Variable Selection in Sequential Hierarchical Regression Imputation

Qiushuang Li * Recai Yucel †

Abstract

We consider the problem of variable selection in the context of sequential (or variable-by-variable) imputation in clustered data. Specifically, we modify the sequential hierarchical regression imputation technique to incorporate variable selection routines using spike-and-slab priors within the Bayesian variable selection routine. Specific choice of these priors allow us to "force" variables of importance (e.g. design variables or variables known to play role in missingness mechanism) into the imputation models. Our ultimate goal is to improve computational speed by removing unnecessary variables. We employ Markov chain Monte Carlo techniques to sample from the implied posterior distributions for model unknowns as well as missing data. We assess the performance of our proposed methodology via simulation study. Our results show that our proposed algorithms lead to satisfactory estimates and in, some instances, outperform some of the existed methods that are available to practitioners.

Key Words: Clustered data, missing data, Markov chain Monte Carlo, multiple imputation, sequential hierarchical regression imputation, spike-and-slab variable selection

1. Introduction

Dealing with missing data in a statistically valid manner has been of interest in many problems in a wide-variety of disciplines. In statistical analysis of high-dimensional data, it is common to encounter large covariance matrix estimation problems for various purposes, such as dimensionality reduction, graphical modeling of conditional independence of random variables via structure learning, image processing. These analytical aims are typically complicated by arbitrary missing values (Lounici et al., 2014). In compressed sensing, one of the interesting problems is the completion of a low-rank matrix in the presence of a noisy matrix with missing entries, and there has been substantial progress in this field for the recent decade thanks to E. Candes and T. Tao (Candès and Recht, 2009, Candès and Tao, 2010). In survey sampling data, it is also typical to collect data when a portion of the subjects fails to provide responses, leading to missing values in the response variable, and missing data can occur in computer experiments as well as biomedical applications due to equipment limitations Bayarri et al. (2007). In short, there are countless examples of missing data in a broad range of fields, and sensible inferences in the presence of missing data have been gaining interest for quite some time.

It is natural for practitioners to enforce imputation of missing values in the presence of missing data. A single imputation process for the missing component of the complete data, which is constituted by concatenating the observed data and the missing data, is hardly adequate to account for the statistical uncertainty and could be potentially severely biased. Idea of multiple imputation (MI), which was first introduced by Rubin (1987), RUBIN (1976), has become a standard method to account for uncertainty due to missing data. Rather than performing a single round of imputations, MI proposes to implement multiple round of imputation of the missing value to account for the uncertainty due to missingness.

*Department of Epidemiology and Biostatistics, University at Albany, SUNY, 1 University Pl, Rensselaer, NY 12144

†Department of Epidemiology and Biostatistics, Temple University, 1301 Cecil B. Moore Ave. Philadelphia, PA 19122

The statistical analysis proceeds by treating each set of the imputed data as a set of complete data, followed by a combined analysis using Rubin's method (Rubin, 1987, RUBIN, 1976). More specifically, the MI is built upon a complete probabilistic model for the complete data, namely, a class of joint distributions of both the observed data and the missing values, from which a simulation-based approach is implemented to perform multiple imputation for the missing portion.

The most widely adopted strategy for MI is based on Bayesian modeling. It begins with first specifying the conditional distribution of the complete data given the unknown parameters, often referred to as the complete-data liklihood, and then a distribution for the unknown parameters, referred to as prior distributions. This is followed by a posterior computation via a Markov chain Monte Carlo sampler that draws random samples from the posterior distribution of the unknown parameters as well as the missing portion of the data given the observed portion of the data. Then each of the random sample drawn from the posterior predictive distribution of the missing data serves as a single round among the MI part of the missing data, and provide a copy of the imputed version of the complete data available for combined subsequent Rubin's analysis (RUBIN, 1976).

Variable selection problem arises in regression models when the number of predictors or covariates that are available to users exceeds the number of the true active predictors, and one aims to recover the correct set of active predictors. There has been vast literature on developing frequentist methods for variable selection. Classical criterion-based approaches include generalized cross-validation (GCV) and the Bayesian information criterion (BIC). These methods become computationally expensive when the number of candidate predictors becomes large as they require exhaustive search of the all possible sub-models, the number of which grows exponentially with the number of predictors. Last decade has also witnessed the progress of penalized-based approaches for variable selection (Bickel et al., 2006), including the LASSO, Smoothly Clipped Absolute Deviation (SCAD) penalty (Zou, 2006), and Adaptive LASSO (ALASSO) (Zou, 2006). These methods translate the problem of variable selection into convex programming problems and there has been relative mature algorithms for solving these mathematical optimization problems, greatly facilitating the use of penalized-likelihood methods. The challenge of these likelihood-based methods is that they require the computation of the likelihood function of the incomplete data when one is faced with missing responses and/or predictors. Such incomplete-data likelihood function is typically intractable to compute and involves high-dimensional integrals (Garcia et al., 2010). It is therefore computationally infeasible to enforce these classical penalty-based methods for variable selection in regression models with missing data.

There has also been significant progress in developing Bayesian methods for variable selection. The most widely adopted method is via the spike-and-slab prior distribution (Castillo et al., 2012, 2015). In particular, Castillo et al. (2015) extensively studied the theoretical properties for Bayesian linear regression model with fixed effects using the spike-and-slab prior distribution. Other forms of the variable selection prior include the Bayesian LASSO (Park and Casella, 2008), the horseshoe prior Carvalho et al. (2010), the Dirichlet-Laplace prior (Bhattacharya et al., 2015), and the spike-and-slab LASSO prior (Ročková et al., 2018, Ročková and George, 2018). This body of literature, however, focus on sparsity recovery and parameter estimation in regression models and do not consider missing data scenario as well as MI, which is the focus of this work.

In this paper, we are interested in methods for variable selection in the presence of missing data for both continuous value responses as well as binary value responses, using generalized linear mixed-effects models.In particular, our contributions are:

1. The proposed method is able to simultaneously perform variable selection and multiple imputation of missing responses for continuous and binary responses via mixed-

effects models. We build this based on generalized linear mixed-effects models and the posterior inference for the unknown parameters, including variable selection, as well as the simulation-based MI for the missing responses. For computations, we utilize a Markov chain Monte Carlo sampler. Coefficients of the underlying regression models are assigned a spike-and-slab prior distribution that allows variable selection *a posteriori*.

2. For the classical linear model with normal errors, we derive the corresponding full conditional distributions of all the parameters involved, together with the full conditional posterior predictive distributions of the missing variables, thanks to the standard conjugate normal model, facilitating the implementation of a computationally accessible Gibbs sampler.

3. For the binary response variables, we consider the generalized linear mixed-effect model with a logit link function, also referred to as the logistic linear mixed-effects model. The full conditional distribution for the linear coefficients as well as the random-effects coefficients are not in closed-forms directly, and we borrow the parameter expansion for data augmentation (PX-DA) idea of Polson et al. (2013) by introducing the cleverly-designed auxiliary Pólya-Gamma random variables, such that the full conditional distributions of the expanded set of the parameters are easily accessible, whereas the marginal distribution of the original (unexpanded) set of parameters is left invariant. As a consequence, we develop an easy-to-implement Gibbs sampler as well.

4. For the simulation-based MI via the MCMC, rather than jointly drawing a set of the random sample from the joint predictive posterior distribution of the missing variables, we follow the idea of Yucel et al. (2018) and draw each of the missing variable sequentially via the full conditional predictive distribution of the corresponding variable, referred to as sequential hierarchical regression imputation (SHRIMP). The fundamental idea of SHRIMP is that one first sort the missing variables by their corresponding percentages of missing portion in the increasing order, and then draw samples from the posterior predictive distribution of the missing variables following this sorted order. The formal description of the SHRIMP strategy will be introduced in Section 2. The advantage of this variable-by-variable MI strategy is that it reduces the computational complexity for high-dimensional data (Yucel et al., 2018) significantly.

A frequentist version for variable selection in regression models in the presence of missing data is fully addressed in Garcia et al. (2010). The major difference is that our approach is built upon a fully Bayesian methodology that allows for parameter estimation and inference, variable selection, and the implementation of MI simultaneously via a coherent Gibbs sampler, whereas Garcia et al. (2010) focused on developing easy-to-compute penalized likelihood approach and focus on the inference goal via point estimators, together with some well-established theoretical properties, and MI needs to be performed separately.

The rest of this paper is organized as follows. Section 2 is devoted to the linear mixed-effects regression model for variable selection with missing responses for continuous value response variables, in which a Gibbs sampler is developed. For binary response variables, we elaborate the logistic mixed-effects model for the same tasks in Section 3, introduce the Pólya-Gamma random variables, leverage them for the PX-DA, and successfully develop a closed-form Gibbs sampler as well. These two Gibbs samplers allow simulatenous inference of the parameters and MI of the missing responses. The advantage of the proposed

approach is illustrated via numerical examples in Section 4, and we conclude the paper with a discussion in Section 5.

2. Linear mixed-effects regression models with missing responses

Let us consider a linear mixed-effects model with random intercept only for continuous response variable y_{ij}, which has also been considered in Yucel et al. (2018):

$$y_{ij} = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + b_i + \epsilon_{ij}, \quad i = 1, \ldots, m, \quad j = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the fixed-effect, $b_1, \ldots, b_m \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma_b^2)$ are the random effects, and $\epsilon_{11}, \ldots, \epsilon_{mn} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma_e^2)$ are the errors. The responses y_{ij}'s are either observed or missing, but the missing portion can be imputed via the last cycle of the SHRIMP strategy, as is suggested in Yucel et al. (2018). Finally, $\mathbf{x}_{ij} \in \mathbb{R}^p$'s are the individual-level covariates that can also be either observed or missing, and the missing values are sampled through the SHRIMP strategy.

We develop a Gibbs sampler to draw posterior samples from the joint distribution of $(\boldsymbol{\beta}, b_1, \ldots, b_m, \sigma_b, \sigma_e)$, as well as to draw samples of the missing data (y_{mis}). To select the variables among $x_{ij1}, \ldots, x_{ijp}$, we assign a spike-and-slab prior distribution, which has been widely applied to Bayesian variable selection (Mitchell and Beauchamp, 1988, George and McCulloch, 1993, Clyde et al., 1996, Geweke, 1996, Kuo and Mallick, 1998), is assigned to the fixed-effects coefficient β_k. In the current context, if we are not certain whether the k^{th} variable is to be selected, then we assign the following spike-and-slab prior to β_k:

$$\beta_k \mid w, \mu_0, \sigma_0 \begin{cases} = 0, & \text{with probability } (1-w), \\ \sim \mathrm{N}(\mu_0, \sigma_0^2), & \text{with probability } w, \end{cases} \tag{2}$$

where $w > 0$ is the prior probability that the kth variable x_{ijk} is selected, and with probability $(1-w)$, β_k is set to 0 so that under the prior distribution, the kth variable is not selected. The spike-and-slab prior distribution (2) can be equivalently written as

$$(\beta_k \mid w, \mu_0, \sigma_0) \sim (1-w)\delta_0 + w\mathrm{N}(\mu_0, \sigma_0^2),$$

where δ_0, point mass at 0, is assigned a normal prior if there is a sure certainty of selection:

$$(\beta_k \mid w, \mu_0, \sigma_0) \sim \mathrm{N}(\mu_0, \sigma_0^2).$$

To reduce the effect of hyperparameters and enhance the robustness of the entire Bayesian model, we further assume that the hyperparameters have the following hyperprior distributions: $w \sim \mathrm{Beta}(a_w, b_w)$, $\mu_0 \sim \mathrm{N}(0, 1)$, and $\sigma^2 \sim \mathrm{Inverse-Gamma}(1, 1)$. For the rest of the parameters (σ_b^2, σ_e^2), we assume the inverse-χ^2 distribution for the sake of conjugacy, which has also been adopted in Yucel et al. (2018): $\sigma_b^2 \sim \chi_{\nu_b}^{-2}$ and $\sigma_e^2 \sim \chi_{\nu_e}^{-2}$.

We provide the detailed full conditional distributions that are needed for the Gibbs sampler in Appendix A. Here we focus on the conditional distribution of the linear coefficients β_k, $k = 1, 2, \ldots, p$. Denote the parameters $\boldsymbol{\theta}_{-k}$ be the set of all parameters except β_k: $\boldsymbol{\theta}_{-k} = (\boldsymbol{\beta}_{-k}, \sigma_b, \sigma_e)$, where $\boldsymbol{\beta}_{-k} = \{\beta_1, \ldots, \beta_p\} \backslash \{\beta_k\}$, and the random effects $\mathbf{b} = [b_1, \ldots, b_m]^{\mathrm{T}}$. Then the full conditional distribution of β_k for $k = 1, 2, \ldots, p$ is given by

$$(\beta_k \mid \mathbf{X}, \boldsymbol{\theta}_{-k}, w, \mu_0, \sigma_0) \sim \begin{cases} w_1^*\delta_0 + w_2^*\mathrm{N}(\widehat{\mu}, \widehat{V}), & \text{if } x_{ijk} \text{ is not required,} \\ \mathrm{N}(\widehat{\mu}, \widehat{V}), & \text{if } x_{ijk} \text{ is required,} \end{cases} \tag{3}$$

where $\mathbf{X}$ denotes the full set of covariates $\mathbf{X} = [\mathbf{x}_{ij}]_{i=1,\ldots,m, j=1,\ldots,n}$,

$$w_1^* \propto (1-w)\mathcal{N}\left(0 \,\middle|\, \frac{\sum_{i,j} x_{ijk}(y_{ij} - \sum_{\ell \neq k} x_{ij\ell}\beta_\ell - b_i)}{\sum_{i,j} x_{ijk}^2}, \frac{\sigma_e^2}{\sum_{i,j} x_{ijk}^2}\right),$$

$$w_2^* \propto w\mathcal{N}\left(\mu_0 \,\middle|\, \frac{\sum_{i,j} x_{ijk}(y_{ij} - \sum_{\ell \neq k} x_{ij\ell}\beta_\ell - b_i)}{\sum_{i,j} x_{ijk}^2}, \sigma_0^2 + \frac{\sigma_e^2}{\sum_{i,j} x_{ijk}^2}\right),$$

$$\widehat{V} = \left(\frac{1}{\sigma_e^2}\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ijk}^2 + \frac{1}{\sigma_0^2}\right)^{-1},$$

$$\widehat{\mu} = \widehat{V}\left[\frac{\mu_0}{\sigma_0^2} + \frac{1}{\sigma_e^2}\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ijk}\left(y_{ij} - \sum_{\ell \neq k} x_{ij\ell}\beta_\ell - b_i\right)\right],$$

The derivation of the rest of the full conditional distributions are routine and are deferred to Appendix A. We also emphasize that (3) presents the nature of variable selection inside a single cycle of the Gibbs sampler: with probability w_1^*, we set $\beta_k = 0$, suggesting that currently the kth variable x_{ijk} is not selected, and with probability w_2^*, we draw β_k from a normal distribution, indicating that $\beta_k \neq 0$, and therefore, the kth variable x_{ijk} needs to be selected.

For a given set of values $\boldsymbol{\beta}, \sigma_b, \sigma_e, b_1, \ldots, b_m$ that are drawn from a single cycle of the Gibbs sampler, each missing response can be drawn from

$$(y_{ij} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma_b, \sigma_e, b_1, \ldots, b_m) \sim \mathrm{N}(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + b_i, \sigma_e^2),$$

where "j" indicates missing value(s) among the i^{th} observational (cluster) unit. Here we adopt the SHRIMP strategy to draw the samples $y_{(mis)}$ given $y_{(obs)}$ and $\boldsymbol{\theta}$ in the following sequential fashion:

- **Step 1:** Order the column indices $\{1, 2, \ldots, n\}$ of the response matrix $\mathbf{Y}$ such that the sorted indices, say $\{j_1, \ldots, j_n\}$, satisfy

  $$\sum_{i=1}^{n} \mathbb{1}(y_{ij_k} = \mathrm{NA}) \leq \sum_{i=1}^{n} \mathbb{1}(y_{ij_{k+1}} = \mathrm{NA}),$$

  i.e., the number of missing values of the j_kth column is always no greater than the number of missing values of the j_{k+1}th column.

- **Step 2:** Sample $\{y_{ij}\}$ where j denotes the missing data value for j^{th} observation in cluster i.

The idea of the SHRIMP strategy is to impute missing values in a variable in an order defined according to the amount of missingness (from least missing to most). By iterating the above cycles for sufficiently large number of times, we are able to obtain a sequence of parameters drawn from the Gibbs sampler $\{\boldsymbol{\theta}_{(1)}, \boldsymbol{\theta}_{(2)}, \ldots\}$, which converges in distribution to $\{y_{\mathrm{mis}(1)}, y_{\mathrm{mis}(2)}, \ldots\}$ as the number of cycles goes to infinity, as well as a sequence of missing responses $y_{\mathrm{mis}} = \{y_{\mathrm{mis}(1)}, y_{\mathrm{mis}(2)}, \ldots\}$, whose limiting distribution is $P(y_{\mathrm{mis}} \mid y_{\mathrm{obs}}, \mathbf{X}, b_1, \ldots, b_m)$, where y_{obs} denotes the observed y-values. After the Gibbs sampler is completed and the Markov chain converges, we sample y_{mis} from its predictive distribution with the final set of drawn values of all the parameters. For the purpose of multiple imputation, one can repeat this procedure for M times to obtain M copies of the imputed data.

3. Logistic mixed-effects regression models with missing responses

We assume that our binary variable follows a logistic mixed-effects regression model:

$$\mathbb{P}\left(y_{ij}=1 \mid \mathbf{x}_{ij}, b_i, \boldsymbol{\beta}\right)=\frac{1}{1+\exp(-\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}-b_i)},$$

where $\boldsymbol{\beta}$ are the fixed-effects coefficients for covariates x_i, and $b_1,\ldots,b_m \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,\sigma_b^2)$ are the random effects. To develop a closed-form Gibbs sampler for the logistic regression model with respect to the random-effects, we adopt a similar strategy suggested by Polson et al. (2013). They suggest introducing a collection of auxiliary variables following the *Pólya-Gamma* distribution, such that the full conditional distributions of all parameters are obtainable in closed-form. Before understanding the mechanism, we first present the definition of the Pólya-Gamma distribution (see Definition 1 in Polson et al., 2013): A random variable X is said to follow a Pólya-Gamma distribution with parameters $b>0$ and $c\in\mathbb{R}$, denoted by $X\sim\mathrm{PG}(b,c)$, if there exists a sequence of independent Gamma random variables $(g_k)_{k=1}^{\infty} \overset{\text{i.i.d.}}{\sim} \mathrm{Gamma}(b,1)$, such that

$$X=\frac{1}{2\pi^2}\sum_{k=1}^{\infty}\frac{g_k}{(k-1/2)^2+c^2/(4\pi^2)}.$$

The key result of the Pólya-Gamma distribution lies in the following theorem, which is established in Polson et al. (2013):

Theorem 1 (Theorem 1 in Polson et al., 2013) *Let $p(\omega)$ be the density function of $\omega\sim\mathrm{PG}(b,0)$, $b>0$. Then the following integral identity holds for all $a\in\mathbb{R}$:*

$$\frac{[\exp(\psi)]^a}{[1+\exp(\psi)]^b}=2^{-b}\exp\left[\left(a-\frac{b}{2}\right)\psi\right]\int_0^{\infty}\exp\left(-\frac{1}{2}\omega\psi^2\right)p(\omega)\mathrm{d}\omega.$$

Moreover, the normalized integrand

$$p(\omega\mid\psi)=\frac{\exp\left(-\omega\psi^2/2\right)p(\omega)}{\int_0^{\infty}\exp\left(-\omega\psi^2/2\right)p(\omega)\mathrm{d}\omega}$$

is the density function of $\omega\sim\mathrm{PG}(b,\psi)$.

We let the following prior distributions reflect the appropriate prior knowledge on the fixed-effects coefficients $\beta_1,\ldots,\beta_p$. In light of the need for variable selection, we assign the spike-and-slab prior (2) to $\beta_1,\ldots,\beta_p$ as follows

$$\begin{aligned}
&(\beta_k\mid w,\mu_0,\sigma_0^2)\sim(1-w)\delta_0+w\mathrm{N}(\mu_0,\sigma_0^2), \quad \text{if the } k\text{th variable is undetermined,}\\
&(\beta_k\mid w,\mu_0,\sigma_0^2)\sim\mathrm{N}(\mu_0,\sigma_0^2), \quad \text{if the } k\text{th variable is forced to be selected,}\\
&w\sim\mathrm{Beta}(a_w,b_w), \quad \mu_0\sim\mathrm{N}(0,1), \quad \sigma^2\sim\mathrm{IG}(1,1).
\end{aligned} \tag{4}$$

The prior distribution on σ_b is same as Section 2: $\sigma_b^2\sim\chi_{\nu_b}^{-2}$.

We now elaborate on the full conditional distributions of the linear coefficients β_k, $k=1,2,\ldots,p$. The rest of the full conditional distributions necessary for deriving the Gibbs sampler that draws posterior samples from the joint distribution of $(\boldsymbol{\beta},b_1,\ldots,b_m)$, together with the samples of the missing data (y_{mis}), are provided in Appendix B. Following the derivation in Polson et al. (2013), we utilize Theorem 1 and derive the likelihood function of $\eta_{ij}:=\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}+b_i$

$$\mathcal{L}(\eta_{ij}\mid y_{ij})\propto\exp\left[\left(y_{ij}-\frac{1}{2}\right)\eta_{ij}\right]\int_0^{\infty}\exp\left(-\frac{\omega_{ij}\eta_{ij}^2}{2}\right)p(\omega_{ij}\mid 1,0)\mathrm{d}\omega_{ij},$$

where $p(\omega_{ij} \mid 1, 0)$ is the density of an auxiliary Pólya-Gamma random variable $\omega_{ij} \sim \text{PG}(1, 0)$. The idea of introducing the auxiliary variables ω_{ij}'s is such that after marginalizing them out, the joint distribution of the rest variables is left invariant. We derive the likelihood of $\boldsymbol{\beta}$ for all mn data points after introducing $\boldsymbol{\Omega} = [\omega_{ij}]_{m\times n}$:

$$\mathcal{L}(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{Y}, \boldsymbol{\Omega}, b_1, \ldots, b_m, \sigma^2) \propto \exp\left\{-\frac{1}{2}(\mathbf{z} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{z} - \mathbf{X}\boldsymbol{\beta})\right\},$$

where $z_{ij} = (y_{ij} - 1/2)/\omega_{ij} - b_i$,

$$\begin{aligned}
\mathbf{z} &= [z_{11}, \ldots, z_{1n}, z_{21}, \ldots, z_{2n}, \ldots, z_{m1}, \ldots, z_{mn}]^{\mathrm{T}} \in \mathbb{R}^{mn}, \\
\mathbf{X} &= [\mathbf{x}_{11}, \ldots, \mathbf{x}_{1n}, \mathbf{x}_{21}, \ldots, \mathbf{x}_{2n}, \ldots, \mathbf{x}_{m1}, \ldots, \mathbf{x}_{mn}]^{\mathrm{T}} \in \mathbb{R}^{mn\times p}, \\
\boldsymbol{\Sigma}^{-1} &= \text{diag}(\omega_{11}, \ldots, \omega_{1n}, \omega_{21}, \ldots, \omega_{2n}, \ldots, \omega_{m1}, \ldots, \omega_{mn}) \in \mathbb{R}^{mn\times mn}.
\end{aligned}$$

We then obtain the following closed-form full conditional distribution of β_k, $k = 1, 2, \ldots, p$:

$$(\beta_k \mid \mathbf{X}, \mathbf{Y}, \boldsymbol{\Omega}, \boldsymbol{\beta}_{-k}, \mathbf{b}, \sigma_b, w, \mu_0, \sigma_0) \sim \begin{cases} w_1^*\delta_0 + w_2^*\text{N}(\widehat{\mu}, \widehat{V}), & \text{if } x_{ijk} \text{ is not required,} \\ \text{N}(\widehat{\mu}, \widehat{V}), & \text{if } x_{ijk} \text{ is required,} \end{cases} \tag{5}$$

where $\mathbf{X}$ denotes the full set of covariates $\mathbf{X} = [\mathbf{x}_{ij}]_{i=1,\ldots,m, j=1,\ldots,n}$, and

$$\begin{aligned}
w_1^* &\propto (1-w)\mathcal{N}\left(0 \,\middle|\, \frac{\sum_{i,j}\omega_{ij}x_{ijk}(z_{ij} - \sum_{\ell\neq k}x_{ij\ell}\beta_\ell)}{\sum_{i,j}\omega_{ij}x_{ijk}^2}, \frac{1}{\sum_{i,j}\omega_{ij}x_{ijk}^2}\right), \\
w_2^* &\propto w\mathcal{N}\left(\mu_0 \,\middle|\, \frac{\sum_{i,j}\omega_{ij}x_{ijk}(z_{ij} - \sum_{\ell\neq k}x_{ij\ell}\beta_\ell)}{\sum_{i,j}\omega_{ij}x_{ijk}^2}, \sigma_0^2 + \frac{1}{\sum_{i,j}\omega_{ij}x_{ijk}^2}\right), \\
\widehat{V} &= \left(\sum_{i=1}^{m}\sum_{j=1}^{n}\omega_{ij}x_{ijk}^2 + \frac{1}{\sigma_0^2}\right)^{-1}, \\
\widehat{\mu} &= \widehat{V}\left[\frac{\mu_0}{\sigma_0^2} + \sum_{i=1}^{m}\sum_{j=1}^{n}\omega_{ij}x_{ijk}\left(z_{ij} - \sum_{\ell\neq k}x_{ij\ell}\beta_\ell\right)\right].
\end{aligned}$$

The full conditional distribution of the auxiliary variables $\boldsymbol{\Omega} = [\omega_{ij}]_{m\times n}$ can be derived similarly as that in Polson et al. (2013):

$$(\omega_{ij} \mid \boldsymbol{\beta}, b_1, \ldots, b_m) \sim \text{PG}(1, \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + b_i), \tag{6}$$

and sampling a random variable following a Pólya-Gamma distribution can be implemented using the algorithm described in Secion 4 in Polson et al. (2013). The derivation of the rest of the full conditional distributions are similar to those in Section 2 and we leave them in Appendix B. Finally, for each missing $y_{ij} \in (y_{\text{mis}})$, one can draw it from the following conditional distribution in a single cyle of the Gibbs sampler:

$$(y_{ij} \mid \mathbf{X}, \boldsymbol{\beta}, b_1, \ldots, b_m) \sim \text{Bernoulli}\left(\frac{1}{1 + \exp(-\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} - b_i)}\right).$$

Similar to our algorithm of Gibbs sampler in Section 2, the above procedure is performed to lead to imputed values for the binary variables (ordered from highest to lowest missing values) with selected as well as forced covariates. The post-MCMC analysis is the same as that in Section 2.

4. Simulated examples

4.1 A linear mixed-effects regression model example

We begin our simulated examples with the classical linear mixed-effects regression model with normal errors. Our simulation proceeds for y under the following linear mixed-effects model:

$$y_{ij} = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + b_i + \epsilon_{ij}, \quad m = 1, \ldots, n, \quad j = 1, \ldots, n,$$

where $(\mathbf{x}_{ij} : i = 1, \ldots, n, j = 1, \ldots, m)$ are p-dimensional predictor vectors, $\boldsymbol{\beta}$ is the p-dimensional fixed-effect linear coefficients, $b_1, \ldots, b_m \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma_b^2)$ are random effects, and $\epsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma_e^2)$ are independent homoscedastic errors. Here we consider $m = 50$ and $j = 100$. The coordinates of the covariates of $\mathbf{x}_{ij}$'s are generated independently from $\mathrm{N}(0, 3^2)$, and the true value of $\boldsymbol{\beta}$ is generated as follows: First set

$$\boldsymbol{\alpha} = [\alpha_1, \alpha_2, 0, 0, \alpha_3, \alpha_4, 0, \alpha_5, \alpha_6, 0],$$

where $\alpha_1, \ldots, \alpha_6 \sim \mathrm{N}(1, 0.1^2)$ independently. Then set $\boldsymbol{\beta} = \boldsymbol{\alpha}/\|\boldsymbol{\alpha}\|_2$. The response matrix $\mathbf{Y} = [y_{ij}]_{m\times n}$ is assumed to be contaminated by missing values, and we consider two scenarios of missingness mechanism:

- Missing completely at random (MCAR): the probability for missingness of y_{ij} follows $\mathrm{Bernoulli}(0.4)$ independently for all $i = 1, \ldots, m$ and $j = 1, \ldots, n$.
- Missing at random (MAR): the missingness of y_{ij} follows $\mathrm{Bernoulli}(p_{ij})$ independently for all $i = 1, \ldots, m$ and $j = 1, \ldots, n$, where $\mathrm{logit}(p_{ij}) = -\sum_{k=1}^{p} |x_{ijk}|/20$. This results in the percentage of missingness around 20%.

We then employ the Gibbs sampler developed in Section 2 for posterior inference and MI, and the number of MIs is set to $M = 5$. For each set of the imputed data, posterior median and standard deviation $\boldsymbol{\beta}$ are computed. Then we combine these estimands using Rubin's rules (RUBIN, 1976). The results are summarized in Table 1 under the MCAR and Table 2 under the MAR, respectively.

Table 1: Linear mixed-effects model with missing completely at random (MCAR) probability 0.4, $m = 50$, $n = 100$, and $p = 10$. Imputation and estimation method is the Bayesian method with the spike-and-slab (SS) prior and sequential hierarchical regression imputation (SHRIMP). Number of multiple imputation is $M = 5$.

β	True values	Estimate	Credible intervals (CI)	CI width	Total variance
β_1	0.3726	0.3395	(0.3747, 0.3043)	0.0704	2.4917×10^{-4}
β_2	0.3667	0.3915	(0.4156, 0.3675)	**0.0481**	1.4139×10^{-4}
β_3	0.0000	**0.0000**	(0.0003, -0.0003)	**0.0005**	1.7423×10^{-8}
β_4	0.0000	**0.0000**	(0.0012, -0.0012)	**0.0023**	3.5239×10^{-7}
β_5	0.4326	0.4280	(0.4595, 0.3965)	0.0630	2.0917×10^{-4}
β_6	0.3934	0.3985	(0.4217, 0.3752)	**0.0465**	1.3361×10^{-4}
β_7	0.0000	**0.0000**	(0.0002, -0.0002)	**0.0004**	1.2171×10^{-8}
β_8	0.4370	0.4325	(0.4690, 0.3959)	0.0731	2.6499×10^{-4}
β_9	0.4401	0.4330	(0.4594, 0.4067)	0.0526	1.6186×10^{-4}
β_{10}	0.0000	**0.0000**	(0.0018, -0.0018)	**0.0036**	8.2747×10^{-7}

We also implement the `pan` package (Zhao and Schafer, 2013) and the `mice` (Buuren and Groothuis-Oudshoorn, 2010) package for comparison. The corresponding

Table 2: Linear mixed-effects model with missing at random (MAR) probability 0.4, $m = 50$, $n = 100$, and $p = 10$. Imputation and estimation method is the Bayesian method with the spike-and-slab (SS) prior and sequential hierarchical regression imputation (SHRIMP). Number of multiple imputation is $M = 5$.

β	True values	Estimate	Credible intervals	CI width	Total variance
β_1	0.3652	0.3868	(0.4081, 0.3655)	0.0426	1.1555×10^{-4}
β_2	0.4289	0.4161	(0.4349, 0.3973)	**0.0376**	9.1991×10^{-5}
β_3	0.0000	**0.0000**	(0.0000, 0.0000)	**0.0000**	0.0000
β_4	0.0000	**0.0000**	(0.0000, 0.0000)	**0.0000**	4.9048×10^{-13}
β_5	0.3985	0.4027	(0.4272, 0.3782)	0.0489	1.4247×10^{-4}
β_6	0.4165	0.4065	(0.4256, 0.3874)	**0.0382**	9.4402×10^{-5}
β_7	0.0000	**0.0000**	(0.0000, 0.0000)	**0.0009**	5.6395×10^{-8}
β_8	0.4088	0.3978	(0.4166, 0.3791)	**0.0375**	9.1407×10^{-5}
β_9	0.4282	0.4224	(0.4453, 0.3994)	0.0459	1.2999×10^{-4}
β_{10}	0.0000	**0.0000**	(0.0000, 0.0000)	**0.0008**	4.2961×10^{-8}

combined Rubin's analysis results for the `pan` package are tabulated in Table 3 under the MCAR and Table 4 under the MAR, and the corresponding results for the `mice` package are listed in Table 5 and Table 6 under the MCAR and MAR, respectively. The numerical results for the three approaches under the MCAR and the MAR are also visualized in Figure 1 and Figure 2, respectively. From both the tables and the plots, one can identify that the `pan` package can estimate $\boldsymbol{\beta}$ well but is unable to detect the sparsity pattern of $\boldsymbol{\beta}$, and hence is not successful in variable selection; It can also be seen that for the `mice` package, it is unable to estimate $\boldsymbol{\beta}$ accurately, losses the coverage of the confidence intervals for $\boldsymbol{\beta}$, and is not successful in variable selection. On the contrary, we can see that the proposed approach outperforms some of the alternatives (e.g., the `pan` package and the `mice` package) in terms of both accuracy for estimating the regression coefficient $\boldsymbol{\beta}$, detecting the sparsity pattern of $\boldsymbol{\beta}$, and uncertainty quantification assessed by the width of the confidence intervals.

Table 3: Linear mixed-effects model with missing completely at random (MCAR) probability 0.4, $m = 50$, $n = 100$, and $p = 10$. Imputation and estimation method is the `pan` package. Number of multiple imputation is $M = 5$.

β	True values	Estimate	Confidence intervals (CI)	CI width	Total variance
β_1	0.3726	0.3387	(0.3632, 0.3142)	0.0491	1.5664×10^{-4}
β_2	0.3667	0.3829	(0.4076, 0.3581)	0.0494	1.5909×10^{-4}
β_3	0.0000	-0.0098	(0.0139, -0.0335)	0.0473	1.4567×10^{-4}
β_4	0.0000	-0.0144	(0.0105, -0.0392)	0.0497	1.6080×10^{-4}
β_5	0.4326	0.4222	(0.4465, 0.3979)	0.0486	1.5397×10^{-4}
β_6	0.3934	0.4010	(0.4255, 0.3766)	0.0489	1.5585×10^{-4}
β_7	0.0000	-0.0102	(0.0141, -0.0345)	0.0485	1.5332×10^{-4}
β_8	0.4370	0.4385	(0.4633, 0.4137)	0.0496	1.6007×10^{-4}
β_9	0.4401	0.4339	(0.4590, 0.4088)	0.0501	1.6351×10^{-4}
β_{10}	0.0000	-0.0184	(0.0064, -0.0431)	0.0495	1.5963×10^{-4}

Table 4: Linear mixed-effects model with missing at random (MAR) probability 0.4, $m = 50$, $n = 100$, and $p = 10$. Imputation and estimation method is the `pan` package. Number of multiple imputation is $M = 5$.

β	True values	Estimate	Confidence intervals (CI)	CI width	Total variance
β_1	0.3652	0.3805	(0.4017, 0.3593)	0.0424	1.1700×10^{-4}
β_2	0.4289	0.4168	(0.4386, 0.3950)	0.0435	1.2336×10^{-4}
β_3	0.0000	0.0057	(0.0271, -0.0157)	0.0428	1.1928×10^{-4}
β_4	0.0000	0.0028	(0.0237, -0.0182)	0.0419	1.1407×10^{-4}
β_5	0.3985	0.4016	(0.4229, 0.3802)	0.0427	1.1875×10^{-4}
β_6	0.4165	0.4053	(0.4265, 0.3842)	0.0423	1.1641×10^{-4}
β_7	0.0000	-0.0096	(0.0126, -0.0318)	0.0444	1.2812×10^{-4}
β_8	0.4088	0.3983	(0.4203, 0.3763)	0.0439	1.2557×10^{-4}
β_9	0.4282	0.4289	(0.4505, 0.4073)	0.0432	1.2130×10^{-4}
β_{10}	0.0000	-0.0121	(0.0097, -0.0340)	0.0437	1.2414×10^{-4}

Table 5: Linear mixed-effects model with missing completely at random (MCAR) probability 0.4, $m = 50$, $n = 100$, and $p = 10$. Imputation and estimation method is the `mice` package. Number of multiple imputation is $M = 5$.

β	True values	Estimate	Confidence intervals (CI)	CI width	Total variance
β_1	0.3726	0.1960	(0.2469, 0.1451)	0.1017	5.9542×10^{-4}
β_2	0.3667	0.2288	(0.2713, 0.1863)	0.0849	4.5158×10^{-4}
β_3	0.0000	-0.0134	(0.0330, -0.0599)	0.0930	5.1194×10^{-4}
β_4	0.0000	-0.0109	(0.0269, -0.0488)	0.0757	3.7076×10^{-4}
β_5	0.4326	0.2553	(0.2985, 0.2120)	0.0865	4.5857×10^{-4}
β_6	0.3934	0.2273	(0.2774, 0.1772)	0.1003	5.8649×10^{-4}
β_7	0.0000	-0.0031	(0.0420, -0.0483)	0.0903	4.9888×10^{-4}
β_8	0.4370	0.2443	(0.2833, 0.2054)	0.0779	3.8950×10^{-4}
β_9	0.4401	0.2675	(0.3093, 0.2258)	0.0835	4.3849×10^{-4}
β_{10}	0.0000	-0.0022	(0.0654, -0.0698)	0.1351	9.2000×10^{-4}

Table 6: Linear mixed-effects model with missing at random (MAR) probability 0.4, $m = 50$, $n = 100$, and $p = 10$. Imputation and estimation method is the `mice` package. Number of multiple imputation is $M = 5$.

β	True values	Estimate	Confidence intervals (CI)	CI width	Total variance
β_1	0.3652	0.3017	(0.3445, 0.2590)	0.0854	4.3491×10^{-4}
β_2	0.4289	0.3328	(0.3717, 0.2939)	0.0778	3.7629×10^{-4}
β_3	0.0000	-0.0024	(0.0388, -0.0436)	0.0824	4.1141×10^{-4}
β_4	0.0000	-0.0029	(0.0328, -0.0387)	0.0716	3.2681×10^{-4}
β_5	0.3985	0.3097	(0.3512, 0.2682)	0.0829	4.1551×10^{-4}
β_6	0.4165	0.3190	(0.3623, 0.2758)	0.0866	4.4447×10^{-4}
β_7	0.0000	-0.0191	(0.0229, -0.0611)	0.0840	4.2886×10^{-4}
β_8	0.4088	0.3062	(0.3475, 0.2650)	0.0825	4.1432×10^{-4}
β_9	0.4282	0.3437	(0.3921, 0.2954)	0.0967	5.2824×10^{-4}
β_{10}	0.0000	-0.0205 (	0.0152, -0.0562)	0.0714	3.2673×10^{-4}

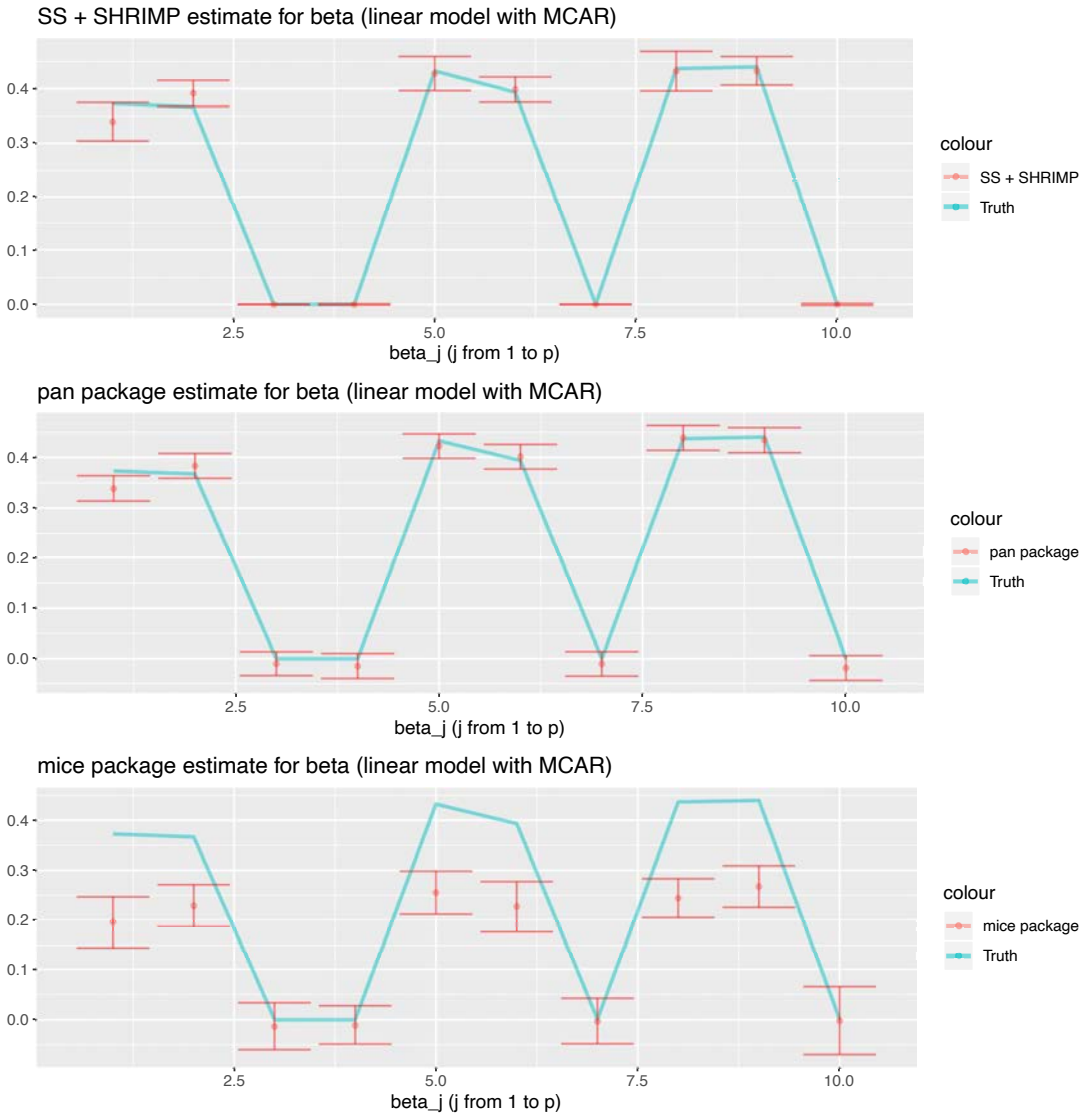

Figure 1: Simulation performance of the SS-SHRIMP method, the pan package, and the mice package, for multipe imputation and estimation of the linear mixed-effects model with MCAR: The red dots are the multiple imputation estimates of β, the red bars in the top panels are estimated 95% confidence intervals for β after multiple imputation, and the blue lines represent the true values of β.

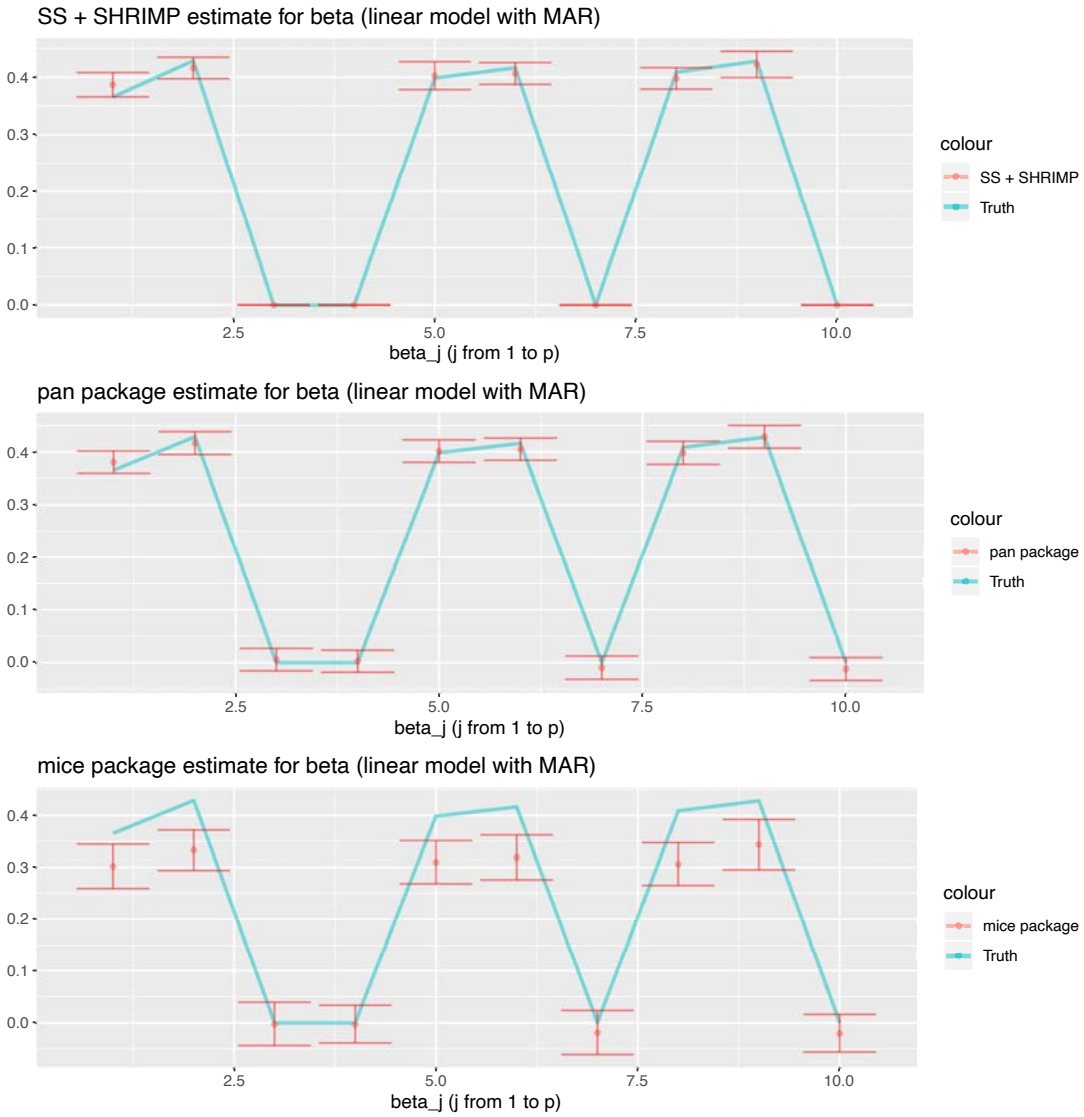

Figure 2: Simulation performance of the SS-SHRIMP method, the pan package, and the mice package, for multipe imputation and estimation of the linear mixed-effects model with MAR: The red dots are the multiple imputation estimates of $\boldsymbol{\beta}$, the red bars in the top panels are estimated 95% confidence intervals for $\boldsymbol{\beta}$ after multiple imputation, and the blue lines represent the true values of $\boldsymbol{\beta}$.

4.2 Logistic Mixed-effects Regression Model

The second simulated example is a generalized linear mixed-effects model with the logit link, i.e., the logistic version of the model in Section 4.1. Simulation proceeds under

$$y_{ij} \sim \text{Bernoulli}(p_{ij}), \quad i = 1, \ldots, m, \quad j = 1, \ldots, n,$$

where

$$\text{logit}(p_{ij}) = \log(p_{ij}/(1 - p_{ij})) = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + b_i,$$

$\mathbf{x}_{ij}$'s are p-dimensional predictors, $\boldsymbol{\beta}$ is the p-dimensional vector of linear coefficients, and $b_1, \ldots, b_m \overset{\text{i.i.d.}}{\sim} \text{N}(0, \sigma_b^2)$ are random effects. We set $m = 50$ and $n = 50$ in this example. The covariates $\mathbf{x}_{ij}$'s and the true value of $\boldsymbol{\beta}$ are generated following the same distribution as those in Section 4.1. The response matrix $\mathbf{Y} = [y_{ij}]_{n\times m}$ is now a $m \times n$ binary matrix, and we assume that it is also contaminated by missing values. The missingness mechanism is set to be the same as the MAR in Section 4.1.

The Gibbs sampler introduced in Section 3 based on the PX-DA strategy is implemented for posterior computation of $\boldsymbol{\beta}$ and MI, and the number of imputation times is set to $M = 5$. A similar MI inference is presented in Table 7.

Table 7: Logistic mixed-effects model with missing at random (MAR) probability 0.4, $m = 50$, $n = 100$, and $p = 10$. Imputation and estimation method is the Bayesian method with the spike-and-slab (SS) prior and sequential hierarchical regression imputation (SHRIMP). Number of multiple imputation is $M = 5$.

β	True values	Estimate	Credible intervals (CI)	CI width	Total variance
β_1	0.4108	0.3900	(0.4368, 0.3433)	**0.0935**	5.6173×10^{-4}
β_2	0.3331	0.3079	(0.3661, 0.2497)	0.1164	7.8233×10^{-4}
β_3	0.0000	**0.0000**	(0.0048, -0.0048)	**0.0096**	6.0540×10^{-6}
β_4	0.0000	**0.0000**	(0.0078, -0.0078)	**0.0155**	1.5680×10^{-5}
β_5	0.4009	0.3948	(0.4539, 0.3357)	0.1182	8.3012×10^{-4}
β_6	0.3659	0.3392	(0.3917, 0.2866)	0.1051	6.7536×10^{-4}
β_7	0.0000	**0.0000**	(0.0140, -0.0140)	**0.0279**	5.0822×10^{-5}
β_8	0.4389	0.3958	(0.4605, 0.3312)	0.1292	9.5072×10^{-4}
β_9	0.4827	0.4411	(0.4971, 0.3851)	0.1120	7.7004×10^{-4}
β_{10}	0.0000	**0.0000**	(0.0060, -0.0060)	**0.0121**	9.4592×10^{-6}

In this example we compare the performance with `lme4` package and the `mice` package. The corresponding results based on the combined Rubin's analysis are given in Table 8 and Table 9, respectively. Visualization of the comparison of the results produced by different methods are presented in Figure 3, and the advantages of the proposed method in terms of the accuracy for estimating $\boldsymbol{\beta}$, sparsity recovery of $\boldsymbol{\beta}$ (i.e., variable selection accuracy), and smaller uncertainty measured by the width of the confidence intervals and the total variances, are demonstrated clearly in both the figure and the Tables. We can also see that the `mice` package is far from satisfactory in this example, which is similar to the case of Section 4.1, and the behavior of the `lme4` package is very similar to the `pan` package in the case of Section 4.1.

4.3 A combined mixed-effects model

The third simulated example pertains to joint aspects of binary and continuous variables. We simulate data under a marginal distribution for $\mathbf{Y}_1$ and conditional distribution for $\mathbf{Y}_2$

Table 8: Logistic mixed-effects model with missing at random (MAR) probability 0.4, $m = 50$, $n = 100$, and $p = 10$. Estimation method is the `lme4` package with missing values deleted. Number of multiple imputation is $M = 5$.

β	True values	Estimate	Confidence intervals (CI)	CI width	Total variance
β_1	0.4108	0.3980	(0.4511, 0.3449)	0.1062	7.3393×10^{-4}
β_2	0.3331	0.2952	(0.3442, 0.2463)	0.0979	6.2366×10^{-4}
β_3	0.0000	0.0067	(0.0499, -0.0365)	0.0864	4.8539×10^{-4}
β_4	0.0000	0.0228	(0.0665, -0.0209)	0.0873	4.9638×10^{-4}
β_5	0.4009	0.4011	(0.4545, 0.3477)	0.1068	7.4230×10^{-4}
β_6	0.3659	0.3298	(0.3797, 0.2799)	0.0998	6.4871×10^{-4}
β_7	0.0000	0.0313	(0.0761, -0.0135)	0.0896	5.2287×10^{-4}
β_8	0.4389	0.3981	(0.4518, 0.3445)	0.1073	7.4994×10^{-4}
β_9	0.4827	0.4501	(0.5058, 0.3944)	0.1114	8.0720×10^{-4}
β_{10}	0.0000	0.0098	(0.0537, -0.0340)	0.0877	5.0085×10^{-4}

Table 9: Logistic mixed-effects model with missing at random (MAR) probability 0.4, $m = 50$, $n = 100$, and $p = 10$. Imputation and estimation method is the `mice` package. Number of multiple imputation is $M = 5$.

β	True values	Estimate	Confidence intervals (CI)	CI width	Total variance
β_1	0.4108	0.2179	(0.2551, 0.1806)	0.0745	3.5552×10^{-4}
β_2	0.3331	0.1447	(0.1819, 0.1075)	0.0744	3.5001×10^{-4}
β_3	0.0000	0.0067	(0.0391, -0.0256)	0.0647	2.7152×10^{-4}
β_4	0.0000	0.0154	(0.0567, -0.0260)	0.0827	4.0954×10^{-4}
β_5	0.4009	0.2129	(0.2520, 0.1739)	0.0781	3.8377×10^{-4}
β_6	0.3659	0.1698	(0.2089, 0.1306)	0.0783	3.8197×10^{-4}
β_7	0.0000	0.0448	(0.0853, 0.0043)	0.0810	3.9683×10^{-4}
β_8	0.4389	0.2036	(0.2393, 0.1679)	0.0714	3.2933×10^{-4}
β_9	0.4827	0.2302	(0.2742, 0.1862)	0.0880	4.6399×10^{-4}
β_{10}	0.0000	0.0151	(0.0554, -0.0251)	0.0805	3.9543×10^{-4}

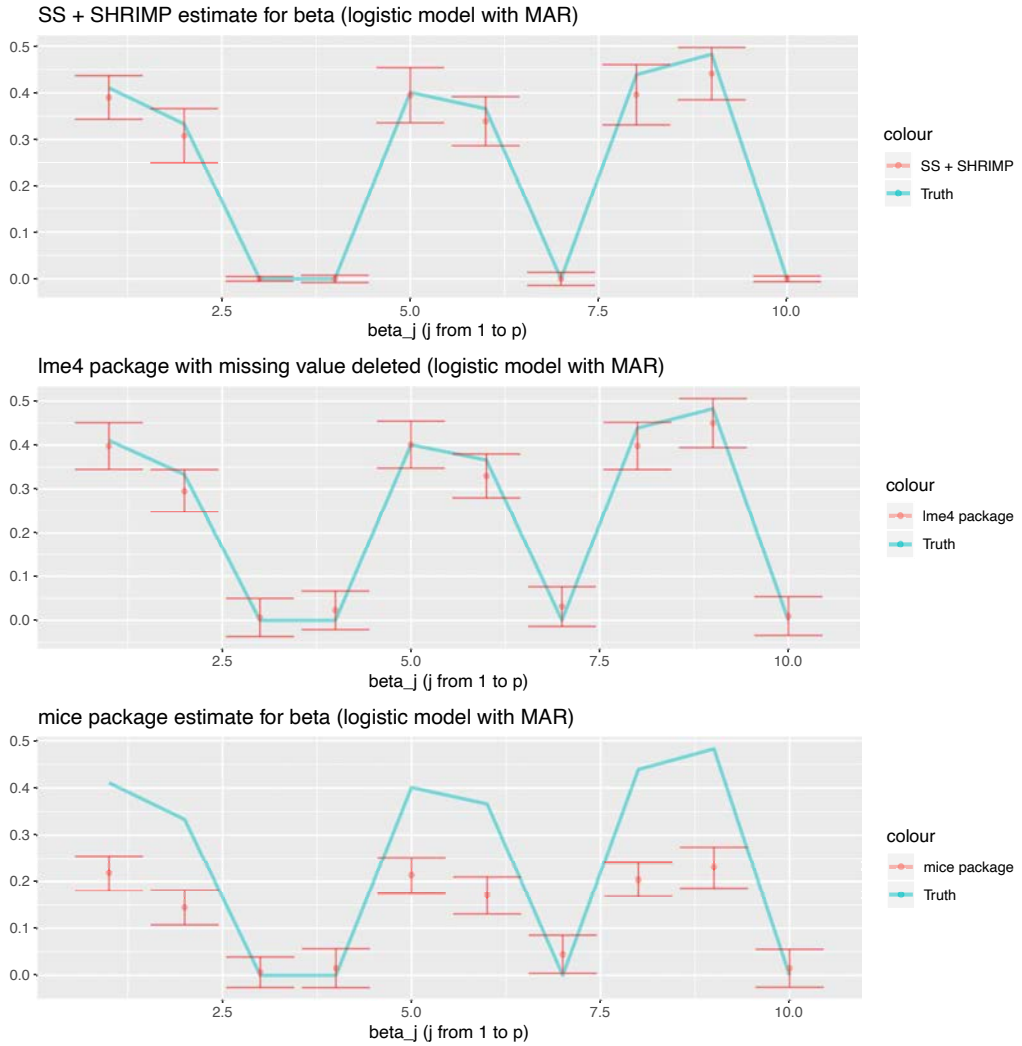

Figure 3: Simulation performance of the SS-SHRIMP method, the lme4 package with missing values deleted, and the mice package, for multipe imputation and/or estimation of the logistic mixed-effects model with MAR: The red dots are the multiple imputation estimates of β, the red bars in the top panels are estimated 95% confidence intervals for β after multiple imputation, and the blue lines represent the true values of β.

given $\mathbf{Y}_1$:

$$p(\mathbf{Y}_1, \mathbf{Y}_2) = p(\mathbf{Y}_2 \mid \mathbf{Y}_1)p(\mathbf{Y}_1),$$

where the distribution of $\mathbf{Y}_1 = [y_{1ij}]_{m\times n}$ is given by a linear mixed-effects model

$$\mathbf{y}_{1ij} = \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_{ij} + b_{1i} + \epsilon_{ij}, \quad i = 1, \ldots, m, \quad j = 1, \ldots, n,$$

where $b_{11}, \ldots, b_{1m} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2_{b_1})$, $\epsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2_e)$, with $\mathbf{x}_{ij}$'s generated following the same distribution as those in Section 4.1, and $\boldsymbol{\beta}_1$ is set to be

$$\boldsymbol{\beta}_1 = [0.3726, 0.3667, 0, 0, 0.4326, 0.3934, 0, 0.4370, 0.4401, 0]^{\mathrm{T}}.$$

The conditional distribution of $\mathbf{Y}_2 = [y_{2ij}]_{m\times n}$, $p(\mathbf{Y}_2 \mid \mathbf{Y}_1)$, can be described as follows: Given $\mathbf{Y}_1$,

$$y_{2ij} \sim \text{Bernoulli}(p_{ij}), \quad i = 1, \ldots, m, \quad j = 1, \ldots, n,$$

where $\text{logit}(p_{ij}) = \log(p_{ij}/(1-p_{ij})) = -1+0.5y_{1ij}+b_{2i}$, and $b_{21}, \ldots, b_{2m} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2_{b_2})$. Here we set $m = 25$ and $j = 100$. The corresponding working model for inference, with the true values of $\sigma^2_{b_1}, \sigma^2_{b_2}, \boldsymbol{\beta}_1$, and $\boldsymbol{\beta}_2 = [-1, 0.5]^{\mathrm{T}}$ as unknown parameters, is set as follows:

$$\begin{aligned}
&y_{1ij} = \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_{ij} + b_{1i} + \epsilon_{ij},\\
&x_{ij} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, 3),\\
&\text{logit}(\mathbb{P}(y_{2ij} = 1)) = \beta_{21} + \beta_{22}y_{1ij} + b_{2i},
\end{aligned}$$

where $i = 1, \ldots, m$ and $j = 1, \ldots, n$. To completely build a hierarchical Bayesian model such that posterior computation for parameter estimation, variable selection, and MI can be performed, we specify the following prior distributions for the unknown parameters: with $p_1 = 10$ and $p_2 = 2$, we assign

$$\begin{aligned}
\beta_{11}, \ldots, \beta_{1p_1} \mid w, \mu_{01}, \sigma^2_{01} &\sim (1-w)\delta_0(\mathrm{d}\beta_1) + w\mathrm{N}(\mu_{01}, \sigma^2_{01})\mathrm{d}\beta_1,\\
\beta_{21}, \ldots, \beta_{2p_2} \mid \mu_{02}, \sigma^2_{02} &\sim \mathrm{N}(\mu_{02}, \sigma^2_{02}),\\
w &\sim \text{Beta}(a_w, b_w),\\
\sigma^2_e &\sim \chi^{-2}_{\nu_e},\\
\sigma_{b_1}, \sigma_{b_2} &\sim \chi^{-2}_{\nu_b},\\
\mu_{01}, \mu_{02} &\sim \mathrm{N}(0, 1),\\
\sigma^2_{01}, \sigma^2_{02} &\sim \text{IG}(1, 1)
\end{aligned}$$

The response matrices $\mathbf{Y}_1$, $\mathbf{Y}_2$, as usual, are also contaminated by missing values. The missingness mechanism adopted here is very similar to that in Section 3.1 of Yucel et al. (2018), and we consider the following missing at random (MAR) scenario: Denote r_{1ij} and r_{2ij} the missingness indicators for y_{1ij} and y_{2ij}. Then we simulate the missingness indicators using the following hierarchical model:

$$\begin{aligned}
&b_i^{z_1}, b_i^{z_2} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, 1),\\
&z_{1ij} \mid b_i^{z_1} \sim \mathrm{N}(b_i^{z_1}, 2),\\
&z_{2ij} \mid b_i^{z_2} \sim \mathrm{N}(b_i^{z_2}, 2)\\
&\text{logit}(\mathbb{P}(z_{1ij} = 1)) = \gamma_{10} + \gamma_{11}z_{1ij},\\
&\text{logit}(\mathbb{P}(z_{2ij} = 1)) = \gamma_{20} + \gamma_{21}z_{2ij},
\end{aligned}$$

where the coefficients $\gamma_{10}, \gamma_{11}, \gamma_{20}, \gamma_{21}$ are selected such that the overall missingness percentage is approximately 18%.

The posterior computation that produces inference results for $\boldsymbol{\beta}$ (including sparsity recovery, i.e., variable selection via detection of zeros in $\boldsymbol{\beta}$) is performed by an implementation of the Gibbs sampler provided in Appendix C. The results are tabulated in Table 10 for $\boldsymbol{\beta}_1$ and Table 11, respectively.

Table 10: Inference for $\boldsymbol{\beta}_1$ under the compound conditional mixed-effects model with missing at random (MAR), $m = 25$, $n = 100$, and $p_1 = 10$. Imputation and estimation method is the SS-SHRIMP.

$\boldsymbol{\beta}_1$	True values	Estimate	Credible intervals (CI)	CI width
β_{11}	0.3726	0.3710	(0.3748, 0.3670)	0.0077
β_{12}	0.3667	0.3653	(0.3687, 0.3623)	0.0063
β_{13}	0.0000	**0.0000**	(0.0000, 0.0000)	**0.0000**
β_{14}	0.0000	**0.0000**	(0.0000, 0.0000)	**0.0000**
β_{15}	0.4326	0.4317	(0.4358, 0.4281)	0.0077
β_{16}	0.3934	0.3927	(0.3960, 0.3889)	0.0072
β_{17}	0.0000	**0.0000**	(0.0000, 0.0000)	**0.0000**
β_{18}	0.4370	0.4366	(0.4404, 0.4326)	0.0078
β_{19}	0.4401	0.4402	(0.4435, 0.4358)	0.0077
$\beta_{1,10}$	0.0000	**0.0000**	(0.0000, 0.0000)	**0.0000**

Table 11: Inference for $\boldsymbol{\beta}_2$ under the compound conditional mixed-effects model with missing at random (MAR), $m = 25$, $n = 100$, and $p_2 = 2$. Imputation and estimation method is SS-SHRIMP.

$\boldsymbol{\beta}_2$	True values	Estimate	Confidence intervals	CI width
β_{21}	-1	**-0.9959**	(-0.8824, -1.1314)	0.2490
β_{22}	0.5	**0.4873**	(0.5330, 0.4472)	0.0859

Here we considered the `pan` package (as an approximate for binary variable) and `mice` package. We summarize the results in Table 12 and Table 13, respectively. We also visualize the comparison of the results provided by our method against the `pan` package in Figure 4 for $\boldsymbol{\beta}_1$ and Figure 5. We can see that both methods can successfully and accurately estimate the non-zero signal of $\boldsymbol{\beta}_1$, but the spike-and-slab approach provides better results in terms of recovering the zero coordinates of $\boldsymbol{\beta}$ compared to the `pan` package, as the latter produces relatively wider confidence intervals for $\beta_{13}, \beta_{14}, \beta_{17}, \beta_{1,10}$. The `mice` package produces inaccurate estimates for $\boldsymbol{\beta}_2$, which can be easily recognized from Table 13 and Figure 5. Therefore, we conclude that the empirical performance of the proposed spike-and-slab variable selection approach embedded in SHRIMP outperforms the competitors in this simulated example.

5. Discussion

We have illustrated that the variable selection problem in the presence of missing response variables in mixed-effects regression models can be done by a hierarchical Bayesian approach with a spike-and-slab prior distribution for the linear coefficients. We successfully derive efficient Gibbs sampler for posterior computation of the corresponding linear and logistic mixed-effects models. The hierarchical Bayesian model itself also permits the inte-

Table 12: Inference for $\boldsymbol{\beta}_1$ under the compound conditional mixed-effects model with missing at random (MAR), $m = 25$, $n = 100$, and $p_1 = 10$. Imputation and estimation method is the `pan` package.

$\boldsymbol{\beta}_1$	True values	Estimate	Confidence intervals (CI)	CI width
β_{11}	0.3726	0.3730	(0.3759, 0.3702)	0.0057
β_{12}	0.3667	0.3653	(0.3682, 0.3624)	0.0058
β_{13}	0.0000	0.0009	(0.0038, -0.0020)	0.0058
β_{14}	0.0000	0.0020	(0.0048, -0.0008)	0.0056
β_{15}	0.4326	0.4323	(0.4353, 0.4293)	0.0060
β_{16}	0.3934	0.3922	(0.3949, 0.3894)	0.0055
β_{17}	0.0000	0.0017	(0.0047, -0.0012)	0.0059
β_{18}	0.4370	0.4374	(0.4404, 0.4345)	0.0059
β_{19}	0.4401	0.4429	(0.4458, 0.4400)	0.0058
$\beta_{1,10}$	0.0000	0.0012	(0.0041, -0.0018)	0.0059

Table 13: Inference for $\boldsymbol{\beta}_2$ under the compound conditional mixed-effects model with missing at random (MAR), $m = 25$, $n = 100$, and $p_2 = 2$. Imputation and estimation method is the mice package.

$\boldsymbol{\beta}_2$	True values	Estimate	Confidence intervals (CI)	CI width
β_{21}	-1	-0.7433	(-0.6218, -0.8647)	0.2430
β_{22}	0.5	0.3307	(0.3726, 0.2888)	0.0838

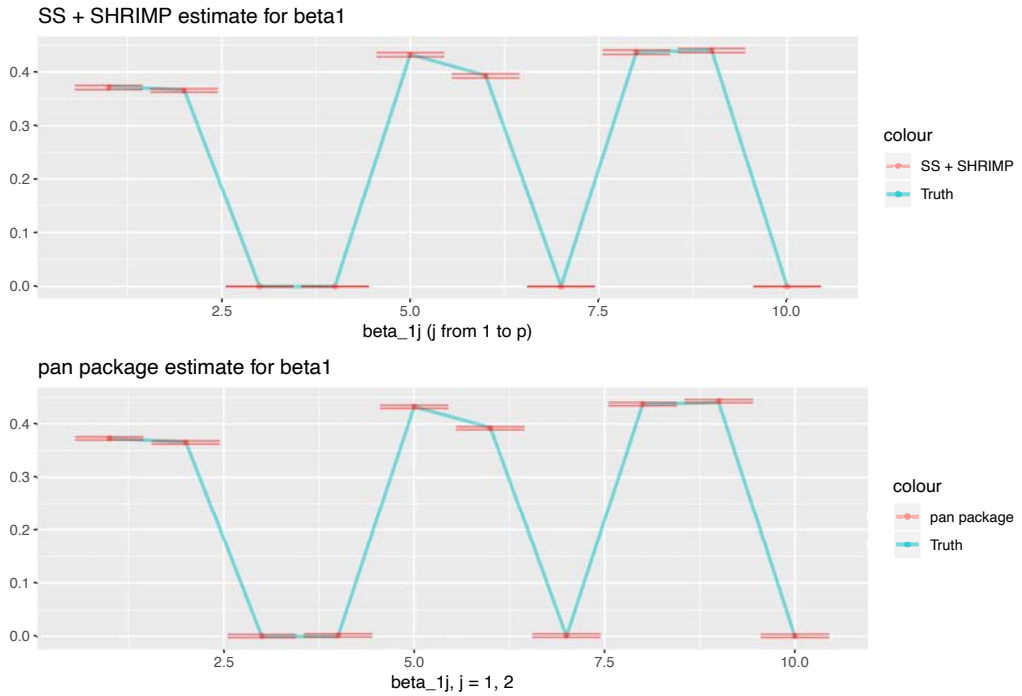

Figure 4: Simulation performance of the SS-SHRIMP method, the pan package, and the mice package, for imputation-based estimation of $\boldsymbol{\beta}_1$ in the compound conditional mixed-effects model with MAR: The red dots are the multiple imputation estimates of $\boldsymbol{\beta}_1$, the red bars in the top panels are estimated 95% confidence intervals for $\boldsymbol{\beta}_1$ after multiple imputation, and the blue lines represent the true values of $\boldsymbol{\beta}_1$.

Figure 5: Simulation performance of the SS-SHRIMP method, the pan package, and the mice package, for imputation-based estimation of β_2 in the compound conditional mixed-effects model with MAR: The red dots are the multiple imputation estimates of β_2, the red bars in the top panels are estimated 95% confidence intervals for β_2 after multiple imputation, and the blue lines represent the true values of β_2.

gration with the sequential hierarchical regression imputation strategy introduced by Yucel et al. (2018) for multiple imputation of the missing responses, further facilitate the computational efficiency of the corresponding MCMC algorithm.

There are some potential future extensions of the current methodology. The numerical examples provided in this work are relatively low-dimensional regression problems. Although the spike-and-slab prior distributions (2) permits the derivation of closed-form Gibbs sampler either by a direct approach or via a PX-DA strategy (e.g., the auxiliary Pólya-Gamma random variable), the corresponding computation expense for the MCMC is still problematic with ultra-high-dimensional data. Even with the help of Monte Carlo sampling methods and the spike-and-slab prior (2), it is still required to explore the entire space of all possible models as much as possible. Nonetheless, the complexity of the space of all possible models grows exponentially with the number of predictors, and in moderately high-dimensional setup, the MCMC could be cumbersome or even infeasible to implement. We have already observed the potential computational difficulty of the MCMC-based MI method involving variable selection in the simulated examples. The comparison among the computation expenses using different methods are tabulated in Table 14, and we note that the computation expense for the spike-and-slab variable selection composite with SHRIMP for MI is much more expensive than the other competitors, but we gain estimation and variable selection accuracy instead. It has also been pointed out in Castillo et al. (2015) that

Table 14: Computation expenses using different MI methods in Section 4

Method	Section 4.1	Section 4.2	Section 4.3
Spike-and-slab SHIRMP	986s	580s	604s
`pan` package	3.43s	N/A	1.73s
`mice` package	124s	19.2s	134s
`lme4` package	N/A	2.00s	N/A

algorithms that can successfully addresses ultra-high-dimensional variable selection problems is beyond the scope of fully Bayesian methods.

In contrast to relying on MCMC-based posterior computation algorithms, which is a class of exact Bayesian inference methods in the sense that the random samples drawn from the Markov chain can be regarded as samples generated from the exact full posterior distribution, a relatively more efficient method is the variational inference (VI). Unlike the

MCMC approach, the VI is an approximate Bayesian inference algorithm that can be much faster but at the cost of certain model bias. Under certain regularity conditions, it has also been proved that the variational posterior distribution is comparable to the exact posterior distribution (Zhang et al., 2020, Pati et al., 2018, Wang and Blei, 2019, Han and Yang, 2019). The use of VI for linear regression models has been restricted on the case of low-dimensional case (You et al., 2014). In the future, we plan to explore the methodology and theory for VI for linear and generalized linear mixed-effects models in the presence of the missing responses for the sake of computational efficiency for high-dimensional data.

Appendix

A. Gibbs sampler for Section 2

In this section, we derive the detailed Gibbs sampling algorithm, which reduces to the following full conditional distributions of the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_b, \sigma_e)$ and the random effects $\mathbf{b} = [b_1, \ldots, b_m]^{\mathrm{T}}$. A single cycle of the Gibbs sampler iterates the following sampling schemes:

$$(\beta_k \mid \mathbf{X}, \boldsymbol{\beta}_{-k}, \mathbf{b}, \sigma_b, \sigma_e, w, \mu_0, \sigma_0) \sim \begin{cases} w_1^* \delta_0 + w_2^* \mathrm{N}(\widehat{\mu}, \widehat{V}), & \text{if } x_{ijk} \text{is not required,} \\ \mathrm{N}(\widehat{\mu}, \widehat{V}), & \text{if } x_{ijk} \text{ is required,} \end{cases} \tag{7}$$

$$(w \mid \mathbf{X}, \boldsymbol{\beta}) \sim \mathrm{Beta}\left(a_w + \sum_{k=1}^{p} z_k, b_w + \sum_{k=1}^{p}(1 - z_k)\right), \tag{8}$$

$$(\mu_0 \mid \boldsymbol{\beta}, \sigma_0) \sim \mathrm{N}\left(\left(1 + \frac{1}{\sigma_0^2}\sum_{k=1}^{p} z_k\right)^{-1} \frac{1}{\sigma_0^2}\sum_{k=1}^{p}\beta_k, \left(1 + \frac{1}{\sigma_0^2}\sum_{k=1}^{p} z_k\right)^{-1}\right), \tag{9}$$

$$(\sigma_0^2 \mid \boldsymbol{\beta}, \mu_0) \sim \mathrm{Inverse-Gamma}\left(1 + \frac{1}{2}\sum_{k=1}^{p} z_k, 1 + \frac{1}{2}\sum_{k=1}^{p} z_k(\beta_k - \mu_0)^2\right), \tag{10}$$

$$(b_i \mid \mathbf{X}, \boldsymbol{\beta}, \sigma_b, \sigma_e) \sim \mathrm{N}(\widehat{b}_i, V(\widehat{b}_i)), \tag{11}$$

$$(\sigma_e^2 \mid \mathbf{X}, b_1, \ldots, b_m, \boldsymbol{\beta}, \sigma_b) \sim \left(1 + \sum_{i=1}^{m}\sum_{j=1}^{n}\widehat{\epsilon}_{ij}^2\right)\chi_{\nu_e + mn - 1}^{-2}, \tag{12}$$

$$(\sigma_b^2 \mid \mathbf{X}, b_1, \ldots, b_m, \boldsymbol{\beta}, \sigma_e) \sim \left(\frac{\nu_b + \sum_{i=1}^{m} b_i^2}{\nu_b + m}\right)\chi_{\nu_b + m}^{-2}, \tag{13}$$

where $\mathbf{X}$ denotes the full set of covariates $\mathbf{X} = [\mathbf{x}_{ij}]_{i=1,\ldots,m, j=1,\ldots,n}$, $z_k = \mathbb{1}(\beta_k \neq 0)$, $w_1^*, w_2^*, \widehat{V}, \widehat{\mu}$ are the same as those given in Section 2,

$$\widehat{\epsilon}_{ij} = y_{ij} - \mathbf{x}_{ij}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}, \quad \widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{x}_{ij}\mathbf{x}_{ij}^{\mathrm{T}}\right)^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{x}_{ij}(y_{ij} - b_i),$$

$$V(\widehat{b}_i) = \left(\frac{n}{\sigma_e^2} + \frac{1}{\sigma_b^2}\right)^{-1}, \quad \widehat{b}_i = \frac{V(\widehat{b}_i)}{\sigma_e^2}\sum_{j=1}^{n}(y_{ij} - \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}).$$

Note that formulas (11), (12), and (13) are the same as those appearing in Section 2.2.1 in Yucel et al. (2018). The last step in one iteration of the Gibbs sampler is to draw the predictive posterior distribution of the missing response $y_{ij} \in (y_{\mathrm{mis}})$ using the SHRIMP strategy described at the end of Section 2.

B. Gibbs sampler for Section 3

We provide the complete full conditional distributions that are required for the Gibbs sampler to draw posterior samples from the joint distribution of $(\boldsymbol{\beta}, b_1, \ldots, b_m)$, together with the samples of the missing data (y_{mis}). Following the derivation in Section 3, we obtain the

following closed-form full conditional distribution of $\boldsymbol{\beta}$, w, μ_0, and σ_0^2:

$$(\beta_k \mid \mathbf{X}, \mathbf{Y}, \boldsymbol{\Omega}, \boldsymbol{\beta}_{-k}, \mathbf{b}, \sigma_b, w, \mu_0, \sigma_0) \sim \begin{cases} w_1^* \delta_0 + w_2^* \mathrm{N}(\widehat{\mu}, \widehat{V}), & \text{if } x_{ijk} \text{ is not required,} \\ \mathrm{N}(\widehat{\mu}, \widehat{V}), & \text{if } x_{ijk} \text{ is required,} \end{cases} \tag{14}$$

$$(w \mid \mathbf{X}, \boldsymbol{\beta}) \sim \mathrm{Beta}\left(a_w + \sum_{k=1}^{p} z_k, b_w + \sum_{k=1}^{p}(1 - z_k)\right), \tag{15}$$

$$(\mu_0 \mid \boldsymbol{\beta}, \sigma_0) \sim \mathrm{N}\left(\left(1 + \frac{1}{\sigma_0^2}\sum_{k=1}^{p} z_k\right)^{-1} \frac{1}{\sigma_0^2}\sum_{k=1}^{p}\beta_k, \left(1 + \frac{1}{\sigma_0^2}\sum_{k=1}^{p} z_k\right)^{-1}\right), \tag{16}$$

$$(\sigma_0^2 \mid \boldsymbol{\beta}, \mu_0) \sim \mathrm{Inverse-Gamma}\left(1 + \frac{1}{2}\sum_{k=1}^{p} z_k, 1 + \frac{1}{2}\sum_{k=1}^{p} z_k(\beta_k - \mu_0)^2\right) \tag{17}$$

where $\mathbf{X}$ denotes the full set of covariates $\mathbf{X} = [\mathbf{x}_{ij}]_{i=1,\ldots,m, j=1,\ldots,n}$, $z_k = \mathbb{1}(\beta_k \neq 0)$, and the formulas for computing $w_1^*, w_2^*, \widehat{V}, \widehat{\mu}$ are provided in Section 3. The full conditional distribution of the auxiliary variables $\boldsymbol{\Omega} = [\omega_{ij}]_{m \times n}$ is given by (6) in Section 3. Similar to the derivation of (5), the full conditional distribution of the random effects $b_1, \ldots, b_m$ can be derived analogously:

$$\begin{aligned} p(b_i \mid \mathbf{X}, \mathbf{Y}, \boldsymbol{\beta}, \sigma_b) &\propto p(b_i) \prod_{j=1}^{n} \mathcal{L}(\eta_{ij} \mid y_{ij}) \\ &\propto p(b_i) \prod_{j=1}^{n} \exp\left\{\left(y_{ij} - \frac{1}{2}\right)(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + b_i) - \frac{\omega_{ij}}{2}(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + b_i)^2\right\} \\ &\propto p(b_i) \prod_{j=1}^{n} \exp\left\{-\frac{\omega_{ij}}{2}\left[b_i^2 - 2\left(\frac{y_{ij} - 1/2}{\omega_{ij}} - \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}\right) b_i\right]\right\} \\ &\propto p(b_i) \prod_{j=1}^{n} \exp\left[-\frac{\omega_{ij}}{2}(b_i - u_{ij})^2\right], \end{aligned}$$

where $u_{ij} = (y_{ij} - 1/2)/\omega_{ij} - \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}$. Since $p(b_i) = (1/\sqrt{2\pi\sigma_b^2})\exp[-b_i^2/(2\sigma_b^2)]$, it follows directly from the normal conjugacy that

$$(b_i \mid \mathbf{X}, \mathbf{Y}, \boldsymbol{\beta}, \sigma_b) \sim \mathrm{N}\left(\left(\frac{1}{\sigma_b^2} + \sum_{j=1}^{n}\omega_{ij}\right)^{-1} \sum_{j=1}^{n}\omega_{ij}u_{ij}, \left(\frac{1}{\sigma_b^2} + \sum_{j=1}^{n}\omega_{ij}\right)^{-1}\right). \tag{18}$$

The full conditional distribution of σ_b is the same as (13):

$$(\sigma_b^2 \mid \mathbf{X}, b_1, \ldots, b_m) \sim \left(\frac{\nu_b + \sum_{i=1}^{m} b_i^2}{\nu_b + m}\right) \chi_{\nu_b + m}^{-2}.$$

The last step in a single iteration of the Gibbs sampler is to draw the predictive posterior distribution of the missing response $y_{ij} \in (y_{\mathrm{mis}})$ following the SHRIMP strategy mentioned at the end of Section 2.

C. Gibbs sampler for Section 4.3

In this appendix we derive the full conditional distributions of the unknown parameters as well as the posterior predictive distribution for the missing response variables that are

needed for the Gibbs sampler for Section 4.3. The full conditional distributions of the parameters for the linear mixed-effects component are listed as follows:

$$(\beta_{1k} \mid -) \sim w_{1k}^* \delta_0 + w_{2k}^* \mathrm{N}(\widehat{\mu}_1, \widehat{V}_1), \quad k = 1, 2, \ldots, p_1 \tag{19}$$

$$(w \mid -) \sim \mathrm{Beta}\left(a_w + \sum_{k=1}^{p_1} \mathbb{1}(\beta_{1k} \neq 0), b_w + \sum_{k=1}^{p_1} (\beta_{1k} = 0)\right), \tag{20}$$

$$(\mu_{01} \mid -) \sim \mathrm{N}\left(\left(1 + \frac{1}{\sigma_0^2} \sum_{k=1}^{p_1} \mathbb{1}(\beta_{1k} \neq 0)\right)^{-1} \sum_{k=1}^{p_1} \frac{\beta_{1k}}{\sigma_{01}^2}, \left(1 + \frac{1}{\sigma_{01}^2} \sum_{k=1}^{p_1} \mathbb{1}(\beta_{1k} \neq 0)\right)^{-1}\right), \tag{21}$$

$$(\sigma_{01}^2 \mid -) \sim \mathrm{IG}\left(1 + \frac{1}{2} \sum_{k=1}^{p_1} \mathbb{1}(\beta_{1k} \neq 0), 1 + \frac{1}{2} \sum_{k=1}^{p_1} \mathbb{1}(\beta_{1k} \neq 0)(\beta_{1k} - \mu_{01})^2\right), \tag{22}$$

$$(b_{1i} \mid -) \sim \mathrm{N}(\widehat{b}_{1i}, V(\widehat{b}_{1i})), \tag{23}$$

$$(\sigma_e^2 \mid \mathbf{X}, b_1, \ldots, b_m, \boldsymbol{\beta}, \sigma_b) \sim \left(1 + \sum_{i=1}^{m} \sum_{j=1}^{n} \widehat{\epsilon}_{ij}^2\right) \chi_{\nu_e + mn - 1}^{-2}, \tag{24}$$

$$(\sigma_{b_1}^2 \mid \mathbf{X}, b_1, \ldots, b_m, \boldsymbol{\beta}, \sigma_e) \sim \left(\frac{\nu_b + \sum_{i=1}^{m} b_{1i}^2}{\nu_b + m}\right) \chi_{\nu_b + m}^{-2}, \tag{25}$$

where

$$w_{1k}^* \propto (1 - w) \mathcal{N}\left(0 \,\middle|\, \frac{\sum_{i,j} x_{ijk}^{(1)} (y_{1ij} - \sum_{\ell \neq k} x_{ij\ell}^{(1)} \beta_{1\ell} - b_{1i})}{\sum_{i,j} (x_{ijk}^{(1)})^2}, \frac{\sigma_e^2}{\sum_{i,j} (x_{ijk}^{(1)})^2}\right),$$

$$w_{2k}^* \propto w \mathcal{N}\left(\mu_{01} \,\middle|\, \frac{\sum_{i,j} x_{ijk}^{(1)} (y_{1ij} - \sum_{\ell \neq k} x_{ij\ell}^{(1)} \beta_{1\ell} - b_{1i})}{\sum_{i,j} (x_{ijk}^{(1)})^2}, \sigma_{01}^2 + \frac{\sigma_e^2}{\sum_{i,j} (x_{ijk}^{(1)})^2}\right),$$

$$\widehat{V}_1 = \left(\frac{1}{\sigma_e^2} \sum_{i=1}^{m} \sum_{j=1}^{n} (x_{ijk}^{(1)})^2 + \frac{1}{\sigma_{01}^2}\right)^{-1},$$

$$\widehat{\mu}_1 = \widehat{V}_1 \left[\frac{\mu_{01}}{\sigma_{01}^2} + \frac{1}{\sigma_e^2} \sum_{i=1}^{m} \sum_{j=1}^{n} x_{ijk}^{(1)} \left(y_{1ij} - \sum_{\ell \neq k} x_{ij\ell}^{(1)} \beta_{1\ell} - b_{1i}\right)\right],$$

$$\widehat{\epsilon}_{ij} = y_{1ij} - (\mathbf{x}_{ij}^{(1)})^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_1, \qquad \widehat{\boldsymbol{\beta}}_1 = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} (\mathbf{x}_{ij}^{(1)})(\mathbf{x}_{ij}^{(1)})^{\mathrm{T}}\right)^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{x}_{ij}^{(1)} (y_{1ij} - b_{1i}),$$

$$V(\widehat{b}_{1i}) = \left(\frac{n}{\sigma_e^2} + \frac{1}{\sigma_{b_1}^2}\right)^{-1}, \quad \widehat{b}_{1i} = \frac{V(\widehat{b}_{1i})}{\sigma_e^2} \sum_{j=1}^{n} (y_{1ij} - (\mathbf{x}_{ij}^{(1)})^{\mathrm{T}} \boldsymbol{\beta}_1),$$

$\mathbf{x}_{ij}^{(1)} = \mathbf{x}_{ij}^{\mathrm{T}}$. Denote $z_{ij} = (y_{2ij} - 1/2)/\omega_{ij} - b_{2i}$,

$$\begin{aligned}
\mathbf{z} &= [z_{11}, \ldots, z_{1n}, z_{21}, \ldots, z_{2n}, \ldots, z_{m1}, \ldots, z_{mn}]^{\mathrm{T}} \in \mathbb{R}^{mn}, \\
\mathbf{X} &= [\mathbf{x}_{11}^{(2)}, \ldots, \mathbf{x}_{1n}^{(2)}, \mathbf{x}_{21}^{(2)}, \ldots, \mathbf{x}_{2n}^{(2)}, \ldots, \mathbf{x}_{m1}^{(2)}, \ldots, \mathbf{x}_{mn}^{(2)}]^{\mathrm{T}} \in \mathbb{R}^{mn \times p_2}, \\
\boldsymbol{\Sigma}^{-1} &= \mathrm{diag}(\omega_{11}, \ldots, \omega_{1n}, \omega_{21}, \ldots, \omega_{2n}, \ldots, \omega_{m1}, \ldots, \omega_{mn}) \in \mathbb{R}^{mn \times mn},
\end{aligned}$$

where $\mathbf{x}_{ij}^{(2)} = [1, y_{1ij}]^{\mathrm{T}}$. We then obtain the following closed-form full conditional distribution of $\boldsymbol{\beta}_2$, μ_{02}, and σ_{02}^2, which are the parameters for the logistic mixed-effects component:

$$(\beta_{2k} \mid -) \sim \mathrm{N}(\widehat{\mu}_2, \widehat{V}_2), \quad k = 1, \ldots, p_2 \tag{26}$$

$$(\mu_{02} \mid -) \sim \mathrm{N}\left(\left(1 + \frac{p_2}{\sigma_{02}^2}\right)^{-1} \frac{1}{\sigma_{02}^2} \sum_{k=1}^{p_2} \beta_{2k}, \left(1 + \frac{p_2}{\sigma_{02}^2}\right)^{-1}\right), \tag{27}$$

$$(\sigma_{02}^2 \mid -) \sim \mathrm{IG}\left(1 + \frac{p_2}{2}, 1 + \frac{1}{2}\sum_{k=1}^{p_2}(\beta_{2k} - \mu_{02})^2\right) \tag{28}$$

where

$$\widehat{V}_2 = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} \omega_{ij}(x_{ijk}^{(2)})^2 + \frac{1}{\sigma_{02}^2}\right)^{-1},$$

$$\widehat{\mu}_2 = \widehat{V}_2 \left[\frac{\mu_{02}}{\sigma_{02}^2} + \sum_{i=1}^{m}\sum_{j=1}^{n} \omega_{ij}(x_{ijk}^{(2)}) \left(z_{ij} - \sum_{\ell \neq k} x_{ij\ell}^{(2)} \beta_{2\ell}\right)\right].$$

The full conditional distribution of the auxiliary variables $\boldsymbol{\Omega} = [\omega_{ij}]_{m \times n}$ can be derived similarly as that in Polson et al. (2013):

$$(\omega_{ij} \mid -) \sim \mathrm{PG}(1, (\mathbf{x}_{ij}^{(2)})^{\mathrm{T}}\boldsymbol{\beta}_2 + b_{2i}), \tag{29}$$

Denote $u_{ij} = (y_{2ij} - 1/2)/\omega_{ij} - (\mathbf{x}_{ij}^{(2)})^{\mathrm{T}}\boldsymbol{\beta}_2$. Then the random effects $b_{21}, \ldots, b_{2m}$ can be sampled from

$$(b_{2i} \mid -) \sim \mathrm{N}\left(\left(\frac{1}{\sigma_{b_2}^2} + \sum_{j=1}^{n} \omega_{ij}\right)^{-1} \sum_{j=1}^{n} \omega_{ij} u_{ij}, \left(\frac{1}{\sigma_{b_2}^2} + \sum_{j=1}^{n} \omega_{ij}\right)^{-1}\right). \tag{30}$$

The full conditional distribution of σ_{b_2} is the same as (13):

$$(\sigma_b^2 \mid -) \sim \left(\frac{\nu_b + \sum_{i=1}^{m} b_{2i}^2}{\nu_b + m}\right) \chi_{\nu_b + m}^{-2}.$$

Finally, sampling missing responses $y_{1ij} \in (y_{1\mathrm{mis}})$ and $y_{2ij} \in (y_{2\mathrm{mis}})$ from

$$(y_{1ij} \mid -) \sim \mathrm{N}((\mathbf{x}_{ij}^{(1)})^{\mathrm{T}}\boldsymbol{\beta}_1 + b_{1i}, \sigma_e^2),$$

$$(y_{2ij} \mid -) \sim \mathrm{Bernoulli}\left(\frac{1}{1 + \exp(-(\mathbf{x}_{ij}^{(2)})^{\mathrm{T}}\boldsymbol{\beta}_2 - b_{2i})}\right)$$

References

Bayarri, M. J., Berger, J. O., Paulo, R., Sacks, J., Cafeo, J. A., Cavendish, J., Lin, C.-H., and Tu, J. (2007). A framework for validation of computer models. *Technometrics*, 49(2):138–154.

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). Dirichlet–laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512):1479–1490. PMID: 27019543.

Bickel, P. J., Li, B., Tsybakov, A. B., van de Geer, S. A., Yu, B., Valdés, T., Rivero, C., Fan, J., and van der Vaart, A. (2006). Regularization in statistics. *Test*, 15(2):271–344.

Buuren, S. v. and Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, pages 1–68.

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717.

Candès, E. J. and Tao, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Castillo, I., Schmidt-Hieber, J., Van der Vaart, A., et al. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018.

Castillo, I., van der Vaart, A., et al. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics*, 40(4):2069–2101.

Clyde, M., Desimone, H., and Parmigiani, G. (1996). Prediction via orthogonalized model mixing. *Journal of the American Statistical Association*, 91(435):1197–1208.

Garcia, R. I., Ibrahim, J. G., and Zhu, H. (2010). Variable selection for regression models with missing data. *Statistica Sinica*, 20(1):149.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Geweke, J. (1996). Variable selection and model comparison in regression. *In Bayesian Statistics 5*.

Han, W. and Yang, Y. (2019). Statistical inference in mean-field variational bayes. *arXiv preprint arXiv:1911.01525*.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)*, 60(1):65–81.

Lounici, K. et al. (2014). High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Pati, D., Bhattacharya, A., and Yang, Y. (2018). On statistical optimality of variational bayes. In *International Conference on Artificial Intelligence and Statistics*, pages 1579–1588.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using Pólya-Gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349.

Ročková, V. et al. (2018). Bayesian estimation of sparse signals with a continuous spike-and-slab prior. *The Annals of Statistics*, 46(1):401–437.

Ročková, V. and George, E. I. (2018). The spike-and-slab lasso. *Journal of the American Statistical Association*, 113(521):431–444.

Rubin, D. (1987). Multiple imputation for nonresponse in surveys. *NY John Wiley & Sons Crossref*.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Wang, Y. and Blei, D. M. (2019). Frequentist consistency of variational bayes. *Journal of the American Statistical Association*, 114(527):1147–1161.

You, C., Ormerod, J. T., and Mueller, S. (2014). On variational bayes estimation and variational information criteria for linear regression models. *Australian & New Zealand Journal of Statistics*, 56(1):73–87.

Yucel, R. M., Zhao, E., Schenker, N., and Raghunathan, T. E. (2018). Sequential hierarchical regression imputation. *Journal of Survey Statistics and Methodology*, 6(1):1–22.

Zhang, F., Gao, C., et al. (2020). Convergence rates of variational posterior distributions. *Annals of Statistics*, 48(4):2180–2207.

Zhao, J. and Schafer, J. (2013). pan: Multiple imputation for multivariate panel or clustered data. *R Foundation for statistical computing*, page 1.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.