mathematics

MDPI

Article

Variational Bayesian Inference in High-Dimensional Linear Mixed Models

Jieyi Yi and Niansheng Tang *

Yunnan Key Laboratory of Statistical Modeling and Data Analysis, Yunnan University, Kunming 650091, China; yijy@mail.ynu.edu.cn

* Correspondence: nstang@ynu.edu.cn; Tel.: +86-871-65032416

Abstract: In high-dimensional regression models, the Bayesian lasso with the Gaussian spike and slab priors is widely adopted to select variables and estimate unknown parameters. However, it involves large matrix computations in a standard Gibbs sampler. To solve this issue, the Skinny Gibbs sampler is employed to draw observations required for Bayesian variable selection. However, when the sample size is much smaller than the number of variables, the computation is rather time-consuming. As an alternative to the Skinny Gibbs sampler, we develop a variational Bayesian approach to simultaneously select variables and estimate parameters in high-dimensional linear mixed models under the Gaussian spike and slab priors of population-specific fixed-effects regression coefficients, which are reformulated as a mixture of a normal distribution and an exponential distribution. The coordinate ascent algorithm, which can be implemented efficiently, is proposed to optimize the evidence lower bound. The Bayes factor, which can be computed with the path sampling technique, is presented to compare two competing models in the variational Bayesian framework. Simulation studies are conducted to assess the performance of the proposed variational Bayesian method. An empirical example is analyzed by the proposed methodologies.

Keywords: Bayesian lasso; evidence lower bound; high-dimensional linear mixed model; spike and slab priors; variational Bayesian inference

Citation: Yi, J.; Tang, N. Variational Bayesian Inference in High-Dimensional Linear Mixed Models. *Mathematics* **2022**, *10*, 463. https://doi.org/10.3390/math10030463

Academic Editors: Vitaly Schetinin, Livija Jakaite and Dayou Li

Received: 24 December 2021
Accepted: 27 January 2022
Published: 31 January 2022

1. Introduction

Linear mixed models are widely used to analyze longitudinal and correlated data by considering the between-subject and within-subject variations and incorporating the random effects to account for heterogeneity among the subjects in many fields, such as psychology, medicine, epidemiology and econometrics. Various methods have been developed to estimate fixed-effects parameters and variance–covariance matrices for unobservable random effects and noises or select fixed-effects and random-effects components, even if it is quite challenging for the problem of variable selection and parameter estimation in linear mixed models. For example, see [1] for restricted maximum likelihood estimation of parameters, ref [2] for EM algorithm of parameter estimation, refs [3,4] for Bayesian parameter estimation, ref [5] for Bayesian random effects selection and [6] for moment-based method for random effects selection. The aforementioned methods mainly focus on low-dimensional linear mixed models, while high-dimensional data have become increasingly common with the rapid development of modern information technologies that facilitate data collection. Thus, the aforementioned methods do not work well in high-dimensional linear mixed models, and so some penalized methods have developed to simultaneously estimate parameters and select variables in high-dimensional linear mixed models. For example, Bondell, Krishna and Ghosh [7] and Ibrahim et al. [8] proposed the penalized likelihood methods for joint selection of fixed and random effects; Schelldorfer, Buhlmann and van De Geer [9] proposed an ℓ_1–penalized estimation procedure; Fan and Li [10] investigated the problem of fixed and random effects selection when the cluster

sizes are balanced; Li et al. [11] presented a doubly regularized approach to simultaneously select fixed and random effects; Bradic, Claeskens and Gueuning [12] considered testing a single parameter of fixed effects in high-dimensional linear mixed models with fixed cluster sizes, fixed numbers of random effects and sub-Gaussian designs; Li, Cai and Li [13] proposed a penalized quasi-likelihood method for statistical inference on unknown parameters in high-dimensional linear mixed-effects models. However, the aforementioned regularization methods are computationally complex and unstable and they do not consider the prior information of fixed-effects parameters and variance–covariance matrices, which may lead to unsatisfactory estimation accuracy of parameters or variance–covariance matrices. Bayesian approaches for variable selection and parameter estimation have received much attention over the past years because they can largely improve the accuracy and efficiency of parameter estimation, consistently select important variables and provide more information for variable selection than the corresponding penalization method with a highly non-convex optimization problem by imposing various priors on model parameters. For example, see [14] for reference prior, ref [15] for normal mixture prior, ref [16] for spike and slab prior, ref [17] for horseshoe prior and [18] for shrinking and diffusing prior. In the high-dimensional setting, Bayesian lasso, Bayesian adaptive lasso or the indicator model method, together with the Markov chain Monte Carlo (MCMC) algorithm, are widely used to select important variables. For example, see [19] for Bayesian lasso, ref [20] for Bayesian adaptive lasso and [21,22] for the EM approach in the Bayesian framework. The above-mentioned literature involves the implementation of the standard Gibbs sampler for posterior computation, which is not so scalable for large numbers of fixed-effects components [23]. To address the issue, Narisetty, Shen and He [23] proposed a Skinny Gibbs algorithm by using a sparse matrix to replace the high-dimensional variance–covariance matrix, which avoids large matrix operations. However, implementing the above MCMC algorithm in the presence of high-dimensional data may still be subject to the well-known ill-posed problems, i.e., low efficiency, slow convergence and huge memory being required.

As an alternative to the MCMC, the variational Bayesian method, also called ensemble learning, is widely adopted to approximate intractable integrals involved in Bayesian inference or machine learning due to its good properties, such as high-speed computation. Its basic idea is to transform the high-dimensional integration problem into an optimization problem in making Bayesian inference and then optimize the evidence lower bound (ELB), which is efficiently computed, and finally utilize the ELB to obtain a variational approximation to the posterior distribution in Bayesian analysis. The variational Bayesian approach has been applied to some familiar models, for example, latent variable models [24], mixtures of factor analysis [25], graphical models [26] and partially linear mean shift models with high-dimensional data [27].

Motivated by the aforementioned variational Bayesian studies, we develop a novel variational Bayesian approach to estimate model parameters and select important variables under the Skinny Gibbs sampling framework in a linear mixed model with low-dimensional random effects and high-dimensional fixed effects. We specify the spike and slab priors for the population-specific fixed-effects regression coefficients with completely different shrinkage parameters, which overcomes the problem of selecting a high-dimensional vector of the shrinkage parameters. We reformulate the spike and slab priors of parameter as a mixture of a normal distribution and an exponential distribution, which avoids the high-dimensional integral problem. The coordinate ascent algorithm, which can be implemented efficiently, is proposed to optimize the ELB. The Bayes factor, which can be computed with the path sampling technique, is presented to compare two competing models in the variational Bayesian framework. The merits of the proposed variational Bayesian method are (i) simultaneously estimating parameters and variance–covariance matrices and select fixed- and random-effects components with quite a low computation cost, (ii) efficiently analyzing high-dimensional data without requiring the non-convex optimization and avoiding the curse of dimensionality problem, (iii) automatically incorporating the shrinkage parameters and (iv) avoiding large matrix computations.

The rest of the article is organized as follows: Section 2 introduces the linear mixed model setup, including the spike and slab priors. Section 3 describes the Skinny Gibbs sampler algorithm for selecting fixed- and random-effects components and estimating parameters in coefficients and variance–covariance matrices via the Bayesian Lasso method. Section 4 develops a variational Bayesian approach to approximate posterior distributions of parameters and random effects and presents the Bayes factor for model comparison. The coordinate ascent algorithm is adopted to optimize the ELB in Section 4. Simulation studies are considered in Section 5. An empirical example is illustrated by the proposed methodologies in Section 6. A simple discussion is given in Section 7. Technique details are presented in the Appendix A, Appendix B and Appendix C.

2. Model

Consider a dataset with n subjects. For the ith subject, let y_{ij} be the observation of the response variable, $\boldsymbol{x}_{ij}$ be a $p \times 1$ vector of covariates associated with the fixed effects and $\boldsymbol{z}_{ij}$ be a $q \times 1$ vector of covariates associated with the random effects, which may be a subvector of $\boldsymbol{x}_{ij}$ for $j = 1, \ldots, n_i$, where n_i is the number of times observed repeatedly for the ith subject. Generally, n_i varies across subjects. For simplicity, we suppose that y_{ij} has been centered at zero for avoiding the requirement of intercept and $n_1 = \ldots = n_n = m$, i.e., the balanced design. It is assumed that $p \gg n$ and only a small number of covariates $\boldsymbol{x}_{ij}$ contribute to response variable y_{ij}, i.e., $\boldsymbol{x}_{ij}$ has sparsity, while q is smaller than n.

For the dataset $\mathcal{D} = \{(y_{ij}, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}) : i = 1, \ldots, n,\ j = 1, \ldots, m\}$, we consider the following linear mixed model:

$$y_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top \boldsymbol{b}_i + \varepsilon_{ij},\ \ i = 1, \ldots, n,\ j = 1, \ldots, m, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is a $p \times 1$ vector of population-specific fixed-effects regression coefficients, $\boldsymbol{b}_i$ is a $q \times 1$ vector of subject-specific random effects and ε_{ij} is measurement error. Here, we assume that $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$ are independent and identically distributed (i.i.d.) as the multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{Q}$ and ε_{ij}'s are independently distributed as $\mathcal{N}(0, \sigma_j^2)$, where $\mathcal{N}(\cdot, \cdot)$ represents the normal distribution. Here, $\sigma_1^2, \ldots, \sigma_m^2$ are not completely different but some of them may be identical.

Under the aforementioned assumptions, a penalized likelihood approach to estimate $\boldsymbol{\beta}$ is implemented by

$$\widehat{\boldsymbol{\beta}} = \arg\max_{\beta \in \mathbb{R}^p} \left[-\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(y_{ij} - \boldsymbol{x}_{ij}^\top \boldsymbol{\beta})^2}{\sigma_j^2 + \boldsymbol{z}_{ij}^\top \boldsymbol{Q} \boldsymbol{z}_{ij}} + f_\lambda(\boldsymbol{\beta}) \right], \tag{2}$$

where $f_\lambda(\boldsymbol{\beta})$ is some appropriate penalty function indexed by the penalty parameter λ. In variable selection literature, it is usually assumed that $f_\lambda(\boldsymbol{\beta})$ has the form: $f_\lambda(\boldsymbol{\beta}) = \sum_{k=1}^{p} f_{\lambda_k}(\beta_k)$, where $f_{\lambda_k}(\beta_k)$ takes the ℓ_0-norm, ℓ_1-norm, MCP penalty [28], SCAD penalty [29] and Elastic-Net penalty [30]. Recently, it was widely recognized that $\widehat{\boldsymbol{\beta}}$ can be regarded as a posterior mode of $\boldsymbol{\beta}$ with some proper prior. Inspired by this idea, we consider Bayesian variable selection procedure based on some proper prior on $\boldsymbol{\beta}$.

Following [31], we consider the following spike and slab prior of $\boldsymbol{\beta}$:

$$f(\boldsymbol{\beta} | \boldsymbol{\gamma}, \lambda_0, \lambda_1) = \prod_{k=1}^{p} \{\gamma_k g_1(\beta_k | \lambda_1) + (1 - \gamma_k) g_0(\beta_k | \lambda_0)\}, \tag{3}$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^\top$, in which γ_k is a binary latent variable and follows a Bernoulli distribution with the probability $\rho_k = \Pr(\gamma_k = 1)$, i.e., $\gamma_k = 1$ indicates that the kth covariate is active and $\gamma_k = 0$ implies that the kth covariate is inactive and $g_1(\beta_k | \lambda_1)$ is usually referred to as a diffuse slab prior reflecting the effect of an active covariate, while $g_0(\beta_k | \lambda_0)$ is called a concentrated spike prior reflecting the negligibly unimportant effect

of an inactive covariate for $k = 1, \ldots, p$. Let $f(\boldsymbol{\gamma}|\boldsymbol{\rho})$ be the prior distribution of $\boldsymbol{\gamma}$ indexed by $\boldsymbol{\rho}$. It is assumed that $f(\boldsymbol{\gamma}|\boldsymbol{\rho})$ has the form

$$f(\boldsymbol{\gamma}|\boldsymbol{\rho}) = \prod_{k=1}^{p} \rho_k^{\gamma_k}(1-\rho_k)^{1-\gamma_k}, \tag{4}$$

where $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_p)^\top$. For simplicity, we assume $\rho_1 = \ldots = \rho_p = \rho$, which is the expected proportion of the active covariates. Generally, one can take $g_0(\cdot)$ and $g_1(\cdot)$ as the normal distribution with a small and a large variance, respectively. However, for the spike and slab lasso, we take the following slab and spike priors

$$g_1(\beta_k|\lambda_1) = \frac{\lambda_1}{2} e^{-\lambda_1|\beta_k|}, \;\; g_0(\beta_k|\lambda_0) = \frac{\lambda_0}{2} e^{-\lambda_0|\beta_k|}, \tag{5}$$

respectively, where λ_1 should tend to zero and λ_0 should tend to ∞ as the sample size is sufficiently large, which implies that the inactive covariates will be detected as zeros in that small values of β_k relative to $1/\lambda_0$ or λ_1 are truncated to zero. Following [32], the density $g_\ell(\beta_k|\lambda_\ell) = \frac{\lambda_\ell}{2}\exp(-\lambda_\ell|\beta_k|)$ can be hierarchically written as a mixture of a normal distribution and an exponential distribution, i.e.,

$$\beta_k|\xi_{\ell k}^2, \gamma_k = \ell \sim \mathcal{N}(0, \xi_{\ell k}^2), \quad \xi_{\ell k}^2|\lambda_\ell^2 \sim \text{Exp}(\lambda_\ell^2/2), \;\; \ell = 0, 1. \tag{6}$$

Incorporating the above idea shows that the posterior distributions of binary latent variables can be employed to distinguish active covariates from inactive ones in the considered model.

For covariance matrix $\boldsymbol{Q}$, the proportion ρ, λ_0^2, λ_1^2 and σ_j^2, we consider the following priors:

$$\boldsymbol{Q} \sim \text{IW}(\boldsymbol{S}_0, \nu_0), \; \rho \sim \text{Beta}(a_\gamma, b_\gamma), \; \lambda_0^2 \sim \Gamma(c_0, d_0), \; \lambda_1^2 \sim \Gamma(c_1, d_1), \; \sigma_j^{-2} \sim \Gamma(c_2, d_2), \tag{7}$$

where $\text{IW}(\cdot, \cdot)$ denotes the inverted Wishart distribution, $\text{Beta}(\cdot, \cdot)$ represents the Beta distribution, $\Gamma(\cdot)$ is the gamma distribution, $\text{IG}(\cdot, \cdot)$ is the inverse gamma distribution and $\boldsymbol{S}_0$, ν_0, a_γ, b_γ, c_0, d_0, c_1, d_1, c_2 and d_2 are the user-specified hyperparameters. As mentioned above, λ_1 should tend to zero and λ_0 should tend to ∞ as the sample size is sufficiently large, which implies that c_1/d_1 is smaller than c_0/d_0. To this end, we assume $c_1 \ll c_0$ and $d_0 \ll d_1$.

Based on the above discussion, we can rewrite the considered linear mixed model together with the spike and slab lasso prior as the following hierarchical models:

$$\begin{cases} y_{ij}|\boldsymbol{b}_i \sim \mathcal{N}(\mu_{ij}, \sigma_j^2), \;\; \mu_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top \boldsymbol{b}_i, \;\; i = 1, \ldots, n, \; j = 1, \ldots, m, \\ \boldsymbol{b}_i \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{Q}), \;\; i = 1, \ldots, n, \\ \beta_k|\xi_{1k}^2, \gamma_k = 1 \sim \mathcal{N}(0, \xi_{1k}^2), \;\; \xi_{1k}^2|\lambda_1^2 \sim \text{Exp}(\lambda_1^2/2), \;\; \lambda_1^2 \sim \Gamma(c_1, d_1), \\ \beta_k|\xi_{0k}^2, \gamma_k = 0 \sim \mathcal{N}(0, \xi_{0k}^2), \;\; \xi_{0k}^2|\lambda_0^2 \sim \text{Exp}(\lambda_0^2/2), \;\; \lambda_0^2 \sim \Gamma(c_0, d_0), \\ \gamma_k \sim \text{Bernoulli}(\rho), \;\; k = 1, \ldots, p, \\ \boldsymbol{Q} \sim \text{IW}(\boldsymbol{S}_0, \nu_0), \;\; \rho \sim \text{Beta}(a_\gamma, b_\gamma), \;\; \sigma_j^{-2} \sim \Gamma(c_2, d_2), \; j = 1, \ldots, m. \end{cases} \tag{8}$$

3. Skinny Gibbs Sampler for Bayesian Lasso

Let $\boldsymbol{Y} = \{y_{ij} : i = 1, \ldots, n, \; j = 1, \ldots, m\}$, $\boldsymbol{X} = \{\boldsymbol{x}_{ij} : i = 1, \ldots, n, \; j = 1, \ldots, m\}$ and $\boldsymbol{Z} = \{z_{ij} : i = 1, \ldots, n, \; j = 1, \ldots, m\}$. From Equation (8), the joint posterior density of parameters $\boldsymbol{\beta}$, $\boldsymbol{Q}$, $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_p\}$, $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_m^2)^\top$ and $\boldsymbol{\vartheta} = \{\rho, \lambda_0, \lambda_1\}$ given the data $\mathcal{D} = \{\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}\}$ is given by

$$f(\boldsymbol{\beta}, \boldsymbol{Q}, \gamma, \sigma^2, \boldsymbol{\vartheta}|\mathcal{D}) \propto \left\{\prod_{i=1}^{n}\prod_{j=1}^{m}\psi(y_{ij}, \boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta}, \boldsymbol{z}_{ij}^{\top}\boldsymbol{Q}^{-1}\boldsymbol{z}_{ij} + \sigma_j^2)\right\}\left\{\prod_{j=1}^{m} f(\sigma_j^{-2})\right\} \times \prod_{k=1}^{p}\{\rho g_1(\beta_k|\lambda_1)\}^{\gamma_k}\{(1-\rho)g_0(\beta_k|\lambda_0)\}^{1-\gamma_k} f_W(\boldsymbol{Q}) f_{\vartheta}(\boldsymbol{\vartheta}), \tag{9}$$

where $\psi(x, \mu, \varsigma^2)$ is the probability density of normal random variable x with mean μ and variance ς^2, $f(\sigma_j^{-2})$ denotes the probability density of random variable σ_j^{-2}, $f_W(\boldsymbol{Q})$ is the inverted Wishart density function of random matrix $\boldsymbol{Q}$ and $f_{\vartheta}(\boldsymbol{\vartheta})$ represents the joint prior density function of random variable vector $\boldsymbol{\vartheta}$. It is rather difficult to sample observations from the joint posterior density given in Equation (9) in the presence of high-dimensional fixed effects because of some non-standard distributions and large matrix computations involved. In what follows, the Gibbs sampler is utilized to sample observations required for Bayesian inference.

To avoid expensive computation in running the Gibbs sampler, similarly to [23], at each Gibbs iteration, we divide parameter vector $\boldsymbol{\beta}$ into two subvectors corresponding to those active (i.e., $\gamma_k = 1$) and inactive (i.e., $\gamma_k = 0$) covariates, respectively. To wit, we define $\boldsymbol{\beta} = (\boldsymbol{\beta}_A^{\top}, \boldsymbol{\beta}_I^{\top})^{\top}$, where $\boldsymbol{\beta}_A$ and $\boldsymbol{\beta}_I$ are the subvectors of $\boldsymbol{\beta}$ associated with $\gamma_k = 1$ and $\gamma_k = 0$, respectively. Suppose that the cardinality of the set A is r. Without loss of generality, it is assumed that the first r components of $\boldsymbol{\beta}$ correspond to $\boldsymbol{\beta}_A$ and the last $p - r$ components of $\boldsymbol{\beta}$ correspond to $\boldsymbol{\beta}_I$. Similarly, we decompose $\boldsymbol{x}_{ij}$ as $\boldsymbol{x}_{ij} = (\boldsymbol{x}_{ijA}^{\top}, \boldsymbol{x}_{ijI}^{\top})^{\top}$. Under the above assumptions, the Gibbs sampler is implemented as follows. Observations required at each Gibbs iteration are iteratively drawn from the following conditional distributions: $f_A(\boldsymbol{\beta}_A|\mathcal{D}, \boldsymbol{b}, \sigma^2)$, $f_I(\boldsymbol{\beta}_I|\mathcal{D})$, $f(\boldsymbol{b}_i|\mathcal{D}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{Q})$, $f(\xi_{0k}^2|\beta_k, \gamma_k)$, $f(\xi_{1k}^2|\beta_k, \gamma_k)$, $f_{\gamma}(\gamma_k|\mathcal{D}, \boldsymbol{b}, \boldsymbol{\xi}_1, \boldsymbol{\xi}_0)$, $f(\boldsymbol{Q}|\boldsymbol{b})$, $f(\sigma_j^{-2}|\mathcal{D}, \boldsymbol{b})$, $f(\rho|\boldsymbol{\gamma})$, $f(\lambda_0^2|\boldsymbol{\xi}_0)$ and $f(\lambda_1^2|\boldsymbol{\xi}_1)$, which are given in Appendix A, where $\boldsymbol{b} = \{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n\}$, $\boldsymbol{\xi}_0 = \{\xi_{01}^2, \ldots, \xi_{0p}^2\}$ and $\boldsymbol{\xi}_1 = \{\xi_{11}^2, \ldots, \xi_{1p}^2\}$.

Although the Skinny Gibbs sampler introduced above can be easily conducted, it is rather time-consuming for a sufficiently large p. To address the issue, we investigate a fast yet efficient approach as follows, i.e., the variational Bayesian method.

4. Variational Bayesian Inference

4.1. Variational Bayes

It follows from the principle of variational inference that it is necessary to first construct a variational set $\mathfrak{F}$ of densities for random variables Ξ having the same support as the posterior density $f(\Xi|\mathcal{D})$, where $\Xi = \{\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\xi}_0, \boldsymbol{\xi}_1, \boldsymbol{Q}, \boldsymbol{\gamma}, \sigma^2, \boldsymbol{\vartheta}\}$. It is assumed that $q(\Xi) \in \mathfrak{F}$ is any variational density for approximating $f(\Xi|\mathcal{D})$. The variational Bayes aims to find the best approximation to $f(\Xi|\mathcal{D})$ in terms of the Kullback–Leibler divergence between $q(\Xi)$ and $f(\Xi|\mathcal{D})$, which is a solution to the optimization problem:

$$q^*(\Xi) = \underset{q(\Xi)\in\mathfrak{F}}{\operatorname{argmin}}\, \mathrm{KL}(q(\Xi) \parallel f(\Xi|\mathcal{D})), \tag{10}$$

where

$$\begin{aligned} \mathrm{KL}(q(\Xi) \parallel f(\Xi|\mathcal{D})) &= \int \log\left\{\frac{q(\Xi)}{f(\Xi|\mathcal{D})}\right\} q(\Xi) d\Xi \\ &= \int \log\left\{\frac{q(\Xi) f(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})}{f(\Xi, \boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})}\right\} q(\Xi) d\Xi \\ &= \mathrm{E}_{q(\Xi)}\{\log q(\Xi)\} - \mathrm{E}_{q(\Xi)}\{\log f(\Xi, \boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})\} \\ &\quad + \log f(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}) \geq 0, \end{aligned} \tag{11}$$

in which $E_{q(\Xi)}(\cdot)$ is the expectation taken with respect to $q(\Xi)$. Here, $\mathrm{KL}(q(\Xi) \parallel f(\Xi|\mathcal{D}))$ equals zero if and only if $q(\Xi) \equiv f(\Xi|\mathcal{D})$. Due to the intractable high-dimensional integral involved, it is quite troublesome to conduct the above optimization problem.

However, it follows from $\mathbb{L}\{q(\Xi)\} = \mathrm{E}_{q(\Xi)}\{\log f(\Xi, \boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})\} - \mathrm{E}_{q(\Xi)}\{\log q(\Xi)\}$ that

$$\log f(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}) = \mathrm{KL}(q(\Xi) \parallel f(\Xi|\mathcal{D})) + \mathbb{L}\{q(\Xi)\} \geq \mathbb{L}\{q(\Xi)\}. \tag{12}$$

Thus, $\mathbb{L}\{q(\Xi)\}$ might be regarded as a lower bound of $\log f(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$ and is usually referred to as the evidence lower bound (ELB). Then, minimizing $\mathrm{KL}(q(\Xi) \parallel f(\Xi|\mathcal{D}))$ is equivalent to maximizing $\mathbb{L}\{q(\Xi)\}$ in that $\log f(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$ is not related to Ξ. That is,

$$q^*(\Xi) = \underset{q(\Xi)\in\mathfrak{F}}{\operatorname{argmin}} \mathrm{KL}(q(\Xi) \parallel f(\Xi|\mathcal{D})) = \underset{q(\Xi)\in\mathfrak{F}}{\operatorname{argmax}} \mathbb{L}\{q(\Xi)\}. \tag{13}$$

Finding the problem of the best approximation to $f(\Xi|\mathcal{D})$ is transformed into an optimization problem of maximizing $\mathbb{L}\{q(\Xi)\}$ over the variational family $\mathfrak{F}$. The complexity of the optimization problem is associated with that of the variational set $\mathfrak{F}$. Thus, it is rather desirable to implement the optimization problem over a relatively simple variational set $\mathfrak{F}$.

Following the widely used methods for constructing a relatively simple variational set, we take $\mathfrak{F}$ as the mean-field variational family in which components of Ξ are mutually independent and each has a distinct factor in the variational density. Thus, we can assume that the variational density $q(\Xi)$ has the form

$$q(\Xi) = q(\boldsymbol{\beta})q(\boldsymbol{b})q(\boldsymbol{\sigma}^{-2})q(\boldsymbol{\gamma})q(\boldsymbol{Q})q(\boldsymbol{\vartheta}) \prod_{k=1}^{p} \{q(\xi_{0k}^2)q(\xi_{1k}^2)\} \equiv \prod_{s=1}^{\mathcal{S}} q_s(\boldsymbol{\zeta}_s), \tag{14}$$

where $q_s(\boldsymbol{\zeta}_s)$s are unspecified but the above assumed factorization across components is prespecified. Similarly to considerable variational literature, the optimal solutions of $q_s(\boldsymbol{\zeta}_s)$s can be obtained by maximizing $\mathbb{L}\{q(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_{\mathcal{S}})\}$ via the coordinate ascent method, where $\Xi = \{\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_{\mathcal{S}}\}$.

Following the idea of the coordinate ascent method given in [33–35], when fixing other variational factors $q_j(\boldsymbol{\zeta}_j)$ for $j \neq s$, i.e., $\boldsymbol{\zeta}_{-s} = \{\boldsymbol{\zeta}_j : j \neq s,\ j = 1, \ldots, \mathcal{S}\}$, the optimal variational density $q_s^*(\boldsymbol{\zeta}_s)$ maximizing $\mathbb{L}\{q(\Xi)\}$ with respect to $q_s(\boldsymbol{\zeta}_s)$ has the form

$$\begin{aligned} q_s^*(\boldsymbol{\zeta}_s) &\propto \exp\left[\mathrm{E}_{-s}\{\log f(\boldsymbol{\zeta}_s|\boldsymbol{\zeta}_{-s}, \mathcal{D})\}\right] \\ &\propto \exp[\mathrm{E}_{-s}\{\log f(\boldsymbol{Y}, \Xi|\boldsymbol{X}, \boldsymbol{Z})\}], \end{aligned} \tag{15}$$

where $f(\boldsymbol{\zeta}_s|\boldsymbol{\zeta}_{-s}, \mathcal{D})$ is the conditional density for $\boldsymbol{\zeta}_s$ given $(\boldsymbol{\zeta}_{-s}\ \mathcal{D})$ and $\mathrm{E}_{-s}(\cdot)$ represents the expectation evaluated for $q_{-s}(\boldsymbol{\zeta}_{-s}) = \prod_{j\neq s} q_j(\boldsymbol{\zeta}_j)$. Equation (15) implies that $\mathrm{E}_{-s}(\cdot)$ is not associated with the sth variational factor $q_s(\boldsymbol{\zeta}_s)$ and the optimal variational density $q_s^*(\boldsymbol{\zeta}_s)$ cannot be obtained in that the $q_{-s}(\boldsymbol{\zeta}_{-s})$ on the right-hand side are not the optimal ones. To address this issue, the coordinate updating algorithm is employed to iteratively update $q_s^*(\boldsymbol{\zeta}_s)$ via Equation (15). After the coordinate updating algorithm converges, we can take mean or mode of the optimal variational density $q_s^*(\boldsymbol{\zeta}_s)$ as a variational Bayesian estimate of parameter vector $\boldsymbol{\zeta}_s$ and regard the covariate as active if its corresponding variational Bayesian estimate deviates from zero.

It is easily shown from Equation (15) that the optimal density $q_\beta^*(\boldsymbol{\beta})$ has the form

$$q_{\beta_A}^*(\boldsymbol{\beta}_A) \sim \mathcal{N}_r(\boldsymbol{\mu}_A, \boldsymbol{\Sigma}_A),\ \ q_{\beta_I}^*(\boldsymbol{\beta}_I) \sim \mathcal{N}_{p-r}(\boldsymbol{0}, \boldsymbol{\Sigma}_I), \tag{16}$$

respectively, where $\boldsymbol{\Sigma}_A^{-1} = \sum_{i=1}^{n}\sum_{j=1}^{m} \boldsymbol{x}_{ijA}\boldsymbol{x}_{ijA}^\top E_{\sigma_j^2}^*(\sigma_j^{-2}) + \mathrm{diag}(\boldsymbol{\xi}_A)$ with $\boldsymbol{\xi}_A = \{E_{\xi_{1k}}^*(\xi_{1k}^{-2}),\ k \in A\}$, $\boldsymbol{\mu}_A = \boldsymbol{\Sigma}_A[\sum_{i=1}^{n}\sum_{j=1}^{m} \boldsymbol{x}_{ijA}\{y_{ij} - \boldsymbol{z}_{ij}^\top E_{b_i}^*(\boldsymbol{b}_i)\}E_{\sigma_j^2}^*(\sigma_j^{-2})]$ and $\boldsymbol{\Sigma}_I^{-1} = \mathrm{diag}(\sum_{i=1}^{n}\sum_{j=1}^{m} \boldsymbol{x}_{ijI}\boldsymbol{x}_{ijI}^\top)$ $+\mathrm{diag}(\boldsymbol{\xi}_I^0) = nm\boldsymbol{I}_{p-r} + \mathrm{diag}(\boldsymbol{\xi}_I^0)$ with $\boldsymbol{\xi}_I^0 = \{E_{\xi_{0k}}^*(\xi_{0k}^{-2}), k \in I\}$, in which $\mathrm{E}_{\sigma_j^2}^*(\cdot)$, $\mathrm{E}_{\xi_{1k}}^*(\cdot)$, $\mathrm{E}_{\xi_{0k}}^*(\cdot)$ and $\mathrm{E}_{b_i}^*(\cdot)$ are the expectations taken with respect to $q_{\sigma_j^2}^*(\sigma_j^{-2})$, $q_{\xi_{1k}}^*(\xi_{1k}^{-2})$, $q_{\xi_{0k}}^*(\xi_{0k}^{-2})$ and $q_{b_i}^*(\boldsymbol{b}_i)$, respectively. Then, the estimated posterior means and variance matrices of $\boldsymbol{\beta}_A$ and $\boldsymbol{\beta}_I$ for the variational densities $q_{\beta_A}^*(\boldsymbol{\beta}_A)$ and $q_{\beta_I}^*(\boldsymbol{\beta}_I)$ are $\mathrm{E}_A^*(\boldsymbol{\beta}_A) = \boldsymbol{\mu}_A$, $\mathrm{var}_A^*(\boldsymbol{\beta}_A) = \boldsymbol{\Sigma}_A$, $\mathrm{E}_I^*(\boldsymbol{\beta}_I) = \boldsymbol{0}$ and

$\text{var}_I^*(\boldsymbol{\beta}_I) = \boldsymbol{\Sigma}_I$, respectively. Moreover, the mode estimator $\boldsymbol{\beta}_A^q$ of $\boldsymbol{\beta}_A$ for the variational density $q_{\beta_A}^*(\boldsymbol{\beta}_A)$ is $\boldsymbol{\beta}_A^q = \boldsymbol{\mu}_A$, while the mode estimator $\boldsymbol{\beta}_I^q$ of $\boldsymbol{\beta}_I$ for the variational density $q_{\beta_I}^*(\boldsymbol{\beta}_I)$ is $\boldsymbol{\beta}_I^q = \mathbf{0}$.

The optimal density $q_{b_i}^*(\boldsymbol{b}_i)$ is the multivariate normal distribution

$$q_{b_i}^*(\boldsymbol{b}_i) \sim \mathcal{N}_q(\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b), \tag{17}$$

where $\boldsymbol{\Sigma}_b^{-1} = E_Q^*(\boldsymbol{Q}) + \sum_{j=1}^m \boldsymbol{z}_{ij}\boldsymbol{z}_{ij}^\top E_{\sigma_j^2}^*(\sigma_j^{-2})$ and $\boldsymbol{\mu}_b = \boldsymbol{\Sigma}_b[\sum_{j=1}^m \boldsymbol{z}_{ij}\{y_{ij} - \boldsymbol{x}_{ijA}^\top \mathrm{E}_A^*(\boldsymbol{\beta}_A)\}\mathrm{E}_{\sigma_j^2}^*(\sigma_j^{-2})]$. Then, the estimated posterior mean and variance matrix of $\boldsymbol{b}_i$ for variational densities $q_{b_i}^*(\boldsymbol{b}_i)$ are $\mathrm{E}_{b_i}^*(\boldsymbol{b}_i) = \boldsymbol{\mu}_b$ and $\text{var}_{b_i}^*(\boldsymbol{b}_i) = \boldsymbol{\Sigma}_b$, respectively. Moreover, the mode estimator $\boldsymbol{b}_i^q$ of $\boldsymbol{b}_i$ for variational density $q_{b_i}^*(\boldsymbol{b}_i)$ is $\boldsymbol{b}_i^q = \boldsymbol{\mu}_b$. The optimal densities $q_{\xi_{0k}}^*(\xi_{0k}^{-2})$ and $q_{\xi_{1k}}^*(\xi_{1k}^{-2})$ are given by

$$q_{\xi_{0k}}^*(\xi_{0k}^{-2}) \sim \text{IvG}(a_{0\xi k}^*, b_{0\xi k}^*) \text{ for } k \in I, \;\; q_{\xi_{1k}}^*(\xi_{1k}^{-2}) \sim \text{IvG}(a_{1\xi k}^*, b_{1\xi k}^*) \text{ for } k \in A, \tag{18}$$

respectively, where $a_{0\xi k}^* = \sqrt{\mathrm{E}_{\lambda_0}^*(\lambda_0^2)/\text{var}_{\beta_k}^*(\beta_k)}$, $a_{1\xi k}^* = \sqrt{\mathrm{E}_{\lambda_1}^*(\lambda_1^2)/[\{\mathrm{E}_{\beta_k}^*(\beta_k)\}^2 + \text{var}_{\beta_k}^*(\beta_k)]}$, $b_{0\xi k}^* = \mathrm{E}_{\lambda_0}^*(\lambda_0^2)$, $b_{1\xi k}^* = \mathrm{E}_{\lambda_1}^*(\lambda_1^2)$ and $\mathrm{E}_{\lambda_0}^*(\cdot)$ and $\mathrm{E}_{\lambda_1}^*(\cdot)$ are the expectations taken with respect to $q_{\lambda_0}^*(\lambda_0^2)$ and $q_{\lambda_1}^*(\lambda_1^2)$, respectively. In this case, we have $\mathrm{E}_{\xi_{0k}}^*(\xi_{0k}^{-2}) = a_{0\xi k}^*$, $\mathrm{E}_{\xi_{1k}}^*(\xi_{1k}^{-2}) = a_{1\xi k}^*$, $\text{var}_{\xi_{0k}}^*(\xi_{1k}^{-2}) = (a_{0\xi k}^*)^3/b_{0\xi k}^*$ and $\text{var}_{\xi_{1k}}^*(\xi_{1k}^{-2}) = (a_{1\xi k}^*)^3/b_{1\xi k}^*$. Moreover, the mode estimators ξ_{0k}^{-2q} and ξ_{1k}^{-2q} of ξ_{0k}^{-2} and ξ_{1k}^{-2} for variational densities $q_{\xi_{0k}}^*(\xi_{0k}^{-2})$ and $q_{\xi_{1k}}^*(\xi_{1k}^{-2})$ are $\xi_{0k}^{-2q} = a_{0\xi k}^*\sqrt{1 + (1.5a_{0\xi k}^*/b_{0\xi k}^*)^2} - 1.5(a_{0\xi k}^*)^2/b_{0\xi k}^*$ for $k \in I$ and $\xi_{1k}^{-2q} = a_{1\xi k}^*\sqrt{1 + (1.5a_{1\xi k}^*/b_{1\xi k}^*)^2} - 1.5(a_{1\xi k}^*)^2/b_{1\xi k}^*$ for $k \in A$, respectively.

To derive the optimal density $q_{\gamma_k}^*(\gamma_k)$, we denote

$$\begin{aligned}
\log(\varrho_k) \;=\; & \mathrm{E}_\rho^*(\log\rho) - \mathrm{E}_\rho^*\{\log(1-\rho)\} + \tfrac{1}{2}\left[\mathrm{E}_{\xi_{1k}}^*\{\log(\xi_{1k}^{-2})\} - \mathrm{E}_{\xi_{0k}}^*\{\log(\xi_{0k}^{-2})\}\right] \\
& + \mathrm{E}_{\beta_k}^*(\beta_k)\sum_{i=1}^n\sum_{j=1}^m \{y_{ij} - \boldsymbol{x}_{ij,C_k}^\top \mathrm{E}_\beta^*(\boldsymbol{\beta}_{C_k}) - \boldsymbol{z}_{ij}^\top \mathrm{E}_{b_i}^*(\boldsymbol{b}_i)\}x_{ijk}\mathrm{E}_{\sigma_j^2}^*(\sigma_j^{-2}) \\
& - \frac{1}{2}\left[\text{var}_{\beta_k}^*(\beta_k) + \{\mathrm{E}_{\beta_k}^*(\beta_k)\}^2\right]\left\{\sum_{i=1}^n\sum_{j=1}^m x_{ijk}^2\mathrm{E}_{\sigma_j^2}^*(\sigma_j^{-2}) - \mathrm{E}_{\xi_{0k}}^*(\xi_{0k}^{-2}) + \mathrm{E}_{\xi_{1k}}^*(\xi_{1k}^{-2})\right\},
\end{aligned} \tag{19}$$

where $C_k = \{\ell : \gamma_\ell = 1, \ell \neq k \in A\} = A\backslash\{k\}$, which is an index set with the kth index deleted from the set A. Thus, latent variable γ_k is sampled from the Bernoulli distribution with the probability $\varsigma_k = \varrho_k/(\varrho_k + 1)$, i.e., $\gamma_k|\mathcal{D}, \boldsymbol{b}, \boldsymbol{\sigma} \sim \text{Bernoulli}(\varsigma_k)$ for $k = 1, \ldots, p$. In this case, the estimated posterior mean and variance of γ_k for variational density $q_{\gamma_k}^*(\gamma_k)$ are $\mathrm{E}_{\gamma_k}^*(\gamma_k) = \varsigma_k$ and $\text{var}_{\gamma_k}^*(\gamma_k) = \varsigma_k(1-\varsigma_k)$, respectively. Thus, the mode estimator γ_k^q of γ_k for variational density $q_{\gamma_k}^*(\gamma_k)$ is $\gamma_k^q = \varsigma_k$ for $k = 1, \ldots, p$.

The optimal density $q_Q^*(\boldsymbol{Q})$ has the form

$$q_Q^*(\boldsymbol{Q}) \sim \text{IW}_q(S_0^*, \nu_0^*), \tag{20}$$

where $\boldsymbol{S}_0^* = \boldsymbol{S}_0 + n\boldsymbol{\mu}_b\boldsymbol{\mu}_b^\top + n\boldsymbol{\Sigma}_b$ with $\boldsymbol{\mu}_b$ and $\boldsymbol{\Sigma}_b$ defined in Equation (17) and $\nu_0^* = \nu_0 + n$. Then, we have $\mathrm{E}_Q^*(\boldsymbol{Q}) = \boldsymbol{S}_0^*/(\nu_0^* - q - 1)$. Moreover, the mode estimator $\boldsymbol{Q}^q$ of $\boldsymbol{Q}$ is given by $\boldsymbol{Q}^q = \boldsymbol{S}_0^*/(\nu_0^* + q + 1)$.

The optimal density $q_{\sigma_j^2}^*(\sigma_j^{-2})$ $(j = 1, \ldots, m)$ has the form

$$q_{\sigma_j^2}^*(\sigma_j^{-2}) \sim \Gamma\left(\frac{n}{2}, b_\sigma^*\right), \tag{21}$$

where $b_\sigma^* = 0.5\sum_{i=1}^n h_{ij}$, $h_{ij} = (y_{ij} - \mu_{ij}^*)^2 + \boldsymbol{x}_{ijA}^\top\boldsymbol{\Sigma}_A\boldsymbol{x}_{ijA} + \boldsymbol{x}_{ijI}^\top\boldsymbol{\Sigma}_I\boldsymbol{x}_{ijI} + \boldsymbol{z}_{ij}^\top\boldsymbol{\Sigma}_b\boldsymbol{z}_{ij}$ and $\mu_{ij}^* = \boldsymbol{x}_{ijA}^\top\boldsymbol{\mu}_A + \boldsymbol{z}_{ij}^\top\boldsymbol{\mu}_b$. Thus, we have $\mathrm{E}_{\sigma_j}^*(\sigma_j^{-2}) = n/\sum_{i=1}^n h_{ij}$ and $\text{var}_{\sigma_j}^*(\sigma_j^{-2}) =$

$2n/(\sum_{i=1}^{n} h_{ij})^2$. In this case, the mode estimator σ_j^{-2q} of σ_j^{-2} for variational density $q^*_{\sigma_j^2}(\sigma_j^{-2})$ is $\sigma_j^{-2q} = (n-2)/\sum_{i=1}^{n} h_{ij}$ for $j = 1, \ldots, m$.

The optimal density $q^*_\rho(\rho)$ can be expressed as

$$q^*_\rho(\rho) \sim \text{Beta}(c_\rho, d_\rho), \tag{22}$$

where $c_\rho = a_\gamma + \sum_{k=1}^{p} \text{E}^*_{\gamma_k}(\gamma_k)$ and $d_\rho = b_\gamma + p - \sum_{k=1}^{p} \text{E}^*_{\gamma_k}(\gamma_k)$. Thus, we have $\text{E}^*_\rho(\rho) = c_\rho/(c_\rho + d_\rho)$ and $\text{var}^*_\rho(\rho) = c_\rho d_\rho/\{(c_\rho + d_\rho)^2(c_\rho + d_\rho - 1)\}$. In this case, the mode estimator of ρ is given as $\rho^q = c_\rho/(c_\rho + d_\rho)$.

The optimal densities $q^*_{\lambda_0}(\lambda_0^2)$ and $q^*_{\lambda_1}(\lambda_1^2)$ are

$$q^*_{\lambda_0}(\lambda_0^2) \sim \Gamma(a^*_{0\lambda}, b^*_{0\lambda}), \;\; q^*_{\lambda_1}(\lambda_1^2) \sim \Gamma(a^*_{1\lambda}, b^*_{1\lambda}), \tag{23}$$

respectively, where $a^*_{0\lambda} = c_0 + p - \sum_{k=1}^{p} \text{E}^*_{\gamma_k}(\gamma_k)$, $b^*_{0\lambda} = d_0 + \sum_{k=1}^{p}\{1 - \text{E}^*_{\gamma_k}(\gamma_k)\}\text{E}^*_{\xi_{0k}}(\xi_{0k}^2)/2$, $a^*_{1\lambda} = c_1 + \sum_{k=1}^{p} \text{E}^*_{\gamma_k}(\gamma_k)$ and $b^*_{1\lambda} = d_1 + \sum_{k=1}^{p} \text{E}^*_{\gamma_k}(\gamma_k)\text{E}^*_{\xi_{1k}}(\xi_{1k}^2)/2$. In this case, we obtain $\text{E}^*_{\lambda_0}(\lambda_0^2) = a^*_{0\lambda}/b^*_{0\lambda}$, $\text{var}^*_{\lambda_0}(\lambda_0^2) = a^*_{0\lambda}/(b^*_{0\lambda})^2$, $\text{E}^*_{\lambda_1}(\lambda_1^2) = a^*_{1\lambda}/b^*_{1\lambda}$ and $\text{var}^*_{\lambda_1}(\lambda_1^2) = a^*_{1\lambda}/(b^*_{1\lambda})^2$. The mode estimators λ_0^{2q} and λ_1^{2q} of λ_0^2 and λ_1^2 for variational densities $q^*_{\lambda_0}(\lambda_0^2)$ and $q^*_{\lambda_1}(\lambda_1^2)$ are $\lambda_0^{2q} = (a^*_{0\lambda} - 1)/b^*_{0\lambda}$ and $\lambda_1^{2q} = (a^*_{1\lambda} - 1)/b^*_{1\lambda}$, respectively.

4.2. Optimizing $\mathbb{L}\{q(\Xi)\}$ via Coordinate Ascent Algorithm

The elaborated steps for optimizing $\mathbb{L}\{q(\Xi)\}$ via the coordinate ascent algorithm are given below:

Step (a) Given the initial values of variational densities $q^*_\beta(\boldsymbol{\beta})$, $q^*_{b_i}(\boldsymbol{b}_i)$, $q^*_{\xi 0k}(\xi_{0k}^{-2})$, $q^*_{\xi 1k}(\xi_{1k}^{-2})$, $q^*_{\gamma_k}(\gamma_k)$, $q^*_Q(\boldsymbol{Q})$, $q^*_{\sigma_j^2}(\sigma_j^{-2})$, $q^*_\rho(\rho)$, $q^*_{\lambda_0}(\lambda_0^2)$ and $q^*_{\lambda_1}(\lambda_1^2)$, compute the lower bound $\mathbb{L}\{q(\Xi)\}$ (denoted as $\mathbb{L}^{(0)}\{q(\Xi)\}$) and set $\kappa = 1$.
Step (b) Compute variational density $q^*_\beta(\boldsymbol{\beta})$ and update $\text{E}^*_\beta(\boldsymbol{\beta})$.
Step (c) Compute variational density $q^*_{b_i}(\boldsymbol{b}_i)$ and update $\text{E}^*_{b_i}(\boldsymbol{b}_i)$.
Step (d) Compute variational density $q^*_{\xi 0k}(\xi_{0k}^{-2})$ and update $\text{E}^*_{\xi_{0k}}(\xi_{0k}^{-2})$.
Step (e) Compute variational density $q^*_{\xi 1k}(\xi_{1k}^{-2})$ and update $\text{E}^*_{\xi_{1k}}(\xi_{1k}^{-2})$.
Step (f) For $k = 1, \ldots, p$, compute variational densities $q^*_{\gamma_k}(\gamma_k)$ and update $\text{E}^*_{\gamma_k}(\gamma_k)$.
Step (g) Compute variational density $q^*_Q(\boldsymbol{Q})$ and update $\text{E}^*_Q(\boldsymbol{Q})$.
Step (h) Compute variational densities $q^*_{\sigma_j^2}(\sigma_j^{-2})$ and update $\text{E}^*_{\sigma_j}(\sigma_j^{-2})$.
Step (i) Compute variational density $q^*_\rho(\rho)$ and update $E^*_\rho(\rho)$.
Step (j) Compute variational density $q^*_{\lambda_0}(\lambda_0^2)$ and update $E^*_{\lambda_0}(\lambda_0^2)$.
Step (k) Compute variational density $q^*_{\lambda_1}(\lambda_1^2)$ and update $E^*_{\lambda_1}(\lambda_1^2)$.
Step (l) Based on variational densities from Steps (b)–(k), compute the ELB $\mathbb{L}\{q(\Xi)\}$ (denoted as $\mathbb{L}^{(\kappa)}\{q(\Xi)\}$) and the relative change

$$\text{RC} = \frac{|\mathbb{L}^{(\kappa)}\{q(\Xi)\} - \mathbb{L}^{(\ell-1)}\{q(\Xi)\}|}{\mathbb{L}^{(\kappa-1)}\{q(\Xi)\}}.$$

Step (m) Given sufficiently small ϵ, if $\text{RC} < \epsilon$, the algorithm is stopped. Otherwise, repeat Steps (b)–(l).

The preceding presented coordinate ascent algorithm for computing variational Bayesian estimates of parameters is summarized as Algorithm 1 and converges to the solution of the optimization problem (13) because it satisfies the well-known KKT condition for the considered model.

Algorithm 1: Variational Bayesian estimation

Input: A data set $D = \{\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}\}$ and an LMM $f(\boldsymbol{\Xi}, \mathcal{D})$
Output: Variational densities $q(\boldsymbol{\Xi})$ and estimates of parameters
Initialize: Variational factors $q^*_{\beta}(\boldsymbol{\beta})$, $q^*_{b_i}(\boldsymbol{b}_i)(i = 1, \ldots, n)$, $q^*_{\xi 0k}(\xi_{0k}^{-2})$ $(k = 1, \ldots, p)$, $q^*_{\xi 1k}(\xi_{1k}^{-2})$ $(k = 1, \ldots, p)$, $q^*_{\gamma_k}(\gamma_k)$ $(k = 1, \ldots, p)$, $q^*_{Q}(\boldsymbol{Q})$, $q^*_{\sigma_j^2}(\sigma_j^{-2})$ $(j = 1, \ldots, m)$, $q^*_{\rho}(\rho)$, $q^*_{\lambda_0}(\lambda_0^2)$ and $q^*_{\lambda_1}(\lambda_1^2)$;
while *the ELBO has not converged* **do**
  Update $q^*_{\beta}(\boldsymbol{\beta})$ via Equation (16) and $\boldsymbol{\beta}^q$;
  for $i \in \{1, \ldots, n\}$ **do**
    Update $q^*_{b_i}(\boldsymbol{b}_i)$ via Equation (17) and $\boldsymbol{b}_i^q$;
  end
  for $k \in \{1, \ldots, p\}$ **do**
    Update $q^*_{\xi 0k}(\xi_{0k}^{-2})$ via Equation (18) and ξ_{0k}^{-2q};
    Update $q^*_{\xi 1k}(\xi_{1k}^{-2})$ via Equation (18) and ξ_{1k}^{-2q};
    Update $q^*_{\gamma_k}(\gamma_k)$ via Equation (19) and γ_k^q;
  end
  Update $q^*_{Q}(\boldsymbol{Q})$ via Equation (20) and $\boldsymbol{Q}^q$;
  for $j \in \{1, \ldots, m\}$ **do**
    Update $q^*_{\sigma_j}(\sigma_j^{-2})$ via Equation (21) and σ_j^{-2q};
  end
  Update $q^*_{\rho}(\rho)$ via Equation (22) and ρ^q;
  Update $q^*_{\lambda_0}(\lambda_0^2)$ via Equation (23) and λ_0^{2q};
  Update $q^*_{\lambda_1}(\lambda_1^2)$ via Equation (23) and λ_1^{2q};
  Compute $\mathbb{L}\{q^*(\boldsymbol{\Xi}^q)\} = \mathrm{E}_{q^*(\boldsymbol{\Xi}^q)}\{\log f(\boldsymbol{Y}, \boldsymbol{\Xi}^q | \boldsymbol{X}, \boldsymbol{Z})\} - \mathrm{E}_{q^*(\boldsymbol{\Xi})}\{\log q^*(\boldsymbol{\Xi}^q)\}$
end
return $q^*(\boldsymbol{\Xi})$ and $\boldsymbol{\Xi}^q$

4.3. Model Comparison

The Bayes factor is a vital statistic for model comparison within the Bayesian framework and is widely employed to choose a better model among the considered competing models due to its merits for model selection: (i) it is a consistent selector; (ii) it plays the part of an Occam's razor, preferring the simpler model for the similar fits; (iii) it does not need the models to be nested. For instance, see [36] for structural equation models and [37] for non-ignorable missing data. Denote $f(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Xi}_h, \mathcal{H}_h)$ as the probability density of the data $\{\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}\}$ associated with the model $\mathcal{H}_h$, where $\boldsymbol{\Xi}_h$ is the parameter vector in the model $\mathcal{H}_h$. Define $f(\boldsymbol{\Xi}_h|\mathcal{H}_h)$ as the prior of $\boldsymbol{\Xi}_h$ for $h = 0, 1$. The Bayes factor for comparing two competing models $\mathcal{H}_0$ and $\mathcal{H}_1$ can be written as

$$B_{10} = \frac{\int f(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Xi}_1, \mathcal{H}_1) f(\boldsymbol{\Xi}_1|\mathcal{H}_1) d\boldsymbol{\Xi}_1}{\int f(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Xi}_0, \mathcal{H}_0) f(\boldsymbol{\Xi}_0|\mathcal{H}_0) d\boldsymbol{\Xi}_0} = \frac{f(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}, \mathcal{H}_1)}{f(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}, \mathcal{H}_0)}, \tag{24}$$

where $f(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}, \mathcal{H}_k)$ is the marginal likelihood for the model $\mathcal{H}_h$ for $h = 0$ and 1. However, computing the Bayes factor B_{10} is a non-trivial task for our considered high-dimensional linear mixed model because of the intractable integral involved. Considerable methods have been developed to compute the marginal likelihood $f(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}, \mathcal{H}_k)$ or the Bayes factor, for example, Laplace's method [38], annealed importance sampling [39], bridge sampling [40], path sampling (also called thermodynamic integration) [41], nested sampling [42], power posteriors [43], hybrid method combining simulation and asymptotic approximations [44]. For a comprehensive review, refer to [45]. Here, a path sampling or thermodynamic integration method is adopted to compute B_{10} via a link model: $\mathcal{H}_{\zeta 01} = (1 - \zeta)\mathcal{H}_0 + \zeta\mathcal{H}_1$, where

ζ is a continuous parameter taking value in the interval $[0,1]$. Thus, we have $\mathcal{H}_{\zeta 01} = \mathcal{H}_0$ when $\zeta = 0$ and $\mathcal{H}_{\zeta 01} = \mathcal{H}_1$ when $\zeta = 1$. Similarly to [41], we define the following class of probability densities:

$$\mathcal{Q}(\zeta) = f(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z},\zeta) = \int f(\boldsymbol{Y},\zeta|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\Xi}) f(\boldsymbol{\Xi}) d\boldsymbol{\Xi}, \tag{25}$$

where $f(\boldsymbol{Y},\zeta|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\Xi})$ is the density of $\boldsymbol{Y}$ given $\boldsymbol{X}$ and $\boldsymbol{Z}$ under $\mathcal{H}_\zeta$ and $f(\boldsymbol{\Xi})$ is the prior of $\boldsymbol{\Xi}$. Under the above definition, it is easily known that $\mathcal{Q}(0) = f(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z},\mathcal{H}_0)$ and $\mathcal{Q}(1) = f(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z},\mathcal{H}_1)$. Following the argument of [41], we obtain

$$\log B_{10} = \log \frac{\mathcal{Q}(1)}{\mathcal{Q}(0)} = \int_0^1 \mathrm{E}\{U(\boldsymbol{Y},\zeta,\boldsymbol{\Xi}|\boldsymbol{X},\boldsymbol{Z})\} d\zeta, \tag{26}$$

where $\mathrm{E}(\cdot)$ represents the expectation taken with respect to the conditional density $f(\boldsymbol{\Xi},\zeta|\boldsymbol{Y},\boldsymbol{X},\boldsymbol{Z})$ and $U(\boldsymbol{Y},\zeta,\boldsymbol{\Xi}|\boldsymbol{X},\boldsymbol{Z}) = d\log f(\boldsymbol{Y},\zeta,\boldsymbol{\Xi}|\boldsymbol{X},\boldsymbol{Z})/d\zeta$. Thus, applying the thermodynamic integration [41] or powered posteriors method [43] to Equation (26), $\log B_{10}$ can be estimated by

$$\widehat{\log B_{10}} = \frac{1}{2}\sum_{\ell=0}^{\mathbb{L}} \Big(\zeta_{(\ell+1)} - \zeta_{(\ell)}\Big)\Big(\bar{U}_{(\ell+1)} + \bar{U}_{(\ell)}\Big), \tag{27}$$

where $0 = \zeta_{(0)} < \zeta_{(1)} < \cdots < \zeta_{(\mathbb{L}+1)} = 1$ and $\bar{U}_{(\ell)} = \mathcal{J}^{-1}\sum_{\tau=1}^{\mathcal{J}} U(\boldsymbol{Y},\zeta_{(\ell)},\boldsymbol{\Xi}_\ell^{(\tau)}|\boldsymbol{X},\boldsymbol{Z})$, in which $\{\boldsymbol{\Xi}_\ell^{(\tau)} : \tau = 1,\ldots,\mathcal{J}\}$ are observations sampled from the variational density $q^*(\boldsymbol{\Xi}|\zeta_{(\ell)})$ for $\ell = 1,\ldots,\mathbb{L}$. Following [46], $\mathcal{H}_1$ is selected when $\widehat{\log B_{10}} > 1$; otherwise, $\mathcal{H}_0$ is selected.

5. Simulation Studies

Several simulation studies are implemented to assess the performance of the introduced variational Bayesian methodologies. For comparison, we also take the Bayesian lasso method into consideration. In this simulation study, response variables y_{ij}s are independently sampled from the normal distribution: $y_{ij} \sim \mathcal{N}(\boldsymbol{x}_{ij}^\top\boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top\boldsymbol{b}_i, \sigma_j^2)$, where $\boldsymbol{x}_{ij}$, $\boldsymbol{z}_{ij}$ and $\boldsymbol{b}_i$ are independently drawn from the multivariate normal distributions $\mathcal{N}_p(\boldsymbol{0},\boldsymbol{\Sigma}_x)$, $\mathcal{N}_q(\boldsymbol{0},\boldsymbol{I})$ and $\mathcal{N}_q(\boldsymbol{0},\boldsymbol{Q})$, respectively, for $i = 1,\ldots,n, j = 1,\ldots,m$. The true value of $\boldsymbol{\beta}$ is taken to be $(-0.5, 0.8, 2, 0.8, 0.5, 0.0, \ldots, 0.0)^\top$, which implies that there are five active variables and $p-5$ inactive variables. As an illustration, we set $m = 6$, $q = 4$, $n = 100$, 200 and 300, and $p = 500, 1000$ and 2000, which indicate that $n \ll p$. The true values of σ_j^2's are set to be $\sigma_1^2 = \sigma_2^2 = 0.8$, $\sigma_3^2 = \sigma_4^2 = 0.9$ and $\sigma_5^2 = \sigma_6^2 = 1.0$. The true value of $\boldsymbol{Q}$ is taken with diagonal elements being 1.0 and remaining components being 0.1.

We consider the following two types of covariance structures for $\boldsymbol{\Sigma}_x = (\sigma_{x_{jk}})_{p\times p}$.

Type I. Components of covariate vector $\boldsymbol{x}_{ij}$ are independent of each other, i.e., $\sigma_{x_{jk}} = 0.0$ when $j \neq k$ and $\sigma_{x_{jj}} = 1.0$ when $1 \le j,k \le p$.

Type II. $\boldsymbol{x}_{ij}$ is an autoregressive correlation, i.e., $\sigma_{x_{jk}} = 0.5^{|j-k|}$ when $\forall j \neq k$ and $\sigma_{x_{jj}} = 1.0$ when $1 \le j,k \le p$.

In implementing the preceding presented variational Bayesian approach together with the spike and slab priors, we take the hyperparameters $\nu_0 = 1$ and $\boldsymbol{S}_0 = 0.02\boldsymbol{I}_{q\times q}$ leading to the flat prior for $\boldsymbol{Q}$ and set $a_\gamma = b_\gamma = 0.5$. For the spike and slab priors of β_ks, to achieve appropriate shrinkage and model selection consistency, we take $c_0 = 500$ and $c_1 = 0.3$, indicating $c_1 \ll c_0$, $d_0 = 5$ and $d_1 = 30$, implying $d_0 \ll d_1$, guaranteeing the sparsity of the model. In this simulation, 100 replications are conducted to select active variables and estimate model parameters. To assess the accuracy of parameter estimation via the proposed variational Bayesian method, we calculate the average value of RMSes for unknown parameters, where "RMS" indicates the root mean square between the Bayesian estimates based on 100 replications and true values of unknown parameters. To assess

the performance of the variable selection procedure, we compute TP and FP, where TP represents the average number of active covariates correctly identified as active and FP denotes the average number of inactive covariates incorrectly detected as active. Generally, the closer to the true number of active covariates TP is or the smaller FP is, the better the variable selection method behaves. Results are reported in Table 1. Examination of Table 1 shows that the proposed variational Bayesian method behaves better than Bayesian lasso method, regardless of the values of p and n and covariance structures, in that TP values for the former are closer to the true number of active covariates and FP values for the former are closer to zero than those for the latter. For parameter estimation, the proposed variational Bayesian method behaves better than the Bayesian lasso method in that the average values of the RMSes for the former are smaller than those for the latter, regardless of the values of p and n and covariance structures. To investigate the sensitivity of the selection of the hyperparameters a_γ and b_γ, we take $a_\gamma = 0.1$ and $b_\gamma = 0.9$ and calculate the corresponding results for the Type I structure of $\boldsymbol{\Sigma}_x$, which results are given in Table 1. These empirical results indicate that the proposed variational Bayesian method is not sensitive to the hyperparameters in that the same pattern is observed regardless of the hyperparameters a_γ and b_γ.

Table 1. Performance of variable selection and parameter estimation in the first simulation study.

(a_γ, b_γ)	$\boldsymbol{\Sigma}_x$	n	Method	p = 500			p = 1000			p = 2000		
				TP	**FP**	**RMS**	**TP**	**FP**	**RMS**	**TP**	**FP**	**RMS**
(0.5, 0.5)	I	100	VB	3.91	0.00	0.11	3.79	0.00	0.08	3.84	0.00	0.06
			LASSO	4.44	0.87	1.90	3.54	1.03	1.66	1.39	0.00	1.91
		200	VB	4.71	0.00	0.11	4.68	0.00	0.08	4.65	0.00	0.06
			LASSO	4.95	0.24	2.24	2.78	1.91	1.36	3.34	0.00	1.64
		300	VB	4.89	0.00	0.11	4.81	0.00	0.08	4.91	0.00	0.06
			LASSO	4.99	0.01	2.12	4.91	0.00	1.41	4.23	0.00	1.45
	II	100	VB	3.79	0.00	0.11	3.84	0.00	0.08	3.76	0.00	0.06
			LASSO	3.48	0.10	2.19	3.01	0.00	1.87	3.00	0.00	2.01
		200	VB	3.97	0.00	0.11	3.96	0.00	0.08	3.98	0.00	0.06
			LASSO	3.59	0.02	2.44	3.12	0.00	1.78	3.00	0.00	1.84
		300	VB	3.98	0.00	0.11	3.96	0.00	0.08	3.98	0.00	0.06
			LASSO	3.63	0.03	2.31	3.20	0.00	1.79	3.01	0.00	1.75
(0.1, 0.9)	I	100	VB	3.88	0.00	0.11	3.79	0.00	0.08	3.84	0.00	0.06
			LASSO	4.44	0.87	1.90	3.54	1.03	1.66	1.39	0.00	1.91
		200	VB	4.71	0.00	0.11	4.66	0.00	0.08	4.64	0.00	0.06
			LASSO	4.95	0.24	2.24	2.78	1.91	1.36	3.34	0.00	1.64
		300	VB	4.89	0.00	0.11	4.81	0.00	0.08	4.91	0.00	0.06
			LASSO	4.99	0.01	2.12	4.91	0.00	1.41	4.23	0.00	1.45

Note: VB represents variational Bayesian method and LASSO denotes Bayesian lasso method.

Table 2. Estimated log Bayes factor in the second simulation study.

Bayes Factor	n	p		
		500	**1000**	**2000**
$\widehat{\log B_{10}}$	100	−194	−102	−86
	200	−372	−272	−294
	300	−506	−544	−588
$\widehat{\log B_{20}}$ $(\times 10^7)$	100	−0.95	−4.03	−1.41
	200	−1.54	−3.68	−2.54
	300	−3.13	−3.58	−2.26

As an illustration for model comparison via the proposed Bayes factor, we consider the second simulation study. In the simulation study, the data $\{(\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, y_{ij}) : i =$

$1,\ldots,n, j = 1,\ldots,m\}$ are generated as those in the first simulation study with covariance structure of $\mathbf{\Sigma}_x$ taken to be Type I. To this end, we consider the following competing models:

$$\begin{aligned}
\mathcal{H}_0: &\quad y_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top \boldsymbol{b}_i + \varepsilon_{ij},\ \ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma_j^2),\\
\mathcal{H}_1: &\quad y_{ij} = \boldsymbol{z}_{ij}^\top \boldsymbol{b}_i + \varepsilon_{ij},\ \ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma_j^2),\\
\mathcal{H}_2: &\quad y_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top \boldsymbol{b}_i + \varepsilon_{ij},\ \ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma_0^2),
\end{aligned}$$

where $\mathcal{H}_0$ represents the true linear mixed model and while $\mathcal{H}_1$ and $\mathcal{H}_2$ are two competing linear mixed models, $\mathcal{H}_1$ only containing random effects without fixed effects, and $\mathcal{H}_2$ misspecifying the distribution of measurement error. We define a path $t \in [0,1]$ to link any two of the above presented three models. For example, $\mathcal{H}_0$ and $\mathcal{H}_1$ can be linked by $\mathcal{H}_{t01}: y_{ij} = (1-t)\boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top \boldsymbol{b}_i + \varepsilon_{ij}$, which indicates that $\mathcal{H}_{t01}$ is just $\mathcal{H}_0$ for $t = 0$ and becomes $\mathcal{H}_1$ for $t = 1$, and $\mathcal{H}_0$ and $\mathcal{H}_2$ are linked by $\mathcal{H}_{t02}: y_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top \boldsymbol{b}_i + \varepsilon_{ij}$ with $\varepsilon_{ij} \overset{\text{ind}}{\sim} \mathcal{N}(0, t^2\sigma_0^2 + (1-t)^2\sigma_j^2)$, which implies that $\mathcal{H}_{t02}$ reduces to $\mathcal{H}_0$ with $t = 0$ and $\mathcal{H}_{t02}$ becomes $\mathcal{H}_2$ with $t = 1$.

To calculate the estimated log Bayes factors (i.e., $\widehat{\log B_{10}}$ and $\widehat{\log B_{20}}$) via the preceding proposed path sampling procedure, we take $\zeta_{(\ell)} = \ell/\mathbb{L}$ for $\ell = 0, 1, \ldots, \mathbb{L}$, $\mathbb{L} = 10$, $\mathcal{J} = 1000$ and $\sigma_0^2 = 0.5$ and the same priors as those given in the first simulation studies. Results are given in Table 2, which indicates that $\mathcal{H}_0$ is strongly selected as expected regardless of n and p.

6. An Empirical Example

As an illustration of the variational Bayesian method developed above, we consider the ADNI-2 data [47] published in 2003 and followed by ADNI-1, ADNI-GO and ADNI-2 groups. This study aims to predict the mini-mental state examination (MMSE) score, which is an important index for detecting Alzheimer's disease (AD) stages in that different MMSE scores indicate different progression of a AD patient. AD is the most common type of dementia for elderly people and the sixth leading cause of death in the United States, and it results in the loss of memory and the impairment of cognitive and language skills. More importantly, there is no effective treatment to slow the progression of the disease [48]. The number of AD patients has grown exponentially with the speed of the aging population, bringing a socioeconomic burden to both families and society [49]. The details on the ADNI database can refer to the website http://adni.loni.usc.edu (accessed on 20 May 2021).

The ADNI-2 data were analyzed by [48] using the factor analysis model to impute missing values. As an illustration, we utilize 340 complete magnetic resonance imaging (MRI) features with 62 samples and 3 medical visits (6-month, 12-month and 24-month), take five features among 340 features as covariates associated with random effects and set the MMSE score as the response variable. That is, $n = 62$, $p = 340$, $q = 5$ and $m = 3$. In this case, covariates are high-dimensional compared with the sample size. Here, we assume that only a small fraction of covariates contribute to the response variable.

The preceding introduced variational Bayesian method together with the linear mixed model and the same priors as those in the first simulation study are utilized to fit the above-mentioned MRI data. Here, the hyperparameters are taken as $\nu_0 = 1$, $\boldsymbol{S}_0 = 0.02\boldsymbol{I}_{q\times q}$, $a_\gamma = b_\gamma = 0.5$, $c_0 = 10$, $d_0 = 1$, $c_1 = 1$ and $d_1 = 10$ for ensuring the sparsity of the model. Thus, the proposed variational Bayesian method selects three features as active variables: thickness average of the right fusiform (denoted as "x_1"), thickness standard deviation of the right posterior cingulate (denoted as "x_2") and thickness standard deviation of the left postcentral (denoted as "x_3"). Their corresponding parameter estimates are 1.9, 0.25 and 0.4, respectively, which show that the three active variables have positive effects on MMSE that are consistent with those given in [48]. Bayesian estimates of random effects $\boldsymbol{b}_i$ are -0.003, -0.0021, -0.0013, -0.0058 and -0.0054, respectively, which imply that the selected five covariates associated with random effects have negative effects on MMSE. Table 3 also presents the RMSE and MAP values for the models with 340 covariates

(denoted as the "Complete" model) and the selected three active covariates (denoted as the "Selected" model), where RMSE and MAP are evaluated by RMSE=$\sqrt{n^{-1}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$ and MAP=$n^{-1}\sum_{i=1}^{n}|\hat{y}_i - y_i|$ and $\hat{y}_i$ is the fitted value of response y_i. Examination of Table 3 shows that the selected model has smaller RMS and MAP values than the complete model, i.e., the selected model fits the ADNI-2 data better than the complete model. For the selected model, we also compute the Bayes factors for three competing models $\mathcal{H}_0$, $\mathcal{H}_1$ and $\mathcal{H}_2$ given in the second simulation study, which are $\widehat{\log B_{10}} = -558$ and $\widehat{\log B_{20}} = -46.93$, leading to the conclusion that $\mathcal{H}_0$ is strongly selected.

Table 3. Performance of variational Bayesian method for the complete and selected models in the ADNI-2 data.

Model	n	p	RMSE	MAP
Complete	62	340	49.17	49.15
Selected	62	3	1.05	0.82

7. Discussions

This paper investigates simultaneously estimating model parameters and selecting variables in linear mixed models with high-dimensional fixed effects and low-dimensional random effects in the Bayesian framework. A novel variational Bayesian approach is developed to address the time-consuming problem of the traditional Bayesian lasso method due to the ill-posited problem and large matrix computation involved in the presence of high-dimensional data. The Gaussian spike and slab priors of population-specific fixed-effects regression coefficients are specified to identify important fixed effects by allowing the tuning parameters to tend to zero. For the sake of sampling observations, the Gaussian spike and slab priors are reformulated as a mixture of a normal distribution and an exponential distribution. In the variational Bayesian framework, the problem of best approximating the posterior density is transformed as an optimization problem, i.e., minimizing the evidence lower bound. For ease of computation, the coordinate ascent algorithm, implemented efficiently, is employed to optimize the evidence lower bound. For model comparison, the Bayes factor is computed by the path sampling method. Simulation studies are conducted to investigate the performance of the proposed variational Bayesian method, and a real example is illustrated by the proposed methodologies. Empirical results show that the proposed variational Bayesian method behaves better than the traditional Bayesian lasso method regardless of the accuracy of parameter estimation, the consistency of variable selection or computational flexibility and complexity.

The proposed variational Bayesian method has the following advantages:

- Overcoming the problem of selecting a high-dimensional vector of shrinkage parameters required for the Bayesian lasso method;
- Simultaneously estimating model parameters and variance–covariance matrices and selecting fixed-effects and random-effects components with a relatively low computational cost;
- Avoiding large matrix computations and the curse of dimensionality problem;
- Providing a flexible and efficient approach to compute the Bayes factor for model comparison.

The proposed variational Bayesian method can be extended to more complicated models, such as generalized linear mixed models with mixed discrete and missing data. However, their extensions have huge challenges, including the closed-form derivation of the optimal variational density, the specification of the priors, the learning of the data-driven hyperparameters and the computational complexity. In addition, this paper does not consider the selection of high-dimensional random effects, which is a rather challenging topic. In addition, to speed up the convergence of the chain, we might consider some important and relevant Gibbs sampling schemes, for example, the herded Gibbs sampling,

which is a deterministic variant of the Gibbs sampling scheme and generates observations by matching the full-conditionals rather than by taking the full-conditionals at random [50], the recycling Gibbs sampler, which generates auxiliary observations whose information is eventually discarded and which can be recycled within the Gibbs algorithm for improving efficiency with no extra cost [51], and the blocking and parameterization method [52].

In addition, we did not consider BIC criterion for model comparison in that BIC is only an approximation to the Bayes factor of marginal likelihood of the data given each hypothesis. Moreover, due to the random effects involved in the considered models, BIC behaves unsteadily.

Author Contributions: Conceptualization, N.T.; methodology, N.T.; software, J.Y.; validation, N.T. and J.Y.; formal analysis, N.T. and J.Y.; investigation, J.Y.; resources, N.T. and J.Y.; data curation, J.Y.; writing—original draft preparation, J.Y.; writing—review and editing, N.T.; visualization, J.Y.; supervision, N.T.; project administration, N.T.; funding acquisition, N.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Key Projects of the National Nautral Science Foundation of China (grant number 11731011).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The ADNI database is available on the website http://adni.loni.usc.edu (accessed on 20 May 2021).

Acknowledgments: The authors are grateful for the associate editor and the three referees for their constructive comments, which largely improved an earlier manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MCMC	Markov chain Monte Carlo algorithm
EM	Expectation Maximization algorithm
ELB	evidence lower bound
TP	average number of active covariates correctly identified as active
FP	average number of inactive covariates incorrectly detected as active
RMS	mean square between the Bayesian estimates based on 100 replications and true value of unknown parameter
VB	variational Bayesian with proposed method
LASSO	Bayesian lasso method
AD	Alzheimer's Disease
ADNI	Alzheimer's Disease Neuroimaging Initiative
MRI	magnetic resonance imaging
MMSE	mini-mental state examination

Appendix A. Conditional Distributions Required in Implementing the Gibbs Sampler

By the definitions and priors of $\boldsymbol{\beta}_A$ and $\boldsymbol{\beta}_I$, it is easily shown from Equation (9) that the conditional distributions $f_A(\boldsymbol{\beta}_A|\mathcal{D}, \boldsymbol{b}, \boldsymbol{\sigma})$ and $f_I(\boldsymbol{\beta}_I|\mathcal{D})$ have the forms

$$\boldsymbol{\beta}_A|\mathcal{D}, \boldsymbol{b}, \boldsymbol{\sigma} \sim \mathcal{N}_r(\boldsymbol{\mu}_A^0, \boldsymbol{\Sigma}_A^0), \quad \boldsymbol{\beta}_I|\mathcal{D} \sim \mathcal{N}_{p-r}(\mathbf{0}, \boldsymbol{\Sigma}_I^0), \tag{A1}$$

respectively, where ${\boldsymbol{\Sigma}_A^0}^{-1} = \sum_{i=1}^n \sum_{j=1}^m \boldsymbol{x}_{ijA}\boldsymbol{x}_{ijA}^\top/\sigma_j^2 + \mathrm{diag}(\boldsymbol{\xi}_A^0)$ with $\boldsymbol{\xi}_A^0 = \{\xi_{1k}^{-2}, k \in A\}$, $\boldsymbol{\mu}_A^0 = \boldsymbol{\Sigma}_A^0\{\sum_{i=1}^n \sum_{j=1}^m \boldsymbol{x}_{ijA}(y_{ij} - \boldsymbol{z}_{ij}^\top \boldsymbol{b}_i)/\sigma_j^2\}$ and ${\boldsymbol{\Sigma}_I^0}^{-1} = \mathrm{diag}(\sum_{i=1}^n \sum_{j=1}^m \boldsymbol{x}_{ijI}\boldsymbol{x}_{ijI}^\top) + \mathrm{diag}(\boldsymbol{\xi}_{CI}^0) = nm\boldsymbol{I}_{p-r} + \mathrm{diag}(\boldsymbol{\xi}_{CI}^0)$ with $\boldsymbol{\xi}_{CI}^0 = \{\xi_{0k}^{-2}, k \in I\}$.

The conditional distribution $f(\boldsymbol{b}_i|\mathcal{D}, \boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{Q})$ has the form

$$\boldsymbol{b}_i|\mathcal{D}, \boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{Q} \sim \mathcal{N}_q(\boldsymbol{\mu}_b^C, \boldsymbol{\Sigma}_b^C), \tag{A2}$$

where $\boldsymbol{\Sigma}_b^{C^{-1}} = \boldsymbol{Q} + \sum_{j=1}^m \boldsymbol{z}_{ij}\boldsymbol{z}_{ij}^\top/\sigma_j^2$ and $\boldsymbol{\mu}_b^C = \boldsymbol{\Sigma}_b^C\{\sum_{j=1}^m \boldsymbol{z}_{ij}(y_{ij} - \boldsymbol{x}_{ij}^\top\boldsymbol{\beta})/\sigma_j^2\}$.

The conditional distributions $f(\xi_{0k}^2|\beta_k, \gamma_k)$ and $f(\xi_{1k}^2|\beta_k, \gamma_k)$ are given by

$$\begin{aligned} f(\xi_{0k}^2|\beta_k, \gamma_k) &\propto (\xi_{0k}^2)^{-(1-\gamma_k)/2} \exp\{-(1-\gamma_k)\beta_k^2/(2\xi_{0k}^2) - \lambda_0^2(1-\gamma_k)\xi_{0k}^2/2\}, \\ f(\xi_{1k}^2|\beta_k, \gamma_k) &\propto (\xi_{1k}^2)^{-\gamma_k/2} \exp\{-\gamma_k\beta_k^2/(2\xi_{1k}^2) - \lambda_1^2\gamma_k\xi_{1k}^2/2\}, \end{aligned} \tag{A3}$$

respectively, which lead to

$$\xi_{0k}^2|\beta_k = 0, \gamma_k = 0 \sim \Gamma(1/2, \lambda_0^2/2), \quad \xi_{1k}^{-2}|\beta_k, \gamma_k = 1 \sim \mathrm{IvG}(\sqrt{\lambda_1^2/\beta_k^2}, \lambda_1^2),$$

where $\mathrm{IvG}(a, b)$ represents the inverse Gaussian distribution with parameters a and b.

The ratio of $\Pr(\gamma_k = 1|\mathcal{D}, \boldsymbol{b}, \boldsymbol{\sigma})$ to $\Pr(\gamma_k = 0|\mathcal{D}, \boldsymbol{b}, \boldsymbol{\sigma})$ is proportional to

$$\frac{\rho\psi(\beta_k, 0, \xi_{1k}^2)}{(1-\rho)\psi(\beta_k, 0, \xi_{0k}^2)} \exp\left\{\beta_k \sum_{i=1}^n \sum_{j=1}^m \frac{(y_{ij} - \boldsymbol{x}_{i,C_k}^\top\boldsymbol{\beta}_{C_k} - \boldsymbol{z}_{ij}^\top\boldsymbol{b}_i)x_{ijk}}{\sigma_j^2} + \frac{\beta_k^2}{2}\sum_{i=1}^n \sum_{j=1}^m x_{ijk}^2(1 - \sigma_j^{-2})\right\}, \tag{A4}$$

which is denoted as ϱ_k, where $C_k = \{\ell : \gamma_\ell = 1, \ell \neq k \in A\}$. Thus, latent variable γ_k is sampled from the Bernoulli distribution with the probability $\varsigma_k = \varrho_k/(\varrho_k + 1)$, i.e., $\gamma_k|\mathcal{D}, \boldsymbol{b}, \boldsymbol{\sigma} \sim \mathrm{Bernoulli}(\varsigma_k)$ for $k = 1, \ldots, p$.

The conditional distribution $f(\boldsymbol{Q}|\boldsymbol{b})$ is shown as

$$\boldsymbol{Q}|\boldsymbol{b} \sim \mathrm{IW}_q\left(\boldsymbol{S}_0 + \sum_{i=1}^n \boldsymbol{b}_i\boldsymbol{b}_i^\top, \nu_0 + n\right). \tag{A5}$$

The conditional distribution $f(\sigma_j^{-2}|\mathcal{D}, \boldsymbol{b})$ $(j = 1, \ldots, m)$ has the form

$$f(\sigma_j^{-2}|\mathcal{D}, \boldsymbol{b}) \propto (\sigma_j^{-2})^{n/2+c_2-1} \exp\left\{-\frac{1}{2\sigma_j^2}\sum_{i=1}^n (y_{ij} - \mu_{ij})^2 - \frac{d_2}{\sigma_j^2}\right\}, \tag{A6}$$

which indicates

$$\sigma_j^{-2}|\mathcal{D}, \boldsymbol{b} \sim \Gamma\left(\frac{n}{2} + c_2, d_2 + \frac{1}{2}\sum_{i=1}^n (y_{ij} - \mu_{ij})^2\right).$$

The conditional distribution $f(\rho|\boldsymbol{\gamma})$ is given as

$$\rho|\boldsymbol{\gamma} \sim \mathrm{Beta}\left(a_\gamma + \sum_{k=1}^p \gamma_k, b_\gamma + p - \sum_{k=1}^p \gamma_k\right). \tag{A7}$$

The conditional distributions $f(\lambda_0^2|\boldsymbol{\xi}_0)$ and $f(\lambda_1^2|\boldsymbol{\xi}_1)$ are shown as

$$\begin{aligned} \lambda_0^2|\boldsymbol{\xi}_0 &\sim \Gamma\left(c_0 + p - \sum_{k=1}^p \gamma_k, d_0 + \frac{1}{2}\sum_{k=1}^p (1-\gamma_k)\xi_{0k}^2\right), \\ \lambda_1^2|\boldsymbol{\xi}_1 &\sim \Gamma\left(c_1 + \sum_{k=1}^p \gamma_k, d_1 + \frac{1}{2}\sum_{k=1}^p \gamma_k\xi_{1k}^2\right), \end{aligned} \tag{A8}$$

respectively.

Appendix B. Calculating the Evidence Lower Bound (ELB)

Denote $q^*(\Xi)$ to be the optimal variational density approximating the posterior density $f(\Xi|\mathcal{D})$ and $f(\Xi)$ to be the prior density of $\Xi = \{\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\xi}_0, \boldsymbol{\xi}_1, \boldsymbol{Q}, \boldsymbol{\gamma}, \boldsymbol{\sigma}^2, \boldsymbol{\vartheta}\}$. Define $\mathrm{E}_{q^*(\Xi)}(\cdot)$ as the expectation taken with respect to $q^*(\Xi)$. Thus, it follows from Equation (12) that ELOB has the form

$$\begin{aligned}\mathbb{L}\{q^*(\Xi)\} &= \mathrm{E}_{q^*(\Xi)}\{\log f(\Xi, \boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})\} - \mathrm{E}_{q^*(\Xi)}\{\log q(\Xi)\} \\ &= \mathrm{E}_{q^*(\Xi)}\{\log f(\boldsymbol{Y}|\Xi, \boldsymbol{X}, \boldsymbol{Z}) + \log f(\Xi)\} - \mathrm{E}_{q^*(\Xi)}\{\log q(\Xi)\},\end{aligned} \tag{A9}$$

where

$$\log f(\boldsymbol{Y}|\Xi, \boldsymbol{X}, \boldsymbol{Z}) \propto \frac{n}{2}\sum_{j=1}^{m}\log\sigma_j^{-2} - \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(y_{ij} - \boldsymbol{x}_{ij}^\top\boldsymbol{\beta} - \boldsymbol{z}_{ij}^\top\boldsymbol{b}_i)^2}{2\sigma_j^2}, \tag{A10}$$

$$\begin{aligned}\log f(\Xi) \propto{}& \frac{1}{2}\sum_{k=1}^{r}\left(r\log\xi_{1k}^{-2} - \frac{\beta_k^2}{\xi_{1k}^2}\right) + \frac{1}{2}\sum_{k=1}^{p-r}\left\{(p-r)\log\xi_{0k}^{-2} - \frac{\beta_k^2}{\xi_{0k}^2}\right\} \\ &-\frac{1}{2}\mathrm{trace}\left\{\left(\boldsymbol{S}_0 + \sum_{i=1}^{n}\boldsymbol{b}_i\boldsymbol{b}_i^\top\right)\boldsymbol{Q}^{-1}\right\} + \frac{\lambda_1^2 + \lambda_0^2}{2} + (c_1 - 1)\log\lambda_1^2 \\ &-d_1\lambda_1^2 + (c_0 - 1)\log\lambda_0^2 - d_0\lambda_0^2 - \frac{n + \nu_0 + q + 1}{2}\log|\boldsymbol{Q}| \\ &+(a_\gamma - 1)\log\rho + (b_\gamma - 1)\log(1-\rho) - \sum_{j=1}^{m}\frac{d_2}{\sigma_j^2} \\ &+(c_2 - 1)\sum_{j=1}^{m}\log(\sigma_j^{-2}) + \sum_{k=1}^{p}\{\gamma_k\log\rho + (1-\gamma_k)\log(1-\rho)\}.\end{aligned} \tag{A11}$$

It follows from the definition of $q(\Xi)$ that

$$\begin{aligned}\mathrm{E}_{q^*(\Xi)}\{\log q(\Xi)\} = {}& \mathrm{E}_{\beta}^*\{\log q(\boldsymbol{\beta})\} + \mathrm{E}_{b}^*\{\log q(\boldsymbol{b})\} + \mathrm{E}_{\xi 1}^*\{\log q(\boldsymbol{\xi}_1^{-2})\} + \mathrm{E}_{\xi 0}^*\{\log q(\boldsymbol{\xi}_0^{-2})\} \\ &+\mathrm{E}_{\gamma}^*\{\log q(\boldsymbol{\gamma})\} + \mathrm{E}_{Q}^*\{\log q(Q)\} + \mathrm{E}_{\sigma}^*\{\log q(\boldsymbol{\sigma}^2)\} + \mathrm{E}_{\rho}^*\{\log q(\rho)\} \\ &+\mathrm{E}_{\lambda_0}^*\{\log q(\lambda_0^2)\} + \mathrm{E}_{\lambda_1}^*\{\log q(\lambda_1^2)\},\end{aligned} \tag{A12}$$

where $\mathrm{E}_{\beta}^*\{\log q(\boldsymbol{\beta})\} \propto -\frac{r}{2}\log|\boldsymbol{\Sigma}_A| - \frac{p-r}{2}\log|\boldsymbol{\Sigma}_I|$, $\mathrm{E}_{b}^*\{\log q(\boldsymbol{b})\} \propto -\frac{n}{2}\log|\boldsymbol{\Sigma}_b|$, $\mathrm{E}_{\xi 1}^*\{\log q(\boldsymbol{\xi}_1^{-2})\} \propto -\frac{1}{2}\sum_{k=1}^{p}[3\{\log a_{1\xi}^* - a_{1\xi k}^*/(2b_{1\xi k}^*)\} + 2b_{1\xi k}^*/a_{1\xi k}^* + 1]$, $\mathrm{E}_{\xi 0}^*\{\log q(\boldsymbol{\xi}_0^{-2})\} \propto -\frac{1}{2}\sum_{k=1}^{p}[3\{\log a_{0\xi}^* - a_{0\xi k}^*/(2b_{0\xi k}^*)\} + 2b_{0\xi k}^*/a_{0\xi k}^* + 1]$, $\mathrm{E}_{\gamma}^*\{\log q(\boldsymbol{\gamma})\} \propto \sum_{k=1}^{p}\{\varsigma_k\log\varsigma_k + (1-\varsigma_k)\log(1-\varsigma_k)\}$, $\mathrm{E}_{Q}^*\{\log q(Q)\} \propto -\frac{\nu_0^*}{2}\log\boldsymbol{S}_0^* + \frac{\nu_0^* - q - 1}{2}\nu_0^*\boldsymbol{S}_0^* - \frac{1}{2}\mathrm{trace}(\nu_0^*\boldsymbol{I}_{q\times q})$, $\mathrm{E}_{\sigma}^*\{\log q(\boldsymbol{\sigma})\} \propto -nd_2\sum_{j=1}^{m}(\sum_{i=1}^{n}h_{ij})^{-1} + (c_2 - 1)\sum_{j=1}^{m}(\log n - \log\sum_{i=1}^{n}h_{ij} - 1/n)$, $\mathrm{E}_{\rho}^*\{\log q(\rho)\} \propto (c_\rho - 1)\{\log(c_\rho) - \log(c_\rho + d_\rho)\} - d_\rho(c_\rho - 1)/\{2c_\rho(c_\rho + d_\rho + 1)\} + (d_\rho - 1)\{\log(d_\rho) - \log(c_\rho + d_\rho) - c_\rho(d_\rho - 1)/\{2d_\rho(c_\rho + d_\rho + 1)\}$, $\mathrm{E}_{\lambda_0}^*\{\log q(\lambda_0^2)\} \propto (a_{0\lambda}^* - 1)\{\dot{\Gamma}(a_{0\lambda}^*)/\Gamma(a_{0\lambda}^*) - \log b_{0\lambda}^*\} - a_{0\lambda}^*$ and $\mathrm{E}_{\lambda_1}^*\{\log q(\lambda_1^2)\} \propto (a_{1\lambda}^* - 1)\{\dot{\Gamma}(a_{1\lambda}^*)/\Gamma(a_{1\lambda}^*) - \log b_{1\lambda}^*] - a_{1\lambda}^*$.

Note that for a random variable ξ with mean $\mathrm{E}(\xi) = \mu$ and variance $\mathrm{D}(\xi) = \sigma^2$, it follows from Taylor expansion that the mean of the function $y = f(\xi)$ is $\mathrm{E}(y) \approx f(\mu) + \frac{1}{2}\ddot{f}(\mu)\mathrm{D}(\xi)$, where $\ddot{f}(\cdot)$ denotes the second derivative of the function $f(\xi)$. Then, we have

$$\begin{aligned}\mathrm{E}_{q^*(\Xi)}\{\log f(\boldsymbol{Y}|\Xi, \boldsymbol{X}, \boldsymbol{Z})\} \propto{}& \frac{n}{2}\sum_{j=1}^{m}\left(\frac{1}{n} - \log\frac{n}{\sum_{i=1}^{n}h_{ij}}\right) - \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{n}{\sum_{i'=1}^{n}h_{i'j}}[y_{ij}^2 - 2y_{ij}\{\boldsymbol{x}_{ij}^\top \\ &\mathrm{E}_{\beta}^*(\boldsymbol{\beta}) + \boldsymbol{z}_{ij}^\top\mathrm{E}_{b_i}^*(\boldsymbol{b}_i)\} + \boldsymbol{x}_{ij}^\top\{\mathrm{var}_{\beta}^*(\boldsymbol{\beta}) + \mathrm{E}_{\beta}^*(\boldsymbol{\beta})\mathrm{E}_{\beta}^*(\boldsymbol{\beta}^\top)\}\boldsymbol{x}_{ij} \\ &+\boldsymbol{z}_{ij}^\top\{\mathrm{var}_{b_i}^*(\boldsymbol{b}_i) + \mathrm{E}_{b_i}^*(\boldsymbol{b}_i)\mathrm{E}_{b_i}^*(\boldsymbol{b}_i^\top)\}\boldsymbol{z}_{ij} + 2\boldsymbol{x}_{ij}^\top\mathrm{E}_{\beta}^*(\boldsymbol{\beta})\mathrm{E}_{b_i}^*(\boldsymbol{b}_i^\top)\boldsymbol{z}_{ij}].\end{aligned} \tag{A13}$$

Note that for a random variable $\xi \sim \Gamma(\alpha, \beta)$, we have $\mathrm{E}\{\log(\xi)\} = \dot{\Gamma}(\alpha)/\Gamma(\alpha) - \log(\beta)$, where $\dot{\Gamma}(\cdot)$ denotes the first derivative of gamma function. Thus, we have

$$
\begin{aligned}
\mathrm{E}_{q^*(\Xi)}\{\log f(\boldsymbol{\Xi})\} \quad \propto \quad & \frac{1}{2}\sum_{k=1}^{r}\left[r\left(\log a^*_{1\xi k}-\frac{a^*_{1\xi k}}{2b^*_{1\xi k}}\right)-\{\mathrm{var}^*_{\beta_k}(\beta_k)+(\mathrm{E}^*_{\beta_k}(\beta_k))^2\}\mathrm{E}^*_{\xi 1k}(\xi_{1k}^{-2})\right] \\
& +\frac{1}{2}\sum_{k=1}^{p-r}\left[(p-r)\left(\log a^*_{0\xi k}-\frac{a^*_{0\xi k}}{2b^*_{0\xi k}}\right)-\{\mathrm{var}^*_{\beta_k}(\beta_k)+(\mathrm{E}^*_{\beta_k}(\beta_k))^2\}\mathrm{E}^*_{\xi 0k}(\xi_{0k}^{-2})\right] \\
& -\frac{1}{2}\sum_{i=1}^{n}\mathrm{E}^*_{b_i}(\boldsymbol{b}_i^\top)\mathrm{E}^*_{Q}(\boldsymbol{Q})\mathrm{E}^*_{b_i}(\boldsymbol{b}_i)+\frac{\mathrm{E}^*_{\lambda_1}(\lambda_1^2)+\mathrm{E}^*_{\lambda_0}(\lambda_0^2)}{2} \\
& +(c_1-1)\left\{\frac{\dot{\Gamma}(a^*_{1\lambda})}{\Gamma(a^*_{1\lambda})}-\log(b^*_{1\lambda})\right\}-d_1\mathrm{E}^*_{\lambda_1}(\lambda_1^2) \\
& +(c_0-1)\left\{\frac{\dot{\Gamma}(a^*_{0\lambda})}{\Gamma(a^*_{0\lambda})}-\log b^*_{0\lambda}\right\}-d_0\mathrm{E}^*_{\lambda_0}(\lambda_0^2) \\
& +\frac{n+\nu_0-q-1}{2}\left(\log|\boldsymbol{S}_0^*\nu_0^*|-\frac{\mathrm{var}^*_{Q}|\boldsymbol{Q}|}{2|\boldsymbol{S}_0^*\nu_0^*|^2}\right)-\frac{1}{2}\mathrm{trace}\{\boldsymbol{S}_0^{-1}\mathrm{E}^*_{Q}(\boldsymbol{Q})\} \\
& +(a_\gamma-1)\left(\log\frac{c_\rho}{c_\rho+d_\rho}-\frac{d_\rho}{2c_\rho(c_\rho+d_\rho+1)}\right) \\
& +(b_\gamma-1)\left(\log\frac{d_\rho}{c_\rho+d_\rho}-\frac{c_\rho}{2d_\rho(c_\rho+d_\rho+1)}\right) \\
& -nd_2\textstyle\sum_{j=1}^{m}(\sum_{i=1}^{n}h_{ij})^{-1}+(c_2-1)\sum_{j=1}^{m}(\log n-\log\sum_{i=1}^{n}h_{ij}-1/n) \\
& +\sum_{k=1}^{p}\left[\mathrm{E}^*_{\gamma_k}(\gamma_k)\left(\log\frac{c_\rho}{c_\rho+d_\rho}-\frac{d_\rho}{2c_\rho(c_\rho+d_\rho+1)}\right)\right. \\
& \left.+(1-\mathrm{E}^*_{\gamma_k}(\gamma_k))\left(\log\frac{d_\rho}{c_\rho+d_\rho}-\frac{c_\rho}{2d_\rho(c_\rho+d_\rho+1)}\right)\right],
\end{aligned}
\tag{A14}
$$

where $|\boldsymbol{Q}|$ represents the determinant of matrix $\boldsymbol{Q}$, $\mathrm{var}^*_Q(Q_{ij})=\nu_0(\sigma^{*2}_{ij}+\sigma^*_{ii}\sigma^*_{jj})$ and σ^*_{ij} is the (i,j)-th component of $\boldsymbol{S}_0^*$.

Appendix C. Calculating the Estimated Bayes Factor in the Second Simulation

For the model $\mathcal{H}_{t01}: y_{ij}=\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}+(1-t)\boldsymbol{z}_{ij}^\top\boldsymbol{b}_i+\varepsilon_{ij}$ for $i=1,\ldots,n$ and $j=1,\ldots,m$, where $t\in[0,1]$, its first-order derivative of log joint density function has the form

$$
U(\boldsymbol{Y},t,\boldsymbol{\Xi}|\boldsymbol{X},\boldsymbol{Z})=-\sum_{i=1}^{n}\sum_{j=1}^{m}\{(y_{ij}-\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}-(1-t)\boldsymbol{z}_{ij}^\top\boldsymbol{b}_i)\boldsymbol{z}_{ij}^\top\boldsymbol{b}_i\}/\sigma_j^2. \tag{A15}
$$

In this case, $U(\boldsymbol{Y},0,\boldsymbol{\Xi}|\boldsymbol{X},\boldsymbol{Z})=-\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij}-\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}-\boldsymbol{z}_{ij}^\top\boldsymbol{b}_i)\boldsymbol{z}_{ij}^\top\boldsymbol{b}_i/\sigma_j^2$ and $U(\boldsymbol{Y},1,\boldsymbol{\Xi}|\boldsymbol{X},\boldsymbol{Z})=-\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij}-\boldsymbol{x}_{ij}^\top\boldsymbol{\beta})\boldsymbol{z}_{ij}^\top\boldsymbol{b}_i/\sigma_j^2$.

For $\mathcal{H}_{t02}: y_{ij}=(1-t)\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}+\boldsymbol{z}_{ij}^\top\boldsymbol{b}_i+\varepsilon_{ij}$ for $i=1,\ldots,n$ and $j=1,\ldots,m$, where $t\in[0,1]$, its first-order derivative of log joint density function has the form

$$
U(\boldsymbol{Y},t,\boldsymbol{\Xi}|\boldsymbol{X},\boldsymbol{Z})=-\textstyle\sum_{i=1}^{n}\sum_{j=1}^{m}\{y_{ij}-(1-t)\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}-\boldsymbol{z}_{ij}^\top\boldsymbol{b}_i\}\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}/\sigma_j^2. \tag{A16}
$$

In this case, $U(\boldsymbol{Y},0,\boldsymbol{\Xi}|\boldsymbol{X},\boldsymbol{Z})=-\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij}-\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}-\boldsymbol{z}_{ij}^\top\boldsymbol{b}_i)\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}/\sigma_j^2$ and $U(\boldsymbol{Y},1,\boldsymbol{\Xi}|\boldsymbol{X},\boldsymbol{Z})=-\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij}-\boldsymbol{z}_{ij}^\top\boldsymbol{b}_i)\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}/\sigma_j^2$.

For $\mathcal{H}_{t03}: y_{ij}=\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}+\boldsymbol{z}_{ij}^\top\boldsymbol{b}_i+\varepsilon_{ij}$ with $\varepsilon_{ij}\overset{\text{i.i.d}}{\sim}\mathcal{N}(0,t^2\sigma_0^2+(1-t)^2\sigma_j^2)$ for $i=1,\ldots,n$ and $j=1,\ldots,m$, where $t\in[0,1]$, its first-order derivative of log joint density function has the form

$$
U(\boldsymbol{Y},t,\boldsymbol{\Xi}|\boldsymbol{X},\boldsymbol{Z})=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\{t\sigma_0^2-(1-t)\sigma_j^2\}\{t^2\sigma_0^2+(1-t)^2\sigma_j^2\}^2-(y_{ij}-\mu_{ij})^2\{(1-t)\sigma_j^2-t\sigma_0^2\}}{\{t^2\sigma_0^2+(1-t)^2\sigma_j^2\}^2}. \tag{A17}
$$

In this case, $U(\boldsymbol{Y},0,\boldsymbol{\Xi}|\boldsymbol{X},\boldsymbol{Z})=-\sum_{i=1}^{n}\sum_{j=1}^{m}\{\sigma_j^4+(y_{ij}-\mu_{ij})^2\}/\sigma_j^2$ and $U(\boldsymbol{Y},1,\boldsymbol{\Xi}|\boldsymbol{X},\boldsymbol{Z})=\sum_{i=1}^{n}\sum_{j=1}^{m}\{\sigma_0^4+(y_{ij}-\mu_{ij})^2\}/\sigma_0^2$.

References

1. Lindstrom, M.J.; Bates, D.M. Newton-raphson and EM algorithms for linear mixed-effects models for repeated measures data. *J. Am. Stat. Assoc.* **1988**, *83*, 1014–1022.
2. Laird, N.; Lange, N.; Stram, D. Maximum likelihood computations with repeated measures: Applications of the EM algorithm. *J. Am. Stat. Assoc.* **1987**, *82*, 97–105. [CrossRef]
3. Zeger, S.L.; Karim, M.R. Generalized linear models with random effects: A Gibbs sampling approach. *J. Am. Stat. Assoc.* **1991**, *3*, 79–86. [CrossRef]
4. Gilks, W.R.; Wang, C.C.; Yvonnet, B.; Coursaget, P. Random-effects models for longitudinal data using Gibbs sampling. *Biometrics* **1993**, *49*, 441–453. [CrossRef] [PubMed]
5. Chen, Z.; Dunson, D.B. Random effects selection in linear mixed models. *Biometrics* **2003**, *59*, 762–769. [CrossRef] [PubMed]
6. Ahn, M.; Zhang, H.H.; Lu, W. Moment-based method for random effects selection in linear mixed models. *Stat. Sin.* **2012**, *22*, 1539–1562. [CrossRef] [PubMed]
7. Bondell, H.D.; Krishna, A.; Ghosh, S.K. Joint variable selection of fixed and random effects in linear mixed-effects models. *Biometrics* **2010**, *66*, 1069–1077. [CrossRef]
8. Ibrahim, J.G.; Zhu, H.; Garcia, R.I.; Guo, R. Fixed and random effects selection in mixed effects models. *Biometrics* **2011**, *67*, 495–503. [CrossRef]
9. Schelldorfer, J.; Buhlmann, P.; Van De Geer, S. Estimation for high-dimensional linear mixed-effects models using ℓ_1–penalization. *Scand. J. Stat.* **2011**, *38*, 197–214. [CrossRef]
10. Fan, Y.; Li, R. Variable selection in linear mixed effects models. *Ann. Stat.* **2012**, *40*, 2043–2068. [CrossRef]
11. Li, Y.; Wang, S.J.; Song, P.X.K.; Wang, N.; Zhou, L.; Zhu, J. Doubly regularized estimation and selection in linear mixed-effects models for high-dimensional longitudinal data. *Stat. Interface* **2018**, *11*, 721–737. [CrossRef] [PubMed]
12. Bradic, J.; Claeskens, G.; Gueuning, T. Fixed effects testing in high-dimensional linear mixed models. *J. Am. Stat. Assoc.* **2020**, *115*, 1835–1850. [CrossRef]
13. Li, S.; Cai, T.T.; Li, H. Inference for high-dimensional linear mixed-effects models: A quasi-likelihood approach. *J. Am. Stat. Assoc.* **2021**, 1–12. [CrossRef]
14. Berger J.; Bernardo, J.M. Reference priors in a variance components problem. In *Bayesian Analysis in Statistics and Econometrics*; Lecture Notes in Statistics; Goel, P., Ed.; Springer: New York, NY, USA, 1992; Volume 75, pp. 177–194.
15. George, E.I.; McCullogh, R.E. Variable selection via Gibbs sampling. *J. Am. Stat. Assoc.* **1993**, *88*, 881–889. [CrossRef]
16. Ishwaran, H.; Rao, J.S. Spike and slab gene selection for multigroup microarray data. *J. Am. Stat. Assoc.* **2005**, *100*, 764–780. [CrossRef]
17. Polson, N.G.; Scott, J.G. Local shrinkage rules, Levy processess and regularized regression. *J. R. Stat. Soc.* **2012**, *74*, 287–311. [CrossRef]
18. Narisetty, N.N.; He, X. Bayesian variable selection with shrinking and diffusing priors. *Ann. Stat.* **2014**, *42*, 789–817. [CrossRef]
19. Park, T.; Casella, G. The Bayesian Lasso. *J. Am. Stat. Assoc.* **2008**, *103*, 681–686. [CrossRef]
20. Griffin, J.E.; Brown, P.J. Bayesian adaptive lassos with non-convex penalization. *Aust. N. Z. J. Stat.* **2011**, *53*, 423–442. [CrossRef]
21. Rockova, V.; George, E.I. EMVS: The EM approach to Bayesian variable selection. *J. Am. Stat. Assoc.* **2014**, *109*, 828–846. [CrossRef]
22. Latouche, P.; Mattei, P.A.; Bouveyron, C.; Chiquet, J. Combining a relaxed EM algorithm with Occam's razor for Bayesian variable selection in high-dimensional regression. *J. Multivar. Anal.* **2016**, *146*, 177–190. [CrossRef]
23. Narisetty, N.N.; Shen, J.; He, X. Skinny Gibbs: A consistent and acalable Gibbs sampler for model selection. *J. Am. Stat. Assoc.* **2019**, *114*, 1205–1217. [CrossRef]
24. Wipf, D.P.; Rao, B.D.; Nagarajan, S. Latent variable Bayesian models for promoting sparsity. *IEEE Trans. Inf. Theory* **2011**, *57*, 6236–6255. [CrossRef]
25. Ghahramani, Z.; Beal, M.J. Variational inference for Bayesian mixtures of factor analysis. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2000; Volume 12, pp. 449–455.
26. Attias H. A variational Bayesian framework for graphical models. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2000; Volume 12, pp. 209–215.
27. Wu, Y.; Tang, N.S. Variational Bayesian partially linear mean shift models for high-dimensional Alzheimer's disease neuroimaging data. *Stat. Med.* **2022**, in press. [CrossRef]
28. Zhang, C.H. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* **2010**, *38*, 894–942. [CrossRef]
29. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [CrossRef]
30. Zou, H. The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **2006**, *101*, 1418–1429. [CrossRef]
31. Rockova, V.; George, E.I. The Spike-and-Slab Lasso. *J. Am. Stat. Assoc.* **2018**, *113*, 431–444. [CrossRef]
32. Leng, C.; Tran, M.N.; Nott, D. Bayesian adaptive Lasso. *Ann. Inst. Stat. Math.* **2014**, *66*, 221–244. [CrossRef]
33. Beal, M.J. Variational Algorithms for Approximate Bayesian Inference. Ph.D. Thesis, University of London, London, UK, 2003.
34. Bishop, C. *Pattern Recognition and Machine Learning*; Springer: New York, NY, USA, 2006.
35. Blei, D.M.; Kucukelbir, A.; McAuliffe, J.D. Variational inference: A review for statisticians. *J. Am. Stat. Assoc.* **2017**, *518*, 859–877. [CrossRef]

36. Lee, S.Y.; Song, X.Y. Model comparison of nonlinear structural equation models with fixed covariates. *Psychometrika* **2003**, *68*, 27–47. [CrossRef]
37. Lee, S.Y.; Tang, N.S. Bayesian analysis of nonlinear structural equation models with nonignorable missing data. *Psychometrika* **2005**, *71*, 541–564. [CrossRef]
38. Tierney, L.; Kadane, J.B. Accurate approximations for posterior moments and marginal densities. *J. Am. Stat. Assoc.* **1986**, *81*, 82–86. [CrossRef]
39. Neal, R.M. Annealed importance sampling. *Stat. Comput.* **2001**, *11*, 125–139. [CrossRef]
40. Meng, X.L.; Wong, W. Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Stat. Sin.* **1996**, *6*, 831–860.
41. Gelman, A.; Meng, X.L. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Stat. Sci.* **1998**, *13*, 163–185. [CrossRef]
42. Skilling, J. Nested sampling for general bayesian computation. *Bayesian Anal.* **2006**, *1*, 833–859. [CrossRef]
43. Friel, N.; Pettitt, A.N. Marginal likelihood estimation via power posterior. *J. R. Stat. Soc.* **2008**, *70*, 589–607. [CrossRef]
44. DiCicio, T.; Kass, R.; Raftery, A.; Wasserman, L. Computing Bayes factor by combining simulation and asymptotic approximations. *J. Am. Stat. Assoc.* **1997**, *92*, 903–915. [CrossRef]
45. LIorente, F.; Martino, L.; Delgado, D.; Lopez-Santiago, J. Marginal likelihood computation for model selection and hypothesis testing: An extensive review. *arXiv* **2022**, arXiv:2005.08334.
46. Kass, R.E.; Raftery, A.E. Bayes factors. *J. Am. Stat. Assoc.* **1995**, *90*, 773–795. [CrossRef]
47. Jack, C.; Bernstein, M.; Fox, N.; Thompson, P.; Alexander, G.; Harvey, D.; Borowski, B.; Britson, P.; Whitwell, J.; Ward, C. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *J. Magn. Reson. Imaging* **2008**, *27*, 685–691. [CrossRef] [PubMed]
48. Zhang, Y.Q.; Tang, N.S.; Qu, A. Imputed factor regression for high-dimensional block-wise missing data. *Stat. Sin.* **2020**, *30*, 631–651. [CrossRef]
49. Brookmeyer, R.; Johnson, E.; Ziegler-Graham, K.; Arrighi, H. Forecasting the global burden of Alzheimer's disease. *Alzheimers Dement.* **2007**, *3*, 186–191. [CrossRef] [PubMed]
50. Chen, Y.; Bornn, L.; De Freitas, N.; Eskelin, M.; Fang, J.; Welling, M. Herded Gibbs sampling. *J. Mach. Learn. Res.* **2016**, *17*, 263–291.
51. Martino, L.; Elvira, V.; Camps-Valls, G. The recycling Gibbs sampler for efficient learning. *Digit. Signal Process.* **2018**, *74*, 1–13. [CrossRef]
52. Roberts, G.O.; Sahu, S.K. Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *J. R. Stat. Soc.* **1997**, *59*, 291–317. [CrossRef]