Statistica Sinica Preprint No: SS-2024-0279	
Title	Estimating Covariance Matrices at Different Levels in Repeated Measurements
Manuscript ID	SS-2024-0279
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0279
Complete List of Authors	Sunpeng Duan, Guo Yu, Juntao Duan and Yuedong Wang
Corresponding Authors	Guo Yu
E-mails	guoyu@ucsb.edu
Notice: Accepted author version.	

Statistica Sinica

Estimating Covariance Matrices at Different Levels in Repeated Measurements

Sunpeng Duan*, Guo Yu*, Juntao Duan, Yuedong Wang

Department of Statistics and Applied Probability

University of California, Santa Barbara, 93106

Abstract: 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 covariance/correlation at different levels. Most existing methods for sparse covariance/correlation matrix estimation assume independent samples. Ignoring the underlying hierarchical structure and correlation within the subject may lead to erroneous scientific conclusions. In this paper, we propose to distinguish between the between-subject covariance structure and the within-subject covariance structure. In the presence of repeated measurement, this leads to the problem of sparse and positive-definite estimation of between-subject and within-subject covariance matrices. 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

*These authors contributed equally to this work

hemodialysis patients.

Key words and phrases: Covariance graph; repeated measurements; ecological fallacy; random effect; sparsity.

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 [Bickel and Levina, 2008b, Fan et al., 2016]. This task becomes particularly challenging in high-dimensional settings, where various regularized estimation methods have been proposed in recent years. However, virtually all current methods require the critical assumption that observations are independent, which could be violated in many applications. Estimating covariance structures in the presence of dependence remains an important and challenging problem due to repeated measurements [Ostroff, 1993].

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. Measurements from the same subject are usually

correlated. However, these correlations are rarely characterized and sometimes even totally ignored in the estimation of covariance structure between variables, which may be practically misguided or even lead to erroneous conclusions [Bae et al., 2016].

Repeated measurements have an underlying hierarchical structure, and it is of scientific interest to define and estimate covariance structures at each level. 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., 2018]. However, in many applications, this distinction of covariance structures at different levels is often ignored. A commonly used approach to estimating covariance structure with repeated measurements is to use the sample covariance estimate based on data aggregated for each subject [Epskamp et al., 2018, Fisher et al., 2018]. This approach results in biased estimates for covariance structures at either level, and conclusions based on the estimated covariances may be misleading, causing ecological fallacy or Simpson's paradox. Specifically, in Section 7, we apply this biased estimation based on aggregated data for estimating correlation structures among clinical variables collected from hemodialysis patients and demonstrate in Figure 2 that it misses some of the crucial correlation structures that can be recovered by our proposed method.

The main premise of this paper is that the dependence structure among random variables within a subject is different from the dependence structure between subjects and, thus, should be estimated differently. We are not the only ones in the literature proposing such a distinction of covariance structures at different levels. Indeed, the importance of distinguishing between-subject and within-subject covariances has been increasingly emphasized in the psychometrics literature. For example, the nomothetic approach is used to study variations between subjects, and the idiographic approach is used to study variations within a subject [Hamaker, 2012].

However, this issue has not yet received significant attention in the statistics and biostatistics community and applications in other fields. In particular, to the best of our knowledge, no statistical methods have been introduced to characterize the bias in using the covariance estimation based on aggregated data formally and provide practical advice on the improved estimators.

This paper aims to formally address this methodological gap within the statistics literature and to provide improved estimators with theoretical guarantees. We emphasize the importance of separating the covariance structure among random variables into two targets of estimation: one for within-subject covariance and one for between-subject covariance, which encode different covariance structures and characterize the bias of using aggregated data in estimating

either of these two covariance matrices. As improved methods, we propose estimators that are sparse and non-negative definite for both between-subject and within-subject covariance structures with much improved theoretical properties.

The rest of this paper is organized as follows. Section 2 formally characterizes the distinction between between-subject and within-subject covariances, and introduces the sample estimates for these two target matrices. Based on these sample estimates, we propose sparse estimates that are guaranteed to be positive-definite. As discussed in Section 3, 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 Section 4. Section 5 compares our proposed between-subject covariance estimator with the MANOVA-type estimator. Sections 6 and 7 investigate the numerical performance of our proposed estimators with comprehensive simulations and an application to a dataset collected from end-stage renal disease (ESRD) patients. The paper ends with a discussion. Detailed technical proofs are provided in the Supplementary Materials.

2. Separate Estimation of Between- and Within-subject Covariance Structures

For simplicity, this paper uses the term "subject" to represent a generic experimental unit. We consider a multivariate one-way random effect model for within-subject and between-subject covariance structures among p random variables:

$$\boldsymbol{Y}_{ij} = \boldsymbol{b}_i + \boldsymbol{\varepsilon}_{ij},\ j = 1, \ldots, n_i;\ i = 1, \ldots, m, \tag{2.1}$$

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 $E(\boldsymbol{Y}_{ij} \mid i)$. On the other hand, the *within-subject covariance* Σ_ε characterizes the covariance structure among components in $\boldsymbol{Y}_{ij} - E(\boldsymbol{Y}_{ij} \mid i)$. Model (2.1) has found wide applications, e.g., in the classical test theory [Algina and Swaminathan, 2015], where the observed score is modeled as the summation of the true score (as a latent variable) and a random error. These definitions of between-subject and within-subject

covariance are in a similar spirit to those in Ostroff [1993] and Piantadosi et al. [1988], where only the sample version of these quantities was defined.

In longitudinal studies, data with repeated measurements are commonly modeled by mixed-effects models. Model (2.1) can be considered as a multidimensional response linear mixed effects model without fixed effects. In mixed-effects model literature, there are two types of estimation methods, where one maximizes likelihood or restricted likelihood [Rao and Sylvestre, 1984, Laffont et al., 2014, Fieuws and Verbeke, 2006], the other uses Bayesian Markov chain Monte Carlo (MCMC) [Bürkner, 2017, Stan Development Team, 2024]. In practice, those likelihood-based methods are restricted to low-dimensional cases as computation becomes prohibitive when the number of responses is moderately large. In this paper, we take a moment estimation approach to greatly simplify the computation of covariance matrices, leading to the possibility of developing a theory for sparse estimations in a high-dimensional setting.

For the cross-sectional data, which is a special case of (2.1) 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,

$$\overline{\Sigma} = \frac{1}{m-1}\sum_{i=1}^{m}\left(\bar{\boldsymbol{Y}}_{i\cdot} - \frac{1}{m}\sum_{i=1}^{m}\bar{\boldsymbol{Y}}_{i\cdot}\right)\left(\bar{\boldsymbol{Y}}_{i\cdot} - \frac{1}{m}\sum_{i=1}^{m}\bar{\boldsymbol{Y}}_{i\cdot}\right)^{\mathrm{T}}, \tag{2.2}$$

is an unbiased estimate of

$$E(\overline{\Sigma}) = \Sigma_b + \sum_{i=1}^{m}\frac{1}{mn_i}\Sigma_\varepsilon. \tag{2.3}$$

Consequently, $\overline{\Sigma}$ is a biased estimate of either Σ_ε or Σ_b. Epskamp et al. [2018] used (2.3) to estimate the between-subject covariance structure. Statistical inferences based on aggregated data may be misinterpreted [Fisher et al., 2018]. In particular, analysis based on aggregated data may result in an issue termed ecological fallacy or Simpson's paradox [Piantadosi et al., 1988, Freedman, 1999, Hamaker, 2012, Epskamp et al., 2018]. Therefore, this paper emphasizes that one should separately estimate within-subject and between-subject covariance matrices.

We consider the following unbiased estimates:

$$\widehat{\Sigma}_\varepsilon = \left(\sum_{i=1}^{m} n_i - m\right)^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\boldsymbol{Y}_{ij} - \bar{\boldsymbol{Y}}_{i\cdot})(\boldsymbol{Y}_{ij} - \bar{\boldsymbol{Y}}_{i\cdot})^{\mathrm{T}}, \tag{2.4}$$

$$\widehat{\Sigma}_b = \overline{\Sigma} - \sum_{i=1}^{m}(mn_i)^{-1}\widehat{\Sigma}_\varepsilon. \tag{2.5}$$

The sample estimate $\widehat{\Sigma}_\varepsilon$ is an unbiased estimate of Σ_ε [Rao and Heckler, 1998]. From (2.2), $\widehat{\Sigma}_b$ is an unbiased estimate of Σ_b, and is a multivariate extension of the unweighted sum-of-squares estimator in Rao and Sylvestre [1984].

There exist other unbiased estimates, one of which for Σ_b will be considered in Section 5 and shown to be suboptimal compared to $\widehat{\Sigma}_b$ considered in (2.5).

3. Sparse Positive Definite Estimation of the Two Covariance Matrices

In high-dimensional settings, where the number of random variables p is larger than the number of subjects m, we observe that $\widehat{\Sigma}_\varepsilon$ in (2.5) may be singular. Furthermore, $\widehat{\Sigma}_b$ may not be positive semi-definite for any dimensions. In particular, the diagonal elements in $\widehat{\Sigma}_b$ could be negative, making neither $\widehat{\Sigma}_\varepsilon$ nor $\widehat{\Sigma}_b$ useful in practice. Furthermore, in high-dimensional settings, the entry-wise sparsity assumption in covariance matrices is commonly imposed for better interpretation. These observations call for regularized estimation of both Σ_b and Σ_ε that are sparse and guaranteed to be positive definite.

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, 2008a,b, Cai and Yuan, 2012], and

further improvements have been developed to ensure positive definiteness in the resulting estimates [Rothman et al., 2009, Rothman, 2012, Xue et al., 2012, Cui et al., 2016]. Bien and Tibshirani [2011] 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. In this section, we adapt a similar strategy and propose a regularized estimate of Σ_b and Σ_ε based on their unbiased estimates in (2.5). Our proposed regularized estimators are defined as solutions to convex optimization problems and are guaranteed to be positive definite.

Specifically, we consider the following optimization problem for estimating a generic covariance matrix Σ with input matrix B [Xue et al., 2012],

$$\min_{\Sigma \succeq \delta I_p} \frac{1}{2}\|\Sigma - B\|_F^2 + \lambda|\Sigma|_1, \tag{3.6}$$

where $\|\cdot\|_F$ is the Frobenius norm and $|\cdot|_1$ is the ℓ_1-norm of the off-diagonal elements of the input matrix. In high-dimensional settings, the unbiased estimator for both of the target covariance matrices may not be non-negative definite. In practice, a reasonable estimate of either Σ_B or Σ_ε needs to be non-negative definite. Indeed, a non-negative definite covariance matrix is essential to ensure mathematical validity, interpretability, valid statistical inference, and computational stability. For repeated measurement data, non-negative definite covariance matrices are essential for interpreting heterogeneity between subjects,

variation within subjects, and correlations between different variables and ensuring the validity of downstream analyses. While in general positive definiteness in the estimated covariance matrix is desired, we note that in some cases positive semidefiniteness is induced (e.g., random effects selection by inducing low rank in the estimate of Σ_b, Pinheiro and Bates [2000]). We propose to enforce non-negative definiteness in the estimate of both target matrices by using the constraint $\Sigma \succeq \delta I_p$ in (3.6), which imposes positive semi-definiteness on $\Sigma - \delta I_p$. In practice, a small value of $\delta > 0$ is used for numeric stability since the minimum eigenvalue of the target matrix is usually unknown. This positive definiteness constraint is essential for a usable and accurate estimate (see Table 1 in the Supplementary Materials for numeric evidence).

A solution to (3.6) is simultaneously sparse, positive definite, and close to the input matrix B, 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 (3.6) with $B = \widehat{\Sigma}_\varepsilon$ and $\lambda = \lambda_\varepsilon$, and $\widehat{\Sigma}_b^+$ be the sparse and positive definite estimates of Σ_b as the solution to (3.6) with $B = \widehat{\Sigma}_b$ 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 (3.6) with $B = \overline{\Sigma}$ and $\lambda = \lambda_0$. The theoretical tuning parameter values λ_ε, λ_b, and λ_0 are discussed in Section 4.

Table 1: Summary of notations for various covariance matrix estimators. Regularized estimates are defined as solutions to (3.6) with corresponding sample estimates.

Target	Sample Estimate	Regularized Estimate	Tuning Parameter	Comments
	$\overline{\Sigma}$	$\overline{\Sigma}^{+}$	λ_0	based on (2.2), no distinction at different levels
Σ_ε	$\widehat{\Sigma}_\varepsilon$	$\widehat{\Sigma}_\varepsilon^{+}$	λ_ϵ	$\widehat{\Sigma}_\varepsilon$ defined in (2.4)
Σ_b	$\widehat{\Sigma}_b$	$\widehat{\Sigma}_b^{+}$	λ_b	$\widehat{\Sigma}_b$ defined in (2.5)
Σ_b	$\widetilde{\Sigma}_b$	$\widetilde{\Sigma}_b^{+}$	$\tilde{\lambda}_b$	discussed in Section 5

In Table 1, we provide a summary of notations for various covariance matrix estimators that are considered in this paper.

The convex optimization problem (3.6) can be written equivalently as

$$\min_{\Sigma,\Theta}\left\{\frac{1}{2}\|\Sigma - B\|_F^2 + \lambda|\Theta|_1 :\ \Sigma = \Theta,\ \Sigma \succeq \delta I_p\right\}, \tag{3.7}$$

which we solve using the alternating direction method of multipliers [Boyd et al., 2010]. Specifically, the algorithm iteratively minimizes the following augmented Lagrangian

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

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

$$\Sigma \leftarrow \text{argmin}_{\Sigma \succeq \delta I_p} L(\Sigma, \Theta; \Lambda) = \tfrac{1}{1+\rho}(B + \rho\Theta - \Lambda, \delta)_+, \tag{3.8}$$

$$\Theta \leftarrow \text{argmin}_{\Theta} L(\Sigma, \Theta; \Lambda) = \mathcal{S}_{\lambda/\rho}\Big(\Sigma + \tfrac{1}{\rho}\Lambda\Big), \tag{3.9}$$

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

The update in (3.8) 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 (3.9) evaluates element-wise soft-thresholding operators, where $\{\mathcal{S}_b(A)\}_{jk} = \text{sign}(A_{jk}) \max(|A_{jk}| - b, 0)$ for any matrix A and scalar $b \geq 0$. We follow Boyd et al. [2010] 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 S1 of the Supplementary Materials for further implementation details. This algorithm has been widely used in the literature on covariance estimation [e.g., Bien and Tibshirani, 2011, Xue et al., 2012] with well-established convergence analysis [Nishihara et al., 2015]. The computational complexity of each update is dominated by the eigendecomposition in (3.8), which requires $O(p^3)$ operations. An approximate alternating direction method of multipliers [Rontsis et al., 2022] could be used to improve the computational complexity by avoiding repeated eigendecompositions.

4. Theoretical Properties

In this section, we first illustrate the pitfall of the sample estimator (2.2) based on the aggregated data to estimate either of the target matrices. In comparison, we then derive the finite-sample estimation error rate of our proposed estimators $\widehat{\Sigma}_{\varepsilon}^{+}$ (in Section 4.3) and $\widehat{\Sigma}_{b}^{+}$ (in Section 4.4), and establish their asymptotic consistency.

4.1 Notations and Assumptions

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 (2.1), where $\boldsymbol{\varepsilon}_{ij}$ and $\boldsymbol{b}_i$ are p-dimensional sub-Gaussian random vectors with the true within and between covariance $\mathrm{cov}(\boldsymbol{\varepsilon}_{ij}) = \Sigma_{\varepsilon}^{0}$ and $\mathrm{cov}(\boldsymbol{b}_i) = \Sigma_{b}^{0}$ 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:

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

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.

4.2 Inconsistency of sample covariance estimator based on aggregated data

We first establish that $\overline{\Sigma}$ is inconsistent in estimating either Σ_b or Σ_ϵ due to a non-vanishing bias in the estimation lower bounds, even with the number of subjects approaching infinity. Recall from (2.3) that $\overline{\Sigma}$ is biased in estimating either Σ_b or Σ_ϵ. In the following theorem, we characterize the finite sample lower bound on the element-wise estimation error.

Theorem 1. *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*

$$\lambda_{0,b} = \max_{k,l} \left| \frac{m(\Sigma_\varepsilon^0)_{k,l}}{(m-1)n^*} + \frac{(\Sigma_b^0)_{k,l}}{m-1} \right| - C_1 \left(\frac{\log p}{m} \right)^{1/2},$$

and

$$\lambda_{0,\varepsilon} = \max_{k,l} \left| \frac{m}{m-1}(\Sigma_b^0)_{k,l} - \left(1 - \frac{\sum_{i=1}^m n_i^{-1}}{(m-1)} \right) (\Sigma_\varepsilon^0)_{k,l} \right| - C_1 \left(\frac{\log p}{m} \right)^{1/2},$$

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

$$pr\left\{ \max_{k,l} \left| (\overline{\Sigma} - \Sigma_b^0)_{k,l} \right| > \lambda_{0,b} \right\} \geq 1 - 4p^{-C_2},$$

and

$$pr\left\{ \max_{k,l} \left| (\overline{\Sigma} - \Sigma_\varepsilon^0)_{k,l} \right| > \lambda_{0,\varepsilon} \right\} \geq 1 - 4p^{-C_2},$$

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

The first terms in both of $\lambda_{0,b}$ and $\lambda_{0,\varepsilon}$ above are not vanishing, even when m approaches infinity. This implies that $\overline{\Sigma}$ is not a good estimate of either Σ_e or Σ_b. For example, in the balanced setting where $n_i = n_1$ for all $i = 1, \ldots, m$, it holds that $n_* = n_1$. The bias term in estimating Σ_b does not vanish even if $m \to \infty$ as long as $n_1 = O(1)$. And the bias term in estimating Σ_ϵ is not going to zero even if both $m \to \infty$ and $n_i \to \infty$, as long as $\Sigma_b^0 = \Sigma_\epsilon^0$ does not hold element-wise. However, in practice, $\overline{\Sigma}$ has been misused to provide a sample estimate for subsequent regularized estimation [Epskamp et al., 2018]. In Section S3.5 in Supplementary Materials, we provide upper bounds on the estimation error rates for $\overline{\Sigma}^+$ for estimating either Σ_b or Σ_ε, which is defined as the solution to (3.6) with input sample matrix $B = \overline{\Sigma}$. In the subsequent subsections, we present upper bounds on the estimation error rate of our proposed estimators and show their consistency.

4.3 Estimation Error Rate for the Within-Subject Covariance Estimator

Theorem 2 (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 (3.6) for a sufficiently large constant* $C_1 > 0$. *If*

$\log p \leq N$, *the proposed within-subject estimator* $\widehat{\Sigma}_\varepsilon^+$ *satisfies*

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

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

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 the sparsity level has also been noted in Rothman et al. [2008] and Xue et al. [2012] 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 2 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 Bickel and Levina [2008b], Rothman et al. [2008, 2009], Cai and Liu [2011], Xue et al. [2012], 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 2 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.

Remark 3. Following the similar presentation as in Xue et al. [2012], the probability for which the estimation error bound in Theorem 2 (and all subsequent theorems) holds solely depends on the ambient dimension p, which implicitly assumes that $p \to \infty$. In the settings where p is fixed, following the proof in S3 in Supplementary Materials, the probability can be expressed in terms of the effective sample size, which implies that the probability goes to 1 as $N^{1/2} - N^{-1/2}m \to \infty$.

Remark 4. We note that although the theoretical analysis depends on the exponential tail condition from the sub-Gaussian assumptions, it can be extended to allow for a polynomial-tail condition as in Xue et al. [2012] using similar arguments. Specifically, a generic random variable Z with mean 0 satisfies the *polynomial-tail condition* with constant K_z if for all $\gamma > 0$ and $\delta > 0$, the

following holds

$$\mathrm{E}\left[|Z|^{4(1+\gamma+\delta)}\right] \leq K_z \tag{4.10}$$

for some $K_z > 0$. The following theorem derives the estimation error bound for $\widehat{\Sigma}_\varepsilon^+$ under the polynomial-tail condition.

Theorem 3. *Suppose that random errors $\boldsymbol{\varepsilon}_{ij} \in \mathbb{R}^p$ are i.i.d. random vectors with mean zero and within-subject covariance $\mathrm{cov}(\boldsymbol{\varepsilon}_{ij}) = \Sigma_\varepsilon \in \mathcal{U}(M_\varepsilon, s_\varepsilon)$. Furthermore, suppose that the entries ε_{ijk} satisfy the polynomial-tail condition* (4.10) *with constant K_ε for $k = 1, \ldots, p$. For any constant $M > 0$, let*

$$\lambda_\varepsilon = 16(K_\varepsilon + 1)(M+2)\frac{(N \log p)^{1/2}}{N-m}.$$

If $p \leq cN^\gamma$ for some c, then the regularized within-subject covariance estimate $\widehat{\Sigma}_\varepsilon^+$ satisfies

$$\begin{aligned} & \mathrm{pr}\left\{\left\|\widehat{\Sigma}_\varepsilon^+ - \Sigma_\varepsilon^0\right\|_F \leq 5\lambda_\varepsilon(ps_\varepsilon)^{1/2}\right\} \\ \geq\ & 1 - O(p^{-M}) - K_\varepsilon p(\log N)^{2(1+\gamma+\delta)}N^{-(\gamma+\delta)} - K_\varepsilon p(\log m)^{2(1+\gamma+\delta)}m^{-(\gamma+\delta)}. \end{aligned}$$

Comparing with Theorem 2, $\widehat{\Sigma}_\varepsilon^+$ attains the same rate of error bound under the polynomial-tail condition, with more stringent scaling conditions on p, m, and N so that the probability for which the error bound holds goes to 1. This pattern is consistent with Theorem 2 in Xue et al. [2012].

4.4 Estimation Error Rate for the Between-Subject Covariance Estimator

Theorem 4 (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*

$$\lambda_b = C_1 \left(\frac{\log p}{m}\right)^{1/2} + C_2 \frac{(N \log p)^{1/2}}{(N-m)n^*} + \frac{M_b}{m} + \frac{M_\varepsilon}{mn^*} \tag{4.11}$$

be the value of the tuning parameter λ *in (3.6) for sufficiently large* $C_1, C_2 > 0$*, where* $n^* = m/\sum_{i=1}^m n_i^{-1}$*. If* $\log p \le m$*, then the proposed between-subject estimator* $\widehat{\Sigma}_b^+$ *satisfies*

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

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

Unlike the estimation error rate for $\widehat{\Sigma}_\varepsilon$ in Theorem 2, the rate for $\widehat{\Sigma}_b$ depends on the values of n_i's via the term n^*. A simple bound $n^* \ge \min_i n_i$ 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 2. The rate in λ_b is thus dominated by $(m^{-1}\log p)^{1/2}$. Furthermore, in Theorem 7 in S3.7 of the Supplementary Materials, we derive the estimation error bound for $\widehat{\Sigma}_b^+$ under the polynomial-tail condition (4.10), which is the same as under the exponential-tail condition.

Finally, 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 S3.6 of the Supplementary Materials.

5. Comparison Between Two Unbiased Estimators of Between-subject Covariance

We consider a commonly used unbiased estimator of Σ_b based on the multivariate analysis of variance [Rao and Heckler, 1998]:

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

where

$$n_0 = \frac{N-N^{-1}\sum_{i=1}^{m}n_i^2}{m-1}, \quad \bar{\boldsymbol{Y}}_{i\cdot} = n_i^{-1}\sum_{j=1}^{n_i}\boldsymbol{Y}_{ij}, \quad \bar{\boldsymbol{Y}}_{\cdot\cdot} = N^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\boldsymbol{Y}_{ij},$$

and $N=\sum_{i=1}^{m}n_i$.

It is straightforward to show that $\mathrm{E}(\widetilde{\Sigma}_b)=\Sigma_b$. However, just like $\widehat{\Sigma}_b$ in (2.5), 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 $\mathrm{pr}\{(\widetilde{\Sigma}_b)_{k,k}<0\}$ 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 [1998] 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 (3.6) with $B = \widetilde{\Sigma}_b$ (see Table 1). The following theorem shows that the performance of $\widetilde{\Sigma}_b^+$ hinges on the data imbalance.

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*

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

be the value of the tuning parameter λ in (3.6) for sufficiently large $C_1, C_2 > 0$. If $\log p \le m$, then $\widetilde{\Sigma}_b^+$ satisfies

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

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 $\max_i n_i / n_0 \ge 1$, where n_0 is defined in (5.12). 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.

To understand the theoretical advantage of using $\hat{\Sigma}_b$ in the imbalanced data, we note that it can be interpreted as a pooling estimator, where each repeated measurement contributes to the final covariance matrix estimate equally. In comparison, the form of $\tilde{\Sigma}_b$ suggests that subjects with more measurements will have a heavier weight in the final estimate. This suggests the possible dependence of $\tilde{\Sigma}_b$ on the data imbalance. Furthermore, $\hat{\Sigma}_b$ is an extension to the Unweighted Sums of Squares (USS) type estimator [Rao et al., 1981, Rao and Sylvestre, 1984] to the multivariate case. Conclusions from Rao and Sylvestre [1984], which focuses on the univariate case, state that " *... there is considerable gain in using the USS estimators instead of the ANOVA estimator...* " which suggests the similar advantage of USS estimator ($\hat{\Sigma}_b$) over the ANOVA type of estimator ($\tilde{\Sigma}_b$) in the multivariate settings. We numerically verify this comparison in Section 6, and demonstrate that the practical performance of $\widetilde{\Sigma}_b^+$ could be very sensitive to the data imbalance.

6. Simulation Studies

In this section, 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).

6.1 Simulation Settings

In each of the subsequent subsections, we generate observations $\boldsymbol{Y}_{ij}$ from model (2.1), where $\boldsymbol{b}_i \sim N(\mathbf{0}, \Sigma_b^0)$ and $\boldsymbol{\varepsilon}_{ij} \sim N(\mathbf{0}, \Sigma_{\varepsilon}^0)$. All estimators in comparison are defined as solutions to the optimization problem (3.6) with corresponding input sample covariance matrices. We use a 5-fold cross-validation procedure to select the optimal tuning parameter value λ in (3.6) for each problem. We refer interested readers to S2 of the Supplementary Materials for the details of the cross-validation procedure in the repeated measurement settings.

6.2 General Comparison

In Sections 4 and Section 5, 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., $\max_i n_i/n_0$. To illustrate the established

theoretical results, 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 [2008a], Rothman [2012], Xue et al. [2012], Cui et al. [2016]. 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.

Fig. 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 4, and Theorem 5, the performance of $\widehat{\Sigma}_b^+$ and

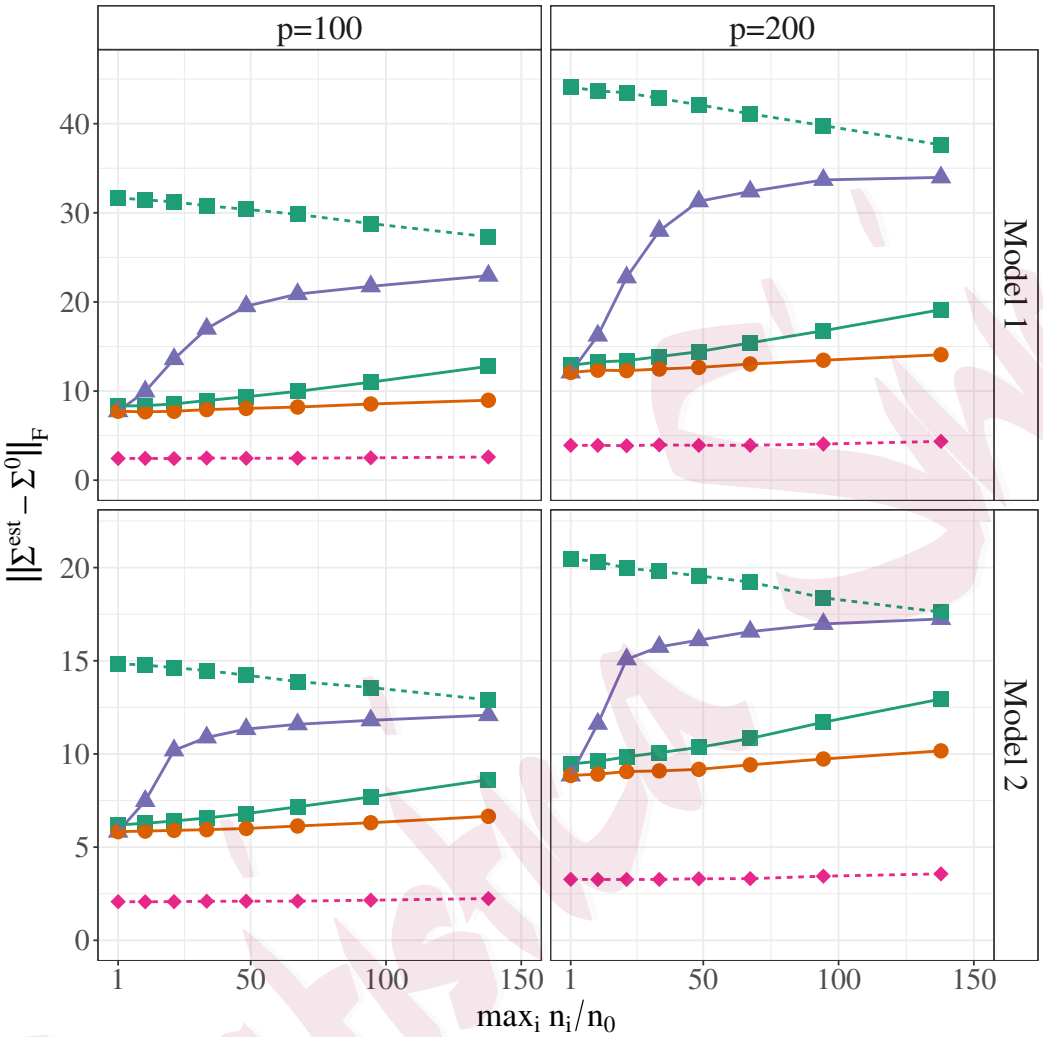

Figure 1: Estimation error (in Frobenius norm, averaged over 100 replicates) for two between-subject (solid) and one within-subject (dash) covariance matrix estimator: $\widetilde{\Sigma}_b^+$ (violet triangle), $\widehat{\Sigma}_b^+$ (orange circle), and $\widehat{\Sigma}_\varepsilon^+$ (pink diamond). 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.

$\overline{\Sigma}^+$ 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 (5.12), performs much worse than $\overline{\Sigma}^+$ which is built on the biased $\overline{\Sigma}$ in (2.2). 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 5 in S3.5 of the Supplementary Materials, which states that the error rate of $\|\overline{\Sigma}^+ - \Sigma_\varepsilon^0\|_F$ is inversely proportional to the imbalance score $\max_i n_i/n_0$. As seen in Fig 1, the estimator $\overline{\Sigma}^+$ based on the biased sample estimate $\overline{\Sigma}$ surprisingly has a relatively acceptable numerical performance. In S4.2 in Supplementary Materials, we conduct additional simulation studies to further demonstrate the limitations of using $\overline{\Sigma}$ and the favorable performance of our proposed estimators.

To demonstrate the effectiveness of regularization, in Fig. 1 in S4.1 of the Supplementary Materials, 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 Fig. 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.

Finally, we note that our proposed methods do not require the assumptions that random effects $\boldsymbol{b}_i$ and random errors $\boldsymbol{\varepsilon}_{ij}$ are normally distributed. To gauge the numeric performance of the proposed estimators when the Gaussian assumption is violated, e.g., in heavy-tailed data, we consider simulation settings where random effects and random errors are both generated from a t_5 distribution, and every other specification remains the same as in Model 1 with $p = 100$. The results, which are summarized in S4.3 in the Supplementary Materials, suggest that our proposed methods still perform favorably in heavy-tailed settings.

7. Covariance Graphs of Clinical Variables 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 treat-

ment 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. [2016], 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 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., 2007]. 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., 2009]. Fig. 2 presents estimates of the within-subject (left panel) and between-subject (middle panel) correlations, which indeed present different correlation structures. We also include the estimate using the aggregated data (right panel) for comparison, which

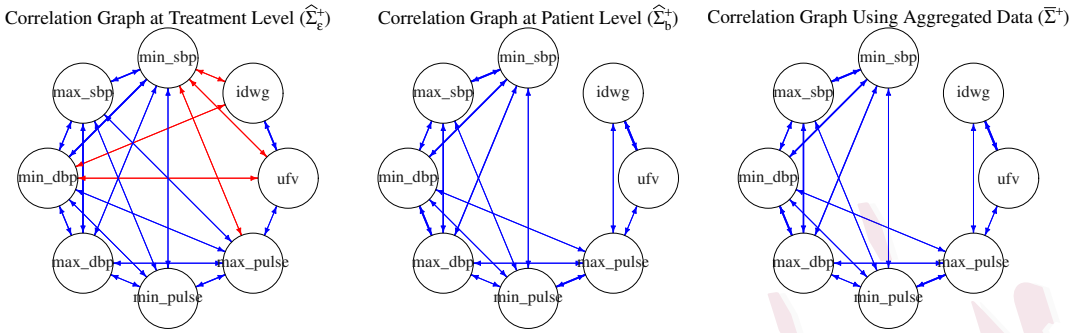

Figure 2: Within-subject (left), between-subject (middle) correlation graphs and correlation graph using the aggregated data (right) for clinical variables from hemodialysis patients. We present correlation matrices with the convention of using bi-directed covariance graphs [Chaudhuri et al., 2007]. 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.

coincides with our between-subject estimate. This is consistent with Theorem 1 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 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 accumulation 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 Fig. 2. 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 Fig. 2. 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 Fig. 2. However, no relationships among `idwg`, `max_pulse`, and `min_sbp` have been observed at the patient level in the middle panel of Fig. 2. 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.

8. Discussion

In this paper, we study the problem of estimating covariance structure among random variables in the presence of dependent observations from repeated measurements. In this challenging setting, we suggest against the commonly used subject-aggregated estimator, which could incur ecological fallacy. Instead, we propose simultaneously studying the between-subject and the within-subject covariance matrices under a random effect model. These two targets of estimation retain different covariance graphs and should be interpreted and estimated differently — a key observation that is commonly ignored in previous literature. Through both theoretical and numerical studies, we show that our proposed sparse and positive definite estimator of both target matrices enjoys favorable estimation error rates and robustness to the imbalance of the data.

We note that this paper emphasizes the importance of separating and estimating covariance structures at different levels in the presence of dependent observations. Our approach is not restricted to either the random effect model in (2.1) or any specific regularized estimation methods such as in (3.6). Following the idea of this paper, extending model (2.1) to more general mixed effects models that allows for fixed effects $X_{ij}\beta$ requires proper treatment of different covariance matrices, and their simultaneous regularized estimation will be an interesting future research direction. In particular, among existing methods in

literature, one strategy [see, e.g., Ahn et al., 2012] is to first deploy a working homoscedastic assumption on the linear model and to obtain a least squares estimate $\hat{\beta}$, which is unbiased but is not statistically efficient. Then the residues $Y_{ij} - X_{ij}\hat{\beta}$ can be used for the task of covariance estimation. Another type of method adapts a joint estimation approach, but has to impose strongly simplifying assumptions on the covariance structure among the random effects [Reisetter and Breheny, 2021]. Most importantly, however, virtually all existing methods only focus on the univariate response setting [Ahn et al., 2012, Reisetter and Breheny, 2021]. The problem of extending this task to the multivariate response settings is of great importance to the literature. However, this problem can be particularly challenging, as the estimation of β (and potentially selection of fixed effects) is largely intertwined with the (regularized) estimation of covariance matrices, both in terms of computation and theoretical analysis. The authors are working on tackling these challenges in a follow-up project.

Acknowledgements

We thank Fresenius Medical Care North America for providing de-identified data and Dr. Hanjie Zhang for discussing real data analysis. We also thank the editor, associate editor, and two referees for constructive comments that substantially improved an earlier draft.

Conflict of interest: None declared.

Funding Statement

This research was partially supported by NIH grant R01DK130067.

Data availability

The R codes that support and reproduce the finding of this study are openly hosted on the Github repository: https://github.com/sunpeng52/GGM. The hemodialysis data are available on the COVID RADx Data Hub.

Supplementary Materials

This Supplementary Materials includes proofs of the theoretical results, computational details, and additional data analyses.

References

Mihye Ahn, Hao Helen Zhang, and Wenbin Lu. Moment-based method for random effects selection in linear mixed models. *Statistica Sinica*, 22(4):1539, 2012.

J. Algina and H. Swaminathan. Psychometrics: Classical test theory. *International Encyclopaedia of the Social and Behavioural Sciences*, 19:423–430, 2015.

H. Bae, S. Monti, M. Montano, M.H. Steinberg, T.T. Perls, and P. Sebastiani. Learning bayesian networks from correlated data. *Scientific Reports*, 6:25156, 2016.

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

P.J. Bickel and E. Levina. Covariance regularization by thresholding. *Ann. Statist.*, 36:2577–2604, 2008b.

J. Bien and R.J. Tibshirani. Sparse estimation of a covariance matrix. *Biometrika*, 98:807–820, 2011.

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3:1–122, 2010.

Paul-Christian Bürkner. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28, 2017.

T. Cai and W. Liu. Adaptive thresholding for sparse covariance matrix estimation. *J. Am. Statist. Ass.*, 106:672–684, 2011.

T. Cai and M. Yuan. Adaptive covariance matrix estimation through block thresholding. *Ann. Statist.*, 40:2014–2042, 2012.

S. Chaudhuri, M. Drton, and T.S. Richardson. Estimation of a covariance matrix with zeros. *Biometrika*, 94:199–216, 2007.

Y. Cui, C. Leng, and D. Sun. Sparse estimation of high-dimensional correlation matrices. *Computational Statistics and Data Analysis*, 93:390–403, 2016.

S. Epskamp, L.J. Waldorp, R. Mõttus, and D. Borsboom. The gaussian graphical model in cross-sectional and time-series data. *Multivariate Behavioral Research*, 53:453–480, 2018.

J. Fan, Y. Liao, and H. Liu. An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal*, 19:C1–C32, 2016.

Steffen Fieuws and Geert Verbeke. Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics*, 62(2):424–431, 2006.

A.J. Fisher, J.D. Medagliab, and B.F. Jeronimusd. Lack of group-to-individual

generalizability is a threat to human subjects research. *Proc. Natn. Acad. Sci. USA*, 115:E6106–E6115, 2018.

D.A. Freedman. Ecological inferences and the ecological fallacy. *International Encyclopaedia of the Social and Behavioural Sciences*, 6:4027–4030, 1999.

E.L. Hamaker. Why researchers should think "within-person": A paradigmatic rationale. In M. R. Mehl and T. S. Conner, editors, *Handbook of Research Methods for Studying Daily Life*, pages 43–61. Guilford Press, New York, 2012.

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

K.J.R. Ipema, J. Kuipers, R. Westerhuis, C.A.J.M. Gaillard, C.P. van der Schans, W.P. Krijnen, and C.F.M. Franssen. Causes and consequences of interdialytic weight gain. *Kidney and Blood Pressure Research*, 41:710–720, 2016.

Celine Marielle Laffont, Marc Vandemeulebroecke, and Didier Concordet. Multivariate analysis of longitudinal ordinal data with mixed effects models, with application to clinical outcomes in osteoarthritis. *Journal of the American Statistical Association*, 109(507):955–966, 2014.

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*, volume 37, pages 343–352, 2015.

C. Ostroff. Comparing correlation based on individual-level and aggregated data. *Journal of Applied Psychology*, 78:569–582, 1993.

S. Piantadosi, D.P. Byar, and S.B. Green. The ecological fallacy. *American Journal of Epidemiology*, 127:893–904, 1988.

José Pinheiro and Douglas Bates. *Mixed-effects models in S and S-PLUS*. Springer science & business media, 2000.

Poduri SRS Rao, Jack Kaplan, and William G Cochran. Estimators for the one-way random effects model with unequal error variances. *Journal of the American Statistical Association*, 76(373):89–97, 1981.

P.S.R.S. Rao and C.E. Heckler. Multivariate one-way random effects model. *American Journal of Mathematical and Management Sciences*, 18:109–130, 1998.

P.S.R.S. Rao and E.A. Sylvestre. Anova and minque type of estimators for the

one-way random effects model. *Communs Statist. Theory Meth.*, 13:1667–1673, 1984.

Anna C Reisetter and Patrick Breheny. Penalized linear mixed models for structured genetic data. *Genetic epidemiology*, 45(5):427–444, 2021.

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

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

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

A.J. Rothman, E. Levina, and J. Zhu. Generalized thresholding of large covariance matrices. *J. Am. Statist. Ass.*, 104:177–186, 2009.

Stan Development Team. *RStan: the R interface to Stan*, 2024. URL `https://mc-stan.org/`. R package version 2.32.6.

L. Xue, S. Ma, and H. Zou. Positive-definite ℓ_1-penalized estimation of large covariance matrices. *J. Am. Statist. Ass.*, 107:1480–1491, 2012.