Statistica Sinica Preprint No: SS-2023-0065	
Title	On Block Cholesky Decomposition for Sparse Inverse Covariance Estimation
Manuscript ID	SS-2023-0065
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0065
Complete List of Authors	Xiaoning Kang, Jiayi Lian and Xinwei Deng
Corresponding Authors	Xinwei Deng
E-mails	xdeng@vt.edu
Notice: Accepted version subject to English editing.	

Statistica Sinica

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

On Block Cholesky Decomposition for Sparse Inverse Covariance Estimation

Xiaoning Kang[†], Jiayi Lian[‡] and Xinwei Deng[‡]

[†]*Institute of Supply Chain Analytics and International Business College, Dongbei University of Finance and Economics, Dalian, China*

[‡]*Department of Statistics, Virginia Tech, Blacksburg, VA*

Abstract: The modified Cholesky decomposition is popular for inverse covariance estimation, but often needs pre-specification on the full information of variable ordering. In this work, we propose a block Cholesky decomposition (BCD) for estimating inverse covariance matrix under the partial information of variable ordering, in the sense that the variables can be divided into several groups with available ordering among groups, but variables within each group have no orderings. The proposed BCD model provides a unified framework for several existing methods including the modified Cholesky decomposition and the Graphical lasso. By utilizing the partial information on variable ordering, the proposed BCD model guarantees the positive definiteness of the estimated matrix with statistically meaningful interpretation. Theoretical results are established under regularity conditions. Simulation and case studies are conducted to evaluate the proposed BCD model.

Key words and phrases: Graphical model, modified Cholesky decomposition, regularization, sparsity, variable ordering.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

1. Introduction

The estimation of covariance and inverse covariance matrices is of fundamental importance in the multivariate statistics with a broad spectrum of applications, such as linear discriminant analysis (Clemmensen et al., 2011), portfolio optimization (Deng and Tsui, 2013), and assimilation (Nino-Ruiz et al., 2019). In high-dimensional data, sparse estimation of an inverse covariance matrix has specially attracted great attention, since it is closely related to a graphical model for inferring the conditional independence between variables of multivariate normal data. However, estimation of a large inverse covariance matrix often encounters two challenges. First, the estimated matrix needs to be positive definite for the valid statistical inferences. Second, the number of parameters in the model increases quickly in a quadratic order in terms of the matrix dimensionality.

Existing studies on the inverse covariance estimation in the literature generally fall into two categories. Denote the random variables of interest by $\boldsymbol{X} = (X_1, \ldots, X_p)'$ with mean $\mathbf{0}$ for simplicity and covariance matrix $\boldsymbol{\Sigma}$. The first category needs the pre-specification on the full ordering information of variables $\boldsymbol{X}$. That is, the variables $X_1, \ldots, X_p$ have a natural ordering, which typically occurs in longitudinal data, time series, spatial data, spectroscopy and so forth. In this situation, banded or tapering estimation for the high-dimensional matrices has been developed (Bickel and Levina, 2008), requiring an assumption that the variables are becoming weakly correlated as

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

their positions in the ordering are far away. Although these methods are straightforward and easy in computation, the resultant estimates may not be positive definite. A better technique for estimating the inverse covariance matrix in this category is the modified Cholesky decomposition (MCD) introduced by Pourahmadi (1999). It not only guarantees the positive definiteness of the estimated matrix, but also utilizes the information of variable ordering, leading to an accurate estimate. Moreover, this decomposition has a meaningful regression interpretation, allowing the use of the regularization for sparse estimation (Huang et al., 2006; Kang and Wang, 2021). That is, the sparse pattern in the Cholesky-based matrix estimates can be induced by the sparsity in the Cholesky factors through linear regressions in the decomposition.

The second category considers that the information of variable ordering is not available. In this case, one strategy is to identify a proper variable ordering based on a certain data driven mechanism, transforming it into the estimation problem of the first category. For example, Wagaman and Levina (2009) determined the variable ordering by the Isomap algorithm and proposed an Isoband matrix estimate. Dellaportas and Pourahmadi (2012) suggested using BIC criterion to seek for the variable ordering before applying the MCD technique. Rajaratnam and Salzman (2013) recovered the variable ordering via the best permutation algorithm, which is very efficient for autoregressive model. However, a potential drawback is obvious that the accuracy of such estimates relies on the accuracy of the estimated orderings of variables. Fur-

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

thermore, some real data, for example gene data and medical data, practically may not have a natural variable ordering, which means that it is not adequate to find a variable ordering for such data. Alternatively, another strategy in this category is to consider a permutation invariant estimation for inverse covariance matrix. One popular method is the Graphical lasso (Glasso) proposed by Meinshausen and Bühlmann (2006) and Yuan and Lin (2007), of which the algorithms and properties have been widely studied (Friedman et al., 2008; Lam and Fan, 2009; Yuan, 2010). Other works on the permutation invariant estimation for inverse covariance matrix can be found in Xue et al. (2012); Wang et al. (2015); Cai et al. (2016); Van Wieringen (2019); Wang et al. (2020), among many others.

By comparing two categories of estimation methods, it is seen that the MCD is an appropriate method for estimating the inverse covariance matrix when the information of variable ordering is fully available, while the Glasso is suitable when the information of variable ordering is not available. However, in many applications, there is only partial information of variable ordering available. Here the partial information means that the variables can be divided into several groups (including one group) with the group ordering known but the variable ordering within each group unknown. For example, in the multi-stage manufacturing process (Shi, 2006), each stage contains a set of variables and the stages have a natural ordering among themselves because of the nature of manufacturing process. Thus, the full information

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

of variable ordering is not available since the variables within the same stage may not have an ordering. In this case the partial information of variable ordering is present since there is an ordering among stages. To estimate the inverse covariance matrix of variables in such a multi-stage manufacturing process, neither the MCD method nor the Glasso method is adequate. Another concrete example is the Covid-19 data in Case Studies Section. The data were weekly collected from 37 continuous weeks at the beginning of the pandemic. In each week, four variables which have no ordering among themselves are recorded. Such four variables in a certain week are correlated with variables in the former weeks (see details in Case Studies). Therefore the $4 \times 37 = 148$ variables can be naturally divided into 37 groups by calendar week, which forms a partial information of the variable ordination.

In this work, we fill in the gap to develop a block Cholesky decomposition (BCD) method for estimating the inverse covariance matrix given the partial information of variable ordering. The proposed BCD method takes advantage of such partial information on the variable ordination to estimate the inverse covariance matrix via the joint estimation of a set of penalized multivariate regressions. It guarantees the positive definiteness of the estimated sparse inverse covariance matrix with statistically meaningful interpretation. Moreover, the proposed method provides a unified framework of estimating the inverse covariance matrix, where the MCD, the Glasso method, the Witten et al. (2011)'s estimator and Rothman et al.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

(2010a)'s estimator can be all considered as special cases of the proposed method. The theoretical results suggest that the proposed model can have faster consistent rate than that of Glasso method when the partial information on the variable ordination is present. The R codes of implementing the proposed model are available at https://github.com/xiaoningmike/BCD.

The remainder of this work is organized as follows. Section 2 develops the proposed method along with its parameter estimation. The asymptotically theoretical property is established in Section 3 under regularity conditions. Sections 4 and 5 demonstrate the merits of the proposed model via simulations and two real data examples. We conclude our work with some discussion in Section 6. All technical proofs are reported in the supplementary materials.

2. The Proposed Method

In this section, we describe the proposed methodology of the BCD method. Different from the MCD technique, the proposed BCD method only requires partial information of the variable ordination.

2.1 Block Cholesky Decomposition for Inverse Covariance Matrix

Suppose that the variables in $\boldsymbol{X}$ can be partitioned into M groups, with the jth group of variables denoted as $\boldsymbol{X}^{(j)} = (X_1^{(j)}, X_2^{(j)}, \ldots, X_{p_j}^{(j)})', j = 1, 2, \ldots, M$, where

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

p_j is the number of variables in the jth group, and $\sum_{j=1}^{M} p_j = p$. Assume that these M groups of variables have a natural ordering of $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \ldots, \boldsymbol{X}^{(M)}$, while there is not any ordering structure among variables $X_1^{(j)}, X_2^{(j)}, \ldots, X_{p_j}^{(j)}$ within each group $\boldsymbol{X}^{(j)}$. We call such ordering information of these M groups as the *partial information of the variable ordination*. Note that when $M = 1$, it reduces to the case that there is no ordering information for the p variables $X_1, \ldots, X_p$, and $M = p$ corresponds to the case that the p variables have a full ordering information. In this work, we assume both of p and M can be diverged as the sample size goes to infinity.

Without loss of generality, we write $\boldsymbol{X} = ((\boldsymbol{X}^{(1)})', (\boldsymbol{X}^{(2)})', \ldots, (\boldsymbol{X}^{(M)})')'$ with its inverse covariance matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. The key idea of the BCD model is to decompose $\boldsymbol{\Omega}$ by a block lower triangular matrix constructed from the multivariate regression coefficients when the variable group $\boldsymbol{X}^{(j)}$ is regressed on its preceding variable groups $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \ldots, \boldsymbol{X}^{(j-1)}$ for $j = 2, 3, \ldots, M$. Specifically, the BCD method considers a series of multivariate regressions

$$\boldsymbol{X}^{(j)} = \sum_{i=1}^{j-1} \boldsymbol{A}_{ji} \boldsymbol{X}^{(i)} + \boldsymbol{\epsilon}_j = \boldsymbol{A}_j \boldsymbol{Z}^{(j)} + \boldsymbol{\epsilon}_j, \ j = 2, \ldots, M, \tag{2.1}$$

where $\boldsymbol{Z}^{(j)} = ((\boldsymbol{X}^{(1)})', (\boldsymbol{X}^{(2)})', \ldots, (\boldsymbol{X}^{(j-1)})')'$ and $\boldsymbol{A}_j = (\boldsymbol{A}_{j1}, \ldots, \boldsymbol{A}_{j,j-1})$ with $\boldsymbol{A}_{ji}$ being the $p_j \times p_i$ coefficient matrix. Here $\boldsymbol{\epsilon}_j$ is the p_j-dimensional vector of error term for the jth multivariate regression with $E\boldsymbol{\epsilon}_j = \boldsymbol{0}$ and $Cov(\boldsymbol{\epsilon}_j) = \boldsymbol{D}_j$. Hence, we can

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

construct a block lower triangular matrix $\boldsymbol{A}$ written as

$$\boldsymbol{A} = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \boldsymbol{A}_{21} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \boldsymbol{A}_{31} & \boldsymbol{A}_{32} & \mathbf{0} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{A}_{M1} & \boldsymbol{A}_{M2} & \dots & \boldsymbol{A}_{M,M-1} & \mathbf{0} \end{pmatrix},$$

with its jth diagonal element being a $p_j \times p_j$ zero matrix, and its lower left part composed of all the regression coefficient matrices in Equation (2.1). Besides, define

$$\boldsymbol{D}_j = Cov(\boldsymbol{\epsilon}_j) = \begin{cases} Cov(\boldsymbol{X}^{(1)}), & j = 1, \\ Cov(\boldsymbol{X}^{(j)} - \sum_{i=1}^{j-1} \boldsymbol{A}_{ji}\boldsymbol{X}^{(i)}), & j = 2, 3, \dots, M. \end{cases}$$

Denote by $\boldsymbol{D} = \text{diag}(\boldsymbol{D}_1, \boldsymbol{D}_2, \dots, \boldsymbol{D}_M)$ the block diagonal covariance matrix of vector $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1', \boldsymbol{\epsilon}_2', \dots, \boldsymbol{\epsilon}_M')'$. Thus the multivariate regressions in (2.1) can be written as

$$\boldsymbol{\epsilon} = \boldsymbol{X} - \boldsymbol{A}\boldsymbol{X} = (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{X} \triangleq \boldsymbol{T}\boldsymbol{X}, \tag{2.2}$$

where $\boldsymbol{I}$ represents the $p \times p$ identity matrix, and $\boldsymbol{T} = \boldsymbol{I} - \boldsymbol{A}$ is a unit block lower triangular matrix having ones on its diagonal. The matrices $\boldsymbol{T}$ and $\boldsymbol{D}$ are called the block Cholesky factors. By taking $Cov(\boldsymbol{\epsilon}) = Cov(\boldsymbol{T}\boldsymbol{X})$ in Equation (2.2), we have

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

$\boldsymbol{D} = \boldsymbol{T}\boldsymbol{\Sigma}\boldsymbol{T}'$, which consequently leads to

$$\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1} = \boldsymbol{T}'\boldsymbol{D}^{-1}\boldsymbol{T}, \tag{2.3}$$

where $\boldsymbol{D}^{-1} = \text{diag}(\boldsymbol{D}_1^{-1}, \boldsymbol{D}_2^{-1}, \ldots, \boldsymbol{D}_M^{-1})$. As a result, the BCD method reduces the challenge of modeling an inverse covariance matrix based on the partial information of the variable ordination into the problem of estimating $(M-1)$ multivariate linear regressions. Because of the decomposition (2.3), the general sparsity in $\boldsymbol{T}$ would induce some sparsity in $\boldsymbol{\Omega}$ (although it may not induce certain structured sparsity), which is easily implemented by the regularization on the linear regressions in (2.1). The decomposition (2.3) also indicates that the BCD method can guarantee the positive definiteness of the resulting estimate of $\boldsymbol{\Omega}$ provided that $\boldsymbol{D}_1^{-1}, \boldsymbol{D}_2^{-1}, \ldots, \boldsymbol{D}_M^{-1}$ are all positive definite. Note that there is no constraint required for parameters in matrix $\boldsymbol{T}$. We would like to remark that although the proposed model needs information on the group ordering, it is invariant to the permutation of variables within each group because the coefficient estimation of multivariate regressions in (2.1) is not affected by the ordering of variables in $\boldsymbol{X}^{(j)}$ and $\boldsymbol{Z}^{(j)}$. This point is verified in the simulation study. Moreover, it is seen that the MCD is a special case of the proposed BCD with $M = p$ (i.e., $p_j = 1$ for $j = 1, 2, \ldots, M$), which represents the case where the p variables have a full ordering information.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

2.2 Parameter Estimation

Denote by $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ the n independently and identically distributed observations from the multivariate normal distribution $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Omega}^{-1})$. Let $\mathbb{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)'$ be the $n \times p$ data matrix. Based on the partial information of the variable ordination, we partition $\mathbb{X}$ by columns and write $\mathbb{X} = (\mathbb{X}^{(1)}, \mathbb{X}^{(2)}, \ldots, \mathbb{X}^{(M)})$, where $\mathbb{X}^{(j)}$ represents the $n \times p_j$ sub-data matrix corresponding to the jth variable group $\boldsymbol{X}^{(j)}$. Based on the methodology of BCD, we need to model a set of multivariate regressions in (2.1) to obtain the estimates of block Cholesky factors matrices $\boldsymbol{T}$ and $\boldsymbol{D}^{-1}$. The negative joint log-likelihood function is expressed as

$$L(\boldsymbol{A}, \boldsymbol{D}^{-1}) = \sum_{j=1}^{M} \left\{ -\log |\boldsymbol{D}_j^{-1}| + \operatorname{tr}\left[\boldsymbol{S}_{\epsilon_j} \boldsymbol{D}_j^{-1}\right] \right\}, \tag{2.4}$$

up to some constant. The symbol $\boldsymbol{S}_{\epsilon_j} = \frac{1}{}$

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

$$= \sum_{j=1}^{M} \left\{ -\log |\boldsymbol{D}_j^{-1}| + \operatorname{tr}\left[\boldsymbol{S}_{\epsilon_j} \boldsymbol{D}_j^{-1}\right] + \lambda_1 \|\boldsymbol{A}_j\|_1 + \lambda_2 \|\boldsymbol{D}_j^{-1}\|_1^- \right\}$$

$$\triangleq \sum_{j=1}^{M} \ell_\lambda(\boldsymbol{A}_j, \boldsymbol{D}_j^{-1}), \tag{2.5}$$

where $\ell_\lambda(\boldsymbol{A}_j, \boldsymbol{D}_j^{-1}) = -\log |\boldsymbol{D}_j^{-1}| + \operatorname{tr}\left[\boldsymbol{S}_{\epsilon_j} \boldsymbol{D}_j^{-1}\right] + \lambda_1 \|\boldsymbol{A}_j\|_1 + \lambda_2 \|\boldsymbol{D}_j^{-1}\|_1^-$ with $\boldsymbol{A}_1 = \boldsymbol{0}$ being zero matrix. Here $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are tuning parameters, the matrix norm $||\boldsymbol{B}||_1 = \sum_{i,j} |b_{ij}|$, and $||\boldsymbol{B}||_1^- = \sum_{i \neq j} |b_{ij}|$ with b_{ij} being the elements of matrix $\boldsymbol{B}$.

Note that different components $\ell_\lambda(\boldsymbol{A}_j, \boldsymbol{D}_j^{-1})$ contain different parameters $\boldsymbol{A}_j$ and $\boldsymbol{D}_j^{-1}$. Hence minimizing $L_\lambda(\boldsymbol{A}, \boldsymbol{D}^{-1})$ is equivalent to minimizing each $\ell_\lambda(\boldsymbol{A}_j, \boldsymbol{D}_j^{-1})$ separately, which facilitates a parallel computing procedure for estimating $(\boldsymbol{A}_j, \boldsymbol{D}_j^{-1})$ simultaneously to save computational time. For each minimization of $\ell_\lambda(\boldsymbol{A}_j, \boldsymbol{D}_j^{-1})$, although the objective function is not convex with respect to $(\boldsymbol{A}_j, \boldsymbol{D}_j^{-1})$, it is a biconvex optimization (Gorski et al., 2007). That is, $\ell_\lambda(\boldsymbol{A}_j, \boldsymbol{D}_j^{-1})$ is convex over $\boldsymbol{A}_j$ when fixing $\boldsymbol{D}_j^{-1}$, and is also convex over $\boldsymbol{D}_j^{-1}$ when fixing $\boldsymbol{A}_j$. This property enables us to apply a coordinate descent algorithm to iteratively estimate $\boldsymbol{A}_j$ by minimizing $\ell_\lambda(\boldsymbol{A}_j, \boldsymbol{D}_j^{-1})$ for a given $\boldsymbol{D}_j^{-1} = (\boldsymbol{D}_j^{-1})_*$, and estimate $\boldsymbol{D}_j^{-1}$ by minimizing $\ell_\lambda(\boldsymbol{A}_j, \boldsymbol{D}_j^{-1})$ for a given $\boldsymbol{A}_j = (\boldsymbol{A}_j)_*$ (Rothman et al., 2010*b*; Sofer et al., 2014). Specifically, for a given $(\boldsymbol{D}_j^{-1})_*$, we solve

$$\hat{\boldsymbol{A}}_j[(\boldsymbol{D}_j^{-1})_*] = \arg\min_{\boldsymbol{A}_j} \ell_\lambda(\boldsymbol{A}_j | (\boldsymbol{D}_j^{-1})_*)$$

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

$$= \arg\min_{\boldsymbol{A}_j}\{\frac{1}{n}(\tilde{\mathbb{X}}^{(j)} - \tilde{\mathbb{Z}}^{(j)}\tilde{\boldsymbol{A}}_j)'(\tilde{\mathbb{X}}^{(j)} - \tilde{\mathbb{Z}}^{(j)}\tilde{\boldsymbol{A}}_j) + \lambda_1\|\tilde{\boldsymbol{A}}_j\|_1\}, \tag{2.6}$$

where $\tilde{\mathbb{X}}^{(j)} = \text{vec}[\mathbb{X}^{(j)}(\boldsymbol{D}_j^{-\frac{1}{2}})_*]$, $\tilde{\mathbb{Z}}^{(j)} = (\boldsymbol{D}_j^{-\frac{1}{2}})_* \otimes \mathbb{Z}^{(j)}$ and $\tilde{\boldsymbol{A}}_j = \text{vec}[\boldsymbol{A}_j']$. Here vec denotes the vectorization operator and $\otimes$ the Kronecker product, and their property $\text{vec}(\boldsymbol{ABC}) = (\boldsymbol{C}' \otimes \boldsymbol{A})\text{vec}(\boldsymbol{B})$ is applied. The optimization problem (2.6) is thus a linear regression with Lasso penalty (Tibshirani, 1996). The initial value for $(\boldsymbol{D}_j^{-1})_*$ is set to be the identity matrix. On the other hand, for a given $(\boldsymbol{A}_j)_*$, we solve

$$\begin{aligned}\hat{\boldsymbol{D}}_j^{-1}[(\boldsymbol{A}_j)_*] &= \arg\min_{\boldsymbol{D}_j^{-1}} \ell_\lambda(\boldsymbol{D}_j^{-1}|(\boldsymbol{A}_j)_*) \\ &= \arg\min_{\boldsymbol{D}_j^{-1}} \left\{-\log|\boldsymbol{D}_j^{-1}| + \text{tr}\left[(\boldsymbol{S}_{\epsilon_j})_*\boldsymbol{D}_j^{-1}\right] + \lambda_2\|\boldsymbol{D}_j^{-1}\|_1^-\right\},\end{aligned} \tag{2.7}$$

where $(\boldsymbol{S}_{\epsilon_j})_* = \frac{1}{n}[\mathbb{X}^{(j)} - \mathbb{Z}^{(j)}(\boldsymbol{A}_j')_*]'[\mathbb{X}^{(j)} - \mathbb{Z}^{(j)}(\boldsymbol{A}_j')_*]$. It has the same form as Glasso estimation. Accordingly, the estimates $\hat{\boldsymbol{A}}_j$ and $\hat{\boldsymbol{D}}_j^{-1}$ are iteratively solved from optimization problems (2.6) and (2.7) until convergence. After obtaining the estimates $\hat{\boldsymbol{A}}_j$ and $\hat{\boldsymbol{D}}_j^{-1}$, we construct the block Choleksy factor estimate $\hat{\boldsymbol{T}}$ with $-\hat{\boldsymbol{A}}_j$ as the jth block row and the identity matrix being the block diagonal. The estimate $\hat{\boldsymbol{D}}^{-1}$ is constructed with $\hat{\boldsymbol{D}}_j^{-1}$ as its jth block diagonal. Then $\hat{\boldsymbol{\Omega}} = \hat{\boldsymbol{T}}'\hat{\boldsymbol{D}}^{-1}\hat{\boldsymbol{T}}$ is a sparse estimate of inverse covariance matrix under the partial information of variable ordination. We briefly summarize the above estimation procedure in Algorithm 1.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Algorithm 1

Input: Data $\mathbb{X}$, tuning parameters λ_1 and λ_2.

Output: Estimate $\hat{\boldsymbol{\Omega}}(\lambda_1, \lambda_2)$ corresponding to λ_1 and λ_2.

For $j = 1$ to M do

Step 0: Set an initial value of $\boldsymbol{D}_j = \boldsymbol{I}$.

Step 1: Given $\boldsymbol{D}_j^{-1} = \hat{\boldsymbol{D}}_{j;t}^{-1}$, solve $\boldsymbol{A}_j$ in (2.6) by the Lasso technique.

Step 2: Given $\boldsymbol{A}_j = \hat{\boldsymbol{A}}_{j;t}$, solve $\boldsymbol{D}_j^{-1}$ in (2.7) by the Glasso technique.

Step 3: Repeat Steps 1 and 2 till both $\hat{\boldsymbol{A}}_j = \hat{\boldsymbol{A}}_{j;t}$ and $\hat{\boldsymbol{D}}_j^{-1} = \hat{\boldsymbol{D}}_{j;t}^{-1}$ converge.

End

Step 4: $\hat{\boldsymbol{T}} = \boldsymbol{I} - \hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{D}}^{-1} = \mathrm{diag}(\hat{\boldsymbol{D}}_1^{-1}, \ldots, \hat{\boldsymbol{D}}_M^{-1})$, then $\hat{\boldsymbol{\Omega}}(\lambda_1, \lambda_2) = \hat{\boldsymbol{T}}'\hat{\boldsymbol{D}}^{-1}\hat{\boldsymbol{T}}$.

Here $\hat{\boldsymbol{A}}_{j;t}$ and $\hat{\boldsymbol{D}}_{j;t}$ represent the estimates of $\boldsymbol{A}_j$ and $\boldsymbol{D}_j$ in the tth iteration. The convergence criteria are $||\hat{\boldsymbol{A}}_{j;t} - \hat{\boldsymbol{A}}_{j;t-1}||_F^2 < \tau_1$ and $||\hat{\boldsymbol{D}}_{j;t} - \hat{\boldsymbol{D}}_{j;t-1}||_F^2 < \tau_2$, where τ_1 and τ_2 are two pre-selected small quantities, and $||\cdot||_F$ stands for the Frobenius norm. Since the objective (2.5) is not joint convex, there is no guarantee of finding the global minimum. However, Algorithm 1 uses a coordinate descent to compute a local solution of (2.5). Steps 1 and 2 both ensure a decrease in the value of objective, leading to the convergence of $\boldsymbol{A}_j$ and $\boldsymbol{D}_j$. In Step 3, we also set a maximum number of iterations as 100 in case that the parameter estimation is not empirically converged. However, in the simulation we have tried, almost all the parameter estimates satisfy the convergence criteria quickly with several tens of iterations.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Note that there are two tuning parameters λ_1 and λ_2 in the objective function (2.5). To choose their optimal values, studies in the literature often suggest cross-validation, information criteria, independent validation set mechanism and so forth. In this work the BIC (Bayesian information criterion) proposed by Yuan and Lin (2007) is adopted to determine the optimal values of tuning parameters as follows

$$\mathrm{BIC}(\lambda_1, \lambda_2) = -\log|\hat{\boldsymbol{\Omega}}(\lambda_1, \lambda_2)| + \mathrm{tr}[\hat{\boldsymbol{\Omega}}(\lambda_1, \lambda_2)\boldsymbol{S}] + \frac{\log n}{}$$

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

2.3 Comparison and Connection with Several Existing Methods

The proposed BCD method obtains a sparse inverse covariance matrix estimate under the partial information of variable ordination. Both of the Cholesky factors' estimates $\hat{\boldsymbol{T}}$ and $\hat{\boldsymbol{D}}^{-1}$ contain the sparsity resulting from a set of penalized multivariate regressions. It makes close connection with several existing methods.

Firstly, we demonstrate that the MCD, Glasso, Rothman et al. (2010*a*) and Witten et al. (2011) are all special cases of the proposed method. The MCD method for estimating $\boldsymbol{\Omega}$ is a special case when $M = p$, indicating that there is a full ordering information. When data have partial ordering information, the MCD is not suitable. The MCD needs a pre-specified full ordering before analyzing data, but the variables have no ordering within each group. Therefore, one needs to identify an ordering before applying the MCD. However, different orderings would lead to different estimates (Chang and Tsay, 2010), and an incorrectly identified ordering would result in an inaccurate estimate. Additionally, the Glasso is a special case of the proposed BCD with $M = 1$, implying that there is no ordering information. Besides, Rothman et al. (2010*a*) studied a banded estimate of $\boldsymbol{\Omega}$ via MCD. Their approach can also be viewed as a special case of the proposed BCD model by $M = p$ and regressing $\boldsymbol{X}^{(j)}$ only on its several nearest previous group variables instead of all the previous group variables in Equation (2.1). Furthermore, the proposed BCD method can be easily extended to estimate the inverse covariance matrix with a banded block structure.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

In addition, Witten et al. (2011) introduced a block diagonal inverse covariance estimation with each block obtained by the Glasso on the corresponding group variables. Their method assumed that variables in different groups are independent. Thus their estimator is also a special case of the proposed BCD, in the sense that the block Cholesky factor $\boldsymbol{T}$ becomes the identity matrix under the independence assumption between group variables, and each $\boldsymbol{D}_j^{-1}$ is then estimated from Glasso on $\mathbb{X}^{(j)}$. The estimate $\hat{\boldsymbol{\Omega}} = \hat{\boldsymbol{T}}'\hat{\boldsymbol{D}}^{-1}\hat{\boldsymbol{T}}$ is the same as that in Witten et al. (2011).

Secondly, we compare several methods from perspective of ordering information. When there is no ordering among variables, apart from Glasso which penalizes likelihood function, some papers investigated matrix estimation through penalized pseudo-likelihood. They solve their optimization problems often in a column-by-column fashion, such as Yuan (2010); Cai et al. (2011); Liu and Luo (2015); Liu and Wang (2017) and so forth. Liu and Luo (2015) extended the idea of Cai et al. (2011) and proposed SCIO estimator as

$$\hat{\boldsymbol{\beta}}_i = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{} \right.$$

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

step is also needed to make their estimates symmetric. In the contrast, the proposed BCD estimate is itself symmetric and positive definite.

On the other hand, when variables have a full information on the ordering, two recent works Yu and Bien (2017) and Khare et al. (2019) proposed to estimate $\boldsymbol{\Omega}$ via the classical Cholesky decomposition $\boldsymbol{\Omega} = \boldsymbol{L}'\boldsymbol{L}$, where the Cholesky factor $\boldsymbol{L}$ is a lower triangular matrix. Both of such decomposition and the MCD induce the sparse estimates by the sparsity in the Cholesky factors. The advantage of classical Cholesky decomposition is to directly result in a convex objective with respect to parameter $\boldsymbol{L}$, hence guaranteeing a global convergence with an appropriate penalty. But it lacks statistical meanings and interpretation as explicit as the MCD technique. Yu and Bien (2017) assumed "local dependence" in the ordered data in the sense that the jth variable is correlated with its K_j nearest variables, where K_j can be different. That is, they assumed structured sparse pattern in the underlying $\boldsymbol{\Omega}$, while the proposed BCD is suitable for a general or unstructured sparsity. Khare et al. (2019) introduced CSCS estimator, which accommodates the unstructured sparsity by imposing a Lasso-type penalty on the likelihood function in terms of $\boldsymbol{L}$, and developed a cyclic coordinate algorithm which leads to a closed form of solution for the estimates of each row of $\boldsymbol{L}$. To the best of our knowledge, few works have contributed to the inverse covariance estimation when variables have partial ordering information.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

3. Theoretical Properties

In this section, we establish the asymptotically theoretical properties for the proposed BCD estimator. To facilitate the presentation and proofs, we introduce some notation and make assumptions on the true model. Let $\boldsymbol{\Omega}_0 = \boldsymbol{T}_0'\boldsymbol{D}_0^{-1}\boldsymbol{T}_0$ be the underlying inverse covariance matrix with its block MCD according to the group variables $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \ldots, \boldsymbol{X}^{(M)}$. Let $\boldsymbol{T}_j = -\boldsymbol{A}_j$, and $\boldsymbol{T}_{j0}$ be the counterpart of $\boldsymbol{T}_j$ in the block Cholesky factor $\boldsymbol{T}_0$. That is, $\boldsymbol{T}_{j0}$ is the jth block row in the lower triangular part of matrix $\boldsymbol{T}_0$. Define $\mathcal{Z}_{T_j} = \{(i,k) : (\boldsymbol{T}_{j0})_{ik} \neq 0\}$ as the collection of nonzero elements in the matrix $\boldsymbol{T}_{j0}$. Similarly, denote the counterpart of $\boldsymbol{D}_j$ in the block Cholesky factor $\boldsymbol{D}$ by the matrix $\boldsymbol{D}_{j0}$, which is the jth block diagonal of matrix $\boldsymbol{D}_0$. Let $\mathcal{Z}_{D_j} = \{(i,k) : i \neq k, (\boldsymbol{D}_{j0}^{-1})_{ik} \neq 0\}$ be the collection of nonzero off-diagonal elements in the matrix $\boldsymbol{D}_0^{-1}$. Denote by s_{T_j} and s_{D_j} the cardinality of $\mathcal{Z}_{T_j}$ and $\mathcal{Z}_{D_j}$, respectively. Let $s_T = \sum_{j=1}^{M} s_{T_j}$ and $s_D = \sum_{j=1}^{M} s_{D_j}$. In order to achieve the asymptotic consistent property of the proposed estimator, a mild condition is needed that there exists a constant $\theta > 0$ such that the singular values of $\boldsymbol{\Omega}_0$ are bounded as

$$1/\theta < \varphi_p(\boldsymbol{\Omega}_0) \leq \varphi_1(\boldsymbol{\Omega}_0) < \theta, \tag{3.8}$$

where $\varphi_1(\boldsymbol{B}), \varphi_2(\boldsymbol{B}), \ldots, \varphi_p(\boldsymbol{B})$ represent the singular values of matrix $\boldsymbol{B}$ in a decreasing order. This assumption is made to guarantee the positive definiteness of $\boldsymbol{\Omega}_0$.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

In addition, assume that there exist constants $0 \leq C_j \leq 1$ and $\gamma_j \geq 1$ such that

$$p_j = \gamma_j p^{C_j} \quad \text{for} \quad j = 1, 2, \ldots, M. \tag{3.9}$$

Assumption (3.9) characterizes the relation between each p_j and p for high-dimensional data. If p diverges to infinity, $C_j = 0$ corresponds to the case where p_j is fixed and controlled by γ_j, while $0 < \tau_c \leq C_j \leq 1$ implies that p_j also diverges, but not faster than p, where τ_c is some positive arbitrarily small number. Denote $\mathcal{Z}_C = \{j : 0 < \tau_c \leq C_j \leq 1\}$, the set of indices that the corresponding groups have a diverging number of variables. Now, we present the main results in Theorems 1 and 2.

Theorem 1. *Suppose that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are n independently and identically distributed observations from $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Omega}^{-1})$. Let $(\boldsymbol{D}_j^{-1})_*$ and $(\boldsymbol{A}_j)_*$ be any estimates of $\boldsymbol{D}_j^{-1}$ and $\boldsymbol{A}_j$ obtained from the path of Algorithm 1. Under* (3.8) *and* (3.9)*, assume that the tuning parameters λ_1 and λ_2 satisfy $\lambda_1 = O(\sqrt{\log(p)/n})$, $\lambda_2 = O(\sqrt{\log(p)/n})$, then*

(a) there exists a local minimum $\hat{\boldsymbol{A}}_j (= -\hat{\boldsymbol{T}}_j)$ of $\ell_\lambda(\boldsymbol{A}_j | (\boldsymbol{D}_j^{-1})_*)$ such that $\|\hat{\boldsymbol{T}}_j - \boldsymbol{T}_{j0}\|_F = O_p(s_{T_j} \log(\sum_{k=1}^{j} p_k)/n)$.

(b) there exists a local minimum $\hat{\boldsymbol{D}}_j$ of $\ell_\lambda(\boldsymbol{D}_j^{-1} | (\boldsymbol{A}_j)_*)$ such that $\|\hat{\boldsymbol{D}}_j - \boldsymbol{D}_{j0}\|_F = O_p((s_{D_j} + p_j) \log p_j / n)$.

Theorem 2. *Let $\hat{\boldsymbol{\Omega}}$ be the estimate of $\boldsymbol{\Omega}$ obtained by Algorithm 1. Assume all the*

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

assumptions in Theorem 1 hold, then we have

$$\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0\|_F = O_p\left(\sqrt{\frac{\qquad\qquad}{n}}\right).$$

Under the condition $s_T \log p + \sum_{j=1}^{M}(s_{D_j} + p_j)\log p_j = o(n)$, $\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0\|_F \xrightarrow{P} 0$.

Theorem 1 provides asymptotic consistent rates of the estimators $\hat{\boldsymbol{T}}_j$ and $\hat{\boldsymbol{D}}_j$ that are obtained by minimizing $\ell_\lambda(\boldsymbol{A}_j|(\boldsymbol{D}_j^{-1})_*)$ in (2.6) and minimizing $\ell_\lambda(\boldsymbol{D}_j^{-1}|(\boldsymbol{A}_j)_*)$ in (2.7). Theorem 2 establishes the property for the inverse covariance estimate $\hat{\boldsymbol{\Omega}}$ obtained from the proposed Algorithm 1, which demonstrates the consistent rate of our estimate in practice. Moreover, we would like to have the following remarks.

First, from Theorem 1 we have $\|\hat{\boldsymbol{D}}^{-1} - \boldsymbol{D}_0^{-1}\|_F^2 = \|\hat{\boldsymbol{D}} - \boldsymbol{D}_0\|_F^2 = \sum_{j=1}^{M}\|\hat{\boldsymbol{D}}_j - \boldsymbol{D}_{j0}\|_F = O_p(\sum_{j=1}^{M}(s_{D_j} + p_j)\log(p_j)/n)$. Such rate is sharper than the theoretical Glasso consistent rate $(\tilde{s} + p)\log(p)/n$ ($\tilde{s}$ is a measure of sparsity), which has been obtained in the literature (Lam and Fan, 2009), because of $\sum_{j=1}^{M} p_j \log p_j \leq p \log p$ resulting from $\prod_{j=1}^{M} p_j^{p_j} \leq \prod_{j=1}^{M} p_{\max}^{p_j} = p_{\max}^{p} \leq p^p$, where $p_{\max} = \max\{p_j\}_{j=1}^{M}$. The equality holds only in the situation that all the p variables are in one group, where it is exactly a Glasso problem. This implies that, as long as there are at least two groups of variables, the proposed BCD model with partial information of variable ordination is useful in reducing the consistent rate.

Second, the estimation of $\boldsymbol{D}$ is decomposed into M separate Glasso estimations in

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

the proposed method. According to the log sum inequality, we have $\sum_{j=1}^{M} p_j \log p_j \geq p \log(p/M)$. The equality holds when $p_j = p/M$ for all $j = 1, 2, \ldots, M$, indicating that the lowest bound of the proposed estimator's consistent rate would be achieved when each group has the equal number of variables. In view of this, we would like to point out that: (i) a larger number of variable groups leads to a smaller value of the lowest bound of the consistent rate; (ii) the more evenly that p variables are assigned into M groups, the more closely that the consistent rate tends to the lowest bound.

Third, the requirement $s_T \log p + \sum_{j=1}^{M}(s_{D_j} + p_j) \log p_j = o(n)$ in Theorem 2 is a relatively weaker condition compared with the assumption for the Glasso model. Note that $\|\hat{\boldsymbol{D}} - \boldsymbol{D}_0\|_F^2 \leq O_p(\sum_{j=1}^{M}(s_{D_j} + p_j) \log(p_{\max})/n) = O_p((s_D + p) \log(p_{\max})/n)$, then we may require a stronger condition $s_T \log p + (s_D + p) \log p_{\max} = o(n)$. Moreover, if we further loose the upper bound of $\|\hat{\boldsymbol{D}} - \boldsymbol{D}_0\|_F^2 \leq O_p((s_D + p) \log(p)/n)$, the proposed BCD model needs an even stronger condition $(s_T + s_D + p) \log p = o(n)$, which however is the similar condition as that for the Glasso estimator.

4. Numerical Study

In this section, we conduct simulation studies to evaluate the performance of the proposed BCD model (Prop) in comparison with several existing methods, including the MCD method, Glasso, SCIO (Liu and Luo, 2015) and CSCS (Khare et al., 2019). The MCD, where $M = p$, reduces Equation (2.1) to $(p - 1)$ univariate linear

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

regressions with their coefficients estimated by Lasso. The Glasso and SCIO are two popular methods, but do not consider ordering information. The CSCS is suitable for data with full ordering information and implemented via classical Cholesky decomposition. The tuning parameters in the Glasso, SCIO and CSCS methods are selected based on BIC. Besides, to examine whether the proposed model is permutation invariant within each group, we also implement the Prop* method, which estimates $\mathbf{\Omega}$ by Algorithm 1 from data which randomly permutate variables within each group.

The data are independently generated from $\mathcal{N}_p(\mathbf{0}, \mathbf{\Omega}^{-1})$ with sample size $n = 50$ and number of variables $p = 200$. We consider two different cases of variable groups: (1) five groups with each containing 40 variables; (2) four groups with each subsequently containing 30, 60, 40 and 70 variables. Let $\mathrm{AR}(\rho)$ represent a squared matrix with autoregressive structure with (i, j)th entry as $\rho^{|i-j|}$, $1 \leq i, j \leq p$. Let $\mathrm{MA}(0.5, 0.4, 0.3)$ indicate a squared banded matrix with the main diagonal elements 1, and the subsequent sub-diagonal elements are 0.5, 0.4 and 0.3 respectively. Denote an $a \times b$ non-squared matrix $\widetilde{\mathrm{AR}}(\rho) = (\mathrm{AR}(\rho), \mathbf{0})$ if $a < b$, and $\widetilde{\mathrm{AR}}(\rho) = \begin{pmatrix} \mathrm{AR}(\rho) \\ \mathbf{0} \end{pmatrix}$ otherwise, where $\mathbf{0}$ represents the matrix with all elements 0. Similarly, Denote an $a \times b$ non-squared matrix $\widetilde{\mathrm{MA}}(0.5, 0.4, 0.3) = (\mathrm{MA}(0.5, 0.4, 0.3), \mathbf{0})$ if $a < b$, and $\widetilde{\mathrm{MA}}(0.5, 0.4, 0.3) = \begin{pmatrix} \mathrm{MA}(0.5, 0.4, 0.3) \\ \mathbf{0} \end{pmatrix}$ otherwise. To systematically investigate the performance of the proposed method, we consider the following different struc-

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

tures of inverse covariance matrix $\boldsymbol{\Omega}$.

- **Scenario 1**. $\boldsymbol{\Omega}_1 = \mathrm{AR}(0.8)$.
- **Scenario 2**. $\boldsymbol{\Omega}_2 = \begin{pmatrix} \mathrm{AR}(0.5) & \dots & \mathbf{0} \\ \vdots & \ddots & \mathbf{0} \\ \mathbf{0} & \dots & \mathrm{AR}(0.5) \end{pmatrix}$.
- **Scenario 3**.
$$\boldsymbol{\Omega}_3 = \begin{pmatrix} \mathrm{MA}(0.5,0.4,0.3) & \widetilde{\mathrm{AR}}(0.5) & \dots & \widetilde{\mathrm{AR}}(0.5) \\ \widetilde{\mathrm{AR}}(0.5) & \mathrm{MA}(0.5,0.4,0.3) & \dots & \widetilde{\mathrm{AR}}(0.5) \\ \vdots & \widetilde{\mathrm{AR}}(0.5) & \ddots & \widetilde{\mathrm{AR}}(0.5) \\ \widetilde{\mathrm{AR}}(0.5) & \widetilde{\mathrm{AR}}(0.5) & \vdots & \mathrm{MA}(0.5,0.4,0.3) \end{pmatrix} + \boldsymbol{\alpha I}.$$
The value of $\boldsymbol{\alpha}$ is gradually increased to ensure that $\boldsymbol{\Omega}_3$ is positive definite.
- **Scenario 4**. $\boldsymbol{\Omega}_4 = \tilde{\boldsymbol{\Omega}}_4 + \boldsymbol{\alpha I}$, where $\tilde{\boldsymbol{\Omega}}_4$ is generated by randomly permuting rows and corresponding columns of each block of
$$\begin{pmatrix} \mathrm{AR}(0.5) & \widetilde{\mathrm{MA}}(0.5,0.4,0.3) & \dots & \widetilde{\mathrm{MA}}(0.5,0.4,0.3) \\ \widetilde{\mathrm{MA}}(0.5,0.4,0.3) & \mathrm{AR}(0.5) & \dots & \widetilde{\mathrm{MA}}(0.5,0.4,0.3) \\ \vdots & \widetilde{\mathrm{MA}}(0.5,0.4,0.3) & \ddots & \widetilde{\mathrm{MA}}(0.5,0.4,0.3) \\ \widetilde{\mathrm{MA}}(0.5,0.4,0.3) & \widetilde{\mathrm{MA}}(0.5,0.4,0.3) & \vdots & \mathrm{AR}(0.5) \end{pmatrix}.$$
The value of $\boldsymbol{\alpha}$ is gradually increased to ensure that $\boldsymbol{\Omega}_4$ is positive definite.
- **Scenario 5**. $\boldsymbol{\Omega}_5 = \boldsymbol{B}'\boldsymbol{H}\boldsymbol{B}$, where $\boldsymbol{H}$ is a block diagonal matrix with its each diagonal block as AR(0.5), and $\boldsymbol{B} = (b_{i,j})$ with $b_{i,i} = 1, b_{i+20,i} = -0.8$ and

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

$b_{i,j} = 0$ otherwise.

- **Scenario 6**. $\boldsymbol{\Omega}_6 = \boldsymbol{B}'\boldsymbol{H}\boldsymbol{B}$, where $\boldsymbol{H}$ is a block diagonal matrix with its each diagonal block being MA(0.5,0.4,0.3), and $\boldsymbol{B} = (b_{i,j})$ with $b_{i,i} = 1, b_{i+20,i} = -0.8, b_{i+21,i} = 0.5$ and $b_{i,j} = 0$ otherwise.

- **Scenario 7**. $\boldsymbol{\Omega}_7 = \tilde{\boldsymbol{\Omega}}_7 + \alpha \boldsymbol{I}$. Here the diagonal elements of $\tilde{\boldsymbol{\Omega}}_7$ are 0, and each off-diagonal element is generated independently as $(\tilde{\boldsymbol{\Omega}}_7)_{ij} = (\tilde{\boldsymbol{\Omega}}_7)_{ji} = b * Unif(-1, 1)$, where b is a Bernoulli random variable with probability 0.15 equal 1. The value of α increases gradually to make sure $\boldsymbol{\Omega}_7$ is positive definite.

Scenario 1 is the AR structure with the variables' correlations decaying when they are far apart from each other. **Scenarios 2** is a block diagonal matrix with multiple groups of variables, where the variables in different groups are independent. **Scenarios 3** and **4** are block matrices with multiple groups of variables with variables in different groups possibly correlated. **Scenarios 5** and **6** are similarly used in Huang et al. (2006). **Scenario 7** is a general sparse matrix with random structure.

To evaluate the accuracy of each estimate $\hat{\boldsymbol{\Omega}} = (\hat{\omega}_{ij})$ for the underlying inverse covariance matrix $\boldsymbol{\Omega} = (\omega_{ij})$, we consider the loss measures L_1, the matrix spectral norm L_2, the Frobenius norm F of $(\boldsymbol{\Omega} - \hat{\boldsymbol{\Omega}})$ as follows

$$L_1 = \max_j \sum_i |\hat{\omega}_{ij} - \omega_{ij}|, \quad L_2 = \lambda_{\max}[(\boldsymbol{\Omega} - \hat{\boldsymbol{\Omega}})], \quad \text{Fnorm} = \sqrt{\sum_{i=1}^{p}\sum_{j=1}^{p}(\hat{\omega}_{ij} - \omega_{ij})^2},$$

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

where $\lambda_{\max}$ is the maximum value of eigenvalues. We also use the Kullback-Leibler loss (KL) as well as the quadratic loss (QL)

$$\mathrm{KL} = \frac{1}{p}\,(\mathrm{tr}[\boldsymbol{\Omega}^{-1}\hat{\boldsymbol{\Omega}}] - \log|\boldsymbol{\Omega}^{-1}\hat{\boldsymbol{\Omega}}| - p), \quad \mathrm{QL} = \frac{1}{p}\,\mathrm{tr}(\boldsymbol{\Omega}^{-1}\hat{\boldsymbol{\Omega}} - \boldsymbol{I})^2.$$

In addition, to gauge the ability of the proposed model to capture the underlying sparse structure, we report the false selection loss FSL = (FP + FN) / p^2 in percentage, where FP is the false positive and FN is the false negative. The simulation results of loss measures for each method are summarized in Tables 1 and 2, reporting their averages and corresponding standard errors (in parenthesis) over 50 replicates.

Overall, the Prop method gives relatively superior performance compared with other methods for the considered loss measures. Especially it performs substantially well with respect to KL and QL for all the scenarios. Note that the results of Prop* are omitted in the tables since they are exactly the same as the results of the Prop method. It further confirms that the proposed BCD model is permutation invariant for the variables within each group. In scenario 1, the CSCS and Prop are the best models. The CSCS method is slight better than Prop regarding L_1, L_2 and Fnorm since it is designed for inverse covariance estimation of data with full information on variable ordering. For other scenarios 2-7, we observe that (1) the Prop is at least comparable with or slightly better than CSCS method; (2) the Prop outperforms

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Table 1: The averages and standard errors (in parenthesis) of estimates for 5 groups of variables.

Scenario		L_1	L_2	Fnorm	KL	QL	FSL (%)
1	MCD	50.25 (6.628)	16.66 (2.187)	196.5 (24.94)	4.000 (0.670)	4.738 (1.027)	34.75 (0.302)
	Glasso	8.852 (0.001)	8.717 (0.001)	28.61 (0.004)	0.518 (0.001)	0.601 (0.004)	27.67 (0.001)
	SCIO	8.843 (0.002)	8.685 (0.001)	28.43 (0.004)	0.451 (0.003)	0.655 (0.005)	27.35 (0.006)
	CSCS	8.750 (0.004)	8.577 (0.002)	28.06 (0.009)	0.415 (0.002)	0.505 (0.005)	27.30 (0.019)
	Prop	8.768 (0.004)	8.581 (0.001)	27.87 (0.006)	0.363 (0.001)	0.397 (0.004)	25.89 (0.022)
2	MCD	41.23 (15.07)	32.46 (4.881)	34.83 (5.265)	0.366 (0.054)	2.921 (0.699)	29.53 (0.580)
	Glasso	2.613 (0.004)	2.379 (0.003)	12.57 (0.009)	0.242 (0.001)	0.410 (0.003)	7.875 (0.001)
	SCIO	2.618 (0.005)	2.363 (0.003)	12.48 (0.010)	0.240 (0.003)	0.419 (0.003)	7.769 (0.003)
	CSCS	2.586 (0.006)	2.324 (0.005)	12.20 (0.031)	0.222 (0.002)	0.390 (0.003)	7.711 (0.010)
	Prop	2.539 (0.007)	2.259 (0.002)	11.78 (0.008)	0.195 (0.001)	0.340 (0.003)	7.343 (0.006)
3	MCD	61.20 (5.725)	205.6 (20.88)	245.2 (25.91)	4.671 (0.303)	28.53 (6.462)	49.24 (0.163)
	Glasso	9.592 (0.001)	9.414 (0.001)	25.87 (0.004)	0.778 (0.001)	2.952 (0.023)	36.35 (0.001)
	SCIO	9.579 (0.002)	9.348 (0.001)	25.62 (0.004)	0.703 (0.003)	3.506 (0.029)	35.99 (0.007)
	CSCS	9.543 (0.007)	9.153 (0.005)	25.14 (0.016)	0.591 (0.003)	1.604 (0.024)	34.72 (0.032)
	Prop	9.530 (0.005)	9.140 (0.001)	25.09 (0.003)	0.424 (0.001)	0.681 (0.009)	36.99 (0.018)
4	MCD	50.70 (6.839)	16.30 (2.144)	19.04 (2.179)	3.108 (0.210)	54.44 (11.78)	47.52 (0.204)
	Glasso	4.221 (0.031)	3.474 (0.012)	11.33 (0.111)	0.568 (0.041)	10.79 (2.661)	21.14 (0.079)
	SCIO	3.936 (0.001)	3.546 (0.001)	11.91 (0.012)	0.597 (0.083)	9.475 (1.973)	19.70 (0.001)
	CSCS	6.880 (0.140)	3.300 (0.006)	10.12 (0.032)	0.566 (0.019)	2.601 (0.067)	22.23 (0.109)
	Prop	4.341 (0.010)	3.373 (0.002)	10.50 (0.008)	0.466 (0.001)	2.075 (0.043)	21.12 (0.014)
5	MCD	136.7 (13.75)	50.75 (5.017)	67.67 (5.552)	1.282 (0.086)	47.12 (6.803)	31.27 (0.281)
	Glasso	9.354 (0.005)	8.612 (0.005)	30.83 (0.029)	0.626 (0.006)	8.278 (0.366)	23.54 (0.013)
	SCIO	9.543 (0.002)	8.841 (0.002)	32.11 (0.010)	0.637 (0.008)	3.559 (0.034)	23.85 (0.001)
	CSCS	9.532 (0.014)	8.540 (0.015)	29.38 (0.075)	0.966 (0.014)	1.786 (0.013)	22.77 (0.043)
	Prop	9.311 (0.007)	8.390 (0.005)	29.77 (0.026)	0.614 (0.003)	0.665 (0.085)	23.09 (0.055)
6	MCD	34.50 (2.309)	13.79 (0.952)	27.24 (0.814)	1.493 (0.016)	20.98 (0.858)	25.59 (0.234)
	Glasso	8.291 (0.004)	5.340 (0.005)	25.74 (0.028)	1.052 (0.009)	1.717 (0.121)	12.45 (0.027)
	SCIO	10.52 (0.713)	7.234 (0.382)	26.68 (0.478)	1.243 (0.083)	1.494 (0.062)	10.70 (0.001)
	CSCS	7.987 (0.017)	4.921 (0.008)	21.62 (0.044)	1.278 (0.039)	1.275 (0.018)	11.33 (0.111)
	Prop	7.724 (0.024)	5.003 (0.015)	22.92 (0.103)	1.062 (0.008)	0.803 (0.258)	10.31 (0.064)
7	MCD	78.65 (9.265)	25.19 (2.910)	28.24 (2.983)	4.042 (0.291)	11.57 (2.259)	48.25 (0.290)
	Glasso	5.112 (0.054)	1.638 (0.005)	10.88 (0.041)	0.537 (0.010)	3.615 (0.313)	16.97 (0.027)
	SCIO	4.502 (0.008)	1.663 (0.001)	10.74 (0.017)	0.547 (0.008)	3.227 (0.412)	14.86 (0.010)
	CSCS	5.776 (0.099)	1.580 (0.005)	9.346 (0.034)	0.524 (0.007)	2.444 (0.081)	17.16 (0.092)
	Prop	4.618 (0.020)	1.540 (0.003)	9.701 (0.027)	0.468 (0.002)	2.129 (0.054)	16.88 (0.028)

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Table 2: The averages and standard errors (in parenthesis) of estimates for 4 groups of variables.

1	MCD	48.46 (6.760)	16.19 (2.258)	178.1 (23.47)	4.400 (0.556)	3.809 (0.841)	33.46 (0.316)
	Glasso	8.850 (0.001)	8.718 (0.001)	28.62 (0.003)	0.518 (0.001)	0.597 (0.003)	27.68 (0.001)
	SCIO	8.845 (0.002)	8.683 (0.001)	28.43 (0.004)	0.489 (0.002)	0.648 (0.004)	27.35 (0.005)
	CSCS	8.771 (0.003)	8.603 (0.002)	28.05 (0.010)	0.451 (0.003)	0.497 (0.004)	27.29 (0.022)
	Prop	8.785 (0.003)	8.610 (0.003)	28.04 (0.016)	0.387 (0.002)	0.421 (0.003)	26.43 (0.030)
2	MCD	23.83 (19.01)	36.34 (6.053)	38.72 (6.322)	0.397 (0.060)	3.993 (0.879)	28.31 (0.577)
	Glasso	2.612 (0.004)	2.384 (0.002)	12.60 (0.008)	0.243 (0.001)	0.418 (0.003)	8.099 (0.001)
	SCIO	2.635 (0.019)	2.367 (0.002)	12.51 (0.008)	0.242 (0.003)	0.430 (0.004)	7.998 (0.003)
	CSCS	2.605 (0.008)	2.331 (0.005)	12.26 (0.036)	0.227 (0.002)	0.400 (0.004)	7.976 (0.015)
	Prop	2.573 (0.008)	2.271 (0.004)	11.88 (0.026)	0.201 (0.002)	0.354 (0.004)	7.687 (0.013)
3	MCD	16.27 (1.882)	60.19 (7.668)	81.22 (7.474)	2.385 (0.119)	16.14 (2.668)	41.68 (0.223)
	Glasso	7.953 (0.003)	7.766 (0.007)	22.23 (0.034)	1.147 (0.033)	7.825 (0.492)	21.93 (0.030)
	SCIO	7.971 (0.055)	7.668 (0.041)	21.14 (0.042)	1.337 (0.054)	9.113 (0.230)	21.38 (0.010)
	CSCS	7.845 (0.011)	7.451 (0.009)	20.01 (0.032)	0.953 (0.048)	1.564 (0.036)	22.53 (0.115)
	Prop	7.817 (0.006)	7.405 (0.002)	20.78 (0.007)	0.486 (0.002)	0.789 (0.016)	23.73 (0.021)
4	MCD	26.29 (4.073)	8.836 (1.365)	11.09 (1.435)	1.883 (0.142)	18.57 (4.659)	39.62 (0.273)
	Glasso	4.284 (0.051)	2.809 (0.008)	9.456 (0.055)	0.347 (0.015)	2.629 (1.172)	15.57 (0.037)
	SCIO	3.576 (0.001)	2.939 (0.001)	10.22 (0.012)	0.405 (0.026)	2.506 (1.088)	14.13 (0.001)
	CSCS	5.846 (0.119)	2.684 (0.005)	8.746 (0.024)	0.369 (0.010)	1.782 (0.043)	15.88 (0.046)
	Prop	4.491 (0.025)	2.788 (0.003)	9.024 (0.008)	0.358 (0.001)	1.487 (0.033)	15.82 (0.017)
5	MCD	69.74 (5.203)	25.87 (1.396)	56.76 (1.174)	1.714 (0.097)	68.35 (9.419)	36.22 (0.221)
	Glasso	25.84 (0.003)	20.54 (0.002)	53.62 (0.008)	0.824 (0.005)	11.44 (0.379)	35.91 (0.014)
	SCIO	25.96 (0.002)	20.65 (0.001)	54.02 (0.002)	0.797 (0.008)	4.084 (0.035)	37.30 (0.001)
	CSCS	25.62 (0.007)	20.35 (0.003)	52.98 (0.010)	1.057 (0.035)	3.884 (0.011)	34.60 (0.035)
	Prop	25.76 (0.004)	20.42 (0.003)	53.07 (0.013)	0.734 (0.002)	3.003 (0.200)	33.54 (0.049)
6	MCD	38.13 (3.725)	15.74 (1.756)	29.40 (1.459)	1.032 (0.022)	22.71 (1.110)	25.72 (0.266)
	Glasso	8.295 (0.005)	5.364 (0.005)	25.81 (0.031)	1.072 (0.010)	1.355 (0.114)	12.55 (0.031)
	SCIO	9.359 (0.798)	6.426 (0.358)	27.73 (0.641)	1.106 (0.094)	1.379 (0.056)	10.99 (0.019)
	CSCS	7.985 (0.022)	4.827 (0.009)	21.68 (0.047)	1.091 (0.018)	1.308 (0.021)	11.65 (0.128)
	Prop	7.659 (0.017)	5.005 (0.009)	22.86 (0.070)	0.902 (0.005)	0.726 (0.159)	11.08 (0.033)
7	MCD	40.16 (5.131)	13.34 (1.641)	16.00 (1.795)	2.658 (0.200)	37.25 (8.799)	47.28 (0.289)
	Glasso	5.944 (0.099)	1.588 (0.006)	10.53 (0.060)	0.480 (0.015)	4.094 (0.968)	16.34 (0.038)
	SCIO	4.682 (0.001)	1.619 (0.001)	10.34 (0.014)	0.483 (0.009)	3.901 (0.876)	15.13 (0.001)
	CSCS	8.205 (0.198)	1.604 (0.038)	9.022 (0.032)	0.451 (0.006)	1.756 (0.068)	17.52 (0.070)
	Prop	4.992 (0.029)	1.503 (0.001)	9.063 (0.010)	0.374 (0.001)	1.312 (0.029)	17.15 (0.016)

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

SCIO, Glasso and MCD methods for most loss measures. In addition, the compared methods also show their advantages in some loss functions. For example, the SCIO produces the lowest values in terms of L_1 and FSL for both of scenarios 4 and 7. The CSCS gives the best performance in terms of Fnorm for scenarios 5, 6 and 7. It is also superior over other methods regarding L_2 for scenarios 4 and 6.

Besides, one can see that the MCD approach does not perform well compared with other methods, possibly due to that most simulated data do not have a valid full information of natural variable ordering. It can make the MCD, the performance of which heavily depends on the variable ordering, being inferior to other methods. Moreover, for Scenario 7 without any variable ordering, we observe that the Glasso can be better in certain criteria such as L_1 and FSL. Especially, the Glasso is superior in FSL as expected since it is good at inducing sparsity. Comparing with CSCS, the Glasso gives similar performance on L_2, although not as good in terms of Fnorm and QL.

5. Case Studies

5.1 Covid-19 Data

We further evaluate the performance of the proposed model through a real data example of Covid-19 pandemic, which is available on the official website of Virginia health department. The data were weekly collected during May 29th, 2020 to February 6th,

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

2021 (37 weeks) from 34 districts in Virginia State such as Arlington, Fairfax, Richmond, Roanoke, Virginia Beach, etc. For each week the data contain four variables: the accumulative number of cases, the accumulative number of people hospitalized, the accumulative number of deaths and the accumulative number of people taking the PCR (polymerase chain reaction) tests, resulting in $4 \times 37 = 148$ variables.

Denote the collected data by $\boldsymbol{N}_i = (N_{i1}, \ldots, N_{i148})', i = 1, ..., 34$, and transform $y_{ij} = \sqrt{(N_{ij} + 1/4)}$ to make the data distribution close to normal (Brown et al., 2005). We then apply the proposed model as well as the MCD, SCIO, CSCS and Glasso methods to estimate the 148×148 inverse covariance matrix. To conduct data analysis, the 148 variables are partitioned into 37 groups with each group naturally corresponding to a calendar week. Hence it is seen that the data have a partial information of the variable ordination in a weekly scale. To examine the performance of methods in comparison, we predict the accumulative number of cases, the accumulative number of people hospitalized, the accumulative number of deaths and the accumulative number of people taking PCR in the last 2 weeks using the inverse covariance estimates obtained from the data in the first 35 weeks. Specifically, let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{i148})' = (\boldsymbol{y}'_{iE}, \boldsymbol{y}'_{iL})'$, where $\boldsymbol{y}_{iE}$ and $\boldsymbol{y}_{iL}$ represent the data in the first 35 weeks and last 2 weeks for the ith district, $i = 1, 2, \ldots, 34$. Hence $\boldsymbol{y}_{iE}$ contains $4 \times 35 = 140$ variables and $\boldsymbol{y}_{iL}$ contains $4 \times 2 = 8$ variables. Accordingly, the mean

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

vector and the inverse covariance matrix are divided as

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Omega}_{11}, & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}'_{12}, & \boldsymbol{\Omega}_{22} \end{bmatrix}.$$

Assuming multivariate normality, we have

$$E(\boldsymbol{y}_{iL}|\boldsymbol{y}_{iE}) = \boldsymbol{\mu}_2 - \boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}'_{12}(\boldsymbol{y}_{iE} - \boldsymbol{\mu}_1). \tag{5.10}$$

The 34 observations are split into a training set and a testing set using the leaving-one-out mechanism. That is, each observation is considered as a testing point with the rest 33 observations as training set. The training set is used to estimate the mean vector $\boldsymbol{\mu}$ and the inverse covariance matrix $\boldsymbol{\Omega}$. The values of $\boldsymbol{y}_{iE} = (y_{i1}, \ldots, y_{i140})'$ in the testing data are used to predict $\boldsymbol{y}_{iL} = (y_{i141}, \ldots, y_{i148})'$ based on Equation (5.10). For each variable in $\boldsymbol{y}_{iL}$, define the average absolute prediction error (APE) by

$$\text{APE}_j = \frac{1}{34}\sum_{i=1}^{34} |\hat{y}_{ij} - y_{ij}|, \quad j = 141, \ldots, 148.$$

where $\hat{y}_{ij}$ is the predicted value.

Table 3 reports the values of APE_j for $j = 141, \ldots, 148$, corresponding to the prediction measurements for weeks 36 and 37. The columns "Hospitalizations" and

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Table 3: The averages ($\times 10^{-1}$) and standard errors ($\times 10^{-1}$) of APE for Covid-19 data.

	36th week				37th week			
	Cases	Hospitalizations	Deaths	PCR	Cases	Hospitalizations	Deaths	PCR
MCD	2.90 (0.26)	0.79 (0.11)	0.79 (0.11)	1.41 (0.16)	1.65 (0.22)	0.87 (0.12)	0.75 (0.09)	2.20 (0.40)
Glasso	3.57 (0.43)	1.03 (0.13)	1.02 (0.13)	5.50 (0.68)	2.41 (0.32)	1.00 (0.10)	0.93 (0.11)	5.16 (0.75)
SCIO	2.48 (0.32)	0.95 (0.13)	1.04 (0.12)	2.65 (0.34)	2.22 (0.31)	0.99 (0.10)	0.93 (0.11)	3.02 (0.51)
CSCS	2.54 (0.24)	1.06 (0.13)	0.79 (0.12)	1.50 (0.19)	1.64 (0.19)	0.85 (0.10)	0.84 (0.08)	1.68 (0.40)
Prop	2.39 (0.33)	0.80 (0.11)	0.75 (0.12)	1.33 (0.17)	1.61 (0.24)	0.80 (0.09)	0.73 (0.08)	2.55 (0.39)

"PCR" represent the variables that the number of people hospitalized, and the number of people attending the PCR tests, respectively. We observe that the Prop outperforms the SCIO and Glasso, and is slightly better than the CSCS and MCD approaches. Specifically, the Prop provides more accurate prediction in the number of cases, which is of practically importance for pandemic study on the risk assessment. It also predicts well in the number of deaths, and performs comparably with the CSCS and MCD in predicting the number of people who will take the PCR tests and the number of patients in hospital due to Covid-19. Additionally, it is seen that the CSCS and MCD methods perform better than the Glasso and SCIO on this data set. One possible reason is that the data, although not very strictly, have a natural time ordering by weeks among the variables. The CSCS and MCD methods utilize this information and show the advantages over the Glasso and SCIO.

In addition, Figure 1 displays the estimated conditional dependency relationship between variables through network plots, which provide a further insight into the analysis results from the compared methods. It is seen that the CSCS and MCD

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

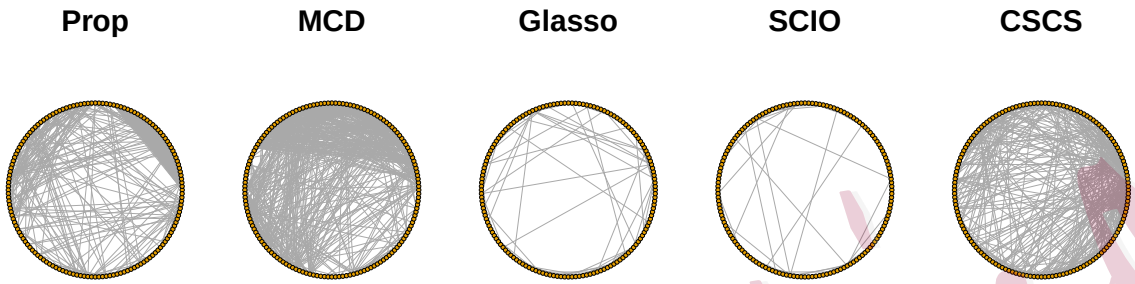

Figure 1: Conditional dependence networks inferred from Covid-19 data for all variables.

Figure 2: Conditional dependence networks inferred from Covid-19 data for variables from 5th week to 20th week.

methods yield an excessive amount of connections, which complicates the estimated model with difficult interpretation. While the graphs constructed from the Glasso and SCIO appear to give few connections, not providing sufficient information to infer the conditional dependency among variables. In contrast, the proposed model identifies a proper number of variable connections with certain sparsity. For a clear presentation, Figure 2 shows the estimated networks for the 60 consecutive variables,

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

corresponding to the 5th week to 20th week. One would expect a meaningful network with nearby variables having some connections while the far-away variables having few connections. From Figure 2, it appears that the proposed model outperforms other methods, since the MCD method presents relatively too many connections of far-away variables, and the Glasso and SCIO provide too little information of dependency relationship between nearby variables. The CSCS displays a proper number of connections as the Prop method, while the Prop method seems to infer more connections of nearby variables, which is indicated by dense connections on the both upper right and bottom right of network plot constructed by the Prop method.

5.2 Call Center Data

In this section, we apply the proposed model to analyze the call center data from Huang et al. (2006). The data set was collected from one of call centers in a major U.S. northeastern financial organization. It recorded the time that every call arrives at the service queue from 7:00am until midnight in each day of 2002, except for 6 days when the data collecting equipment was out of order. The 17-hour period is divided into 102 10-minute intervals, and the number of calls during each time interval is counted. Since the arrival patterns of calls are different between weekdays and weekends, the analysis focuses on weekdays here. In addition, screening out some outliers that include holidays and days when the recording equipment was faulty, we

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

are left with 239 observations.

Denote such data by $\boldsymbol{N}_i = (N_{i1}, \ldots, N_{i102})', i = 1, ..., 239$, where N_{ij} is the number of calls received by the call center for the jth 10-minute interval on day i. Do the transformation $y_{ij} = \sqrt{(N_{ij} + 1/4)}, i = 1, \ldots, 239, j = 1, \ldots, 102$ to make the data distribution close to normal (Brown et al., 2005). We apply the compared methods to estimate the 102×102 inverse covariance matrix. The 102 variables are partitioned into three groups with each group subsequently containing 30, 30 and 42 variables, corresponding to three time intervals of 7:00am - 12:00am, 12:00am - 5:00pm and 5:00pm - midnight. To compare the performance of different methods, we predict the number of arriving calls later in a day based on the arrival patterns at earlier times of that day. As in Section 5.1, we use notation $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{i102})' = (\boldsymbol{y}'_{iE}, \boldsymbol{y}'_{iL})'$, where $\boldsymbol{y}_{iE}$ and $\boldsymbol{y}_{iL}$ measure the arrival patterns in the early and later times of day i. To examine the performance of the proposed model, the 239 observations are split into a training set which contains the first 205 data corresponding to dates from January to October, and a testing set with the rest 34 data corresponding to dates from November and December. The values of $\boldsymbol{y}_{iE}, i = 206, \ldots, 239$ in the testing data are used to predict $\boldsymbol{y}_{iL}$ based on Equation (5.10). Two different settings of $\boldsymbol{y}_{iE}$ are considered as (A): $\boldsymbol{y}_{iE} = (y_{i1}, \ldots, y_{i51})'$ and (B): $\boldsymbol{y}_{iE} = (y_{i1}, \ldots, y_{i60})'$. Setting A is used in Huang et al. (2006) and represents using the data from the early half of a day to predict the call numbers in the later half of the day. Setting B means using

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

the data from the daytime to predict the call numbers in the evenings. For each time interval in $\boldsymbol{y}_{iL}$, the APE is computed as

$$\mathrm{APE}_j = \frac{1}{}$$

Setting	MCD	Glasso	SCIO	CSCS	Prop
A	1.240 (0.052)	1.258 (0.053)	1.194 (0.063)	1.152 (0.045)	0.991 (0.031)
B	1.159 (0.045)	1.163 (0.044)	1.083 (0.030)	1.098 (0.039)	1.023 (0.030)

Table 4 shows the averages and corresponding standard errors of APE over j from each compared method for Settings A and B. The proposed model generally gives superior performance over other methods with lowest values of APE. The SCIO appears to be comparable with the CSCS method, and they perform better than MCD and Glasso. The application results demonstrate the merits of the proposed model when data have partial information of the variable ordination.

6. Discussion

In this work, we propose a block Cholesky decomposition (BCD) method for inverse covariance estimation when the partial information of variable ordination is available.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

The proposed model adopts the MCD to induce the sparsity by imposing the regularization on the multivariate regressions. The proposed method provides a unified framework for several existing methods. including the MCD, the Glasso, and the estimators of Rothman et al. (2010*a*) and Witten et al. (2011). The theoretical results indicate that the proposed model enjoys a faster consistent rate than the Glasso if the data have the partial ordination among the variables.

There are several directions for future research. First, the objective function (2.5) is not jointly convex, hence the theoretical global minimizer is not guaranteed. A potential way to address this issue is to employ the classical Cholesky decomposition for the inverse covariance estimation (Yu and Bien, 2017; Khare et al., 2019), which leads to a convex optimization. However, the statistical interpretation of the classical Cholesky decomposition may not be as explicit as the MCD, where the Cholesky factor matrix can be constructed from regressions. It will be interesting to further investigate how to incorporate the partial information of variable ordering into the classical Cholesky decomposition. Second, the proposed method can be extended to investigate the multivariate response regression, where the multivariate response variables have certain ordering information, such as data with outcomes collected in a time sequence order. Third, the error bounds in Theorem 1 and Theorem 2 could be further improved. Note that $(\boldsymbol{D}_j^{-1})_*$ and $(\boldsymbol{A}_j)_*$ in Theorem 1 are generated on the path of Algorithm 1. When they are close to the underlying true parameters, several

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

terms of $o_p(1)$ in the proof could be close to 0, making a smaller constant term in the error bound.

Acknowledgements

The authors thank the editor, associate editors and referees for their constructive comments. Deng's research was supported by the National Science Foundation CISE Expedition (CCF-1918770). Kang's work was supported by Natural Science Foundation of Liaoning Province (2022-MS-179), Department of Education of Liaoning Province (LJKMZ20221565), and National Natural Science Foundation of China (NSFC-72232001).

References

Bickel, P. J., and Levina, E. (2008), "Regularized estimation of large covariance matrices," *The Annals of Statistics*, 36(1), 199–227.

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L. (2005), "Statistical analysis of a telephone call center: A queueing-science perspective," *Journal of the American statistical association*, 100(469), 36–50.

Cai, T., Liu, W., and Luo, X. (2011), "A constrained ℓ_1 minimization approach to sparse precision matrix estimation," *Journal of the American Statistical Association*, 106(494), 594–607.

Cai, T. T., Li, H., Liu, W., and Xie, J. (2016), "Joint estimation of multiple high-dimensional precision matrices," *Statistica Sinica*, 26(2), 445.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

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

Clemmensen, L., Hastie, T., Witten, D., and Ersbøll, B. (2011), "Sparse discriminant analysis," *Technometrics*, 53(4), 406–413.

Dellaportas, P., and Pourahmadi, M. (2012), "Cholesky-GARCH models with applications to finance," *Statistics and Computing*, 22(4), 849–855.

Deng, X., and Tsui, K.-W. (2013), "Penalized covariance matrix estimation using a matrix-logarithm transformation," *Journal of Computational and Graphical Statistics*, 22(2), 494–512.

Friedman, J., Hastie, T., and Tibshirani, R. (2008), "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, 9(3), 432–441.

Gorski, J., Pfeuffer, F., and Klamroth, K. (2007), "Biconvex sets and optimization with biconvex functions: a survey and extensions," *Mathematical methods of operations research*, 66(3), 373–407.

Huang, J. Z., Liu, N., Pourahmadi, M., and Liu, L. (2006), "Covariance matrix selection and estimation via penalised normal likelihood," *Biometrika*, 93(1), 85–98.

Kang, X., and Deng, X. (2020), "An improved modified cholesky decomposition approach for precision matrix estimation," *Journal of Statistical Computation and Simulation*, 90(3), 443–464.

Kang, X., and Wang, M. (2021), "Ensemble sparse estimation of covariance structure for exploring genetic disease data," *Computational Statistics and Data Analysis*, 159, 107220.

Khare, K., Oh, S.-Y., Rahman, S., and Rajaratnam, B. (2019), "A scalable sparse Cholesky based approach for learning high-dimensional covariance matrices in ordered data," *Machine Learning*, 108(12), 2061–2086.

Lam, C., and Fan, J. (2009), "Sparsistency and rates of convergence in large covariance matrix estimation,"

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Annals of Statistics, 37(6B), 4254–4278.

Liu, H., and Wang, L. (2017), "Tiger: A tuning-insensitive approach for optimally estimating gaussian graphical models," *Electronic Journal of Statistics*, 11(1), 241–294.

Liu, W., and Luo, X. (2015), "Fast and adaptive sparse precision matrix estimation in high dimensions," *Journal of multivariate analysis*, 135, 153–162.

Meinshausen, N., and Bühlmann, P. (2006), "High-dimensional graphs and variable selection with the lasso," *Annals of statistics*, 34(3), 1436–1462.

Nino-Ruiz, E. D., Sandu, A., and Deng, X. (2019), "A parallel implementation of the ensemble Kalman filter based on modified Cholesky decomposition," *Journal of Computational Science*, 36, 100654.

Pourahmadi, M. (1999), "Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation," *Biometrika*, 86(3), 677–690.

Rajaratnam, B., and Salzman, J. (2013), "Best permutation analysis," *Journal of Multivariate Analysis*, 121(10), 193–223.

Rothman, A. J., Levina, E., and Zhu, J. (2010*a*), "A new approach to Cholesky-based covariance regularization in high dimensions," *Biometrika*, 97(3), 539–550.

Rothman, A. J., Levina, E., and Zhu, J. (2010*b*), "Sparse multivariate regression with covariance estimation," *Journal of Computational and Graphical Statistics*, 19(4), 947–962.

Shi, J. (2006), *Stream of variation modeling and analysis for multistage manufacturing processes* CRC press.

Sofer, T., Dicker, L., and Lin, X. (2014), "Variable selection for high dimensional multivariate outcomes," *Statistica Sinica*, 24(4), 1633–1654.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 267–288.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Van Wieringen, W. N. (2019), "The generalized ridge estimator of the inverse covariance matrix," *Journal of Computational and Graphical Statistics*, 28(4), 932–942.

Wagaman, A. S., and Levina, E. (2009), "Discovering sparse covariance structures with the Isomap," *Journal of Computational and Graphical Statistics*, 18(3), 551–572.

Wang, C., Pan, G., Tong, T., and Zhu, L. (2015), "Shrinkage estimation of large dimensional precision matrix using random matrix theory," *Statistica Sinica*, 25, 993–1008.

Wang, L., Chen, Z., Wang, C. D., and Li, R. (2020), "Ultrahigh dimensional precision matrix estimation via refitted cross validation," *Journal of Econometrics*, 215(1), 118–130.

Witten, D. M., Friedman, J. H., and Simon, N. (2011), "New insights and faster computations for the graphical Lasso," *Journal of Computational and Graphical Statistics*, 20(4), 892–900.

Xue, L., Zou, H. et al. (2012), "Regularized rank-based estimation of high-dimensional nonparanormal graphical models," *The Annals of Statistics*, 40(5), 2541–2571.

Yu, G., and Bien, J. (2017), "Learning local dependence in ordered data," *The Journal of Machine Learning Research*, 18(1), 1354–1413.

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

Yuan, M., and Lin, Y. (2007), "Model selection and estimation in the Gaussian graphical model," *Biometrika*, 94(1), 19–35.

Institute of Supply Chain Analytics and Interantional Business College, Dongbei University of Finance and Economics, Dalian, China

E-mail: (xiaoningmike@126.com)

Department of Statistics, Virginia Tech, Blacksburg, VA

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

E-mail: (lianjiayi@vt.edu)

Department of Statistics, Virginia Tech, Blacksburg, VA

E-mail: (xdeng@vt.edu)

Statistica Sinica