Statistica Sinica Preprint No: SS-2023-0131	
Title	Adaptive Block Banding Precision Matrix Estimation For Multivariate Longitudinal Data
Manuscript ID	SS-2023-0131
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0131
Complete List of Authors	Chunhui Liang, Wenqing Ma and Yanyuan Ma
Corresponding Authors	Wenqing Ma
E-mails	wenqingma@cnu.edu.cn
Notice: Accepted author version.	

Statistica Sinica

ADAPTIVE BLOCK BANDING PRECISION MATRIX ESTIMATION FOR MULTIVARIATE LONGITUDINAL DATA

Chunhui Liang[1], Wenqing Ma[2] and Yanyuan Ma[3]

[1] *Tianjin University of Commerce*

[2] *Capital Normal University*

[3] *The Pennsylvania State University*

Abstract: We propose an estimator for precision matrices with the structure of Banded Kronecker Sparse forms (BKS). BKS takes advantage of the special feature of a precision matrix, which has the form of the kronecker product of an adaptively banded matrix and a sparse matrix, both are positive definite. Such precision matrix arises frequently in practice in finance data, medical data and time series data. We achieve the adaptive bandedness via a specially designed penalty, and enforce the sparsity via lasso. We apply a computationally efficient procedure named Alternative Convex Search (ACS) algorithm to implement BKS. We establish the computational convergence and show the statistical guarantee through establishing the asymptotic rate. Our extensive simulation studies indicate the superior finite sample performance of BKS in comparison to existing methods. Additionally, we apply BKS to EEG and ADHD datasets, wherein it outperforms other methods in capturing the banding sparsity characteristics of the precision matrix.

Key words and phrases: Adaptive banding structure; Biconvex function; Convex optimization; Kronecker product; Sparse precision matrix.

1. Introduction

Matrix-valued data are becoming increasingly common in modern data collection procedures. This type of data can be found, for instance, in neuro image data, finance data, and time series data. In this paper, we focus on multivariate longitudinal data, which is a special type of matrix-valued data. This type of data contains multiple outcomes of interest for each subject, with repeated measurements taken over time. It is natural to represent the resulting data in a matrix format with two dimensions corresponding to variables and time. We are aware that the estimation of covariance or precision matrices is a fundamental problem in multivariate data analysis, including techniques such as principle component analysis, discriminant analysis, and regression analysis. Our objective is to estimate the precision matrix of the vectorized version of multivariate longitudinal data. However, considering the characteristics of multivariate longitudinal data, the covariance matrix in the two dimensions of the observed matrix contains different structural information. It not only incorporates the structural information between multiple outcomes at a fixed time point but also incorporates the time-series correlation structure among different time points for each response variable. The precision matrix exhibits the same characteristics. Due to the presence of complex structural information and the typically large dimension of the observed matrices, obtaining an efficient estimator for the precision matrix becomes increasingly challenging as the number of unknown parameters increases quadratically with the vector length.

To address the challenges posed by high-dimensionality when estimating a large di-

mensional precision matrix based on high-dimensional vector observations, the existing literature has proposed two general approaches by incorporating sparse structural assumptions. The first approach involves directly imposing sparsity on the precision matrix through methods, like the graphical lasso (e.g., Yuan and Lin (2007); Banerjee et al. (2008); Friedman et al. (2008)). These methods are suitable for estimating an unstructured precision matrix. In the case of variables with a natural ordering, such as time-series data, the second approach involves directly imposing a banded structure on the precision matrix. This approach has been explored in works like Yu and Bien (2017) and Furrer and Bengtsson (2007). The asymptotic validity of the inversion procedure was later established by Bickel and Levina (2008b) under the assumption of equal bandwidth for all rows and by Cai et al. (2010) under a general bandedness assumption. However, these methods treat the vector data as a whole and assume a sparse structure based on the characteristics of the vector data. Consequently, they are not suitable for modeling multivariate longitudinal data. This is because the aforementioned sparse structures cannot simultaneously capture the correlation structures among response variables and observed time points.

To estimate the precision matrix for multivariate longitudinal data, various methods have been proposed. These methods have shown success when the dimension of the observed matrix is not very large ((Kim and Zimmerman, 2012; Lee et al., 2020)). One approach involves utilizing modified Cholesky block decomposition to reparameterize the covariance. The purpose of this decomposition is to ensure positive definiteness, rather

than reducing the number of parameters. However, when dealing with high-dimensional data, structural assumptions are necessary to reduce the complexity of the model. Taking into consideration the characteristic that the correlation between measurements for any two time points decays with increasing time distance, Qian et al. (2020) and Qian et al. (2021) propose a regularized estimator for the precision matrix. This is achieved through a modified Cholesky block decomposition and by penalizing the log-likelihood with a penalty function that encourages a block banded structure in the lower triangle block matrix. However, the adaptive block banded precision matrix estimator (ABR) introduced by Qian et al. (2021) has a significant computational drawback. The running time of ABR increases dramatically as the number of rows or columns in the observation matrix increases. Additionally, the theoretical properties of ABR are based on the assumption that the repeated measurements are finite.

To address the challenges involved in estimating the precision matrix for high-dimensional multivariate longitudinal data, it is common to introduce a separability assumption on the covariance matrix. This assumption represents the precision matrix as a Kronecker product structure with two smaller matrices. To capture various types of structural information, different approaches have been proposed for these smaller matrices (Tsiligkaridis and Hero (2013),Greenewald and Hero (2015),Leng and Tang (2012), Leng and Pan (2018),Zhang et al. (2023),Dai et al. (2023)). Tsiligkaridis and Hero (2013) and Greenewald and Hero (2015) assume both matrices to be low-rank, Leng and Tang (2012) assume both matrices to be sparse, and further assume normality and propose the Sparse

Matrix Graphical Model (SMGM) estimator by penalizing the log-likelihood. However, these methods only consider sparsity in the precision matrix or correlation matrix, overlooking the potential bandedness property. Zhang et al. (2023) and Dai et al. (2023) assume a banded structure for both of these matrices. These methods are particularly suitable for space-time data where both the row and column variables in the observed matrix have a natural ordering. In this paper, we address the situation where the observed variables of interest may not have a natural ordering. To capture this complexity, we assume that one of the two matrices is sometimes sparse, while the other is banded. For instance, each column of the original matrix data may have a sparse precision matrix, such as in cases where a matrix column represents measurements at different brain locations. Conversely, different columns may correspond to measurements taken at different times, resulting in a banded precision matrix due to the decreasing time relation. Estimating a large-dimensional precision matrix is a challenging task due to the quadratic increase in the number of unknown parameters along the vector length. However, the kronecker product form mentioned earlier effectively reduces the number of parameters and enables contemporary methods to simultaneously consider bandedness and sparsity.

The remainder of this paper is structured as follows. Section 2 introduces the specific model setting, as well as the BKS estimator and the ACS algorithm. The algorithmic convergence of ACS is also established in this section. The theoretical properties of BKS are provided in Section 3. Section 4 presents simulation studies, while Section 5 presents real data analysis. Lastly, Section 6 offers concluding remarks. For the proofs

and technical derivations, please refer to the Supplement Materials.

2. Model and Estimation

In this section, we will provide a description of the model and the motivation behind our statistical model. Additionally, we delve into the computational aspects of the estimation procedure.

2.1 Model setup

Let $\mathbf{Y}_i \equiv (\mathbf{Y}_{i1\cdot}, \ldots, \mathbf{Y}_{iJ\cdot})$ be the $K \times J$ random matrix associated with individual i across all time. We assume $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are independent and identically distributed (iid). Each matrix $\mathbf{Y}_i$ is vectorized to form $\mathrm{vec}(\mathbf{Y}_i) \equiv (\mathbf{Y}_{i1\cdot}^{\mathrm{T}}, \ldots, \mathbf{Y}_{iJ\cdot}^{\mathrm{T}})^{\mathrm{T}} \in \mathcal{R}^{KJ}$, and we assume that the mean and covariance of $\mathrm{vec}(\mathbf{Y}_i)$ are $\mathrm{E}\{\mathrm{vec}(\mathbf{Y}_i)\} = \mathbf{0}$ and $\mathrm{cov}\{\mathrm{vec}(\mathbf{Y}_i)\} = \mathbf{\Sigma}$, respectively. The covariance matrix $\mathbf{\Sigma}$ captures the correlations between any two responses at different times, including the temporal correlation for a fixed response variable $\mathrm{cor}(Y_{ijk}, Y_{ij'k})$, the variable correlation for a fixed time point $\mathrm{cor}(Y_{ijk}, Y_{ijk'})$, and the correlation between any two response variables at different time points $\mathrm{cor}(Y_{ijk}, Y_{ij'k'})$, where $j \neq j' \in \{1, \cdots, J\}$, $k \in \{1, \cdots, K\}$. Let us divide $\mathbf{\Sigma}$ into J^2 size $K \times K$ block matrices. We denote the (j, l) block as $\mathbf{\Sigma}^{(jl)}$, where j and l range from 1 to J. The diagonal block matrix $\mathbf{\Sigma}^{(jj)} = \mathrm{cov}(\boldsymbol{Y}_{ij\cdot})$ represents the variance-covariance structure between the K responses at the jth time point. Similarly, $\mathbf{\Sigma}^{(jl)} = \mathrm{cov}(\boldsymbol{Y}_{ij\cdot}, \boldsymbol{Y}_{il\cdot})$ denotes the covariance between the K responses at the jth and the lth time points. We assume that

the correlation structure information can be separated into two dimensions: variable and time. Specifically, the assumed temporal correlation is the same for all responses, and the assumed variable correlation is the same for all time points. Thus, the covariance matrix for multivariate longitudinal data can be represented as a Kronecker product, $\mathbf{\Sigma} = \mathbf{R} \otimes \mathbf{W}$, where $\mathbf{R} \in \mathcal{R}^{J \times J}$ and $\mathbf{W} \in \mathcal{R}^{K \times K}$. Under this assumption, we can write $\mathbf{\Sigma}^{(jl)}$ as $r_{jl}\mathbf{W}$ for all j, l range from 1 to J. Here, r_{jl} represents a constant that denotes the signal amplication at different time points. In this case, the Kronecker structure $\mathbf{R} \otimes \mathbf{W}$ is non-identifiability. For convenience, we set $\mathbf{W} = \mathbf{\Sigma}^{(11)}/\mathbf{\Sigma}_{1,1}$, where $\mathbf{\Sigma}_{1,1}$ represents the entry at position $(1, 1)$ in the matrix $\mathbf{\Sigma}$. Our interest is in estimating the precision matrix $\mathbf{R}^{-1} \otimes \mathbf{W}^{-1}$. However, in high-dimensional situations, the number of variables and time points both may exceed the sample size. To obtain a stable and efficient estimator for the precision matrix, it is necessary to introduce additional structure assumptions for $\mathbf{R}^{-1}$ and $\mathbf{W}^{-1}$. To obtain a positive definite estimator for $\mathbf{R}^{-1}$, we perform a cholesky decomposition such that $\mathbf{R}^{-1} \equiv \mathbf{L}^{\mathrm{T}}\mathbf{L}$, where $\mathbf{L}$ is a lower triangle matrix and $L_{j,j} > 0$, $j = 1, \cdots, J$. The elements in $\mathbf{R}^{-1}$ represent the conditional correlation between any two time points, given the other observed time points, for the response variable. In practice, as the distance between two time points increases, the conditional correlation will decrease. Therefore, we assume that $\mathbf{R}^{-1}$ is a banded matrix with a bandwidth of d, where d is significantly smaller than J. The elements in $\mathbf{W}^{-1}$ represent the conditional correlation between any two response variables, given the other variables for a specific time point. For $\mathbf{W}^{-1}$, we assume that it is a sparse matrix. Furthermore, we provide

a detailed introduction to the Kronecker structure through two examples in Supplement Material S5.

Remark 1. The assumption that the true covariance or precision matrix is separable plays a crucial role in our model framework and should be evaluated during the data preprocessing stage. In this study, we follow Zhang et al. (2023) and employ the projection-based bootstrap test method introduced by Aston et al. (2017). This method is theoretically guaranteed and computationally fast in high-dimensional settings. Moreover, as a distribution-free approach, it is suitable for our framework.

2.2 Estimation

To estimate $\mathbf{R}^{-1} \otimes \mathbf{W}^{-1}$, we only need to estimate $\mathbf{R}^{-1}$ and $\mathbf{W}^{-1}$ based on the iid observations $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$. We consider minimizing the target function

$$l(\mathbf{W}^{-1}, \mathbf{R}^{-1}) = -J \log |\mathbf{W}^{-1}| - K \log |\mathbf{R}^{-1}| + \frac{1}{n} \sum_{i=1}^{n} \mathrm{tr}(\mathbf{Y}_i^{\mathrm{T}} \mathbf{W}^{-1} \mathbf{Y}_i \mathbf{R}^{-1}).$$

Obviously, up to a constant, $l(\mathbf{W}^{-1}, \mathbf{R}^{-1})$ is the negative loglikelihood of $(\mathbf{W}^{-1}, \mathbf{R}^{-1})$ under the assumption that $\mathrm{vec}(\mathbf{Y}_i)$'s are normally distributed with mean zero. However, we do not assume normality here, so we view $l(\mathbf{W}^{-1}, \mathbf{R}^{-1})$ as a general loss function.

We incorporate a lasso penalty to take into account of the sparsity of $\mathbf{W}^{-1}$. Specifically, we add the penalty term $\lambda_1 \|\mathrm{vec}(\mathbf{W}^{-1})\|_1$ to the loss function. To account for the positive definiteness and banded structure of $\mathbf{R}^{-1}$, we adopt the same methodology as Yu and Bien (2017). Begin with, we utilize the Cholesky decomposition of $\mathbf{R}^{-1}$ to express

it as $\mathbf{R}^{-1} = \mathbf{L}^{\mathrm{T}}\mathbf{L}$, where $\mathbf{L}$ is a lower triangle matrix with $L_{j,j} > 0$ and $L_{j,l} = 0$ for all $j - l > d$. To incorporate the banded structure for $\mathbf{L}$, we consider each row of $\mathbf{L}$ separately. In the lower triangle form of $\mathbf{L}$, the jth row exclusively includes potentially non-zero elements $L_{j,1}, \ldots, L_{j,j-1}$ and a positively definite element $L_{j,j}$. Roughly speaking, the banded structure implies that a smaller column index, denoted by l, suggests a higher probability for the entry $L_{j,l}$ to be zero. Thus to encourage more likely zeros corresponding to smaller column index l, we consider the penalty

$$p(\mathbf{L}_{j,\cdot}) \equiv \sum_{l=1}^{j-1} \sqrt{\sum_{q=1}^{l} L_{j,q}^2}. \tag{2.1}$$

We can see that this is actually a group lasso penalty, where the lth group is the subvector formed by the first l elements of the jth row. Because of the relation $(\sum_{q=1}^{l} L_{j,q}^2)^{1/2} \leq (\sum_{q=1}^{l+1} L_{j,q}^2)^{1/2}$, a zero value corresponding to an index l automatically implies a zero value for all indices $< l$. In other words, the sparsity penalty in (2.1) automatically leads to a banded structure on the jth row. Taking into account all rows, we thus incorporate a penalty $\lambda_2 \sum_{j=2}^{J} p(\mathbf{L}_{j,\cdot})$ to factor in the banded structure on $\mathbf{L}$, or equivalently, the banded structure on $\mathbf{R}^{-1}$.

Combining the above analysis, we propose to estimate $\mathbf{W}^{-1}$ and $\mathbf{L}$ through minimizing

$$\begin{aligned} Q(\mathbf{W}^{-1}, \mathbf{L}, \boldsymbol{\lambda}) = &-J \log |\mathbf{W}^{-1}| - 2K \log |\mathbf{L}| + \frac{1}{n} \sum_{i=1}^{n} \mathrm{tr}(\mathbf{Y}_i^{\mathrm{T}} \mathbf{W}^{-1} \mathbf{Y}_i \mathbf{L}^{\mathrm{T}} \mathbf{L}) \\ &+ \lambda_1 J \|\mathrm{vec}(\mathbf{W}^{-1})\|_1 + \lambda_2 K \sum_{j=2}^{J} p(\mathbf{L}_{j,\cdot}) \end{aligned} \tag{2.2}$$

subject to the positive-definite constraint on $\mathbf{W}^{-1}$, where $p(\mathbf{L}_{j,\cdot})$ is defined in (2.1), and $\boldsymbol{\lambda} \equiv (\lambda_1, \lambda_2)^{\mathrm{T}}$ contains the turning parameters.

At any $\boldsymbol{\lambda}$, $Q(\mathbf{W}^{-1}, \mathbf{L}, \boldsymbol{\lambda})$ is biconvex in $(\mathbf{W}^{-1}, \mathbf{L})$, i.e., it is a convex function of $\mathbf{W}^{-1}$ when $\mathbf{L}$ is fixed, and is a convex function of $\mathbf{L}$ when $\mathbf{W}^{-1}$ is fixed. We thus solve the optimization problem in (2.2) by alternate convex search (ACS), i.e., we alternately minimize $Q(\mathbf{W}^{-1}, \mathbf{L}, \boldsymbol{\lambda})$ with respect to $\mathbf{W}^{-1}$ or $\mathbf{L}$ while keeping the other matrix fixed. Specifically, at the sth step, we first fix $\mathbf{L}$ at $\widehat{\mathbf{L}}^{(s)}$, and update the estimator of $\mathbf{W}^{-1}$ by solving

$$(\widehat{\mathbf{W}}^{-1})^{(s+1)} = \underset{\mathbf{W}^{-1}>0}{\operatorname{argmin}}\{-\log|\mathbf{W}^{-1}| + \frac{1}{nJ}\sum_{i=1}^{n}\operatorname{tr}(\mathbf{Y}_i^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{Y}_i(\widehat{\mathbf{L}}^{(s)})^{\mathrm{T}}\widehat{\mathbf{L}}^{(s)}) + \lambda_1\|\mathrm{vec}(\mathbf{W}^{-1})\|_1\}, \tag{2.3}$$

where $\mathbf{W}^{-1} > 0$ means $\mathbf{W}^{-1}$ is positive definite. We then fix $\mathbf{W}^{-1}$ at $(\widehat{\mathbf{W}}^{-1})^{(s+1)}$ and update the estimator of $\mathbf{L}$ by minimizing

$$-2\log|\mathbf{L}| + \frac{1}{nK}\sum_{i=1}^{n}\operatorname{tr}(\mathbf{Y}_i^{\mathrm{T}}(\widehat{\mathbf{W}}^{-1})^{(s+1)}\mathbf{Y}_i\mathbf{L}^{\mathrm{T}}\mathbf{L}) + \lambda_2\sum_{j=1}^{J}p(\mathbf{L}_{j,\cdot}), \tag{2.4}$$

subject to $L_{j,j} > 0$, and $L_{j,j'} = 0, 1 \le j < j' \le J$. We repeat the above optimization steps (2.3)-(2.4) until $\|(\widehat{\mathbf{W}}^{-1})^{(s+1)} - (\widehat{\mathbf{W}}^{-1})^{(s)}\|_F + \|\widehat{\mathbf{L}}^{(s+1)} - \widehat{\mathbf{L}}^{(s)}\|_F < \varepsilon$, where ε is a pre-determined sufficiently small constant. We then set $\widehat{\mathbf{W}}^{-1} = (\widehat{\mathbf{W}}^{-1})^{(s+1)}$ and $\widehat{\mathbf{L}} = \widehat{\mathbf{L}}^{(s+1)}$ as the final estimators.

Next, we discuss the details in sovling (2.3) and (2.4) respectively. (2.3) is a well studied problem (Yuan and Lin, 2006; Rothman et al., 2008) and here, we adopt the graphical

lasso (glasso) algorithm (Friedman et al., 2008), which guarantees the positive definiteness of the matrix $(\widehat{\mathbf{W}}^{-1})^{(s+1)}$. To investigate the minimization problem of (2.4), we write $\mathbf{Y}_i^* \equiv ((\widehat{\mathbf{W}}^{-1})^{(s+1)})^{1/2}\mathbf{Y}_i$, $\mathbf{Y}^* \equiv \{(\mathbf{Y}_1^*)^{\mathrm{T}}, \ldots, (\mathbf{Y}_n^*)^{\mathrm{T}}\}^{\mathrm{T}}$, and $\mathbf{L}_j \equiv \mathbf{L}_{j,1:j}^{\mathrm{T}}$ as a j-dimensional column vector formed by the first j elements on the jth row of the matrix $\mathbf{L}$ for a generic matrix $\mathbf{L}$. We then obtain $\sum_{i=1}^n \mathrm{tr}(\mathbf{Y}_i^{\mathrm{T}}(\widehat{\mathbf{W}}^{-1})^{(s+1)}\mathbf{Y}_i\mathbf{L}^{\mathrm{T}}\mathbf{L}) = \sum_{i=1}^n \mathrm{tr}((\mathbf{Y}_i^*)^{\mathrm{T}}(\mathbf{Y}_i^*)\mathbf{L}^{\mathrm{T}}\mathbf{L}) = \sum_{j=1}^J \|\mathbf{Y}_{\cdot,1:j}^*\mathbf{L}_j\|_2^2$. Thus (2.4) can be equivalently written as

$$-2\sum_{j=1}^J \log L_{j,j} + \frac{1}{nK}\sum_{j=1}^J \|\mathbf{Y}_{\cdot,1:j}^*\mathbf{L}_j\|_2^2 + \lambda_2 \sum_{j=1}^J p(\mathbf{L}_{j,\cdot}).$$

We can now decompose the optimization with respect to $\mathbf{L}$ into J separately optimization problems with respect to $\mathbf{L}_{1,\cdot}, \ldots \mathbf{L}_{J,\cdot}$ respectively. Specifically,

$$\begin{aligned}
\widehat{L}_{1,1}^{(s+1)} &= \underset{L_{1,1}>0}{\mathrm{argmin}}\{-2K\log L_{1,1} + \frac{1}{n}\|\mathbf{Y}_{\cdot,1}^* L_{1,1}\|_2^2\} = \{(\mathbf{Y}_{\cdot,1}^*)^{\mathrm{T}}\mathbf{Y}_{\cdot,1}^*/nK\}^{-1/2}, \\
\widehat{\mathbf{L}}_j^{(s+1)} &= \underset{L_{j,j}>0, \mathbf{L}_j \in \mathcal{R}^j}{\mathrm{argmin}}\{-2\log L_{j,j} + \frac{1}{nK}\|\mathbf{Y}_{\cdot,1:j}^*\mathbf{L}_j\|_2^2 + \lambda_2 p(\mathbf{L}_j)\}, \quad j = 2, \ldots, J.
\end{aligned} \tag{2.5}$$

Note that $p(\mathbf{L}_{j,\cdot}) = p(\mathbf{L}_j)$. To solve each of the $J-1$ optimization problems above, we adopt the alternating direction method of multipliers (ADMM) algorithm (Boyd et al. (2011)). To this end, to obtain $\widehat{\mathbf{L}}_j^{(s+1)}$, we introduce the constrains $\mathbf{L}_j = \mathbf{\Psi}_j$, and modify the objective function for $(\mathbf{L}_j, \mathbf{\Psi}_j)$ into

$$\begin{aligned}
Q^*(\mathbf{L}_j, \lambda_2, \mathbf{U}_j, \rho, \mathbf{\Psi}_j) = &-2\log L_{j,j} + \frac{1}{nK}\|\mathbf{Y}_{\cdot,1:j}^*\mathbf{L}_j\|_2^2 \\
&+ \lambda_2 p(\mathbf{\Psi}_j) + \mathbf{U}_j^{\mathrm{T}}(\mathbf{L}_j - \mathbf{\Psi}_j) + \frac{\rho}{2}\|\mathbf{L}_j - \mathbf{\Psi}_j\|_2^2.
\end{aligned} \tag{2.6}$$

Given $\widehat{\mathbf{\Psi}}_j^{(r)}$ and $\widehat{\mathbf{U}}_j^{(r)}$, we compute the derivative of (2.6) with respect to $\mathbf{L}_j$ and obtain the estimation equation:

$$-2\frac{1}{L_{j,j}}\mathbf{e}_j + \frac{2}{nK}(\mathbf{Y}^*_{\cdot,1:j})^{\mathrm{T}}\mathbf{Y}^*_{\cdot,1:j}\mathbf{L}_j + \widehat{\mathbf{U}}_j^{(r)} + \rho(\mathbf{L}_j - \widehat{\mathbf{\Psi}}_j^{(r)}) = \mathbf{0},$$

where $\mathbf{e}_j$ denotes the j-dimensional indicator vector, with 1 on the jth element and the other elements 0. The above estimation equation can be written as

$$\begin{cases} \mathbf{L}_{j,1:j-1}\left\{\frac{2}{nK}(\mathbf{Y}^*_{\cdot,1:j-1})^{\mathrm{T}}\mathbf{Y}^*_{\cdot,1:j-1} + \rho\mathbf{I}\right\} + \frac{2}{nK}(\mathbf{Y}^*_{\cdot,j})^{\mathrm{T}}\mathbf{Y}^*_{\cdot,1:j-1}L_{j,j} + \widehat{\mathbf{U}}^{(r)}_{j,1:j-1} = \rho\widehat{\mathbf{\Psi}}^{(r)}_{j,1:j-1}, \\ -\frac{2}{L_{j,j}} + \left\{\frac{2}{nK}(\mathbf{Y}^*_{\cdot,j})^{\mathrm{T}}\mathbf{Y}^*_{\cdot,j} + \rho\right\}L_{j,j} + \frac{2}{nK}(\mathbf{Y}^*_{\cdot,j})^{\mathrm{T}}\mathbf{Y}^*_{\cdot,1:j-1}\mathbf{L}^{\mathrm{T}}_{j,1:j-1} + \widehat{U}^{(r)}_{j,j} = \rho\widehat{\Psi}^{(r)}_{j,j}. \end{cases}$$

The first equation leads to that $\mathbf{L}_{j,1:j-1} = -\{2(nK)^{-1}(\mathbf{Y}^*_{\cdot,j})^{\mathrm{T}}\mathbf{Y}^*_{\cdot,1:j-1}L_{j,j} + \widehat{\mathbf{U}}^{(r)}_{j,1:j-1} - \rho\widehat{\mathbf{\Psi}}^{(r)}_{j,1:j-1}\}\{2(nK)^{-1}(\mathbf{Y}^*_{\cdot,1:j-1})^{\mathrm{T}}\mathbf{Y}^*_{\cdot,1:j-1} + \rho\mathbf{I}\}^{-1}$. Inserting this into the second equation yields an equation of the form $AL^2_{j,j} + BL_{j,j} + 2 = 0$, where

$$\begin{aligned} A &= \left\{\frac{4}{nK}(\mathbf{Y}^*_{\cdot,j})^{\mathrm{T}}\mathbf{Y}^*_{\cdot,1:j-1}\right\}\left\{\frac{2}{nK}(\mathbf{Y}^*_{\cdot,1:j-1})^{\mathrm{T}}\mathbf{Y}^*_{\cdot,1:j-1} + \rho\mathbf{I}\right\}^{-1} \\ &\quad \left\{\frac{1}{nK}(\mathbf{Y}^*_{\cdot,1:j-1})^{\mathrm{T}}\mathbf{Y}^*_{\cdot,j}\right\} - \frac{2}{nK}(\mathbf{Y}^*_{\cdot,j})^{\mathrm{T}}\mathbf{Y}^*_{\cdot,j} - \rho, \\ B &= \left\{\frac{2}{nK}(\mathbf{Y}^*_{\cdot,j})^{\mathrm{T}}\mathbf{Y}^*_{\cdot,1:j-1}\right\}\left\{\frac{2}{nK}(\mathbf{Y}^*_{\cdot,1:j-1})^{\mathrm{T}}\mathbf{Y}^*_{\cdot,1:j-1} + \rho\mathbf{I}\right\}^{-1} \\ &\quad (\widehat{\mathbf{U}}^{(r)}_{j,1:j-1} - \rho\widehat{\mathbf{\Psi}}^{(r)}_{j,1:j-1}) - \widehat{\mathbf{U}}^{(r)}_{j,j} + \rho\widehat{\Psi}^{(r)}_{j,j}. \end{aligned}$$

Note that $-1/A$ is the lower-right entry of the matrix $\{(\mathbf{Y}^*_{\cdot,1:j})^{\mathrm{T}}\mathbf{Y}^*_{\cdot,1:j}/nK + \rho\mathbf{I}/2\}^{-1}$,

hence $A < 0$. Further $(B^2 - 8A)^{1/2} > |B|$. Thus, to satisfy the positive requirement of $L_{j,j}$, we get $\widehat{\mathbf{L}}_{j,j}^{(r+1)} = -\{B + (B^2 - 8A)^{1/2}\}/2A$. This provides a closed-form solution for $\widehat{L}_{j,j}^{(r+1)}$, and subsequently a closed-form solution for $\widehat{\mathbf{L}}_{j,1:j-1}^{(r+1)}$. We next update $\mathbf{\Psi}_j$ based on $\widehat{\mathbf{L}}_j^{(r+1)}$ and $\widehat{\mathbf{U}}_j^{(r)}$ by minimizing the objective function

$$\frac{\rho}{2}\|\mathbf{\Psi}_j - \widehat{\mathbf{L}}_j^{(r+1)} - \frac{1}{\rho}\widehat{\mathbf{U}}_j^{(r)}\|_2^2 + \lambda_2 p(\mathbf{\Psi}_j), \tag{2.7}$$

which has the group lasso penalty. To minimize (2.7), we consider its dual problem.

Theorem 1. *A dual of* (2.7) *is given by*

$$\min_{\mathbf{A}} \|\mathbf{Z}_j - \frac{\lambda_2}{\rho}\sum_{l=1}^{j-1}\mathbf{A}_{\cdot,l}\|_2^2, \quad s.t. \quad \|\mathbf{A}_{1:l,l}\|_2 \le 1, \mathbf{A}_{l+1:j,l} = \mathbf{0}, \ for\ l = 1, \ldots, j-1, \tag{2.8}$$

where $\mathbf{Z}_j = \widehat{\mathbf{L}}_j^{(r+1)} + \widehat{\mathbf{U}}_j^{(r)}/\rho$, $\mathbf{A}$ *is a* $j \times (j-1)$ *matrix. Given* $\widehat{\mathbf{A}}$, *the optimizer of* (2.7) *is*

$$\widehat{\mathbf{\Psi}}_j^{(r+1)} = \mathbf{Z}_j - \frac{\lambda_2}{\rho}\sum_{l=1}^{j-1}\widehat{\mathbf{A}}_{\cdot,l}.$$

The proof of Theorem 1 is given in the Supplement Materials. Following Yu and Bien (2017), we use the blockwise coordinate descent (BCD) method to solve (2.8), where we sequentially perform elliptical projection to update each column of $\mathbf{A} \in \mathcal{R}^{j\times(j-1)}$. This strategy is developed in Bien et al. (2016) and Jenatton et al. (2011). It takes advantage of the upper triangle struture of $\mathbf{A}$, and only requires a single pass of BCD. Specifically, we first set $\widehat{\mathbf{A}} = \mathbf{0}$, then for $l = 1, \cdots, j-1$, we sequentially update the lth column of

$\widehat{\mathbf{A}}$. Following (2.8), the lth column of $\mathbf{A}$ is obtained by solving

$$\min_{\mathbf{A}_{\cdot,l}} \|\mathbf{\Gamma}_l - \frac{\lambda_2}{\rho}\mathbf{A}_{\cdot,l}\|_2^2, \quad s.t. \quad \|\mathbf{A}_{1:l,l}\|_2 \leq 1 \text{ and } \mathbf{A}_{l+1:j,l} = \mathbf{0},$$

where $\mathbf{\Gamma}_l \equiv \mathbf{Z}_j - \rho^{-1}\lambda_2 \sum_{q=1}^{l-1} \widehat{\mathbf{A}}_{\cdot,q}$ is a j-dimensional vector. Obviously, if $\|(\mathbf{\Gamma}_l)_{1:l}\|_2 \leq \lambda_2/\rho$, then $\widehat{\mathbf{A}}_{1:l,l} = \rho(\mathbf{\Gamma}_l)_{1:l}/\lambda_2$. Otherwise, $\widehat{\mathbf{A}}_{1:l,l} = (\mathbf{\Gamma}_l)_{1:l}/\|(\mathbf{\Gamma}_l)_{1:l}\|_2$. Combining the two situations, we can write that $\widehat{\mathbf{A}}_{1:l,l} = (\mathbf{\Gamma}_l)_{1:l}/\max\{\lambda_2/\rho, \|(\mathbf{\Gamma}_l)_{1:l}\|_2\}$. $\widehat{\mathbf{\Psi}}_j^{(r+1)} = \mathbf{Z}_j - \rho^{-1}\lambda_2 \sum_{l=1}^{j-1} \widehat{\mathbf{A}}_{\cdot,l}$. Finally, based on $\widehat{\mathbf{L}}_j^{(r+1)}$ and $\widehat{\mathbf{\Psi}}_j^{(r+1)}$, we follow the ADMM procedure to update Lagrange multiplier $\widehat{\mathbf{U}}_j^{(r+1)}$ via $\widehat{\mathbf{U}}_j^{(r+1)} = \widehat{\mathbf{U}}_j^{(r)} + \rho(\widehat{\mathbf{L}}_j^{(r+1)} - \widehat{\mathbf{\Psi}}_j^{(r+1)})$. The detailed process of solving the objective function (2.6) are provided in Algorithm 1. Algorithm 1 is applied to all $j = 1, \ldots, J$ to yield $\widehat{\mathbf{L}}^{(s+1)}$ defined in (2.4). We iteratively update the estimation for $\mathbf{W}^{-1}$ and $\mathbf{L}$ as described in Algorithm 2 to obtain $\widehat{\mathbf{W}}^{-1}$ and $\widehat{\mathbf{L}}$, and form $\widehat{\mathbf{\Sigma}}^{-1} = \widehat{\mathbf{L}}^{\mathrm{T}}\widehat{\mathbf{L}} \otimes \widehat{\mathbf{W}}^{-1}$ as out final estimator for the precision matrix. Although there is no guarantee that the algorithm converges to the global minimum, the algorithm converges to a local stationary point of (2.2).

Remark 2. The optimization process (Algorithm 2) requires an initial precision matrix for the time dimension, $\mathbf{R}^{-1}$. As the sample covariance matrix for the time dimension depends on $\mathbf{W}^{-1}$, we initialize $\mathbf{R}^{-1}$ as the identity matrix. For the variable dimension, $\mathbf{W}^{-1}$ is estimated using the glasso function, which defaults to the inverse of the sample covariance matrix. In Algorithm 1, there requires initial values of $\mathbf{L}$, $\mathbf{\Phi}$ and $\mathbf{U}$. Given that $\mathbf{R}^{-1}$ is initialized as the identity matrix, we set the initial values as $\mathbf{L}^{(0)} = \mathbf{I}$, $\mathbf{\Phi}^{(0)} = \mathbf{I}$,

Algorithm 1 ADMM algorithm to solve (2.6)

Input: Initial values $\widehat{\mathbf{L}}_j^{(0)}$, $\widehat{\mathbf{\Psi}}_j^{(0)}$, $\widehat{\mathbf{U}}_j^{(0)}$, $\lambda_2, \rho > 0$, $r = 0$.
Main procedure:

Step 1. Update $\widehat{L}_{j,j}^{(r+1)} = \frac{-B-\sqrt{B^2-8A}}{2A}$ and

$$\widehat{\mathbf{L}}_{j,1:j-1}^{(r+1)} = \left\{\rho\widehat{\mathbf{\Psi}}_{j,1:j-1}^{(r)} - \frac{2}{nK}(\mathbf{Y}_{\cdot,j}^*)^{\mathrm{T}}\mathbf{Y}_{\cdot,1:j-1}^*\widehat{L}_{j,j}^{(r+1)} - \widehat{\mathbf{U}}_{j,1:j-1}^{(r)}\right\}\left\{\frac{2}{nK}(\mathbf{Y}_{\cdot,1:j-1}^*)^{\mathrm{T}}\mathbf{Y}_{\cdot,1:j-1}^* + \rho\mathbf{I}\right\}^{-1}$$

Step 2. Let $\mathbf{Z}_j = \widehat{\mathbf{L}}_j^{(r+1)} + \widehat{\mathbf{U}}_j^{(r)}/\rho$. For $l = 1, \ldots, j-1$, let $\mathbf{\Gamma}_l = \mathbf{Z}_j - (\lambda_2/\rho)\sum_{q=1}^{l-1}\widehat{\mathbf{A}}_{\cdot,q}$, and $\widehat{\mathbf{A}}_{1:l,l} = (\mathbf{\Gamma}_l)_{1:l}/\max\{\frac{\lambda_2}{\rho}, \|(\mathbf{\Gamma}_l)_{1:l}\|_2\}$, $\widehat{\mathbf{A}}_{l+1:j,l} = \mathbf{0}$. Set $\widehat{\mathbf{\Psi}}_j^{(r+1)} = \mathbf{Z}_j - \frac{\lambda_2}{\rho}\sum_{l=1}^{j-1}\widehat{\mathbf{A}}_{\cdot,l}$.

Step 3. Set $\widehat{\mathbf{U}}_j^{(r+1)} = \widehat{\mathbf{U}}_j^{(r)} + \rho(\widehat{\mathbf{L}}_j^{(r+1)} - \widehat{\mathbf{\Psi}}_j^{(r+1)})$;

Step 4. Increase r by 1 and go back to Step 1 until $\|\widehat{\mathbf{L}}_j^{(r+1)} - \widehat{\mathbf{\Psi}}_j^{(r+1)}\|_2 < \varepsilon_{prime}$, and $\|\rho(\widehat{\mathbf{\Psi}}_j^{(r+1)} - \widehat{\mathbf{\Psi}}_j^{(r)})\|_2 < \varepsilon_{dual}$, where $\varepsilon_{prime} = \sqrt{j}\varepsilon_{abs} + \varepsilon_{rel}\max\{\|\widehat{\mathbf{L}}_j^{(r+1)}\|_2, \|\widehat{\mathbf{\Psi}}_j^{(r+1)}\|_2\}$, $\varepsilon_{dual} = \sqrt{j}\varepsilon_{abs} + \varepsilon_{rel}\|\widehat{\mathbf{U}}^{(r+1)}\|_2$, and $\varepsilon_{abs}, \varepsilon_{rel}$ are predetermined constants.

Output: $\widehat{\mathbf{L}}_j^{(r+1)}, \widehat{\mathbf{\Psi}}_j^{(r+1)}$.

and $\mathbf{U}^{(0)} = \mathbf{0}$.

We divide the entire optimization process into two stages. First, given $\widehat{\mathbf{L}}^{(s)}$, we apply the graphical lasso algorithm to obtain $(\widehat{\mathbf{W}}^{-1})^{(s+1)}$. Since the objective function is convex and the initial $\widehat{\mathbf{L}}^{(s)}$ ensures that $(\widehat{\mathbf{R}}^{-1})^{(s)}$ remains positive definite, thus, the optimization process always converges. In the second stage, based on $(\widehat{\mathbf{W}}^{-1})^{(s+1)}$, we impose the condition $\mathbf{\Psi} = \mathbf{L}$ and employ the ADMM algorithm to compute $\widehat{\mathbf{L}}^{(s+1)}$. As the objective function is convex with respect to $\mathbf{L}$ and $\mathbf{\Psi}$, and given an appropriate tuning parameter λ_2, this computation process also converges. Since both subroutines are convergent, we leverage the convergence properties of bi-convex functions, as discussed in Gorski et al. (2007). By iteratively alternating between these two stages, the optimization process converges to a locally optimal solution, $\{(\widehat{\mathbf{W}}^{(-1)})^{(s+1)}, \widehat{\mathbf{L}}^{(s+1)}\}$.

Algorithm 2 The complete algorithm to solve (2.2)

Input: Initial values $\widehat{\mathbf{L}}^{(0)}$, $(\widehat{\mathbf{W}}^{-1})^{(0)}$, $\lambda_1, \lambda_2, \rho > 0$, $s = 0$.

Main procedure:

Step 1. At the given $\widehat{\mathbf{L}}^{(s)}$, obtain $(\widehat{\mathbf{W}}^{-1})^{(s+1)}$ by applying the glasso method to solve (2.3).

Step 2. At the given $(\widehat{\mathbf{W}}^{-1})^{(s+1)}$, obtain $\widehat{\mathbf{L}}^{(s+1)}$ by solving (2.4). (2.4) is solved row-wise, where for $j = 1, \ldots, J$, the nonzero part of the jth row, i.e., $\widehat{\mathbf{L}}_j^{(s+1)}$, is obtained through solving (2.5) via Algorithm 1.

Step 3. Increase s by 1 and go back to Step 1 until $\|(\widehat{\mathbf{W}}^{-1})^{(s+1)} - (\widehat{\mathbf{W}}^{-1})^{(s)}\|_F + \|\widehat{\mathbf{L}}^{(s+1)} - \widehat{\mathbf{L}}^{(s)}\|_F < \varepsilon$, where ε is a pre-determined constant.

Output: $\widehat{\mathbf{W}}^{-1} \equiv (\widehat{\mathbf{W}}^{-1})^{(s+1)}$ and $\widehat{\mathbf{L}} \equiv \widehat{\mathbf{L}}^{(s+1)}$.

3. Statistical Properties

We now study the statistical properties of BKS, through establishing the converge rates of $\widehat{\mathbf{L}}$ and $\widehat{\mathbf{W}}^{-1}$ respectively. All technical proofs are provided in the Supplement Materials. Let d_j be the true bandwidth of the jth row in $\mathbf{L}$, we will show the consistence of the estimators $\widehat{d}_j$, $j = 2, \ldots, J$. We now explain some notations that will be used throughout the paper. For a $n \times p$ real matrix $\mathbf{M} = (M_{ij})$, the l_1 norm is defined as $|\mathbf{M}|_1 = \sum_{i,j} |M_{ij}|$, and the Frobenius norm is $\|\mathbf{M}\|_F = (\sum_{i,j} M_{i,j}^2)^{1/2}$. We make the following assumptions.

(C1) The true lower triangular matrix $\mathbf{L} \in \mathcal{R}^{J \times J}$ has bandwidth d_j on row j for $j = 2, \ldots, J$, and has positive diagonal elements. Therefore, $L_{j,q} = 0$ for $1 \le q < j - d_j$

and $q > j$. $\mathbf{W}^{-1} \in \mathcal{R}^{K\times K}$ is a sparse positive definition matrix. Let $\max_{j=2,\ldots,J} d_j = O(1)$. Let w and $v \equiv \sum_{j=2}^{J} d_j$ represent the total numbers of non-zero off-diagonal elements in $\mathbf{W}^{-1}$ and $\mathbf{L}$ respectively, and satisfy $w = O(K)$, $v = O(J)$.

(C2) There exist positive constants τ_1 and τ_2 such that

$$0 < \tau_1^{-1} \leq \sigma_{\min}(\mathbf{L}) \leq \sigma_{\max}(\mathbf{L}) \leq \tau_1 < \infty,$$
$$0 < \tau_2^{-1} \leq \sigma_{\min}(\mathbf{W}^{-1}) \leq \sigma_{\max}(\mathbf{W}^{-1}) \leq \tau_2 < \infty,$$

where $\sigma_{\min}(\cdot)$ and $\sigma_{\max}(\cdot)$ denote the minimum and maximum singular values of a matrix.

(C3) Let the square of the jth component of $\text{vec}(\mathbf{Y}_i)$ have distribution function G_j, then

$$\max_{1\leq j\leq KJ} \int_0^\infty \exp(\psi t) dG_j(t) < \infty, \text{ for all } \psi \in (0, \psi_0),$$

where $\psi_0 > 0$ is a constant.

For (C1), it means that the true precision matrix $\mathbf{\Sigma}^{-1} = \mathbf{R}^{-1} \otimes \mathbf{W}^{-1}$ has a sparse block banded structure. (C2) suggests that the singular value of $\mathbf{L}$ is bounded, which is equivalent to the bounded eigenvalue condition generally. In reference to (C3), which is defined similarly to the condition in Bickel and Levina (2008), to accommodate the departure from normality, we establish that the maximum difference between $\mathbf{S}_{j,l}$ and $\mathbf{\Sigma}_{j,l}$, denoted as $\max_{1\leq j,l\leq KJ} |\mathbf{S}_{j,l} - \mathbf{\Sigma}_{j,l}|$, always satisfies the inequality $O_p(\log\{KJ\}/n)$.

Here, $\mathbf{S}$ represents the sample covariance matrix for the random samples $\{\mathbf{y}_i\}_{i=1}^n$.

3.1 Precision matrix estimation consistency

We now consider the convergence rate of the precision matrix estimator $\widehat{\mathbf{\Sigma}}^{-1}$ when $m \equiv KJ$ and n both diverge to infinity. Denote $a \asymp b$ as $c_1 \leq |a/b| \leq c_2$, where c_1 and c_2 are positive constants.

Theorem 2. *Assuming that Assumptions (C1),(C2) and (C3) hold, and the tuning parameters satisfy* $\lambda_1 \asymp (\log m/n)^{1/2}$ *and* $\lambda_2 \asymp (\log m/n)^{1/2}$. *If* $(K+J)\log m = o(n)$, *then there exists a local minimizer of* (2.2). *Moreover, the estimators* $\widehat{\mathbf{W}}^{-1}$, $\widehat{\mathbf{L}}_j$ *and* $\widehat{\mathbf{\Sigma}}^{-1}$ *converge in the sense that*

$$\|\widehat{\mathbf{W}}^{-1} - \mathbf{W}^{-1}\|_F = O_p\{(K\log m/n)^{1/2}\}, \quad \|\widehat{\mathbf{L}}_j - \mathbf{L}_j\|_2 = O_p\{(\log m/n)^{1/2}\},$$

and

$$\|\widehat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}^{-1}\|_F = O_p\{(m\log m/n)^{1/2}\}.$$

Furthermore, if $\mathrm{vec}(\mathbf{Y}_i) \sim N(\mathbf{0}, \mathbf{R}\otimes\mathbf{W})$, *then the estimators* $\widehat{\mathbf{W}}^{-1}, \widehat{\mathbf{L}}_j$ *and* $\widehat{\mathbf{\Sigma}}^{-1}$ *will converge at a faster rate. Specifically, assume Assumptions (C1), (C2) and (C3) to hold and the tuning parameters to satisfy* $\lambda_1 \asymp \{\log K/(nJ)\}^{1/2}$ *and* $\lambda_2 \asymp \{\log J/(nK)\}^{1/2}$. *If* $J\log J \asymp K\log K$, *and* $\log K = o(n)$, $\log J = o(n)$, *then the estimators* $\widehat{\mathbf{W}}^{-1}, \widehat{\mathbf{L}}_j$ *and*

$\widehat{\mathbf{\Sigma}}^{-1}$ *converge in the sense that*

$$\|\widehat{\mathbf{W}}^{-1} - \mathbf{W}^{-1}\|_F = O_p[\{K \log K/(nJ)\}^{1/2}], \quad \|\widehat{\mathbf{L}}_j - \mathbf{L}_j\|_2 = O_p[\{\log J/(nK)\}^{1/2}],$$

and

$$\|\widehat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}^{-1}\|_F = O_p[\max\{(J \log J/n)^{1/2}, (K \log K/n)^{1/2}\}].$$

Theorem 2 establishes the convergence rate of the precision matrix estimation. This rate agrees with that of SMGM (Leng and Tang, 2012), and is better than that of ABR (Qian et al., 2020), which will be reflected in the simulation studies. As noted by Leng and Tang (2012), when multiple local minimizers exist, identifying the optimal solution in practice becomes challenging, and there does not seem to be an algorithm that can consistently find the optimal solution.

3.2 Uniqueness of the banded estimator

In handling the optimization problem in (2.4), we have decomposed it into J separate optimization problems. It is worth noting that without the banded requirement, the individual estimator $\widehat{\mathbf{L}}_j$ may not be unique. For example, when the dimension of $\mathbf{L}_j$ satisfies $j > nK$, the objective function in (2.5) may not be strictly convex as a function of $\mathbf{L}_j$ which leads to multiple minimizers. However, when λ_2 is sufficiently large, we will show that the additional banded requirement will lead to sufficient sparseness so that the optimizor $\widehat{\mathbf{L}}_j$ will be unique. In order to show this, we first establish two lemmas.

Lemma 1. *For any given $\lambda_2 > 0$ and any given $\widehat{\mathbf{W}}^{-1}$, $\widehat{\mathbf{L}}_j$ is a solution to the objective function $\min_{L_{j,j}>0,\mathbf{L}_j\in\mathcal{R}^j} f(\mathbf{L}_j)$, where*

$$f(\mathbf{L}_j) \equiv -2\log L_{j,j} + \frac{1}{nK}\|\mathbf{Y}^*_{\cdot,1:j}\mathbf{L}_j\|_2^2 + \lambda_2 p(\mathbf{L}_j), \tag{3.1}$$

iff there exists $\widehat{\mathbf{A}} \in \mathcal{R}^{j\times(j-1)}$ such that

$$-\frac{2}{\widehat{L}_{j,j}}\mathbf{e}_j + \frac{2}{nK}\mathbf{Y}^{*\mathrm{J}}_{\cdot,1:j}\mathbf{Y}^*_{\cdot,1:j}\widehat{\mathbf{L}}_j + \lambda_2\sum_{l=1}^{j-1}\widehat{\mathbf{A}}_{\cdot,l} = 0, \tag{3.2}$$

where for $l = 1,\ldots,j-1$,

$$\widehat{\mathbf{A}}_{l+1:j,l} = \mathbf{0}, \widehat{\mathbf{A}}_{1:l,l} = (\widehat{\mathbf{L}}_j)_{1:l}/\|(\widehat{\mathbf{L}}_j)_{1:l}\|_2, \text{ if } (\widehat{\mathbf{L}}_j)_{1:l} \neq \mathbf{0}, \text{ and } \|\widehat{\mathbf{A}}_{1:l,l}\|_2 \leq 1. \tag{3.3}$$

Further, if the tuning parameter $\lambda_2 = C(\log m/n)^{1/2}$, where C is a sufficiently large constant, then under the conditions in Theorem 2, the estimator $\widehat{\mathbf{L}}_j$ is sparse with bandwidth $\widehat{d}_j$, and $\|\widehat{\mathbf{A}}_{1:l,l}\|_2 < 1$ for $l = 1,\ldots,j-1-\widehat{d}_j$.

Lemma 2. *Let $\widehat{\mathbf{L}}_j$ and $\widehat{\mathbf{A}}$ be as defined in Lemma 1. Assume that $\|\widehat{\mathbf{A}}_{1:l,l}\|_2 < 1$ for $l = 1,\ldots,j-\widehat{d}_j-1$. Then, any other solution $\widetilde{\mathbf{L}}_j$ has a bandwidth at most that of $\widehat{\mathbf{L}}_j$, i.e., $\widetilde{d}_j \leq \widehat{d}_j$.*

Theorem 3. *For any given $\lambda_2 > 0$ and $\widehat{\mathbf{W}}^{-1}$, let $\widehat{\mathbf{L}}_j$ be a solution to the objective function $\min_{L_{j,j}>0,\mathbf{L}_j\in\mathcal{R}^j} f(\mathbf{L}_j)$ with bandwidth $\widehat{d}_j$, where $f(\mathbf{L}_j)$ is defined in Lemma 1. Let $\widehat{\mathbf{A}}$ be as defined in Lemma 1, and define the non-zero index set $\widehat{D} \equiv \{l : \widehat{L}_{j,l} \neq 0\}$. Let*

$\lambda_2 = C(\log m/n)^{1/2}$, where C is a sufficiently large constant, and assume $\mathbf{Y}^*_{\cdot,\widehat{D}}$ has full column rank, i.e., $\text{rank}(\mathbf{Y}^*_{\cdot,\widehat{D}}) = \widehat{d}_j + 1$. Then, $\widehat{\mathbf{L}}_j$ is unique.

3.3 True bandwidth recovery

In this section, we show that our estimator $\widehat{\mathbf{L}}$ can correctly recover the true bandwidth of each row uniformly with probability approaching 1 under mild conditions. To show this, following the primal-dual witness procedure in Yu and Bien (2017), we first construct the primal-dual witness solution pairs $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{L}})$ for the optimal problem assuming the true bandwidth d_j of each row is known. We then prove that this solution is identical to the solution to (2.4), which in turn implies that the estimated bandwidths are identical to the true bandwidths.

Theorem 4. *Assume the conditions required in Lemmas 1, 2 and Theorems 2, 3 are satisfied, if the condition $\min_{j\in\{2,\ldots,J\}} \min_{l\geq j-d_j} |L_{j,l}| > \lambda_2$ is satisfied, then*

$$pr(\sup_{j=2,\ldots,J} |\widehat{d}_j - d_j| = 0) \to 1.$$

Remark 3. Theorem 4 holds under the assumption that the nonzero entries in $\mathbf{L}$ are uniformly bounded below by $(\log m/n)^{1/2}$. This implies that the minimal signal strength of $\mathbf{L}$ that is detectable is determined by the relation between the matrix size K, J and the sample size n. A larger matrix requires stronger minimal signal.

4. Simulation Studies

We now conduct simulation studies to study the finite sample performance of BKS. For comparison, in addition to BKS, we also implement some competitive methods, including SMGM (Leng and Tang, 2012) and ABR (Qian et al., 2021), both are designed to handle matrix-valued data. In addition, we also implement VB and Unweighted VB (UVB) proposed in Yu and Bien (2017), which is suitable when the vector vec($\mathbf{Y}_i$) is ordered, i.e. nearby components of vec($\mathbf{Y}_i$) have larger correlations. To facilitate the comparison to VB, we introduce a weighted version of BKS, where the penalty in (2.1) is modified to $p_w(\mathbf{L}_{j\cdot}) \equiv \sum_{l=1}^{j-1}(\sum_{q=1}^{l} w_{jq}^2 L_{j,q}^2)^{1/2}$, where $w_{jq} = 1/(j-q+1)^2$. Please note that the weights w_{jq} are the same as that of VB. We name the corresponding method Weighted BKS (WBKS). Note that these figures in simulation studies both can be found in Supplement Material S6.

4.1 Multivariate normal distribution

We generate the precision matrix by setting $\mathbf{\Sigma}^{-1} = (\mathbf{L}^{\mathrm{T}}\mathbf{L}) \otimes \mathbf{W}^{-1}$, where $\mathbf{L}$ is a lower triangular matrix with ones on the diagonal and row specific bandwidth $d_j, j = 1, \ldots, J$, and $\mathbf{W}^{-1}$ is a sparse positive definite matrix. To generate $\mathbf{L}$, we consider two cases.

- Case 1. $L_{j,l} = I(j - l = 0) + 0.8I(j - l = 1) + 0.6I(j - l = 2) + 0.4I(j - l = 3) + 0.2I(j - l = 4)$, where $j = 1, \ldots, J$, and $1 \le l \le j$.
- Case 2. $L_{j,l} = 0.7^{|j-l|}$, where $j = 1, \ldots, J$, and $j - d_j \le l \le j$. Here, d_j is randomly generated from a discrete uniform distribution on $[1, j/2]$.

We can see that $\mathbf{L}$ in Case 1 is a banded matrix with equal values 1, 0.8, 0.6, 0.4, 0.2 on the lower bands starting from the diagonal, while in Case 2, $\mathbf{L}$ has row-specific bandwidth, and in each row, the values of the nonzero elements are decreasing while moving away from the diagonal position. To generate $\mathbf{W}^{-1}$, we set $\mathbf{W}^{-1} = 0.5(\mathbf{B} + \mathbf{B}^{\mathrm{T}}) + c\mathbf{I}_K$, where $\mathbf{B}$ is a strictly upper triangular matrix with its elements independently generated from a Bernoulli distribution with parameter 0.1, and c is chosen such that the condition number of $\mathbf{W}^{-1}$ is K. We illustrate the structure of the precision matrix in the two cases when $K = 20, J = 10$ in Figure 1.

We then proceed to generate the n independent longitudinal data $\{\mathrm{vec}(\mathbf{Y}_i)\}_{i=1}^n$ from the $m = KJ$ dimensional multivariate normal distribution $N(\mathbf{0}, \mathbf{\Sigma})$. We consider sample sizes $n = 10, 50$ and 100, and repeat 100 times under each sample size. We consider four combinations of (K, J).

We report the estimation accuracy of the estimator $\widehat{\mathbf{\Sigma}}^{-1}$ in terms of two criterions, Frobenius norm (FN) and Kullback-Leibler (KL) loss, defined as

$$\Delta_{\mathrm{FN}}(\widehat{\mathbf{\Sigma}}^{-1}, \mathbf{\Sigma}^{-1}) \equiv \frac{1}{m}\|\widehat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}^{-1}\|_F^2, \quad \Delta_{\mathrm{KL}}(\widehat{\mathbf{\Sigma}}^{-1}, \mathbf{\Sigma}^{-1}) \equiv \frac{1}{m}\{\mathrm{tr}(\mathbf{\Sigma}^{-1}\widehat{\mathbf{\Sigma}}) - \ln(\mathbf{\Sigma}^{-1}\widehat{\mathbf{\Sigma}}) - m\}$$

respectively. We also report the performance of bandwidth recovery using the true negative rate (TNR) and the true positive rate (TPR) (Leng and Tang, 2012), that is,

$$\mathrm{TNR} = \frac{\#\{\widehat{\mathbf{\Sigma}}_{ij} = 0 \quad \& \quad \mathbf{\Sigma}_{ij} = 0\}}{\#\{\mathbf{\Sigma}_{ij} = 0\}}, \quad \mathrm{TPR} = \frac{\#\{\widehat{\mathbf{\Sigma}}_{ij} \neq 0 \quad \& \quad \mathbf{\Sigma}_{ij} \neq 0\}}{\#\{\mathbf{\Sigma}_{ij} \neq 0\}}.$$

Table 1: Comparison in terms of FN and KL for different estimators of the precision matrix $\mathbf{\Sigma}^{-1}$, in the form average$_{\text{standarderror}}$, over 100 replications in Case 1.

	(K, J)	(10,10)	(10,20)	(20,10)	(20,20)	(10,10)	(10,20)	(20,10)	(20,20)
	Methods	FN				KL			
10	WBKS	$\mathbf{1.276}_{0.315}$	$3.256_{0.352}$	$2.620_{0.527}$	$4.782_{0.499}$	$0.106_{0.021}$	$0.120_{0.014}$	$0.075_{0.012}$	$0.078_{0.008}$
	BKS	$1.499_{0.349}$	$\mathbf{2.865}_{0.330}$	$\mathbf{2.412}_{0.458}$	$\mathbf{4.157}_{0.467}$	$\mathbf{0.096}_{0.017}$	$\mathbf{0.093}_{0.011}$	$\mathbf{0.073}_{0.012}$	$\mathbf{0.065}_{0.007}$
	SMGM	$7.714_{3.912}$	$7.287_{2.823}$	$12.39_{4.107}$	$12.45_{2.833}$	$0.764_{0.538}$	$0.403_{0.449}$	$0.365_{0.325}$	$0.215_{0.208}$
50	WBKS	$0.454_{0.150}$	$0.539_{0.114}$	$\mathbf{0.541}_{0.138}$	$0.881_{0.135}$	$0.018_{0.003}$	$0.022_{0.003}$	$0.015_{0.002}$	$0.014_{0.001}$
	BKS	$\mathbf{0.328}_{0.113}$	$\mathbf{0.477}_{0.103}$	$0.634_{0.158}$	$\mathbf{0.644}_{0.112}$	$\mathbf{0.017}_{0.003}$	$\mathbf{0.019}_{0.003}$	$\mathbf{0.014}_{0.002}$	$\mathbf{0.013}_{0.001}$
	SMGM	$0.404_{0.094}$	$0.506_{0.095}$	$1.252_{0.176}$	$1.862_{0.181}$	$0.023_{0.003}$	$0.026_{0.004}$	$0.027_{0.002}$	$0.022_{0.002}$
	ABR	$10.59_{0.093}$	$12.75_{0.070}$	$20.12_{0.092}$	—	$0.386_{0.009}$	$0.429_{0.007}$	$0.481_{0.007}$	—
	VB	$10.73_{0.077}$	$12.96_{0.055}$	$18.59_{0.102}$	$22.64_{0.061}$	$0.396_{0.015}$	$0.468_{0.012}$	$0.442_{0.010}$	$0.520_{0.007}$
	UVB	$11.03_{0.067}$	$12.61_{0.050}$	$20.30_{0.071}$	$23.54_{0.051}$	$0.475_{0.012}$	$0.498_{0.010}$	$0.568_{0.008}$	$0.598_{0.005}$
100	WBKS	$0.172_{0.059}$	$0.327_{0.076}$	$0.261_{0.071}$	$0.308_{0.068}$	$0.009_{0.001}$	$0.011_{0.001}$	$0.007_{0.001}$	$0.007_{0.001}$
	BKS	$\mathbf{0.152}_{0.052}$	$\mathbf{0.324}_{0.075}$	$\mathbf{0.215}_{0.049}$	$\mathbf{0.273}_{0.062}$	$\mathbf{0.009}_{0.001}$	$\mathbf{0.009}_{0.001}$	$\mathbf{0.007}_{0.001}$	$\mathbf{0.006}_{0.001}$
	SMGM	$0.299_{0.078}$	$0.476_{0.085}$	$1.078_{0.145}$	$1.880_{0.143}$	$0.011_{0.002}$	$0.013_{0.001}$	$0.014_{0.001}$	$0.014_{0.001}$
	ABR	$10.59_{0.093}$	$11.78_{0.046}$	$17.98_{0.096}$	—	$0.386_{0.009}$	$0.327_{0.003}$	$0.364_{0.004}$	—
	VB	$10.73_{0.077}$	$12.96_{0.055}$	$17.13_{0.079}$	$20.72_{0.052}$	$0.396_{0.014}$	$0.468_{0.012}$	$0.314_{0.005}$	$0.366_{0.004}$
	UVB	$11.03_{0.067}$	$12.60_{0.050}$	$18.21_{0.057}$	$21.63_{0.040}$	$0.475_{0.012}$	$0.498_{0.010}$	$0.441_{0.006}$	$0.476_{0.004}$

During the estimation process, it is common for the estimated values of elements in $\mathbf{\Sigma}$ that should be 0 to be very small in absolute value but not exactly zero. In order to examine the accuracy of structure recovery, we assign a value of zero to all estimates below 0.01 in absolute value across all methods. In Table 1, we present the matrix estimation performance in Case 1. The results show that the average of FN and KL both decrease when the sample size increases for all estimators.

Table 2: Comparison in terms of TNR and TPR for different estimators of the precision matrix $\mathbf{\Sigma}^{-1}$, in the form average$_{\text{standarderror}}$, over 100 replications in Case 1.

	(K, J)	(10,10)	(10,20)	(20,10)	(20,20)	(10,10)	(10,20)	(20,10)	(20,20)
	Methods	TNR				TPR			
10	WBKS	$70.89_{0.054}$	$85.17_{0.017}$	$86.56_{0.023}$	$92.45_{0.008}$	$\mathbf{98.34}_{0.018}$	$93.50_{0.022}$	$97.13_{0.020}$	$94.24_{0.018}$
	BKS	$77.21_{0.044}$	$89.20_{0.018}$	$84.49_{0.023}$	$93.78_{0.008}$	$97.44_{0.021}$	$\mathbf{94.73}_{0.019}$	$\mathbf{97.27}_{0.019}$	$\mathbf{94.79}_{0.018}$
	SMGM	$\mathbf{88.51}_{0.133}$	$\mathbf{92.69}_{0.038}$	$\mathbf{89.85}_{0.054}$	$\mathbf{93.96}_{0.011}$	$41.25_{0.395}$	$62.98_{0.290}$	$54.09_{0.255}$	$67.83_{0.099}$
50	WBKS	$85.88_{0.030}$	$84.16_{0.020}$	$87.31_{0.016}$	$92.95_{0.007}$	$99.61_{0.009}$	$99.19_{0.010}$	$99.94_{0.004}$	$99.82_{0.003}$
	BKS	$82.39_{0.035}$	$86.32_{0.021}$	$87.96_{0.015}$	$93.40_{0.007}$	$\mathbf{99.77}_{0.007}$	$\mathbf{99.37}_{0.008}$	$\mathbf{99.94}_{0.004}$	$\mathbf{99.91}_{0.002}$
	SMGM	$56.72_{0.043}$	$77.27_{0.021}$	$59.00_{0.020}$	$82.08_{0.010}$	$98.48_{0.018}$	$96.70_{0.020}$	$95.72_{0.031}$	$91.36_{0.025}$
	ABR	$81.96_{0.009}$	$92.86_{0.003}$	$89.75_{0.007}$	—	$39.56_{0.009}$	$33.33_{0.005}$	$30.76_{0.009}$	—
	VB	$\mathbf{93.51}_{0.005}$	$\mathbf{97.89}_{0.002}$	$94.74_{0.003}$	$\mathbf{98.43}_{0.001}$	$33.25_{0.010}$	$25.88_{0.006}$	$30.62_{0.008}$	$22.57_{0.005}$
	UVB	$86.19_{0.005}$	$92.63_{0.001}$	$\mathbf{95.87}_{0.002}$	$98.12_{0.001}$	$31.21_{0.007}$	$29.51_{0.004}$	$14.79_{0.003}$	$12.81_{0.002}$
100	WBKS	$76.73_{0.034}$	$88.70_{0.012}$	$86.30_{0.015}$	$91.10_{0.013}$	$99.94_{0.003}$	$99.85_{0.004}$	$100.00_{0.00}$	$100.00_{0.00}$
	BKS	$76.56_{0.034}$	$90.05_{0.014}$	$85.58_{0.015}$	$\mathbf{93.05}_{0.012}$	$\mathbf{99.97}_{0.002}$	$\mathbf{99.89}_{0.003}$	$\mathbf{100.00}_{0.00}$	$\mathbf{100.00}_{0.00}$
	SMGM	$65.31_{0.038}$	$82.97_{0.019}$	$66.97_{0.020}$	$87.41_{0.013}$	$99.12_{0.015}$	$98.06_{0.015}$	$97.11_{0.023}$	$92.01_{0.019}$
	ABR	$74.90_{0.009}$	$90.96_{0.003}$	$82.30_{0.004}$	—	$52.11_{0.013}$	$39.84_{0.008}$	$38.33_{0.006}$	—
	VB	$\mathbf{84.39}_{0.006}$	$\mathbf{93.51}_{0.003}$	$\mathbf{92.01}_{0.003}$	$91.40_{0.001}$	$48.44_{0.012}$	$37.59_{0.007}$	$38.66_{0.008}$	$31.96_{0.005}$
	UVB	$76.98_{0.005}$	$87.19_{0.002}$	$82.54_{0.003}$	$92.25_{0.001}$	$46.57_{0.011}$	$47.53_{0.007}$	$34.41_{0.005}$	$29.47_{0.003}$

WBKS and BKS perform the best on average in every circumstance, because these methods fully take into account the sparsity of $\mathbf{W}^{-1}$, the bandedness of $\mathbf{L}$ and the Kronecker product structure. In contrast, SMGM does not take into account the banded matrix feature for $\mathbf{L}$, ABR does not utilize the Kronecker product nature of the precision matrix, and VB and UVB ignore sparsity and the Kronecker product structure. We also find that WBKS and BKS tend to have rather small variability. In fact, they have the smallest variability when sample size $n = 10$, reflecting superior estimation efficiency.

Table 3: Comparison in terms of FN and KL for different estimators of the precision matrix $\mathbf{\Sigma}^{-1}$, in the form $\text{average}_{\text{standarderror}}$, over 100 replications in Case 2.

	(K, J)	(10,10)	(10,20)	(20,10)	(20,20)	(10,10)	(10,20)	(20,10)	(20,20)
	Methods	FN				KL			
10	WBKS	$1.651_{0.220}$	$2.333_{0.210}$	$1.708_{0.220}$	$2.220_{0.286}$	$0.104_{0.018}$	$0.109_{0.012}$	$0.078_{0.012}$	$0.072_{0.008}$
	BKS	$\mathbf{1.308}_{0.210}$	$\mathbf{1.687}_{0.199}$	$\mathbf{1.610}_{0.205}$	$\mathbf{1.934}_{0.254}$	$\mathbf{0.091}_{0.017}$	$\mathbf{0.084}_{0.010}$	$\mathbf{0.075}_{0.011}$	$\mathbf{0.060}_{0.007}$
	SMGM	$3.976_{1.825}$	$5.042_{2.019}$	$5.153_{2.788}$	$6.521_{0.778}$	$0.827_{0.416}$	$0.883_{0.435}$	$0.485_{0.418}$	$0.138_{0.086}$
50	WBKS	$0.195_{0.064}$	$0.268_{0.058}$	$0.313_{0.076}$	$0.511_{0.091}$	$\mathbf{0.015}_{0.003}$	$0.021_{0.002}$	$\mathbf{0.015}_{0.002}$	$0.014_{0.001}$
	BKS	$\mathbf{0.180}_{0.060}$	$\mathbf{0.242}_{0.054}$	$\mathbf{0.285}_{0.069}$	$\mathbf{0.390}_{0.077}$	$\mathbf{0.015}_{0.003}$	$\mathbf{0.018}_{0.002}$	$\mathbf{0.015}_{0.002}$	$\mathbf{0.013}_{0.001}$
	SMGM	$0.190_{0.042}$	$0.243_{0.044}$	$0.410_{0.058}$	$0.935_{0.109}$	$0.021_{0.003}$	$0.024_{0.002}$	$0.025_{0.003}$	$0.020_{0.002}$
	ABR	$4.797_{0.062}$	$6.080_{0.041}$	$8.882_{0.051}$	—	$0.350_{0.010}$	$0.351_{0.006}$	$0.455_{0.006}$	—
	VB	$4.635_{0.058}$	$6.310_{0.031}$	$7.982_{0.052}$	$12.91_{0.042}$	$0.340_{0.016}$	$0.397_{0.009}$	$0.421_{0.012}$	$0.480_{0.008}$
	UVB	$4.836_{0.051}$	$6.351_{0.034}$	$8.719_{0.043}$	$13.61_{0.041}$	$0.425_{0.014}$	$0.447_{0.007}$	$0.534_{0.008}$	$0.562_{0.006}$
100	WBKS	$0.124_{0.037}$	$0.214_{0.042}$	$\mathbf{0.167}_{0.038}$	$0.203_{0.044}$	$\mathbf{0.008}_{0.001}$	$0.011_{0.001}$	$\mathbf{0.007}_{0.001}$	$0.007_{0.001}$
	BKS	$\mathbf{0.113}_{0.035}$	$\mathbf{0.177}_{0.038}$	$0.186_{0.042}$	$\mathbf{0.187}_{0.042}$	$\mathbf{0.008}_{0.001}$	$\mathbf{0.009}_{0.001}$	$\mathbf{0.007}_{0.001}$	$\mathbf{0.006}_{0.000}$
	SMGM	$0.097_{0.025}$	$0.190_{0.035}$	$0.254_{0.045}$	$0.881_{0.087}$	$0.010_{0.002}$	$0.013_{0.001}$	$0.011_{0.001}$	$0.011_{0.001}$
	ABR	$3.719_{0.056}$	$4.194_{0.054}$	$7.831_{0.037}$	—	$0.226_{0.006}$	$0.257_{0.005}$	$0.339_{0.004}$	—
	VB	$3.821_{0.047}$	$5.600_{0.030}$	$6.804_{0.046}$	$11.70_{0.038}$	$0.216_{0.007}$	$0.267_{0.004}$	$0.272_{0.005}$	$0.330_{0.004}$
	UVB	$3.799_{0.041}$	$5.188_{0.028}$	$7.733_{0.037}$	$12.36_{0.025}$	$0.275_{0.008}$	$0.299_{0.005}$	$0.409_{0.005}$	$0.428_{0.004}$

Between the weighted and unweighted versions of BKS, we would recommend BKS, due to its simplicity and its better performance under larger J. To demonstrate the bandwidth recovery performance of different estimators, we further report the averages and standard errors of TNR and TPR in Table 2. Once again, WBKS and BKS consistently demonstrate superior performance compared to their competitors, as evidenced by their large and balanced TNR and TPR values, and the superior ROC curves presented in Figure 4. Due to the similar performance from BKS and WBKS, we only choose to show

Table 4: Comparison in terms of TNR and TPR for different estimators of the precision matrix $\mathbf{\Sigma}^{-1}$, in the form $\text{average}_{\text{standarderror}}$, over 100 replications in Case 2.

	(K, J)	(10,10)	(10,20)	(20,10)	(20,20)	(10,10)	(10,20)	(20,10)	(20,20)
	Methods	TNR				TPR			
10	WBKS	$82.17_{0.052}$	$84.48_{0.023}$	$80.40_{0.035}$	$80.08_{0.023}$	$98.02_{0.028}$	$95.84_{0.026}$	$99.37_{0.013}$	$98.47_{0.011}$
	BKS	$83.39_{0.045}$	$87.71_{0.021}$	$81.62_{0.031}$	$82.70_{0.019}$	$\mathbf{99.14}_{0.017}$	$\mathbf{97.69}_{0.020}$	$\mathbf{99.43}_{0.013}$	$\mathbf{98.73}_{0.010}$
	SMGM	$\mathbf{85.90}_{0.235}$	$\mathbf{94.16}_{0.100}$	$\mathbf{84.00}_{0.131}$	$\mathbf{90.74}_{0.014}$	$32.53_{0.385}$	$26.63_{0.352}$	$54.27_{0.383}$	$65.10_{0.075}$
50	WBKS	$75.38_{0.041}$	$77.57_{0.029}$	$74.08_{0.029}$	$81.45_{0.015}$	$\mathbf{100.00}_{0.00}$	$\mathbf{99.95}_{0.006}$	$\mathbf{100.00}_{0.00}$	$99.03_{0.010}$
	BKS	$75.76_{0.039}$	$76.82_{0.028}$	$76.00_{0.026}$	$81.71_{0.015}$	$99.96_{0.004}$	$99.48_{0.007}$	$\mathbf{100.00}_{0.00}$	$\mathbf{99.32}_{0.007}$
	SMGM	$36.65_{0.056}$	$57.35_{0.045}$	$42.13_{0.041}$	$65.51_{0.021}$	$99.89_{0.006}$	$97.27_{0.015}$	$99.90_{0.006}$	$94.02_{0.019}$
	ABR	$66.55_{0.023}$	$80.64_{0.009}$	$62.07_{0.013}$	—	$66.25_{0.040}$	$45.89_{0.020}$	$72.32_{0.021}$	—
	VB	$\mathbf{75.86}_{0.015}$	$\mathbf{93.44}_{0.004}$	$85.03_{0.007}$	$93.22_{0.003}$	$58.98_{0.025}$	$30.36_{0.013}$	$48.30_{0.015}$	$28.20_{0.009}$
	UVB	$72.62_{0.010}$	$90.62_{0.003}$	$\mathbf{89.29}_{0.004}$	$\mathbf{94.93}_{0.001}$	$58.63_{0.019}$	$27.09_{0.007}$	$28.74_{0.010}$	$15.55_{0.003}$
100	WBKS	$\mathbf{72.52}_{0.038}$	$75.44_{0.030}$	$79.28_{0.020}$	$81.86_{0.013}$	$\mathbf{100.00}_{0.00}$	$99.28_{0.009}$	$\mathbf{100.00}_{0.00}$	$\mathbf{99.67}_{0.006}$
	BKS	$72.46_{0.039}$	$76.45_{0.028}$	$80.03_{0.019}$	$82.27_{0.012}$	$\mathbf{100.00}_{0.00}$	$\mathbf{99.42}_{0.007}$	$\mathbf{100.00}_{0.00}$	$99.65_{0.006}$
	SMGM	$53.73_{0.040}$	$70.07_{0.033}$	$52.12_{0.025}$	$75.65_{0.013}$	$100.00_{0.00}$	$96.80_{0.015}$	$99.96_{0.004}$	$94.51_{0.019}$
	ABR	$36.57_{0.018}$	$54.04_{0.011}$	$\mathbf{84.28}_{0.007}$	—	$94.96_{0.013}$	$86.75_{0.013}$	$47.41_{0.013}$	—
	VB	$57.40_{0.015}$	$\mathbf{83.85}_{0.007}$	$70.25_{0.008}$	$\mathbf{86.15}_{0.003}$	$77.54_{0.018}$	$45.46_{0.015}$	$66.50_{0.013}$	$40.71_{0.008}$
	UVB	$47.37_{0.011}$	$74.52_{0.005}$	$72.04_{0.005}$	$85.67_{0.002}$	$89.91_{0.015}$	$59.48_{0.011}$	$58.67_{0.011}$	$35.33_{0.005}$

the ROC curve of BKS. We also experiment with the data generated from Case 2 under sample sizes $n = 10, 50$ and 100, and present the corresponding results in Tables 3 and 4. A similar conclusion can be drawn as in Case 1. Besides, we provide the boxplot figures of FN and KL values, as well as the corresponding boxplots of TNR and TPR in two Cases in the Supplement Materials.

Further, we again find BKS tend to outperform WBKS, especially when J is large. This observation agrees with the relative performance of VB and UVB in Yu and Bien

Table 5: Average running time (seconds) of six precision matrix estimators under different combinations of (K, J). The results are based on 100 replicates with sample size $n = 100$.

Method	$(K, J) = (10, 5)$	$(K, J) = (10, 10)$	$(K, J) = (20, 10)$	$(K, J) = (10, 20)$
WBKS	1.22	1.55	2.20	2.95
BKS	**1.19**	**1.46**	**2.05**	**2.72**
SMGM	1.41	1.91	3.03	3.44
ABR	8.13	118.86	1531.42	3044.67
VB	20.41	84.89	776.71	792.86
UVB	2.79	14.62	138.18	122.68

(2017). Intuitively, this is because the banded property is a special sparseness, where on each row, the elements farther away from the diagonal is more likely to be zero. $p(\mathbf{L}_j)$ incorporates this feature by repeated penalization, while $p_w(\mathbf{L}_j)$ downweights the penalty in each repetition, hence somewhat reduces the heavier penalty for elements farther away from the diagonal imposed by $p(\mathbf{L}_j)$. Such downweighting is especially harmful when J is large due to larger sparseness. Further, we provide the ROC curves in the right Figure 4.

Finally, we compare the computational complexity of these methods by examining their respective running times, measured in seconds. The results are presented in Table 5. It is evident that BKS is the fastest, and this advantage becomes particularly significant as J increases.

4.2 Multivariate t distribution

We will analyze the precision matrix used in Case 1 of the multivariate t distribution. We generate n independent longitudinal data $\{\text{vec}(\mathbf{Y}_i)\}_{i=1}^n$ from the $m = KJ$ dimensional

multivariate t distribution with $df = 4$ degrees of freedom. We consider sample sizes of $n = 10, 50$ and 100. Similar to the multivariate normal distribution, we repeat the process 100 times for each sample size across four combinations of (K, J).

We evaluate the accuracy of the $\mathbf{\Sigma}^{-1}$ estimator using FN and KL loss as metrics. Furthermore, we evaluate the performance of bandwidth recovery through the use of TNR and TPR measures. Table 6 presents the performance of matrix estimation accuracy in Case 1 for the multivariate t distribution. The results reveal that, for all estimators, the average values of FN and KL decrease as the sample size increases. Additionally, WBKS and BKS consistently exhibit superior performance across all scenarios, mirroring their performance with the multivariate normal distribution. To illustrate the bandwidth recovery performance of different estimators, we also report the averages and standard errors of TNR and TPR in Table 7. It is evident that WBKS and BKS consistently outperform the other methods, as reflected by their high and well-balanced TNR and TPR values.

5. Real Data Analysis

5.1 EEG data

We apply our method to analyze a public data set EEG from the UCI machine learning repository dataset(`http://archive.ics.uci.edu/ml/datasets/EEG+Database`). This data set contains $n = 122$ subjects, including $n_0 = 45$ alcoholic subjects ($z = 0$) and $n_1 = 77$ controls ($z = 1$). Each subject had 64 electrodes placed on his/her scalp,

Table 6: Comparison in terms of FN and KL for different estimators of the precision matrix $\mathbf{\Sigma}^{-1}$ over 100 replications, in the form $\text{average}_{\text{standarderror}}$.

	(K, J)	(10,10)	(10,20)	(20,10)	(20,20)	(10,10)	(10,20)	(20,10)	(20,20)
	Methods	FN				KL			
10	WBKS	$7.453_{1.324}$	$13.376_{0.65}$	$17.117_{1.57}$	$14.598_{2.80}$	$0.299_{0.153}$	$0.521_{0.177}$	$0.366_{0.158}$	$0.276_{0.176}$
	BKS	$\mathbf{7.154}_{1.424}$	$\mathbf{11.584}_{1.68}$	$\mathbf{16.944}_{1.67}$	$\mathbf{13.467}_{3.24}$	$\mathbf{0.284}_{0.153}$	$\mathbf{0.395}_{0.184}$	$\mathbf{0.360}_{0.159}$	$\mathbf{0.263}_{0.185}$
	SMGM	$10.482_{0.54}$	$13.064_{0.48}$	$18.866_{0.99}$	$20.281_{2.56}$	$0.510_{0.316}$	$0.552_{0.229}$	$0.483_{0.225}$	$0.574_{0.372}$
50	WBKS	$4.789_{1.337}$	$\mathbf{5.424}_{1.183}$	$\mathbf{7.228}_{2.481}$	$9.638_{2.551}$	$0.224_{0.123}$	$0.195_{0.103}$	$\mathbf{0.198}_{0.121}$	$0.217_{0.126}$
	BKS	$\mathbf{4.748}_{1.361}$	$5.544_{1.204}$	$7.646_{2.484}$	$\mathbf{9.137}_{2.635}$	$\mathbf{0.223}_{0.124}$	$\mathbf{0.194}_{0.104}$	$0.201_{0.122}$	$\mathbf{0.212}_{0.126}$
	SMGM	$6.003_{3.175}$	$8.339_{4.785}$	$12.862_{7.49}$	$18.173_{7.68}$	$0.451_{0.472}$	$0.742_{0.581}$	$0.566_{0.428}$	$0.753_{0.420}$
	ABR	$12.643_{0.11}$	$14.021_{0.17}$	$22.438_{0.23}$	—	$0.638_{0.034}$	$0.573_{0.023}$	$0.728_{0.030}$	—
	VB	$12.485_{0.12}$	$14.155_{0.11}$	$21.570_{0.20}$	$25.040_{0.23}$	$0.596_{0.041}$	$0.586_{0.023}$	$0.616_{0.032}$	$0.672_{0.028}$
	UVB	$12.576_{0.15}$	$14.176_{0.14}$	$22.107_{0.29}$	$25.482_{0.34}$	$0.633_{0.047}$	$0.611_{0.027}$	$0.705_{0.057}$	$0.747_{0.056}$
100	WBKS	$\mathbf{4.028}_{1.181}$	$\mathbf{4.884}_{1.172}$	$7.057_{1.706}$	$\mathbf{8.644}_{2.069}$	$\mathbf{0.196}_{0.093}$	$0.189_{0.082}$	$0.191_{0.072}$	$0.200_{0.087}$
	BKS	$4.120_{0.171}$	$4.925_{1.175}$	$\mathbf{7.001}_{1.704}$	$8.816_{2.059}$	$\mathbf{0.196}_{0.093}$	$\mathbf{0.187}_{0.083}$	$\mathbf{0.190}_{0.072}$	$\mathbf{0.199}_{0.087}$
	SMGM	$4.412_{2.728}$	$5.712_{3.485}$	$8.219_{4.455}$	$16.386_{1.96}$	$0.307_{0.365}$	$0.375_{0.437}$	$0.269_{0.261}$	$0.317_{0.193}$
	ABR	$11.459_{0.15}$	$13.346_{0.12}$	$20.905_{0.21}$	—	$0.455_{0.028}$	$0.484_{0.024}$	$0.532_{0.021}$	—
	VB	$11.398_{0.03}$	$13.675_{0.12}$	$20.497_{0.20}$	$24.212_{0.16}$	$0.423_{0.033}$	$0.492_{0.029}$	$0.486_{0.029}$	$0.553_{0.024}$
	UVB	$11.479_{0.14}$	$13.430_{0.13}$	$22.065_{0.23}$	$25.337_{0.21}$	$0.452_{0.030}$	$0.488_{0.028}$	$0.660_{0.042}$	$0.688_{0.041}$

and the measurements were taken at 256 Hz (3.9ms epoch) for 1 second. The electrode positions were located at standard sites, see Zhang et al. (1995) for specific names of these standard sites. In addition, each subject was exposed to two situations, either a single stimulus (S1) or two stimuli (S1 and S2), which were pictures of objects chosen from the 1980 Snodgrass and Vanderwart picture set. When two stimuli were shown, they were presented in either a matched condition where S1 was identical to S2 or in a non-matched condition, where S1 differed from S2. In this dataset, each subject completed 120

Table 7: Comparison in terms of TNR and TPR for different estimators of the precision matrix $\mathbf{\Sigma}^{-1}$ over 100 replications, in the form $\text{average}_{\text{standarderror}}$.

	(K, J)	(10,10)	(10,20)	(20,10)	(20,20)	(10,10)	(10,20)	(20,10)	(20,20)
	Methods	TNR				TPR			
10	WBKS	$90.64_{0.027}$	$\mathbf{99.23}_{0.005}$	$95.61_{0.020}$	$93.42_{0.013}$	$86.47_{0.043}$	$37.02_{0.033}$	$\mathbf{57.77}_{0.053}$	$83.40_{0.031}$
	BKS	$90.51_{0.031}$	$98.71_{0.006}$	$95.76_{0.020}$	$96.06_{0.010}$	$\mathbf{86.71}_{0.043}$	$\mathbf{56.53}_{0.052}$	$56.99_{0.041}$	$\mathbf{87.00}_{0.028}$
	SMGM	$\mathbf{97.81}_{0.013}$	$98.80_{0.007}$	$\mathbf{98.14}_{0.011}$	$\mathbf{98.67}_{0.011}$	$45.50_{0.178}$	$35.28_{0.093}$	$35.43_{0.133}$	$36.83_{0.263}$
50	WBKS	$81.25_{0.051}$	$88.16_{0.022}$	$83.03_{0.046}$	$92.84_{0.017}$	$97.64_{0.022}$	$\mathbf{95.59}_{0.021}$	$\mathbf{98.36}_{0.019}$	$97.21_{0.018}$
	BKS	$84.77_{0.046}$	$90.83_{0.024}$	$87.51_{0.041}$	$93.46_{0.018}$	$\mathbf{97.91}_{0.020}$	$95.37_{0.021}$	$98.18_{0.019}$	$\mathbf{97.97}_{0.017}$
	SMGM	$78.30_{0.135}$	$89.04_{0.116}$	$79.14_{0.210}$	$94.03_{0.090}$	$72.28_{0.393}$	$49.20_{0.455}$	$51.69_{0.465}$	$32.73_{0.419}$
	ABR	$89.41_{0.030}$	$93.47_{0.015}$	$93.57_{0.038}$	—	$32.45_{0.028}$	$31.98_{0.025}$	$24.02_{0.045}$	—
	VB	$\mathbf{95.98}_{0.017}$	$\mathbf{97.21}_{0.009}$	$\mathbf{97.04}_{0.012}$	$98.62_{0.005}$	$25.49_{0.021}$	$23.12_{0.014}$	$21.71_{0.018}$	$16.89_{0.014}$
	UVB	$91.67_{0.014}$	$95.23_{0.006}$	$96.76_{0.008}$	$\mathbf{98.68}_{0.003}$	$23.41_{0.020}$	$23.50_{0.013}$	$13.08_{0.006}$	$11.00_{0.004}$
100	WBKS	$84.51_{0.045}$	$89.08_{0.021}$	$88.65_{0.038}$	$93.38_{0.018}$	$\mathbf{99.11}_{0.015}$	$98.06_{0.015}$	$\mathbf{99.61}_{0.011}$	$\mathbf{99.24}_{0.017}$
	BKS	$85.59_{0.044}$	$91.47_{0.023}$	$88.32_{0.039}$	$94.58_{0.017}$	$98.09_{0.016}$	$\mathbf{99.42}_{0.015}$	$\mathbf{99.61}_{0.012}$	$99.17_{0.011}$
	SMGM	$69.60_{0.125}$	$84.99_{0.073}$	$68.42_{0.114}$	$\mathbf{98.30}_{0.006}$	$86.62_{0.312}$	$81.10_{0.355}$	$88.29_{0.293}$	$78.04_{0.186}$
	ABR	$80.21_{0.029}$	$90.80_{0.013}$	$85.78_{0.024}$	—	$45.57_{0.036}$	$40.91_{0.032}$	$35.54_{0.020}$	—
	VB	$\mathbf{87.19}_{0.025}$	$\mathbf{95.38}_{0.012}$	$93.79_{0.014}$	$97.64_{0.007}$	$42.45_{0.022}$	$31.82_{0.015}$	$33.00_{0.013}$	$25.81_{0.009}$
	UVB	$81.36_{0.013}$	$91.70_{0.006}$	$\mathbf{95.61}_{0.007}$	$97.92_{0.004}$	$39.02_{0.020}$	$33.61_{0.017}$	$15.79_{0.009}$	$13.80_{0.009}$

trials under each situation. Taking averages over 120 trials, each subject has a 64×256 measurement matrix. Following Qian et al. (2021), we take the average of every 32 measurements to reduce the time dimension from 256 to 8 and obtain a 64×8 matrix for each subject. This eventually leads to a data set $\mathbf{Y}_i \in \mathcal{R}^{K \times J} (i = 1, \ldots, n)$ with $K = 64$, $J = 8$ and $n = 122$. In addition, we also have a class label $z_i \in \{0, 1\}$ for $i = 1, \ldots, n$.

Similar to Qian et al. (2021), we aim to classify these subjects into two classes, alcoholic (class 0) and control (class 1), based on the information in $\mathbf{Y}_i$'s. We consider two

methods, linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) to perform the classification. LDA classifies subject i to class 0 if $\delta^{(0)}_{\text{LDA}}(\mathbf{y}_i) > \delta^{(1)}_{\text{LDA}}(\mathbf{y}_i)$, otherwise to class 1, where

$$\delta^{(c)}_{\text{LDA}}(\mathbf{y}_i) = \mathbf{y}_i^{\text{T}}\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}^{(c)} - \frac{1}{2}(\widehat{\boldsymbol{\mu}}^{(c)})^{\text{T}}\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}^{(c)} + \log\widehat{\pi}^{(c)}$$

for $c = 0, 1$ and $\widehat{\pi}^{(c)}$ is the estimated proportion of group c. Here, $\widehat{\boldsymbol{\Sigma}}$ is the estimated overall covariance matrix, and $\widehat{\boldsymbol{\mu}}^{(c)}$ is the estimated mean in group c. Similarly, QDA classifies subject i to class 0 if $\delta^{(0)}_{\text{QDA}}(\mathbf{y}_i) > \delta^{(1)}_{\text{QDA}}(\mathbf{y}_i)$, otherwise to class 1, where

$$\delta^{(c)}_{\text{QDA}}(\mathbf{y}_i) = \mathbf{y}_i^{\text{T}}(\widehat{\boldsymbol{\Sigma}}^{(c)})^{-1}\widehat{\boldsymbol{\mu}}^{(c)} - \frac{1}{2}(\widehat{\boldsymbol{\mu}}^{(c)})^{\text{T}}(\widehat{\boldsymbol{\Sigma}}^{(c)})^{-1}\widehat{\boldsymbol{\mu}}^{(c)} + \log\widehat{\pi}^{(c)}.$$

for $c = 0, 1$. Here, $\widehat{\boldsymbol{\Sigma}}^{(c)}$ is the estimated covariance matrix in group c. To implement these methods, we randomly sample 70% of the data to form a training set and use the remaining 30% as the testing data. We used sample proportions to form $\widehat{\pi}^{(c)}$, sample averages to form $\widehat{\mu}^{(c)}$, $(c = 0, 1)$, and used WBKS, BKS, SMGM, ABR, VB and UVB to estimate $\boldsymbol{\Sigma}^{-1}$ in LDA and $(\boldsymbol{\Sigma}^{(c)})^{-1}, (c = 0, 1)$ in QDA. To select the tuning parameters in these methods, we used a five-fold crossvalidation.

The upper row of Figure 5 in Supplement shows the estimated precision matrix $\widehat{\boldsymbol{\Sigma}}^{-1}$ obtained from WBKS, BKS and SMGM. The plots from ABR, VB, and UVB are excluded since they are already provided in Qian et al. (2021).

WBKS and BKS lead to a block-banded precision matrix, which agrees with the

Table 8: The average classification errors of different methods over 10 random train-test splits.

	WBKS	BKS	SMGM	ABR	VB	UVB
LDA	0.19	0.18	0.24	0.25	0.24	0.28
QDA	0.18	0.18	–	0.24	0.29	0.29

general conclusion from ABR (Qian et al., 2021). In contrast, SMGM exhibits a lack of time correlation during the first three time points, followed by inter-correlation within the remaining five time points. This pattern is counter-intuitive. Finally, UVB and VB lead to simple banded precision matrix estimation, which is also unrealistic due to the spacial correlation that tends to persist across time. We further plot the resulting $\widehat{\mathbf{R}}^{-1}$ and $\widehat{\mathbf{W}}^{-1}$ by BKS in the lower row of Figure 5. We can clearly see the banded feature of $\widehat{\mathbf{R}}^{-1}$ and the sparseness of $\widehat{\mathbf{W}}^{-1}$. To evaluate the performance of these methods, we compute the classification errors on the testing data. The results in Table 6 contain the average testing data classification errors over 10 random train-test splits. It is clear that WBKS and BKS outperform the other methods. Among the remaining methods, ABR is the winner, indicating that the true precision matrix is close to have the block-banded structure. However, ABR is inferior to WBKS and BKS, possibly because it contains too many parameters (Qian et al., 2021). Note that due to the relative small sample size, SMGM fails to produce a result when performing QDA.

5.2 ADHD data

In this section, we will analyze a dataset pertaining Attention Deficit Hyperactivity Disorder (ADHD). ADHD is a prevalent mental disorder observed in children and adolescents, characterized by symptoms such as distractibility, impulsivity, and restlessness. Functional Magnetic Resonance Imaging (fMRI) data at rest from the ADHD-200 sample dataset (`http://www.nitrc.org/frs/?group_id=383`) were collected by Oregon Health and Science University. The data were processed using the Automated Anatomical Labeling (AAL) software package and a dedicated digital atlas designed for the human brain.

The ADHD dataset consists of 42 subjects who belong to the typical developmental control groups. These children are used as a baseline for comparing with individuals diagnosed with Attention Deficit Hyperactivity Disorder (ADHD). Brain activity is measured by detecting changes in blood flow correlated with low-frequency Blood Oxygen Level Dependent (BOLD) signals. Tzourio-Mazoyer et al. (2002) provided a detailed description of brain region segmentation. Each individual's brain was monitored in 116 regions of interest (ROIs), and the signals from these 116 ROIs for each child were recorded over 74 scans. Consequently, we obtained multivariate longitudinal data $\boldsymbol{Y}_i \in \mathcal{R}^{K\times J}(i = 1, \ldots, n)$, where $K = 116$, $J = 74$, $n = 42$. Leng and Pan (2018) assumed a Kronecker structure for the covariance matrix of this data and estimated it using large-dimensional random matrix theory. Their analysis revealed that the temporal covariance matrix exhibits a banded structure, where correlations are strongest near the main diagonal and gradually

decrease as they move away from it. However, the covariance matrix of the 116 brain regions demonstrates sparsity. Additionally, their method does not incorporate the banded structure information that exists in the longitudinal data.

To further explore the structural information among the K brain regions and J measurements in the ADHD data, we applied our proposed method to estimate the precision matrix of this dataset. Obtaining the ABR, UVB, and VB estimators for the data is challenging due to the high temporal and variable dimensions. Furthermore, we utilized the sparse matrix graphical model with Kronecker structure, as proposed by Leng and Tang (2012), to estimate the precision matrix for SMGM. Figure 6 shows the plotted structural information of the precision matrix estimator obtained using our method, while Figure 7 presents the structural information of the precision matrix estimator for SMGM. Due to the high dimensionality of $KJ = 8584$, which is conducive to detailed structural representation, we present the precision matrix separately for the temporal dimension (left) and the variable dimension (right).

From Figure 6, it is evident that the precision matrix in the temporal dimension exhibits an adaptive banded structure. The band width starts wider at the beginning and gradually narrows. The conditional correlations among the 116 brain regions in the variable dimension exhibit sparsity, with a distinct block structure observed in the top-left and bottom-right corners. However, the structural information in other regions appears relatively scattered, which can be attributed to the division of brain regions. Figure 7 displays the estimated precision matrix by SMGM, which is observed to be a diagonal

matrix, indicating a lack of captured structural information within the data. This finding aligns with the conclusions of Leng and Pan (2018). Additionally, Figure 8 presents the correlation matrix in the temporal and variable dimensions obtained by inverting the precision matrix. The results reveal a banded correlation structure among the 74 time points and a localized block-structured correlation pattern among the 116 brain regions. This can be attributed to the collective influence of specific local brain regions on certain human behaviors. In conclusion, the precision matrix estimator obtained from our proposed method effectively captures the conditional correlations among the 74 time points. Furthermore, we observe that the conditional correlations among the 116 brain regions exhibit sparsity, which aligns with the observed characteristics in reality.

6. Conclusions

We have proposed a new precision matrix estimator named BKS. BKS takes advantage of the fact that the precision matrix is the Kronecker product of a banded matrix and a sparse matrix. It incorporates the matrix bandedness by considering its Cholesky decomposition and imposing a new penalty which increasingly encourages zeros for elements farther away from the matrix diagonal. Matrix sparsity is enforced by applying the standard lasso penalty. BKS also guarantees a positive definite estimator of the precision matrix. BKS is easy to implement. It optimizes a biconvex objective function, and is achieved through an alternative optimization algorithm named ACS. ACS comprises of two repeating steps, with each step solving a convex optimization problem using the glas-

so and ADMM algorithms, respectively. This approach ensures computational efficiency. We show the algorithmic convergence of ACS and establish the statistical convergence rate of BKS. We find that BKS exhibits the same convergence rate as SMGM when the data is normally distributed. However, the application and advantageous characteristics of BKS extend beyond normality. We also demonstrate that BKS has the ability to recover the true bandwidths of the banded matrix with a probability close to 1. Both simulation studies and real data applications consistently indicate that BKS exhibits favorable performance overall.

Supplementary Materials

The supplementary materials provides the proofs of the lemmas, Theorem 1, Theorem 2, Theorem 3 and Theorem 4, and these figures in simulation and real data studies.

Acknowledgments

We are very grateful to the Editor,Associate Editor and referees, as well as our financial sponsors for their insightful comments and suggestions that have improved the manuscript significantly. This work was supported by the Key Program of the National Natural Science Foundation of China [grant number 12431009].

References

Aston, J. A., P. Davide, and T. Shahin (2017). Tests for separability in nonparametric covariance operators of random surfaces. *The Annals of Statistics 45*(4), 1431–1461.

Banerjee, O., L. El Ghaoui, and A. d'Aspremont (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research 9*(3), 485–516.

Bickel, P. J. and E. Levina (2008). Covariance regularization by thresholding. *Annals of Statistics 36*, 2577–2604.

Bien, J., F. Bunea, and L. Xiao (2016). Convex banding of the covariance matrix. *Journal of the American Statistical Association 111*(514), 834–845.

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

Dai, D., C. Hao, S. Jin, and Y. Liang (2023). Regularized estimation of kronecker structured covariance matrix using modified cholesky decomposition. *Journal of Statistical Computation and Simulation 95*(5), 1–26.

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

Furrer, R. and T. Bengtsson (2007). Estimation of high-dimensional prior and posterior covariance matrices in kalman filter variants. *Journal of Multivariate Analysis 98*(1), 227–255.

Gorski, J., F. Pfeuffer, and K. Klamroth (2007). Biconvex sets and optimization with biconvex functions: a survey and extensions. *Mathematical Methods of Operations Research 66*, 373408.

Greenewald, K. and A. O. Hero (2015). Robust kronecker product pca for spatio-temporal covariance estimation. *IEEE Transactions on Signal Processing 63*(23), 6368–6378.

Jenatton, R., J.-Y. Audibert, and F. Bach (2011). Structured variable selection with sparsity-inducing norms. *Journal of Machine Learning Research 12*, 2777–2824.

Kim, C. and D. L. Zimmerman (2012). Unconstrained models for the covariance structure of multivariate longitudinal data. *Journal of Multivariate Analysis 107*, 104–118.

Lee, K., H. Cho, M.-S. Kwak, and E. J. Jang (2020). Estimation of covariance matrix of multivariate longitudinal data using modified choleksky and hypersphere decompositions. *Journal of Multivariate Analysis 76*(1), 75–86.

Leng, C. and G. Pan (2018). Covariance estimation via sparse kronecker structures. *Bernoulli 24*(4B), 3833–3863.

Leng, C. and C. Tang (2012). Sparse matrix graphical models. *Journal of the American Statistical Association 107*(499), 1187–1200.

Qian, F., Y. Chen, and W. Zhang (2020). Regularized estimation of precision matrix for high-dimensional multivariate longitudinal data. *Journal of Multivariate Analysis 176*, 104580.

Qian, F., W. Zhang, and Y. Chen (2021). Adaptive banding covariance estimation for high-dimensional multivariate longitudinal data. *Canadian Journal of Statistics 49*(3), 906–938.

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

Tsiligkaridis, T. and A. O. Hero (2013). Covariance estimation in high dimensions via kronecker product expansions. *IEEE Transactions on Signal Processing 61*(21), 5347–5360.

Yu, G. and J. Bien (2017). Learning local dependence in ordered data. *Journal of Machine Learning Research 18*(42), 1–60.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Statistical Methodology Series B 68*, 49–67.

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

Zhang, X. L., H. Begleiter, and B. Porjesz (1995). Event related potentials during object recognition tasks. *Brain research bulletin 6*, 531–538.

Zhang, Y., W. Shen, and D. Kong (2023). Covariance estimation for matrix-valued data. *Journal of the American Statistical Association 118*(554), 2620–2631.

School of Sciences, Tianjin University of Commerce, 409 Guangrong Road, Tianjin, China

E-mail: (liangch722@163.com)

School of Mathematical Sciences, Capital Normal University, 105 North Xisanhuan Road, Beijing, China

E-mail: (wenqingma@cnu.edu.cn)

Department of Statistics, The Pennsylvania State University, University Park, Pennsylvania 16802, U.S.A.

E-mail: (yanyuanma@gmail.com)