Regularized Parameter Estimation in Mixed Model Trace Regression

Ian Hultman
Department of Statistics and Actuarial Science, University of Iowa
and
Sanvesh Srivastava
Department of Statistics and Actuarial Science, University of Iowa

March 19, 2025

Abstract

We introduce mixed model trace regression (MMTR), a mixed model linear regression extension for scalar responses and high-dimensional matrix-valued covariates. MMTR's fixed effects component is equivalent to trace regression, with an element-wise lasso penalty imposed on the regression coefficients matrix to facilitate the estimation of a sparse mean parameter. MMTR's key innovation lies in modeling the covariance structure of matrix-variate random effects as a Kronecker product of low-rank row and column covariance matrices, enabling sparse estimation of the covariance parameter through low-rank constraints. We establish identifiability conditions for the estimation of row and column covariance matrices and use them for rank selection by applying group lasso regularization on the columns of their respective Cholesky factors. We develop an Expectation-Maximization (EM) algorithm extension for numerically stable parameter estimation in high-dimensional applications. MMTR achieves estimation accuracy comparable to leading regularized quasi-likelihood competitors across diverse simulation studies and attains the lowest mean square prediction error compared to its competitors on a publicly available image dataset.

Keywords: High-dimensional regularization; matrix normal distribution; mixed model; separable covariance; trace regression

arXiv:2503.13782v1 [stat.ME] 18 Mar 2025

混合模型追踪回归中的正则化参数估计

Ian Hultman
Department of Statistics and Actuarial Science, University of Iowa
and
Sanvesh Srivastava
统计与精算学系，爱荷华大学

2025年3月19日

摘要

我们介绍了混合模型追踪回归（MMTR），这是一种用于标量响应和高维矩阵值协变量的混合模型线性回归扩展。MMTR的固定效应分量等效于迹回归，对回归系数矩阵施加元素级Lasso惩罚，以促进稀疏均值参数的估计。MMTR的关键创新在于将矩阵变量随机效应的协方差结构建模为低秩行和列协方差矩阵的Kronecker积，通过低秩约束实现协方差参数的稀疏估计。我们建立了行和列协方差矩阵估计的识别条件，并通过对各自的Cholesky因子列应用组Lasso正则化来使用这些条件进行秩选择。我们开发了一种期望最大化（EM）算法扩展，用于高维应用中的数值稳定参数估计。MMTR在多种模拟研究中实现了与领先的正则化准似然竞争对手相当估计精度，并在公开可用的图像数据集上与竞争对手相比获得了最低均方预测误差。

关键词：高维正则化；矩阵正态分布；混合模型；可分离协方差；迹回归

1 Introduction

We introduce trace regression models that include matrix-valued random effects, referred to as *mixed model trace regression* (MMTR). MMTR extends mixed model linear regression to include matrix-valued fixed and random effects covariates, but the response remains a vector of correlated observations. The parameter dimension in these models easily exceeds the number of observations, requiring specific sparsity assumptions on the model parameters. MMTR models the mean parameter by assuming that the regression coefficients matrix is sparse. The random effects' row and column covariance matrices in MMTR are assumed to have low ranks, leading to sparse covariance parameters. We develop an EM algorithm for regularized parameter estimation in MMTR, which exploits the sparsity patterns in the parameter components for efficiency and stability in the high-dimensional regime.

Consider the MMTR setup. Let n be the number of subjects or clusters, m_i be the number of observations specific to cluster i $(i = 1, \ldots, n)$, and $N = \sum_{i=1}^{n} m_i$ be the total number of observations. The response vector for cluster i is $\mathbf{y}_i \in \mathbb{R}^{m_i}$, with jth element $y_{ij} \in \mathbb{R}$. The fixed and random effects covariates matrices specific to y_{ij} are $\mathbf{X}_{ij} \in \mathbb{R}^{P_1 \times P_2}$ and $\mathbf{Z}_{ij} \in \mathbb{R}^{Q_1 \times Q_2}$ $(i = 1, \ldots, n;\ j = 1, \ldots, m_i)$. Then, MMTR sets

$$y_{ij} = \mathrm{tr}(\mathbf{X}_{ij}^{\top}\mathbf{B}) + \mathrm{tr}(\mathbf{Z}_{ij}^{\top}\mathbf{A}_i) + e_{ij}, \quad \mathbf{A}_i \sim N_{Q_1,Q_2}(\mathbf{0}, \tau^2\,\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2), \quad \mathbf{e}_i \sim N_{m_i}(\mathbf{0}, \tau^2\,\mathbf{I}_{m_i}), \quad (1)$$

where tr is the *trace* operator, $\mathbf{B} \in \mathbb{R}^{P_1 \times P_2}$ is the fixed effects parameter matrix, $\mathbf{A}_i \in \mathbb{R}^{Q_1 \times Q_2}$ is the random effects matrix, $N_{Q_1,Q_2}(\mathbf{0}, \tau^2\,\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$ is the matrix variate Gaussian distribution with row and column covariance matrices $\tau^2\,\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, respectively, $\mathbf{e}_i = (e_{i1}, \ldots, e_{im_i})^{\top}$ is the ith idiosyncratic error vector, $\mathbf{I}_{m_i}$ is an $m_i \times m_i$ identity matrix, and $\mathbf{A}_i, \mathbf{e}_i$ are mutually independent for every i. The model parameters are $\mathbf{B}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$, and τ^2.

Mixed model linear and trace regression models are special cases of MMTR in (1). MMTR reduces to mixed model linear regression if $P_2 = Q_2 = 1$. This model is widely

1 引言

我们介绍包含矩阵值随机效应的追踪回归模型，称为 混合模型追踪回归 (MMTR)。MMTR 将混合模型线性回归扩展到包含矩阵值固定效应和随机效应协变量，但响应仍然是一组相关的观测值。在这些模型中，参数维度很容易超过观测值数量，需要对模型参数进行特定的稀疏性假设。MMTR 通过假设回归系数矩阵是稀疏的来建模均值参数。MMTR 中随机效应的行和列协方差矩阵被假定为低秩，导致稀疏协方差参数。我们开发了一种用于 MMTR 正则化参数估计的 EM 算法，该算法利用参数组件中的稀疏模式，以提高高维情况下的效率和稳定性。

考虑 MMTR 设置。令 n 为受试者或集群的数量， m_i 为特定于集群 i $(i = 1, \ldots, n)$ 的观测值数量， $N = \sum_{i=1}^{n} m_i$ 为总观测值数量。集群 i 的响应向量为 $\mathbf{y}_i \in \mathbb{R}^{m_i}$，其 j 个元素为 $y_{ij} \in \mathbb{R}$。特定于 y_{ij} 的固定效应和随机效应协变量矩阵分别为 $\mathbf{X}_{ij} \in \mathbb{R}^{P_1 \times P_2}$ 和 $\mathbf{Z}_{ij} \in \mathbb{R}^{Q_1 \times Q_2}$ $(i = 1, \ldots, n;\ j = 1, \ldots, m_i)$。然后，MMTR 设置

$$y_{ij} = \mathrm{tr}(\mathbf{X}_{ij}^{\top}\mathbf{B}) + \mathrm{tr}(\mathbf{Z}_{ij}^{\top}\mathbf{A}_i) + e_{ij}, \quad \mathbf{A}_i \sim N_{Q_1,Q_2}(\mathbf{0}, \tau^2\,\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2), \quad \mathbf{e}_i \sim N_{m_i}(\mathbf{0}, \tau^2\,\mathbf{I}_{m_i}), \quad (1)$$

其中 tr 是 迹 算子， $\mathbf{B} \in \mathbb{R}^{P_1 \times P_2}$ 是固定效应参数矩阵， $\mathbf{A}_i \in \mathbb{R}^{Q_1 \times Q_2}$ 是随机效应矩阵， $N_{Q_1\,Q_2}(\mathbf{0}, \tau^2\,\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$ ，是具有行和列协方差矩阵 $\tau^2\,\boldsymbol{\Sigma}_1$ 和 $\boldsymbol{\Sigma}_2$ 的矩阵变差高斯分布， $\mathbf{e}_i = (e_{i1}, \ldots, e_{im_i})^{\top}$ 是第 i 个特殊误差向量， $\mathbf{I}_{m_i}$ 是一个 $m_i \times m_i$ 单位矩阵，并且$\mathbf{A}^i$, $\mathbf{e}^i$ 对于每个 i 是相互独立的。模型参数是 $\mathbf{B}$ $\boldsymbol{\Sigma}_1$， $\boldsymbol{\Sigma}_2$， 和 τ^2。

混合模型线性回归和迹回归模型是 (1) 中 MMTR 的特例。如果 $P_2 = Q_2 = 1$, MMTR 简化为混合模型线性回归。该模型被广泛

used for analyzing repeated measures and clustered data, but its application is limited to vector-valued covariates (Verbeke et al., 1997). If $\mathbf{A}_i = \mathbf{0}$ and $m_i = 1$ for every i in (1), then MMTR is equivalent to the trace regression model (Fan et al., 2019). This model accommodates matrix-valued covariates but fails to model the correlation in the responses. MMTR integrates the advantages of these two frameworks and enables realistic models for repeated measures and clustered data with matrix covariates and scalar responses.

用于分析重复测量和聚类数据，但其应用仅限于向量值协变量（Verbeke 等人，1997 年）。如果对于（1）中的每个 i， $\mathbf{A}_i = \mathbf{0}$ 和 $m_i = 1$ 成立，则 MMTR 等同于迹回归模型（Fan 等人，2019 年）。该模型支持矩阵值协变量，但无法对响应中的相关性进行建模。MMTR 结合了这两个框架的优点，并能够为具有矩阵协变量和标量响应的重复测量和聚类数据建立实际模型。

1.1 Prior Literature

The vectorization of covariates reduces MMTR to a "structured" mixed model linear regression. If $P = P_1P_2$ and $Q = Q_1Q_2$, then (1) is equivalent to a mixed model with fixed and random effects covariates $\text{vec}(\mathbf{X}_{ij}) \in \mathbb{R}^P$ and $\text{vec}(\mathbf{Z}_{ij}) \in \mathbb{R}^Q$, P-dimensional regression coefficient $\text{vec}(\mathbf{B})$, and Q-dimensional random effects $\text{vec}(\mathbf{A}_i)$ with a $Q \times Q$ covariance matrix $\tau^2\,\boldsymbol{\Sigma}_2 \otimes \boldsymbol{\Sigma}_1$, where vec is the column-wise vectorization of a matrix and $\otimes$ is the Kronecker product. Fitting these structured mixed models using classical algorithms, such as those in `lme4` (Bates et al., 2015), has poor empirical performance for two main reasons. First, the estimation algorithms do not exploit the sparse and the separable matrix-variate structures of $\mathbf{B}$ and $\tau^2\,\boldsymbol{\Sigma}_2 \otimes \boldsymbol{\Sigma}_1$, respectively. Second, they are prohibitively slow because the parameter dimension grows rapidly as $O(P + Q^2)$.

1.1 先前文献

协变量的向量化将 MMTR 简化为“结构化”混合模型线性回归。如果 $P = P_1P_2$ 和 $Q = Q_1Q_2$，则（1）等同于具有固定效应和随机效应协变量 $\text{vec}(\mathbf{X}_{ij}) \in \mathbb{R}^P$ 和 $\text{vec}(\mathbf{Z}_{ij}) \in \mathbb{R}^Q$，$P$-维回归系数 $\text{vec}(\mathbf{B})$ 以及 Q-维随机效应 $\text{vec}(\mathbf{A}_i)$ 的混合模型，其协方差矩阵为 $\tau^2\,\boldsymbol{\Sigma}_2 \otimes \boldsymbol{\Sigma}_1$，其中 vec 是矩阵的按列向量化，$\otimes$ 是 Kronecker 积。使用经典算法（如 `lme4`(Bates 等人，2015 年) 中的算法）拟合这些结构化混合模型，在经验性能方面表现不佳，主要原因有两个。首先，估计算法没有利用 $\mathbf{B}$ 和 $\tau^2\,\boldsymbol{\Sigma}_2 \otimes \boldsymbol{\Sigma}_1$ 的稀疏和可分离的矩阵变量结构。其次，它们非常慢，因为参数维度随着 $O(P + Q^2)$ 的增加而迅速增长。

High-dimensional mixed model extensions address these issues through penalization but fail to exploit the separable covariance structures. The main idea of these methods is to define a quasi-likelihood by replacing the covariance matrix with a proxy matrix (Fan and Li, 2012; Hui et al., 2017; Bradic et al., 2020; Li et al., 2022). A sparse estimate of $\mathbf{B}$ is obtained using lasso-type penalties, which are optimal under various high-dimensional asymptotic regimes. The estimate of the random effects covariance matrix is only available when $Q < \min_i m_i$, a condition often violated in practice (Li et al., 2022). Furthermore, even when the estimate exists, it does not have the separable structure of the true covariance

高维混合模型扩展通过惩罚来解决这些问题，但未能利用可分离协方差结构。这些方法的主要思想是通过用代理矩阵替换协方差矩阵来定义准似然（Fan and Li, 2012; Hui et al., 2017; Bradic et al., 2020; Li et al., 2022）。使用Lasso型惩罚获得$\mathbf{B}$的稀疏估计，这些惩罚在各种高维渐近情形下是最优的。随机效应协方差矩阵的估计仅在 $Q < \min_i m_i$时可用，而这一条件在实践中通常不成立（Li et al., 2022）。此外，即使估计存在，它也不具有真实协方差的可分离结构

matrix.

The literature on separable covariance estimation addresses such problems. Consider a modification of (1) with $\mathbf{B} = \mathbf{0}$ and $\tau^2 = 1$ that sets $\mathbf{Y}_i = \mathbf{A}_i \in \mathbb{R}^{Q_1 \times Q_2}$ for $i = 1, \ldots, n$. In this model, $\text{Cov}\{\text{vec}(\mathbf{A}_i)\}$ has the separable form $\mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1$, where $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are defined in (1). In the low dimensional setting, the two parameters are estimated using a "flip-flop" algorithm that estimates $\mathbf{\Sigma}_2$ given $\mathbf{\Sigma}_1$ and vice versa (Dutilleul, 1999; Lu and Zimmerman, 2005; Srivastava et al., 2008). Hoff (2011) develops tensor-variate extensions of these algorithms; however, all these approaches assume that $Q \ll n$. Extending methods from the literature on low-rank covariance estimation, Zhang et al. (2023) estimate $\text{Cov}\{\text{vec}(\mathbf{A}_i)\}$ via regularized banded estimates of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$. Unlike the setup in these models, the random effects $\mathbf{A}_i$'s are unobserved in (1), making it impossible to apply these algorithms directly for parameter estimation in MMTR.

Tensor regression directly handles matrix-variate covariates but assumes $\mathbf{A}_i = \mathbf{0}$ and $m_i = 1$ for every i in (1). The main focus is on estimating $\mathbf{B}$ under sparsity inducing penalties. If the correlation in $\mathbf{y}_i$ induced by $\mathbf{A}_i$ in (1) is ignored, then existing trace regression algorithms estimate low-rank, row-sparse, or column-sparse estimates of $\mathbf{B}$ in (1) using nuclear norm and group lasso penalties (Zhao et al., 2017; Slawski et al., 2015; Fan et al., 2019). While such estimates remain under-studied, similar estimates, which assume the independence of $\mathbf{y}_i$ entries, in mixed model linear regression have poor inferential and predictive performance (Hui et al., 2021; Li et al., 2022). We conjecture that such results extend to the MMTR due to the equivalence between the two model classes.

MMTR belongs to the class of tensor mixed models, where the responses and covariates are structured as tensors. The simplest models in this class have no random effects and estimate the regression coefficients under low rank and sparse constraints (Zhou et al., 2013; Zhou and Li, 2014). There is limited literature on tensor mixed models that jointly estimate mean and covariance parameters. Yue et al. (2020) develop one such model, but

矩阵。

可分协方差估计的文献研究了这类问题。考虑对(1)的修改，使用$\mathbf{B} = \mathbf{0}$和$\tau^2 = 1$，使得$\mathbf{Y}_i = \mathbf{A}_i \in \mathbb{R}^{Q_1 \times Q_2}$为$i = 1, \ldots, n$。在此模型中，$\text{Cov}\{\text{vec}(\mathbf{A}_i)\}$具有可分形式$\mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1$，其中$\mathbf{\Sigma}_1$和$\mathbf{\Sigma}_2$在(1)中定义。在低维设置中，使用“翻转-翻转”算法估计这两个参数，该算法根据$\mathbf{\Sigma}_1$估计$\mathbf{\Sigma}_2$，反之亦然（Dutilleul，1999；陆和齐默曼，2005；Srivastava等人，2008）。霍夫（2011）开发了这些算法的张量变量扩展；然而，所有这些方法都假设$Q \ll n$。扩展低秩协方差估计文献中的方法，张等人（2023）通过$\mathbf{\Sigma}_1$和$\mathbf{\Sigma}_2$的正则化带状估计来估计$\text{Cov}\{\text{vec}(\mathbf{A}_i)\}$。与这些模型中的设置不同，随机效应$\mathbf{A}_i$’s在(1)中是未观测的，这使得无法直接将这些算法应用于MMTR的参数估计。

张量回归直接处理矩阵变量协变量，但假设（1）中的每个i都有$\mathbf{A}_i = \mathbf{0}$和$m_i = 1$。主要关注点是在稀疏性诱导惩罚下估计$\mathbf{B}$。如果忽略了（1）中$\mathbf{A}_i$在$\mathbf{y}_i$中诱导的相关性，则现有的迹回归算法使用核范数和组套索惩罚估计（1）中$\mathbf{B}$的低秩、行稀疏或列稀疏估计（Zhao等人，2017；Slawski等人，2015；Fan等人，2019）。虽然这些估计仍研究不足，但在混合模型线性回归中假设$\mathbf{y}_i$条目独立的类似估计具有较差的推断和预测性能（Hui等人，2021；Li等人，2022）。我们推测由于这两种模型类之间的等价性，这些结果扩展到MMTR。

MMTR属于张量混合模型类别，其中响应和协变量结构化为张量。此类中最简单的模型没有随机效应，并在低秩和稀疏约束下估计回归系数（Zhou等人，2013；Zhou和Li，2014）。关于联合估计均值和协方差参数的张量混合模型的文献有限。Yue等人（2020）开发了一个此类模型，但

it excludes random effects covariates. The random effects tensor has a separable structure in this model and the estimation algorithm is a variant of the flip-flop algorithm; however, their approach is only applicable when $n \ll Q$ and $\mathbf{Z}_{ij} = \mathbf{I}_{Q_1}$ for every i and j, implying that this method cannot be used for parameter estimation in MMTR.

MMTR is closely related to tensor mixed models based on generalized estimating equations (GEE), which are tuned for modeling longitudinal imaging data (Zhang et al., 2019). The GEE-based model has scalar responses and tensor covariates. For two dimensional tensors, the GEE-based models and MMTR have the same mean parametrization; however, the GEE-based model assumes that the regression coefficients have a low tensor rank, whereas MMTR assumes they are sparse. The random effects terms are absent in the GEE-based model, which replaces them with sample-specific marginal "working" covariance matrices. The GEE-based model requires selection of the rank, working covariance matrix form, and penalty tuning parameters for parameter estimation using a minorization maximization (MM) algorithm. In contrast, MMTR's parameter estimation algorithm only requires two tuning parameters that determine the sparsity of $\mathbf{B}$ and ranks of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$.

1.2 Our Contributions

MMTR employs a regularized EM for estimating the parameters in (1). Let $\mathbf{L}_1$ and $\mathbf{L}_2$ be two matrices such that $\mathbf{\Sigma}_1 = \mathbf{L}_1\mathbf{L}_1^\top$ and $\mathbf{\Sigma}_2 = \mathbf{L}_2\mathbf{L}_2^\top$. Then, the parameters in (1) are $\boldsymbol{\theta} = \{\mathbf{B}, \mathbf{L}_1, \mathbf{L}_2, \tau^2\}$. The E step treats the random effects $\mathbf{A}_1, \ldots, \mathbf{A}_n$ as missing data and replaces the log-likelihood with a minorizer, which is analytically tractable due to the set up in (1). The M step maximizes the E step minorizer through a series of closed-form conditional maximizations; however, this approach requires $P + Q_1^2 + Q_2^2 \ll n$ for numerical stability. For vector-valued covariates, this algorithm reduces to the EM algorithm in van Dyk (2000) for parameter estimation in mixed model linear regression.

In the high-dimensional settings where $n \ll P + Q_1^2 + Q_2^2$, the previous EM algorithm

它排除了随机效应协变量。在该模型中，随机效应张量具有可分离结构，估计算法是翻转算法的一种变体；然而，他们的方法仅适用于对于每个 i 和 j， $n \ll Q$ 和 $\mathbf{Z}_{ij} = \mathbf{I}_{Q_1}$ 的情况，这意味着这种方法不能用于MMTR的参数估计。

MMTR与基于广义估计方程（GEE）的张量混合模型密切相关，后者被调优用于建模纵向成像数据（Zhang等人，2019年）。基于GEE的模型具有标量响应和张量协变量。对于二维张量，基于GEE的模型和MMTR具有相同的均值参数化；然而，基于GEE的模型假设回归系数具有低张量秩，而MMTR假设它们是稀疏的。在基于GEE的模型中不存在随机效应项，该模型用样本特定的边际“工作”协方差矩阵来替换它们。基于GEE的模型需要选择秩、工作协方差矩阵形式和惩罚调优参数，以使用最小化最大化（MM）算法进行参数估计。相比之下，MMTR的参数估计算法只需要两个调优参数，这些参数决定了$\mathbf{B}$ 的稀疏性和 $\mathbf{\Sigma}_1$ 和 $\mathbf{\Sigma}_2$的秩。

1.2 我们的贡献

MMTR 采用正则化 EM 算法来估计公式 (1) 中的参数。令 $\mathbf{L}_1$ 和 $\mathbf{L}_2$ 为两个矩阵，使得 $\mathbf{\Sigma}_1 = \mathbf{L}_1\mathbf{L}_1^\top$ 和 $\mathbf{\Sigma}_2 = \mathbf{L}_2\mathbf{L}_2^\top$。然后，公式 (1) 中的参数是 $\boldsymbol{\theta} = \{\mathbf{B}, \mathbf{L}_1, \mathbf{L}_2, \tau^2\}$。E 步将随机效应 $\mathbf{A}_1, \ldots, \mathbf{A}_n$ 视为缺失数据，并用一个最小化函数替换对数似然，该最小化函数由于公式 (1) 的设置而解析上可处理。M 步通过一系列闭式条件最大化来最大化 E 步的最小化函数；然而，这种方法需要 $P + Q_1^2 + Q_2^2 \ll n$ 以确保数值稳定性。对于向量值协变量，该算法简化为 van Dyk (2000) 中用于混合模型线性回归参数估计的 EM 算法。

在高维设置中 $n \ll P + Q_1^2 + Q_2^2$，之前的 EM 算法

requires regularization. We assume that $\mathbf{B}$ is sparse and that $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are approximately low rank; that is, $\mathbf{\Sigma}_1 \approx \mathbf{L}_1\mathbf{L}_1^\top$ and $\mathbf{\Sigma}_2 \approx \mathbf{L}_2\mathbf{L}_2^\top$, where $\mathbf{L}_k \in \mathbb{R}^{Q_k \times S_k}$, and $S_k \ll Q_k$ for $k = 1, 2$, reducing the parameter dimension from $O(P_1P_2 + Q_1^2 + Q_2^2)$ to $O(P_1P_2 + Q_1S_1 + Q_2S_2)$. The $\mathbf{\Sigma}_k$ estimates depend on S_k choice, so we use group lasso penalties on the columns of $\mathbf{L}_k$ for a data-driven choice of S_k. The low-rank factorization of $\mathbf{\Sigma}_k$ resembles a factor-analytic structure, so we set $S_k = O(\log Q_k)$ initially (Ročková and George, 2016; Srivastava et al., 2017). This idea extends the low-rank random effects in mixed model linear regression to low-rank matrix-variate random effects in MMTR (James et al., 2000; Heiling et al., 2024). This setup implies that the maximization of the E step objective reduces to a series of regularized least squares problems, enabling estimation of $\mathbf{B}, \tau^2, \mathbf{L}_1$ and $\mathbf{L}_2$ via `scalreg` (Sun, 2019) and `gglasso` (Yang et al., 2024) *R* packages; see Section 3 for the details.

Our main contributions are threefold. First, MMTR is a novel extension of mixed models for matrix-valued covariates. We establish conditions for the separate estimation of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$, given that only $\mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1$ is identified in (1); see Section 2.1. Second, random effects $\mathbf{A}_i$'s in (1) have a low-rank separable covariance structure, which is a natural extension of separable covariance arrays to high-dimensional settings (Hoff, 2011). MMTR is also related to separable factor analysis (SFA) in that SFA adds diagonal matrices with positive entries to $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$, implying that MMTR's covariance structure is more parsimonious than SFA when $\min(P_1, P_2)$ is large (Fosdick and Hoff, 2014); see Section 2.2. Finally, we develop an EM algorithm for efficient parameter regularized estimation and data-driven choice of the ranks of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$; see Section 3. MMTR achieves lower estimation and prediction errors than its high-dimensional mixed model competitors across diverse simulated and real data analyses.

需要正则化。我们假设$\mathbf{B}$ 是稀疏的，并且 $\mathbf{\Sigma}_1$ 和 $\mathbf{\Sigma}_2$ 大致是低秩的；即， $\mathbf{\Sigma}_1 \approx \mathbf{L}_1\mathbf{L}_1^\top$ 和 $\mathbf{\Sigma}_2 \approx \mathbf{L}_2\mathbf{L}_2^\top$，其中 $\mathbf{L}_k \in \mathbb{R}^{Q_k \times S_k}$，以及 $S_k \ll Q_k$ 对于 k =1， 2，将参数维度从 $O(P_1P_2 + Q_1^2 + Q_2^2)$ 减少到 $O(P_1P_2 + Q_1S_1 + Q_2S_2)$。 $\mathbf{\Sigma}_k$ 估计取决于 S_k 选择，因此我们对$\mathbf{L}_k$的列使用组套索惩罚，以进行数据驱动的 S_k选择。 $\mathbf{\Sigma}_k$ 的低秩分解类似于因子分析结构，因此我们最初设置 $S_k = O(\log Q_k)$ （Roˇckov´a 和 George，2016；Srivastava 等人，2017）。这个想法将混合模型线性回归中的低秩随机效应扩展到 MMTR（ James 等人，2000；Heiling 等人，2024）中的低秩矩阵变量随机效应。这种设置意味着 E步目标的最大化减少为一系列正则化最小二乘问题，从而能够通过 `scalreg` (Sun，2019) 和 `gglasso` (Yang 等人，2024) *R* 包估计 $\mathbf{B}$ τ^2， $\mathbf{L}_1$和 $\mathbf{L}_2$；有关详细信息，请参阅第 3 节。

我们的主要贡献有三方面。首先，MMTR 是一种针对矩阵值协变量的混合模型的创新扩展。我们建立了在（1）中仅识别 $\mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1$ 的情况下，对 $\mathbf{\Sigma}_1$ 和 $\mathbf{\Sigma}_2$ 进行分别估计的条件；参见第 2.1 节。其次，(1) 中的随机效应 $\mathbf{A}_i$'s 具有低秩可分离协方差结构，这是将可分离协方差数组自然扩展到高维设置（Hoff, 2011）的延伸。MMTR 还与可分离因子分析（SFA）相关，因为 SFA 向 $\mathbf{\Sigma}_1$ 和 $\mathbf{\Sigma}_2$ 添加具有正条目的对角矩阵，这意味着当 $\min(P_1, P_2)$ 较大时，MMTR 的协方差结构比 SFA 更简洁（Fosdick and Hoff, 2014）；参见第 2.2 节。最后，我们开发了一种 EM 算法，用于高效的参数正则化估计和数据驱动选择 $\mathbf{\Sigma}_1$ 和 $\mathbf{\Sigma}_2$; 的秩；参见第 3 节。在多种模拟和真实数据分析中，MMTR 比其高维混合模型竞争对手实现了更低的估计和预测误差。

2 Mixed Model Trace Regression

2.1 Model

Consider the parameter expanded form of (1). Let $\mathbf{L}_k \in \mathbb{R}^{Q_k \times S_k}$ be the square root matrix factor of $\mathbf{\Sigma}_k$ for $k = 1, 2$. If $\mathbf{L}_k$ has full rank for $k = 1, 2$, then $Q_k = S_k$ and $\mathbf{\Sigma}_k = \mathbf{L}_k \mathbf{L}_k^\top$. MMTR imposes a low-rank structure on $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ by assuming that $S_1 \ll Q_1$ and $S_2 \ll Q_2$, so that $\mathbf{L}_1 \mathbf{L}_1^\top \approx \mathbf{\Sigma}_1$ and $\mathbf{L}_2 \mathbf{L}_2^\top \approx \mathbf{\Sigma}_2$. The random effects are expressed using $\mathbf{L}_1$ and $\mathbf{L}_2$ as

$$\mathbf{A}_i = \mathbf{L}_1 \mathbf{C}_i \mathbf{L}_2^\top, \quad \mathbf{C}_i \sim N_{S_1,S_2}(\mathbf{0}, \tau^2 \mathbf{I}_{S_1}, \mathbf{I}_{S_2}), \tag{2}$$

where $\mathbf{C}_i$ is a $S_1 \times S_2$ matrix of independent normal random variables with mean 0 and variance τ^2. Based on these assumptions, the parameter expanded form of MMTR in (1) is

$$y_{ij} = \mathrm{tr}(\mathbf{X}_{ij}^\top \mathbf{B}) + \mathrm{tr}(\mathbf{Z}_{ij}^\top \mathbf{L}_1 \mathbf{C}_i \mathbf{L}_2^\top) + e_{ij}, \quad e_{ij} \sim N(0, \tau^2). \tag{3}$$

The model parameters are $\boldsymbol{\theta} = (\mathbf{B}, \mathbf{L}_1, \mathbf{L}_2, \tau^2)$, and $\mathbf{L}_k$ for $k = 1, 2$ is non-identified because replacing $\mathbf{L}_k$ with $\mathbf{L}_k \mathbf{O}_k$ for any orthonormal matrix $\mathbf{O}_k$ leaves the decomposition of $\mathbf{\Sigma}_k$ in (2) unchanged. We do not impose identifiability conditions on $\mathbf{L}_1$ and $\mathbf{L}_2$ because they do not impact the estimates of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$.

MMTR in (3) reduces to the trace regression model when $m_i = 1$ and $\mathbf{A}_i = \mathbf{0}$ for every $i = 1, \ldots, n$ (Slawski et al., 2015; Zhao et al., 2017; Fan et al., 2019). These methods estimate $\mathbf{B}$ under different regularization schemes using the lasso-type and nuclear norm penalties. For example, $\| \mathbf{B} \|_1 = \sum_{ij} |b_{ij}|$, $\| \mathbf{B} \|_{1,2} = \sum_{i=1}^{P_1} \| \mathbf{b}_{i:} \|_2$, and $\| \mathbf{B} \|_{2,1} = \sum_{j=1}^{P_2} \| \mathbf{b}_{:j} \|_2$, where b_{ij}, $\mathbf{b}_{i:}$, and $\mathbf{b}_{:j}$ are the (i, j)th entry, ith row, and jth column of $\mathbf{B}$. The application of these three penalties results in sparse, row sparse, and column sparse $\mathbf{B}$ estimates, which

2 混合模型追踪回归

2.1 模型

考虑(1)的参数扩展形式。令 $\mathbf{L}_k \in \mathbb{R}^{Q_k \times S_k}$ 是 $\mathbf{\Sigma}_k$的平方根矩阵因子，对于 $k = 1$,2。如果 $\mathbf{L}_k$对于 $k = 1$ 2,具有满秩，那么 $Q_k = S_k$和 $\mathbf{\Sigma}_k = \mathbf{L}_k \mathbf{L}_k^\top$。MMTR通过对 $\mathbf{\Sigma}_1$ 和 $\mathbf{\Sigma}_2$ 施加低秩结构来假设 $S_1 \ll Q_1$ 和$S_2 \ll Q_2$，以便 $\mathbf{L}_1 \mathbf{L}_1^\top \approx \mathbf{\Sigma}_1$和 $\mathbf{L}_2 \mathbf{L}_2^\top \approx \mathbf{\Sigma}_2$。随机效应使用$\mathbf{L}_1$ 和$\mathbf{L}_2$ 表示

$$\mathbf{A}_i = \mathbf{L}_1 \mathbf{C}_i \mathbf{L}_2^\top, \quad \mathbf{C}_i \sim N_{S_1,S_2}(\mathbf{0}, \tau^2 \mathbf{I}_{S_1}, \mathbf{I}_{S_2}), \tag{2}$$

其中 $\mathbf{C}_i$ 是一个 $S_1 \times S_2$ 矩阵，由均值为0、方差为 τ^2的独立正态随机变量组成。基于这些假设，(1)中MMTR的参数扩展形式是

$$y_{ij} = \mathrm{tr}(\mathbf{X}_{ij}^\top \mathbf{B}) + \mathrm{tr}(\mathbf{Z}_{ij}^\top \mathbf{L}_1 \mathbf{C}_i \mathbf{L}_2^\top) + e_{ij}, \quad e_{ij} \sim N(0, \tau^2). \tag{3}$$

模型参数是 $\boldsymbol{\theta} = (\mathbf{B}, \mathbf{L}_1, \mathbf{L}_2, \tau^2)$，和$\mathbf{L}_k$对于 $k = 1, 2$是非识别的，因为用 $\mathbf{L}_k \mathbf{O}_k$ 替换任何正交矩阵 $\mathbf{O}_k$ 中的 $\mathbf{L}_k$，都会使(2)中的 $\mathbf{\Sigma}_k$ 分解保持不变。我们对$\mathbf{L}_1$ 和$\mathbf{L}_2$ 不施加可识别性条件，因为它们不会影响 $\mathbf{\Sigma}_1$ 和 $\mathbf{\Sigma}_2$的估计。

MMTR在(3)中简化为追踪回归模型，当 $m_i = 1$ 和 $\mathbf{A}_i = \mathbf{0}$ 对于每个$i = 1 \ \ldots \ n$，，(Slawski等人，2015年；Zhao等人，2017年；Fan等人，2019年)。这些方法在不同的正则化方案下使用Lasso型和核范数惩罚来估计$\mathbf{B}$。例如，$\| \mathbf{B} \|_1 = \sum_{ij} |b_{ij}|$, $\| \mathbf{B} \|_{1,2} = \sum_{i=1}^{P_1} \| \mathbf{b}_{i:} \|_2$和 $\| \mathbf{B} \|_{2,1} = \sum_{j=1}^{P_2} \| \mathbf{b}_{:j} \|_2$，其中 b_{ij}，$\mathbf{b}_{i:}$和$\mathbf{b}_{:j}$ 是$\mathbf{B}$的$(i, \ j)$项、 i行和 j列。应用这三种惩罚产生了稀疏、行稀疏和列稀疏的$\mathbf{B}$ 估计，

are efficiently estimated using the alternating direction method of multipliers or projected gradient methods. If $\mathbf{L}_1, \mathbf{L}_2, \tau^2$ are known, then we can use these algorithms for sparse estimation of $\mathbf{B}$. MMTR extensions with $\|\mathbf{B}\|_{1,2}$ and $\|\mathbf{B}\|_{2,1}$ penalties can leverage these estimation algorithms; however, we employ only the $\|\mathbf{B}\|_1$ penalty for simplicity, enabling the direct application of existing algorithms designed for scaled sparse regression (Sun and Zhang, 2012).

The covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ in (1) are identified up to a constant. For $i = 1, \ldots, n$, $\mathbf{A}_i$ in (1) has a separable covariance array; that is, $\text{Cov}\{\text{vec}(\mathbf{A}_i)\} = \tau^2 \mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1$ (Dutilleul, 1999; Lu and Zimmerman, 2005; Hoff, 2011). The covariance array remains unchanged if $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are modified to $c'\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2/c'$ for any $c' > 0$. If σ^2 is the positive "overall variance" parameter, then $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are identified by assuming that $\det(\mathbf{\Sigma}_1) = \det(\mathbf{\Sigma}_2) = 1$ and $\text{Cov}\{\text{vec}(\mathbf{A}_i)\} = \tau^2\sigma^2 \mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1$ (Gerard and Hoff, 2015). This condition is inapplicable for identifying $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ in (1) because they are low-rank matrices.

We enforce the identifiability of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ based on $\mathbf{L}_2$. Let $(\mathbf{M})_{ij}$ denote the (i,j)th element of some matrix $\mathbf{M}$. A popular identifiability condition assumes that $(\mathbf{\Sigma}_2)_{11} = 1$, implying that the first $Q_1 \times Q_1$ block of $\mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1$ equals $\mathbf{\Sigma}_1$ (Dutilleul, 1999). We can extend this condition to the case when $(\mathbf{\Sigma}_2)_{jj} = 1$, which results in the jth diagonal $Q_1 \times Q_1$ block of $\mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1$ equaling $\mathbf{\Sigma}_1$. This condition is useful when $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ have full ranks and are estimated directly; however, MMTR estimates $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ via $\mathbf{L}_1$ and $\mathbf{L}_2$, so this condition requires modification. The following proposition restates this identifiability condition using $\mathbf{L}_2$, which is suited for high-dimensional MMTR applications.

Proposition 2.1 *Assume that* $(\mathbf{L}_2)_{j1} = 1$ *and* $(\mathbf{L}_2)_{j2} = \cdots = (\mathbf{L}_2)_{jS_2} = 0$. *Then,* $(\mathbf{\Sigma}_2)_{jj} = 1$ *and the* j*th diagonal* $Q_1 \times Q_1$ *block of* $\mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1$ *is* $\mathbf{\Sigma}_1$.

The proof of this proposition is given in the supplementary material along with other proofs. Proposition 1.1 does not impose any assumption on the ranks of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$, so it is applicable when $\mathbf{\Sigma}_1$ or $\mathbf{\Sigma}_2$ are rank-deficient. Furthermore, the following proposition

可以使用交替方向乘子法或投影梯度法高效地估计。如果 $\mathbf{L}_1, \mathbf{L}_2, \tau^2$ 已知，那么我们可以使用这些算法对 $\mathbf{B}$ 进行稀疏估计。具有 $\|\mathbf{B}\|_{1\,2}$,和 $\|\mathbf{B}\|_{2\,1}$,惩罚的 MMTR 扩展可以利用这些估计算法；然而，我们仅使用 $\|\mathbf{B}\|_1$ 惩罚以简化，从而能够直接应用为缩放稀疏回归设计的现有算法（孙和张，2012）。

公式 (1) 中的协方差矩阵 $\mathbf{\Sigma}_1$ 和 $\mathbf{\Sigma}_2$ 被识别到常数。对于 i =1, …, n, $\mathbf{A}_i$ 在 (1) 中具有可分离的协方差数组；即，$\text{Cov}\{\text{vec}(\mathbf{A}_i)\} = \tau^2 \mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1$(Dutilleul, 1999; Lu and Zimmerman, 2005; Hoff, 2011)。如果 $\mathbf{\Sigma}_1$ 和 $\mathbf{\Sigma}_2$ 被修改为 $c'\mathbf{\Sigma}_1$ 和 $\mathbf{\Sigma}_2/c'$，则协方差数组保持不变，对于任何 $c' > 0$。如果 σ^2 是正的“总体方差”参数，那么 $\mathbf{\Sigma}_1$ 和 $\mathbf{\Sigma}_2$ 通过假设 $\det(\mathbf{\Sigma}_1)$ =$\det(\mathbf{\Sigma}_2) = 1$ 和 $\text{Cov}\{\text{vec}(\mathbf{A}_i)\} = \tau^2\sigma^2 \mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1$ (杰拉德和霍夫, 2015) 被识别。由于 $\mathbf{\Sigma}_1$ 和 $\mathbf{\Sigma}_2$ 在 (1) 中是低秩矩阵，因此该条件不适用于识别它们。

我们基于 $\mathbf{L}_2$确保 $\mathbf{\Sigma}_1$ 和 $\mathbf{\Sigma}_2$ 的可识别性。令 $(\mathbf{M})_{ij}$ 表示某个矩阵 <code>M</code>的(i,j)个元素。一个流行的可识别性条件假设 $(\mathbf{\Sigma}_2)_{11} = 1$，这意味着 $\mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1$ 的第一个 $Q_1 \times Q_1$ 块等于 $\mathbf{\Sigma}_1$ (Dutilleul, 1999)。我们可以将此条件扩展到 $(\mathbf{\Sigma}_2)_{jj} = 1$的情况，这导致 $\mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1$ 的 j个对角 $Q_1 \times Q_1$ 块等于 $\mathbf{\Sigma}_1$。当 $\mathbf{\Sigma}_1$ 和 $\mathbf{\Sigma}_2$ 满秩且直接估计时，此条件很有用；然而，MMTR通过<code>L</code>$_1$ 和 $\mathbf{L}_2$估计 $\mathbf{\Sigma}_1$ 和 $\mathbf{\Sigma}_2$，因此此条件需要修改。以下命题使用<code>L</code>$_2$重新表述了此可识别性条件，适用于高维MMTR应用。

Proposition 2.1 *Assume that* $(\mathbf{L}_2)_{j1} = 1$ *and* $(\mathbf{L}_2)_{j2} = \cdots = (\mathbf{L}_2)_{jS_2} = 0$. *Then,* $(\mathbf{\Sigma}_2)_{jj} = 1$ *and the* j*th diagonal* $Q_1 \times Q_1$ *block of* $\mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1$ *is* $\mathbf{\Sigma}_1$.

此命题的证明在补充材料中与其他证明一起给出。命题1.1不对 $\mathbf{\Sigma}_1$ 和 $\mathbf{\Sigma}_2$的秩作任何假设，因此当 $\mathbf{\Sigma}_1$ 或 $\mathbf{\Sigma}_2$ 秩亏时适用。此外，以下命题

shows that we can convert any square-root covariance matrix $\mathbf{L}_2$ into the form described in Proposition 1.1

Proposition 2.2 *For any matrix* $\mathbf{L} \in \mathbb{R}^{Q\times S}$*, there exists an orthonormal matrix* $\mathbf{Q}_j \in \mathbb{R}^{S\times S}$ *and constant* c *such that* $c\,\mathbf{L}\,\mathbf{Q}_j^\top$ *is a matrix whose* j*th row is the first standard basis vector in* $\mathbb{R}^S$*.*

Proposition 1.2 states that $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are identified if $\mathbf{\Sigma}_2$ is scaled so that any one of its diagonal values equals 1. We use Proposition 1.2 to evaluate the estimation accuracy of our estimation algorithm. We compare our estimates $\hat{\mathbf{\Sigma}}_1$ and $\hat{\mathbf{\Sigma}}_2$ of the true parameters $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ in the simulations by scaling $\hat{\mathbf{\Sigma}}_1$, $\hat{\mathbf{\Sigma}}_2$, $\mathbf{\Sigma}_1$, and $\mathbf{\Sigma}_2$ so that their respective largest diagonal entries are 1. We also use Proposition 1.2 to scale the $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ estimates at the end of each iteration to further improve the stability of the estimation algorithm.

2.2 Reduction to a Structured Mixed Model

The reduction of MMTR to a structured mixed model linear regression requires the following definitions. Let $\mathbf{b} = \mathrm{vec}(\mathbf{B})$, $\mathbf{x}_{ij} = \mathrm{vec}(\mathbf{X}_{ij})$, $\mathbf{Z}_{ij(1)} = \mathbf{Z}_{ij}$, $\mathbf{Z}_{ij(2)} = \mathbf{Z}_{ij}^\top$, $\mathbf{z}_{ij(1)} = \mathrm{vec}(\mathbf{Z}_{ij(1)})$, $\mathbf{z}_{ij(2)} = \mathrm{vec}(\mathbf{Z}_{ij(2)})$, $\mathbf{a}_{i(1)} = \mathrm{vec}(\mathbf{A}_i)$, $\mathbf{a}_{i(2)} = \mathrm{vec}(\mathbf{A}_i^\top)$, $\mathbf{c}_{i(1)} = \mathrm{vec}(\mathbf{C}_i)$, $\mathbf{c}_{i(2)} = \mathrm{vec}(\mathbf{C}_i^\top)$, $\mathbf{l}_1 = \mathrm{vec}(\mathbf{L}_1)$, $\mathbf{l}_2 = \mathrm{vec}(\mathbf{L}_2)$. Then, identities relating tr, vec, and $\otimes$ operators imply that $\mathbf{A}_i = \mathbf{L}_1\,\mathbf{C}_i\,\mathbf{L}_2^\top$ is equivalent to $\mathbf{a}_{i(1)} = \mathbf{L}_{(1)}\,\mathbf{c}_{(1)}$ and $\mathbf{a}_{i(2)} = \mathbf{L}_{(2)}\,\mathbf{c}_{i(2)}$, where $\mathbf{L}_{(1)} = \mathbf{L}_2 \otimes \mathbf{L}_1$ and $\mathbf{L}_{(2)} = \mathbf{L}_1 \otimes \mathbf{L}_2$; see Seber (2007). The vectors $\mathbf{a}_{i(1)}$'s and $\mathbf{a}_{i(2)}$'s are used in estimating $\mathbf{L}_1$ given $\mathbf{L}_2$ and $\mathbf{L}_2$ given $\mathbf{L}_1$, respectively.

MMTR in (3) is equivalent to a mixed model linear regression with the mean parameter $\mathbf{b}$ and random effects covariance $\tau^2\,\mathbf{L}_2\,\mathbf{L}_2^\top \otimes \mathbf{L}_1\,\mathbf{L}_1^\top$. Using (3) and previous identities,

$$y_{ij} = \mathbf{x}_{ij}^\top\,\mathbf{b} + \mathbf{z}_{ij(1)}^\top(\mathbf{L}_2 \otimes \mathbf{L}_1)\,\mathbf{c}_{i(1)} + e_{ij}, \quad \mathbf{y}_i = \mathbf{X}_i\,\mathbf{b} + \mathbf{Z}_{i(1)}(\mathbf{L}_2 \otimes \mathbf{L}_1)\,\mathbf{c}_{i(1)} + \mathbf{e}_i, \qquad (4)$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$, where $\mathbf{X}_i$ is the matrix whose jth row is $\mathbf{x}_{ij}^\top$ and

表明我们可以将任何平方根协方差矩阵 $\mathbf{L}_2$ 转换为命题1.1中描述的形式

命题2.2 对于任何矩阵 $\mathbf{L} \in \mathbb{R}^{Q\times S}$，都存在一个正交矩阵 $\mathbf{Q}_j \in \mathbb{R}^{S\times S}$ 和常数 c，使得 $c\,\mathbf{L}\,\mathbf{Q}_j^\top$ 是一个矩阵，其 j行是 $\mathbb{R}^S$ 中的第一个标准基向量。

命题1.2指出，如果 $\mathbf{\Sigma}_2$ 被缩放，使得其任意一个对角线值等于1，则 $\mathbf{\Sigma}_1$ 和 $\mathbf{\Sigma}_2$ 被识别。我们使用命题1.2来评估我们估计算法的估计精度。我们在模拟中通过缩放 $\hat{}\mathbf{\Sigma}_1$、$\hat{}\mathbf{\Sigma}_2$、$\mathbf{\Sigma}_1$ 和 $\mathbf{\Sigma}_2$，使它们各自的最大对角线条目为1，来比较我们对真实参数 $\mathbf{\Sigma}_1$ 和 $\mathbf{\Sigma}_2$ 的估计 $\hat{\mathbf{\Sigma}}_1$ 和 $\hat{\mathbf{\Sigma}}_2$。我们还使用命题1.2来缩放每次迭代结束时的 $\mathbf{\Sigma}_1$ 和 $\mathbf{\Sigma}_2$ 估计，以进一步提高估计算法的稳定性。

2.2 降级为结构化混合模型

将MMTR降级为结构化混合模型线性回归需要以下定义。令 $\mathrm{b} = \mathrm{vec(B)}$, $\mathrm{x}_{ij} = \mathrm{vec(X}_{ij})$, $\mathrm{Z}_{ij(1)} = \mathrm{Z}_{ij}$, $\mathrm{Z}_{ij(2)} = \mathrm{Z}_{ij}^\top$, $\mathrm{z}_{ij(1)} = \mathrm{vec(Z}_{ij(1)})$, $\mathrm{z}_{ij(2)} = \mathrm{vec(Z}_{ij(2)})$, $\mathrm{a}_{i(1)} = \mathrm{vec(A}_i)$, $\mathrm{a}_{i(2)} = \mathrm{vec(A}_i^\top)$, $\mathrm{c}_{i(1)} = \mathrm{vec(C}_i)$, $\mathrm{c}_{i(2)} = \mathrm{vec(C}_i^\top)$, $\mathrm{l}_1 = \mathrm{vec(L1)}$, $\mathrm{l}_2 = \mathrm{vec(L2)}$。然后，与tr、vec和 $\otimes$ 运算符相关的恒等式意味着 $\mathbf{A}_i = \mathbf{L}_1\,\mathbf{C}_i\,\mathbf{L}_2^\top$ 等效于 $\mathrm{a}_{i(1)} = \mathrm{L}_{(1)}\,\mathrm{c}_{(1)}$ 和 $\mathrm{a}_{i(2)} = \mathrm{L}_{(2)}\,\mathrm{c}_{i(2)}$，其中 $\mathbf{L}_{(1)} = \mathbf{L}_2 \otimes \mathbf{L}_1$ 和 $\mathbf{L}_{(2)} = \mathbf{L}_1 \otimes \mathbf{L}_2$；参见Seber (2007)。向量 $\mathrm{a}_{i(1)}$'s和 $\mathbf{a}_{i(2)}$'s 分别用于根据L2估计L1和根据L1估计L2。

MMTR 在 (3) 中等效于具有均值参数 $\mathbf{b}$ 和随机效应协方差 $\tau^2\,\mathbf{L}_2\,\mathbf{L}_2^\top \otimes \mathbf{L}_1\,\mathbf{L}_1^\top$ 的混合模型线性回归。使用 (3) 和先前的恒等式，

$$y_{ij} = \mathbf{x}_{ij}^\top\,\mathbf{b} + \mathbf{z}_{ij(1)}^\top(\mathbf{L}_2 \otimes \mathbf{L}_1)\,\mathbf{c}_{i(1)} + e_{ij}, \quad \mathbf{y}_i = \mathbf{X}_i\,\mathbf{b} + \mathbf{Z}_{i(1)}(\mathbf{L}_2 \otimes \mathbf{L}_1)\,\mathbf{c}_{i(1)} + \mathbf{e}_i, \qquad (4)$$

对于 $i = 1, \ldots, n$ 和 $j = 1, \ldots, m_i$，其中 $\mathbf{X}_i$ 是一个矩阵，其 j 行是 $\mathbf{x}_{ij}^\top$ 并且

$\mathbf{Z}_{i(1)}$ is the matrix whose jth row is $\mathbf{z}_{ij(1)}^\top$. The first equation uses $\mathrm{tr}(\mathbf{Z}_{ij}^\top \mathbf{L}_1 \mathbf{C}_i \mathbf{L}_2^\top) = \mathbf{z}_{ij(1)}^\top \mathrm{vec}(\mathbf{L}_1 \mathbf{C}_i \mathbf{L}_2^\top) = \mathbf{z}_{ij(1)}^\top (\mathbf{L}_2 \otimes \mathbf{L}_1)\, \mathbf{c}_{i(1)}$, which implies that the covariance matrix of random effects is $\mathrm{Cov}\{(\mathbf{L}_2 \otimes \mathbf{L}_1)\, \mathbf{c}_{i(1)}\} = \tau^2\, \mathbf{L}_2 \mathbf{L}_2^\top \otimes \mathbf{L}_1 \mathbf{L}_1^\top$. If $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are full-rank and n and $m_1, \ldots, m_n$ are sufficiently large, then we estimate $\mathbf{B} = \mathrm{vec}^{-1}(\mathbf{b})$ and $\mathbf{\Sigma} = \mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1$ using existing software such as `lme4`. If $\|\cdot\|_F$ is the Frobenius norm, then the $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ estimates minimize $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1\|_F$ and are obtained using the singular value decomposition (SVD) of a matrix formed by reordering $\hat{\mathbf{\Sigma}}$ entries (Van Loan, 2000). The complexity of computing $\hat{\mathbf{\Sigma}}_1$ and $\hat{\mathbf{\Sigma}}_2$ using this method scales as $O(Q_1^3 Q_2^3)$, which leads to inefficiency in practice even for small Q_1 and Q_2. Furthermore, high-dimensional extensions of mixed models face similar issues.

The quasi-likelihood approaches bypass these problems by replacing $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ with proxy matrices that have certain asymptotic properties. The resulting loss for the estimation of $\mathbf{b}$ is equivalent to that of weighted least squares regression. The penalized $\mathbf{b}$ estimate is obtained using a lasso-type penalty on $\mathbf{b}$ that has optimal theoretical properties depending on the choice of proxy matrices and tuning parameter (Fan and Li, 2012; Li et al., 2022). The estimation of $\mathbf{\Sigma}$ is still challenging in that it requires that $Q_1 Q_2 < \min_i m_i$, limiting their practical applications (Li et al., 2022).

Due to these reasons, the structured mixed model for MMTR has two forms that enable estimation of $\mathbf{L}_1$ given $\mathbf{L}_2$ and vice versa. These two models facilitate the estimation of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ without any previous restrictions. For $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$, $\mathrm{tr}(\mathbf{X}_{ij}^\top \mathbf{B}) = \mathbf{x}_{ij}^\top \mathbf{b}$, $\mathrm{tr}(\mathbf{Z}_{ij}^\top \mathbf{A}_i) = \mathbf{z}_{ij(1)}^\top \mathbf{a}_{i(1)} = \mathbf{z}_{ij(1)}^\top \mathbf{L}_{(1)}\, \mathbf{c}_{(1)}$, and $\mathrm{tr}(\mathbf{Z}_{ij}\, \mathbf{A}_i^\top) = \mathbf{z}_{ij(2)}^\top \mathbf{a}_{i(2)} = \mathbf{z}_{ij(2)}^\top \mathbf{L}_{(2)}\, \mathbf{c}_{(2)}$. Substituting these identities in (3) gives

$$y_{ij} = \mathbf{x}_{ij}^\top \mathbf{b} + \mathbf{z}_{ij(1)}^\top \mathbf{L}_{(1)}\, \mathbf{c}_{i(1)} + e_{ij} = \mathbf{x}_{ij}^\top \mathbf{b} + \mathbf{z}_{ij(2)}^\top \mathbf{L}_{(2)}\, \mathbf{c}_{i(2)} + e_{ij}, \quad e_{ij} \sim N(0, \tau^2), \qquad (5)$$

$\mathbf{Z}_{i(1)}$ 是矩阵，其 j行是 $\mathbf{z}_{ij(1)}^\top$。第一个方程使用 $\mathrm{tr}(\mathbf{Z}_{ij}^\top \mathbf{L}_1 \mathbf{C}_i \mathbf{L}_2^\top) = \mathbf{z}_{ij(1)}^\top \mathrm{vec}(\mathbf{L}_1 \mathbf{C}_i \mathbf{L}_2^\top) = \mathbf{z}_{ij(1)}^\top (\mathbf{L}_2 \otimes \mathbf{L}_1)\, \mathbf{c}_{i(1)}$，这意味着随机效应的协方差矩阵是 $\mathrm{Cov}\{(\mathbf{L}_2 \otimes \mathbf{L}_1)\, \mathbf{c}_{i(1)}\} = \tau^2\, \mathbf{L}_2 \mathbf{L}_2^\top \otimes \mathbf{L}_1 \mathbf{L}_1^\top$。如果 $\mathbf{\Sigma}_1$和 $\mathbf{\Sigma}_2$是满秩的，并且 n和 $m_1, \ldots, m_n$足够大，那么我们使用现有的软件（如 `lme4`）来估计 $\mathbf{B} = \mathrm{vec}^{-1}(\mathbf{b})$ 和 $\mathbf{\Sigma} = \mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1$。如果 $\|\cdot\|_F$是Frobenius范数，那么 $\mathbf{\Sigma}_1$ 和 $\mathbf{\Sigma}_2$估计最小化 $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1\|_F$，并且通过重新排序 $\hat{\mathbf{\Sigma}}$ 条目形成的矩阵的奇异值分解（SVD）获得。使用这种方法计算 $\hat{\mathbf{\Sigma}}_1$ 和 $\hat{\mathbf{\Sigma}}_2$ 的复杂度按 $O(Q_1^3 Q_2^3)$缩放，即使对于小的 Q_1 和 Q_2，在实践中也会导致低效。此外，混合模型的高维扩展也面临类似问题。

准似然方法通过用具有某些渐近性质的代理矩阵替换 $\mathbf{\Sigma}_1$ 和 $\mathbf{\Sigma}_2$ 来绕过这些问题。由此产生的估计 $\mathbf{b}$ 的损失等效于加权最小二乘回归的损失。使用对 $\mathbf{b}$ 的 Lasso 型惩罚来获得惩罚后的 $\mathbf{b}$ 估计，该惩罚具有最优的理论性质，这取决于代理矩阵和调整参数的选择（Fan and Li, 2012; Li et al., 2022）。$\mathbf{\Sigma}$ 的估计仍然具有挑战性，因为它需要 $Q_1 Q_2 < \min_i m_i$，这限制了它们的实际应用（Li et al., 2022）。

由于这些原因，MMTR 的结构化混合模型有两种形式，它们能够根据 $\mathbf{L}_1$ 估计 $\mathbf{L}_2$，反之亦然。这两种模型促进了 $\mathbf{\Sigma}_1$ 和 $\mathbf{\Sigma}_2$ 的估计，而无需任何先前的限制。对于 $i = 1 \ldots n$，，和 $j = 1, \ldots, m_i$，$\mathrm{tr}(\mathbf{X}_{ij}^\top \mathbf{B}) = \mathbf{x}_{ij}^\top \mathbf{b}$，$\mathrm{tr}(\mathbf{Z}_{ij}^\top \mathbf{A}_i) = \mathbf{z}_{ij(1)}^\top \mathbf{a}_{i(1)} = \mathbf{z}_{ij(1)}^\top \mathbf{L}_{(1)}\, \mathbf{c}_{(1)}$，和 $\mathrm{tr}(\mathbf{Z}_{ij}\, \mathbf{A}_i^\top) = \mathbf{z}_{ij(2)}^\top \mathbf{a}_{i(2)} = \mathbf{z}_{ij(2)}^\top \mathbf{L}_{(2)}\, \mathbf{c}_{(2)}$。将这些恒等式代入 (3) 得到

$$y_{ij} = \mathbf{x}_{ij}^\top \mathbf{b} + \mathbf{z}_{ij(1)}^\top \mathbf{L}_{(1)}\, \mathbf{c}_{i(1)} + e_{ij} = \mathbf{x}_{ij}^\top \mathbf{b} + \mathbf{z}_{ij(2)}^\top \mathbf{L}_{(2)}\, \mathbf{c}_{i(2)} + e_{ij}, \quad e_{ij} \sim N(0, \tau^2), \qquad (5)$$

where $\mathbf{c}_{i(1)}$ and $\mathbf{c}_{i(2)}$ are distributed as $N(0, \tau^2 \mathbf{I}_{S_1 S_2})$. Combining (5) for $j = 1, \ldots, m_i$ gives

$$\mathbf{y}_i = \mathbf{X}_i \, \mathbf{b} + \mathbf{Z}_{i(1)} \, \mathbf{L}_{(1)} \, \mathbf{c}_{i(1)} + \mathbf{e}_i, \quad \mathbf{y}_i = \mathbf{X}_i \, \mathbf{b} + \mathbf{Z}_{i(2)} \, \mathbf{L}_{(2)} \, \mathbf{c}_{i(2)} + \mathbf{e}_i, \quad \mathbf{e}_i \sim N(0, \tau^2 \mathbf{I}_{m_i}), \quad (6)$$

for $i = 1, \ldots, n,$, where the jth row of $\mathbf{X}_i$ is $\mathbf{x}_{ij}^\top$, the jth row of $\mathbf{Z}_{i(1)}$ is $\mathbf{z}_{ij(1)}^\top$, and the jth row of $\mathbf{Z}_{i(2)}$ is $\mathbf{z}_{ij(2)}^\top$. The two models in (6) are used for estimating $\mathbf{L}_1$ given $\mathbf{L}_2$ and vice versa.

We regularize $\mathbf{B}, \mathbf{L}_1, \mathbf{L}_2$ to ensure numerically stable parameter estimation in the high-dimensional settings. As noted earlier, the trace regression literature offers a range of penalty options for $\mathbf{B}$, but we use the lasso penalty because it facilitates efficient estimation via the `scalreg` package. An additional advantage of using `scalreg` is that its τ^2 estimate has a smaller bias compared to other alternatives, including `glmnet` (Friedman et al., 2010). The penalties on $\mathbf{L}_1$ and $\mathbf{L}_2$ are chosen to avoid the need for the exact specification of S_1 and S_2, which are typically unknown. The $\mathbf{L}_1$ and $\mathbf{L}_2$ matrices define low-rank row and column covariance matrices in that $\mathbf{\Sigma}_k \approx \mathbf{L}_k \, \mathbf{L}_k^\top$ and $S_k \ll Q_k$ for $k = 1, 2$. Leveraging similar results in factor analysis that bypass the number of latent factors specification, we set $S_1 = O(\log Q_1)$ and $S_2 = O(\log Q_2)$ and use the group lasso penalty on the columns of $\mathbf{L}_1$ and $\mathbf{L}_2$ (Ročková and George, 2016; Srivastava et al., 2017). This penalty offers a data-driven approach for estimating S_1 and S_2, where their estimates correspond to the highest column index with nonzero entries.

3 Regularized Parameter Estimation

We employ a regularized alternating expected-conditional maximization (AECM) algorithm with three cycles for parameter estimation. This algorithm is an EM extension that estimates $(\mathbf{b}, \tau^2)$, $\mathbf{L}_1$, and $\mathbf{L}_2$ across its three cycles, with each cycle conditioning on the remaining parameters. The first cycle estimates $(\mathbf{b}, \tau^2)$ using regularized weighted least

在 $\mathbf{c}_{i(1)}$和 $\mathbf{c}_{i(2)}$分布为 $N(0, \tau^2 \mathbf{I}_{S_1 S_2})$的地方。结合 (5) 对于 $j = 1, \ldots, m_i$得到

$$\mathbf{y}_i = \mathbf{X}_i \, \mathbf{b} + \mathbf{Z}_{i(1)} \, \mathbf{L}_{(1)} \, \mathbf{c}_{i(1)} + \mathbf{e}_i, \quad \mathbf{y}_i = \mathbf{X}_i \, \mathbf{b} + \mathbf{Z}_{i(2)} \, \mathbf{L}_{(2)} \, \mathbf{c}_{i(2)} + \mathbf{e}_i, \quad \mathbf{e}_i \sim N(0, \tau^2 \mathbf{I}_{m_i}), \quad (6)$$

对于 $i = 1 \ \ldots \ n,$, 其中 $\mathbf{X}_i$的第 j行是 $\mathbf{x}_{ij}^\top$, $\mathbf{Z}_{i(1)}$的第 j行是 $\mathbf{z}_{ij(1)}^\top$, 以及 $\mathbf{Z}_{i(2)}$ 的第 j 行是 $\mathbf{z}_{ij(2)}^\top$。公式 (6) 中的两个模型用于估计 $\mathbf{L}_1$ 给定$\mathbf{L}_2$ 以及反过来。

我们对 $\mathbf{B}$,$\mathbf{L}_1$,$\mathbf{L}_2$ 进行正则化，以确保在高维设置中的数值稳定参数估计。如前所述，迹回归文献为$\mathbf{B}$提供了一系列惩罚选项，但我们使用 Lasso 惩罚，因为它通过 `scalreg` 包促进高效估计。使用 `scalreg` 的另一个优点是，其 τ^2 估计与其他替代方案相比具有更小的偏差，包括 `glmnet` (Friedman 等人，2010)。对 $\mathbf{L}_1$ 和 $\mathbf{L}_2$ 的惩罚选择是为了避免需要精确指定 S_1和 S_2，它们通常是未知的。矩阵 $\mathbf{L}_1$ 和 $\mathbf{L}_2$ 定义了低秩行和列协方差矩阵，在 $\mathbf{\Sigma}_k \approx \mathbf{L}_k \, \mathbf{L}_k^\top$ 和 $S_k \ll Q_k$ 对于 $k = 1,2$。利用因子分析中类似的绕过潜在因子数量的指定的结果，我们设置 $S_1 = O(\log Q_1)$ 和 $S_2 = O(\log Q_2)$，并对 $\mathbf{L}_1$ 和 $\mathbf{L}_2$ 的列使用组 Lasso 惩罚（Ro$^\vee$ckov´a 和 George，2016；Srivastava 等人，2017）。这种惩罚提供了一种数据驱动的方法来估计 S_1 和 S_2，其中它们的估计对应于具有非零条目的最高列索引。

3 正则化参数估计

我们采用一种具有三个周期的正则化交替期望条件最大化（AECM）算法进行参数估计。该算法是EM扩展，在它的三个周期中估计 ($\mathbf{b}$ τ^2)、$\mathbf{L}_1$ 和 $\mathbf{L}_2$，每个周期都基于剩余参数进行条件估计。第一个周期使用正则化加权最小来估计 ($\mathbf{b}$ τ^2)

squares regression. The second and third cycles treat $\mathbf{c}_{i(2)}$'s and $\mathbf{c}_{i(1)}$'s in (6) as missing data and estimate $\mathbf{L}_1$ and $\mathbf{L}_2$, respectively. The derivation of the parameter updates is in the supplementary material.

Consider the $(t+1)$th AECM iteration. Let $\boldsymbol{\theta}^{(t)} = (\mathbf{b}^{(t)}, \mathbf{L}_1^{(t)}, \mathbf{L}_2^{(t)}, \tau^{2(t)})$ be the parameter estimate at the end of the tth AECM iteration. Define $\boldsymbol{\Lambda}_i^{(t)} = \mathbf{Z}_{i(1)}\, \mathbf{L}_{(1)}^{(t)}\, \mathbf{L}_{(1)}^{(t)\top}\, \mathbf{Z}_{i(1)}^{\top} + \mathbf{I}_{m_i}$ for $i = 1, \ldots, n$, where, as defined in Section 2.2, $\mathbf{L}_{(1)}^{(t)} = \mathbf{L}_2^{(t)} \otimes \mathbf{L}_1^{(t)}$ and $\mathbf{Z}_{i(1)}$ is the matrix whose jth row is $\mathbf{z}_{ij(1)}^{\top}$. The first AECM cycle updates $(\mathbf{b}, \tau^2)$ given $(\mathbf{L}_1^{(t)}, \mathbf{L}_2^{(t)})$. Marginalizing over $\mathbf{c}_{i(1)}$ in (4) implies that

$$\mathbf{y}_i = \mathbf{X}_i\, \mathbf{b} + \bar{\mathbf{e}}_i, \quad \mathbb{E}(\bar{\mathbf{e}}_i) = \mathbf{0}, \quad \mathrm{Cov}(\bar{\mathbf{e}}_i) = \tau^2\, \boldsymbol{\Lambda}_i^{(t)}, \quad i = 1, \ldots, n. \tag{7}$$

Let $\boldsymbol{\Lambda}_i^{1/2(t)}$ be any matrix such that $\boldsymbol{\Lambda}_i^{1/2(t)}\, \boldsymbol{\Lambda}_i^{1/2(t)\top} = \boldsymbol{\Lambda}_i^{(t)}$, and $\breve{\mathbf{y}}^{(t)} \in \mathbb{R}^N$ and $\breve{\mathbf{X}}^{(t)} \in \mathbb{R}^{N \times P_1 P_2}$ be the scaled response and design matrix with $\boldsymbol{\Lambda}_i^{-1/2(t)}\, \mathbf{y}_i \in \mathbb{R}^{m_i}$ and $\boldsymbol{\Lambda}_i^{-1/2(t)}\, \mathbf{X}_i$ as their ith row blocks. Then, (7) implies that the loss for estimating $\mathbf{b}$ is $\| \breve{\mathbf{y}}^{(t)} - \breve{\mathbf{X}}^{(t)}\, \mathbf{b} \|^2$. We jointly estimate $(\mathbf{b}^{(t+1)}, \tau^{2(t+1)})$ by solving the following scaled lasso optimization problem (Sun and Zhang, 2012):

$$(\mathbf{b}^{(t+1)}, \tau^{(t+1)}) = \underset{\mathbf{b} \in \mathbb{R}^{P_1 P_2}, \tau \in \mathbb{R}}{\operatorname{argmin}} \frac{\| \breve{\mathbf{y}}^{(t)} - \breve{\mathbf{X}}^{(t)}\, \mathbf{b} \|^2}{2N\tau} + \frac{\tau}{2} + \lambda_{\mathbf{B}} \| \mathbf{b} \|_1, \tag{8}$$

where $\lambda_{\mathbf{B}} \geq 0$ is a tuning parameter and $\| \mathbf{b} \|_1 = \sum_{i=1}^{P_1} \sum_{j=1}^{P_2} |b_{ij}|$ is the lasso penalty. Given $\lambda_{\mathbf{B}}$, we estimate $\mathbf{b}^{(t+1)}$ and $\tau^{2(t+1)}$ using the `scalreg` package and update $\boldsymbol{\theta}^{(t)}$ to $\boldsymbol{\theta}^{(t+1/3)} = (\mathbf{b}^{(t+1)}, \mathbf{L}_1^{(t)}, \mathbf{L}_2^{(t)}, \tau^{2(t+1)})$, ending the first AECM cycle.

The second cycle treats $\mathbf{c}_{1(2)}, \ldots, \mathbf{c}_{n(2)}$ in (6) as missing data and updates $\mathbf{L}_1^{(t)}$ to $\mathbf{L}_1^{(t+1)}$ given $(\mathbf{b}^{(t+1)}, \mathbf{L}_2^{(t)}, \tau^{2(t+1)})$. For $i = 1, \ldots, n$, (6) implies that the complete data are $(\mathbf{y}_i, \mathbf{c}_{i(2)})$ and $\boldsymbol{\mu}_{i(2)}^{(t)} = \mathbb{E}(\mathbf{c}_{i(2)} \mid \mathbf{y}_i, \boldsymbol{\theta}^{(t+1/3)})$ and $\boldsymbol{\Gamma}_{i(2)}^{(t)} = \mathbb{E}(\mathbf{c}_{i(2)}\, \mathbf{c}_{i(2)}^{\top} \,|\, \mathbf{y}_i, \boldsymbol{\theta}^{(t+1/3)})$ are analytically

平方回归。第二和第三个周期将(6)中的 $\mathbf{c}_{i(2)}$'s和$\mathbf{c}_{i(1)}$视为缺失数据，并分别估计$\mathbf{L}_1$ 和 $\mathbf{L}_2$。参数更新的推导在补充材料中。

考虑$(t+1)$次AECM迭代。令 $\boldsymbol{\theta}^{(t)} = (\mathbf{b}^{(t)}, \mathbf{L}_1^{(t)}, \mathbf{L}_2^{(t)}, \tau^{2(t)})$ 为 t次AECM迭代结束时的参数估计。定义 $\boldsymbol{\Lambda}_i^{(t)} = \mathbf{Z}_{i(1)}\, \mathbf{L}_{(1)}^{(t)}\, \mathbf{L}_{(1)}^{(t)\top}\, \mathbf{Z}_{i(1)}^{\top} + \mathbf{I}_{m_i}$ 对于$i = 1\ \ldots\ n,,,$ 其中，如第2节所述，$2, \mathbf{L}_{(1)}^{(t)} = \mathbf{L}_2^{(t)} \otimes \mathbf{L}_1^{(t)}$ 和 $\mathbf{Z}_{i(1)}$ 是j行是 $\mathbf{z}_{ij(1)}^{\top}$的矩阵。第一个AECM周期更新$(\mathbf{b}\ \tau^2)$给定 $(\mathbf{L}_1^{(t)}, \mathbf{L}_2^{(t)})$。在(4)中对$\mathbf{c}_{i(1)}$ 进行边缘化意味着

$$\mathbf{y}_i = \mathbf{X}_i\, \mathbf{b} + \bar{\mathbf{e}}_i, \quad \mathbb{E}(\mathbf{e}_i) = \mathbf{0}, \quad \mathrm{Cov}(\mathbf{e}_i) = \tau^2\, \boldsymbol{\Lambda}_i^{(t)}, \quad i = 1, \ldots, n. \tag{7}$$

设 $\boldsymbol{\Lambda}_i^{1/2(t)}$ 是任意矩阵，满足 $\boldsymbol{\Lambda}_i^{1/2(t)}\, \boldsymbol{\Lambda}_i^{1/2(t)\top} = \boldsymbol{\Lambda}_i^{(t)}$，且 $\breve{\mathbf{y}}^{(t)} \in \mathbb{R}^N$ 和 $\breve{\mathbf{X}}^{(t)} \in \mathbb{R}^{N \times P_1 P_2}$ 是具有 $\boldsymbol{\Lambda}_i^{-1/2(t)}\, \mathbf{y}_i \in \mathbb{R}^{m_i}$ 和 $\boldsymbol{\Lambda}_i^{-1/2(t)}\, \mathbf{X}_i$ 作为其 i 行块的缩放响应矩阵和设计矩阵。那么，(7) 表明估计 $\mathbf{b}$ 的损失为 $\| \breve{\mathbf{y}}^{(t)} - \breve{\mathbf{X}}^{(t)}\, \mathbf{b} \|^2$。我们通过求解以下缩放 Lasso 优化问题（孙和张，2012）来联合估计 $(\mathbf{b}^{(t+1)}, \tau^{2(t+1)})$：

$$(\mathbf{b}^{(t+1)}, \tau^{(t+1)}) = \underset{\mathbf{b} \in \mathbb{R}^{P_1 P_2}, \tau \in \mathbb{R}}{\operatorname{argmin}} \frac{\| \breve{\mathbf{y}}^{(t)} - \breve{\mathbf{X}}^{(t)}\, \mathbf{b} \|^2}{2N\tau} + \frac{\tau}{2} + \lambda_{\mathbf{B}} \| \mathbf{b} \|_1, \tag{8}$$

其中 $\lambda_{\mathbf{B}} \geq 0$ 是调整参数，$\| \mathbf{b} \|_1 = \sum_{i=1}^{P_1} \sum_{j=1}^{P_2} |b_{ij}|$ 是 Lasso 惩罚。给定 $\lambda_{\mathbf{B}}$，我们使用 `scalreg` 包估计 $\mathbf{b}^{(t+1)}$ 和 $\tau^{2(t+1)}$，并将 $\boldsymbol{\theta}^{(t)}$ 更新为 $\boldsymbol{\theta}^{(t+1/3)} = (\mathbf{b}^{(t+1)}, \mathbf{L}_1^{(t)}, \mathbf{L}_2^{(t)}, \tau^{2(t+1)})$，结束第一个 AECM 周期。

第二个周期将 $\mathbf{c}_{1(2)}\ \ldots\ \mathbf{c}_{n(2)}\,,,$在 (6) 中视为缺失数据，并根据 $(\mathbf{b}^{(t+1)}, \mathbf{L}_2^{(t)}, \tau^{2(t+1)})$ 更新 $\mathbf{L}_1^{(t)}$ 为 $\mathbf{L}_1^{(t+1)}$。对于 $i = 1\ \ldots\ n,,,$ (6) 意味着完整数据是 $(\mathbf{y}_i, \mathbf{c}_{i(2)})$ 和 $\boldsymbol{\mu}_{i(2)}^{(t)} = \mathbb{E}(\mathbf{c}_{i(2)} \mid \mathbf{y}_i, \boldsymbol{\theta}^{(t+1/3)})$，而 $\boldsymbol{\Gamma}_{i(2)}^{(t)} = \mathbb{E}(\mathbf{c}_{i(2)}\, \mathbf{c}_{i(2)}^{\top} \,|\, \mathbf{y}_i, \boldsymbol{\theta}^{(t+1/3)})$ 是通过解析得到的

tractable. If $\mathbf{l}_1 = \text{vec}(\mathbf{L}_1)$, then the second cycle E-step's objective is

$$\begin{aligned}
\mathcal{Q}_{(1)}(\mathbf{l}_1) &\propto -(\mathbf{H}_{(1)}^{-1/2(t)}\,\mathbf{g}_{(1)}^{(t)} - \mathbf{H}_{(1)}^{1/2(t)\top}\,\mathbf{l}_1)^\top(\mathbf{H}_{(1)}^{-1/2(t)}\,\mathbf{g}_{(1)}^{(t)} - \mathbf{H}_{(1)}^{1/2(t)\top}\,\mathbf{l}_1), \qquad (9)\\
\mathbf{H}_{(1)}^{(t)} &= \sum_{i=1}^{n}\sum_{j=1}^{m_i}(\mathbf{I}_{S_1}\otimes\mathbf{Z}_{ij(1)}\,\mathbf{L}_2^{(t)})\,\boldsymbol{\Gamma}_{i(2)}^{(t)}(\mathbf{I}_{S_1}\otimes\mathbf{L}_2^{(t)\top}\,\mathbf{Z}_{ij(1)}^\top),\\
\mathbf{g}_{(1)}^{(t)} &= \sum_{i=1}^{n}\left[\mathbf{I}_{S_1}\otimes\left\{\sum_{j=1}^{m_i}(y_{ij}-\mathbf{x}_{ij}^\top\,\mathbf{b}^{(t+1)})\,\mathbf{Z}_{ij(1)}\right\}\mathbf{L}_2^{(t)}\right]\boldsymbol{\mu}_{i(2)}^{(t)},
\end{aligned}$$

where $\mathbf{H}_{(1)}^{1/2(t)}$ is any matrix such that $\mathbf{H}_{(1)}^{1/2(t)}\,\mathbf{H}_{(1)}^{1/2(t)\top} = \mathbf{H}_{(1)}^{(t)}$ and $\mathbf{H}_{(1)}^{-1/2(t)}$ is any generalized inverse of $\mathbf{H}_{(1)}^{1/2(t)}$. Using (9), $-\mathcal{Q}_{(1)}(\mathbf{l}_1)$ is equivalent to a squared loss for estimating $\mathbf{l}_1$.

The conditional maximization (CM) step in the second cycle minimizes $-\mathcal{Q}_{(1)}(\mathbf{l}_1)$ using the group lasso penalty on the columns of $\mathbf{L}_1$. Let $\mathbf{l}_{(1):i}$ is the ith Q_1-dimensional block of $\mathbf{l}_1$, which corresponds to the ith column of $\mathbf{L}_1$. Then, $\mathbf{l}_1$ is estimated using the `gglasso` package as

$$\mathbf{l}_1^{(t+1)} = \underset{\mathbf{l}_1\in\mathbb{R}^{Q_1 S_1}}{\text{argmin}}\ \|\,\mathbf{H}_{(1)}^{-1/2(t)}\,\mathbf{g}_{(1)}^{(t)} - \mathbf{H}_{(1)}^{1/2(t)\top}\,\mathbf{l}_1\,\|_2^2 + \lambda_{\mathbf{L}}\sum_{i=1}\|\mathbf{l}_{(1):i}\|_2, \qquad (10)$$

where $\mathbf{L}_1^{(t+1)} = \text{vec}^{-1}(\mathbf{l}_1^{(t+1)})$, $\lambda_{\mathbf{L}} \geq 0$ is a tuning parameter, and $\|\cdot\|_2$ is the Euclidean norm. The second cycle ends by updating $\boldsymbol{\theta}^{(t+1/3)}$ to $\boldsymbol{\theta}^{(t+2/3)} = (\mathbf{b}^{(t+1)}, \mathbf{L}_1^{(t+1)}, \mathbf{L}_2^{(t)}, \tau^{2(t+1)})$.

Finally, the third AECM cycle updates $\boldsymbol{\theta}^{(t+2/3)}$ to $\boldsymbol{\theta}^{(t+1)}$ by updating $\mathbf{L}_2$ given $(\mathbf{b}^{(t+1)}, \mathbf{L}_1^{(t+1)}, \tau^{2(t+1)})$. The E and CM steps in this cycle change the indices 1 to 2 and vice versa in the second cycle E and CM steps. Specifically, for $i = 1, \ldots, n$, the complete data are $(\mathbf{y}_i, \mathbf{c}_{i(1)})$ and $\boldsymbol{\mu}_{i(1)}^{(t)} = \mathbb{E}(\mathbf{c}_{i(1)} \mid \mathbf{y}_i, \boldsymbol{\theta}^{(t+2/3)})$ and $\boldsymbol{\Gamma}_{i(1)}^{(t)} = \mathbb{E}(\mathbf{c}_{i(1)}\,\mathbf{c}_{i(1)}^\top \,|\, \mathbf{y}_i, \boldsymbol{\theta}^{(t+2/3)})$ are analytically tractable. The E step objective for the third cycle, $\mathcal{Q}_{(2)}(\mathbf{l}_2)$, swaps the indices 1 and 2 in $\mathcal{Q}_{(1)}(\mathbf{l}_1)$ as defined in (9), where $\mathbf{l}_2 = \text{vec}(\mathbf{L}_2)$ and $\mathbf{H}_{(2)}^{(t)}$ and $\mathbf{g}_{(2)}^{(t)}$ are obtained by swapping indices 1

可处理的。如果 $\mathbf{l}_1 = \text{vec}(\mathbf{L}_1)$，则第二个周期 E 步的目标是

$$\begin{aligned}
\mathcal{Q}_{(1)}(\mathbf{l}_1) &\propto -(\mathbf{H}_{(1)}^{-1/2(t)}\,\mathbf{g}_{(1)}^{(t)} - \mathbf{H}_{(1)}^{1/2(t)\top}\,\mathbf{l}_1)^\top(\mathbf{H}_{(1)}^{-1/2(t)}\,\mathbf{g}_{(1)}^{(t)} - \mathbf{H}_{(1)}^{1/2(t)\top}\,\mathbf{l}_1), \qquad (9)\\
\mathbf{H}_{(1)}^{(t)} &= \sum_{i=1}^{n}\sum_{j=1}^{m_i}(\mathbf{I}_{S_1}\otimes\mathbf{Z}_{ij(1)}\,\mathbf{L}_2^{(t)})\,\boldsymbol{\Gamma}_{i(2)}^{(t)}(\mathbf{I}_{S_1}\otimes\mathbf{L}_2^{(t)\top}\,\mathbf{Z}_{ij(1)}^\top),\\
\mathbf{g}_{(1)}^{(t)} &= \sum_{i=1}^{n}\left[\mathbf{I}_{S_1}\otimes\left\{\sum_{j=1}^{m_i}(y_{ij}-\mathbf{x}_{ij}^\top\,\mathbf{b}^{(t+1)})\,\mathbf{Z}_{ij(1)}\right\}\mathbf{L}_2^{(t)}\right]\boldsymbol{\mu}_{i(2)}^{(t)},
\end{aligned}$$

其中 $\mathbf{H}_{(1)}^{1/2(t)}$ 是任何矩阵，使得 $\mathbf{H}_{(1)}^{1/2(t)}\,\mathbf{H}_{(1)}^{1/2(t)\top} = \mathbf{H}_{(1)}^{(t)}$ 且 $\mathbf{H}_{(1)}^{-1/2(t)}$ 是 $\mathbf{H}_{(1)}^{1/2(t)}$ 的任何广义逆。使用 (9) $-\mathcal{Q}_{(1)}(\mathbf{l}_1)$ 等价于对估计 $\mathbf{l}_1$ 的平方损失。

第二个周期中的条件最大化（CM）步使用组 Lasso 惩罚在 $\mathbf{L}_1$ 的列上最小化 $-\mathcal{Q}_{(1)}(\mathbf{l}_1)$。设 $\mathbf{l}_{(1):i}$ 是 $\mathbf{l}_1$ 的第 i 个 Q_1-维块，它对应于 $\mathbf{L}_1$ 的第 i 列。然后，使用 `gglasso` 包估计 $\mathbf{l}_1$ 为

$$\mathbf{l}_1^{(t+1)} = \underset{\mathbf{l}_1\in\mathbb{R}^{Q_1 S_1}}{\text{argmin}}\ \|\,\mathbf{H}_{(1)}^{-1/2(t)}\,\mathbf{g}_{(1)}^{(t)} - \mathbf{H}_{(1)}^{1/2(t)\top}\,\mathbf{l}_1\,\|_2^2 + \lambda_{\mathbf{L}}\sum_{i=1}\|\mathbf{l}_{(1):i}\|_2, \qquad (10)$$

其中 $\mathbf{L}_1^{(t+1)} = \text{vec}^{-1}(\mathbf{l}_1^{(t+1)})$，$\lambda_{\mathbf{L}} \geq 0$ 是调整参数，$\|\cdot\|_2$ 是欧几里得范数。第二个周期通过更新 $\boldsymbol{\theta}^{(t+1/3)}$ to $\boldsymbol{\theta}^{(t+2/3)} = (\mathbf{b}^{(t+1)}, \mathbf{L}_1^{(t+1)}, \mathbf{L}_2^{(t)}, \tau^{2(t+1)})$. 结束。

最后，第三AECM循环通过更新<code id='3'>L</code><code id='5'>2 </code>给定 $(\mathbf{b}^{(t+1)}, \mathbf{L}_1^{(t+1)}, \tau^{2(t+1)})$来更新 $\boldsymbol{\theta}^{(t+2/3)}$ 到 $\boldsymbol{\theta}^{(t+1)}$。在这个周期中的E步和CM步改变了第二周期E步和CM步中的索引1和2的顺序。具体来说，对于 $i = 1 \ldots n$<code id='9'>， ，</code>，完整数据是 $(\mathbf{y}_i, \mathbf{c}_{i(1)})$ 和 $\boldsymbol{\mu}_{i(1)}^{(t)} = \mathbb{E}(\mathbf{c}_{i(1)} \mid \mathbf{y}_i, \boldsymbol{\theta}^{(t+2/3)})$，而 $\boldsymbol{\Gamma}_{i(1)}^{(t)} = \mathbb{E}(\mathbf{c}_{i(1)}\,\mathbf{c}_{i(1)}^\top \,|\, \mathbf{y}_i, \boldsymbol{\theta}^{(t+2/3)})$ 是解析上可处理的。第三周期的E步目标 $\mathcal{Q}_{(2)}(\mathbf{l}_2)$在 $\mathcal{Q}_{(1)}(\mathbf{l}_1)$ 中交换了索引1和2，如(9)中定义的，其中 $\mathbf{l}_2 = \text{vec}(\mathbf{L}_2)$ 、$\mathbf{H}_{(2)}^{(t)}$ 和 $\mathbf{g}_{(2)}^{(t)}$ 是通过交换索引1获得的。

and 2 in $\mathbf{H}_{(1)}^{(t)}$ and $\mathbf{g}_{(1)}^{(t)}$. Following (10), the CM step in the third cycle is

$$\mathbf{l}_2^{(t+1)} = \underset{\mathbf{l}_2 \in \mathbb{R}^{Q_2 S_2}}{\operatorname{argmin}} \left\| \mathbf{H}_{(2)}^{-1/2(t)} \mathbf{g}_{(2)}^{(t)} - \mathbf{H}_{(2)}^{1/2(t)\top} \mathbf{l}_2 \right\|_2^2 + \lambda_{\mathbf{L}} \sum_{i=1}^{S_2} \|\mathbf{l}_{(2):i}\|_2, \tag{11}$$

where $\mathbf{L}_2^{(t+1)} = \text{vec}^{-1}(\mathbf{l}_2^{(t+1)})$, $\mathbf{l}_{(2):i}$ is the ith Q_2-dimensional block of $\mathbf{l}_2$, which corresponds to the ith column of $\mathbf{L}_2$, and $\mathbf{l}_2^{(t+1)}$ is estimated using the `gglasso` package. This cycle finishes the $(t+1)$th iteration of the AECM algorithm, updating the $\boldsymbol{\theta}^{(t)}$ to $\boldsymbol{\theta}^{(t+3/3)} = \boldsymbol{\theta}^{(t+1)} = (\mathbf{b}^{(t+1)}, \mathbf{L}_1^{(t+1)}, \mathbf{L}_2^{(t+1)}, \tau^{2(t+1)})$. Algorithm 1 summarizes the AECM algorithm.

Algorithm 1: MMTR AECM Algorithm

First cycle:

- **Regularized weighted least squares:** Given $\mathbf{L}_1^{(t)}$ and $\mathbf{L}_2^{(t)}$, compute $\mathbf{B}^{(t+1)}$ and $\tau^{2(t+1)}$ by solving (8). Update $\boldsymbol{\theta}^{(t)}$ to $\boldsymbol{\theta}^{(t+1/3)} = (\mathbf{b}^{(t+1)}, \mathbf{L}_1^{(t)}, \mathbf{L}_2^{(t)}, \tau^{2(t+1)})$.

Second cycle:

- **E step:** Compute $\mathbf{H}_{(1)}^{(t)}$ and $\mathbf{g}_{(1)}^{(t)}$ in (9).
- **CM step:** Given $\boldsymbol{\theta}^{(t+1/3)}$, compute $\mathbf{L}_1^{(t+1)}$ by solving (10). Update $\boldsymbol{\theta}^{(t+1/3)}$ to $\boldsymbol{\theta}^{(t+2/3)} = (\mathbf{b}^{(t+1)}, \mathbf{L}_1^{(t+1)}, \mathbf{L}_2^{(t)}, \tau^{2(t+1)})$.

Third cycle:

- **E step:** Compute $\mathbf{H}_{(2)}^{(t)}$ and $\mathbf{g}_{(2)}^{(t)}$ by swapping 1 and 2 in (9).
- **CM step:** Given $\boldsymbol{\theta}^{(t+2/3)}$, compute $\mathbf{L}_2^{(t+1)}$ by solving (11). Update $\boldsymbol{\theta}^{(t+2/3)}$ to $\boldsymbol{\theta}^{(t+1)} = (\mathbf{b}^{(t+1)}, \mathbf{L}_1^{(t+1)}, \mathbf{L}_2^{(t+1)}, \tau^{2(t+1)})$.

We post-process the $\mathbf{L}_1$ and $\mathbf{L}_2$ estimates at the end of every iteration to improve the efficiency and stability of Algorithm 1. For $t = 1, 2, \ldots, \infty$, any zero columns in $\mathbf{L}_1^{(t)}$ and $\mathbf{L}_2^{(t)}$ are removed. Then, $\mathbf{L}_1^{(t)}$ and $\mathbf{L}_2^{(t)}$ are updated following Proposition 1.2 such that the maximum diagonal values of $\boldsymbol{\Sigma}_1^{(t)}$ and $\boldsymbol{\Sigma}_2^{(t)}$ equal 1 and the "overall variance" $\sigma^{2(t)}$ is split equally between $\boldsymbol{\Sigma}_1^{(t)}$ and $\boldsymbol{\Sigma}_2^{(t)}$. This update maintains similar ranks for $\mathbf{L}_1^{(t)}$ and $\mathbf{L}_2^{(t)}$ across different values of tuning parameter $\lambda_{\mathbf{L}}$ and random initializations $\mathbf{L}_1^{(0)}$ and $\mathbf{L}_2^{(0)}$.

The MMTR's parameter estimate $\boldsymbol{\theta}^{(\infty)}$ is a regularized maximum likehood estimate of

和 $\mathbf{H}_{(1)}^{(t)}$ 和 $\mathbf{g}_{(1)}^{(t)}$ 中的 2。根据 (10)，第三个周期中的 CM 步是

$$\mathbf{l}_2^{(t+1)} = \underset{\mathbf{l}_2 \in \mathbb{R}^{Q_2 S_2}}{\operatorname{argmin}} \left\| \mathbf{H}_{(2)}^{-1/2(t)} \mathbf{g}_{(2)}^{(t)} - \mathbf{H}_{(2)}^{1/2(t)\top} \mathbf{l}_2 \right\|_2^2 + \lambda_{\mathbf{L}} \sum_{i=1}^{S_2} \|\mathbf{l}_{(2):i}\|_2, \tag{11}$$

其中 $\mathbf{L}_2^{(t+1)} = \text{vec}^{-1}(\mathbf{l}_2^{(t+1)})$, $\mathbf{l}_{(2):i}$ 是 $\mathbf{l}_2$ 的第 i 个 Q_2-维块，它对应于 $\mathbf{L}_2$ 的第 i 列，并且 $\mathbf{l}_2^{(t+1)}$ 是使用 `gglasso` 软件包估计的。这个周期完成了 AECM 算法的第 $(t+1)$ 次迭代，将 $\boldsymbol{\theta}^{(t)}$ 更新为 $\boldsymbol{\theta}^{(t+3/3)} = \boldsymbol{\theta}^{(t+1)} = (\mathbf{b}^{(t+1)}, \mathbf{L}_1^{(t+1)}, \mathbf{L}_2^{(t+1)}, \tau^{2(t+1)})$。算法1总结了 AECM 算法。

Algorithm 1: MMTR AECM Algorithm

First cycle:

- **Regularized weighted least squares:** Given $\mathbf{L}_1^{(t)}$ and $\mathbf{L}_2^{(t)}$, compute $\mathbf{B}^{(t+1)}$ and $\tau^{2(t+1)}$ by solving (8). Update $\boldsymbol{\theta}^{(t)}$ to $\boldsymbol{\theta}^{(t+1/3)} = (\mathbf{b}^{(t+1)}, \mathbf{L}_1^{(t)}, \mathbf{L}_2^{(t)}, \tau^{2(t+1)})$.

Second cycle:

- **E step:** Compute $\mathbf{H}_{(1)}^{(t)}$ and $\mathbf{g}_{(1)}^{(t)}$ in (9).
- **CM step:** Given $\boldsymbol{\theta}^{(t+1/3)}$, compute $\mathbf{L}_1^{(t+1)}$ by solving (10). Update $\boldsymbol{\theta}^{(t+1/3)}$ to $\boldsymbol{\theta}^{(t+2/3)} = (\mathbf{b}^{(t+1)}, \mathbf{L}_1^{(t+1)}, \mathbf{L}_2^{(t)}, \tau^{2(t+1)})$.

Third cycle:

- **E step:** Compute $\mathbf{H}_{(2)}^{(t)}$ and $\mathbf{g}_{(2)}^{(t)}$ by swapping 1 and 2 in (9).
- **CM step:** Given $\boldsymbol{\theta}^{(t+2/3)}$, compute $\mathbf{L}_2^{(t+1)}$ by solving (11). Update $\boldsymbol{\theta}^{(t+2/3)}$ to $\boldsymbol{\theta}^{(t+1)} = (\mathbf{b}^{(t+1)}, \mathbf{L}_1^{(t+1)}, \mathbf{L}_2^{(t+1)}, \tau^{2(t+1)})$.

我们在每次迭代的结束时对 $\mathbf{L}_1$ 和 $\mathbf{L}_2$ 估计值进行后处理，以提高算法1的效率和稳定性。对于 $t = 1, 2, \ldots, \infty$，从 $\mathbf{L}_1^{(t)}$ 和 $\mathbf{L}_2^{(t)}$ 中删除任何零列。然后，$\mathbf{L}_1^{(t)}$ 和 $\mathbf{L}_2^{(t)}$ 根据命题1.2进行更新，使得 $\boldsymbol{\Sigma}_1^{(t)}$ 和 $\boldsymbol{\Sigma}_2^{(t)}$ 的最大对角值等于 1，并且“总体方差” $\sigma^{2(t)}$ 平均分配在 $\boldsymbol{\Sigma}_1^{(t)}$ 和 $\boldsymbol{\Sigma}_2^{(t)}$ 之间。此更新在不同值的调整参数 $\lambda_{\mathbf{L}}$ 和随机初始化 $\mathbf{L}_1^{(0)}$ 和 $\mathbf{L}_2^{(0)}$ 下，保持了 $\mathbf{L}_1^{(t)}$ 和 $\mathbf{L}_2^{(t)}$ 的相似排名。

MMTR的参数估计 $\boldsymbol{\theta}^{(\infty)}$ 是一个正则化最大似然估计

$\boldsymbol{\theta}$. Let $\ell(\boldsymbol{\theta})$ be the negative log likehood function implied by (3), $\boldsymbol{\theta}_1 = (\mathbf{B}, \tau^2)$, $\boldsymbol{\theta}_2 = \mathbf{L}_1$, $\boldsymbol{\theta}_3 = \mathbf{L}_2$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$, and $\mathcal{P}_{\lambda_\mathbf{B}}(\boldsymbol{\theta}_1)$, $\mathcal{P}_{\lambda_\mathbf{L}}(\boldsymbol{\theta}_2)$, and $\mathcal{P}_{\lambda_\mathbf{L}}(\boldsymbol{\theta}_3)$ are the penalties on $\mathbf{B}$, $\mathbf{L}_1$, and $\mathbf{L}_2$ in (8), (10), and (11), respectively. Then, the following proposition shows that $\{\boldsymbol{\theta}^{(t)}\}_{t=1}^{\infty}$ sequence produced by Algorithm 1 converges under weak assumptions.

Proposition 3.1 *The MMTR objective is* $f(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) + \mathcal{P}_{\lambda_\mathbf{B}}(\boldsymbol{\theta}_1) + \mathcal{P}_{\lambda_\mathbf{L}}(\boldsymbol{\theta}_2) + \mathcal{P}_{\lambda_\mathbf{L}}(\boldsymbol{\theta}_3)$. *Let* $\mathcal{M}(\cdot)$ *be the function that maps* $\boldsymbol{\theta}^{(t)}$ *to* $\boldsymbol{\theta}^{(t+1)}$ *using Algorithm 1. Then, each iteration of Algorithm 1 does not increase* $f(\boldsymbol{\theta})$. *Furthermore, assume that the parameter space* $\boldsymbol{\Theta}$ *is compact and* $f(\boldsymbol{\theta}) = f\{\mathcal{M}(\boldsymbol{\theta})\}$ *only for the stationary points of* $f(\boldsymbol{\theta})$. *Then, the* $\{\boldsymbol{\theta}^{(t)}\}_{t=1}^{\infty}$ *sequence converges to a stationary point.*

4 Experiments

4.1 Setup

MMTR's performance was evaluated using simulated and real data. The high-dimensional penalized competitors were based on quasi-likelihood (Li et al., 2022), trace regression (Zhao et al., 2017), posterior mode estimation (Heiling et al., 2023), and generalized estimating equations (GEE) (Zhang et al., 2019). The posterior mode estimation algorithm implemented in `glmmPen` R package failed with an error in the pre-screening step, so we do not present comparisons with this method. In low-dimensional settings, we used `lme4` (Bates et al., 2013) for maximum likelihood estimation, resulting in total four competing methods. The penalized quasi-likelihood (PQL) method, implemented using the `scalreg` R package, and `lme4` used MMTR's vectorized form in (5) for parameter estimation. In contrast, trace regression and GEE did not have random effects, so we employed `TensorReg` (Zhou, 2017) and `SparseReg` (Zhou and Gaines, 2017) MATLAB packages to fit trace regression and GEE models. MMTR used Algorithm 1 for parameter estimation.

$\boldsymbol{\theta}$. 令 $\ell(\boldsymbol{\theta})$ 是由 (3) 意味的负对数似然函数， $\boldsymbol{\theta}_1 = (\mathbf{B}\ \tau^2)$ ， $\boldsymbol{\theta}_2 = \mathbf{L}_1$，$\boldsymbol{\theta}_3 = \mathbf{L}_2$，$\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$ ，以及 $\mathcal{P}_{\lambda_\mathbf{B}}(\boldsymbol{\theta}_1)$ ， $\mathcal{P}_{\lambda_\mathbf{L}}(\boldsymbol{\theta}_2)$ ，和 $\mathcal{P}_{\lambda_\mathbf{L}}(\boldsymbol{\theta}_3)$ 分别是 (8)，(10)，和 (11) 中对 $\mathbf{B}$，$\mathbf{L}_1$，和 $\mathbf{L}_2$ 的惩罚。那么，以下命题表明，在弱假设下，算法 1 产生的序列收敛。

命题3.1 *MMTR*目标 是 $f(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) + \mathcal{P}_{\lambda_\mathbf{B}}(\boldsymbol{\theta}_1) + \mathcal{P}_{\lambda_\mathbf{L}}(\boldsymbol{\theta}_2) + \mathcal{P}_{\lambda_\mathbf{L}}(\boldsymbol{\theta}_3)$。令 $\mathcal{M}(\cdot)$ 是使用算法*1*将 $\boldsymbol{\theta}^{(t)}$ 映射到 $\boldsymbol{\theta}^{(t+1)}$ 的函数。那么，算法*1*的每次迭代都不会增加 $f(\boldsymbol{\theta})$。此外，假设参数空间 $\boldsymbol{\Theta}$ 是紧致的，并且 $f(\boldsymbol{\theta}) = f\{\mathcal{M}(\boldsymbol{\theta})\}$ 仅对 $f(\boldsymbol{\theta})$ 的驻点成立。那么，$\{\boldsymbol{\theta}^{(t)}\}_{t=1}^{\infty}$序列收敛到一个驻点。

4 实验

4.1 设置

MMTR性能使用模拟和真实数据进行了评估。高维惩罚竞争者基于准似然（Li等人，2022年）、迹回归（Zhao等人，2017年）、后验模式估计（Heiling等人，2023年）和广义估计方程（GEE）（Zhang等人，2019年）。在 `glmmPen` R包中实现的后验模式估计算法在预筛选步骤中失败，因此我们没有展示与该方法的比较。在低维设置中，我们使用 `lme4` (Bates等人，2013年)进行最大似然估计，从而产生了总共四种竞争方法。惩罚准似然（PQL）方法使用 `scalreg` R包实现，而 `lme4` 使用MMTR在(5)中的向量化形式进行参数估计。相比之下，迹回归和GEE没有随机效应，因此我们采用 `TensorReg` (Zhou，2017年)和 `SparseReg`(Zhou和Gaines，2017年)MATLAB包来拟合迹回归和GEE模型。MMTR使用算法1进行参数估计。

The marginal model in (7) was used for simulations. Specifically,

$$\mathbf{y} = \mathbf{X}\,\mathbf{b} + \mathbf{e}, \quad \mathbf{e} \sim N(\mathbf{0}, \mathbf{\Lambda}), \quad \mathbf{y} \in \mathbb{R}^N,\ \mathbf{X} \in \mathbb{R}^{N\times P},\ \mathbf{b} \in \mathbb{R}^P,\ \mathbf{e} \in \mathbb{R}^N, \tag{12}$$

where $\mathbf{\Lambda}$ was either a diagonal or block diagonal covariance matrix. We used MMTR and GEE as the true models for simulating the data. The GEE model specified $\mathbf{\Lambda}$ directly, whereas MMTR specified $\mathbf{\Lambda}$ through random effects design matrix, $\mathbf{\Sigma}_1, \mathbf{\Sigma}_2$, and τ^2.

Except for `lme4`, the performance of MMTR and its competitors depended on model-specific choices. The PQL method replaced $\mathbf{\Lambda}$ with a proxy matrix $c^2\,\mathbf{I}$ and used a lasso penalty on $\mathbf{b}$. Sparse trace regression set $\mathbf{\Lambda} = \tau^2\,\mathbf{I}$ and $\mathbf{B} = \text{vec}^{-1}(\mathbf{b})$ to be a low-rank and sparse matrix. We used two sparse GEE models: one with an unstructured covariance and the other with an equicorrelation covariance structure. Both these models employed a lasso penalty on $\mathbf{b}$. The unstructured covariance model assumed that $\mathbf{e}_i$ was distributed as $N(\mathbf{0}, \tilde{\mathbf{\Lambda}})$ for every i, implying that $m_i = m$ and $\mathbf{\Lambda} = \mathbf{I}_n \otimes \tilde{\mathbf{\Lambda}}$. Due to this constraint, we used this model only in simulations. The sparse GEE equicorrelation model assumed that $\mathbf{\Lambda}$ was a correlation matrix with off-diagonal elements equal to $\alpha \in (0, 1)$. Unlike the previous GEE model, the m_is in this model were allowed to be unequal, enabling its use for simulated and real data analyses. The PQL, trace regression, and GEE tuning parameters were selected using the recommended approach or cross-validation. MMTR had two tuning parameters: $\lambda_{\mathbf{B}}$ in (8), which controlled the sparsity of $\mathbf{B}$, and $\lambda_{\mathbf{L}}$ in (10) and (11), which determined the ranks of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$. We varied $\log \lambda_{\mathbf{B}}$ and $\log \lambda_{\mathbf{L}}$ on a grid and chose the best $(\lambda_{\mathbf{B}}, \lambda_{\mathbf{L}})$ pair with the minimum extended Bayesian information criterion (Chen and Chen, 2008).

The empirical performance in estimation and prediction was quantified using relative estimation error and mean squared prediction error. Let $\hat{\boldsymbol{\xi}}$ be the estimate of the parameter $\boldsymbol{\xi}$, where $\boldsymbol{\xi} \in \{\mathbf{B}, \mathbf{\Lambda}, \mathbf{\Sigma}_1\,.\,\mathbf{\Sigma}_2\}$, and $\hat{\mathbf{y}}_i$ be the predicted value of $\mathbf{y}_i$ in the test data. Then,

公式(7)中的边缘模型用于模拟。具体来说，

$$\mathbf{y} = \mathbf{X}\,\mathbf{b} + \mathbf{e}, \quad \mathbf{e} \sim N(\mathbf{0}, \mathbf{\Lambda}), \quad \mathbf{y} \in \mathbb{R}^N,\ \mathbf{X} \in \mathbb{R}^{N\times P},\ \mathbf{b} \in \mathbb{R}^P,\ \mathbf{e} \in \mathbb{R}^N, \tag{12}$$

其中 $\mathbf{\Lambda}$ 要么是对角协方差矩阵或块对角协方差矩阵。我们使用 MMTR 和 GEE 作为模拟数据的真实模型。GEE 模型直接指定了 $\mathbf{\Lambda}$，而 MMTR 通过随机效应设计矩阵 $\mathbf{\Sigma}_1$, $\mathbf{\Sigma}_2$ 和 τ^2 指定了 $\mathbf{\Lambda}$。

除了 `lme4` 之外，MMTR 及其竞争对手的性能取决于模型特定的选择。PQL 方法用代理矩阵 $c^2\,\mathbf{I}$ 替换了 $\mathbf{\Lambda}$ 并在 $\mathbf{b}$ 上使用了 Lasso 惩罚。稀疏迹回归将 $\mathbf{\Lambda} = \tau^2\,\mathbf{I}$ 和 $\mathbf{B} = \text{vec}^{-1}(\mathbf{b})$ 设置为低秩稀疏矩阵。我们使用了两个稀疏 GEE 模型：一个具有非结构化协方差，另一个具有等相关性协方差结构。这两个模型都在 $\mathbf{b}$ 上使用了 Lasso 惩罚。非结构化协方差模型假设$\mathbf{e}_i$ 对每个 i 都以 $N(\mathbf{0}, \tilde{\mathbf{\Lambda}})$ 分布，这意味着 $m_i = m$ 和 $\mathbf{\Lambda} = \mathbf{I}_n \otimes \tilde{\mathbf{\Lambda}}$。由于这个约束，我们只在模拟中使用该模型。稀疏 GEE 等相关性模型假设 $\mathbf{\Lambda}$ 是一个相关矩阵，其非对角元素等于 $\alpha \in (0, 1)$。与之前的 GEE 模型不同，该模型中的 m_is 允许不相等，因此可用于模拟和真实数据分析。PQL、迹回归和 GEE 调整参数是使用推荐方法或交叉验证选择的。MMTR 有两个调整参数：公式(8)中的 $\lambda_{\mathbf{B}}$，它控制了 $\mathbf{B}$ 的稀疏性，以及公式(10)和(11)中的 $\lambda_{\mathbf{L}}$，它确定了 $\mathbf{\Sigma}_1$ 和 $\mathbf{\Sigma}_2$ 的秩。我们对 $\log\ \lambda_{\mathbf{B}}$ 和 $\log \lambda_{\mathbf{L}}$ 进行网格搜索，并选择了具有最小扩展贝叶斯信息准则 (Chen and Chen, 2008) 的最佳 $(\lambda_{\mathbf{B}}, \lambda_{\mathbf{L}})$ 对。

经验性能在估计和预测中通过相对估计误差和均方预测误差进行量化。设$\hat{}\boldsymbol{\xi}$ 为参数$\boldsymbol{\xi}$的估计值，其中 $\boldsymbol{\xi} \in \{\mathbf{B}, \mathbf{\Lambda}, \mathbf{\Sigma}_1\,.\,\mathbf{\Sigma}_2\}$和 $\hat{\mathbf{y}}_i$ 是测试数据中 $\mathbf{y}_i$ 的预测值。然后，

the relative estimation error and mean square prediction errors are

$$\text{err}_{\boldsymbol{\xi}} = \|\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}\|_F / \|\boldsymbol{\xi}\|_F, \quad \boldsymbol{\xi} \in \{\mathbf{B}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2\}, \quad \text{MSPE} = \sum_{i=1}^{n^*} \|\mathbf{y}_i - \hat{\mathbf{y}}_i\|_2^2 / n^*, \tag{13}$$

where n^* is test data sample size and $\|\cdot\|_F$ and $\|\cdot\|_2$ are the Frobenius and Euclidean norms.

4.2 Simulation Using the MMTR Model

We simulated the data from MMTR in (3), resulting in a misspecified model for the MMTR's competitors. Two choices of parameter dimensions were used: $\mathbf{B} \in \mathbb{R}^{5\times5}$, $\mathbf{L}_1 \in \mathbb{R}^{5\times2}$, $\mathbf{L}_2 \in \mathbb{R}^{5\times2}$ (**Case 1**), and (2) $\mathbf{B} \in \mathbb{R}^{10\times10}$, $\mathbf{L}_1 \in \mathbb{R}^{10\times3}$, $\mathbf{L}_2 \in \mathbb{R}^{10\times3}$ (**Case 2**). For the two cases, $\lfloor 0.4 P_1 P_2 \rfloor$ entries in $\mathbf{B}$ were randomly set to zero, and the remaining $\mathbf{B}$ entries were sampled with replacement from a uniform grid in $[-10, 10]$ with a step size of 0.5. The entries in $\mathbf{L}_1$ and $\mathbf{L}_2$ were sampled from a continuous uniform distribution on $[-1, 1]$, implying that the row and column covariance matrices were $\boldsymbol{\Sigma}_1 = \mathbf{L}_1 \mathbf{L}_1^\top$ and $\boldsymbol{\Sigma}_2 = \mathbf{L}_2 \mathbf{L}_2^\top$. The covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ were identified by scaling their nonzero singular values such that their product equaled one. The τ^2 parameter was fixed at 0.5 for all simulations.

We varied m_i, n, and N to evaluate their impact on MMTR's performance. For Cases 1 and 2, we simulated data with $n \in \{18, 27, 54, 81\}$ and $n \in \{64, 96, 192, 288\}$, respectively, ensuring that the ratio of n to the parameter dimension was 0.4, 0.6, 1.2, and 1.8. For every case and n combination, the number of observations per group was set to $m_i \in \{2, 6\}$, resulting in 16 different simulation scenarios. The `lme4` package required $N \geq nQ_1Q_2$ for estimating covariance parameters, so it was excluded as a competitor. We performed 100 replications for each scenario and evaluated MMTR's performance using three competitors.

We estimated $\mathbf{B}$ and $\boldsymbol{\Lambda}$ in (12) by selecting tuning parameters via ten-fold cross-validation. MMTR's tuning parameters, $\lambda_{\mathbf{B}}$ and $\lambda_{\mathbf{L}}$, ranged from 10^{-4} and 0.04 and 10^{-4}

相对估计误差和均方预测误差是

$$\text{err}_{\boldsymbol{\xi}} = \|\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}\|_F / \|\boldsymbol{\xi}\|_F, \quad \boldsymbol{\xi} \in \{\mathbf{B}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2\}, \quad \text{MSPE} = \sum_{i=1}^{n^*} \|\mathbf{y}_i - \hat{\mathbf{y}}_i\|_2^2 / n^*, \tag{13}$$

其中 n^* 是测试数据样本大小， $\|\cdot\|_F$ 和 $\|\cdot\|_2$ 是 Frobenius 范数和欧几里得范数。

4.2 使用 MMTR 模型进行仿真

我们对 (3) 中的 MMTR 数据进行了仿真，导致 MMTR 的竞争对手模型未指定。使用了两种参数维度选择： $\mathbf{B} \in \mathbb{R}^{5\times5}$,$\mathbf{L}_1 \in \mathbb{R}^{5\times2}$, $\mathbf{L}_2 \in \mathbb{R}^{5\times2}$ (**案例1**)，以及 (2) $\mathbf{B} \in \mathbb{R}^{10\times10}$, $\mathbf{L}_1 \in \mathbb{R}^{10\times3}$, $\mathbf{L}_2 \in \mathbb{R}^{10\times3}$ (**案例 2**)。对于这两种情况， $\lfloor 0.4 P_1 P_2 \rfloor$ 个 $\mathbf{B}$ 条目被随机设置为 0，其余 $\mathbf{B}$ 条目从 $[-10, 10]$ 中的均匀网格中以步长 0.5 进行有放回抽样。$\mathbf{L}_1$ 和 $\mathbf{L}_2$ 中的条目从 $[-1, 1]$ 上的连续均匀分布中抽样，这意味着行和列协方差矩阵是 $\boldsymbol{\Sigma}_1 = \mathbf{L}_1 \mathbf{L}_1^\top$ 和 $\boldsymbol{\Sigma}_2 = \mathbf{L}_2 \mathbf{L}_2^\top$。协方差矩阵 $\boldsymbol{\Sigma}_1$ 和 $\boldsymbol{\Sigma}_2$ 通过缩放其非零奇异值，使得它们的乘积等于 1 来识别。 τ^2 参数在所有仿真中固定为 0.5。

我们改变了 m_i、 n和 N，以评估它们对MMTR性能的影响。对于案例1和案例2，我们分别模拟了具有 $n \in \{18, 27, 54, 81\}$ 和 $n \in \{64, 96, 192, 288\}$的数据，确保 n 与参数维度的比率是0.4、0.6、1.2和1.8。对于每个案例和 n 组合，每个组的观察数量设置为 $m_i \in \{2, 6\}$，从而产生了16种不同的模拟场景。 `lme4` 包需要 $N \geq nQ_1Q_2$ 来估计协方差参数，因此它被排除为竞争对手。我们对每个场景进行了100次重复实验，并使用三个竞争对手评估了MMTR的性能。

我们通过十折交叉验证选择调整参数，估计了(12)中的$\mathbf{B}$ 和 $\boldsymbol{\Lambda}$ 。MMTR的调整参数 $\lambda_{\mathbf{B}}$ 和 $\lambda_{\mathbf{L}}$的范围从 10^{-4} 和0.04和 10^{-4}

and 1, respectively, on an equally spaced 10-by-10 log-scale grid. Sparse Kruskal regression estimated only $\mathbf{B}$ by varying its tuning parameter from 10^{-2} to 10 on an equally spaced log-scale grid. The quasi-likelihood method also required $N \geq nQ_1Q_2$ for estimating $\mathbf{\Lambda}$, so it estimated only $\mathbf{B}$. For the two GEE models and the quasi-likelihood method, the tuning parameters for estimating $\mathbf{B}$ varied evenly over 10 values from 10^{-4} and 10 on a log-scale grid. This resulted in $\mathbf{B}$ estimates for four methods and $\mathbf{\Lambda}$ estimates for only two.

We evaluated the performance of MMTR and its competitors using the metrics in (13) (Tables 1 and 2). For $\mathbf{B}$ estimation, all four methods showed similar estimation accuracy across all simulation scenarios, demonstrating that mean parameter estimation was robust to model misspecification. In contrast, accurate $\mathbf{\Lambda}$ estimation was more challenging and required m_i and n to be sufficiently large. Furthermore, MMTR outperformed both GEE models across all scenarios in $\mathbf{\Lambda}$ estimation, indicating that covariance estimation was sensitive to model misspecification. MMTR's relative accuracy increased with n and m_i, whereas both GEE models' performance remained unaffected by increasing n and m_i.

Unlike its competitors, MMTR estimated $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ using Proposition 1.2 (Tables 3 and 4). The results showed that m_i had a greater impact on estimation accuracy of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ than N. Specifically, for a fixed N, MMTR achieved higher accuracy when $m_i = 6$ than when $m_i = 2$; see the blue and red highlighted rows in Tables 3 and 4. In summary, while all the methods had similar accuracy in estimating $\mathbf{B}$, MMTR outperformed its competitors in accurately estimating $\mathbf{\Sigma}_1, \mathbf{\Sigma}_2$, and $\mathbf{\Lambda}$, particularly for large n and m_i.

和1，分别在等距的10×10对数刻度网格上。稀疏Kruskal回归仅通过改变其调优参数从 10^{-2} 到10在等距的对数刻度网格上估计了$\mathbf{B}$ 。准似然法也需要 $N \geq nQ_1Q_2$ 来估计 $\mathbf{\Lambda}$，因此它仅估计了$\mathbf{B}$。对于两个GEE模型和准似然法，估计$\mathbf{B}$ 的调优参数在对数刻度网格上均匀分布在10个值之间，从 10^{-4} 和10。这导致四个方法有$\mathbf{B}$ 估计，而只有两个方法有 $\mathbf{\Lambda}$ 估计。

我们使用(13)中的指标（表1和表2）评估了MMTR及其竞争对手的性能。对于$\mathbf{B}$估计，所有四种方法在所有模拟场景中都表现出相似的估计精度，表明均值参数估计对模型设定错误具有鲁棒性。相比之下，准确的 $\mathbf{\Lambda}$ 估计更具挑战性，需要 m_i 和 n 足够大。此外，MMTR在所有场景中的 $\mathbf{\Lambda}$ 估计都优于两个GEE模型，表明协方差估计对模型设定错误敏感。MMTR的相对精度随着 n 和 m_i的增加而提高，而两个GEE模型的表现不受增加 n 和 m_i的影响。

与竞争对手不同，MMTR 使用命题1.2（表3和表4）估计 $\mathbf{\Sigma}_1$ 和 $\mathbf{\Sigma}_2$ 。结果表明，m_i 对 $\mathbf{\Sigma}_1$ 和$\mathbf{\Sigma}_2$的估计精度的影响大于 N。具体来说，对于固定的 N，当 $m_i = 6$时，MMTR 的精度更高，而当 $m_i = 2;$ 时，请参见表3和表4中蓝色和红色突出显示的行。总之，虽然所有方法在估计 $\mathbf{B}$ 时的精度相似，但 MMTR 在准确估计 $\mathbf{\Sigma}_1, \mathbf{\Sigma}_2$和 $\mathbf{\Lambda}$方面优于竞争对手，特别是在大型 n 和 m_i的情况下。

Table 1: Relative estimation error in Case 1. The simulation follows MMTR in (3). Every estimation error is averaged over 100 simulation replications, with the Monte Carlo errors in parenthesis. A small relative error indicates high estimation accuracy. GEE un. and GEE eq. denote the GEE models with unstructured and equicorrelated covariance matrix. Kruskal and Quasi denote the sparse Kruskal and quasi-likelihood methods.

group size	$\text{err}_\mathbf{B}$					$\text{err}_\mathbf{A}$		
	MMTR	Kruskal	GEE un.	GEE eq.	Quasi	MMTR	GEE un.	GEE eq.
n = 18								
$m_i = 2$	0.130 (0.0440)	0.130 (0.0500)	0.170 (0.0650)	0.110 (0.0360)	0.620 (0.2000)	0.91 (0.052)	0.95 (0.094)	0.86 (0.052)
$m_i = 6$	0.044 (0.0100)	0.051 (0.0120)	0.082 (0.0340)	0.047 (0.0120)	0.110 (0.0310)	0.74 (0.330)	0.93 (0.027)	0.91 (0.037)
n = 27								
$m_i = 2$	0.099 (0.0290)	0.077 (0.0250)	0.087 (0.0430)	0.073 (0.0230)	0.310 (0.1500)	0.99 (0.250)	0.81 (0.160)	0.86 (0.055)
$m_i = 6$	0.028 (0.0076)	0.037 (0.0100)	0.043 (0.0120)	0.035 (0.0097)	0.083 (0.0220)	0.49 (0.150)	1.20 (0.470)	0.91 (0.036)
n = 54								
$m_i = 2$	0.051 (0.0110)	0.048 (0.0110)	0.046 (0.0120)	0.045 (0.0120)	0.110 (0.0280)	0.92 (0.290)	0.81 (0.071)	0.85 (0.056)
$m_i = 6$	0.016 (0.0040)	0.026 (0.0060)	0.026 (0.0066)	0.024 (0.0058)	0.052 (0.0110)	0.29 (0.090)	0.88 (0.046)	0.91 (0.037)
n = 81								
$m_i = 2$	0.034 (0.0066)	0.037 (0.0096)	0.035 (0.0092)	0.035 (0.0091)	0.082 (0.0210)	0.59 (0.220)	0.83 (0.062)	0.85 (0.054)
$m_i = 6$	0.013 (0.0034)	0.021 (0.0057)	0.021 (0.0056)	0.020 (0.0056)	0.042 (0.0094)	0.24 (0.074)	0.89 (0.043)	0.91 (0.036)

Table 2: Relative estimation error in Case 2. The simulation follows MMTR in (3). Every estimation error is averaged over 100 simulation replications, with the Monte Carlo errors in parenthesis. A small relative error indicates high estimation accuracy. GEE un. and GEE eq. denote the GEE models with unstructured and equicorrelated covariance matrix. Kruskal and Quasi denote the sparse Kruskal and quasi-likelihood methods.

group size	$\text{err}_\mathbf{B}$					$\text{err}_\mathbf{A}$		
	MMTR	Kruskal	GEE un.	GEE eq.	Quasi	MMTR	GEE un.	GEE eq.
n = 64								
$m_i = 2$	0.120 (0.0200)	0.55 (0.097)	0.120 (0.0230)	0.096 (0.0160)	0.890 (0.0440)	0.92 (0.019)	0.98 (0.0045)	0.90 (0.019)
$m_i = 6$	0.040 (0.0066)	0.36 (0.050)	0.039 (0.0091)	0.033 (0.0050)	0.150 (0.0350)	0.87 (0.240)	0.95 (0.0110)	0.93 (0.015)
n = 96								
$m_i = 2$	0.085 (0.0130)	0.42 (0.062)	0.071 (0.0240)	0.057 (0.0090)	0.740 (0.0980)	1.10 (0.140)	0.76 (0.0700)	0.90 (0.018)
$m_i = 6$	0.022 (0.0028)	0.34 (0.044)	0.028 (0.0081)	0.025 (0.0037)	0.084 (0.0130)	0.46 (0.098)	0.84 (0.1800)	0.93 (0.015)
n = 192								
$m_i = 2$	0.043 (0.0075)	0.36 (0.049)	0.033 (0.0057)	0.033 (0.0059)	0.150 (0.0280)	1.10 (0.200)	0.86 (0.0260)	0.90 (0.019)
$m_i = 6$	0.012 (0.0018)	0.33 (0.043)	0.018 (0.0026)	0.018 (0.0026)	0.049 (0.0063)	0.24 (0.041)	0.88 (0.0240)	0.93 (0.015)
n = 288								
$m_i = 2$	0.028 (0.0045)	0.34 (0.045)	0.025 (0.0036)	0.025 (0.0036)	0.086 (0.0130)	0.63 (0.130)	0.88 (0.0230)	0.90 (0.019)
$m_i = 6$	0.010 (0.0015)	0.32 (0.042)	0.014 (0.0021)	0.014 (0.0021)	0.038 (0.0049)	0.18 (0.032)	0.90 (0.0200)	0.93 (0.015)

表1：案例1中的相对估计误差。模拟遵循(3)中的MMTR。每个估计误差在100次模拟复制上取平均值，括号内为蒙特卡洛误差。小的相对误差表示高估计精度。GEE非结构化和GEE等方差表示具有非结构化和等方差协方差矩阵的GEE模型。Kruskal和Quasi表示稀疏Kruskal和准似然方法。

组规模	$\text{err}_\mathbf{B}$					$\text{err}_\mathbf{A}$		
	MMTR	Kruskal	GEE非结构化	GEE等方差	Quasi	MMTR	GEE非结构化	GEE eq.
n = 18								
$m_i = 2$	0.130 (0.0440)	0.130 (0.0500)	0.170 (0.0650)	0.110 (0.0360)	0.620 (0.2000)	0.91 (0.052)	0.95 (0.094)	0.86 (0.052)
$m_i = 6$	0.044 (0.0100)	0.051 (0.0120)	0.082 (0.0340)	0.047 (0.0120)	0.110 (0.0310)	0.74 (0.330)	0.93 (0.027)	0.91 (0.037)
n = 27								
$m_i = 2$	0.099 (0.0290)	0.077 (0.0250)	0.087 (0.0430)	0.073 (0.0230)	0.310 (0.1500)	0.99 (0.250)	0.81 (0.160)	0.86 (0.055)
$m_i = 6$	0.028 (0.0076)	0.037 (0.0100)	0.043 (0.0120)	0.035 (0.0097)	0.083 (0.0220)	0.49 (0.150)	1.20 (0.470)	0.91 (0.036)
n = 54								
$m_i = 2$	0.051 (0.0110)	0.048 (0.0110)	0.046 (0.0120)	0.045 (0.0120)	0.110 (0.0280)	0.92 (0.290)	0.81 (0.071)	0.85 (0.056)
$m_i = 6$	0.016 (0.0040)	0.026 (0.0060)	0.026 (0.0066)	0.024 (0.0058)	0.052 (0.0110)	0.29 (0.090)	0.88 (0.046)	0.91 (0.037)
n = 81								
$m_i = 2$	0.034 (0.0066)	0.037 (0.0096)	0.035 (0.0092)	0.035 (0.0091)	0.082 (0.0210)	0.59 (0.220)	0.83 (0.062)	0.85 (0.054)
$m_i = 6$	0.013 (0.0034)	0.021 (0.0057)	0.021 (0.0056)	0.020 (0.0056)	0.042 (0.0094)	0.24 (0.074)	0.89 (0.043)	0.91 (0.036)

表2：案例2中的相对估计误差。模拟遵循(3)中的MMTR。每个估计误差在100次模拟复制上取平均值，括号中的为蒙特卡洛误差。小的相对误差表示高估计精度。GEE非结构化和GEE等方差表示具有非结构化和等方差协方差矩阵的GEE模型。Kruskal和Quasi表示稀疏Kruskal和准似然方法。

组规模	$\text{err}_\mathbf{B}$					$\text{err}_\mathbf{A}$		
	MMTR	Kruskal	GEE非结构化	GEE等方差	Quasi	MMTR	GEE非结构化	GEE eq.
n = 64								
$m_i = 2$	0.120 (0.0200)	0.55 (0.097)	0.120 (0.0230)	0.096 (0.0160)	0.890 (0.0440)	0.92 (0.019)	0.98 (0.0045)	0.90 (0.019)
$m_i = 6$	0.040 (0.0066)	0.36 (0.050)	0.039 (0.0091)	0.033 (0.0050)	0.150 (0.0350)	0.87 (0.240)	0.95 (0.0110)	0.93 (0.015)
n = 96								
$m_i = 2$	0.085 (0.0130)	0.42 (0.062)	0.071 (0.0240)	0.057 (0.0090)	0.740 (0.0980)	1.10 (0.140)	0.76 (0.0700)	0.90 (0.018)
$m_i = 6$	0.022 (0.0028)	0.34 (0.044)	0.028 (0.0081)	0.025 (0.0037)	0.084 (0.0130)	0.46 (0.098)	0.84 (0.1800)	0.93 (0.015)
n = 192								
$m_i = 2$	0.043 (0.0075)	0.36 (0.049)	0.033 (0.0057)	0.033 (0.0059)	0.150 (0.0280)	1.10 (0.200)	0.86 (0.0260)	0.90 (0.019)
$m_i = 6$	0.012 (0.0018)	0.33 (0.043)	0.018 (0.0026)	0.018 (0.0026)	0.049 (0.0063)	0.24 (0.041)	0.88 (0.0240)	0.93 (0.015)
n = 288								
$m_i = 2$	0.028 (0.0045)	0.34 (0.045)	0.025 (0.0036)	0.025 (0.0036)	0.086 (0.0130)	0.63 (0.130)	0.88 (0.0230)	0.90 (0.019)
$m_i = 6$	0.010 (0.0015)	0.32 (0.042)	0.014 (0.0021)	0.014 (0.0021)	0.038 (0.0049)	0.18 (0.032)	0.90 (0.0200)	0.93 (0.015)

Table 3: MMTR's relative estimation error for $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ in Case 1. Every estimation error is averaged over 100 simulation replications, with the Monte Carlo errors in parenthesis. The red and blue lines represent simulations with equal N values.

group size	$\text{err}_{\mathbf{\Sigma}_1}$	$\text{err}_{\mathbf{\Sigma}_2}$
n = 18		
$m_i = 2$	1.10 (0.230)	1.10 (0.270)
$m_i = 6$	0.57 (0.260)	0.54 (0.250)
n = 27		
$m_i = 2$	1.00 (0.290)	1.10 (0.300)
$m_i = 6$	0.40 (0.170)	0.39 (0.150)
n = 54		
$m_i = 2$	0.84 (0.290)	0.83 (0.340)
$m_i = 6$	0.21 (0.100)	0.21 (0.085)
n = 81		
$m_i = 2$	0.64 (0.320)	0.63 (0.280)
$m_i = 6$	0.17 (0.069)	0.16 (0.066)

Table 4: MMTR's relative estimation error for $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ in Case 2. Every estimation error is averaged over 100 simulation replications, with the Monte Carlo errors in parenthesis. The red and blue lines represent simulations with equal N values.

group size	$\text{err}_{\mathbf{\Sigma}_1}$	$\text{err}_{\mathbf{\Sigma}_2}$
n = 64		
$m_i = 2$	1.20 (0.200)	1.20 (0.170)
$m_i = 6$	0.80 (0.230)	0.80 (0.210)
n = 96		
$m_i = 2$	1.20 (0.190)	1.20 (0.190)
$m_i = 6$	0.47 (0.160)	0.49 (0.160)
n = 192		
$m_i = 2$	1.10 (0.200)	1.10 (0.210)
$m_i = 6$	0.24 (0.056)	0.23 (0.048)
n = 288		
$m_i = 2$	0.86 (0.280)	0.87 (0.260)
$m_i = 6$	0.17 (0.041)	0.18 (0.048)

4.3 Simulations Using the Equicorrelation GEE model

We simulated the data using the equicorrelation GEE model based on (12) to evaluate the impact of misspecification on MMTR's performance. The mean parameter in (12) was generated as in the previous simulation, resulting in two cases: $\mathbf{B} \in \mathbb{R}^{5\times5}$ (**Case 1**) and $\mathbf{B} \in \mathbb{R}^{10\times10}$(**Case 2**). The n choices for these two cases were the same as in the previous simulation to facilitate comparisons. Based on the previous results, we set $m_i = 6$ in both cases because MMTR demonstrated satisfactory performance only with sufficiently large m_i. This setup resulted in eight different simulation scenarios. The covariance parameter $\mathbf{\Lambda}$ in this simulation was specified by one correlation parameter α. For each of the 100 simulation replications across the eight scenarios, α was drawn from a continuous uniform distribution on $[0.2, 0.8]$. The three MMTR competitors (GEE, Kruskal, and quasi-likelihood), tuning parameter selection methods, and error metrics were identical to those used in the previous simulation.

All methods demonstrated similar accuracy in $\mathbf{B}$ estimation, but the equicorrelation GEE model outperformed the others in $\mathbf{\Lambda}$ estimation (Tables 5 and 6). For both cases, $\mathbf{B}$ estimation accuracy improved as n increased, providing further evidence that the mean

表3：案例1中MMTR的相对估计误差对于 $\mathbf{\Sigma}_1$ 和 $\mathbf{\Sigma}_2$ 。每个估计误差在100次模拟重复实验中取平均值，括号内为蒙特卡洛误差。红色和蓝色线条表示具有相等 N值的模拟。

组规模	$\text{err}_{\mathbf{\Sigma}_1}$	$\text{err}_{\mathbf{\Sigma}_2}$
n = 18		
$m_i = 2$	1.10 (0.230)	1.10 (0.270)
$m_i = 6$	0.57 (0.260)	0.54 (0.250)
n = 27		
$m_i = 2$	1.00 (0.290)	1.10 (0.300)
$m_i = 6$	0.40 (0.170)	0.39 (0.150)
n = 54		
$m_i = 2$	0.84 (0.290)	0.83 (0.340)
$m_i = 6$	0.21 (0.100)	0.21 (0.085)
n = 81		
$m_i = 2$	0.64 (0.320)	0.63 (0.280)
$m_i = 6$	0.17 (0.069)	0.16 (0.066)

表4：案例2中MMTR的相对估计误差对于 $\mathbf{\Sigma}_1$ 和 $\mathbf{\Sigma}_2$ 。每个估计误差在100次模拟重复实验中取平均值，括号内为蒙特卡洛误差。红色和蓝色线表示具有相等 N值的模拟。

组规模	$\text{err}_{\mathbf{\Sigma}_1}$	$\text{err}_{\mathbf{\Sigma}_2}$
n = 64		
$m_i = 2$	1.20 (0.200)	1.20 (0.170)
$m_i = 6$	0.80 (0.230)	0.80 (0.210)
n = 96		
$m_i = 2$	1.20 (0.190)	1.20 (0.190)
$m_i = 6$	0.47 (0.160)	0.49 (0.160)
n = 192		
$m_i = 2$	1.10 (0.200)	1.10 (0.210)
$m_i = 6$	0.24 (0.056)	0.23 (0.048)
n = 288		
$m_i = 2$	0.86 (0.280)	0.87 (0.260)
$m_i = 6$	0.17 (0.041)	0.18 (0.048)

4.3 使用等相关性GEE模型进行模拟

我们使用基于(12)的等相关性GEE模型模拟数据，以评估模型设定错误对MMTR性能的影响。公式(12)中的均值参数与先前模拟中生成的方式相同，导致两种情况：$\mathbf{B} \in \mathbb{R}^{5\times5}$ (**案例1**)和 $\mathbf{B} \in \mathbb{R}^{10\times10}$(**Case 2**)。这两种情况的选择与先前模拟相同，以便进行比较。根据先前结果，我们在两种情况下都设置了$m_i = 6$，因为MMTR只有在 m_i足够大时才表现出令人满意的性能。这种设置产生了八个不同的模拟场景。此模拟中的协方差参数 $\mathbf{\Lambda}$ 由一个相关参数 α指定。对于八个场景中的每个100次模拟复制， α 从 $[0.2, 0.8]$上的连续均匀分布中抽取。三个MMTR竞争对手（GEE、Kruskal和准似然）、调优参数选择方法和误差指标与先前模拟中使用的相同。

所有方法在 $\mathbf{B}$ 估计方面表现出相似的精度，但等相关性GEE模型在 $\mathbf{\Lambda}$ 估计方面表现优于其他模型（表5和表6）。对于两种情况，$\mathbf{B}$ 估计精度随着 n 的增加而提高，这进一步证明了均值

parameter estimation was robust to model misspecification. Although MMTR's covariance parameter was misspecified relative to the simulation model, its accuracy in estimating $\mathbf{\Lambda}$ improved up to a point with increasing n. Furthermore, MMTR's covariance estimation errors were significantly smaller for lower α values than higher values. A larger α corresponds to a heavily misspecified covariance model, making it more challenging to approximate using a Kronecker product structure; see supplementary materials for additional simulation results. In summary, while model misspecification implied that a larger n and smaller α were required for MMTR's accurate covariance estimation, MMTR showed excellent performance in estimating the mean parameter across all n choices.

Table 5: Relative estimation error in Case 1 with a misspecified equicorrelated covariance structure in (12). Every estimation error is averaged over 100 simulation replications, with the Monte Carlo errors in parenthesis. A small relative error indicates high estimation accuracy. GEE un. and GEE eq. denote the GEE models with unstructured and equicorrelated covariance matrix. Kruskal and Quasi denote the sparse Kruskal and quasi-likelihood methods.

group size	$\text{err}_\mathbf{B}$					$\text{err}_\mathbf{\Lambda}$		
	MMTR	Kruskal	GEE un.	GEE eq.	Quasi	MMTR	GEE un.	GEE eq.
$n = 18, m_i = 6$	0.038 (0.0110)	0.025 (0.0071)	0.0480 (0.0270)	0.0200 (0.0053)	0.062 (0.0160)	1.90 (0.75)	1.20 (2.200)	0.150 (0.110)
$n = 27, m_i = 6$	0.022 (0.0065)	0.020 (0.0046)	0.0220 (0.0061)	0.0160 (0.0038)	0.046 (0.0100)	1.10 (0.39)	5.20 (3.000)	0.110 (0.074)
$n = 54, m_i = 6$	0.013 (0.0031)	0.013 (0.0031)	0.0120 (0.0030)	0.0110 (0.0026)	0.029 (0.0057)	0.76 (0.12)	0.68 (0.130)	0.088 (0.066)
$n = 81, m_i = 6$	0.011 (0.0029)	0.011 (0.0024)	0.0095 (0.0022)	0.0088 (0.0020)	0.023 (0.0048)	0.73 (0.11)	0.40 (0.085)	0.071 (0.056)

Table 6: Relative estimation error in Case 2 with a misspecified equicorrelated covariance structure in (12). Every estimation error is averaged over 100 simulation replications, with the Monte Carlo errors in parenthesis. A small relative error indicates high estimation accuracy. GEE un. and GEE eq. denote the GEE models with unstructured and equicorrelated covariance matrix. Kruskal and Quasi denote the sparse Kruskal and quasi-likelihood methods.

group size	$\text{err}_\mathbf{B}$					$\text{err}_\mathbf{\Lambda}$		
	MMTR	Kruskal	GEE un.	GEE eq.	Quasi	MMTR	GEE un.	GEE eq.
$n = 64, m_i = 6$	0.0180 (0.0023)	0.36 (0.042)	0.0140 (0.00490)	0.0100 (0.00130)	0.060 (0.0170)	1.20 (0.30)	0.42 (0.090)	0.130 (0.077)
$n = 96, m_i = 6$	0.0110 (0.0015)	0.34 (0.039)	0.0094 (0.00130)	0.0081 (0.00089)	0.033 (0.0039)	0.90 (0.13)	2.50 (2.800)	0.098 (0.054)
$n = 192, m_i = 6$	0.0079 (0.0016)	0.33 (0.037)	0.0059 (0.00070)	0.0057 (0.00066)	0.019 (0.0016)	0.72 (0.12)	0.77 (0.082)	0.054 (0.033)
$n = 288, m_i = 6$	0.0066 (0.0015)	0.32 (0.036)	0.0047 (0.00053)	0.0046 (0.00046)	0.014 (0.0013)	0.70 (0.13)	0.42 (0.045)	0.043 (0.031)

4.4 Real Data: Labeled Faces in the Wild

We used the Labeled Faces in the Wild (LFW) database for our real data analysis, a benchmark dataset widely used to evaluate the empirical performance of array-variate

参数估计对模型设定错误具有鲁棒性。尽管MMTR的协方差参数相对于模拟模型存在未指定问题，但随着 n的增加，其在估计 $\mathbf{\Lambda}$方面的精度提升到了一定程度。此外，MMTR的协方差估计误差在较低的 α 值时显著小于较高值。较大的 α 对应于严重未指定的协方差模型，使用Kronecker积结构进行近似更加困难；有关更多模拟结果，请参见补充材料。总之，虽然模型设定错误意味着MMTR的准确协方差估计需要更大的 n 和更小的 α，但MMTR在所有 n 选择中均表现出优异的均值参数估计性能。

表5：案例1中（12）中错误指定的等协方差结构的相对估计误差。每个估计误差在100次模拟复制上取平均值，括号内为蒙特卡洛误差。较小的相对误差表示较高的估计精度。GEE un.和GEE eq.分别表示具有非结构化和等协方差矩阵的GEE模型。Kruskal和Quasi分别表示稀疏Kruskal和准似然方法。

组规模	$\text{err}_\mathbf{B}$					$\text{err}_\mathbf{\Lambda}$		
	MMTR	Kruskal	GEE非结构化	GEE eq.	Quasi	MMTR	GEE非结构化	GEE eq.
$n = 18, m_i = 6$	0.038 (0.0110)	0.025 (0.0071)	0.0480 (0.0270)	0.0200 (0.0053)	0.062 (0.0160)	1.90 (0.75)	1.20 (2.200)	0.150 (0.110)
$n = 27, m_i = 6$	0.022 (0.0065)	0.020 (0.0046)	0.0220 (0.0061)	0.0160 (0.0038)	0.046 (0.0100)	1.10 (0.39)	5.20 (3.000)	0.110 (0.074)
$n = 54, m_i = 6$	0.013 (0.0031)	0.013 (0.0031)	0.0120 (0.0030)	0.0110 (0.0026)	0.029 (0.0057)	0.76 (0.12)	0.68 (0.130)	0.088 (0.066)
$n = 81, m_i = 6$	0.011 (0.0029)	0.011 (0.0024)	0.0095 (0.0022)	0.0088 (0.0020)	0.023 (0.0048)	0.73 (0.11)	0.40 (0.085)	0.071 (0.056)

表6：案例2中错误指定的等协方差结构(12)的相对估计误差。每个估计误差在100次模拟复制上取平均值，括号内为蒙特卡洛误差。小的相对误差表示高估计精度。GEE非结构化和GEE等协方差矩阵分别表示非结构化和等协方差矩阵的GEE模型。Kruskal和Quasi分别表示稀疏Kruskal和准似然方法。

组规模	$\text{err}_\mathbf{B}$					$\text{err}_\mathbf{\Lambda}$		
	MMTR	Kruskal	GEE非结构化	GEE等协方差结构	Quasi	MMTR	GEE非结构化	GEE eq.
$n = 64, m_i = 6$	0.0180 (0.0023)	0.36 (0.042)	0.0140 (0.00490)	0.0100 (0.00130)	0.060 (0.0170)	1.20 (0.30)	0.42 (0.090)	0.130 (0.077)
$n = 96, m_i = 6$	0.0110 (0.0015)	0.34 (0.039)	0.0094 (0.00130)	0.0081 (0.00089)	0.033 (0.0039)	0.90 (0.13)	2.50 (2.800)	0.098 (0.054)
$n = 192, m_i = 6$	0.0079 (0.0016)	0.33 (0.037)	0.0059 (0.00070)	0.0057 (0.00066)	0.019 (0.0016)	0.72 (0.12)	0.77 (0.082)	0.054 (0.033)
$n = 288, m_i = 6$	0.0066 (0.0015)	0.32 (0.036)	0.0047 (0.00053)	0.0046 (0.00046)	0.014 (0.0013)	0.70 (0.13)	0.42 (0.045)	0.043 (0.031)

4.4 真实数据：标记人脸在野外

我们使用了标记人脸在野外（LFW）数据库进行真实数据分析，这是一个广泛用于评估数组变量经验性能的基准数据集。

models (Huang et al., 2007; Lock, 2018). The LFW data contains multiple images of famous individuals and 73 real-valued image attributes for every image (Kumar et al., 2009). For this analysis, we focused on a subset of 5749 individuals whose images were aligned using deep funneling (Huang et al., 2012). To ensure sufficiently large m_i values for MMTR, we selected individuals with 4 to 50 images, resulting in a final dataset of 594 individuals and a total of 5124 images.

The data were pre-processed to define the covariates. Each image was converted to grayscale and resized to 32×32 dimensions using *grayscale* and *resize* functions in the `imager` library (Barthelme, 2024). This process generated 32×32 fixed and random effects covariate matrices for all the 5124 images. The response variable was defined as the first principal component of the 73 image attributes for each image.

We randomly split the images into ten training and testing sets. Each test set consisted of four randomly selected images from each of the 115 individuals with at least twelve images, resulting in ten training sets of 4664 images and ten testing sets of 460 images. We fit MMTR and its four competitors to each training set. Tuning parameter selection methods remain unchanged from the simulations. Unlike the simulations, `lme4` and quasi-likelihood methods included random intercepts to account for dependencies within images of the same individual. The GEE model with an unstructured covariance parameter was inapplicable because the m_i values varied across the selected individuals. Their performance was evaluated on the test sets using MSPE in (13) and prediction R^2.

MMTR achieved the lowest MSPE and highest prediction R^2 among all the methods (Table 7). The averaged estimate of $\mathbf{B}$ across the ten training sets revealed key facial features in the images (Figure 1). Based on the simulation results, MSPE, and prediction R^2, we concluded that MMTR outperforms its competitors in accurately predicting resized LFW images.

模型（黄等人，2007；洛克，2018）。LFW数据包含多位著名人士的多张图像，每张图像具有73个实值图像属性（库马尔等人，2009）。在此分析中，我们关注了5749个个体子集，其图像使用深度漏斗对齐（黄等人，2012）。为确保MMTR具有足够大的 m_i 值，我们选择了拥有4到50张图像的个体，最终得到594个个体和5124张图像的数据集。

数据进行了预处理以定义协变量。每张图像被转换为灰度并使用<code>grayscale</code>和<code>resize</code>函数在`imager`库（巴塞梅，2024）中调整大小到 32×32 维度。该过程为所有5124张图像生成了固定和随机效应协变量矩阵。响应变量被定义为每张图像73个图像属性的第一主成分。

我们将图像随机分成十个训练集和测试集。每个测试集由从至少有十二张图像的115个人中随机选择的四张图像组成，最终得到四个训练集，每个训练集包含4664张图像，十个测试集，每个测试集包含460张图像。我们将MMTR及其四个竞争对手拟合到每个训练集上。调优参数选择方法与模拟保持不变。与模拟不同，`lme4`和准似然方法包括了随机截距，以解释同一个人的图像之间的依赖关系。具有非结构化协方差参数的GEE模型不适用，因为 m_i 值在选定的个人之间有所不同。它们在测试集上的性能使用公式(13)中的MSPE和预测 R^2进行评估。

MMTR在所有方法中实现了最低的MSPE和最高的预测 R^2 （表7）。在十个训练集上对$\mathbf{B}$ 的均值估计揭示了图像中的关键面部特征（图1）。根据模拟结果、MSPE和预测R^2，我们得出结论，MMTR在其竞争对手中在准确预测调整大小的LFW图像方面表现最佳。

Table 7: MSPE and prediction R^2. The MSPE and prediction R^2 values are averaged across ten testing sets of the LFW data, with the Monte Carlo errors in parenthesis. GEE equicorr. Kruskal, and Quasi denote the GEE with equicorrelated covariance matrix, sparse Kruskal, and quasi-likelihood methods.

	MMTR	Quasi	LME4	GEE equicorr.	Kruskal
MSPE	2.19 (0.13)	2.32 (0.14)	2.81 (0.20)	3.18 (0.24)	5.47 (0.42)
R^2	0.84 (0.012)	0.83 (0.013)	0.79 (0.009)	0.76 (0.012)	0.60 (0.018)

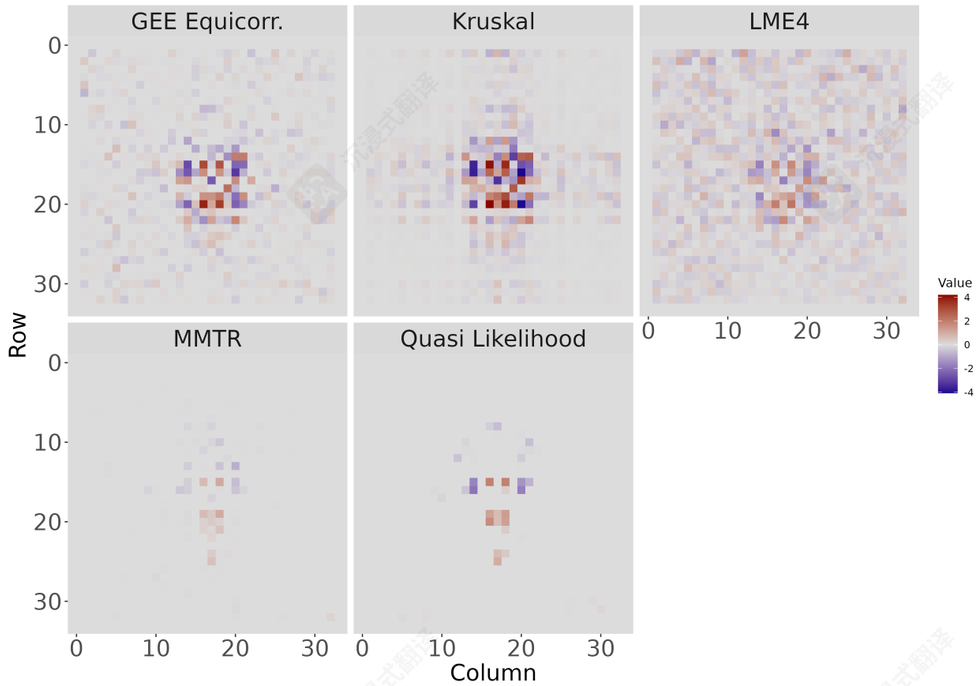

Figure 1: Mean parameter $\mathbf{B}$ estimated by MMTR, quasi-likelihood, lme4, GEE with equicorrelation covariance matrix, and sparse Kruskal methods. The $\mathbf{B}$ estimates are averaged across ten training sets of the LFW data.

5 Discussion

We have focused on a correlated scalar responses and matrix-variate covariates, but MMTR generalizes to array-variate covariates. If the random effect $\mathbf{A}_i$ in (1) is a $Q_1 \times \cdots \times Q_d$ array, then its separable covariance structure generalizes as $\mathbb{E}\{\text{vec}(\mathbf{A}_i)\} = \mathbf{0}$ and $\text{Cov}\{\text{vec}(\mathbf{A}_i)\} =$

表7：MSPE和预测 R^2。MSPE和预测 R^2 值在LFW数据的十个测试集上取平均值，括号内为蒙特卡洛误差。GEE等权重协方差矩阵、稀疏Kruskal和准似然方法分别表示GEE的等权重协方差矩阵、稀疏Kruskal和准似然方法。

	MMTR	Quasi	LME4	GEE equicorr.	Kruskal
MSPE	2.19 (0.13)	2.32 (0.14)	2.81 (0.20)	3.18 (0.24)	5.47 (0.42)
R^2	0.84 (0.012)	0.83 (0.013)	0.79 (0.009)	0.76 (0.012)	0.60 (0.018)

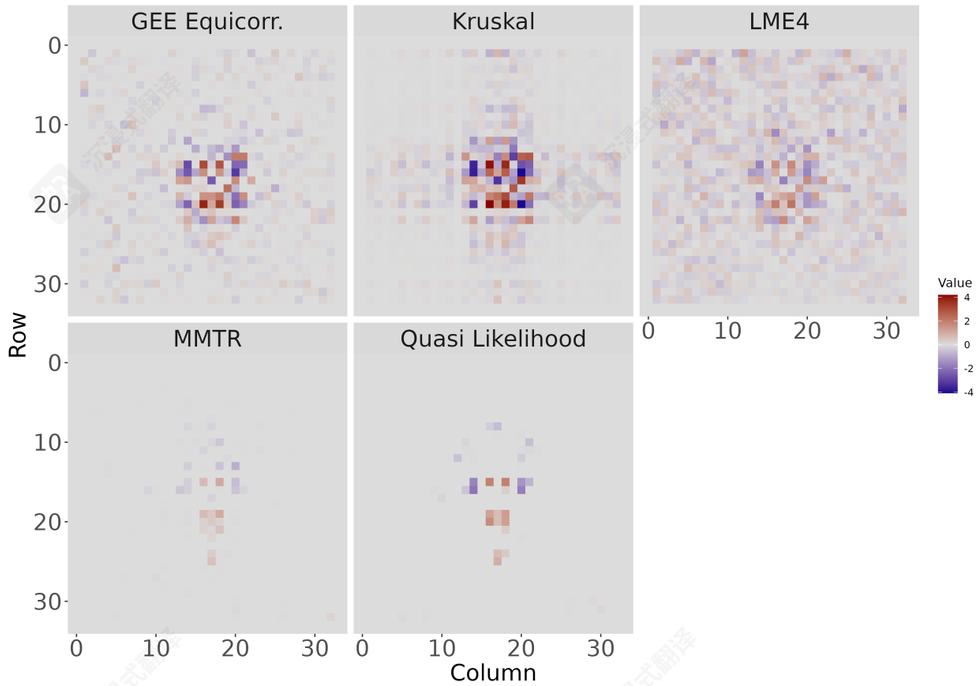

图1：由MMTR、准似然、lme4、具有等相关性协方差矩阵的GEE以及稀疏Kruskal方法估计的均值参数 $\mathbf{B}$ 。该 $\mathbf{B}$ 估计值是在LFW数据的十个训练集上平均得到的。

5 讨论

我们关注的是相关标量响应和矩阵变量协变量，但MMTR可以推广到数组变量协变量。如果(1)中的随机效应 $\mathbf{A}_i$ 是 $Q_1 \times \cdots \times Q_d$ 数组，则其可分离协方差结构推广为 $\mathbb{E}\{\text{vec}(\mathbf{A}_i)\} = \mathbf{0}$ ，并且$\text{Cov}\{\text{vec}(\mathbf{A}_i)\} =$

$\tau^2\,\mathbf{\Sigma}_d \otimes \cdots \otimes \mathbf{\Sigma}_1$, where $\mathbf{A}_i$ is an array-variate normal distribution with zero mean array and dimension j-specific covariance matrix $\mathbf{\Sigma}_j$ $(j = 1, \ldots, d)$ (Hoff, 2011). For $j = 1, \ldots, d$ and $S_j \ll Q_j$, we assume that there exists $\mathbf{L}_j \in \mathbb{R}^{Q_j \times S_j}$ such that $\mathbf{\Sigma}_j \approx \mathbf{L}_j\,\mathbf{L}_j^\top$, $\mathbf{C}_i$ in (3) is an $S_1 \times \cdots \times S_d$ array, and $\mathbf{L}_j$ is regularized by imposing group lasso penalty on its columns to enable sparse estimation. Similarly, Propositions 1.1, 1.2, and 1.3 naturally generalize to the array-variate MMTR extension. Finally, the array-variate extension of Algorithm 1 has $d + 1$ cycles, where the first cycle estimates $\mathbf{B}$ and τ^2 using regularized array-variate regression (Zhou et al., 2013) and the next d cycles estimate $\mathbf{L}_j$ given the remaining parameters using a version of (10) for $j = 1, \ldots, d$. We are developing an MMTR extension for correlated array-variate responses.

The objectives for estimating $\mathbf{L}_1$ and $\mathbf{L}_2$ in Algorithm 1 rely on the assumption that the random effects and error terms are Gaussian, but our parameter estimation algorithm generalizes to the case where we only assume the first and second moments of the random effects exist. In this case, Algorithm 1 iteratively replaces the first two moments of the random effects, $\mathbb{E}(\mathbf{A}_i \mid \mathbf{y}_i, \boldsymbol{\theta}^{(t)})$ and $\mathbb{E}\{\mathrm{vec}(\mathbf{A}_i)\mathrm{vec}(\mathbf{A}_i)^\top \mid \mathbf{y}_i, \boldsymbol{\theta}^{(t)}\}$ $(i = 1, \ldots, n)$, by those based on the assumption that $\mathbf{A}_i$ follows a matrix normal distribution, resulting in the objectives (10) and (11).

Acknowledgments

Ian Hultman and Sanvesh Srivastava were partially supported by grants from the National Institutes of Health (1DP2MH126377-01) and the National Science Foundation (DMS-1854667). The authors thank Kshitij Khare, Joe Lang, Boxiang Wang, and Dale Zimmerman for their valuable feedback on an earlier version of this manuscript. The code used in the experiments is available at `https://github.com/IHultman/MMTR`.

$\tau^2\,\mathbf{\Sigma}_d \otimes \cdots \otimes \mathbf{\Sigma}_1$, 其中 $\mathbf{A}_i$ 是一个具有零均值数组和特定维度 j 协方差矩阵 $\mathbf{\Sigma}_j$ $(j = 1, \ldots, d)$ 的数组变量正态分布（Hoff, 2013）。对于 $j = 1 \ldots d$,,和 $S_j \ll Q_j$, 我们假设存在 $\mathbf{L}_j \in \mathbb{R}^{Q_j \times S_j}$ 使得 $\mathbf{\Sigma}_j \approx \mathbf{L}_j\mathbf{L}_j^\top$, $\mathbf{C}_i$ 在 (3) 中是一个 $S_1 \times \cdots \times S_d$ 数组, 并且 $\mathbf{L}_j$ 通过对其列施加组Lasso惩罚进行正则化，以实现稀疏估计。类似地，命题1.1、1.2和1.3 自然地推广到数组变量MMTR扩展。最后，算法1的数组变量扩展有 $d + 1$ 个周期，其中第一个周期使用正则化数组变量回归（Zhou等人，2013）估计 $\mathbf{B}$ 和 τ^2, 接下来的 d 个周期使用 (10) 的一个版本来估计 $\mathbf{L}_j$, 给定其余参数，其中 $j = 1 \ldots d$,,。我们正在开发一个用于相关数组变量响应的MMTR扩展。

用于估计算法1中的$\mathbf{L}_1$ 和 $\mathbf{L}_2$ 的目标依赖于随机效应和误差项服从高斯的假设，但我们的参数估计算法推广到仅假设随机效应的第一和第二矩存在的情况。在这种情况下，算法1迭代地将随机效应的前两个矩 $\mathbb{E}(\mathbf{A}_i \mid \mathbf{y}_i, \boldsymbol{\theta}^{(t)})$ 和 $\mathbb{E}\{\mathrm{vec}(\mathbf{A}_i)\mathrm{vec}(\mathbf{A}_i)^\top \mid \mathbf{y}_i, \boldsymbol{\theta}^{(t)}\}$ $(i = 1, \ldots, n)$替换为基于假设 $\mathbf{A}_i$ 服从矩阵正态分布的矩, 从而得到目标(10)和(11)。

致谢

Ian Hultman和Sanvesh Srivastava的部分研究由国立卫生研究院（1DP2MH126377-01）和国家科学基金会（DMS- 1854667）的资助。作者感谢Kshitij Khare、Joe Lang、Boxiang Wang和Dale Zimmerman对本文早期版本的宝贵反馈。实验中使用的代码可在 `https://github.com/IHultman/MMTR`处获取。

References

Barthelme, S. (2024). *imager: Image Processing Library Based on 'CImg'*. R package version 1.0.2.

Bates, D., M. Mächler, B. Bolker, and S. Walker (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software 67*(1), 1–48.

Bates, D., M. Maechler, B. Bolker, and S. Walker (2013). lme4: Linear mixed-effects models using Eigen and S4. *R package version 1.1-9*.

Bradic, J., G. Claeskens, and T. Gueuning (2020). Fixed effects testing in high-dimensional linear mixed models. *Journal of the American Statistical Association 115*(532), 1835–1850.

Chen, J. and Z. Chen (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika 95*(3), 759–771.

Dutilleul, P. (1999). The MLE algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation 64*(2), 105–123.

Fan, J., W. Gong, and Z. Zhu (2019). Generalized high-dimensional trace regression via nuclear norm regularization. *Journal of Econometrics 212*(1), 177–202.

Fan, Y. and R. Li (2012). Variable selection in linear mixed effects models. *Annals of statistics 40*(4), 2043.

Fosdick, B. K. and P. D. Hoff (2014). Separable factor analysis with applications to mortality data. *The annals of applied statistics 8*(1), 120.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1.

参考文献

Barthelme, S. (2024). *imager: Image Processing Library Based on 'CImg'*. R package version 1.0.2.

Bates, D., M. Mächler, B. Bolker, and S. Walker (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software 67*(1), 1–48.

Bates, D., M. Maechler, B. Bolker, and S. Walker (2013). lme4: Linear mixed-effects models using Eigen and S4. *R package version 1.1-9*.

Bradic, J., G. Claeskens, and T. Gueuning (2020). Fixed effects testing in high-dimensional linear mixed models. *Journal of the American Statistical Association 115*(532), 1835–1850.

Chen, J. and Z. Chen (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika 95*(3), 759–771.

Dutilleul, P. (1999). The MLE algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation 64*(2), 105–123.

Fan, J., W. Gong, and Z. Zhu (2019). Generalized high-dimensional trace regression via nuclear norm regularization. *Journal of Econometrics 212*(1), 177–202.

Fan, Y. and R. Li (2012). Variable selection in linear mixed effects models. *Annals of statistics 40*(4), 2043.

Fosdick, B. K. and P. D. Hoff (2014). Separable factor analysis with applications to mortality data. *The annals of applied statistics 8*(1), 120.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1.

Gerard, D. and P. Hoff (2015). Equivariant minimax dominators of the mle in the array normal model. *Journal of multivariate analysis 137*, 32–49.

Heiling, H. M., N. U. Rashid, Q. Li, and J. G. Ibrahim (2023). glmmpen: High dimensional penalized generalized linear mixed models. *The R journal 15*(4), 106.

Heiling, H. M., N. U. Rashid, Q. Li, X. L. Peng, J. J. Yeh, and J. G. Ibrahim (2024). Efficient computation of high-dimensional penalized generalized linear mixed models by latent factor modeling of the random effects. *Biometrics 80*(1), ujae016.

Hoff, P. D. (2011). Separable covariance arrays via the tucker product, with applications to multivariate relational data. *Bayesian Analysis 6*(2), 179–196.

Huang, G. B., M. Mattar, H. Lee, and E. Learned-Miller (2012). Learning to align from scratch. In *NIPS*.

Huang, G. B., M. Ramesh, T. Berg, and E. Learned-Miller (2007, October). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst.

Hui, F. K., S. Müller, and A. Welsh (2017). Joint selection in mixed models using regularized pql. *Journal of the American Statistical Association 112*(519), 1323–1333.

Hui, F. K., S. Müller, and A. H. Welsh (2021). Random effects misspecification can have severe consequences for random effects inference in linear mixed models. *International Statistical Review 89*(1), 186–206.

James, G. M., T. J. Hastie, and C. A. Sugar (2000). Principal component models for sparse functional data. *Biometrika 87*(3), 587–602.

Kumar, N., A. C. Berg, P. N. Belhumeur, and S. K. Nayar (2009). Attribute and simile classifiers for face verification. In *2009 IEEE 12th International Conference on Computer Vision*, pp. 365–372.

Gerard, D. and P. Hoff (2015). Equivariant minimax dominators of the mle in the array normal model. *Journal of multivariate analysis 137*, 32–49.

Heiling, H. M., N. U. Rashid, Q. Li, and J. G. Ibrahim (2023). glmmpen: High dimensional penalized generalized linear mixed models. *The R journal 15*(4), 106.

Heiling, H. M., N. U. Rashid, Q. Li, X. L. Peng, J. J. Yeh, and J. G. Ibrahim (2024). Efficient computation of high-dimensional penalized generalized linear mixed models by latent factor modeling of the random effects. *Biometrics 80*(1), ujae016.

Hoff, P. D. (2011). Separable covariance arrays via the tucker product, with applications to multivariate relational data. *Bayesian Analysis 6*(2), 179–196.

Huang, G. B., M. Mattar, H. Lee, and E. Learned-Miller (2012). Learning to align from scratch. In *NIPS*.

Huang, G. B., M. Ramesh, T. Berg, and E. Learned-Miller (2007, October). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst.

Hui, F. K., S. Müller, and A. Welsh (2017). Joint selection in mixed models using regularized pql. *Journal of the American Statistical Association 112*(519), 1323–1333.

Hui, F. K., S. Müller, and A. H. Welsh (2021). Random effects misspecification can have severe consequences for random effects inference in linear mixed models. *International Statistical Review 89*(1), 186–206.

James, G. M., T. J. Hastie, and C. A. Sugar (2000). Principal component models for sparse functional data. *Biometrika 87*(3), 587–602.

Kumar, N., A. C. Berg, P. N. Belhumeur, and S. K. Nayar (2009). Attribute and simile classifiers for face verification. In *2009 IEEE 12th International Conference on Computer Vision*, pp. 365–372.

Li, S., T. T. Cai, and H. Li (2022). Inference for high-dimensional linear mixed-effects models: A quasi-likelihood approach. *Journal of the American Statistical Association 117*(540), 1835–1846.

Lock, E. F. (2018). Tensor-on-tensor regression. *Journal of Computational and Graphical Statistics 27*(3), 638–647.

Lu, N. and D. L. Zimmerman (2005). The likelihood ratio test for a separable covariance matrix. *Statistics & Probability Letters 73*(4), 449–457.

Ročková, V. and E. I. George (2016). Fast bayesian factor analysis via automatic rotations to sparsity. *Journal of the American Statistical Association 111*(516), 1608–1622.

Seber, G. (2007). *Special Products and Operators*, Chapter 11, pp. 233–255. John Wiley & Sons, Ltd.

Slawski, M., P. Li, and M. Hein (2015). Regularization-free estimation in trace regression with symmetric positive semidefinite matrices. *Advances in neural information processing systems 28*.

Srivastava, M. S., T. von Rosen, and D. Von Rosen (2008). Models with a kronecker product covariance structure: estimation and testing. *Mathematical methods of statistics 17*, 357–370.

Srivastava, S., B. E. Engelhardt, and D. B. Dunson (2017). Expandable factor analysis. *Biometrika 104*(3), 649–663.

Sun, T. (2019). *scalreg: Scaled Sparse Linear Regression*. R package version 1.0.1.

Sun, T. and C.-H. Zhang (2012). Scaled sparse linear regression. *Biometrika 99*(4), 879–898.

Li, S., T. T. Cai, and H. Li (2022). Inference for high-dimensional linear mixed-effects models: A quasi-likelihood approach. *Journal of the American Statistical Association 117*(540), 1835–1846.

Lock, E. F. (2018). Tensor-on-tensor regression. *Journal of Computational and Graphical Statistics 27*(3), 638–647.

Lu, N. and D. L. Zimmerman (2005). The likelihood ratio test for a separable covariance matrix. *Statistics & Probability Letters 73*(4), 449–457.

Ročková, V. and E. I. George (2016). Fast bayesian factor analysis via automatic rotations to sparsity. *Journal of the American Statistical Association 111*(516), 1608–1622.

Seber, G. (2007). *Special Products and Operators*, Chapter 11, pp. 233–255. John Wiley & Sons, Ltd.

Slawski, M., P. Li, and M. Hein (2015). Regularization-free estimation in trace regression with symmetric positive semidefinite matrices. *Advances in neural information processing systems 28*.

Srivastava, M. S., T. von Rosen, and D. Von Rosen (2008). Models with a kronecker product covariance structure: estimation and testing. *Mathematical methods of statistics 17*, 357–370.

Srivastava, S., B. E. Engelhardt, and D. B. Dunson (2017). Expandable factor analysis. *Biometrika 104*(3), 649–663.

Sun, T. (2019). *scalreg: Scaled Sparse Linear Regression*. R package version 1.0.1.

Sun, T. and C.-H. Zhang (2012). Scaled sparse linear regression. *Biometrika 99*(4), 879–898.

27

van Dyk, D. A. (2000). Fitting mixed-effects models using efficient em-type algorithms. *Journal of Computational and Graphical Statistics 9*(1), 78–98.

Van Loan, C. F. (2000). The ubiquitous kronecker product. *Journal of computational and applied mathematics 123*(1-2), 85–100.

Verbeke, G., G. Molenberghs, and G. Verbeke (1997). *Linear mixed models for longitudinal data*. Springer.

Yang, Y., H. Zou, and S. Bhatnagar (2024). *gglasso: Group Lasso Penalized Learning Using a Unified BMD Algorithm*. R package version 1.5.1.

Yue, X., J. G. Park, Z. Liang, and J. Shi (2020). Tensor mixed effects model with application to nanomanufacturing inspection. *Technometrics 62*(1), 116–129.

Zhang, X., L. Li, H. Zhou, Y. Zhou, D. Shen, et al. (2019). Tensor generalized estimating equations for longitudinal imaging analysis. *Statistica Sinica 29*(4), 1977.

Zhang, Y., W. Shen, and D. Kong (2023). Covariance estimation for matrix-valued data. *Journal of the American Statistical Association 118*(544), 2620–2631.

Zhao, J., L. Niu, and S. Zhan (2017). Trace regression model with simultaneously low rank and row(column) sparse parameter. *Computational Statistics & Data Analysis 116*, 1–18.

Zhou, H. (2017). *Matlab TensorReg Toolbox Version 1.0*. `https://hua-zhou.github.io/TensorReg`.

Zhou, H. and B. Gaines (2017). *Matlab SparseReg Toolbox Version 1.0.0*. `https://hua-zhou.github.io/SparseReg`.

Zhou, H. and L. Li (2014). Regularized matrix regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology 76*(2), 463–483.

van Dyk, D. A. (2000). Fitting mixed-effects models using efficient em-type algorithms. *Journal of Computational and Graphical Statistics 9*(1), 78–98.

Van Loan, C. F. (2000). The ubiquitous kronecker product. *Journal of computational and applied mathematics 123*(1-2), 85–100.

Verbeke, G., G. Molenberghs, and G. Verbeke (1997). *Linear mixed models for longitudinal data*. Springer.

Yang, Y., H. Zou, and S. Bhatnagar (2024). *gglasso: Group Lasso Penalized Learning Using a Unified BMD Algorithm*. R package version 1.5.1.

Yue, X., J. G. Park, Z. Liang, and J. Shi (2020). Tensor mixed effects model with application to nanomanufacturing inspection. *Technometrics 62*(1), 116–129.

Zhang, X., L. Li, H. Zhou, Y. Zhou, D. Shen, et al. (2019). Tensor generalized estimating equations for longitudinal imaging analysis. *Statistica Sinica 29*(4), 1977.

Zhang, Y., W. Shen, and D. Kong (2023). Covariance estimation for matrix-valued data. *Journal of the American Statistical Association 118*(544), 2620–2631.

Zhao, J., L. Niu, and S. Zhan (2017). Trace regression model with simultaneously low rank and row(column) sparse parameter. *Computational Statistics & Data Analysis 116*, 1–18.

Zhou, H. (2017). *Matlab TensorReg Toolbox Version 1.0*. `https://hua-zhou.github.io/TensorReg`.

Zhou, H. and B. Gaines (2017). *Matlab SparseReg Toolbox Version 1.0.0*. `https://hua-zhou.github.io/SparseReg`.

Zhou, H. and L. Li (2014). Regularized matrix regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology 76*(2), 463–483.

Zhou, H., L. Li, and H. Zhu (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association 108*(502), 540–552.

周，朱，李和李（2013）。张量回归在神经影像数据分析中的应用。美国统计协会杂志 *108*(502)，540–552。

29

Supplementary Material: Regularized Parameter Estimation in Mixed Model Trace Regression

1 Proof of the Propositions

1.1 Proof of Proposition 2.1

Recall that Proposition 2.1 from the main manuscript.

Proposition 1.1 *Assume that* $(\mathbf{L}_2)_{j1} = 1$ *and* $(\mathbf{L}_2)_{j2} = \cdots = (\mathbf{L}_2)_{jS_2} = 0$. *Then,* $(\mathbf{\Sigma}_2)_{jj} = 1$ *and the* j*th diagonal* $Q_1 \times Q_1$ *block of* $\mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1$ *is* $\mathbf{\Sigma}_1$.

Proof [Proof of Proposition 1.1] Let $\mathbf{L}_2 \in \mathbb{R}^{Q_2 \times S_2}$ denote the matrix such that $\mathbf{L}_2 \mathbf{L}_2^\top = \mathbf{\Sigma}_2$, let $\mathbf{l}_{i:}$ denote the ith row vector of $\mathbf{L}_2$ and let $\mathbf{e}_1 \in \mathbb{R}^{S_2}$ denote the first standard basis vector in $\mathbb{R}^{S_2}$. If the jth row of $\mathbf{L}_2$ is $\mathbf{e}_1$, then:

$$\mathbf{L}_2 = \begin{bmatrix} \mathbf{l}_{1:}^\top \\ \vdots \\ \mathbf{e}_1^\top \\ \vdots \\ \mathbf{l}_{Q_2:}^\top \end{bmatrix} \implies \mathbf{L}_2 \mathbf{L}_2^\top = \begin{bmatrix} \mathbf{l}_{1:}^\top \mathbf{l}_{1:} & \dots & (\mathbf{L}_2)_{11} & \dots & \mathbf{l}_{1:}^\top \mathbf{l}_{Q_2:} \\ \vdots & \ddots & \vdots & & \vdots \\ (\mathbf{L}_2)_{11} & \dots & 1 & \dots & (\mathbf{L}_2)_{Q_2 1} \\ \vdots & & \vdots & \ddots & \vdots \\ \mathbf{l}_{Q_2:}^\top \mathbf{l}_{1:} & \dots & (\mathbf{L}_2)_{Q_2 1} & \dots & \mathbf{l}_{Q_2:}^\top \mathbf{l}_{Q_2:} \end{bmatrix},$$

demonstrating that $(\mathbf{\Sigma}_2)_{jj} = 1$ which implies, by definition of the Kronecker product, that the jth diagonal $Q_1 \times Q_1$ block of $\mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1$ is $\mathbf{\Sigma}_1$. ■

1.2 Proof of Proposition 2.2

Recall that Proposition 2.2 from the main manuscript.

补充材料：混合模型追踪回归中的正则化参数估计

1 命题的证明

1.1 命题2.1的证明

回想一下主文稿中的命题2.1。

命题1.1 假设 $(\mathbf{L}_2)_{j1} = 1$ 和 $(\mathbf{L}_2)_{j2} = \cdots = (\mathbf{L}_2)_{jS_2} = 0$。然后， $(\mathbf{\Sigma}_2)_{jj} = 1$ 并且 j*th* 对角线 $Q_1 \times Q_1$ 块 *of* $\mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1$ 是 $\mathbf{\Sigma}_1$。

证明 [命题1.1] 的证明。令 $\mathbf{L}_2 \in \mathbb{R}^{Q_2 \times S_2}$ 表示满足 $\mathbf{L}_2 \mathbf{L}_2^\top = \mathbf{\Sigma}_2$的矩阵，令 $\mathbf{l}_{i:}$ 表示 $\mathbf{L}_2$ 的 ith 行向量，并令 $\mathbf{e}_1 \in \mathbb{R}^{S_2}$ 表示 $\mathbb{R}^{S_2}$中的第一个标准基向量。如果 $\mathbf{L}_2$ 的 jth 行是 $\mathbf{e}_1$，则：

$$\mathbf{L}_2 = \begin{bmatrix} \mathbf{l}_{1:}^\top \\ \vdots \\ \mathbf{e}_1^\top \\ \vdots \\ \mathbf{l}_{Q_2:}^\top \end{bmatrix} \implies \mathbf{L}_2 \mathbf{L}_2^\top = \begin{bmatrix} \mathbf{l}_{1:}^\top \mathbf{l}_{1:} & \dots & (\mathbf{L}_2)_{11} & \dots & \mathbf{l}_{1:}^\top \mathbf{l}_{Q_2:} \\ \vdots & \ddots & \vdots & & \vdots \\ (\mathbf{L}_2)_{11} & \dots & 1 & \dots & (\mathbf{L}_2)_{Q_2 1} \\ \vdots & & \vdots & \ddots & \vdots \\ \mathbf{l}_{Q_2:}^\top \mathbf{l}_{1:} & \dots & (\mathbf{L}_2)_{Q_2 1} & \dots & \mathbf{l}_{Q_2:}^\top \mathbf{l}_{Q_2:} \end{bmatrix},$$

证明 $(\mathbf{\Sigma}_2)_{jj} = 1$ ，这意味着，根据 Kronecker 积的定义， jth 对角线 $Q_1 \times Q_1$ 块 of $\mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1$ 是 $\mathbf{\Sigma}_1$。 ■

1.2 命题2.2的证明

回想一下主文稿中的命题2.2。

Proposition 1.2 *For any matrix* $\mathbf{L} \in \mathbb{R}^{Q \times S}$*, there exists an orthonormal matrix* $\mathbf{Q}_j \in \mathbb{R}^{S \times S}$ *and constant* c *such that* $c\,\mathbf{L}\,\mathbf{Q}_j^\top$ *is a matrix whose* j*th row is the first standard basis vector in* $\mathbb{R}^S$*.*

Proof [Proof of Proposition 1.2] Let $\mathbf{L}$ be any matrix in $\mathbb{R}^{Q \times S}$ such that $\mathbf{L}\,\mathbf{L}^\top = \mathbf{\Sigma}$, and let $\mathbf{Q}_j \in \mathbb{R}^{S \times S}$ be the Householder matrix that reflects the jth row vector of $\mathbf{L}$ onto the first standard basis vector in $\mathbb{R}^S$, denoted as $\mathbf{e}_1$. Let $\mathbf{l}_{i:}$ denote the ith row vector of $\mathbf{L}$, then

$$\mathbf{L} = \begin{bmatrix} \mathbf{l}_{1:}^\top \\ \vdots \\ \mathbf{l}_{j:}^\top \\ \vdots \\ \mathbf{l}_{Q:}^\top \end{bmatrix} \implies \mathbf{L}\,\mathbf{Q}_j^\top = \begin{bmatrix} \mathbf{l}_{1:}^\top\,\mathbf{Q}_j^\top \\ \vdots \\ \|\mathbf{l}_{j:}\|_2\,\mathbf{e}_1^\top \\ \vdots \\ \mathbf{l}_{Q:}^\top\,\mathbf{Q}_j^\top \end{bmatrix} \implies \frac{1}{\|\mathbf{l}_{j:}\|_2}\,\mathbf{L}\,\mathbf{Q}_j^\top = \begin{bmatrix} \frac{1}{\|\mathbf{l}_{j:}\|_2}\,\mathbf{l}_{1:}^\top\,\mathbf{Q}_j^\top \\ \vdots \\ \mathbf{e}_1^\top \\ \vdots \\ \frac{1}{\|\mathbf{l}_{j:}\|_2}\,\mathbf{l}_{Q:}^\top\,\mathbf{Q}_j^\top \end{bmatrix},$$

which has the form described in Proposition 1.1. Because Householder transformations are orthonormal,

$$\frac{1}{\|\mathbf{l}_{j:}\|_2^2}\,\mathbf{L}\,\mathbf{Q}_j^\top\,\mathbf{Q}_j\,\mathbf{L}^\top = \frac{1}{\|\mathbf{l}_{j:}\|_2^2}\,\mathbf{\Sigma},$$

which is just the initial covariance matrix $\mathbf{\Sigma}$ scaled such that its jth diagonal entry is equal to one. ■

1.3 Proof of Proposition 3.1

Recall Proposition 3.1 from the main manuscript. Let $\ell(\boldsymbol{\theta})$ be the negative log likehood function implied by the MMTR, $\boldsymbol{\theta}_1 = (\mathbf{B}, \tau^2)$, $\boldsymbol{\theta}_2 = \mathbf{L}_1$, $\boldsymbol{\theta}_3 = \mathbf{L}_2$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$, and $\mathcal{P}_{\lambda_\mathbf{B}}(\boldsymbol{\theta}_1)$, $\mathcal{P}_{\lambda_\mathbf{L}}(\boldsymbol{\theta}_2)$, and $\mathcal{P}_{\lambda_\mathbf{L}}(\boldsymbol{\theta}_3)$ are the penalties on $\mathbf{B}$, $\mathbf{L}_1$, and $\mathbf{L}_2$, respectively. Then, the Proposition 3.1 from the main manuscript is as follows.

命题1.2 对于任何矩阵 $\mathbf{L} \in \mathbb{R}^{Q \times S}$，都存在一个正交矩阵 $\mathbf{Q}_j \in \mathbb{R}^{S \times S}$和常数 c，使得 $c\,\mathbf{L}\,\mathbf{Q}_j^\top$ 是一个矩阵，其 j 行是 $\mathbb{R}^S$ 中的第一个标准基向量。

证明 [命题1.2] 的证明。令 $\mathbf{L}$ 是 $\mathbb{R}^{Q \times S}$ 中的任意矩阵，使得 $\mathbf{L}\,\mathbf{L}^\top = \mathbf{\Sigma}$，并且令 $\mathbf{Q}_j \in \mathbb{R}^{S \times S}$ 是一个Householder矩阵，它将 $\mathbf{L}$ 的 j 行向量反射到 $\mathbb{R}^S$ 中的第一个标准基向量上，记为 $\mathbf{e}_1$。令 $\mathbf{l}_{i:}$ 表示 $\mathbf{L}$的 i 行向量，则

$$\mathbf{L} = \begin{bmatrix} \mathbf{l}_{1:}^\top \\ \vdots \\ \mathbf{l}_{j:}^\top \\ \vdots \\ \mathbf{l}_{Q:}^\top \end{bmatrix} \implies \mathbf{L}\,\mathbf{Q}_j^\top = \begin{bmatrix} \mathbf{l}_{1:}^\top\,\mathbf{Q}_j^\top \\ \vdots \\ \|\mathbf{l}_{j:}\|_2\,\mathbf{e}_1^\top \\ \vdots \\ \mathbf{l}_{Q:}^\top\,\mathbf{Q}_j^\top \end{bmatrix} \implies \frac{1}{\|\mathbf{l}_{j:}\|_2}\,\mathbf{L}\,\mathbf{Q}_j^\top = \begin{bmatrix} \frac{1}{\|\mathbf{l}_{j:}\|_2}\,\mathbf{l}_{1:}^\top\,\mathbf{Q}_j^\top \\ \vdots \\ \mathbf{e}_1^\top \\ \vdots \\ \frac{1}{\|\mathbf{l}_{j:}\|_2}\,\mathbf{l}_{Q:}^\top\,\mathbf{Q}_j^\top \end{bmatrix},$$

其形式如命题1.1所述。因为Householder变换是正交的，

$$\frac{1}{\|\mathbf{l}_{j:}\|_2^2}\,\mathbf{L}\,\mathbf{Q}_j^\top\,\mathbf{Q}_j\,\mathbf{L}^\top = \frac{1}{\|\mathbf{l}_{j:}\|_2^2}\,\mathbf{\Sigma},$$

这仅仅是初始协方差矩阵 $\mathbf{\Sigma}$ 的缩放，使得其 j 对角线元素等于1。 ■

1.3 命题3.1的证明

回忆主文稿中的命题3.1。令 $\ell(\boldsymbol{\theta})$是由MMTR隐含的负对数似然函数，$\boldsymbol{\theta}_1 = (\mathbf{B}\ \tau^2)$，$\boldsymbol{\theta}_2 = \mathbf{L}_1$，$\boldsymbol{\theta}_3 = \mathbf{L}_2$，$\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$，$\mathcal{P}_{\lambda_\mathbf{B}}(\boldsymbol{\theta}_1)$，$\mathcal{P}_{\lambda_\mathbf{L}}(\boldsymbol{\theta}_2)$，$\mathcal{P}_{\lambda_\mathbf{L}}(\boldsymbol{\theta}_3)$ 分别是关于 $\mathbf{B}$，$\mathbf{L}_1$，和 $\mathbf{L}_2$ 的惩罚，分别。那么，主文稿中的命题3.1如下。

Proposition 1.3 *The MMTR objective is* $f(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) + \mathcal{P}_{\lambda_\mathbf{B}}(\boldsymbol{\theta}_1) + \mathcal{P}_{\lambda_\mathbf{L}}(\boldsymbol{\theta}_2) + \mathcal{P}_{\lambda_\mathbf{L}}(\boldsymbol{\theta}_3)$. *Let* $\mathcal{M}(\cdot)$ *be the function that maps* $\boldsymbol{\theta}^{(t)}$ *to* $\boldsymbol{\theta}^{(t+1)}$ *using Algorithm 1 in the main manuscript. Then, each iteration of Algorithm 1 does not increase* $f(\boldsymbol{\theta})$. *Furthermore, assume that the parameter space* $\boldsymbol{\Theta}$ *is compact and* $f(\boldsymbol{\theta}) = f\{\mathcal{M}(\boldsymbol{\theta})\}$ *only for the stationary points of* $f(\boldsymbol{\theta})$. *Then, the* $\{\boldsymbol{\theta}^{(t)}\}_{t=1}^{\infty}$ *sequence converges to a stationary point.*

Proof [Proof of Proposition 1.3] Let $\boldsymbol{\theta}^{(t)} = (\boldsymbol{\theta}_1^{(t)}, \boldsymbol{\theta}_2^{(t)}, \boldsymbol{\theta}_3^{(t)})$ for any $t = 0, 1, 2, \ldots, \infty$. Then, for any non-negative $\lambda_\mathbf{B}$ and $\lambda_\mathbf{L}$,

$$f(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t)}, \boldsymbol{\theta}_3^{(t)}) \leq f(\boldsymbol{\theta}_1^{(t)}, \boldsymbol{\theta}_2^{(t)}, \boldsymbol{\theta}_3^{(t)}) \tag{1}$$

because the first cycle implies that $\boldsymbol{\theta}_1^{(t+1)} = \underset{\boldsymbol{\theta}_1}{\operatorname{argmin}} f(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2^{(t)}, \boldsymbol{\theta}_3^{(t)})$; see (8) in the main manuscript. We use two properties of the AECM algorithm's second cycle. First, $-\mathcal{Q}_{(1)}(\boldsymbol{\theta}_2) + \mathcal{P}_{\lambda_\mathbf{L}}(\boldsymbol{\theta}_2)$ majorizes $f(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3^{(t)})$ for every $\boldsymbol{\theta}_2$, where $\mathcal{Q}_{(1)}$ is defined in (9) in the main manuscript; therefore, $-\mathcal{Q}_{(1)}(\boldsymbol{\theta}_2) + \mathcal{P}_{\lambda_\mathbf{L}}(\boldsymbol{\theta}_2) \leq f(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3^{(t)})$. Second, $-\mathcal{Q}_{(1)}(\boldsymbol{\theta}_2^{(t)}) + \mathcal{P}_{\lambda_\mathbf{L}}(\boldsymbol{\theta}_2^{(t)}) = f(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t)}, \boldsymbol{\theta}_3^{(t)})$; see (9) and (10) in the main manuscript for details. We use these properties to obtain that

$$\begin{aligned}
f(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t+1)}, \boldsymbol{\theta}_3^{(t)}) &= -\mathcal{Q}_{(1)}(\boldsymbol{\theta}_2^{(t+1)}) + \mathcal{P}_{\lambda_\mathbf{L}}(\boldsymbol{\theta}_2^{(t+1)}) + \\
&\quad f(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t+1)}, \boldsymbol{\theta}_3^{(t)}) - \{-\mathcal{Q}_{(1)}(\boldsymbol{\theta}_2^{(t+1)}) + \mathcal{P}_{\lambda_\mathbf{L}}(\boldsymbol{\theta}_2^{(t+1)})\} \\
&\overset{(i)}{\leq} -\mathcal{Q}_{(1)}(\boldsymbol{\theta}_2^{(t+1)}) + \mathcal{P}_{\lambda_\mathbf{L}}(\boldsymbol{\theta}_2^{(t+1)}) \\
&\overset{(ii)}{\leq} -\mathcal{Q}_{(1)}(\boldsymbol{\theta}_2^{(t)}) + \mathcal{P}_{\lambda_\mathbf{L}}(\boldsymbol{\theta}_2^{(t)}) \\
&\overset{(iii)}{=} f(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t)}, \boldsymbol{\theta}_3^{(t)}),
\end{aligned} \tag{2}$$

where (i) follows because $f(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t+1)}, \boldsymbol{\theta}_3^{(t)}) - \{-\mathcal{Q}_{(1)}(\boldsymbol{\theta}_2^{(t+1)}) + \mathcal{P}_{\lambda_\mathbf{L}}(\boldsymbol{\theta}_2^{(t+1)}) \geq 0$ using the first property of AECM algorithm's second cycle, (ii) follows because $\boldsymbol{\theta}_2^{(t+1)} = \underset{\boldsymbol{\theta}_2}{\operatorname{argmin}} -\mathcal{Q}_{(1)}(\boldsymbol{\theta}_2) + \mathcal{P}_{\lambda_\mathbf{L}}(\boldsymbol{\theta}_2)$, and (iii) follows from the second property of AECM algorithm's second

命题1.3 *MMTR*目标为 $f(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) + \mathcal{P}_{\lambda_\mathbf{B}}(\boldsymbol{\theta}_1) + \mathcal{P}_{\lambda_\mathbf{L}}(\boldsymbol{\theta}_2) + \mathcal{P}_{\lambda_\mathbf{L}}(\boldsymbol{\theta}_3)$。设 $\mathcal{M}(\cdot)$ 是使用主文稿中算法*1*将 $\boldsymbol{\theta}^{(t)}$ 映射到 $\boldsymbol{\theta}^{(t+1)}$ 的函数。那么，算法*1*的每次迭代都不会增加 $f(\boldsymbol{\theta})$。此外，假设参数空间 $\boldsymbol{\Theta}$ 是紧致的，并且 $f(\boldsymbol{\theta}) = f\{\mathcal{M}(\boldsymbol{\theta})\}$ 仅对于 $f(\boldsymbol{\theta})$ 的驻点。那么，$\{\boldsymbol{\theta}^{(t)}\}_{t=1}^{\infty}$ 序列收敛到一个驻点。

证明 [命题1.3] 的证明。设 $\boldsymbol{\theta}^{(t)} = (\boldsymbol{\theta}_1^{(t)}, \boldsymbol{\theta}_2^{(t)}, \boldsymbol{\theta}_3^{(t)})$ 对于任何 $t = 0\ 1\ 2\ \ldots\ \infty$,,,,。那么，对于任何非负的 $\lambda_\mathbf{B}$ 和 $\lambda_\mathbf{L}$，

$$f(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t)}, \boldsymbol{\theta}_3^{(t)}) \leq f(\boldsymbol{\theta}_1^{(t)}, \boldsymbol{\theta}_2^{(t)}, \boldsymbol{\theta}_3^{(t)}) \tag{1}$$

因为第一个周期意味着 $\boldsymbol{\theta}_1^{(t+1)} = \underset{\boldsymbol{\theta}_1}{\operatorname{argmin}} f(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2^{(t)}, \boldsymbol{\theta}_3^{(t)})$；参见主文稿中的(8)。我们使用AECM算法第二个周期的两个性质。首先，$-\mathcal{Q}_{(1)}(\boldsymbol{\theta}_2) + \mathcal{P}_{\lambda_\mathbf{L}}(\boldsymbol{\theta}_2)$ 对每个 $\boldsymbol{\theta}_2$ 支配 $f(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3^{(t)})$，其中 $\mathcal{Q}_{(1)}$ 在主文稿的(9)中定义；因此，$-\mathcal{Q}_{(1)}(\boldsymbol{\theta}_2) + \mathcal{P}_{\lambda_\mathbf{L}}(\boldsymbol{\theta}_2) \leq f(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3^{(t)})$。其次，$-\mathcal{Q}_{(1)}(\boldsymbol{\theta}_2^{(t)}) + \mathcal{P}_{\lambda_\mathbf{L}}(\boldsymbol{\theta}_2^{(t)}) = f(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t)}, \boldsymbol{\theta}_3^{(t)})$；参见主文稿中的(9)和(10)以获取详细信息。我们使用这些性质来获得

$$\begin{aligned}
f(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t+1)}, \boldsymbol{\theta}_3^{(t)}) &= -\mathcal{Q}_{(1)}(\boldsymbol{\theta}_2^{(t+1)}) + \mathcal{P}_{\lambda_\mathbf{L}}(\boldsymbol{\theta}_2^{(t+1)}) + \\
&\quad f(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t+1)}, \boldsymbol{\theta}_3^{(t)}) - \{-\mathcal{Q}_{(1)}(\boldsymbol{\theta}_2^{(t+1)}) + \mathcal{P}_{\lambda_\mathbf{L}}(\boldsymbol{\theta}_2^{(t+1)})\} \\
&\overset{(i)}{\leq} -\mathcal{Q}_{(1)}(\boldsymbol{\theta}_2^{(t+1)}) + \mathcal{P}_{\lambda_\mathbf{L}}(\boldsymbol{\theta}_2^{(t+1)}) \\
&\overset{(ii)}{\leq} -\mathcal{Q}_{(1)}(\boldsymbol{\theta}_2^{(t)}) + \mathcal{P}_{\lambda_\mathbf{L}}(\boldsymbol{\theta}_2^{(t)}) \\
&\overset{(iii)}{=} f(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t)}, \boldsymbol{\theta}_3^{(t)}),
\end{aligned} \tag{2}$$

在i之后，因为 $f(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t+1)}, \boldsymbol{\theta}_3^{(t)}) - \{-\mathcal{Q}_{(1)}(\boldsymbol{\theta}_2^{(t+1)}) + \mathcal{P}_{\lambda_\mathbf{L}}(\boldsymbol{\theta}_2^{(t+1)}) \geq 0$ 使用了AECM算法第二周期的第一个属性，所以(ii)之后，因为 $\boldsymbol{\theta}_2^{(t+1)} = \underset{\boldsymbol{\theta}_2}{\operatorname{argmin}} -\mathcal{Q}_{(1)}(\boldsymbol{\theta}_2) + \mathcal{P}_{\lambda_\mathbf{L}}(\boldsymbol{\theta}_2)$，并且$(iii)$是从AECM算法第二周期的第二个属性得出的

cycle, which implies that $\mathcal{Q}_{(1)}(\boldsymbol{\theta}_2^{(t)}) + \mathcal{P}_{\lambda_L}(\boldsymbol{\theta}_2^{(t)}) = f(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t)}, \boldsymbol{\theta}_3^{(t)})$.

We follow a similar sequence of arguments after swapping the subscripts 1 and 2 to obtain lower bound for $f(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t+1)}, \boldsymbol{\theta}_3^{(t)})$ in (2) at the end for the AECM algorithm's third cycle. The objective at the end of the third cycle satisfies

$$
\begin{aligned}
f(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t+1)}, \boldsymbol{\theta}_3^{(t+1)}) &= -\mathcal{Q}_{(2)}(\boldsymbol{\theta}_3^{(t+1)}) + \mathcal{P}_{\lambda_L}(\boldsymbol{\theta}_3^{(t+1)}) + \\
&\quad f(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t+1)}, \boldsymbol{\theta}_3^{(t+1)}) - \{-\mathcal{Q}_{(2)}(\boldsymbol{\theta}_3^{(t+1)}) + \mathcal{P}_{\lambda_L}(\boldsymbol{\theta}_3^{(t+1)})\} \\
&\overset{(i)}{\leq} -\mathcal{Q}_{(2)}(\boldsymbol{\theta}_3^{(t+1)}) + \mathcal{P}_{\lambda_L}(\boldsymbol{\theta}_3^{(t+1)}) \\
&\overset{(ii)}{\leq} -\mathcal{Q}_{(2)}(\boldsymbol{\theta}_3^{(t)}) + \mathcal{P}_{\lambda_L}(\boldsymbol{\theta}_3^{(t)}) \\
&\overset{(iii)}{=} f(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t+1)}, \boldsymbol{\theta}_3^{(t)}),
\end{aligned} \tag{3}
$$

where $(i), (ii)$, and (iii) follow from the same arguments used in (2), except we swap 1 and 2. Finally, (1), (2), and (3) imply that

$$
f(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t+1)}, \boldsymbol{\theta}_3^{(t+1)}) \leq f(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t+1)}, \boldsymbol{\theta}_3^{(t)}) \leq f(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t)}, \boldsymbol{\theta}_3^{(t)}) \leq f(\boldsymbol{\theta}_1^{(t)}, \boldsymbol{\theta}_2^{(t)}, \boldsymbol{\theta}_3^{(t)}), \tag{4}
$$

showing that the AECM algorithm does not increase the objective in every iteration. The objective function $f(\boldsymbol{\theta})$ is bounded for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ because $\boldsymbol{\Theta}$ is compact; therefore, $f(\boldsymbol{\theta}^{(t)})$ is a bounded non-increasing sequence, so $f(\boldsymbol{\theta}^{(t)})$ converges to $f(\boldsymbol{\theta}^{(\infty)})$. Our assumption implies that convergence happens only for stationary points, so $\boldsymbol{\theta}^{(\infty)}$ is a stationary point of $\boldsymbol{\theta}^{(t)}$ sequence. The proposition is proved. ■

周期，这意味着 $\mathcal{Q}_{(1)}(\boldsymbol{\theta}_2^{(t)}) + \mathcal{P}_{\lambda_L}(\boldsymbol{\theta}_2^{(t)}) = f(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t)}, \boldsymbol{\theta}_3^{(t)})$.

我们在交换下标 1 和 2 后，遵循类似的论证序列，以在 AECM 算法的第三次周期结束时为 $f(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t+1)}, \boldsymbol{\theta}_3^{(t)})$ 在 (2) 中获得下界。第三次周期结束时的目标满足

$$
\begin{aligned}
f(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t+1)}, \boldsymbol{\theta}_3^{(t+1)}) &= -\mathcal{Q}_{(2)}(\boldsymbol{\theta}_3^{(t+1)}) + \mathcal{P}_{\lambda_L}(\boldsymbol{\theta}_3^{(t+1)}) + \\
&\quad f(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t+1)}, \boldsymbol{\theta}_3^{(t+1)}) - \{-\mathcal{Q}_{(2)}(\boldsymbol{\theta}_3^{(t+1)}) + \mathcal{P}_{\lambda_L}(\boldsymbol{\theta}_3^{(t+1)})\} \\
&\overset{(i)}{\leq} -\mathcal{Q}_{(2)}(\boldsymbol{\theta}_3^{(t+1)}) + \mathcal{P}_{\lambda_L}(\boldsymbol{\theta}_3^{(t+1)}) \\
&\overset{(ii)}{\leq} -\mathcal{Q}_{(2)}(\boldsymbol{\theta}_3^{(t)}) + \mathcal{P}_{\lambda_L}(\boldsymbol{\theta}_3^{(t)}) \\
&\overset{(iii)}{=} f(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t+1)}, \boldsymbol{\theta}_3^{(t)}),
\end{aligned} \tag{3}
$$

其中 (i),(ii), 和 (iii) 源自 (2) 中使用的相同论证，只是我们交换了 1 和 2。最后，(1)、(2) 和 (3) 意味着

$$
f(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t+1)}, \boldsymbol{\theta}_3^{(t+1)}) \leq f(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t+1)}, \boldsymbol{\theta}_3^{(t)}) \leq f(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t)}, \boldsymbol{\theta}_3^{(t)}) \leq f(\boldsymbol{\theta}_1^{(t)}, \boldsymbol{\theta}_2^{(t)}, \boldsymbol{\theta}_3^{(t)}), \tag{4}
$$

这表明 AECM 算法并非在每次迭代中都增加目标。目标函数 $f(\boldsymbol{\theta})$ 对每个 $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ 都是有限的，因为 $\boldsymbol{\Theta}$ 是紧致的；因此，$f(\boldsymbol{\theta}^{(t)})$ 是一个有界非增序列，所以 $f(\boldsymbol{\theta}^{(t)})$ 收敛到 $f(\boldsymbol{\theta}^{(\infty)})$。我们的假设意味着收敛仅发生在驻点，所以 $\boldsymbol{\theta}^{(\infty)}$ 是 $\boldsymbol{\theta}^{(t)}$ 序列的一个驻点。命题得证。

■

2 Derivation of the AECM Algorithm Updates

2.1 Derivation of $\mathbf{B}^{(t+1)}$ and $\tau^{2(t+1)}$

To estimate $\mathbf{B}^{(t+1)}$ and $\tau^{2(t+1)}$, we directly use the observed data log likelihood, $\log f_{\mathbf{y}\,|\,\boldsymbol{\theta}}(\mathbf{y})$, while conditioning on $\mathbf{L}_1^{(t)}$ and $\mathbf{L}_2^{(t)}$ which amounts to solving a weighted least squares; see Equation (7) in the main manuscript. Given $\mathbf{L}_1^{(t)}$ and $\mathbf{L}_2^{(t)}$, the log likelihood as a function of $(\mathbf{B}, \tau^2)$ is

$$\begin{aligned}\log f_{\mathbf{y}\,|\,\boldsymbol{\theta}}(\mathbf{y}) &= \sum_{i=1}^{n} \log f_{\mathbf{y}_i\,|\,\boldsymbol{\theta}}(\mathbf{y}_i) \\ &= -\frac{N}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{n} \operatorname{logdet}(\boldsymbol{\Sigma}_{\mathbf{y}_i}^{(t)}) - \frac{1}{2}\sum_{i=1}^{n} (\mathbf{y}_i - \mathbf{X}_i\,\mathbf{b})^\top\, \boldsymbol{\Sigma}_{\mathbf{y}_i}^{-1(t)} (\mathbf{y}_i - \mathbf{X}_i\,\mathbf{b}),\end{aligned}$$

where $\log f_{\mathbf{y}_i\,|\,\boldsymbol{\theta}}(\mathbf{y}_i)$ is the log likelihood sample i, $\boldsymbol{\Sigma}_{\mathbf{y}_i}^{(t)} = \tau^2(\mathbf{Z}_{i(1)}\,\mathbf{L}_{(1)}^{(t)}\,\mathbf{L}_{(1)}^{(t)\top}\,\mathbf{Z}_{i(1)}^\top + \mathbf{I}_{m_i})$, $\mathbf{Z}_{i(1)}$ is the matrix whose jth row is $\operatorname{vec}(\mathbf{Z}_{ij})^\top$, $\mathbf{b} = \operatorname{vec}(\mathbf{B})$, and $\mathbf{L}_{(1)}^{(t)} = \mathbf{L}_2^{(t)} \otimes \mathbf{L}_1^{(t)}$; see Section 2.1 in the main manuscript about the notation. Let $\boldsymbol{\Lambda}_i^{(t)} = \mathbf{Z}_{i(1)}\,\mathbf{L}_{(1)}^{(t)}\,\mathbf{L}_{(1)}^{(t)\top}\,\mathbf{Z}_{i(1)}^\top + \mathbf{I}_{m_i}$, $\boldsymbol{\Lambda}_i^{1/2(t)}$ be any matrix such that $\boldsymbol{\Lambda}_i^{1/2(t)}\,\boldsymbol{\Lambda}_i^{1/2(t)\top} = \boldsymbol{\Lambda}_i^{(t)}$, and the scaled response and predictor matrices be

$$\breve{\mathbf{y}} = \begin{bmatrix} \boldsymbol{\Lambda}_1^{-1/2(t)}\,\mathbf{y}_1 \\ \boldsymbol{\Lambda}_2^{-1/2(t)}\,\mathbf{y}_2 \\ \vdots \\ \boldsymbol{\Lambda}_n^{-1/2(t)}\,\mathbf{y}_n \end{bmatrix}, \quad \breve{\mathbf{X}} = \begin{bmatrix} \boldsymbol{\Lambda}_1^{-1/2(t)}\,\mathbf{X}_1 \\ \boldsymbol{\Lambda}_2^{-1/2(t)}\,\mathbf{X}_2 \\ \vdots \\ \boldsymbol{\Lambda}_n^{-1/2(t)}\,\mathbf{X}_n \end{bmatrix}.$$

Then, maximizing $\log f_{\mathbf{y}\,|\,\boldsymbol{\theta}}(\mathbf{y})$ with respect to $\mathbf{b}$ while conditioning on $\mathbf{L}_1^{(t)}$ and $\mathbf{L}_2^{(t)}$ amounts to minimizing the quadratic form $(\breve{\mathbf{y}} - \breve{\mathbf{X}}\,\mathbf{b})^\top(\breve{\mathbf{y}} - \breve{\mathbf{X}}\,\mathbf{b})$ with respect to $\mathbf{b}$. We estimate $\mathbf{B}^{(t+1)}$ and $\tau^{2(t+1)}$ simultaneously by adding the scaled lasso penalty terms to this quadratic form as in Equation (8) in the main manuscript to enforce sparsity in $\mathbf{B}^{(t+1)}$ and reduce the bias in estimating $\tau^{2(t+1)}$. We also update $\boldsymbol{\theta}^{(t)}$ to $\boldsymbol{\theta}^{(t+1/3)} = (\mathbf{B}^{(t+1)}, \mathbf{L}_1^{(t)}, \mathbf{L}_2^{(t)}, \tau^{2(t+1)})$.

2 AECM算法更新推导

2.1 $\mathbf{B}^{(t+1)}$ 和 $\tau^{2(t+1)}$的推导

为估计 $\mathbf{B}_{(t+1)}$和 $\tau^{2(t+1)}$，我们直接使用观测数据的对数似然log $f_{\mathbf{y}\,|\,\boldsymbol{\theta}}(\mathbf{y})$，同时条件在 $\mathbf{L}_1^{(t)}$ 和 $\mathbf{L}_2^{(t)}$ 上，这相当于求解加权最小二乘；参见主文稿中方程(7)。给定 $\mathbf{L}_1^{(t)}$ 和 $\mathbf{L}_2^{(t)}$，对数似然作为 $(\mathbf{B}\ \tau^2)$ 的函数是

$$\begin{aligned}\log f_{\mathbf{y}\,|\,\boldsymbol{\theta}}(\mathbf{y}) &= \sum_{i=1}^{n} \log f_{\mathbf{y}_i\,|\,\boldsymbol{\theta}}(\mathbf{y}_i) \\ &= -\frac{N}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{n} \operatorname{logdet}(\boldsymbol{\Sigma}_{\mathbf{y}_i}^{(t)}) - \frac{1}{2}\sum_{i=1}^{n} (\mathbf{y}_i - \mathbf{X}_i\,\mathbf{b})^\top\, \boldsymbol{\Sigma}_{\mathbf{y}_i}^{-1(t)} (\mathbf{y}_i - \mathbf{X}_i\,\mathbf{b}),\end{aligned}$$

其中log $f_{\mathbf{y}_i\,|\,\boldsymbol{\theta}}(\mathbf{y}_i)$是对数似然样本 i，$\boldsymbol{\Sigma}_{\mathbf{y}_i}^{(t)} = \tau^2(\mathbf{Z}_{i(1)}\,\mathbf{L}_{(1)}^{(t)}\,\mathbf{L}_{(1)}^{(t)\top}\,\mathbf{Z}_{i(1)}^\top + \mathbf{I}_{m_i})$ $\mathbf{Z}_{i(1,)}$是矩阵，其第 j行是$\operatorname{vec}(\mathbf{Z}_{ij})^\top$，$\mathbf{b} = \operatorname{vec}(\mathbf{B})$，和 $\mathbf{L}_{(1)}^{(t)} = \mathbf{L}_2^{(t)} \otimes \mathbf{L}_1^{(t)}$；参见主文稿中关于符号的2.1节。设 $\boldsymbol{\Lambda}_i^{(t)} = \mathbf{Z}_{i(1)}\,\mathbf{L}_{(1)}^{(t)}\,\mathbf{L}_{(1)}^{(t)\top}\,\mathbf{Z}_{i(1)}^\top + \mathbf{I}_{m_i}$ $\boldsymbol{\Lambda}_i^{1/2(t)}$，是任何矩阵，使得 $\boldsymbol{\Lambda}_i^{1/2(t)}\,\boldsymbol{\Lambda}_i^{1/2(t)\top} = \boldsymbol{\Lambda}_i^{(t)}$，和缩放响应和预测矩阵为

$$\breve{\mathbf{y}} = \begin{bmatrix} \boldsymbol{\Lambda}_1^{-1/2(t)}\,\mathbf{y}_1 \\ \boldsymbol{\Lambda}_2^{-1/2(t)}\,\mathbf{y}_2 \\ \vdots \\ \boldsymbol{\Lambda}_n^{-1/2(t)}\,\mathbf{y}_n \end{bmatrix}, \quad \breve{\mathbf{X}} = \begin{bmatrix} \boldsymbol{\Lambda}_1^{-1/2(t)}\,\mathbf{X}_1 \\ \boldsymbol{\Lambda}_2^{-1/2(t)}\,\mathbf{X}_2 \\ \vdots \\ \boldsymbol{\Lambda}_n^{-1/2(t)}\,\mathbf{X}_n \end{bmatrix}.$$

然后，在给定 $\mathbf{L}_1^{(t)}$和 $\mathbf{L}_2^{(t)}$的条件下，最大化 log $f_{\mathbf{y}\,|\,\boldsymbol{\theta}}(\mathbf{y})$关于 $\mathbf{b}$的值相当于最小化二次型 $(\breve{\mathbf{y}} - \breve{\mathbf{X}}\,\mathbf{b})^\top(\breve{\mathbf{y}} - \breve{\mathbf{X}}\,\mathbf{b})$关于 $\mathbf{b}$的值。我们通过将缩放Lasso惩罚项添加到此二次型中（如主文稿中方程(8)所示）来同时估计$\mathbf{B}_{(t+1)}$和 $\tau^{2(t+1)}$，以在 $\mathbf{B}^{(t+1)}$ 中强制稀疏性并减少估计 $\tau^{2(t+1)}$时的偏差。我们还更新 $\boldsymbol{\theta}^{(t)}$为 $\boldsymbol{\theta}^{(t+1/3)} = (\mathbf{B}^{(t+1)}, \mathbf{L}_1^{(t)}, \mathbf{L}_2^{(t)}, \tau^{2(t+1)})$。

2.2 Derivation of $\mathbf{L}_1^{(t+1)}$ and $\mathbf{L}_2^{(t+1)}$

The AECM algorithm treats the matrices $\mathbf{C}_i$ $(i = 1, \ldots, n)$ as missing data and considers the joint distributions of the observed and missing data $(\mathbf{y}_i, \mathbf{c}_{i(k)})$, for $i = 1, \ldots, n$ and $k = 1, 2$; see Equations (9), (10), and (11) in the main manuscript. The AECM algorithm then maximizes $\mathbb{E}[\log f_{\mathbf{y}\,|\,\mathbf{c},\boldsymbol{\theta}}(\mathbf{y})|\,\mathbf{y}, \boldsymbol{\theta}^{(t+1/3)}]$ with respect to $\mathbf{L}_1$, where $\log f_{\mathbf{y}\,|\,\mathbf{c},\boldsymbol{\theta}}(\mathbf{y})|\,\mathbf{y}, \boldsymbol{\theta}^{(t+1/3)}$ is the complete data log likelihood in Equation of (6) of the main manuscript; see also Section 3 and Equation (8) in the main manuscript.

Consider the estimation of $\mathbf{L}_1$ given $\boldsymbol{\theta}^{(t+1/3)}$ in the second cycle of the AECM algorithm. The E step in the second cycle computes the conditional expectation $\mathbb{E}[\log f_{\mathbf{y}\,|\,\mathbf{c},\boldsymbol{\theta}}(\mathbf{y})|\,\mathbf{y}, \boldsymbol{\theta}^{(t+1/3)}]$. The CM step in the second cycle maximizes this objective with respect to $\mathbf{L}_1$. This is equivalent to maximizing $\mathbb{E}[\log f_{\tilde{\mathbf{y}}|\,\mathbf{c},\boldsymbol{\theta}}(\tilde{\mathbf{y}})|\tilde{\mathbf{y}}, \boldsymbol{\theta}^{(t+1/3)}]$ with respect to $\mathbf{L}_1$, where $\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{X}\,\mathbf{b}^{(t+1)}$. The complete data log likelihood of $(\tilde{\mathbf{y}}, \mathbf{c})$ is

$$\log f_{\tilde{\mathbf{y}}|\,\mathbf{c},\boldsymbol{\theta}}(\tilde{\mathbf{y}}) = -\frac{N}{2}\log 2\pi\tau^2 - \frac{1}{2\tau^2}\sum_{i=1}^{n}(\tilde{\mathbf{y}}_i - \mathbf{Z}_{i(2)}\,\mathbf{L}_{(2)}\,\mathbf{c}_{i(2)})^\top(\tilde{\mathbf{y}}_i - \mathbf{Z}_{i(2)}\,\mathbf{L}_{(2)}\,\mathbf{c}_{i(2)}), \qquad (5)$$

where, as defined in Section 2.1 of the main manuscript, $\mathbf{Z}_{i(2)}$ is the matrix whose jth row is $\mathbf{z}_{ij(2)} = \mathrm{vec}(\mathbf{Z}_{ij(2)}) = \mathrm{vec}(\mathbf{Z}_{ij}^\top), \mathbf{L}_{(2)} = \mathbf{L}_1 \otimes \mathbf{L}_2$ and $\mathbf{c}_{i(2)} = \mathrm{vec}(\mathbf{C}_i^\top)$.

We derive the loss function in the E step of the second cycle. For $i = 1 \ldots n$, the joint distribution of $(\tilde{\mathbf{y}}_i, \mathbf{c}_{i(2)})$ is

$$\begin{bmatrix} \tilde{\mathbf{y}}_i \\ \mathbf{c}_{i(2)} \end{bmatrix} \sim N_{m_i+S_1S_2}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \tau^2 \begin{bmatrix} \mathbf{Z}_{i(2)}\,\mathbf{L}_{(2)}\,\mathbf{L}_{(2)}^\top\,\mathbf{Z}_{i(2)}^\top + \mathbf{I}_{m_i} & \mathbf{Z}_{i(2)}\,\mathbf{L}_{(2)} \\ \mathbf{L}_{(2)}^\top\,\mathbf{Z}_{i(2)}^\top & \mathbf{I}_{S_1S_2} \end{bmatrix}\right).$$

Using the analytic form of the conditional distribution of $\mathbf{c}_{i(2)}$ given $\tilde{\mathbf{y}}_i$ in (5), gives the loss function $-\mathcal{Q}_{(1)}$ for estimating $\mathbf{L}_1$. Specifically, the negative of the conditional expectation

2.2 $\mathbf{L}_1^{(t+1)}$ 和$\mathbf{L}_2^{(t+1)}$ 的推导

AECM算法将矩阵 $\mathbf{C}_i$ $(i = 1, \ldots, n)$ 视为缺失数据，并考虑观测数据和缺失数据的联合分布 $(\mathbf{y}_i, \mathbf{c}_{i(k)})$，对于 $i = 1 \ldots n$， ，和$k = 1\ 2$; ,参见主文稿中方程(9)、(10)和(11)。然后，AECM算法最大化 $\mathbb{E}[\log\ f_{\mathbf{y}\,|\,\mathbf{c}\,\boldsymbol{\theta}}(\mathbf{y})|\,\mathbf{y}\ \boldsymbol{\theta}^{(t+1/3)}]$.,关于$\mathbf{L}_1$，其中log $f_{\mathbf{y}\,|\,\mathbf{c}\,\boldsymbol{\theta}}(\mathbf{y})|\,\mathbf{y}\ \boldsymbol{\theta}^{(t+1/3)}$.,是主文稿中方程(6)的完整数据对数似然；另见主文稿第3节和方程(8)。

考虑AECM算法第二周期中给定 $\boldsymbol{\theta}^{(t+1/3)}$ 的$\mathbf{L}_1$ 估计。第二周期的E步计算条件期望 $\mathbb{E}[\log f_{\mathbf{y}\,|\,\mathbf{c},\boldsymbol{\theta}}(\mathbf{y})|\,\mathbf{y}, \boldsymbol{\theta}^{(t+1/3)}]$。第二周期的CM步最大化此目标关于$\mathbf{L}_1$。这相当于最大化 $\mathbb{E}[\log\ f_{\tilde{\mathbf{y}}|\,\mathbf{c}\,\boldsymbol{\theta}}(\tilde{\mathbf{y}})|\tilde{\mathbf{y}}\ \boldsymbol{\theta}^{(t+1/3)}]$.,关于 $\mathbf{L}_1$，其中 $\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{X}\,\mathbf{b}^{(t+1)}$。(˜y,c)的完整数据对数似然是

$$\log f_{\tilde{\mathbf{y}}|\,\mathbf{c},\boldsymbol{\theta}}(\tilde{\mathbf{y}}) = -\frac{N}{2}\log 2\pi\tau^2 - \frac{1}{2\tau^2}\sum_{i=1}^{n}(\tilde{\mathbf{y}}_i - \mathbf{Z}_{i(2)}\,\mathbf{L}_{(2)}\,\mathbf{c}_{i(2)})^\top(\tilde{\mathbf{y}}_i - \mathbf{Z}_{i(2)}\,\mathbf{L}_{(2)}\,\mathbf{c}_{i(2)}), \qquad (5)$$

在主文稿第2.1节中定义的位置， $\mathbf{Z}_{i(2)}$ 是矩阵，其第 j行是 $\mathbf{z}_{ij(2)} = \mathrm{vec}(\mathbf{Z}_{ij(2)}) = \mathrm{vec}(\mathbf{Z}_{ij}^\top)\ \mathbf{L}_{(2)} = \mathbf{L}_1 \otimes \mathbf{L}_2$ and $\mathbf{c}_{i(2)} = \mathrm{vec}(\mathbf{C}_i^\top)$,。

我们在第二周期的E步推导出损失函数。对于 $i = 1 \ldots n$， $(\tilde{\mathbf{y}}_i, \mathbf{c}_{i(2)})$ 的联合分布是

$$\begin{bmatrix} \tilde{\mathbf{y}}_i \\ \mathbf{c}_{i(2)} \end{bmatrix} \sim N_{m_i+S_1S_2}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \tau^2 \begin{bmatrix} \mathbf{Z}_{i(2)}\,\mathbf{L}_{(2)}\,\mathbf{L}_{(2)}^\top\,\mathbf{Z}_{i(2)}^\top + \mathbf{I}_{m_i} & \mathbf{Z}_{i(2)}\,\mathbf{L}_{(2)} \\ \mathbf{L}_{(2)}^\top\,\mathbf{Z}_{i(2)}^\top & \mathbf{I}_{S_1S_2} \end{bmatrix}\right).$$

使用(5)中给出的条件分布$\mathbf{c}_{i(2)}$ 关于$\mathbf{y}_i$ 的分析形式，得到用于估计$\mathbf{L}_1$的损失函数 $-\mathcal{Q}_{(1)}$ 。具体来说，是条件期望的负值

of the term in (5) that depends on $\mathbf{L}_1$ is

$$-\mathcal{Q}_{(1)}(\mathbf{L}_1) = \mathbb{E}\left[\sum_{i=1}^{n}(\mathbf{c}_{i(2)}^\top \mathbf{L}_{(2)}^\top \mathbf{Z}_{i(2)}^\top \mathbf{Z}_{i(2)} \mathbf{L}_{(2)} \mathbf{c}_{i(2)} - 2\tilde{\mathbf{y}}_i^\top \mathbf{Z}_{i(2)} \mathbf{L}_{(2)} \mathbf{c}_{i(2)})|\tilde{\mathbf{y}}, \boldsymbol{\theta}^{(t+1/3)}\right],$$

and maximizing $\mathbb{E}[\log f_{\tilde{\mathbf{y}}|\,\mathbf{c},\boldsymbol{\theta}}(\tilde{\mathbf{y}})|\tilde{\mathbf{y}}, \boldsymbol{\theta}^{(t+1/3)}]$ with respect to $\mathbf{L}_1$ amounts to minimizing $-\mathcal{Q}_{(1)}(\mathbf{L}_1)$ with respect to $\mathbf{L}_1$. Let

$$\begin{aligned}
\boldsymbol{\mu}_{i(2)}^{(t)} &= \mathbb{E}(\mathbf{c}_{i(2)}\,|\tilde{\mathbf{y}}_i, \boldsymbol{\theta}^{(t+1/3)}) &&= \mathbf{L}_{(2)}^{(t)\top} \mathbf{Z}_{i(2)}^\top (\mathbf{Z}_{i(2)} \mathbf{L}_{(2)}^{(t)} \mathbf{L}_{(2)}^{(t)\top} \mathbf{Z}_{i(2)}^\top + \mathbf{I}_{m_i})^{-1} \tilde{\mathbf{y}}_i, \\
\boldsymbol{\Sigma}_{i(2)}^{(t)} &= \mathrm{Cov}(\mathbf{c}_{i(2)}\,|\tilde{\mathbf{y}}_i, \boldsymbol{\theta}^{(t+1/3)}) &&= \tau^{2(t+1)} (\mathbf{I}_{S_1 S_2} + \mathbf{L}_{(2)}^{(t)\top} \mathbf{Z}_{i(2)}^\top \mathbf{Z}_{i(2)} \mathbf{L}_{(2)}^{(t)})^{-1}, \\
\boldsymbol{\Gamma}_{i(2)}^{(t)} &= \mathbb{E}(\mathbf{c}_{i(2)}\, \mathbf{c}_{i(2)}^\top |\tilde{\mathbf{y}}_i, \boldsymbol{\theta}^{(t+1/3)}) &&= \boldsymbol{\Sigma}_{i(2)}^{(t)} + \boldsymbol{\mu}_{i(2)}^{(t)}\, \boldsymbol{\mu}_{i(2)}^{(t)\top}.
\end{aligned}$$

Then, we can reexpress $-\mathcal{Q}_{(1)}(\mathbf{L}_1)$ as

$$\begin{aligned}
-\mathcal{Q}_{(1)}(\mathbf{L}_1) &= \mathbb{E}\left\{\sum_{i=1}^{n}\left[\mathrm{tr}(\mathbf{L}_{(2)}^\top \mathbf{Z}_{i(2)}^\top \mathbf{Z}_{i(2)} \mathbf{L}_{(2)} \mathbf{c}_{i(2)} \mathbf{c}_{i(2)}^\top) - 2\tilde{\mathbf{y}}_i^\top \mathbf{Z}_{i(2)} \mathbf{L}_{(2)} \mathbf{c}_{i(2)}\right] |\tilde{\mathbf{y}}, \boldsymbol{\theta}^{(t+1/3)}\right\} \\
&= \sum_{i=1}^{n}\left[\mathrm{tr}(\mathbf{L}_{(2)}^\top \mathbf{Z}_{i(2)}^\top \mathbf{Z}_{i(2)} \mathbf{L}_{(2)} \boldsymbol{\Gamma}_{i(2)}^{(t)}) - 2\tilde{\mathbf{y}}_i^\top \mathbf{Z}_{i(2)} \mathbf{L}_{(2)} \boldsymbol{\mu}_{i(2)}^{(t)}\right] \\
&= \sum_{i=1}^{n}\left[\mathrm{tr}(\mathbf{Z}_{i(2)} \mathbf{L}_{(2)} \boldsymbol{\Gamma}_{i(2)}^{(t)} \mathbf{L}_{(2)}^\top \mathbf{Z}_{i(2)}^\top) - 2\,\boldsymbol{\mu}_{i(2)}^{(t)\top} \mathbf{L}_{(2)}^\top \mathbf{Z}_{i(2)}^\top \tilde{\mathbf{y}}_i\right]. \qquad (6)
\end{aligned}$$

We simplify this expression by reexpressing the first component of this sum, $\mathrm{tr}(\mathbf{Z}_{i(2)} \mathbf{L}_{(2)} \boldsymbol{\Gamma}_{i(2)}^{(t)} \mathbf{L}_{(2)}^\top \mathbf{Z}_{i(2)}^\top)$.

在(5)中的术语中，取决于 $\mathbf{L}_1$ 的项是

$$-\mathcal{Q}_{(1)}(\mathbf{L}_1) = \mathbb{E}\left[\sum_{i=1}^{n}(\mathbf{c}_{i(2)}^\top \mathbf{L}_{(2)}^\top \mathbf{Z}_{i(2)}^\top \mathbf{Z}_{i(2)} \mathbf{L}_{(2)} \mathbf{c}_{i(2)} - 2\tilde{\mathbf{y}}_i^\top \mathbf{Z}_{i(2)} \mathbf{L}_{(2)} \mathbf{c}_{i(2)})|\tilde{\mathbf{y}}, \boldsymbol{\theta}^{(t+1/3)}\right],$$

并且最大化 $\mathbb{E}[\log f_{\tilde{\mathbf{y}}|\,\mathbf{c},\boldsymbol{\theta}}(\tilde{\mathbf{y}})|\tilde{\mathbf{y}}, \boldsymbol{\theta}^{(t+1/3)}]$ 关于 $\mathbf{L}_1$ 的量相当于最小化 $-\mathcal{Q}_{(1)}(\mathbf{L}_1)$ 关于 $\mathbf{L}_1$。令

$$\begin{aligned}
\boldsymbol{\mu}_{i(2)}^{(t)} &= \mathbb{E}(\mathbf{c}_{i(2)}\,|\tilde{\mathbf{y}}_i, \boldsymbol{\theta}^{(t+1/3)}) &&= \mathbf{L}_{(2)}^{(t)\top} \mathbf{Z}_{i(2)}^\top (\mathbf{Z}_{i(2)} \mathbf{L}_{(2)}^{(t)} \mathbf{L}_{(2)}^{(t)\top} \mathbf{Z}_{i(2)}^\top + \mathbf{I}_{m_i})^{-1} \tilde{\mathbf{y}}_i, \\
\boldsymbol{\Sigma}_{i(2)}^{(t)} &= \mathrm{Cov}(\mathbf{c}_{i(2)}\,|\tilde{\mathbf{y}}_i, \boldsymbol{\theta}^{(t+1/3)}) &&= \tau^{2(t+1)} (\mathbf{I}_{S_1 S_2} + \mathbf{L}_{(2)}^{(t)\top} \mathbf{Z}_{i(2)}^\top \mathbf{Z}_{i(2)} \mathbf{L}_{(2)}^{(t)})^{-1}, \\
\boldsymbol{\Gamma}_{i(2)}^{(t)} &= \mathbb{E}(\mathbf{c}_{i(2)}\, \mathbf{c}_{i(2)}^\top |\tilde{\mathbf{y}}_i, \boldsymbol{\theta}^{(t+1/3)}) &&= \boldsymbol{\Sigma}_{i(2)}^{(t)} + \boldsymbol{\mu}_{i(2)}^{(t)}\, \boldsymbol{\mu}_{i(2)}^{(t)\top}.
\end{aligned}$$

然后，我们可以重新表示 $-\mathcal{Q}_{(1)}(\mathbf{L}_1)$ 为

$$\begin{aligned}
-\mathcal{Q}_{(1)}(\mathbf{L}_1) &= \mathbb{E}\left\{\sum_{i=1}^{n}\left[\mathrm{tr}(\mathbf{L}_{(2)}^\top \mathbf{Z}_{i(2)}^\top \mathbf{Z}_{i(2)} \mathbf{L}_{(2)} \mathbf{c}_{i(2)} \mathbf{c}_{i(2)}^\top) - 2\tilde{\mathbf{y}}_i^\top \mathbf{Z}_{i(2)} \mathbf{L}_{(2)} \mathbf{c}_{i(2)}\right] |\tilde{\mathbf{y}}, \boldsymbol{\theta}^{(t+1/3)}\right\} \\
&= \sum_{i=1}^{n}\left[\mathrm{tr}(\mathbf{L}_{(2)}^\top \mathbf{Z}_{i(2)}^\top \mathbf{Z}_{i(2)} \mathbf{L}_{(2)} \boldsymbol{\Gamma}_{i(2)}^{(t)}) - 2\tilde{\mathbf{y}}_i^\top \mathbf{Z}_{i(2)} \mathbf{L}_{(2)} \boldsymbol{\mu}_{i(2)}^{(t)}\right] \\
&= \sum_{i=1}^{n}\left[\mathrm{tr}(\mathbf{Z}_{i(2)} \mathbf{L}_{(2)} \boldsymbol{\Gamma}_{i(2)}^{(t)} \mathbf{L}_{(2)}^\top \mathbf{Z}_{i(2)}^\top) - 2\,\boldsymbol{\mu}_{i(2)}^{(t)\top} \mathbf{L}_{(2)}^\top \mathbf{Z}_{i(2)}^\top \tilde{\mathbf{y}}_i\right]. \qquad (6)
\end{aligned}$$

我们通过重新表示此求和的第一分量， $\mathrm{tr}(\mathbf{Z}_{i(2)} \mathbf{L}_{(2)} \boldsymbol{\Gamma}_{i(2)}^{(t)} \mathbf{L}_{(2)}^\top \mathbf{Z}_{i(2)}^\top)$，来简化此表达式。

If $\mathbf{l}_1 = \text{vec}(\mathbf{L}_1)$, then

$$\begin{aligned}
\mathbf{L}_{(2)}^\top \mathbf{Z}_{i(2)}^\top &= \mathbf{L}_{(2)}^\top \begin{bmatrix} \mathbf{z}_{i1(2)} & \mathbf{z}_{i2(2)} & \dots & \mathbf{z}_{im_i(2)} \end{bmatrix} \\
&= \begin{bmatrix} (\mathbf{L}_1^\top \otimes \mathbf{L}_2^\top)\, \mathbf{z}_{i1(2)} & (\mathbf{L}_1^\top \otimes \mathbf{L}_2^\top)\, \mathbf{z}_{i2(2)} & \dots & (\mathbf{L}_1^\top \otimes \mathbf{L}_2^\top)\, \mathbf{z}_{im_i(2)} \end{bmatrix} \\
&= \begin{bmatrix} \text{vec}(\mathbf{L}_2^\top \mathbf{Z}_{i1(2)}\, \mathbf{L}_1) & \text{vec}(\mathbf{L}_2^\top \mathbf{Z}_{i2(2)}\, \mathbf{L}_1) & \dots & \text{vec}(\mathbf{L}_2^\top \mathbf{Z}_{im_i(2)}\, \mathbf{L}_1) \end{bmatrix} \\
&= \begin{bmatrix} \text{vec}(\mathbf{L}_2^\top \mathbf{Z}_{i1(1)}^\top\, \mathbf{L}_1) & \text{vec}(\mathbf{L}_2^\top \mathbf{Z}_{i2(1)}^\top\, \mathbf{L}_1) & \dots & \text{vec}(\mathbf{L}_2^\top \mathbf{Z}_{im_i(1)}^\top\, \mathbf{L}_1) \end{bmatrix}, \\
\text{tr}(\mathbf{Z}_{i(2)}\, \mathbf{L}_{(2)}\, \boldsymbol{\Gamma}_{i(2)}^{(t)}\, \mathbf{L}_{(2)}^\top\, \mathbf{Z}_{i(2)}^\top) &= \sum_{j=1}^{m_i} \text{vec}(\mathbf{L}_2^\top\, \mathbf{Z}_{ij(1)}^\top\, \mathbf{L}_1)^\top\, \boldsymbol{\Gamma}_{i(2)}^{(t)}\, \text{vec}(\mathbf{L}_2^\top\, \mathbf{Z}_{ij(1)}^\top\, \mathbf{L}_1) \\
&= \sum_{j=1}^{m_i} \text{vec}(\mathbf{L}_1)^\top (\mathbf{I}_{S_1} \otimes \mathbf{Z}_{ij(1)}\, \mathbf{L}_2)\, \boldsymbol{\Gamma}_{i(2)}^{(t)} (\mathbf{I}_{S_1} \otimes \mathbf{L}_2^\top\, \mathbf{Z}_{ij(1)}^\top) \text{vec}(\mathbf{L}_1) \\
&= \mathbf{l}_1^\top \left[\sum_{j=1}^{m_i} (\mathbf{I}_{S_1} \otimes \mathbf{Z}_{ij(1)}\, \mathbf{L}_2)\, \boldsymbol{\Gamma}_{i(2)}^{(t)} (\mathbf{I}_{S_1} \otimes \mathbf{L}_2^\top\, \mathbf{Z}_{ij(1)}^\top) \right] \mathbf{l}_1 \,.
\end{aligned}$$

Next, we pull $\mathbf{L}_1$ out of the second component of (6), $\boldsymbol{\mu}_{i(2)}^{(t)\top}\, \mathbf{L}_{(2)}^\top\, \mathbf{Z}_{i(2)}^\top\, \tilde{\mathbf{y}}_i$ to obtain that

$$\begin{aligned}
\mathbf{Z}_{i(2)}^\top\, \tilde{\mathbf{y}}_i &= \sum_{j=1}^{m_i} \tilde{y}_{ij}\, \mathbf{z}_{ij(2)}, \\
\mathbf{L}_{(2)}^\top\, \mathbf{Z}_{i(2)}^\top\, \tilde{\mathbf{y}}_i &= (\mathbf{L}_1^\top \otimes \mathbf{L}_2^\top) \sum_{j=1}^{m_i} \tilde{y}_{ij}\, \mathbf{z}_{ij(2)} \\
&= \text{vec} \left[\mathbf{L}_2^\top (\sum_{j=1}^{m_i} \tilde{y}_{ij}\, \mathbf{Z}_{ij(2)})\, \mathbf{L}_1 \right] \\
&= \text{vec} \left[\mathbf{L}_2^\top (\sum_{j=1}^{m_i} \tilde{y}_{ij}\, \mathbf{Z}_{ij(1)}^\top)\, \mathbf{L}_1 \right] \\
&= \left[\mathbf{I}_{S_1} \otimes \mathbf{L}_2^\top (\sum_{j=1}^{m_i} \tilde{y}_{ij}\, \mathbf{Z}_{ij(1)}^\top) \right] \text{vec}(\mathbf{L}_1), \\
\boldsymbol{\mu}_{i(2)}^{(t)\top}\, \mathbf{L}_{(2)}^\top\, \mathbf{Z}_{i(2)}^\top\, \tilde{\mathbf{y}}_i &= \boldsymbol{\mu}_{i(2)}^{(t)\top} \left[\mathbf{I}_{S_1} \otimes \mathbf{L}_2^\top (\sum_{j=1}^{m_i} \tilde{y}_{ij}\, \mathbf{Z}_{ij(1)}^\top) \right] \mathbf{l}_1 \,.
\end{aligned}$$

When estimating $\mathbf{L}_1$, we condition on $\mathbf{L}_2^{(t)}$, thus we replace $\mathbf{L}_2$ with $\mathbf{L}_2^{(t)}$ in our expression

如果 $\mathbf{l}_1 = \text{vec}(\mathbf{L}_1)$，那么

$$\begin{aligned}
\mathbf{L}_{(2)}^\top \mathbf{Z}_{i(2)}^\top &= \mathbf{L}_{(2)}^\top \begin{bmatrix} \mathbf{z}_{i1(2)} & \mathbf{z}_{i2(2)} & \dots & \mathbf{z}_{im_i(2)} \end{bmatrix} \\
&= \begin{bmatrix} (\mathbf{L}_1^\top \otimes \mathbf{L}_2^\top)\, \mathbf{z}_{i1(2)} & (\mathbf{L}_1^\top \otimes \mathbf{L}_2^\top)\, \mathbf{z}_{i2(2)} & \dots & (\mathbf{L}_1^\top \otimes \mathbf{L}_2^\top)\, \mathbf{z}_{im_i(2)} \end{bmatrix} \\
&= \begin{bmatrix} \text{vec}(\mathbf{L}_2^\top \mathbf{Z}_{i1(2)}\, \mathbf{L}_1) & \text{vec}(\mathbf{L}_2^\top \mathbf{Z}_{i2(2)}\, \mathbf{L}_1) & \dots & \text{vec}(\mathbf{L}_2^\top \mathbf{Z}_{im_i(2)}\, \mathbf{L}_1) \end{bmatrix} \\
&= \begin{bmatrix} \text{vec}(\mathbf{L}_2^\top \mathbf{Z}_{i1(1)}^\top\, \mathbf{L}_1) & \text{vec}(\mathbf{L}_2^\top \mathbf{Z}_{i2(1)}^\top\, \mathbf{L}_1) & \dots & \text{vec}(\mathbf{L}_2^\top \mathbf{Z}_{im_i(1)}^\top\, \mathbf{L}_1) \end{bmatrix}, \\
\text{tr}(\mathbf{Z}_{i(2)}\, \mathbf{L}_{(2)}\, \boldsymbol{\Gamma}_{i(2)}^{(t)}\, \mathbf{L}_{(2)}^\top\, \mathbf{Z}_{i(2)}^\top) &= \sum_{j=1}^{m_i} \text{vec}(\mathbf{L}_2^\top\, \mathbf{Z}_{ij(1)}^\top\, \mathbf{L}_1)^\top\, \boldsymbol{\Gamma}_{i(2)}^{(t)}\, \text{vec}(\mathbf{L}_2^\top\, \mathbf{Z}_{ij(1)}^\top\, \mathbf{L}_1) \\
&= \sum_{j=1}^{m_i} \text{vec}(\mathbf{L}_1)^\top (\mathbf{I}_{S_1} \otimes \mathbf{Z}_{ij(1)}\, \mathbf{L}_2)\, \boldsymbol{\Gamma}_{i(2)}^{(t)} (\mathbf{I}_{S_1} \otimes \mathbf{L}_2^\top\, \mathbf{Z}_{ij(1)}^\top) \text{vec}(\mathbf{L}_1) \\
&= \mathbf{l}_1^\top \left[\sum_{j=1}^{m_i} (\mathbf{I}_{S_1} \otimes \mathbf{Z}_{ij(1)}\, \mathbf{L}_2)\, \boldsymbol{\Gamma}_{i(2)}^{(t)} (\mathbf{I}_{S_1} \otimes \mathbf{L}_2^\top\, \mathbf{Z}_{ij(1)}^\top) \right] \mathbf{l}_1 \,.
\end{aligned}$$

接下来，我们从(6)的第二分量中提取 $\mathbf{L}_1$， $\boldsymbol{\mu}_{i(2)}^{(t)\top}\, \mathbf{L}_{(2)}^\top\, \mathbf{Z}_{i(2)}^\top\, \tilde{\mathbf{y}}_i$ 得到

$$\begin{aligned}
\mathbf{Z}_{i(2)}^\top\, \tilde{\mathbf{y}}_i &= \sum_{j=1}^{m_i} \tilde{y}_{ij}\, \mathbf{z}_{ij(2)}, \\
\mathbf{L}_{(2)}^\top\, \mathbf{Z}_{i(2)}^\top\, \tilde{\mathbf{y}}_i &= (\mathbf{L}_1^\top \otimes \mathbf{L}_2^\top) \sum_{j=1}^{m_i} \tilde{y}_{ij}\, \mathbf{z}_{ij(2)} \\
&= \text{vec} \left[\mathbf{L}_2^\top (\sum_{j=1}^{m_i} \tilde{y}_{ij}\, \mathbf{Z}_{ij(2)})\, \mathbf{L}_1 \right] \\
&= \text{vec} \left[\mathbf{L}_2^\top (\sum_{j=1}^{m_i} \tilde{y}_{ij}\, \mathbf{Z}_{ij(1)}^\top)\, \mathbf{L}_1 \right] \\
&= \left[\mathbf{I}_{S_1} \otimes \mathbf{L}_2^\top (\sum_{j=1}^{m_i} \tilde{y}_{ij}\, \mathbf{Z}_{ij(1)}^\top) \right] \text{vec}(\mathbf{L}_1), \\
\boldsymbol{\mu}_{i(2)}^{(t)\top}\, \mathbf{L}_{(2)}^\top\, \mathbf{Z}_{i(2)}^\top\, \tilde{\mathbf{y}}_i &= \boldsymbol{\mu}_{i(2)}^{(t)\top} \left[\mathbf{I}_{S_1} \otimes \mathbf{L}_2^\top (\sum_{j=1}^{m_i} \tilde{y}_{ij}\, \mathbf{Z}_{ij(1)}^\top) \right] \mathbf{l}_1 \,.
\end{aligned}$$

在估计 $\mathbf{L}_1$ 时，我们基于 $\mathbf{L}_2^{(t)}$，因此我们在表达式中用 $\mathbf{L}_2^{(t)}$ 替换 $\mathbf{L}_2$

for $-\mathcal{Q}_{(1)}(\mathbf{L}_1)$. Let

$$\mathbf{H}_{(1)}^{(t)} = \sum_{i=1}^{n}\sum_{j=1}^{m_i}(\mathbf{I}_{S_1}\otimes\mathbf{Z}_{ij(1)}\,\mathbf{L}_2^{(t)})\,\mathbf{\Gamma}_{i(2)}^{(t)}(\mathbf{I}_{S_1}\otimes\mathbf{L}_2^{(t)\top}\,\mathbf{Z}_{ij(1)}^{\top}),$$
$$\mathbf{g}_{(1)}^{(t)} = \sum_{i=1}^{n}\left[\mathbf{I}_{S_1}\otimes(\sum_{j=1}^{m_i}\tilde{y}_{ij}\,\mathbf{Z}_{ij(1)})\,\mathbf{L}_2^{(t)}\right]\boldsymbol{\mu}_{i(2)}^{(t)},$$

then $-\mathcal{Q}_{(1)}(\mathbf{L}_1) = \mathbf{l}_1^{\top}\,\mathbf{H}_{(1)}^{(t)}\,\mathbf{l}_1 - 2\,\mathbf{g}_{(1)}^{(t)\top}\,\mathbf{l}_1$, which is expressed as the quadratic form

$$-\mathcal{Q}_{(1)}(\mathbf{L}_1) \propto (\mathbf{H}_{(1)}^{-1/2(t)}\,\mathbf{g}_{(1)}^{(t)} - \mathbf{H}_{(1)}^{1/2(t)\top}\,\mathbf{l}_1)^{\top}(\mathbf{H}_{(1)}^{-1/2(t)}\,\mathbf{g}_{(1)}^{(t)} - \mathbf{H}_{(1)}^{1/2(t)\top}\,\mathbf{l}_1), \tag{7}$$

where $\mathbf{H}_{(1)}^{1/2(t)}$ is any matrix such that $\mathbf{H}_{(1)}^{1/2(t)}\,\mathbf{H}_{(1)}^{1/2(t)\top} = \mathbf{H}_{(1)}^{(t)}$ and $\mathbf{H}_{(1)}^{-1/2(t)}$ is any generalized inverse of $\mathbf{H}_{(1)}^{1/2(t)}$. It can be shown that $\mathbf{g}_{(1)}^{(t)}$ is in the column space of $\mathbf{H}_{(1)}^{(t)}$, and thus minimizing $-\mathcal{Q}_{(1)}(\mathbf{L}_1)$ with respect to $\mathbf{l}_1$ is equivalent to minimizing (7) with respect to $\mathbf{l}_1$. This statement remains true even when $\mathbf{H}_{(1)}^{(t)}$ is a singular matrix. We further add to (7) a group lasso penalty as in Equation (10) in the main manuscript to enforce rank constraints on our estimate of $\mathbf{L}_1$. This completes the derivation of the objective in Equation (10) of the main manuscript.

The objective for estimation $\mathbf{L}_2^{(t+1)}$ in Equation (11) of the main manuscript is obtained by swapping indices 1 and 2 in the derivation above for $\mathbf{L}_1$ updates.

3 MMTR Misspecification Plots

We present additional plots for each of the simulation scenarios described in Section 4.2 showing how the relative errors in estimating $\mathbf{\Lambda}$ by MMTR change depending on the GEE's equicorrelation model parameter α (Figures 1 and 2). The α term in the GEE's equicorrelation model determines the correlation between all pairs of responses within each group of observations. The greater the α value, the more misspecified the model is to the MMTR setting, which assumes the covariance between pairs of responses in each group to be a

对于 $-\mathcal{Q}_{(1)}(\mathbf{L}_1)$。令

$$\mathbf{H}_{(1)}^{(t)} = \sum_{i=1}^{n}\sum_{j=1}^{m_i}(\mathbf{I}_{S_1}\otimes\mathbf{Z}_{ij(1)}\,\mathbf{L}_2^{(t)})\,\mathbf{\Gamma}_{i(2)}^{(t)}(\mathbf{I}_{S_1}\otimes\mathbf{L}_2^{(t)\top}\,\mathbf{Z}_{ij(1)}^{\top}),$$
$$\mathbf{g}_{(1)}^{(t)} = \sum_{i=1}^{n}\left[\mathbf{I}_{S_1}\otimes(\sum_{j=1}^{m_i}\tilde{y}_{ij}\,\mathbf{Z}_{ij(1)})\,\mathbf{L}_2^{(t)}\right]\boldsymbol{\mu}_{i(2)}^{(t)},$$

则 $-\mathcal{Q}_{(1)}(\mathbf{L}_1) = \mathbf{l}_1^{\top}\,\mathbf{H}_{(1)}^{(t)}\,\mathbf{l}_1 - 2\,\mathbf{g}_{(1)}^{(t)\top}\,\mathbf{l}_1$，其表示为二次型

$$-\mathcal{Q}_{(1)}(\mathbf{L}_1) \propto (\mathbf{H}_{(1)}^{-1/2(t)}\,\mathbf{g}_{(1)}^{(t)} - \mathbf{H}_{(1)}^{1/2(t)\top}\,\mathbf{l}_1)^{\top}(\mathbf{H}_{(1)}^{-1/2(t)}\,\mathbf{g}_{(1)}^{(t)} - \mathbf{H}_{(1)}^{1/2(t)\top}\,\mathbf{l}_1), \tag{7}$$

其中 $\mathbf{H}_{(1)}^{1/2(t)}$ 是任何矩阵，使得 $\mathbf{H}_{(1)}^{1/2(t)}\,\mathbf{H}_{(1)}^{1/2(t)\top} = \mathbf{H}_{(1)}^{(t)}$ 和 $\mathbf{H}_{-1/2(t)\ (1)}$ 是 $\mathbf{H}_{(1)}^{1/2(t)}$ 的任何广义逆。可以证明 $\mathbf{g}_{(1)}^{(t)}$ 在 $\mathbf{H}_{(1)}^{(t)}$ 的列空间中，因此最小化 $-\mathcal{Q}_{(1)}(\mathbf{L}_1)$ 关于 $\mathbf{l}_1$ 等价于最小化 (7) 关于 $\mathbf{l}_1$。即使 $\mathbf{H}_{(1)}^{(t)}$ 是奇异矩阵，这一陈述仍然成立。我们在 (7) 中进一步添加一个组Lasso惩罚，如主文稿中方程 (10) 所示，以对 $\mathbf{L}_1$ 的估计施加秩约束。这完成了主文稿中方程 (10) 的目标的推导。

主文稿中方程（11）的目标 $\mathbf{L}_2^{(t+1)}$ 是通过在上面的推导中交换索引 1 和 2 以获得 $\mathbf{L}_1$ 更新而获得的。

3 MMTR 未指定图

我们展示了每个模拟场景（见第 4.2 节）的附加图，显示了 MMTR 估计 $\mathbf{\Lambda}$ 的相对误差如何根据 GEE 的等相关性模型参数 α （图 1 和图 2）而变化。GEE 的等相关性模型中的 α 项确定了每个观测组内所有响应对之间的相关性。α 值越大，模型对 MMTR 设置就越未指定，MMTR 假设每个组中响应对之间的协方差为

function of that group's random-effects covariates. These plots demonstrate that the more misspecified the correlation structure is from a random-effects setting, the worse MMTR will perform in estimating the true covariance matrix.

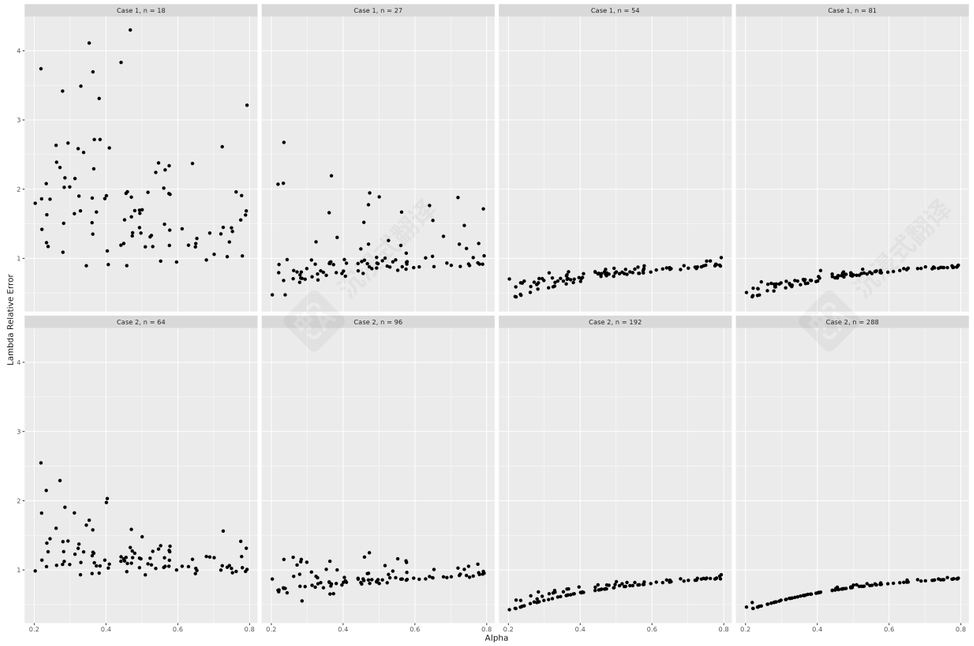

Figure 1: Relative errors in estimating $\mathbf{\Lambda}$ by MMTR against α. All plots here have the same range on the y axis.

该组随机效应协变量的函数。这些图表明，从随机效应设置来看，相关结构越未指定，MMTR 在估计真实协方差矩阵方面的表现就越差。

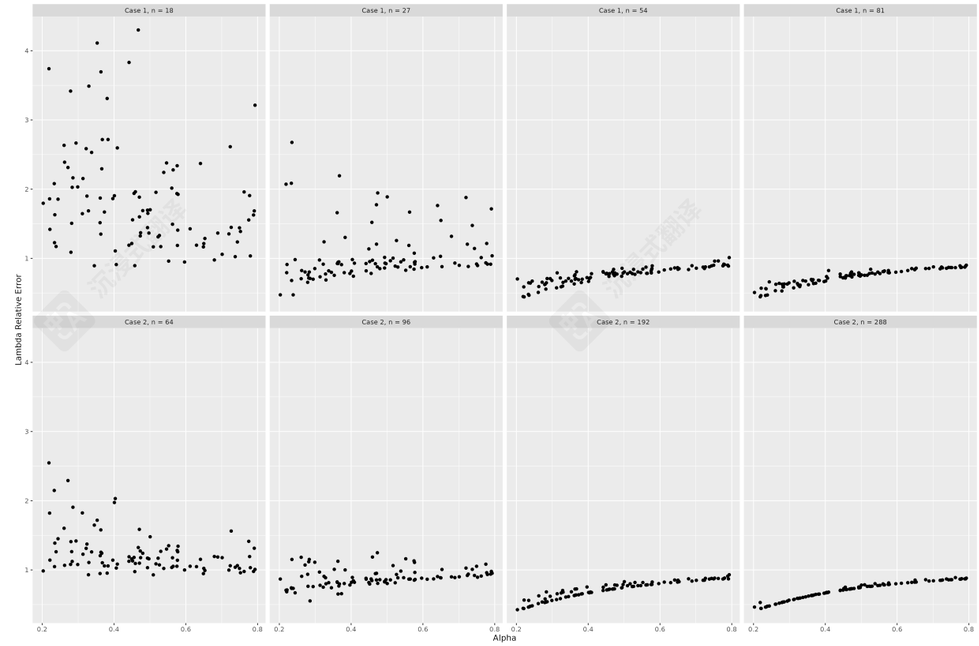

图 1：MMTR 估计 $\mathbf{\Lambda}$ 相对误差与 α 的关系。这里所有图在 y 轴上的范围相同。

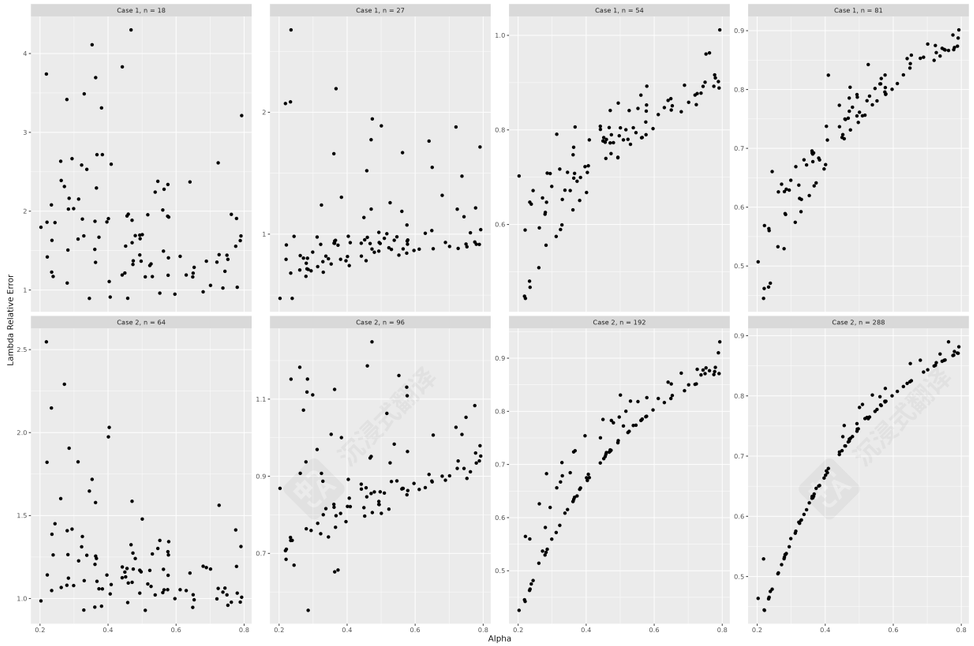

Figure 2: Relative errors in estimating $\mathbf{\Lambda}$ by MMTR against α. These plots are zoomed in from Figure 1.

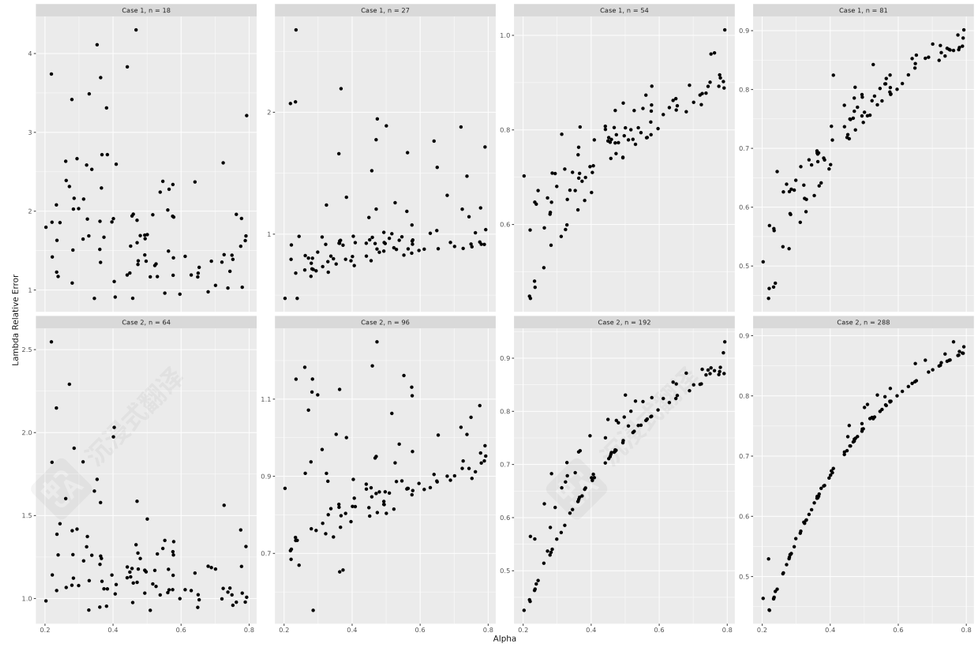

图2：使用MMTR估计 $\mathbf{\Lambda}$ 相对于 α的相对误差。这些图是从图1中放大的。