Variational Multiple Imputation in High-Dimensional Regression Models With Missing Responses

Qiushuang Li [*] Recai Yucel [†]

Abstract
Multiple imputation has become one of the standard methods in drawing inferences in many incomplete data applications. Applications of multiple imputation in relatively more complex settings such as high-dimensional clustered data require specialized methods to overcome the computational burden. Using mixed-effects models , we develop methods that can be applied to continuous, binary, or categorical incomplete data. We overcome the computational burden by employing variational Bayesian inference for sampling the posterior predictive distribution missing data. These methods specifically target high-dimensional covariates and work with spike-and-slab priors, which force the variables of importance to be in the imputation model. The individual regression computation is then incorporated in an increasingly popular variable-by-variable imputation algorithm. Finally, we use calibration-based algorithms to adopt these methods to multiply-impute categorical variables. We present a simulation study to assess the performance of these methods in a repetitive sampling framework.

Key Words: Clustered data, missing data, variational inference, multiple imputation, sequential hierarchical regression imputation, calibration-based imputation, spike-and-slab variable selection

1. Introduction

Missing data is typically seen as norma rather than exception in a wide range of areas ranging from survey data analysis to signal processing, compressed sensing (Candès and Recht, 2009, Candes and Tao, 2010, Candes and Plan, 2010, Gross, 2011), collaborative filtering, and recommendation systems (Koren et al., 2009). For example, in the Netflix Prize competition, some movies which are not rated can be treated as missing. The participants must predict grades on the entire qualifying set with the scores for half of the data. The missing data problems also occur in the computer experiments and biomedical applications because of equipment limitations (Bayarri et al., 2007). The analysis of numerous missing data problems have been attracting an increasing attention.

To deal with missing data, statisticians have relied many imputation methods. However, a recurring problem of imputation is the impact on the statistical uncertainty. Multiple imputation (MI) aims to solve this aspect of imputation. A sensible strategy of MI is to sample missing data from their underlying distribution, and by doing so, statisticians hope to account for the uncertainty inherent to missing values in contrast to a single imputation (Rubin, 1987). The MI strategy is usually implemented within a fully Bayesian model, which additionally incorporates the uncertainty in the unknown parameters. And computational aspects are typically based on Markov Chain Monte Carlo techniques.

In this work, we are particularly interested in the problem of variable selection in linear mixed-effect models in the presence of missing responses. Classically, variable selection in general linear models was addressed by certain information-criterion-based model selection

[*] 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

approaches, such as Akaike information criterion (AIC) (Akaike, 1974) and Bayesian information criterion (BIC) (Schwarz, 1978). These model selection approaches suffers from computation inefficiency, which is exponential in the number of variables. The computation bottleneck of variable selection for linear models was successfully tackled by LASSO (Tibshirani, 1996) and its variants (Zou, 2006, Zou and Hastie, 2005) through a collection of seminal convex optimization approach.

These methods are frequentist by their nature and are not user-friendly from the perspective of missing data and MI. To address this issue, Li and Yucel (2020) proposed to apply the spike-and-slab prior for variable selection in the context of a fully Bayesian model, such that simulation-based MI can be performed through sampling the posterior predictive distribution of the missing responses. Theoretical properties of the spike-and-slab prior for standard linear regression model with high dimensionality has been explored in Castillo et al. (2015). However, as is commented in Castillo et al. (2015), high-dimensional variable selection problems are out of scope of fully Bayesian models at the present time due to the need to explore the entire space of all possible models is exponential in the number of variables. This problem becomes particularly challenging when MCMC is implemented.

The main purpose of this work is to propose a computational efficient method that can deal with high-dimensional variable selection problem in the linear mixed-effect model and MI of missing responses. This is completed by an optimization-based approximate inference method, referred to as the *variational inference* (Bishop, 2006). In contrast to MCMC methods, which are simulation-based inference algorithms, and the resulting Markov chains could be time-consuming to converge or exhibits poor mixing behavior, variational inference is a collection of approximate Bayesian inference methods that are formulated as a mathematical optimization problem. Specifically, we develop a computational-efficient variational inference algorithm for approximate inference of high-dimensional linear mixed-effect model, which can be applied for MI of missing responses. In particular, we combine the proposed variational inference method for variable selection with various MI method, including the sequential hierarchical regression imputation (SHRIMP) (Yucel et al., 2017) for continous data, and the calibration-based imputation (Yucel et al., 2008, 2011) for binary and ordinal data. The advantage of the proposed method is that it addresses the computation bottleneck of variable selection problem in Bayesian models through approximate inference and also allows a collection of MI approaches that deal with missing data.

The rest of this working paper is arranged as follows. In Section 2, we first briefly review the high-dimensional linear mixed-effect model with missing responses, and then elaborate on the Bayesian model with the spike-and-slab prior. In Section 3, we develop the proposed variational inference algorithm with the spike-and-slab prior for variable selection in the presence of missing responses. This section is the core of the entire work. Section 4 demonstrates how the variational inference algorithm serves as a building block that can be embedded for different MI methods for missing responses, including the sequential hierarchical regression imputation method for continuous data and the calibration-based routine for categorical data. The usefulness of the proposed methodology is empirically presented in Section 5 through the analyses of simulated examples. We conclude the work with a discussion in Section 6.

2. High-Dimensional Linear Mixed-effect Model 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. (2017):

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

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the fixed-effect, $\mathbf{b}_1, \ldots, \mathbf{b}_m \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \boldsymbol{\Psi})$ 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 sequential hierarchical regression imputation (SHRIMP) strategy, as is suggested in Yucel et al. (2017). 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 portion can be imputed using some other imputation method.

In this work we consider the scenario where the number of covariates p is comparable or even larger than the sample size. The fixed-effect regression coefficient vector $\boldsymbol{\beta}$ is assumed to be sparse, namely, the number of non-zero coordinates of $\boldsymbol{\beta}$ is comparably smaller than the sample size. Consequently, the number of active covariates, namely, those coordinates of $\mathbf{x}_{ij}$'s corresponding the non-zero coordinates of $\boldsymbol{\beta}$, is also significantly smaller than the sample size presumably. The inference focus here is to "recover" the missing portion of the missing responses y_{ij}'s (i.e., multiple imputation with the appropriate uncertainty) but also to account for the variable selection structure due to the sparsity of $\boldsymbol{\beta}$ and recover the variable selection structure. Leveraging a fully Bayesian model, in Li and Yucel (2020), the authors developed a Gibbs sampler to draw posterior samples from the joint distribution of $(\boldsymbol{\beta}, \mathbf{b}_1, \ldots, \mathbf{b}_m, \sigma_{\mathbf{b}_i}, \sigma_e)$, as well as to draw samples of the missing data $y_{,\text{is}}$ from the corresponding posterior predictive distribution. To select the variables among $x_{ij1}, \ldots, x_{ijp}$, a spike-and-slab prior distribution is assigned to the regression coefficient $\boldsymbol{\beta}$, and this spike-and-slab variable selection approach has been broadly applied to Bayesian variable selection. When we have to deal with the high dimensional model where the dimension p of the regression coefficient vector $\boldsymbol{\beta}$ is comparable or significantly larger than m or n, the variable selction process in the Markov Chain Monte Carlo can be extremely slow because the algorithm requires randomly searching the model space with 2^p probabilities. In what follows we develop an optimization-based variational inference for variable selection in the presence of missing data.

2.1 Background on variational inference

We first briefly review the generic variational inference method, also referred to as variational Bayes. For a detailed description, we refer the readers to Chapter 10 of Bishop (2006). It is a family of approximate Bayesian inference methods that differ from classical simulation-based Bayesian inference method (e.g., MCMC or approximate Bayesian computation). Compared with Markov Chain Monte Carlo samplers, the variational inference is formulated as a mathematical optimization problem and is comparably faster than MCMC.

Specialized to the linear mixed-effect model of interest, the variational inference begins with a fully Bayesian model by specifying appropriate prior distributions of the model parameters. Since the linear mixed-effect model has the mixed-effect coefficient $\mathbf{z}_1, \ldots, \mathbf{z}_m$ and the missing portion of the responses (denoted by $\mathbf{Y}_{(\text{mis})}$ generically) as latent variables in addition to the model parameters (denoted by $\boldsymbol{\Theta}$ generically), the posterior inference also takes the latent variables into account and we shall denote the set of all latent variables and parameters by $\boldsymbol{\Phi}$. Meanwhile we denote the set of all observed variables by $\mathbf{Y}_{(\text{obs})}$, and

$$\boldsymbol{\Phi} = \{\mathbf{Y}_{(\text{mis})}, \boldsymbol{\beta}, \mathbf{b}_1, \ldots, \mathbf{b}_n, \tau, \boldsymbol{\Psi}, \mathbf{V}, \nu, w, \gamma, \sigma_e, \sigma_0, \mu_0\}.$$

We first specificy the complete Bayesian model through the prior distribution $p(\boldsymbol{\Theta})$, and our goal is to find a distribution $q(\boldsymbol{\Phi})$ as an approximation for the posterior distribution $p(\boldsymbol{\Phi} \mid \mathbf{Y})$. The distribution q is referred to as the *variational distribution*. To begin with, we first observe the following decomposition of the log marginal distribution of the observed responses $\ln p(\mathbf{Y}_{(\text{obs})})$:

$$\ln p(\mathbf{Y}_{(\text{obs})}) = \mathcal{L}(q) + \mathrm{KL}(q\|p)$$

where $\mathcal{L}(q)$ is referred to as the *evidence lower bound* (ELBO) that can be written as

$$\mathcal{L}(q) = \int q(\mathbf{\Phi}) \ln \left\{ \frac{p(\mathbf{Y}, \mathbf{\Phi})}{q(\mathbf{\Phi})} \right\} \mathrm{d}\mathbf{\Phi},$$

and $\mathrm{KL}(q\|p)$ is the Kullback-Leibler (KL) divergence between the the variational distribution q and the posterior distribution $p(\mathbf{\Phi} \mid \mathbf{Y}_{(\text{obs})})$ of $\mathbf{\Phi}$ given $\mathbf{Y}_{(\text{obs})}$:

$$\mathrm{KL}(q\|p) = -\int q(\mathbf{\Phi}) \ln \left\{ \frac{p(\mathbf{Y}, \mathbf{\Phi})}{q(\mathbf{\Phi})} \right\} \mathrm{d}\mathbf{\Phi}.$$

To obtain an approximation q for the posterior distribution, a reasonable choice is to minimize the Kullback-Leibler divergence $\mathrm{KL}(q\|p)$, and this in turn is equivalent to maximize the ELBO $\mathcal{L}(q)$.

2.2 Bayesian linear mixed-effect model with a spike-and-slab prior

To lay the foundation of the variational inference in the context of the linear mixed-effect model with missing responses, we first specify the fully Bayesian model by assigning a hiearchical prior distribution to $\mathbf{\Theta}$. First note that there are 9 sets of latent latent variables in total:

$$\{\mathbf{Y}_{(\text{obs})}, \boldsymbol{\beta}, \mathbf{B}, \sigma_e^2, \mathbf{\Psi}, \mu_0, \sigma_0^2, w, \gamma\},$$

where $\mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_m] \in \mathbb{R}^{l \times m}$. Denote $\mathcal{I}_{(\text{mis})} = \{(i,j) : y_{ij} \text{ is NA}, i = 1, \dots, m, j = 1, \dots, n\}$ the set of indices (i,j) corresponding to $\mathbf{Y}_{(\text{mis})}$ and $\mathcal{I}_{(\text{obs})}$ the indices corresponding to $\mathbf{Y}_{(\text{obs})}$. The sampling model of the complete data $(\mathbf{Y}_{(\text{obs})}, \mathbf{Y}_{(\text{mis})}, \mathbf{B})$ can be described as

$$\begin{aligned} p(\mathbf{Y}_{(\text{obs})}, \mathbf{Y}_{(\text{mis})}, \mathbf{Z} \mid \mathbf{\Theta}) &= p(\mathbf{Y}_{(\text{obs})} \mid \mathbf{B}, \mathbf{\Theta}) p(\mathbf{Y}_{(\text{mis})} \mid \mathbf{B}, \mathbf{\Theta}) p(\mathbf{B} \mid \mathbf{\Theta}), \\ p(\mathbf{Y}_{(\text{obs})} \mid \mathbf{B}) &= \prod_{(i,j) \in \mathcal{I}_{(\text{obs})}} p(y_{ij} \mid \boldsymbol{\beta}, \mathbf{b}_i, \sigma_e^2), \\ p(\mathbf{Y}_{(\text{mis})} \mid \mathbf{B}) &= \prod_{(i,j) \in \mathcal{I}_{(\text{mis})}} p(y_{ij} \mid \boldsymbol{\beta}, \mathbf{b}_i, \sigma_e^2), \\ p(\mathbf{B} \mid \mathbf{\Psi}) &= \prod_{i=1}^{n} p(\mathbf{b}_i \mid \mathbf{\Psi}). \end{aligned} \tag{2}$$

For each $i = 1, \dots, m$ and $j = 1, \dots, n$, we have

$$\begin{aligned} p(y_{ij} \mid \boldsymbol{\beta}, \mathbf{b}_i, \sigma_e^2) &= \mathrm{N}(y_{ij} \mid \mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta} + \mathbf{z}_{ij}^{\mathrm{T}} \mathbf{b}_i, \sigma_e^2) = \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left\{ -\frac{(y_{ij} - \mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta} - \mathbf{z}_{ij}^{\mathrm{T}} \mathbf{b}_i)^2}{2\sigma_e^2} \right\} \\ p(\mathbf{b}_i | \mathbf{\Psi}) &= \mathrm{N}(\mathbf{b}_i \mid 0, \mathbf{\Psi}) = \prod_{i=1}^{m} \frac{1}{(\sqrt{2\pi})^l |\mathbf{\Psi}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \mathbf{b}_i^{\mathrm{T}} \mathbf{\Psi}^{-1} \mathbf{b}_i \right). \end{aligned}$$

The key to enforce sparsity in the fixed-effect regression coefficient vector $\boldsymbol{\beta}$ lies in the spike-and-slab prior distribution (Castillo et al., 2012, 2015). Specifically, for each coordinate β_k, $k = 1, \dots, p$, the spike-and-slab distribution allows β_k to take zero with a strictly positive probability w, and with probability $1 - w$, β_k is generated from an absolutely continuous distribution supported on $\mathbb{R}$ (here we specifically take the continuous component to be a normal). Formally, given the zero selection probability w, an auxiliary component assignment variable γ_k is generated from $\text{Bernoulli}(1-w)$. The auxiliary variable γ_k has the following interpretation: if $\gamma_k = 0$, then we set $\beta_k = 0$, and if $\gamma_k = 1$, then we draw β_k from the continuous component of the spike-and-slab distribution. The prior samples each coordinate $\beta_1, \ldots, \beta_p$ independently given the selection probability w. Consequently, the conditional prior of $\boldsymbol{\beta}$ given w can be described as follows:

$$\begin{aligned} p(\boldsymbol{\beta}|\gamma_k, \mu_0, \sigma_0^2)\mathrm{d}\boldsymbol{\beta} &= \prod_{k=1}^{p} \{\mathrm{N}(\boldsymbol{\beta}_k|\mu_0, \sigma_0^2)\mathrm{d}\boldsymbol{\beta}_k\}^{\gamma_k}\{\delta_0 \mathrm{d}\boldsymbol{\beta}_k\}^{1-\gamma_k} \\ &= \prod_{k=1}^{p}\left[\frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left\{-\frac{1}{2\sigma_0^2}(\boldsymbol{\beta}_k-\mu_0)^2\right\}\mathrm{d}\boldsymbol{\beta}_k\right]^{\gamma_k}[\delta_0\mathrm{d}\boldsymbol{\beta}_k]^{1-\gamma_k}, \quad (3) \\ p(\gamma_k|w) &= \prod_{k=1}^{p} w^{\gamma_k}(1-w)^{1-\gamma_k}, \end{aligned}$$

where (μ_0, σ_0^2) are the hyperparameters. The complete hierarchical prior distribution is completed by assigning the following hyperprior distributions to the hyperparameters as well as σ_e^2 for the sake of conjugacy:

$$\begin{aligned} p(w) &= \mathrm{Beta}(w \mid a_w, b_w) \propto w^{a_w-1}(1-w)^{b_w-1}, \\ p(\sigma_e^2) &= \mathrm{IG}(\sigma_e^2 \mid a_1, b_1) = \frac{b_1^{a_1}}{\Gamma(a_1)}\left(\frac{1}{\sigma_e^2}\right)^{a_1+1}\exp\left\{-\frac{b_1}{\sigma_e^2}\right\}, \\ p(\mu) &= \mathrm{N}(\mu \mid 0,1) = \frac{1}{\sqrt{2\pi}}e^{-\mu^2/2}, \quad (4) \\ p(\sigma_0^2) &= \mathrm{IG}(\sigma_0^2 \mid 1,1) = \left(\frac{1}{\sigma_0^2}\right)^2\exp\left\{-\frac{1}{\sigma_0^2}\right\}, \\ p(\boldsymbol{\Psi}^{-1}) &= \mathcal{W}(\boldsymbol{\Psi}^{-1} \mid \nu, \mathbf{V}^{-1}) \propto |\boldsymbol{\Psi}^{-1}|^{\frac{\nu-l-1}{2}}\exp\left\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Psi}^{-1}\mathbf{V}^{-1})\right\}. \end{aligned}$$

Then the entire hierarchical Bayesian model is completed by the distributions (2), (3), and (4).

3. Variational Inference With a Spike-and-Slab Prior

Leveraging the hierarchical Bayesian model in Section 2.2, in this section, we are now in a position to describe the framework of variational inference in the context of the linear mixed-effect model with missing responses. We follow the commonly-adopted mean-field approximation assumption and set the variational distributions in the following factorized form (Bishop, 2006):

$$q(\boldsymbol{\Phi}) = q(\mathbf{Y}_{(\mathrm{mis})})q(\boldsymbol{\beta}, \gamma)q(\mathbf{B})q(\sigma_e^2)q(\boldsymbol{\Psi})q(\mu_0)q(\sigma_0^2)q(w). \quad (5)$$

Here, $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_p]^{\mathrm{T}}$ is the p-dimensional auxiliary variable that specify the sparsity of $\boldsymbol{\beta}$ in the spike-and-slab distribution setup. Note, however, that the mathematical optimization problem

$$\max_q \mathcal{L}\{q(\boldsymbol{\Phi})\} \quad \text{subject to} \quad q(\boldsymbol{\Phi}) \text{ satisfies constraint (5)}$$

is an infinite-dimensional optimization problem because constraint (5) is still an infinite-dimensional statistical manifold. Although an easy-to-implement coordinate-ascent variational algorithm can be obtained when the likelihood is in the exponential family form as suggested by Bishop (2006), our case brings additional computational bottleneck as we introduce the spike-and-slab prior (3) for $\boldsymbol{\beta}$ with singularity. Alternatively, it is also reasonable to posit certain parametric form of the variational distributions q such that (5) can be further reduced to a finite-dimensional statistical manifold (see, for example, Blei et al., 2003). Hence, we further assume the following parametric form of the variational distributions for the sake of conjugacy:

$$\begin{aligned}
q(\mathbf{Y}_{(\mathrm{mis})} \mid (\widehat{\mu}_{y_{ij}}, \widehat{\sigma}^2_{y_{ij}})_{(i,j)\in\mathcal{I}_{(\mathrm{mis})}}) &= \prod_{(i,j)\in\mathcal{I}_{(\mathrm{mis})}} \mathrm{N}(y_{ij} \mid \widehat{\mu}_{y_{ij}}, \widehat{\sigma}^2_{y_{ij}}),\\
q(\boldsymbol{\beta}, \gamma \mid (\theta_k, \widehat{\mu}_{\beta_k}, \widehat{\sigma}^2_{\beta_k})_{k=1}^{p})\mathrm{d}\boldsymbol{\beta} &= \prod_{k=1}^{p} \left\{\theta_k \mathrm{N}(\boldsymbol{\beta}_k \mid \hat{\mu}_{\beta_k}, \hat{\sigma}^2_{\beta_k})\mathrm{d}\beta_k\right\}^{\gamma_k} \left\{(1-\theta_k)\delta_0 \mathrm{d}\beta_k\right\}^{1-\gamma_k},\\
q(\mathbf{B} \mid \widehat{\boldsymbol{\mu}}_{\mathbf{b}_1}, \widehat{\boldsymbol{\Psi}}_1, \dots, \widehat{\boldsymbol{\mu}}_{\mathbf{b}_m}, \widehat{\boldsymbol{\Psi}}_m) &= \prod_{i=1}^{m} \mathrm{N}(\mathbf{b}_i \mid \widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}, \widehat{\boldsymbol{\Psi}}_{\mathbf{b}_i}),\\
q(\sigma_e^2 \mid \hat{a}_{\sigma_e^2}, \hat{b}_{\sigma_e^2}) &= \mathrm{IG}(\sigma_e^2 \mid \hat{a}_{\sigma_e^2}, \hat{b}_{\sigma_e^2}),\\
q(\boldsymbol{\Psi}^{-1} \mid \widehat{\boldsymbol{\Psi}})\mathrm{d}\boldsymbol{\Psi}^{-1} &= \delta_{\widehat{\boldsymbol{\Psi}}^{-1}}(\mathrm{d}\boldsymbol{\Psi}^{-1}),\\
q(\mu_0 \mid \hat{\mu}_{\mu_0}, \hat{\sigma}^2_{\mu_0}) &= \mathrm{N}(\mu_0 \mid \hat{\mu}_{\mu_0}, \hat{\sigma}^2_{\mu_0}),\\
q(\sigma_0^2 \mid \hat{a}_{\sigma_0^2}, \hat{b}_{\sigma_0^2}) &= \mathrm{IG}(\sigma_0^2 \mid \hat{a}_{\sigma_0^2}, \hat{b}_{\sigma_0^2}),\\
q(w \mid \hat{a}_w, \hat{b}_w) &= \mathrm{Beta}(w \mid \hat{a}_w, \hat{b}_w) = \frac{w^{\hat{a}_w - 1}(1-w)^{\hat{b}_w - 1}}{B(\hat{a}_w, \hat{b}_w)}.
\end{aligned} \tag{6}$$

Note that we could use a Wishart variational distribution to approximate the posterior of the precision matrix $\boldsymbol{\Psi}^{-1}$ for the random effect coefficient $\mathbf{b}_1, \dots, \mathbf{b}_m$. Here we use a Dirac point mass at $\widehat{\boldsymbol{\Psi}}^{-1}$ to indicate that a point estimator for $\boldsymbol{\Psi}$ is taken, and the resulting solution is a maximum *a posteriori* estimator for $\boldsymbol{\Psi}$. We also remark that the set of parameters

$$\begin{aligned}
\boldsymbol{\Xi} := \Big\{ & (\widehat{\mu}_{y_{ij}}, \widehat{\sigma}^2_{y_{ij}})_{i,j\in\mathcal{I}_{(\mathrm{mis})}}, (\theta_k, \widehat{\mu}_{\beta_k}, \widehat{\sigma}^2_{\beta_k})_{k=1}^{p}, (\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}, \widehat{\boldsymbol{\Psi}}_{\mathbf{b}_i})_{i=1}^{m},\\
& (\widehat{a}_{\sigma_e^2}, \widehat{b}_{\sigma_e^2}), (\widehat{a}_{\sigma_e^2}, \widehat{b}_{\sigma_e^2}), \widehat{\boldsymbol{\Psi}}, (\widehat{\mu}_{\mu_0}, \widehat{\sigma}^2_{\mu_0}), (\widehat{a}_w, \widehat{b}_w) \Big\}.
\end{aligned}$$

are the variables to be learned by optimizing the objective function $\mathcal{L}(q)$. The the variational inference is implemented by iteratively maximizing the objective function $\mathcal{L}(q)$ with respect to the variational parameters $\boldsymbol{\Xi}$. Below, we present the complete variational inference algorithm. The detailed derivation is deferred to Appendix.

- **Input:**

  Response matrix $\mathbf{Y} \in \mathbb{R}^{m\times n}$ (with potentially missing entries)

  Fixed effect covariate tensor $\mathbb{X} = [x_{ijk}]_{m\times n\times p}$

  Random effect covariate tensor $\mathbb{Z} = [z_{ijt}]_{m\times n\times l}$

1. **Step 1:** Randomly initialize the variational parameters

   $$\begin{aligned}
   \boldsymbol{\Xi} := \Big\{ & (\widehat{\mu}_{y_{ij}}, \widehat{\sigma}^2_{y_{ij}})_{i,j\in\mathcal{I}_{(\mathrm{mis})}}, (\theta_k, \widehat{\mu}_{\beta_k}, \widehat{\sigma}^2_{\beta_k})_{k=1}^{p}, (\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}, \widehat{\boldsymbol{\Psi}}_{\mathbf{b}_i})_{i=1}^{m},\\
   & (\widehat{a}_{\sigma_e^2}, \widehat{b}_{\sigma_e^2}), (\widehat{a}_{\sigma_e^2}, \widehat{b}_{\sigma_e^2}), \widehat{\boldsymbol{\Psi}}, (\widehat{\mu}_{\mu_0}, \widehat{\sigma}^2_{\mu_0}), (\widehat{a}_w, \widehat{b}_w) \Big\}.
   \end{aligned}$$

   and compute the following matrices related to the fixed-effect covariate tensor $\mathbf{X}$:

   $$\mathbf{D}_1(\mathbf{X}) = \sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{x}_{ij}\mathbf{x}_{ij}^{\mathrm{T}}, \quad \mathbf{D}_2(\mathbf{X}) = \sum_{i=1}^{m}\sum_{j=1}^{n} \mathrm{diag}(x_{ij1}^2, \dots, x_{ijp}^2).$$

2. **Step 2:** Loop until the objective function Ω converges:

    - Update $\widehat{\boldsymbol{\Psi}}$ by maximizing Ω with respect to $\widehat{\boldsymbol{\Psi}}$: This yields

    $$\widehat{\boldsymbol{\Psi}} \longleftarrow \frac{1}{(m+\nu-l-1)}\left\{\sum_{i=1}^{m}(\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}^{\mathrm{T}}+\widehat{\boldsymbol{\Psi}}_{\mathbf{b}_i})+\mathbf{V}^{-1}\right\}$$

    - Update $\widehat{\mu}_{y_{ij}}, \widehat{\sigma}^2_{y_{ij}}$ for all $(i,j) \in \mathcal{I}_{(\mathrm{mis})}$ using the following formula:

    $$\widehat{\mu}_{ij} \longleftarrow \sum_{k=1}^{p} x_{ijk}\theta_k\widehat{\mu}_{\beta_k} + \mathbf{z}_{ij}^{\mathrm{T}}\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}, \quad \widehat{\sigma}^2_{y_{ij}} \longleftarrow \frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}}.$$

    - Update $(\theta_k, \widehat{\mu}_{\beta_k}, \widehat{\sigma}^2_{\beta_k})$ for all $k = 1, \dots, p$. This step is the major computation bottleneck of the entire algorithm and we seek an updating rule that can be vectorized. For all $i = 1, \dots, m$ and $j = 1, \dots, n$, we set

    $$\widehat{y}_{ij} \longleftarrow \begin{cases} y_{ij}, & \text{if } (i,j) \notin \mathcal{I}_{(\mathrm{mis})}, \\ \widehat{\mu}_{y_{ij}}, & \text{if } (i,j) \in \mathcal{I}_{(\mathrm{mis})}, \end{cases},$$

    $$\langle y_{ij}^2\rangle_q \longleftarrow \begin{cases} y_{ij}^2, & \text{if } (i,j) \notin \mathcal{I}_{(\mathrm{mis})}, \\ \widehat{\mu}^2_{y_{ij}} + \widehat{\sigma}^2_{y_{ij}}, & \text{if } (i,j) \in \mathcal{I}_{(\mathrm{mis})}. \end{cases}$$

    Set

    $$\mathbf{v} \longleftarrow \sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{x}_{ij}(\widehat{y}_{ij} - \mathbf{z}_{ij}^{\mathrm{T}}\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}).$$

    Next, we consider vectorized updating formula for $\boldsymbol{\theta}, \widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}, \widehat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}}$, where $\boldsymbol{\theta} = [\theta_1, \dots, \theta_p]^{\mathrm{T}}$, $\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}} = [\widehat{\mu}_{\beta_1}, \dots, \widehat{\mu}_{\beta_p}]^{\mathrm{T}}$, and $\widehat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}} = [\widehat{\sigma}^2_{\beta_1}, \dots, \widehat{\sigma}^2_{\beta_p}]^{\mathrm{T}}$. Denote $\mathbf{1}_p = [1, \dots, 1]^{\mathrm{T}} \in \mathbb{R}^p$. For any function $f : \mathcal{D} \subset \mathbb{R} \to \mathbb{R}$, we denote $f(\mathbf{x})$ generically as the entrywise application of f to the elements of $\mathbf{x}$, namely, $f([x_1, \dots, x_p]^{\mathrm{T}}) := [f(x_1), \dots, f(x_p)]^{\mathrm{T}}$. Denote $\mathbf{x}_1 \circ \mathbf{x}_2$ as the entrywise product between two vector $\mathbf{x}_1, \mathbf{x}_2$ of the same dimension, and $\mathbf{x}_1/\mathbf{x}_2$ as the entrywise ratio between $\mathbf{x}_1$ and $\mathbf{x}_2$. Finally, we use $\psi(\cdot)$ to denote the digamma function $\psi(x) = (\mathrm{d}/\mathrm{d}x)\ln[\Gamma(x)]$, where $\Gamma(\cdot)$ is the Gamma function. Then $\boldsymbol{\theta}$ can be obtained by first computing the update for $\mathrm{logit}(\boldsymbol{\theta})$:

    $$\begin{aligned}\mathrm{logit}(\boldsymbol{\theta}) \longleftarrow & \left[\psi(\widehat{a}_w) - \psi(\widehat{b}_w) + \frac{1}{2}\psi(\widehat{a}_{\sigma_0^2}) - \frac{1}{2}\ln(\widehat{b}_{\sigma_0^2})\right]\mathbf{1}_p + \frac{1}{2}\ln(\widehat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}}) + \frac{1}{2}\mathbf{1}_p \\ & - \frac{1}{2}\left(\frac{\widehat{a}_{\sigma_0^2}}{\widehat{b}_{\sigma_0^2}}\right)[(\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}_k} - \widehat{\mu}_{\mu_0}\mathbf{1}_p) \circ (\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}_k} - \widehat{\mu}_{\mu_0}\mathbf{1}_p) + \widehat{\sigma}^2_{\mu_0}\mathbf{1}_p + \widehat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}}] \\ & + \widehat{\boldsymbol{\mu}}^2_{\boldsymbol{\beta}} \circ \left(\frac{1}{\widehat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}}}\right) - \left(\frac{\widehat{a}_{\sigma_0^2}}{\widehat{b}_{\sigma_0^2}}\right)\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}\widehat{\mu}_{\mu_0} - \left(\frac{\widehat{a}_{\sigma_0^2}}{\widehat{b}_{\sigma_0^2}}\right)\widehat{\boldsymbol{\mu}}^2_{\boldsymbol{\beta}} \\ & - \frac{1}{2}\left(\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}}\right)\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{x}_{ij} \circ \mathbf{x}_{ij} \circ (\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}} \circ \widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}} + \boldsymbol{\sigma}^2_{\boldsymbol{\beta}}).\end{aligned}$$

    Then $\boldsymbol{\theta}$ can be directly obtained by taking $\boldsymbol{\theta} \longleftarrow \mathrm{logit}^{-1}(\mathrm{logit}(\boldsymbol{\theta}))$. Set

    $$\boldsymbol{\Theta} = \mathrm{diag}(\theta_1, \dots, \theta_p).$$

We then compute the updates for $\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}}$:

$$\widehat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}} \longleftarrow \left(\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}} \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{x}_{ij} \circ \mathbf{x}_{ij} + \frac{\widehat{a}_{\sigma_0^2}}{\widehat{b}_{\sigma_0^2}} \mathbf{1}_p \right)^{-1},$$

$$\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}} \longleftarrow \left[\left(\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}} \right) \mathbf{D}_1(\mathbb{X}) \boldsymbol{\Theta} + \left(\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}} \right) \mathbf{D}_2(\mathbb{X})(\mathbf{I}_p - \boldsymbol{\Theta}) + \frac{\widehat{a}_{\sigma_0^2}}{\widehat{b}_{\sigma_0^2}} \mathbf{I}_p \right]^{-1} \times \left(\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}} \mathbf{v} + \frac{\widehat{a}_{\sigma_0^2}}{\widehat{b}_{\sigma_0^2}} \widehat{\mu}_{\mu_0} \mathbf{1}_p \right).$$

Note that the above vectorized updating rules can be easily implemented in `R`, `MATLAB`, or `Python`.

- Update $(\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}, \widehat{\boldsymbol{\Psi}}_{\mathbf{b}_i})$ for all $i = 1, \ldots, m$:

$$\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i} \longleftarrow \widehat{\boldsymbol{\Psi}}_{\mathbf{b}_i} \left[\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}} \sum_{j=1}^{n} (\widehat{y}_{ij} - \mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}) \mathbf{z}_{ij} \right],$$

$$\widehat{\boldsymbol{\Psi}}_{\mathbf{b}_i} \longleftarrow \left(\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}} \sum_{j=1}^{n} \mathbf{z}_{ij} \mathbf{z}_{ij}^{\mathrm{T}} + \widehat{\boldsymbol{\Psi}}^{-1} \right)^{-1}.$$

- Update $(\widehat{a}_{\sigma_e^2}, \widehat{b}_{\sigma_e^2}), (\widehat{a}_{\sigma_0^2}, \widehat{b}_{\sigma_0^2}), (\widehat{a}_w, \widehat{b}_w), (\widehat{\mu}_{\mu_0}, \widehat{\sigma}^2_{\mu_0})$:

$$\begin{aligned}
\mathrm{SSR} \longleftarrow & \sum_{i=1}^{m} \sum_{j=1}^{n} \left\{ \langle y_{ij}^2 \rangle_q + \mathbf{z}_{ij}^{\mathrm{T}} (\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i} \widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}^{\mathrm{T}} + \widehat{\boldsymbol{\Psi}}_{\mathbf{b}_i}) \mathbf{z}_{ij} + \left(\sum_{k=1}^{p} x_{ijk} \theta_k \widehat{\mu}_{\beta_k} \right)^2 \right\} \\
& + \sum_{i=1}^{m} \sum_{j=1}^{n} \left\{ \sum_{k=1}^{p} x_{ijk}^2 [\theta_k (1 - \theta_k) \widehat{\mu}^2_{\beta_k} + \theta_k \widehat{\sigma}^2_{\beta_k}] \right\} \\
& - 2 \sum_{i=1}^{m} \sum_{j=1}^{n} \widehat{y}_{ij} \left(\sum_{k=1}^{p} x_{ijk} \theta_k \widehat{\mu}_{\beta_k} + \mathbf{z}_{ij}^{\mathrm{T}} \widehat{\boldsymbol{\mu}}_{\mathbf{b}_i} \right) \\
& + 2 \sum_{i=1}^{m} \sum_{j=1}^{n} \widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}^{\mathrm{T}} \mathbf{z}_{ij} \left(\sum_{k=1}^{p} x_{ijk} \theta_k \widehat{\mu}_{\beta_k} \right),
\end{aligned}$$

$$\widehat{a}_{\sigma_e^2} \longleftarrow a_1 + \frac{mn}{2}, \quad \widehat{b}_{\sigma_e^2} \longleftarrow b_1 + \frac{1}{2}\mathrm{SSR},$$

$$\widehat{a}_{\sigma_0^2} \longleftarrow 1 + \frac{1}{2} \sum_{k=1}^{p} \theta_k, \quad \widehat{b}_{\sigma_0^2} \longleftarrow 1 + \frac{1}{2} \sum_{k=1}^{p} \theta_k [(\widehat{\mu}_{\mu_0} - \widehat{\mu}_{\beta_k})^2 + \widehat{\sigma}^2_{\mu_0} + \widehat{\sigma}^2_{\beta_k}],$$

$$\widehat{a}_w = a_w + \sum_{k=1}^{p} \theta_k, \quad \widehat{b}_w = b_w + \sum_{k=1}^{p} (1 - \theta_k)$$

$$\widehat{\sigma}^2_{\mu_0} \longleftarrow \left(1 + \frac{\widehat{a}_{\sigma_0^2}}{\widehat{b}_{\sigma_0^2}} \sum_{k=1}^{p} \theta_k \right)^{-1}, \quad \widehat{\mu}_{\mu_0} \longleftarrow \widehat{\sigma}^2_{\mu_0} \frac{\widehat{a}_{\sigma_0^2}}{\widehat{b}_{\sigma_0^2}} \sum_{k=1}^{p} \theta_k \widehat{\mu}_{\beta_k}.$$

3. **Step 3:** Output variational parameters

$$\left\{ (\widehat{\mu}_{y_{ij}}, \widehat{\sigma}^2_{y_{ij}})_{i,j \in \mathcal{I}_{(\mathrm{mis})}}, (\theta_k, \widehat{\mu}_{\beta_k}, \widehat{\sigma}^2_{\beta_k})_{k=1}^{p}, (\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}, \widehat{\boldsymbol{\Psi}}_{\mathbf{b}_i})_{i=1}^{m}, (\widehat{a}_{\sigma_e^2}, \widehat{b}_{\sigma_e^2}), (\widehat{a}_{\sigma_e^2}, \widehat{b}_{\sigma_e^2}), \widehat{\boldsymbol{\Psi}}, (\widehat{\mu}_{\mu_0}, \widehat{\sigma}^2_{\mu_0}), (\widehat{a}_w, \widehat{b}_w) \right\}.$$

4. Application in Multiple Imputation of Missing Data

In this section, we discuss the application of the variational inference algorithm developed in Section 3 to mulitiple imputation of missing data. Although in Section 2 the response variable y_{ij}'s are necessarily continuous and the sampling model of interest posits a linear mixed-effect model with Gaussian noise, we will see next that the variational inference algorithm is also valid even if the sampling model is misspecified. For example, in Section 4.2, the response variable is categorical but we can still apply a misspecified linear mixed-effect model for variational inference. In this case, a calibration-based method due to Yucel et al. (2008) and Yucel et al. (2011) can be applied improve the multiple imputation method based on rounding continuous imputation method.

4.1 Sequential hierarchical regression imputation for continuous data

We first consider the case where the responses y_{ij}'s are continuous. Recall that the linear mixed-effect model (1) assumes that the noise ϵ_{ij}'s are Gaussian. Therefore, the responses in the sampling model (1) is necessarily continuous, so that the variational inference algorithm developed in Section 3 are directly applicable. Note, however, that the inference task in Section 3 was to compute an approximation to the exact posterior distribution, referred to as the variational distribution, by solving a mathematical optimization problem. In contrast, the focus here is on the multiple imputation of the missing portion $\mathbf{Y}_{(\text{mis})}$ of the response matrix $\mathbf{Y}$. The two goals can be simultaneously accomplished when the MCMC is applied for posterior computation. In an MCMC sampler, posterior samples of the model parameters $\mathbf{\Theta}$ are drawn from a Markov chain that converges to the exact posterior distribution of $\mathbf{\Theta}$ given $\mathbf{Y}_{(\text{obs})}$, whereas posterior predictive samples of the $\mathbf{Y}_{(\text{mis})}$ serves as the imputed values of the missing portion of $\mathbf{Y}$. In what follows, we slightly explore the strategy for imputation of $\mathbf{Y}_{(\text{mis})}$ using the output of the variational inference algorithm in Section 3.

We follow the sequential hierarchical regression imputation (SHRIMP) strategy developed in Yucel et al. (2017), which has also been applied in Li and Yucel (2020). The basic idea is that the response variables y_{ij}'s are sorted by columns according to the missing ratio. The left-most column of $\mathbf{Y}$ after sorting has the least number of missing values, whereas the right-most column of $\mathbf{Y}$ after sorting has the most number of missing values. Formally, for a column $j \in \{1, \dots, n\}$, we define the missing ratio as the percentage of missingness:

$$R_j^{(\text{mis})} := \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}(y_{ij} \text{ is NA}).$$

Then SHRIMP relables the column indices $[n] = \{1, 2, \dots, n\}$ as $\{j_1, \dots, j_n\}$ such that $R_{j_r}^{(\text{mis})} \le R_{j_{r+1}}^{(\text{mis})}$ for all $r = 1, 2, \dots, n-1$. Namely, the ratio of the missing percentage of a column in $\mathbf{Y}$ is always no greater than that of the next column after sorting. Then for each fixed i, SHRIMP sequentially impute missing values of y_{ij} from $j = j_1$ to $j = j_n$.

Below, we outline the step-by-step procedure for imputation of $\mathbf{Y}_{(\text{mis})}$ using the output of the variational inference algorithm.

- **Input:**

  Response matrix $\mathbf{Y} \in \mathbb{R}^{m \times n}$ (with potentially missing entries)

  Fixed effect covariate tensor $\mathbb{X} = [x_{ijk}]_{m \times n \times p}$

  Random effect covariate tensor $\mathbb{Z} = [z_{ijt}]_{m \times n \times l}$

  Number of MIs $M \in \mathbb{N}_+$

1. **Step 1: Run variational inference.** Call the variational inference algorithm in Section 3 to obtain the variational parameters

$$\Big\{(\widehat{\mu}_{y_{ij}},\widehat{\sigma}^2_{y_{ij}})_{i,j\in\mathcal{I}_{(\text{mis})}},(\theta_k,\widehat{\mu}_{\beta_k},\widehat{\sigma}^2_{\beta_k})_{k=1}^p,(\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i},\widehat{\boldsymbol{\Psi}}_{\mathbf{b}_i})_{i=1}^m,$$
$$(\widehat{a}_{\sigma_e^2},\widehat{b}_{\sigma_e^2}),(\widehat{a}_{\sigma_e^2},\widehat{b}_{\sigma_e^2}),\widehat{\boldsymbol{\Psi}},(\widehat{\mu}_{\mu_0},\widehat{\sigma}^2_{\mu_0}),(\widehat{a}_w,\widehat{b}_w)\Big\}.$$

2. **Step 2: Sort column indices.** For each column $j=1,\dots,n$, compute the percentage of missingness

$$R_j^{(\text{mis})}:=\frac{1}{m}\sum_{i=1}^m \mathbb{1}(y_{ij}\text{ is NA}).$$

Order the column indices $\{1,2,\dots,n\}$ of the response matrix $\mathbf{Y}$ such that the sorted indices $\{j_1,\dots,j_n\}$ satisfy $R_{j_r}^{(\text{mis})}\le R_{j_{r+1}}^{(\text{mis})}$ for all $r=1,\dots,n-1$.

3. **Step 3: Draw missing values.** For $t=1,\dots,M$: Sample

$$\begin{aligned}&(\sigma_e^2)^{(t)}\sim \text{IG}(\widehat{a}_{\sigma_e^2},\widehat{b}_{\sigma_e^2}),\\&\beta_k^{(t)}\sim\theta_k\text{N}(\widehat{\mu}_{\beta_k}^{(t)},(\widehat{\sigma}^2_{\beta_k})^{(t)})+(1-\theta_k^{(t)})\delta_0,\quad k=1,\dots,p.\end{aligned}$$

For $i=1,\dots,m$, sample

$$\mathbf{b}_i\sim\text{N}(\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i},\widehat{\boldsymbol{\Psi}}_{\mathbf{b}_i});$$

For $r=1,\dots,n$: If $(i,j)\in\mathcal{I}_{(\text{mis})}$, sample

$$y_{ij_r}^{(t)}\sim\text{N}(\mathbf{x}_{ij_r}^{\text{T}}\boldsymbol{\beta}^{(t)}+\mathbf{z}_{ij_r}^{\text{T}}\mathbf{b}_i,(\sigma_e^2)^{(t)}).$$

End For

End for

Set $\widetilde{\mathbf{Y}}^{(t)}=[\widetilde{y}_{ij}^{(t)}]_{m\times n}$, where

$$\widetilde{y}_{ij}^{(t)}=\begin{cases}y_{ij}, & \text{if }(i,j)\notin\mathcal{I}_{(\text{mis})},\\ y_{ij}^{(t)}, & \text{if }(i,j)\in\mathcal{I}_{(\text{mis})}\end{cases}$$

4. **Step 4:** Output

$$\widetilde{\mathbf{Y}}^{(1)},\dots,\widetilde{\mathbf{Y}}^{(M)}.$$

4.2 Calibration-based imputation for categorical data

We now turn to the case where the response variable y_{ij}'s are categorical. According to the aforementioned linear mixed-effect model (1) with Gaussian noise ϵ_{ij}'s, the response matrix $\mathbf{Y}$ consists of continuous data. Here we adopt a technique, referred to as the calibration-based imputation method (Yucel et al., 2008, 2011), that allows for the model misspecification when the specified model is designed for continuous data but the actual data is categorical. The technique is a quite general imputation method for missing categorical data. The basic requirement is that there exists a surrogate imputation model $p(\mathbf{Y}_C)$ that generates continuous data such that the categorical data $\mathbf{Y}$ can be viewed as a discretized version of the continuous data $\mathbf{Y}_C$ generated from the surrogate imputation model $p(\mathbf{Y}_C)$.

The key idea of the calibration-based imputation method can be loosely stated as follows. We generate two copies $\mathbf{Y}_C,\mathbf{Y}_{(\text{dup}),C}$ using an imputation method based on the continuous surrogate model $p(\mathbf{Y}_C)$. Let $\mathbf{Y}_{(\text{obs}),C}$, $\mathbf{Y}_{(\text{mis}),C}$ be the portion of $\mathbf{Y}_C$ corresponding to the observed or missing locations of $\mathbf{Y}_{(\text{obs})}$ or $\mathbf{Y}_{(\text{mis})}$ in $\mathbf{Y}$, respectively, and we define $\mathbf{Y}_{(\text{obs,dup}),C}$ and $\mathbf{Y}_{(\text{mis,dup}),C}$ for $\mathbf{Y}_{(\text{dup}),C}$ similarly. Suppose the number of total categories in $\mathbf{Y}$ is G and the categories are labeled as $g = 1, 2, \ldots, G$. Then the calibration-based imputation method proposes to compute a sequence of cut-off values $-\infty = c_0, c_1, \ldots, c_{g-1}$ that can be determined as follows. For each (i, j) pair corresponding to an observed y_{ij}, we let $y_{ij}^{(\text{dup})} = g$ if the (i, j)th element of $\mathbf{Y}_{(\text{obs,dup}),C}$ lies in the interval $(c_{g-1}, c_g]$. The cut-off values $c_1, \ldots, c_{g-1}$ are selected such that for each category $g = 1, \ldots, G$,

$$\sum_{(i,j)\in\mathcal{I}_{(\text{obs})}} \mathbb{1}\{y_{ij}^{(\text{dup})} = g\} = \sum_{(i,j)\in\mathcal{I}_{(\text{obs})}} \mathbb{1}\{y_{ij} = g\}.$$

In other words, the proportions of different categories after applying the cut-off values $c_0, c_1, \ldots, c_{g-1}$ to $\mathbf{Y}_{(\text{obs,dup}),C}$ matches with the proportions of different categories in $\mathbf{Y}_{(\text{obs})}$. These cut-off values can be computed explicitly using quantiles of $\mathbf{Y}_{(\text{obs,dup}),C}$.

We now describe the outline of the calibration-based imputation method using the output of the variational inference algorithm developed in Section 3.

■ **Input:**

Categorical response matrix $\mathbf{Y} \in \mathbb{R}^{m\times n}$ (with potentially missing entries)

Fixed effect covariate tensor $\mathbb{X} = [x_{ijk}]_{m\times n\times p}$

Random effect covariate tensor $\mathbb{Z} = [z_{ijt}]_{m\times n\times l}$

Number of MIs $M \in \mathbb{N}_+$

1. **Step 1: Run variational inference.** Call the variational inference algorithm in Section 3 to obtain the variational parameters

$$\Big\{(\widehat{\mu}_{y_{ij}}, \widehat{\sigma}^2_{y_{ij}})_{i,j\in\mathcal{I}_{(\text{mis})}}, (\theta_k, \widehat{\mu}_{\beta_k}, \widehat{\sigma}^2_{\beta_k})_{k=1}^{p}, (\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}, \widehat{\boldsymbol{\Psi}}_{\mathbf{b}_i})_{i=1}^{m},$$
$$(\widehat{a}_{\sigma_e^2}, \widehat{b}_{\sigma_e^2}), (\widehat{a}_{\sigma_e^2}, \widehat{b}_{\sigma_e^2}), \widehat{\boldsymbol{\Psi}}, (\widehat{\mu}_{\mu_0}, \widehat{\sigma}^2_{\mu_0}), (\widehat{a}_w, \widehat{b}_w)\Big\}.$$

2. **Step 2: Generate imputed data for missing responses.**

For $t = 1, \ldots, M$:

Sample

$$(\sigma_e^2)^{(t)} \sim \text{IG}(\widehat{a}_{\sigma_e^2}, \widehat{b}_{\sigma_e^2}),$$
$$\beta_k^{(t)} \sim \theta_k \text{N}(\widehat{\mu}_{\beta_k}^{(t)}, (\widehat{\sigma}^2_{\beta_k})^{(t)}) + (1 - \theta_k^{(t)})\delta_0, \quad k = 1, \ldots, p.$$

For $i = 1, \ldots, m$, sample

$$\mathbf{b}_i \sim \text{N}(\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}, \widehat{\boldsymbol{\Psi}}_{\mathbf{b}_i});$$

For $j = 1, \ldots, n$: sample

$$(y_{ij}^{(\text{dup}),C})^{(t)} \sim \text{N}(\mathbf{x}_{ij}^{\text{T}}\boldsymbol{\beta}^{(t)} + \mathbf{z}_{ij}^{\text{T}}\mathbf{b}_i, (\sigma_e^2)^{(t)}),$$
$$(y_{ij}^{C})^{(t)} \sim \text{N}(\mathbf{x}_{ij}^{\text{T}}\boldsymbol{\beta}^{(t)} + \mathbf{z}_{ij}^{\text{T}}\mathbf{b}_i, (\sigma_e^2)^{(t)}).$$

End for

End for

For each $g = 1, 2, \ldots, G-1$, compute

$$s_g^{(t)} = \frac{1}{|\mathcal{I}_{(\mathrm{obs})}|} \sum_{(i,j)\in\mathcal{I}_{(\mathrm{obs})}} \mathbb{1}\{y_{ij} \leq g\}.$$

Compute the $s_g^{(t)}$-sample quantile $c_g^{(t)}$ of $\{y_{(\mathrm{dup}),C}^{(t)}(i,j) : (i,j) \notin \mathcal{I}_{(\mathrm{mis})}\}$ such that

$$s_g^{(t)} = \frac{1}{|\mathcal{I}_{(\mathrm{obs})}|} \sum_{(i,j)\in\mathcal{I}_{(\mathrm{obs})}} \mathbb{1}\{(y_{ij}^{(\mathrm{dup}),C})^{(t)} \leq c_g^{(t)}\}, \quad g = 1, \ldots, G-1.$$

Set $c_0^{(t)} = -\infty$ and $c_G^{(t)} = +\infty$. Set the tth imputed complete data to be $\widetilde{\mathbf{Y}}^{(t)} = [\widetilde{y}_{ij}^{(t)}]$, where

$$\widetilde{y}_{ij}^{(t)} = \begin{cases} y_{ij}, & \text{if } (i,j) \notin \mathcal{I}_{(\mathrm{mis})}, \\ g, & \text{if } (i,j) \in \mathcal{I}_{(\mathrm{mis})} \text{ and } (y_{ij}^{C})^{(t)} \in (c_{g-1}^{(t)}, c_g^{(t)}]. \end{cases}$$

3. **Step 3:** Output

$$\widetilde{\mathbf{Y}}^{(1)}, \ldots, \widetilde{\mathbf{Y}}^{(M)}$$

5. Numerical examples

5.1 Continuous data example

We first consider a well-specified example with continuous responses. The generative model for the synthetic data is the same as (1):

$$y_{ij} = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_{ij}^{\mathrm{T}}\mathbf{b}_i + \epsilon_{ij}, \quad \epsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma_e^2), \quad \mathbf{b}_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mathbf{0}_l, \boldsymbol{\Psi}),$$

where $i = 1, \ldots, m$, $j = 1, \ldots, n$. We set $m = 50$, $n = 20$, $p = 100$, and $l = 3$. The coordinates of the fixed-effect covariates $\mathbf{x}_{ij}$'s are independently generated from $\mathrm{N}(0, 3^2)$ for all i, j, and the coordinates of the random-effect covariances $\mathbf{z}_{ij}$'s are independently drawn from $\mathrm{N}(0, 1)$ for all i, j. The fixed-effect regression coefficient $\boldsymbol{\beta}$ is assumed to have a weakly sparse structure. We adopt the "three-peak curve"example constructed in Johnstone and Lu (2009) and set the coordinates of $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^{\mathrm{T}} \in \mathbb{R}^p$ as follows:

$$\begin{aligned} \beta_k =& \frac{0.7}{20}\mathrm{Beta}\left(k/p \mid 1500, 3000\right) + \frac{0.5}{20}\mathrm{Beta}(k/p \mid 1200, 900) \\ &+ \frac{0.5}{20}\mathrm{Beta}(k/p \mid 600, 160), \end{aligned}$$

where $\mathrm{Beta}(t \mid a, b)$ is the density of the $\mathrm{Beta}(a, b)$ distribution evaluated at $t \in [0, 1]$. Figure 1 below visualizes the true values of β_j as a function of $j = 1, \ldots, p$, from which we can see that a significant portion of the coordinates of $\boldsymbol{\beta}$ are rather close to 0, whereas only several coordinates are bounded away from 0. The covariance matrix $\boldsymbol{\Psi}$ for the random effect is set to be the 3×3 identity matrix.

The response matrix $\mathbf{Y} \in \mathbb{R}^{m\times n}$ is contaminated by missing values. We consider the missing at random mechanism, meaning that the distribution of the missingness depends on the observed variables but not the missing values. To this end, we generate a collection of binary random variables $\boldsymbol{\Gamma} = [\gamma_{ij}]_{m\times n}$ to assign missing values to $\mathbf{Y}$. Specifically, we take

$$\gamma_{ij} \sim \mathrm{Bernoulli}(p_{ij}), \quad \mathrm{where} \mathrm{logit}(p_{ij}) = \alpha_R + \beta_{\mathrm{mis}} x_{ij1}.$$

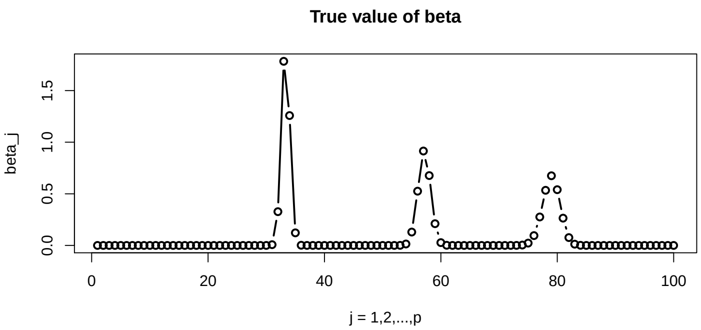

Figure 1: True value of $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]$ in Section 5.1.

Here the parameters α_R and β_{mis} are tunned such that the overall missing percentage of $\mathbf{Y}$ is approximately 27%, i.e., $(1/mn)\sum_{i=1}^{m}\sum_{j=1}^{n}\gamma_{ij} \approx 0.27$. We then set y_{ij} to be `NA` if $\gamma_{ij} = 1$, and maintain the orignal value of y_{ij} if $\gamma_{ij} = 0$.

The inference tasks here is two fold: Parameter estimation and multiple imputation. We first focus on the performance of the parameter estimation for the fixed-effect regression coefficient $\boldsymbol{\beta}$ in the presence of the potential missing values in the response matrix $\mathbf{Y}$. To this end, we implement the proposed variational inference algorithm in Section 3, together with the classical lasso method (Tibshirani, 1996) implemented in the `glmnet` package, and the `pan` package (Zhao and Schafer, 2013) for comparison. The entire experiment is repeated for 1000 Monte Carlo replicates. For each replicate of the synthetic dataset, we compute three types of estimation error for the estimate $\widehat{\boldsymbol{\beta}}$ obtained from the three methods: The 2-norm error $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 = [\sum_k(\widehat{\beta}_k - \beta_k)^2]^{1/2}$, the 1-norm error $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 = \sum_k|\widehat{\beta}_k - \beta_k|$, and the infinity-norm error $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_\infty = \max_k|\widehat{\beta}_k - \beta_k|$. Here $\widehat{\boldsymbol{\beta}}$ represents a generic estimate for $\boldsymbol{\beta}$. For the variational inference method, $\widehat{\boldsymbol{\beta}}$ is taken to be the variational posterior mean; For the pan package, we take $\widehat{\boldsymbol{\beta}}$ to be the posterior mean. Below, Figure 2 visualizes the boxplots of the three types of the estimation errors across the 1000 Monte Carlo replicates for the three methods. We can see that in terms of the 2-norm errors and the 1-norm errors, the variational posterior mean is significantly smaller than the other two competitors. In terms of the infinity-norm errors, the variational posterior mean is slightly better than the `pan` package, and both are also significantly better than the lasso estimate. We also remark that both the variational inference method and the `pan` package provide

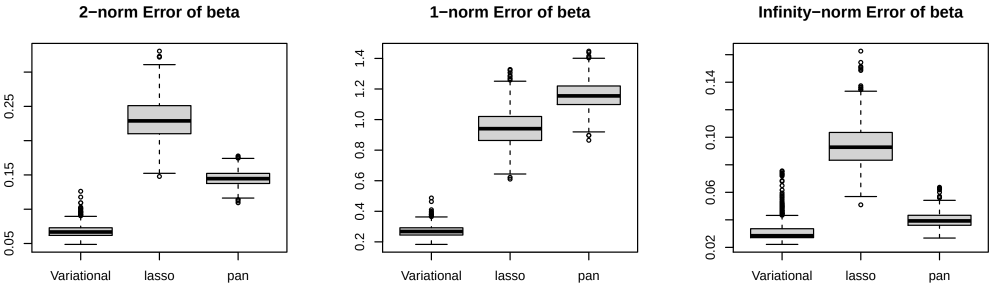

Figure 2: The boxplots of the 2-norm errors, the 1-norm errors, and the infinity-norm errors for estimating $\boldsymbol{\beta}$ using the variational inference method, the lasso, and the `pan` package, for the simulated example in Section 5.1.

natural environments for dealing with missing responses and are able to perform multiple imputation. This part of the analysis demonstrates that advantage of the proposed method in terms of the parameter estimation in the presence of missing responses and sparsity. In addition, we also provide the visualization of the point estimates using the variational inference method, the lasso, and the pan package in a randomly selected replicate in Figure 3 below. From the perspective of a single synthetic dataset, the performance of the lasso is similar to the variational inference method, but the panpackage provides a worse estimate because the sparsity structure of $\boldsymbol{\beta}$ is not captured.

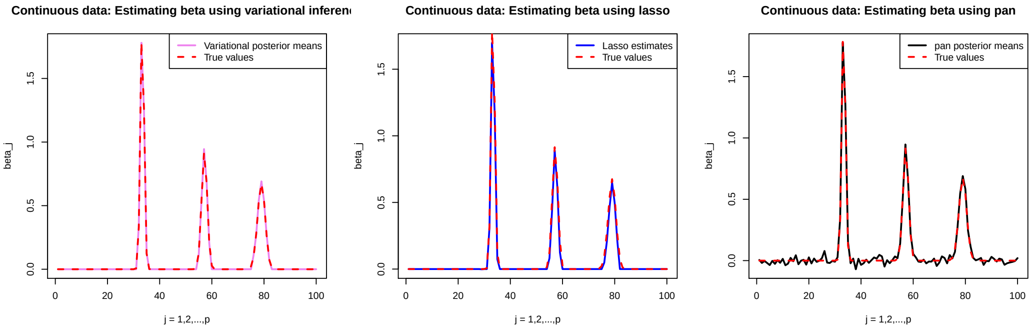

Figure 3: Visualization of the estimation for $\boldsymbol{\beta}$ using the variational inference method, the lasso, and the pan package, for the simulated example in Section 5.1.

We next investigate the performance of the multiple imputation strategy elaborated on Section 4.1. In preparation for doing so, we develop another layer of the conditional model using the response matrix $\mathbf{Y}$ as the covariate. Specifically, we consider the following (conditional) linear mixed-effect model:

$$u_{ij} = \theta_0 + \theta_1 y_{ij} + v_i + e_{ij}, \quad v_1, \ldots, v_m \overset{\text{i.i.d.}}{\sim} \text{N}(0,1), \quad e_{ij} \overset{\text{i.i.d.}}{\sim} \text{N}(0, 0.3^2),$$

$i = 1, \ldots, m$, $j = 1, \ldots, n$, where we set the fixed-effect coefficient as $\theta_0 = -2$ and $\theta_1 = 4$. Here, the matrix $\mathbf{Y}$ contains missing values but the response variable u_{ij} are generated using the complete data before $\mathbf{Y}$ is contaminated by NA's. The aforementioned 1000 Monte Carlo replicates are applied here to generate the corresponding second layer responses u_{ij}'s, $i = 1, \ldots, m, j = 1, \ldots, n$. For the imputation methods, we implement the proposed SHRIMP strategy in Section 4.1, together with the pan package, for comparison. The number of imputation is set to $M = 5$. As the focus here is to evaluate the performance of the multiple imputation methods, we choose to use the plain-vanilla lm function in R to perform the regression analysis for θ_0 and θ_1. After obtaining the estimates for θ_0 and θ_1 using the lm function in R, we apply the Rubin's rule for combined analysis (Rubin, 1987). Figure 4 below demonstrates the boxplots of the point estimates for θ_0 and θ_1 based on the variational Bayes multiple imputation (VBMI) method developed in Section 4.1, and the pan package imputation estimates. We can see that in terms of the point estimates, the two methods provide similar peformance. Furthermore, Figure 5 presents the widths of the confidence interval for θ_0 and θ_1 computed using the Rubin's combined analysis across the 1000 Monte Carlo replicates. The performance here is also similar, with the pan package producing some slightly wider confidence intervals. However, the key difference between the two methods is in the coverage probability of the confidence interval, which are tabulated in Table 1 below. The covarage probabilities are comparatively smaller than the nominal coverage probability 95% due to the model misspecification. Still, the VBMI method outperforms the pan package with higher coverage probabilities for both θ_0 and θ_1, demonstrating the power of the proposed method.

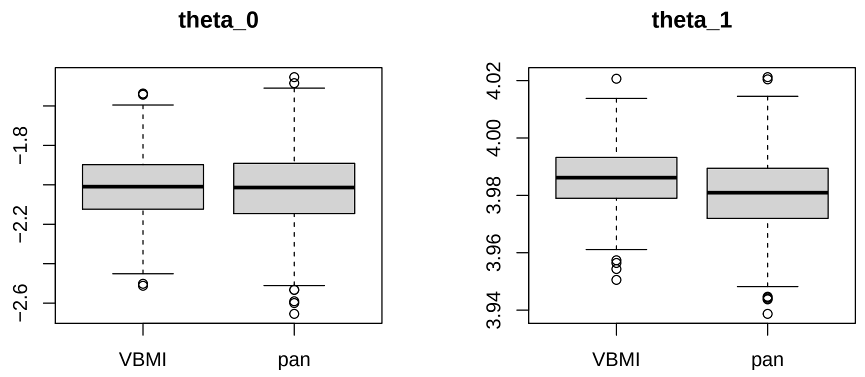

Figure 4: Visualization of the boxplots of θ_0 and θ_1 using the variational Bayes multiple imputation (VBMI) method and the pan package, for the simulated example in Section 5.1.

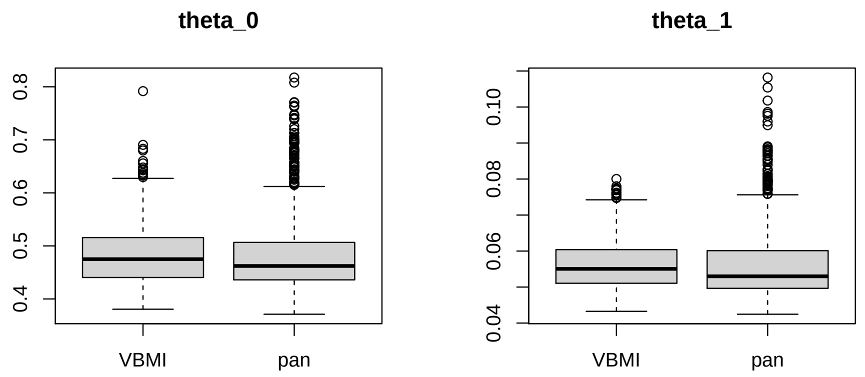

Figure 5: Visualization of the boxplots of the confidence interval widths for θ_0 and θ_1 using the variational Bayes multiple imputation (VBMI) method and the pan package, for the simulated example in Section 5.1.

Table 1: Coverage probability of the confidence intervals for θ_0 and θ_1 using the variational Bayes multiple imputation (VBMI) method and the pan package, for the simulated example in Section 5.1.

θ_j	VBMI coverage probability	pan coverage probability
θ_0	**86.6%**	82.9%
θ_1	**89.6%**	75.3%

5.2 Categorical data example

We next consider examples with categorical responses. Unlike the setup in Section 5.1, the sampling model (1) is doomed to be misspecified because the responses y_{ij}'s are necessarily continuous. Nevertheless, the misspecified model (1) is still valid as a continuous approximation to the underlying categorical model, and can be particularly useful for multiple imputation when combined with the calibration-based routine, as discussed in Section 4.2. Specialized to the synthetic data analysis in this section, we consider the following two simulation setups:

- **Setup 1: Binary responses.** The generative model of the synthetic data is a logistic mixed-effect model:

  $$y_{ij} \sim \text{Bernoulli}(p_{ij}), \quad \text{logit}(y_{ij}) = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_{ij}^{\mathrm{T}}\mathbf{b}_i, \quad \mathbf{b}_1, \ldots, \mathbf{b}_m \overset{\text{i.i.d.}}{\sim} \mathrm{N}_l(\mathbf{0}_l, \boldsymbol{\Psi}),$$

  where $i = 1, \ldots, m$, $j = 1, \ldots, n$, $m = 50$, $n = 20$, $p = 100$, $l = 3$, and $\boldsymbol{\Psi} = \mathbf{I}_l$. We follow the setup in Section 5.1 and generate the coordinates of $\mathbf{x}_{ij}$ from $\mathrm{N}(0, 3^2)$ and the coordinates of $\mathbf{z}_{ij}$ from $\mathrm{N}(0, 1)$, independently. The fixed-effect regression coefficient $\boldsymbol{\beta}$ is generated as follows: We first generate the coordinates of $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_{10}]^{\mathrm{T}}$ independently from $\mathrm{N}(0, 0.1^2)$, and then set $\beta_k = \alpha_k / \|\boldsymbol{\alpha}\|_2$ with $k = 1, \ldots, 10$, and $\beta_k = 0$ for $k = 11, \ldots, p = 100$.

- **Setup 2: Categorical responses with 5 categories.** The generative model of the synthetic data is a multiclass logistic mixed-effect model:

  $$\mathbb{P}(y_{ij} = s \mid \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \mathbf{B}) = \frac{\exp(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_s + \mathbf{z}_{ij}^{\mathrm{T}}\mathbf{b}_i)}{\sum_{t=1}^{K}\exp(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_t + \mathbf{z}_{ij}^{\mathrm{T}}\mathbf{b}_i)},$$

  where $i = 1, \ldots, m$, $j = 1, \ldots, n$, $K = 5$ is the number of categories, $m = 50$, $n = 20$, $p = 100$, $l = 3$, and $\boldsymbol{\Psi} = \mathbf{I}_l$. The setup here is similar to Setup 1 above: The coordinates of $\mathbf{x}_{ij}$ are generated from $\mathrm{N}(0, 3^2)$ and the coordinates of $\mathbf{z}_{ij}$ are simulated from $\mathrm{N}(0, 1)$, independently. The fixed-effect regression coefficients $\boldsymbol{\beta}_s = [\beta_{s1}, \ldots, \beta_{sp}]^{\mathrm{T}}$ for $s = 1, \ldots, K$ are set similar to the $\boldsymbol{\beta}$ in Setup 1 above: We first draw the coordinates of $\boldsymbol{\alpha}_s = [\alpha_{s1}, \ldots, \alpha_{s,10}]^{\mathrm{T}}$ independently from $\mathrm{N}(0, 0.1^2)$, and then set $\beta_{sk} = \alpha_{sk}/\|\boldsymbol{\alpha}_s\|_2$ with $k = 1, \ldots, 10$, and $\beta_{sk} = 0$ for $k = 11, \ldots, p = 100$. The generative process is repeated independently for each $s = 1, 2, 3, 4, 5$.

For the two simulation setups above, the response matrix $\mathbf{Y}$ is also contaminated by missing values, where the distribution of the missingness is the missing at random (MAR). Specifically, missingness mechanism is set the same as in Section 5.1 to ensure that the overall missing percentage of the response matrix $\mathbf{Y}$ is approximately 27%. The entire experiment for each setup above is repeated for 1000 Monte Carlo replicates.

As the generative model of the synthetic data differs from the working model (1) employed here (i.e., model misspecification), we turn our focus to the performance of the multiple imputation methods rather than the parameter estimation. To this end, we follow the idea in Section 5.1 and use the response matrix $\mathbf{Y}$ as the covariate to construct a second layer linear model. Specifically, given $\mathbf{Y}$, we consider

$$\mathbf{u} = \mathbf{Y}\boldsymbol{\theta} + \mathbf{e}, \quad \mathbf{e} \sim \mathrm{N}_m(\mathbf{0}_m, 5^2\mathbf{I}_m). \tag{7}$$

Here, the second layer response variable $\mathbf{u} = [u_1, \ldots, u_m]^{\mathrm{T}}$ is generated from the complete data $\mathbf{Y}$, but the observable $\mathbf{Y}$ is contaminated by the missing values. The regression coefficient $\boldsymbol{\theta}$ is gnerated from $\mathrm{N}_n(\mathbf{0}_n, 5^2\mathbf{I}_n)$. We then apply the aforementioned 1000 Monte

Carlo replicates for $\mathbf{Y}$ to further generate the second layer model (7). Since the observed second layer covariate matrix $\mathbf{Y}$ contains missing values and the generative model for $\mathbf{Y}$ can be approximated by the continuous working model (1), we apply the imputation method elaborated on Section 4.2. For comparison, we consider the `pan` package and the rounding to the nearest integer strategy based on `pan` package adopted in Yucel et al. (2011). The number of imputation datasets is set as $M = 5$, followed by the Rubin's combined analysis for inference on $\boldsymbol{\theta}$.

Below, Figure 6 and Figure 7 present the boxplots of the 2-norm errors, the 1-norm errors, and the infinity-norm errors for estimating $\boldsymbol{\theta}$ using different imputation methods under setup 1 and setup 2, respectively. It is clear that estimation error using the VBMI method is significantly smaller than the other two competitors in both setups for all three types of errors. Furthermore, Figure 8 and Figure 9 visualize the empirical coverage probabilities of the confidence intervals across the 1000 Monte Carlo replicates using the three imputation methods for setup 1 and setup 2, respectively. We can see that the coverage probabilities of the confidence intervals using the proposed VBMI method is consistent and satisfactory in comparison with the other two competitors. In particular, the `pan` package the rounding strategy is rather inconsistent and produces confidence intervals that are not reliable because the coverage probabilities are significantly lower than the nominal coverage 95%. Therefore, through empirical demonstration via the analyses of synthetic datasets, we show

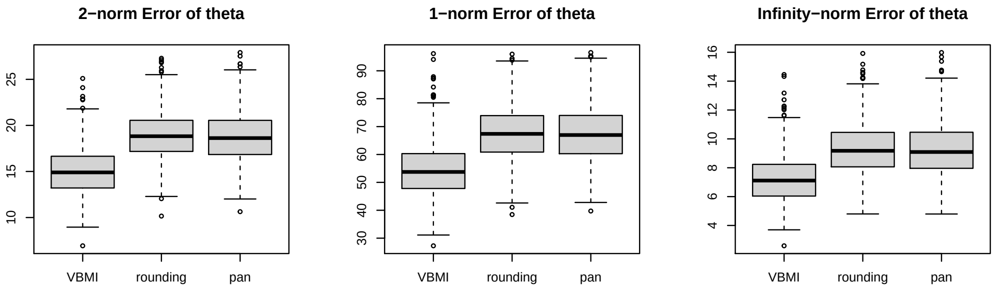

Figure 6: The boxplots of the 2-norm errors, the 1-norm errors, and the infinity-norm errors for estimating $\boldsymbol{\theta}$ using the VBMI method, the rounding strategy, and the `pan` package, for the simulated example (setup 1) in Section 5.2.

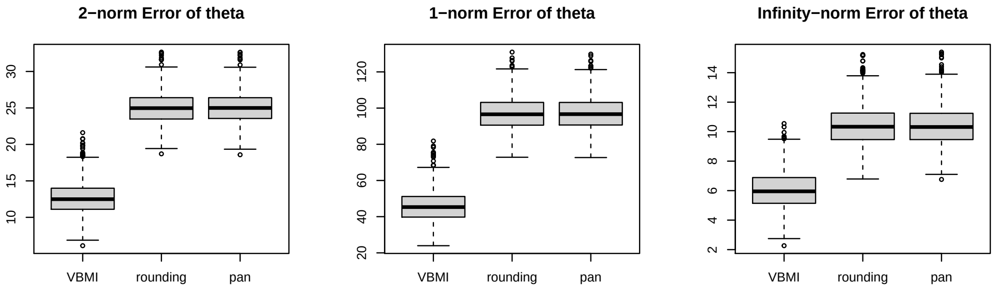

Figure 7: The boxplots of the 2-norm errors, the 1-norm errors, and the infinity-norm errors for estimating $\boldsymbol{\theta}$ using the VBMI method, the rounding strategy, and the `pan` package, for the simulated example (setup 2) in Section 5.2.

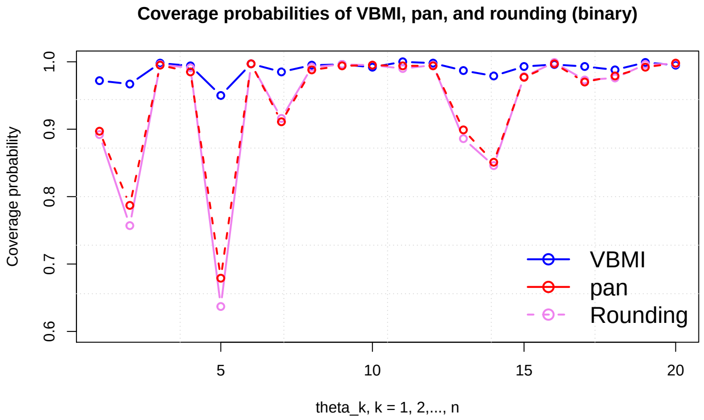

Figure 8: Coverage probabilities of the confidence intervals for $\boldsymbol{\theta}$ using the variational Bayes multiple imputation (VBMI) method, the `pan` package, and the rounding strategy, for the simulated example (setup 1) in Section 5.2.

that the proposed VBMI method is powerful and robust in terms of the imputation for categorical responses even when the model is misspecified.

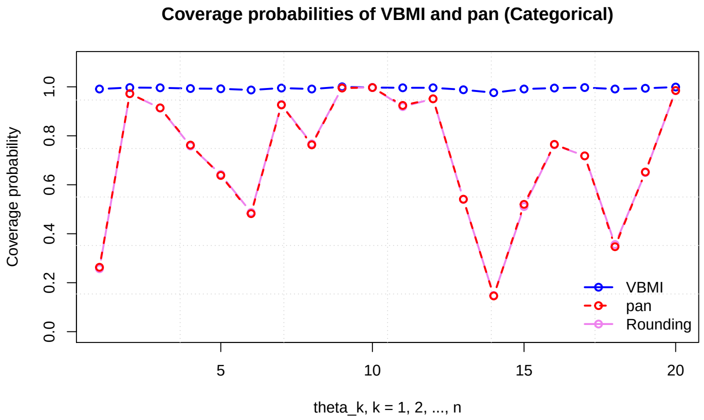

Figure 9: Coverage probabilities of the confidence intervals for $\boldsymbol{\theta}$ using the variational Bayes multiple imputation (VBMI) method, the `pan` package, and the rounding strategy, for the simulated example (setup 2) in Section 5.2.

6. Discussion

In this work, we developed a variational inference method for approximate Bayesian inference of the high-dimensional linear mixed-effect model in the presence of missing responses. The sparsity structure of the regression coefficient vector can be modeled by a spike-and-slab prior on the coordinates of the regression coefficient. The computation algorithm is easy-to-implement, efficient, and can be faster than the classical Markov chain Monte Carlo samplers because of the vectorized updating formula for $\boldsymbol{\beta}$, circumventing the need to explore the entire model selection space with exponential possibilities. The variational inference method can be further incorporated with the sequential hierarchical regression imputation strategy for continuous data and the calibration-based imputation strategy for categorical data to improve the performance of imputing missing values, which is quite flexible and powerful.

There are, however, some future extensions. The linear mixed-effect model only serves as a continuous approximation to the categorical model, and this step of the approximation can be rather sloppy in high-dimensions. The underlying reasons is that the categorical model could be potentially highly nonlinear and the working model (1) may not be rich enough to capture the nonlinearity happening inside the categorical model. It would be interesting to further expand the model structure of (1) by considering a nonparametric component, so that it can provides a better approximation to the black-box categorical data in practice. This step can be further optimized by coping with the variational inference algorithm, so that an easy-to-implement, computationally efficient, and sufficient sophisticated methodology can be developed to deal with the complex missing data scheme happening in the contemporary statistics world. We defer this topic to the future research direction.

APPENDIX: Detailed derivation of the variational inference algorithm in Section 3

In this Appendix, we provide the step-by-step derivation of the updating formulas for the variational inference algorithm outlined in Section 3. Recall that the objective function is given by

$$\mathcal{L}(q) = \mathbb{E}_q \left[\ln \frac{p(\mathbf{Y}_{(\mathrm{obs})}, \mathbf{Y}_{(\mathrm{mis})}, \mathbf{B}, \mathbf{\Theta})}{q(\mathbf{Y}_{(\mathrm{mis})}, \mathbf{B}, \mathbf{\Theta})} \right],$$

where $p(\mathbf{Y}_{(\mathrm{obs})}, \mathbf{Y}_{(\mathrm{mis})}, \mathbf{B}, \mathbf{\Theta})$ is given by the joint model (2), (3), and (4), and $q(\mathbf{Y}_{(\mathrm{mis})}, \mathbf{B}, \mathbf{\Theta})$ is the variational distribution

$$\begin{aligned} q(\mathbf{Y}_{(\mathrm{mis})}, \mathbf{B}, \mathbf{\Theta}) \mathrm{d}\mathbf{Y}_{(\mathrm{mis})} \mathrm{d}\mathbf{B} \mathrm{d}\mathbf{\Theta} = {} & q(\mathbf{Y}_{(\mathrm{mis})} \mid (\widehat{\mu}_{y_{ij}}, \widehat{\sigma}^2_{y_{ij}})_{(i,j)\in\mathcal{I}_{(\mathrm{mis})}}) \mathrm{d}\mathbf{Y}_{(\mathrm{mis})} \\ & \times q(\boldsymbol{\beta}, \gamma \mid (\theta_k, \widehat{\mu}_{\beta_k}, \widehat{\sigma}^2_{\beta_k})_{k=1}^p) \mathrm{d}\boldsymbol{\beta} \\ & \times q(\mathbf{B} \mid \widehat{\boldsymbol{\mu}}_{\mathbf{b}_1}, \widehat{\mathbf{\Psi}}_1, \dots, \widehat{\boldsymbol{\mu}}_{\mathbf{b}_m}, \widehat{\mathbf{\Psi}}_m) \mathrm{d}\mathbf{B} \\ & \times q(\sigma_e^2 \mid \hat{a}_{\sigma_e^2}, \hat{b}_{\sigma_e^2}) \mathrm{d}\sigma_e^2 q(\mathbf{\Psi}^{-1} \mid \widehat{\mathbf{\Psi}}) \mathrm{d}\mathbf{\Psi}^{-1} \\ & \times q(\mu_0 \mid \hat{\mu}_{\mu_0}, \hat{\sigma}^2_{\mu_0}) \mathrm{d}\mu_0 q(\sigma_0^2 \mid \hat{a}_{\sigma_0^2}, \hat{b}_{\sigma_0^2}) \mathrm{d}\sigma_0^2 \\ & \times q(w \mid \hat{a}_w, \hat{b}_w) \mathrm{d}w. \end{aligned}$$

The specific form of q is given in (6). The computation of the entire objection function $\mathcal{L}(q)$ is tedious and unnecessary for deriving the coordinate-ascent variational inference updating formula for each block of the variational parameters. Instead, when focusing on the derivation of a fixed block of the variational parameter, we only need to consider the likelihood involving the latent variable and its variational distribution. Below, we discuss each updating rule separately.

A. Updating $\mathbf{\Psi}$

By construction, we have

$$\arg\max_{\widehat{\mathbf{\Psi}}} \mathcal{L}(q) = \arg\max_{\widehat{\mathbf{\Psi}}} \left[p(\widehat{\mathbf{\Psi}}) + \sum_{i=1}^m \ln p(\mathbf{b}_i \mid \widehat{\mathbf{\Psi}}) \right] := \arg\max_{\widehat{\mathbf{\Psi}}} \Omega(\widehat{\mathbf{\Psi}}),$$

where

$$
\begin{aligned}
\Omega(\widehat{\mathbf{\Psi}}) &= -\frac{m}{2}\ln|\widehat{\mathbf{\Psi}}| - \frac{1}{2}\sum_{i=1}^{m}\mathbb{E}_q[\mathbf{b}_i^{\mathrm{T}}\widehat{\mathbf{\Psi}}^{-1}\mathbf{b}_i] + \frac{\nu-l-1}{2}\ln|\widehat{\mathbf{\Psi}}|^{-1} - \frac{1}{2}\mathrm{tr}(\widehat{\mathbf{\Psi}}^{-1}\mathbf{V}^{-1}) \\
&= \frac{m+\nu-l-1}{2}\ln|\mathbf{\Psi}|^{-1} - \frac{1}{2}\mathrm{tr}(\mathbf{\Psi}^{-1}(\mathbf{V}^{-1} + \sum_i \mathbb{E}_q[\mathbf{b}_i\mathbf{b}_i^{\mathrm{T}}])).
\end{aligned}
$$

Maximizing $\Omega(\widehat{\mathbf{\Psi}})$ over the positive definite cone in $\mathbb{R}^{l\times l}$ yields that

$$
\begin{aligned}
\widehat{\mathbf{\Psi}} &= \frac{1}{m+\nu-l-1}\left(\sum_{i=1}^{m}\mathbb{E}_q[\mathbf{b}_i\mathbf{b}_i^{\mathrm{T}}] + \mathbf{V}^{-1}\right) \\
&= \frac{1}{m+\nu-l-1}\left\{\sum_{i=1}^{m}(\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}^{\mathrm{T}} + \widehat{\mathbf{\Psi}}_{\mathbf{b}_i}) + \mathbf{V}^{-1}\right\}.
\end{aligned}
$$

B. Updating variational parameters for $\mathbf{Y}_{(\mathrm{mis})}$

We first fix the index (i,j). This reduces to solving the problem

$$
\underset{(\widehat{\mu}_{y_{ij}},\widehat{\sigma}^2_{y_{ij}})}{\arg\max}\,\mathcal{L}(q) = \underset{(\widehat{\mu}_{y_{ij}},\widehat{\sigma}^2_{y_{ij}})}{\arg\max}\,\mathbb{E}_q\left[\ln\frac{p(y_{ij}\mid\boldsymbol{\beta},\mathbf{b}_i,\sigma_e^2)}{q(y_{ij}\mid\widehat{\mu}_{y_{ij}},\widehat{\sigma}^2_{y_{ij}})}\right].
$$

Observe that

$$
\begin{aligned}
\Omega(\widehat{\mu}_{y_{ij}},\widehat{\sigma}^2_{y_{ij}}) &:= \mathbb{E}_q\left[\ln\frac{p(y_{ij}\mid\boldsymbol{\beta},\mathbf{b}_i,\sigma_e^2)}{q(y_{ij}\mid\widehat{\mu}_{y_{ij}},\widehat{\sigma}^2_{y_{ij}})}\right] \\
&= -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\mathbb{E}_q\left[\ln(\sigma_e^2)\right] - \mathbb{E}_q\left[\frac{1}{2\sigma_e^2}\right]\mathbb{E}_q\left[(y_{ij} - \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} - \mathbf{z}_{ij}^{\mathrm{T}}\mathbf{b}_i)^2\right] \\
&\quad + \frac{1}{2}\ln(2\pi) + \frac{1}{2}\ln(\widehat{\sigma}^2_{y_{ij}}) + \frac{1}{2\widehat{\sigma}^2_{y_{ij}}}\mathbb{E}_q\left[(y_{ij} - \widehat{\mu}_{y_{ij}})^2\right] \\
&= -\frac{\widehat{a}_{\sigma_e^2}}{2\widehat{b}_{\sigma_e^2}}\left[\mathbb{E}_q(y_{ij}^2) - \mathbb{E}_q(y_{ij})\mathbb{E}_q(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_{ij}^{\mathrm{T}}\mathbf{b}_i)\right] + \frac{1}{2}\ln(\widehat{\sigma}^2_{y_{ij}}) + \text{constant} \\
&= -\frac{\widehat{a}_{\sigma_e^2}}{2\widehat{b}_{\sigma_e^2}}\left[\widehat{\mu}^2_{y_{ij}} + \widehat{\sigma}^2_{y_{ij}} - \widehat{\mu}^2_{y_{ij}}\left(\sum_{k=1}^{p}x_{ijk}\theta_k\widehat{\mu}_{\beta_k} + \mathbf{z}_{ij}^{\mathrm{T}}\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}\right)\right] \\
&\quad + \frac{1}{2}\ln(\widehat{\sigma}^2_{y_{ij}}) + \text{constant.}
\end{aligned}
$$

Now we proceed to solve

$$
\begin{aligned}
\frac{\partial}{\partial\widehat{\mu}_{y_{ij}}}\Omega(\widehat{\mu}_{y_{ij}},\widehat{\sigma}^2_{y_{ij}}) = 0 \quad &\Longrightarrow \quad \widehat{\mu}_{y_{ij}} = \sum_{k=1}^{p}x_{ijk}\theta_k\widehat{\mu}_{\beta_k} + \mathbf{z}_{ij}^{\mathrm{T}}\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}, \\
\frac{\partial}{\partial\widehat{\sigma}^2_{y_{ij}}}\Omega(\widehat{\mu}_{y_{ij}},\widehat{\sigma}^2_{y_{ij}}) = 0 \quad &\Longrightarrow \quad \widehat{\sigma}^2_{y_{ij}} = \frac{\widehat{a}^2_{\sigma_e}}{\widehat{b}^2_{\sigma_e}}.
\end{aligned}
$$

C. Updating variational parameters for β

This part is the most challenging part as there does not exist a closed-form updating formula when we optimize with regard to $\boldsymbol{\theta} = [\theta_1,\ldots,\theta_p]^{\mathrm{T}}$. Denote $\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}} = [\widehat{\mu}_{\beta_1},\ldots,\widehat{\mu}_{\beta_p}]^{\mathrm{T}}$ and

$\widehat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}} = [\widehat{\sigma}^2_{\beta_1}, \ldots, \widehat{\sigma}^2_{\beta_p}]^{\mathrm{T}}$. First write

$$\begin{aligned}\Omega(\boldsymbol{\theta}, \widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}, \widehat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}}) &= \mathbb{E}_q\left[\ln \frac{p(\mathbf{Y} \mid \boldsymbol{\beta}, \mathbf{B}, \sigma_0^2) p(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \mathbf{w}, \mu_0, \sigma_0)\mathrm{d}\boldsymbol{\beta}}{q(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \boldsymbol{\theta}, \widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}, \widehat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}})\mathrm{d}\boldsymbol{\beta}}\right] \\ &= \mathbb{E}_q\left[\ln p(\mathbf{Y} \mid \boldsymbol{\beta}, \mathbf{B}, \sigma_0^2)\right] + \mathbb{E}_q\left[\ln \frac{p(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \mathbf{w}, \mu_0, \sigma_0)\mathrm{d}\boldsymbol{\beta}}{q(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \boldsymbol{\theta}, \widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}, \widehat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}})\mathrm{d}\boldsymbol{\beta}}\right].\end{aligned}$$

We first consider the expected value of the log-likelihood. Write

$$\mathbb{E}_q\left[\ln p(\mathbf{Y} \mid \boldsymbol{\beta}, \mathbf{B}, \sigma_0^2)\right] = -\frac{1}{2}\left(\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}}\right)\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbb{E}_q[(y_{ij} - \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} - \mathbf{z}_{ij}^{\mathrm{T}}\mathbf{b}_i)^2] + \text{constant}.$$

The keystone computation is the quadratic form $\mathbb{E}_q[(y_{ij} - \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} - \mathbf{z}_{ij}^{\mathrm{T}}\mathbf{b}_i)^2]$. When we focus on the parameter $\boldsymbol{\theta}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}}$, we have

$$\begin{aligned}\mathbb{E}_q[(y_{ij} - \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} - \mathbf{z}_{ij}^{\mathrm{T}}\mathbf{b}_i)^2] &= \mathbb{E}_q(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{ij}\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}) - 2\mathbb{E}_q(\boldsymbol{\beta})^{\mathrm{T}}\mathbf{x}_{ij}\mathbb{E}_q(y_{ij} - \mathbf{z}_{ij}^{\mathrm{T}}\mathbf{b}_i) + \text{constant} \\ &= \operatorname{tr}\left\{\mathbf{x}_{ij}\mathbf{x}_{ij}^{\mathrm{T}}\mathbb{E}_q(\boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}})\right\} - 2\mathbb{E}_q(\boldsymbol{\beta})^{\mathrm{T}}\mathbf{x}_{ij}\mathbb{E}_q(y_{ij} - \mathbf{z}_{ij}^{\mathrm{T}}\mathbf{b}_i) \\ &\quad + \text{constant}.\end{aligned}$$

Let $\boldsymbol{\Theta} = \operatorname{diag}(\boldsymbol{\theta})$ and $\mathbf{U} = \operatorname{diag}(\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}})$. By definition of $q(\boldsymbol{\beta} \mid \boldsymbol{\theta}) = \sum_{\boldsymbol{\gamma}} q(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \boldsymbol{\theta})$, we have

$$\mathbb{E}_q(\boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}}) = \boldsymbol{\Theta}(\mathbf{I} - \boldsymbol{\Theta})\mathbf{U} + \boldsymbol{\Theta}\operatorname{diag}(\widehat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}}) + \mathbf{U}\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}.$$

It follows that

$$\begin{aligned}\operatorname{tr}\left\{\mathbf{x}_{ij}\mathbf{x}_{ij}^{\mathrm{T}}\mathbb{E}_q(\boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}})\right\} &= \left(\sum_{k=1}^{p} x_{ijk}\theta_k\widehat{\mu}_{\boldsymbol{\beta}_k}\right)^2 + \sum_{k=1}^{p} x_{ijk}^2[\theta_k(1-\theta_k)\widehat{\mu}^2_{\beta_k} + \theta_k\widehat{\sigma}^2_{\beta_k}] \\ &= \left(\sum_{k=1}^{p} x_{ijk}\theta_k\widehat{\mu}_{\boldsymbol{\beta}_k}\right)^2 + \sum_{k=1}^{p}\theta_k x_{ijk}^2(\widehat{\mu}^2_{\beta_k} + \widehat{\sigma}^2_{\beta_k}) - \sum_{k=1}^{p}\theta_k^2 x_{ijk}^2\widehat{\mu}^2_{\beta_k} \\ &= (\mathbf{x}_{ij}^{\mathrm{T}}\mathbf{U}\boldsymbol{\theta})^2 - \boldsymbol{\theta}^{\mathrm{T}}\mathbf{U}\operatorname{diag}(\mathbf{x}_{ij}^2)\mathbf{U}\boldsymbol{\theta} + (\widehat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}} + \widehat{\boldsymbol{\mu}}^2_{\boldsymbol{\beta}})^{\mathrm{T}}\operatorname{diag}(\mathbf{x}_{ij}^2)\boldsymbol{\theta}.\end{aligned}$$

Therefore, we obtain:

$$\begin{aligned}\mathbb{E}_q[(y_{ij} - \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} - \mathbf{z}_{ij}^{\mathrm{T}}\mathbf{b}_i)^2] &= (\mathbf{x}_{ij}^{\mathrm{T}}\mathbf{U}\boldsymbol{\theta})^2 - \boldsymbol{\theta}^{\mathrm{T}}\mathbf{U}\operatorname{diag}(\mathbf{x}_{ij}^2)\mathbf{U}\boldsymbol{\theta} + (\widehat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}} + \widehat{\boldsymbol{\mu}}^2_{\boldsymbol{\beta}})^{\mathrm{T}}\operatorname{diag}(\mathbf{x}_{ij}^2)\boldsymbol{\theta} \\ &\quad - 2\boldsymbol{\theta}^{\mathrm{T}}\mathbf{U}\mathbf{x}_{ij}(\langle y_{ij}\rangle - \mathbf{z}_{ij}^{\mathrm{T}}\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}).\end{aligned}$$

Here we set

$$\widehat{y}_{ij} = \begin{cases} y_{ij}, & \text{if } (i,j) \notin \mathcal{I}_{(\text{mis})}, \\ \widehat{\mu}_{y_{ij}}, & \text{if } (i,j) \in \mathcal{I}_{(\text{mis})}. \end{cases}$$

Hence, we obtain

$$\begin{aligned}\mathbb{E}_q\left[\ln p(\mathbf{Y} \mid \boldsymbol{\beta}, \mathbf{B}, \sigma_0^2)\right] &= -\frac{1}{2}\left(\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}}\right)\sum_{i=1}^{m}\sum_{j=1}^{n}(\mathbf{x}_{ij}^{\mathrm{T}}\mathbf{U}\boldsymbol{\theta})^2 \\ &\quad + \frac{1}{2}\left(\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}}\right)\sum_{i=1}^{m}\sum_{j=1}^{n}\boldsymbol{\theta}^{\mathrm{T}}\mathbf{U}\operatorname{diag}(\mathbf{x}_{ij}^2)\mathbf{U}\boldsymbol{\theta} \\ &\quad - \frac{1}{2}\left(\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}}\right)\sum_{i=1}^{m}\sum_{j=1}^{n}(\widehat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}} + \widehat{\boldsymbol{\mu}}^2_{\boldsymbol{\beta}})^{\mathrm{T}}\operatorname{diag}(\mathbf{x}_{ij}^2)\boldsymbol{\theta} \\ &\quad + \left(\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}}\right)\sum_{i=1}^{m}\sum_{j=1}^{n}\boldsymbol{\theta}^{\mathrm{T}}\mathbf{U}\mathbf{x}_{ij}(\langle y_{ij}\rangle - \mathbf{z}_{ij}^{\mathrm{T}}\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}) + \text{constant}.\end{aligned} \tag{8}$$

We then compute the second term in $\Omega(\boldsymbol{\theta}, \widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}, \widehat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}})$:

$$
\begin{aligned}
&\mathbb{E}_q\left[\ln \frac{p(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \mathbf{w}, \mu_0, \sigma_0^2)\mathrm{d}\boldsymbol{\beta}}{q(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \boldsymbol{\theta}, \widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}, \widehat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}})\mathrm{d}\boldsymbol{\beta}}\right] \\
&\quad = \sum_{k=1}^{p} \theta_k \left\{[\psi(\widehat{a}_w) - \psi(\widehat{a}_w + \widehat{b}_w)] - \frac{1}{2}\ln(2\pi) + \frac{1}{2}[\psi(\widehat{a}_{\sigma_0^2}) - \ln(\widehat{b}_{\sigma_0^2})]\right\} \\
&\qquad - \sum_{k=1}^{p} \mathbb{E}_q\left[\frac{(\beta_k - \mu_0)^2}{2\sigma_0^2}\gamma_k\right] + \sum_{k=1}^{p}(1-\theta_k)[\psi(\widehat{b}_w) - \psi(\widehat{a}_w + \widehat{b}_w)] \\
&\qquad - \sum_{k=1}^{p} \theta_k \left[\ln\theta_k - \frac{1}{2}\ln(2\pi) + \frac{1}{2}\ln\frac{1}{\widehat{\sigma}^2_{\beta_k}}\right] + \sum_{k=1}^{p} \mathbb{E}_q\left[\frac{(\beta_k - \widehat{\mu}_{\beta_k})^2}{2\widehat{\sigma}^2_{\beta_k}}\gamma_k\right] \\
&\qquad - \sum_{k=1}^{p}(1-\theta_k)\ln(1-\theta_k) \\
&\quad = \sum_{k=1}^{p} \theta_k \left\{[\psi(\widehat{a}_w) - \psi(\widehat{a}_w + \widehat{b}_w)] - \frac{1}{2}\ln(2\pi) + \frac{1}{2}[\psi(\widehat{a}_{\sigma_0^2}) - \ln(\widehat{b}_{\sigma_0^2})]\right\} \\
&\qquad - \sum_{k=1}^{p} \frac{\widehat{a}_{\sigma_0^2}}{2\widehat{b}_{\sigma_0^2}}\theta_k[(\widehat{\mu}_{\beta_k} - \widehat{\mu}_{\mu_0})^2 + \widehat{\sigma}^2_{\mu_0} + \widehat{\sigma}^2_{\beta_k}] + \sum_{k=1}^{p}(1-\theta_k)[\psi(\widehat{b}_w) - \psi(\widehat{a}_w + \widehat{b}_w)] \\
&\qquad - \sum_{k=1}^{p} \theta_k \left[\ln\theta_k - \frac{1}{2}\ln(2\pi) + \frac{1}{2}\ln\frac{1}{\widehat{\sigma}^2_{\beta_k}}\right] + \sum_{k=1}^{p}\frac{\theta_k}{2} - \sum_{k=1}^{p}(1-\theta_k)\ln(1-\theta_k)
\end{aligned}
\tag{9}
$$

♦ We first optimize over $(\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}, \widehat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}})$. This step is relatively straightforward because of the closed-form solution to the stationary point. It is straightforward to obtain the derivative of $\widehat{\mu}_{\beta_k}$ using (8) and (9):

$$
\begin{aligned}
\frac{\partial\Omega}{\partial\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}} = &-\frac{1}{2}\left(\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}}\right)\sum_{i=1}^{m}\sum_{j=1}^{n} 2\boldsymbol{\Theta}\mathbf{x}_{ij}\mathbf{x}_{ij}\boldsymbol{\Theta}\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}} \\
&+\frac{1}{2}\left(\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}}\right)\sum_{i=1}^{m}\sum_{j=1}^{n} 2\boldsymbol{\Theta}\mathrm{diag}(\mathbf{x}_{ij}^2)\boldsymbol{\Theta}\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}} \\
&-\frac{1}{2}\left(\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}}\right)\sum_{i=1}^{m}\sum_{j=1}^{n} 2\boldsymbol{\Theta}\mathrm{diag}(\mathbf{x}_{ij}^2)\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}} \\
&+\left(\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}}\right)\sum_{i=1}^{m}\sum_{j=1}^{n}(\langle y_{ij}\rangle - \mathbf{z}_{ij}^{\mathrm{T}}\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i})\boldsymbol{\Theta}\mathbf{x}_{ij}.
\end{aligned}
$$

Setting the gradient to $\mathbf{0}$ yields

$$
\begin{aligned}
\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}} = &\left\{\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}}\sum_{i=1}^{m}\sum_{j=1}^{n}[\mathbf{x}_{ij}\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\Theta} + \mathrm{diag}(\mathbf{x}_{ij}^2)(\mathbf{I} - \boldsymbol{\Theta})] + \frac{\widehat{a}_{\sigma_0^2}}{\widehat{b}_{\sigma_0^2}}\mathbf{I}_p\right\}^{-1} \\
&\times\left[\frac{\widehat{a}_{\sigma_0^2}}{\widehat{b}_{\sigma_0^2}}\mu_0\mathbf{1}_p + \frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}}\sum_{i=1}^{m}\sum_{j=1}^{n}(\widehat{y}_{ij} - \mathbf{z}_{ij}^{\mathrm{T}}\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i})\mathbf{x}_{ij}\right].
\end{aligned}
$$

We also take the gradient with regard to $\boldsymbol{\sigma}^2_{\boldsymbol{\beta}}$ to obtain

$$\frac{\partial\Omega}{\partial\widehat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}}} = -\frac{1}{2}\left(\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}}\right)\sum_{i=1}^{m}\sum_{j=1}^{n}\operatorname{diag}(\mathbf{x}_{ij}^2)\boldsymbol{\theta} - \frac{1}{2}\boldsymbol{\theta}\circ\left(\frac{1}{\widehat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}}}\right) - \frac{\widehat{a}_{\sigma_0^2}}{2\widehat{b}_{\sigma_0^2}}\boldsymbol{\theta}\circ\mathbf{1}_p.$$

Therefore, setting the gradient to $\mathbf{0}$ yields

$$\widehat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}} = \left[\left(\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}}\right)\sum_{i=1}^{m}\sum_{j=1}^{n}\operatorname{diag}(\mathbf{x}_{ij}^2) + \frac{\widehat{a}_{\sigma_0^2}}{2\widehat{b}_{\sigma_0^2}}\mathbf{1}_p\right]^{-1}$$

- We next consider optimizing over $\boldsymbol{\theta}$, which is comparably more challenging. In this case, we can view $\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}, \widehat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}}$ as constants. Invoking (8) and (9), we write

$$\begin{aligned}
\Omega_{\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}},\widehat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}}}(\boldsymbol{\theta}) &:= \Omega(\boldsymbol{\theta}, \widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}, \widehat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}}) \\
&= \sum_{k=1}^{p}\theta_k\left\{[\psi(\widehat{a}_w) - \psi(\widehat{a}_w + \widehat{b}_w)] + \frac{1}{2}[\psi(\widehat{a}_{\sigma_0^2}) - \ln(\widehat{b}_{\sigma_0^2})] + \frac{1}{2}\ln(\widehat{\sigma}^2_{\beta_k})\right\} \\
&\quad + \sum_{k=1}^{p}\theta_k\left\{\frac{1}{2} - \frac{\widehat{a}_{\sigma_0^2}}{2\widehat{b}_{\sigma_0^2}}\theta_k[(\widehat{\mu}_{\beta_k} - \widehat{\mu}_{\mu_0})^2 + \widehat{\sigma}^2_{\mu_0} + \widehat{\sigma}^2_{\beta_k}]\right\} - \sum_{k=1}^{p}\theta_k\ln\theta_k \\
&\quad + \sum_{k=1}^{p}(1-\theta_k)[\psi(\widehat{b}_w) - \psi(\widehat{a}_w + \widehat{b}_w)] - \sum_{k=1}^{p}(1-\theta_k)\ln(1-\theta_k) \\
&\quad - \frac{1}{2}\left(\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}}\right)\boldsymbol{\theta}^{\mathrm{T}}\mathbf{U}\left\{\sum_{i=1}^{m}\sum_{j=1}^{n}[\mathbf{x}_{ij}\mathbf{x}_{ij}^{\mathrm{T}} - \operatorname{diag}(\mathbf{x}_{ij}^2)]\right\}\mathbf{U}\boldsymbol{\theta} \\
&\quad - \frac{1}{2}\left(\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}}\right)\sum_{i=1}^{m}\sum_{j=1}^{n}(\widehat{\boldsymbol{\mu}}^2_{\boldsymbol{\beta}} + \widehat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}})^{\mathrm{T}}\operatorname{diag}(\mathbf{x}_{ij}^2)\boldsymbol{\theta} \\
&\quad + \left(\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}}\right)\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{x}_{ij}(\langle y_{ij}\rangle - \mathbf{z}_{ij}^{\mathrm{T}}\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i})\mathbf{x}_{ij}^{\mathrm{T}}\mathbf{U}\boldsymbol{\theta}.
\end{aligned}$$

The technical challenge in optimizing the function above over $\boldsymbol{\theta}$ is that the stationary point cannot be explicitly computed in a closed-form formula. This is due to the quadratic function

$$g(\boldsymbol{\theta}) = \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\Delta}\mathbf{U}, \quad \text{where } \boldsymbol{\Delta} = \mathbf{U}\left\{\sum_{i=1}^{m}\sum_{j=1}^{n}[\mathbf{x}_{ij}\mathbf{x}_{ij}^{\mathrm{T}} - \operatorname{diag}(\mathbf{x}_{ij}\circ\mathbf{x}_{ij})]\right\}\mathbf{U}.$$

We borrow the idea of Huang et al. (2016) and consider the following linear approximation of the quadratic function g at the last updated value $\boldsymbol{\theta}^{(\text{old})}$:

$$\begin{aligned}
g(\boldsymbol{\theta}) &\approx g(\boldsymbol{\theta}^{(\text{old})}) + 2(\boldsymbol{\theta}^{(\text{old})})^{\mathrm{T}}\boldsymbol{\Delta}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(\text{old})}) \\
&= 2(\boldsymbol{\theta}^{(\text{old})})^{\mathrm{T}}\boldsymbol{\Delta}\boldsymbol{\theta} + g(\boldsymbol{\theta}^{(\text{old})}) - 2(\boldsymbol{\theta}^{(\text{old})})^{\mathrm{T}}\boldsymbol{\Delta}\boldsymbol{\theta}^{(\text{old})}.
\end{aligned}$$

Namely, we can approximate $\Omega_{\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}},\widehat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}}}(\boldsymbol{\theta})$ by

$$\Omega_{\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}},\widehat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}}}(\boldsymbol{\theta}) \approx \sum_{k=1}^{p}\theta_k\left\{[\psi(\widehat{a}_w) - \psi(\widehat{a}_w + \widehat{b}_w)] + \frac{1}{2}[\psi(\widehat{a}_{\sigma_0^2}) - \ln(\widehat{b}_{\sigma_0^2})] + \frac{1}{2}\ln(\widehat{\sigma}^2_{\beta_k})\right\}$$

$$
\begin{aligned}
&+\sum_{k=1}^{p}\theta_k\left\{\frac{1}{2}-\frac{\widehat{a}_{\sigma_0^2}}{2\widehat{b}_{\sigma_0^2}}\theta_k[(\widehat{\mu}_{\beta_k}-\widehat{\mu}_{\mu_0})^2+\widehat{\sigma}^2_{\mu_0}+\widehat{\sigma}^2_{\beta_k}]\right\}-\sum_{k=1}^{p}\theta_k\ln\theta_k\\
&+\sum_{k=1}^{p}(1-\theta_k)[\psi(\widehat{b}_w)-\psi(\widehat{a}_w+\widehat{b}_w)]-\sum_{k=1}^{p}(1-\theta_k)\ln(1-\theta_k)\\
&-\left(\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}}\right)(\boldsymbol{\theta}^{(\mathrm{old})})^{\mathrm{T}}\mathbf{U}\left\{\sum_{i=1}^{m}\sum_{j=1}^{n}[\mathbf{x}_{ij}\mathbf{x}_{ij}^{\mathrm{T}}-\mathrm{diag}(\mathbf{x}_{ij}\circ\mathbf{x}_{ij})]\right\}\mathbf{U}\boldsymbol{\theta}\\
&-\frac{1}{2}\left(\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}}\right)\sum_{i=1}^{m}\sum_{j=1}^{n}(\widehat{\boldsymbol{\mu}}^2_{\boldsymbol{\beta}}+\widehat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}})^{\mathrm{T}}\mathrm{diag}(\mathbf{x}_{ij}\circ\mathbf{x}_{ij})\boldsymbol{\theta}\\
&+\left(\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}}\right)\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{x}_{ij}(\langle y_{ij}\rangle-\mathbf{z}_{ij}^{\mathrm{T}}\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i})\mathbf{x}_{ij}^{\mathrm{T}}\mathbf{U}\boldsymbol{\theta}+\text{constant.}
\end{aligned}
$$

Here $\boldsymbol{\theta}^{(\mathrm{old})}$ is the last iterate of the $\boldsymbol{\theta}$ value that can be treated as a constant. Let $\mathbf{1}_p=[1,\ldots,1]^{\mathrm{T}}\in\mathbb{R}^p$ be the p-dimensional vector of all ones. Hence, we take the gradient to obtain

$$
\begin{aligned}
\frac{\partial}{\partial\boldsymbol{\theta}}\Omega_{\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}},\widehat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}}}(\boldsymbol{\theta})\approx&\left\{[\psi(\widehat{a}_w)-\psi(\widehat{a}_w+\widehat{b}_w)]+\frac{1}{2}[\psi(\widehat{a}_{\sigma_0^2})-\ln(\widehat{b}_{\sigma_0^2})]\right\}\mathbf{1}_p+\frac{1}{2}\ln(\widehat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}})\\
&+\frac{1}{2}\mathbf{1}_p-\frac{\widehat{a}_{\sigma_0^2}}{2\widehat{b}_{\sigma_0^2}}[(\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}-\widehat{\mu}_{\mu_0}\mathbf{1}_p)^2+\widehat{\sigma}^2_{\mu_0}\mathbf{1}_p+\widehat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}}]-\mathbf{1}_p-\mathrm{logit}(\boldsymbol{\theta})\\
&-[\psi(\widehat{b}_w)-\psi(\widehat{a}_w+\widehat{b}_w)]\mathbf{1}_p+\mathbf{1}_p\\
&-\frac{1}{2}\left(\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}}\right)\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{x}^2_{ij}\circ(\widehat{\boldsymbol{\mu}}^2_{\boldsymbol{\beta}}+\widehat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}})\\
&+\left(\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}}\right)\sum_{i=1}^{m}\sum_{j=1}^{n}(\widehat{y}_{ij}-\mathbf{z}_{ij}^{\mathrm{T}}\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i})(\mathbf{x}_{ij}\circ\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}})\\
&-\left(\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}}\right)\mathbf{U}\left\{\sum_{i=1}^{m}\sum_{j=1}^{n}[\mathbf{x}_{ij}\mathbf{x}_{ij}^{\mathrm{T}}-\mathrm{diag}(\mathbf{x}_{ij}\circ\mathbf{x}_{ij})]\right\}\mathbf{U}\boldsymbol{\theta}^{(\mathrm{old})}.
\end{aligned}
$$

We now focus on the last two lines of the preceeding display. We use the updating formula for $\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}$ obtained earlier to write

$$
\begin{aligned}
&\sum_{i=1}^{m}\sum_{j=1}^{n}(\widehat{y}_{ij}-\mathbf{z}_{ij}^{\mathrm{T}}\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i})\mathbf{x}_{ij}\circ\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}-\mathbf{U}\left\{\sum_{i=1}^{m}\sum_{j=1}^{n}[\mathbf{x}_{ij}\mathbf{x}_{ij}^{\mathrm{T}}-\mathrm{diag}(\mathbf{x}_{ij}\circ\mathbf{x}_{ij})]\right\}\mathbf{U}\boldsymbol{\theta}^{(\mathrm{old})}\\
&\quad=\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{U}\mathbf{x}_{ij}\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\Theta}\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}+\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{U}\mathrm{diag}(\mathbf{x}^2_{ij})(\mathbf{I}-\boldsymbol{\Theta})\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}-\frac{\widehat{a}_{\sigma_0^2}\widehat{b}_{\sigma_e^2}}{\widehat{b}_{\sigma_0^2}\widehat{a}_{\sigma_e^2}}\widehat{\mu}_{\mu_0}\mathbf{U}\mathbf{I}_p\\
&\qquad+\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{U}\mathrm{diag}(\mathbf{x}^2_{ij})\boldsymbol{\Theta}\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}-\sum_{i=1}^{m}\sum_{j=1}^{m}\mathbf{U}\mathbf{x}_{ij}\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\Theta}\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}+\frac{\widehat{b}_{\sigma_e^2}\widehat{a}_{\sigma_0^2}}{\widehat{a}_{\sigma_e^2}\widehat{b}_{\sigma_0^2}}\widehat{\boldsymbol{\mu}}^2_{\boldsymbol{\beta}}\\
&\quad=\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{U}\mathrm{diag}(\mathbf{x}^2_{ij})\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}-\frac{\widehat{a}_{\sigma_0^2}\widehat{b}_{\sigma_e^2}}{\widehat{b}_{\sigma_0^2}\widehat{a}_{\sigma_e^2}}\widehat{\mu}_{\mu_0}\mathbf{U}\mathbf{I}_p\\
&\quad=\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{x}^2_{ij}\circ\widehat{\boldsymbol{\mu}}^2_{\boldsymbol{\beta}}-\frac{\widehat{a}_{\sigma_0^2}\widehat{b}_{\sigma_e^2}}{\widehat{b}_{\sigma_0^2}\widehat{a}_{\sigma_e^2}}\widehat{\mu}_{\mu_0}\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}+\frac{\widehat{b}_{\sigma_e^2}\widehat{a}_{\sigma_0^2}}{\widehat{a}_{\sigma_e^2}\widehat{b}_{\sigma_0^2}}\widehat{\boldsymbol{\mu}}^2_{\boldsymbol{\beta}}
\end{aligned}
$$

$$= \frac{\widehat{b}_{\sigma_e^2}}{\widehat{a}_{\sigma_e^2}} \left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}^2 \circ \frac{1}{\widehat{\boldsymbol{\sigma}}_{\boldsymbol{\beta}}^2} - \frac{\widehat{a}_{\sigma_0^2}}{\widehat{b}_{\sigma_0^2}} \widehat{\mu}_{\mu_0} \widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}} \right).$$

This allows us to remove the last iterate $\boldsymbol{\theta}^{(\text{old})}$ and hence, leads to the following approximation of the gradient

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}} \Omega_{\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}, \widehat{\boldsymbol{\sigma}}_{\boldsymbol{\beta}}^2}(\boldsymbol{\theta}) \approx{} & \left\{ [\psi(\widehat{a}_w) - \psi(\widehat{a}_w + \widehat{b}_w)] + \frac{1}{2}[\psi(\widehat{a}_{\sigma_0^2}) - \ln(\widehat{b}_{\sigma_0^2})] \right\} \mathbf{1}_p + \frac{1}{2} \ln(\widehat{\boldsymbol{\sigma}}_{\boldsymbol{\beta}}^2) \\
& + \frac{\widehat{a}_{\sigma_0^2}}{2\widehat{b}_{\sigma_0^2}} [(\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}} - \widehat{\mu}_{\mu_0} \mathbf{1}_p)^2 + \widehat{\sigma}_{\mu_0}^2 \mathbf{1}_p + \widehat{\boldsymbol{\sigma}}_{\boldsymbol{\beta}}^2] + \frac{1}{2} \mathbf{1}_p - \operatorname{logit}(\boldsymbol{\theta}) \\
& - [\psi(\widehat{b}_w) - \psi(\widehat{a}_w + \widehat{b}_w)] \mathbf{1}_p - \frac{1}{2} \left(\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}} \right) \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{x}_{ij}^2 \circ (\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}^2 + \widehat{\boldsymbol{\sigma}}_{\boldsymbol{\beta}}^2) \\
& + \widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}^2 \circ \frac{1}{\widehat{\boldsymbol{\sigma}}_{\boldsymbol{\beta}}^2} - \frac{\widehat{a}_{\sigma_0^2}}{\widehat{b}_{\sigma_0^2}} \widehat{\mu}_{\mu_0} \widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}.
\end{aligned}$$

Now setting the gradient to $\mathbf{0}$ gives rise to

$$\begin{aligned}
\operatorname{logit}(\boldsymbol{\theta}) ={} & \left\{ [\psi(\widehat{a}_w) - \psi(\widehat{a}_w + \widehat{b}_w)] + \frac{1}{2}[\psi(\widehat{a}_{\sigma_0^2}) - \ln(\widehat{b}_{\sigma_0^2})] \right\} \mathbf{1}_p + \frac{1}{2} \ln(\widehat{\boldsymbol{\sigma}}_{\boldsymbol{\beta}}^2) \\
& + \frac{\widehat{a}_{\sigma_0^2}}{2\widehat{b}_{\sigma_0^2}} [(\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}} - \widehat{\mu}_{\mu_0} \mathbf{1}_p)^2 + \widehat{\sigma}_{\mu_0}^2 \mathbf{1}_p + \widehat{\boldsymbol{\sigma}}_{\boldsymbol{\beta}}^2] + \frac{1}{2} \mathbf{1}_p \\
& - [\psi(\widehat{b}_w) - \psi(\widehat{a}_w + \widehat{b}_w)] \mathbf{1}_p - \frac{1}{2} \left(\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}} \right) \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{x}_{ij}^2 \circ (\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}^2 + \widehat{\boldsymbol{\sigma}}_{\boldsymbol{\beta}}^2) \\
& + \widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}^2 \circ \frac{1}{\widehat{\boldsymbol{\sigma}}_{\boldsymbol{\beta}}^2} - \frac{\widehat{a}_{\sigma_0^2}}{\widehat{b}_{\sigma_0^2}} \widehat{\mu}_{\mu_0} \widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}
\end{aligned}$$

D. Updating variational parameters for $\mathbf{B}$

We now move on to the updating formula for the variational parameters for the random effect $\mathbf{B}$. This part still requires some work but is significantly simpler than the previous set of parameters $(\boldsymbol{\theta}, \widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}, \widehat{\boldsymbol{\sigma}}_{\boldsymbol{\beta}}^2)$. To begin with, we first compute the expected value of the quadratic form $(y_{ij} - \mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta} - \mathbf{z}_{ij}^{\mathrm{T}} \mathbf{b}_i)^2$ and view the variational parameters of y_{ij}, $\boldsymbol{\beta}$, and σ_e^2 as constants:

$$\begin{aligned}
\mathbb{E}_q[(y_{ij} - \mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta} - \mathbf{z}_{ij}^{\mathrm{T}} \mathbf{b}_i)^2] &= \operatorname{tr}[\mathbf{z}_{ij} \mathbf{z}_{ij}^{\mathrm{T}} \mathbb{E}_q(\mathbf{b}_i \mathbf{b}_i^{\mathrm{T}})] - 2\mathbb{E}_q(\mathbf{b}_i)^{\mathrm{T}} \mathbf{z}_{ij} (\widehat{y}_{ij} - \mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\Theta} \widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}) \\
&\quad + \text{constant} \\
&= \widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}^{\mathrm{T}} \mathbf{z}_{ij} \mathbf{z}_{ij}^{\mathrm{T}} \widehat{\boldsymbol{\mu}}_{\mathbf{b}_i} + \mathbf{z}_{ij}^{\mathrm{T}} \widehat{\boldsymbol{\Psi}}_{\mathbf{b}_i} \mathbf{z}_{ij} - 2\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}^{\mathrm{T}} \mathbf{z}_{ij} (\widehat{y}_{ij} - \mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\Theta} \widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}) \\
&\quad + \text{constant}.
\end{aligned}$$

Hence, we can proceed to write

$$\begin{aligned}
\Omega(\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}, \widehat{\boldsymbol{\Psi}}_{\mathbf{b}_i}) &= \mathbb{E}_q \left[\ln \frac{p(\mathbf{Y} \mid \boldsymbol{\beta}, \mathbf{B}, \sigma_e^2) p(\mathbf{b}_i \mid \widehat{\boldsymbol{\Psi}})}{q(\mathbf{b}_i \mid \widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}, \widehat{\boldsymbol{\Psi}}_{\mathbf{b}_i})} \right] \\
&= \sum_{j=1}^{n} \mathbb{E}_q[\ln p(y_{ij} \mid \boldsymbol{\beta}, \mathbf{b}_i, \sigma_e^2)] + \mathbb{E}_q[\ln p(\mathbf{b}_i \mid \widehat{\boldsymbol{\Psi}})] - \mathbb{E}_q[\ln q(\mathbf{b}_i \mid \widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}, \widehat{\boldsymbol{\Psi}}_{\mathbf{b}_i})]
\end{aligned}$$

$$
\begin{aligned}
&+ \text{constant} \\
&= -\mathbb{E}_q\left(\frac{1}{2\sigma_e^2}\right)\sum_{j=1}^{m}\mathbb{E}_q[(y_{ij}-\mathbf{x}_{ij}\boldsymbol{\beta}-\mathbf{z}_{ij}\mathbf{b}_i)^2] - \frac{1}{2}\mathbb{E}_q(\mathbf{b}_i^{\mathrm{T}}\widehat{\boldsymbol{\Psi}}^{-1}\mathbf{b}_i) \\
&\quad + \frac{1}{2}\ln\det\widehat{\boldsymbol{\Psi}}_i + \frac{1}{2}\mathbb{E}_q[(\mathbf{b}_i-\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i})^{\mathrm{T}}\widehat{\boldsymbol{\Psi}}_{\mathbf{b}_i}^{-1}(\mathbf{b}_i-\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i})] + \text{constant} \\
&= -\frac{1}{2}\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}}\sum_{j=1}^{n}\{\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}^{\mathrm{T}}\mathbf{z}_{ij}\mathbf{z}_{ij}^{\mathrm{T}}\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i} + \mathbf{z}_{ij}^{\mathrm{T}}\widehat{\boldsymbol{\Psi}}_{\mathbf{b}_i}\mathbf{z}_{ij} - 2\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}^{\mathrm{T}}\mathbf{z}_{ij}(\widehat{y}_{ij}-\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\Theta}\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}})\} \\
&\quad - \frac{1}{2}\mathrm{tr}(\hat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{\Psi}}_{\mathbf{b}_i}) - \frac{1}{2}\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}^{\mathrm{T}}\hat{\boldsymbol{\Psi}}^{-1}\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i} + \frac{1}{2}\mathrm{tr}(\widehat{\boldsymbol{\Psi}}_{\mathbf{b}_i}^{-1}\widehat{\boldsymbol{\Psi}}_{\mathbf{b}_i}) + \frac{1}{2}\ln\det\widehat{\boldsymbol{\Psi}}_{\mathbf{b}_i} \\
&\quad + \text{constant}
\end{aligned}
$$

Optimizing over $\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}$, we obtain

$$
\begin{aligned}
\frac{\partial\Omega}{\partial\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}} &= -\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}}\sum_{j=1}^{n}[\mathbf{z}_{ij}\mathbf{z}_{ij}^{\mathrm{T}}\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i} - (\widehat{y}_{ij}-\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\Theta}\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}})\widehat{\mathbf{z}}_{ij}] - \widehat{\boldsymbol{\Psi}}^{-1}\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i} = 0 \\
\Longrightarrow \quad \widehat{\boldsymbol{\mu}}_{\mathbf{b}_i} &= \left(\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}}\sum_{j=1}^{n}\mathbf{z}_{ij}\mathbf{z}_{ij}^{\mathrm{T}} + \widehat{\boldsymbol{\Psi}}^{-1}\right)^{-1}\left[\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}}\sum_{j=1}^{n}(y_{ij}-\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta})\mathbf{z}_{ij}\right].
\end{aligned}
$$

Optimizing over $\widehat{\boldsymbol{\Psi}}_{\mathbf{b}_i}$ in the positive definite cone in $\mathbb{R}^{l\times l}$, we obtain

$$
\widehat{\boldsymbol{\Psi}}_{\mathbf{b}_i} = \left(\frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}}\sum_{j=1}^{n}\mathbf{z}_{ij}\mathbf{z}_{ij}^{\mathrm{T}} + \hat{\boldsymbol{\Psi}}^{-1}\right)^{-1}.
$$

E. Updating variational parameters for σ_e^2

The updating formula for $\widehat{a}_{\sigma_e^2}$ and $\widehat{b}_{\sigma_e^2}$ involves a delicate and complete computation of the expected value of the quadratic form $\mathbb{E}_q[(y_{ij}-\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}-\mathbf{z}_{ij}^{\mathrm{T}}\mathbf{b}_i)^2]$. Write

$$
\begin{aligned}
&\mathbb{E}_q[(y_{ij}-\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}-\mathbf{z}_{ij}^{\mathrm{T}}\mathbf{b}_i)^2] \\
&\quad = \mathbb{E}_q[(y_{ij})^2] + \mathrm{tr}[\mathbf{z}_{ij}\mathbf{z}_{ij}^{\mathrm{T}}\mathbb{E}_q(\mathbf{b}_i\mathbf{b}_i^{\mathrm{T}})] + \mathrm{tr}[\mathbf{x}_{ij}\mathbf{x}_{ij}^{\mathrm{T}}\mathbb{E}_q(\boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}})] \\
&\qquad - 2\mathbb{E}_q(y_{ij})[\mathbf{x}_{ij}^{\mathrm{T}}\mathbb{E}_q(\boldsymbol{\beta}) + \mathbf{z}_{ij}^{\mathrm{T}}\mathbb{E}_q(\mathbf{b}_i)] + 2\mathbb{E}_q(\boldsymbol{\beta})^{\mathrm{T}}\mathbf{x}_{ij}\mathbb{E}_q(\mathbf{b}_i)^{\mathrm{T}}\mathbf{z}_{ij} \\
&\quad = \langle y_{ij}^2\rangle + (\mathbf{z}_{ij}^{\mathrm{T}}\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i})^2 + \mathbf{z}_{ij}^{\mathrm{T}}\widehat{\boldsymbol{\Psi}}_{\mathbf{b}_i}\mathbf{z}_{ij} + (\mathbf{x}_{ij}\boldsymbol{\Theta}\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}})^2 + \mathbf{x}_{ij}^2\boldsymbol{\Theta}(\mathbf{I}_p-\boldsymbol{\Theta})\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}^2 + \mathbf{x}_{ij}^2\boldsymbol{\Theta}\widehat{\boldsymbol{\sigma}}_{\boldsymbol{\beta}}^2 \\
&\qquad - 2\widehat{y}_{ij}[\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\Theta}\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}} + \mathbf{z}_{ij}^{\mathrm{T}}\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}] + 2\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{\Theta}\mathbf{x}_{ij}\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}^{\mathrm{T}}\mathbf{z}_{ij}.
\end{aligned}
$$

Namely, by setting

$$
\begin{aligned}
\mathrm{SSR} &= \sum_{i=1}^{m}\sum_{j=1}^{n}\mathbb{E}_q[(y_{ij}-\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}-\mathbf{z}_{ij}^{\mathrm{T}}\mathbf{b}_i)^2] \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n}\left\{\langle y_{ij}^2\rangle + (\mathbf{z}_{ij}^{\mathrm{T}}\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i})^2 + \mathbf{z}_{ij}^{\mathrm{T}}\widehat{\boldsymbol{\Psi}}_{\mathbf{b}_i}\mathbf{z}_{ij}\right\} \\
&\quad + \sum_{i=1}^{m}\sum_{j=1}^{n}\left\{(\mathbf{x}_{ij}\boldsymbol{\Theta}\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}})^2 + \mathbf{x}_{ij}^2\boldsymbol{\Theta}(\mathbf{I}_p-\boldsymbol{\Theta})\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}^2 + \mathbf{x}_{ij}^2\boldsymbol{\Theta}\widehat{\boldsymbol{\sigma}}_{\boldsymbol{\beta}}^2\right\} \\
&\quad - 2\sum_{i=1}^{m}\sum_{j=1}^{n}\widehat{y}_{ij}[\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\Theta}\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}} + \mathbf{z}_{ij}^{\mathrm{T}}\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}] + 2\sum_{i=1}^{m}\sum_{j=1}^{n}\widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{\Theta}\mathbf{x}_{ij}\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}^{\mathrm{T}}\mathbf{z}_{ij},
\end{aligned}
$$

we can write

$$\begin{aligned}
\Omega(\widehat{a}_{\sigma_e^2}, \widehat{b}_{\sigma_e^2}) &= \mathbb{E}_q\left[\ln p(\mathbf{Y} \mid \boldsymbol{\beta}, \mathbf{B}, \sigma_e^2)\right] + \mathbb{E}_q[\ln p(\sigma_e^2 \mid a_1, b_1)] - \mathbb{E}_q[\ln q(\sigma_e^2 \mid \widehat{a}_{\sigma_e^2}, \widehat{b}_{\sigma_e^2})] \\
&= \frac{mn}{2}\mathbb{E}_q\left[\ln\left(\frac{1}{\sigma_e^2}\right)\right] - \frac{1}{2}\mathbb{E}_q\left(\frac{1}{\sigma_e^2}\right)SSR + (a_1 - 1)\mathbb{E}_q\left[\ln\left(\frac{1}{\sigma_e^2}\right)\right] \\
&\quad - b_1\mathbb{E}_q\left(\frac{1}{\widehat{\sigma}_e^2}\right) + \ln\Gamma(\widehat{a}^2_{\sigma_e}) - \widehat{a}_{\sigma_e^2}\ln(\widehat{b}_{\sigma_e^2}) \\
&\quad - (\widehat{a}_{\sigma_e^2} - 1)\mathbb{E}_q\left[\ln\left(\frac{1}{\sigma_e^2}\right)\right]\widehat{b}_{\sigma_e^2}\mathbb{E}_q\left(\frac{1}{\widehat{\sigma}_e^2}\right) \\
&= \left(\frac{mn}{2} + a_1 - \widehat{a}_{\sigma_e^2}\right)[\psi(\hat{a}_{\sigma_e^2}) - \ln(\widehat{b}_{\sigma_e^2})] - \frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}}\left[\frac{1}{2}\mathrm{SSR} + b_1 - \hat{b}_{\sigma_e^2}\right] \\
&\quad + \ln\Gamma(\widehat{a}_{\sigma_e^2}) - \widehat{a}_{\sigma_e^2}\ln(\widehat{b}_{\sigma_e^2}).
\end{aligned}$$

We now optimize over $(\widehat{a}_{\sigma_e^2}, \widehat{b}_{\sigma_e^2})$ by taking the derivative of Ω and setting them to zero:

$$\begin{aligned}
&\frac{\partial\Omega}{\partial\widehat{a}_{\sigma_e^2}} = \left(\frac{mn}{2} + a_1 - \widehat{a}_{\sigma_e^2}\right)\psi'(\widehat{a}_{\sigma_e^2}) - \frac{1}{\widehat{b}_{\sigma_e^2}}\left(\frac{1}{2}\mathrm{SSR} + b_1 - \hat{b}_{\sigma_e^2}\right), \\
&\frac{\partial\Omega}{\partial\widehat{b}_{\sigma_e^2}} = -\left(\frac{mn}{2} + a_1 - \widehat{a}_{\sigma_e^2}\right)\frac{1}{\widehat{b}_{\sigma_e^2}} + \frac{\widehat{a}_{\sigma_e^2}}{\widehat{b}_{\sigma_e^2}}\left(\frac{1}{2}\mathrm{SSR} + b_1 - \widehat{b}_{\sigma_e^2}\right), \\
&\frac{\partial\Omega}{\partial\widehat{a}_{\sigma_e^2}} = \frac{\partial\Omega}{\partial\widehat{b}_{\sigma_e^2}} = 0 \quad\Longrightarrow\quad \widehat{a}_{\sigma_e^2} = a_1 + \frac{mn}{2}, \quad \widehat{b}_{\sigma_e^2} = \frac{1}{2}\mathrm{SSR} + b_1.
\end{aligned}$$

F. Updating the rest of the variational parameters

The updating formulas for the rest of the variational parameters can be obtained using routine methods. Write

$$\begin{aligned}
&\Omega(\widehat{a}_{\sigma_0^2}, \widehat{b}_{\sigma_0^2}) \\
&\quad = \mathbb{E}_q\left[\ln\frac{p(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid w, \mu_0, \sigma_0^2)p(\sigma_0^2)}{q(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \boldsymbol{\theta}, \widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}, \widehat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}})q(\sigma_0^2 \mid \widehat{a}_{\sigma_0^2}, \widehat{b}_{\sigma_0^2})}\right] \\
&\quad = \frac{1}{2}\sum_{k=1}^{p}\theta_k[\psi(\widehat{a}_{\sigma_0^2}) - \ln(\widehat{b}_{\sigma_0^2})] \\
&\qquad - \frac{\widehat{a}_{\sigma_0^2}}{2\widehat{b}_{\sigma_0^2}}\sum_{k=1}^{p}\theta_k\left\{(\widehat{\mu}^2_{\mu_0} + \widehat{\sigma}^2_{\mu_0})(1 - \theta_k) + [(\widehat{\mu}_{\mu_0} - \widehat{\mu}_{\boldsymbol{\beta}_k})^2 + \widehat{\sigma}^2_{\mu_0} + \widehat{\sigma}^2_{\boldsymbol{\beta}_k}]\theta_k\right\} \\
&\qquad + 2[\psi(\widehat{a}_{\sigma_0^2}) - \ln(\widehat{b}_{\sigma_0^2})] - \frac{\widehat{a}_{\sigma_0^2}}{\widehat{b}_{\sigma_0^2}} - (\widehat{a}_{\sigma_0^2} + 1)[\psi(\widehat{a}_{\sigma_0^2}) - \ln(\widehat{b}_{\sigma_0^2})] + \widehat{a}_{\sigma_0^2} \\
&\qquad + \ln\Gamma(\widehat{a}_{\sigma_0^2}) - \widehat{a}_{\sigma_0^2}\ln\widehat{b}_{\sigma_0^2}.
\end{aligned}$$

Setting the derivatives to zero yields

$$\begin{aligned}
&\frac{\partial\Omega}{\partial\widehat{a}_{\sigma_0^2}} = \frac{\partial\Omega}{\partial\widehat{b}_{\sigma_0^2}} = 0 \quad\Longrightarrow \\
&\widehat{a}_{\sigma_0^2} = 1 + \frac{1}{2}\sum_{k=1}^{p}\theta_k,
\end{aligned}$$

$$\widehat{b}_{\sigma_0^2} = 1 + \frac{1}{2}\sum_{k=1}^{p} \theta_k\{(\widehat{\mu}_{\mu_0}^2 + \widehat{\sigma}_{\mu_0}^2)(1-\theta_k) + [(\widehat{\mu}_{\mu_0} - \widehat{\mu}_{\boldsymbol{\beta}_k})^2 + \widehat{\sigma}_{\mu_0}^2 + \widehat{\sigma}_{\boldsymbol{\beta}_k}^2]\theta_k\}.$$

For $(\widehat{a}_w, \widehat{b}_w)$, we have

$$\begin{aligned}
\Omega(\widehat{a}_w, \widehat{b}_w) &= \mathbb{E}_q\left[\ln \frac{p(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid w, \mu_0, \sigma_0^2)p(w \mid a_w, b_w)}{q(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \boldsymbol{\theta}, \widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}, \widehat{\boldsymbol{\sigma}}_{\boldsymbol{\beta}}^2)q(w \mid \widehat{a}_w, \widehat{b}_w)}\right] \\
&= \sum_{k=1}^{p}\mathbb{E}_q[\gamma_k]\mathbb{E}_q[\ln w] + \sum_{k=1}^{p}\mathbb{E}_q[1-\gamma_k]\mathbb{E}_q[\ln(1-w)] + (a_w - 1)\mathbb{E}_q[\ln w] \\
&\quad + (b_w - 1)\mathbb{E}_q[\ln(1-w)] - (\widehat{a}_w - 1)\mathbb{E}_q[\ln w] - (\widehat{b}_w - 1)\mathbb{E}_q[\ln(1-w)] \\
&\quad + \ln\Gamma(\widehat{a}_w) + \ln\Gamma(\widehat{b}_w) - \ln\Gamma(\widehat{a}_w + \widehat{b}_w) \\
&= \sum_{k=1}^{p}\theta_k[\psi(\widehat{a}_w) - \psi(\widehat{a}_w + \widehat{b}_w)] + \sum_{k=1}^{p}(1-\theta_k)[\psi(\widehat{b}_w) - \psi(\widehat{a}_w + \widehat{b}_w)] \\
&\quad + (a_w - 1)[\psi(\widehat{a}_w) - \psi(\widehat{a}_w + \widehat{b}_w)] + (b_w - 1)[\psi(\widehat{b}_w) - \psi(\widehat{a}_w + \widehat{b}_w)] \\
&\quad - (\widehat{a}_w - 1)[\psi(\widehat{a}_w) - \psi(\widehat{a}_w + \widehat{b}_w)] - (\widehat{b}_w - 1)[\psi(\widehat{b}_w) - \psi(\widehat{a}_w + \widehat{b}_w)] \\
&\quad + \ln\Gamma(\widehat{a}_w) + \ln\Gamma(\widehat{b}_w) - \ln\Gamma(\widehat{a}_w + \widehat{b}_w) \\
&= \left(\sum_{k=1}^{p}\theta_k + a_w - \widehat{a}_w\right)\psi(\widehat{a}_w) \\
&\quad - \left[\sum_{k=1}^{p}\theta_k + \sum_{k=1}^{p}(1-\theta_k) + a_w + b_w - \widehat{a}_w - \widehat{b}_w\right]\psi(\widehat{a}_w + \widehat{b}_w) \\
&\quad + \left[\sum_{k=1}^{p}(1-\theta_k) + b_w - \widehat{b}_w\right]\psi(\widehat{b}_w) + \ln\Gamma(\widehat{a}_w) + \ln\Gamma(\widehat{b}_w) \\
&\quad - \ln\Gamma(\widehat{a}_w + \widehat{b}_w).
\end{aligned}$$

Now we optimize over $\widehat{a}_w$ and $\widehat{b}_w$ by taking the derivatives and setting them to zero:

$$\begin{aligned}
\frac{\partial\Omega}{\partial\widehat{a}_w} &= \left(\sum_{k=1}^{p}\theta_k + a_w - \widehat{a}_w\right)\psi'(\widehat{a}_w) - \psi'(\widehat{a}_w + \widehat{b}_w)\left(p + a_w + b_w - \widehat{a}_w - \widehat{b}_w\right), \\
\frac{\partial\Omega}{\partial\widehat{b}_w} &= \left[\sum_{k=1}^{p}(1-\theta_k) + b_w - \widehat{b}_w\right]\psi'(\widehat{b}_w) - \psi'(\widehat{a}_w + \widehat{b}_w)\left(p + a_w + b_w - \widehat{a}_w - \widehat{b}_w\right), \\
\frac{\partial\Omega}{\partial\widehat{a}_w} &= \frac{\partial\Omega}{\partial\widehat{b}_w} = 0 \quad \Longrightarrow \quad \widehat{a}_w = \sum_{k=1}^{p}\theta_k + a_w, \quad \widehat{b}_w = \sum_{k=1}^{p}(1-\theta_k) + b_w.
\end{aligned}$$

Finally, for $(\widehat{\mu}_{\mu_0}, \widehat{\sigma}_{\mu_0}^2)$, we have

$$\begin{aligned}
\Omega(\widehat{\mu}_{\mu_0}, \widehat{\sigma}_{\mu_0}^2) &= \mathbb{E}_q\left[\ln \frac{p(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid w, \mu_0, \sigma_0^2)p(\mu_0)}{q(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \boldsymbol{\theta}, \widehat{\boldsymbol{\mu}}_{\boldsymbol{\beta}}, \widehat{\boldsymbol{\sigma}}_{\boldsymbol{\beta}}^2)q(\mu_0 \mid \widehat{\mu}_{\mu_0}, \widehat{\sigma}_{\mu_0}^2)}\right] \\
&= -\frac{1}{2}\mathbb{E}_q\left[\frac{1}{\sigma_0^2}\right]\sum_{k=1}^{p}\mathbb{E}_q[\gamma_k]\mathbb{E}_q[(\boldsymbol{\beta}_k - \mu_0)^2] - \frac{1}{2}\mathbb{E}_q[\mu_0^2] \\
&\quad + \frac{1}{2\widehat{\sigma}_{\mu_o}^2}\mathbb{E}_q[(\mu_0 - \widehat{\mu}_{\mu_0})^2] + \text{constant}
\end{aligned}$$

$$= -\frac{\widehat{a}_{\sigma_0^2}}{2\widehat{b}_{\sigma_0^2}} \sum_{k=1}^{p} \theta_k \{ (\widehat{\mu}_{\mu_0}^2 + \widehat{\sigma}_{\mu_0}^2)(1-\theta_k) + [(\widehat{\mu}_{\mu_0} - \widehat{\mu}_{\boldsymbol{\beta}_k})^2 + \widehat{\sigma}_{\mu_0}^2 + \widehat{\sigma}_{\boldsymbol{\beta}_k}^2]\theta_k \}$$
$$- \frac{1}{2}(\widehat{\sigma}_{\mu_0}^2 + \widehat{\mu}_{\mu_0}^2) - \frac{1}{2}\ln\frac{1}{\widehat{\sigma}_{\mu_0}^2} + \text{constant}.$$

Setting the derivative to zero, we have

$$\frac{\partial\Omega}{\partial\widehat{\mu}_{\mu_0}} = -\frac{\widehat{a}_{\sigma_0^2}}{2\widehat{b}_{\sigma_0^2}} \sum_{k=1}^{p} \theta_k(\widehat{\mu}_{\mu_0} - \theta_k\widehat{\mu}_{\boldsymbol{\beta}_k}) - \widehat{\mu}_{\mu_0}, \quad \frac{\partial\Omega}{\partial\widehat{\sigma}_{\mu_0}^2} = -\frac{\widehat{a}_{\sigma_0^2}}{2\widehat{b}_{\sigma_0^2}} \sum_{k=1}^{p} \frac{\theta_k}{2} - \frac{1}{2} + \frac{1}{2\widehat{\sigma}_{\mu_0}^2},$$
$$\frac{\partial\Omega}{\partial\widehat{\mu}_{\mu_0}} = \frac{\partial\Omega}{\partial\widehat{\sigma}_{\mu_0}^2} = 0 \Rightarrow$$
$$\widehat{\mu}_{\mu_0} = \left(1 + \frac{\widehat{a}_{\sigma_0^2}}{\widehat{b}_{\sigma_0^2}} \sum_{k=1}^{p} \theta_k\right)^{-1} \left(\frac{\widehat{a}_{\sigma_0^2}}{\widehat{b}_{\sigma_0^2}} \sum_{k=1}^{p} \theta_k^2 \widehat{\mu}_{\boldsymbol{\beta}_k}\right), \quad \widehat{\sigma}_{\mu_0}^2 = \left(1 + \frac{\widehat{a}_{\sigma_0^2}}{\widehat{b}_{\sigma_0^2}} \sum_{k=1}^{p} \theta_k\right)^{-1}.$$

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

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.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Candes, E. J. and Plan, Y. (2010). Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936.

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

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

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.

Gross, D. (2011). Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566.

Huang, X., Wang, J., and Liang, F. (2016). A variational algorithm for bayesian variable selection. *arXiv preprint arXiv:1602.07640*.

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693.

Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

Li, Q. and Yucel, R. M. (2020). Variable selection in sequential hierarchical regression imputation. Technical report, University at Albany, the State University of New York.

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

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Yucel, R. M., He, Y., and Zaslavsky, A. M. (2008). Using calibration to improve rounding in imputation. *The American Statistician*, 62(2):125–129.

Yucel, R. M., He, Y., and Zaslavsky, A. M. (2011). Gaussian-based routines to impute categorical variables in health surveys. *Statistics in Medicine*, 30(29):3447–3460.

Yucel, R. M., Zhao, E., Schenker, N., and Raghunathan, T. E. (2017). Sequential hierarchical regression imputation. *Journal of Survey Statistics and Methodology*, 6(1):1–22.

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.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.