equation (1) does not imply that $\boldsymbol{\Sigma}$ must be sparse since Equation (1) involves an inverse of $\boldsymbol{T}$ and hence fails to ensure that the resulting estimate of $\boldsymbol{\Sigma}$ will be sparse, as required. Alternatively, one could consider using a latent variable regression model based on the MCD. Writing $\boldsymbol{X} = \boldsymbol{L}\boldsymbol{\epsilon}$ would lead to

$$Var(\boldsymbol{X}) = Var(\boldsymbol{L}\boldsymbol{\epsilon})$$
$$\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}'.$$

This decomposition can be interpreted as resulting from a new sequence of regressions, where each variable X_j is regressed on all the previous latent variables $\epsilon_1, \ldots, \epsilon_{j-1}$ rather than the $\boldsymbol{X}$'s themselves. This approach results in a sequence of regression models

$$X_j = \boldsymbol{l}_j'\boldsymbol{\epsilon} = \sum_{k<j} l_{jk}\epsilon_k + \epsilon_j, \quad j = 2, \ldots, p,$$

where $\boldsymbol{l}_j = (l_{jk})$ is the jth row of $\boldsymbol{L}$. Here $l_{jj} = 1$ and $l_{jk} = 0$ for $k > j$.

With the data matrix $\mathbb{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$, define its jth column to be $\boldsymbol{x}^{(j)}$. Denote by $\boldsymbol{e}^{(j)}$ the residuals of $\boldsymbol{x}^{(j)}$ for $j \geq 2$, and $\boldsymbol{e}^{(1)} = \boldsymbol{x}^{(1)}$. Let $\mathbb{Z}^{(j)} = (\boldsymbol{e}^{(1)}, \ldots, \boldsymbol{e}^{(j-1)})$ be the matrix containing the first $(j-1)$ residuals. Now the Lasso regularization (Tibshirani, 1996) can be used to induce sparsity in $\hat{\boldsymbol{L}}$ (Huang et al., 2006; Rothman, Levina & Zhu, 2010; Chang & Tsay, 2010; Kang et al., 2019), i.e.,

$$\hat{\boldsymbol{l}}_j = \arg\min_{\boldsymbol{l}_j} \|\boldsymbol{x}^{(j)} - \mathbb{Z}^{(j)}\boldsymbol{l}_j\|_2^2 + \eta_j\|\boldsymbol{l}_j\|_1, \ j = 2, \ldots, p, \tag{2}$$

where $\eta_j \geq 0$ is a tuning parameter and selected by cross-validation. The symbol $\|\cdot\|_1$ stands for the vector L_1 norm. The variable $\boldsymbol{e}^{(j)} = \boldsymbol{x}^{(j)} - \mathbb{Z}^{(j)}\boldsymbol{l}_j$ is used to construct the residuals for the last column of $\mathbb{Z}^{(j+1)}$. Then d_j^2 is estimated via

$$\hat{d}_j^2 = \widehat{Var}(\hat{\boldsymbol{e}}^{(j)}) = \widehat{Var}(\boldsymbol{x}^{(j)} - \mathbb{Z}^{(j)}\hat{\boldsymbol{l}}_j), \tag{3}$$

the sample variance of $\boldsymbol{e}^{(j)}$, when constructing matrix $\hat{\boldsymbol{D}} = diag(\hat{d}_1^2, \ldots, \hat{d}_p^2)$. As a result, $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{L}}\hat{\boldsymbol{D}}\hat{\boldsymbol{L}}'$ will be a sparse covariance matrix estimate.

3. THE PROPOSED METHOD

Clearly, the estimate $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{L}}\hat{\boldsymbol{D}}\hat{\boldsymbol{L}}'$ depends on the order of the variables $X_1, \ldots, X_p$, which means that different orders would lead to different estimates of $\boldsymbol{\Sigma}$. To address this order-dependence issue, we consider an order-averaged method of estimating $\boldsymbol{\Sigma}$ by exploiting the idea of permutation. Specifically, generate M different permutations of $\{1, \ldots, p\}$ as orders of the variables, denoted by π_k's, $k = 1, \ldots, M$. Let $\boldsymbol{P}_{\pi_k}$ be the corresponding permutation matrix. Under an order π_k, the corresponding estimate of $\boldsymbol{\Sigma}$ is $\hat{\boldsymbol{\Sigma}}_{\pi_k} = \hat{\boldsymbol{L}}_{\pi_k}\hat{\boldsymbol{D}}_{\pi_k}\hat{\boldsymbol{L}}'_{\pi_k}$, where $\hat{\boldsymbol{L}}_{\pi_k}$ and $\hat{\boldsymbol{D}}_{\pi_k}$ are calculated using Equations (2) and (3). Then transforming back to the original variable order, we have

$$\hat{\boldsymbol{\Sigma}}_k = \boldsymbol{P}_{\pi_k}\hat{\boldsymbol{\Sigma}}_{\pi_k}\boldsymbol{P}'_{\pi_k}.$$

To obtain a proper estimator for $\boldsymbol{\Sigma}$, Zheng et al. (2017) proposed $\bar{\boldsymbol{\Sigma}} = \frac{1}{M}\sum_{k=1}^{M}\hat{\boldsymbol{\Sigma}}_k$; however, the resulting estimator is clearly not sparse since any sparse structure in $\hat{\boldsymbol{\Sigma}}_k$ is destroyed by computing the average.

In order to simultaneously achieve a sparse estimate that is also positive definite, we propose using

$$\hat{\boldsymbol{\Sigma}} = \arg\min_{\boldsymbol{\Sigma}\succeq\nu\boldsymbol{I}} \frac{1}{2}\sum_{k=1}^{M}\|\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}}_k\|_F^2 + \tilde{\lambda}|\boldsymbol{\Sigma}|_1, \tag{4}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, $\tilde{\lambda}\geq 0$ is a tuning parameter, and $|\cdot|_1$ is the L_1 norm for all the off-diagonal elements. Here ν is some positive arbitrarily small number. The constraint $\boldsymbol{\Sigma}\succeq\nu\boldsymbol{I}$ is introduced to guarantee that $\hat{\boldsymbol{\Sigma}}$ is positive definite, whereas the penalty term ensures that $\hat{\boldsymbol{\Sigma}}$ is suitably sparse. It is worth pointing out that, if $\tilde{\lambda}=0$ in (4), the solution for $\hat{\boldsymbol{\Sigma}}$ would be $\bar{\boldsymbol{\Sigma}} = \frac{1}{M}\sum_{k=1}^{M}\hat{\boldsymbol{\Sigma}}_k$, which is the estimator of Zheng et al. (2017). If we omit the constraint $\boldsymbol{\Sigma}\succeq\nu\boldsymbol{I}$ in (4), the solution for $\hat{\boldsymbol{\Sigma}}$ would be the soft-threshold estimate of $\bar{\boldsymbol{\Sigma}}$. The objective (4) is similar to that adopted in Xue, Ma & Zou (2012), but their implications are different. Xue, Ma & Zou (2012) used the sample covariance matrix $\boldsymbol{S}$ instead of $\hat{\boldsymbol{\Sigma}}_k$ in (4). Hence, their estimate can be considered to attain the minimum distance from $\boldsymbol{S}$ with respect to the Frobenius norm. However, our proposed estimator minimizes the averaged distance to all $\hat{\boldsymbol{\Sigma}}_k$'s, while ensuring that $\hat{\boldsymbol{\Sigma}}$ is both positive definite and sparse. Thus, $\hat{\boldsymbol{\Sigma}}$ can be interpreted as the "centre" of estimates $\hat{\boldsymbol{\Sigma}}_k$'s with respect to the Frobenius norm. As evidenced in the simulation study we report in Section 5, our proposed estimator can be more accurate than the estimator proposed by Xue, Ma & Zou (2012).

For ease of theoretical deduction, we re-write Equation (4) as

$$\hat{\boldsymbol{\Sigma}} = \arg\min_{\boldsymbol{\Sigma}\succeq\nu\boldsymbol{I}} \frac{1}{2M}\sum_{k=1}^{M}\|\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}}_k\|_F^2 + \lambda|\boldsymbol{\Sigma}|_1, \tag{5}$$

where $\lambda=\tilde{\lambda}/M$. To efficiently solve the optimization problem (5), we employ the alternating direction method of multipliers (ADMM) (Boyd et al., 2011), which has been widely used in solving the convex optimization of L_1 penalized covariance matrix estimation. Let us first introduce a new variable $\boldsymbol{\Phi}$ and an equality constraint via

$$(\hat{\boldsymbol{\Sigma}},\hat{\boldsymbol{\Phi}}) = \arg\min_{\boldsymbol{\Sigma},\boldsymbol{\Phi}}\left\{\frac{1}{2M}\sum_{k=1}^{M}\|\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}}_k\|_F^2 + \lambda|\boldsymbol{\Sigma}|_1 \,:\, \boldsymbol{\Sigma}=\boldsymbol{\Phi},\boldsymbol{\Phi}\succeq\nu\boldsymbol{I}\right\}. \tag{6}$$

Note that the solution of (6) provides solution for the corresponding problem in (5). To solve the former problem, we minimize its augmented Lagrangian function

$$\begin{aligned} L(\boldsymbol{\Sigma},\boldsymbol{\Phi};\boldsymbol{\Lambda}) =& \frac{1}{2M}\sum_{k=1}^{M}\|\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}}_k\|_F^2 + \lambda|\boldsymbol{\Sigma}|_1 \\ & - \langle\boldsymbol{\Lambda},\boldsymbol{\Phi}-\boldsymbol{\Sigma}\rangle + \frac{1}{2\tau}\|\boldsymbol{\Phi}-\boldsymbol{\Sigma}\|_F^2 \end{aligned} \tag{7}$$

for some given penalty parameter τ, where $\boldsymbol{\Lambda}$ is the Lagrangian multiplier. The notation $\langle\cdot,\cdot\rangle$ represents the matrix inner product $\langle\boldsymbol{A},\boldsymbol{B}\rangle = \sum_{i,j}a_{ij}b_{ij}$, where a_{ij} and b_{ij} are the elements of

matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. The ADMM iteratively solves the following steps sequentially for $i = 0, 1, 2, \ldots$ till convergence

$$\boldsymbol{\Phi} \text{ step} : \boldsymbol{\Phi}^{i+1} = \arg\min_{\boldsymbol{\Phi} \geq \nu \boldsymbol{I}} L(\boldsymbol{\Sigma}^i, \boldsymbol{\Phi}; \boldsymbol{\Lambda}^i) \tag{8}$$

$$\boldsymbol{\Sigma} \text{ step} : \boldsymbol{\Sigma}^{i+1} = \arg\min_{\boldsymbol{\Sigma}} L(\boldsymbol{\Sigma}, \boldsymbol{\Phi}^{i+1}; \boldsymbol{\Lambda}^i) \tag{9}$$

$$\boldsymbol{\Lambda} \text{ step} : \boldsymbol{\Lambda}^{i+1} = \boldsymbol{\Lambda}^i - \frac{1}{\tau}(\boldsymbol{\Phi}^{i+1} - \boldsymbol{\Sigma}^{i+1}). \tag{10}$$

Assume the eigenvalue decomposition of a matrix $\boldsymbol{A}$ is $\sum_{i=1}^{p} \lambda_i \boldsymbol{\xi}' \boldsymbol{\xi}$, and define $(\boldsymbol{A})_+ = \sum_{i=1}^{p} \max(\lambda_i, \nu) \boldsymbol{\xi}_i' \boldsymbol{\xi}_i$. Then we develop the closed form for the $\boldsymbol{\Phi}$ step in Equation (8) as

$$\begin{aligned} \frac{\partial L(\boldsymbol{\Sigma}^i, \boldsymbol{\Phi}; \boldsymbol{\Lambda}^i)}{\partial \boldsymbol{\Phi}} &= -\boldsymbol{\Lambda}^i + \frac{1}{\tau}(\boldsymbol{\Phi} - \boldsymbol{\Sigma}^i) \triangleq 0 \\ \boldsymbol{\Phi} &= \boldsymbol{\Sigma}^i + \tau \boldsymbol{\Lambda}^i \\ \boldsymbol{\Phi}^{i+1} &= (\boldsymbol{\Sigma}^i + \tau \boldsymbol{\Lambda}^i)_+. \end{aligned}$$

Next, define an element-wise soft threshold for each entry z_{ij} in matrix $\boldsymbol{Z}$ as $\boldsymbol{s}(\boldsymbol{Z}, \delta) = \{s(z_{ij}, \delta)\}_{1 \leq i,j \leq p}$ with

$$s(z_{ij}, \delta) = \text{sign}(z_{ij}) \max(|z_{ij}| - \delta, 0) I_{\{i \neq j\}} + z_{ij} I_{\{i=j\}}.$$

Then the solution of Equation (9) is

$$\begin{aligned} \frac{\partial L(\boldsymbol{\Sigma}, \boldsymbol{\Phi}^{i+1}; \boldsymbol{\Lambda}^i)}{\partial \boldsymbol{\Sigma}} &= \frac{1}{M} \sum_{k=1}^{M} (\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}_k) + \boldsymbol{\Lambda}^i + \frac{1}{\tau}(\boldsymbol{\Sigma} - \boldsymbol{\Phi}^{i+1}) + \lambda \text{sign}^*(\boldsymbol{\Sigma}) \triangleq 0 \\ (\tau + 1)\boldsymbol{\Sigma} &= \tau\left(\frac{1}{M} \sum_{k=1}^{M} \hat{\boldsymbol{\Sigma}}_k - \boldsymbol{\Lambda}^i\right) + \boldsymbol{\Phi}^{i+1} - \lambda\tau \text{sign}^*(\boldsymbol{\Sigma}) \\ \boldsymbol{\Sigma}^{i+1} &= \{s(\tau(\frac{1}{M} \sum_{k=1}^{M} \hat{\boldsymbol{\Sigma}}_k - \boldsymbol{\Lambda}^i) + \boldsymbol{\Phi}^{i+1}, \lambda\tau)\}/(\tau + 1), \end{aligned}$$

where $\text{sign}^*(\boldsymbol{\Sigma})$ means $\text{sign}(\boldsymbol{\Sigma})$ with the diagonal elements replaced by the $\boldsymbol{0}$ vector. Algorithm 1 summarizes the procedure we have outlined above to solve Equation (5) using the ADMM technique.

Algorithm 1.

Step 1 *: Input initial values* $\boldsymbol{\Sigma}_{init}$, $\boldsymbol{\Lambda}_{init}$ *and* τ.
Step 2 *:* $\boldsymbol{\Phi}^{i+1} = (\boldsymbol{\Sigma}^i + \tau \boldsymbol{\Lambda}^i)_+$.
Step 3 *:* $\boldsymbol{\Sigma}^{i+1} = \{s(\tau(\frac{1}{M} \sum_{k=1}^{M} \hat{\boldsymbol{\Sigma}}_k - \boldsymbol{\Lambda}^i) + \boldsymbol{\Phi}^{i+1}, \lambda\tau)\}/(\tau + 1)$.
Step 4 *:* $\boldsymbol{\Lambda}^{i+1} = \boldsymbol{\Lambda}^i - \frac{1}{\tau}(\boldsymbol{\Phi}^{i+1} - \boldsymbol{\Sigma}^{i+1})$.
Step 5 *: Repeat Step 2–4 until the algorithm converges numerically.*

This algorithm converges quickly and produces the optimal solution $\arg\min L(\boldsymbol{\Sigma}, \boldsymbol{\Phi}; \boldsymbol{\Lambda})$ in Equation (7). In practice, the initial value $\boldsymbol{\Sigma}_{init}$ is set equal to $\bar{\boldsymbol{\Sigma}} = \frac{1}{M}\sum_{k=1}^{M} \hat{\boldsymbol{\Sigma}}_k$. The initial value $\boldsymbol{\Lambda}_{init}$ is set equal to the zero matrix, and $\tau = 2$ as well as $\nu = 10^{-4}$. The optimal value of the tuning parameter λ in Equation (7) is chosen based on the Bayesian information criterion (BIC) (Yuan & Lin, 2007)

$$\mathrm{BIC}(\lambda) = -\log|\hat{\boldsymbol{\Sigma}}_\lambda^{-1}| + \mathrm{tr}\left[\hat{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{S}\right] + \frac{\log n}{n}\sum_{i\le j}\hat{e}_{ij}(\lambda),$$

where $\boldsymbol{S}$ is the sample covariance matrix, $\hat{\boldsymbol{\Sigma}}_\lambda = (\hat{\sigma}_{ij}^{(\lambda)})_{p\times p}$ indicates the estimate of $\boldsymbol{\Sigma}$ obtained by applying Algorithm 1 with tuning parameter λ. The quantities $\hat{e}_{ij}(\lambda) = 0$ if $\hat{\sigma}_{ij}^{(\lambda)} = 0$, and $\hat{e}_{ij}(\lambda) = 1$ otherwise.

4. THEORETICAL CONVERGENCE

In this section, Theorem 1 states that the sequence $(\boldsymbol{\Sigma}^i, \boldsymbol{\Phi}^i, \boldsymbol{\Lambda}^i)$ generated by Algorithm 1 from any initial value converges numerically to an optimal minimizer $(\hat{\boldsymbol{\Sigma}}^+, \hat{\boldsymbol{\Phi}}^+, \hat{\boldsymbol{\Lambda}}^+)$ of Equation (7), where $\hat{\boldsymbol{\Lambda}}^+$ is the optimal dual variable. Theorem 2 demonstrates that our proposed estimator is asymptotically consistent under certain regularity conditions. The details of the proofs of Theorems 1 and 2 can be found in the Appendix. To facilitate the presentation of the proofs, we first introduce some notation. Define a $2p$ by $2p$ matrix $\boldsymbol{J}$ as

$$\boldsymbol{J} = \begin{pmatrix} \tau\boldsymbol{I}_{p\times p} & 0 \\ 0 & \tau^{-1}\boldsymbol{I}_{p\times p} \end{pmatrix}.$$

Let the notation $\|\cdot\|_J^2$ be $\|\boldsymbol{U}\|_J^2 = \langle \boldsymbol{U}, \boldsymbol{J}\boldsymbol{U}\rangle$ and $\langle \boldsymbol{U}, \boldsymbol{V}\rangle_J = \langle \boldsymbol{U}, \boldsymbol{J}\boldsymbol{V}\rangle$. Let $\boldsymbol{\Sigma}_0 = (\sigma_{ij}^0)_{p\times p} = \boldsymbol{L}_0\boldsymbol{D}_0\boldsymbol{L}_0'$ be the true covariance matrix for the observations $\mathbb{X} = (x_{ij})_{n\times p}$, and define the number of non-zero off-diagonal elements of $\boldsymbol{\Sigma}_0$ as s_0. Denote the maximal true variance of $\boldsymbol{\Sigma}_0$ by $\sigma_{\max}$. Let $Z_{\pi_k} = \{(j,k) : k < j, l_{0jk}^{(\pi_k)} \neq 0\}$ be the collection of non-zero elements in the lower triangular part of the matrix $\boldsymbol{L}_{0\pi_k}$. Denote by s_1 the maximum of the cardinality of Z_{π_k} for $k = 1, 2, \ldots, M$. We now state the following lemmas, together with Theorems 1 and 2.

Lemma 1. *Assume that* $(\hat{\boldsymbol{\Sigma}}^+, \hat{\boldsymbol{\Phi}}^+)$ *is an optimal solution of Equation (6) and* $\hat{\boldsymbol{\Lambda}}^+$ *is the corresponding optimal dual variable with the equality constraint* $\boldsymbol{\Sigma} = \boldsymbol{\Phi}$. *Then the sequence* $(\boldsymbol{\Sigma}^i, \boldsymbol{\Phi}^i, \boldsymbol{\Lambda}^i)$ *generated by Algorithm 1 satisfies*

$$\|\boldsymbol{W}^+ - \boldsymbol{W}^i\|_J^2 - \|\boldsymbol{W}^+ - \boldsymbol{W}^{i+1}\|_J^2 \ge \|\boldsymbol{W}^i - \boldsymbol{W}^{i+1}\|_J^2,$$

where $\boldsymbol{W}^+ = (\hat{\boldsymbol{\Lambda}}^+, \hat{\boldsymbol{\Sigma}}^+)'$ *and* $\boldsymbol{W}^i = (\boldsymbol{\Lambda}^i, \boldsymbol{\Sigma}^i)'$.

Theorem 1 (Algorithmic convergence). *Suppose* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ *are n independent and identically distributed observations from* $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$. *Then the sequence* $(\boldsymbol{\Sigma}^i, \boldsymbol{\Phi}^i, \boldsymbol{\Lambda}^i)$ *generated by Algorithm 1 from any initial value converges numerically to an optimal minimizer of the objective function (7).*

Theorem 1 demonstrates the convergence of Algorithm 1. It automatically indicates that the sequence $\boldsymbol{\Sigma}^i, i = 1, 2, \ldots$, produced by Algorithm 1 converges to an optimal solution of the objective (5). We prove Lemma 1 and Theorem 1 following the ideas of Xue, Ma & Zou (2012) via the Karush–Kuhn–Tucker conditions (Karush, 1939; Kuhn & Tucker, 1951). The proofs are outlined in the Appendix.

In order to establish the asymptotic consistency of our proposed estimator, we assume that there exists a constant $\theta > 1$ such that the singular values of the true covariance matrix are bounded, i.e.,

$$1/\theta < sv_p(\boldsymbol{\Sigma}_0) \leq sv_1(\boldsymbol{\Sigma}_0) < \theta, \tag{11}$$

where we use $sv_1(A), sv_2(A), \ldots, sv_p(A)$ to indicate the singular values of matrix $\boldsymbol{A}$ in decreasing order. They are the square roots of the eigenvalues of matrix $\boldsymbol{AA'}$. This same assumption was used in Rothman et al. (2008), Lam & Fan (2009) and Guo et al. (2011), which guarantees the positive definiteness and makes inverting the covariance matrix meaningful. The following lemma and theorem establish that our proposed estimator is asymptotically consistent with respect to the Frobenius norm.

Lemma 2. *Let $\boldsymbol{\Sigma}_0 = \boldsymbol{L}_0\boldsymbol{D}_0\boldsymbol{L}_0^{'}$ be the MCD of the true covariance matrix. If the singular values of $\boldsymbol{\Sigma}_0$ are bounded, so that there exist constants θ_1 and θ_2 such that $0 < \theta_1 < sv_p(\boldsymbol{\Sigma}_0) \leq sv_1(\boldsymbol{\Sigma}_0) < \theta_2 < \infty$, then there exist constants h_1 and h_2 such that*

$$h_1 < sv_p(\boldsymbol{L}_0) \leq sv_1(\boldsymbol{L}_0) < h_2,$$

and

$$h_1 < sv_p(\boldsymbol{D}_0) \leq sv_1(\boldsymbol{D}_0) < h_2.$$

Lemma 3. *Suppose $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are n independent and identically distributed observations from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$. Let $\boldsymbol{\Sigma}_{0\pi_k} = \boldsymbol{L}_{0\pi_k}\boldsymbol{D}_{0\pi_k}\boldsymbol{L}_{0\pi_k}^{'}$ be the MCD of the true covariance matrix resulting from an order of variables π_k. Under (11), assume that the tuning parameters η_j in Equation (2) satisfy $\sum_{j=1}^{p}\eta_j = O(\sqrt{\log(p)/n})$ and $(s_1 + p)\log(p) = o(n)$; then*

$$\|\hat{\boldsymbol{L}}_{\pi_k} - \boldsymbol{L}_{0\pi_k}\|_F \xrightarrow{P} 0 \quad \textit{and} \quad \|\hat{\boldsymbol{D}}_{\pi_k} - \boldsymbol{D}_{0\pi_k}\|_F \xrightarrow{P} 0.$$

Lemma 3 demonstrates the asymptotical convergence of the Cholesky factor matrices $\hat{\boldsymbol{L}}_{\pi_k}$ and $\hat{\boldsymbol{D}}_{\pi_k}$. Based on this result, we can derive the theoretical property of $\hat{\boldsymbol{\Sigma}}_{\pi_k}$ under variable order π_k, which is then used to prove the following theorem.

Theorem 2 (Asymptotic convergence). *Assume all the conditions in Lemma 3 hold, and $\lambda = o((s_0 + p)^{-1/2})$. Under the condition that for all $|t| \leq \rho$ and $1 \leq i \leq n, 1 \leq j \leq p$*

$$E\{\exp(tx_{ij}^2)\} \leq K.$$

For any $m > 0$, set

$$\lambda = c_0^2\frac{\log\ p}{n} + c_1\left(\frac{\log\ p}{n}\right)^{1/2},$$

where

$$c_0 = \frac{1}{2}eK\rho^{1/2} + \rho^{-1/2}(m+1)$$

and

$$c_1 = 2K\left(\rho^{-1} + \frac{1}{4}\rho\sigma_{\max}^2\right)\ \exp\left(\frac{1}{2}\rho\sigma_{\max}\right)\ +\ 2\rho^{-1}(m+2).$$

Then $\|\hat{\boldsymbol{\Sigma}}^{+} - \boldsymbol{\Sigma}_0\|_F \xrightarrow{P} 0$.

Theorem 2 demonstrates that our proposed estimator is asymptotically consistent with respect to the Frobenius norm under certain regularity conditions. Together with Theorem 1, this property implies that the estimate obtained from Algorithm 1 is consistent. We would like to remark that the constraint $\mathbf{\Sigma} \succeq \nu \boldsymbol{I}$ could increase the computational cost of iterations in Algorithm 1; however, it guarantees that our proposed estimator is positive definite.

Without this constraint, the solution of the optimization problem (4) would become the soft-threshold estimate of $\bar{\mathbf{\Sigma}}$. Moreover, such a constraint helps to establish the convergence of Algorithm 1 as well as the proposed estimator. For details, see the proof outlined in the Appendix.

5. SIMULATION STUDY

In this section, we conduct a comprehensive simulation study to evaluate the performance of the proposed method. We consider the following five covariance matrix structures:

Model 1. $\mathbf{\Sigma}_1$ = MA(0.5, 0.3), where MA stands for "moving average." The diagonal elements are 1 with the first sub-diagonal elements equal to 0.5 and the second sub-diagonal elements equal to 0.3.

Model 2. $\mathbf{\Sigma}_2$ = AR(0.5), where AR stands for "autoregressive." The conditional covariance between any two random variables X_i and X_j equals $0.5^{|i-j|}$, $1 \leq i, j \leq p$.

Model 3. $\mathbf{\Sigma}_3$ is generated by randomly permuting the rows and corresponding columns of $\mathbf{\Sigma}_1$.

Model 4. $\mathbf{\Sigma}_4$ is generated by randomly permuting the rows and corresponding columns of $\mathbf{\Sigma}_2$.

Model 5. $\mathbf{\Sigma}_5 = \mathbf{\Theta} + \alpha \boldsymbol{I}$, where the diagonal elements of $\mathbf{\Theta}$ are zeroes and $\mathbf{\Theta}_{ij} = \mathbf{\Theta}_{ji} = b * Unif(-1, 1)$ for $i \neq j$, where b has a Bernoulli distribution and equals 1 with probability 0.15, or 0 otherwise. Each off-diagonal element of $\mathbf{\Theta}$ is generated independently. The value of α is gradually increased to ensure that $\mathbf{\Sigma}_5$ is positive definite.

Note that Models 1 and 2 represent a banded or nearly banded structure for the covariance matrix, whereas the covariance matrices of Models 3 and 4 do not have structured sparsity due to the random permutations. Model 5 is a more general sparse matrix with no particular structure. Hence from the perspective of sparse pattern, Model 5 represents the most general case and Models 1 and 2 correspond to the least general cases. For each case, we generate the data independently from the normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ with three settings of different sample sizes and variable sizes: (1) $n = 50$, $p = 30$; (2) $n = 50$, $p = 50$ and (3) $n = 50$, $p = 100$. For the implementation of the proposed method in this study, we choose to set $M = 100$ in the simulation. We have also tried $M = 10, 30, 50, 100$ and 150 as the number of randomly selected permutations from all $p!$ possible permutations. The resulting performances appeared to be marginally better when M is larger than 30. Please refer to Kang & Deng (2020) for a detailed discussion and justification of the choice of M. In practice, we would suggest to choose a relatively large value of M to ensure the accuracy of the estimate, provided the computational resources are available. Otherwise, a moderate value of M is recommended to balance the accuracy and computation efficiency for the proposed model.

The performance of the proposed estimator is examined in comparison with several other approaches, which are divided into three classes. The first class is the sample covariance matrix $\boldsymbol{S}$ that serves as the benchmark. The second class is composed of three methods that deal with the variable order used in the MCD, including the MCD-based method with BIC order selection (BIC) (Dellaportas & Pourahmadi, 2012), the best permutation algorithm (BPA) (Rajaratnam & Salzman, 2013) and the proposed method (Proposed). The third class of competing methods consists of five approaches, including Bien and Tibshirani's estimate (BT) (Bien & Tibshirani, 2011), Bickel and Levina's estimate (BL) (Bickel & Levina, 2009), Xue, Ma and Zou's estimate

(XMZ) (Xue, Ma & Zou, 2012), Wagaman and Levina's Isoband estimate (IB) (Wagaman & Levina, 2009) and Rothman et al.'s estimate (RLZ) (Rothman, Levina & Zhu, 2010).

To assess the accuracy of the covariance matrix estimates $\hat{\mathbf{\Sigma}} = (\hat{\sigma}_{ij})_{p\times p}$ obtained from each approach, we use the F norm, entropy loss (EN), L_1 norm and mean absolute error (MAE), defined as

$$\mathrm{F} = \sqrt{\sum_{i=1}^{p}\sum_{j=1}^{p}(\hat{\sigma}_{ij} - \sigma_{ij})^2},$$

$$\mathrm{EN} = \mathrm{tr}[\mathbf{\Sigma}^{-1}\hat{\mathbf{\Sigma}}] - \log|\mathbf{\Sigma}^{-1}\hat{\mathbf{\Sigma}}| - p,$$

$$L_1 \text{ norm} = \max_j \sum_i |\hat{\sigma}_{ij} - \sigma_{ij}|,$$

$$\mathrm{MAE} = \frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{p}|\hat{\sigma}_{ij} - \sigma_{ij}|.$$

In addition, to gauge the performance concerning sparsity, we use the notion of false selection loss (FSL), which is the summation of false positive (FP) and false negative (FN). Here we say a FP occurs if a non-zero element in the true matrix is incorrectly estimated as a zero. Similarly, a FN occurs if a zero element in the true matrix is incorrectly identified as a non-zero. The FSL is computed in percentage as $(\mathrm{FP} + \mathrm{FN})/p^2$, expressed as a percentage value. For each loss function mentioned above, Tables 1 and 2 and Tables A1–A3 in the Appendix report the average values of the performance measures and their corresponding standard errors in the parentheses over 100 replicates. For each model, the two methods with lowest averages for each measure are shown in bold. Dashed lines in the tables represent cases where the corresponding values could not be determined due to singularity of the estimated matrix.

For a short summary of the numerical results, it shows that the proposed method generally provides more reliable estimates of a large covariance matrix than other approaches in comparison. It is able to accurately reflect the underlying sparse structure of the covariance matrix. Although the IB method exhibits good estimation performance, it could not ensure that the resulting estimate was positive definite. When the underlying covariance matrix is banded or tapered, the proposed method is not as good as the RLZ. The reason for this result is that the RLZ method targets a banded covariance matrix. When the underlying structure of covariance matrix is more general without any specification, the proposed method still performs well. Furthermore, the advantage of the proposed method is even more evident in the high-dimensional cases.

We first analyze the performance results and demonstrate the mechanism of several methods from the perspective of covariance structures using F loss. Since the IB method assumes the true matrix has banded structure after re-ordering the variables, this method exhibits good performance for Models 1–4, since Models 1 and 2 are banded matrices and Models 3 and 4 also possess banded structure after certain permutations of the variables. However, the IB method is inferior to the proposed method for Model 5, since this case represents a general sparse covariance matrix with no particular banded structure, even if we permute the variable order. The RLZ method performs well for Models 1 and 2 under F loss, since this method is designated to estimate the banded or tapered matrices. But it is not suitable for Models 3, 4 and 5. In addition, we observe that the BPA method yields relatively low F loss for Model 4. The reason is that this method is good at recovering the variable order for the AR models. Therefore, although the BPA, IB and RLZ methods all perform well for the banded or tapered matrices, they are inferior to the proposed model when the true covariance is a general matrix with no sparse pattern.

TABLE 1: The averages and standard errors of estimates for Model 1.

		F	EN	L_1	MAE	FSL (%)
$p = 30$	***S***	4.43 (0.05)	12.42 (0.08)	5.21 (0.07)	3.51 (0.03)	83.96 (0.01)
	BIC	3.25 (0.03)	7.22 (0.08)	2.96 (0.04)	1.76 (0.02)	52.32 (0.55)
	BPA	2.98 (0.03)	**6.02 (0.09)**	2.74 (0.05)	1.55 (0.02)	46.53 (0.68)
	BT	4.74 (0.01)	7.78 (0.03)	2.14 (0.01)	1.85 (0.00)	6.92 (0.10)
	BL	3.32 (0.05)	–	2.33 (0.05)	1.17 (0.02)	6.81 (0.14)
	XMZ	3.35 (0.04)	10.61 (0.17)	**1.88 (0.01)**	1.25 (0.01)	7.13 (0.18)
	IB	**2.95 (0.05)**	–	2.22 (0.05)	**1.14 (0.03)**	8.86 (0.46)
	RLZ	**2.90 (0.02)**	9.36 (0.07)	**1.55 (0.01)**	**1.04 (0.01)**	**6.22 (0.00)**
	Proposed	3.26 (0.03)	**7.10 (0.11)**	1.92 (0.02)	1.22 (0.01)	**6.75 (0.15)**
$p = 50$	***S***	7.25 (0.05)	–	8.26 (0.07)	5.75 (0.03)	90.19 (0.01)
	BIC	4.51 (0.03)	15.77 (0.21)	3.85 (0.06)	1.98 (0.01)	43.30 (0.47)
	BPA	4.30 (0.03)	**12.98 (0.16)**	3.62 (0.08)	1.84 (0.02)	41.49 (0.63)
	BT	6.10 (0.07)	15.07 (0.28)	2.42 (0.02)	1.93 (0.02)	10.68 (0.46)
	BL	4.67 (0.05)	–	2.41 (0.05)	1.27 (0.01)	4.80 (0.06)
	XMZ	4.64 (0.05)	20.45 (0.29)	**1.98 (0.01)**	1.36 (0.01)	5.13 (0.07)
	IB	**4.08 (0.05)**	–	2.52 (0.04)	**1.21 (0.02)**	6.05 (0.27)
	RLZ	**3.79 (0.02)**	16.20 (0.07)	**1.63 (0.01)**	**1.06 (0.00)**	**3.84 (0.00)**
	Proposed	4.58 (0.03)	**13.68 (0.10)**	2.02 (0.01)	1.35 (0.01)	**4.36 (0.06)**
$p = 100$	***S***	14.40 (0.06)	–	16.15 (0.12)	11.43 (0.03)	95.01 (0.00)
	BIC	6.87 (0.03)	42.56 (0.45)	5.26 (0.07)	2.24 (0.01)	32.75 (0.36)
	BPA	6.74 (0.03)	35.68 (0.38)	5.29 (0.11)	2.20 (0.02)	33.60 (0.38)
	BT	8.54 (0.13)	**28.78 (0.33)**	2.40 (0.03)	1.87 (0.02)	4.08 (0.33)
	BL	7.19 (0.04)	–	2.67 (0.06)	1.42 (0.01)	2.83 (0.02)
	XMZ	14.39 (0.06)	364.15 (0.18)	16.14 (0.12)	11.42 (0.03)	94.96 (0.01)
	IB	**5.89 (0.05)**	–	2.80 (0.05)	**1.23 (0.01)**	2.71 (0.07)
	RLZ	**5.41 (0.02)**	33.40 (0.12)	**1.69 (0.01)**	**1.08 (0.00)**	**1.96 (0.00)**
	Proposed	7.06 (0.02)	**31.28 (0.14)**	**2.12 (0.01)**	1.49 (0.00)	**2.38 (0.02)**

Next, we provide some insights of the results through methods and other loss functions. Tables 1 and 2 summarize the comparison results for Models 1 and 2, respectively. From the perspective of competing methods, the sample covariance matrix ***S*** does not necessarily yield a sparse estimate, and exhibits poor performance with respect to all the loss measures. The BIC and BPA in the second class of approaches provide sparse covariance matrix estimates compared with ***S***, but their FSL are considerably larger than that of the proposed method. Moreover, our proposed method is clearly superior to both BIC and BPA with respect to L_1 and MAE for all settings that we consider. Although the proposed method is comparable to the BIC and BPA under EN criterion when $p = 30$, it performs slightly better when $p = 50$ and much better in the case of $p = 100$.

TABLE 2: The averages and standard errors of estimates for Model 2.

		F	EN	L_1	MAE	FSL (%)
$p = 30$	***S***	4.39 (0.04)	12.58 (0.08)	5.10 (0.07)	3.49 (0.03)	46.66 (0.01)
	BIC	3.35 (0.03)	5.35 (0.09)	2.94 (0.04)	1.91 (0.01)	43.47 (0.30)
	BPA	3.16 (0.03)	4.60 (0.07)	2.82 (0.04)	1.78 (0.01)	42.53 (0.31)
	BT	4.70 (0.01)	5.40 (0.03)	2.52 (0.01)	2.19 (0.00)	43.82 (0.08)
	BL	3.43 (0.04)	–	2.61 (0.04)	1.57 (0.01)	43.70 (0.20)
	XMZ	3.48 (0.04)	5.82 (0.11)	**2.27 (0.01)**	1.64 (0.01)	42.24 (0.20)
	IB	**3.02 (0.04)**	–	2.61 (0.04)	**1.56 (0.24)**	**41.88 (0.45)**
	RLZ	**2.76 (0.02)**	**3.16 (0.03)**	**1.89 (0.01)**	**1.34 (0.01)**	43.56 (0.00)
	Proposed	3.47 (0.03)	**4.09 (0.06)**	2.30 (0.01)	1.66 (0.01)	**41.82 (0.16)**
$p = 50$	***S***	7.36 (0.05)	–	8.54 (0.08)	5.84 (0.03)	65.59 (0.01)
	BIC	4.57 (0.02)	11.08 (0.20)	3.93 (0.08)	2.18 (0.01)	42.28 (0.24)
	BPA	4.38 (0.03)	9.17 (0.15)	3.70 (0.06)	2.08 (0.02)	41.73 (0.28)
	BT	6.06 (0.05)	10.07 (0.18)	2.69 (0.02)	2.26 (0.01)	30.34 (0.13)
	BL	4.73 (0.04)	–	2.91 (0.05)	1.68 (0.01)	29.45 (0.07)
	XMZ	4.73 (0.04)	11.26 (0.20)	**2.36 (0.01)**	1.76 (0.01)	28.85 (0.07)
	IB	**4.20 (0.05)**	–	2.92 (0.04)	**1.64 (0.02)**	**27.99 (0.25)**
	RLZ	**3.59 (0.01)**	**5.48 (0.04)**	**1.96 (0.01)**	**1.38 (0.00)**	28.68 (0.00)
	Proposed	4.70 (0.02)	**7.22 (0.06)**	2.40 (0.01)	1.77 (0.01)	**28.60 (0.07)**
$p = 100$	***S***	14.40 (0.07)	–	16.04 (0.12)	11.43 (0.04)	81.86 (0.00)
	BIC	6.92 (0.03)	29.70 (0.55)	5.32 (0.08)	2.47 (0.01)	33.77 (0.25)
	BPA	6.78 (0.03)	23.65 (0.36)	5.16 (0.11)	2.44 (0.02)	34.41 (0.31)
	BT	8.34 (0.13)	21.09 (0.36)	2.80 (0.03)	2.24 (0.02)	17.01 (0.23)
	BL	7.18 (0.04)	–	3.04 (0.05)	1.82 (0.01)	16.07 (0.02)
	XMZ	14.39 (0.07)	369.27 (0.19)	16.03 (0.12)	11.42 (0.04)	81.83 (0.00)
	IB	**6.53 (0.05)**	–	3.25 (0.05)	**1.64 (0.01)**	**15.22 (0.06)**
	RLZ	**5.13 (0.02)**	**11.03 (0.07)**	**2.05 (0.02)**	**1.41 (0.00)**	**15.12 (0.00)**
	Proposed	7.11 (0.02)	**16.29 (0.09)**	**2.49 (0.01)**	1.90 (0.00)	15.63 (0.02)

In comparison to the BT method, our proposed approach is clearly superior in capturing the sparse structure for both $p = 50$ and $p = 100$. Furthermore, the proposed method gives superior performance to the BT with respect to all the other loss criteria. In comparison with the BL method, the performance of our proposed method appears to be quite similar. It is well known that the BL method is asymptotically optimal for sparse covariance matrix (Bickel & Levina, 2009). However, the resulting estimate is not necessarily positive definite, which gives rise to problems when computing the EN loss function. Compared with the XMZ approach, our proposed method is superior or comparable with respect to all the loss measures both for $p = 30$ and 50. In the high-dimensional case when $p = 100$, the proposed method performs much better than the XMZ

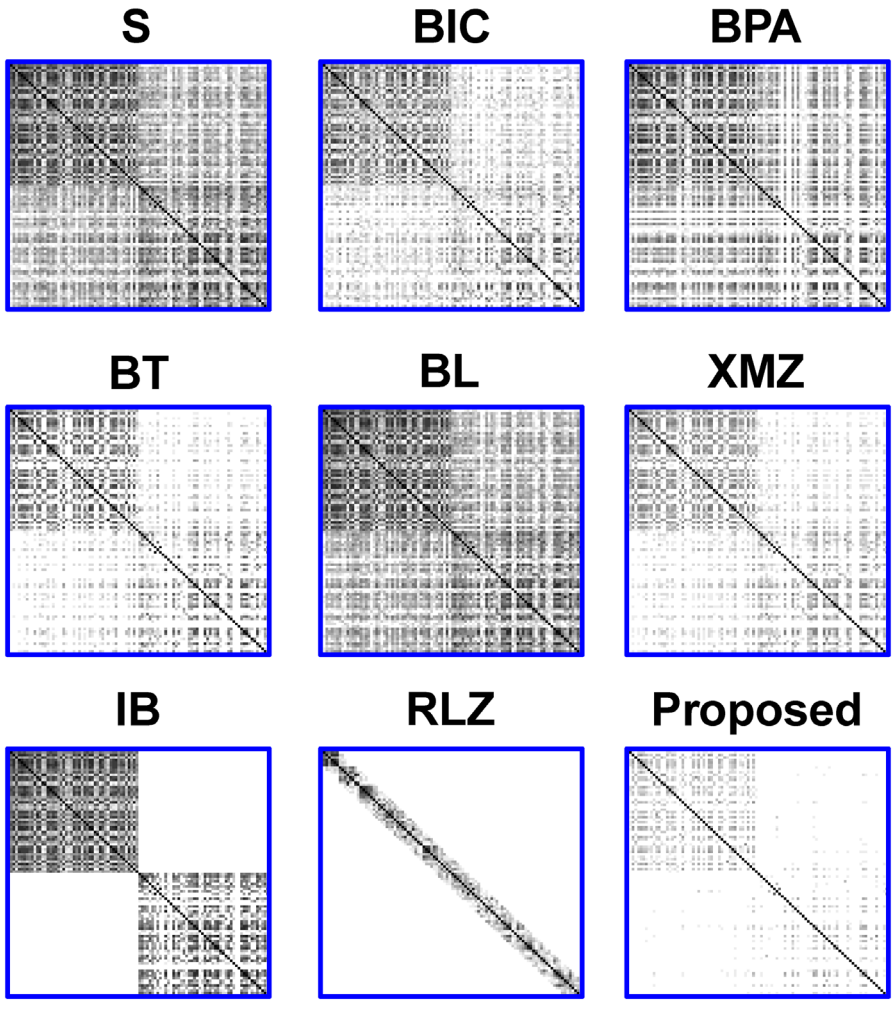

FIGURE 1: Heatmaps of the absolute values of the estimated correlation matrices obtained from the proposed method and other approaches for prostate cancer data. A darker shade indicates a higher density, and a lighter shade a lower density.

approach. The IB method performs well regarding MAE and comparably to the proposed method under FSL. But it has the singularity issue. Finally, as we have already remarked, when the true covariance matrix is either banded or tapered, i.e., Models 1 and 2, the RLZ method outperforms our proposed model.

Tables A1 and A2 in the Appendix present the comparison results for Models 3 and 4, respectively. Different from Models 1 and 2, the covariance matrices under Models 3 and 4 are unstructured. This implies that the RLZ does not have the advantage. Hence, it is clearly seen that the proposed method performs much better than the RLZ approach, especially at capturing the sparse structure and with respect to EN loss. Overall, the proposed method provides superior performance to other approaches, with similar comparison results as described under Models 1 and 2. For the most general covariance matrix with no sparse pattern of Model 5, our proposed method shows even better performance; see Table A3 in the Appendix. Finally, note that for the cases when $p = 100$, our method usually exhibits very good performance.

TABLE 3: Estimated percent misclassification errors for LDA obtained in the prostate cancer data.

Methods	BIC	BPA	BT	XMZ	RLZ	Proposed
ME	31.7	15.6	14.9	16.2	14.4	14.7
SE	1.58	1.06	1.01	0.97	1.04	0.95

6. APPLICATION

In this section, a real prostate cancer data set (Glaab et al., 2012) is used to evaluate the performance of the proposed method in comparison with other approaches described in Section 5. The data contain two classes with 50 normal samples and 52 prostate cancer samples, as well as 2,135 gene expression values recorded for each sample. Data are available online at http://ico2s.org/datasets/microarray.html. Because it includes a large number of variables, we adopt a variable screening procedure using a two sample t-statistic to identify the 50 variables that generate the largest observed values (Group 1), together with the corresponding 50 variables that give rise to the smallest observed values (Group 2). The underlying intention is to identify two groups of variables that would likely be somewhat mutually correlated within a group, but otherwise rather weakly dependent (Rothman, Levina & Zhu, 2009; Xue, Ma & Zou, 2012). Data are centred within each class and then used for the analysis. In this section, to make each variable at the same scale, we focus on the correlation matrix rather than the covariance matrix.

Figure 1 shows the heatmaps of the absolute values of the estimated correlation matrices obtained from each method. It can be seen that for this data set, the IB method appears to have a leading performance for identifying the expected sparse pattern with clear blocks, followed by the proposed method, the BT and XMZ approaches, which are comparable to capture the sparse structure with two diagonal blocks. All the rest approaches either result in a much sparser matrix as diagonal matrix (i.e., RLZ) or fail to identify the sparsity pattern (i.e., $\boldsymbol{S}$, BIC, BPA and BL). We also observe that the IB and BL estimators yield negative eigenvalues, while the other estimators guarantee the positive definiteness.

Next, we further examine the performance of the proposed method by means of classification of the linear discriminant analysis (LDA). The whole data set is randomly split into the training set with 50 observations and the testing set with the rest 52 observations. For this analysis, we screen all 2,135 gene expressions by the two sample t-test based on the training data to select the top 100 significant variables. Then all the compared methods use the training data to estimate the covariance matrix of these 100 variables. Finally, each estimate is plugged into the LDA rule to classify the testing data. Table 3 displays the averaged misclassification errors in percentage and corresponding standard errors by each method for the above split procedure of 50 times. We see that although the proposed method is slightly inferior to the RLZ, it performs better than others in classification for this set of data. Since we order the 100 variables by their significance from two sample t-test, the variables far apart in distance from each other may have weak correlations. Hence, the RLZ performs well as the covariance matrix of such 100 variables may be banded.

7. DISCUSSION

In this article, we consider a positive definite estimate of covariance matrix based on the MCD. The proposed method resolves the order dependency issue in the MCD by exploiting the multiple estimates obtained from different variable orders. The positive definite constraint and L_1 penalty are added to the objective function to guarantee the positive definiteness and encourage the sparse structure of the estimated covariance matrix. An efficient algorithm is developed to solve the

constraint optimization problem. The proposed estimator does not require any prior knowledge of the variable order used in the MCD, and performs well in the high-dimensional cases. Simulation studies and an application from a prostate cancer study demonstrate the superiority of our proposed method of estimation with respect to existing alternative methods.

The idea of addressing variable ordination in this research may also be applied in other estimation problems, such as the inverse covariance matrix estimate. However, one potential issue in practice is that the variables involved may have relations among themselves, e.g., a causal relationship, or perhaps spatial information. This could mean that some orders of variables are meaningful and reflect such relations, while others may not. This issue can be clearly identified in the performances of the BIC and BPA methods in our simulation study. Hence, ruling out the meaningless orders and only using those that are meaningful would improve the performance of the proposed method. How to implement this idea in practice needs further study.

APPENDIX

Proof of Lemma 1. Since $(\boldsymbol{\Sigma}^+, \boldsymbol{\Phi}^+, \boldsymbol{\Lambda}^+)$ is the optimal minimizer of Equation (7), based on the Karush—Kuhn–Tucker conditions we have

$$\left(-\hat{\boldsymbol{\Sigma}}^+ + \frac{1}{M}\sum_{k=1}^{M}\hat{\boldsymbol{\Sigma}}_k - \hat{\boldsymbol{\Lambda}}^+ \right)_{jl} \in \lambda\partial|\hat{\boldsymbol{\Sigma}}^+_{jl}|,\ j = 1, \ldots, p, l = 1, \ldots, p, \text{and } j \neq l \tag{A1}$$

$$\left(-\hat{\boldsymbol{\Sigma}}^+ + \frac{1}{M}\sum_{k=1}^{M}\hat{\boldsymbol{\Sigma}}_k \right)_{jj} + \hat{\boldsymbol{\Lambda}}^+_{jj} = 0,\ j = 1, \ldots, p \tag{A2}$$

$$\hat{\boldsymbol{\Phi}}^+ = \hat{\boldsymbol{\Sigma}}^+ \tag{A3}$$

$$\hat{\boldsymbol{\Phi}}^+ \succeq \nu \boldsymbol{I}, \tag{A4}$$

and

$$\langle \hat{\boldsymbol{\Lambda}}^+, \boldsymbol{\Phi} - \hat{\boldsymbol{\Phi}}^+ \rangle \leq 0,\ \forall \boldsymbol{\Phi} \succeq \nu \boldsymbol{I}. \tag{A5}$$

The expressions in Equations (A1) and (A2) result from the stationarity, and the results in (A3) and (A4) are valid because of the primal feasibility. By the optimality conditions of the problem (8) with respect to $\boldsymbol{\Phi}$, we obtain

$$\left\langle \boldsymbol{\Lambda}^i - \frac{1}{\tau}(\boldsymbol{\Phi}^{i+1} - \boldsymbol{\Sigma}^i), \boldsymbol{\Phi} - \boldsymbol{\Phi}^{i+1} \right\rangle \leq 0,\ \forall \boldsymbol{\Phi} \succeq \nu \boldsymbol{I}.$$

This, together with $\boldsymbol{\Lambda}$ step (10), yields

$$\left\langle \boldsymbol{\Lambda}^{i+1} - \frac{1}{\tau}(\boldsymbol{\Sigma}^{i+1} - \boldsymbol{\Sigma}^i), \boldsymbol{\Phi} - \boldsymbol{\Phi}^{i+1} \right\rangle \leq 0,\ \forall \boldsymbol{\Phi} \succeq \nu \boldsymbol{I}. \tag{A6}$$

Now by setting $\boldsymbol{\Phi} = \boldsymbol{\Phi}^{i+1}$ in Expression (A5) and $\boldsymbol{\Phi} = \hat{\boldsymbol{\Phi}}^+$ in Expression (A6), it follows that

$$\langle \hat{\boldsymbol{\Lambda}}^+, \boldsymbol{\Phi}^{i+1} - \hat{\boldsymbol{\Phi}}^+ \rangle \leq 0, \tag{A7}$$

and

$$\langle \boldsymbol{\Lambda}^{i+1} - \frac{1}{\tau}(\boldsymbol{\Sigma}^{i+1} - \boldsymbol{\Sigma}^i), \hat{\boldsymbol{\Phi}}^+ - \boldsymbol{\Phi}^{i+1} \rangle \leq 0. \tag{A8}$$

Summing Expressions (A7) and (A8) gives

$$\left\langle (\boldsymbol{\Lambda}^{i+1} - \hat{\boldsymbol{\Lambda}}^{+}) - \frac{1}{\tau}(\boldsymbol{\Sigma}^{i+1} - \boldsymbol{\Sigma}^{i}), \boldsymbol{\Phi}^{i+1} - \hat{\boldsymbol{\Phi}}^{+} \right\rangle \geq 0. \tag{A9}$$

On the other hand, by the optimality conditions of the problem (9) with respect to $\boldsymbol{\Sigma}$, we have

$$0 \in \left[\frac{1}{M} \sum_{k=1}^{M} (\boldsymbol{\Sigma}^{i+1} - \hat{\boldsymbol{\Sigma}}_k) + \boldsymbol{\Lambda}^{i} + \frac{1}{\tau}(\boldsymbol{\Sigma}^{i+1} - \boldsymbol{\Phi}^{i+1}) \right]_{jl} + \lambda \partial |\boldsymbol{\Sigma}_{jl}^{i+1}|, \ j \neq l, \tag{A10}$$

and

$$\left[\frac{1}{M} \sum_{k=1}^{M} (\boldsymbol{\Sigma}^{i+1} - \hat{\boldsymbol{\Sigma}}_k) + \boldsymbol{\Lambda}^{i} + \frac{1}{\tau}(\boldsymbol{\Sigma}^{i+1} - \boldsymbol{\Phi}^{i+1}) \right]_{jj} = 0, \ j = 1, \ldots, p. \tag{A11}$$

Plugging $\boldsymbol{\Lambda}$ step (10) into Expressions (A10) and (A11) respectively results in

$$\left(-\boldsymbol{\Sigma}^{i+1} + \frac{1}{M} \sum_{k=1}^{M} \hat{\boldsymbol{\Sigma}}_k - \boldsymbol{\Lambda}^{i+1} \right)_{jl} \in \lambda \partial |\boldsymbol{\Sigma}_{jl}^{i+1}|, \quad \text{for } j = 1, \ldots, p, l = 1, \ldots, p, \text{and } j \neq l, \tag{A12}$$

and

$$\left(\boldsymbol{\Sigma}^{i+1} - \frac{1}{M} \sum_{k=1}^{M} \hat{\boldsymbol{\Sigma}}_k \right)_{jj} + \boldsymbol{\Lambda}_{jj}^{i+1} = 0, \ j = 1, \ldots, p. \tag{A13}$$

Since $\partial |\cdot|$ is monotonically non-decreasing, for $j \neq l$, Expressions (A1) and (A12) yield

$$\left(-\boldsymbol{\Sigma}^{i+1} + \frac{1}{M} \sum_{k=1}^{M} \hat{\boldsymbol{\Sigma}}_k - \boldsymbol{\Lambda}^{i+1} \right)_{jl} \geq \left(-\hat{\boldsymbol{\Sigma}}^{+} + \frac{1}{M} \sum_{k=1}^{M} \hat{\boldsymbol{\Sigma}}_k - \hat{\boldsymbol{\Lambda}}^{+} \right)_{jl}$$

when $\boldsymbol{\Sigma}_{jl}^{i+1} \geq \hat{\boldsymbol{\Sigma}}_{jl}^{+}$, and

$$\left(-\boldsymbol{\Sigma}^{i+1} + \frac{1}{M} \sum_{k=1}^{M} \hat{\boldsymbol{\Sigma}}_k - \boldsymbol{\Lambda}^{i+1} \right)_{jl} \leq \left(-\hat{\boldsymbol{\Sigma}}^{+} + \frac{1}{M} \sum_{k=1}^{M} \hat{\boldsymbol{\Sigma}}_k - \hat{\boldsymbol{\Lambda}}^{+} \right)_{jl}$$

when $\boldsymbol{\Sigma}_{jl}^{i+1} < \hat{\boldsymbol{\Sigma}}_{jl}^{+}$. That is,

$$(\hat{\boldsymbol{\Sigma}}^{+} - \boldsymbol{\Sigma}^{i+1} + \hat{\boldsymbol{\Lambda}}^{+} - \boldsymbol{\Lambda}^{i+1})_{jl} \begin{cases} \geq 0, \text{ if } \boldsymbol{\Sigma}_{jl}^{i+1} \geq \hat{\boldsymbol{\Sigma}}_{jl}^{+} \\ \leq 0, \text{ if } \boldsymbol{\Sigma}_{jl}^{i+1} < \hat{\boldsymbol{\Sigma}}_{jl}^{+} \end{cases}.$$

As a result, we obtain

$$(\boldsymbol{\Sigma}^{i+1} - \hat{\boldsymbol{\Sigma}}^{+})_{jl} (\hat{\boldsymbol{\Sigma}}^{+} - \boldsymbol{\Sigma}^{i+1} + \hat{\boldsymbol{\Lambda}}^{+} - \boldsymbol{\Lambda}^{i+1})_{jl} \geq 0 \quad \text{for } j = 1, \ldots, p, l = 1, \ldots, p, \text{ and } j \neq l. \tag{A14}$$

In addition, subtracting Expression (A13) from Expression (A2) implies

$$(\hat{\boldsymbol{\Sigma}}^{+} - \boldsymbol{\Sigma}^{i+1} + \hat{\boldsymbol{\Lambda}}^{+} - \boldsymbol{\Lambda}^{i+1})_{jj} = 0,\ j = 1, \ldots, p. \tag{A15}$$

Then combining Expressions (A14) and (A15) leads to

$$\langle \boldsymbol{\Sigma}^{i+1} - \hat{\boldsymbol{\Sigma}}^{+}, \hat{\boldsymbol{\Sigma}}^{+} - \boldsymbol{\Sigma}^{i+1} + \hat{\boldsymbol{\Lambda}}^{+} - \boldsymbol{\Lambda}^{i+1} \rangle \geq 0. \tag{A16}$$

By summing Expressions (A9) and (A16), we have

$$\begin{aligned} &\langle \boldsymbol{\Sigma}^{i+1} - \hat{\boldsymbol{\Sigma}}^{+}, \hat{\boldsymbol{\Lambda}}^{+} - \boldsymbol{\Lambda}^{i+1} \rangle + \langle \boldsymbol{\Lambda}^{i+1} - \hat{\boldsymbol{\Lambda}}^{+}, \boldsymbol{\Phi}^{i+1} - \hat{\boldsymbol{\Phi}}^{+} \rangle \\ &\quad - \frac{1}{\tau} \langle \boldsymbol{\Sigma}^{i+1} - \hat{\boldsymbol{\Sigma}}^{i}, \boldsymbol{\Phi}^{i+1} - \hat{\boldsymbol{\Phi}}^{+} \rangle \geq \| \boldsymbol{\Sigma}^{i+1} - \hat{\boldsymbol{\Sigma}}^{+} \|_F^2. \end{aligned}$$

This result, together with Expression (A3) and $\boldsymbol{\Phi}^{i+1} = \tau(\boldsymbol{\Lambda}^{i} - \boldsymbol{\Lambda}^{i+1}) + \boldsymbol{\Sigma}^{i+1}$ from the $\boldsymbol{\Lambda}$ step (10), gives

$$\begin{aligned} &\tau \langle \boldsymbol{\Lambda}^{i+1} - \hat{\boldsymbol{\Lambda}}^{+}, \boldsymbol{\Lambda}^{i} - \boldsymbol{\Lambda}^{i+1} \rangle + \frac{1}{\tau} \langle \boldsymbol{\Sigma}^{i+1} - \hat{\boldsymbol{\Sigma}}^{+}, \boldsymbol{\Sigma}^{i} - \boldsymbol{\Sigma}^{i+1} \rangle \\ &\qquad \geq \| \boldsymbol{\Sigma}^{i+1} - \hat{\boldsymbol{\Sigma}}^{+} \|_F^2 - \langle \boldsymbol{\Lambda}^{i} - \boldsymbol{\Lambda}^{i+1}, \boldsymbol{\Sigma}^{i} - \boldsymbol{\Sigma}^{i+1} \rangle. \end{aligned} \tag{A17}$$

By $\hat{\boldsymbol{\Phi}}^{+} - \boldsymbol{\Phi}^{i+1} = (\hat{\boldsymbol{\Phi}}^{+} - \boldsymbol{\Phi}^{i}) + (\boldsymbol{\Phi}^{i} - \boldsymbol{\Phi}^{i+1})$ and $\hat{\boldsymbol{\Sigma}}^{+} - \boldsymbol{\Sigma}^{i+1} = (\hat{\boldsymbol{\Sigma}}^{+} - \boldsymbol{\Sigma}^{i}) + (\boldsymbol{\Sigma}^{i} - \boldsymbol{\Sigma}^{i+1})$, Expression (A17) is reduced to

$$\begin{aligned} &\tau \langle \boldsymbol{\Lambda}^{i} - \hat{\boldsymbol{\Lambda}}^{+}, \boldsymbol{\Lambda}^{i} - \boldsymbol{\Lambda}^{i+1} \rangle + \frac{1}{\tau} \langle \boldsymbol{\Sigma}^{i} - \hat{\boldsymbol{\Sigma}}^{+}, \boldsymbol{\Sigma}^{i} - \boldsymbol{\Sigma}^{i+1} \rangle \geq \tau \| \boldsymbol{\Lambda}^{i} - \boldsymbol{\Lambda}^{i+1} \|_F^2 \\ &\quad + \frac{1}{\tau} \| \boldsymbol{\Sigma}^{i} - \boldsymbol{\Sigma}^{i+1} \|_F^2 + \| \boldsymbol{\Sigma}^{i+1} - \hat{\boldsymbol{\Sigma}}^{+} \|_F^2 - \langle \boldsymbol{\Lambda}^{i} - \boldsymbol{\Lambda}^{i+1}, \boldsymbol{\Sigma}^{i} - \boldsymbol{\Sigma}^{i+1} \rangle. \end{aligned} \tag{A18}$$

Using the notation $\boldsymbol{W}^{+}$ and $\boldsymbol{W}^{i}$, the left-hand side of Expression (A18) becomes

$$\begin{aligned} &\langle (\boldsymbol{\Lambda}^{i} - \hat{\boldsymbol{\Lambda}}^{+}, \boldsymbol{\Sigma}^{i} - \hat{\boldsymbol{\Sigma}}^{+})', [\tau(\boldsymbol{\Lambda}^{i} - \boldsymbol{\Lambda}^{i+1}), \frac{1}{\tau}(\boldsymbol{\Sigma}^{i} - \boldsymbol{\Sigma}^{i+1})]' \rangle \\ &\quad = \langle (\boldsymbol{\Lambda}^{i}, \boldsymbol{\Sigma}^{i})' - (\hat{\boldsymbol{\Lambda}}^{+}, \hat{\boldsymbol{\Sigma}}^{+})', \boldsymbol{J}[(\boldsymbol{\Lambda}^{i}, \boldsymbol{\Sigma}^{i})' - (\hat{\boldsymbol{\Lambda}}^{i+1}, \hat{\boldsymbol{\Sigma}}^{i+1})'] \rangle \\ &\quad = \langle \boldsymbol{W}^{i} - \boldsymbol{W}^{+}, \boldsymbol{J}(\boldsymbol{W}^{i} - \boldsymbol{W}^{i+1}) \rangle \\ &\quad = \langle \boldsymbol{W}^{i} - \boldsymbol{W}^{+}, \boldsymbol{W}^{i} - \boldsymbol{W}^{i+1} \rangle_{J}. \end{aligned}$$

The first two terms on the right-hand side of Expression (A18) becomes

$$\begin{aligned} &\tau \| \boldsymbol{\Lambda}^{i} - \boldsymbol{\Lambda}^{i+1} \|_F^2 + \frac{1}{\tau} \| \boldsymbol{\Sigma}^{i} - \boldsymbol{\Sigma}^{i+1} \|_F^2 \\ &\quad = \tau \langle \boldsymbol{\Lambda}^{i} - \boldsymbol{\Lambda}^{i+1}, \boldsymbol{\Lambda}^{i} - \boldsymbol{\Lambda}^{i+1} \rangle + \frac{1}{\tau} \langle \boldsymbol{\Sigma}^{i} - \boldsymbol{\Sigma}^{i+1}, \boldsymbol{\Sigma}^{i} - \boldsymbol{\Sigma}^{i+1} \rangle \\ &\quad = \langle (\boldsymbol{\Lambda}^{i} - \boldsymbol{\Lambda}^{i+1}, \boldsymbol{\Sigma}^{i} - \boldsymbol{\Sigma}^{i+1})', [\tau(\boldsymbol{\Lambda}^{i} - \boldsymbol{\Lambda}^{i+1}), \frac{1}{\tau}(\boldsymbol{\Sigma}^{i} - \boldsymbol{\Sigma}^{i+1})]' \rangle \\ &\quad = \langle (\boldsymbol{\Lambda}^{i}, \boldsymbol{\Sigma}^{i})' - (\boldsymbol{\Lambda}^{i+1}, \boldsymbol{\Sigma}^{i+1})', \boldsymbol{J}[(\boldsymbol{\Lambda}^{i}, \boldsymbol{\Sigma}^{i})' - (\boldsymbol{\Lambda}^{i+1}, \boldsymbol{\Sigma}^{i+1})'] \rangle \end{aligned}$$

$$= \langle \boldsymbol{W}^i - \boldsymbol{W}^{i+1}, \boldsymbol{J}(\boldsymbol{W}^i - \boldsymbol{W}^{i+1}) \rangle$$
$$= \|\boldsymbol{W}^i - \boldsymbol{W}^{i+1}\|_J^2.$$

As a result, Expression (A18) can be rewritten as

$$\langle \boldsymbol{W}^i - \boldsymbol{W}^+, \boldsymbol{W}^i - \boldsymbol{W}^{i+1} \rangle_J \geq \|\boldsymbol{W}^i - \boldsymbol{W}^{i+1}\|_J^2 + \|\boldsymbol{\Sigma}^{i+1} - \hat{\boldsymbol{\Sigma}}^+\|_F^2 - \langle \boldsymbol{\Lambda}^i - \boldsymbol{\Lambda}^{i+1}, \boldsymbol{\Sigma}^i - \boldsymbol{\Sigma}^{i+1} \rangle.$$

Note that

$$\|\boldsymbol{W}^+ - \boldsymbol{W}^{i+1}\|_J^2 = \|\boldsymbol{W}^+ - \boldsymbol{W}^i\|_J^2 - 2\langle \boldsymbol{W}^+ - \boldsymbol{W}^i, \boldsymbol{W}^{i+1} - \boldsymbol{W}^i \rangle_J + \|\boldsymbol{W}^i - \boldsymbol{W}^{i+1}\|_J^2.$$

Therefore,

$$\begin{aligned}
&\|\boldsymbol{W}^+ - \boldsymbol{W}^i\|_J^2 - \|\boldsymbol{W}^+ - \boldsymbol{W}^{i+1}\|_J^2 \\
&= 2\langle \boldsymbol{W}^+ - \boldsymbol{W}^i, \boldsymbol{W}^{i+1} - \boldsymbol{W}^i \rangle_J - \|\boldsymbol{W}^i - \boldsymbol{W}^{i+1}\|_J^2 \\
&\geq 2\|\boldsymbol{W}^i - \boldsymbol{W}^{i+1}\|_J^2 + 2\|\boldsymbol{\Sigma}^{i+1} - \hat{\boldsymbol{\Sigma}}^+\|_F^2 - 2\langle \boldsymbol{\Lambda}^i - \boldsymbol{\Lambda}^{i+1}, \boldsymbol{\Sigma}^i - \boldsymbol{\Sigma}^{i+1} \rangle \\
&\quad - \|\boldsymbol{W}^i - \boldsymbol{W}^{i+1}\|_J^2 \\
&= \|\boldsymbol{W}^i - \boldsymbol{W}^{i+1}\|_J^2 + 2\|\boldsymbol{\Sigma}^{i+1} - \hat{\boldsymbol{\Sigma}}^+\|_F^2 + 2\langle \boldsymbol{\Lambda}^{i+1} - \boldsymbol{\Lambda}^i, \boldsymbol{\Sigma}^i - \boldsymbol{\Sigma}^{i+1} \rangle. \qquad \text{(A19)}
\end{aligned}$$

Hence, next we only need to show $\langle \boldsymbol{\Lambda}^{i+1} - \boldsymbol{\Lambda}^i, \boldsymbol{\Sigma}^i - \boldsymbol{\Sigma}^{i+1} \rangle \geq 0$. Now replacing i instead of $i+1$ in Expressions (A12) and (A13) yields

$$\left(-\boldsymbol{\Sigma}^i + \frac{1}{M} \sum_{k=1}^{M} \hat{\boldsymbol{\Sigma}}_k - \boldsymbol{\Lambda}^i \right)_{jl} \in \lambda \partial |\boldsymbol{\Sigma}_{jl}^i|, \ j = 1, \ldots, p, l = 1, \ldots, p, \text{and } j \neq l, \qquad \text{(A20)}$$

and

$$\left(\boldsymbol{\Sigma}^i - \frac{1}{M} \sum_{k=1}^{M} \hat{\boldsymbol{\Sigma}}_k \right)_{jj} + \boldsymbol{\Lambda}_{jj}^i = 0, \ j = 1, \ldots, p. \qquad \text{(A21)}$$

So Expressions (A12), (A13), (A20) and (A21), together with the monotonically non-decreasing property of $\partial|\cdot|$, imply that

$$\langle \boldsymbol{\Sigma}^i - \boldsymbol{\Sigma}^{i+1}, \boldsymbol{\Lambda}^{i+1} - \boldsymbol{\Lambda}^i + \boldsymbol{\Sigma}^{i+1} - \boldsymbol{\Sigma}^i \rangle \geq 0, \qquad \text{(A22)}$$

from which it follows that

$$\langle \boldsymbol{\Sigma}^i - \boldsymbol{\Sigma}^{i+1}, \boldsymbol{\Lambda}^{i+1} - \boldsymbol{\Lambda}^i \rangle \geq \|\boldsymbol{\Sigma}^{i+1} - \boldsymbol{\Sigma}^i\|_F^2 \geq 0.$$

Hence the last two terms on the right-hand side of Expression (A19) are both non-negative, which proves Lemma 1. ■

Proof of Theorem 1. According to Lemma 1, we have

(a) $\|\boldsymbol{W}^i - \boldsymbol{W}^{i+1}\|_J^2 \to 0$, as $i \to +\infty$;

(b) $\|\boldsymbol{W}^+ - \boldsymbol{W}^i\|_J^2$ is non-increasing and thus bounded.

Result (a) indicates that $\boldsymbol{\Sigma}^i - \boldsymbol{\Sigma}^{i+1} \to 0$ and $\boldsymbol{\Lambda}^i - \boldsymbol{\Lambda}^{i+1} \to 0$. Based on Expression (10), it is easy to see that $\boldsymbol{\Phi}^i - \boldsymbol{\Sigma}^i \to 0$. On the other hand, the result in (b) indicates that $\boldsymbol{W}^i$ lies in a compact region. Accordingly, there exists a subsequence $\boldsymbol{W}^{i_j}$ of $\boldsymbol{W}^i$ such that $\boldsymbol{W}^{i_j} \to \boldsymbol{W}^* = (\boldsymbol{\Lambda}^*, \boldsymbol{\Sigma}^*)$. In addition, we also have $\boldsymbol{\Phi}^{i_j} \to \boldsymbol{\Phi}^* \triangleq \boldsymbol{\Sigma}^*$. Therefore, $\lim_{i\to\infty}(\boldsymbol{\Sigma}^i, \boldsymbol{\Phi}^i, \boldsymbol{\Lambda}^i) = (\boldsymbol{\Sigma}^*, \boldsymbol{\Phi}^*, \boldsymbol{\Lambda}^*)$.

Next we show that $(\boldsymbol{\Sigma}^*, \boldsymbol{\Phi}^*, \boldsymbol{\Lambda}^*)$ is an optimal solution of Equation (4). By letting $i \to +\infty$ in Expressions (A12), (A13) and (A6), we have

$$\left(-\boldsymbol{\Sigma}^* + \frac{1}{M}\sum_{k=1}^{M} \hat{\boldsymbol{\Sigma}}_k - \boldsymbol{\Lambda}^* \right)_{jl} \in \lambda \partial |\boldsymbol{\Sigma}^*_{jl}|,\ j = 1, \ldots, p, l = 1, \ldots, p, \text{and } j \neq l, \tag{A23}$$

$$\left(\boldsymbol{\Sigma}^* - \frac{1}{M}\sum_{k=1}^{M} \hat{\boldsymbol{\Sigma}}_k \right)_{jj} + \boldsymbol{\Lambda}^*_{jj} = 0,\ j = 1, \ldots, p, \tag{A24}$$

and

$$\langle \boldsymbol{\Lambda}^*, \boldsymbol{\Phi} - \boldsymbol{\Phi}^* \rangle \leq 0,\ \forall \boldsymbol{\Phi} \succeq \nu \boldsymbol{I}. \tag{A25}$$

Expressions (A23), (A24) and (A25), together with $\boldsymbol{\Phi}^* = \boldsymbol{\Sigma}^*$, imply that $(\boldsymbol{\Sigma}^*, \boldsymbol{\Phi}^*, \boldsymbol{\Lambda}^*)$ is an optimal solution of $\arg\min L(\boldsymbol{\Sigma}, \boldsymbol{\Phi}; \boldsymbol{\Lambda})$ in Equation (7). Hence, we prove that the sequence produced by Algorithm 1 from any initial point converges numerically to an optimal minimizer of Equation (7). ■

Proof of Lemma 2. The proof is very similar to that of Lemma A.2 in Jiang (2012), so we omit it here. ■

Proof of Lemma 3. We prove this lemma using the idea found in Jiang (2012) by constructing a function $G(\cdot, \cdot)$ via the likelihood function, then decomposing $G(\cdot, \cdot)$ into several parts and bounding each part separately.

To simplify the notation, we prove the results using the original order without the symbol π_k. Note that the estimates $\hat{\boldsymbol{L}}$ and $\hat{\boldsymbol{D}}$ are based on a sequence of regressions derived from $\boldsymbol{\epsilon} = \boldsymbol{L}^{-1}\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D})$. The loss functions for the sequence of regressions can be written as the negative log likelihood, $\sum_{i=1}^{n}[\log|\boldsymbol{D}| + \text{tr}(\boldsymbol{x}_i'\boldsymbol{L}'^{-1}\boldsymbol{D}^{-1}\boldsymbol{L}^{-1}\boldsymbol{x}_i)]$, up to some constant. Consequently, adding the penalty terms to the negative log likelihood leads to the objective function

$$\sum_{i=1}^{n}[\log|\boldsymbol{D}| + \text{tr}(\boldsymbol{x}_i'\boldsymbol{L}'^{-1}\boldsymbol{D}^{-1}\boldsymbol{L}^{-1}\boldsymbol{x}_i)] + \sum_{j=1}^{p}\eta_j \sum_{k<j}|l_{jk}|.$$

Denote

$$Q(\boldsymbol{D}, \boldsymbol{L}) = (\log|\boldsymbol{D}| + \text{tr}(\boldsymbol{L}'^{-1}\boldsymbol{D}^{-1}\boldsymbol{L}^{-1}\boldsymbol{S}) + \sum_{j=1}^{p}\eta_j \sum_{k<j}|l_{jk}|.$$

TABLE A1: The averages and standard errors of estimates for Model 3.

		F	EN	L_1	MAE	FSL (%)
$p = 30$	$\boldsymbol{S}$	4.39 (0.04)	12.46 (0.09)	5.08 (0.06)	3.48 (0.02)	83.96 (0.01)
	BIC	3.28 (0.03)	7.28 (0.10)	2.93 (0.04)	1.76 (0.01)	52.71 (0.51)
	BPA	3.34 (0.03)	**5.87 (0.09)**	2.60 (0.05)	1.52 (0.02)	46.59 (0.78)
	BT	4.73 (0.01)	7.77 (0.04)	2.14 (0.01)	1.85 (0.00)	6.94 (0.10)
	BL	3.28 (0.05)	–	2.24 (0.05)	**1.14 (0.01)**	**6.62 (0.15)**
	XMZ	3.33 (0.04)	10.53 (0.14)	**1.87 (0.01)**	1.23 (0.01)	7.00 (0.15)
	IB	**2.99 (0.04)**	–	2.24 (0.04)	1.26 (0.03)	9.37 (0.44)
	RLZ	4.37 (0.01)	16.75 (0.11)	2.28 (0.02)	1.68 (0.00)	15.55 (0.00)
	Proposed	**3.22 (0.04)**	**6.96 (0.10)**	**1.89 (0.02)**	**1.21 (0.01)**	**6.84 (0.13)**
$p = 50$	$\boldsymbol{S}$	7.25 (0.05)	–	8.28 (0.08)	5.75 (0.03)	90.19 (0.01)
	BIC	4.91 (0.03)	16.00 (0.21)	3.85 (0.08)	1.98 (0.02)	43.07 (0.50)
	BPA	4.89 (0.03)	**12.92 (0.16)**	3.55 (0.08)	1.83 (0.02)	40.90 (0.65)
	BT	6.18 (0.06)	15.58 (0.26)	2.45 (0.02)	1.96 (0.01)	11.21 (0.47)
	BL	4.65 (0.05)	–	2.46 (0.05)	**1.27 (0.01)**	**4.79 (0.06)**
	XMZ	4.63 (0.06)	20.40 (0.32)	**1.97 (0.01)**	1.36 (0.01)	5.18 (0.07)
	IB	**4.35 (0.04)**	–	2.53 (0.05)	**1.25 (0.02)**	6.48 (0.32)
	RLZ	6.03 (0.01)	31.67 (0.17)	2.43 (0.01)	1.90 (0.00)	11.36 (0.00)
	Proposed	**4.60 (0.03)**	**13.74 (0.11)**	**2.00 (0.01)**	1.36 (0.01)	**4.39 (0.07)**
$p = 100$	$\boldsymbol{S}$	14.41 (0.05)	–	16.18 (0.10)	11.43 (0.03)	95.01 (0.00)
	BIC	7.59 (0.03)	42.88 (0.52)	5.36 (0.10)	2.26 (0.01)	32.65 (0.37)
	BPA	7.14 (0.03)	35.37 (0.43)	5.12 (0.12)	2.19 (0.02)	33.30 (0.41)
	BT	8.62 (0.14)	**28.78 (0.37)**	**2.36 (0.02)**	1.89 (0.02)	3.91 (0.30)
	BL	7.18 (0.05)	–	2.58 (0.05)	**1.41 (0.01)**	2.83 (0.02)
	XMZ	14.40 (0.05)	364.17 (0.16)	16.17 (0.10)	11.42 (0.03)	94.96 (0.00)
	IB	**7.11 (0.04)**	–	2.80 (0.05)	**1.21 (0.01)**	**2.61 (0.07)**
	RLZ	8.57 (0.01)	64.68 (0.22)	2.50 (0.01)	1.92 (0.00)	5.72 (0.00)
	Proposed	**7.06 (0.03)**	**31.33 (0.15)**	**2.10 (0.01)**	1.49 (0.00)	**2.39 (0.02)**

Define $G(\Delta_L, \Delta_D) = Q(\boldsymbol{D}_0 + \Delta_D, \boldsymbol{L}_0 + \Delta_L) - Q(\boldsymbol{D}_0, \boldsymbol{L}_0)$. Let $\mathcal{A}_{U_1} = \{\Delta_L : \|\Delta_L\|_F^2 \leq U_1^2 s_1 \log(p)/n\}$ and $\mathcal{B}_{U_2} = \{\Delta_D : \|\Delta_D\|_F^2 \leq U_2^2 p \log(p)/n\}$, where U_1 and U_2 are constants. We will show that for each $\Delta_L \in \partial\mathcal{A}_{U_1}$ and $\Delta_D \in \partial\mathcal{B}_{U_2}$, probability $P(G(\Delta_L, \Delta_D) > 0)$ is tending to 1 as $n \to \infty$ for sufficiently large U_1 and U_2, where $\partial\mathcal{A}_{U_1}$ and $\partial\mathcal{B}_{U_2}$ are the boundaries of $\mathcal{A}_{U_1}$ and $\mathcal{B}_{U_2}$, respectively. Additionally, since $G(\Delta_L, \Delta_D) = 0$ when $\Delta_L = 0$ and $\Delta_D = 0$, the minimum point of $G(\Delta_L, \Delta_D)$ is achieved when $\Delta_L \in \mathcal{A}_{U_1}$ and $\Delta_D \in \mathcal{B}_{U_2}$; that is $\|\Delta_L\|_F^2 = O_p(s_1 \log(p)/n)$ and $\|\Delta_D\|_F^2 = O_p(p \log(p)/n)$.

TABLE A2: The averages and standard errors of estimates for Model 4.

		F	EN	L_1	MAE	FSL (%)
$p = 30$	$\boldsymbol{S}$	4.41 (0.05)	12.48 (0.08)	5.16 (0.07)	3.49 (0.03)	46.66 (0.01)
	BIC	3.35 (0.03)	5.22 (0.08)	2.89 (0.04)	1.90 (0.01)	43.03 (0.33)
	BPA	**3.13 (0.03)**	**4.48 (0.08)**	2.79 (0.04)	1.76 (0.01)	42.66 (0.35)
	BT	4.69 (0.01)	5.36 (0.04)	2.50 (0.01)	2.19 (0.00)	43.79 (0.08)
	BL	3.47 (0.05)	–	2.69 (0.06)	**1.58 (0.02)**	43.86 (0.20)
	XMZ	3.51 (0.03)	5.86 (0.09)	**2.24 (0.01)**	1.64 (0.01)	42.31 (0.19)
	IB	**3.08 (0.04)**	7.91 (0.21)	2.62 (0.05)	**1.57 (0.02)**	**41.28 (0.39)**
	RLZ	4.55 (0.01)	11.57 (0.10)	2.67 (0.02)	2.15 (0.00)	50.67 (0.00)
	Proposed	3.48 (0.03)	**4.06 (0.05)**	**2.27 (0.01)**	1.66 (0.01)	**42.06 (0.16)**
$p = 50$	$\boldsymbol{S}$	7.24 (0.05)	–	8.24 (0.08)	5.74 (0.03)	65.58 (0.01)
	BIC	4.59 (0.03)	11.16 (0.21)	3.95 (0.08)	2.17 (0.01)	42.10 (0.24)
	BPA	**4.40 (0.03)**	**9.22 (0.15)**	3.72 (0.08)	2.07 (0.01)	41.28 (0.30)
	BT	6.09 (0.04)	10.41 (0.21)	2.71 (0.01)	2.27 (0.01)	30.31 (0.14)
	BL	4.71 (0.04)	–	2.79 (0.06)	**1.68 (0.01)**	29.50 (0.07)
	XMZ	4.72 (0.04)	10.96 (0.19)	**2.37 (0.01)**	1.75 (0.01)	28.78 (0.08)
	IB	**4.21 (0.04)**	14.23 (0.31)	2.86 (0.04)	**1.62 (0.02)**	**27.98 (0.20)**
	RLZ	5.94 (0.01)	19.21 (0.11)	2.77 (0.01)	2.24 (0.00)	33.60 (0.00)
	Proposed	4.74 (0.03)	**7.29 (0.06)**	**2.41 (0.01)**	1.78 (0.01)	**28.56 (0.07)**
$p = 100$	$\boldsymbol{S}$	14.41 (0.08)	–	16.10 (0.13)	11.44 (0.04)	81.85 (0.00)
	BIC	6.92 (0.02)	30.15 (0.51)	5.39 (0.10)	2.46 (0.01)	33.63 (0.25)
	BPA	**6.80 (0.03)**	23.53 (0.35)	5.39 (0.13)	2.43 (0.01)	33.92 (0.29)
	BT	8.51 (0.11)	**20.62 (0.33)**	**2.73 (0.02)**	2.24 (0.02)	16.43 (0.17)
	BL	7.20 (0.04)	–	3.03 (0.06)	**1.82 (0.01)**	16.08 (0.02)
	XMZ	14.40 (0.08)	369.42 (0.22)	16.09 (0.13)	11.43 (0.04)	81.82 (0.01)
	IB	**6.76 (0.05)**	27.14 (0.39)	3.23 (0.04)	**1.63 (0.01)**	**15.12 (0.06)**
	RLZ	8.54 (0.01)	40.19 (0.21)	2.89 (0.01)	2.31 (0.00)	18.44 (0.00)
	Proposed	7.12 (0.03)	**16.39 (0.08)**	**2.49 (0.01)**	1.90 (0.00)	**15.64 (0.02)**

Assume $\|\Delta_L\|_F^2 = U_1^2 s_1 \log(p)/n$ and $\|\Delta_D\|_F^2 = U_2^2 p \log(p)/n$. From assumption (11) and by Lemma 2, without loss of generality, there exists a constant h such that $0 < 1/h < sv_p(\boldsymbol{L}_0) \leq sv_1(\boldsymbol{L}_0) < h < \infty$ and $0 < 1/h < sv_p(\boldsymbol{D}_0) \leq sv_1(\boldsymbol{D}_0) < h < \infty$. Write $\boldsymbol{D} = \boldsymbol{D}_0 + \Delta_D$ and $\boldsymbol{L} = \boldsymbol{L}_0 + \Delta_L$. We decompose $G(\Delta_L, \Delta_D)$ into three parts and then consider them separately.

$$
\begin{aligned}
G(\Delta_L, \Delta_D) &= Q(\boldsymbol{D}, \boldsymbol{L}) - Q(\boldsymbol{D}_0, \boldsymbol{L}_0) \\
&= \log|\boldsymbol{D}| - \log\left|\boldsymbol{D}_0\right| + \operatorname{tr}(\boldsymbol{L}'^{-1}\boldsymbol{D}^{-1}\boldsymbol{L}^{-1}\boldsymbol{S}) - \operatorname{tr}(\boldsymbol{L}_0'^{-1}\boldsymbol{D}_0^{-1}\boldsymbol{L}_0^{-1}\boldsymbol{S})
\end{aligned}
$$

TABLE A3: The averages and standard errors of estimates for Model 5.

		F	EN	L_1	MAE	FSL (%)
$p = 30$	***S***	4.38 (0.03)	12.52 (0.08)	4.90 (0.05)	3.47 (0.02)	83.72 (0.01)
	BIC	**2.73 (0.02)**	**6.80 (0.12)**	2.76 (0.07)	1.34 (0.02)	38.29 (0.51)
	BPA	2.76 (0.03)	**8.20 (0.13)**	3.06 (0.09)	1.46 (0.02)	55.20 (0.78)
	BT	3.49 (0.01)	10.15 (0.15)	2.06 (0.01)	1.21 (0.01)	**10.23 (0.09)**
	BL	2.88 (0.02)	–	1.92 (0.03)	0.96 (0.01)	11.65 (0.08)
	XMZ	2.78 (0.02)	19.72 (0.28)	**1.78 (0.02)**	**0.94 (0.01)**	11.45 (0.09)
	IB	2.90 (0.03)	–	2.01 (0.04)	1.08 (0.02)	14.97 (0.50)
	RLZ	3.21 (0.01)	23.74 (0.19)	2.18 (0.01)	1.22 (0.01)	18.44 (0.01)
	Proposed	**2.48 (0.02)**	11.80 (0.29)	**1.62 (0.02)**	**0.86 (0.01)**	**10.73 (0.11)**
$p = 50$	***S***	7.25 (0.03)	–	8.03 (0.06)	5.75 (0.02)	84.43 (0.01)
	BIC	3.92 (0.02)	**21.74 (0.41)**	4.09 (0.11)	1.71 (0.01)	30.76 (0.37)
	BPA	**3.75 (0.02)**	22.41 (0.28)	3.78 (0.09)	1.76 (0.02)	47.28 (0.72)
	BT	4.46 (0.01)	**20.83 (0.21)**	2.58 (0.01)	1.48 (0.01)	**13.25 (0.13)**
	BL	3.79 (0.01)	–	2.44 (0.02)	1.26 (0.01)	13.37 (0.02)
	XMZ	3.68 (0.01)	38.03 (0.23)	**2.38 (0.01)**	**1.24 (0.01)**	13.49 (0.01)
	IB	4.07 (0.01)	–	2.63 (0.02)	1.44 (0.01)	14.57 (0.05)
	RLZ	3.90 (0.01)	37.04 (0.25)	2.48 (0.02)	1.44 (0.01)	16.96 (0.01)
	Proposed	**3.63 (0.01)**	32.08 (0.26)	**2.31 (0.01)**	**1.23 (0.01)**	**13.21 (0.04)**
$p = 100$	***S***	14.26 (0.04)	–	15.56 (0.09)	11.32 (0.03)	84.32 (0.01)
	BIC	5.87 (0.02)	43.75 (0.47)	5.38 (0.11)	2.27 (0.01)	24.49 (0.22)
	BPA	5.77 (0.02)	37.55 (0.28)	5.67 (0.13)	2.39 (0.02)	38.77 (0.45)
	BT	6.76 (0.02)	**31.42 (0.30)**	3.37 (0.02)	2.12 (0.01)	15.99 (0.26)
	BL	**5.38 (0.02)**	–	**3.09 (0.02)**	**1.87 (0.01)**	**14.69 (0.01)**
	XMZ	14.26 (0.04)	368.48 (0.19)	15.55 (0.09)	11.31 (0.03)	84.28 (0.01)
	IB	6.07 (0.01)	–	3.35 (0.02)	2.05 (0.01)	15.62 (0.02)
	RLZ	5.51 (0.01)	38.82 (0.12)	3.23 (0.02)	1.94 (0.01)	16.22 (0.01)
	Proposed	**5.23 (0.01)**	**33.60 (0.09)**	**3.06 (0.01)**	**1.76 (0.01)**	**14.62 (0.01)**

$$
\begin{aligned}
&+ \sum_{j=1}^{p} \eta_j \sum_{k<j} |l_{jk}| - \sum_{j=1}^{p} \eta_j \sum_{k<j} |l_{0jk}| \\
&= \log |\boldsymbol{D}| - \log |\boldsymbol{D}_0| + \mathrm{tr}[(\boldsymbol{D}^{-1} - \boldsymbol{D}_0^{-1})\boldsymbol{D}_0] - \mathrm{tr}[(\boldsymbol{D}^{-1} - \boldsymbol{D}_0^{-1})\boldsymbol{D}_0] \\
&\quad + \mathrm{tr}(\boldsymbol{L}'^{-1}\boldsymbol{D}^{-1}\boldsymbol{L}^{-1}\boldsymbol{S}) - \mathrm{tr}(\boldsymbol{L}_0'^{-1}\boldsymbol{D}_0^{-1}\boldsymbol{L}_0^{-1}\boldsymbol{S}) \\
&\quad + \sum_{j=1}^{p} \eta_j \sum_{k<j} |l_{jk}| - \sum_{j=1}^{p} \eta_j \sum_{k<j} |l_{0jk}| \\
&= M_1 + M_2 + M_3,
\end{aligned}
$$

where

$$M_1 = \log|\boldsymbol{D}| - \log|\boldsymbol{D}_0| + \mathrm{tr}[(\boldsymbol{D}^{-1} - \boldsymbol{D}_0^{-1})\boldsymbol{D}_0],$$
$$M_2 = \mathrm{tr}(\boldsymbol{L}'^{-1}\boldsymbol{D}^{-1}\boldsymbol{L}^{-1}\boldsymbol{S}) - \mathrm{tr}(\boldsymbol{L}_0'^{-1}\boldsymbol{D}_0^{-1}\boldsymbol{L}_0^{-1}\boldsymbol{S}) - \mathrm{tr}[(\boldsymbol{D}^{-1} - \boldsymbol{D}_0^{-1})\boldsymbol{D}_0],$$
$$M_3 = \sum_{j=1}^{p}\eta_j\sum_{k<j}|l_{jk}| - \sum_{j=1}^{p}\eta_j\sum_{k<j}|l_{0jk}|.$$

Based on the proof of Theorem 3.1 in Jiang (2012), we can show that $M_1 \geq \frac{\|\Delta_D\|_F^2}{8h^4}$. For the second term,

$$\begin{aligned}
M_2 &= \mathrm{tr}(\boldsymbol{L}'^{-1}\boldsymbol{D}^{-1}\boldsymbol{L}^{-1}\boldsymbol{S}) - \mathrm{tr}(\boldsymbol{L}'^{-1}\boldsymbol{D}_0^{-1}\boldsymbol{L}^{-1}\boldsymbol{S}) + \mathrm{tr}(\boldsymbol{L}'^{-1}\boldsymbol{D}_0^{-1}\boldsymbol{L}^{-1}\boldsymbol{S}) \\
&\quad - \mathrm{tr}(\boldsymbol{L}_0'^{-1}\boldsymbol{D}_0^{-1}\boldsymbol{L}_0^{-1}\boldsymbol{S}) - \mathrm{tr}[(\boldsymbol{D}^{-1} - \boldsymbol{D}_0^{-1})\boldsymbol{D}_0] \\
&= \mathrm{tr}(\boldsymbol{D}^{-1} - \boldsymbol{D}_0^{-1})[\boldsymbol{L}^{-1}(\boldsymbol{S} - \boldsymbol{\Sigma}_0)\boldsymbol{L}'^{-1}] + \mathrm{tr}\boldsymbol{D}_0^{-1}(\boldsymbol{L}^{-1}\boldsymbol{S}\boldsymbol{L}'^{-1} - \boldsymbol{L}_0^{-1}\boldsymbol{S}\boldsymbol{L}_0'^{-1}) \\
&\quad + \mathrm{tr}(\boldsymbol{D}^{-1} - \boldsymbol{D}_0^{-1})(\boldsymbol{L}^{-1}\boldsymbol{\Sigma}_0\boldsymbol{L}'^{-1} - \boldsymbol{D}_0) \\
&= \mathrm{tr}(\boldsymbol{D}^{-1} - \boldsymbol{D}_0^{-1})[\boldsymbol{L}^{-1}(\boldsymbol{S} - \boldsymbol{\Sigma}_0)\boldsymbol{L}'^{-1}] + \mathrm{tr}[\boldsymbol{D}_0^{-1}(\boldsymbol{L}^{-1}(\boldsymbol{S} - \boldsymbol{\Sigma}_0)\boldsymbol{L}'^{-1} \\
&\quad - \boldsymbol{L}_0^{-1}(\boldsymbol{S} - \boldsymbol{\Sigma}_0)\boldsymbol{L}_0'^{-1})] + \mathrm{tr}[\boldsymbol{D}_0^{-1}(\boldsymbol{L}^{-1}\boldsymbol{\Sigma}_0\boldsymbol{L}'^{-1} - \boldsymbol{L}_0^{-1}\boldsymbol{\Sigma}_0\boldsymbol{L}_0'^{-1})] \\
&\quad + \mathrm{tr}(\boldsymbol{D}^{-1} - \boldsymbol{D}_0^{-1})(\boldsymbol{L}^{-1}\boldsymbol{\Sigma}_0\boldsymbol{L}'^{-1} - \boldsymbol{D}_0) \\
&= \mathrm{tr}(\boldsymbol{D}^{-1} - \boldsymbol{D}_0^{-1})[\boldsymbol{L}^{-1}(\boldsymbol{S} - \boldsymbol{\Sigma}_0)\boldsymbol{L}'^{-1}] + \mathrm{tr}[\boldsymbol{D}_0^{-1}(\boldsymbol{L}^{-1}(\boldsymbol{S} - \boldsymbol{\Sigma}_0)\boldsymbol{L}'^{-1} \\
&\quad - \boldsymbol{L}_0^{-1}(\boldsymbol{S} - \boldsymbol{\Sigma}_0)\boldsymbol{L}_0'^{-1})] + \mathrm{tr}[\boldsymbol{D}^{-1}(\boldsymbol{L}^{-1}\boldsymbol{\Sigma}_0\boldsymbol{L}'^{-1} - \boldsymbol{L}_0^{-1}\boldsymbol{\Sigma}_0\boldsymbol{L}_0'^{-1})] \\
&= M_2^{(1)} + M_2^{(2)} + M_2^{(3)},
\end{aligned}$$

where the fourth equality uses the results $\boldsymbol{L}_0^{-1}\boldsymbol{\Sigma}_0\boldsymbol{L}_0'^{-1} = \boldsymbol{L}_0^{-1}(\boldsymbol{L}_0\boldsymbol{D}_0\boldsymbol{L}_0')\boldsymbol{L}_0'^{-1} = \boldsymbol{D}_0$. The quantities $M_2^{(1)}$, $M_2^{(2)}$ and $M_2^{(3)}$ are defined in the following

$$M_2^{(1)} = \mathrm{tr}(\boldsymbol{D}^{-1} - \boldsymbol{D}_0^{-1})[\boldsymbol{L}^{-1}(\boldsymbol{S} - \boldsymbol{\Sigma}_0)\boldsymbol{L}'^{-1}],$$
$$M_2^{(2)} = \mathrm{tr}[\boldsymbol{D}_0^{-1}(\boldsymbol{L}^{-1}(\boldsymbol{S} - \boldsymbol{\Sigma}_0)\boldsymbol{L}'^{-1} - \boldsymbol{L}_0^{-1}(\boldsymbol{S} - \boldsymbol{\Sigma}_0)\boldsymbol{L}_0'^{-1})],$$
$$M_2^{(3)} = \mathrm{tr}[\boldsymbol{D}^{-1}(\boldsymbol{L}^{-1}\boldsymbol{\Sigma}_0\boldsymbol{L}'^{-1} - \boldsymbol{L}_0^{-1}\boldsymbol{\Sigma}_0\boldsymbol{L}_0'^{-1})].$$

Based on the proof of Theorem 3.1 in Jiang (2012), for any $\epsilon > 0$, there exists $V_1 > 0$ and $V_2 > 0$ such that

$$|M_2^{(1)}| \leq V_1\sqrt{p\log(p)/n}\|\Delta_D\|_F$$

and

$$\begin{aligned}
M_2^{(3)} - |M_2^{(2)}| > 1/2h^4\|\Delta_L\|_F^2 - V_2\sqrt{\log(p)/n}\sum_{(j,k)\in Z^c}|l_{jk}| \\
- V_2\sqrt{s_1\log(p)/n}\|\Delta_L\|_F,
\end{aligned}$$

where $Z = \{(j,k) : k < j, l_{0jk} \neq 0\}$, and l_{0jk} represents the element (j,k) of the matrix $\boldsymbol{L}_0$. Next, for the penalty term,

$$M_3 = \sum_{j=1}^{p} \eta_j \sum_{(j,k)\in Z^c} |l_{jk}| + \sum_{j=1}^{p} \eta_j \sum_{(j,k)\in Z} (|l_{jk}| - |l_{0jk}|) = M_3^{(1)} + M_3^{(2)},$$

where

$$M_3^{(1)} = \sum_{j=1}^{p} \eta_j \sum_{(j,k)\in Z^c} |l_{jk}|,$$

and

$$\begin{aligned} |M_3^{(2)}| = |\sum_{j=1}^{p} \eta_j \sum_{(j,k)\in Z} (|l_{jk}| - |l_{0jk}|)| &\leq \sum_{j=1}^{p} \eta_j \sum_{(j,k)\in Z} |(|l_{jk}| - |l_{0jk}|)| \\ &\leq \sum_{j=1}^{p} \eta_j \sum_{(j,k)\in Z} |l_{jk} - l_{0jk}| \\ &\leq \sum_{j=1}^{p} \eta_j \sqrt{s_1} \|\Delta_L\|_F, \end{aligned}$$

where the last inequality uses the fact that $(a_1 + a_2 + \cdots + a_m)^2 \leq m(a_1^2 + a_2^2 + \cdots + a_m^2)$. Combining all the terms above together, with probability greater than $1 - 2\epsilon$, we have

$$\begin{aligned} |G(\Delta_L, \Delta_D)| &\geq M_1 - |M_2^{(1)}| + M_2^{(3)} - |M_2^{(2)}| + M_3^{(1)} - |M_3^{(2)}| \\ &\geq \frac{\|\Delta_D\|_F^2}{8h^4} - V_1\sqrt{p\log(p)/n}\|\Delta_D\|_F + \frac{\|\Delta_L\|_F^2}{2h^4} - V_2\sqrt{\log(p)/n} \sum_{(j,k)\in Z^c} |l_{jk}| \\ &\quad - V_2\sqrt{s_1\log(p)/n}\|\Delta_L\|_F + \sum_{j=1}^{p} \eta_j \sum_{(j,k)\in Z^c} |l_{jk}| - \sum_{j=1}^{p} \eta_j \sqrt{s_1}\|\Delta_L\|_F \\ &= \frac{U_2^2}{8h^4} p\log(p)/n - V_1U_2 p\log(p)/n + \frac{U_1^2}{2h^4} s_1\log(p)/n \\ &\quad - V_2\sqrt{\log(p)/n} \sum_{(j,k)\in Z^c} |l_{jk}| - V_2U_1 s_1\log(p)/n + \sum_{j=1}^{p} \eta_j \sum_{(j,k)\in Z^c} |l_{jk}| \\ &\quad - s_1U_1\sqrt{\log(p)/n} \sum_{j=1}^{p} \eta_j \\ &= \frac{U_2 p\log(p)}{n}\left(\frac{U_2}{8h^4} - V_1\right) + \frac{U_1 s_1\log(p)}{n}\left(\frac{U_1}{2h^4} - \frac{\sum_{j=1}^{p}\eta_j}{\sqrt{\log(p)/n}} - V_2\right) \\ &\quad + \sum_{(j,k)\in Z^c} |l_{jk}|\left(\sum_{j=1}^{p} \eta_j - V_2\sqrt{\log(p)/n}\right). \end{aligned}$$

Here V_1 and V_2 are only related to the sample size n and ϵ. Assume $\sum_{j=1}^{p} \eta_j = K(\sqrt{\log(p)/n})$ where $K > V_2$ and choose $U_1 > 2h^4(K + V_2)$, $U_2 > 8h^4 V_1$, then $G(\Delta_L, \Delta_D) > 0$. This establishes the lemma. ■

Proof of Theorem 2. Based on the proof of Theorem 3.2 in Jiang (2012), it follows that

$$\begin{aligned}\|\hat{\boldsymbol{\Sigma}}_{\pi_k} - \boldsymbol{\Sigma}_{0\pi_k}\|_F^2 &= O_p(\|\hat{\boldsymbol{L}}_{\pi_k} - \boldsymbol{L}_{0\pi_k}\|_F^2) + O_p(\|\hat{\boldsymbol{D}}_{\pi_k} - \boldsymbol{D}_{0\pi_k}\|_F^2) \\ &= O_p(s_1 \log(p)/n) + O_p(p \log(p)/n) \\ &= O_p((s_1 + p)\log(p)/n),\end{aligned}$$

where the second equality is provided by the proof of Lemma 3. Then

$$\begin{aligned}\|\hat{\boldsymbol{\Sigma}}_k - \boldsymbol{\Sigma}_0\|_F^2 &= \|\boldsymbol{P}_{\pi_k}\hat{\boldsymbol{\Sigma}}_{\pi_k}\boldsymbol{P}'_{\pi_k} - \boldsymbol{P}_{\pi_k}\boldsymbol{\Sigma}_{0\pi_k}\boldsymbol{P}'_{\pi_k}\|_F^2 \\ &= \|\boldsymbol{P}_{\pi_k}(\hat{\boldsymbol{\Sigma}}_{\pi_k} - \boldsymbol{\Sigma}_{0\pi_k})\boldsymbol{P}'_{\pi_k}\|_F^2 \\ &= \|\hat{\boldsymbol{\Sigma}}_{\pi_k} - \boldsymbol{\Sigma}_{0\pi_k}\|_F^2 \\ &= O_p((s_1 + p)\log(p)/n),\end{aligned}$$

where the third equality uses the fact that the Frobenius norm of a matrix is invariant on the permutation matrix.

Since $\boldsymbol{\Sigma}_0$ is positive definite, there exists $\epsilon > 0$ such that $\epsilon < \lambda_{\min}(\boldsymbol{\Sigma}_0)$, where $\lambda_{\min}(\boldsymbol{\Sigma}_0)$ is the smallest eigenvalue of $\boldsymbol{\Sigma}_0$. By introducing $\boldsymbol{\Delta} = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}_0$, the expression of (4) can be rewritten in terms of $\boldsymbol{\Delta}$ as

$$\begin{aligned}\hat{\boldsymbol{\Delta}} &= \arg \min_{\boldsymbol{\Delta}=\boldsymbol{\Delta}', \boldsymbol{\Delta}+\boldsymbol{\Sigma}_0 \geq \epsilon \boldsymbol{I}} \frac{1}{2M} \sum_{k=1}^{M} \|\boldsymbol{\Delta} + \boldsymbol{\Sigma}_0 - \hat{\boldsymbol{\Sigma}}_k\|_F^2 + \lambda|\boldsymbol{\Delta} + \boldsymbol{\Sigma}_0|_1 \\ &\triangleq \mathcal{F}(\boldsymbol{\Delta}).\end{aligned}$$

Note that it is easy to see $\hat{\boldsymbol{\Delta}} = \hat{\boldsymbol{\Sigma}}^{+} - \boldsymbol{\Sigma}_0$. Now consider $\boldsymbol{\Delta} \in \{\boldsymbol{\Delta} : \boldsymbol{\Delta} = \boldsymbol{\Delta}', \boldsymbol{\Delta} + \boldsymbol{\Sigma}_0 \geq \epsilon \boldsymbol{I}, \|\boldsymbol{\Delta}\|_F = 5\lambda\sqrt{s_0 + p}\}$. Define the active set of $\boldsymbol{\Sigma}_0$ as $A_0 = \{(i,j) : \sigma_{ij}^0 \neq 0, i \neq j\}$, and $\boldsymbol{B}_{A_0} = (b_{ij} \cdot \boldsymbol{I}_{\{(i,j)\in A_0\}})_{1\leq i,j\leq p}$. Let A_0^c be the complement set of A_0. Denote element (i,j) of matrix $\boldsymbol{\Delta}$ by $\boldsymbol{\Delta}_{ij}$. Under the probability event $\{|\hat{\sigma}_{ij}^k - \sigma_{ij}^0| \leq \lambda\}$ where $\hat{\boldsymbol{\Sigma}}_k = (\hat{\sigma}_{ij}^k)_{p\times p}$, we have

$$\begin{aligned}\mathcal{F}(\boldsymbol{\Delta}) - \mathcal{F}(0) &= \frac{1}{2M}\sum_{k=1}^{M}\|\boldsymbol{\Delta} + \boldsymbol{\Sigma}_0 - \hat{\boldsymbol{\Sigma}}_k\|_F^2 - \frac{1}{2M}\sum_{k=1}^{M}\|\boldsymbol{\Sigma}_0 - \hat{\boldsymbol{\Sigma}}_k\|_F^2 \\ &\quad + \lambda|\boldsymbol{\Delta} + \boldsymbol{\Sigma}_0|_1 - \lambda|\boldsymbol{\Sigma}_0|_1 \\ &= \frac{1}{2}\|\boldsymbol{\Delta}\|_F^2 + \frac{1}{M}\sum_{k=1}^{M} < \boldsymbol{\Delta}, \boldsymbol{\Sigma}_0 - \hat{\boldsymbol{\Sigma}}_k > + \lambda|\boldsymbol{\Delta}_{A_0^c}|_1 \\ &\quad + \lambda(|\boldsymbol{\Delta}_{A_0} + (\boldsymbol{\Sigma}_0)_{A_0}|_1 - |(\boldsymbol{\Sigma}_0)_{A_0}|_1) \\ &\geq \frac{1}{2}\|\boldsymbol{\Delta}\|_F^2 - \lambda(|\boldsymbol{\Delta}|_1 + \sum_i \boldsymbol{\Delta}_{ii}) + \lambda|\boldsymbol{\Delta}_{A_0^c}|_1 - \lambda|\boldsymbol{\Delta}_{A_0}|_1\end{aligned}$$

$$\begin{aligned}
&\geq \frac{1}{2}\|\boldsymbol{\Delta}\|_F^2 - 2\lambda\left(|\boldsymbol{\Delta}_{A_0}|_1 + \sum_i \boldsymbol{\Delta}_{ii}\right) \\
&\geq \frac{1}{2}\|\boldsymbol{\Delta}\|_F^2 - 2\lambda\sqrt{s_0+p}\|\boldsymbol{\Delta}\|_F \\
&= \frac{5}{2}\lambda^2(s_0+p) \\
&> 0.
\end{aligned}$$

Note that $\hat{\boldsymbol{\Delta}}$ is also the optimal solution to the convex optimization problem

$$\hat{\boldsymbol{\Delta}} = \arg \min_{\boldsymbol{\Delta}=\boldsymbol{\Delta}', \boldsymbol{\Delta}+\boldsymbol{\Sigma}_0 \geq \epsilon \boldsymbol{I}} \mathcal{F}(\boldsymbol{\Delta}) - \mathcal{F}(0).$$

The rest of proof is the same as that of Theorem 2 in Xue, Ma & Zou (2012), and hence is omitted. ■

ACKNOWLEDGEMENTS

The authors thank the editor and referees for their insightful and helpful comments that have greatly improved the original manuscript. The authors would like to acknowledge the support from the Ministry of Education of the People's Republic of China (20YJC910007), the National Natural Science Foundation of China (71903090) and the Science Education Foundation of Liaoning Province (LN2019Q21).

BIBLIOGRAPHY

Aubry, A., De Maio, A., Pallotta, L., & Farina, A. (2012). Maximum likelihood estimation of a structured covariance matrix with a condition number constraint. *IEEE Transactions on Signal Processing*, 60, 3004–3021.

Bickel, P. J. & Levina, E. (2009). Covariance regularization by thresholding. *The Annals of Statistics*, 36, 2577–2604.

Bien, J. & Tibshirani, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika*, 98, 807–820.

Boyd, S., Parikh, N., Chu, E., Peleato, B., & Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3, 1–122.

Cai, T. & Yuan, M. (2012). Adaptive covariance matrix estimation through block thresholding. *The Annals of Statistics*, 40, 2014–2042.

Cai, T., Ren, Z., & Zhou, H. H. (2016). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics*, 10, 1–59.

Chang, C. & Tsay, R. (2010). Estimation of covariance matrix via the sparse Cholesky factor with Lasso. *Journal of Statistical Planning and Inference*, 140, 3858–3873.

Dellaportas, P. & Pourahmadi, M. (2012). Cholesky-GARCH models with applications to finance. *Statistics and Computing*, 22, 849–855.

Deng, X. & Tsui, K. W. (2013). Penalized covariance matrix estimation using a matrix-logarithm transformation. *Journal of Computational and Graphical Statistics*, 22, 494–512.

Deng, X. & Yuan, M. (2009). Large Gaussian covariance matrix estimation with Markov structure. *Journal of Computational and Graphical Statistics*, 18, 640–657.

Dey, D. K. & Srinivasan, C. (1985). Estimation of a covariance matrix under Stein's loss. *The Annals of Statistics*, 13, 1581–1591.

Fan, J., Liao, Y., & Liu, H. (2016). An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal*, 19, 1–32.

Fan, J., Liao, Y., & Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society Series B*, 75, 603–680.

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

Glaab, E., Bacardit, J., Garibaldi, J. M., & Krasnogor, N. (2012). Using rule-based machine learning for candidate disease gene prioritization and sample classification of cancer gene expression data. *PLoS One*, 7, e39932.

Guo, J., Levina, E., Michailidis, G., & Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika*, 98, 1–15.

Haff, L. R. (1991). The variational form of certain Bayes estimators. *The Annals of Statistics*, 19, 1163–1190.

Huang, C., Farewell, D., & Pan, J. (2017). A calibration method for non-positive definite covariance matrix in multivariate data analysis. *Journal of Multivariate Analysis*, 157, 45–52.

Huang, J. Z., Liu, N., Pourahmadi, M., & Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93, 85–98.

Jiang, X. (2012). *Joint estimation of covariance matrix via Cholesky decomposition*, Ph.D. dissertation, Department of Statistics and Applied Probability, National University of Singapore.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29, 295–327.

Kang, X. & Deng, X. (2020). An improved modified Cholesky decomposition approach for precision matrix estimation. *Journal of Statistical Computation and Simulation*, 90, 443–464.

Kang, X., Xie, C., & Wang, M. (2020). A Cholesky-based estimation for large-dimensional covariance matrices. *Journal of Applied Statistics*, 47, 1017–1030.

Kang, X., Deng, X., Tsui, K. W., & Pourahmadi, M. (2019). On variable ordination of modified Cholesky decomposition for estimating time-varying covariance matrices. *International Statistical Review*, 10.1111/insr.12357, (to appear in print).

Karush, W. (1939). *Minima of functions of several variables with inequalities as side conditions*, Master's dissertation, Department of Mathematics, University of Chicago, Chicago, IL.

Kuhn, H. & Tucker, A. (1951). Nonlinear programming. In Neyman, J. (Ed.) *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, 481–492.

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

Ledoit, O. & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88, 365–411.

Liu, H., Wang, L., & Zhao, T. (2014). Sparse covariance matrix estimation with eigenvalue constraints. *Journal of Computational and Graphical Statistics*, 23, 439–459.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika*, 86, 677–690.

Pourahmadi, M. (2013). *High-Dimensional Covariance Estimation with High-Dimensional Data*. John Wiley & Sons, Chichester.

Pourahmadi, M., Daniels, M. J., & Park, T. (2007). Simultaneous modelling of the Cholesky decomposition of several covariance matrices. *Journal of Multivariate Analysis*, 98, 568–587.

Rajaratnam, B. & Salzman, J. (2013). Best permutation analysis. *Journal of Multivariate Analysis*, 121, 193–223.

Rocha, G. V., Zhao, P., & Yu, B. (2008). *A path following algorithm for sparse pseudo-likelihood inverse covariance estimation*, Technical report, Statistics Department, UC Berkeley, Berkeley, CA.

Rothman, A., Bickel, P., Levina, E., & Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2, 494–515.

Rothman, A. J., Levina, E., & Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104, 177–186.

Rothman, A. J., Levina, E., & Zhu, J. (2010). A new approach to Cholesky-based covariance regularization in high dimensions. *Biometrika*, 97, 539–550.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B*, 58, 267–288.

Wagaman, A. & Levina, E. (2009). Discovering sparse covariance structures with the Isomap. *Journal of Computational and Graphical Statistics*, 18, 551–572.

Won, J. H., Lim, J., Kim, S. J., & Rajaratnam, B. (2013). Condition number regularized covariance estimation. *Journal of the Royal Statistical Society Series B*, 75, 427–450.

Wu, W. B. & Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90, 831–844.

Xiao, L., Zipunnikov, V., Ruppert, D., & Crainiceanu, C. (2016). Fast covariance estimation for high-dimensional functional data. *Statistics and Computing*, 26, 409–421.

Xue, L., Ma, S., & Zou, H. (2012). Positive-definite L_1-penalized estimation of large covariance matrices. *Journal of the American Statistical Association*, 107, 1480–1491.

Yu, P. L. H., Wang, X., & Zhu, Y. (2017). High dimensional covariance matrix estimation by penalizing the matrix-logarithm transformed likelihood. *Computational Statistics and Data Analysis*, 114, 12–25.

Yuan, M. (2008). Efficient computation of the ℓ_1 regularized solution path in Gaussian graphical models. *Journal of Computational and Graphical Statistics*, 17, 809–826.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *The Journal of Machine Learning Research*, 11, 2261–2286.

Yuan, M. & Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94, 19–35.

Zheng, H., Tsui, K., Kang, X., & Deng, X. (2017). Cholesky-based model averaging for covariance matrix estimation. *Statistical Theory and Related Fields*, 1, 48–58.

Received 18 January 2019
Accepted 29 March 2020