University of California
Santa Barbara

Learning graphs for dependence and conditional dependence at different levels

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Statistics and Applied Probability

by

Sunpeng Duan

Committee in charge:

Professor Yuedong Wang, Committee Co-Chair
Professor Guo Yu, Committee Co-Chair
Professor Sang-Yun Oh

March 2024

The Dissertation of Sunpeng Duan is approved.

Professor Sang-Yun Oh

Professor Guo Yu, Committee Co-Chair

Professor Yuedong Wang, Committee Co-Chair

March 2024

Learning graphs for dependence and conditional dependence at different levels

by

Sunpeng Duan

To my beloved parents

Acknowledgements

I would like to express my deepest gratitude to my advisors, Professor Yuedong Wang and Professor Guo Yu, for their patience, motivation, expertise and continuous support throughout my Ph.D. study. Their patience and constructive feedbacks have been invaluable to me. And their insights into my research have been incredibly beneficial. Without their precious support, it would not be possible to complete this dissertation.

My gratitude also extends to Professor Sang-Yun Oh for his insightful comments and encouragement as my dissertation committee member. His valuable ideas and feedback have greatly contributed to my work.

In addition, I want to thank Fresenius Medical Care North America for providing de-identified data and Dr. Hanjie Zhang from Renal Research Institute for discussing real data analysis. I also owe special thanks to the National Institute of Diabetes and Digestive and Kidney Diseases (R01DK130067) for their support.

Lastly, I want to express my heartfelt thanks to my parents and friends for their unwavering love and support during my six years at University of California, Santa Barbara. Their presence has been a cornerstone of my life's journey.

Curriculum Vitæ

Sunpeng Duan

Education

2024	Ph.D. in Statistics and Applied Probability (Expected), University of California, Santa Barbara.
2022	M.A. in Statistics, University of California, Santa Barbara.
2018	M.S. in Statistics, Xiamen University, China.
2015	B.E. in Aircraft Propulsion Engineering, Xiamen University, China.

Experience

2018-2024	Teaching Assistant and Teaching Associate, Department of Statistics and Applied Probability, University of California, Santa Barbara.
2023	Data Scientist Intern, Capital One, Dallas.
2022	Research Assistant, Department of Statistics and Applied Probability, University of California, Santa Barbara.

Publications

1. **Duan, S.**, Wang, Y., Kotanko, P. and Zhang, H., *Network Analysis of Spread of SARS-CoV-2 Within Dialysis Clinics: A Multi-center Network Analysis*, PLOS One, **19** (2024) e0299855.
2. Zhong, W., Guo, W., **Duan, S.**, and Cui, H., *Conditional Test for Ultrahigh Dimensional Linear Regression Coefficients*, Statistica Sinica, **32** (2022) 1381–1409.
3. Zhong, W., **Duan, S.**, and Zhu, L., *Forward Additive Regression for Ultrahigh Dimensional Nonparametric Additive Models.* Statistica Sinica, **30** (2020), 175–192.

Abstract

Learning graphs for dependence and conditional dependence at different levels

by

Sunpeng Duan

Repeated measurements are common in many fields, where random variables are observed repeatedly across different subjects. Such data have an underlying hierarchical structure, and it is of interest to learn dependence structures at different levels. Most existing methods for sparse estimation of dependence and conditional dependence structures assume independent samples. Ignoring the underlying hierarchical structure within the subject may lead to erroneous scientific conclusion.

In Part I, we study the problem of sparse and positive-definite estimation of between-subject and within-subject covariance matrices for repeated measurements. Our estimators are solutions to convex optimization problems that can be solved efficiently. We establish estimation error rates for the proposed estimators and demonstrate their favorable performance through theoretical analysis and comprehensive simulation studies. We further apply our methods to construct between-subject and within-subject covariance graphs of clinical variables from hemodialysis patients.

Part II shifts the focus towards learning temporal, contemporaneous and between-subjects conditional dependence graphs with a graphical vector autoregression model. We propose a two-stage procedure for the simultaneous estimation of these three graphs. Furthermore, Bayesian information criteria are formulated for tuning parameters selection in our two-stage method. The performance of the proposed method is evaluated through extensive simulation studies and one real data application.

Contents

Chapter 1

Introduction

1.1 Undirected and Directed Graphical Models

Capturing dependencies and conditional dependencies among a set of random variables is fundamental in modern multivariate analysis. Covariance graphs and graphical models are powerful tools for the estimation of covariance and concentration or precision matrices in many fields such as biology, neuroscience, genomics, medicine, economics, and finance (Drton and Maathuis [1], Fan et al. [2], Hastie et al. [3], and Lauritzen [4]). A covariance graph or graphical model for a p-dimensional random vector $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_p)^{\mathrm{T}}$ is represented by a graph $G = (V, E)$, where the vertex set V contains p vertices corresponding to the p coordinates of $\mathbf{Y}$ and the edge set $E \subset V \times V$ reveals a set of marginal or conditional dependencies among $Y_1, Y_2, \ldots, Y_p$.

Undirected graphical models, known as concentration graphs or Markov networks, have been heavily investigated over the past decades, in which E reveals conditional dependencies between variables. See, for example, Friedman et al. [5], Rothman et al. [6], and Yuan and Lin [7]. In a concentration graph, the edge between Y_i and Y_j is absent if and only if Y_i and Y_j are independent conditional on the other variables, denoted by

$Y_i \perp\!\!\!\perp Y_j \mid \mathbf{Y}_{V\setminus\{i,j\}}$, where $\perp\!\!\!\perp$ represents independency and $\mathbf{Y}_{V\setminus\{i,j\}}$ indicates all variables in $\mathbf{Y}$ except for Y_i and Y_j. And there is no distinction between an edge $(i,j) \in E$ and the edge (j,i). Due to its factorization and Markov properties, storage and computation for a concentration graph are substantially efficient.

Unlike a concentration graph, a covariance graph or relevance network encodes marginal dependence between variables, which is popular in genomics (Butte et al. [8]). By convention, vertices i and j in a covariance graph are joined by a bi-directed edge if Y_i and Y_j are not marginally independent. See, for example, Chaudhuri et al. [9]. That is, $(i,j) \notin E$ when $Y_i \perp\!\!\!\perp Y_j$. Although the edge $(i,j) \in E$ and the edge (j,i) are equivalent in a covariance graph, the large number of arrowheads inside the graph increases computational complexity and burden (Drton and Richardson [10]).

1.2 Gaussian Graphical Models

Among various parametric graphical models, the Gaussian graphical model is the most popular one, due to its mathematical simplicity and the central limit theorem (Uhler [11]). It can be applied in many fields, ranging from machine learning to computational biology and finance.

For a p-dimensional random vector $\mathbf{Y} \in \mathbb{R}^p$ that follows a multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$, it has the joint density function

$$f_{\boldsymbol{\mu},\Sigma}(\boldsymbol{y}) = \frac{1}{(2\pi)^{\frac{p}{2}}(\det\Sigma)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right\}.$$

Let $\Omega = \Sigma^{-1}$ be the concentration matrix. Given an independent and identically distributed (i.i.d.) sample of $\mathbf{Y}$ with size n, i.e., $\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)}, \ldots, \boldsymbol{Y}^{(n)}$, we define the

sample mean $\bar{\boldsymbol{Y}} = \sum_{i=1}^{n} \boldsymbol{Y}^{(i)}/n$ and the empirical covariance matrix

$$S = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{Y}^{(i)} - \bar{\boldsymbol{Y}})(\boldsymbol{Y}^{(i)} - \bar{\boldsymbol{Y}})^{\mathrm{T}}. \tag{1.1}$$

Without loss of generality, assuming $\boldsymbol{\mu} = \mathbf{0}$, the negative log-likelihood function for $\Sigma = \{\Sigma_{i,j}\} \in \mathbb{R}^{p\times p}$, up to a constant, can be written as

$$\ell(\Sigma) = \frac{n}{2}\log\det\Sigma + \frac{n}{2}\mathrm{tr}(\Sigma^{-1}S). \tag{1.2}$$

We can also rewrite $\ell(\Sigma)$ in terms of the concentration matrix $\Omega = \{\Omega_{i,j}\} \in \mathbb{R}^{p\times p}$ as

$$\ell(\Omega) = -\frac{n}{2}\log\det\Omega + \frac{n}{2}\mathrm{tr}(\Omega S). \tag{1.3}$$

Under the normality assumption, in a concentration graph, $\Omega_{i,j} = 0$ if and only if $(i,j) \notin E$ (Lauritzen [4], Edwards [12] and Whittaker [13]). With contemporary data, a sparse concentration matrix or covariance matrix is usually assumed in high-dimensional settings (Johnstone [14] and Rothman et al. [15]). To achieve a sparse graph structure, Friedman et al. [5], Rothman et al. [6], and Yuan and Lin [7] minimize (1.3) with a lasso type penalty. That is to seek the solution to

$$\min_{\Omega\succ 0}\Big\{ -\log\det\Omega + \mathrm{tr}(\Omega S) + \lambda\|P * \Omega\|_1\Big\}, \tag{1.4}$$

where $\|\cdot\|_1$ is the ℓ_1-norm of the input matrix, i.e., $\|A\|_1 = \sum_{i=1}^{p}\sum_{j=1}^{q}|A_{i,j}|$ for any matrix $A = \{A_{i,j}\} \in \mathbb{R}^{p\times q}$, $*$ represents element-wise multiplication, and λ is a tuning parameter that controls the sparsity of Ω: the larger the value of λ, the sparser the estimation of Ω. P is a matrix with zeros on the diagonal to avoid the shrinkage of the diagonal elements of Ω. The minimization (1.4) is a convex optimization problem of Ω.

Yuan and Lin [7] use the interior point algorithm for the maxdet problem in Vandenberghe et al. [16], while Friedman et al. [5] transform the problem into a lasso-type regression (Tibshirani [17]) and propose a faster block coordinate descent algorithm. Rothman et al. [6] develop a Cholesky-based iterative algorithm for solving (1.4) which may contain more complicated penalties, such as a bridge penalty (Fu [18]) or a SCAD penalty (Fan and Li [19]). Additionally, Cai et al. [20] introduce the constrained ℓ_1-minimization for inverse matrix estimation (CLIME) with the following optimization problem:

$$\min \ \|\Omega\|_1 \qquad \text{subject to: } |S\Omega - I|_\infty \leq \lambda_n, \qquad \Omega \in \mathbb{R}^{p\times p}, \tag{1.5}$$

where $|A|_\infty = \max_{1\leq i\leq p, 1\leq j\leq q} |A_{i,j}|$ for any matrix $A = \{A_{i,j}\} \in \mathbb{R}^{p\times q}$, and λ_n is the tuning parameter. They further decompose the problem (1.5) into p vector minimization problems and solve these column problems as linear programs with the primal-dual interior method approach in Boyd et al. [21]. The solution to (1.5) is not symmetric in general, and additional symmetrization should be applied.

1.3 Covariance Estimation

In a covariance graph, $\Sigma_{i,j} = 0$ if and only if $(i, j) \notin E$. And covariance graphs in the Gaussian setting are also studied by some researchers. The function $\ell(\Sigma)$ in (1.2) is non-convex in Σ, which makes the minimization problem more challenging. Chaudhuri et al. [9] develop an iterative conditional fitting algorithm, using simple least squares computations. Drton and Richardson [10] further transform the Gaussian covariance graph into a minimally oriented graph to avoid unnecessary computations in Chaudhuri et al. [9]. To accommodating the sparsity of the covariance matrix in high-dimensional

settings, Bien and Tibshirani [22] propose a majorize-minimize approach to solve

$$\min_{\Sigma \succ 0} \Big\{ \log\det\Sigma + \mathrm{tr}(\Sigma^{-1}S) + \lambda \|P * \Sigma\|_1 \Big\}. \tag{1.6}$$

Wang [23] derives a block-wise coordinate descent algorithm to find the minimizer of (1.6), which is analogous to the method in Friedman et al. [5].

However, the penalized likelihood optimization in (1.6) is non-convex and computationally challenging. Thresholding methods for sparse estimation of covariance matrices have also been developed for recovering the covariance graphs, in which the normality assumption of the data is not required (see, for example, Bickel and Levina [24, 25]). Bickel and Levina [24] propose to taper the empirical sample covariance matrix S in (1.1), i.e.,

$$B_k(S) = \{S_{i,j}\mathbf{1}(|i-j| \le k)\} \in \mathbb{R}^{p\times p},$$

where k is the banding (tuning) parameter. Bickel and Levina [25] and Rothman et al. [26] later recommend the soft-thresholding operator,

$$\mathcal{S}_b(S) = \{\mathrm{sign}(S_{i,j})\max(|S_{i,j}| - b, 0)\} \in \mathbb{R}^{p\times p},$$

where b is the universal thresholding (tuning) parameter. Cai and Liu [27] and Cai and Yuan [28] propose the adaptive thresholding procedure for sparse covariance matrix estimation, which are adaptive to the variability of individual entries of the covariance matrix. They also prove that the adaptive thresholding estimators achieve the optimal rate of convergence over a large class of sparse covariance matrices under the spectral norm, while the commonly used universal thresholding estimators are sub-optimal over the same parameter spaces.

The thresholding methods do not guarantee the positive definiteness of the covariance matrix estimator. Several researchers consider the positive definite constraint for the thresholding methods. To obtain a positive definite covariance matrix estimator, Xue et al. [29] suggest the alternating direction method of multipliers (ADMM) to solve the following optimization problem,

$$\min_{\Sigma \succeq \delta I_p} \frac{1}{2}\|\Sigma - S\|_F^2 + \lambda|\Sigma|_1, \tag{1.7}$$

where S is the empirical sample covariance matrix defined in (1.1), δ is a small positive number, $\|\cdot\|_F$ is the Frobenius norm, and $|\cdot|_1$ is the ℓ_1-norm of the off-diagonal elements of the input matrix. For a matrix $A \in \mathbb{R}^{p\times p}$, $\|A\|_F = \sqrt{\sum_{i=1}^p \sum_{j=1}^p A_{i,j}^2}$, and $|A|_1 = \sum_{i\neq j} |A_{i,j}|$. Similarly, Rothman [30] considers a slightly perturbed version of (1.7) by adding a log-determinant barrier function, i.e.,

$$\min_{\Sigma \succ 0} \frac{1}{2}\|\Sigma - S\|_F^2 - \tau \log\det\Sigma + \lambda|\Sigma|_1, \tag{1.8}$$

where the barrier parameter τ is a fixed small positive constant. The optimization algorithm for (1.8) is similar to the graphical lasso (Glasso) algorithm (Friedman et al. [5]). Cui et al. [31] follow Xue et al. [29] and consider a positive definite correlation matrix estimator with an adaptive ℓ_1-penalty over the off-diagonal elements of the correlation matrix.

1.4 Graphical Vector Autoregressive (VAR) Models

The analysis of time series presents distinct challenges in statistical modeling and inference – most notably, the issue of the temporal correlation resulting from the sampling of points in close temporal proximity (Shumway and Stoffer [32]). The introduction of

the temporal correlation may be generated through lagged linear relations among the observed data. The vector autoregression (VAR) model provides a classical framework to model the lagged linear relationship among multivariate time series, which has gained widespread popularity across various fields, including economics, finance, psychology, and clinical research (e.g., Bringmann et al. [33], Sims [34], Stock and Watson [35], and Wild et al. [36]). The VAR modeling technique was well described and studied in Shumway and Stoffer [32], Hamilton [37] , Lütkepohl [38], Zivot and Wang [39], etc. Let $\boldsymbol{Y}_t = (Y_{t1}, \ldots, Y_{tk}, \ldots, Y_{tp})^{\mathrm{T}}$ be the observation of the p-variate random variable $\mathbf{Y} \in \mathbb{R}^p$ at time t, $t \in \mathbb{Z}$. The standard VAR model of order p, denoted by VAR(p) is given by

$$\boldsymbol{Y}_t = \boldsymbol{\beta}_1^{\mathrm{T}} \boldsymbol{Y}_{t-1} + \boldsymbol{\beta}_2^{\mathrm{T}} \boldsymbol{Y}_{t-2} + \cdots + \boldsymbol{\beta}_p^{\mathrm{T}} \boldsymbol{Y}_{t-p} + \boldsymbol{\varepsilon}_t, \tag{1.9}$$

where $\boldsymbol{\beta}_i$'s are $p \times p$ coefficient matrices, and $\boldsymbol{\varepsilon}_t$ is a p-variate white noise process with mean $\mathbf{0}$ and covariance matrix $\Sigma_\varepsilon = \Omega_\varepsilon^{-1}$.

The graphical vector autoregression (graphical VAR) model combines the principles of the VAR model with graphical models, which has been extensively applied in a wide range of fields, including economics, finance, environmental studies, neuroscience, psychology, epidemiology (e.g., Wild et al. [36], Ahelegbey et al. [40], Barnett and Seth [41] and Eichler [42]). The graphical VAR model with (1.9) allows for the analysis of multivariate time series data with a focus on uncovering the directional relationships and conditional dependencies among multiple variables over time. We further assume that $\boldsymbol{\varepsilon}_t$, $t \in \mathbb{Z}$, are i.i.d. normal. Then the relationships among variables in $\mathbf{Y}$ can be represented by the graphs generated from $\boldsymbol{\beta}_i$'s and Ω_ε, simultaneously. In these graphs, the non-zero elements in $\boldsymbol{\beta}_i$'s correspond to the edges in directed graphs, which indicate possible Granger-causal relationships among variables in $\mathbf{Y}$, while the non-zero elements in Ω_ε reveal the unconditional contemporaneous dependencies among variables in $\mathbf{Y}$ (Eichler

[43]).

For illustration, we consider the following four-dimensional VAR(1) model based on the electronic diary data of obese patients with binge eating disorder (BED) in Wild et al. [36],

$$\boldsymbol{Y}_t = \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{Y}_{t-1} + \boldsymbol{\varepsilon}_t,$$

where $\boldsymbol{Y}_t = (\mathtt{eat}_t, \mathtt{dep}_t, \mathtt{anx}_t, \mathtt{ctl}_t)^{\mathrm{T}}$ represents the observation of eating behaviour, depression, anxiety and eating control at time t, and $\boldsymbol{\varepsilon}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \Omega_{\varepsilon}^{-1})$, $t \in \mathbb{Z}$. We also have the following parameter matrices,

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_{11} & 0 & 0 & \beta_{14} \\ \beta_{21} & \beta_{22} & \beta_{23} & 0 \\ 0 & 0 & \beta_{33} & \beta_{34} \\ \beta_{41} & \beta_{42} & 0 & \beta_{44} \end{pmatrix} \quad \text{and} \quad \Omega_{\varepsilon} = \begin{pmatrix} \Omega_{11} & \Omega_{12} & 0 & \Omega_{14} \\ \Omega_{21} & \Omega_{22} & \Omega_{23} & \Omega_{24} \\ 0 & \Omega_{32} & \Omega_{33} & 0 \\ \Omega_{41} & \Omega_{42} & 0 & \Omega_{44} \end{pmatrix}. \tag{1.10}$$

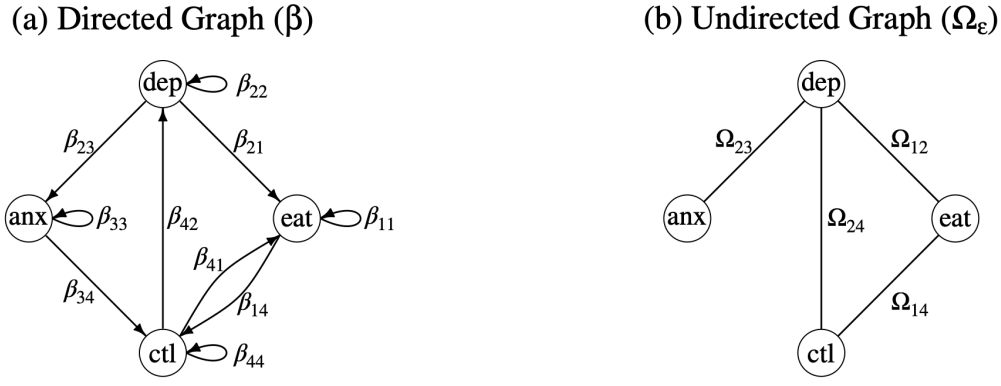

Figure 1.1: Two graphs obtained from $\boldsymbol{\beta}$ (left panel) and Ω_{ε} (right panel) with a four-dimensional VAR(1) process that satisfies the parameter constraints in (1.10).

With the parameter constraints in (1.10), we can generate two corresponding graphs in Figure 1.1. For example, the directed edge between `eat` and `dep` encoded by β_{21} means that depression Granger-causes eating behaviour for patients with BED. We also notice the self-loops for `eat`, `dep`, `anx` and `ctl` in Figure 1.1, which represent the autocorrelations in eating behaviour, depression, anxiety and eating control. Usually, these self-loops are omitted, since they do not play any role in the graphical analysis of Granger-causal relationships (Wild et al. [36]).

1.5 Multilevel Gaussian Graphical Models

Clustered data arise in many areas, such as economics, education, epidemiology, medicine, psychology, and social science. Most existing methods for learning the structure of graphical models assume independent samples. It is well-known that ignoring the correlation between observations may lead to flawed insights (Bae et al. [44]). In addition, researchers are often interested in correlations at different levels and comparing them (Epskamp et al. [45] and Ostroff [46]). For example, in psychology, variations in the measurements between subjects are studied using a nomothetic approach, whereas variations within a subject are examined through an idiographic approach (Hamaker [47]). Network psychometrics has emerged as useful additions to the psychometric toolbox in recent years (Bringmann et al. [33]). Correlations at group and sub-group levels may be different, an issue termed ecological fallacy or Simpson's paradox (Epskamp et al. [45], Hamaker [47], Freedman [48], and Piantadosi [49]). For example, typing faster than one's average speed tends to result in more errors (within a subject), while individuals who generally type quickly often make fewer spelling errors (between subjects) (Epskamp et al. [45] and Hamaker [47]).

This thesis uses the terms "subject" and "individual" to represent a generic experi-

mental unit and each observation for simplicity. Consider p random variables $Y_1, \ldots, Y_p$ with observations $\boldsymbol{Y}_{ij} = (Y_{ij1}, \cdots, Y_{ijp})^{\mathrm{T}}$ at group level i and individual level j. We are interested in correlations or conditional correlations at the individual level (within-subject) as well as correlations or conditional correlations at the group level (between-subject). We assume that the observed groups are a random sample from a population of all group levels. The between-subject correlations measure the correlations among variables at the group level $\mathbb{E}(\boldsymbol{Y}_{ij} \mid i)$ while within-subject correlations measure the correlations among variables at the individual level $\boldsymbol{Y}_{ij} - \mathbb{E}(\boldsymbol{Y}_{ij} \mid i)$. There are different ways to define within and between group correlations (see, for example, Bland and Altman [50, 51]). We define them using a simple multivariate linear mixed effects model.

For simplicity, we will first consider a multivariate one-way random effect model. Assume that $\boldsymbol{Y}_{ij}$ equals to a random mean vector for subject i, $\boldsymbol{b}_i$, plus random error for individual j, $\boldsymbol{\varepsilon}_{ij}$:

$$\boldsymbol{Y}_{ij} = \boldsymbol{b}_i + \boldsymbol{\varepsilon}_{ij},\ j = 1, \ldots, n_i;\ i = 1, \ldots, m, \tag{1.11}$$

where $\boldsymbol{b}_i = (b_{i1}, \cdots, b_{ip})^{\mathrm{T}}$ are i.i.d. random vectors with mean $\mathbf{0}$ and covariance matrix Σ_b, and $\boldsymbol{\varepsilon}_{ij} = (\varepsilon_{ij1}, \ldots, \varepsilon_{ijp})^{\mathrm{T}}$ are i.i.d. random vectors with mean $\mathbf{0}$ and covariance matrix Σ_ε, and $\boldsymbol{b}_i$ and $\boldsymbol{\varepsilon}_{ij}$ are mutually independent. For now, we assume that the observations are centered such that $\mathbb{E}(\boldsymbol{Y}_{ij}) = \mathbf{0}$ (Yuan and Lin [7]). Denote $\sigma^2_{b,k} = \mathbb{V}\mathrm{ar}(b_{ik})$ and $\sigma^2_{\varepsilon,k} = \mathbb{V}\mathrm{ar}(\varepsilon_{ijk})$ for $k = 1, \cdots, p$; and $\rho_{k_1,k_2} = \mathbb{C}\mathrm{orr}(Y_{ijk_1}, Y_{ijk_2})$, $\rho_{b,k_1,k_2} = \mathbb{C}\mathrm{orr}(b_{ik_1}, b_{ik_2})$, and $\rho_{\varepsilon,k_1,k_2} = \mathbb{C}\mathrm{orr}(\varepsilon_{ijk_1}, \varepsilon_{ijk_2})$ for $k_1, k_2 = 1, \cdots, p$ and $k_1 \neq k_2$. Then, $(\Sigma_b)_{k_1,k_2} = \rho_{b,k_1,k_2}\sigma_{b,k_1}\sigma_{b,k_2}$ and $(\Sigma_b)_{k_1,k_2} = \rho_{\varepsilon,k_1,k_2}\sigma_{\varepsilon,k_1}\sigma_{\varepsilon,k_2}$. Note that ρ_{b,k_1,k_2} and $\rho_{\varepsilon,k_1,k_2}$ represent between-subject and within-subject correlations. These definitions of between-subject and within-subject correlations are in a spirit similar to those in Ostroff [46] and Piantadosi [49] where only the sample version of these quantities was defined.

Furthermore, we have

$$\rho_{k_1,k_2} = \frac{\sigma_{b,k_1}\sigma_{b,k_2}}{\sqrt{(\sigma^2_{b,k_1}+\sigma^2_{\varepsilon,k_1})(\sigma^2_{b,k_2}+\sigma^2_{\varepsilon,k_2})}}\rho_{b,k_1,k_2} + \frac{\sigma_{\epsilon,k_1}\sigma_{\epsilon,k_2}}{\sqrt{(\sigma^2_{b,k_1}+\sigma^2_{\varepsilon,k_1})(\sigma^2_{b,k_2}+\sigma^2_{\varepsilon,k_2})}}\rho_{\varepsilon,k_1,k_2}$$

which can be regarded as the population version of equation (3) in Piantadosi [49].

Based on the estimators of between-subject and with-subject covariance matrices with Model (1.11), we could construct two separate covariance graphs as in Bien and Tibshirani [22]. The edge between Y_{k_1} and Y_{k_2} exists at the group level (between-subject covariance graph) if and only if $(\Sigma_b)_{k_1,k_2} \neq 0$, and the edge between Y_{k_1} and Y_{k_2} exists at the individual level (within-subject covariance graph) if and only if $(\Sigma_\varepsilon)_{k_1,k_2} \neq 0$. When an edge is present, we will use blue and red to represent positive and negative correlations (Epskamp et al. [45]), which will allow us to detect potential ecological fallacy.

In addition to the multivariate one-way random effects model (1.11), for multivariate time series from multiple subjects, we will consider the following graphical VAR model of order 1, denoted as GVAR(1),

$$\boldsymbol{Y}_{it} = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Y}_{i(t-1)} + \boldsymbol{b}_i + \boldsymbol{\varepsilon}_{it},\ t = 1,\ldots,n_i;\ i = 1,\ldots,m, \tag{1.12}$$

where $\boldsymbol{Y}_{i(t-1)}$ is the design matrix, and $\boldsymbol{b}_i$ and $\boldsymbol{\varepsilon}_{it}$ follow the same assumptions as those in (1.11). In addition, we assume that $\boldsymbol{b}_i$ and $\boldsymbol{\varepsilon}_{it}$ follow multivariate Gaussian distribution. To make model (1.12) consistent to the notation in time series analysis, we use index t to represent the observation in each subject i. Obviously, model (1.11) is a special case of model (1.12) with $\boldsymbol{\beta} = \boldsymbol{0}$. Epskamp et al. [45] consider a similar model as model (1.12),

$$\boldsymbol{Y}_{it} = \boldsymbol{\beta}_i^{\mathrm{T}}\boldsymbol{Y}_{i(t-1)} + (I - \boldsymbol{\beta}_i^{\mathrm{T}})\boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_{it},\ t = 1,\ldots,n_i;\ i = 1,\ldots,m, \tag{1.13}$$

where $\boldsymbol{\mu}_i = (\mu_{i1}, \cdots, \mu_{ip})^{\mathrm{T}}$ are i.i.d. Gaussian random vectors with mean $\mathbf{0}$ and covariance matrix $\Sigma_\mu = \Omega_\mu^{-1}$, and $\boldsymbol{\varepsilon}_{ij} = (\varepsilon_{ij1}, \ldots, \varepsilon_{ijp})^{\mathrm{T}}$ are independent Gaussian random vectors with mean $\mathbf{0}$ and covariance matrix $\Sigma_{\varepsilon i} = \Omega_{\varepsilon i}^{-1}$, and $\boldsymbol{\mu}_i$ and $\boldsymbol{\varepsilon}_{ij}$ are mutually independent. In addition, they also assume that both $\boldsymbol{\beta}_i$ and $\Omega_{\varepsilon i}$ are random matrices with $\mathbb{E}(\boldsymbol{\beta}_i) = \boldsymbol{\beta}$ and $\mathbb{E}(\Omega_{\varepsilon i}) = \Omega_\varepsilon$. Model (1.13) is a non-linear mixed effects model due to the product of two random effects $\boldsymbol{\beta}_i$ and $\boldsymbol{\mu}_i$. When $\boldsymbol{\beta}_i$ and $\Omega_{\varepsilon i}$ are fixed matrices, i.e., $\boldsymbol{\beta}_i = \boldsymbol{\beta}$ and $\Omega_{\varepsilon i} = \Omega_\varepsilon$ for $i = 1, \ldots, m$, model (1.12) is equivalent to model (1.13), and $(I - \boldsymbol{\beta}^{\mathrm{T}})\mu_i$ in (1.13) plays the same role as $\boldsymbol{b}_i$ in (1.12).

A graphical VAR model in (1.12) has three graphical structures: the temporal network decided by $\boldsymbol{\beta}$, the contemporaneous network decided by Ω_ε, and the between-subjects network decided by Ω_b. As illustrated in Epskamp et al. [45], the graphical models at different levels provide a powerful addition to the exploratory toolbox in many research areas. We are interested in recovering these three graphs simultaneously. However, existing estimation methods reviewed in Epskamp et al. [45] have the following limitations: (a) some procedures are ad hoc since they contain several steps such that existing methods can be used in each step; (b) some procedures use sample means from each subject for the analysis of conditional between-subject dependence structure which could lead to an erroneous structure; and (c) the computation is only feasible for small data sets and up to eight variables.

1.6 Dissertation Outline

The goal of this dissertation is to develop new methods for learning dependence and conditional dependence structures at different levels for clustered data. We focus on clustered data with multilevel dependencies and conditional dependencies and construct multilevel graphs under the sparsity assumptions simultaneously. The rest of this disser-

tation is organized in two main parts as follows.

In Part I, we aim at estimating within-subject and between-subject covariance matrices simultaneously in Chapter 2, 3 and 4. Chapter 2 discusses sparse covariance graphs at within-subject and between-subject levels. We first introduce the sample estimates for between-subject and within-subject covariance matrices. Based on these sample estimates, we propose sparse estimates that are guaranteed to be positive-definite. Our proposed estimators are defined as solutions to convex optimization problems, which can be solved efficiently using an ADMM algorithm. The statistical properties of our proposed estimators are presented in Chapter 3. Chapter 3 also includes the comparison between our proposed between-subject covariance estimator and the MANOVA-type estimator. Chapter 4 investigates the numerical performance of our proposed two covariance estimators with comprehensive simulations and an application to a dataset collected from end-stage renal disease (ESRD) patients.

In Part II, we focus on estimating the fixed effect coefficient matrix, within-subject, and between-subject precision matrices based on a GVAR(1) model. Chapter 5 introduces a two-stage procedure for recovering sparse fixed effect coefficient, within-subject and between-subject precision matrices. In the first stage, we iteratively estimate the fixed effect coefficient and within-subject precision matrix with efficient moment methods based on the group-centered data. Subsequently, the sparse between-subject precision matrix is learned by the CLIME method in the second stage. The corresponding Bayesian information criteria (BIC) are also developed for tuning parameter selection in both stages. A comprehensive numerical study is conducted in Chapter 6, which also includes a real data example. Future studies are discussed in Chapter 7.

Chapter 2

Sparse Estimation of Multilevel Covariances with Repeated Measurements

2.1 Introduction

Understanding the covariance structure among random variables is one of the most fundamental tasks in statistics with applications in a wide range of fields, including economics, biology, and biomedical sciences (Fan et al. [2] and Bickel and Levina [25]). Various sparse estimation methods have been proposed in high-dimensional settings. However, virtually all current methods require the critical assumption of independent samples, which could be violated in many applications. This paper considers a special correlated data structure where observations are repeated measurements.

In many fields, such as medicine, psychology, and neuroscience, random variables of interest are often measured repeatedly across different subjects, which leads to dependence among observations within each subject. For example, vital signs such as pulse

and blood pressure are usually measured in multiple physical exams for each subject, and these measurements from the same subject are correlated. Conclusions drawn from ignoring such dependence structures among observations may be practically misguided or even erroneous (Bae et al. [44]). Therefore, it is important to estimate covariance structures in the presence of dependence due to repeated measurements (Ostroff [46]).

Repeated measurements have an underlying hierarchical structure, and it is of scientific interest to define and estimate covariance structures at each level. In psychology, the nomothetic approach is used to study variations between subjects, and the idiographic approach is used to study variations within a subject (Hamaker [47]). Covariance structures between subjects and within a subject may thus be different. For example, physical activity tends to increase the heart rate of a person (within a subject), while physically active people tend to have a lower average heart rate (between subjects) (Epskamp et al. [45]). This chapter aims to develop new methods to estimate the within-subject and between-subject covariance structures simultaneously.

Recall model (1.11) in Section 1.5, we consider a multivariate one-way random effect model for within-subject and between-subject covariance structures among p random variables:

$$\boldsymbol{Y}_{ij} = \boldsymbol{b}_i + \boldsymbol{\varepsilon}_{ij},\ j = 1, \ldots, n_i;\ i = 1, \ldots, m,$$

where $\boldsymbol{Y}_{ij} = (Y_{ij1}, \ldots, Y_{ijp})^{\mathrm{T}} \in \mathbb{R}^p$ is the j-th (out of n_i) observation of the i-th subject, $\boldsymbol{b}_i = (b_{i1}, \cdots, b_{ip})^{\mathrm{T}} \in \mathbb{R}^p$ are independent and identically distributed random vectors with mean $\mathbf{0}$ and covariance matrix $\Sigma_b \in \mathbb{R}^{p\times p}$, and $\boldsymbol{\varepsilon}_{ij} = (\varepsilon_{ij1}, \ldots, \varepsilon_{ijp})^{\mathrm{T}} \in \mathbb{R}^p$ are independent and identically distributed random vectors with mean $\mathbf{0}$ and covariance matrix $\Sigma_\varepsilon \in \mathbb{R}^{p\times p}$. Additionally, $\boldsymbol{b}_i$ and $\boldsymbol{\varepsilon}_{ij}$ are mutually independent. The between-subject covariance Σ_b measures the covariance structure among variables at the group level $\mathbb{E}(\boldsymbol{Y}_{ij} \mid i)$. On

the other hand, the within-subject covariance Σ_ε characterizes the covariance structure among components in $\boldsymbol{Y}_{ij} - \mathbb{E}(\boldsymbol{Y}_{ij} \mid i)$. Model (1.11) has found wide applications, e.g., in the classical test theory (Algina and Swaminathan [52]), where the observed score is modeled as the summation of the true score (as a latent variable) and a random error.

For the cross-sectional data, which is a special case of (1.11) with $n_i = 1$ for $i = 1, \ldots, m$, it is clear that one can only estimate the overall covariance $\Sigma_b + \Sigma_\varepsilon$, which does not separate the within-subject and between-subject covariance structures. When $n_i \geq 2$ for at least some $i \in \{1, \ldots, m\}$, a common approach is to aggregate data across subjects and obtain $\{\bar{\boldsymbol{Y}}_{1\cdot}, \ldots, \bar{\boldsymbol{Y}}_{m\cdot}\}$, where $\bar{\boldsymbol{Y}}_{i\cdot} = \sum_{j=1}^{n_i} \boldsymbol{Y}_{ij}/n_i$. The sample covariance estimate based on this aggregated data,

$$\overline{\Sigma} = \frac{1}{m-1} \sum_{i=1}^{m} \left(\bar{\boldsymbol{Y}}_{i\cdot} - \frac{1}{m} \sum_{i=1}^{m} \bar{\boldsymbol{Y}}_{i\cdot} \right) \left(\bar{\boldsymbol{Y}}_{i\cdot} - \frac{1}{m} \sum_{i=1}^{m} \bar{\boldsymbol{Y}}_{i\cdot} \right)^{\mathrm{T}}, \tag{2.1}$$

is an unbiased estimate of

$$\mathbb{E}(\overline{\Sigma}) = \Sigma_b + \sum_{i=1}^{m} \frac{1}{m n_i} \Sigma_\varepsilon. \tag{2.2}$$

Consequently, $\overline{\Sigma}$ is a biased estimate of either Σ_ε or Σ_b. Epskamp et al. [45] used (2.2) to estimate the between-subject covariance structure. Statistical inferences based on aggregated data may be misinterpreted (Fisher et al. [53]). In particular, analysis based on aggregated data may result in an issue termed ecological fallacy or Simpson's paradox (Epskamp et al. [45], Hamaker [47], Freedman [48], and Piantadosi [49]).

Furthermore, in high-dimensional settings where p could be much larger than m or N, the sample covariance estimate is no longer positive definite, making it less amenable for interpretation or downstream statistical tasks. To our knowledge, there is no research on covariance structure learning for high-dimensional repeated measures data. We fill in this

methodological gap in this thesis by emphasizing the importance of treating the target of estimation separately and proposing two new sparse positive definite estimators, one for the within-subject covariance Σ_ε and one for the between-subject covariance matrix Σ_b. We demonstrate the benefit of our proposed estimators by comparing both theoretically and numerically with other estimators that have been previously studied in different settings.

2.2 Sparse Estimation of Within-subject and Between-subject Covariance Matrices

Most recent approaches to estimating a large covariance matrix involve regularized estimation based on an unbiased estimate of the target covariance matrix. In a setting with independent and identically distributed samples, it is straightforward to use the sample covariance matrix as an unbiased estimate, and methods in the literature differ in various approaches to imposing regularization. Specifically, methods based on thresholding the sample covariance matrix have been well-studied (Bickel and Levina [24, 25], and Cai and Yuan [28]), and further improvements have been developed to ensure positive definiteness in the resulting estimates (Rothman et al. [26], Xue et al. [29, 30], and Cui et al. [31]). Bien and Tibshirani [22] proposed a penalized likelihood procedure for estimating a sparse covariance matrix, which could be computationally intensive due to the non-convexity of the likelihood in the covariance matrix.

There are several unbiased estimates of the two covariance matrices in the model

(1.11). We first consider the following unbiased estimates:

$$\widehat{\Sigma}_\varepsilon = (\sum_{i=1}^{m} n_i - m)^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\boldsymbol{Y}_{ij} - \bar{\boldsymbol{Y}}_{i\cdot})(\boldsymbol{Y}_{ij} - \bar{\boldsymbol{Y}}_{i\cdot})^{\mathrm{T}}, \tag{2.3}$$

$$\widehat{\Sigma}_b = \overline{\Sigma} - \sum_{i=1}^{m} (mn_i)^{-1} \widehat{\Sigma}_\varepsilon. \tag{2.4}$$

The sample estimate $\widehat{\Sigma}_\varepsilon$ is an unbiased estimate of Σ_ε (Rao and Heckler [54]). From (2.1), $\widehat{\Sigma}_b$ is an unbiased estimate of Σ_b, and is a multivariate extension of the unweighted sum-of-squares estimator in Rao and Sylvestre [55]. We will consider another commonly used unbiased estimate of Σ_b in Section 3.5 and demonstrate that it is suboptimal for estimation.

Note that $\widehat{\Sigma}_\varepsilon$ in (2.3) may be singular in high-dimensional settings where $p > m$, and $\widehat{\Sigma}_b$ may not be positive semi-definite for any dimensions. In particular, the diagonal elements in $\widehat{\Sigma}_b$ could be negative. To derive sparse and positive-definite estimates of Σ_ε and Σ_b, we follow Xue et al. [29] and consider the following optimization problem for estimating a generic covariance matrix Σ with input matrix D,

$$\min_{\Sigma \succeq \delta I_p} \frac{1}{2} \|\Sigma - D\|_F^2 + \lambda |\Sigma|_1, \tag{2.5}$$

where $\|\cdot\|_F$ is the Frobenius norm and $|\cdot|_1$ is the ℓ_1-norm of the off-diagonal elements of the input matrix. The constraint $\Sigma \succeq \delta I_p$ imposes positive semi-definiteness on $\Sigma - \delta I_p$, which results in a positive definite solution to (2.5) with a small value of $\delta > 0$. This positive definiteness constraint is essential to provide a usable and accurate estimate. A solution to (2.5) is simultaneously sparse, positive definite, and close to the input matrix D, which is usually set as an unbiased sample estimate. Let $\widehat{\Sigma}_\varepsilon^+$ be the sparse and positive definite estimate of Σ_ε as the solution to (2.5) with $D = \widehat{\Sigma}_\varepsilon$ and $\lambda = \lambda_\varepsilon$, and $\widehat{\Sigma}_b^+$

be the sparse and positive definite estimates of Σ_b as the solution to (2.5) with $D = \widehat{\Sigma}_b$ and $\lambda = \lambda_b$. We study $\widehat{\Sigma}_\varepsilon^+$ and $\widehat{\Sigma}_b^+$ both theoretically and numerically. In addition, to illustrate the suboptimality of using group aggregation in estimating either covariance matrix, we further study $\overline{\Sigma}^+$, which is defined as the solution to (2.5) with $D = \overline{\Sigma}$ and $\lambda = \lambda_0$. The theoretical tuning parameter values λ_ε, λ_b, and λ_0 are discussed in Chapter 3.

The convex optimization problem (2.5) can be written equivalently as

$$\min_{\Sigma,\Theta} \left\{ \frac{1}{2}\|\Sigma - D\|_F^2 + \lambda|\Theta|_1 : \ \Sigma = \Theta, \ \Sigma \succeq \delta I_p \right\}, \tag{2.6}$$

which we solve using the alternating direction method of multipliers (Boyd et al. [21]). Specifically, the algorithm iteratively minimizes the following augmented Lagrangian

$$L(\Sigma, \Theta; \Lambda) = \frac{1}{2}\|\Sigma - D\|_F^2 + \lambda|\Theta|_1 + \langle \Lambda, \Sigma - \Theta \rangle + \frac{\rho}{2}\|\Sigma - \Theta\|_F^2,$$

over Σ, Θ, and the dual variable Λ using the following updates until convergence:

$$\Sigma \leftarrow \operatorname{argmin}_{\Sigma \succeq \delta I_p} L(\Sigma, \Theta; \Lambda) = \tfrac{1}{1+\rho}(D + \rho\Theta - \Lambda, \delta)_+, \tag{2.7}$$

$$\Theta \leftarrow \operatorname{argmin}_{\Theta} L(\Sigma, \Theta; \Lambda) = \mathcal{S}_{\lambda/\rho}\Big(\Sigma + \tfrac{1}{\rho}\Lambda\Big), \tag{2.8}$$

$$\Lambda \leftarrow \Lambda + \rho(\Sigma - \Theta).$$

The update in (2.7) computes the projection onto a positive semi-definite cone, where $(A, \delta)_+ = \sum_{j=1}^p \max(\lambda_j, \delta) v_j v_j^{\mathrm{T}}$ for a generic matrix $A \in \mathbb{R}^{p\times p}$ with the eigendecomposition $A = \sum_{j=1}^p \lambda_j v_j v_j^{\mathrm{T}}$. The update in (2.8) evaluates element-wise soft-thresholding operators, where $\{\mathcal{S}_b(A)\}_{j,k} = \operatorname{sign}(A_{j,k}) \max(|A_{j,k}| - b, 0)$ for any matrix A and scalar $b \geq 0$. We follow Boyd et al. [21] for practical considerations in this algorithm, including

the initial values, the stopping criterion, and the updating strategy for the optimization parameter ρ, and refer to Section 2.3 for further implementation details. This algorithm has been widely used in the literature on covariance estimation, (e.g., Bien and Tibshirani[22] and Xue et al. [29]) with well-established convergence analysis (Nishihara et al. [56]). The computational complexity of each update is dominated by the eigendecomposition in (2.7), which requires $O(p^3)$ operations. An approximate alternating direction method of multipliers (Rontsis et al. [57]) could be used to improve the computational complexity by avoiding repeated eigendecompositions.

2.3 Further Details on Optimization Algorithm Implementation

The complete algorithm solving the convex optimization problem (2.6) in Section 2.2 is summarized in Algorithm 1.

Algorithm 1 Alternating direction method of multipliers for solving (2.6) in Section 2.2.

Require: δ, λ, $\rho^{(0)}$, D, $\Sigma^{(0)}$, $\Theta^{(0)}$, $\Lambda^{(0)}$, and $l = 0$.

1: Repeat
2: $\quad \Sigma^{(l+1)} \leftarrow \frac{1}{1+\rho^{(l)}} \left(D + \rho\Theta^{(l)} - \Lambda^{(l)}, \delta\right)_+$
3: $\quad \Theta^{(l+1)} \leftarrow \mathcal{S}_{\lambda/\rho^{(l)}} \left(\Sigma^{(l+1)} + \frac{1}{\rho^{(l)}}\Lambda^{(l)}\right)$
4: $\quad \Lambda^{(l+1)} \leftarrow \Lambda^{(l)} + \rho^{(l)} \left(\Sigma^{(l+1)} - \Theta^{(l+1)}\right)$
5: $\quad$ Update $\rho^{(l+1)}$ based on equation (3.13) in Boyd et al. [21]
6: Until convergence

A reasonable stopping criterion suggested by Boyd et al. [21] is

$$\|\Sigma^{(l+1)} - \Theta^{(l+1)}\|_F \leq \epsilon^{\text{pri}} \qquad \text{and} \qquad \|\rho(\Theta^{(l+1)} - \Theta^{(l)})\|_F \leq \epsilon^{\text{dual}}.$$

where ϵ^{pri} and ϵ^{dual} are positive feasibility tolerances for the primal and dual feasibility

conditions, which are controlled by an absolute criterion ϵ^{abs} and a relative criterion ϵ^{rel}:

$$\begin{aligned}\epsilon^{\text{pri}} &= p\epsilon^{\text{abs}} + \epsilon^{\text{rel}} \max\{\|\Sigma^{(l+1)}\|_F, \|\Theta^{(l+1)}\|_F\},\\ \epsilon^{\text{dual}} &= p\epsilon^{\text{abs}} + \epsilon^{\text{rel}}\|\Lambda^{(l+1)}\|_F,\end{aligned}$$

where $\epsilon^{\text{abs}} > 0$ and $\epsilon^{\text{rel}} > 0$. In the numerical studies, we choose $\epsilon^{\text{abs}} = \epsilon^{\text{rel}} = 10^{-8}$. The choice of ρ can greatly impact the practical convergence of the alternating direction method procedure. To improve the convergence, we adopt an adaptive strategy described in Boyd et al. [21] for varying penalty parameter ρ. In practice, we use the soft-thresholding estimators based on the sample estimates as the initial $(\Sigma^{(0)}, \Theta^{(0)})$. And the initial input for $\Lambda^{(0)}$ is a zero matrix. The initial penalty parameter ρ is 0.1. Without the positive semi-definite constraints of Σ_ε and Σ_b in (2.6), the unconstrained solutions will be $\mathcal{S}_\lambda(\widehat{\Sigma}_\varepsilon)$ and $\mathcal{S}_\lambda(\widehat{\Sigma}_b)$ with $D = \widehat{\Sigma}_\varepsilon$ and $D = \widehat{\Sigma}_b$, respectively. For efficient computation, we always first check the positive semi-definiteness of $\mathcal{S}_\lambda(\widehat{\Sigma}_\varepsilon)$ and $\mathcal{S}_\lambda(\widehat{\Sigma}_b)$. If $\mathcal{S}_\lambda(\widehat{\Sigma}_\varepsilon)$ and $\mathcal{S}_\lambda(\widehat{\Sigma}_b)$ are positive semi-definite, they are the final solutions to (2.6), respectively. Otherwise, we will use Algorithm 1 to solve (2.6).

2.4 Cross-validation Procedure for Tuning Parameters Selection

The main optimization problem (2.6) defines various estimators that we study in this thesis, where λ is the tuning parameter that controls the level of regularization of the sample estimates. We present in this section a cross-validation procedure for selecting the tuning parameter (Bickel and Levina [25], Rothman et al. [26] and Cai and Liu [27]) specifically in the presence of repeated measurements.

For each (of the K) split in a K-fold cross-validation procedure, we randomly partition

the m groups into a set of m_1 groups of training set, i.e., $\mathcal{T}_r = \{\boldsymbol{Y}_{ij} : i \in \mathcal{A}\}$ with $|\mathcal{A}| = m_1$ and a set of $m - m_1$ groups of validation set, i.e., $\mathcal{T}_e = \{\boldsymbol{Y}_{ij} : i \in \mathcal{A}^c\}$ with $|\mathcal{A}^c| = m - m_1$.

Let $\widehat{S}^+\{\lambda, \widehat{S}(\mathcal{T})\}$ denote a generic estimator, which is defined as a solution to the optimization problem (2.6) with the tuning parameter value λ and input sample matrix $\widehat{S}(\mathcal{T})$ evaluated using a dataset $\mathcal{T}$. Specifically, the estimator $\widehat{S}^+\{\lambda, \widehat{S}(\mathcal{T})\}$ could refer to $\widehat{\Sigma}_b^+$, $\widehat{\Sigma}_\varepsilon^+$, $\widetilde{\Sigma}_b^+$, and $\overline{\Sigma}^+$. And $\widehat{S}(\mathcal{T})$ refers to the unbiased estimator $\widehat{\Sigma}_b$, $\widehat{\Sigma}_\varepsilon$, $\widetilde{\Sigma}_b$, and the biased estimator $\overline{\Sigma}$. The cross-validation procedure is presented in the following Algorithm 2 to choose the tuning parameter from a path of candidate tuning parameter values $\{\lambda_1 > \lambda_2 > \ldots > \lambda_L\}$.

Algorithm 2 A K-fold Cross-Validation Procedure

Require: $\{\boldsymbol{Y}_{ij} : 1 \le i \le m, 1 \le j \le n_i\}$ and $\{\lambda_1 > \lambda_2 > \ldots > \lambda_L\}$.

1: **for** $\ell = 1, \ldots, L$ **do**
2: **for** $\nu = 1, \ldots, K$ **do**
3: Divide $\{\boldsymbol{Y}_{ij} : 1 \le i \le m, 1 \le j \le n_i\}$ into training set $\mathcal{T}_r^{(\nu)}$ and validation set $\mathcal{T}_e^{(\nu)}$;
4: Compute the sample covariance matrix $\widehat{S}(\mathcal{T}_e^{(\nu)})$ on the validation set $\mathcal{T}_e^{(\nu)}$;
5: Compute the estimator $\widehat{S}^+\{\lambda_\ell, \widehat{S}(\mathcal{T}_r^{(\nu)})\}$ on the training set $\mathcal{T}_r^{(\nu)}$.
6: **end for**
7: Compute CV estimate of error $E_\ell = \sum_{\nu=1}^K \|\widehat{S}^+\{\lambda_\ell, \widehat{S}(\mathcal{T}_r^{(\nu)})\} - \widehat{S}(\mathcal{T}_e^{(\nu)})\|_F^2 / K$.
8: **end for**
9: Let $\hat{\ell} = \text{argmin}_{\ell=1,\ldots,L} E_\ell$, and return the selected tuning parameter $\lambda_{\hat{\ell}}$.

Chapter 3

Theoretical Properties of Sparse Covariance Estimation with Repeated Measurements

3.1 Notations and Assumptions

In this chapter, we derive the finite-sample estimation error rate of our proposed estimators $\widehat{\Sigma}_{\varepsilon}^{+}$ (in Section 3.3) and $\widehat{\Sigma}_{b}^{+}$ (in Section 3.4), and establish their asymptotic consistency. In comparison, we further establish that $\overline{\Sigma}^{+}$ is inconsistent in estimating Σ_b due to a non-vanishing bias even with an infinite number of subjects, thus illustrating the pitfall of the sample estimator (2.1) based on the aggregated data.

We observe $\boldsymbol{Y}_{ij} \in \mathbb{R}^p$, which is the j-th repeated measurement of the i-th subject for $j = 1, \ldots, n_i$ and $i = 1, \ldots, m$, following the model (1.11), where $\boldsymbol{\varepsilon}_{ij}$ and $\boldsymbol{b}_i$ are p-dimensional sub-Gaussian random vectors with the true within and between covariance $\mathbb{V}\mathrm{ar}(\boldsymbol{\varepsilon}_{ij}) = \Sigma_{\varepsilon}^{0}$ and $\mathbb{V}\mathrm{ar}(\boldsymbol{b}_i) = \Sigma_{b}^{0}$ respectively, and $\boldsymbol{b}_i$ and $\boldsymbol{\varepsilon}_{ij}$ are mutually independent. Let $N = \sum_{i=1}^{m} n_i$ be the total number of observations. We consider the following class of

sparse covariance matrices:

$$\mathcal{U}(M,s) = \Big\{\Sigma \in \mathbb{S}_{++}^{p\times p} : \max_k \Sigma_{k,k} \le M, \max_k \sum_{\ell=1}^{p} 1(\Sigma_{k,\ell} \neq 0) \le s\Big\},$$

where $\mathbb{S}_{++}^{p\times p}$ is the set of all p-by-p symmetric positive definite matrices, and $\Sigma_{k,\ell}$ is the (k,ℓ)-th entry of Σ. A matrix in $\mathcal{U}(M,s)$ has diagonals bound M and maximum row-wise (and by symmetry, column-wise) sparsity level s.

3.2 Lemmas

We begin with several lemmas essential for the proof of the main results. In Lemma 1 and 2, we establish the entry-wise convergence rate for our unpenalized within-subject and between-subject covariance estimator, i.e., $\widehat{\Sigma}_\varepsilon$ and $\widehat{\Sigma}_b$ in (2.3) and (2.4). According to (2.2) in Section 2.1, $\overline{\Sigma}$ in (2.1) is a biased estimate for both within-subject and between-subject covariance matrices. Therefore, we also compare $\overline{\Sigma}$ with Σ_b and Σ_ε in Lemma 3 and 4.

Lemma 1 *Consider the true within-subject covariance Σ_ε^0 with $\max_k(\Sigma_\varepsilon^0)_{k,k} \le M_\varepsilon$. Let $\lambda_\varepsilon = C_1\{N\log p\}^{1/2}/(N-m)$ for a sufficiently large constant C_1. If $\log p \le N$, then the unbiased within-subject sample estimate $\widehat{\Sigma}_\varepsilon$ satisfies*

$$\mathbb{P}\mathrm{r}\left\{\max_{k,l}\left|(\widehat{\Sigma}_\varepsilon - \Sigma_\varepsilon^0)_{k,l}\right| > \lambda_\varepsilon\right\} \le 4p^{-C_2},$$

where $C_2 > 0$ only depends on C_1 and M_ε.

Proof. We first rewrite $\widehat{\Sigma}_\varepsilon$ as follows,

$$\begin{aligned}\widehat{\Sigma}_\varepsilon &= \frac{1}{N-m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\boldsymbol{Y}_{ij}-\bar{\boldsymbol{Y}}_{i\cdot})(\boldsymbol{Y}_{ij}-\bar{\boldsymbol{Y}}_{i\cdot})^{\mathrm{T}} \\ &= \frac{1}{N-m}\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i}\boldsymbol{\varepsilon}_{ij}\boldsymbol{\varepsilon}_{ij}^{\mathrm{T}}-\sum_{i=1}^{m}n_i\bar{\boldsymbol{\varepsilon}}_{i\cdot}\bar{\boldsymbol{\varepsilon}}_{i\cdot}^{\mathrm{T}}\right).\end{aligned}$$

Then,

$$\begin{aligned}(\widehat{\Sigma}_\varepsilon)_{k,l} &= \frac{1}{N-m}\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i}\varepsilon_{ijk}\varepsilon_{ijl}-\sum_{i=1}^{m}n_i\bar{\varepsilon}_{i\cdot k}\bar{\varepsilon}_{i\cdot l}\right) \\ &= \frac{1}{N-m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\varepsilon_{ijk}\varepsilon_{ijl}-\frac{1}{N-m}\sum_{i=1}^{m}n_i\bar{\varepsilon}_{i\cdot k}\bar{\varepsilon}_{i\cdot l} \\ &= \frac{1}{N-m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left\{\varepsilon_{ijk}\varepsilon_{ijl}-(\Sigma_\varepsilon^0)_{k,l}\right\} \\ &\quad -\frac{1}{N-m}\sum_{i=1}^{m}\left\{\frac{1}{n_i}S_{i\cdot k}S_{i\cdot l}-(\Sigma_\varepsilon^0)_{k,l}\right\}+(\Sigma_\varepsilon^0)_{k,l}, \qquad (3.1)\end{aligned}$$

where $S_{i\cdot k}=\sum_{j=1}^{n_i}\varepsilon_{ijk}$.

By (3.1),

$$\begin{aligned}&\max_{k,l}\left|(\widehat{\Sigma}_\varepsilon-\Sigma_\varepsilon^0)_{k,l}\right| \\ \leq\ &\max_{k,l}\frac{1}{N-m}\left|\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left\{\varepsilon_{ijk}\varepsilon_{ijl}-(\Sigma_\varepsilon^0)_{k,l}\right\}\right| \\ &+\frac{1}{N-m}\max_{k,l}\left|\sum_{i=1}^{m}\left\{\frac{1}{n_i}S_{i\cdot k}S_{i\cdot l}-(\Sigma_\varepsilon^0)_{k,l}\right\}\right|. \qquad (3.2)\end{aligned}$$

Now, we assume that $\varepsilon_{ijk}\in\mathcal{SG}(\sigma_{\varepsilon,k}^2)$, i.e., ε_{ijk} is sub-Gaussian with a variance factor $\sigma_{\varepsilon,k}^2$ for $1\leq i\leq m, 1\leq j\leq n_i, 1\leq k\leq p$. It is easy to check that $n_i^{-1/2}S_{i\cdot k}\in\mathcal{SG}(\sigma_{\varepsilon,k}^2)$.

Let $\psi:\mathbb{R}_+\to\mathbb{R}_+$ be a convex function with $\psi(0)=0$, especially, $\psi_q(v)=\exp(|v|^q)-1$, for $q\in[1,2]$. Then for an $\mathbb{R}$-valued random variable X, the Orlicz norm of X is $\|X\|_\psi=$

$\inf\{t \in \mathbb{R}_+ : \mathbb{E}\{\psi(|X|/t)\} \leq 1\}$. And by the properties of Orlicz norms, for any random variable X and any increasing convex $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ with $\psi(0) = 0$, we have

$$\|X - E(X)\|_\psi \leq 2\|X\|_\psi. \tag{3.3}$$

Moreover, if $X \in \mathcal{SG}(\sigma^2)$, then

$$\|X\|_{\psi_2} \leq c_0\sigma, \tag{3.4}$$

for some $c_0 \leq (8/3)^{1/2}$.

Since $\varepsilon_{ijk} \in \mathcal{SG}(\sigma^2_{\varepsilon,k})$ and $n_i^{-1/2}S_{i\cdot k} \in \mathcal{SG}(\sigma^2_{\varepsilon,k})$, by Lemma 2.7.7 in Vershyin [58], $\varepsilon_{ijk}\varepsilon_{ijl}$ and $n_i^{-1}S_{i\cdot k}S_{i\cdot l}$ are sub-Exponential random variables. Let $\max_k \sigma^2_{\varepsilon,k} = M_\varepsilon$. Combining (3.3) and (3.4), Lemma 2.7.7 in Vershyin [58] implies that

$$\left\|\varepsilon_{ijk}\varepsilon_{ijl} - (\Sigma^0_\varepsilon)_{k,l}\right\|_{\psi_1} \leq 2\left\|\varepsilon_{ijk}\varepsilon_{ijl}\right\|_{\psi_1} \leq 2\left\|\varepsilon_{ijk}\right\|_{\psi_2}\left\|\varepsilon_{ijl}\right\|_{\psi_2} \leq c_1 M_\varepsilon,$$

and

$$\left\|n_i^{-1}S_{i\cdot k}S_{i\cdot l} - (\Sigma^0_\varepsilon)_{k,l}\right\|_{\psi_1} \leq 2\left\|n_i^{-1}S_{i\cdot k}S_{i\cdot l}\right\|_{\psi_1} \leq 2\left\|n_i^{-1/2}S_{i\cdot k}\right\|_{\psi_2}\left\|n_i^{-1/2}S_{i\cdot l}\right\|_{\psi_2} \leq c_1 M_\varepsilon,$$

where $c_1 = 2c_0^2$.

Hence, for the first term in (3.2), by the union sum inequality and Bernstein's inequality (Theorem 2.8.2 in Vershyin [58]), we can get

$$\begin{aligned} &\mathbb{P}\mathrm{r}\left[\max_{k,l}\frac{1}{N-m}\left|\sum_{i=1}^{m}\sum_{j=1}^{n_i}\{\varepsilon_{ijk}\varepsilon_{ijl} - (\Sigma^0_\varepsilon)_{k,l}\}\right| \geq t\right] \\ \leq\ & 2p^2\exp\left[-c_2\min\left\{\frac{t^2(N-m)^2}{NK_1^2}, \frac{t(N-m)}{K_1}\right\}\right], \end{aligned} \tag{3.5}$$

where $c_2 > 0$, $K_1 = \max_{i,k,l} \|\varepsilon_{ijk}\varepsilon_{ijl} - (\Sigma_\varepsilon^0)_{k,l}\|_{\psi_1} \leq c_1 M_\varepsilon$.

Similarly,

$$\begin{aligned} & \mathbb{P}\mathrm{r}\left[\frac{1}{N-m}\max_{k_1}\left|\sum_{i=1}^{m}\left\{\frac{1}{n_i}S_{i\cdot k_1}S_{i\cdot k_2} - (\Sigma_\varepsilon^0)_{k,l}\right\}\right| \geq t\right] \\ \leq \; & 2p^2\exp\left[-c_3\min\left\{\frac{t^2(N-m)^2}{mK_2^2}, \frac{t(N-m)}{K_2}\right\}\right], \end{aligned} \tag{3.6}$$

where $c_3 > 0$, $K_2 = \max_{i,k,l} \|n_i^{-1}S_{i\cdot k}S_{i\cdot l} - (\Sigma_\varepsilon^0)_{k,l}\|_{\psi_1} \leq c_1 M_\varepsilon$.

By (3.5) and (3.6), take $t = C_1(N\log p)^{1/2}/\{2(N-m)\}$ for a sufficiently large constant $C_1 > 0$, with $N > \log p$, we will have

$$\begin{aligned} & \mathbb{P}\mathrm{r}\left[\max_{k,l}\frac{1}{N-m}\left|\sum_{i=1}^{m}\sum_{j=1}^{n_i}\{\varepsilon_{ijk}\varepsilon_{ijl} - (\Sigma_\varepsilon^0)_{k,l}\}\right| \geq t\right] \\ \leq \; & 2\exp\left[\max\left\{\left(2 - \frac{c_2NC_1^2}{4mK_1^2}\right)\log p, 2\log p - \frac{c_2C_1}{2K_1}(N\log p)^{1/2}\right\}\right] \\ \leq \; & 2\exp\left\{\max\left(2 - \frac{c_2C_1^2}{4c_1^2M_\varepsilon^2}, 2 - \frac{c_2C_1}{2c_1M_\varepsilon}\right)\log p\right\}, \end{aligned} \tag{3.7}$$

and

$$\begin{aligned} & \mathbb{P}\mathrm{r}\left[\frac{1}{N-m}\max_{k,l}\left|\sum_{i=1}^{m}\left\{\frac{1}{n_i}S_{i\cdot k}S_{i\cdot l} - (\Sigma_\varepsilon^0)_{k,l}\right\}\right| \geq t\right] \\ \leq \; & 2\exp\left[\max\left\{\left(2 - \frac{c_3NC_1^2}{4mK_2^2}\right)\log p, 2\log p - \frac{c_3C_1}{2K_2}(N\log p)^{1/2}\right\}\right] \\ \leq \; & 2\exp\left\{\max\left(2 - \frac{c_3C_1^2}{4c_1^2M_\varepsilon^2}, 2 - \frac{c_3C_1}{2c_1M_\varepsilon}\right)\log p\right\}. \end{aligned} \tag{3.8}$$

Combining (3.7) and (3.8), with $\lambda_\varepsilon = C_1(N\log p)^{1/2}/(N-m)$, we have

$$\begin{aligned}
& \mathbb{P}\mathrm{r}\left\{\max_{k,l}\left|(\widehat{\Sigma}_\varepsilon - \Sigma_\varepsilon^0)_{k,l}\right| > \lambda_\varepsilon\right\} \\
\leq\ & \mathbb{P}\mathrm{r}\left[\frac{1}{N-m}\max_{k_1}\left|\sum_{i=1}^{m}\left\{\frac{1}{n_i}S_{i\cdot k}S_{i\cdot l} - (\Sigma_\varepsilon^0)_{k,l}\right\}\right| \geq \frac{C_1(N\log p)^{1/2}}{2(N-m)}\right] \\
& \quad + \mathbb{P}\mathrm{r}\left[\max_{k,l}\frac{1}{N-m}\left|\sum_{i=1}^{m}\sum_{j=1}^{n_i}\{\varepsilon_{ijk}\varepsilon_{ijl} - (\Sigma_\varepsilon^0)_{k,l}\}\right| \geq \frac{C_1(N\log p)^{1/2}}{2(N-m)}\right] \\
\leq\ & 2\exp\left\{\max\left(2 - \frac{c_3C_1^2}{4c_1^2M_\varepsilon^2}, 2 - \frac{c_3C_1}{2c_1M_\varepsilon}\right)\log p\right\} \\
& \quad + 2\exp\left\{\max\left(2 - \frac{c_2C_1^2}{4c_1^2M_\varepsilon^2}, 2 - \frac{c_2C_1}{2c_1M_\varepsilon}\right)\log p\right\} \\
\leq\ & 4p^{-C_2},
\end{aligned}$$

where $C_2 = \min\{c_3C_1(2c_1M_\varepsilon)^{-1}, c_3, c_2C_1(2c_1M_\varepsilon)^{-1}, c_2\}(2c_1M_\varepsilon)^{-1}C_1 - 2$. □

Lemma 2 *Consider the true within-subject covariance* Σ_ε^0 *with* $\max_k(\Sigma_\varepsilon^0)_{k,k} \leq M_\varepsilon$ *and the true between-subject covariance* Σ_b^0 *with* $\max_k(\Sigma_b^0)_{k,k} \leq M_b$. *Let*

$$\lambda_b = C_1\left(\frac{\log p}{m}\right)^{1/2} + C_2\frac{(N\log p)^{1/2}}{(N-m)n^*} + \frac{M_b}{m} + \frac{M_\varepsilon}{mn^*}$$

for sufficiently large $C_1, C_2 > 0$, *where* $n^* = m/\sum_{i=1}^m n_i^{-1}$. *If* $\log p \leq m$, *then the unbiased between-subject sample estimate* $\widehat{\Sigma}_b$ *satisfies*

$$\mathbb{P}\mathrm{r}\left\{\max_{k,l}\left|(\widehat{\Sigma}_b - \Sigma_b^0)_{k,l}\right| > 2\lambda_b\right\} \leq 8p^{-C_3},$$

where $C_3 > 0$ *only depends on* C_1, C_2 *and* $\max(M_\varepsilon, M_b)$.

Proof. Let $\bar{Y}_{i\cdot k} = b_{ik} + n_i^{-1}\sum_{j=1}^{n_i}\varepsilon_{ijk} = b_{ik} + n_i^{-1}S_{i\cdot k} = W_{ik}$, then by decomposition,

$$\begin{aligned}
(\widehat{\Sigma}_b - \Sigma_b^0)_{k,l} &= \{\overline{\Sigma} - (n^*)^{-1}\widehat{\Sigma}_\varepsilon - \Sigma_b^0\}_{k,l} \\
&= [\overline{\Sigma} - \{\Sigma_b^0 + (n^*)^{-1}\Sigma_\varepsilon^0\}]_{k,l} - (n^*)^{-1}(\widehat{\Sigma}_\varepsilon - \Sigma_\varepsilon^0)_{k,l} \\
&= \frac{1}{m-1}\sum_{i=1}^{m}\left\{W_{ik}W_{il} - \left(\Sigma_b^0 + n_i^{-1}\Sigma_\varepsilon^0\right)_{k,l}\right\} \\
&\quad -\frac{m}{m-1}\left(\frac{1}{m}\sum_{i=1}^{m}W_{ik}\right)\left(\frac{1}{m}\sum_{i=1}^{m}W_{il}\right) \\
&\quad +\frac{(\Sigma_b^0)_{k,l}}{m-1} + \frac{(\Sigma_\varepsilon^0)_{k,l}}{(m-1)n^*} - \frac{(\widehat{\Sigma}_\varepsilon - \Sigma_\varepsilon^0)_{k,l}}{n^*}. \qquad (3.9)
\end{aligned}$$

Then, with $|(\Sigma_b^0)_{k,l}| \le M_b$ and $|(\Sigma_\varepsilon^0)_{k,l}| \le M_\varepsilon$, we have

$$\begin{aligned}
\max_{k,l}\left|(\widehat{\Sigma}_b - \Sigma_b^0)_{k,l}\right| &\le 2\max_{k,l}\left|\frac{1}{m}\sum_{i=1}^{m}\left\{W_{ik}W_{il} - \left(\Sigma_b^0 + n_i^{-1}\Sigma_\varepsilon^0\right)_{k,l}\right\}\right| \\
&\quad +2\max_{k,l}\left|\left(\frac{1}{m}\sum_{i=1}^{m}W_{ik}\right)\left(\frac{1}{m}\sum_{i=1}^{m}W_{il}\right)\right| \\
&\quad +\max_{k,l}(n^*)^{-1}\left|(\widehat{\Sigma}_\varepsilon - \Sigma_\varepsilon^0)_{k,l}\right| \\
&\quad +\frac{2M_b}{m} + \frac{2M_\varepsilon}{mn^*}. \qquad (3.10)
\end{aligned}$$

Assume that $b_{ik} \in \mathcal{SG}(\sigma_{b,k}^2)$, i.e., b_{ik} is sub-Gaussian with a variance factor $\sigma_{b,k}^2$ for $1 \le i \le m, 1 \le k \le p$. Then $W_{ik} \in \mathcal{SG}(\sigma_{b,k}^2 + n_i^{-1}\sigma_{\varepsilon,k}^2)$. Let $\max_k \sigma_{b,k}^2 = M_b$. Then, by Lemma 2.7.7 in Vershyin [58], we obtain

$$\begin{aligned}
\left\|W_{ik}W_{il} - \left(\Sigma_b^0 + n_i^{-1}\Sigma_\varepsilon^0\right)_{k,l}\right\|_{\psi_1} &\le 2\left\|W_{ik}\right\|_{\psi_2}\left\|W_{il}\right\|_{\psi_2} \\
&\le c_1\left(\Sigma_b^0 + n_i^{-1}\Sigma_\varepsilon^0\right)_{k,l} \\
&\le c_1\left(1 + n_l^{-1}\right)M_* \\
&\le 2c_1M_*,
\end{aligned}$$

where $n_l = \min n_i$ and $M_* = \max(M_\varepsilon, M_b)$. And with the Bernstein's inequality, we have

$$\mathbb{P}\mathrm{r}\left[\frac{1}{m}\left|\sum_{i=1}^{m}\left\{W_{ik}W_{il} - (\Sigma_b^0 + n_i^{-1}\Sigma_\varepsilon^0)_{k,l}\right\}\right| \geq t\right] \leq 2\exp\left\{-c_4 \min\left(\frac{mt^2}{K_3^2}, \frac{mt}{K_3}\right)\right\},$$

where $c_4 > 0$, $K_3 = \max_{i,k,l} \|W_{ik}W_{il} - (\Sigma_b^0 + n_i^{-1}\Sigma_\varepsilon^0)_{k,l}\|_{\psi_1} \leq 2c_1 M_*$.

By the union sum inequality and taking $t = 2^{-1}C_1(\log p/m)^{1/2}$ for a sufficiently large constant $C_1 > 0$, if $m \geq \log p$, we have

$$\begin{aligned}
& \mathbb{P}\mathrm{r}\left[\max_{k,l}\frac{1}{m}\left|\sum_{i=1}^{m}\left\{W_{ik}W_{il} - (\Sigma_b^0 + n_i^{-1}\Sigma_\varepsilon^0)_{k,l}\right\}\right| \geq t\right] \\
\leq\ & 2p^2\exp\left[-c_4 \min\left\{\frac{C_1^2 \log p}{4K_3^2}, \frac{C_1(m\log p)^{1/2}}{2K_3}\right\}\right] \\
\leq\ & 2\exp\left[\left\{2 - \min\left(\frac{c_4 C_1^2}{16 c_1^2 M_*^2}, \frac{c_4 C_1}{4c_1 M_*}\right)\right\}\log p\right]. \qquad (3.11)
\end{aligned}$$

We use a union bound with the general Hoeffding's inequality (Theorem 2.6.2 by Vershyin [58]) to bound the second term in (3.10). Specifically, with $m \geq \log p$ and taking $t = 2^{-1}C_1(\log p/m)^{1/2}$, we have

$$\begin{aligned}
\mathbb{P}\mathrm{r}\left(\max_{k,l}\left|\frac{1}{m}\sum_{i=1}^{m}W_{ik}\right|^2 \geq t\right) &= \mathbb{P}\mathrm{r}\left(\max_{k,l}\left|\sum_{i=1}^{m}W_{ik}\right| \geq mt^{1/2}\right) \\
&\leq 2p\exp\left(-\frac{c_5 m^2 t}{\sum_{i=1}^{m}\|W_{ik}\|_{\psi_2}^2}\right) \\
&\leq 2p\exp\left(-\frac{c_5 mt}{c_1 M_*}\right) \\
&= 2p\exp\left\{-\frac{c_5 C_1}{2c_1 M_*}(m\log p)^{1/2}\right\} \\
&\leq 2\exp\left\{\left(1 - \frac{c_5 C_1}{2c_1 M_*}\right)\log p\right\}, \qquad (3.12)
\end{aligned}$$

where $c_5 > 0$.

For the third term in (3.10), by Lemma 1, for a sufficiently large constant $C_2 > 0$, we have

$$\mathbb{P}\mathrm{r}\left\{\max_{k,l}\frac{\left|(\widehat{\Sigma}_\varepsilon - \Sigma_\varepsilon^0)_{k,l}\right|}{n^*} \geq 2C_2\frac{(N\log p)^{1/2}}{(N-m)n^*}\right\} \leq 4p^{-C_3'}, \tag{3.13}$$

where $C_3' > 0$ only depends on C_2 and M_ε.

Collecting (3.11)-(3.13), with

$$\lambda_b = C_1\left(\frac{\log p}{m}\right)^{1/2} + C_2\frac{(N\log p)^{1/2}}{(N-m)n_*} + \frac{M_b}{m} + \frac{M_\varepsilon}{mn_*},$$

we have

$$\begin{aligned}
& \mathbb{P}\mathrm{r}\left\{\max_{k,l}\left|(\widehat{\Sigma}_b - \Sigma_b^0)_{k,l}\right| \geq 2\lambda_b\right\} \\
\leq\ & \mathbb{P}\mathrm{r}\left[\max_{k,l}\frac{2}{m}\left|\sum_{i=1}^m\left\{W_{ik}W_{il} - (\Sigma_b^0 + n_i^{-1}\Sigma_\varepsilon^0)_{k,l}\right\}\right| \geq C_1\left(\frac{\log p}{m}\right)^{1/2}\right] \\
& \quad + \mathbb{P}\mathrm{r}\left\{2\max_{k,l}\left|\left(\frac{1}{m}\sum_{i=1}^m W_{ik}\right)\left(\frac{1}{m}\sum_{i=1}^m W_{il}\right)\right| \geq C_1\left(\frac{\log p}{m}\right)^{1/2}\right\} \\
& \quad + \mathbb{P}\mathrm{r}\left\{\max_{k,l}\frac{\left|(\widehat{\Sigma}_\varepsilon - \Sigma_\varepsilon^0)_{k,l}\right|}{n^*} \geq 2C_2\frac{(N\log p)^{1/2}}{(N-m)n^*}\right\} \\
\leq\ & 4p^{-C_3'} + 4p^{-C_3''} \\
\leq\ & 8p^{-C_3},
\end{aligned}$$

where $C_3'' = \min\{c_4C_1^2(16c_1^2M_*^2)^{-1}, c_4C_1(4c_1M_*)^{-1}, c_5C_1(2c_1M_*)^{-1} + 1\} - 2$ and $C_3 = \min(C_3', C_3'')$.

□

Lemma 3 *Consider the true within-subject covariance* Σ_ε^0 *with* $\max_k(\Sigma_\varepsilon^0)_{k,k} \leq M_\varepsilon$ *and*

the true between-subject covariance Σ_b^0 *with* $\max_k(\Sigma_b^0)_{k,k} \leq M_b$. *Let*

$$\lambda_0 = C_1 \left(\frac{\log p}{m}\right)^{1/2} + \frac{M_b}{m} + \frac{M_\varepsilon}{n^*}$$

for sufficiently large $C_1 > 0$, *where* $n^* = m/\sum_{i=1}^m n_i^{-1}$. *If* $\log p \leq m$, *then the naive between-subject sample estimate* $\overline{\Sigma}$ *satisfies*

$$\Pr\left\{\max_{k,l}\left|(\overline{\Sigma} - \Sigma_b^0)_{k,l}\right| > 2\lambda_b\right\} \leq 8p^{-C_2}$$

where $C_2 > 0$ *only depends on* C_1 *and* $\max(M_\varepsilon, M_b)$.

Proof. Now, we will consider the convergence rate of $\max_{k,l}|(\overline{\Sigma} - \Sigma_b^0)_{k,l}|$. By (3.9), we have

$$\begin{aligned}(\overline{\Sigma} - \Sigma_b^0)_{k,l} &= \frac{1}{m-1}\sum_{i=1}^m \left\{W_{ik}W_{il} - \left(\Sigma_b^0 + n_i^{-1}\Sigma_\varepsilon^0\right)_{k,l}\right\} \\ &\quad - \frac{m}{m-1}\left(\frac{1}{m}\sum_{i=1}^m W_{ik}\right)\left(\frac{1}{m}\sum_{i=1}^m W_{il}\right) \\ &\quad + \frac{(\Sigma_b^0)_{k,l}}{m-1} + \frac{m(\Sigma_\varepsilon^0)_{k,l}}{(m-1)n^*}. \end{aligned} \tag{3.14}$$

Then, with $|(\Sigma_b^0)_{k,l}| \leq M_b$ and $|(\Sigma_\varepsilon^0)_{k,l}| \leq M_\varepsilon$, we have

$$\begin{aligned}\max_{k,l}\left|(\overline{\Sigma} - \Sigma_b^0)_{k,l}\right| &\leq 2\max_{k,l}\left|\frac{1}{m}\sum_{i=1}^m \left\{W_{ik}W_{il} - \left(\Sigma_b^0 + n_i^{-1}\Sigma_\varepsilon^0\right)_{k,l}\right\}\right| \\ &\quad + 2\max_{k,l}\left|\left(\frac{1}{m}\sum_{i=1}^m W_{ik}\right)\left(\frac{1}{m}\sum_{i=1}^m W_{il}\right)\right| \\ &\quad + \frac{2M_b}{m} + \frac{2M_\varepsilon}{n^*}. \end{aligned} \tag{3.15}$$

Following the steps in Lemma 2, with

$$\lambda_0 = C_1 \left(\frac{\log p}{m} \right)^{1/2} + \frac{M_b}{m} + \frac{M_\varepsilon}{n^*}$$

for a sufficiently large constant $C_1 > 0$, we have

$$\mathbb{P}\mathrm{r} \left\{ \max_{k,l} \left| (\overline{\Sigma} - \Sigma_b^0)_{k,l} \right| > 2\lambda_0 \right\} \leq 4p^{-C_2},$$

where $C_2 > 0$ only depends on C_1 and $\max(M_\varepsilon, M_b)$. □

Lemma 4 *Consider the true within-subject covariance* Σ_ε^0 *with* $\max_k(\Sigma_\varepsilon^0)_{k,k} \leq M_\varepsilon$ *and the true between-subject covariance* Σ_b^0 *with* $\max_k(\Sigma_b^0)_{k,k} \leq M_b$. *Let*

$$\lambda_1 = C_1 \left(\frac{\log p}{m} \right)^{1/2} + M_b + \frac{(2 - n^*)M_\varepsilon}{2n^*}$$

for sufficiently large $C_1 > 0$, *where* $n^* = m/\sum_{i=1}^m n_i^{-1}$. *If* $\log p \leq m$, *then* $\overline{\Sigma}$ *satisfies*

$$\mathbb{P}\mathrm{r} \left\{ \max_{k,l} \left| (\overline{\Sigma} - \Sigma_\varepsilon^0)_{k,l} \right| > 2\lambda_1 \right\} \leq 4p^{-C_2}$$

where $C_2 > 0$ *only depends on* C_1 *and* $\max(M_\varepsilon, M_b)$.

Proof. Now, we will consider the convergence rate of $\max_{k,l} |(\overline{\Sigma} - \Sigma_\varepsilon^0)_{k,l}|$. Note that $(\overline{\Sigma} - \Sigma_\varepsilon^0)_{k,l} = (\overline{\Sigma} - \Sigma_b^0)_{k,l} + (\Sigma_b^0)_{k,l} - (\Sigma_\varepsilon^0)_{k,l}$. Then, by (3.14), with $|(\Sigma_b^0)_{k,l}| \leq M_b$ and

$|(\Sigma_\varepsilon^0)_{k,l}| \le M_\varepsilon$, we have

$$\begin{aligned}\max_{k,l}\left|(\overline{\Sigma}-\Sigma_\varepsilon^0)_{k,l}\right| &\le 2\max_{k,l}\left|\frac{1}{m}\sum_{i=1}^{m}\left\{W_{ik}W_{il}-\left(\Sigma_b^0+n_i^{-1}\Sigma_\varepsilon^0\right)_{k,l}\right\}\right| \\ &\quad +2\max_{k,l}\left|\left(\frac{1}{m}\sum_{i=1}^{m}W_{ik}\right)\left(\frac{1}{m}\sum_{i=1}^{m}W_{il}\right)\right| \\ &\quad +2M_b+\frac{(2-n^*)M_\varepsilon}{n^*}. \end{aligned} \tag{3.16}$$

Following the steps in Lemma 2, with

$$\lambda_1 = C_1\left(\frac{\log p}{m}\right)^{1/2}+M_b+\frac{(2-n^*)M_\varepsilon}{2n^*}$$

for a sufficiently large constant $C_1>0$, we have

$$\mathbb{P}\mathrm{r}\left\{\max_{k,l}\left|(\overline{\Sigma}-\Sigma_\varepsilon^0)_{k,l}\right|>2\lambda_1\right\}\le 4p^{-C_2},$$

where $C_2>0$ only depends on C_1 and $\max(M_\varepsilon, M_b)$. □

3.3 Estimation Error Rate for the Within-Subject Covariance Estimator

Theorem 1 (Estimation error rate of $\widehat{\Sigma}_\varepsilon^+$) *Assume that the true within-subject covariance matrix $\Sigma_\varepsilon^0 \in \mathcal{U}(M_\varepsilon, s_\varepsilon)$. Let $\lambda_\varepsilon = C_1(N\log p)^{1/2}/(N-m)$ be the value of the tuning parameter λ in (2.5) for a sufficiently large constant $C_1>0$. If $\log p \le N$, the proposed within-subject estimator $\widehat{\Sigma}_\varepsilon^+$ satisfies*

$$\left\|\widehat{\Sigma}_\varepsilon^+-\Sigma_\varepsilon^0\right\|_F \le 5\lambda_\varepsilon(ps_\varepsilon)^{1/2}$$

with probability at least $1 - 4p^{-C_2}$*, where* $C_2 > 0$ *only depends on* C_1 *and* M_ε*.*

Proof. Define $\Delta_\varepsilon = \Sigma_\varepsilon - \Sigma_\varepsilon^0$ and $F_\varepsilon(\Delta_\varepsilon) = \|\Delta_\varepsilon + \Sigma_\varepsilon^0 - \widehat{\Sigma}_\varepsilon\|_F^2/2 + \lambda_\varepsilon|\Delta_\varepsilon + \Sigma_\varepsilon^0|_1$, then the objective function (2.6) is equivalent to

$$\min_{\Delta_\varepsilon : \Delta_\varepsilon = \Delta_\varepsilon^{\mathrm{T}}, \Delta_\varepsilon + \Sigma_\varepsilon^0 \succeq \delta I} F_\varepsilon(\Delta_\varepsilon).$$

Consider the set

$$\{\Delta_\varepsilon : \Delta_\varepsilon = \Delta_\varepsilon^{\mathrm{T}}, \Delta_\varepsilon + \Sigma_\varepsilon^0 \succeq \delta I, \|\Delta_\varepsilon\|_F = 5\lambda_\varepsilon(ps_\varepsilon)^{1/2}\}. \tag{3.17}$$

According to Xue et al. [29], under the probability event $\{|(\widehat{\Sigma}_\varepsilon - \Sigma_\varepsilon^0)_{,k,l}| \le \lambda_\varepsilon, \forall(i,j)\}$, we have

$$\begin{aligned} F_\varepsilon(\Delta_\varepsilon) - F_\varepsilon(\mathbf{0}) &\ge \frac{1}{2}\|\Delta_\varepsilon\|_F^2 - 2\lambda_\varepsilon \left[\sum_{k,l=1}^{p} 1\{(\Sigma_\varepsilon^0)_{k,l} \neq 0\}\right]^{1/2} \|\Delta_\varepsilon\|_F \\ &\ge \frac{1}{2}\|\Delta_\varepsilon\|_F^2 - 2\lambda_\varepsilon(ps_\varepsilon)^{1/2}\|\Delta_\varepsilon\|_F \\ &= \frac{5}{2}\lambda_\varepsilon^2 ps_\varepsilon \\ &> 0. \end{aligned}$$

Note that $F_\varepsilon(\Delta_\varepsilon)$ is a convex function and $F_\varepsilon(\widehat{\Delta}_\varepsilon) \le F_\varepsilon(\mathbf{0}) = 0$. Then, the minimizer $\widehat{\Delta}_\varepsilon$ must be inside the sphere (3.17). Hence, we have

$$\begin{aligned} & \mathbb{P}\mathrm{r}\left\{\left\|\widehat{\Sigma}_\varepsilon^+ - \Sigma_\varepsilon^0\right\|_F \le 5\lambda_\varepsilon(ps_\varepsilon)^{1/2}\right\} \\ \ge\ & 1 - \mathbb{P}\mathrm{r}\left\{\max_{k,l}\left|(\widehat{\Sigma}_\varepsilon - \Sigma_\varepsilon^0)_{k,l}\right| > \lambda_\varepsilon\right\} \\ \ge\ & 1 - 4p^{-C_2}. \end{aligned}$$

□

The term $(ps_\varepsilon)^{1/2}$ in the error rate above represents the overall sparsity of the true covariance matrix Σ_ε^0. This dependence on sparsity level has also been noted in Rothman et al. [6] and Xue et al. [29] over slightly different matrix classes. Notably, the estimation error rate does not depend on M_ε or on the exact values of n_i for $i = 1, ..., m$. Instead, the effective sample size in λ is $N^{1/2} - N^{-1/2}m$, which only depends on the total observation number N and the number of subjects m.

Remark 1. When the number of subject m is relatively small compared with the total number of observations N in the scale of $m = o(N^{1/2})$, Theorem 1 implies that

$$\left\|\widehat{\Sigma}_\varepsilon^+ - \Sigma_\varepsilon^0\right\|_F = O_P\left\{(ps_\varepsilon N^{-1}\log p)^{1/2}\right\},$$

where $X_n = O_P(a_n)$ means that for a set of random variables X_n and a corresponding set of constants a_n, X_n/a_n is bounded by a positive constant with probability approaching 1. This rate coincides with those in Rothman et al. [6], Bickel and Levina [25], Rothman et al. [26], Cai and Liu [27] and Xue et al. [29], which are derived based on the assumption of independent and identically distributed observations.

Remark 2. On the other hand, with $m = O(N)$, e.g., when the number of repeated measurements of each subject is bounded by a constant, Theorem 1 implies that

$$\left\|\widehat{\Sigma}_\varepsilon^+ - \Sigma_\varepsilon^0\right\|_F = O_P\left\{(ps_\varepsilon m^{-1}\log p)^{1/2})\right\}.$$

In this scenario, m plays the role of the effective sample size, and estimation consistency is achieved when m approaches infinity.

3.4 Estimation Error Rate for the Between-Subject Covariance Estimator

Theorem 2 (Estimation error rate of $\widehat{\Sigma}_b^+$) *Assume that the true between-subject covariance matrix $\Sigma_b^0 \in \mathcal{U}(M_b, s_b)$ and the true within-subject covariance matrix $\Sigma_\varepsilon^0 \in \mathcal{U}(M_\varepsilon, s_\varepsilon)$. Let*

$$\lambda_b = C_1 \left(\frac{\log p}{m}\right)^{1/2} + C_2 \frac{(N \log p)^{1/2}}{(N-m)n^*} + \frac{M_b}{m} + \frac{M_\varepsilon}{mn^*} \tag{3.18}$$

be the value of the tuning parameter λ in (2.5) for sufficiently large $C_1, C_2 > 0$, where $n^* = m/\sum_{i=1}^m n_i^{-1}$. If $\log p \leq m$, then the proposed between-subject estimator $\widehat{\Sigma}_b^+$ satisfies

$$\left\| \widehat{\Sigma}_b^+ - \Sigma_b^0 \right\|_F \leq 10\lambda_b (ps_b)^{1/2}$$

with probability at least $1 - 8p^{-C_3}$, where $C_3 > 0$ only depends on C_1, C_2 and $\max(M_\varepsilon, M_b)$.

Unlike the estimation error rate for $\widehat{\Sigma}_\varepsilon$ in Theorem 1, the rate for $\widehat{\Sigma}_b$ depends on the values of n_i's via the term n^*. A simple bound $n^* \geq \min_i n_i$ implies that the second term in λ_b converges to 0 at a rate that is at least not slower than λ_ε in Theorem 1. The rate in λ_b is thus dominated by $(m^{-1} \log p)^{1/2}$.

Recall from (2.2) that $\overline{\Sigma}$ has a bias of Σ_ε / n^* in estimating Σ_b. In practice, $\overline{\Sigma}$ has been misused to provide a sample estimate for subsequent regularized estimation (Epskamp et al. [45]). We establish the following estimation error rate for $\overline{\Sigma}^+$, which is defined as the solution to (2.5) with input sample matrix $D = \overline{\Sigma}$, to illustrate that the bias in the sample estimate is carried over to the regularized estimation.

Theorem 3 (Estimation error rate of $\overline{\Sigma}^+$) *Assume that the true between-subject covariance matrix $\Sigma_b^0 \in \mathcal{U}(M_b, s_b)$ and the true within-subject covariance matrix $\Sigma_\varepsilon^0 \in$*

$\mathcal{U}(M_\varepsilon, s_\varepsilon)$. *Let*

$$\lambda_0 = C_1 \left(\frac{\log p}{m}\right)^{1/2} + \frac{M_b}{m} + \frac{M_\varepsilon}{n^*}$$

be the value of the tuning parameter λ *in (2.5) for sufficiently large* $C_1 > 0$, *and the same* n^* *defined in Theorem 2. If* $\log p \leq m$, *then the aggregated between-subject estimator* $\overline{\Sigma}^+$ *satisfies*

$$\left\|\overline{\Sigma}^+ - \Sigma_b^0\right\|_F \leq 10\lambda_0 (p s_b)^{1/2}$$

with probability at least $1 - 4p^{-C_2}$, *where* $C_2 > 0$ *only depends on* C_1 *and* $\max(M_\varepsilon, M_b)$.

The upper bound of the estimation error rate in $\overline{\Sigma}^+$ is strictly larger than that of $\widehat{\Sigma}_b^+$ due to the dominant term M_ε/n^* in λ_0, which corresponds to the bias in (2.2). For example, in the balanced setting where $n_i = n_1$ for all $i = 1, \ldots, m$, it holds that $n^* = n_1$ and this bias term M_ε/n_1 does not vanish even if $m \to \infty$ as long as $n_1 = O(1)$. We also show that $\overline{\Sigma}^+$ is inconsistent in estimating the within-subject covariance Σ_ε in Theorem 4.

Theorem 4 *Consider the true between-subject covariance matrix* $\Sigma_b^0 \in \mathcal{U}(M_b, s_b)$ *and the true within-subject covariance matrix* $\Sigma_\varepsilon^0 \in \mathcal{U}(M_\varepsilon, s_\varepsilon)$. *Let*

$$\lambda_1 = C_1 \left(\frac{\log p}{m}\right)^{1/2} + M_b + \frac{(2 - n^*) M_\varepsilon}{2n^*}$$

be the value of the tuning parameter λ *in* (2.6) *for sufficiently large* $C_1 > 0$, *and the same* n^* *defined in Theorem 2. If* $\log p \leq m$, *then the naive estimator* $\overline{\Sigma}^+$ *satisfies*

$$\left\|\overline{\Sigma}^+ - \Sigma_\varepsilon^0\right\|_F \leq 10\lambda_1 (p s_\varepsilon)^{1/2}$$

with probability at least $1 - 4p^{-C_2}$, *where* $C_2 > 0$ *only depends on* C_1 *and* $\max(M_\varepsilon, M_b)$.

The proof of Theorems 2, 3 and 4 follow straightforwardly from Theorem 1. In some scenarios, the estimation of between-subject and within-subject correlation matrices, instead of covariance matrices, is of interest and can be obtained similarly in the proposed framework. We provide estimation error rates of the sparse positive definite estimators of two correlation matrices in Corollaries 1 and 2. And the sparse and positive-definite estimate of R_ε, denoted as $\widehat{R}_\varepsilon^+$, and of Σ_b, denoted as $\widehat{R}_b^+$ are defined as solution of (2.6) with $D = \widehat{R}_\varepsilon = D_\varepsilon^{-1/2}\widehat{\Sigma}_\varepsilon D_\varepsilon^{-1/2}$ and $D = \widehat{R}_b = D_b^{-1/2}\widehat{\Sigma}_b D_b^{-1/2}$, where $D_\varepsilon = \text{diag}\{(\widehat{\Sigma}_\varepsilon)_{1,1}, \ldots, (\widehat{\Sigma}_\varepsilon)_{p,p}\}$ and $D_b = \text{diag}\{(\widehat{\Sigma}_b)_{1,1}, \ldots, (\widehat{\Sigma}_b)_{p,p}\}$.

Corollary 1 *Under conditions of Theorem 1, if* $\min_k(\Sigma_\varepsilon^0)_{k,k}$ *is bounded from below, then*

$$\left\|\widehat{R}_\varepsilon^+ - R_\varepsilon^0\right\|_F = O_P\left\{\frac{(ps_\varepsilon N \log p)^{1/2}}{N-m}\right\},$$

uniformly on $\Sigma_\varepsilon^0 \in \mathcal{U}(M_\varepsilon, s_\varepsilon)$*, as* $N, m \to \infty$*.*

Proof. By Lemma 1, we have

$$\mathbb{P}\text{r}\left\{\max_{k,l}\left|(\widehat{\Sigma}_\varepsilon - \Sigma_\varepsilon^0)_{k,l}\right| > C_1\frac{(N\log p)^{1/2}}{N-m}\right\} = o(1). \tag{3.19}$$

According to Lemma 2 in Cui et al. [31], with (3.19) and the fact that $(\widehat{R}_\varepsilon)_{k,l} = (\widehat{\Sigma}_\varepsilon)_{k,l}/\{(\widehat{\Sigma}_\varepsilon)_{k,k}(\widehat{\Sigma}_\varepsilon)_{l,l}\}^{1/2}$, for a sufficiently large constant $C_1' > 0$, we have

$$\mathbb{P}\text{r}\left\{\max_{k,l}\left|(\widehat{R}_\varepsilon - R_\varepsilon^0)_{k,l}\right| > C_1'\frac{(N\log p)^{1/2}}{N-m}\right\} = o(1).$$

Following the steps in the proof of Theorem 1, it is easily shown that

$$\left\|\widehat{R}_\varepsilon^+ - R_\varepsilon^0\right\|_F = O_P\left\{\frac{(ps_\varepsilon N \log p)^{1/2}}{N-m}\right\}.$$

□

Corollary 2 *Under conditions of Theorem 2, if* $\min_k(\Sigma_\varepsilon^0)_{k,k}$ *and* $\min_k(\Sigma_b^0)_{k,k}$ *are bounded from below, then*

$$\left\|\widehat{R}_b^+ - R_b^0\right\|_F = O_P\left[(ps_b)^{1/2}\left\{C_1'\left(\frac{\log p}{m}\right)^{1/2} + C_2'\frac{(N\log p)^{1/2}}{(N-m)n_0}\right\}\right],$$

uniformly on $\Sigma_\varepsilon^0 \in \mathcal{U}(M_\varepsilon, s_\varepsilon)$ *and* $\Sigma_b^0 \in \mathcal{U}(M_b, s_b)$*, for some large* $C_1', C_2' > 0$*, as* $m, n \to \infty$.

3.5 Comparison Between Two Unbiased Between-Subject Covariance Estimators

We consider a commonly used unbiased estimator of Σ_b based on the multivariate analysis of variance (Rao and Heckler [54]):

$$\widetilde{\Sigma}_b = \frac{1}{n_0}\left\{\sum_{i=1}^m \frac{n_i}{m-1}(\bar{\boldsymbol{Y}}_{i\cdot} - \bar{\boldsymbol{Y}}_{\cdot\cdot})(\bar{\boldsymbol{Y}}_{i\cdot} - \bar{\boldsymbol{Y}}_{\cdot\cdot})^{\mathrm{T}} - \widehat{\Sigma}_\varepsilon\right\},$$
$$\text{where } n_0 = \frac{N - N^{-1}\sum_{i=1}^m n_i^2}{m-1}, \tag{3.20}$$

$\bar{\boldsymbol{Y}}_{i\cdot} = n_i^{-1}\sum_{j=1}^{n_i}\boldsymbol{Y}_{ij}$, $\bar{\boldsymbol{Y}}_{\cdot\cdot} = N^{-1}\sum_{i=1}^m\sum_{j=1}^{n_i}\boldsymbol{Y}_{ij}$, and $N = \sum_{i=1}^m n_i$.

It is straightforward to show that $\mathbb{E}(\widetilde{\Sigma}_b) = \Sigma_b$. However, just like $\widehat{\Sigma}_b$ in (2.4), the diagonal elements of $\widetilde{\Sigma}_b$ could be negative, which is undesirable for an estimate of Σ_b. Specifically, in the setting where $\boldsymbol{b}_i$ and $\boldsymbol{\varepsilon}_{ij}$ follow Gaussian distributions and n_i's are all equal, it can be shown that $\Pr\{(\widetilde{\Sigma}_b)_{k,k} < 0\}$ decreases with $(\Sigma_b^0)_{k,k}/(\Sigma_\varepsilon^0)_{k,k}$. An adjustment for negative diagonal values of $\widetilde{\Sigma}_b$ is proposed in Rao and Heckler [54] based on the assumption that $\widehat{\Sigma}_\varepsilon$ is positive definite, which is violated in the high-dimensional settings.

We demonstrate an additional limitation of using $\widetilde{\Sigma}_b$, in comparison with $\widehat{\Sigma}_b$, in obtaining a sparse positive definite estimate of Σ_b. Define $\widetilde{\Sigma}_b^+$ as a solution of (2.5) with $D = \widetilde{\Sigma}_b$. The following lemma and theorem show that the performance of $\widetilde{\Sigma}_b^+$ hinges on the data imbalance.

Lemma 5 *Consider the true within-subject covariance* Σ_ε^0 *with* $\max_k(\Sigma_\varepsilon^0)_{k,k} \leq M_\varepsilon$ *and the true between-subject covariance* Σ_b^0 *with* $\max_k(\Sigma_b^0)_{k,k} \leq M_b$. *Let*

$$\widetilde{\lambda}_b = C_1 \frac{\max_i n_i}{n_0} \left(\frac{\log p}{m}\right)^{1/2} + C_2 \frac{(N \log p)^{1/2}}{n_0(N-m)} + \frac{(2N - n_0 m)M_b}{2n_0 m} + \frac{M_\varepsilon}{n_0 m}$$

for sufficiently large $C_1, C_2 > 0$. *If* $\log p \leq m$, *then* $\widetilde{\Sigma}_b$ *satisfies*

$$\mathbb{P}\mathrm{r}\left[\max_{k,l} \left|(\widetilde{\Sigma}_b - \Sigma_b^0)_{k,l}\right| > 2\widetilde{\lambda}_b\right] \leq 8p^{-C_3},$$

where $C_3 > 0$ *only depends on* C_1, C_2 *and* $\max(M_\varepsilon, M_b)$.

Proof. Consider

$$\begin{aligned}
& \max_{k,l} |(\widetilde{\Sigma}_b - \Sigma_b^0)_{k,l}| \\
= & \max_{k,l} \left| \left(\frac{\overline{\Sigma} - \widehat{\Sigma}_\varepsilon}{n_0} - \Sigma_b^0 \right)_{k,l} \right| \\
= & \max_{k,l} \left| \frac{(\overline{\Sigma} - n_0\Sigma_b^0 - \Sigma_\varepsilon^0)_{k,l}}{n_0} - \frac{(\widehat{\Sigma}_\varepsilon - \Sigma_\varepsilon^0)_{k,l}}{n_0} \right| \\
\leq & \max_{k,l} \left| \frac{(\overline{\Sigma} - n_0\Sigma_b^0 - \Sigma_\varepsilon^0)_{k,l}}{n_0} \right| + \max_{k,l} \left| \frac{(\widehat{\Sigma}_\varepsilon - \Sigma_\varepsilon^0)_{k,l}}{n_0} \right|. \qquad (3.21)
\end{aligned}$$

With $\bar{Y}_{i\cdot k} = W_{ik}$, we have

$$\begin{aligned}
\bar{Y}_{\cdot\cdot k} &= \frac{1}{N}\sum_{i=1}^{m} n_i W_{ik}, \\
(\overline{\Sigma})_{k,l} &= \frac{1}{m-1}\sum_{i=1}^{m} n_i \left(W_{ik} - \frac{1}{N}\sum_{i=1}^{m} n_i W_{ik}\right)\left(W_{il} - \frac{1}{N}\sum_{i=1}^{m} n_i W_{il}\right) \\
&= \frac{1}{m-1}\sum_{i=1}^{m} n_i W_{ik} W_{il} - \frac{1}{(m-1)N}\left(\sum_{i=1}^{m} n_i W_{ik}\right)\left(\sum_{i=1}^{m} n_i W_{il}\right).
\end{aligned}$$

Thus, we obtain

$$\begin{aligned}
&\frac{(\overline{\Sigma} - n_0\Sigma_b^0 - \Sigma_\varepsilon^0)_{k,l}}{n_0} \\
= \;& \frac{1}{n_0(m-1)}\sum_{i=1}^{m}\{n_i W_{ik} W_{il} - (n_i\Sigma_b^0 + \Sigma_\varepsilon^0)_{k,l}\} \\
& -\frac{1}{n_0(m-1)N}\left(\sum_{i=1}^{m} n_i W_{ik}\right)\left(\sum_{i=1}^{m} n_i W_{il}\right) \\
& +\left\{\frac{N}{n_0(m-1)} - 1\right\}(\Sigma_b^0)_{k,l} + \frac{1}{n_0(m-1)}(\Sigma_\varepsilon^0)_{k,l}.
\end{aligned}$$

Then, for the first term in (3.21), with $|(\Sigma_b^0)_{k,l}| \le M_b$ and $|(\Sigma_\varepsilon^0)_{k,l}| \le M_\varepsilon$, we have

$$\begin{aligned}
&\max_{k,l}\left|\frac{(\overline{\Sigma} - n_0\Sigma_b^0 - \Sigma_\varepsilon^0)_{k,l}}{n_0}\right| \\
\le \;& \max_{k,l}\frac{2}{n_0 m}\left|\sum_{i=1}^{m}\{n_i W_{ik} W_{il} - (n_i\Sigma_b^0 + \Sigma_\varepsilon^0)_{k,l}\}\right| \\
& + \max_{k}\frac{2}{n_0 m N}\left|\sum_{i=1}^{m} n_i W_{ik}\right|^2 + \left\{\frac{2N}{n_0 m} - 1\right\} M_b + \frac{2}{n_0 m} M_\varepsilon. \qquad (3.22)
\end{aligned}$$

Recall that by assumptions $b_{ik} \in \mathcal{SG}(\sigma_{b,k}^2)$, i.e., b_{ik} is sub-Gaussian with a variance factor $\sigma_{b,k}^2$ for $1 \le i \le m, 1 \le k \le p$. Then we have $W_{ik} \in SG(\sigma_{b,k}^2 + n_i^{-1}\sigma_{\varepsilon,k}^2)$. Let $\max_k \sigma_{\varepsilon,k}^2 = M_\varepsilon$ and $\max_k \sigma_{b,k}^2 = M_b$. Together with (3.3) and (3.4), we get $\|n_i W_{ik} W_{il} -$

$(n_i\Sigma_b^0+\Sigma_\varepsilon^0)_{k,l}\|_{\psi_1}\le c_1(n_iM_b+M_\varepsilon)$. Then, $n_iW_{ik}W_{il}-(n_i\Sigma_b^0+\Sigma_\varepsilon^0)_{k,l}$ is sub-Exponential. With Bernstein's inequality, for any k,l, we have

$$\mathbb{P}\mathrm{r}\left[\left|\frac{1}{n_0m}\sum_{i=1}^m\left\{n_iW_{ik}W_{il}-(n_i\Sigma_b^0+\Sigma_\varepsilon^0)_{k,l}\right\}\right|\ge t\right]\le 2\exp\left\{-c_6\min\left(\frac{t^2n_0^2m}{K_4^2},\frac{tn_0m}{K_4}\right)\right\},$$

where $c_6>0$, $K_4=\max_{i,k,l}\|n_iW_{ik}W_{il}-(n_i\Sigma_b^0+\Sigma_\varepsilon^0)_{k,l}\|_{\psi_1}\le 2c_1n_uM_*$, $n_u=\max_i n_i$, $M_*=\max(M_\varepsilon,M_b)$.

Take $t=C_1n_u(2n_0)^{-1}(\log p/m)^{1/2}$ for a sufficiently large constant $C_1>0$. With $m\ge\log p$ and the union sum inequality, we obtain

$$\begin{aligned}
&\mathbb{P}\mathrm{r}\left[\max_{k,l}\left|\frac{1}{n_0m}\sum_{i=1}^m\left\{n_iW_{ik}W_{il}-(n_i\Sigma_b^0+\Sigma_\varepsilon^0)_{k,l}\right\}\right|\ge t\right]\\
\le\ & 2p^2\exp\left\{-c_6\min\left(\frac{t^2n_0^2m}{4c_1^2n_u^2M_*^2},\frac{tn_0m}{2c_1n_uM_*}\right)\right\}\\
=\ & 2\exp\left[2\log p-\min\left\{\frac{c_6C_1^2}{16c_1^2M_*^2}\log p,\frac{c_6C_1}{4c_1M_*}(m\log p)^{1/2}\right\}\right]\\
\le\ & 2\exp\left[\left\{2-\min\left(\frac{c_6C_1^2}{16c_1^2M_*^2},\frac{c_6C_1}{4c_1M_*}\right)\right\}\log p\right].
\end{aligned}\tag{3.23}$$

Then we will bound the second term in (3.22). By the property of sub-Gaussian assumption, $n_iW_{ik}=n_ib_{ik}+\sum_{j=1}^{n_i}\varepsilon_{ijk}\in\mathcal{SG}(n_i^2M_b+n_iM_\varepsilon)$. Then, according to the general Hoeffding's inequality (Theorem 2.6.2 by Vershyin [58]), we have

$$\begin{aligned}
\mathbb{P}\mathrm{r}\left(\frac{1}{n_0mN}\left|\sum_{i=1}^m n_iW_{ik}\right|^2\ge t\right) &\le \mathbb{P}\mathrm{r}\left\{\left|\sum_{i=1}^m n_iW_{ik}\right|\ge(tn_0mN)^{1/2}\right\}\\
&\le 2\exp\left\{-\frac{c_7tn_0mN}{\sum_{i=1}^m c_0^2(n_i^2M_b+n_iM_\varepsilon)}\right\}\\
&\le 2\exp\left\{-\frac{c_7tn_0mN}{\sum_{i=1}^m c_0^2(n_in_uM_b+n_iM_\varepsilon)}\right\}\\
&\le 2\exp\left(-\frac{c_7tn_0m}{c_1n_uM_*}\right),
\end{aligned}$$

where $c_7 > 0$.

Then, take $t = C_1 n_u (2n_0)^{-1} (\log p/m)^{1/2}$, with $m \geq \log p$, by the union sum inequality, we have

$$\begin{aligned}
\mathbb{P}\mathrm{r}\left(\max_k \frac{1}{n_0 m N} \left| \sum_{i=1}^{m} n_i W_{ik} \right|^2 \geq t \right) &\leq 2p\exp\left(-\frac{c_7 t n_0 m}{c_1 n_u M_*} \right) \\
&\leq 2p\exp\left\{ -\frac{c_7 C_1}{2c_1 M_*} (m \log p)^{1/2} \right\} \\
&\leq 2\exp\left\{ \left(1 - \frac{c_7 C_1}{2c_1 M_*} \right) \log p \right\} \qquad (3.24)
\end{aligned}$$

for a sufficiently large constant $C_1 > 0$.

To bound the second term in (3.21), by Lemma 1, for a sufficiently large constant $C_2 > 0$, we have

$$\mathbb{P}\mathrm{r}\left\{ \max_{k,l} \left| \frac{(\widehat{\Sigma}_\varepsilon - \Sigma_\varepsilon^0)_{k,l}}{n_0} \right| > 2C_2 \frac{(N \log p)^{1/2}}{n_0 (N-m)} \right\} \leq 4p^{-C_3'}, \qquad (3.25)$$

where $C_3' > 0$ only depends on C_2 and M_ε.

Then, with

$$\widetilde{\lambda}_b = C_1 \frac{\max n_i}{n_0} \left(\frac{\log p}{m} \right)^{1/2} + C_2 \frac{(N \log p)^{1/2}}{n_0 (N-m)} + \frac{(2N - n_0 m) M_b}{2 n_0 m} + \frac{M_\varepsilon}{n_0 m}$$

for sufficiently large $C_1, C_2 > 0$, combining (3.21)-(3.25), we obtain

$$\begin{aligned}
& \mathbb{P}\mathrm{r}\left\{\max_{k,l}\left|(\widetilde{\Sigma}_b - \Sigma_b^0)_{k,l}\right| > 2\widetilde{\lambda}_b\right\} \\
\leq\ & \mathbb{P}\mathrm{r}\left[\max_{k,l}\left|\frac{2}{n_0 m}\sum_{i=1}^{m}\left\{n_i W_{ik}W_{il} - (n_i\Sigma_b^0 + \Sigma_\varepsilon^0)_{k,l}\right\}\right| \geq C_1\frac{\max_i n_i}{n_0}\left(\frac{\log p}{m}\right)^{1/2}\right] \\
& \quad + \mathbb{P}\mathrm{r}\left\{\max_{k}\frac{2}{n_0 m N}\left|\sum_{i=1}^{m} n_i W_{ik}\right|^2 \geq C_1\frac{\max_i n_i}{n_0}\left(\frac{\log p}{m}\right)^{1/2}\right\} \\
& \quad + \mathbb{P}\mathrm{r}\left\{\max_{k,l}\left|\frac{(\widehat{\Sigma}_\varepsilon - \Sigma_\varepsilon^0)_{k,l}}{n_0}\right| > 2C_2\frac{(N\log p)^{1/2}}{n_0(N-m)}\right\} \\
\leq\ & 4p^{-C_3''} + 4p^{-C_3'} \\
\leq\ & 8p^{-C_3},
\end{aligned}$$

where $C_3'' = \min\{c_6 C_1^2(16c_1^2 M_*^2)^{-1}, c_6 C_1(4c_1 M_*)^{-1}, c_7 C_1(2c_1 M_*)^{-1} + 1\} - 2$ and $C_3 = \min\{C_3', C_3''\}$.

□

Theorem 5 (Estimation error rate of $\widetilde{\Sigma}_b^+$) *Assume that the true between-subject covariance matrix $\Sigma_b^0 \in \mathcal{U}(M_b, s_b)$ and the true within-subject covariance matrix $\Sigma_\varepsilon^0 \in \mathcal{U}(M_\varepsilon, s_\varepsilon)$. Let*

$$\begin{aligned}
\widetilde{\lambda}_b \ =\ & C_1\frac{\max_i n_i}{n_0}\left(\frac{\log p}{m}\right)^{1/2} + C_2\frac{(N\log p)^{1/2}}{n_0(N-m)} \\
& + \frac{(2N - n_0 m)M_b}{2n_0 m} + \frac{M_\varepsilon}{n_0 m}
\end{aligned}$$

be the value of the tuning parameter λ in (2.5) for sufficiently large $C_1, C_2 > 0$. If $\log p \leq m$, then $\widetilde{\Sigma}_b^+$ satisfies

$$\left\|\widetilde{\Sigma}_b^+ - \Sigma_b^0\right\|_F \leq 10\widetilde{\lambda}_b(p s_b)^{1/2}$$

with probability at least $1-8p^{-C_3}$*, where* $C_3 > 0$ *only depends on* C_1*,* C_2 *and* $\max(M_\varepsilon, M_b)$*.*

We define a measure of data imbalance as $\max_i n_i/n_0 \geq 1$, where n_0 is defined in (3.20). In the balanced dataset where all n_i's are equal, we have $\max_i n_i/n_0 = 1$ and the two estimators coincide $\widehat{\Sigma}_b^+ = \widetilde{\Sigma}_b^+$. This equivalence is also reflected by the same estimation error rate since $\lambda_b = \widetilde{\lambda}_b$. When n_i's are not all equal, the imbalance $\max_i n_i/n_0 > 1$ increases with $\max_i n_i$ for fixed m and N. Comparing the first term in λ_b and $\widetilde{\lambda}_b$, the estimation error rate of $\widetilde{\Sigma}_b^+$ in the dimension p is strictly worse than that of $\widehat{\Sigma}_b^+$, which does not depend on the imbalance of the dataset. We numerically verify this comparison in Section 4, and demonstrate that the practical performance of $\widetilde{\Sigma}_b^+$ could be very sensitive to the imbalance of the data.

Chapter 4

Numerical Study for Sparse Covariance Estimations with Repeated Measurements

4.1 General Settings

In this chapter, we evaluate the numeric performance of our proposed estimators $\widehat{\Sigma}_{\varepsilon}^{+}$ (for the within-subject covariance Σ_{ε}) and $\widehat{\Sigma}_{b}^{+}$ (for the between-subject covariance Σ_{b}), and compare with $\overline{\Sigma}^{+}$ (in estimating either Σ_{b} or Σ_{ε}) and $\widetilde{\Sigma}_{b}^{+}$ (in estimating Σ_{b}).

In each of the subsequent subsections, we generate observations $\boldsymbol{Y}_{ij}$ from model (1.11), where $\boldsymbol{b}_{i} \sim \mathcal{N}(\boldsymbol{0}, \Sigma_{b}^{0})$ and $\boldsymbol{\varepsilon}_{ij} \sim \mathcal{N}(\boldsymbol{0}, \Sigma_{\varepsilon}^{0})$. All estimators in comparison are defined as solutions to the optimization problem (2.5) with corresponding input sample covariance matrices. We use a 5-fold cross-validation procedure in Section 2.4 to select the optimal tuning parameter value λ in (2.5) for each problem.

To illustrate the established theoretical results in Chapter 3, we consider the following models:

Model 1 *Banded matrices with bandwidth* 10*: set* $(\Sigma_b^0)_{j,k} = (1 - |j-k|/10)_+$ *and* $(\Sigma_\varepsilon^0)_{j,k} = (-1)^{|k_1-k_2|}(1 - |k_1-k_2|/10)_+$*;*

Model 2 *Covariance matrices corresponding to an* AR(1) *series: set* $(\Sigma_b^0)_{j,k} = 0.6^{|j-k|}$ *and* $(\Sigma_\varepsilon^0)_{j,k} = (-0.6)^{|j-k|}$*.*

We note that the same covariance structures had been used in Bickel and Levina [24], Xue et al. [29], Rothman [30], and Cui et al. [31].

4.2 Sanity Check for Positive-definiteness

We generate 100 independent data sets for both balanced Model 1 and Model 2 with $n_i = 2$, $m = 100$, and $p = 100$ or 200. We compare the performance of the unconstrained estimators, $\mathcal{S}_\lambda(\widehat{\Sigma}_\varepsilon)$ and $\mathcal{S}_\lambda(\widehat{\Sigma}_b)$, and the constrained estimators, $\widehat{\Sigma}_\varepsilon^+$ and $\widehat{\Sigma}_b^+$, in terms of estimation errors and the percentage of positive definite estimators, where $\mathcal{S}_\lambda()$ is the soft-thresholding operator defined in Section 2.2. The simulation results are summarized in Table 4.1. In general, the constrained estimators have slightly better performance in terms of estimation errors. In addition, we demonstrate that the positive definite constraint is crucial by observing that in most cases, the unconstrained estimators are not guaranteed to be positive definite, making them less qualified for interpretation or downstream statistical tasks.

4.3 General Comparison

In Chapter 3, we have shown that the estimation error rates of the estimators we study in this paper depend on various factors: the number of subjects m, the total number of observations N, the ambient dimension p, and for $\widetilde{\Sigma}_b^+$ the data imbalance, i.e.,

Table 4.1: Comparison of the unconstrained and constrained estimators under the balanced setting. Each metric is averaged over 100 replicates with the standard error shown in the parentheses. Comparisons are in terms of the estimation errors (F-error and L_2-error) and the percentage of positive definite estimators.

		Model 1		Model 2	
	p	100	200	100	200
		Within-Subject			
F-error	$\mathcal{S}_\lambda(\widehat{\Sigma}_\varepsilon)$	7.1804 (0.0562)	11.4040 (0.0490)	5.3956 (0.0202)	8.3116 (0.0159)
	$\widehat{\Sigma}_\varepsilon^+$	7.0548 (0.0552)	11.1804 (0.0490)	5.3956 (0.0202)	8.3116 (0.0159)
L_2-error	$\mathcal{S}_\lambda(\widehat{\Sigma}_\varepsilon)$	3.6179 (0.0451)	4.2217 (0.0282)	2.7131 (0.0115)	2.1257 (0.0083)
	$\widehat{\Sigma}_\varepsilon^+$	3.5553 (0.0438)	4.1564 (0.0286)	2.7131 (0.0115)	2.1257 (0.0083)
PD%	$\mathcal{S}_\lambda(\widehat{\Sigma}_\varepsilon)$	18%	3%	100%	100%
	$\widehat{\Sigma}_\varepsilon^+$	100%	100%	100%	100%
		Between-Subject			
F-error	$\mathcal{S}_\lambda(\widehat{\Sigma}_b)$	10.8195 (0.0611)	17.0538 (0.0416)	7.6064 (0.0212)	11.6116 (0.0187)
	$\widehat{\Sigma}_b^+$	10.1304 (0.0635)	16.1446 (0.0436)	7.5382 (0.0222)	11.6005 (0.0139)
L_2-error	$\mathcal{S}_\lambda(\widehat{\Sigma}_b)$	4.5419 (0.0447)	5.3739 (0.0258)	2.3508 (0.0104)	2.5681 (0.0051)
	$\widehat{\Sigma}_b^+$	4.2857 (0.0467)	5.0994 (0.0257)	2.3143 (0.0104)	2.5358 (0.0046)
PD%	$\mathcal{S}_\lambda(\widehat{\Sigma}_b)$	0%	0%	7%	12%
	$\widehat{\Sigma}_b^+$	100%	100%	100%	100%

$\mathcal{S}_\lambda(\widehat{\Sigma}_\varepsilon)$ and $\widehat{\Sigma}_\varepsilon^+$: unconstrained and constrained estimators for within-subject covariance;
$\mathcal{S}_\lambda(\widehat{\Sigma}_b)$ and $\widehat{\Sigma}_b^+$: unconstrained and constrained estimators for between-subject covariance;
PD%, percentage of positive definite estimators;
F-error: the Frobenius norm of $\widehat{Q} - Q^0$, i.e., $\|\widehat{Q} - Q^0\|_F$, where $\widehat{Q}$ is an estimate of the a generic parameter matrix Q^0;
L_2-error: the spectral norm of $\widehat{Q} - Q^0$, i.e., $\|\widehat{Q} - Q^0\|_2$.

$\max_i n_i/n_0$. We consider Model 1 and Model 2 in Section 4.1. In each setting, we let $N = 1000$ and $m = 100$ and consider $p = 100$ and $p = 200$. Furthermore, to study the effect of data imbalance on the estimation error, we set $n_i = a$ for $i = 1, 2, \ldots, 99$, where $a = \{3, 4, \cdots, 10\}$, and $n_{100} = N - 99a$. By doing so, we generate settings where the measure of data imbalance, $\max_i n_i/n_0$, varies.

Figure 4.1 summarizes the estimation error in the Frobenius norm averaged over 100 replications. We present the performance of four estimators: the proposed within-subject estimator $\widehat{\Sigma}_\varepsilon^+$ for estimating Σ_ε^0, and three between-subject estimators $\widehat{\Sigma}_b^+$ (our proposed method), $\widetilde{\Sigma}_b^+$ (the ANOVA type estimator), and $\overline{\Sigma}^+$ (the aggregated estimator) for estimating either Σ_b^0 or Σ_ε^0. Among the three between-subject estimators, our proposed method $\widehat{\Sigma}_b^+$ achieves the lowest estimation error in all simulation settings. Furthermore, being consistent with the results in Theorem 2, Theorem 3 and Theorem 5, the performance of $\widehat{\Sigma}_b^+$ and $\overline{\Sigma}^+$ are much less sensitive to the data imbalance $\max_i n_i/n_0$ while the error of $\widetilde{\Sigma}_b^+$ dramatically increases as the data become less balanced. Surprisingly, in all but the perfectly balanced case ($\max_i n_i/n_0 = 1$), we observe that $\widetilde{\Sigma}_b^+$, which is built on the unbiased sample estimate (3.20), performs much worse than $\overline{\Sigma}^+$ which is built on the biased $\overline{\Sigma}$ in (2.1). This suggests the dominating role of data imbalance in the estimation error of $\widetilde{\Sigma}_b^+$. Our proposed method $\widehat{\Sigma}_\varepsilon^+$ also achieves much lower estimation errors than $\overline{\Sigma}^+$ in estimating within-subject covariance in all simulation settings. The decreasing error of $\overline{\Sigma}^+$ in estimating Σ_ε^0 is consistent with Theorem 4, which states that the error rate of $\|\overline{\Sigma}^+ - \Sigma_\varepsilon^0\|_F$ is inversely proportional to the imbalance score $\max_i n_i/n_0$.

We also measure the support recovery performance of an estimator $\widehat{Q}$ for the true

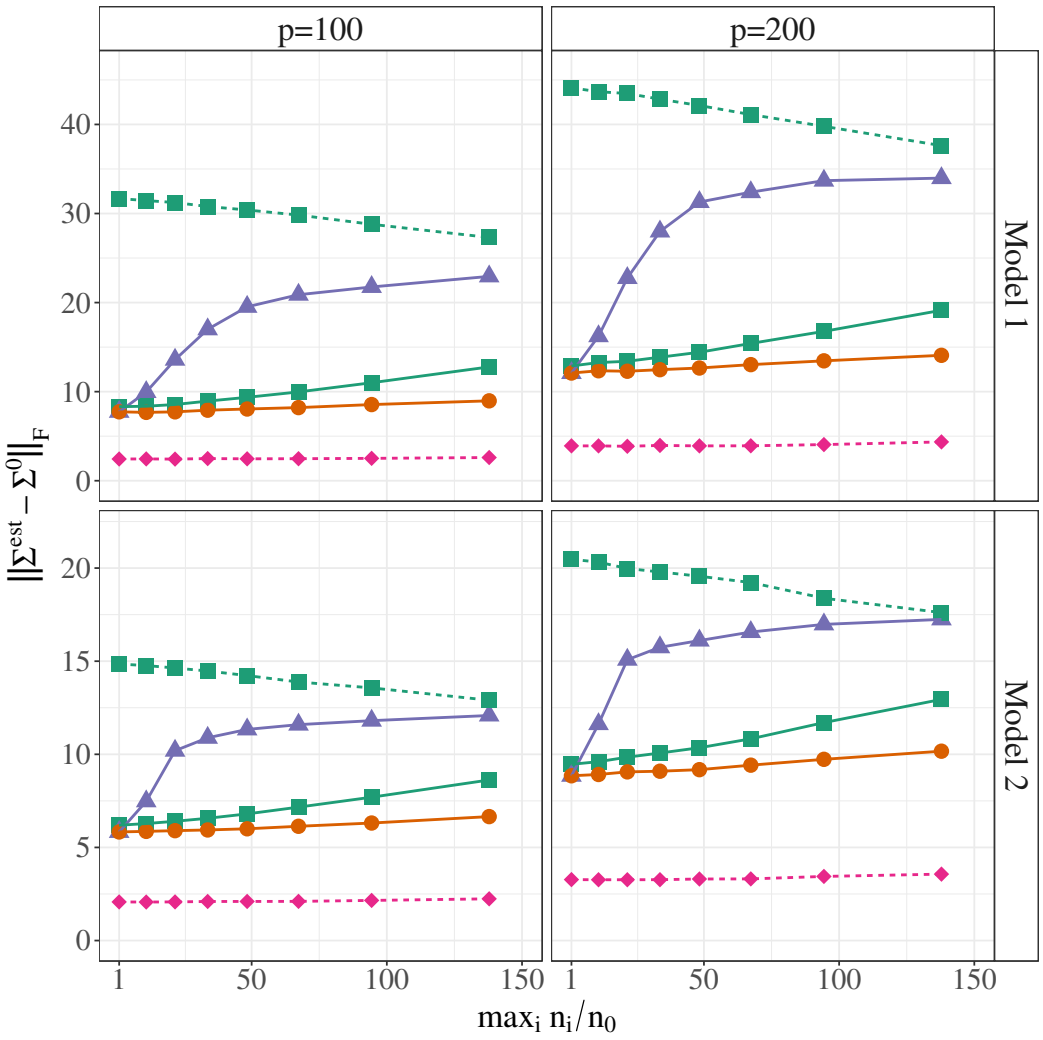

Figure 4.1: Estimation error (in Frobenius norm, averaged over 100 replicates) for two between-subject (solid) and one within-subject (dash) covariance matrix estimator: $\widetilde{\Sigma}_b^+$ (violet triangle), $\widehat{\Sigma}_b^+$ (orange circle), and $\widehat{\Sigma}_\varepsilon^+$ (pink diamond). The estimation error of the aggregated estimator ($\overline{\Sigma}^+$, green square) is evaluated in estimating the within-subject (dash) and the between-subject (solid) covariance matrices. The x-axis is $\max_i n_i/n_0$, which characterizes the imbalance of the data.

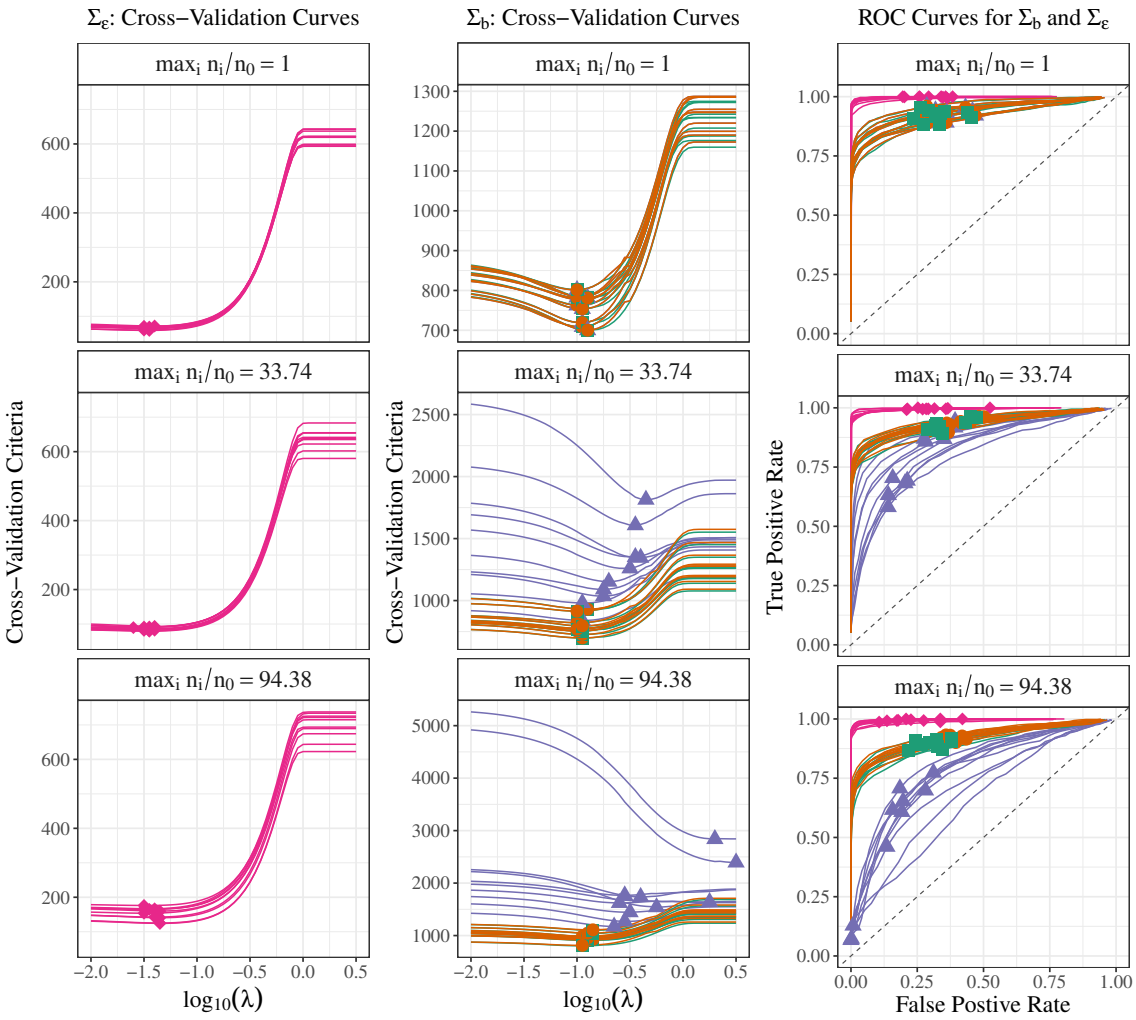

Figure 4.2: Cross-validation curves and receiver operating characteristic (ROC) curves between-subject and within-subject covariance sparsity recovery in Model 1 with $p = 100$ and different values of $\max_i n_i/n_0$. The top, middle and bottom rows correspond to different levels of data imbalance (with $a = 10$, 7, and 4, respectively). For simplicity of presentation, we randomly select 10 out of the 100 replicates. The left and middle panels exhibit 5-fold cross-validation curves of $\widehat{\Sigma}_{\varepsilon}^{+}$ (pink) for within-subject covariance, $\widehat{\Sigma}_{b}^{+}$ (orange), $\widetilde{\Sigma}_{b}^{+}$ (violet), and $\overline{\Sigma}^{+}$ (green) for between-subject covariance. Diamonds ($\widehat{\Sigma}_{\varepsilon}^{+}$), circles ($\widehat{\Sigma}_{b}^{+}$), triangles ($\widetilde{\Sigma}_{b}^{+}$), and squares ($\overline{\Sigma}^{+}$) in these two panels mark the minimum points on these curves. The right panels present the ROC curves. The diamonds ($\widehat{\Sigma}_{\varepsilon}^{+}$), circles ($\widehat{\Sigma}_{b}^{+}$), triangles ($\widetilde{\Sigma}_{b}^{+}$), and squares ($\overline{\Sigma}^{+}$) represent the true positive rate and false positive rate with λ values selected by the 5-fold cross-validation.

parameter matrix Q^0 using true positive rate (TPR) and false positive rate (FPR),

$$\mathrm{TPR}(\widehat{Q}, Q^0) = \frac{\#\{(i,j) : \widehat{Q}_{i,j} \neq 0 \text{ and } Q^0_{i,j} \neq 0\}}{\#\{(i,j) : Q^0_{i,j} \neq 0\}}, \tag{4.1}$$

$$\mathrm{FPR}(\widehat{Q}, Q^0) = \frac{\#\{(i,j) : \widehat{Q}_{i,j} \neq 0 \text{ and } Q^0_{i,j} = 0\}}{\#\{(i,j) : Q^0_{i,j} = 0\}}. \tag{4.2}$$

To demonstrate the effectiveness of regularization, in Figure 4.2, we present the cross-validation curves and the receiver operating characteristic (ROC) of the sparsity recovery of these estimators in Model 1 with $p = 100$ and under three different levels of data imbalance. The optimal values of λ for $\widehat{\Sigma}^+_\varepsilon$, $\widehat{\Sigma}^+_b$, and $\overline{\Sigma}^+$ are relatively stable across different levels of data imbalance, while the optimal value of λ for $\widetilde{\Sigma}^+_b$ sharply fluctuates and generally increases with $\max_i n_i/n_0$. This indicates that large values of $\max_i n_i/n_0$ tend to result in more shrinkage of the off-diagonal entries in $\widetilde{\Sigma}^+_b$ towards 0. This observation is aligned with the larger error of $\widetilde{\Sigma}^+_b$ in Frobenius norm in Figure 4.1 for large values of $\max_i n_i/n_0$. While the theoretical guarantees of support recovery would be an interesting and challenging problem for future research, we observe numerically that the data imbalance seems not to affect the support recovery performance of $\widehat{\Sigma}^+_\varepsilon$, $\widehat{\Sigma}^+_b$, and $\overline{\Sigma}^+$, which is an established favorable properties of these estimators in terms of estimation error. In contrast, just as in estimation error, $\widetilde{\Sigma}^+_b$ suffers in sparsity recovery performance from the data imbalance.

4.4 Understanding the Effects of the Bias in Sample Estimates

As seen in Figure 4.1 and Figure 4.2, the estimator $\overline{\Sigma}^+$ based on the biased sample estimate $\overline{\Sigma}$ surprisingly has relatively acceptable numerical performance. This subsection

investigates this observation by comparing our proposed between-subject estimator $\widehat{\Sigma}_b^+$ with $\overline{\Sigma}^+$. We consider two modifications of Model 1 as follows:

Model 3 *For any given $a > 0$, we set $(\Sigma_b^0)_{j,k} = (1 - |j-k|/10)_+$ and $(\Sigma_\varepsilon^0)_{j,k} = a(1 - |j-k|/10)_+$.*

Model 4 *For any given $a > 0$, we set $(\Sigma_b^0)_{j,k} = (1 - |j-k|/10)_+$ and $(\Sigma_\varepsilon^0)_{j,k} = a(-1)^{|j-k|}(1 - |j-k|/10)_+$.*

From (2.2), the matrix of Σ_ε can be considered as the additive noise for the task of estimating Σ_b. We thus define the inverse signal-to-noise ratio as $|\Sigma_\varepsilon^0|_\infty/|\Sigma_b^0|_\infty$. By varying $|\Sigma_\varepsilon^0|_\infty/|\Sigma_b^0|_\infty = a \in \{1, 2, \ldots, 10\}$ in Model 3 and Model 4, we construct settings where the relative signal strength from Σ_ε and Σ_b is different. In comparison with Model 3, we alternate the signs of sub-diagonal elements in Σ_ε^0 in Model 4. In both models, we generate balanced data with $n_i = 5$ for $i = 1, \ldots, m = 100$ and $p = 50$. Estimation errors in Frobenius norm are summarized (over 100 replications) in Figure 4.3.

In general, our proposed between-subject sample estimate $\widehat{\Sigma}_b$ significantly outperforms $\overline{\Sigma}$ in both examples. This demonstrates the effect of the bias correction as in (2.2). Moreover, for both sample estimators, their regularized versions (dash lines) achieve lower estimation errors, indicating the benefit of regularization.

Surprisingly, as $|\Sigma_\varepsilon^0|_\infty/|\Sigma_b^0|_\infty$ gets relatively small, $\overline{\Sigma}^+$ achieves an even smaller estimation error than $\widehat{\Sigma}_b^+$. This is an interesting cancellation of two biases with opposite signs: the estimation bias in the sample estimate $\overline{\Sigma}$ and the shrinkage bias in the ℓ_1-penalty. Specifically, for any index pair (j, k), (2.2) indicates that the bias of $\overline{\Sigma}_{j,k}$ in estimating $(\Sigma_b^0)_{j,k}$ is $\sum_i (mn_i)^{-1}(\Sigma_\varepsilon^0)_{j,k}$. In cases where $(\Sigma_\varepsilon^0)_{j,k}$ and $(\Sigma_b^0)_{j,k}$ have the same signs (as in Model 3), this sample estimation bias has the opposite effect from the shrinkage bias from the ℓ_1 penalty. Consequently, these two biases could cancel each other when they have

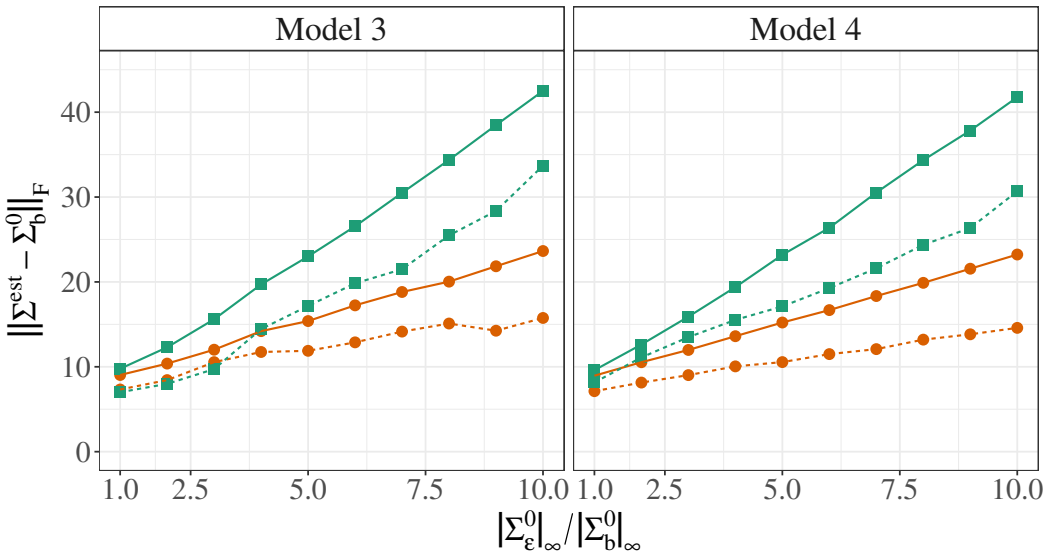

Figure 4.3: Estimation error (in Frobenius norm, averaged over 100 replicates) of the two between-subject sample covariance (solid) estimators ($\overline{\Sigma}$ and $\widehat{\Sigma}_b$) and their corresponding sparse and positive definite (dash) covariance estimators ($\overline{\Sigma}^+$ and $\widehat{\Sigma}_b^+$). The horizontal axis is the inverse signal-to-noise ratio, i.e., $|\Sigma_\varepsilon^0|_\infty / |\Sigma_b^0|_\infty$. The estimation errors of $\overline{\Sigma}$ and $\overline{\Sigma}^+$ are marked in green, and the estimation errors of $\widehat{\Sigma}_b$ and $\widehat{\Sigma}_b^+$ are marked in orange.

similar magnitudes, which is achieved when $(\Sigma_\varepsilon^0)_{j,k}$ is on a similar scale as λ, and thus resulting in a surprisingly better performance of $\widehat{\Sigma}_b$ than $\widehat{\Sigma}_b^+$. Notably, when the estimation bias (as characterized by $|\Sigma_\varepsilon^0|_\infty/|\Sigma_b^0|_\infty$) is too large to be canceled by the shrinkage bias, or when both biases have the same signs (as in Model 4), the performance of $\widehat{\Sigma}_b^+$ is dominating that of $\overline{\Sigma}^+$.

4.5 Covariance Graphs of Clinical Measurements Collected from Hemodialysis Patients

We apply our proposed methods to estimate the between-subject and within-subject covariance structures among some clinical variables collected from hemodialysis patients. Hemodialysis is a treatment that filters wastes and fluid from patients' blood when the kidneys no longer function well. Hemodialysis patients usually follow a strict schedule by visiting a dialysis center about three times a week. Clinical variables, such as blood pressure and pulse, are measured during each treatment. Since numerous metabolic changes accompanying impaired kidney function affect all organ systems of the human body, it is imperative to study correlations among clinical variables. Those clinical variables are measured repeatedly for each hemodialysis patient at each treatment. We will investigate correlation structures at the patient (between-patient) and treatment (within-patient) levels.

We use a dataset of measurements of several clinical and laboratory variables during 2018 and 2021 from 5,000 hemodialysis patients. For homogeneity, we consider white, non-diabetic, and non-Hispanic male patients who never had a COVID-19-positive polymerase chain reaction test. We use the measurements starting from the second year to avoid large fluctuations in the first year of dialysis. The dataset contains 276 patients

with at least three complete treatment records every 30 days. The data imbalance is $\max_i n_i/n_0 = 2.54$. For simplicity, we focus on the relationships among interdialytic weight gain, blood pressure, and heart rate. Based on Ipema et al. [59], we consider the following eight variables: `idwg` (interdialytic weight gain, kg), `ufv` (ultrafiltration volume, L), `min_sbp` (minimum systolic blood pressure, mmHg), `min_dbp` (minimum diastolic blood pressure, mmHg), `max_sbp` (maximum systolic blood pressure, mmHg), `max_dbp` (maximum diastolic blood pressure, mmHg), `min_pulse` (minimum pulse, beats/min), and `max_pulse` (maximum pulse, beats/min). In our analysis, `ufv` is set to be the difference between predialysis and postdialysis weight within a hemodialysis session.

We are interested in recovering the correlation structures at the patient and the treatment levels. Estimating the correlation matrix corresponds to recovering the correlation graph, where the nodes represent the random variables of interest and the edges present the marginal correlation between the nodes (Chaudhuri et al. [9]). We apply our method to repeated clinical measurements from these 276 patients. The regularization parameters are chosen by 5-fold cross-validation with the one standard error rule (Hastie et al. [60]). Figure 4.4 presents estimates of the within-subject (top right panel) and between-subject (top left panel) correlations, which indeed present different correlation structures. We also include the estimate using the aggregated data (bottom left panel) for comparison, which coincides with our between-subject estimate. This is consistent with Theorem 3 for this dataset's small value of $\max_i n_i/n_0$.

It is important to realize that covariance structures at the treatment and patient levels could differ and should be estimated separately. Existing biological studies based on the aggregated measurements ignore such a difference in estimation and thus may lead to erroneous conclusions. In particular, our estimated correlation graph at the treatment level (within-subject) reveals much insight for hemodialysis treatment that cannot be recovered using the aggregate data. Specifically, we discuss several important recovered

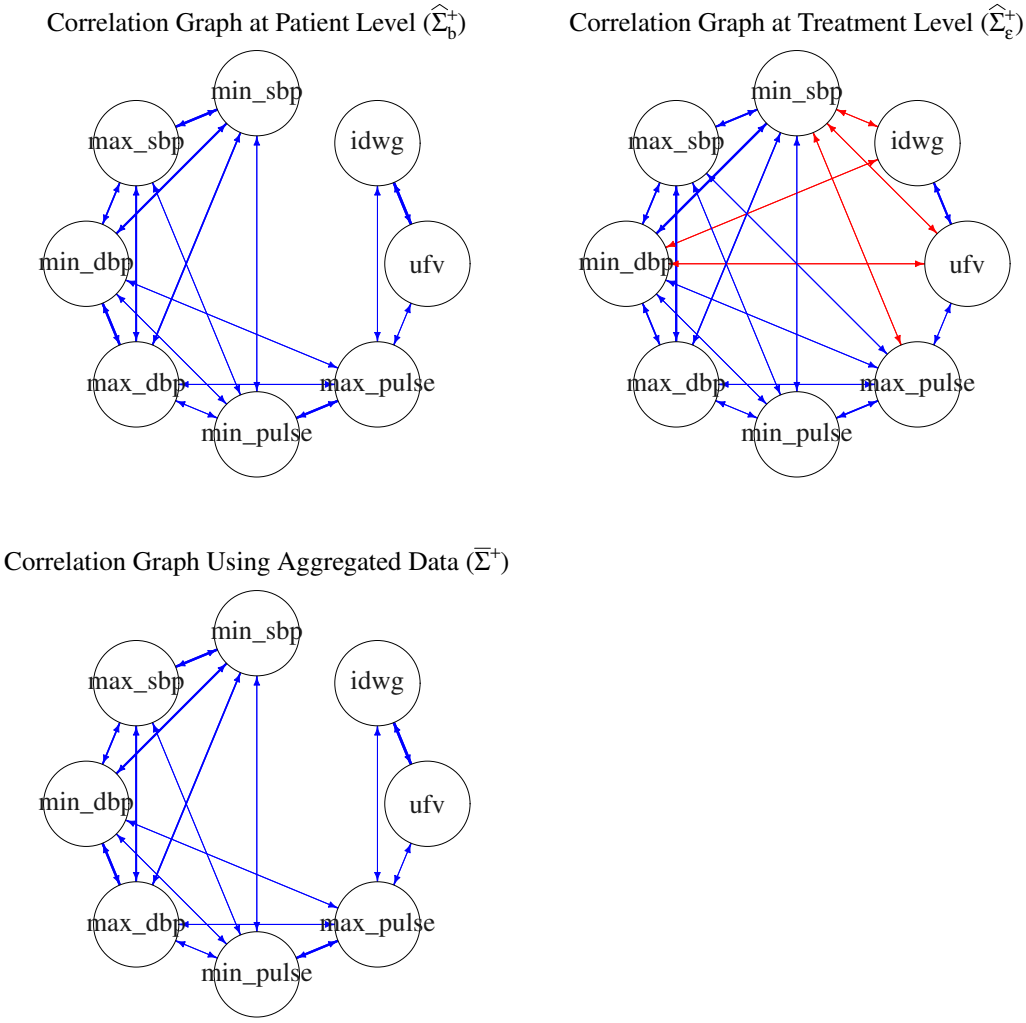

Figure 4.4: Between-subject (top left) correlation graphs, within-subject (top right), and correlation graph using the aggregated data (bottom left) for clinical variables from hemodialysis patients. We present correlation matrices with the convention of using bi-directed covariance graphs (Chaudhuri et al. [9]). The blue edges correspond to the positive correlations, while the red edges represent the negative correlations. The width of an edge corresponds to the strength of the correlation.

correlations in $\widehat{\Sigma}_{\varepsilon}^{+}$ that have been missed in either $\widehat{\Sigma}_{b}^{+}$ or $\overline{\Sigma}^{+}$. Specifically, salt and fluid intake between two hemodialysis sessions leads to interdialytic weight gain. A dialyzer, an artificial kidney, should filter the cumulation of waste and fluid. Ultrafiltration volume measures the waste and fluid removed from patients' blood. Consequently, higher `idwg` leads to larger `ufv`, confirmed by the positive correlation between `idwg` and `ufv` at the treatment level in Figure 4.4. A rapid removal of fluid from a patient's blood results in the depletion of blood volume and subsequently leads to a decrease in systolic blood pressure, confirmed by the negative correlation between `ufv` and `min_sbp` at the treatment level in Figure 4.4. The lowered blood pressure will be compensated by heart functionality, which elevates the heart rate, again confirmed by the negative correlation between `min_sbp` and `max_pulse` at the treatment level in Figure 4.4. However, no relationships among `idwg`, `max_pulse` and `min_sbp` have been observed at patient level in the middle panel of Figure 4.4. This implies that we should focus on correlations between clinical measurements at the treatment level rather than the patient level when evaluating the effectiveness of hemodialysis.

Chapter 5

Sparse Graph Estimation for Graphical VAR Models with Repeated Measurements

5.1 Introduction

The graphical VAR model is a framework that merges the principles of graphical models with those of VAR models to analyze multivariate time series data. This synthesis allows for identifying the dynamic interrelationships and conditional dependencies among several time-dependent variables, which has wide applications in economics, finance, neuroscience, environmental science, and clinical research. For example, in Wild et al. [36], the graphical VAR model was used to model dynamic dependence structures and feedback mechanisms between symptom-relevant variables with the electronic diary data from 35 obese German patients. In this study, symptom-relevant variables were monitored daily for each patient. Despite the temporal dependencies among symptom-relevant variables, these repeated daily measurements could also result in the dependence among

observations within each patient. Therefore, in this chapter, we are interested in uncovering the potential dynamic Granger-causal relationships, conditional within-subject, and between-subject dependencies among multiple variables simultaneously.

Let $\boldsymbol{Y}_{it} = (Y_{it1}, Y_{it2}, \ldots, Y_{itp})^{\mathrm{T}} \in \mathbb{R}^p$ be the observation of p random variables $Y_1, \ldots, Y_p$ at time t from subject i. Recall the GVAR(1) model with random effect in (1.12) in Section 1.5,

$$\boldsymbol{Y}_{it} = \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{Y}_{i(t-1)} + \boldsymbol{b}_i + \boldsymbol{\varepsilon}_{it}, \quad i = 1, \ldots, m;\ t = 1, \ldots, n_i,$$

where $\boldsymbol{\beta} \in \mathbb{R}^{p \times p}$ is the fixed-effect coefficient matrix, $\boldsymbol{b}_i = (b_{i1}, b_{i2}, \ldots, b_{ip})^{\mathrm{T}} \in \mathbb{R}^p$ is a p-variate i.i.d. normal random variable with mean $\mathbf{0}$ and variance $\Sigma_b = \Omega_b^{-1}$, $\boldsymbol{\varepsilon}_{it} = (\varepsilon_{it1}, \varepsilon_{it2}, \ldots, \varepsilon_{itp})^{\mathrm{T}} \in \mathbb{R}^p$ is a p-variate i.i.d. normal random variable with mean $\mathbf{0}$ and variance $\Sigma_\varepsilon = \Omega_\varepsilon^{-1}$. Moreover, $\boldsymbol{b}_i$ and $\boldsymbol{\varepsilon}_{it}$ are mutually independent. In model (1.12), we use $\boldsymbol{b}_i$ to account for the dependence among variables of interest introduced by repeated measurements. Recall model (1.13) with fixed $\boldsymbol{\beta}$,

$$\boldsymbol{Y}_{it} = \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{Y}_{i(t-1)} + (I - \boldsymbol{\beta}^{\mathrm{T}}) \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_{it}, \quad i = 1, \ldots, m;\ t = 1, \ldots, n_i, \tag{5.1}$$

$\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{ip})^{\mathrm{T}} \in \mathbb{R}^p$ is a p-variate i.i.d. normal random variable with mean $\mathbf{0}$ and variance $\Sigma_\mu = \Omega_\mu^{-1}$, and $\boldsymbol{\mu}_i$ and $\boldsymbol{\varepsilon}_{it}$ are mutually independent. Then, we know that

$$(I - \boldsymbol{\beta}^{\mathrm{T}}) \boldsymbol{\mu}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\{\mathbf{0}, (I - \boldsymbol{\beta}^{\mathrm{T}}) \Omega_\mu^{-1} (I - \boldsymbol{\beta})\}.$$

The random effect term, $(I - \boldsymbol{\beta}^{\mathrm{T}}) \boldsymbol{\mu}_i$, in (5.1) plays the same role as $\boldsymbol{b}_i$ in model (1.12). Epskamp et al. [45] use the aggregated data across subjects $\{\bar{\boldsymbol{Y}}_{1\cdot}, \ldots, \bar{\boldsymbol{Y}}_{m\cdot}\}$ to find the

between-subject precision matrix Ω_μ via Glasso, in which they estimate

$$\left\{\Omega_\mu^{-1} + \frac{1}{n}(I - \boldsymbol{\beta}^{\mathrm{T}})^{-1}\Sigma_\varepsilon(I - \boldsymbol{\beta})^{-1}\right\}^{-1}$$

in place of Ω_μ as the between-subject precision matrix, when $I - \boldsymbol{\beta}^{\mathrm{T}}$ is invertible and $n_i = n$ for $i = 1, \ldots, m$. It could lead to erroneous conditional dependencies among variables of interest. Besides the bias introduced by the aggregated data, the estimator of the precision matrix based on the maximum likelihood method could also be biased (Sun and Sun [61]).

Let $N = \sum_{i=1}^{m} n_i$. For the i-th subject, with model (1.12), we have the following matrix form,

$$\underbrace{\begin{pmatrix} \boldsymbol{Y}_{i1}^{\mathrm{T}} \\ \boldsymbol{Y}_{i2}^{\mathrm{T}} \\ \vdots \\ \boldsymbol{Y}_{in_i}^{\mathrm{T}} \end{pmatrix}}_{\boldsymbol{Y}_i} = \underbrace{\begin{pmatrix} \boldsymbol{Y}_{i0}^{\mathrm{T}} \\ \boldsymbol{Y}_{i1}^{\mathrm{T}} \\ \vdots \\ \boldsymbol{Y}_{i(n_i-1)}^{\mathrm{T}} \end{pmatrix}}_{\boldsymbol{X}_i} \underbrace{\begin{pmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1p} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ \beta_{p1} & \beta_{p2} & \cdots & \beta_{pp} \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}}_{\mathbf{1}_{n_i}} \underbrace{\begin{pmatrix} b_{i1} & b_{i2} & \cdots & b_{ip} \end{pmatrix}}_{\boldsymbol{b}_i^{\mathrm{T}}} + \underbrace{\begin{pmatrix} \boldsymbol{\varepsilon}_{i1}^{\mathrm{T}} \\ \boldsymbol{\varepsilon}_{i2}^{\mathrm{T}} \\ \vdots \\ \boldsymbol{\varepsilon}_{in_i}^{\mathrm{T}} \end{pmatrix}}_{\boldsymbol{E}_i},$$

That is

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \mathbf{1}_{n_i}\boldsymbol{b}_i^{\mathrm{T}} + \boldsymbol{E}_i, \tag{5.2}$$

where $\mathbf{1}_{n_i}$ is the vector of all ones with length n_i. Finally, we concatenate the above

matrix expressions for m groups,

$$\underbrace{\begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \\ \vdots \\ \boldsymbol{Y}_m \end{pmatrix}}_{\boldsymbol{Y}} = \underbrace{\begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_m \end{pmatrix}}_{\boldsymbol{X}} \boldsymbol{\beta} + \underbrace{\begin{pmatrix} \mathbf{1}_{n_1} & & & \\ & \mathbf{1}_{n_2} & & \\ & & \ddots & \\ & & & \mathbf{1}_{n_m} \end{pmatrix}}_{\boldsymbol{Z}} \underbrace{\begin{pmatrix} \boldsymbol{b}_1^{\mathrm{T}} \\ \boldsymbol{b}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{b}_m^{\mathrm{T}} \end{pmatrix}}_{\boldsymbol{B}} + \underbrace{\begin{pmatrix} \boldsymbol{E}_1 \\ \boldsymbol{E}_2 \\ \vdots \\ \boldsymbol{E}_m \end{pmatrix}}_{\boldsymbol{E}},$$

i.e.,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{B} + \boldsymbol{E}. \tag{5.3}$$

For simplicity, we assume that columns of $\boldsymbol{X}$ and $\boldsymbol{Y}$ have been centered. In this chapter and the remainder of this dissertation, with model (5.3), we are interested in recovering temporal, contemporaneous, and between-subjects networks obtained from $\boldsymbol{\beta}$, Ω_ε and Ω_b, respectively.

5.2 A Two-stage Estimation Method for GVAR Models

With the normality assumption of $\boldsymbol{b}_i$ and $\boldsymbol{\varepsilon}_{it}$, we have that

$$\mathrm{vec}(\boldsymbol{Y}_i^{\mathrm{T}}) \sim \mathcal{N}[\mathrm{vec}\{(\boldsymbol{X}_i\boldsymbol{\beta})^{\mathrm{T}}\}, \mathbf{1}_{n_i}\mathbf{1}_{n_i}^{\mathrm{T}} \otimes \Omega_b^{-1} + \boldsymbol{I}_{n_i} \otimes \Omega_\varepsilon^{-1}].$$

Let $\boldsymbol{A}_i = \boldsymbol{I}_{n_i} - \frac{1}{n_i}\boldsymbol{J}_{n_i}$ and $\boldsymbol{P} = \mathrm{diag}(\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m)$, where $\boldsymbol{I}_{n_i}$ is a $n_i \times n_i$ identity matrix, and $\boldsymbol{J}_{n_i}$ is a $n_i \times n_i$ matrix of ones. Then, given $\{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m\}$, the log-likelihood function

$\tilde{\ell}(\boldsymbol{\beta}, \Omega_\varepsilon, \Omega_b)$ is

$$\begin{aligned}\tilde{\ell}(\boldsymbol{\beta}, \Omega_\varepsilon, \Omega_b) &= -\frac{Np}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{m}\log\det(\Omega_\varepsilon^{-1} + n_i\Omega_b^{-1}) + \frac{N-m}{2}\log\det\Omega_\varepsilon \\ &\quad -\frac{1}{2}\mathrm{tr}\left\{\Omega_\varepsilon(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{P}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})\right\} \\ &\quad -\frac{1}{2}\mathrm{tr}\left\{\sum_{i=1}^{m}\frac{1}{n_i}\left(\Omega_\varepsilon^{-1}+n_i\Omega_b^{-1}\right)^{-1}(\boldsymbol{Y}_i-\boldsymbol{X}_i\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{J}_{n_i}(\boldsymbol{Y}_i-\boldsymbol{X}_i\boldsymbol{\beta})\right\}. \quad (5.4)\end{aligned}$$

We note that the likelihood function is not jointly convex in $\boldsymbol{\beta}$, Ω_ε and Ω_b. Consequently, solving penalized likelihood to simultaneously estimate these three matrices is computationally expensive. Therefore, we develop a projection and correct pooling framework for GVAR(1) model (PCP-GVAR) to recover the multilevel networks in (1.12).

5.2.1 Estimation of Fixed Effect and Within-Subject Precision Matrix

In the first stage (Stage 1), we will estimate the fixed effect $\boldsymbol{\beta}$ and the within-subject precision matrix Ω_ε simultaneously. We first remove the random effect part by within-group centering, i.e.,

$$\boldsymbol{A}_i\boldsymbol{Y}_i = \boldsymbol{A}_i\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{A}_i\boldsymbol{E}_i.$$

With the projection matrix $\boldsymbol{P}$, we have $\boldsymbol{P}\boldsymbol{Y} = \boldsymbol{P}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{P}\boldsymbol{E}$. By the normality assumption, it is easy to show that

$$\mathrm{vec}(\boldsymbol{P}\boldsymbol{Y})|\boldsymbol{X} \sim \mathcal{N}\{\mathrm{vec}(\boldsymbol{P}\boldsymbol{X}\boldsymbol{\beta}), \Sigma_\varepsilon \otimes \boldsymbol{P}\}.$$

Note that $\Sigma_\varepsilon \otimes \boldsymbol{P}$ is non-invertible, we will have a degenerate distribution with the following probability density function,

$$\begin{aligned} f\{\mathrm{vec}(\boldsymbol{PY})\} &= (2\pi)^{-Np/2}\{\mathrm{pdet}(\Sigma_\varepsilon \otimes \boldsymbol{P})\}^{-1/2} \times \\ &\qquad \exp\left[-\frac{1}{2}\mathrm{vec}\{\boldsymbol{P}(\boldsymbol{Y}-\boldsymbol{X\beta})\}^{\mathrm{T}}(\Sigma_\varepsilon \otimes \boldsymbol{P})^{+}\mathrm{vec}\{\boldsymbol{P}(\boldsymbol{Y}-\boldsymbol{X\beta})\}\right], \end{aligned}$$

where $\mathrm{pdet}(A)$ and A^+ are the pseudo-determiant and pseudo-inverse matrix for a generic matrix A. Then two times the negative log-likelihood function for Σ_ε and $\boldsymbol{\beta}$, and ignoring a constant term, is

$$\ell(\Sigma_\varepsilon, \boldsymbol{\beta}) = \log\ \mathrm{pdet}(\Sigma_\varepsilon \otimes \boldsymbol{P}) + \mathrm{vec}\{\boldsymbol{P}(\boldsymbol{Y}-\boldsymbol{X\beta})\}^{\mathrm{T}}(\Sigma_\varepsilon \otimes \boldsymbol{P})^{+}\mathrm{vec}\{\boldsymbol{P}(\boldsymbol{Y}-\boldsymbol{X\beta})\}. \quad (5.5)$$

According to Castañeda and Nossek [62] and Petersen and Pedersen [63], we have

$$\begin{aligned} \mathrm{pdet}(\Sigma_\varepsilon \otimes \boldsymbol{P}) &= \mathrm{pdet}(\Sigma_\varepsilon)^{\mathrm{rank}(\boldsymbol{P})}\mathrm{pdet}(\boldsymbol{P})^{\mathrm{rank}(\Sigma_\varepsilon)} \\ &= \det(\Omega_\varepsilon)^{-(N-m)}\mathrm{pdet}(\boldsymbol{P})^{p}, \\ (\Sigma_\varepsilon \otimes \boldsymbol{P})^{+} &= \Sigma_\varepsilon^{+} \otimes \boldsymbol{P}^{+} = \Omega_\varepsilon \otimes \boldsymbol{P}^{+}. \end{aligned}$$

By Theorem 16.2.2 in Harville [64], we obtain

$$\begin{aligned} & [\mathrm{vec}\{\boldsymbol{P}(\boldsymbol{Y}-\boldsymbol{X\beta})\}]^{\mathrm{T}}(\Omega_\varepsilon \otimes \boldsymbol{P}^{+})\mathrm{vec}\{\boldsymbol{P}(\boldsymbol{Y}-\boldsymbol{X\beta})\} \\ = \ & \mathrm{tr}[\{\boldsymbol{P}(\boldsymbol{Y}-\boldsymbol{X\beta})\}^{\mathrm{T}}\boldsymbol{P}^{+}\boldsymbol{P}(\boldsymbol{Y}-\boldsymbol{X\beta})\Omega_\varepsilon] \\ = \ & \mathrm{tr}\{(\boldsymbol{Y}-\boldsymbol{X\beta})^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P}^{+}\boldsymbol{P}(\boldsymbol{Y}-\boldsymbol{X\beta})\Omega_\varepsilon\} \\ = \ & \mathrm{tr}\{(\boldsymbol{Y}-\boldsymbol{X\beta})^{\mathrm{T}}\boldsymbol{P}(\boldsymbol{Y}-\boldsymbol{X\beta})\Omega_\varepsilon\} \\ = \ & \mathrm{tr}\{(\boldsymbol{Y}-\boldsymbol{X\beta})^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P}(\boldsymbol{Y}-\boldsymbol{X\beta})\Omega_\varepsilon\} \\ = \ & \mathrm{tr}\{(\widetilde{\boldsymbol{Y}}-\widetilde{\boldsymbol{X}}\boldsymbol{\beta})^{\mathrm{T}}(\widetilde{\boldsymbol{Y}}-\widetilde{\boldsymbol{X}}\boldsymbol{\beta})\Omega_\varepsilon\}, \end{aligned} \quad (5.6)$$

where $\widetilde{\boldsymbol{Y}} = \boldsymbol{P}\boldsymbol{Y}$ and $\widetilde{\boldsymbol{X}} = \boldsymbol{P}\boldsymbol{X}$. Therefore, the negative log-likelihood function for Ω_ε and $\boldsymbol{\beta}$ in (5.5) can be simplified to

$$\ell(\Omega_\varepsilon, \boldsymbol{\beta}) \;=\; \mathrm{tr}\left\{(\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta})^{\mathrm{T}}(\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta})\Omega_\varepsilon\right\} - (N-m)\log\det\Omega_\varepsilon. \tag{5.7}$$

The objective function $\ell(\Omega_\varepsilon, \boldsymbol{\beta})$ in (5.7) is equivalent to (1.1) in Rothman et al. [15], differing only by a constant multiplier. To obtain a spare Ω_ε and $\boldsymbol{\beta}$, we consider the following minimization problem,

$$(\widehat{\Omega}_\varepsilon, \widehat{\boldsymbol{\beta}}) = \mathrm{argmin}_{\Omega_\varepsilon, \boldsymbol{\beta}}\, \ell(\Omega_\varepsilon, \boldsymbol{\beta}) + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\Omega_\varepsilon\|_1, \tag{5.8}$$

where $\|\cdot\|_1$ is the ℓ_1-norm of the input matrix, which is defined in Section 1.2. The negative log-likelihood function $\ell(\Omega_\varepsilon, \boldsymbol{\beta})$ is biconvex rather than jointly convex in $(\boldsymbol{\beta}, \Omega_\varepsilon)$. Thus, Rothman et al. [15] propose the alternating two-step estimators, which do not rely on the convexity of the full likelihood function to solve (5.8).

With a fixed $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, the optimization problem (5.8) is equivalent to

$$\widehat{\Omega}_\varepsilon(\boldsymbol{\beta}_0) = \mathrm{argmin}_{\Omega_\varepsilon}\, \mathrm{tr}\left\{S_\varepsilon(\boldsymbol{\beta}_0)\Omega_\varepsilon\right\} - \log\det\Omega_\varepsilon + \tilde{\lambda}_2\|\Omega_\varepsilon\|_1, \tag{5.9}$$

where $S_\varepsilon(\boldsymbol{\beta}_0) = \frac{1}{N-m}(\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta}_0)^{\mathrm{T}}(\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta}_0)$, and $\tilde{\lambda}_2 = (N-m)^{-1}\lambda_2$. The optimization problem in (5.9) is considered in Yuan and Lin [7], Friedman et al. [5], and Rothman et al. [6]. The solution to (5.9) satisfies

$$\left\{\widehat{\Omega}_\varepsilon(\boldsymbol{\beta}_0)\right\}^{-1} - S_\varepsilon(\boldsymbol{\beta}_0) = \tilde{\lambda}_2\widehat{Z},$$

where $\widehat{Z} \in \mathbb{R}^{p\times p}$ with $\widehat{Z}_{k_1,k_2} = \mathrm{sign}\left\{\widehat{\Omega}_{\varepsilon,k_1,k_2}(\boldsymbol{\beta}_0)\right\}$ if $\widehat{\Omega}_{\varepsilon,k_1,k_2}(\boldsymbol{\beta}_0) \neq 0$ $(k_1, k_2 = 1, \ldots, p)$.

It leads to the following CLIME optimization problem in Cai et al. [20],

$$\min \ \|\Omega_\varepsilon\|_1 \qquad \text{subject to: } |S_\varepsilon(\boldsymbol{\beta}_0)\Omega_\varepsilon - I|_\infty \le \tilde{\lambda}_2, \qquad \Omega_\varepsilon \in \mathbb{R}^{p\times p}, \tag{5.10}$$

which could be solved efficiently with linear programmings (Cai et al. [20]) in a column-by-column fashion.

With a fixed $\Omega_\varepsilon = \Omega_{\varepsilon 0}$, the optimization problem (5.8) yields to

$$\min_{\boldsymbol{\beta}} \operatorname{tr}\left\{(\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta})^{\mathrm{T}}(\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta})\Omega_{\varepsilon 0}\right\} + \lambda_1\|\boldsymbol{\beta}\|_1, \tag{5.11}$$

which has a global minimizer $\widehat{\boldsymbol{\beta}}(\Omega_{\varepsilon 0}) = \left\{\hat{\beta}_{k_1,k_2}\right\} \in \mathbb{R}^{p\times p}$ that satisfies the optimality condition:

$$\widehat{\boldsymbol{\beta}}(\Omega_{\varepsilon 0}) = \underbrace{\left(\widetilde{\boldsymbol{X}}^{\mathrm{T}}\widetilde{\boldsymbol{X}}\right)^{-1}\widetilde{\boldsymbol{X}}^{\mathrm{T}}\widetilde{\boldsymbol{Y}}}_{\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}} - \lambda_1\left(2\widetilde{\boldsymbol{X}}^{\mathrm{T}}\widetilde{\boldsymbol{X}}\right)^{-1}\Gamma\Omega_{\varepsilon 0}^{-1}, \tag{5.12}$$

where Γ is a $p\times p$ matrix with the (k_1,k_2)-th entry $\Gamma_{k_1,k_2} = \operatorname{sign}(\hat{\beta}_{k_1,k_2})$ if $\hat{\beta}_{k_1,k_2} \neq 0$ and otherwise $\Gamma_{k_1,k_2} \in [-1,1]$ with specific values chosen to solve (5.12)(Rothman et al. [15]). The optimization problem in (5.11) could be easily solved with Algorithm 1 in Rothman et al. [15].

5.2.2 Estimation of Between-Subject Precision Matrix

After obtaining the sparse $\widehat{\Omega}_\varepsilon$ and $\widehat{\boldsymbol{\beta}}$, we work on the estimation of Ω_b in the second stage (Stage 2). We first aggregate the original observations by subjects and consider the following model:

$$\bar{\boldsymbol{Y}}_{i\cdot} = \boldsymbol{\beta}^{\mathrm{T}}\bar{\boldsymbol{X}}_{i\cdot} + \boldsymbol{b}_i + \bar{\boldsymbol{\varepsilon}}_{i\cdot}, \quad i = 1,\ldots,m, \tag{5.13}$$

where $\bar{\boldsymbol{Y}}_{i\cdot} = \sum_{t=1}^{n_i} \boldsymbol{Y}_{it}/n_i$, $\bar{\boldsymbol{X}}_{i\cdot} = \sum_{t=0}^{n_i-1} \boldsymbol{Y}_{it}/n_i$ and $\bar{\boldsymbol{\varepsilon}}_{i\cdot} = \sum_{t=1}^{n_i} \boldsymbol{\varepsilon}_{it}/n_i$. Under the normality assumption of $\boldsymbol{b}_i$ and $\boldsymbol{\varepsilon}_{it}$, the new p-variate response also follows normal distribution,

$$\bar{\boldsymbol{Y}}_{i\cdot}|\bar{\boldsymbol{X}}_{i\cdot} = \bar{\boldsymbol{x}}_{i\cdot} \sim \mathcal{N}(\boldsymbol{\beta}^{\mathrm{T}}\bar{\boldsymbol{x}}_{i\cdot}, \Omega_b^{-1} + \Omega_\varepsilon^{-1}/n_i).$$

Thus, we have

$$\begin{aligned}
& \mathbb{E}\left\{\left(\bar{\boldsymbol{Y}}_{i\cdot} - \boldsymbol{\beta}^{\mathrm{T}}\bar{\boldsymbol{x}}_{i\cdot}\right)\left(\bar{\boldsymbol{Y}}_{i\cdot} - \boldsymbol{\beta}^{\mathrm{T}}\bar{\boldsymbol{x}}_{i\cdot}\right)^{\mathrm{T}}\right\} = \Omega_b^{-1} + \Omega_\varepsilon^{-1}/n_i, \text{ for } i = 1, \ldots, m \\
\Rightarrow \quad & \mathbb{E}\left\{\left(\bar{\boldsymbol{Y}}_{i\cdot} - \boldsymbol{\beta}^{\mathrm{T}}\bar{\boldsymbol{x}}_{i\cdot}\right)\left(\bar{\boldsymbol{Y}}_{i\cdot} - \boldsymbol{\beta}^{\mathrm{T}}\bar{\boldsymbol{x}}_{i\cdot}\right)^{\mathrm{T}}\Omega_b\right\} = I_p + \Omega_\varepsilon^{-1}\Omega_b/n_i, \text{ for } i = 1, \ldots, m \\
\Rightarrow \quad & \mathbb{E}\left\{\frac{1}{m}\sum_{i=1}^{m}\left(\bar{\boldsymbol{Y}}_{i\cdot} - \boldsymbol{\beta}^{\mathrm{T}}\bar{\boldsymbol{x}}_{i\cdot}\right)\left(\bar{\boldsymbol{Y}}_{i\cdot} - \boldsymbol{\beta}^{\mathrm{T}}\bar{\boldsymbol{x}}_{i\cdot}\right)^{\mathrm{T}}\Omega_b\right\} = I_p + \left(\sum_{i=1}^{m}\frac{1}{mn_i}\right)\Omega_\varepsilon^{-1}\Omega_b \\
\Rightarrow \quad & \mathbb{E}\left[\left\{\frac{1}{m}\sum_{i=1}^{m}\left(\bar{\boldsymbol{Y}}_{i\cdot} - \boldsymbol{\beta}^{\mathrm{T}}\bar{\boldsymbol{x}}_{i\cdot}\right)\left(\bar{\boldsymbol{Y}}_{i\cdot} - \boldsymbol{\beta}^{\mathrm{T}}\bar{\boldsymbol{x}}_{i\cdot}\right)^{\mathrm{T}} - \sum_{i=1}^{m}\frac{1}{mn_i}\Omega_\varepsilon^{-1}\right\}\Omega_b - I_p\right] = \mathbf{0}.
\end{aligned}$$

To estimate Ω_b, we adopt the CLIME method in Cai et al. [20]. That is

$$\begin{aligned}
& \min \|\Omega_b\|_1 \qquad \text{subject to :} \\
& |\check{S}_b\Omega_b - I_p|_\infty \le \lambda_3, \qquad \Omega_b \in \mathbb{R}^{p\times p},
\end{aligned} \tag{5.14}$$

where

$$\check{S}_b = \underbrace{\frac{1}{m}\sum_{i=1}^{m}\left(\bar{\boldsymbol{Y}}_{i\cdot} - \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\bar{\boldsymbol{x}}_{i\cdot}\right)\left(\bar{\boldsymbol{Y}}_{i\cdot} - \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\bar{\boldsymbol{x}}_{i\cdot}\right)^{\mathrm{T}}}_{\check{S}} - \sum_{i=1}^{m}\frac{1}{mn_i}\widehat{\Sigma}_\varepsilon, \tag{5.15}$$

and $\widehat{\Sigma}_\varepsilon = \frac{1}{N-m}(\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}(\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}})$.

5.3 BIC for Tuning parameters Selection

An appropriate choice of tuning parameters involved in our proposed two-stage estimators is important to control the sparsity and estimation error in $\boldsymbol{\beta}$, Ω_ε and Ω_b. The K-fold cross-validation has been successfully applied in tuning hyper-parameters in multivariate regression with precision matrix estimation (Rothman et al. [15] and Lee and Liu [65]) and graphical models (see, for example, Cai et al. [20] and Fan et al. [66]). Besides the cross-validation procedure, the Bayesian information criterion (BIC) also gains popularity in selecting the tuning parameters for mixed effect models (Delattre et al. [67] and Müler et al. [68]) and fixed-effect models with precision matrix estimation (Abegaz and Wit[69], Yin and Li [70], and Wang [71]), which has demonstrated effectiveness in tuning penalized log-likelihood models (Wang et al. [72]).

Combining the BIC's in Delattre et al. [67] and Abegaz and Wit [69], we first come up with the following joint BIC for Stage 1 and Stage 2,

$$\begin{aligned} \text{BIC}(\lambda_1, \lambda_2, \lambda_3) &= -2\tilde{\ell}(\widehat{\boldsymbol{\beta}}_{\lambda_{1:3}}, \widehat{\Omega}_{\varepsilon,\lambda_{1:3}}, \widehat{\Omega}_{b,\lambda_{1:3}}) \\ &\quad + \Big(\frac{a_n}{2} + b_n + p\Big)\log(N) + \Big(\frac{c_n}{2} + p\Big)\log(m), \end{aligned} \tag{5.16}$$

where $\widehat{\boldsymbol{\beta}}_{\lambda_{1:3}}$,$\widehat{\Omega}_{\varepsilon,\lambda_{1:3}}$ and $\widehat{\Omega}_{b,\lambda_{1:3}}$ are the corresponding estimators with the specific values of $(\lambda_1, \lambda_2, \lambda_3)$, p is the number of variables, a_n and c_n is the number of nonzero off-diagonal elements of $\widehat{\Omega}_{\varepsilon,\lambda_{1:3}}$ and $\widehat{\Omega}_{b,\lambda_{1:3}}$, respectively, and b_n is the number of nonzero elements of $\widehat{\boldsymbol{\beta}}_{\lambda_{1:3}}$. We are supposed to select the triplet of values of λ_1, λ_2 and λ_3 that minimizes $\text{BIC}(\lambda_1, \lambda_2, \lambda_3)$ in (5.16) with a grid search. However, it would be computationally expensive to consider all combinations of λ_1, λ_2 and λ_3.

Note that, after within-group centering, random effects are removed in Stage 1. And Stage 2 only involves the estimation of the between-subject precision matrix. Therefore,

we consider the following separate BIC's, i.e., $\mathrm{BIC}_{\mathrm{Stage}_1}(\lambda_1, \lambda_2)$ and $\mathrm{BIC}_{\mathrm{Stage}_2}(\lambda_3)$ for Stages 1 and 2, respectively. In Stage 1, we will employ $\mathrm{BIC}_{\mathrm{Stage}_1}$,

$$\mathrm{BIC}_{\mathrm{Stage}_1}(\lambda_1, \lambda_2) = \ell(\widehat{\Omega}_{\varepsilon,\lambda_{1:2}}, \widehat{\boldsymbol{\beta}}_{\lambda_{1:2}}) + \Big(\frac{a_n}{2} + b_n + p\Big)\log(N-m), \tag{5.17}$$

while in Stage 2, the following $\mathrm{BIC}_{\mathrm{Stage}_2}(\lambda_3)$,

$$\begin{aligned}\mathrm{BIC}_{\mathrm{Stage}_2}(\lambda_3) = {} & \mathrm{tr}\left\{\sum_{i=1}^{m}\left(\frac{1}{n_i}\widehat{\Omega}^{-1}_{\varepsilon,\lambda_{1:2}} + \widehat{\Omega}^{-1}_{b,\lambda_3}\right)^{-1}(\bar{\boldsymbol{Y}}_{i\cdot} - \widehat{\boldsymbol{\beta}}^{\mathrm{T}}_{\lambda_{1:2}}\bar{\boldsymbol{X}}_{i\cdot})(\bar{\boldsymbol{Y}}_{i\cdot} - \widehat{\boldsymbol{\beta}}^{\mathrm{T}}_{\lambda_{1:2}}\bar{\boldsymbol{X}}_{i\cdot})^{\mathrm{T}}\right\} \\ & + \sum_{i=1}^{m}\log\det\left(\frac{1}{n_i}\widehat{\Omega}^{-1}_{\varepsilon,\lambda_{1:2}} + \widehat{\Omega}^{-1}_{b,\lambda_3}\right) + \Big(\frac{c_n}{2} + p\Big)\log(m),\end{aligned} \tag{5.18}$$

will be considered. Thus, we only need to separately select the pair of (λ_1, λ_2) and λ_3 that minimizes the criterion in $\mathrm{BIC}_{\mathrm{Stage}_1}(\lambda_1, \lambda_2)$ and $\mathrm{BIC}_{\mathrm{Stage}_2}(\lambda_3)$. The minimization of $\mathrm{BIC}_{\mathrm{Stage}_1}(\lambda_1, \lambda_2)$ and $\mathrm{BIC}_{\mathrm{Stage}_2}(\lambda_3)$ with respect to $\lambda_1, \lambda_2,$ and λ_3 is achieved by a grid search, which will reduce the computational expense by using $\mathrm{BIC}(\lambda_1, \lambda_2, \lambda_3)$ in (5.16). The complete algorithm is summarized in Algorithm 3. In our PCP-GVAR algorithm in Algorithm 3, $\boldsymbol{\lambda}_1$, $\boldsymbol{\lambda}_2$ and $\boldsymbol{\lambda}_3$ are the candidate tuning parameters of λ_1, λ_2 and λ_3, which control the sparsity of $\boldsymbol{\beta}$, Ω_ε and Ω_b, respectively. In practice, these candidate tuning parameters are typically set using a 10^x resolution with a sequence of evenly-spaced x's. The implementation of the PCP-GVAR algorithm is built on two R packages: `MRCE` and `flare`. The `flare` package is used for updating $\widehat{\Omega}_\varepsilon$ and $\widehat{\Omega}_b$ with the CLIME method, while the update of $\widehat{\boldsymbol{\beta}}$ is implemented with the `MRCE` package.

Algorithm 3 Projection and Correct Pooling for Graphical VAR Model (PCP-GVAR)

Require: $\boldsymbol{\lambda}_1$, $\boldsymbol{\lambda}_2$, $\boldsymbol{\lambda}_3$ and $\{\boldsymbol{Y}_{ij} : 1 \leq i \leq m, 1 \leq j \leq n_i\}$.

Ensure: $\widehat{\boldsymbol{\beta}}_{\lambda^*_{1:2}}$, $\widehat{\Omega}_{\varepsilon,\lambda^*_{1:2}}$ and $\widehat{\Omega}_{b,\lambda^*_3}$

Create the design matrix $\boldsymbol{X}$ by lagging the response
Remove random effect by within-group centering and obtain $\widetilde{\boldsymbol{Y}}$ and $\widetilde{\boldsymbol{X}}$
for $\lambda_1 \in \boldsymbol{\lambda}_1$ **do**
    for $\lambda_2 \in \boldsymbol{\lambda}_2$ **do**
        repeat
            Update $\widehat{\Omega}_\varepsilon$ with (5.10) by the CLIME method in Cai et al. [20]
            Update $\widehat{\boldsymbol{\beta}}$ with (5.11) using Algorithm 1 in Rothman et al. [15]
        until convergence
        Compute and record $\mathrm{BIC}_{\mathrm{Stage}_1}(\lambda_1, \lambda_2)$ in (5.17)
    end for
end for
Re-estimate $\widehat{\boldsymbol{\beta}}_{\lambda^*_{1:2}}$ and $\widehat{\Omega}_{\varepsilon,\lambda^*_{1:2}}$ in Stage 1, where $\lambda^*_{1:2} = \mathrm{argmin}_{\lambda_1,\lambda_2} \mathrm{BIC}_{\mathrm{Stage}_1}(\lambda_1, \lambda_2)$
for $\lambda_3 \in \boldsymbol{\lambda}_3$ **do**
    Compute $\widehat{\Omega}_{b,\lambda_3}$ with (5.14), where $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{\lambda^*_{1:2}}$ in $\check{S}_b$
    Compute and record $\mathrm{BIC}_{\mathrm{Stage}_2}(\lambda_3)$ in (5.18)
end for
Re-estimate $\widehat{\Omega}_{b,\lambda^*_3}$ in Stage 2, where $\lambda^*_3 = \mathrm{argmin}_{\lambda_3} \mathrm{BIC}_{\mathrm{Stage}_2}(\lambda_3)$

Chapter 6

Numerical Study for Graphical VAR Model with Repeated Measurements

6.1 General Setting

In this chapter, we evaluate the numeric performance of our proposed estimators $\widehat{\boldsymbol{\beta}}$ (for the fixed-effect coefficient matrix $\boldsymbol{\beta}$), $\widehat{\Omega}_\varepsilon$ (for the within-subject precision matrix Ω_ε) and $\widehat{\Omega}_b$ (for the between-subject covariance Ω_b), and compare with the corresponding estimators obtained with the pooled and individual LASSO (PIL-GVAR) estimation procedure in Epskamp et al. [45]. The PIL-GVAR algorithm has the similar two-stage structure as our proposed PCP-GVAR algorithm. However, there are several differences between these two algorithms. Firstly, when estimating Ω_ε in Stage 1, the efficient column-wise CLIME method is employed in our PCP-GVAR algorithm, while the Glasso method is applied for the estimation of Ω_ε in the PIL-GVAR method. Secondly, even though both algorithms use BIC to select the tuning parameters, our BIC in (5.17) is derived exactly from the group-centered data, while the BIC in Stage 1 in the PIL-GVAR algorithm ignores the fact that the data are group-centered. Thirdly, we first obtain a

bias corrected sample estimate for Ω_b, i.e., $\check{S}_b$ in (5.15) and use it as the plug-in for the CLIME method in the second stage of our PCP-GVAR framework. However, the PIL-GVAR algorithm solves the following minization problem,

$$\widehat{\Omega} = \text{argmin}_{\Omega} \, \text{tr}\left\{\check{S}\Omega\right\} - \log \det\Omega + \lambda_3 \|\Omega\|_1,$$

via the Glasso algorithm, in which $\check{S}$ in (5.15) is wrongly treated as the sample estimate of the between-subject covariance matrix Σ_b.

In our simulation study, we generate observations $\boldsymbol{Y}_{ij}$ from model (1.12). We follow Rothman et al. [15] to consider the following setting for the fixed-effect coefficient matrix $\boldsymbol{\beta}$,

$$\boldsymbol{\beta} = U * H * L,$$

where $*$ represents the matrix element-wise product, the entries of U are drawn independently from $\mathcal{N}(0, 1/p)$, the entries of H follow i.i.d. Bernoulli distribution with success probability $s_1 = 0.7$, and L has rows that are either all one or all zero, which are determined by p i.i.d. Bernoulli draws with success probability $s_2 = 0.8$. H and L control the entry-wise and column-wise sparsity of $\boldsymbol{\beta}$, correspondingly. Two different precision matrix structures are considered for Ω_ε and Ω_b.

Model 5 *Covariance matrices corresponding to an* AR(1) *series: set* $(\Sigma_b^0)_{j,k} = c_e \cdot 0.5^{|j-k|}$ *and* $(\Sigma_\varepsilon^0)_{j,k} = 0.6^{|j-k|}$.

The AR(1) structure of covariances in Model 5 results in two tri-diagonal sparse precision

matrices. Generally, if a generic matrix Σ has an AR(1) structure, i.e., $\Sigma_{j,k} = \rho^{|j-k|}$, then

$$\Sigma^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix} 1 & -\rho & 0 & \cdots & 0 \\ -\rho & 1+\rho^2 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 1+\rho^2 & -\rho \\ 0 & \cdots & 0 & -\rho & 1 \end{pmatrix}.$$

Model 6 *Random sparse precision matrices: set* $\Omega_\varepsilon = C_\varepsilon + \delta_\varepsilon I$ *and* $\Omega_b = C_b + \delta_b I$, *where each off-diagonal entry in* C_ε *and* C_b *is generated independently and equals 0.5 with probability 0.1 or 0 with probability 0.9.* δ_ε *and* δ_b *are chosen such that the conditional number (the ratio of maximal and minimal singular values of a matrix) is equal to* p.

We note that the same precision structures in Model 6 had been used in Rothman et al. [6] and Cai et al. [20].

To evaluate the performance of the estimators, we measure the estimation performance by Frobenius norm,

$$F\text{-error} \;=\; \left\| \widehat{Q} - Q^0 \right\|_F,$$

where $\widehat{Q}$ is an estimate of the a generic parameter matrix Q^0. Besides the F-error, we also compute the mean squared error (MSE) of $\widehat{Q}$, and decompose the MSE into the bias and variance terms. The sparsity recognition performance is examined by true positive rate (TPR) and false positive rate (FPR). TPR is also termed sensitivity. Recall the

TPR in (4.1) and FPR in (4.2) defined in Chapter 4,

$$\begin{aligned} \text{TPR}(\widehat{Q}, Q^0) &= \frac{\#\{(i,j) : \widehat{Q}_{i,j} \neq 0 \text{ and } Q^0_{i,j} \neq 0\}}{\#\{(i,j) : Q^0_{i,j} \neq 0\}}, \\ \text{FPR}(\widehat{Q}, Q^0) &= \frac{\#\{(i,j) : \widehat{Q}_{i,j} \neq 0 \text{ and } Q^0_{i,j} = 0\}}{\#\{(i,j) : Q^0_{i,j} = 0\}}. \end{aligned}$$

We also calculated the F_1-score,

$$F_1\text{-score}(\widehat{Q}, Q^0) = \frac{2\text{TP}}{2\text{TP} + 2\text{FP} + 2\text{FN}},$$

where $\text{TP} = \#\{(i,j) : \widehat{Q}_{i,j} \neq 0 \text{ and } Q^0_{i,j} \neq 0\}$, $\text{FP} = \#\{(i,j) : \widehat{Q}_{i,j} \neq 0 \text{ and } Q^0_{i,j} = 0\}$ and $\text{FN} = \#\{(i,j) : \widehat{Q}_{i,j} = 0 \text{ and } Q^0_{i,j} \neq 0\}$. The F_1-score unifies the precision and sensitivity of a classification problem by the harmonic mean, which penalizes extreme values of precision and sensitivity (Hicks et al. [73]).

6.2 Simulation Study

We consider Model 5 and Model 6 under the simulation setting in Section 6.1. In each setting, we let $m = 100$ and $n_i = 10$ for $i = 1, \ldots, m$ and consider $p = 20$ and $p = 40$. Let $c_e = 1$ in Model 5. In our implementation, we set $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \boldsymbol{\lambda}_3 = 10^{\boldsymbol{x}}$, where $\boldsymbol{x} = \{-2, \ldots, \log_{10}(0.5)\}$. Table 6.1 shows the average metrics over 100 replications, and Table 6.2 summarizes the MSE and its decomposition for all the estimators. The estimators of $\boldsymbol{\beta}$ and Ω_ε obtained from the PCP-GVAR algorithm tend to have larger estimation errors, which can potentially result from the larger biases inherent in our estimator as observed in Table 6.2. Nevertheless, our proposed estimator of Ω_ε yields superior selection performance. For all models with different p, the F_1-scores uniformly approximate 0.8. In the context of the estimation of the between-subject precision matrix

Table 6.1: Comparison of the two sets of estimators for Model 5 and 6.

		$p = 20$		$p = 40$	
		PCP-GVAR	PIL-GVAR	PCP-GVAR	PIL-GVAR
		Model 1			
$\boldsymbol{\beta}$	F-error	0.7222 (0.0496)	**0.6449** (0.0465)	1.4442 (0.0595)	**1.3493** (0.0600)
	TPR	0.9300 (0.0165)	**0.9325** (0.0153)	0.8208 (0.0174)	**0.8294** (0.0164)
	FPR	**0.3744** (0.0544)	0.3781 (0.0544)	**0.3436** (0.0362)	0.3626 (0.0323)
	F_1-score	0.8114 (0.0152)	0.8114 (0.0154)	**0.7849** (0.0068)	0.7843 (0.0072)
Ω_ε	F-error	**1.0219** (0.1473)	1.1326 (0.1505)	**1.9779** (0.1925)	2.0595 (0.2071)
	TPR	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
	FPR	**0.0827** (0.0331)	0.2098 (0.0449)	**0.0450** (0.0154)	0.1197 (0.0190)
	F_1-score	**0.8088** (0.0620)	0.6220 (0.0525)	**0.7841** (0.0571)	0.5736 (0.0396)
Ω_b	F-error	**2.8493** (0.3485)	3.1528 (0.3125)	**5.0350** (0.3178)	5.4812 (0.2985)
	TPR	0.9969 (0.0099)	**1.0000** (0.0000)	0.9917 (0.0129)	0.9988 (0.0043)
	FPR	**0.1464** (0.0572)	0.2244 (0.0545)	**0.0549** (0.0163)	0.0965 (0.0265)
	F_1-score	**0.7062** (0.0792)	0.6073 (0.0585)	**0.7441** (0.0563)	0.6286 (0.0637)
		Model 2			
$\boldsymbol{\beta}$	F-error	0.6528 (0.0451)	**0.6141** (0.0404)	1.4523 (0.0712)	**1.3913** (0.0610)
	TPR	0.9230 (0.0188)	**0.9269** (0.0169)	0.8158 (0.0204)	**0.8305** (0.0173)
	FPR	**0.3666** (0.0566)	0.3839 (0.0581)	**0.3034** (0.0343)	0.3418 (0.0373)
	F_1-score	**0.8104** (0.0156)	0.8066 (0.0163)	**0.7940** (0.0084)	0.7911 (0.0084)
Ω_ε	F-error	0.9447 (0.1499)	**0.8097** (0.0944)	2.6871 (0.2544)	**1.9891** (0.1607)
	TPR	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
	FPR	**0.0740** (0.0440)	0.1137 (0.0325)	**0.0694** (0.0167)	0.1461 (0.0218)
	F_1-score	**0.8122** (0.0913)	0.7287 (0.0565)	**0.8044** (0.0365)	0.6856 (0.0360)
Ω_b	F-error	**2.8887** (0.2933)	3.2684 (0.2348)	**7.3452** (0.5002)	7.4941 (0.2656)
	TPR	0.9503 (0.0419)	**0.9619** (0.0340)	0.6801 (0.0797)	**0.2779** (0.0658)
	FPR	**0.1751** (0.0706)	0.2218 (0.0687)	**0.0768** (0.0304)	0.1126 (0.0325)
	F_1-score	**0.6722** (0.0767)	0.6226 (0.0634)	**0.6046** (0.0327)	0.5948 (0.0311)

Table 6.2: Mean squared error, bias and variance of the two sets of estimators for Model 5 and 6

		PCP-GVAR			PIL-GVAR		
		$\boldsymbol{\beta}$	Ω_ε	Ω_b	$\boldsymbol{\beta}$	Ω_ε	Ω_b
				Model 1			
	MSE	0.00131	0.00266	0.02060	0.00104	0.00326	0.02509
$p = 20$	Bias^2	0.00099	0.00164	0.01405	0.00071	0.00240	0.02238
	Variance	0.00032	0.00102	0.00655	0.00033	0.00086	0.00271
	MSE	0.00120	0.00019	0.01009	0.00087	0.00015	0.01721
$p = 40$	Bias^2	0.00104	0.00015	0.00968	0.00069	0.00012	0.01711
	Variance	0.00016	0.00005	0.00041	0.00018	0.00003	0.00010
				Model 2			
	MSE	0.00107	0.00229	0.02108	0.00095	0.00166	0.02684
$p = 20$	Bias^2	0.00074	0.00097	0.01662	0.00061	0.00126	0.02518
	Variance	0.00033	0.00132	0.00446	0.00034	0.00044	0.00166
	MSE	0.00118	0.00933	0.01756	0.00080	0.00870	0.02451
$p = 40$	Bias^2	0.00101	0.00928	0.01708	0.00062	0.00868	0.02443
	Variance	0.00017	0.00005	0.00048	0.00018	0.00002	0.00007

Ω_b, the estimators $\widehat{\Omega}_b$ based on our method surpasses those obtained from the PIL-GVAR in both estimation and selection performance, which indicates the success of the bias correction in $\check{S}$.

Note that, when $n_i = n$ for $i = 1, \ldots, m$, $\check{S}_b$ in (5.15) will be simplified to

$$\underbrace{\frac{1}{m}\sum_{i=1}^{m}\left(\bar{\boldsymbol{Y}}_{i\cdot} - \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\bar{\boldsymbol{x}}_{i\cdot}\right)\left(\bar{\boldsymbol{Y}}_{i\cdot} - \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\bar{\boldsymbol{x}}_{i\cdot}\right)^{\mathrm{T}}}_{\check{S}} - \frac{1}{n}\widehat{\Sigma}_\varepsilon.$$

We want to evaluate the effect of $\widehat{\Sigma}_\varepsilon/n$ with a more accurate $\check{S}$. Therefore, we consider Model 5 with $m = 500$ and $n_i = 5$ for $i = 1, \ldots, m$, and let $p = 20$ and $p = 60$. And we also set $c_e \in \{1, 2, 3\}$. By doing so, we generate settings where the strength of the error term $\widehat{\Sigma}_\varepsilon/n$ varies. We summarize the MSE in Table 6.3 and the average metrics over 100 replicates in Table 6.4. The simulation results are similar to the results in Table 6.1 and Table 6.2. As m increases, our estimator $\widehat{\Omega}_\varepsilon$ tends to be more conservative in

Table 6.3: Mean squared error, bias, and variance of the two sets of estimators for Model 5 with varying c_e.

		PCP-GVAR			PIL-GVAR		
		$\boldsymbol{\beta}$	Ω_ε	Ω_b	$\boldsymbol{\beta}$	Ω_ε	Ω_b
				$p = 20$			
	MSE	0.00198	0.00094	0.01501	0.00153	0.00083	0.02602
$c_e = 1$	Bias^2	0.00180	0.00046	0.01354	0.00135	0.00034	0.02542
	Variance	0.00018	0.00048	0.00147	0.00018	0.00050	0.00006
	MSE	0.00216	0.00024	0.01716	0.00173	0.00021	0.03696
$c_e = 2$	Bias^2	0.00196	0.00012	0.01509	0.00153	0.00009	0.03645
	Variance	0.00020	0.00012	0.00207	0.00020	0.00012	0.00051
	MSE	0.00222	0.00013	0.01955	0.00183	0.00010	0.04646
$c_e = 3$	Bias^2	0.00202	0.00008	0.01703	0.00163	0.00004	0.04606
	Variance	0.00020	0.00006	0.00252	0.00020	0.00006	0.00039
				$p = 60$			
	MSE	0.00112	0.00090	0.00836	0.00100	0.00088	0.01276
$c_e = 1$	Bias^2	0.00096	0.00068	0.00803	0.00085	0.00074	0.01261
	Variance	0.00016	0.00022	0.00033	0.00015	0.00015	0.00015
	MSE	0.00122	0.00022	0.01050	0.00111	0.00021	0.01713
$c_e = 2$	Bias^2	0.00105	0.00017	0.01011	0.00095	0.00017	0.01700
	Variance	0.00017	0.00005	0.00039	0.00016	0.00004	0.00012
	MSE	0.00129	0.00011	0.01238	0.00119	0.00010	0.02069
$c_e = 3$	Bias^2	0.00113	0.00009	0.01194	0.00103	0.00008	0.02059
	Variance	0.00017	0.00002	0.00043	0.00016	0.00002	0.00010

selecting true fixed-effect coefficients, whose false positive rates are all above 0.5. These high FPRs may contribute to the large biases in $\widehat{\Omega}_\varepsilon$. A better tuning parameter selection method is desirable to reduce the false positive rate and bias in the estimates of $\boldsymbol{\beta}$. As shown in both Table 6.3 and Table 6.1, our proposed $\widehat{\Omega}_b$ are more consistent in the estimation error compared with the between-subject precision estimator obtained from the PIL-GVAR algorithm. As c_e increases from 1 to 3, the MSEs of our between-subject precision estimators increase by 30% and 48% for $p = 20$ and $p = 60$, respectively. However, the MSEs of the between-subject precision estimators based on the PIL-GVAR algorithm increase by 79% and 62% for $p = 20$ and $p = 60$, respectively.

6.3 A Real Data Example with Clinical Measurements Collected from Hemodialysis Patients

We apply our proposed methods to estimate the fixed-effect coefficients, within-subject and between-subject precision structures with several clinical variables gathered from patients undergoing hemodialysis. While employing the identical dataset introduced in Chapter 4, we opt for a distinct set of clinical variables. For homogeneity, we exclusively consider patients who have never received a positive COVID-19 diagnosis via polymerase chain reaction testing and who have survived beyond the second year of observation. We use the measurements in the second year to avoid large fluctuations in the first year of dialysis. The robust monthly measurements are generated by calculating their medians for each of the 1074 patients. Finally, we have a balanced dataset in which all patients have 12 monthly consecutive records. Hypertension is common in hemodialysis patients, which also serves as a major prognostic factor for cardiovascular disease (Inrig et al. [74] and Schillaci and Pucci [75]). The risk of cardiovascular disease,

Table 6.4: Comparison of the two sets of estimators for Model 5 with varying c_e.

		$p = 20$		$p = 60$	
		PCP-GVAR	PIL-GVAR	PCP-GVAR	PIL-GVAR
Model 1: $c_e = 1$					
$\boldsymbol{\beta}$	F-error	0.8902 (0.0254)	**0.7819** (0.0210)	2.0053 (0.0515)	**1.8929** (0.0600)
	TPR	**0.9548** (0.0138)	0.9440 (0.0140)	**0.8828** (0.0131)	0.8416 (0.0127)
	FPR	0.5553 (0.0555)	**0.4347** (0.0424)	0.5053 (0.0383)	**0.3411** (0.0253)
	F_1-score	0.7661 (0.0152)	**0.7982** (0.0125)	0.8023 (0.0043)	**0.8186** (0.0036)
Ω_ε	F-error	0.6040 (0.1074)	**0.5731** (0.0705)	**1.7772** (0.2737)	1.7795 (0.1379)
	TPR	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
	FPR	**0.0799** (0.0223)	0.3234 (0.0587)	**0.0354** (0.0236)	0.1204 (0.0193)
	F_1-score	**0.8117** (0.0441)	0.5157 (0.0442)	**0.7631** (0.1060)	0.4669 (0.0406)
Ω_b	F-error	**2.4425** (0.1945)	3.2233 (0.1330)	**5.4816** (0.2254)	6.7754 (0.1478)
	TPR	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
	FPR	**0.2743** (0.0603)	0.3101 (0.0388)	**0.0784** (0.0149)	0.1241 (0.0222)
	F_1-score	**0.5582** (0.0559)	0.5242 (0.0313)	**0.5742** (0.0478)	0.4601 (0.0438)
Model 1: $c_e = 2$					
$\boldsymbol{\beta}$	F-error	0.9289 (0.0202)	**0.8320** (0.0194)	2.0960 (0.0506)	**1.9955** (0.0528)
	TPR	**0.9519** (0.0151)	0.9411 (0.0164)	**0.8783** (0.0125)	0.8333 (0.0127)
	FPR	0.5575 (0.0610)	**0.4398** (0.0713)	0.5040 (0.0325)	**0.3407** (0.0220)
	F_1-score	0.7640 (0.0160)	**0.7954** (0.0191)	0.8002 (0.0035)	**0.8139** (0.0036)
Ω_ε	F-error	0.3063 (0.0458)	**0.0294** (0.0334)	0.8802 (0.1059)	**0.8706** (0.0589)
	TPR	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
	FPR	**0.0988** (0.0323)	0.3209 (0.0637)	**0.0335** (0.0169)	0.1276 (0.0196)
	F_1-score	**0.7783** (0.0560)	0.5184 (0.0484)	**0.7657** (0.0816)	0.4524 (0.0390)
Ω_b	F-error	**2.6121** (0.2036)	3.8421 (0.1564)	**6.1396** (0.3058)	7.8506 (0.1604)
	TPR	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
	FPR	**0.3331** (0.0964)	0.3380 (0.0556)	**0.0834** (0.0181)	0.1484 (0.0277)
	F_1-score	**0.5144** (0.0732)	0.5039 (0.0382)	**0.5604** (0.0570)	0.4171 (0.0462)
Model 1: $c_e = 3$					
$\boldsymbol{\beta}$	F-error	0.9429 (0.0249)	**0.8556** (0.0201)	2.1586 (0.0345)	**2.0706** (0.0596)
	TPR	**0.9475** (0.0133)	0.9396 (0.0152)	**0.8705** (0.0121)	0.8251 (0.0138)
	FPR	0.5470 (0.0576)	**0.4452** (0.0542)	0.4874 (0.0351)	**0.3299** (0.0273)
	F_1-score	0.7649 (0.0138)	**0.7926** (0.0142)	0.7997 (0.0041)	**0.8118** (0.0037)
Ω_ε	F-error	0.2294 (0.0374)	**0.1945** (0.0229)	0.6352 (0.0535)	**0.5906** (0.0504)
	TPR	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
	FPR	**0.0871** (0.0326)	0.3187 (0.0521)	**0.0260** (0.0082)	0.1336 (0.0173)
	F_1-score	**0.8001** (0.0603)	0.5185 (0.0385)	**0.8031** (0.0502)	0.4398 (0.0277)
Ω_b	F-error	**2.7859** (0.2436)	4.3089 (0.1260)	**6.6692** (0.2924)	8.6299 (0.1341)
	TPR	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
	FPR	0.3554 (0.1079)	**0.3522** (0.0492)	**0.0840** (0.0204)	0.1621 (0.0258)
	F_1-score	**0.4995** (0.0759)	0.4930 (0.0350)	**0.5594** (0.0590)	0.3945 (0.0389)

such as, arterial stiffness, is reflected by the absolute blood pressure values (Schillaci and Pucci[75] and Li et al. [76]). To understand the dynamic relationship among blood pressures for hemodialysis patients, we focus on the relationships among interdialytic weight gain and blood pressure, and consider the following six variables: `idwg` (interdialytic weight gain, kg), `ufv` (ultrafiltration volume, L), `pre_sbp` (predialysis systolic blood pressure, mmHg), `post_sbp` (postdialysis systolic blood pressure, mmHg), `pre_dbp` (predialysis diastolic blood pressure, mmHg), `post_dbp` (postdialysis diastolic blood pressure, mmHg). The `ufv` has the same definition as in Chapter 4.

We are interested in recovering the temporal graph ($\boldsymbol{\beta}$), contemporaneous graph (Ω_ε) and between-patients graph (Ω_ε). In these three graphs, the nodes represent the random variables of interest. In the temporal graph, a directed edge connecting two nodes signifies Granger-causality, whereas an undirected edge present in the contemporaneous graph and between-patients graph represents the conditional correlation between the nodes. Prior to applying our method to the repeated clinical measurements obtained from these 1074 patients, we first standardize the data. And the regularization parameters are chosen by minimizing $\mathrm{BIC}_{\mathrm{Stage}_1}(\lambda_1, \lambda_2)$ in (5.17) and $\mathrm{BIC}_{\mathrm{Stage}_2}(\lambda_3)$ in (5.18), where $\lambda_1, \lambda_2, \lambda_3 \in \{0.01, \ldots, 0.3\}$ on the logarithmic scale.

Figure 6.1 presents estimates of the temporal graph (top left panel), within-subject (top right panel), and between-subject (bottom left panel) correlations. Based on these graphs, we can see all the blood pressure measurements are connected to each other with both directed and undirected edges. Moreover, `idwg` and ufv are also connected by an edge in these three graphs. However, the relationships between blood pressures and `idwg` or `ufv` are different among these three graphs. We observe that `idwg` only negatively Granger-causes two postdialysis blood pressures, while `ufv` negatively Granger-causes all blood pressures except the predialysis systolic blood pressure. In both the contemporaneous graph and between-patients graph, `idwg` are only connected to `post_sbp`. In

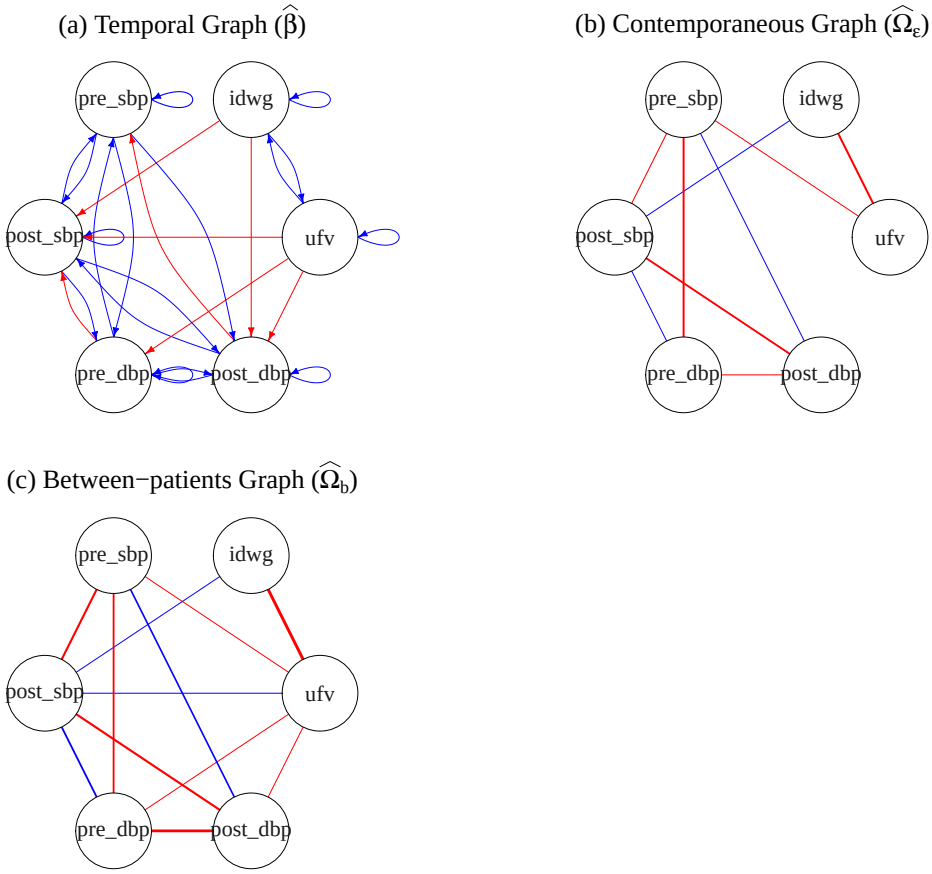

Figure 6.1: Temporal (top left) graph, within-subject (top right) and between-subject (bottom left) precision graphs obtained from the PCP-GVAR Algorithm 3 based on monthly measurements from 1074 hemodialysis patients. The blue edges correspond to the positive relationships, while the red edges represent the negative relationships. The width of an edge corresponds to the strength of the relationships.

the between-patients graph, `ufv` has a more complicated relationship among blood pressures, while it is only related to predialysis systolic blood pressure in the contemporaneous graph.

Chapter 7

Future Studies

7.1 Construction of a New BIC for Graphical VAR Model

As indicated by the simulation study in Section 6, our proposed PCP-GVAR method tends to recover the temporal graph $\boldsymbol{\beta}$ in an anti-conservative manner with a high false positive rate. Hence, we will construct a novel BIC for the selection of tuning parameters in Stage 1.

One possible modification is to add an extra penalty term in our current first-stage BIC in (5.17). For example, Foygel and Drton [77] propose an extended Bayesian information criterion (BIC) for Gaussian graphical models. We propose a potential BIC as follows,

$$\mathrm{BIC}_{new}(\lambda_1, \lambda_2, \gamma) = \mathrm{BIC}_{\mathrm{Stage}_1}(\lambda_1, \lambda_2) + 4\gamma b_n \log(p), \tag{7.1}$$

where $\gamma \in [0, 1]$ has the Bayesian interpretation in Chen and Chen [78]. Positive γ leads to stronger penalization in $\boldsymbol{\beta}$ and could reduce the false positive rate. However, the new

BIC in (7.1) introduces a new hyper-parameter, which may increase the computational burden for finding a suitable γ.

7.2 New Implementation Algorithm for Graphical VAR Model

Tuning the parameters in Stage 1 is computationally expensive. We will consider reconstructing the estimation procedure in Stage 1. Recall the minimization problem in (5.7),

$$\min_{\boldsymbol{\beta},\Omega_\varepsilon \succ 0} \operatorname{tr}\left\{(\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta})^{\mathrm{T}}(\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta})\Omega_\varepsilon\right\} - (N-m)\log\det\Omega_\varepsilon + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\Omega_\varepsilon\|_1.$$

As described in Section 5.2.1, solving (5.7) requires iteratively updating $\boldsymbol{\beta}$ with Ω_ε held fixed and vice versa. This alternating update can be time-consuming in high-dimensional settings (Molstad [79]). It would be plausible to estimate $\boldsymbol{\beta}$, Ω_ε, and Ω_b with three separate but sequential sub-optimization problems. For example, we could construct the following multivariate square-root Lasso optimization problem (Molstad [79] and Van de Geer and Stucky [80]),

$$\min_{\boldsymbol{\beta},\Sigma_\varepsilon^{1/2} \succ 0} \operatorname{tr}\left\{\frac{1}{N-m}(\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta})^{\mathrm{T}}(\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta})\Sigma_\varepsilon^{-1/2}\right\} + \operatorname{tr}\left(\Sigma_\varepsilon^{1/2}\right) + 2\tilde{\lambda}_1\|\boldsymbol{\beta}\|_1, \quad (7.2)$$

where $\tilde{\lambda}_1$ is the tuning parameter and controls the sparsity of $\boldsymbol{\beta}$. The optimization problem in (7.2) will give us a reliable estimator for $\boldsymbol{\beta}$. With a precise estimate of $\boldsymbol{\beta}$, we can obtain an estimate for Σ_ε, which we can use as the plug-in for $S_\varepsilon(\boldsymbol{\beta}_0)$ in (5.10) when estimating Ω_ε in the following step. Therefore, we can replace the minimization problem in (5.7) with two separate sequential optimization problems.

Bibliography

[1] M. Drton and M. H. Maathuis, *Structure learning in graphical modeling*, *Annu. Rev. Stat. Appl.* **4** (2017) 365–393.

[2] J. Fan, Y. Liao, and H. Liu, *An overview of the estimation of large covariance and precision matrices*, *Econom. J.* **19** (2016) C1–C32.

[3] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical learning with sparsity: The lasso and generalizations*. CRC Press, 2015.

[4] S. L. Lauritzen, *Graphical models*. Oxford: Clarendon Press, 1996.

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

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

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

[8] A. J. Butte, P. Tamayo, D. Slonim, T. R. Golub, and I. S. Kohane, *Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks*, *Proc. Natl. Acad. Sci. U.S.A.* **97** (2000) 12182–12186.

[9] S. Chaudhuri, M. Drton, and T. S. Richardson, *Estimation of a covariance matrix with zeros*, *Biometrika* **94** (2007) 199–216.

[10] M. Drton and T. S. Richardson, *Graphical methods for efficient likelihood inference in gaussian covariance models*, *J. Mach. Learn. Res.* **4** (2008) 365–393.

[11] C. Uhler, *Gaussian graphical models: an algebraic and geometric perspective*, *arXiv preprint arXiv:1707.04345* (2009).

[12] D. Edwards, *Introduction to graphical modelling*. New York: Springer, 2000.

[13] J. Whittaker, *Graphical models in applied multivariate statistics*. Chichester: John Wiley and Sons, 1990.

[14] I. M. Johnstone, *On the distribution of the largest eigenvalue in principal components analysis*, *Ann. Stat.* **29** (2001) 295–327.

[15] A. J. Rothman, E. Levina, and J. Zhu, *Sparse multivariate regression with covariance estimation*, *J. Comput. Graph. Stat.* **19** (2010) 947–962.

[16] L. Vandenberghe, S. Boyd, and S.-P. Wu, *Determinant maximization with linear matrix inequality constraints*, *SIAM J. Matrix Anal. Appl.* **19** (1998) 499–533.

[17] R. Tibshirani, *Regression shrinkage and selection via the lasso*, *J. R. Stat. Soc. Series B Stat. Methodol.* **58** (1996) 267–288.

[18] W. J. Fu, *Penalized regressions: the bridge versus the lasso*, *J. Comput. Graph. Stat.* **7** (1998) 397–416.

[19] J. Fan and R. Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, *J. Am. Stat. Assoc.* **96** (2001) 1348–1360.

[20] T. Cai, W. Liu, and X. Luo, *A constrained ℓ_1 minimization approach to sparse precision matrix estimation*, *J. Am. Stat. Assoc.* **106** (2011) 594 – 607.

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

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

[23] H. Wang, *Coordinate descent algorithm for covariance graphical lasso*, *Stat. Comput.* **24** (2014) 521–529.

[24] P. J. Bickel and E. Levina, *Regularized estimation of large covariance matrices*, *Ann. Stat.* **36** (2008a) 199–227.

[25] P. J. Bickel and E. Levina, *Covariance regularization by thresholding*, *Ann. Stat.* **36** (2008b) 2577–2604.

[26] A. J. Rothman, E. Levina, and J. Zhu, *Generalized thresholding of large covariance matrices*, *J. Am. Stat. Assoc.* **104** (2009) 177–186.

[27] T. Cai and W. Liu, *Adaptive thresholding for sparse covariance matrix estimation*, *J. Am. Stat. Assoc.* **106** (2011) 672–684.

[28] T. Cai and M. Yuan, *Adaptive covariance matrix estimation through block thresholding*, *Ann. Stat.* **40** (2012) 2014–2042.

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

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

[31] Y. Cui, C. Leng, and D. Sun, *Sparse estimation of high-dimensional correlation matrices*, *Comput. Stat. Data Anal.* **93** (2016) 390–403.

[32] R. H. Shumway and D. S. Stoffer, *Time series analysis and its applications: with R examples*. New York: Springer, 4th ed., 2017.

[33] L. F. Bringmann, N. Vissers, M. Wichers, N. Geschwind, P. Kuppens, F. Peeters, D. Borsboom, and F. Tuerlinckx, *Correction: A network approach to psychopathology: New insights into clinical longitudinal data*, *PLoS One* **9** (2014) e60188.

[34] C. A. Sims, *Macroeconomics and reality*, *Econometrica* **48** (1980) 1–48.

[35] J. H. Stock and M. W. Watson, *Vector autoregressions*, *J. Econ. Perspect.* **15** (2001) 101–115.

[36] B. Wild, M. Eichler, H.-C. Friederich, M. Hartmann, S. Zipfel, and W. Herzog, *A graphical vector autoregressive modelling approach to the analysis of electronic diary data*, *BMC Medical Res. Methodol.* **10** (2010) 28.

[37] J. D. Hamilton, *Time series analysis*. Princeton: Princeton University Press, 1994.

[38] H. Lütkepohl, *New introduction to multiple time series analysis*. New York: Springer, 2nd ed., 2005.

[39] E. Zivot and J. Wang, *Modeling financial time series with S-Plus*. New York: Springer, 2nd ed., 2006.

[40] D. F. Ahelegbey, M. Billio, and R. Casarin, *Sparse graphical vector autoregression: a bayesian approach*, *Ann. Econ. Stat.* (2016), no. 123/124 333–361.

[41] L. Barnett and A. K. Seth, *The mvgc multivariate granger causality toolbox: A new approach to granger-causal inference*, *J. Neurosci. Methods* **223** (2014) 50–68.

[42] M. Eichler, *Granger causality and path diagrams for multivariate time series*, *J. Econom.* **137** (2007) 334–353.

[43] M. Eichler, *Graphical modelling of multivariate time series*, *Probab. Theory Relat. Fields* **153** (2012) 233–268.

[44] H. Bae, S. Monti, M. Montano, M. H. Steinberg, T. T. Perls, and P. Sebastiani, *Learning bayesian networks from correlated data*, *Sci. Rep.* **6** (2016) 25156.

[45] S. Epskamp, L. J. . Waldorp, R. Mõttus, and D. Borsboom, *The gaussian graphical model in cross-sectional and time-series data*, *Multivar. Behav. Res.* **53** (2018) 453–480.

[46] C. Ostroff, *Comparing correlation based on individual-level and aggregated data*, *J. Appl. Psychol.* **78** (1993) 569–582.

[47] E. L. Hamaker, *Why researchers should think "within-person": A paradigmatic rationale*, in *Handbook of Research Methods for Studying Daily Life* (M. R. Mehl and T. S. Conner, eds.), pp. 43–61. New York: Guilford Press, 2012.

[48] D. A. Freedman, *Ecological inferences and the ecological fallacy*, in *International Encyclopaedia of the Social and Behavioural Sciences* (N. J. Smelser and B. P. B., eds.), vol. 6, pp. 4027–4030. New York: Elsevier, 1999.

[49] S. Piantadosi, D. P. Byar, and S. B. Green, *The ecological fallacy*, *Am. J. Epidemiol.* **127** (1988) 893–904.

[50] J. M. Bland and D. G. Altman, *Calculating correlation coefficients with repeated observations: Part 1 – correlation within subjects*, *BMJ* **310** (1995a) 446.

[51] J. M. Bland and D. G. Altman, *Calculating correlation coefficients with repeated observations: Part 2 – correlation within subjects*, *BMJ* **310** (1995b) 633.

[52] J. Algina and H. Swaminathan, *Psychometrics: Classical test theory*, in *International Encyclopaedia of the Social and Behavioural Sciences* (J. D. Wright, ed.), vol. 19, pp. 423–430. New York: Elsevier, 2nd ed., 2015.

[53] A. J. Fisher, J. D. Medaglia, and B. F. Jeronimus, *Lack of group-to-individual generalizability is a threat to human subjects research*, *Proc. Natl. Acad. Sci. U.S.A.* **115** (2018) E6106–E6115.

[54] P. S. Rao and C. E. Heckler, *Multivariate one-way random effects model*, *Am. J. Math. Manag. Sci.* **18** (1998) 109–130.

[55] P. S. Rao and E. A. Sylvestre, *Anova and minque type of estimators for the one-way random effects model*, *Commun. Stat. – Theory Methods* **13** (1984) 1667–1673.

[56] R. Nishihara, L. Lessard, B. Recht, A. Packard, and M. I. Jordan, *A general analysis of the convergence of alternating direction method*, in *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37, pp. 343–352, 2015.

[57] N. Rontsis, P. Goulart, and Y. Nakatsukasa, *Efficient semidefinite programming with approximate admm*, *Journal of Optimization Theory and Applications* **192** (2022) 292–320.

[58] R. Vershynin, *High-dimensional probability: an introduction with applications in data science*. Cambridge: Cambridge University Press, 2018.

[59] K. J. Ipema, J. Kuipers, R. Westerhuis, C. A. Gaillard, C. P. van der Schans, W. P. Krijnen, and C. F. Franssen, *Causes and consequences of interdialytic weight gain*, *Kidney Blood Press. Res.* **41** (2016) 710–720.

[60] T. J. Hastie, R. J. Tibshirani, and J. H. Friedman, *The elements of statistical learning: prediction, inference and data mining*. New York: Springer, 2nd ed., 2009.

[61] D. Sun and X. Sun, *Estimation of the multivariate normal precision and covariance matrices in a star-shape model*, *Ann. Inst. Statist. Math.* **57** (2005) 455–484.

[62] J. A. Castañeda, Mario H. & Nossek, *Estimation of rank deficient covariance matrices with kronecker structure*, in *2014 IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, pp. 394–398, 2014.

[63] K. B. Petersen and M. S. Pedersen, *The Matrix Cookbook*. Technical University of Denmark, 2012.

[64] D. A. Harville, *Matrix algebra from a statistician's perspective*. New York, NY: Springer, 1997.

[65] W. Lee and Y. Liu, *Simultaneous multiple response regression and inverse covariance matrix estimation via penalized gaussian maximum likelihood*, *J. Multivar. Anal.* **111** (2012) 241–255.

[66] J. Fan, Y. Feng, and Y. Wu, *Network exploration via the adaptive lasso and scad penalties*, *Ann. Appl. Stat.* **3** (2009) 521–541.

[67] M. Delattre, M. Lavielle, and M.-A. Poursat, *A note on bic in mixed-effects models*, *Electron. J. Statist.* **8** (2014) 456–475.

[68] S. Müller, M. Lavielle, and A. H. Welsh, *Model selection in linear mixed models*, *Stat. Sci.* **28** (2013) 135–167.

[69] F. Abegaz and E. Wit, *Sparse time series chain graphical models for reconstructing genetic networks*, *Biostatistics* **14** (2013) 586–599.

[70] J. Yin and H. Li, *A sparse conditional gaussian graphical model for analysis of genetical genomics data*, *Ann. Appl. Stat.* **14** (2011) 2630–2650.

[71] J. Wang, *Joint estimation of sparse multivariate regression and conditional graphical models*, *Stat. Sin.* **25** (2015) 831–851.

[72] H. Wang, R. Li, and C.-L. Tsai, *Tuning parameter selectors for the smoothly clipped absolute deviation method*, *Biometrika* **94** (2007) 553—-568.

[73] S. A. Hicks, I. Strümke, V. Thambawita, M. Hammou, M. A. Riegler, P. Halvorsen, and S. Parasa, *On evaluation metrics for medical applications of artifcial intelligence*, *Sci. Rep.* **12** (2022) 5979.

[74] J. K. Inrig, U. D. Patel, B. S. Gillespie, V. Hasselblad, J. Himmelfarb, D. Reddan, R. M. Lindsay, J. F. Winchester, J. Stivelman, R. Toto, and L. A. Szczech, *Relationship between interdialytic weight gain and blood pressure among prevalent hemodialysis patients*, *Am. J. Kidney Dis.* **50** (2007) 108–118.

[75] G. Schillaci and G. Pucci, *The dynamic relationship between systolic and diastolic blood pressure: yet another marker of vascular aging?*, *Hypertens. Res.* **33** (2010) 659–661.

[76] Y. Li, J.-G. Wang, E. Dolan, P.-J. Gao, H.-F. Guo, T. Nawrot, A. V. Stanton, D.-L. Zhu, E. O'Brien, and J. A. Staessen, *Ambulatory aterial stiffness index derived from 24-hour ambulatory blood pressure monitoring*, *Hypertens.* **47** (2006) 359–364.

[77] R. Foygel and M. Drton, *Extended bayesian information criteria for gaussian graphical models*, *Advances in Neural Information Processing Systems 23 (NIPS 2010)* **1** (2010) 604–612.

[78] J. Chen and Z. Chen, *Extended bayesian information criteria for model selection with large model spaces*, *Biometrika* **95** (2008) 759–771.

[79] A. J. Molstad, *New insights for the multivariate square-root lasso*, *J. Mach. Learn. Res.* **23** (2022) 1–52.

[80] S. Van de Geer and B. Stucky, χ^2*-confidence sets in high-dimensional regression*, in *Statistical Analysis for High-Dimensional Data*, pp. 279–306. Springer, 2016.