

A Bayesian Approach for Robust Longitudinal Envelope Models

Peng Zeng and Yushan Mu

Department of Mathematics and Statistics, Auburn University

December 9, 2025

Abstract

The envelope model provides a dimension-reduction framework for multivariate linear regression. However, existing envelope methods typically assume normally distributed random errors and do not accommodate repeated measures in longitudinal studies. To address these limitations, we propose the robust longitudinal envelope model (RoLEM). RoLEM employs a scale mixture of matrix-variate normal distributions to model random errors, allowing it to handle potential outliers, and incorporates flexible correlation structures for repeated measurements. In addition, we introduce new prior and proposal distributions on the Grassmann manifold to facilitate Bayesian inference for RoLEM. Simulation studies and real data analysis demonstrate the superior performance of the proposed method.

1 Introduction

In longitudinal studies, multivariate responses are often collected repeatedly from the same subjects over time. Beyond modeling the means and variances of these responses, it is crucial to account for the correlations of the same response across time as well as the correlations among different responses at a given time point. As the number of responses increases, the number of parameters grows quadratically. Efficient modeling strategies are therefore essential for capturing and interpreting these complex dependence structures.

The envelope model is a dimension-reduction framework for multivariate linear regression [5]. It assumes that the response vector can be decomposed into two components: one that depends on the covariates and one that does not. Estimation efficiency can be improved by identifying the component associated with the covariates. The envelope method has been extended to numerous settings [6, 8, 9, 4, 17]. A brief review is provided in Section 2.1.

Existing envelope models, which typically assume independence among observations, are not directly applicable to longitudinal studies. By ignoring within-subject correlation, these models may yield inaccurate estimates of variability and lead to flawed statistical inference. More recently, a mixed-effects envelope model has been proposed to incorporate random effects, and an EM algorithm has been developed for model fitting [19].

Bayesian methods for envelope models with independent observations have been studied previously [16, 2], which represent two different parameterizations of the covariance matrix; see Section 2.1 for details. To this end, we aim to develop a Bayesian framework for envelope models to analyze longitudinal data. Additionally, we introduce a new prior distribution that facilitates the incorporation of existing knowledge, together with a novel proposal distribution that enables efficient sampling within a Metropolis-Hastings algorithm.

The assumption of normal random errors is commonly adopted in existing envelope models for theoretical and computational convenience. However, this assumption makes the models sensitive to outliers and heavy-tailed error distributions. There is therefore a need to relax the assumption of normal random errors to enhance the robustness and flexibility of envelope models.

In this paper, we propose a novel Bayesian approach for robust longitudinal envelope models (RoLEM). This approach has three key features: it accommodates various correlation structures for repeated measures, it relaxes the assumption of normal random errors by employing a scale mixture of matrix-variate normal distributions to achieve robustness, and it introduces a flexible prior that readily incorporates existing knowledge. Simulation studies and real data analysis demonstrate the superior performance of the proposed RoLEM. The most comparable existing method is the mixed-effects envelope model [19]. However, it is equivalent to RoLEM only under a compound-symmetric correlation structure, and it neither extends to other correlation structures nor accounts for potential outliers.

The remainder of the paper is organized as follows. Section 2 reviews envelope models and related distributions. Section 3 presents the model specification in detail and derives the posterior sampling algorithm. Section 4 discusses implementation issues, including a Metropolis-Hastings algorithm, choices of hyperparameters, initialization, model selection, and software. Section 5 illustrates the proposed methods through simulation studies and a real data analysis. Section 6 concludes the paper with additional discussion. Additional proofs, tables, and figures are provided in the supplementary material.

2 Review of Envelope Models and Related Distributions

This section provides a brief review of relevant literature. Section 2.1 reviews envelope models; Section 2.2 reviews scale mixtures of matrix-variate normal distributions; and Section 2.3 reviews distributions on the Grassmann manifold.

2.1 Envelope models

Consider the multivariate linear regression model

$$y = \alpha + \beta x + \varepsilon,$$

where $y \in \mathbb{R}^r$ is the response vector, $x \in \mathbb{R}^p$ is the covariate, $\alpha \in \mathbb{R}^r$ and $\beta \in \mathbb{R}^{r \times p}$ are unknown parameters, and $\varepsilon \in \mathbb{R}^r$ is a random error. For an $r \times r$ orthogonal matrix (Γ, Γ_0) , where $\Gamma \in \mathbb{R}^{r \times u}$ and $\Gamma_0 \in \mathbb{R}^{r \times (r-u)}$, the envelope model [5] assumes that (a) $\Gamma_0^T y \mid x \sim \Gamma_0^T y$, where \sim denotes identical distribution, and (b) $\Gamma^T y$ is uncorrelated with $\Gamma_0^T y$ given x . This implies that the response vector y can be decomposed into two components: one that depends on the covariates x and one that does not. Under this assumption, the regression coefficient β and the error covariance matrix $\Sigma_\varepsilon = \text{var}(\varepsilon) \in \mathbb{R}^{r \times r}$ are structured as

$$\beta = \Gamma \eta, \quad \Sigma_\varepsilon = \Gamma \Omega \Gamma^T + \Gamma_0 \Omega_0 \Gamma_0^T,$$

where $\eta \in \mathbb{R}^{u \times p}$, $\Omega \in \mathbb{R}^{u \times u}$, and $\Omega_0 \in \mathbb{R}^{(r-u) \times (r-u)}$. The integer u is referred to as the dimension of the envelope model.

The parameters to be estimated are $\theta = \{\alpha, \beta, \Sigma_\varepsilon\} = \{\alpha, \eta, \Gamma, \Gamma_0, \Omega, \Omega_0\}$. Both Khare et al. [16] and Chakraborty and Su [2] recommend a non-informative uniform prior for α and a normal

prior for $\eta \mid \Gamma$. Constraints on $\Gamma, \Gamma_0, \Omega, \Omega_0$ are required to ensure identifiability. The two studies adopt different strategies for imposing these constraints, resulting in distinct prior specifications.

Khare et al. [16] assume that (Γ, Γ_0) forms an orthogonal matrix and that Ω and Ω_0 are diagonal matrices with ordered positive entries. Under this parameterization, the $r(r+1)/2$ parameters in Σ_ε are represented by $r(r-1)/2$ parameters in (Γ, Γ_0) and r parameters in $\{\Omega, \Omega_0\}$. The prior distribution for (Γ, Γ_0) is a matrix-variate Bingham distribution, while the diagonal entries of Ω and Ω_0 are assigned independent inverse-Gamma priors. Khare et al. [16] also develop an algorithm for posterior sampling, although it is complex and challenging to implement.

Chakraborty and Su [2] parameterize (Γ, Γ_0) using a matrix $A \in \mathbb{R}^{(r-u) \times u}$. Specifically, for any given A , one can construct column-orthogonal matrices $\Gamma = C_A(C_A^T C_A)^{-1/2}$ and $\Gamma_0 = D_A(D_A^T D_A)^{-1/2}$ satisfying $\Gamma^T \Gamma_0 = 0$, where $C_A = (I_u, A^T)^T$ and $D_A = (-A, I_{r-u})^T$. They further assume that Ω and Ω_0 are positive definite. Under this parameterization, the $r(r+1)/2$ parameters in Σ_ε are represented by $u(r-u)$ parameters in (Γ, Γ_0) or equivalently A , $u(u+1)/2$ parameters in Ω , and $(r-u)(r-u+1)/2$ parameters in Ω_0 . Chakraborty and Su [2] assign a matrix-variate normal prior to A and independent inverse-Wishart priors to Ω and Ω_0 . While computationally convenient, using a normal prior for A makes it difficult to incorporate prior knowledge about (Γ, Γ_0) .

The parameterization of our method is similar to that of Chakraborty and Su [2]. However, instead of placing a prior on A , we directly specify a prior on Γ , or more precisely, on $P = \Gamma\Gamma^T$. Note that there is a one-to-one correspondence between A and $P = \Gamma\Gamma^T$. Since P is idempotent and uniquely defines a linear subspace, this establishes a one-to-one correspondence between A and u -dimensional linear subspaces, which form a Grassmann manifold. By specifying the prior directly on the Grassmann manifold, it becomes straightforward to incorporate existing knowledge about (Γ, Γ_0) . Section 2.3 provides a brief review of the Grassmann manifold.

Both Khare et al. [16] and Chakraborty and Su [2] assume that the random errors are independently distributed according to a multivariate normal distribution. In this paper, we instead assume that the random errors follow a scale mixture of matrix-variate normal distributions, a flexible class capable of capturing outliers. Moreover, our approach accommodates correlation structures arising from repeated measurements.

2.2 Scale mixture of matrix-variate normal distributions

This subsection introduces the scale mixture of matrix-variate normal (SMMN) distributions, which extend the normal distribution to accommodate outliers. In addition, a stochastic representation of the SMMN distribution provides a hierarchical model that can be used in practice for sampling.

Let $M \in \mathbb{R}^{a \times b}$ be an arbitrary matrix, and let $\Lambda_1 \in \mathbb{R}^{a \times a}$ and $\Lambda_2 \in \mathbb{R}^{b \times b}$ be positive definite matrices. A random matrix $Y \in \mathbb{R}^{a \times b}$ is said to follow a matrix-variate normal distribution [14], denoted by $\text{MN}(M, \Lambda_1, \Lambda_2)$, if its density is

$$(2\pi)^{-ab/2} |\Lambda_1|^{-b/2} |\Lambda_2|^{-a/2} \text{etr} \left(-\frac{1}{2} \Lambda_1^{-1} (Y - M) \Lambda_2^{-1} (Y - M)^T \right),$$

where $\text{etr}(X) = \exp(\text{tr}(X))$ for a matrix X . Λ_1 and Λ_2 are referred to as the row and column covariance matrices, respectively, and they capture the covariance structures of the rows and columns of Y . It can be shown that $Y \sim \text{MN}(M, \Lambda_1, \Lambda_2)$ if and only if $\text{vec}(Y) \sim N(\text{vec}(M), \Lambda_2 \otimes \Lambda_1)$, where $\text{vec}(\cdot)$ stacks the columns of a matrix into a vector, and \otimes denotes the Kronecker product.

Let $G(\cdot; \nu)$ be a univariate probability distribution with parameter ν and support on $(0, \infty)$. A random matrix $Y \in \mathbb{R}^{a \times b}$ is said to follow a scale mixture of matrix-variate normal distributions,

denoted by $\text{SMMN}(M, \Lambda_1, \Lambda_2, G)$, if its density is

$$(2\pi)^{-ab/2} |\Lambda_1|^{-b/2} |\Lambda_2|^{-a/2} \int_0^\infty u^{ab/2} \text{etr} \left(-\frac{u}{2} \Lambda_1^{-1} (Y - M) \Lambda_2^{-1} (Y - M)^T \right) dG(u; \nu).$$

A stochastic representation of the SMMN distribution [1] is

$$Y = M + \tau^{-1/2} Z,$$

where $Z \in \mathbb{R}^{a \times b} \sim \text{MN}(0, \Lambda_1, \Lambda_2)$ is independent of $\tau \sim G(\cdot; \nu)$. This leads to the hierarchical representation

$$Y \mid \tau \sim \text{MN}(M, \tau^{-1} \Lambda_1, \Lambda_2), \quad \tau \sim G(\cdot; \nu),$$

for which the marginal distribution of Y is exactly $\text{SMMN}(M, \Lambda_1, \Lambda_2, G)$. The first two central moments of Y are

$$\begin{aligned} E(Y) &= E[E(Y \mid \tau)] = M, \\ \text{var}(\text{vec}(Y)) &= \text{var}[E(\text{vec}(Y) \mid \tau)] + E[\text{var}(\text{vec}(Y) \mid \tau)] = (\Lambda_2 \otimes \Lambda_1) \cdot E(\tau^{-1}). \end{aligned}$$

By choosing an appropriate distribution $G(\cdot; \nu)$, the SMMN distribution can accommodate outliers. For example, when τ is a finite mixture of point masses, Y becomes a finite mixture of normal distributions. When $\tau \sim \text{Gamma}(\nu/2, \nu/2)$, Y follows a matrix-variate t -distribution [14], denoted by $\text{MT}(\nu, M, \Lambda_1, \Lambda_2)$, with density

$$\frac{\Gamma((ab + \nu)/2)}{\Gamma(\nu/2) \nu^{ab/2} \pi^{ab/2}} |\Lambda_1|^{-b/2} |\Lambda_2|^{-a/2} \left[1 + \frac{1}{\nu} \text{tr} (\Lambda_1^{-1} (Y - M) \Lambda_2^{-1} (Y - M)^T) \right]^{-(ab + \nu)/2},$$

where ν is the degrees of freedom and $\Gamma(\cdot)$ denotes the Gamma function. When $\nu > 1$, the mean of Y is M , and when $\nu > 2$, the variance of $\text{vec}(Y)$ is $(\nu/(\nu - 2))(\Lambda_2 \otimes \Lambda_1)$.

In this paper, we present the general model using the SMMN distribution, but focus on the matrix-variate t -distribution in our algorithms and examples due to its wide applicability. Extensive studies on t -regression have been conducted in literature [11, 12, 20, 7].

2.3 Distributions on Grassmann manifold

This subsection introduces several distributions on the Grassmann manifold. For a comprehensive discussion, see Chikuse [3]. Here, we focus only on the distributions relevant to this paper.

The Grassmann manifold $\mathcal{G}_{m,k}$ is the space of all k -dimensional linear subspaces of \mathbb{R}^m . Each k -dimensional subspace corresponds uniquely to an $m \times m$ idempotent matrix of rank k , so the Grassmann manifold $\mathcal{G}_{m,k}$ can be equivalently represented by

$$\mathcal{P}_{m,k} = \{P \in \mathbb{R}^{m \times m} \mid PP = P, \text{rank}(P) = k\}.$$

We typically write $P = \Gamma\Gamma^T$, where the columns of Γ form an orthogonal basis for the subspace corresponding to P . The choice of Γ is not unique: for any $k \times k$ orthogonal matrix U , ΓU spans the same subspace as Γ , because $P = (\Gamma U)(\Gamma U)^T = \Gamma\Gamma^T$.

We introduce two distributions on $\mathcal{P}_{m,k}$ that exhibit a concentration property, analogous to how a normal distribution is concentrated around its mode. Note that the density function of a

distribution on $\mathcal{P}_{m,k}$ is defined with respect to the Haar measure or uniform distribution on $\mathcal{P}_{m,k}$. The matrix Langevin distribution on $\mathcal{P}_{m,k}$ [3] has density

$$f(P) = \frac{1}{{}_1F_1(\frac{1}{2}k; \frac{1}{2}m; M)} \text{etr}(MP), \quad (1)$$

where ${}_1F_1(\cdot; \cdot; \cdot)$ is a hypergeometric function with a matrix argument, and $M \in \mathbb{R}^{m \times m}$ is symmetric. This distribution is closely related to the matrix Bingham distribution: if $\Gamma \in \mathbb{R}^{m \times k}$ is a random column-orthogonal matrix following a matrix Bingham distribution with parameter M , then the density of $P = \Gamma\Gamma^T$ is exactly (1). Another common distribution on $\mathcal{P}_{m,k}$ [3] has density

$$f(P) = |M|^{-k/2} |I_m - P + M^{-1}P|^{-m/2}, \quad (2)$$

where $M \in \mathbb{R}^{m \times m}$ is symmetric. This distribution characterizes the linear space spanned by a normal random matrix. The connection between the distribution (2) and the normal distribution makes it particularly suitable as a proposal distribution in a Metropolis-Hastings algorithm, since sampling from it is straightforward.

Proposition 1. *If Z is an $m \times k$ random matrix following a matrix-variate normal distribution $MN(0, M, I_k)$, then the corresponding projection matrix $P = Z(Z^T Z)^{-1}Z^T$ has the density given in (2).*

For both densities (1) and (2), the modes can be explicitly determined from M . Let the spectral decomposition of M be $M = U\Lambda U^T$, where U is an $m \times m$ orthogonal matrix and $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_m)$ with $\lambda_1 \geq \dots \geq \lambda_m$. The following proposition characterizes the modes. The proof is provided in the supplementary material.

Proposition 2. *The mode of (1) or (2) is given by $P_0 = U_1 U_1^T$, where U_1 contains the first k columns of U . The mode is unique if and only if $\text{rank}(M) \geq k$ and $\lambda_k > \lambda_{k+1}$.*

This proposition implies that when (1) or (2) is used as a prior distribution, the dominant eigenvectors of M can be chosen to encode prior knowledge about P , and the magnitudes of the associated eigenvalues indicate the strength of this prior belief. For example, setting $M = 0$ yields the uniform distribution on $\mathcal{P}_{m,k}$, corresponding to a non-informative prior. A vague prior can be specified by taking $M = sP_0$ with a small positive s and an initial guess P_0 . If we believe that a particular direction γ lies in the subspace spanned by P , we may set $M = a\gamma\gamma^T + sI_m$ for $a > 0$, where larger values of a represent stronger confidence in this prior information. For multiple directions, one may use $M = a\gamma_1\gamma_1^T + b\gamma_2\gamma_2^T + sI_m$ with a and b tuned to reflect the relative strength of confidence in each direction. The effects of these choices of M are further illustrated by simulation in Section 5.4.

3 Model Specification

This section presents the model specification, recommends prior distributions, and derives the corresponding posterior distributions.

Let the observations be $\{(y_{ij}, x_{ij}), y_{ij} \in \mathbb{R}^r, x_{ij} \in \mathbb{R}^p, i = 1, \dots, n, j = 1, \dots, J_i\}$, where (y_{ij}, x_{ij}) denotes the data collected from subject i at time point j . Let $Y_i = (y_{i1}, \dots, y_{iJ_i}) \in \mathbb{R}^{r \times J_i}$ and

$X_i = (x_{i1}, \dots, x_{iJ_i}) \in \mathbb{R}^{p \times J_i}$ be all responses and covariates for the i th subject. Consider the following linear regression model with latent variables, for $i = 1, \dots, n$,

$$Y_i = \Gamma Z_i + \Gamma_0 Z_{0,i}, \quad Z_i = \mu 1_{J_i}^T + \eta X_i + \zeta_i, \quad Z_{0,i} = \mu_0 1_{J_i}^T + \zeta_{0,i}, \quad (3)$$

where (Γ, Γ_0) is an $r \times r$ orthogonal matrix, $\Gamma \in \mathbb{R}^{r \times u}$, $\Gamma_0 \in \mathbb{R}^{r \times (r-u)}$, $Z_i \in \mathbb{R}^{u \times J_i}$, $Z_{0,i} \in \mathbb{R}^{(r-u) \times J_i}$, $\mu \in \mathbb{R}^u$, $\eta \in \mathbb{R}^{u \times p}$, $\zeta_i \in \mathbb{R}^{u \times J_i}$, $\mu_0 \in \mathbb{R}^{r-u}$, and $\zeta_{0,i} \in \mathbb{R}^{(r-u) \times J_i}$. The model (3) explicitly writes the response Y_i as the sum of two components, where Z_i depends on the covariate X_i whereas $Z_{0,i}$ does not. Following the envelope-model framework, we assume that ζ_i and $\zeta_{0,i}$ are uncorrelated. Specifically, we assume that they jointly follow a scalar mixture of matrix-variate normal distributions as follows.

$$\begin{pmatrix} \zeta_i \\ \zeta_{0,i} \end{pmatrix} \sim \text{SMMN}(0_{r \times J_i}, \text{diag}\{\Omega, \Omega_0\}, R_i, G), \quad (4)$$

where $\Omega \in \mathbb{R}^{u \times u}$ and $\Omega_0 \in \mathbb{R}^{(r-u) \times (r-u)}$ are positive definite, and $R_i = R_i(\rho) \in \mathbb{R}^{J_i \times J_i}$ is a correlation matrix parameterized by ρ . It follows that the marginal distribution of Y_i is

$$Y_i \sim \text{SMMN}(\alpha 1_{J_i}^T + \beta X_i, \Sigma_\varepsilon, R_i, G), \quad (5)$$

where

$$\alpha = \Gamma \mu + \Gamma_0 \mu_0 \in \mathbb{R}^r, \quad \beta = \Gamma \eta \in \mathbb{R}^{r \times p}, \quad \Sigma_\varepsilon = \Gamma \Omega \Gamma^T + \Gamma_0 \Omega_0 \Gamma_0^T \in \mathbb{R}^{r \times r}. \quad (6)$$

We further assume that observations from different subjects are independent, that is, Y_i and $Y_{i'}$ are independent for $i \neq i'$. The latent model (3)-(4), or equivalently the marginal model (5)-(6), are referred to as the **Robust Longitudinal Envelope Model** (RoLEM). As a special case, when the distribution in (4) or (5) is normal, the model reduces to the **Longitudinal Envelope Model** (LEM).

The parameter Σ_ε quantifies the variance and covariance of different responses at each time point. For any subject i , each column of Y_i , for example, $\text{col}_j(Y_i)$, contains all responses at time point j . Its variance is

$$\text{var}(\text{col}_j(Y_i)) = c \Sigma_\varepsilon,$$

where $c = E(\tau^{-1})$ for $\tau \sim G(\cdot; \nu)$ is a constant only depending on G . Thus, the variance of the multivariate outcome is the same at every time point. Similarly, each row of Y_i , for example, $\text{row}_k(Y_i)$, contains the k th response across all time points for the i th subject. Its variance is

$$\text{var}(\text{row}_k(Y_i)) = c \sigma_{kk} R_i,$$

where σ_{kk} is the (k, k) th entry of Σ_ε . Here, R_i accounts for the temporal correlation of repeated measurements of the same response through a parameter ρ . It implies that all responses share the same temporal correlation structure. Common choices for R_i include compound symmetry (CS), where $R_i = (1 - \rho)I_{J_i} + \rho 1_{J_i} 1_{J_i}^T$, or AR(1), where $R_i = (\rho^{|s-l|})$.

To ensure identifiability, we impose constraints on $\{\Gamma, \Gamma_0, \Omega, \Omega_0\}$. Recall that Γ is an orthogonal matrix, and the projection matrix $P = \Gamma \Gamma^T$ is identifiable as long as the linear space spanned by Γ remains unchanged. Because the values of Ω and Ω_0 depend on the particular choice of Γ and Γ_0 , it suffices to specify a unique and deterministic procedure to compute Γ from a given P . Let

(U_1, U_2) be any chosen orthogonal matrix, where $U_1 \in \mathbb{R}^{r \times u}$ and $U_2 \in \mathbb{R}^{r \times (r-u)}$. For any given $P \in \mathcal{P}_{r,u}$, compute $\tilde{\Gamma} \in \mathbb{R}^{r \times u}$, whose columns are the first u eigenvectors of P . Because the first u eigenvalues of P are all ones, $\tilde{\Gamma}$ may not be unique. Next, let $A = (U_2^T \tilde{\Gamma})(U_1^T \tilde{\Gamma})^{-1} \in \mathbb{R}^{(r-u) \times u}$, which is uniquely determined whenever $U_1^T \tilde{\Gamma}$ is invertible. Finally, compute

$$\Gamma = (U_1 + U_2 A)(I_u + A^T A)^{-1/2}, \quad \Gamma_0 = (-U_1 A^T + U_2)(I_{r-u} + A A^T)^{-1/2}. \quad (7)$$

This construction parameterizes P through the matrix A , which reduces to the method of Chakraborty and Su [2] when $U_1 = (I_u, 0)^T$ and $U_2 = (0, I_{r-u})^T$. This general form provides greater flexibility, particularly when $(I_u, 0)^T \tilde{\Gamma}$ is not invertible.

3.1 Prior distributions

This subsection describes the prior distributions for the model parameters. For ease of presentation, we focus on the case where the SMMN follows a matrix-variate t -distribution; that is, $G(\cdot; \nu)$ is taken to be $\text{Gamma}(\nu/2, \nu/2)$, so that ν represents the degrees of freedom. The full parameter vector is

$$\theta = \{\alpha, \beta, \Sigma_\varepsilon, \rho, \nu\} = \{\alpha, \eta, P, \Omega, \Omega_0, \rho, \nu\}.$$

The dimension u and the correlation structure are assumed to be known. Model selection for these components will be addressed in Section 4.4.

- The prior distribution of the intercept $\alpha \in \mathbb{R}^r$ is non-informative, that is,

$$\pi(\alpha) \propto 1.$$

It is the same prior specification used in Khare et al [16] and Chakraborty and Su [2].

- The prior distribution of $\eta \in \mathbb{R}^{u \times p}$, conditional on Γ and Ω , is a matrix-variate normal distribution $\text{MN}(\Gamma^T \xi, \Omega, H^{-1})$, where $\xi \in \mathbb{R}^{r \times p}$ and $H \in \mathbb{R}^{p \times p}$ are user-specified matrices.

$$\pi(\eta \mid \Gamma, \Omega) \propto |\Omega|^{-p/2} \text{etr} \left(-\frac{1}{2} \Omega^{-1} (\eta - \Gamma^T \xi) H (\eta - \Gamma^T \xi)^T \right).$$

This prior is conjugate and is also used in Khare et al [16] and Chakraborty and Su [2]. The prior is vague when H is close to zero and becomes non-informative when $H = 0$.

- The matrices $\Omega \in \mathbb{R}^{u \times u}$ and $\Omega_0 \in \mathbb{R}^{(r-u) \times (r-u)}$ are both positive definite. Their prior distributions are independent inverse-Wishart distributions with different parameters.

$$\begin{aligned} \pi(\Omega) &\propto |\Omega|^{-(k+u+1)/2} \text{etr} \left(-\frac{1}{2} \Psi \Omega^{-1} \right), \\ \pi(\Omega_0) &\propto |\Omega_0|^{-(k_0+r-u+1)/2} \text{etr} \left(-\frac{1}{2} \Psi_0 \Omega_0^{-1} \right), \end{aligned}$$

where the degrees of freedom $k > 0$ and $k_0 > 0$, and scale parameters $\Psi \in \mathbb{R}^{u \times u}$ and $\Psi_0 \in \mathbb{R}^{(r-u) \times (r-u)}$ are positive definite.

- The prior distribution of $P \in \mathcal{P}_{r,u}$ is a matrix Langevin distribution with parameter $M \in \mathbb{R}^{r \times r}$, where M is symmetric.

$$\pi(P) \propto \text{etr}(MP).$$

Setting $M = 0$ yields a non-informative prior, whereas choosing M to be close to zero leads to a vague prior. Prior knowledge can also be incorporated through an appropriate choice of M , as discussed in Section 2.3.

- We consider two correlation structures, CS and AR(1), in this paper. Let $\rho \in (0, 1)$ and the prior distribution is uniform(0, 1).

$$\pi(\rho) \propto 1.$$

- The degrees of freedom ν are positive. Assume $\nu > 2$ to ensure the existence of the second moment of Y . The prior distribution of ν is Gamma(a, b) with $a > 0$ and $b > 0$.

$$\pi(\nu) \propto \nu^{a-1} e^{-b\nu}.$$

There exists a large literature on choosing a prior for ν in t -regression [11, 12, 20, 7]. The choice of Gamma distribution is a compromise between simplicity and flexibility.

The parameters in the prior distributions are referred to as hyperparameters. They are specified by the user to reflect prior knowledge about θ , the model parameter. Section 4.2 provides a discussion on the selection of hyperparameters.

3.2 Posteriors

The density of the joint posterior distribution is

$$\pi(\theta | Y, X) \propto \pi(Y | X, \theta) \pi(\alpha) \pi(\eta | \Gamma, \Omega) \pi(\Omega) \pi(\Omega_0) \pi(P) \pi(\rho) \pi(\nu).$$

It is difficult to work with this posterior distribution directly. We introduce a latent variable $\tau = \{\tau_1, \dots, \tau_n\}$ and consider the hierarchical model

$$\begin{aligned} Y_i | \tau_i &\sim \text{MN}(\alpha 1_{J_i}^T + \beta X_i, \tau_i^{-1} \Sigma_\varepsilon, R_i), \\ \tau_i &\sim \text{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right). \end{aligned}$$

The latent variable τ_i can be incorporated into either Σ_ε or R_i , both of which yield the same marginal distribution. We choose to include τ_i in Σ_ε in order to keep R_i as a correlation matrix. When $\tau_i = 1$, Y_i follows a normal distribution. A large value of τ_i indicates that Y_i is a potential outlier. The density of the joint posterior for the augmented data is

$$\pi(\theta, \tau | Y, X) \propto \pi(Y | X, \theta, \tau) \pi(\tau) \pi(\alpha) \pi(\eta | \Gamma, \Omega) \pi(\Omega) \pi(\Omega_0) \pi(P) \pi(\rho) \pi(\nu).$$

We derive the conditional posterior distribution of each parameter given the remaining ones, Y , and X as follows. Some of them are standard distributions, while others are not and require an implementation of the Metropolis-Hastings algorithm.

Let

$$\Delta_i = \text{tr}(\Sigma_\varepsilon^{-1} (Y_i - \alpha 1_{J_i}^T - \beta X_i) R_i^{-1} (Y_i - \alpha 1_{J_i}^T - \beta X_i)^T).$$

- The conditional posterior of τ_i is a Gamma($(\nu + J_i r)/2$, $(\nu + \Delta_i)/2$).
- The conditional posterior of α is a multivariate normal distribution $N(\mu_\alpha, \Sigma_\alpha)$, where

$$\mu_\alpha = \left(\sum_{i=1}^n \tau_i (1_{J_i}^T R_i^{-1} 1_{J_i}) \right)^{-1} \sum_{i=1}^n \tau_i [(Y_i - \beta X_i) R_i^{-1} 1_{J_i}],$$

$$\Sigma_\alpha = \left(\sum_{i=1}^n \tau_i (1_{J_i}^T R_i^{-1} 1_{J_i}) \right)^{-1} \Sigma_\varepsilon.$$

- The conditional posterior of η is a matrix-variate normal distribution $MN(\Gamma^T \tilde{\xi}, \Omega, \tilde{H}^{-1})$, where

$$\tilde{H} = H + \sum_{i=1}^n \tau_i X_i R_i^{-1} X_i^T,$$

$$\tilde{\xi} = \left(\xi H + \sum_{i=1}^n \tau_i (Y_i - \alpha 1_{J_i}^T) R_i^{-1} X_i^T \right) \tilde{H}^{-1}.$$

- The conditional posterior of Ω is an inverse-Wishart distribution $IW(k + p + \sum_i J_i, \tilde{\Psi})$, where

$$\tilde{\Psi} = \Psi + (\eta - \Gamma^T \xi) H (\eta - \Gamma^T \xi)^T + \sum_{i=1}^n \tau_i \Gamma^T (Y_i - \alpha 1_{J_i}^T - \beta X_i) R_i^{-1} (Y_i - \alpha 1_{J_i}^T - \beta X_i)^T \Gamma.$$

- The conditional posterior of Ω_0 is an inverse-Wishart distribution $IW(k_0 + \sum_i J_i, \tilde{\Psi}_0)$, where

$$\tilde{\Psi}_0 = \Psi_0 + \sum_{i=1}^n \tau_i \Gamma_0^T (Y_i - \alpha 1_{J_i}^T) R_i^{-1} (Y_i - \alpha 1_{J_i}^T)^T \Gamma_0.$$

- The conditional posterior of P is not a standard distribution. Its density is

$$\pi(P \mid \dots) \propto \exp \left(-\frac{1}{2} \sum_{i=1}^n \tau_i \Delta_i - \frac{1}{2} \text{tr}(\Omega^{-1} (\eta - \Gamma^T \xi) H (\eta - \Gamma^T \xi)^T) + \text{tr}(MP) \right),$$

where \dots means the remaining parameters, Y , and X .

- The conditional posterior of ρ is not a standard distribution. Its density is

$$\pi(\rho \mid \dots) \propto \exp \left(-\frac{r}{2} \sum_{i=1}^n \log |R_i| - \frac{1}{2} \sum_{i=1}^n \tau_i \Delta_i \right).$$

- The conditional posterior of ν is not a standard distribution. Its density is

$$\pi(\nu \mid \dots) \propto \exp \left(\frac{n\nu}{2} \log \frac{\nu}{2} - n \log \Gamma \left(\frac{\nu}{2} \right) + \left(\frac{\nu}{2} - 1 \right) \sum_{i=1}^n \log \tau_i - \frac{\nu}{2} \sum_{i=1}^n \tau_i + (a - 1) \log \nu - b\nu \right).$$

4 Implementation

This section addresses practical aspects of implementation, including the Metropolis-Hastings algorithm, specification of hyper-parameters, selection of initial values, model selection, and software.

4.1 Metropolis-Hastings algorithm

Samples of $\{\theta, \tau\}$ are drawn from the posterior distribution in the order of τ_i , ν , Ω , Ω_0 , P , ρ , α , and η . We can compute Γ and Γ_0 from P by (7) whenever is necessary. The conditional posterior distributions of τ_i , Ω , Ω_0 , α , and η are standard distributions and hence they can be sampled directly. We implement Metropolis-Hastings algorithms [13] to sample ν , P , and ρ from their conditional posteriors.

- When sampling ρ with the current value being $\rho^{(s)}$, a proposed value ρ^* is drawn from $\text{uniform}(\rho^{(s)} - \delta_\rho, \rho^{(s)} + \delta_\rho)$, where $\delta_\rho > 0$ is a pre-specified constant. Because $\rho \in (0, 1)$, we adjust ρ^* if it falls outside this interval: replace ρ^* by $|\rho^*|$ if $\rho^* < 0$, and by $2 - \rho^*$ if $\rho^* > 1$. Let $f(\rho | \rho^{(s)})$ be the density of the proposal distribution. This proposal distribution is symmetric in the sense that $f(\rho | \rho^{(s)}) = f(\rho^{(s)} | \rho)$, and hence it yields a random walk update.
- When sampling ν with the current value being $\nu^{(s)}$, a proposed value ν^* is drawn from $\text{uniform}(\nu^{(s)} - \delta_\nu, \nu^{(s)} + \delta_\nu)$, where $\delta_\nu > 0$ is a pre-specified constant. Because $\nu > 2$, we adjust ν^* if it falls below 2 by replacing it by $4 - \nu^*$ if $\nu^* < 2$. Let $f(\nu | \nu^{(s)})$ be the density of the proposal distribution. This proposal distribution is symmetric in the sense that $f(\nu | \nu^{(s)}) = f(\nu^{(s)} | \nu)$, resulting in a random walk update.
- When sampling P with the current value $P^{(s)}$, a proposed matrix P^* is drawn from the density $f(P | P^{(s)}) = |W|^{-u/2} |I_r - P + W^{-1}P|^{-r/2}$, where $W = \sigma^2 I_r + P^{(s)}$ and σ^2 is a pre-specified constant. This proposed matrix can be generated as $P^* = Z(Z^T Z)^{-1} Z^T$, where Z is drawn from a matrix-variate normal distribution $\text{MN}(0, W, I)$. This proposal distribution is symmetric in the sense that $f(P | P^{(s)}) = f(P^{(s)} | P)$, which is proved in the supplementary material. Hence, this step is also a random walk.

The values of δ_ρ , δ_ν , σ^2 control the acceptance rates of the Metropolis-Hastings algorithm. It is recommended that the values be chosen to ensure the acceptance rate falls within $(0.2, 0.5)$. The convergence of the MCMC chains can be assessed using the standard diagnostic tools, such as trace plots, autocorrelation plots, and formal convergence tests.

4.2 Hyper-parameters in prior distributions

The hyperparameters of the prior distributions need to be specified. Users may choose them based on prior knowledge, following common practice in Bayesian inference. When limited prior information is available, we recommend the following settings for vague priors:

- set $\xi = 0_{u \times p}$ and $H = 10^{-3} I_p$ for the prior of η ;
- set $k = u + 1$ and $\Psi = 10^{-3} I_u$ for the prior of Ω ;
- set $k_0 = r - u + 1$ and $\Psi_0 = 10^{-3} I_{r-u}$ for the prior of Ω_0 ; and
- set $M = 10^{-3} I_r$ for the prior of P .

In some cases, even weaker prior information may be desired, in which case one may instead set $H = 10^{-6} I_p$, $\Psi = 10^{-6} I_u$, $\Psi_0 = 10^{-6} I_{r-u}$, and $M = 10^{-6} I_r$ to obtain nearly non-informative priors. Finally, for the Gamma-prior for the degrees of freedom ν , set $a = 1.4$ and $b = 0.04$, which makes the prior probability of $P(2 < \nu < 110) = 0.95$.

4.3 Initialization

Under mild conditions, an MCMC chain starting from any value will converge if the chain is long enough. In practice, to ensure faster convergence, we recommend setting the initial value of the chain as a raw estimate of the parameters, assuming the distribution of Y_i is normal. Notice that if the distribution of Y_i is normal, the mean and variance of Y_i are

$$\mu_i = \alpha 1_{J_i}^T + \beta X_i \in \mathbb{R}^{r \times J_i}, \quad \Lambda_i = R_i \otimes \Sigma_\varepsilon \in \mathbb{R}^{J_i^r \times J_i^r}.$$

Hence, the initial values can be obtained by the steps below.

1. Set $\tau_i^{(0)} = 1$ for all $i = 1, \dots, n$. Set $\nu^{(0)} = 10$.
2. Set $\alpha^{(0)}$ and $\beta^{(0)}$ as the least squares estimates of the intercepts and slopes when fitting individual regression with each component of Y_i as response and X_i as covariates.
3. Set $\Sigma_\varepsilon^{(0)}$ as the variance matrix of all r_{ij} , where residual $r_{ij} = y_{ij} - \alpha^{(0)} - \beta^{(0)} x_{ij}$.
4. Set $\Gamma^{(0)}$ as the matrix formed by the first u left singular vectors from the singular value decomposition of $\beta^{(0)}$.
5. Compute $A^{(0)}$ from $\Gamma^{(0)}$ and then compute $\Gamma_0^{(0)}$ from $A^{(0)}$ as in (7).
6. Set $\eta^{(0)} = \Gamma^{(0)T} \beta^{(0)}$, $\Omega^{(0)} = \Gamma^{(0)T} \Sigma_\varepsilon^{(0)} \Gamma^{(0)}$, and $\Omega_0^{(0)} = \Gamma_0^{(0)T} \Sigma_\varepsilon^{(0)} \Gamma_0^{(0)}$.
7. Set $\rho^{(0)}$ as the correlation between \tilde{r}_{ijk} and $\tilde{r}_{i,j+1,k}$, where $\tilde{r}_{ijk} = r_{ijk} / \sqrt{\sigma_{kk}}$, r_{ijk} is the k th component of r_{ij} , and σ_{kk} is the (k, k) th entry of $\Sigma_\varepsilon^{(0)}$.

Note that the last step is valid because, for both the CS and AR(1) correlation structures, the correlation between r_{ijk} and $r_{i,j+1,k}$ equals ρ for each k th component of residuals.

We also need to choose an orthogonal matrix U , which is used in (7) to compute Γ and Γ_0 from A , or to compute A from P . This matrix remains fixed throughout the sampling procedure. A simple choice is $U = I_r$ or a randomly generated orthogonal matrix. If the trace plot of A exhibits many extreme values, it may indicate that the current choice of U is not appropriate, as it leads to inverting a near-singular matrix. In such cases, select a different U and restart the sampling procedure. The following two strategies usually lead to a good choice.

- Set U as the Q -matrix from the QR factorization of an initial estimate of β , for example, the least squares estimate as in Step 2 above. If $r > p$, append additional columns to make U a full orthogonal matrix.
- Set $U = (\hat{\Gamma}, \hat{\Gamma}_0)$, where $\hat{\Gamma}$ and $\hat{\Gamma}_0$ are the posterior means of Γ and Γ_0 from the initial sampling using $U = I_r$ or other choice of matrix.

4.4 Model selection

When applying RoLEM, it is necessary to determine the envelope dimension u , the correlation structure, and whether to assume t -distributed or normally distributed random errors. These decisions can all be formulated as a model selection problem. There is a rich literature on this topic in Bayesian analysis; in this paper, we focus on BIC and WAIC.

The BIC (Bayesian or Schwarz information criterion) [18] is defined as

$$\text{BIC} = -2 \log L + p_{\text{BIC}} \log n,$$

where L is the maximized likelihood, p_{BIC} is the effective number of parameters, and n is the number of subjects. In practice, L can be approximated by the maximum of the likelihood observed across MCMC draws. For RoLEM,

$$p_{\text{BIC}} = r + up + (r - u)u + u(u + 1)/2 + (r - u)(r - u + 1)/2 + 1 + 1,$$

which corresponds to α , η , P , Ω , Ω_0 , ρ , and ν . BIC achieves a trade-off between goodness-of-fit and model complexity, with a smaller value indicating a better model. Comparing two models using BIC is asymptotically equivalent to computing the Bayes factor [15]. Specifically, the Bayes factor BF_{12} of two models M_1 and M_2 is approximated by $\text{BF}_{12} \approx \exp(-(\text{BIC}_1 - \text{BIC}_2)/2)$, where BIC_1 and BIC_2 are BIC values of M_1 and M_2 , respectively. A difference of $\text{BIC}_1 - \text{BIC}_2 \leq -6$ corresponds to a Bayes factor larger than 20, and provides strong evidence in favor of M_1 over M_2 [15].

Out-of-sample prediction performance is commonly used for model comparison. Although cross-validation can estimate out-of-sample prediction, it is often time-consuming. The WAIC (widely applicable or Watanabe-Akaike information criterion) [13, 21] is shown to be asymptotically equivalent to cross-validation. It is defined by

$$\text{WAIC} = -2 \sum_{i=1}^n \log E_{\text{post}}[p(y_i | \theta)] + 2 \sum_{i=1}^n \text{var}_{\text{post}}(\log p(y_i | \theta)),$$

where E_{post} and var_{post} denote the expectation and variance with respect to the posterior distribution of θ . The first term represents the log pointwise predictive density, while the second term corresponds to the effective number of parameters. In practice, these quantities can be approximated by their sample counterparts using posterior samples. Models with smaller WAIC values are preferred.

In practice, BIC and WAIC can be computed for different candidate models, which may differ in terms of the envelope dimension u , the correlation structure, or both. The model with the smallest BIC or WAIC is preferred.

4.5 Software

The algorithm has been implemented in an R package, with the core computational components written in C++. The package is available at <https://github.com/pzengauburn/benvlp>. Table 1 reports the average computation time (in seconds) required to generate 1,000 samples for selected scenarios. All computations were performed on a MacBook Pro with an Apple M2 chip and 8 GB of RAM, running macOS Tahoe 26.1. An AR(1) correlation structure was assumed in these computations; results are similar for other correlation structures. The computation time increases

as r , p , n , or J grows. Although the total number of observations is the same for $(n, J) = (100, 5)$ and $(50, 10)$, as well as for $(n, J) = (200, 5)$ and $(100, 10)$, computation time increases as J becomes larger.

Table 1: Average Computation Time (in seconds) Required to Generate 1,000 Samples

(r, p)	u	$J = 5$			$J = 10$		
		$n = 50$	$n = 100$	$n = 200$	$n = 50$	$n = 100$	$n = 200$
(5, 6)	2	0.14	0.27	0.52	0.31	0.61	1.18
	3	0.15	0.27	0.52	0.31	0.61	1.20
	4	0.15	0.27	0.53	0.30	0.60	1.16
(20, 30)	2	0.89	1.68	3.28	1.71	3.35	6.55
	3	0.89	1.64	3.28	1.72	3.36	6.57
	4	0.89	1.70	3.29	1.73	3.35	6.57

In both the simulation studies and the real data analysis, we use equal time points for each subject for simplicity and clarity of presentation. However, the model can accommodate the imbalanced case. Recall that for each subject, the response $Y_i \in \mathbb{R}^{r \times J_i}$ contains J_i time points, and J_i need not be the same across subjects. The implementation provided in the associated R package can also handle data with unequal time points.

5 Examples

This section discusses some numerical experiments and a real data analysis to demonstrate the performance of RoLEM.

5.1 Simulation study

This subsection reports simulation studies to compare the performance of RoLEM with several alternative methods under various settings.

Follow the steps below to generate synthetic datasets. Set the number of responses to $r = 20$, the number of covariates to $p = 30$, and the number of time points to $J = 5$ or 10. Set the envelope dimension to $u = 3$. Randomly generate a matrix $A \in \mathbb{R}^{(r-u) \times u}$, whose entries are independently sampled from $\text{uniform}(-1, 1)$. Let $U = I_r$ and compute Γ and Γ_0 using (7). Randomly generate Ω and Ω_0 as diagonal matrices whose diagonal entries are independently drawn from $\text{uniform}(0, 1)$ and $\text{uniform}(5, 10)$, respectively. Finally, compute $\Sigma_\varepsilon = \Gamma\Omega\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T$. For each subject, generate the covariate matrix $X_i \in \mathbb{R}^{p \times J}$ with entries independently drawn from $N(0, 1)$. Compute the response as $Y_i = \alpha 1_J^T + \Gamma\eta X_i + \varepsilon_i$, where the entries of α and η are independently sampled from $\text{uniform}(-5, 5)$ and ε_i is drawn from $\text{MT}(4, 0, \Sigma_\varepsilon, R(\rho))$ with an AR(1) correlation structure and $\rho = 0.5$. The number of subjects is $n = 50, 100$, or 200. All parameters, covariates X , and response Y are regenerated for each synthetic dataset.

For each synthetic dataset, we apply three methods separately:

- RLMM: Robust Bayesian linear mixed-effects model assuming t -distributed random errors,
- LEM: Longitudinal envelope model assuming normally distributed random errors, and
- RoLEM: Robust longitudinal envelope model assuming t -distributed random errors.

Note that RLMM does not include an envelope structure. We therefore compare the following scenarios in the simulation study:

- Number of subjects: $n = 50, 100, 200$.
- Number of time points: $J = 5, 10$.
- Estimation method: RLMM, LEM, RoLEM.

Assume that the envelope dimension is known to be 3 and the correlation structure is known to be AR(1). Select the hyperparameters according to the nearly non-informative priors described in Section 4.2. Discard the first 10,000 burn-in samples, and then retain one out of every ten from the subsequent 100,000 samples. The estimates of the parameters are the posterior means. We evaluate the performance of the estimates of β and Σ_ε , which are denoted by $\hat{\beta}$ and $\hat{\Sigma}_\varepsilon$, using the Frobenius norm of their differences,

$$D(\hat{\beta}, \beta) = \|\hat{\beta} - \beta\|_F, \quad D(\hat{\Sigma}_\varepsilon, \Sigma_\varepsilon) = \|\hat{\Sigma}_\varepsilon - \Sigma_\varepsilon\|_F,$$

where $\|M\|_F = \sqrt{\text{tr}(M^T M)}$ is the Frobenius norm of a matrix M . Additionally, we compute the 95% highest posterior density (HPD) intervals for entries of β .

Based on 500 synthetic datasets, Figure 1 compares the performance of estimating β . Their numerical values are provided in the supplementary material. Figure 1 shows that estimation accuracy improves as either the number of subjects or the number of time points increases, reflecting the increase in total sample size. RoLEM consistently outperforms LEM because the true random errors follow a t -distribution, whereas LEM assumes normally distributed errors. Both RoLEM and LEM outperform RLMM, demonstrating that identification of the underlying envelope structure improves estimation efficiency. When $n = 50$ and $J = 5$, the performance of RLMM is substantially worse than in other scenarios, highlighting the importance of capturing the envelope structure, particularly for smaller sample sizes.

Figure 2 compares the performance of estimating Σ_ε . LEM exhibits much larger variability than RLMM and RoLEM because it assumes normal random errors, whereas the synthetic datasets are generated with t -distributed random errors. The performances of RLMM and RoLEM are comparable in most scenarios, except when $n = 200$, where RoLEM performs better due to the efficiency gained from exploiting the envelope structure.

Figure 3 displays boxplots of the lengths of the 95% HPD intervals for β . Since β has rp components, each boxplot contains all rp components across 500 replicates. The results show that RoLEM generally yields slightly shorter HPD intervals than LEM, and both have substantially shorter intervals than RLMM. This highlights the advantage of the envelope structure, as it effectively removes variability in the responses that is unrelated to the covariates once the envelope is correctly identified.

We then compare the empirical coverage probabilities of the HPD intervals. For each component of β , the empirical coverage probability is computed as the proportion of replicates in which the true value falls within the corresponding 95% HPD interval. Figure 4 displays the empirical coverage probabilities for all components of β . Ideally, these probabilities should be close to the nominal value of 0.95. The empirical coverage of LEM is consistently below 0.95, due to its assumption of normal random errors. RLMM shows slightly lower coverage when $n = 50$ and $J = 5$ but aligns with the nominal level as the sample size increases. RoLEM performs as expected, with empirical coverages closely matching the nominal value.

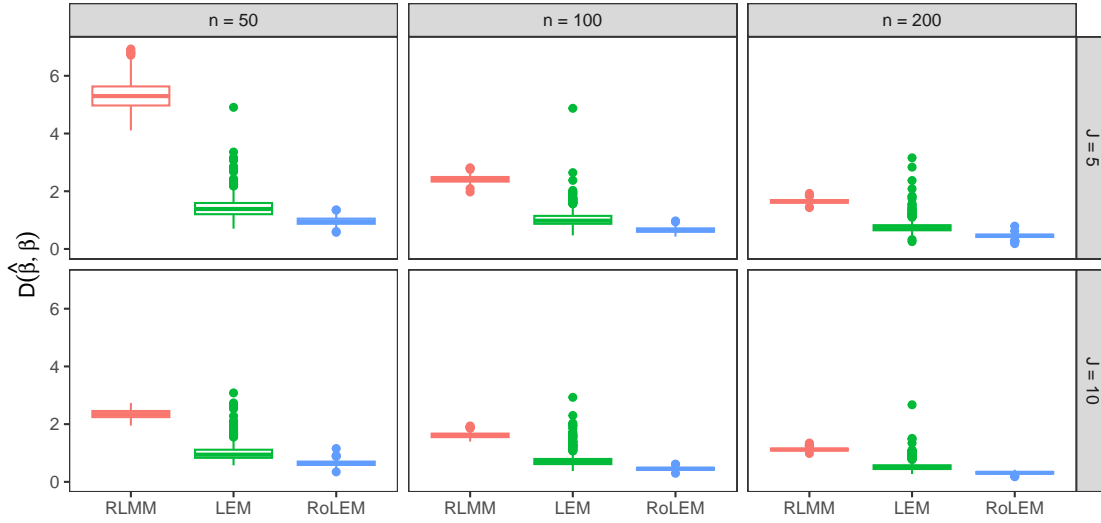


Figure 1: Performance of estimating β using RLMM, LEM, and RoLEM. Columns correspond to different sample sizes $n = 50, 100, 200$, and rows correspond to different time points $J = 5$ and 10 .

In summary, RoLEM outperforms LEM because it accounts for t -distributed random errors, and it outperforms RLMM because leveraging the envelope structure enhances estimation efficiency. The strength of envelope models lies not in improving the estimation of Σ_ε , but in identifying the component of variability in responses that is relevant to β . This, in turn, leads to improved estimation of β and substantially shorter HPD intervals.

5.2 Robustness

To evaluate the robustness of the proposed methods, we follow the data generation procedure in Section 5.1, but consider alternative ways to generate random errors. Set $\varepsilon_i = L\epsilon_i B \in \mathbb{R}^{r \times J}$, where $\Sigma_\varepsilon = LL^T$ and $R(\rho) = BB^T$ and

- Normal: the components of ϵ_i are iid $N(0, 2)$. In this case, $\varepsilon_i \sim \text{MT}(0, 2\Sigma_\varepsilon, R(\rho))$.
- Mixture normal: the components of ϵ_i are iid from a mixture-normal distribution $0.9N(0, 1) + 0.1N(0, 11)$.

Both the normal distribution and mixture-normal distribution are selected to have the same covariance matrix as the t -distribution used in Section 5.1.

Repeat the simulation 500 times and evaluate the performance of estimating β and Σ_ε , as in Section 5.1. Figures 5 and 6 correspond to scenarios with normally distributed random errors, while Figures 7 and 8 correspond to scenarios with mixture-normal random errors. The numerical results are provided in the supplementary material. Because the performance of estimating Σ_ε and the coverage probabilities of the HPD intervals exhibit similar patterns, they are provided in the supplementary material to save space.

When the random errors follow a normal distribution, Figure 5 shows that RoLEM and LEM perform comparably in estimating β , since the t -distribution includes the normal distribution as a special case. Both RoLEM and LEM still outperform RLMM, highlighting the advantage of

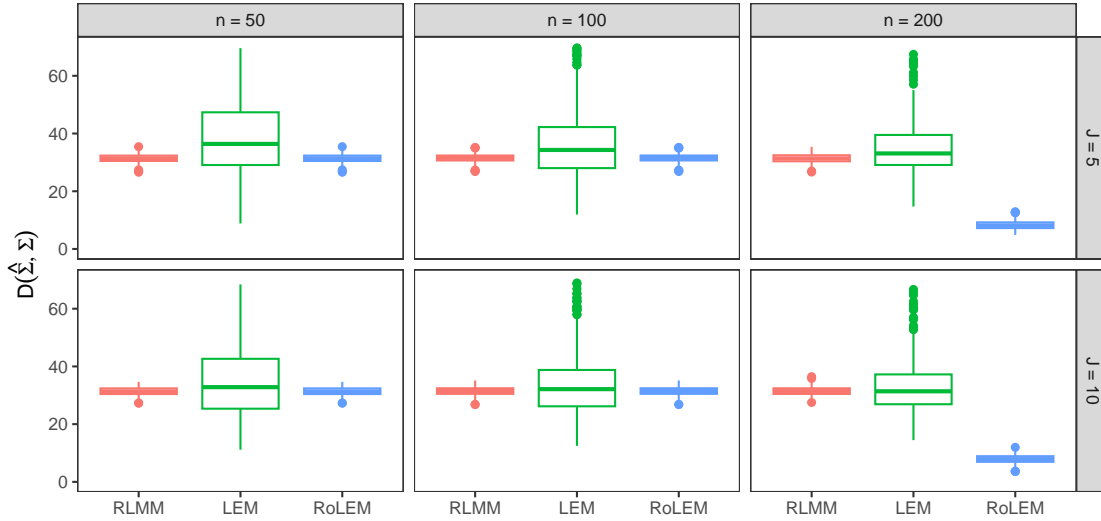


Figure 2: Performance of estimating Σ_ε using RLMM, LEM, or RoLEM. Columns correspond to different sample sizes $n = 50, 100, 200$, and rows correspond to different time points $J = 5$ and 10 .

leveraging the envelope structure to improve estimation efficiency. RLMM has slightly better performance than RoLEM and LEM in estimating Σ_ε , although this advantage decreases as the sample size increases. Nevertheless, Figure 6 shows that RLMM consistently produces much longer HPD intervals than RoLEM and LEM across all scenarios. The empirical coverage probabilities of the HPD intervals for all three methods are generally close to the nominal value of 0.95, with slight deviations observed when $n = 50$ and $J = 5$.

When the random errors follow a mixture-normal distribution, Figure 7 shows that RoLEM and LEM perform comparably in estimating β , and both outperform RLMM, highlighting the advantage of identifying the envelope structure. RoLEM slightly outperforms LEM and RLMM in estimating Σ_ε . Figure 8 shows that RoLEM and LEM provide much shorter HPD intervals than RLMM. The empirical coverage probabilities of the HPD intervals for all three methods are generally close to the nominal value of 0.95, with slight deviations observed when $n = 50$.

In summary, RoLEM demonstrates strong overall performance. When the random errors follow a normal distribution, it automatically adapts and achieves performance comparable to LEM. Even when the random errors follow a mixture-normal distribution, RoLEM maintains its robust performance. By identifying the envelope structure, RoLEM provides substantially shorter HPD intervals while preserving empirical coverage probabilities compared to standard methods.

5.3 Model selection

This subsection illustrates the performance of model selection using BIC and WAIC. We generate synthetic datasets as described in Section 5.1. The true envelope dimension is $u = 3$, and the correlation structure is AR(1).

We first assume that the correlation structure is known to be AR(1) and focus on selecting the envelope dimension u . For each synthetic dataset, we fit models with different values of u and compute the corresponding BIC and WAIC. The model with the smallest information criterion is selected. The left part of Table 2 reports the proportion of times each u is selected across 500

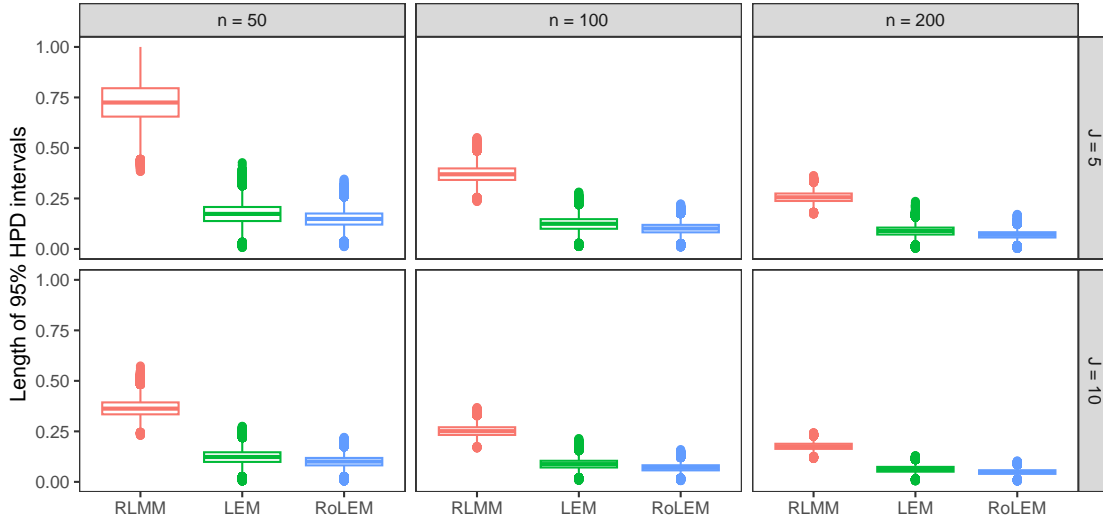


Figure 3: Length of the 95% HPD intervals for β estimated by RLMM, LEM, and RoLEM. Columns correspond to different sample sizes $n = 50, 100, 200$, and rows correspond to different time points $J = 5$ and 10 .

replicates. Both BIC and WAIC correctly identify $u = 3$ in most scenarios.

Next, we assume that the envelope dimension is known to be $u = 3$ and focus on selecting the correlation structure from UNCOR, CS, and AR(1), where UNCOR denotes the uncorrelated case or $\rho = 0$. The right part of Table 2 shows the proportion of times each correlation structure is selected across 500 replicates. Both BIC and WAIC correctly identify AR(1) in all scenarios.

Table 2: Proportion of Selected Models Using BIC and WAIC

(n, J)	Criterion	1	2	3	4	5	6+	UNCOR	CS	AR(1)
(50, 5)	BIC	0	0	0.998	0.002	0	0	0	0	1.000
	WAIC	0	0	0.996	0.004	0	0	0	0	1.000
(50, 10)	BIC	0	0	1.000	0	0	0	0	0	1.000
	WAIC	0	0	0.998	0.002	0	0	0	0	1.000
(100, 5)	BIC	0	0	1.000	0	0	0	0	0	1.000
	WAIC	0	0	1.000	0	0	0	0	0	1.000
(100, 10)	BIC	0	0	1.000	0	0	0	0	0	1.000
	WAIC	0	0	1.000	0	0	0	0	0	1.000
(200, 5)	BIC	0	0	1.000	0	0	0	0	0	1.000
	WAIC	0	0	1.000	0	0	0	0	0	1.000
(200, 10)	BIC	0	0	1.000	0	0	0	0	0	1.000
	WAIC	0	0	1.000	0	0	0	0	0	1.000

When both u and the correlation structure are unknown, BIC and WAIC can be compared for models with each possible combination of u and correlation structure. Although this approach is feasible, it requires substantially more computation. We do not explore it in the simulation studies but demonstrate it in the real data analysis.

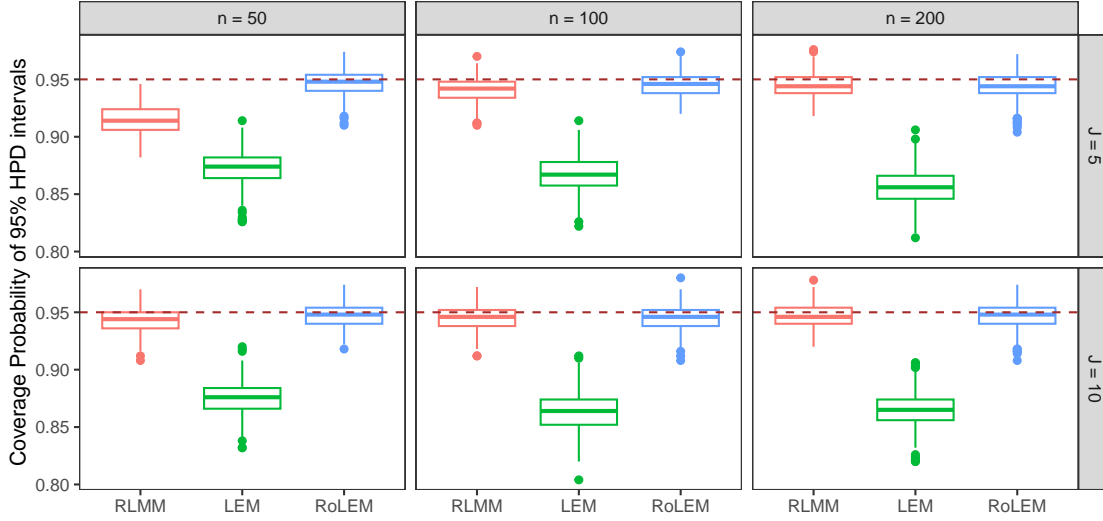


Figure 4: Empirical coverage probabilities of the 95% HPD intervals for β using RLMM, LEM, and RoLEM. Columns correspond to different sample sizes $n = 50, 100, 200$, and rows correspond to different time points $J = 5$ and 10 .

5.4 Prior distribution

This subsection demonstrates the impact of incorporating prior knowledge on the estimation of β . We modify the data generation procedure in Section 5.1 slightly by fixing A to be $A = (a, I_3, a, I_3, a, I_3, a, I_3, a)^T \in \mathbb{R}^{17 \times 3}$, where $a = (-1, 1, 1)^T$. With this setup, the envelope is spanned by $\gamma_1 = 1_5 \otimes (1, 1, 1, 1)^T / \sqrt{20}$, $\gamma_2 = 1_5 \otimes (1, -1, 1, -1)^T / \sqrt{20}$, and $\gamma_3 = 1_5 \otimes (1, 1, -1, -1)^T / \sqrt{20}$, which are three mutually orthogonal unit-length vectors in \mathbb{R}^{20} . Consider the following four different hyperparameters in the prior for P .

$$\begin{aligned} M_1 &= s_0 I_{20}, & M_2 &= s_1 \gamma_1 \gamma_1^T + s_0 I_{20}, \\ M_3 &= s_1 (\gamma_1 \gamma_1^T + \gamma_2 \gamma_2^T) + s_0 I_{20}, & M_4 &= s_1 (\gamma_1 \gamma_1^T + \gamma_2 \gamma_2^T + \gamma_3 \gamma_3^T) + s_0 I_{20}. \end{aligned}$$

These four scenarios represent gradually increasing amounts of prior knowledge about P . In the simulation, we set $s_1 = 10^5$ and $s_0 = 10^{-6}$.

Figure 9 displays the performance of estimating β based on 500 replicates. As the prior distribution incorporates more information about P , the accuracy of estimating β improves across all scenarios, demonstrating that prior knowledge has been effectively leveraged in the analysis.

5.5 Real data analysis

Existing literature has suggested that diabetes mellitus may be a potential risk factor for knee osteoarthritis [10]. This study aims to evaluate the impact of diabetes mellitus on the physical and psychological health outcomes of patients with osteoarthritis in a longitudinal setting, accounting for potential confounding variables such as body mass index (BMI), osteoarthritis severity, and pain medication use.

The dataset is extracted from the Osteoarthritis Initiative (OAI) database (<https://nda.nih.gov/oai>), which is a longitudinal cohort study of individuals with or at increased risk for knee osteoarthritis.

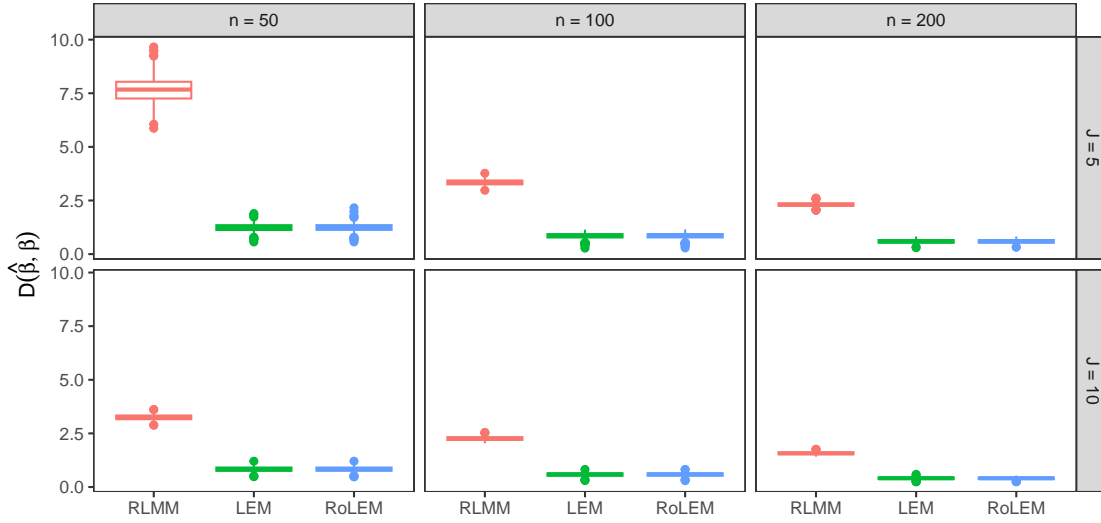


Figure 5: Performance of estimating β using RLMM, LEM, and RoLEM when the random errors follow a normal distribution. Columns correspond to different sample sizes $n = 50, 100, 200$, and rows correspond to different time points $J = 5$ and 10 .

tis. The impact of osteoarthritis on patients can be evaluated through both physical and mental health measures. In this study, we focus on knee pain (KOOS and NRS), physical activity level (PASE), physical health status (PCS), and mental health status (MCS and CESD). The primary covariate of interest is the presence of diabetes, and additional covariates include sex, age, BMI, pain medication use, and osteoarthritis severity (KL grade). Descriptions of all variables are provided in Table 3. After excluding observations with missing values, the final dataset consists of 221 osteoarthritis patients assessed at baseline and at 12-, 24-, 36-, and 48-month follow-ups. Thus, the analysis involves $r = 6$ response variables and $p = 6$ covariates, with each patient contributing $J = 5$ visits.

Table 3: List of Variables in the OAI Data

Variable	Description	Values
CESD	Center for Epidemiologic Studies Depression Scale	0 – 60
KOOS	Knee Injury and Osteoarthritis Outcome Score	0 – 100
MCS	Mental Component Summary from SF-12	0 – 100
NRS	Pain severity over past 30 days	0 – 10
PASE	Physical Activity Scale for the Elderly	0 – 793
PCS	Physical Component Summary from SF-12	0 – 100
Age	Age in years	45 – 83
BMI	Body Mass Index	18.50 – 41.10
Diabetes	Diabetes presence, 0 = No, 1 = Yes	0, 1
KL	Kellgren-Lawrence grade for severity of osteoarthritis	0 – 4
Medication	Pain Medication use, 0 = No, 1 = Yes	0, 1
Sex	Sex, 0 = Male, 1 = Female	0, 1

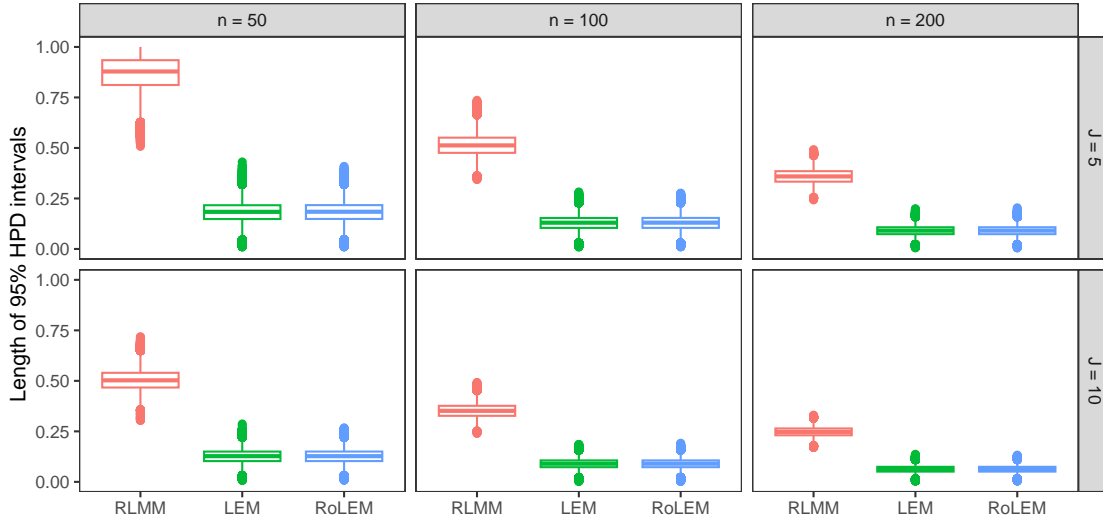


Figure 6: Length of the 95% HPD intervals for β estimated by RLMM, LEM, and RoLEM when the random errors follow a normal distribution. Columns correspond to different sample sizes $n = 50, 100, 200$, and rows correspond to different time points $J = 5$ and 10 .

Figure 10 presents the trajectories of each patient for each response variable, with the thick lines indicating the mean trajectories. Although individual patient trajectories exhibit substantial variability, the mean trajectories remain relatively stable over time. Several individual trajectories also appear to be potential outliers.

All response and covariate variables were standardized to zero-mean and unit-variance before analysis. We applied the RoLEM to analyze the data, discarding the first 50,000 MCMC samples as burn-in and then generating 250,000 additional samples, retaining one out of every 50. We compared WAIC and BIC values across models with different choices of u and correlation structures in Table 4. Models with a compound-symmetric (CS) correlation structure consistently yielded smaller WAIC and BIC values than models with other correlation structures. The model with $u = 5$ attained the lowest WAIC overall, while the model with $u = 2$ achieved the smallest BIC. The WAIC approximates the out-of-sample prediction, and the difference between the $u = 2$ and $u = 5$ models is roughly $1 - (14157.48/14167.93) = 0.07\%$. On the other hand, comparing BIC corresponds to evaluating the Bayes factor. The difference of BIC values for the $u = 2$ model and the $u = 5$ model is large enough to claim very strong evidence in favor of the former. Hence, we proceed with the analysis using the model with $u = 2$ and the CS correlation structure.

Figure 11 presents the trace plots and autocorrelation plots for the posterior draws of $A_{1,1}$ and $\eta_{1,1}$. The trace plots indicate good convergence of the MCMC chains. The draws of $\eta_{1,1}$ appear nearly independent, while the autocorrelation plot of $A_{1,1}$ decreases rapidly with increasing lag. Similar plots for the remaining parameters are provided in the supplementary material.

The posterior means of β and their associated 95% HPD intervals are shown in Figure 12, together with the corresponding estimates obtained from LEM and RLMM. The numerical values are provided in the supplementary material. The three methods yield similar estimates for some coefficients, but the lengths of the HPD intervals differ, reflecting varying levels of uncertainty. For example, there is strong evidence that the presence of diabetes is associated with higher NRS scores

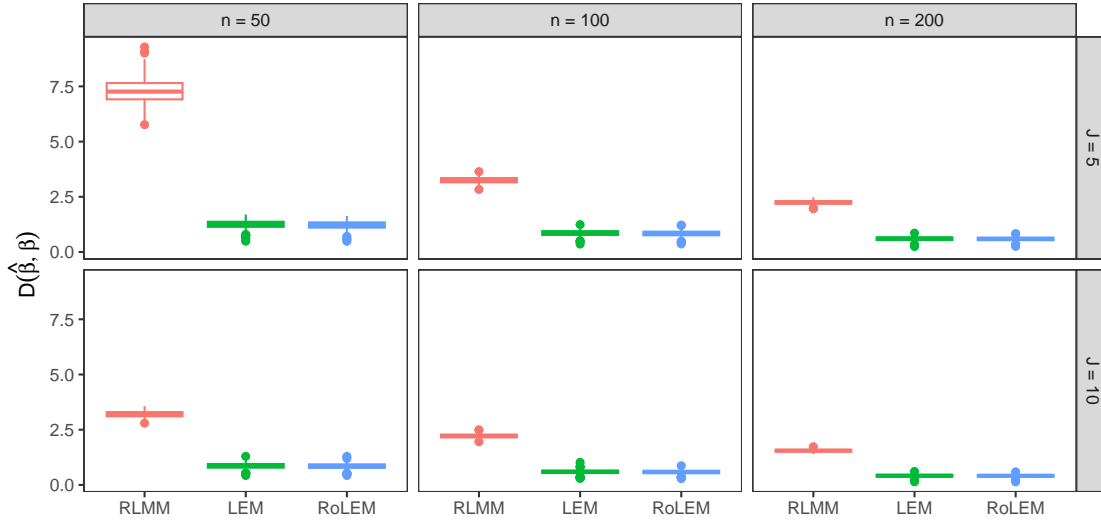


Figure 7: Performance of estimating β using RLMM, LEM, and RoLEM when the random errors follow a mixture-normal distribution. Columns correspond to different sample sizes $n = 50, 100, 200$, and rows correspond to different time points $J = 5$ and 10 .

Table 4: WAIC and BIC for Models with Different u and Correlation Structure

		Dimension				
	Correlation	1	2	3	4	5
WAIC	UNCOR	16223.55	16080.04	16081.65	16074.36	16066.37
	AR(1)	14657.11	14600.87	14599.85	14599.18	14584.64
	CS	14208.63	14167.93	14170.81	14169.99	14157.48
BIC	UNCOR	16284.09	16165.80	16184.14	16191.24	16198.57
	AR(1)	14757.39	14728.52	14746.45	14772.02	14780.39
	CS	14319.96	14301.85	14325.45	14351.61	14362.29

under RoLEM, whereas LEM and RLMM do not provide strong evidence for this association, as their 95% HPD intervals include zero.

To compare the three models, LEM, RLMM, and RoLEM, Table 5 reports their WAIC, BIC, and five-fold cross-validation (CV) scores. The CV performance is evaluated using both the mean log posterior density (MLPD) and the mean absolute error (MAE), where

$$\text{MLPD} = \frac{1}{n} \sum_{i=1}^n \log f_{\text{post}}(\tilde{y}_i), \quad \text{MAE} = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^J |\tilde{y}_{ij} - \hat{y}_{ij}|,$$

where $\tilde{y}_i = (\tilde{y}_{i1}, \dots, \tilde{y}_{iJ})$, $i = 1, \dots, n$, denote the test data, $f_{\text{post}}(\cdot)$ is the posterior predictive density and can be approximated using posterior draws, and \hat{y}_{ij} denotes the posterior mean prediction. The results show that LEM performs worse than both RLMM and RoLEM, which is expected given the presence of outliers suggested by the trajectories in Figure 10, since LEM does not accommodate outliers. RLMM and RoLEM exhibit nearly identical cross-validation performance. RLMM has a smaller WAIC, however, the difference in the WAIC is negligible compared with the magnitude of the WAIC values. RoLEM attains a smaller BIC, and the difference can translate to a Bayes

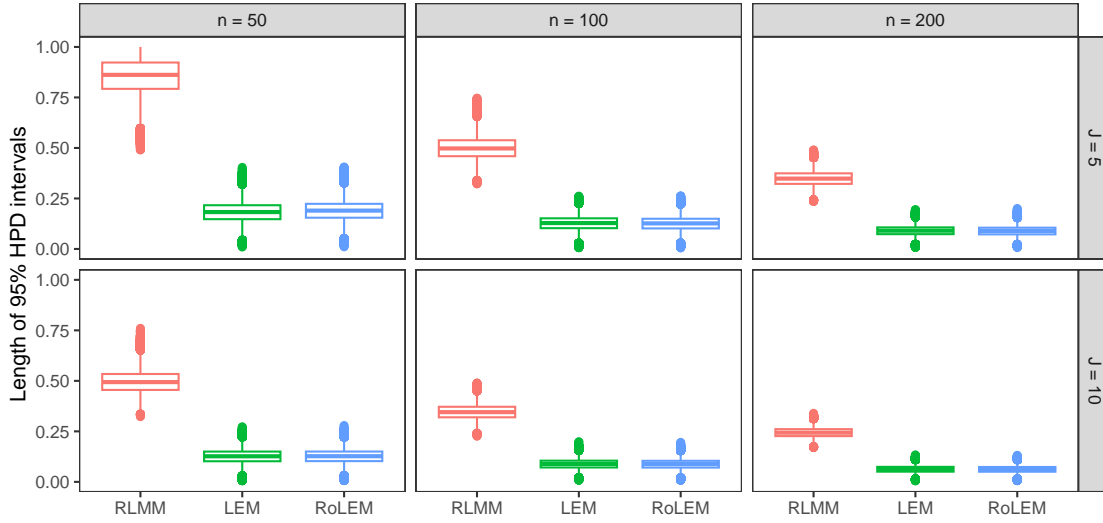


Figure 8: Length of the 95% HPD intervals for β estimated by RLMM, LEM, and RoLEM when the random errors follow a mixture-normal distribution. Columns correspond to different sample sizes $n = 50, 100, 200$, and rows correspond to different time points $J = 5$ and 10 .

factor that indicating strong evidence in favor of RoLEM. Additionally, RoLEM consistently produces shorter HPD intervals than RLMM, as shown in Figure 12 and a plot in the supplementary material. Thus, RoLEM offers a better performance over both LEM and RLMM.

Table 5: WAIC, BIC, and CV Scores for LEM, RLMM, and RoLEM

Method	WAIC	BIC	CV-MLPD	CV-MAE
LEM	14856.00	14973.44	-33.55	0.74
RLMM	14162.97	14393.41	-32.00	0.72
RoLEM	14166.57	14302.92	-32.01	0.72

We interpret the results based on the RoLEM estimates of β , which are shown in Figure 12 with numerical values provided in the supplementary material. If a 95% HPD interval excludes zero, there exists strong evidence of an association between the covariate and the response; otherwise, we conclude that no association is detected. CESD measures depressive symptoms. None of the covariates show strong evidence of association with CESD. KOOS evaluates knee pain, with higher scores indicating less pain. Higher age, lower BMI, absence of diabetes, lower KL grade, and no pain medication use are all associated with less knee pain. MCS measures overall mental health status, with higher scores reflecting better mental health. Higher BMI is associated with better mental health. NRS measures non-activity-specific pain severity over the past 30 days, with higher scores indicating more pain. Younger age, higher BMI, presence of diabetes, higher KL grade, and pain medication use are associated with more pain. PASE assesses physical activity level, where higher scores reflect greater activity. Younger age and pain medication use are associated with higher activity levels. PCS measures physical health status, with higher values indicating better health. Lower BMI, absence of diabetes, and no pain medication use are associated with better physical health. Sex is not associated with any of the response variables.

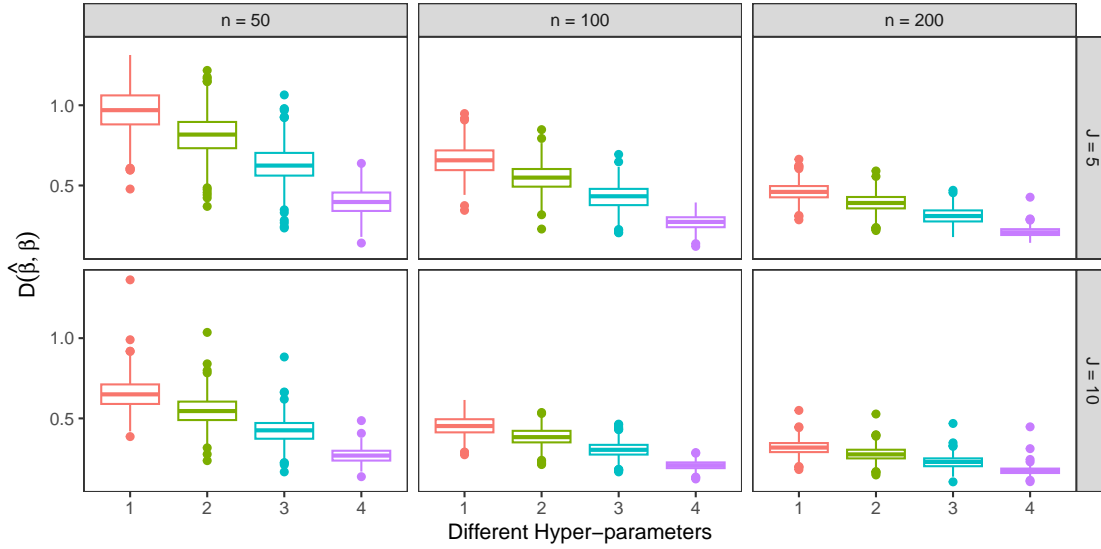


Figure 9: Performance of estimating β under different hyper-parameters in the prior for P . Columns correspond to different sample sizes $n = 50, 100, 200$, and rows correspond to different time points $J = 5$ and 10 .

Using the RoLEM, the posterior mean of ρ is 0.51 with a 95% HPD interval of (0.48, 0.53), which indicates the existence of correlation among data collected from repeated visits for the same patients. The posterior mean of ν , the degrees of freedom parameter, is 6.12 with a 95% HPD interval of (4.78, 7.45), which provides evidence for the presence of outliers in the dataset.

Finally, we explore whether model fit could be improved by including pairwise interactions among the covariates. For each candidate interaction, we added a single interaction term to the model and computed the corresponding WAIC and BIC values. The results are summarized in Table 6. Although adding certain interactions, for example, the BMI-medication interaction, yields models with smaller WAIC than the current model, none of the models with an interaction achieve a smaller BIC. Recall that a smaller WAIC indicates better out-of-sample predictive performance, whereas comparing BIC corresponds to evaluating a Bayes factor. Given that no model with interactions improves the BIC and the gains in WAIC are limited, we conclude that including interaction terms is not sufficiently beneficial and therefore do not pursue them further in this analysis.

6 Conclusion

This paper proposes RoLEM, which extends existing methods in three key aspects: (1) modeling the correlation structure among repeated measurements from the same subject; (2) modeling random errors using a scale mixture of matrix-variate normal distributions to accommodate potential outliers; and (3) introducing novel prior and proposal distributions for the envelope model. A central challenge in the Bayesian inference for RoLEM is the parameterization of Σ_ϵ and the modeling of Γ and Γ_0 . We address this by working directly with the projection matrix corresponding to the subspace spanned by Γ and imposing a prior distribution on the Grassmann manifold. Both the simulation studies and real data analysis demonstrate the superior performance of RoLEM.

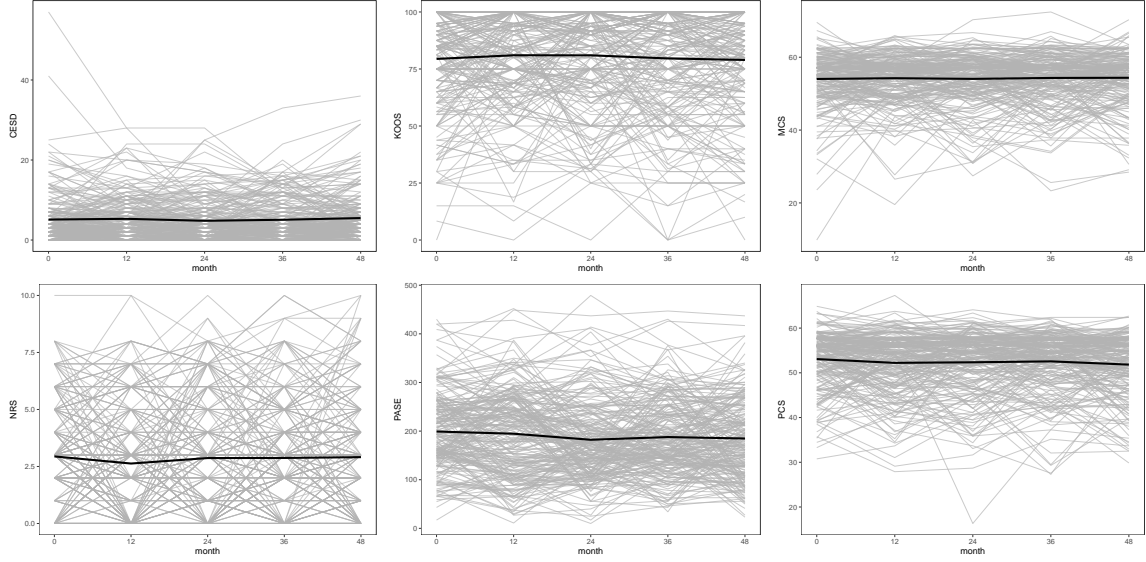


Figure 10: The trajectories of each patient for each response variable, with the thick lines indicating the mean trajectories.

Table 6: WAIC and BIC for Models with an Interaction						
		Age	BMI	Diabetes	KL	Medication
WAIC	BMI	14164.85				
	Diabetes	14166.47	14171.97			
	KL	14171.53	14163.59	14172.79		
	Medication	14169.76	14161.05	14169.61	14172.59	
	Sex	14171.77	14171.97	14166.57	14162.28	14170.04
BIC	BMI	14304.78				
	Diabetes	14310.60	14310.82			
	KL	14314.39	14305.87	14311.12		
	Medication	14312.12	14303.06	14314.86	14312.32	
	Sex	14314.32	14314.36	14302.92	14304.60	14308.64

The model relies on certain intrinsic assumptions. In the most general setting, the variance matrix of $\text{vec}(Y_i)$ is an $rJ_i \times rJ_i$ matrix. However, RoLEM assumes that it admits the structured form $R_i \otimes \Sigma_\varepsilon$, which implies that the variance of each response is constant across time points and that the temporal correlation structure stays the same for all responses. The model also assumes that the kurtosis is identical across all outcomes. When this homogeneous structure does not hold, more flexible and complex variance structures may be required. If the variances differ across time, we may introduce a time-specific error covariance $\Sigma_{\varepsilon,j}$ for the j th time point. In this case, we can impose an envelope structure of the form $\Sigma_{\varepsilon,j} = \Gamma\Omega_j\Gamma^T + \Gamma_0\Omega_{0,j}\Gamma_0^T$, which maintains a common envelope basis across time while allowing Ω_j and $\Omega_{0,j}$ to vary with time j . Under this structure, we may still assume $\beta = \Gamma\eta$. We will investigate this issue further in future research.

Missing data are also common in longitudinal studies. For simplicity, we remove subjects with incomplete observations. In practice, however, various imputation methods can be applied prior to fitting RoLEM.

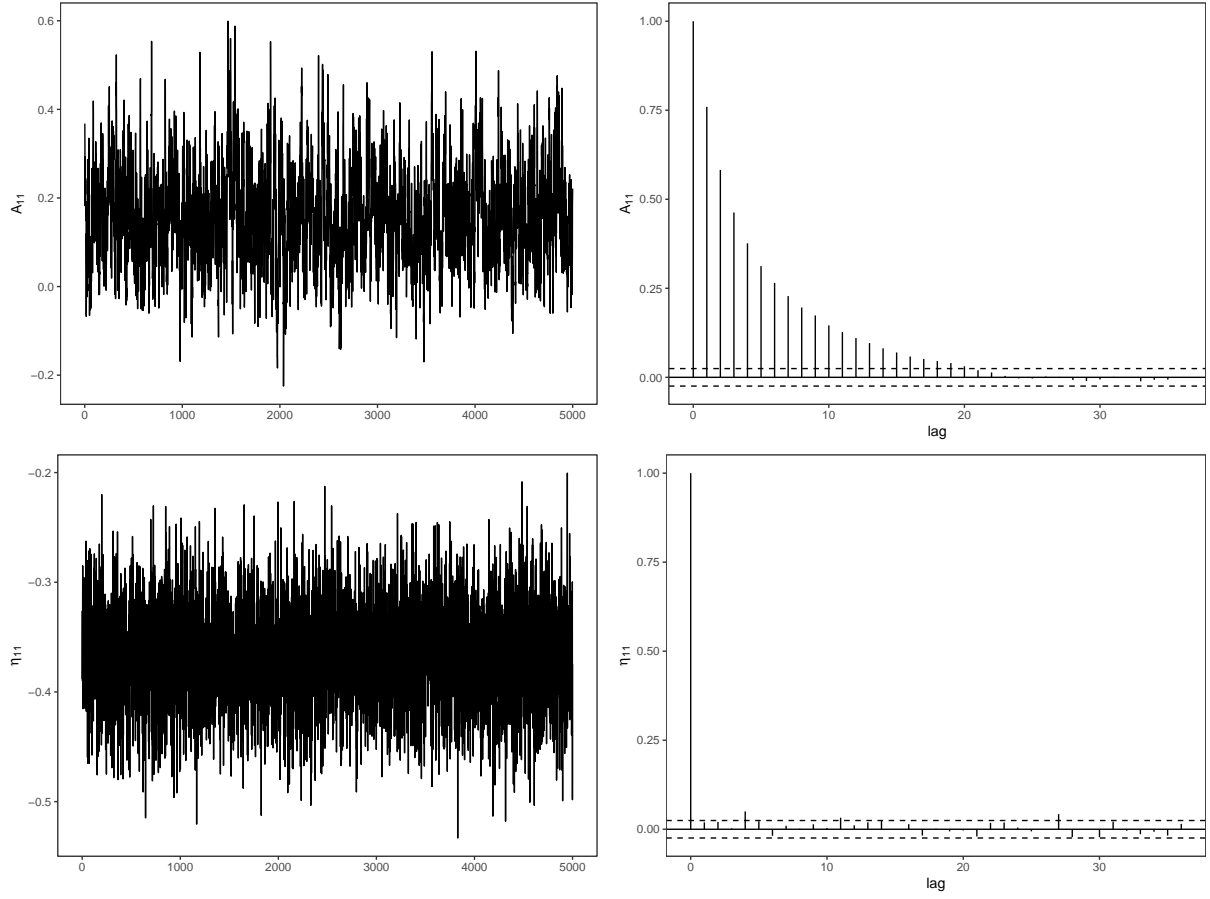


Figure 11: The trace plots and autocorrelation plots for the posterior draws of $A_{1,1}$ and $\eta_{1,1}$.

References

- [1] Adelchi Azzalini and Antonella Capitanio. *The Skew-Normal and Related Families*. Cambridge University Press, 2014.
- [2] Saptarshi Chakraborty and Zhihua Su. A comprehensive Bayesian framework for envelope models. *Journal of the American Statistical Association*, 119(547):2129–2139, 2024.
- [3] Yasuko Chikuse. *Statistics on Special Manifolds*. Springer, 2003.
- [4] R Dennis Cook. *An Introduction to Envelopes: Dimension Reduction for Efficient Estimation in Multivariate Statistics*. John Wiley & Sons, 2018.
- [5] R Dennis Cook, Bing Li, and Francesca Chiaromonte. Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica*, 20(3):927–960, 2010.
- [6] R Dennis Cook and Xin Zhang. Foundations for envelope models and methods. *Journal of the American Statistical Association*, 110(510):599–611, 2015.
- [7] Peng Ding. Bayesian robust inference of sample selection using selection- t models. *Journal of Multivariate Analysis*, 124:451–464, 2014.

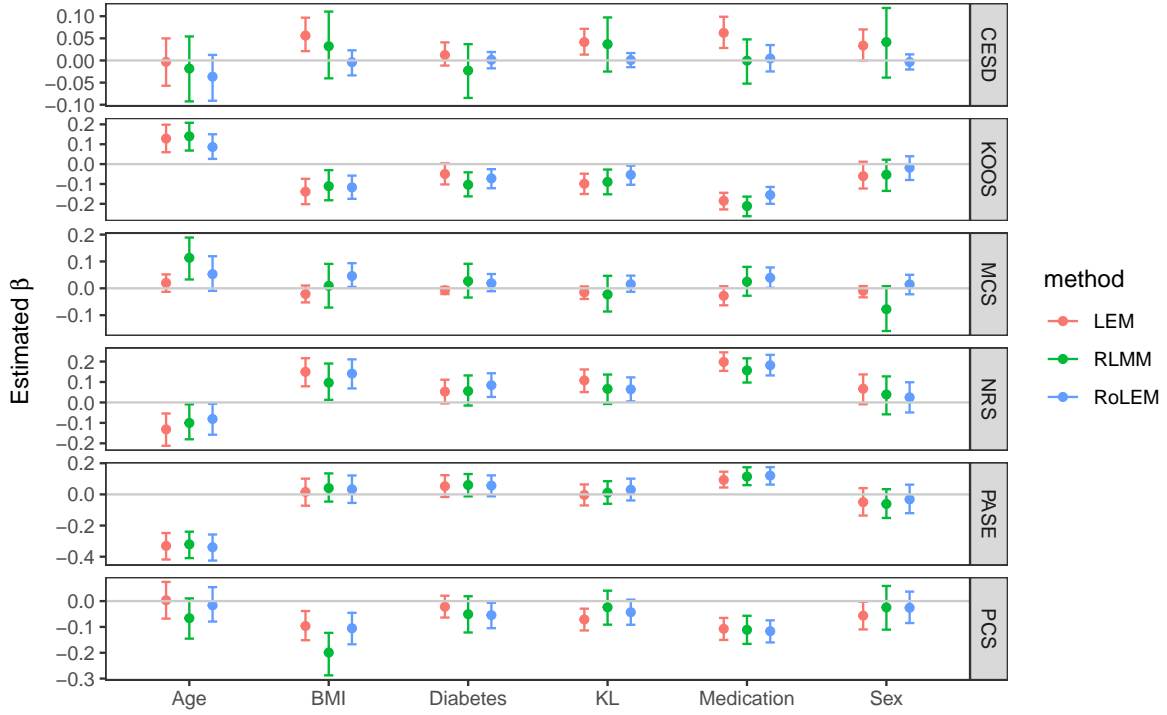


Figure 12: The posterior means of β and their associated 95% HPD intervals obtained from LEM, RLMM, and RoLEM.

- [8] Shanshan Ding and R Dennis Cook. Matrix variate regressions and envelope models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(2):387–408, 2018.
- [9] Shanshan Ding, Zhihua Su, Guangyu Zhu, and Lan Wang. Envelope quantile regression. *Statistica Sinica*, 31(1):79–105, 2021.
- [10] Annett Eitner, Adam G. Culvenor, Wolfgang Wirth, Hans-Georg Schaible, and Felix Eckstein. Impact of diabetes mellitus on knee osteoarthritis pain and physical and mental status: Data from the osteoarthritis initiative. *Arthritis Care & Research*, 73(4):540–548, 2021.
- [11] Carmen Fernández and Mark F J Steel. Multivariate student- t regression models: Pitfalls and inference. *Biometrika*, 86(1):153–167, 1999.
- [12] Thaís C O Fonseca, Marco A R Ferreira, and Helio S Migon. Objective bayesian analysis for the student- t regression model. *Biometrika*, 95(2):325–333, 2008.
- [13] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 3rd edition, 2013.
- [14] A. K. Gupta and D. K. Nagar. *Matrix Variate Distributions*. Chapman & Hall/CRC, Boca Raton, 2000.
- [15] Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

- [16] Kshitij Khare, Subhadip Pal, and Zhihua Su. A Bayesian approach for envelope models. *The Annals of Statistics*, 45(1):196–222, 2017.
- [17] Minji Lee and Zhihua Su. A review of envelope models. *International Statistical Review*, 88(3):658–676, 2020.
- [18] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, pages 461–464, 1978.
- [19] Yuyang Shi, Linquan Ma, and Lan Liu. Mixed effects envelope models. *Stat*, 9(1):e313, 2020.
- [20] Wan-Lun Wang and Tsai-Hung Fan. Bayesian analysis of multivariate t -linear mixed models using a combination of IBF and Gibbs samplers. *Journal of Multivariate Analysis*, 105(1):300–310, 2012.
- [21] Sumio Watanabe. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(12):3571–3594, 2010.