BAYESIAN GENERALIZED ADDITIVE MODEL SELECTION

by **Virginia Xiaojun He**

Thesis submitted in fulfillment of the requirements for the degree of

Doctor of Philosophy in Statistics

under the supervision of **Prof. Matt Wand** and **Dr. Matias Quiroz**

School of Mathematical and Physical Sciences

Faculty of Science

University of Technology Sydney

April 2025

Certificate of Original Authorship

I, **Virginia Xiaojun He** declare that this thesis is submitted in fulfillment of the requirements for the award of Doctor of Philosophy, in the School of Mathematical and Physical Sciences at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:
Signature removed prior to publication.

Virginia Xiaojun He

Date: April 30, 2025

Acknowledgements

This thesis would not have been possible without the support of a number of individuals.

Many thanks to my principal supervisor Professor Matt Wand. Thank you for your patience and kindness. I really appreciate it. This achievement would not have been possible without you. It is my honor to have Professor Matt Wand as my PhD principal supervisor.

To my colleagues and friends at UTS, thank you for your help, constant support and for listening to me.

Finally to Mum and Dad, I am eternally grateful for your support.

List of papers/publications/presentations

The following list of papers, publications and presentations relate to the work presented in this thesis:

- He, V. X., & Wand, M. P. (2024). Bayesian generalized additive model selection including a fast variational option, Journal of the Advances in Statistical Analysis, 108:639-668, DOI:10.1007/s10182-023-00490-y.
- He, V. X., & Wand, M. P. (2023). gamselBayes: Bayesian generalized additive model selection. R package (version 2.0). Comprehensive R Archive Network. Retrieved from https://doi.org/10.32614//CRAN.package.gamselBayes.
- He, V. X., & Wand, M. P. (2025). The grouped horseshoe distribution and its statistical properties. Journal of the Stat. Volume 14, Issue 1. DOI:10.1002/sta4.70050.
- Award for Excellence in Postgraduate Research at 24th Annual J.B. Douglas Award, New South Wales Branch of Statistical Society of Australia (2023).
- Presented at 12th conference of the Asian Regional Section of the International Association for Statistical Computing (2023).

Notation

In this chapter, important notations are introduced that we refer to throughout this thesis.

Acronyms

Table 1: Table with acronyms used in the thesis with their meanings.

Acronym	Meaning
CDF	Cumulative density function
DAG	Directed acyclic graph
GAM	Generalized additive model
GAMM	Generalized additive mixed model
KL	Kullback-Leibler
LASSO	Least absolute shrinkage and selection operator
MCMC	Markov chain Monte Carlo
MFVB	Mean field variational Bayes
PDF	Probability density function
PRSS	Penalised residual sum of squares

Contents

Abstract

Generalized additive models (GAMs) offer a parsimonious, flexible and interpretable framework for regression, particularly when handling a large numbers of candidate predictors. This thesis addresses the GAM variable selection problem: categorizing each candidate predictor's effect type to be linear, non-linear or zero on the mean response. We use Bayesian model selection paradigms and group least absolute shrinkage and selection operator (LASSO) priors. Two types of priors are explored for the sparse fits. The first, Laplace-Zero and Grouped Lasso-Zero priors, is applied to Gaussian and binary responses. The second, (Grouped) Horseshoe priors, is used for Gaussian, Poisson and Negative-Binomial responses. For both prior types, tailored auxiliary variable representations enable develop practical implementation of the Markov chain Monte Carlo (MCMC) sampling. Specifically, the sampling procedure reduces to the Gibbs sampling for Gaussian and binary responses, and to slice sampling for count responses. To improve computational scalability, particularly for large datasets, we also derive the mean field variational Bayes (MFVB) algorithms under the Laplace-Zero and Grouped Lasso-Zero priors. Although less accurate than MCMC, this variational approach offers substantial gains in speed. The GAM selection framework is further extended to generalized additive mixed model (GAMM) with random intercept, for Gaussian and binary responses using both MCMC and MFVB. Finally, the properties of the Grouped Horseshoe distribution and its use in Bayesian GAM selection are investigated. While many characteristics of univariate Horseshoe distribution are carried over, some distinctions arise in the grouped case.

Chapter 1

Introduction and Background

This chapter provides an introduction to the thesis, background information and theory required for the upcoming chapters.

1.1 Introduction

The need for dealing with large and complex data sets will continue to grow with rapid technological advancement. Continuous development of statistical methodologies is essential to keep pace with the growing variety and volume of data. Flexible regression methods are crucial for deriving insights and guiding actions from such data. Ongoing improvements are necessary due to the changing nature of large data sets. Generalized additive models (GAMs) are powerful for regression, providing attractive solutions to the problem of obtaining parsimonious and easy to interpret regression fits, with flexible predictor functions to reveal hidden patterns in the data (e.g. Hastie and Tibshirani, 1990; Wood, 2017). GAMs are able to model highly complex non-linear relationships when the number of potential predictors is large. Whilst there has been great progress in the development of methodology and software for GAMs in the past 40 years, there is still room for improvement in achieving principled, scalable and reliable model selection.

The critical challenges when using non-linear regression methods, such as GAMs, are categorizing each candidate predictor into one of three classes:

- zero effect,
- linear effect,

- non-linear effect

on the mean response and determining the forms of their potentially non-linear effects. We call this the three-category GAM selection problem, which is the focus of this thesis. Several approaches to this problem have been proposed, for instance, Shively et al. (1999), Ravikumar et al. (2009), Reich et al. (2009), Scheipl et al. (2012) and Chouldechova and Hastie (2015). Penalty-based model selection methods are particularly useful for large data sets with many candidate predictors. The main penalty-based model selection strategy uses the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996). LASSO is for selecting a scalar coefficient, to decide a predictor having a zero or linear effect. In contrast, Group LASSO is an attractive mechanism for selecting a vector of coefficients corresponding to spline basis functions, to decide a continuous predictor having a linear or non-linear effect.

We now summarize each of the existing approaches:

- Ravikumar et al. (2009) with the R package SAM is based on the regularization path, which is a grid of regularization amounts, and k-fold cross-validation idea. The methodology is for sparse additive models. It does not achieve the three-category GAM selection.

- Reich et al. (2009) is with a Bayesian formulation and use of the Markov chain Monte Carlo (MCMC).

- Shively et al. (1999) is with a Brownian motion approach.

- Scheipl et al. (2012) proposed a Bayesian approach for the three-category GAM selection with the R package spikeSlabGAM, with use of the "spike-and-slab" type priors for the sparse fit.

- Chouldechova and Hastie (2015) with the R package gamsel is a frequentist approach, called regularization path approach and it supports Gaussian and binary responses.

Both Reich et al. (2009) and Shively et al. (1999) are distinctly "beyond additive" in flavour and motivation. In short, Ravikumar et al. (2009), Reich et al. (2009) and Shively et al. (1999) do not consider a Bayesian three-category GAM selection. The "spike-and-slab" type priors used in Scheipl et al. (2012) has a Gaussian density function being the "slab" and has a mean zero, very small variance normal distribution

being the "spike", for selecting a scalar coefficient, to decide a predictor having a zero or linear effect. spikeSlabGAM uses the grouped extension of its "spike-and-slab" type priors, for selection a vector of coefficients corresponding to spline basis functions, to decide a continuous predictor having a linear or non-linear effect. spikeSlabGAM uses Gibbs sampling on Gaussian responses and uses Metropolis-Hasting sampling on binary responses.

The regularization path approach proposed in Chouldechova and Hastie (2015) can be viewed as one possible extension of the LASSO to the additive model setting. It does selection and shrinkage via convex optimization. There are four key features in their regularization path approach:

- use of the group LASSO,
- use of the Demmler-Reinsch spline bases to maintain orthogonal property,
- use of the regularization paths to compute a family of selected models over a grid of the regularization parameter values,
- use of the cross-validatory selection of the regularization parameter to choose a single model.

Since the cross-validation mean function often being monotonic rather than U-shaped, a lot of manual adjustment is required to the regularization parameter grids. Hence, the users have practical difficulties of finding the cross-validation minima to choose a single model. After running the simulation study, it was found that gamsel approach has a tendency to choose larger models. This leads to its poor classification performance with considerably higher misclassification rate on the three-category GAM selection compared to spikeSlabGAM. The misclassification rate is concerned with the proportion of times that a predictor is misclassified as zero, linear or non-linear effect based on the nature of the true effects according to how the data are simulated.

Apart from the "spike-and-slab" type priors, there are other types of priors for model selection. One of them is the Horseshoe-type prior. Xu et al. (2016) proposed the univariate Horseshoe distribution. The univariate Horseshoe distribution is for selecting a scalar coefficient, to decide a predictor having a zero or linear effect. Carvalho et al. (2010) proved the univariate Horseshoe distribution having a pole at

the origin, its score function converges to zero for large signals and the risk of its estimator is bounded. These features show that the univariate Horseshoe prior is heavy-tailed in constructing robust estimators, to prevent overshrinkage of obvious signals. Therefore, the ordinary Horseshoe distribution can be used as a prior. Carvalho et al. (2010) also proved, in sparse situations, the univariate Horseshoe prior will ensure the Bayes estimator for the sampling density converges to the right answer at a super-efficient rate.

1.2 Thesis Aim

Although spikeSlabGAM (Scheipl, 2022) achieved the three-category GAM selection, but it is comparatively slow and does not scale well to large data sets. Our methodology is inspired by the regularization path approach in Chouldechova and Hastie (2015) using the group LASSO and the Demmler-Reinsch spline bases. To avoid the practical difficulties of finding the cross-validation minima to choose a single model in regularization path approach, and to traverse a bigger model space compared with regularization path approach, this thesis presents a completely automatic and accurate Bayesian three-category GAM selection methodology with improved speed and scalability, on Gaussian and binary responses in the MCMC and the mean field variational Bayes (MFVB) as a faster option, with use of the Laplace Zero/-Grouped Lasso-Zero priors. We implemented the MCMC and MFVB algorithms in public available R package gamselBayes, with application on mortgage data, car auction data, educational data, Sydney real estate data and other areas. Further, this thesis investigates the grouped extensions of the various Bayesian statistical inference properties of univariate horseshoe priors in Carvalho et al. (2010) and the Grouped Horseshoe distribution in Bayesian GAM selection. We explored the three-category GAM selection with use of the (Grouped) Horseshoe priors on Gaussian, Poisson and Negative-Binomial responses in the MCMC. Additionally, we extended the Bayesian three-category GAM selection to generalized additive mixed model (GAMM) selection with random intercept, with use of the Laplace-Zero/Grouped Lasso-Zero priors to solve additional challenges arise with repeated measures in grouped data, with application to the Australian Red Cross Blood Service data.

1.3 Notations and Functions

The notations and functions used in this thesis are as follows:

- $\overset{\text{ind.}}{\sim}$ is an abbreviation for "distributed independently as".
- For a logical proposition $\mathcal{P}$, we let

  $I(\mathcal{P}) = 1$, if $\mathcal{P}$ is true and

  $I(\mathcal{P}) = 0$, if $\mathcal{P}$ is false.
- If $\boldsymbol{v}$ is a random vector then $\mathfrak{p}(\boldsymbol{v})$ denotes the density function of $\boldsymbol{v}$.
- If f is a smooth function that maps $\mathbb{R}^d$ to $\mathbb{R}$ then $\nabla_{\boldsymbol{x}} f(\boldsymbol{x})$ denotes the $d \times 1$ vector of partial derivatives of $f(\boldsymbol{x})$ with respect to the entries of $\boldsymbol{x}$.

For any column vector $\boldsymbol{a}$, column vector $\boldsymbol{b}$ having the same number of rows as $\boldsymbol{a}$ and square matrix $\boldsymbol{A}$, we have:

$$
\begin{aligned}
&\boldsymbol{a}_{-j} \equiv \text{the column vector } \boldsymbol{a} \text{ with the } j\text{th entry of } \boldsymbol{a} \text{ omitted},\\
&\boldsymbol{a} \odot \boldsymbol{b} \equiv \text{column vectors formed from } \boldsymbol{a} \text{ and } \boldsymbol{b} \text{ by obtaining element-wise products},\\
&\boldsymbol{a}/\boldsymbol{b} \equiv \text{column vectors formed from } \boldsymbol{a} \text{ and } \boldsymbol{b} \text{ by obtaining element-wise quotients},\\
&\|\boldsymbol{a}\| \equiv (\boldsymbol{a}^T\boldsymbol{a})^{1/2}, \text{ the Euclidean norm of } \boldsymbol{a} \text{ and}\\
&\text{diagonal}(\boldsymbol{A}) \equiv \text{ the column vector containing the diagonal entries of } \boldsymbol{A},\\
&\qquad\qquad \text{for any square matrix } \boldsymbol{A}.
\end{aligned}
\tag{1.1}
$$

The main algorithms use the following functions:

$$\begin{aligned}
&\text{logit}(x) \equiv \log\left(\frac{x}{1-x}\right), \\
&\text{expit}(x) \equiv \text{logit}^{-1}(x) = \frac{1}{1+\exp(-x)} \text{ and} \\
&\zeta(x) = \log\left\{2\Phi(x)\right\}, \text{ that leads to } \zeta'(x) = \phi(x)/\Phi(x), \text{ where } \phi \text{ is the } N(0,1) \\
&\qquad \text{density function and } \Phi \text{ is the } N(0,1) \text{ cumulative distribution function.}
\end{aligned} \tag{1.2}$$

Stable computation of $\zeta'(x)$ when x is a large negative number is not straightforward. Azzalini (2021) and Wand and Ormerod (2012), for example, provide practical solutions to this problem. Lastly, and expression of the form $\zeta'(\boldsymbol{v})$, where $\boldsymbol{v}$ is a column vector, is such that function evaluation is element-wise.

1.4 Notation for Spaces

We denote spaces as follows:

$\mathbb{R}$	Set of real numbers.
$\mathbb{R}^d$	Real coordinate space of d dimensions.
$\mathbb{Z}_{\geq 0}$	Coordinate line with all positive integers (integers greater than or equal to zero).

1.5 Probability Distributions

We now present the distributions used in this thesis in Table 1.1.

distribution	density/probability function in x	abbreviation
Bernoulli	$\wp^x (1-\wp)^{1-x};\ \ x=0,1; 0<\wp<1$	Bernoulli$(\wp)$
Multivariate Normal	$\left\|2\pi\boldsymbol{\Sigma}\right\|^{-1/2} \exp\left\{(\boldsymbol{x}-\boldsymbol{\mu})^T \times\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$	$N(\boldsymbol{\mu},\boldsymbol{\Sigma})$
Inverse Gamma	$\frac{\lambda^\kappa x^{-\kappa-1}e^{-\lambda/x}}{\Gamma(\kappa)};\quad x>0;\kappa,\lambda>0$	Inverse-Gamma(κ,λ)
Inverse Gaussian	$\frac{\sqrt{\lambda}\exp\left\{\frac{-\lambda(x-\mu)^2}{2\mu^2 x}\right\}}{\sqrt{2\pi x^3}};\ \ x>0;$ $\mu,\lambda>0$	Inverse-Gaussian(μ,λ)
Half-Cauchy	$\frac{2}{\pi\sigma\left((x/\sigma)^2+1\right)};\ \ x>0;\sigma>0$	Half-Cauchy(σ)
Laplace	$\frac{1}{2s}\exp\left(-\frac{\left\|x-\mu\right\|}{s}\right), s>0$	Laplace(μ,s)
Normal-Zero	$\frac{\rho}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{\sigma}\right)$ $+(1-\rho)\,\delta_0(x);\sigma>0;0\le\rho\le1$	Normal-Zero(μ,σ^2,ρ)
Laplace-Zero	$\frac{\rho}{2\sigma}\exp\left(-\frac{\left\|x-\mu\right\|}{\sigma}\right)$ $+(1-\rho)\,\delta_0(x);\sigma>0;0\le\rho\le1$	Laplace-Zero(μ,σ^2,ρ)
Truncated-Normal$_+$	$\frac{\exp\left\{-(x-\mu)^2/(2\sigma^2)\right\}}{\Phi(\mu/\sigma)\sqrt{2\pi\sigma^2}};$ $x>0;\sigma>0$	Truncated-Normal$_+$ (μ,σ^2)

Table 1.1: *Distributions used in this thesis and their corresponding density/probability functions.*

distribution	density/probability function in x	abbreviation
Truncated-Normal$_-$	$\frac{\exp\left\{-(x-\mu)^2/(2\sigma^2)\right\}}{\Phi(-\mu/\sigma)\sqrt{2\pi\sigma^2}}$; $x>0;\sigma>0$	Truncated-Normal$_-$ (μ,σ^2)
Poisson	$\frac{\lambda^x\exp(-\lambda)}{\Gamma(x+1)}$; $\Gamma(x)=(x-1)!$	Poisson(λ)
Negative-Binomial	$\frac{\Gamma(x+\kappa)}{\Gamma(x+1)\Gamma(\kappa)}\left(\frac{\mu}{\mu+\kappa}\right)^x\left(\frac{\kappa}{\mu+\kappa}\right)^\kappa$; $x=0,1,\ldots;\ \kappa,\mu>0$ $\Gamma(x)=(x-1)!$	NB(μ,κ)
Horseshoe	$\mathfrak{p}_{\text{HS}}\left(\frac{x-\mu}{\sigma}\right)\Big/\sigma$; $\mathfrak{p}_{\text{HS}}(x)\equiv(2\pi^3)^{-1/2}\exp(x^2/2)\times E_1(x^2/2)$; $E_1(x)\equiv\int_x^\infty\frac{e^{-t}}{t}dt, x\neq 0$	Horseshoe(μ,σ)
Moon Rock	$\left[\int_0^\infty\left\{t^t/\Gamma(t)\right\}^\alpha\exp(-\beta t)dt\right]^{-1}\times\left\{x^x/\Gamma(x)\right\}^\alpha\exp(-\beta x)$; $x,\alpha>0,\beta>\alpha$	Moon-Rock(α,β)

Table 1.1 continued: *Distributions used in this thesis and their corresponding density/probability functions.*

1.6 Useful Inverse Gaussian Density Function Results

Result 1.6.1 *Suppose the density function of x takes the form*

$$\mathfrak{p}(x)\propto x^{-3/2}\exp\left(-Sx-T/x\right),\quad x>0.$$

Then

$$x\sim\text{Inverse-Gaussian}\left(\sqrt{T/S},2T\right).$$

Result 1.6.2 *If* $x \sim$ *Inverse-Gaussian* (μ, λ) *, then*

$$E\left(1/x\right) = \frac{1}{\mu} + \frac{1}{\lambda}.$$

1.7 Useful Truncated Normal Distribution Results

ϕ and Φ are defined in (1.2).

Result 1.7.1 *Truncated-Normal*$_+(\mu, 1) \equiv N(\mu, 1)$ *with truncation over* $(0, +\infty)$*. If* $X \sim$ *Truncated-Normal*$_+\left(\mu, 1\right)$*, then the normalizing factor is* $\Phi(\mu)$*, and*

$$\int_0^{+\infty} \frac{\phi(x-\mu)}{\Phi(\mu)} dx = 1, \;\; E(X) = \mu + \frac{\phi(\mu)}{\Phi(\mu)}.$$

Result 1.7.2 *Truncated-Normal*$_-(\mu, 1) \equiv N(\mu, 1)$ *with truncation over* $(-\infty, 0)$*. If* $X \sim$ *Truncated-Normal*$_-\left(\mu, 1\right)$*, then the normalizing factor is* $\Phi(-\mu)$*, and*

$$\int_{-\infty}^0 \frac{\phi(x-\mu)}{\Phi(-\mu)} dx = 1, \;\; E(X) = \mu - \frac{\phi(-\mu)}{\Phi(-\mu)}.$$

Result 1.7.3 *If* $S \sim$ *Truncated-Normal*$_-(\mu, \sigma^2)$*, then*

$$-S \sim \text{Truncated-Normal}_+(-\mu, \sigma^2).$$

1.8 Fundamental Auxiliary Variable Results

Using standard distributional manipulations, we have the following fundamental auxiliary variable results listed in **Result 1.8.1–Result 1.8.6**, which are useful for the models being considered in this thesis. Derivation of **Result 1.8.2** is given in Appendix 1.A and Derivation of **Result 1.8.3** is given in Appendix 1.B.

Result 1.8.1 *Let* x *and* a *be random variables such that*

$$x\big|a \sim \text{Inverse-Gamma}\left(\tfrac{1}{2}, 1/a\right) \;\; \textit{and} \;\; a \sim \text{Inverse-Gamma}\left(\tfrac{1}{2}, 1/s^2\right).$$

Then

$$\sqrt{x} \sim \text{Half-Cauchy}\left(s\right).$$

Result 1.8.2 *Let x and a be random variables such that*

$$x|a \sim N\left(0, \sigma^2/a\right) \text{ and } a \sim \text{Inverse-Gamma}\left(1, \tfrac{1}{2}\right).$$

Then

$$x \sim \text{Laplace}\left(0, \sigma\right).$$

Result 1.8.3 *Let $\boldsymbol{x}$ be a $d \times 1$ random vector and a be a random variable such that*

$$\boldsymbol{x}|a \sim N\left(\mathbf{0}, \left(1/a\right)\sigma^2 \boldsymbol{I}_d\right) \text{ and } a \sim \text{Inverse-Gamma}\left(\frac{d+1}{2}, \frac{1}{2}\right).$$

Then

$$\mathfrak{p}\left(\boldsymbol{x}\right) = \frac{\exp\left(-\frac{\left\|\boldsymbol{x}\right\|_2}{\sigma}\right)}{\left(2\sigma^2\right)^d \pi^{(d-1)/2}\Gamma\left(\frac{d+1}{2}\right)}.$$

Result 1.8.4 *Suppose $\widetilde{x} \sim$ Laplace (μ, σ) and $b \sim$ Bernoulli (ρ) and let $x = b\widetilde{x}$.*

Then

$$x \sim \text{Laplace-Zero}\left(\mu, \sigma^2, \rho\right).$$

Result 1.8.5 *Suppose $x|a \sim$ Poisson(a) and $a \sim$ Gamma$(\kappa, \kappa/\mu)$.*

Then

$$x \sim \text{Negative-Binomial}\left(\mu, \kappa\right).$$

Result 1.8.6 *Let x, f_1, f_2 be random variables such that*

$$x|f_2 \sim N\left(0, 1/f_2\right), \quad f_2|f_1 \sim \text{Gamma}\left(\tfrac{1}{2}, f_1\right) \text{ and } f_1 \sim \text{Gamma}\left(\tfrac{1}{2}, 1\right).$$

Then

$$x \sim \text{Horseshoe}\left(0, 1\right).$$

1.9 Directed Acyclic Graphs

Graphical models, known as *probability graphical models*, are diagrammatic representations of probability distributions. Graphical models combine ideas from graph theory and probability (e.g. Bishop, 2006; Jordan, 2004). A graph contains a set of *nodes*

connected by *edges*. Adapting the conventions of Bishop (2006), nodes are shown as circles and edges are shown as line segments. Each node represents a random variable or a set of random variables. Each directed edge represents a "parent-child" relationship between the corresponding nodes.

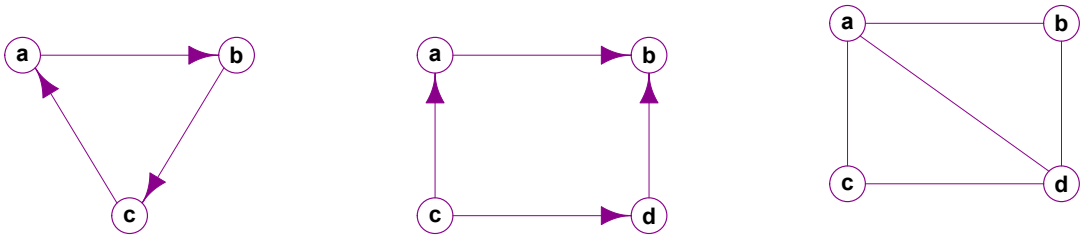
There are two major types of graphical models: *directed acyclic graphs* (DAGs), also known as the *Bayesian networks*, and *undirected graphs,* known as the *Markov random fields.* DAGs are distinguished from the undirected graphs by their directed edges. Of the two types, DAGs are more relevant to the theme of this thesis for our Bayesian graphical models. Therefore, we restrict our attention to DAGs. A directed acyclic graph requires all edges to be directed and free of cycles. Double-headed arrows are not allowed in DAGs. No sequence of connected directed edges can lead back to the starting point (see Figure 1.1b). The *Markov blanket* of a node is the set containing the node's parents, children and co-parents.

In a DAG, a parent node is the node connected away from the directed edge and the child node is on the directed edge. For example, in Figure 1.1b, nodes $\boldsymbol{a}$ and $\boldsymbol{d}$ are the parent nodes of node $\boldsymbol{b}$ and node $\boldsymbol{b}$ is the child of nodes $\boldsymbol{a}$ and $\boldsymbol{d}$. The Markov blanket of node $\boldsymbol{a}$ would be nodes $\boldsymbol{b}$ (child), $\boldsymbol{c}$ (parent) and $\boldsymbol{d}$ (co-parent). The Markov blanket of node $\boldsymbol{c}$ would be nodes $\boldsymbol{a}$ (child) and $\boldsymbol{d}$ (child).

Figure 1.1c shows the moralization of directed acyclic graph of Figure 1.1b. To moralize Figure 1.1b, we link nodes $\boldsymbol{a}$ and $\boldsymbol{d}$ with an undirected edge. However, no further links are required for the parent node of $\boldsymbol{a}$ and $\boldsymbol{d}$ (node $\boldsymbol{c}$) since it does not have a co-parent. In Figure 1.1c, the Markov blanket of node $\boldsymbol{a}$ would be nodes $\boldsymbol{b}$, $\boldsymbol{c}$ and $\boldsymbol{d}$, and the Markov blanket of node $\boldsymbol{c}$ would be nodes $\boldsymbol{a}$ and $\boldsymbol{d}$ (Pearl, 1988).

Theorem 1.9.1 For each node on a probabilistic DAG, the conditional distribution of the node given the rest of the nodes is the same as the conditional distribution of the node given its Markov blanket (Dechter and Pearl, 1988).

Figure 1.2 is another example of a DAG. In the graphical models literature, the variable X_6 shaded in red is an observed ("*evidence*") node. The variables X_3, X_4, X_5, X_8, X_9 and X_{10} in shaded blue and X_1, X_2, X_7, X_{11}, X_{12} and X_{13} in white are all random ("*hidden*") nodes. Based on **Theorem 1.9.1**, the full conditional distribution of the observed node X_6 denoted as $\mathfrak{p}\left(X_6 \middle| \text{rest}\right)$ is the same as the conditional



(a) A non DAG.

(b) A DAG.

(c) Moralized DAG from Figure 1.1b.

Figure 1.1: *Three basic graphical models. Figure 1.1b is a directed acyclic graph (DAG) since all edges are directed without cycles. Conversely, Figure 1.1a is not a directed acyclic graph, since it contains a cycle, although it contains directed edges. The moralized version of the directed acyclic graph from Figure 1.1b is Figure 1.1c. No edges have a direction and the parent nodes* $\boldsymbol{a}$ *and* $\boldsymbol{d}$ *are linked.*

distribution of X_6 given its Markov blanket, where "rest" denotes all of the random variables excluding X_6, i.e. $\{X_1, \ldots, X_5, X_7, \ldots, X_{13}\}$. The Markov blanket of the observed node X_6 contains its parents nodes X_3 and X_4, its children nodes X_9 and X_{10}, and its co-parents nodes X_5 and X_8, which are the six shaded nodes in blue. The observed node X_6 is conditionally independent of every other node in white in Figure 1.2 given its Markov blanket in Figure 1.3.

The Markov blanket of X_6 is $\{X_3, X_4, X_5, X_8, X_9, X_{10}\}$. The full conditional distribution of X_6 can be expressed as follows:

$$\begin{aligned}
\mathfrak{p}\left(X_6 \middle| \text{rest}\right) &= \mathfrak{p}\left(X_6 \middle| \text{Markov blanket of } X_6\right) = \mathfrak{p}\left(X_6 \middle| X_3, X_4, X_9, X_{10}, X_5, X_8\right) \\
&= \frac{\mathfrak{p}\left(X_3, X_4, X_5, X_6, X_8, X_9, X_{10}\right)}{\mathfrak{p}\left(X_3, X_4, X_5, X_8, X_9, X_{10}\right)} \propto \mathfrak{p}\left(X_3, X_4, X_5, X_6, X_8, X_9, X_{10}\right) \\
&\propto \mathfrak{p}\left(X_6 \middle| X_3, X_4\right) \mathfrak{p}\left(X_9 \middle| X_5, X_6\right) \mathfrak{p}\left(X_{10} \middle| X_6, X_8\right).
\end{aligned}$$

1.10 Generalized Additive Models

Generalized additive models (GAMs) are a flexible class of regression models that extend generalized linear models (GLMs) by allowing for non-linear relationships between the predictors and the response variable. GAMs achieve this by modeling the response as the sum of smooth functions of the predictors, more flexible and interpretable.

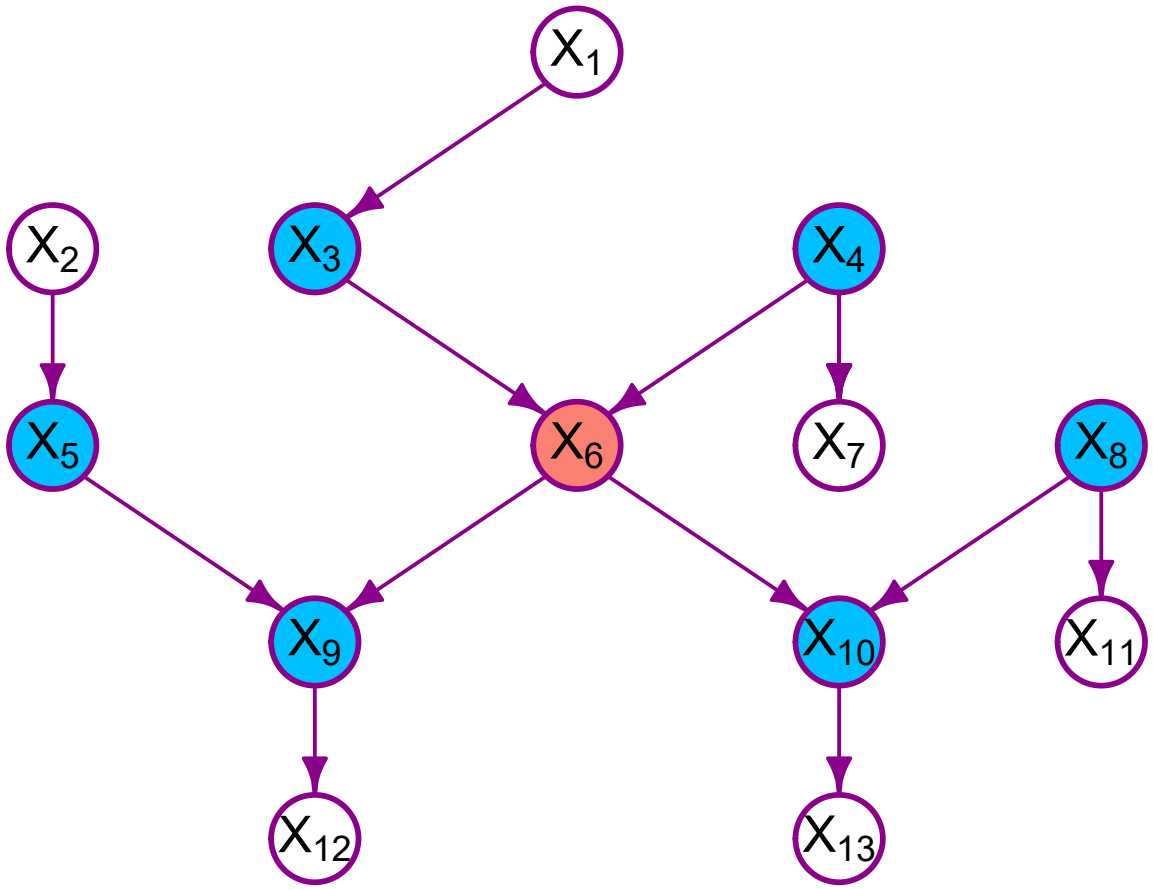


Figure 1.2: *Directed acyclic graph involving thirteen random variables:* $X_1, \ldots, X_{13}$. *The shaded node in red* X_6 *is the observed node. The other open circles in blue and white represent the random ("hidden") nodes. The nodes in blue represent the Markov blanket of the observed node* X_6.

We assume that the expectation of y, which is $E(y)$, denoted by μ, via the relationship $g(\mu) = \eta$, where $g(\cdot)$ is the link function and $F(\cdot) = g^{-1}(\cdot)$ is the inverse link function, also called the mean function. $x_1, \ldots, x_d$ are all fixed. In other words, although the mean is not directly the sum of smooth functions of the predictors, some function of the mean is.

Traditional linear regression assumes a linear relationship between η and the predictors as:

$$\eta = g(E(y)) = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d.$$

To accommodate possibly non-linear relationships between η and the predictors,

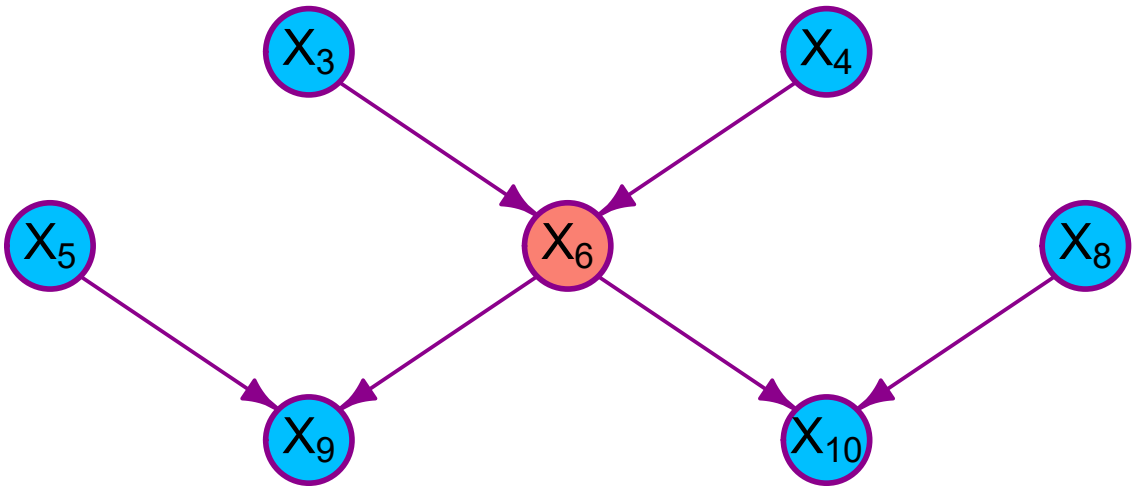


Figure 1.3: *The Markov blanket of observed node X_6, shown by blue shaded nodes.*

GAMs use smooth function of x_j, denoted as $f_j(x_j), 1 \leq j \leq d$. Thus, the GAM extension to model η is

$$\eta = g(\mu) = g(E(y)) = \beta_0 + f_1(x_1) + \ldots + f_d(x_d). \tag{1.3}$$

Since η in (1.3) can potentially be any real number, ideally g maps the domain of μ on to the entire real line. In a GAM, η is no longer linear in the predictor variables as in a GLM, but instead η is an additive function of the predictors. The corresponding smooth function model takes the form:

$$f_j(x_j) \equiv \beta_j x_j + \sum_{k=1}^{K_j} u_{jk} z_{jk}(x_j), \quad 1 \leq j \leq d,$$

for coefficients β_j and $\boldsymbol{u}_j \equiv \left(u_{j1}, \ldots, u_{jK_j}\right)$. Here $\{z_{jk}(\cdot) : 1 \leq k \leq K_j\}$ is an appropriate spline basis over an interval containing the x_j.

1.10.1 Binary Response Models

For a binary response variable y, where $y \in \{0, 1\}$, then $\mu = E(y)$ is in the interval $(0, 1)$ to the real line. The popular inverse link functions are as follows:

$$F = \begin{cases} \text{expit} & \text{logistic inverse link,} \\ \Phi & \text{probit inverse link} \end{cases}$$

where expit and Φ are defined in (1.2). The expit inverse link is the inverse of the logit link. The probit link is the inverse of the cumulative distribution function. For a binary response, the expected response is also the probability that the response equals 1. The logit link transforms this probability to the log-odds. Hence, the expit inverse link provides an log-odds interpretation and algebraic simplifications. The probit links often facilitates more tractable solutions. However, their model coefficients do not have a direct log-odds ratio interpretation in the same way that logit links do.

1.10.2 Count Response Models

Two models are commonly used to model count responses where $y \in \mathbb{Z}_{\geq 0}$. They are the Poisson model and negative binomial model.

The Poisson model (i.e. the response is Poisson distributed), we have the inverse link function being $F = \exp$. Since Poisson distributed variables are required to have an equal mean and variance, it is not appropriate for over-dispersed data where the variance is greater than the mean.

The negative binomial model is similar to the Poisson model, where the inverse link function is $F = \exp$. Compared to the Poisson model, it includes an additional parameter κ to account for over-dispersed data. As the shape parameter $\kappa \to \infty$ the negative binomial distribution converges to the Poisson distribution.

1.11 Generalized Additive Mixed Models

GAMs introduced in Section 1.10 only include fixed effects to model the response. Generalized additive mixed models (GAMMs) extend GAMs by incorporating ran-

dom effects, making them suitable for modeling complex data with hierarchical, grouped, or repeated measures structures. GAMMs combine the flexibility of GAMs in modeling non-linear relationships with the hierarchical and random-effects modeling capabilities of mixed models. The relationship between additive models (AM), additive mixed models (AMM), GAMs and GAMMs is clearly demonstrated in Table 1.2.

Table 1.2: *The relationship between four types of additive models, where AM refers to additive models, GAM refers to generalized additive models, AMM refers to additive mixed models and GAMM refers to generalized additive mixed models.*

	Response is normally distributed	Response is not normally distributed
Fixed effects only	AM	GAM
Fixed and random effects	AMM	GAMM

The expected value of the response y conditional on the fixed and random effects is related to the predictors using a link function $g(\cdot)$, where $F(\cdot) = g^{-1}(\cdot)$ is the inverse link function, also called the mean function. Although the mean is not directly the sum of smooth functions and random effects of the predictors, some function of the mean is. In Chapter 6 of this thesis, we investigate the GAMM selection with random intercept.

GAMM extends GLMM (generalized linear mixed model) by allowing continuous predictors to have a nonparametric functional impact on the mean response. Suppose that, in a longitudinal data set, y_{ij} is the jth measurement of the ith subject, with $1 \leq j \leq n_i$, $1 \leq i \leq m$. Here, both $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are fixed. A GAMM for the data is

$$y_{ij}\big|U_i \overset{\text{ind.}}{\sim} \text{Bernoulli}\left(\text{logit}^{-1}\left(U_i + f_1\left(x_{1ij}\right) + f_2\left(x_{2ij}\right)\right)\right), \;\; U_i \overset{\text{ind.}}{\sim} N\left(\mu, \sigma^2\right)$$

where U_i is a random subject intercept, f_1 and f_2 are smooth functions of the predictors, capturing non-linear relationships. The notation $\overset{\text{ind.}}{\sim}$ is defined in Section 1.3. The Bernoulli distribution is described in Table 1.1. $\text{logit}^{-1}(x)$ is defined in (1.2). $f_1(x_1) + f_2(x_2)$ is modelled using mixed model-based penalized splines as follows:

$$f_1(\boldsymbol{x}_1) + f_2(\boldsymbol{x}_2) = \beta_0 + \beta_{x_1}\boldsymbol{x}_1 + \sum_{\ell=1}^{L_1} u_{1\ell} z_{1\ell}(\boldsymbol{x}_1) + \beta_{x_2}\boldsymbol{x}_2 + \sum_{\ell=1}^{L_2} u_{2\ell} z_{2\ell}(\boldsymbol{x}_2)$$

where $u_{1\ell} \overset{\text{ind.}}{\sim} N(0, \sigma_1^2), \quad 1 \leq \ell \leq L_1$ and $u_{2\ell} \overset{\text{ind.}}{\sim} N(0, \sigma_2^2), \quad 1 \leq \ell \leq L_2$. Here $\{z_{k\ell}(\cdot) : 1 \leq \ell \leq L_k, \quad 1 \leq k \leq 2\}$ is an appropriate spline basis over an interval containing the $\boldsymbol{x}_k$.

1.12 Semiparametric Regression

Throughout this thesis, we use a direct generalization of smoothing splines, namely the *O'Sullivan splines* (O'Sullivan, 1986, Section 3). O'Sullivan splines are a class of penalised splines based on B-spline basis functions. They have the attractions of requiring considerably fewer basis functions than the ones described in Wahba (1990) and Green and Silverman (1994), in which the number of basis functions roughly equals the sample size. In this thesis, the default value of number of basis functions we used is 12. Moreover, they have the attractions of their smoothness, numerical stability and natural boundary conditions makes them the most widely used class of penalised splines in standard statistical software, for example, the smooth.spline( ) function in R. Wand and Ormerod (2008) give a detailed description of O'Sullivan penalized splines. A brief sketch of description is provided here for reference.

Notations used in this subsection are as follows:

- $\boldsymbol{x} \equiv (x_1, \ldots, x_n) \equiv$ a continuous univariate data set and $n \equiv$ sample size.
- x_is correspond to values of a continuous candidate predictor, with $1 \leq i \leq n$.
- Integer $K \equiv$ the number of knots, with $K \leq n$ and $\boldsymbol{\kappa}_{\text{inter.}} \equiv (\kappa_1, \ldots, \kappa_{K-2})$ to be a set of interior knots.
- $\boldsymbol{B}_x \equiv [B_1(x), \ldots, B_{K+4}(x)]$, cubic B-spline basis functions designed by knots.
- $\boldsymbol{B} \equiv$ design matrix, a $n \times (K+4)$ matrix.
- $B_{ik} \equiv$ design matrix $\boldsymbol{B}$ with (i,k)th entry, written as $B_{ik} = B_k(x_i)$.
- $\boldsymbol{v} \equiv$ the coefficients of $\boldsymbol{B}$ and $\hat{\boldsymbol{v}} \equiv$ the estimates of $\boldsymbol{v}$.
- $\lambda \equiv$ smoothing parameter, with $\lambda > 0$.
- $\boldsymbol{\Omega} \equiv$ penalty matrix, a $(K+4) \times (K+4)$ matrix.
- $\boldsymbol{\Omega}_{kk'} \equiv$ penalty matrix $\boldsymbol{\Omega}$ with (k,k')th entry.

- $\hat{f}(\boldsymbol{x}) \equiv$ an estimate of f.
- $n_U \equiv$ the number of unique x_is.

Consider the simple non-parametric regression setting as follows:

$$y_i = f(x_i) + \epsilon_i, \;\; 1 \leq i \leq n,$$

where $(x_i, y_i) \in \mathbb{R} \times \mathbb{R}$. Suppose we use a set of cubic B-spline basis functions as follows:

$$\boldsymbol{B}_x \equiv [B_1(x), \ldots, B_{K+4}(x)], \;\; \text{for} \;\; K \leq n$$

to estimate f over the interval $[a, b]$ containing the x_is. The corresponding knot sequence is defined by:

$$a = \kappa_1 = \kappa_2 = \kappa_3 = \kappa_4 < \kappa_5 < \ldots < \kappa_{K+4} < \kappa_{K+5} = \kappa_{K+6} = \kappa_{K+7} = \kappa_{K+8} = b.$$

The coefficients $\boldsymbol{v}$ can be estimated in a number of ways. One of the simple way is the *penalized residual sum of squares* (PRSS). PRSS chooses the coefficients $\boldsymbol{v}$ by minimizing:

$$PRSS(f, \lambda) = \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \int_a^b f''(x)^2 \, dx. \tag{1.4}$$

The second term $\lambda \int_a^b f''(x)^2 \, dx$ in (1.4) is referred to as the penalty term. It penalizes overly rough fits, leading to a smoother result. λ controls the amount of smoothing. The solution to (1.4) is the O'Sullivan penalised spline as follows:

$$f(\boldsymbol{x}) = \boldsymbol{B}\boldsymbol{v}.$$

Therefore, (1.4) can be written as:

$$PRSS(\boldsymbol{v}, \lambda) = (\boldsymbol{y} - \boldsymbol{B}\boldsymbol{v})^T (\boldsymbol{y} - \boldsymbol{B}\boldsymbol{v}) + \lambda \boldsymbol{v}^T \boldsymbol{\Omega} \boldsymbol{v} \tag{1.5}$$

where

$$B_{ik} = B_k(x_i) \;\; \text{and} \;\; \Omega_{kk'} = \int_a^b B_k''(x) \, B_{k'}''(x) \, dx.$$

Straightforward algebraic manipulation results in the following fitted O'Sullivan penalised spline, which provides a solution to (1.5):

$$\hat{f}(\boldsymbol{x}) = \boldsymbol{B}\hat{\boldsymbol{v}}, \;\; \text{where} \;\; \hat{\boldsymbol{v}} = \left(\boldsymbol{B}^T\boldsymbol{B} + \lambda\boldsymbol{\Omega}\right)^{-1} \boldsymbol{B}^T\boldsymbol{y}. \tag{1.6}$$

The design matrix $\boldsymbol{B}$ is easy to compute and is readily accessible in the R environment. However, the calculation of the penalty matrix $\boldsymbol{\Omega}$ requires an additional effect. In Section 6 of Wand and Ormerod (2008), the penalty matrix $\boldsymbol{\Omega}$ is derived using an exact algebraic expression by applying the Simpson's rule to each inter-knot interval as follows:

$$\boldsymbol{\Omega} = \left(\widetilde{\boldsymbol{B}}^{''}\right)^{T} \operatorname{diag}(\boldsymbol{\omega})\, \widetilde{\boldsymbol{B}}^{''},$$

where $\widetilde{\boldsymbol{B}}^{''}$ is the $3(K+7) \times (K+4)$ matrix with the (i,j)th entry $\widetilde{\boldsymbol{B}}_j^{''}(\widetilde{x}_i)$ and $\widetilde{x}_i$ is the ith entry of the vector given by:

$$\widetilde{x}_i = \left(\kappa_1, \frac{\kappa_1+\kappa_2}{2}, \kappa_2, \kappa_2, \frac{\kappa_2+\kappa_3}{2}, \kappa 3, \ldots, \kappa_{K+7}, \frac{\kappa_{K+7}+\kappa_{K+8}}{2}, \kappa_{K+8}\right),$$

and $\boldsymbol{\omega}$ is the $3(K+7) \times 1$ vector given by:

$$\begin{aligned}\boldsymbol{\omega} = \big\{ &\tfrac{1}{6}(\triangle\boldsymbol{\kappa})_1, \tfrac{4}{6}(\triangle\boldsymbol{\kappa})_1, \tfrac{1}{6}(\triangle\boldsymbol{\kappa})_1, \tfrac{1}{6}(\triangle\boldsymbol{\kappa})_2, \tfrac{4}{6}(\triangle\boldsymbol{\kappa})_2, \\ &\tfrac{1}{6}(\triangle\boldsymbol{\kappa})_2, \ldots, \tfrac{1}{6}(\triangle\boldsymbol{\kappa})_{K+7}, \tfrac{4}{6}(\triangle\boldsymbol{\kappa})_{K+7}, \tfrac{1}{6}(\triangle\boldsymbol{\kappa})_{K+7}\big\},\end{aligned}$$

where $(\triangle\boldsymbol{\kappa})_k \equiv \kappa_{K+1} - \kappa_K, 1 \le k \le K+7$. A common default choice for the number of knots is $K = \min\left(n_U/4, 35\right)$. The distribution of knots can either be quantile-based or equally spaced (e.g. Ruppert et al., 2003).

1.13 Bayesian Inference

Bayesian inference focuses on determining the joint posterior distribution of the parameters of interest based on the observed data. In many cases, exact inference is not practical since the posterior distributions are intractable. Therefore, it is necessary to use the approximate inference methods in practice. Examples include BUGS (Ligges et al., 2009) and Stan (Stan Development Team, 2015) , which are based on Markov chain Monte Carlo (MCMC) methods. MCMC, described in Section 1.13.1, is one of the standard methods for estimating Bayesian statistical models, which has been proven effective across a wide range of problems. For large models with complex posterior distributions, MCMC can be computationally demanding and prone to poor mixing, resulting in slow convergence. We therefore propose an alternative mean field variational Bayes (MFVB) approach as a faster option, which we will elaborate on further in Section 1.13.2. MFVB, whilst not as accurate as MCMC, this fast

variational option enhances scalability to very large data sets.

Consider a generic Bayesian model with parameter vector $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_n\} \in \Theta$ and observed data vector $\boldsymbol{y}$. Bayesian inference updates the probability of a posterior $\boldsymbol{\theta}$ based on prior knowledge and data. The posterior density function is as follows:

$$\mathfrak{p}\left(\boldsymbol{\theta}|\boldsymbol{y}\right) = \frac{\mathfrak{p}\left(\boldsymbol{y}, \boldsymbol{\theta}\right)}{\mathfrak{p}\left(\boldsymbol{y}\right)} = \frac{\mathfrak{p}\left(\boldsymbol{y}|\boldsymbol{\theta}\right)\mathfrak{p}\left(\boldsymbol{\theta}\right)}{\mathfrak{p}\left(\boldsymbol{y}\right)}$$

For many practical models, the posterior density functions are intractable, making it impossible to directly compute the marginal densities of parameters or other quantities of interest.

1.13.1 Markov Chain Monte Carlo

The MCMC approach is used in Chapter 2, Chapter 3, Chapter 5, Chapter 6 and Chapter 7 in this thesis.

For a long time, MCMC sampling techniques were the only broadly applicable method available. A Markov chain is a sequence of random variables $\boldsymbol{\theta}^{(0)}, \ldots, \boldsymbol{\theta}^{(T)}$. The following expression

$$\mathfrak{p}\left(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(0)}, \ldots, \boldsymbol{\theta}^{(t-1)}, \boldsymbol{y}\right) = \mathfrak{p}\left(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t-1)}, \boldsymbol{y}\right)$$

shows that for a Markov chain, the probability of transitioning to the current state depends only on the preceding state.

We denote d as the number of predictors, $1 \leq j \leq d$. Gibbs sampling is the most basic component-wise algorithm in MCMC methods (Gelfand and Smith, 1990; Geman and Geman, 1984; Robert and Casella, 2004). The Markov chain of samples of $\boldsymbol{\theta}^{(t)}, 0 \leq t \leq T$, is obtained by sampling from the full conditional distributions $\mathfrak{p}\left(\theta_j^{(t)}|\theta_{-j}^{(t-1,t)}, \boldsymbol{y}\right)$, which is the density of θ_j conditional on other parameters θ_{-j}. Here, $\theta_{-j}^{(t-1,t)}$ denotes all variables except θ_d. It relies on the assumption that the conditional distributions $\mathfrak{p}\left(\theta_j^{(t)}|\theta_{-j}^{(t-1,t)}, \boldsymbol{y}\right)$ are easier to sample from than the full joint distribution $\mathfrak{p}\left(\theta_1, \ldots, \theta_d\right)$.

Each $\theta_j^{(t)}$ is the updated conditional on the latest value of the other parameters

θ_{-j}. For the parameters already updated, it refers to the iteration t values for these parameters. For the parameters not updated, it refers to the iteration $t-1$ values for these parameters. The sequence of samples forms a Markov chain, where the next state depends only on the current state. The chain eventually converges to the target distribution. The iterative process is as follows:

- Starting with an initial guess for the variables: $\theta_1^{(0)}, \ldots, \theta_d^{(0)}$.
- For $t = 0, \ldots, T$:

$$\theta_1^{(t)} \sim \mathfrak{p}\left(\theta_1 \middle| \theta_2^{(t-1)}, \ldots, \theta_d^{(t-1)}, \boldsymbol{y}\right)$$
$$\theta_2^{(t)} \sim \mathfrak{p}\left(\theta_2 \middle| \theta_1^{(t)}, \ldots, \theta_d^{(t-1)}, \boldsymbol{y}\right)$$
$$\vdots$$
$$\theta_{d-1}^{(t)} \sim \mathfrak{p}\left(\theta_{d-1} \middle| \theta_1^{(t)}, \ldots, \theta_d^{(t-1)}, \boldsymbol{y}\right)$$
$$\theta_d^{(t)} \sim \mathfrak{p}\left(\theta_d \middle| \theta_1^{(t)}, \ldots, \theta_{d-1}^{(t)}, \boldsymbol{y}\right)$$

  until convergence reaches.

Gibbs sampling, by iteratively sampling from the conditional distributions of each variable, is accurate and effective for models with tractable conditional distributions. It can be efficient if conditional distributions are simple. It is computationally costly and it may take a long time to converge to the target distribution for complex models. It has applications in Bayesian inference, latent variable models, image processing, spatial statistics and more.

1.13.2 A Brief Introduction to Variational Approximations

Variational approximations are a class of fast and deterministic methods. They are originally used in the statistical physics and computer science literature (Bishop, 2006; Jordan et al., 1999; Titterington, 2004; Wainwright and Jordan, 2008). Since around 2005, variational methods have gained increasing attention in the statistical literature. For example, McGrory and Titterington (2007), Wand et al. (2012) and Wand and Ormerod (2011) propose variational methodology for a various rage of applications, from finite mixture models, complex models with elaborate distributions (such as asymmetric Laplace and skew normal) to spline and wavelet regression models. Moreover, Wang and Titterington (2006) prove convergence of variational

algorithms for normal mixture models and You et al. (2014) present several information criteria that are useful for model selection. Here, we offer a brief overview of variational approximations and emphasize their key concepts. Comprehensive summaries can be found in Bishop (2006) and Ormerod and Wand (2010).

Variational approximations are used to approximate complex probability distributions, such as the posterior distributions in Bayesian inference. MCMC produces random samples. Variational methods yield a fixed approximating distribution $\mathfrak{q}(\theta)$, which is a deterministic output.

1.13.3 Mean Field Variational Bayes

The mean field variational Bayes (MFVB) approach is used in Chapter 4, Chapter 5, Chapter 6 and Chapter 7 in this thesis.

MFVB is a specific type of variational approximation method used to approximate complex posterior distributions in Bayesian inference. The advantage is its simplicity and scalability, making it suitable for large-scale data and high-dimensional models.

The notation used in this subsection are as follows:

- $d \equiv$ number of predictors, with $1 \leq j \leq d$.
- Observed data vector $\boldsymbol{y}$, parameter vector $\boldsymbol{\theta}$ and parameter space $\boldsymbol{\Theta}$.
- $\mathfrak{p}\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right) \equiv$ true posterior density function.
- $\mathfrak{q}(\boldsymbol{\theta}) \equiv$ variational distributions, also called approximating density function.
- $\mathcal{Q} \equiv$ mean field restriction or product restriction. The posterior $\mathfrak{q}(\boldsymbol{\theta})$ can be factorized into independent distributions of each parameter or groups of parameters, without any restriction on the functional form on the factors (Bishop, 2006), expressed as follows:

$$\mathcal{Q} = \left\{ \mathfrak{q}(\boldsymbol{\theta}) : \mathfrak{q}(\boldsymbol{\theta}) = \prod_{j=1}^{d} \mathfrak{q}_j(\boldsymbol{\theta}_j) \quad \text{for some partition} \quad \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d\} \quad \text{of} \quad \boldsymbol{\theta} \right\}.$$

- $\mathrm{KL}\left\{\mathfrak{q}\left(\boldsymbol{\theta}\right) \middle\| \mathfrak{p}\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right)\right\} \equiv$ the difference, also called the Kullback-Leibler (KL) divergence, between the approximating density function $\mathfrak{q}\left(\boldsymbol{\theta}\right)$ and the actual posterior $\mathfrak{p}\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right)$, as a measure of how well our approximation fits the true posterior, expressed as follows:

$$\mathrm{KL}\left\{\mathfrak{q}\left(\boldsymbol{\theta}\right) \middle\| \mathfrak{p}\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right)\right\} = \int_{\boldsymbol{\Theta}} \mathfrak{q}\left(\boldsymbol{\theta}\right) \log\left\{\frac{\mathfrak{q}\left(\boldsymbol{\theta}\right)}{\mathfrak{p}\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right)}\right\}$$

  with $\mathrm{KL}\left\{\mathfrak{q}\left(\boldsymbol{\theta}\right) \middle\| \mathfrak{p}\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right)\right\} \geq 0$ for all densities $\mathfrak{q}$.

- $\mathfrak{q}^{*}\left(\boldsymbol{\theta}\right) \equiv$ optimal $\mathfrak{q}$-density function, a particular density function over the elements of the partition $\mathcal{Q}$ with minimum KL divergence, also called mean field variational Bayes approximation to the actual posterior, expressed as follows:

$$\mathfrak{q}^{*}\left(\boldsymbol{\theta}\right) = \underset{\mathfrak{q}\in\mathcal{Q}}{\operatorname{argmin}} \ \mathrm{KL}\left\{\mathfrak{q}\left(\boldsymbol{\theta}\right) \middle\| \mathfrak{p}\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right)\right\}.$$

- $\underline{p}\left(\boldsymbol{y};\mathfrak{q}\right) \equiv$ lower bound on the marginal likelihood .

- $\log p\left(\boldsymbol{y}\right) \equiv$ marginal log-likelihood.

- $E_{\mathfrak{q}(-\boldsymbol{\theta}_j)} \equiv$ expectation with respect to the $\mathfrak{q}$-densities of all parameters except $\boldsymbol{\theta}_j, 1 \leq j \leq d$.

The essence of variational inference is to search over $\mathfrak{q} \in \mathcal{Q}$, for some set $\mathcal{Q}$ of density functions. The goal is to find out the optimal $\mathfrak{q}$-density function $\mathfrak{q}^{*}\left(\boldsymbol{\theta}\right)$, with the minimum KL divergence with the actual posterior $\mathfrak{p}\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right)$.

MFVB essentially assumes independence among parameters in the posterior, which may not exist in the actual posterior distribution. The independence assumption, based on the chosen parameter partition, can be unrealistic in cases with high posterior correlations among parameters, resulting in poor approximations. The specific parametric families used for each approximating density factor $\mathfrak{q}_j\left(\boldsymbol{\theta}_j\right)$ are determined through the variational approach.

Since the marginal log-likelihood can be expressed as follows:

$$\begin{aligned}\log p(\boldsymbol{y}) &= \log p(\boldsymbol{y}) \int_{\boldsymbol{\Theta}} \mathfrak{q}(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \int_{\boldsymbol{\Theta}} \mathfrak{q}(\boldsymbol{\theta}) \log p(\boldsymbol{y})\, d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\Theta}} \mathfrak{q}(\boldsymbol{\theta}) \log \left\{ \frac{\mathfrak{p}(\boldsymbol{y}, \boldsymbol{\theta}) / \mathfrak{q}(\boldsymbol{\theta})}{\mathfrak{p}(\boldsymbol{\theta}|\boldsymbol{y}) / \mathfrak{q}(\boldsymbol{\theta})} \right\} d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\Theta}} \mathfrak{q}(\boldsymbol{\theta}) \log \left\{ \frac{\mathfrak{p}(\boldsymbol{y}, \boldsymbol{\theta})}{\mathfrak{q}(\boldsymbol{\theta})} \right\} d\boldsymbol{\theta} + \int_{\boldsymbol{\Theta}} \mathfrak{q}(\boldsymbol{\theta}) \log \left\{ \frac{\mathfrak{q}(\boldsymbol{\theta})}{\mathfrak{p}(\boldsymbol{\theta}|\boldsymbol{y})} \right\} d\boldsymbol{\theta} \\ &= \log \underline{p}(\boldsymbol{y}; \mathfrak{q}) + \mathrm{KL}\left\{ \mathfrak{q}(\boldsymbol{\theta}) \,\middle\|\, \mathfrak{p}(\boldsymbol{\theta}|\boldsymbol{y}) \right\}.\end{aligned}$$

Therefore, minimizing the KL divergence is equivalent to maximizing the lower bound on the marginal likelihood $\underline{p}(\boldsymbol{y}; \mathfrak{q})$. With the product restriction being

$$\mathfrak{q}(\boldsymbol{\theta}) = \prod_{j=1}^{d} \mathfrak{q}_j(\boldsymbol{\theta}_j),$$

the optimal $\mathfrak{q}$-density functions can be obtained via an iterative coordinate ascent algorithm (Bishop, 2006; Ormerod and Wand, 2010). Define

$$E_{\mathfrak{q}(-\boldsymbol{\theta}_j)} \{\log \mathfrak{p}(\boldsymbol{y}, \boldsymbol{\theta})\} = \int \log p(\boldsymbol{y}, \boldsymbol{\theta}) \prod_{k \neq j} \mathfrak{q}(\boldsymbol{\theta}_k)\, d\boldsymbol{\theta}_k.$$

to be the log posterior averaged over the current estimates of the approximating density functions for all but the i-th parameter vector.

An updated value of the lower bound on the marginal log-likelihood at the end of each iteration is given by:

$$\log \underline{p}(\boldsymbol{y}; \mathfrak{q}) = E_{\mathfrak{q}} [\log \{\mathfrak{p}(\boldsymbol{y}, \boldsymbol{\theta})\}] - E_{\mathfrak{q}} [\log \{\mathfrak{q}(\boldsymbol{\theta})\}].$$

The steps of MFVB algorithm are as follows:

- Select a parametric family for the variational distributions $\mathfrak{q}_j(\boldsymbol{\theta}_j)$, such as a Gaussian with mean and variance as parameters and initialize the variational distributions $\mathfrak{q}_1(\boldsymbol{\theta}_1), \ldots, \mathfrak{q}_d(\boldsymbol{\theta}_d)$.
- Iterative updates are used:

$$\mathfrak{q}_1(\boldsymbol{\theta}_1) \longleftarrow \frac{\exp\left[E_{\mathfrak{q}(-\boldsymbol{\theta}_1)}\left\{\log \mathfrak{p}(\boldsymbol{y}, \boldsymbol{\theta})\right\}\right]}{\int \exp\left[E_{\mathfrak{q}(-\boldsymbol{\theta}_1)}\left\{\log \mathfrak{p}(\boldsymbol{y}, \boldsymbol{\theta})\right\}\right] d\boldsymbol{\theta}_1}$$

$$\vdots$$

$$\mathfrak{q}_d(\boldsymbol{\theta}_d) \longleftarrow \frac{\exp\left[E_{\mathfrak{q}(-\boldsymbol{\theta}_d)}\left\{\log \mathfrak{p}(\boldsymbol{y}, \boldsymbol{\theta})\right\}\right]}{\int \exp\left[E_{\mathfrak{q}(-\boldsymbol{\theta}_d)}\left\{\log \mathfrak{p}(\boldsymbol{y}, \boldsymbol{\theta})\right\}\right] d\boldsymbol{\theta}_d}$$

until the increase in lower bound on the marginal likelihood $\underline{p}(\boldsymbol{y}; \mathfrak{q})$ is negligible.

Although the forms of the approximating density functions $\mathfrak{q}_j(\boldsymbol{\theta}_j)$ are not specified, the structure of the statistical model ensures that the solution falls within a specific parametric family for each of the approximating density functions $\mathfrak{q}_j(\boldsymbol{\theta}_j)$. If all parameters in a model are conditionally conjugate, the approximating density functions can be expressed in closed forms. If the form of the approximating density function $\mathfrak{q}_j(\boldsymbol{\theta}_j)$ can not be recognized as a standard distribution, numerical integration methods need to be used to estimate the marginal likelihood, making the computation more demanding.

Upon convergence, the MFVB algorithm shows that the optimal occurs when

$$\mathfrak{q}_j^*(\boldsymbol{\theta}_j) \propto \exp\left[E_{\mathfrak{q}(-\boldsymbol{\theta}_j)}\left\{\log \mathfrak{p}(\boldsymbol{\theta}, \boldsymbol{y})\right\}\right], \ \ 1 \leq j \leq d.$$

This implies that the logarithm of the optimal solution for each factor is obtained by taking the logarithm of the joint density function over all random and observed variables and calculating its expectation with respect to all other factors $\mathfrak{q}_k(\boldsymbol{\theta}_k), k \neq j$ (Ormerod and Wand, 2010). An alternative expression for $\mathfrak{q}_j^*(\boldsymbol{\theta}_j)$ is

$$\mathfrak{q}_j^*(\boldsymbol{\theta}_j) \propto \exp\left[E_{\mathfrak{q}(-\boldsymbol{\theta}_j)}\left\{\log \mathfrak{p}\left(\boldsymbol{\theta}_j \middle| \text{rest}\right)\right\}\right], \ \ 1 \leq j \leq d,$$

where

- "rest" $\equiv$ all of the random variables excluding $\boldsymbol{\theta}_j$, i.e. $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{j-1}, \boldsymbol{\theta}_{j+1}, \ \ldots, \boldsymbol{\theta}_d\}$

  and

- $\mathfrak{p}\left(\boldsymbol{\theta}_j \middle| \text{rest}\right) \equiv$ the full conditional density function of $\boldsymbol{\theta}_j$ based on DAG related concept in **Theorem 1.9.1**.

It immediately follows that

$$\mathfrak{q}_j^*(\boldsymbol{\theta}_j) \propto \exp\left[E_{\mathfrak{q}(-\boldsymbol{\theta}_j)}\left\{\log \mathfrak{p}\left(\boldsymbol{\theta}_j \middle| \text{Markov blanket of } \boldsymbol{\theta}_j\right)\right\}\right], \ 1 \leq j \leq d.$$

This is knows as the *locality property of DAGs*. For large DAGs, this property offers significant algebraic advantages. In particular, it shows that the optimal $\mathfrak{q}$-density function $\mathfrak{q}_j^*(\boldsymbol{\theta}_j)$ rely solely on local computations on the model's DAG.

For the description of the DAG, please refer to Section 1.9.

1.14 Matrix Results

From Searle (1982), we have the following matrix results using in this thesis.

Result 1.14.1 If $\boldsymbol{A}$ and $\boldsymbol{B}$ are $m \times n$ matrices, then $\text{tr}\left(\boldsymbol{A}^T\boldsymbol{B}\right) = \text{vec}(\boldsymbol{A})^T \text{vec}(\boldsymbol{B})$.

Result 1.14.2 Let $\boldsymbol{v}$ be a random vector. Then

$$E\left(\boldsymbol{v}\boldsymbol{v}^T\right) = E(\boldsymbol{v})E(\boldsymbol{v})^T + \text{Cov}(\boldsymbol{v}) \ \text{ and } \ E\left(\left\|\boldsymbol{v}\right\|^2\right) = \left\|E(\boldsymbol{v})\right\|^2 + \text{tr}\left\{\text{Cov}(\boldsymbol{v})\right\}.$$

Result 1.14.3 Let $\boldsymbol{v}$ be a random vector and let $\boldsymbol{A}$ be a fixed matrix with the same number of rows as $\boldsymbol{v}$. Then

$$E\left(\boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v}\right) = E(\boldsymbol{v})\boldsymbol{A}E(\boldsymbol{v})^T + \text{tr}\left\{\boldsymbol{A}\text{Cov}(\boldsymbol{v})\right\}.$$

Result 1.14.4 Let $\boldsymbol{A}$ be a symmetric invertible matrix and $\boldsymbol{x}$ and $\boldsymbol{b}$ be column vectors with the same number of rows as $\boldsymbol{A}$. Then

$$\frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x} = -\frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{A}^{-1}\boldsymbol{b}\right)^T\boldsymbol{A}\left(\boldsymbol{x} - \boldsymbol{A}^{-1}\boldsymbol{b}\right) + \frac{1}{2}\boldsymbol{b}^T\boldsymbol{A}^{-1}\boldsymbol{b}.$$

Result 1.14.5 If $\boldsymbol{a}$ and $\boldsymbol{b}$ are vectors of the same length, then

$$\left\|\boldsymbol{a} - \boldsymbol{b}\right\|^2 = \left\|\boldsymbol{a}\right\|^2 - 2\boldsymbol{a}^T\boldsymbol{b} + \left\|\boldsymbol{b}\right\|^2.$$

1.15 Thesis Structure

Having provided the necessary background information, we now proceed to present our novel research on the three-category GAM selection. Chapter 2 explores the

MCMC approach with use of the Laplace-Zero/Grouped Lasso-Zero priors on Gaussian and binary responses and the application on real data sets with methodology implemented. Chapter 3 investigates the Grouped Horseshoe distribution properties and explores the MCMC approach with use of the (Grouped) Horseshoe priors on Gaussian and count responses with simulation study. Chapter 4 continuous from Chapter 2 by applying the MFVB approach as a faster option and the application on real data sets with methodology implemented. Chapter 5 shows the simulation study of Chapter 2 and Chapter 4. Chapter 6 extends the work in Chapter 2 and Chapter 4 to the three-category GAMM selection with random intercept in grouped data with simulation study. Chapter 7 applies the methodology given in Chapter 6 to the Australian Red Cross Blood Service data.

1.A Derivation of Result 1.8.2

Suppose that $x|a \sim N\left(0, \sigma^2/a\right), a \sim$ Inverse-Gamma $\left(1, \frac{1}{2}\right)$, we will show that

$$\mathfrak{p}(x) = \frac{1}{2\sigma} \exp\left(-\frac{|x|}{\sigma}\right),$$

which is equivalent to

$$x \sim \text{Laplace}(0, \sigma),$$

with 0 as location parameter and σ as scale parameter.

Derivation:
Firstly,

$$x|a \sim N\left(0, \sigma^2/a\right), a \sim \text{Inverse-Gamma}\left(1, \frac{1}{2}\right).$$

is equivalent to

$$x|a \sim N\left(0, a\sigma^2\right), a \sim \text{Gamma}\left(1, \frac{1}{2}\right).$$

Then

$$\begin{aligned}
\mathfrak{p}(x) &= \int_{-\infty}^{+\infty} \mathfrak{p}(x, a)\, da = \int_{-\infty}^{+\infty} \mathfrak{p}\left(x|a\right) \mathfrak{p}(a)\, da \\
&= \int_{0}^{+\infty} \frac{1}{\sqrt{2\pi a\sigma^2}} \exp\left(-\frac{x^2}{2a\sigma^2}\right) \frac{\left(\frac{1}{2}\right)^1}{\Gamma(1)} \left(a\right)^{1-1} \exp\left(-\frac{a}{2}\right) da \\
&= \frac{1}{2\sigma\sqrt{2\pi}} \int_{0}^{+\infty} a^{\left(-\frac{1}{2}\right)} \exp\left\{-\frac{1}{2}\left(\frac{x^2}{a\sigma^2} + a\right)\right\} da.
\end{aligned}$$

Let us look at the term $-\frac{1}{2}\left(\frac{x^2}{a\sigma^2} + a\right)$ inside the integral. We have

$$-\frac{1}{2}\left(\frac{x^2}{a\sigma^2} + a\right) = -\frac{\left(a\sigma - |x|\right)^2}{2a\sigma^2} - \frac{|x|}{\sigma}.$$

Therefore

$$\mathfrak{p}(x) = \frac{1}{2\sigma}\exp\left(-\frac{|x|}{\sigma}\right)\int_0^{+\infty}\frac{1}{\sqrt{2\pi}}(a)^{\left(-\frac{1}{2}\right)}\exp\left(-\frac{\left(a\sigma-|x|\right)^2}{2a\sigma^2}\right)da.$$

The result is proven if we can show that, for all $x \in \mathbb{R}$,

$$\int_0^{+\infty}\frac{1}{\sqrt{2\pi}}(a)^{\left(-\frac{1}{2}\right)}\exp\left(-\frac{(a\sigma-|x|)^2}{2a\sigma^2}\right)da = 1. \tag{1.7}$$

In the next part, firstly, we show (1.7) for $x = 0$. Then we show (1.7) also holds for $x \neq 0$.

The $x = 0$ case:

When $x = 0$, the left-hand side of (1.7) becomes

$$\begin{aligned}
&\int_0^{+\infty}\frac{1}{\sqrt{2\pi}}a^{\left(-\frac{1}{2}\right)}\exp\left(-\frac{a}{2}\right)da = \frac{1}{\sqrt{2\pi}}\int_0^{+\infty}a^{\left(\frac{1}{2}-1\right)}\exp\left(-\frac{a}{2}\right)da \\
&= \frac{1}{\sqrt{2\pi}}\left[\int_0^{+\infty}\frac{\frac{1}{2}^{\left(\frac{1}{2}\right)}}{\Gamma\left(\frac{1}{2}\right)}a^{\left(\frac{1}{2}-1\right)}\exp\left(-\frac{a}{2}\right)da\right]\Gamma\left(\frac{1}{2}\right)2^{\left(\frac{1}{2}\right)} \\
&= \frac{\sqrt{2}\Gamma\left(\frac{1}{2}\right)}{\sqrt{2}\sqrt{\pi}} = 1.
\end{aligned}$$

Since the last-written integral is that of the Gamma $\left(\frac{1}{2},\frac{1}{2}\right)$ distribution and $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$, (1.7) holds for $x = 0$.

The $x \neq 0$ case:

Make the following substitution for the integral in (1.7):

$$b = a^{-1} \text{ then } a = b^{-1} \text{ and}$$

$$\frac{da}{db} = -b^{-2} \text{ then } da = -b^{-2}db.$$

The integral in (1.7) becomes:

$$\begin{aligned}
&\int_0^{+\infty} \frac{1}{\sqrt{2\pi}} a^{\left(-\frac{1}{2}\right)} \exp\left(-\frac{\left(a\sigma - \left|x\right|\right)^2}{2a\sigma^2}\right) da \\
&= \int_{+\infty}^{0} \frac{1}{\sqrt{2\pi}} b^{\left(\frac{1}{2}\right)} \exp\left(\frac{-\left(b^{-1}\sigma - \left|x\right|\right)^2}{2b^{-1}\sigma^2}\right) \left(-b^{-2}\right) db \\
&= \int_0^{+\infty} \frac{1}{\sqrt{2\pi b^3}} \exp\left(\frac{-\left(b - \frac{\sigma}{|x|}\right)^2}{2b\left(\frac{\sigma}{|x|}\right)^2}\right) db.
\end{aligned}$$

The last-written integral is that of an Inverse-Gaussian density function with mean $\frac{\sigma}{|x|}$. Hence, the integral is 1. Therefore, (1.7) holds for all $x \neq 0$ as well.

Hence, $\mathfrak{p}(x) = \frac{1}{2\sigma} \exp\left(-\frac{|x|}{\sigma}\right)$, which is equivalent to $x \sim \text{Laplace}(0, \sigma)$.

From the above two cases, the following is approved:

if $x|a \sim N\left(0, \sigma^2/a\right)$ and $a \sim \text{Inverse-Gamma}\left(1, \frac{1}{2}\right)$, then

$$x \sim \text{Laplace}(0, \sigma) \text{ and } \mathfrak{p}(x) = \frac{1}{2\sigma} \exp\left(-\frac{|x|}{\sigma}\right).$$

1.B Derivation of Result 1.8.3

Suppose that $\boldsymbol{x}\big|a \sim N\left(\mathbf{0}, \left(1/a\right)\sigma^2\boldsymbol{I}_d\right)$. $\boldsymbol{x}$ is a $d \times 1$ vector and

$$a \sim \text{Inverse-Gamma}\left(\frac{d+1}{2}, \frac{1}{2}\right).$$

We will show that

$$\mathfrak{p}\left(x\right) \propto \exp\left(-\frac{\left\|\boldsymbol{x}\right\|_2}{\sigma}\right)$$

where $\left\|\boldsymbol{x}\right\|_2 \equiv \sqrt{x_1^2 + x_2^2 + \ldots + x_d^2}$.

Derivation:
Firstly,

$$\boldsymbol{x}\big|a \sim N\left(\mathbf{0}, \left(1/a\right)\sigma^2\boldsymbol{I}_d\right) \text{ and } a \sim \text{Inverse-Gamma}\left(\frac{d+1}{2}, \frac{1}{2}\right)$$

is equivalent to

$$\boldsymbol{x}\big|a \sim N\left(\mathbf{0}, a\sigma^2\boldsymbol{I}_d\right) \text{ and } a \sim \text{Gamma}\left(\frac{d+1}{2}, \frac{1}{2}\right).$$

Then

$$\begin{aligned}
\mathfrak{p}\left(\boldsymbol{x}\right) &= \int_{-\infty}^{+\infty} \mathfrak{p}\left(\boldsymbol{x}, a\right) da = \int_{-\infty}^{+\infty} \mathfrak{p}\left(\boldsymbol{x}\big|a\right) \mathfrak{p}\left(a\right) da \\
&= \int_{-\infty}^{+\infty} \left(2\pi a\sigma^2\right)^{-\frac{d}{2}} \exp\left(-\frac{\left\|\boldsymbol{x}\right\|_2^2}{2a\sigma^2}\right) \frac{\frac{1}{2}^{\left(\frac{d+1}{2}\right)}}{\Gamma\left(\frac{d+1}{2}\right)} a^{\left(\frac{d+1}{2}-1\right)} \exp\left(-\frac{a}{2}\right) da \\
&= \frac{\exp\left(-\frac{\left\|\boldsymbol{x}\right\|_2}{\sigma}\right)}{\left(2\sigma^2\right)^{(d)} \pi^{\left(\frac{d-1}{2}\right)} \Gamma\left(\frac{d+1}{2}\right)} \int_0^{+\infty} \frac{1}{\sqrt{2\pi}} a^{\left(-\frac{1}{2}\right)} \exp\left\{-\frac{\left(a\sigma - \left\|\boldsymbol{x}\right\|_2\right)^2}{2a\sigma^2}\right\} da.
\end{aligned}$$

Now we make the claim that

$$\int_0^{+\infty} \frac{1}{\sqrt{2\pi}} a^{\left(-\frac{1}{2}\right)} \exp\left\{-\frac{\left(a\sigma - \left\|\boldsymbol{x}\right\|_2\right)^2}{2a\sigma^2}\right\} da = 1 \tag{1.8}$$

for all $d \times 1$ vector of $\boldsymbol{x}$.

Let $\left|h\right| \equiv \left\|\boldsymbol{x}\right\|_2$, then (1.8) is equivalent to

$$\int_0^{+\infty} \frac{1}{\sqrt{2\pi}} a^{\left(-\frac{1}{2}\right)} \exp\left\{-\frac{\left(a\sigma - \left|h\right|\right)^2}{2a\sigma^2}\right\} da = 1 \tag{1.9}$$

for all $h \in \mathbb{R}$.

(1.9) is approved (with x instead of h) in (1.7) for **Result 1.8.2**.

In summary,

$$\mathfrak{p}\left(\boldsymbol{x}\right) = \frac{\exp\left(-\frac{\left\|\boldsymbol{x}\right\|_2}{\sigma}\right)}{\left(2\sigma^2\right)^{(d)} \pi^{\left(\frac{d-1}{2}\right)} \Gamma\left(\frac{d+1}{2}\right)}$$

where

$$\left\|\boldsymbol{x}\right\|_2 = \sqrt{x_1^2 + x_2^2 + \ldots + x_d^2}.$$

Therefore

$$\mathfrak{p}\left(\boldsymbol{x}\right) \propto \exp\left(-\frac{\left\|\boldsymbol{x}\right\|_2}{\sigma}\right)$$

where $\left\|\boldsymbol{x}\right\|_2 \equiv \sqrt{x_1^2 + x_2^2 + \ldots + x_d^2}$ and

$$\mathfrak{p}\left(\boldsymbol{x}\right) \sim \text{Laplace}\left(\boldsymbol{0}, \sigma \boldsymbol{I}_d\right).$$

Chapter 2

Generalized Additive Model Selection Via Markov Chain Monte Carlo

Declaration The method and results presented in this chapter are presented in the paper *Bayesian Generalized Additive Model Selection Including a Fast Variational Option* (He and Wand 2024). I certify that this publication was a direct result of my research towards this PhD, under the supervision of my principal supervisor Prof Matt Wand, and the reproduction in this thesis does not breach any copyright regulations.

2.1 Introduction

Nowadays many data sets have a large number of predictors. When we use non-linear regression analytic methods, such as generalized additive models, one of the most important concerns is to find out which predictors impact to the response. In this chapter, we aim to categorize each predictor's effect type into one of three classes: having zero effect, having a linear effect or having a non-linear effect on the mean response, and the forms of their possibly non-linear effects. We call this as the three-category model selection problem of generalized additive models.

For additive models, there are penalty-based model selection algorithms similar to the least absolute shrinkage and selection operator (LASSO) that are available in several R packages. For example, cosso (Zhang and Lin, 2013) uses component selection and smoothing operation. It can determine predictor's effect is either non-linear or zero, but designation of linear effects are not part of the methodology.

Chouldechova and Hastie (2015) achieved categorizing predictors' effect types with the frequentist approach by the LASSO, Demmler-Reinsch spline bases, regularization paths and cross-validatory selection of the regularization parameter with R package gamsel (Chouldechova and Hastie, 2022). We call this as the regularization path approach in this thesis. It has the attraction of obtaining a family of selected models along the regularization path. If one wants to choose a single model then the cross-validation needs to be used to get the global cross-validation minima, which is not always easy in practice.

Our work is to produce new algorithms with two approaches under Bayesian fitting and inference for both Gaussian and binary responses, for the three-category model selection problem for generalized additive models. We build on the LASSO and Demmler-Reinsch spline bases to maintain the orthogonality. The first approach is Markov chain Monte Carlo (MCMC), described in Chapter 2. The second approach is mean field variational Bayes (MFVB), which is the fast option for model selection, described in Chapter 4. Our algorithms of both MCMC and MFVB are implemented in an R package gamselBayes Simulation studies and application on actual data are described in Chapter 5.

To make our methodology being independent of units of measurement, we use standardization forms of the original input data in the Bayesian fitting and inference, described in Section 2.2.2.

To allow sparseness of coefficient vectors and get the resulting sparse fits, we use

the Laplace-Zero priors. However, the Laplace-Zero distribution is not amenable to Markov chain Monte Carlo and mean field variational Bayes fitting algorithms. To solve this problem, we use tailored auxiliary variable representations, described in Section 2.3.4, Section 2.3.5 and Section 2.3.6.

For Gaussian responses, the description of the full model is described in Section 2.4.2. For binary responses, the description of the full model is described in Section 2.4.3.

To implement practical fitting algorithms for Markov chain Monte Carlo, we use Markov blanket to work out the posterior distributions of each parameters in our full model, described in Section 2.5.3. The posterior distributions are all in standard forms for practical implementation. Therefore, the Markov chain Monte Carlo sampling reduces to Gibbs sampling for the model. The MCMC fitting algorithm listing is given in Algorithm 2 in Section 2.5.3 for implementation into R. The sufficient statistic matrices described in Section 2.5.1 offers computational speed-ups. Notation for fitting algorithms is described in Section 2.5.2. Based on posterior distributions of MCMC, the model selection strategies and simulation studies for default value of threshold parameter described in Section 2.6. Further simulation study of classification performance comparison among our MCMC, MFVB and other methods described in Chapter 5.

Throughout Chapter 2, "rest" denotes the set of other random variables in model (2.28).

The additive structure of generalized additive model (GAM) described in this chapter is given in Section 2.3. The independence assumptions of GAM is given in (2.28) for Gaussian responses and Section 2.4.3 for binary responses. The additive part uses splines for non-linear effects. The predictors include binary predictors and continuous predictors. The number of binary predictors is $d_\circ$. The number of continuous predictors is $d_\bullet$.

For generalized additive model described in Section 2.2–2.5, the sample size is n. The response $\boldsymbol{y}$ is a $n \times 1$ vector. The linear design matrix $\boldsymbol{X}$ is a $n \times (d_\circ + d_\bullet)$ matrix. The $\boldsymbol{Z}_j$ matrix containing the canonical Demmler-Reinsch basis is a $n \times K_j$ matrix, with $1 \leq j \leq d_\bullet$, K_j as the number of basis functions for the j-th continuous predictor with default value of 12. Throughout this chapter, $\boldsymbol{y}$ vector, $\boldsymbol{X}$ matrix and $\boldsymbol{Z}$ matrix are all fixed.

Our simulation study suggests a good default threshold value τ as 0.5 for MCMC given in Section 2.6.3. The data were generated with 30 continuous predictors, 10

having "true" linear effects, 10 having "true" non-linear effects and 10 having "true" zero effects. The sample size varied as $n \in \{500, 1000, 2000\}$. The error standard deviation varied as $\sigma_\varepsilon \in \{0.25, 0.5, 1, 2\}$. The threshold varied as $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. For each combination of sample size and error standard deviation, we ran simulation on 100 independently generated data sets.

We applied Algorithm 2 to two real-world datasets – the mortgage and the educational dataset given in Section 2.7, to demonstrate the practical applicability of the approach. The mortgage dataset used in Section 2.7.1 contains $2,380$ observations and 18 predictors. The educational dataset used in Section 2.7.2 contains 420 observations and 9 predictors.

We conducted a simulation study on a prediction accuracy evaluation metric given in Figure 2.4 of Section 2.6.3. The threshold varied as $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. The sample size varied as $n \in \{500, 1000, 2000\}$. The error standard deviation varied as $\sigma_\varepsilon \in \{0.25, 0.5, 1, 2\}$. Moreover, hyperparameters sensitivity checks are given in Section 5.4.

2.2 Original Input Data and Transformed Data

2.2.1 The Original Input Data

The original input data are as follows:

$$\left(\mathring{\boldsymbol{x}}_i^{\text{orig}}, \dot{\boldsymbol{x}}_i^{\text{orig}}, y_i^{\text{orig}}\right), \quad 1 \leq i \leq n,$$

where

$\mathring{\boldsymbol{x}}_i^{\text{orig}} \equiv d_\circ \times 1$ vector of predictors that can only enter the model linearly (e.g. binary predictors),

$\dot{\boldsymbol{x}}_i^{\text{orig}} \equiv d_\bullet \times 1$ vector of predictors that can enter the model either linearly or non-linearly

and

$y_i^{\text{orig}} \equiv$ the ith entry of $\boldsymbol{y}$, $\boldsymbol{y}$ is a $n \times 1$ vector of original response data.

For $1 \leq i \leq n$, the jth original predictor data is denoted to be:

$\mathring{x}_{ij}^{\text{orig}} \equiv$ the jth entry of $\mathring{\boldsymbol{x}}_i^{\text{orig}}$, the jth original predictor can only enter the model linearly, $1 \leq j \leq d_\circ$

and

$\dot{x}_{ij}^{\text{orig}} \equiv$ the jth entry of $\dot{\boldsymbol{x}}_i^{\text{orig}}$, the jth original predictor can enter the model either linearly or non-linearly, $1 \leq j \leq d_\bullet$.

2.2.2 Standardization of the Input Data for Fitting and Model Selection

Standardization of the original input data makes our methodology independent of the units of measurement for fixed hyperparameter settings and also provides improved numerical stability.

The full data to be used for fitting and model selection are

$$(\mathring{\boldsymbol{x}}_i, \dot{\boldsymbol{x}}_i, y_i), \quad 1 \leq i \leq n,$$

where

$\mathring{\boldsymbol{x}}_i \equiv$ standardized data versions of $\mathring{\boldsymbol{x}}_i^{\text{orig}}$,

$\dot{\boldsymbol{x}}_i \equiv$ standardized data versions of $\dot{\boldsymbol{x}}_i^{\text{orig}}$

and

$y_i \equiv$ standardized data versions of y_i^{orig}.

For $1 \leq i \leq n$, after standardization, the jth transformed predictor data is de-

noted to be:

$\mathring{x}_{ij} \equiv$ the jth entry of $\mathring{\boldsymbol{x}}_i, 1 \leq j \leq d_\circ$, the jth transformed predictor can only enter the model linearly

and

$\dot{x}_{ij} \equiv$ the jth entry of $\dot{\boldsymbol{x}}_i, 1 \leq j \leq d_\bullet$, the jth transformed predictor can enter the model either linearly or non-linearly.

Then for fitting and model selection we replace the $\mathring{x}_{ij}^{\text{orig}}$ by $\mathring{x}_{ij} = \left(\mathring{x}_{ij}^{\text{orig}} - a_x\right)/b_x$ where

$$a_x \equiv \frac{1}{n}\sum_{i=1}^{n} \mathring{x}_{ij}^{\text{orig}} \quad \text{and} \quad b_x \equiv \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\mathring{x}_{ij}^{\text{orig}} - a_x\right)^2}, \tag{2.1}$$

with $1 \leq j \leq d_\circ$ for predictors can only enter the model linearly, $1 \leq i \leq n$.

And we replace the $\dot{x}_{ij}^{\text{orig}}$ by $\dot{x}_{ij} = \left(\dot{x}_{ij}^{\text{orig}} - c_x\right)/d_x$ where

$$c_x \equiv \frac{1}{n}\sum_{i=1}^{n} \dot{x}_{ij}^{\text{orig}} \quad \text{and} \quad d_x \equiv \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\dot{x}_{ij}^{\text{orig}} - c_x\right)^2}, \tag{2.2}$$

with $1 \leq j \leq d_\bullet$ for predictors can enter the model either linearly or non-linearly, $1 \leq i \leq n$.

For $1 \leq i \leq n$, this transformation of the original predictor data implies that the transformed predictor data are *centred* in that

$$\sum_{i=1}^{n} \mathring{x}_{ij} = 0, \;\; 1 \leq j \leq d_\circ \;\; \text{and} \;\; \sum_{i=1}^{n} \dot{x}_{ij} = 0, \;\; 1 \leq j \leq d_\bullet. \tag{2.3}$$

For Gaussian responses, the standardized response data for fitting and model selection are made scale-free as follows:

$$y_i = y_i^{\text{orig}}/s_y, \;\; s_y \equiv \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i^{\text{orig}} - \overline{y}^{\text{orig}}\right)^2}, \;\; 1 \leq i \leq n, \tag{2.4}$$

is the sample standard deviation of the y_i^{orig}s.

For binary responses, the y_is are not pre-processed and remain as values in $\{0, 1\}$.

Algorithm 1 in Section 2.5.1 provides the operational details of the standardiza-

tion process.

2.3 Design Matrices and Auxiliary Variable Representations for Parameters

2.3.1 Generalized Additive Model Form for the Full Model

For Gaussian responses, we have $g(\mu) = \mu$ (referred to Section 1.10 for the choice of $g(\cdot)$). This is called the identity link. It directly models the mean of the response without transformation. Here, $g(\cdot)$ is of the mean of $\boldsymbol{y}$, with $\mathring{x}_{ij}$ fixed, with $1 \leq i \leq n$ and $1 \leq j \leq d_{\circ}$, and with $\dot{\bullet}_{ij}$ fixed, with $1 \leq i \leq n$ and $1 \leq j \leq d_{\bullet}$.

Generalized additive models involve linear predictors having generic form

$$y_{ij} = \beta_0 + \sum_{j=1}^{d_{\circ}} \beta_j \mathring{x}_{ij} + \sum_{j=1}^{d_{\bullet}} f_j\left(\dot{\bullet}_{ij}\right) \tag{2.5}$$

for scalar parameters $\beta_0, ..., \beta_{d_{\circ}}$ and the f_j are smooth real-valued functions over an interval containing the $\dot{\bullet}_{ij}$ data, $1 \leq i \leq n$.

From (2.5), we have below detailed form for our generalized additive full model

$$\begin{aligned}
&\beta_0 + \sum_{j=1}^{d_{\circ}} \left\{(\mathbf{1}_n \beta_j) \odot \mathring{\boldsymbol{x}}_j\right\} + \sum_{j=1}^{d_{\bullet}} \left\{(\mathbf{1}_n \beta_j) \odot \dot{\boldsymbol{x}}_j\right\} + \sum_{j=1}^{d_{\bullet}} \left\{\sum_{k=1}^{K_j} (\mathbf{1}_n u_{jk}) \odot \dot{\boldsymbol{Z}}_{jk}\right\} \\
&= \beta_0 + \sum_{j=1}^{d_{\circ}+d_{\bullet}} \left\{(\mathbf{1}_n \beta_j) \odot \dot{\boldsymbol{X}}_j\right\} + \sum_{j=1}^{d_{\bullet}} \left\{\sum_{k=1}^{K_j} (\mathbf{1}_n u_{jk}) \odot \dot{\boldsymbol{Z}}_{jk}\right\} \\
&= \beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \sum_{j=1}^{d_{\bullet}} \boldsymbol{Z}_j \boldsymbol{u}_j \\
&= \beta_0 + \boldsymbol{X}\left(\boldsymbol{\gamma}_{\beta} \odot \widetilde{\boldsymbol{\beta}}\right) + \sum_{j=1}^{d_{\bullet}} \boldsymbol{Z}_j \left(\gamma_{uj} \widetilde{\boldsymbol{u}}_j\right)
\end{aligned} \tag{2.6}$$

where

$\boldsymbol{X} \equiv$ the *linear design matrix,* a $n \times (d_\circ + d_\bullet)$ matrix;

$\boldsymbol{\beta} \equiv \boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}$, $\boldsymbol{\beta}$ is the linear term coefficients vector for matrix $\boldsymbol{X}$, a $(d_\circ + d_\bullet) \times 1$ vector;

$\boldsymbol{\gamma}_\beta \equiv$ auxiliary probability parameters for $\boldsymbol{\beta}$, a $(d_\circ + d_\bullet) \times 1$ vector;

$\widetilde{\boldsymbol{\beta}} \equiv$ auxiliary parameters for $\boldsymbol{\beta}$, a $(d_\circ + d_\bullet) \times 1$ vector;

$\boldsymbol{Z}_j \equiv$ matrix containing the *canonical Demmler-Reinsch basis* for $\dot{\boldsymbol{x}}_j$, a $n \times K_j$ matrix , $1 \leq j \leq d_\bullet$;

$K_j \equiv$ The number of columns in $\boldsymbol{Z}_j$;

$\boldsymbol{u}_j \equiv \gamma_{uj}\widetilde{\boldsymbol{u}}_j$, $\boldsymbol{u}_j$ is the spline coefficients vector for matrix $\boldsymbol{Z}_j$, a $K_j \times 1$ vector , $1 \leq j \leq d_\bullet$;

$\gamma_{uj} \equiv$ auxiliary probability parameters for $\boldsymbol{u}_j$, a scalar , $1 \leq j \leq d_\bullet$

and

$\widetilde{\boldsymbol{u}}_j \equiv$ auxiliary parameters for $\boldsymbol{u}_j$, a $K_j \times 1$ vector , $1 \leq j \leq d_\bullet$.

Also denote

$$\dot{\boldsymbol{X}}_j = \begin{bmatrix} X_{1j} \\ \vdots \\ X_{ij} \\ \vdots \\ X_{nj} \end{bmatrix} \text{ with } 1 \leq i \leq n, \ 1 \leq j \leq d_\circ + d_\bullet,$$

and

$$\dot{\boldsymbol{Z}}_{jk} = \begin{bmatrix} Z_{1jk} \\ \vdots \\ Z_{ijk} \\ \vdots \\ Z_{njk} \end{bmatrix} \text{ with } 1 \leq i \leq n,\ \ 1 \leq j \leq d_{\bullet},\ \ 1 \leq k \leq K_j.$$

Throughout Chapter 2, $\boldsymbol{y}$ vector, $\boldsymbol{X}$ matrix and $\boldsymbol{Z}$ matrix are all fixed.

The linear design matrix $\boldsymbol{X}$, its linear term coefficients vector $\boldsymbol{\beta}$, and $\boldsymbol{\beta}$'s auxiliary parameters $\gamma_{\beta j}$, $\widetilde{\beta}_j, 1 \leq j \leq d_{\circ} + d_{\bullet}$, used in our full model are described in Section 2.3.2 and Section 2.3.4 respectively. The matrix containing the *canonical Demmler-Reinsch basis* $\boldsymbol{Z}_j$, its spline coefficient vector $\boldsymbol{u}_j$, and $\boldsymbol{u}_j$'s auxiliary parameters $\boldsymbol{\gamma}_{u\,j}$, $\widetilde{\boldsymbol{u}}_j, 1 \leq j \leq d_{\bullet}$, used in our full model are described in Section 2.3.3 and Section 2.3.5 respectively.

2.3.2 The Linear Design Matrix

The linear design matrix is the following $n \times (d_{\circ} + d_{\bullet})$ matrix:

$$\boldsymbol{X} \equiv \begin{bmatrix} \mathring{\boldsymbol{x}}_1^T & \dot{\boldsymbol{x}}_1^T \\ \vdots & \vdots \\ \mathring{\boldsymbol{x}}_n^T & \dot{\boldsymbol{x}}_n^T \end{bmatrix}.$$

Let $\boldsymbol{\beta}$ be the $(d_{\circ} + d_{\bullet}) \times 1$ vector of corresponding coefficients for linear effects. $\mathring{\boldsymbol{x}}_i$ is a $d_{\circ} \times 1$ vector, which stands for standardized predictors can only enter the model linearly, and $\dot{\boldsymbol{x}}_i$ is a $d_{\bullet} \times 1$ vector, which stands for standardized predictors can enter the model either linearly or non-linearly, $1 \leq i \leq n$.

From the standardization of original input data descried in Section 2.2.2 and (2.3), it offers simplification to be

$$\mathbf{1}_n^T \boldsymbol{X} = \mathbf{0} \tag{2.7}$$

2.3.3 Canonical Demmler-Reinsch Basis Design Matrices

We denote $\dot{\boldsymbol{x}}_j$ as a $n \times 1$ vector as follows, with $1 \leq j \leq d_\bullet$:

$$\dot{\boldsymbol{x}}_j \equiv \begin{bmatrix} \dot{x}_{1j} \\ \vdots \\ \dot{x}_{nj} \end{bmatrix}.$$

The construction of $\boldsymbol{Z}_j$ matrix, containing the canonical Demmler-Reinsch basis for the predictor data $\dot{\boldsymbol{x}}_j$, is described in Appendix 2.A, $1 \leq j \leq d_\bullet$. We will refer to the columns of $\boldsymbol{Z}_j$ as the canonical Demmler-Reinsch basis for the $\dot{x}_j$s.

The $\boldsymbol{Z}$ matrix we construct according to Appendix 2.A satisfies:

$$\boldsymbol{Z}_j^T \mathbf{1}_n = \boldsymbol{Z}_j^T \dot{\boldsymbol{x}}_j = \mathbf{0}_{K_j} \quad \text{and} \quad \boldsymbol{Z}_j^T \boldsymbol{Z}_j \text{ is a diagonal matrix}, \quad 1 \leq j \leq d_\bullet. \tag{2.8}$$

Spline bases satisfying (2.8) are referred to as having a *Demmler-Reinsch* form.

We require a version of $\boldsymbol{Z}_j$ for $\dot{\boldsymbol{x}}_j$, $1 \leq j \leq d_\bullet$. Denote the canonical Demmler-Reinsch basis matrices by

$$\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{d_\bullet}.$$

with $\boldsymbol{Z}_j$ having dimension $n \times K_j$. Let

$$\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{d_\bullet}$$

be the corresponding coefficient vectors, where $\boldsymbol{u}_j$ has dimension $K_j \times 1$.

2.3.4 Linear Term Coefficients with Auxiliary Variable Representations

In our full model, let $\boldsymbol{\beta}$ denote a $(d_\circ + d_\bullet) \times 1$ vector of linear term coefficients for the linear design matrix $\boldsymbol{X}$ described in Section 2.3.2. For each $\beta_j, 1 \leq j \leq d_\circ + d_\bullet$, it has the following *Laplace-Zero* family of prior distributions for the sparseness of coefficients:

$$\mathfrak{p}\left(\beta_j | \sigma_\beta, \rho_\beta\right) = \rho_\beta \left(2\sigma_\beta\right)^{-1} \exp\left(-|\beta_j|/\sigma_\beta\right) + \left(1 - \rho_\beta\right) \delta_0\left(\beta_j\right), 1 \leq j \leq d_\circ + d_\bullet \tag{2.9}$$

for parameters $\sigma_\beta > 0$ and $0 \leq \rho_\beta \leq 1$. δ_0 denotes the Dirac delta function at zero. The Laplace-Zero distribution is a "spike-and-slab" mixture of a Laplace density function and a point mass at zero (e.g. Ishwaran and Rao, 2005).

When $\rho_\beta = 1$, (2.9) corresponds to the Bayesian LASSO approach descried in Park and Casella (2008). In that article it is pointed out Bayes estimation does not lead to sparse fits for the purely Laplace situation. Therefore, we use Laplace-Zero priors with the attraction of having a point mass at zero of posterior distributions that induces a sparse solution. This aspect is further exploited in Chapter 4 for principled model selection strategies.

The Laplace-Zero distribution is not amenable to Markov chain Monte Carlo and mean field variational Bayes fitting algorithms due to its non-standard full conditional distributions. We re-express (2.9) using auxiliary parameters $b_{\beta j}$, and $\gamma_{\beta j}, 1 \leq j \leq d_\circ + d_\bullet$, and auxiliary parameter $\widetilde{\boldsymbol{\beta}}$. In our full model, we set

$$\boldsymbol{\beta} \equiv \boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}. \tag{2.10}$$

$$\widetilde{\beta}_j | b_{\beta j}, \sigma_\beta^2 \sim N\left(0, \sigma_\beta^2 / b_{\beta j}\right), \;\; b_{\beta j} \sim \text{Inverse-Gamma}\left(1, \frac{1}{2}\right), \;\; 1 \leq j \leq d_\circ + d_\bullet. \tag{2.11}$$

$$\gamma_{\beta j} \sim \text{Bernoulli}\left(\rho_\beta\right), 1 \leq j \leq d_\circ + d_\bullet. \tag{2.12}$$

From (2.11), using auxiliary variable representation stated in **Result 1.8.2**, we have

$$\widetilde{\beta}_j \sim \text{Laplace}\left(0, \sigma_\beta\right), \;\; 1 \leq j \leq d_\circ + d_\bullet. \tag{2.13}$$

From (2.13) and (2.12), using auxiliary variable representation stated in **Result 1.8.4**, we have

$$\beta_j \sim \text{Laplace-Zero}\left(0, \sigma_\beta^2, \rho_\beta\right), \;\; 1 \leq j \leq d_\circ + d_\bullet. \tag{2.14}$$

Therefore, (2.14) is equivalent to (2.9), which means each β_j in linear term coefficients vector $\boldsymbol{\beta}$, having Laplace-Zero family of prior distributions, $1 \leq j \leq d_\circ + d_\bullet$.

The scale and mixture parameters in (2.9) have the following prior distributions:

$$\sigma_\beta \sim \text{Half-Cauchy}(s_\beta), \tag{2.15}$$

for hyperparameters $s_\beta > 0, \;\; \rho_\beta = \frac{1}{2}$.

The Half-Cauchy distribution is not amenable to Markov chain Monte Carlo and mean field variational Bayes fitting algorithms due to its non-standard full condi-

tional distributions. In our full model, we set

$$\sigma_\beta^2|a_\beta \sim \text{Inverse-Gamma}\left(\frac{1}{2}, 1/a_\beta\right) \text{ and } a_\beta \sim \text{Inverse-Gamma}\left(\frac{1}{2}, 1/s_\beta^2\right). \quad (2.16)$$

Using auxiliary variable representation stated in **Result 1.8.1**, (2.16) is equivalent to (2.15), which means $\sigma_\beta \sim \text{Half-Cauchy}\,(s_\beta)$.

2.3.5 Spline Coefficient Vectors with Auxiliary Variable Representations

In our full model, let $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{d_\bullet}$ be the full set of spline coefficient vectors for the canonical Demmler-Reinsch basis matrices $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{d_\bullet}$, described in Section 2.3.3, where $\boldsymbol{u}_j$ is a $K_j \times 1$ vector for corresponding matrix $\boldsymbol{Z}_j$, where $\boldsymbol{Z}_j$ is a $n \times K_j$ matrix with $1 \leq j \leq d_\bullet$. For each $\boldsymbol{u}_j$, it has the following *group Laplace-Zero* family of prior distributions for the sparseness of spline coefficient vector:

$$\mathfrak{p}\left(\boldsymbol{u}_j|\sigma_{uj}, \rho_u\right) = \rho_u \left(C_{K_j}\sigma_{uj}\right)^{-1} \exp\left(-\left\|\boldsymbol{u}_j\right\|/\sigma_{uj}\right) + (1-\rho_u)\,\boldsymbol{\delta}_0\left(\boldsymbol{u}_j\right), 1 \leq j \leq d_\bullet. \quad (2.17)$$

for parameters $\sigma_{uj} > 0$ and $0 \leq \rho_{uj} \leq 1$ and with $C_{K_j} \equiv 2^{K_j}\pi^{(K_j-1)/2}\Gamma(\frac{1}{2}(K_j+1))$. Here $\left\|\boldsymbol{u}_j\right\| \equiv (\boldsymbol{u}_j^T\boldsymbol{u}_j)^{\frac{1}{2}}$ and δ_0 denotes the K_j-variate Dirac delta function at $\mathbf{0}_{K_j}$, the $K_j \times 1$ vector of zeros.

When $\rho_u = 1$, (2.17) is described in Kyung et al. (2010) using the phrase *group LASSO* for the family of priors. In Yuan and Lin (2006), it describes the extension of the ordinary LASSO is treating the elements of vector coefficients together as an entity. When penalty terms meet certain conditions, penalty terms of the form $\lambda\left\|\boldsymbol{\theta}\right\|$, for some $\lambda > 0$, all entries of the vector coefficients can be estimated as exactly zero. Chouldechova and Hastie (2015) using the frequentist approach to generalized additive model selection apply this idea to the spline coefficient vectors. Let $\widehat{\boldsymbol{u}}$ be an estimate of $\boldsymbol{u}$, and $\boldsymbol{u}$ is vector of spline coefficients. Then $\widehat{\boldsymbol{u}} = \mathbf{0}$ or $\widehat{\boldsymbol{u}} \neq \mathbf{0}$ allows for smooth function effects to be categorized as either linear or non-linear.

In order to invoke the sparse fits, keeping with (2.9) in Section 2.3.4, we extend the *group LASSO* distribution to a K_j-variate "spike-slab-form", $1 \leq j \leq d_\bullet$. Therefore, we have the *group Laplace-Zero* priors with the attraction of having a point mass of a $K_j \times 1$ vector of zeros.

The group LASSO distribution is not amenable to Markov chain Monte Carlo and mean field variational Bayes fitting algorithms due to its non-standard full con-

ditional distributions. We re-express (2.17) using auxiliary parameters b_{uj}, $\boldsymbol{\gamma}_{u\,j}$, and $\widetilde{\boldsymbol{u}}_j, 1 \leq j \leq d_\bullet$. In our full model, we set

$$\boldsymbol{u}_j \equiv \gamma_{uj}\widetilde{\boldsymbol{u}}_j, \quad 1 \leq j \leq d_\bullet. \tag{2.18}$$

$$\widetilde{\boldsymbol{u}}_j | b_{uj}, \sigma^2_{uj} \overset{\text{ind.}}{\sim} N\left(\mathbf{0}, \left(\sigma^2_{uj}/b_{uj}\right)\boldsymbol{I}_{K_j}\right), b_{uj} \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(\frac{1}{2}\left(K_j+1\right), \frac{1}{2}\right). \tag{2.19}$$

$$\gamma_{uj} \sim \text{Bernoulli}\left(\rho_u\right), 1 \leq j \leq d_\bullet. \tag{2.20}$$

From (2.19), using auxiliary variable representation stated in **Result 1.8.3**, we have

$$\widetilde{\boldsymbol{u}}_j \sim \text{Laplace}\left(\mathbf{0}, \sigma_{uj}\boldsymbol{I}_{K_j}\right). \tag{2.21}$$

From (2.21) and (2.20), using auxiliary variable representation stated in **Result 1.8.4**, we have

$$\boldsymbol{u}_j \sim \text{Laplace-Zero}\left(\mathbf{0}, \sigma^2_{uj}\boldsymbol{I}_{K_j}, \rho_u\right), 1 \leq j \leq d_\bullet. \tag{2.22}$$

Therefore, (2.22) is equivalent to (2.17), which means spline coefficient vector $\boldsymbol{u}_j$ having group Laplace-Zero family of prior distributions, $1 \leq j \leq d_\bullet$.

The scale and mixture parameters have the following prior distributions:

$$\sigma_{uj} \overset{\text{ind.}}{\sim} \text{Half-Cauchy}(s_u), \quad 1 \leq j \leq d_\bullet. \tag{2.23}$$

for hyperparameters $s_u > 0, \rho_u = \dfrac{1}{2}$.

The Half Cauchy distribution is not amenable to Markov chain Monte Carlo and mean field variational Bayes fitting algorithms due to its non-standard full conditional distributions. In our full model, we set

$$\sigma^2_{uj} | a_{uj} \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(\frac{1}{2}, 1/a_{uj}\right), a_{uj} \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(\frac{1}{2}, 1/s^2_u\right), 1 \leq j \leq d_\bullet. \tag{2.24}$$

Using auxiliary variable representation stated in **Result 1.8.1**, (2.24) is equivalent to (2.23), which means $\sigma_{uj} \overset{\text{ind.}}{\sim} \text{Half-Cauchy}\left(s_u\right), 1 \leq j \leq d_\bullet$.

2.3.6 Standard Deviation with Auxiliary Variable Representations

Our prior distribution for standard deviation σ_ε is:

$$\sigma_\varepsilon \sim \text{Half-Cauchy}\left(s_\varepsilon\right), \tag{2.25}$$

The Half Cauchy distribution is not amenable to Markov chain Monte Carlo and mean field variational Bayes fitting algorithms due to their non-standard full conditional distributions. In our full model, we set

$$\sigma_\varepsilon^2|a_\varepsilon \sim \text{Inverse-Gamma}\left(\frac{1}{2}, 1/a_\varepsilon\right) \text{ and } a_\varepsilon \sim \text{Inverse-Gamma}\left(\frac{1}{2}, 1/s_\varepsilon^2\right). \tag{2.26}$$

Using auxiliary variable representation stated in **Result 1.8.1**, (2.26) is equivalent to (2.25), which means $\sigma_\varepsilon \sim \text{Half-Cauchy}\left(s_\varepsilon\right)$.

2.4 Essence of the Bayesian Model

In this section we describe, in essence, the Bayesian model.

2.4.1 The Likelihood of the Response

Based on the linear design matrix $\boldsymbol{X}$ and its linear term coefficients vector $\boldsymbol{\beta}$ described in Section 2.3.2 and Section 2.3.4 respectively, and based on the matrix containing the canonical Demmler-Reinsch basis $\boldsymbol{Z}_j$ and its spline coefficient vector $\boldsymbol{u}_j, 1 \leq j \leq d_\bullet$, described in Section 2.3.3 and Section 2.3.5 respectively, and the standard deviation σ_ε described in Section 2.3.6, the likelihood of the response of our full model is:

$$\boldsymbol{y}|\beta_0, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}_1, \ldots, \widetilde{\boldsymbol{u}}_{d_\bullet}, \gamma_{u1}, \ldots, \gamma_{ud_\bullet}, \sigma_\varepsilon^2 \sim N\left(\mathbf{1}_n\beta_0, +\boldsymbol{X}(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) + \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j), \sigma_\varepsilon^2 \boldsymbol{I}_n\right). \tag{2.27}$$

2.4.2 The Full Bayesian Gaussian Response Additive Model Selection

For Gaussian response, based on the likelihood function in (2.27) given in Section 2.4.1 and model description described in Section 2.3, we set up the full model to be:

$$\boldsymbol{y}\big|\beta_0, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{\beta}}, \gamma_{u1}, \ldots, \gamma_{d_\bullet}, \widetilde{\boldsymbol{u}}_1, \ldots, \widetilde{\boldsymbol{u}}_{d_\bullet}, \sigma_\varepsilon^2 \sim$$
$$N\left(\mathbf{1}_n\beta_0 + \boldsymbol{X}(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) + \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j), \sigma_\varepsilon^2 \boldsymbol{I}_n\right), \;\; \beta_0 \sim N(0, \sigma_{\beta_0}^2),$$

$$\sigma_\varepsilon^2\big|a_\varepsilon \sim \text{Inverse-Gamma}\left(\tfrac{1}{2}, 1/a_\varepsilon\right), \;\; a_\varepsilon \sim \text{Inverse-Gamma}\left(\tfrac{1}{2}, 1/s_\varepsilon^2\right),$$
$$\gamma_{\beta j} \overset{\text{ind.}}{\sim} \text{Bernoulli}\left(\rho_\beta\right), \;\; \widetilde{\beta}_j\big|\sigma_\beta^2, b_{\beta j} \overset{\text{ind.}}{\sim} N\left(0, \sigma_\beta^2/b_{\beta j}\right), \;\; 1 \le j \le d_\circ + d_\bullet,$$
$$b_{\beta j} \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(1, \tfrac{1}{2}\right), \quad 1 \le j \le d_\circ + d_\bullet,$$
$$\gamma_{uj} \overset{\text{ind.}}{\sim} \text{Bernoulli}\left(\rho_u\right), \;\; 1 \le j \le d_\bullet,$$
$$\widetilde{\boldsymbol{u}}_j\big|\sigma_{uj}^2, b_{uj} \overset{\text{ind.}}{\sim} N\left(\mathbf{0}, (\sigma_{uj}^2/b_{uj})\boldsymbol{I}_{K_j}\right), \;\; b_{uj} \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(\tfrac{1}{2}(K_j+1), \tfrac{1}{2}\right), \;\; 1 \le j \le d_\bullet,$$
$$\sigma_\beta^2\big|a_\beta \sim \text{Inverse-Gamma}\left(\tfrac{1}{2}, 1/a_\beta\right), \;\; a_\beta \sim \text{Inverse-Gamma}\left(\tfrac{1}{2}, 1/s_\beta^2\right),$$
$$\sigma_{uj}^2\big|a_{uj} \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(\tfrac{1}{2}, 1/a_{uj}\right), \;\; a_{uj} \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(\tfrac{1}{2}, 1/s_u^2\right), \;\; 1 \le j \le d_\bullet. \tag{2.28}$$

The full set of hyperparameters in model (2.28) is:

$$\sigma_{\beta_0} = 10^5, \;\; s_\beta = s_\varepsilon = s_u = 1000, \;\; \rho_\beta = \rho_u = \tfrac{1}{2}.$$

The distributional assumptions on the parameters using tailored auxiliary variable representations, stated in (2.28), enable Gibbs sampling for the MCMC. Gibbs sampling is fairly straightforward since the full conditional distributions of parameters all have standard forms, which is a must for the derivation of the closed forms updates for the MFVB algorithm described in Chapter 4. The specific values of hyperparameters and the initial value used enable non-informative priors, to ensure that the posterior inference is driven primarily by the data, rather than subjective or arbitrary prior assumptions.

Figure 2.1 shows the directed acyclic graph corresponding to model (2.28).

The directed acyclic graph (DAG) Figure 2.1 is useful for determination of the full conditional distributions of parameters for Markov chain Monte Carlo fitting algorithms. This is due to the fact that the full conditional distribution of any node on the graph is the same as the distribution of the node conditional on its Markov

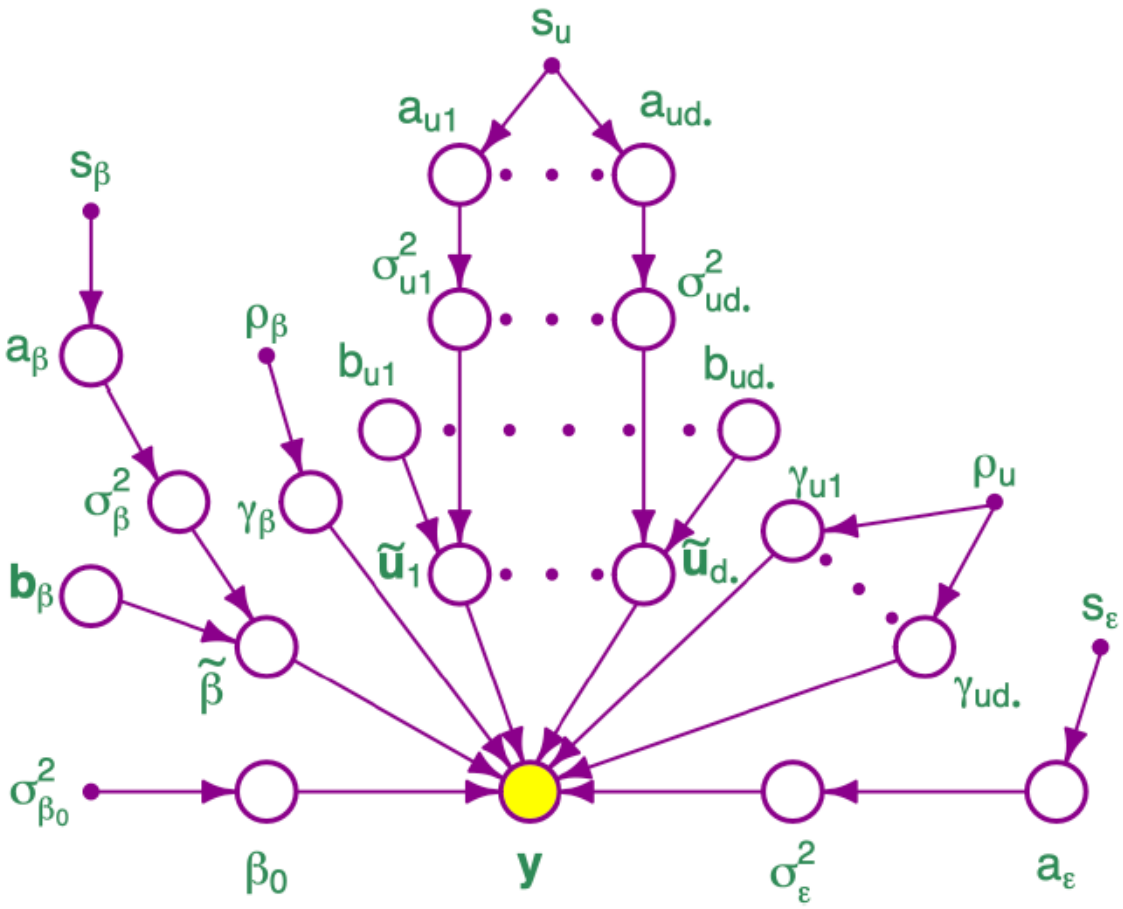


Figure 2.1: *Directed acyclic graph representation of Bayesian model (2.28) for Gaussian response. Random variables and vectors are shown as larger open circles, with shading indicating to the observed response data. The small closed circles are user-specified hyperparameters.*

blanket (e.g. Pearl, 1988). The Markov blanket of a node consists of its parent node, co-parent nodes and child nodes.

For example, when computing full conditional distributions of $\widetilde{\boldsymbol{\beta}}$, we look for its parent nodes: σ^2_β and $\boldsymbol{b}_\beta$,, its child node: $\boldsymbol{y}$, and its co-parent nodes: β_0, $\boldsymbol{\gamma}_\beta$, $\widetilde{\boldsymbol{u}}_1, \ldots,$ $\widetilde{\boldsymbol{u}}_{d_\bullet}, \gamma_{u_1}, \ldots, \gamma_{u_{d_\bullet}}$, and σ^2_ε. Then we work out the full conditional distribution of $\widetilde{\boldsymbol{\beta}}$ as following:

$$\begin{aligned}\mathfrak{p}(\widetilde{\boldsymbol{\beta}}|\text{rest}) &= \mathfrak{p}(\widetilde{\boldsymbol{\beta}}|\boldsymbol{\beta}_0, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}_1, \ldots, \widetilde{\boldsymbol{u}}_{d_\bullet}, \gamma_{u_1}, \ldots, \gamma_{u_{d_\bullet}}, \sigma^2_\varepsilon) \\ &\propto \mathfrak{p}(\boldsymbol{y}|\beta_0, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}_1, \ldots, \widetilde{\boldsymbol{u}}_{d_\bullet}, \gamma_{u_1}, \ldots, \gamma_{u_{d_\bullet}}, \sigma^2_\varepsilon)\mathfrak{p}(\widetilde{\boldsymbol{\beta}}|\sigma^2_\beta, \boldsymbol{b}_\beta).\end{aligned}$$

2.4.3 Bayesian Probit Additive Model Selection

For binary responses, below is the adjustment to our likelihood to model (2.28) described in Section 2.4.2, with $1 \leq i \leq n$:

$$\begin{aligned} y_i \big| \beta_0, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{\beta}}, \gamma_{u1}, \ldots, \gamma_{u_{d_\bullet}}, \widetilde{\boldsymbol{u}}_1, \ldots, \widetilde{\boldsymbol{u}}_{d_\bullet} \\ \overset{\text{ind.}}{\sim} \text{Bernoulli} \left(\Phi \left(\beta_0 + \left(X(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) + \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j(\gamma_{uj} \widetilde{\boldsymbol{u}}_j) \right)_i \right) \right). \end{aligned} \tag{2.29}$$

Following Albert and Chib (1993), we introduce an auxiliary random variables vector $\boldsymbol{c} \equiv (c_1, \ldots, c_n)$ such that

$$\begin{aligned} & y_i = 1 \quad \text{if and only if } c_i \geq 0, \quad 1 \leq i \leq n, \\ & y_i | c_i \overset{\text{ind.}}{\sim} \text{Bernoulli} \left(I \left(c_i \geq 0 \right) \right). \end{aligned} \tag{2.30}$$

and impose the following conditional distribution on $\boldsymbol{c} \equiv (c_1, \ldots, c_n)$:

$$\begin{aligned} \boldsymbol{c} \big| \beta_0, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{\beta}}, \gamma_{u1}, \ldots, \gamma_{u_{d_\bullet}}, \widetilde{\boldsymbol{u}}_1, \ldots, \widetilde{\boldsymbol{u}}_{d_\bullet} \\ \sim N\Big(\mathbf{1}_n \beta_0 + \boldsymbol{X}(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) + \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j(\gamma_{uj} \widetilde{\boldsymbol{u}}_j), \boldsymbol{I}_n \Big). \end{aligned} \tag{2.31}$$

In Figure 2.2 for binary responses, the prior distributions on $\beta_0, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{\beta}}, \gamma_{u1}, \ldots, \gamma_{u_{d_\bullet}}$ and $\widetilde{\boldsymbol{u}}_1, \ldots, \widetilde{\boldsymbol{u}}_{d_\bullet}$ are the same as in the Gaussian response case in Figure 2.1. However, the error variance variables σ_ε^2 and a_ε are not present for the binary response data model in Figure 2.2. Therefore, our binary response model is modified to Figure 2.2 from Figure 2.1, where $\boldsymbol{y}$ distributional specification is replaced by (2.30) and (2.31).

2.5 Posterior Distributions for Markov chain Monte Carlo Practical Fitting

We use sufficient statistic matrices described in Section 2.5.1 for both Markov chain Monte Carlo and mean field variational Bayes fitting algorithms. These sufficient statistic matrices offer computational speed-ups in implementation to facilitate scalability to large data sets.

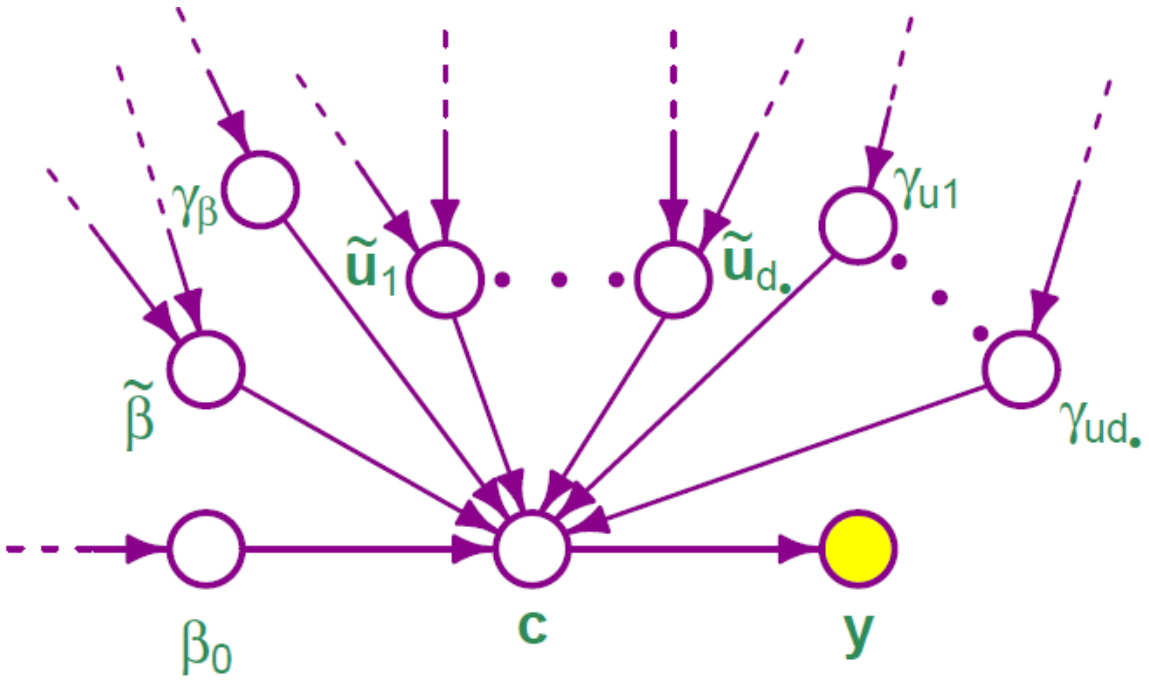


Figure 2.2: *Sub-graph of the directed acyclic graph for the binary responses adjustment to (2.28). This graph is the same as that shown in Figure 2.1 except for locations near the response variables node. The new graph has the following modifications: (1) the σ_ε^2 and a_ε nodes are absent, (2) a hidden node $\boldsymbol{c}$ corresponding to the the Albert-Chib auxiliary variables is added to the position held by $\boldsymbol{y}$ in the Gaussian responses graph and the binary responses observed data node $\boldsymbol{c}$ is a parent of $\boldsymbol{y}$.*

Section 2.5.2 lists some notation used in our algorithm listing, for both Markov chain Monte Carlo and mean field variational Bayes.

The most accurate practical approach is Markov chain Monte Carlo (e.g. Gelfand and Smith, 1990). In order to implement the practical fitting algorithm of Markov chain Monte Carlo, we need to work out the approximation of the posterior distributions of each parameter, which are the hidden nodes on the directed acyclic graphs. Using Markov blanket, based on **Theorem 1.9.1**, we work out the full conditional distributions of parameters. The full conditional distribution results are listed in (2.36) – (2.48) in Section 2.5.3, with derivation details given in Appendix 2.B.

2.5.1 Pre-Processing and Storage of Key Matrices

Algorithm 1 in this subsection provides detailed steps of standardization of original input data described in Section 2.2.2. It makes our methodology independent of units of measurement. Conversion of final results to the original units is done after

running our Bayesian inference algorithms.

Algorithm 1 also provides detailed steps in computing and storing the sufficient statistic matrices for computational speed-ups. It helps to maintain our Bayesian generalized additive models fitting in a stable and efficient manner. Once the linear design matrix $\boldsymbol{X}$ and matrices containing the canonical Demmler-Reinsch basis $\boldsymbol{Z}$ are computed, the sufficient statistic matrices, such as $\boldsymbol{X}^T\boldsymbol{y}, \boldsymbol{X}^T\boldsymbol{X}, \boldsymbol{Z}^T\boldsymbol{y}, \boldsymbol{Z}^T\boldsymbol{X}, \boldsymbol{Z}^T\boldsymbol{Z}$ are able to compute. These matrices appear a couple of times, in both posterior distributions of parameters for Markov chain Monte Carlo described in Section 2.5.3 and $\mathfrak{q}$-density approximation for mean field variational Bayes described in Chapter 4. These sufficient statistic matrices only need to be computed once and stored, then they can be used repeatedly when needed for computational speed-ups. Therefore, pre-processing of the original data is an important part of our implementation.

2.5.2 Notation Used in the Fitting Algorithms

In this subsection, we define the notations used in Algorithm 2 in Section 2.5.3. The notation in (2.32) in this subsection is used to locate a particular row or a particular column of the sufficient statistic matrices described in Section 2.5.1. These notation appear in the algorithm listings of our two approaches, Algorithm 2 in Section 2.5.3 for Markov chain Monte Carlo sample drawing, and Algorithm 7 in Chapter 4 for $\mathfrak{q}$-density approximation of parameters for mean field variational Bayes.

$$
\begin{aligned}
&\boldsymbol{K}_j \equiv \text{the number of columns in } \boldsymbol{Z}_j, 1 \le j \le d_\bullet,\\
&\mathfrak{c} \text{ is the } (d_\bullet + 1) \times 1 \text{ vector with entries } \mathfrak{c}_1 \equiv 0 \text{ and } \mathfrak{c}_{j+1} \equiv \sum_{k=1}^{j} K_k, 1 \le j \le d_\bullet,\\
&\mathbf{ZTy}^{\langle j \rangle} \equiv \text{the sub-block of } \mathbf{ZTy} \text{ corresponding to rows } (\mathfrak{c}_j + 1) \text{ to } \mathfrak{c}_{j+1}, 1 \le j \le d_\bullet,\\
&\mathbf{ZTX}^{\langle j \rangle} \equiv \text{the sub-block of } \mathbf{ZTX} \text{ corresponding to rows } (\mathfrak{c}_j + 1) \text{ to } \mathfrak{c}_{j+1}, 1 \le j \le d_\bullet,\\
&\mathbf{ZTZ}^{\langle j,j' \rangle} \equiv \text{the sub-block of } \mathbf{ZTZ} \text{ corresponding to rows } (\mathfrak{c}_j + 1) \text{ to } \mathfrak{c}_{j+1}\\
&\qquad\qquad \text{and columns } (\mathfrak{c}_{j'} + 1) \text{ to } (\mathfrak{c}_{j'+1}), 1 \le j, j' \le d_\bullet.
\end{aligned}
\tag{2.32}
$$

Note that, according to the notation in (2.32),

$$
\mathbf{ZTX}^{\langle j \rangle} = \boldsymbol{Z}_j^T\boldsymbol{X} \quad \text{and} \quad \mathbf{ZTZ}^{\langle j.j' \rangle} = \boldsymbol{Z}_j^T\boldsymbol{Z}_{j'}.
$$

Algorithm 1: *Pre-processing of original data and creation of key matrices for input into Bayesian generalized additive model algorithms.*

Inputs: $\boldsymbol{y}^{\text{orig}}(n \times 1)$; $\mathring{\boldsymbol{x}}_j^{\text{orig}}(n \times 1), 1 \leq j \leq d_\circ$; $\dot{\boldsymbol{x}}_j^{\text{orig}}(n \times 1), 1 \leq j \leq d_\bullet$

mean($\boldsymbol{y}^{\text{orig}}$) $\longleftarrow$ sample mean of $\boldsymbol{y}^{\text{orig}}$; st.dev($\boldsymbol{y}^{\text{orig}}$) $\longleftarrow$ sample standard dev'n of $\boldsymbol{y}^{\text{orig}}$

If $\boldsymbol{y}^{\text{orig}}$ is continuous then $\boldsymbol{y} \longleftarrow \{\boldsymbol{y}^{\text{orig}} - \text{mean}(\boldsymbol{y}^{\text{orig}})\mathbf{1}_n\}/\text{st.dev}(\boldsymbol{y}^{\text{orig}})$

If $\boldsymbol{y}^{\text{orig}}$ is binary then $\boldsymbol{y} \longleftarrow \boldsymbol{y}^{\text{orig}}$

For $j = 1, \ldots, d_\circ$:

mean($\mathring{\boldsymbol{x}}_j^{\text{orig}}$)$\longleftarrow$ sample mean of $\mathring{\boldsymbol{x}}_j^{\text{orig}}$; st.dev($\mathring{\boldsymbol{x}}_j^{\text{orig}}$)$\longleftarrow$ sample standard dev'n of $\mathring{\boldsymbol{x}}_j^{\text{orig}}$

$\mathring{\boldsymbol{x}}_j \longleftarrow \{\mathring{\boldsymbol{x}}_j^{\text{orig}} - \text{mean}(\mathring{\boldsymbol{x}}_j^{\text{orig}})\mathbf{1}_n\}/\text{st.dev}(\mathring{\boldsymbol{x}}_j^{\text{orig}})$

For $j = 1, \ldots, d_\bullet$:

mean($\dot{\boldsymbol{x}}_j^{\text{orig}}$)$\longleftarrow$ sample mean of $\dot{\boldsymbol{x}}_j^{\text{orig}}$; st.dev($\dot{\boldsymbol{x}}_j^{\text{orig}}$)$\longleftarrow$ sample standard dev'n of $\dot{\boldsymbol{x}}_j^{\text{orig}}$

$\dot{\boldsymbol{x}}_j \longleftarrow \{\dot{\boldsymbol{x}}_j^{\text{orig}} - \text{mean}(\dot{\boldsymbol{x}}_j^{\text{orig}})\mathbf{1}_n\}/\text{st.dev}(\dot{\boldsymbol{x}}_j^{\text{orig}})$

$\boldsymbol{X} \longleftarrow \left[\mathring{\boldsymbol{x}}_1 \ldots \mathring{\boldsymbol{x}}_{d_\circ} \dot{\boldsymbol{x}}_1 \ldots \dot{\boldsymbol{x}}_{d_\bullet}\right]$

For $j = 1, \ldots, d_\bullet$:

$\boldsymbol{Z}_j \longleftarrow n \times K_j$ matrix containing canonical Demmler-Reinsch basis for the data vector $\dot{\boldsymbol{x}}_j$, predictor using the construction described in Appendix 2.A.

$\boldsymbol{Z} \longleftarrow [\boldsymbol{Z}_1 \ldots \boldsymbol{Z}_{d_\bullet}]$; **XTy** $\longleftarrow \boldsymbol{X}^T\boldsymbol{y}$; **XTX** $\longleftarrow \boldsymbol{X}^T\boldsymbol{X}$; **ZTX** $\longleftarrow \boldsymbol{Z}^T\boldsymbol{X}$; **ZTZ** $\longleftarrow \boldsymbol{Z}^T\boldsymbol{Z}$;

Outputs: $\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{d_\bullet}$, **XTy**, **XTX**, **ZTX**, **ZTZ**, mean($\boldsymbol{y}^{\text{orig}}$),st.dev($\boldsymbol{y}^{\text{orig}}$),

$$\left\{(\text{mean}(\mathring{\boldsymbol{x}}_j^{\text{orig}}), \text{st.dev}(\mathring{\boldsymbol{x}}_j^{\text{orig}})) : 1 \leq j \leq d_\circ\right\}, \left\{(\text{mean}(\dot{\boldsymbol{x}}_j^{\text{orig}})), \text{st.dev}(\dot{\boldsymbol{x}}_j^{\text{orig}}) : 1 \leq j \leq d_\bullet\right\}$$

The notational convention $\boldsymbol{e}_r$ will be used in our algorithm listings of two approaches.

$$\boldsymbol{e}_r \equiv \text{a column vector of appropriate length with } r\text{th entry equal to 1 and zeros elsewhere.} \tag{2.33}$$

For example, the kth row of $\mathbf{ZTX}^{<j>}$ is $\boldsymbol{e}_k^T\mathbf{ZTX}^{<j>}$ where $\boldsymbol{e}_k$ is the $K_j \times 1$ vector with kth entry 1 and 0 elsewhere.

The implementation of MCMC following Algorithm 2 in Section 2.5.3 and the implementation of MFVB following Algorithm 7 in Chapter 4 both would not require

explicit calculation and storage of $\boldsymbol{e}_r$ vectors. The use of notation $\boldsymbol{e}_r$ is for array subsetting code specific for programming steps in algorithm listings, and it avoids further subscripting in our algorithm listing.

In the binary response case, we use the Albert-Chib auxiliary variables vector $\boldsymbol{c}$ to replace vector $\boldsymbol{y}$ described in Section 2.4.3. Then the notation of (2.32) for extraction of sub-blocks of $\mathbf{ZTy}$ is replaced by $\mathbf{ZTy}_{\text{adj}}$.

Algorithm 2 for MCMC in Section 2.5.3 and Algorithm 7 for MFVB in Chapter 4, handle both Gaussian response and binary response together. In order to show programming steps for both responses, we use notations $\mathbf{yT1}_{\text{adj}}$, $\mathbf{XTy}_{\text{adj}}$ and $\mathbf{ZTy}_{\text{adj}}$ denoted to be:

$$\begin{aligned}
\mathbf{yT1}_{\text{adj}} &\equiv \text{adjustment of } \mathbf{yT1} \equiv \begin{cases} \mathbf{1}_n^T \boldsymbol{y} & \text{for Gaussian responses,} \\ \mathbf{1}_n^T \boldsymbol{c} & \text{for binary responses,} \end{cases} \\
\mathbf{XTy}_{\text{adj}} &\equiv \text{adjustment of } \mathbf{XTy} \equiv \begin{cases} \boldsymbol{X}^T \boldsymbol{y} & \text{for Gaussian responses,} \\ \boldsymbol{X}^T \boldsymbol{c} & \text{for binary responses,} \end{cases} \\
\mathbf{ZTy}_{\text{adj}} &\equiv \text{adjustment of } \mathbf{ZTy} \equiv \begin{cases} \boldsymbol{Z}^T \boldsymbol{y} & \text{for Gaussian responses,} \\ \boldsymbol{Z}^T \boldsymbol{c} & \text{for binary responses.} \end{cases}
\end{aligned} \tag{2.34}$$

The main algorithms also uses the following functions:

$$\begin{aligned}
&\text{logit}(x) \equiv \log\left(\frac{x}{1-x}\right), \\
&\text{expit}(x) \equiv \text{logit}^{-1} = \frac{1}{1+\exp(-x)}, \\
&\zeta(x) = \log\left\{2\Phi(x)\right\}, \zeta'(x) = \phi(x)/\Phi(x)
\end{aligned} \tag{2.35}$$

where Φ is the $N(0,1)$ cumulative distribution function and ϕ is the $N(0,1)$ density function.

2.5.3 Markov Chain Monte Carlo

For Gaussian response, based on full model description in Section 2.4.2 and Figure 2.1, for binary responses, based on full model description in Section 2.4.3 and Figure 2.2, both provide full algorithmic details for Markov chain Monte Carlo-based ap-

proximate full conditional distributions of parameters. Using Markov blanket, we obtain the full conditional distributions of parameters, which are the posterior distributions, given in (2.36) – (2.48) in Section 2.5.3, with derivation details given in Appendix 2.B.

$$\beta_0 \big| \text{rest} \sim N\left(\frac{\frac{1}{\sigma_\varepsilon^2}\mathbf{1}_n^T(\boldsymbol{y})}{(n/\sigma_\varepsilon^2)+(1/\sigma_{\beta_0}^2)}, \frac{1}{(n/\sigma_\varepsilon^2)+(1/\sigma_{\beta_0}^2)} \right), \tag{2.36}$$

$$\begin{aligned}\widetilde{\boldsymbol{\beta}} \big| \text{rest} \sim N\Bigg(& \left(\frac{1}{\sigma_\varepsilon^2}(\boldsymbol{X}^T\boldsymbol{X}) \odot (\boldsymbol{\gamma}_\beta \boldsymbol{\gamma}_\beta^T) + \frac{1}{\sigma_\beta^2}\text{diag}(\boldsymbol{b}_\beta) \right)^{-1} \\ & \times \frac{1}{\sigma_\varepsilon^2}\left[\boldsymbol{\gamma}_\beta \odot \left\{ (\boldsymbol{X}^T\boldsymbol{y}) - \sum_{j=1}^{d_\bullet} (\boldsymbol{X}^T\boldsymbol{Z}_j)(\gamma_{uj}\widetilde{\boldsymbol{u}}_j) \right\} \right], \\ & \left(\frac{1}{\sigma_\varepsilon^2}(\boldsymbol{X}^T\boldsymbol{X}) \odot (\boldsymbol{\gamma}_\beta \boldsymbol{\gamma}_\beta^T) + \frac{1}{\sigma_\beta^2}\text{diag}(\boldsymbol{b}_\beta) \right)^{-1} \Bigg),\end{aligned} \tag{2.37}$$

$$b_{\beta j} \big| \text{rest} \sim \text{Inverse-Gaussian}\left(\frac{\sigma_\beta}{|\widetilde{\beta}_j|}, 1 \right), \;\; 1 \le j \le d_\circ + d_\bullet, \tag{2.38}$$

$$\sigma_\beta^2 \big| \text{rest} \sim \text{Inverse-Gamma}\left(\frac{d_\circ + d_\bullet + 1}{2}, \frac{\widetilde{\boldsymbol{\beta}}^T \text{diag}(\boldsymbol{b}_\beta)\widetilde{\boldsymbol{\beta}}}{2} + \frac{1}{a_\beta} \right), \tag{2.39}$$

$$a_\beta \big| \text{rest} \sim \text{Inverse-Gamma}\left(1, \frac{1}{\sigma_\beta^2} + \frac{1}{s_\beta^2} \right), \tag{2.40}$$

$$\begin{aligned}\gamma_{\beta j} \big| \text{rest} \sim \text{Bernoulli}\left(\frac{1}{1+\exp(-\eta_{\beta j})} \right), \;\; 1 \le j \le d_\circ + d_\bullet, \text{ where} \\ \eta_{\beta j} = \text{logit}(\rho_\beta) - \frac{1}{2\sigma_\varepsilon^2}\Bigg(\widetilde{\beta}_j^2 \|\boldsymbol{X}_j\|^2 - 2\widetilde{\beta}_j \boldsymbol{X}_j^T \big[\boldsymbol{y} - \boldsymbol{X}_{-j}\{(\boldsymbol{\gamma}_\beta)_{-j} \odot (\widetilde{\boldsymbol{\beta}})_{-j}\} \\ - \sum_{j'=1}^{d_\bullet} \boldsymbol{Z}_{j'}(\gamma_{uj'}\widetilde{\boldsymbol{u}}_{j'}) \big] \Bigg),\end{aligned} \tag{2.41}$$

$$\widetilde{\boldsymbol{u}}_j \big| \text{rest} \sim N\Bigg(\text{diag}\left(\frac{\gamma_{uj}\boldsymbol{\omega}_{\boldsymbol{Z}_j}}{\sigma_\varepsilon^2} + \frac{b_{uj}\mathbf{1}}{\sigma_{uj}^2} \right)^{-1} \left(\frac{1}{\sigma_\varepsilon^2}\right) \Big[\gamma_{uj} \left(\boldsymbol{Z}_j^T\boldsymbol{y} - \boldsymbol{Z}_j^T\boldsymbol{X}\left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}} \right) \right.$$

$$-\sum_{j' \neq j}^{d_\bullet} \boldsymbol{Z}_j^T \boldsymbol{Z}_{j'} \left(\gamma_{uj'} \widetilde{\boldsymbol{u}}_{j'}\right)\Bigg)\Bigg], \operatorname{diag}\left(\frac{\gamma_{uj}\boldsymbol{\omega}_{\boldsymbol{Z}_j}}{\sigma_\varepsilon^2} + \frac{b_{uj}\mathbf{1}}{\sigma_{uj}^2}\right)^{-1}\Bigg), \;\; 1 \leq j \leq d_\bullet, \quad (2.42)$$

$$b_{uj}\big|\text{rest} \sim \text{Inverse-Gaussian}\left(\frac{\sigma_{uj}}{\left\|\widetilde{\boldsymbol{u}}_j\right\|}, 1\right), \;\; 1 \leq j \leq d_\bullet, \quad (2.43)$$

$$\sigma_{uj}^2\big|\text{rest} \sim \text{Inverse-Gamma}\left(\frac{K_j+1}{2}, \frac{b_{uj}\left\|\widetilde{\boldsymbol{u}}_j\right\|^2}{2} + \frac{1}{a_{uj}}\right), \;\; 1 \leq j \leq d_\bullet, \quad (2.44)$$

$$a_{uj}\big|\text{rest} \sim \text{Inverse-Gamma}\left(1, \frac{1}{\sigma_{uj}^2} + \frac{1}{s_u^2}\right), \;\; 1 \leq j \leq d_\bullet \quad (2.45)$$

and

$$\gamma_{uj}\big|\text{rest} \sim \text{Bernoulli}\left(\frac{1}{1+\exp(-\eta_{uj})}\right), \;\; 1 \leq j \leq d_\bullet, \text{ where}$$

$$\eta_{uj} = \text{logit}(\rho_u) - \frac{1}{2\sigma_\varepsilon^2}\Bigg[\boldsymbol{\omega}_{\boldsymbol{Z}_j}^T\left(\widetilde{\boldsymbol{u}}_j \odot \widetilde{\boldsymbol{u}}_j\right) - 2\widetilde{\boldsymbol{u}}_j^T \boldsymbol{Z}_j^T\Bigg\{\boldsymbol{y} - \boldsymbol{X}\left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}\right) - \sum_{j' \neq j}^{d_\bullet} \boldsymbol{Z}_{j'}(\gamma_{uj'}\widetilde{\boldsymbol{u}}_{j'})\Bigg\}\Bigg]. \quad (2.46)$$

If $\boldsymbol{y}$ is Gaussian response, then

$$\sigma_\varepsilon^2\big|\text{rest} \sim \text{Inverse-Gamma}\left(\frac{n+1}{2}, \frac{1}{a_\varepsilon} + \frac{1}{2}\left\|\boldsymbol{y} - \mathbf{1}_n\beta_0 - \boldsymbol{X}(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j)\right\|^2\right) \quad (2.47)$$

and

$$a_\varepsilon\big|\text{rest} \sim \text{Inverse-Gamma}\left(1, \frac{1}{\sigma_\varepsilon^2} + \frac{1}{s_\varepsilon^2}\right). \quad (2.48)$$

If $\boldsymbol{y}$ is binary responses, then

$$y_i\big|c_i \overset{\text{ind.}}{\sim} \text{Bernoulli}\left(I\left(c_i \geq 0\right)\right), \quad 1 \leq i \leq n,$$
$$c_i\big|\text{rest} \sim (2y_i - 1)\,\text{Truncated-Normal}_+$$

$$\left((2y_i - 1) \left[\mathbf{1}_n \beta_0 + \boldsymbol{X} \left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}} \right) + \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j \left(\gamma_{uj} \widetilde{\boldsymbol{u}}_j \right) \right]_i , 1 \right), \quad y_i = 1 \text{ or } y_i = 0 \quad (2.49)$$

The full conditional distributions all have standard forms for practical implementation of the Markov chain Monte Carlo approach. These include Multivariate Normal, Inverse Gamma, Bernoulli and Beta distributions. Some full conditional distributions are Inverse Gaussian, Michael et al. (1976) provides an algorithm for drawing from Inverse Gaussian distributions. Some full conditional distributions are Truncated-Normal, similarly, Robert (1995) provides effective solutions for drawing from Truncated-Normal distributions. Therefore, the posterior distributions for Markov chain Monte Carlo are fairly straightforward. The Markov chain Monte Carlo sampling reduces to Gibbs sampling for our model.

Algorithm 2 in this subsection lists the full set of steps needed to draw samples from the posterior distributions of the model parameters. The fact that most of the draws only require the sufficient statistic matrices from Algorithm 1 given in Section 2.2.2 means that the sampling can be done quite rapidly regardless of sample size.

Following Algorithm 2, the Markov chain Monte Carlo algorithm is implemented as the R package gamselBayes.

Algorithm 2: *Markov chain Monte Carlo generation of samples from the posterior distributions of the parameters in (2.28) using Laplace-Zero priors.*

Data Inputs: $\boldsymbol{y}(n \times 1)$; $\boldsymbol{X}\,(n \times (d_\circ + d_\bullet))$; $\boldsymbol{Z}_j(n \times K_j), 1 \le j \le d_\bullet$.

Response Type Input: **responseType** $\in$ {Gaussian, Bernoulli} .

Sufficient Statistics Inputs: **XTy**, **XTX**, **ZTy**, **ZTX**, **ZTZ**.

Hyperparameter Inputs: $\sigma_{\beta_0}, s_\beta, s_\varepsilon, s_u > 0,\ 0 < \rho_\beta, \rho_u < 1$.

Chain Length Inputs: N_{warm} and N_{kept} both positive integers.

Initialize: $\boldsymbol{\gamma}_\beta^{[0]} \longleftarrow \frac{1}{2}\mathbf{1}_{d_\circ+d_\bullet}; \gamma_{uj}^{[0]} \longleftarrow \frac{1}{2}, 1 \le j \le d_\bullet;\ \widetilde{\boldsymbol{\beta}}^{[0]} \longleftarrow \mathbf{0}_{d_\circ+d_\bullet}$

$\widetilde{\boldsymbol{u}}_j^{[0]} \longleftarrow \mathbf{0}_{K_j}, 1 \le j \le d_\bullet;\ (\sigma_\varepsilon^2)^{[0]} \longleftarrow 1;\ a_\varepsilon^{[0]} \longleftarrow 1;\ (\sigma_\beta^2)^{[0]} \longleftarrow 1;\ a_\beta^{[0]} \longleftarrow 1$

$\boldsymbol{b}_\beta^{[0]} \longleftarrow \mathbf{1}_{d_\circ+d_\bullet};\ b_{uj}^{[0]} \longleftarrow 1, 1 \le j \le d_\bullet;$

$a_{uj}^{[0]} \longleftarrow 1, 1 \le j \le d_\bullet;\ (\sigma_{uj}^2)^{[0]} \longleftarrow 1, 1 \le j \le d_\bullet.$

$\mathbf{yT1}_{\text{adj}} \longleftarrow 0\ ;\ \mathbf{XTy}_{\text{adj}} \longleftarrow \mathbf{XTy}\ ;\ \mathbf{ZTy}_{\text{adj}} \longleftarrow \mathbf{ZTy}$

For $j = 1, \ldots, d_\bullet:\ \boldsymbol{\omega}_{Z_j} \longleftarrow \text{diagonal}\left(\mathbf{ZTZ}^{\langle j,j\rangle}\right)$

For $g = 1, \ldots, N_{\text{warm}} + N_{\text{kept}}$:

$\omega_1 \longleftarrow \mathbf{yT1}_{\text{adj}}$

$\omega_2 \longleftarrow \left\{n \Big/ \left(\sigma_\varepsilon^2\right)^{[g-1]} + \left(1/\sigma_{\beta_0}^2\right)\right\} \quad ; \quad \beta_0^{[g]} \sim N\Big(\frac{\omega_1}{\left(\sigma_\varepsilon^2\right)^{[g-1]}\omega_2}, \frac{1}{\omega_2}\Big)$

$\boldsymbol{\Omega} \longleftarrow \left(\boldsymbol{\gamma}_\beta^{[g-1]}\boldsymbol{\gamma}_\beta^{[g-1]T}\right) \odot (\mathbf{XTX}) \Big/ \left(\sigma_\varepsilon^2\right)^{[g-1]} + \text{diag}\left(\boldsymbol{b}_\beta^{[g-1]}\right) \Big/ \left(\sigma_\beta^2\right)^{[g-1]}$

$\boldsymbol{\omega}_3 \longleftarrow \mathbf{XTy}_{\text{adj}} - \sum_{j=1}^{d_\bullet} \mathbf{ZTX}^{\langle j\rangle T}\left(\gamma_{uj}^{[g-1]}\widetilde{\boldsymbol{u}}_j^{[g-1]}\right)$

Decompose $\boldsymbol{\Omega} = \boldsymbol{U}_{\boldsymbol{\Omega}}\text{diag}(d_{\boldsymbol{\Omega}})\boldsymbol{U}_{\boldsymbol{\Omega}}^T$ where $\boldsymbol{U}_{\boldsymbol{\Omega}}\boldsymbol{U}_{\boldsymbol{\Omega}}^T = \boldsymbol{I}$

$\boldsymbol{z} \sim N(\mathbf{0}, \boldsymbol{I})((d_\circ + d_\bullet) \times 1) \quad ; \quad \widetilde{\boldsymbol{\beta}}^{[g]} \longleftarrow \boldsymbol{U}_{\boldsymbol{\Omega}}\left(\frac{\boldsymbol{U}_{\boldsymbol{\Omega}}^T\boldsymbol{z}}{\sqrt{d_{\boldsymbol{\Omega}}}} + \frac{\boldsymbol{U}_\Omega^T\left(\boldsymbol{\gamma}_\beta^{[g-1]} \odot \boldsymbol{\omega}_3\right)}{\boldsymbol{d}_{\boldsymbol{\Omega}}\left(\sigma_\varepsilon^2\right)^{[g-1]}}\right)$

$\left(\boldsymbol{b}_\beta^{[g]}\right)_j \sim \text{Inverse-Gaussian}\left(\sigma_\beta^{[g-1]} \Big/ \left|\left(\widetilde{\boldsymbol{\beta}}^{[g]}\right)_j\right|\right),\ 1 \le j \le d_\circ + d_\bullet$

continued on a subsequent page ...

Algorithm 2 continued: *This is a continuation of the description of this algorithm that commences on a preceding page.*

$$\left(\sigma_\beta^2\right)^{[g]} \sim \text{Inverse-Gamma}\left(\frac{1}{2}(d_\circ + d_\bullet + 1), 1/a_\beta^{[g-1]} + \frac{1}{2}\widetilde{\boldsymbol{\beta}}^{[g]T}\text{diag}\left(\boldsymbol{b}_\beta^{[g]}\right)\widetilde{\boldsymbol{\beta}}^{[g]}\right)$$

$$a_\beta^{[g]} \sim \text{Inverse-Gamma}\left(1, \left\{1/(\sigma_\beta^2)^{[g]}\right\} + (1/s_\beta^2)\right) \; ; \; \boldsymbol{\beta}^{\text{curr}} \longleftarrow \boldsymbol{\gamma}_\beta^{[g-1]} \odot \widetilde{\boldsymbol{\beta}}^{[g]}$$

For $j = 1, \ldots, d_\bullet$: $\quad \boldsymbol{u}_j^{\text{curr}} \longleftarrow \gamma_{uj}^{[g-1]}\widetilde{\boldsymbol{u}}_j^{[g-1]}$

For $j = 1, \ldots, d_\circ + d_\bullet$:

$$\omega_4 \longleftarrow \boldsymbol{e}_j^T\mathbf{XTy}_{\text{adj}} - \left(\mathbf{XTX}\boldsymbol{e}_j\right)_{-j}^T\left(\boldsymbol{\beta}^{\text{curr}}\right)_{-j} - \sum_{j'=1}^{d_\bullet}\left(\mathbf{ZTX}^{\langle j'\rangle}\boldsymbol{e}_j\right)^T\boldsymbol{u}_{j'}^{\text{curr}}$$

$$\omega_5 \longleftarrow \text{logit}\left(\rho_\beta\right) - \frac{1}{2}\left\{(\widetilde{\beta}_j^{[g]})^2\boldsymbol{e}_j^T\mathbf{XTX}\boldsymbol{e}_j - 2\widetilde{\beta}_j^{[g]}\omega_4\right\}\Big/(\sigma_\varepsilon^2)^{[g-1]}$$

$$\left(\boldsymbol{\gamma}_\beta^{[g]}\right)_j \sim \text{Bernoulli}\left(\text{expit}(\omega_5)\right)$$

$\boldsymbol{\beta}^{\text{curr}} \longleftarrow \boldsymbol{\gamma}_\beta^{[g]} \odot \widetilde{\boldsymbol{\beta}}^{[g]}$ $\quad ; \quad$ For $j = 1, \ldots, d_\bullet$: $\quad \widetilde{\boldsymbol{u}}_j^{\text{curr}} \longleftarrow \widetilde{\boldsymbol{u}}_j^{[g-1]}$

For $j = 1, \ldots, d_\bullet$:

$$\boldsymbol{\omega}_6 \longleftarrow \mathbf{ZTy}_{\text{adj}}^{\langle j\rangle} - \mathbf{ZTX}^{\langle j\rangle}\boldsymbol{\beta}^{\text{curr}} - \sum_{j'\neq j}^{d_\bullet}\mathbf{ZTZ}^{\langle j,j'\rangle}\left(\gamma_{uj'}^{[g-1]}\widetilde{\boldsymbol{u}}_{j'}^{\text{curr}}\right)$$

$$\boldsymbol{\omega}_7 \longleftarrow \left\{\gamma_{uj}^{[g-1]}\boldsymbol{\omega}_{\boldsymbol{Z}_j}\Big/(\sigma_\varepsilon^2)^{[g-1]}\right\} + \left\{b_{uj}^{[g-1]}\mathbf{1}_{K_j}\Big/(\sigma_{uj}^2)^{[g-1]}\right\}$$

$$\boldsymbol{z} \sim N(\mathbf{0}, \boldsymbol{I}_{K_j}) \quad ; \quad \widetilde{\boldsymbol{u}}_j^{\text{curr}} \longleftarrow \left(\boldsymbol{z}\Big/\sqrt{\boldsymbol{\omega}_7}\right) + \left[\gamma_{uj}^{[g-1]}\boldsymbol{\omega}_6\Big/\left\{\boldsymbol{\omega}_7(\sigma_\varepsilon^2)^{[g-1]}\right\}\right]$$

For $j = 1, \ldots, d_\bullet$: $\quad \widetilde{\boldsymbol{u}}_j^{[g]} \longleftarrow \widetilde{\boldsymbol{u}}_j^{\text{curr}}$

For $j = 1, \ldots, d_\bullet$:

$$b_{uj}^{[g]} \sim \text{Inverse-Gaussian}\left(\sigma_{uj}^{[g-1]}\Big/\left\|\widetilde{\boldsymbol{u}}_j^{[g]}\right\|, 1\right)$$

$$(\sigma_{uj}^2)^{[g]} \sim \text{Inverse-Gamma}\left(\tfrac{1}{2}(K_j+1), \left\{1\Big/a_{uj}^{[g-1]}\right\} + \tfrac{1}{2}\left\|\widetilde{\boldsymbol{u}}_j^{[g]}\right\|^2 b_{uj}^{[g]}\right)$$

$$a_{uj}^{[g]} \sim \text{Inverse-Gamma}\left(1, \left\{1\Big/(\sigma_{uj}^2)^{[g]}\right\} + (1\Big/s_u^2)\right)$$

For $j = 1, \ldots, d_\bullet$: $\quad \gamma_{uj}^{\text{curr}} \longleftarrow \gamma_{uj}^{[g-1]}$

continued on a subsequent page ...

Algorithm 2 continued: *This is a continuation of the description of this algorithm that commences on a preceding page.*

For $j = 1, \ldots, d_\bullet$:

$$\boldsymbol{\omega}_8 \longleftarrow \mathbf{ZTy}^{\langle j \rangle}_{\text{adj}} - \mathbf{ZTX}^{\langle j \rangle}\boldsymbol{\beta}^{\text{curr}} - \sum_{j' \neq j}^{d_\bullet} \mathbf{ZTZ}^{\langle j,j' \rangle}\left(\gamma^{\text{curr}}_{uj'}\widetilde{\boldsymbol{u}}^{[g]}_{j'}\right)$$

$$\omega_9 \longleftarrow \text{logit}(\rho_u) - \frac{1}{2}\left\{\boldsymbol{\omega}^T_{\boldsymbol{Z}_j}\left(\widetilde{\boldsymbol{u}}^{[g]}_j \odot \widetilde{\boldsymbol{u}}^{[g]}_j\right) - 2\left(\widetilde{\boldsymbol{u}}^{[g]}_j\right)^T \boldsymbol{\omega}_8\right\} \Big/ \left(\sigma^2_\varepsilon\right)^{[g-1]}$$

$$\gamma^{\text{curr}}_{uj} \sim \text{Bernoulli}\left(\text{expit}\left(\omega_9\right)\right)$$

For $j = 1, \ldots, d_\bullet$: $\quad \gamma^{[g]}_{uj} \longleftarrow \gamma^{\text{curr}}_{uj}$

$$\boldsymbol{\omega}_{10} \longleftarrow \mathbf{1}_n\beta^{[g]}_0 + \boldsymbol{X}\left(\boldsymbol{\gamma}^{[g]}_\beta \odot \widetilde{\boldsymbol{\beta}}^{[g]}\right) + \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j(\gamma^{[g]}_{uj}\widetilde{\boldsymbol{u}}^{[g]}_j)$$

If **responseType** is Gaussian then

$$(\sigma^2_\varepsilon)^{[g]} \sim \text{Inverse-Gamma}\left(\tfrac{1}{2}(n+1), (1\Big/a^{[g-1]}_\varepsilon) + \tfrac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{\omega}_{10}\right\|^2\right)$$

$$a^{[g]}_\varepsilon \sim \text{Inverse-Gamma}\left(1, \left\{1\Big/(\sigma^2_\varepsilon)^{[g]}\right\} + (1\Big/s^2_\varepsilon)\right)$$

If **responseType** is Bernoulli then

$(\sigma^2_\varepsilon)^{[g]} \longleftarrow 1$

For $i = 1, \ldots, n$:

$\omega_{11} \longleftarrow \text{Truncated-Normal}_+\left((2y_i - 1)(\boldsymbol{\omega}_{10})_i, 1\right)$; $\; c_i \longleftarrow (2y_i - 1)\omega_{11}$

$\mathbf{yT1}_{\text{adj}} \longleftarrow \mathbf{1}^T\boldsymbol{c}$; $\; \mathbf{XTy}_{\text{adj}} \longleftarrow \boldsymbol{X}^T\boldsymbol{c}$; $\; \mathbf{ZTy}_{\text{adj}} \longleftarrow \boldsymbol{Z}^T\boldsymbol{c}$

Outputs: All chains after omission of the first N_{warm} values.

2.6 Model Selection Strategies with Laplace-Zero Priors for Markov Chain Monte Carlo

After obtaining the full conditional distributions of parameters for Markov chain Monte Carlo, we need to set up the model selection strategies to decide the effect type of a predictor is zero, linear or non-linear effect. The general rule is that zeroing of coefficient corresponds to $E\left(\gamma_{\beta j}\middle|\boldsymbol{y}\right) \leq 1-\tau$, with $1 \leq j \leq d_{\circ}+d_{\bullet}$, or $E\left(\gamma_{u_j}\middle|\boldsymbol{y}\right) \leq 1-\tau$, with $1 < j < d_{\bullet}$. τ is the threshold parameter, $\tau \in (0,1)$.

2.6.1 Deciding Between an Effect Being Zero or Linear

For Gaussian and binary response generalized additive models described in Section 2.4, recalling that (2.10) in Section 2.3.4 to be $\boldsymbol{\beta} \equiv \boldsymbol{\gamma}_{\beta} \odot \widetilde{\boldsymbol{\beta}}$, where $\widetilde{\boldsymbol{\beta}}$ is continuous and $\boldsymbol{\gamma}_{\beta}$ is binary. Therefore

$$P(\beta_j = 0\big|\boldsymbol{y}) = P(\gamma_{\beta j} = 0\big|\boldsymbol{y}) = 1 - E(\gamma_{\beta j}\big|\boldsymbol{y}),\ \ 1 \leq j \leq d_{\circ}+d_{\bullet}. \tag{2.50}$$

We can use the posterior distribution of $\gamma_{\beta j}$ to decide $\beta_j = 0$ or $\beta_j \neq 0,\ 1 \leq j \leq d_{\circ}+d_{\bullet}$. We use a general rule to be:

$$\beta_j = 0 \ \text{ if and only if } P(\gamma_{\beta j} = 0\big|\boldsymbol{y}) > \tau, \text{ and } \tau = \frac{1}{2},\ \ 1 \leq j \leq d_{\circ}+d_{\bullet}, \tag{2.51}$$

which is equivalent to

$$E(\gamma_{\beta j}\big|\boldsymbol{y}) \leq 1-\tau \ \text{ and } \ \tau = \frac{1}{2},\ \ 1 \leq j \leq d_{\circ}+d_{\bullet}.$$

Our "effect is zero" rule between zero effect and linear effect is

$$\begin{aligned} &\text{the effect is zero if } E(\gamma_{\beta j}\big|\boldsymbol{y}) \leq 1-\tau,\ \ 1 \leq j \leq d_{\circ}+d_{\bullet}, \\ &\text{otherwise the effect is linear.} \end{aligned} \tag{2.52}$$

For Markov chain Monte Carlo approach, $E(\gamma_{\beta_j}\big|\boldsymbol{y})$, $1 \leq j \leq d_{\circ}+d_{\bullet}$, is approximate the sample mean of $\gamma_{\beta_j}^{[g]}$ in Algorithm 2, $1 \leq g \leq N_{\text{kept}}$. Simulation study of deciding default values for τ for the Markov chain Monte Carlo approach are described in Section 2.6.3.

2.6.2 Deciding Between an Effect Being Zero, Linear or Non-Linear

For Gaussian and binary response generalized additive models described in Section 2.4, recalling that (2.18) in Section 2.3.5 to be $\boldsymbol{u}_j \equiv \gamma_{uj}\widetilde{\boldsymbol{u}}_j, 1 \leq j \leq d_\bullet$, where $\widetilde{\boldsymbol{u}}_j$ is continuous and $\boldsymbol{\gamma}_{u\,j}$ is binary. Therefore, we can use the posterior distribution of γ_{uj} to express the posterior distribution of $\boldsymbol{u}_j$:

$$P(\boldsymbol{u}_j = \mathbf{0}\big|\boldsymbol{y}) = P(\gamma_{uj} = 0\big|\boldsymbol{y}) = 1 - E(\gamma_{uj}\big|\boldsymbol{y}),\ \ 1 \leq j \leq d_\bullet.$$

We use a general rule to be:

$$\boldsymbol{u}_j = 0 \ \ \text{if and only if } P(\gamma_{uj} = 0\big|\boldsymbol{y}) > \tau, \text{ and } \tau = \frac{1}{2},\ \ 1 \leq j \leq d_\bullet, \tag{2.53}$$

which is equivalent to

$$E(\gamma_{uj}\big|\boldsymbol{y}) \leq 1 - \tau \ \ \text{and} \ \ \tau = \frac{1}{2},\ \ 1 \leq j \leq d_\bullet.$$

Recalling that $\boldsymbol{\beta}$ is a $(d_\circ + d_\bullet) \times 1$ linear coefficient vector, and $\boldsymbol{u}_j, 1 \leq j \leq d_\bullet$, being a $K_j \times 1$ spline coefficient vector attached to $\dot{\boldsymbol{x}}_i, 1 \leq i \leq n$, under the Bayesian group LASSO approach to deciding between linear and nonlinear effects based on the maximum a posteriori estimate of $\boldsymbol{u}_j$, our strategy of deciding between an effect being zero, linear or non-linear is:

$$\begin{aligned}
&\text{the effect is zero if } \begin{cases} E(\gamma_{\beta j}\big|\boldsymbol{y}) \leq 1 - \tau,\ \ 1 \leq j \leq d_\circ + d_\bullet, \text{ and} \\ E(\gamma_{uj}\big|\boldsymbol{y}) \leq 1 - \tau,\ \ 1 \leq j \leq d_\bullet, \end{cases} \\
&\text{the effect is linear if } \begin{cases} E(\gamma_{\beta j}\big|\boldsymbol{y}) > 1 - \tau,\ \ 1 \leq j \leq d_\circ + d_\bullet, \text{ and} \\ E(\gamma_{uj}\big|\boldsymbol{y}) \leq 1 - \tau,\ \ 1 \leq j \leq d_\bullet, \end{cases} \\
&\text{otherwise the effect is non-linear.}
\end{aligned} \tag{2.54}$$

For Markov chain Monte Carlo approach, $E(\gamma_{\beta_j}\big|\boldsymbol{y}), 1 \leq j \leq d_\circ + d_\bullet$, is approximate the sample mean of $\gamma_{\beta_j}^{[g]}$ in Algorithm 2, $1 \leq g \leq N_{\text{kept}}$. $E\left(\gamma_{uj}\big|\boldsymbol{y}\right), 1 \leq j \leq d_\bullet$, is approximate the sample mean of $\gamma_{uj}^{[g]}$ in Algorithm 2, $1 \leq g \leq N_{\text{kept}}$.

From the rules (2.52) of Section 2.6.1 and (2.54) of this subsection, we know that the lower values of τ producing sparse fits, which means τ controls the degree of sparsity in the model selection. Therefore, we call τ as the sparsity threshold pa-

rameter. Even though different values of τ can be used in model selection, for a completely automatic model selection methodology, a reasonable and good choice of default value of sparsity threshold parameter τ is desirable. We will discuss this problem in the next subsection.

2.6.3 Choice of Default Value of Sparsity Threshold Parameter

We use the misclassification rate to measure the classification performance of Markov chain Monte Carlo. For each simulation run, the misclassification rate is the proportion, expressed as a percentage, of the predictors that are misclassified by the methodology into the classes: zero, linear and non-linear. The lower of the misclassification rate, is the better.

Using both Gaussian and binary response generalized additive models data sets for simulation, we have

$$\begin{aligned} &\text{30 continuous predictors:} \\ &\text{10 having zero effect;} \\ &\text{10 having linear effects;} \\ &\text{10 having non-linear effects.} \end{aligned} \tag{2.55}$$

For each simulation run, the misclassification rate is the proportion, expressed as a percentage, of the 30 predictors that are misclassified by the methodology into the classes: {zero, linear, non-linear} . The generation of data are as following:

$$\begin{aligned} &\text{Each of the predictors were generated from independent Normal distributions.} \\ &\text{The non-linear effects corresponded to quintic polynomials with} \\ &\qquad \text{random generated coefficients.} \end{aligned} \tag{2.56}$$

The design of simulation run for each combination of sample size n and error standard deviation σ_ε to be:

$$\begin{aligned} &\text{Running simulation on 100 data sets generated;} \\ &\text{Involving generation of new coefficients.} \end{aligned} \tag{2.57}$$

The settings of the number of continuous predictors $d_\bullet$, sample size n, error standard deviations σ_ε, and sparsity threshold parameter τ are to be

$$\begin{aligned} & d_\circ = 0, \;\; d_\bullet = 30, \\ & n \in \{500, 1000, 2000\}, \\ & \sigma_\varepsilon \in \{0.25, 0.5, 1, 2\} \text{ for Gaussian response,} \\ & \tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}. \end{aligned} \tag{2.58}$$

For each combination of sample size n and error standard deviation σ_ε, we run simulation on 100 data sets generated. For Markov chain Monte Carlo approach, we use $N_{\text{warm}} = N_{\text{kept}} = 1000$ from Algorithm 2 in Section 2.5.3. The rules of model selection follow (2.52) of Section 2.6.1 and (2.54) of Section 2.6.2.

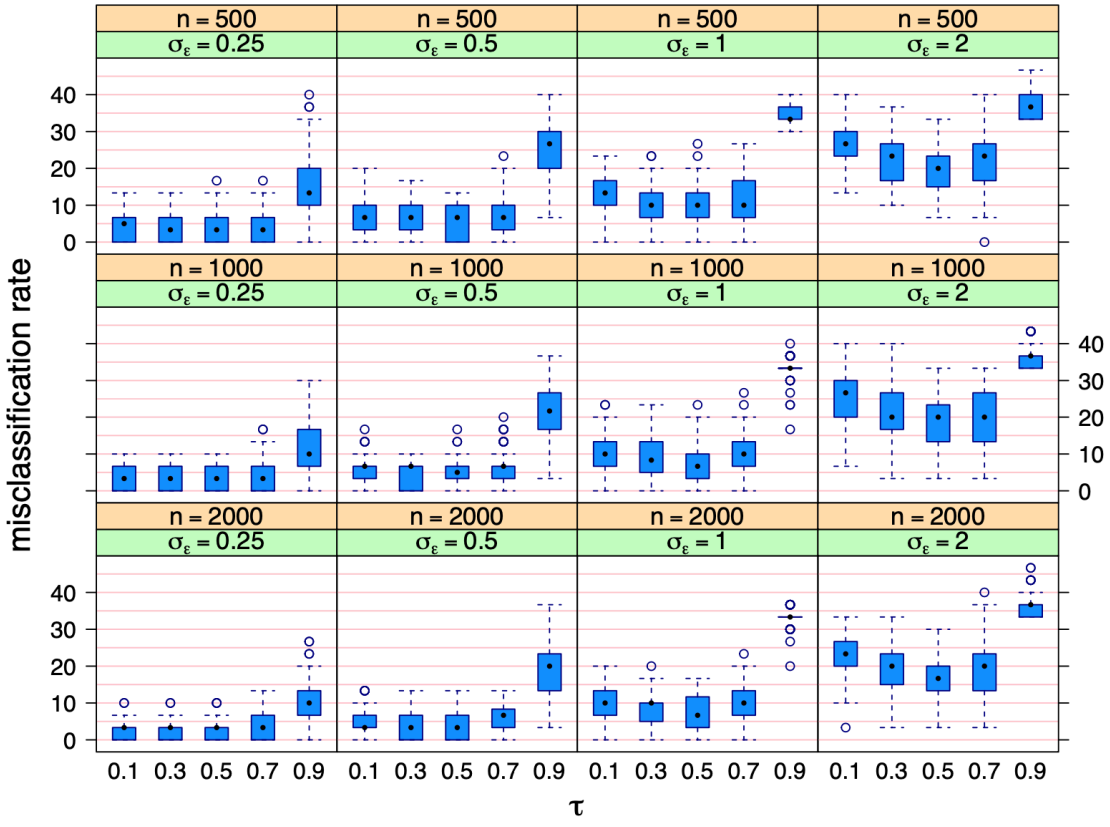


Figure 2.3: *Side-by-side boxplots of the misclassification rates for the Markov chain Monte Carlo Algorithm 2 for the simulation study described in the text. Each panel corresponds to a different combination of sample size and error standard deviation. Within each panel, the side-by-side boxplots compare misclassification rate as a function of the sparsity threshold parameter τ.*

Figure 2.3 shows the misclassification rate as a function of sparsity threshold parameter τ under Markov chain Monte Carlo. Each panel corresponds to a different combination of sample size n and error standard deviation σ_ε. Within each panel, the side-by-side box plots show the misclassification rate under different values of sparsity threshold parameter τ. For low noise level with $\sigma_\varepsilon < 1$, there is not much of a difference in misclassification rates. However, for high noise level with $\sigma_\varepsilon \geq 1$, it has the advantages of having τ being 0.5 for lower misclassification rate. Therefore, the default value of sparsity threshold parameter τ for Markov chain Monte Carlo is 0.5.

Our simulation study that suggests a good default threshold value of 0.5 for the MCMC is conducted on a somewhat restricted set of data generated. We list some trouble-shooting tips that may aid successful use of our R gamselBayes to minimize the possibility of model misspecification:

- Ensure that the input data are free of missing values and non-numerical objects.
- Predictors containing gross outliers can lead to problems with penalized spline fitting, so checks and pre-processing to avoid such an issue may be worthwhile.
- If a predictor is strongly skewed then problems with penalized spline fitting are also more likely. Hence, depending on the strength of the skewness, pre-transformation of such predictors may be beneficial.
- In theory, continuous variables have a large number of unique values. However, if its measurements have been discretized to the point that the number of unique values is smaller than around 15–25 then this can lead to problems with penalized spline fitting. Such a predictor should be restricted to zero or linear, but not non-linear.
- If the sample size is smaller than around 15–25 then all candidate predictors have the limitation described there and it may not be feasible to include them in the input.

To achieve interpretable and parsimonious models, the effect type misclassification rate is one of the measurements for model selection. Another measurement is relative test error, which is one of the five evaluation metrics from the simulation studies of Hastie et al., 2020.

The settings of the number of continuous predictors $d_\bullet$, sample size n, error standard deviations σ_ε, and sparsity threshold parameter τ, are as in (2.58). The continuous predictors generation for data sets are as in Section 2.6.3.

Then we re-ran the Gaussian response simulation studies of Section 2.6.3. Suppose

$$\begin{aligned} &\text{selected model based on data set } \mathscr{D} \text{ corresponds to some additive function } \widehat{f}, \\ &\text{true model corresponds to } f_{\text{true}}, \\ &\text{continuous predictor } \boldsymbol{x} \text{ is a } d_\bullet \times 1 \text{ random vector with density function } \mathfrak{p}(\boldsymbol{x}). \end{aligned} \tag{2.59}$$

Then the relative test error is

$$E\left[\left\{y - \widehat{f}(\boldsymbol{x})\right\}^2 \middle| \mathscr{D}\right] \Big/ \sigma_\varepsilon^2 \quad \text{where} \quad y \sim N(f_{\text{true}}(\boldsymbol{x}), \sigma_\varepsilon^2). \tag{2.60}$$

In (2.60), the expectation is over the predictor distribution corresponding to $\mathfrak{p}(\boldsymbol{x})$. The denominator σ_ε^2 is the *Bayes error*, corresponding to the situation where $\widehat{f} = f_{\text{true}}$. Therefore, the relative test error measures the expected test error relative to the Bayes error rate. The perfect score is 1 when f_{true} is estimated perfectly.

Our simulation study used 100,000 draws from the predictor distribution. To aid visualization, we use the $\log_{10}$ transformation to the relative test error values. Figure 2.4 shows that the relative test errors are lower for higher sample sizes under the same noise level. And it also shows that the relative test errors are lower for higher noise level under the same sample size. However, we cannot just simply compare the relative test errors across different noise levels by σ_ε. As the estimators are subject to bias, there is no clear-cut comparisons of relative test errors across various noise levels. Figure 2.4 shows that the choice of sparsity threshold parameter τ does not affect the relative test errors.

The simulation study in this subsection replace the effect type misclassification rate used in Section 2.6.3 with the relative test error. It does not change our recommendation concerning the choice of sparsity threshold parameter τ being 0.5 for Markov chain Monte Carlo described in Section 2.6.3.

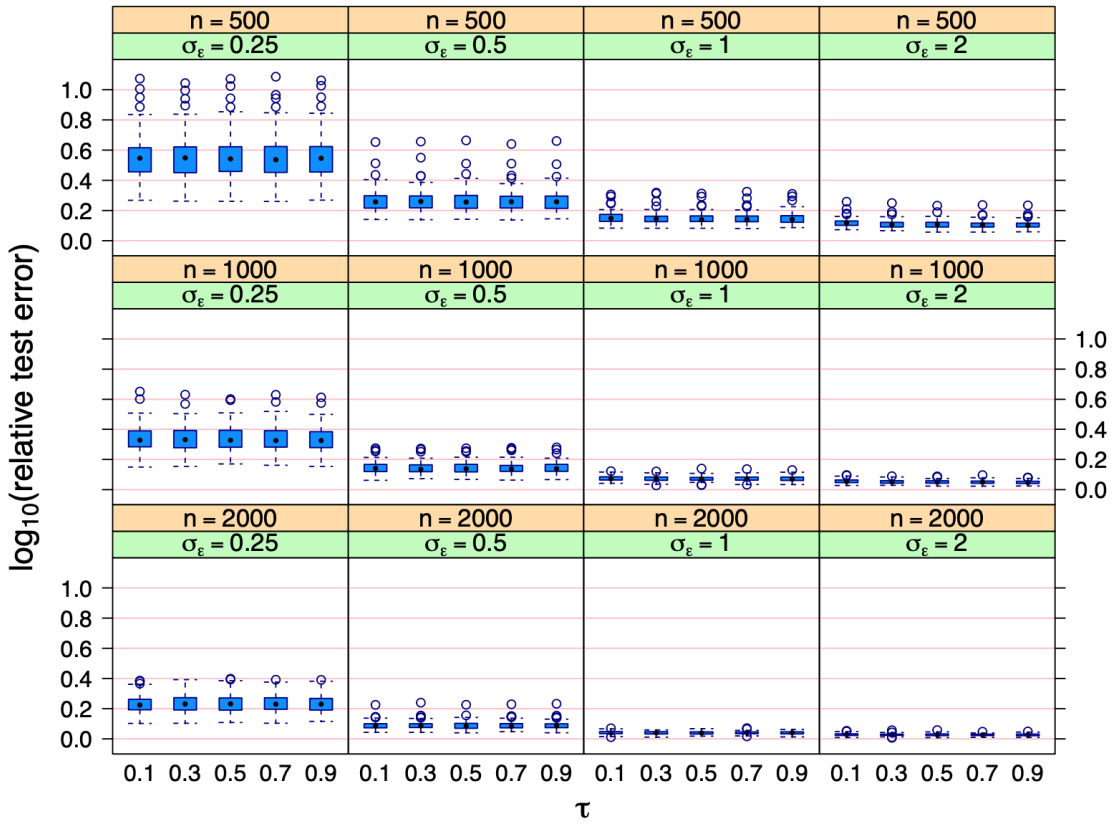


Figure 2.4: *Side-by-side boxplots of the algorithms, to base 10, of relative test error for the Markov chain Monte Carlo Algorithm 2 for the simulation study described in the text. Each panel corresponds to a different combination of sample size and error standard deviation. Within each panel, the side-by-side boxplots compare relative test error as a function of the sparsity threshold parameter τ.*

2.7 Data Illustration for Markov Chain Monte Carlo

We implemented our MCMC approach as given by Algorithm 2 in Section 2.5.3 in our R package gamselBayes. In this subsection, we apply our MCMC approach to two real data sets. The first data illustration involves a relative small mortgage application data set using the MCMC approach. The second data illustration involves data concerning schools in California. For the linear effects variables, we want to know their positive or negative impacts on the response. For the non-linear effects, we want to know their forms. These are achieved by using our R package gamselBayes.

2.7.1 Application to Mortgage Applications Data

The data frame BostonMortgages within the HRW package (Harezlak et al., 2018) contains a data set from the Federal Bank of Boston, U.S.A.. There are 2,380 records in the mortgage application data. There are 18 predictors, some of which are obtained by conversion from categorical variables to indicator forms. Fourteen are binary predictors and four are continuous predictors. The fourteen binary predictors can be categorized as having linear or zero effects. The credit score ranges from one to six, where a low value indicates low credit risk. The remaining four continuous predictors can be categorized as having linear, non-linear or zero effects. One of the continuous predictor has only 10 unique values. As penalized spline models have borderline viability, this predictor can only be considered having linear or zero effect. The binary response variable is the indicator of whether the mortgage application was denied.

Since the number of predictors is not large, we use the MCMC approach as given by Algorithm 2 in Section 2.5.3 for model selection, using default warm-up and kept chain sizes both as $1,000,$ and with the threshold parameter τ being 0.5 described in Section 2.6.3.

Using gamselBayes, we obtained the categorization results shown in Table 2.1. It shows that 7 predictors have linear effects, and 2 predictors have non-linear effects. Note that 9 predictors were discarded.

We obtained an inferential summary in Table 2.2 for all the selected linear fits. It shows that being self-employed, single or African-American denoted as self, single and black respectively, each have positive impacts on the probability of mortgage denial. Of potential interest from a social justice standpoint is whether there is a significant effect on mortgage denial probability for applicants that are either black or single. It also shows that an applicant having bad public credit record is more likely to have their mortgage application denied, which is in keeping with financial common sense.

There are two predictors having non-linear effects on mortgage denial, which are ratio of the debt payments to the total income and ratio of the loan size to the assessed value of the property. We obtained Figure 2.5 for the non-linear effects of these two predictors on mortgage denial. The left panel of Figure 2.5 shows the effect of debt payment to income ratio is quite a striking non-monotonic curve. Both low debt payments to income ratios and especially high values of this ratio increase the probability of mortgage application denial. The left curve shows that the prob-

Predictors selected as having linear effects:
credit score being 1, credit score being 2, is the applicant black?
was the applicant single?
was the applicant denied mortgage insurance?
did the application have a public bad credit record?
was the applicant self-employed?
Predictors selected as having non-linear effects:
ratio of the debt payments to the total income,
ratio of the loan size to the assessed value of property.

Table 2.1: *Categorization results obtained from Boston mortgage application data set using MCMC approach as given by Algorithm 2.*

	posterior mean	95% credible interval
credit score being 1	–0.6906	$(-0.898, -0.4513)$
credit score being 2	–0.3238	$(-0.5869, -0.0000)$
is the applicant black?	0.3461	$(0.08425, 0.5404)$
was the applicant denied mortgage insurance?	2.762	$(2.143, 3.517)$
did the application have a public bad credit record?	0.735	$(0.4926, 0.9848)$
was the applicant self-employed?	0.1703	$(0.0000, 0.4327)$
was the applicant single?	0.1368	$(0.0000, 0.3417)$

Table 2.2: *Inferential summary of the linear coefficients obtained from Boston mortgage application data set using MCMC approach as given by Algorithm 2.*

ability that a mortgage is denied is decreasing in the range from 0 to 0.3 of the debt payments to income ratio, and is increasing beyond 0.3. It indicates that the best chance of not being denied a mortgage is when one's debt payments to income ratio is between 0.2 and 0.4 and the loan size to assessed valued is low, between 0.2 and

0.8.

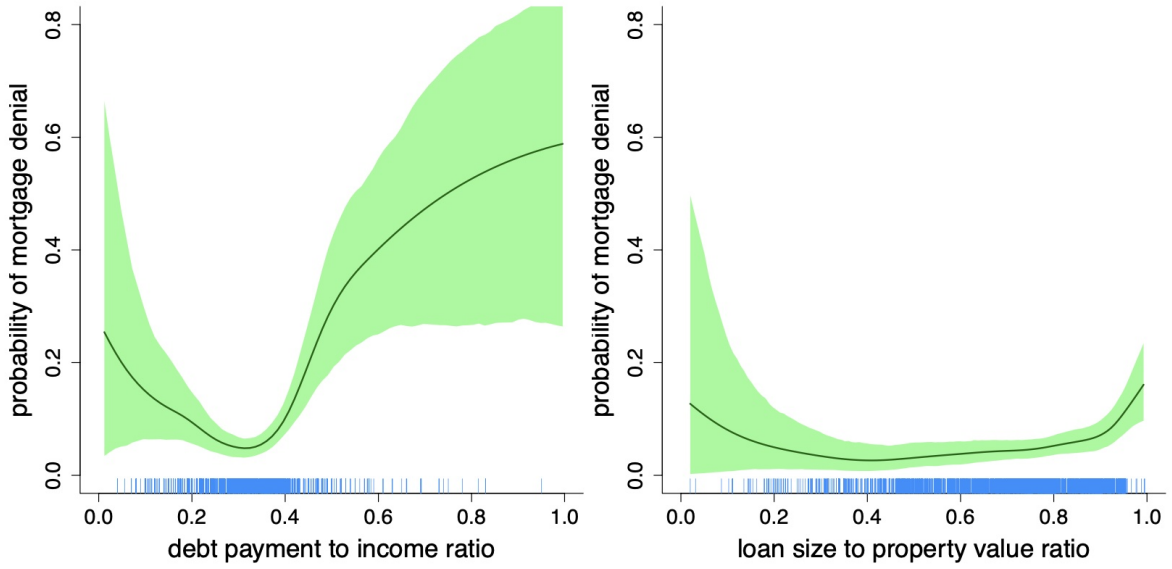


Figure 2.5: *The two estimated non-linear effects for the Boston mortgage example from application of Algorithm 2 and effect type estimation rules of Section 2.6 with* $\tau = 0.5$*. Each curve is the slice of estimated probability of mortgage denial as a function of the predictor, with all other selected predictors set to their median values. The shade region corresponds to pointwise approximate* 95% *credible intervals.*

2.7.2 Application to California Schools Data

The data are from R package Ecdat (Crossaint, 2020) in the data frame named Cashchool. The data frame contains 9 predictors for 420 school districts in the state of California, U.S.A.. We use the average mathematics score for the district as the response. The Gaussian assumption seems reasonable. We choose the data for schools have grades spanning from kindergarten to year 8.

Using the MCMC approach in our R package gamselBayes, as given by Algorithm 2 in Section 2.5.3 for model selection, using default warm-up and kept chain sizes both as $1,000$, with the threshold parameter τ being 0.5 described in Section 2.6.3, we obtained the results shown in Table 2.3. It shows that 4 predictors out of 9 were selected and considered having linear effects.

We display the inferential summary of all the selected linear fits in Table 2.4. It shows that, predictor district average income has the largest positive effect with mean mathematics score.

Predictors selected as having linear effects:
average income of the district, percentage of English learners.
percentage of students qualifying for reduced-price lunches,
school has grades spanning from kindergarten to year 8.

Table 2.3: *Categorization results obtained from California Schools application data set using MCMC approach as given by Algorithm 2, with defaulted warm-up and kept chain sizes of* $1,000$.

	posterior mean	95% credible interval
average income of the district	0.7158	$(0.525, 0.8946)$
percentage of English learners	–0.1391	$(-0.2241, 0.0000)$
percentage of students qualifying for reduced-price lunches	–0.3504	$(-0.4322, -0.2532)$
school has grades spanning from kindergarten to year 8	–3.088	$(-6.571, 0.0000)$

Table 2.4: *Inferential summary of the linear coefficients obtained from California Schools data set, with defaulted warm-up and kept chain sizes of* $1,000$.

We may consider increasing the warm-up and kept chain sizes from $1,000$ as default to $2,000$. It leads to a new MCMC fit with a two-fold increase in the chain sizes compared to their default values. We obtained the coefficient summary in Table 2.5.

It shows that variables percentage of students qualifying for reduced-price lunches, percentage of English learners and school going up to Year 8, all have significant negative effect on mean mathematics score. The average income of the district denoted as avginc has a significant positive effect on mean mathematics score.

Three continuous predictors: total enrolment, average income of the district and

	posterior mean	95% credible interval
average income of the district	0.7189	$(0.5284, 0.9122)$
percentage of English learners	–0.1387	$(-0.2197, -0.03157)$
percentage of students qualifying for reduced-price lunches	–0.3516	$(-0.4383, -0.2519)$
school has grades spanning from kindergarten to year 8	–3.026	$(-6.553, 0.0000)$

Table 2.5: *Inferential summary of the linear coefficients obtained from California Schools data set, with increased warm-up and kept chain sizes of* $2,000$.

percentage of English learners for analysis via the MCMC fitting are quite skewed. We use the logarithmic transformations on these three predictors to see if skewness would be reducing.

We obtained Table 2.6 for the categorical results with logarithmic transformations. Compared to Table 2.3 without logarithmic transformation, the main changes of Table 2.6 is the presence of a non-linear effect for the logarithm of the district average income. Table 2.6 shows that 5 predictors out of 9 were selected. Four predictors were considered having linear effect on the mean response of mathematics score. We display the summary for all the selected linear fits in Table 2.7. Figure 2.6 shows the logarithm of predictor the district average income denoted has two increasing steadily ramps and a plateau.

Predictors selected as having linear effects:
percentage of students qualifying for the CalWORKS welfare program,
logarithm of percentage of English learners,
logarithm of percentage of students qualifying for reduced-price lunches,
school has grades spanning from kindergarten to year 8.

Predictors selected as having non-linear effects:
logarithm of average income of the district.

Table 2.6: *Categorization results obtained from California Schools data set with logarithmic transformations for MCMC fitting as given by Algorithm 2.*

	posterior mean	95% credible interval
percentage of students qualifying for the CalWORKS welfare program	–0.08745	$(-0.26, 0.0000)$
logarithm of percentage of English learners	–1.848	$(-2.93, -0.7177)$
percentage of students qualifying for reduced-price lunches	–0.3888	$(-0.4344, -0.2933)$
school has grades spanning from kindergarten to year 8	–4.584	$(-7.582, 0.0000)$

Table 2.7: *Inferential summary of the linear coefficients obtained from California Schools data set, with logarithmic transformation on three predictors, using MCMC approach as given by Algorithm 2.*

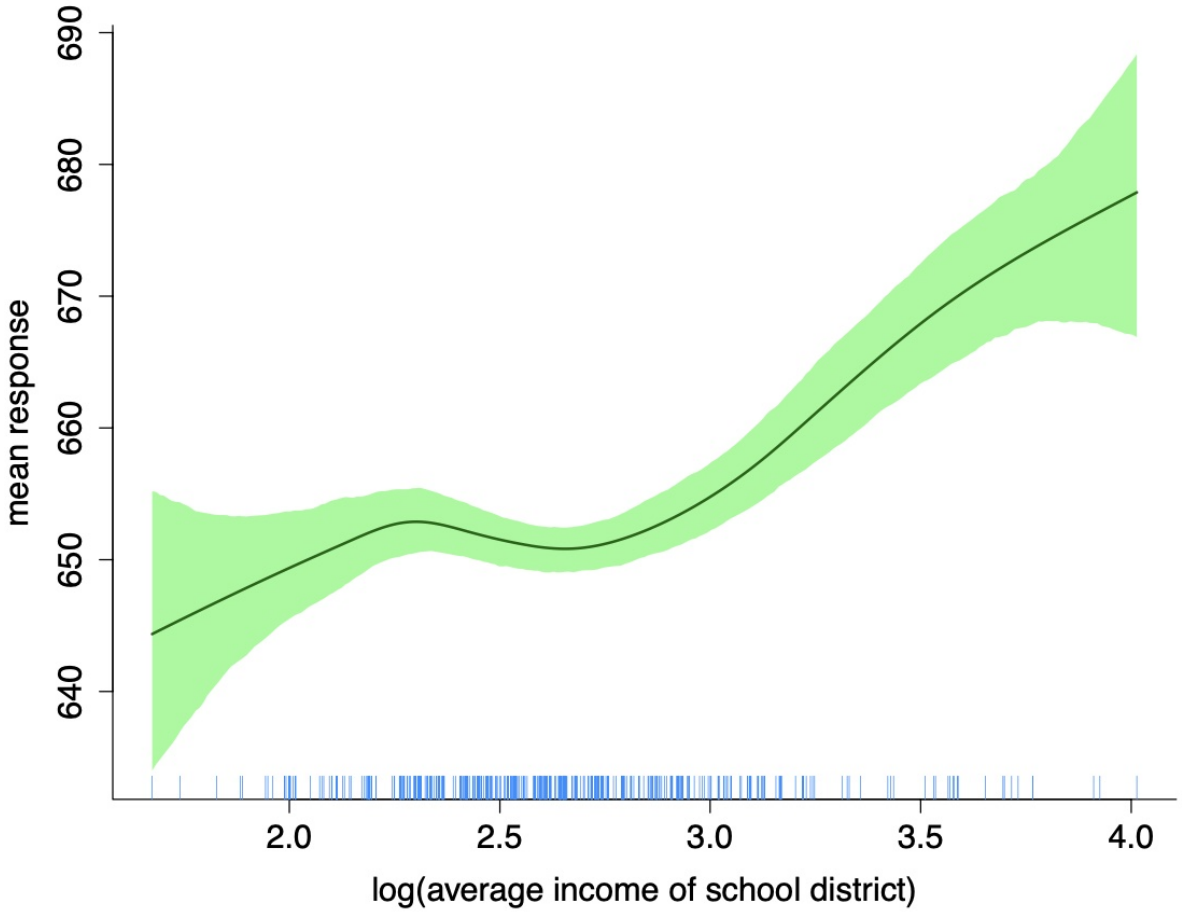


Figure 2.6: *The plot produced from the MCMC approach as given by Algorithm 2 on California Schools data set. The shaded regions correspond to pointwise* 95% *credible intervals. The rug at the base of the plot show values of each predictor.*

2.A The Canonical Demmler-Reinsch Spline Basis

Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be a continuous univariate data set. In the context of this thesis, the x_is correspond to values of a continuous candidate predictor. Let $[a, b]$ be an interval containing the x_is. For an integer $K \leq n - 2$, let $\boldsymbol{\kappa}_{\text{inter.}} \equiv (\kappa_1, \ldots, \kappa_{K-2})$ be a set of so-called interior knots such that

$$a < \kappa_1 < \ldots < \kappa_{K-2} < b.$$

A reasonable default value for K is around 30, or smaller values if the number of unique x_is is lower. It is common to place the interior knots at sample quantiles of the x_is.

We now list steps for construction of the matrix $\boldsymbol{Z}$ containing the canonical Demmler-Reinsch basis functions of the entries of $\boldsymbol{x}$. The justification for Steps (3)-(6) is given in Section 9.1.1 of Ngo and Wand (2004).

(1) Use the steps described in Section 4 of Wand and Ormerod (2008) to obtain the matrix denoted by $\boldsymbol{Z}$ in that section's equation (6), which contains canonical O'Sullivan spline basis functions. Denote this matrix by $\boldsymbol{Z}_{\text{OS}}$ and note that it has dimension $n \times K$.

(2) Form the matrix $\boldsymbol{C}_{\text{OS}} \equiv [\mathbf{1}_n \; \boldsymbol{x} \; \boldsymbol{Z}_{\text{OS}}]$ and set $\boldsymbol{D} = \text{diag}(0, 0, \mathbf{1}_K)$.

(3) Obtain the singular value decomposition of $\boldsymbol{C}_{\text{OS}}$:

$\boldsymbol{C}_{\text{OS}} = \boldsymbol{U}_{\text{C}}\text{diag}(\boldsymbol{d}_{\text{C}})\boldsymbol{V}_{\text{C}}^T$ where $\boldsymbol{U}_{\text{C}}$ is $n \times (K+2)$ and $\boldsymbol{V}_{\text{C}}$ is $(K+2) \times (K+2)$ such that $\boldsymbol{U}_{\text{C}}^T\boldsymbol{U}_{\text{C}} = \boldsymbol{V}_{\text{C}}^T\boldsymbol{V}_{\text{C}} = \boldsymbol{I}_{K+2}$.

(4) Form the symmetric matrix $\text{diag}(\mathbf{1}/\boldsymbol{d}_{\text{C}})\boldsymbol{V}_{\text{C}}^T\boldsymbol{D}\boldsymbol{V}_{\text{C}}\text{diag}(\mathbf{1}/\boldsymbol{d}_{\text{C}})$ and obtain its singular value decomposition:

$\text{diag}(\mathbf{1}/\boldsymbol{d}_{\text{C}})\boldsymbol{V}_{\text{C}}^T\boldsymbol{D}\boldsymbol{V}_{\text{C}}\text{diag}(\mathbf{1}/\boldsymbol{d}_{\text{C}}) = \boldsymbol{U}_D\text{diag}(\boldsymbol{d}_D)\boldsymbol{V}_D^T$ where $\boldsymbol{U}_D$ is $(K+2) \times (K+2)$

and $\boldsymbol{V}_D$ is $(K+2) \times (K+2)$ such that $\boldsymbol{U}_D^T\boldsymbol{U}_D = \boldsymbol{V}_D^T\boldsymbol{V}_D = \boldsymbol{I}_{K+2}$.

(5) Set the full canonical Demmler-Reinsch matrix as follows: $\boldsymbol{C}_{\text{DR}} \longleftarrow \boldsymbol{U}_C\boldsymbol{U}_D$.

(6) The next steps assume that the singular value decompositions follow the convention that $\boldsymbol{d}_D$ is a $(K+2) \times 1$ vector with its entries in non-increasing order.

Adjustments to the singular value decompositions are needed if this convention is not used.

(7) Set the $(K+2) \times 1$ vector $\boldsymbol{s}_D$ as follows:

$$\omega_{21} \longleftarrow \sqrt{K\text{th entry of } \boldsymbol{d}_D}, \quad ; \quad \boldsymbol{s}_D \longleftarrow \omega_{21}\mathbf{1}_{K+2} \Big/ \sqrt{\boldsymbol{d}_D}$$

and then set the last two entries of $\boldsymbol{s}_D$ to equal 1.

(8) Set the full canonical Demmler-Reinsch design matrix as follows:

$$\boldsymbol{C}_{\text{cDR}} \longleftarrow \boldsymbol{C}_{\text{DR}}\text{diag}(\boldsymbol{s}_D).$$

(9) Set the O'Sullivan to canonical Demmler-Reinsch transformation matrix as follows:

$$\boldsymbol{L}_{\text{OS.to.cDR}} \longleftarrow \boldsymbol{V}_C\text{diag}(\mathbf{1}/\boldsymbol{d}_C)\boldsymbol{U}_D\text{diag}(\boldsymbol{s}_D).$$

This $(K+2)\,(K+2)$ matrix has the following property:

$$\boldsymbol{C}_{\text{OS}}\boldsymbol{L}_{\text{OS.to.cDR}} = \boldsymbol{C}_{\text{cDR}}$$

and is very useful for prediction and plotting purposes. This is because gridwise analogues of $\boldsymbol{C}_{\text{OS}}$ are readily computed using the structures described in Wand and Ormerod (2008) involving cubic B-spline basis functions.

(10) Reverse the order of the columns of $\boldsymbol{C}_{\text{cDR}}$. Reverse the order of the columns of $\boldsymbol{L}_{\text{OS.to.cDR}}$.

(11) The matrix containing canonical spline basis functions of the inputs $\boldsymbol{x}$ and $\boldsymbol{\kappa}_{\text{inter.}}$. is

$$\boldsymbol{Z} \longleftarrow \text{the } n \times K \text{ matrix consisting of columns 3 to } K+2 \text{ of } \boldsymbol{C}_{\text{cDR}}.$$

A function in the R language for computing $\boldsymbol{Z}$ and $\boldsymbol{L}_{\text{OS.to.cDR}}$ for given $\boldsymbol{x}$ and $\boldsymbol{\kappa}_{\text{inter.}}$ can be accessed by downloading the accompanying gamselBayes package. Assuming that the gamselBayes package is installed, the relevant function is `gamselBayes:::ZcDR()`.

2.B Derivations of Full Conditional Distributions

2.B.1 Full Conditional Distribution of β_0

First note that

$$\mathfrak{p}(\beta_0|\text{rest}) \propto \mathfrak{p}(\boldsymbol{y}|\beta_0, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u, \sigma_\varepsilon^2)\mathfrak{p}(\beta_0).$$

Recalling that

$$\boldsymbol{y}|\beta_0, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u, \sigma_\varepsilon^2 \sim N\left(\mathbf{1}_n\beta_0 + \boldsymbol{X}(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) + \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j), \sigma_\varepsilon^2 \boldsymbol{I}_n\right)$$

and

$$\beta_0 \sim N(0, \sigma_{\beta_0}^2),$$

if we define $\boldsymbol{r}_{\beta_0} \equiv \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j)$, we have

$$\mathfrak{p}(\boldsymbol{y}|\beta_0, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u, \sigma_\varepsilon^2) \propto \exp\left\{\left(-\frac{1}{2\sigma_\varepsilon^2}\right)\left\|\boldsymbol{y} - \mathbf{1}_n\beta_0 - \boldsymbol{X}(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j)\right\|^2\right\}$$

$$\propto \exp\left\{\left(-\frac{1}{2\sigma_\varepsilon^2}\right)\left\|\boldsymbol{r}_{\beta_0} - \mathbf{1}_n\beta_0\right\|^2\right\} \propto \exp\left\{\left(-\frac{1}{2\sigma_\varepsilon^2}\right)\left(-2\beta_0^T\mathbf{1}_n^T\boldsymbol{r}_{\beta_0} + n\beta_0^2\right)\right\}.$$

Therefore

$$\mathfrak{p}(\boldsymbol{y}|\beta_0, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u, \sigma_\varepsilon^2) \propto \exp\left\{\begin{bmatrix}\beta_0 \\ \beta_0^2\end{bmatrix}^T \times \begin{bmatrix}\mathbf{1}_n^T\left(\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j)\right) \\ -n/2\end{bmatrix}(1/\sigma_\varepsilon^2)\right\}.$$

Also we have

$$\mathfrak{p}(\beta_0) \propto \exp\left\{\begin{bmatrix}\beta_0 \\ \beta_0^2\end{bmatrix}^T \begin{bmatrix}0 \\ -1/2\end{bmatrix}(1/\sigma_{\beta_0}^2)\right\}.$$

Then

$$\mathfrak{p}(\beta_0|\text{rest}) \propto \exp\left\{ \begin{bmatrix} \beta_0 \\ \beta_0^2 \end{bmatrix}^T \begin{bmatrix} \left(\frac{1}{\sigma_\varepsilon^2}\right) \mathbf{1}_n^T \left(\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j) \right) \\ -\frac{1}{2}\left(\frac{n}{\sigma_\varepsilon^2} + \frac{1}{\sigma_{\beta_0}^2}\right) \end{bmatrix} \right\}. \tag{2.61}$$

Therefore $\mathfrak{p}(\beta_0|\text{rest})$ has a Normal distribution with natural parameter vector

$$\begin{bmatrix} \mathbf{1}_n^T \left(\boldsymbol{y} - \boldsymbol{X}\left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}\right) - \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j) \right) \Big/ \sigma_\varepsilon^2 \\ \frac{-n}{2\sigma_\varepsilon^2} - \frac{1}{2\sigma_{\beta_0}^2} \end{bmatrix}.$$

The variance is

$$\frac{1}{(n/\sigma_\varepsilon^2) + (1/\sigma_{\beta_0}^2)}.$$

The mean is

$$\frac{\frac{1}{\sigma_\varepsilon^2}\mathbf{1}_n^T \left(\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j) \right)}{(n/\sigma_\varepsilon^2) + (1/\sigma_{\beta_0}^2)}.$$

In summary, we obtain (2.36) in Section 2.5.3 to be

$$\beta_0|\text{rest} \sim N\left(\frac{\frac{1}{\sigma_\varepsilon^2}\mathbf{1}_n^T \left(\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j) \right)}{(n/\sigma_\varepsilon^2) + (1/\sigma_{\beta_0}^2)}, \frac{1}{(n/\sigma_\varepsilon^2) + (1/\sigma_{\beta_0}^2)} \right).$$

If we define

$$\omega_1 \equiv \mathbf{1}_n^T \left(\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j) \right),$$

recalling that (2.7) of Section 2.2.2 to be $\mathbf{1}_n^T \boldsymbol{X} = \mathbf{0}$ and (2.8) of Section 2.3.3 to be $\boldsymbol{Z}_j^T \mathbf{1}_n = \mathbf{0}_{K_j}, 1 \le j \le d_\bullet$, we have $\omega_1 = \mathbf{1}_n^T \boldsymbol{y}$. If define $\omega_2 \equiv (n/\sigma_\varepsilon^2) + (1/\sigma_{\beta_0}^2)$, then $\beta_0|\text{rest} \sim N\left(\frac{\omega_1}{(\sigma_\varepsilon^2)\,\omega_2}, \frac{1}{\omega_2}\right)$ is given in Algorithm 2 in Section 2.5.3.

2.B.2 Full Conditional Distribution of $\widetilde{\boldsymbol{\beta}}$

To aid the calculations of the full conditional distribution of $\widetilde{\boldsymbol{\beta}}$ we note that

$$\boldsymbol{X}(\boldsymbol{\gamma}_{\beta} \odot \widetilde{\boldsymbol{\beta}}) = \boldsymbol{X}\text{diag}(\boldsymbol{\gamma}_{\beta})\widetilde{\boldsymbol{\beta}} = \boldsymbol{X}_{\boldsymbol{\gamma}_{\beta}}\widetilde{\boldsymbol{\beta}}$$

where

$$\boldsymbol{X}_{\boldsymbol{\gamma}_{\beta}} \equiv \boldsymbol{X}\text{diag}(\boldsymbol{\gamma}_{\beta}).$$

Then note that

$$\mathfrak{p}(\widetilde{\boldsymbol{\beta}}|\text{rest}) \propto \mathfrak{p}(\boldsymbol{y}\big|\beta_0, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_{\beta}, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u, \sigma_{\varepsilon}^2)\mathfrak{p}(\widetilde{\boldsymbol{\beta}}|\sigma_{\beta}^2, \boldsymbol{b}_{\beta})$$

where

$$\boldsymbol{b}_{\beta} \equiv (b_{\beta_1}, \ldots, b_{\beta_{(d_\circ + d_\bullet)}}).$$

Recalling that

$$\boldsymbol{y}\big|\beta_0, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_{\beta}, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u, \sigma_{\varepsilon}^2 \sim N\left(\mathbf{1}_n\beta_0 + \boldsymbol{X}(\boldsymbol{\gamma}_{\beta} \odot \widetilde{\boldsymbol{\beta}}) + \sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j), \sigma_{\varepsilon}^2\boldsymbol{I}_n\right)$$

and

$$\widetilde{\beta}_j\big|\sigma_{\beta}^2, b_{\beta_j} \sim N(0, \sigma_{\beta}^2/b_{\beta_j}), \;\; 1 \leq j \leq d_\circ + d_\bullet,$$

if we define $\boldsymbol{r}_{\widetilde{\boldsymbol{\beta}}} \equiv \boldsymbol{y} - \mathbf{1}_n\beta_0 - \sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j)$, we have

$$\begin{aligned}
\mathfrak{p}(\boldsymbol{y}\big|\beta_0, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_{\beta}, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u, \sigma_{\varepsilon}^2) &\propto \exp\left\{\left(-\frac{1}{2\sigma_{\varepsilon}^2}\right)\left\|\boldsymbol{y} - \mathbf{1}_n\beta_0 - \boldsymbol{X}_{\boldsymbol{\gamma}_{\beta}}\widetilde{\boldsymbol{\beta}} - \sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j)\right\|^2\right\} \\
&\propto \exp\left\{\left(-\frac{1}{2\sigma_{\varepsilon}^2}\right)\left\|\boldsymbol{r}_{\widetilde{\boldsymbol{\beta}}} - \boldsymbol{X}_{\boldsymbol{\gamma}_{\beta}}\widetilde{\boldsymbol{\beta}}\right\|^2\right\} \\
&\propto \exp\left\{\left(-\frac{1}{2\sigma_{\varepsilon}^2}\right)\left(-2\widetilde{\boldsymbol{\beta}}^T\boldsymbol{X}_{\boldsymbol{\gamma}_{\beta}}^T\boldsymbol{r}_{\widetilde{\boldsymbol{\beta}}} + \left(\text{vec}(\widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}^T)\right)^T\text{vec}\left(\boldsymbol{X}_{\boldsymbol{\gamma}_{\beta}}^T\boldsymbol{X}_{\boldsymbol{\gamma}_{\beta}}\right)\right)\right\}.
\end{aligned}$$

Therefore

$$\mathfrak{p}(\boldsymbol{y}\big|\beta_0,\widetilde{\boldsymbol{\beta}},\boldsymbol{\gamma}_\beta,\widetilde{\boldsymbol{u}},\boldsymbol{\gamma}_u,\sigma_\varepsilon^2)$$

$$\propto \exp\left\{\begin{bmatrix}\widetilde{\boldsymbol{\beta}}\\ \\ \mathrm{vec}(\widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}^T)\end{bmatrix}^T\begin{bmatrix}\boldsymbol{X}_{\boldsymbol{\gamma}_\beta}^T\left(\boldsymbol{y}-\mathbf{1}_n\beta_0-\sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j)\right)\\ \\ -\frac{1}{2}\mathrm{vec}(\boldsymbol{X}_{\boldsymbol{\gamma}_\beta}^T\boldsymbol{X}_{\boldsymbol{\gamma}_\beta})\end{bmatrix}(1/\sigma_\varepsilon^2)\right\}.$$

We also have

$$\mathfrak{p}(\widetilde{\boldsymbol{\beta}}\big|\sigma_\beta^2,\boldsymbol{b}_\beta)\propto\exp\left(-\frac{\widetilde{\boldsymbol{\beta}}^T\mathrm{diag}(\boldsymbol{b}_\beta)\widetilde{\boldsymbol{\beta}}}{2\sigma_\beta^2}\right)$$

$$\propto\exp\left\{\begin{bmatrix}\widetilde{\boldsymbol{\beta}}\\ \mathrm{vec}(\widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}^T)\end{bmatrix}^T\begin{bmatrix}\mathbf{0}\\ -\frac{1}{2}\mathrm{vec}(\mathrm{diag}(\boldsymbol{b}_\beta)\end{bmatrix}(1/\sigma_\beta^2)\right\}.$$

Therefore

$$\mathfrak{p}(\boldsymbol{y}\big|\beta_0,\widetilde{\boldsymbol{\beta}},\boldsymbol{\gamma}_\beta,\widetilde{\boldsymbol{u}},\boldsymbol{\gamma}_u,\sigma_\varepsilon^2)$$

$$\propto\exp\left\{\begin{bmatrix}\widetilde{\boldsymbol{\beta}}\\ \\ \mathrm{vec}(\widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}^T)\end{bmatrix}^T\begin{bmatrix}\frac{\boldsymbol{X}_{\boldsymbol{\gamma}_\beta}^T}{\sigma_\varepsilon^2}\left(\boldsymbol{y}-\mathbf{1}_n\beta_0-\sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j)\right)\\ \\ -\frac{1}{2}\mathrm{vec}\left(\frac{\boldsymbol{X}_{\boldsymbol{\gamma}_\beta}^T\boldsymbol{X}_{\boldsymbol{\gamma}_\beta}}{\sigma_\varepsilon^2}+\frac{\mathrm{diag}\,(\boldsymbol{b}_\beta)}{\sigma_\beta^2}\right)\end{bmatrix}\right\}.$$

Then the first component of the natural parameter vector of $\mathfrak{p}(\widetilde{\boldsymbol{\beta}}|\text{rest})$ is

$$\boldsymbol{X}_{\boldsymbol{\gamma}_\beta}^T\left(\boldsymbol{y}-\mathbf{1}_n\beta_0-\sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j)\right)(1/\sigma_\varepsilon^2).$$

The inverse vec of the second component of the natural parameter vector is

$$-\frac{1}{2}\left(\frac{1}{\sigma_\varepsilon^2}\boldsymbol{X}_{\boldsymbol{\gamma}_\beta}^T\boldsymbol{X}_{\boldsymbol{\gamma}_\beta}+\frac{1}{\sigma_\beta^2}\mathrm{diag}(\boldsymbol{b}_\beta)\right).$$

It follows that the covariance matrix of $\widetilde{\boldsymbol{\beta}}|\text{rest}$ is

$$\left(\frac{1}{\sigma_\varepsilon^2}\boldsymbol{X}_{\boldsymbol{\gamma}_\beta}^T\boldsymbol{X}_{\boldsymbol{\gamma}_\beta}+\frac{1}{\sigma_\beta^2}\text{diag}(\boldsymbol{b}_\beta)\right)^{-1},$$

and the mean vector is

$$\left(\frac{1}{\sigma_\varepsilon^2}\boldsymbol{X}_{\boldsymbol{\gamma}_\beta}^T\boldsymbol{X}_{\boldsymbol{\gamma}_\beta}+\frac{1}{\sigma_\beta^2}\text{diag}(\boldsymbol{b}_\beta)\right)^{-1}\frac{1}{\sigma_\varepsilon^2}\boldsymbol{X}_{\boldsymbol{\gamma}_\beta}^T\left(\boldsymbol{y}-\mathbf{1}_n\beta_0-\sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j)\right).$$

Therefore

$$\widetilde{\boldsymbol{\beta}}\big|\text{rest}\sim N\Bigg(\left(\frac{1}{\sigma_\varepsilon^2}\boldsymbol{X}_{\boldsymbol{\gamma}_\beta}^T\boldsymbol{X}_{\boldsymbol{\gamma}_\beta}+\frac{1}{\sigma_\beta^2}\text{diag}(\boldsymbol{b}_\beta)\right)^{-1}\frac{1}{\sigma_\varepsilon^2}\boldsymbol{X}_{\boldsymbol{\gamma}_\beta}^T\left(\boldsymbol{y}-\mathbf{1}_n\beta_0-\sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j)\right),$$

$$\left(\frac{1}{\sigma_\varepsilon^2}\boldsymbol{X}_{\boldsymbol{\gamma}_\beta}^T\boldsymbol{X}_{\boldsymbol{\gamma}_\beta}+\frac{1}{\sigma_\beta^2}\text{diag}(\boldsymbol{b}_\beta)\right)^{-1}\Bigg).$$

Next note that $\boldsymbol{X}_{\boldsymbol{\gamma}_\beta}^T\boldsymbol{X}_{\boldsymbol{\gamma}_\beta}=\text{diag}(\boldsymbol{\gamma}_\beta)\boldsymbol{X}^T\boldsymbol{X}\text{diag}(\boldsymbol{\gamma}_\beta)=(\boldsymbol{X}^T\boldsymbol{X})\odot(\boldsymbol{\gamma}_\beta\boldsymbol{\gamma}_\beta^T)$. Also

$$\begin{aligned}&\boldsymbol{X}_{\boldsymbol{\gamma}_\beta}^T\left(\boldsymbol{y}-\mathbf{1}_n\beta_0-\sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j)\right)\\&\quad=\boldsymbol{\gamma}_\beta\odot\left\{(\boldsymbol{X}^T\boldsymbol{y})-(\boldsymbol{X}^T\mathbf{1}_n\beta_0)-\sum_{j=1}^{d_\bullet}(\boldsymbol{X}^T\boldsymbol{Z}_j)(\gamma_{uj}\widetilde{\boldsymbol{u}}_j)\right\}.\end{aligned}$$

Then

$$\begin{aligned}\mathfrak{p}(\widetilde{\boldsymbol{\beta}}\big|\text{rest})\propto\exp\Bigg\{&\begin{bmatrix}\widetilde{\boldsymbol{\beta}}\\ \\ \text{vec}(\widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}^T)\end{bmatrix}^T\\&\times\begin{bmatrix}\left(\frac{1}{\sigma_\varepsilon^2}\right)\left[\boldsymbol{\gamma}_\beta\odot\left\{\boldsymbol{X}^T\left(\boldsymbol{y}-\mathbf{1}_n\beta_0-\sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j)\right)\right\}\right]\\ -\frac{1}{2}\text{vec}\left(\frac{(\boldsymbol{\gamma}_\beta\boldsymbol{\gamma}_\beta^T)\odot(\boldsymbol{X}^T\boldsymbol{X})}{\sigma_\varepsilon^2}+\frac{\text{diag}\,(\boldsymbol{b}_\beta)}{\sigma_\beta^2}\right)\end{bmatrix}\Bigg\}.\end{aligned}\tag{2.62}$$

Recalling that (2.7) of Section 2.2.2 to be $\boldsymbol{X}^T\mathbf{1}_n = \mathbf{0}$, and the matrices

$$\boldsymbol{X}^T\boldsymbol{y}, \quad \boldsymbol{X}^T\boldsymbol{X} \quad \text{and} \quad \boldsymbol{X}^T\boldsymbol{Z}_j, \quad 1 \leq j \leq d_\bullet,$$

only need to be computed once and can be stored and not re-computed inside Markov chain Monte Carlo iterative algorithms, then the full conditional distribution of $\widetilde{\boldsymbol{\beta}}$ is then obtained to be (2.37) of Section 2.5.3

$$\widetilde{\boldsymbol{\beta}}|\text{rest} \sim N\Bigg(\left(\frac{1}{\sigma_\varepsilon^2}(\boldsymbol{X}^T\boldsymbol{X}) \odot (\boldsymbol{\gamma}_\beta\boldsymbol{\gamma}_\beta^T) + \frac{1}{\sigma_\beta^2}\text{diag}(\boldsymbol{b}_\beta)\right)^{-1} \frac{1}{\sigma_\varepsilon^2}\Bigg[\boldsymbol{\gamma}_\beta \odot \Big\{\boldsymbol{X}^T\boldsymbol{y} - \sum_{j=1}^{d_\bullet}\left(\boldsymbol{X}^T\boldsymbol{Z}_j\right)(\gamma_{uj}\widetilde{\boldsymbol{u}}_j)\Big\}\Bigg], \left(\frac{1}{\sigma_\varepsilon^2}\left(\boldsymbol{X}^T\boldsymbol{X}\right) \odot \left(\boldsymbol{\gamma}_\beta\boldsymbol{\gamma}_\beta^T\right) + \frac{1}{\sigma_\beta^2}\text{diag}\left(\boldsymbol{b}_\beta\right)\right)^{-1}\Bigg).$$

If we define

$$\boldsymbol{\omega}_3 \equiv \boldsymbol{X}^T\boldsymbol{y} - \sum_{j=1}^{d_\bullet}\left(\boldsymbol{X}^T\boldsymbol{Z}_j\right)(\gamma_{uj}\widetilde{\boldsymbol{u}}_j),$$

and define

$$\boldsymbol{\Omega} \equiv \frac{1}{\sigma_\varepsilon^2}\left(\boldsymbol{X}^T\boldsymbol{X}\right) \odot \left(\boldsymbol{\gamma}_\beta\boldsymbol{\gamma}_\beta^T\right) + \frac{1}{\sigma_\beta^2}\text{diag}\left(\boldsymbol{b}_\beta\right),$$

then decompose $\boldsymbol{\Omega} = \boldsymbol{U}_{\boldsymbol{\Omega}}\text{diag}\left(\boldsymbol{d}_{\boldsymbol{\Omega}}\right)\boldsymbol{U}_{\boldsymbol{\Omega}}^T$ where $\boldsymbol{U}_{\boldsymbol{\Omega}}^T\boldsymbol{U}_{\boldsymbol{\Omega}} = \boldsymbol{I}$. Draw

$$\boldsymbol{z} \sim N\left(\mathbf{0}, \boldsymbol{I}\right)\left((d_\circ + d_\bullet) \times 1\right);$$

then

$$\boldsymbol{U}_{\boldsymbol{\Omega}}\left(\frac{\boldsymbol{U}_{\boldsymbol{\Omega}}^T\boldsymbol{z}}{\sqrt{\boldsymbol{d}_{\boldsymbol{\Omega}}}} + \frac{\boldsymbol{U}_{\boldsymbol{\Omega}}^T\left(\boldsymbol{\gamma}_\beta \odot \boldsymbol{\omega}_3\right)}{\boldsymbol{d}_{\boldsymbol{\Omega}}\left(\sigma_\varepsilon^2\right)}\right) \text{ has the same distribution as } \widetilde{\boldsymbol{\beta}}|\text{rest},$$

given in Algorithm 2 of Section 2.5.3.

2.B.3 Full Conditional Distribution of $\widetilde{\boldsymbol{u}}_j$

To aid the calculations of the full conditional distribution of $\widetilde{\boldsymbol{u}}_j, 1 \leq j \leq d_\bullet$, we note that

$$\boldsymbol{Z}_j\left(\gamma_{uj}\widetilde{\boldsymbol{u}}_j\right) = \boldsymbol{Z}_{\boldsymbol{\gamma}_{uj}}\widetilde{\boldsymbol{u}}_j,$$

where

$$\boldsymbol{Z}_{\boldsymbol{\gamma}_{uj}} \equiv \boldsymbol{Z}_j\left(\gamma_{uj}\right).$$

First note that

$$\mathfrak{p}\left(\widetilde{\boldsymbol{u}}_j\middle|\text{rest}\right) \propto \mathfrak{p}\left(\boldsymbol{y}\middle|\beta_0,\widetilde{\boldsymbol{\beta}},\boldsymbol{\gamma}_\beta,\widetilde{\boldsymbol{u}},\boldsymbol{\gamma}_u,\sigma_\varepsilon^2\right)\mathfrak{p}\left(\widetilde{\boldsymbol{u}}_j\middle|\sigma_{uj}^2,b_{uj}\right).$$

Recalling that

$$\boldsymbol{y}\middle|\beta_0,\widetilde{\boldsymbol{\beta}},\boldsymbol{\gamma}_\beta,\widetilde{\boldsymbol{u}},\boldsymbol{\gamma}_u,\sigma_\varepsilon^2 \sim N\left(\mathbf{1}_n\beta_0+\boldsymbol{X}(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}})+\sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j),\sigma_\varepsilon^2\boldsymbol{I}_n\right)$$

and

$$\widetilde{\boldsymbol{u}}_j\middle|\sigma_{uj}^2,b_{uj} \sim N\Big(0,(\sigma_{uj}^2/b_{uj})\boldsymbol{I}\Big),\ \ 1\le j\le d_\bullet,$$

if we define $\boldsymbol{r}_{\widetilde{\boldsymbol{u}}_j}\equiv\boldsymbol{y}-\mathbf{1}_n\beta_0-\boldsymbol{X}(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}})-\sum_{j'\neq j}^{d_\bullet}\boldsymbol{Z}_{j'}(\gamma_{uj'}\widetilde{\boldsymbol{u}}_{j'})$, we have

$$\begin{aligned}
&\mathfrak{p}\left(\boldsymbol{y}\middle|\beta_0,\widetilde{\boldsymbol{\beta}},\boldsymbol{\gamma}_\beta,\widetilde{\boldsymbol{u}},\boldsymbol{\gamma}_u,\sigma_\varepsilon^2\right)\\
&\quad\propto\exp\left\{\left(-\frac{1}{2\sigma_\varepsilon^2}\right)\left\|\boldsymbol{y}-\mathbf{1}_n\beta_0-\boldsymbol{X}(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}})-\sum_{j'\neq j}^{d_\bullet}\boldsymbol{Z}_{j'}(\gamma_{uj'}\widetilde{\boldsymbol{u}}_{j'})-\boldsymbol{Z}_{\boldsymbol{\gamma}_{uj}}\widetilde{\boldsymbol{u}}_j\right\|^2\right\}\\
&\quad\propto\exp\left\{\left(-\frac{1}{2\sigma_\varepsilon^2}\right)\left\|\boldsymbol{r}_{\widetilde{\boldsymbol{u}}_j}-\boldsymbol{Z}_{\boldsymbol{\gamma}_{uj}}\widetilde{\boldsymbol{u}}_j\right\|^2\right\}.\\
&\quad\propto\exp\left\{\left(-\frac{1}{2\sigma_\varepsilon^2}\right)\left(-2\widetilde{\boldsymbol{u}}_j^T\boldsymbol{Z}_{\boldsymbol{\gamma}_{uj}}^T\boldsymbol{r}_{\widetilde{\boldsymbol{u}}_j}+\text{vec}\left(\widetilde{\boldsymbol{u}}_j\widetilde{\boldsymbol{u}}_j^T\right)\text{vec}\left(\boldsymbol{Z}_{\boldsymbol{\gamma}_{uj}}^T\boldsymbol{Z}_{\boldsymbol{\gamma}_{uj}}\right)\right)\right\}.
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathfrak{p}\left(\boldsymbol{y}\middle|\beta_0,\widetilde{\boldsymbol{\beta}},\boldsymbol{\gamma}_\beta,\widetilde{\boldsymbol{u}},\boldsymbol{\gamma}_u,\sigma_\varepsilon^2\right)\propto\exp\Bigg\{&\begin{bmatrix}\widetilde{\boldsymbol{u}}_j\\ \\ \text{vec}(\widetilde{\boldsymbol{u}}_j\widetilde{\boldsymbol{u}}_j^T)\end{bmatrix}^T\\
&\times\begin{bmatrix}\boldsymbol{Z}_{\boldsymbol{\gamma}_{uj}}^T\left(\boldsymbol{y}-\mathbf{1}_n\beta_0-\boldsymbol{X}(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}})-\sum_{j'\neq j}^{d_\bullet}\boldsymbol{Z}_{j'}(\gamma_{uj'}\widetilde{\boldsymbol{u}}_{j'})\right)\\ \\ -\frac{1}{2}\text{vec}\left(\boldsymbol{Z}_{\boldsymbol{\gamma}_{uj}}^T\boldsymbol{Z}_{\boldsymbol{\gamma}_{uj}}\right)\end{bmatrix}(1/\sigma_\varepsilon^2)\Bigg\}.
\end{aligned}$$

We also have

$$\mathfrak{p}\left(\widetilde{\boldsymbol{u}}_j\middle|\sigma^2_{uj}, b_{uj}\right) \propto \exp\left\{-\frac{1}{2\sigma^2_{uj}}\text{vec}\left(\widetilde{\boldsymbol{u}}_j\widetilde{\boldsymbol{u}}_j^T\right)\text{vec}\left(b_{uj}\boldsymbol{I}_{K_j}\right)\right\}$$

$$\propto \exp\left\{\begin{bmatrix}\widetilde{\boldsymbol{u}}_j \\ \\ \text{vec}\left(\widetilde{\boldsymbol{u}}_j\widetilde{\boldsymbol{u}}_j^T\right)\end{bmatrix}^T\begin{bmatrix}\boldsymbol{0} \\ \\ -\frac{1}{2}b_{uj}\text{vec}\left(\boldsymbol{I}_{K_j}\right)\end{bmatrix}\left(1/\sigma^2_{uj}\right)\right\}.$$

Then

$$\mathfrak{p}\left(\widetilde{\boldsymbol{u}}_j\middle|\text{rest}\right)$$
$$\sim \exp\left\{\begin{bmatrix}\widetilde{\boldsymbol{u}}_j \\ \\ \text{vec}\left(\widetilde{\boldsymbol{u}}_j\widetilde{\boldsymbol{u}}_j^T\right)\end{bmatrix}^T\begin{bmatrix}\frac{\boldsymbol{Z}^T_{\boldsymbol{\gamma}_{uj}}}{\sigma^2_\varepsilon}\left(\boldsymbol{y}-\boldsymbol{1}_n\beta_0-\boldsymbol{X}(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}})-\sum_{j'\neq j}^{d_\bullet}\boldsymbol{Z}_{j'}(\gamma_{uj'}\widetilde{\boldsymbol{u}}_{j'})\right) \\ -\frac{1}{2}\text{vec}\left(\frac{\boldsymbol{Z}^T_{\boldsymbol{\gamma}_{uj}}\boldsymbol{Z}_{\boldsymbol{\gamma}_{uj}}}{\sigma^2_\varepsilon}+\frac{b_{uj}}{\sigma^2_{uj}}\boldsymbol{I}_{K_j}\right)\end{bmatrix}\right\}. \tag{2.63}$$

Therefore, the first component of the natural parameter vector of $\mathfrak{p}(\widetilde{\boldsymbol{u}}_j|\text{rest})$ is

$$\boldsymbol{Z}^T_{\boldsymbol{\gamma}_{uj}}\left(\boldsymbol{y}-\boldsymbol{1}_n\beta_0-\boldsymbol{X}(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}})-\sum_{j'\neq j}^{d_\bullet}\boldsymbol{Z}_{j'}(\gamma_{uj'}\widetilde{\boldsymbol{u}}_{j'})\right)\left(1/\sigma^2_\varepsilon\right).$$

The inverse vec of the second component of the natural parameter vector is

$$-\frac{1}{2}\left(\frac{1}{\sigma^2_\varepsilon}\boldsymbol{Z}^T_{\boldsymbol{\gamma}_{uj}}\boldsymbol{Z}_{\boldsymbol{\gamma}_{uj}}+\frac{b_{uj}}{\sigma^2_{uj}}\boldsymbol{I}_{K_j}\right).$$

It follows that the covariance matrix of $\widetilde{\boldsymbol{u}}_j\big|$rest is:

$$\left(\frac{1}{\sigma^2_\varepsilon}\boldsymbol{Z}^T_{\boldsymbol{\gamma}_{uj}}\boldsymbol{Z}_{\boldsymbol{\gamma}_{uj}}+\frac{b_{uj}}{\sigma^2_{uj}}\boldsymbol{I}_{K_j}\right)^{-1}.$$

The mean vector is

$$\left(\frac{1}{\sigma^2_\varepsilon}\boldsymbol{Z}^T_{\boldsymbol{\gamma}_{uj}}\boldsymbol{Z}_{\boldsymbol{\gamma}_{uj}}+\frac{b_{uj}}{\sigma^2_{uj}}\boldsymbol{I}_{K_j}\right)^{-1}\left(\frac{1}{\sigma^2_\varepsilon}\right)\boldsymbol{Z}^T_{\boldsymbol{\gamma}_{uj}}\left(\boldsymbol{y}-\boldsymbol{1}_n\beta_0-\boldsymbol{X}(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}})-\sum_{j'\neq j}^{d_\bullet}\boldsymbol{Z}_{j'}(\gamma_{uj'}\widetilde{\boldsymbol{u}}_{j'})\right).$$

Next, we simplify $\left(\frac{1}{\sigma_\varepsilon^2}\boldsymbol{Z}_{\boldsymbol{\gamma}_{u\,j}}^T\boldsymbol{Z}_{\boldsymbol{\gamma}_{u\,j}}+\frac{b_{uj}}{\sigma_{uj}^2}\boldsymbol{I}_{K_j}\right)^{-1}$. Note that $\boldsymbol{Z}_j^T\boldsymbol{Z}_j$ is a diagonal matrix from (2.8), if we define $\boldsymbol{Z}_j^T\boldsymbol{Z}_j \equiv \text{diag}\left(\boldsymbol{\omega}_{\boldsymbol{Z}_j}\right),\ \ 1\le j\le d_\bullet$, and note that $\gamma_{uj}^2=\gamma_{uj}$, then

$$\boldsymbol{Z}_{\boldsymbol{\gamma}_{u\,j}}^T\boldsymbol{Z}_{\boldsymbol{\gamma}_{u\,j}}=(\gamma_{uj})\boldsymbol{Z}_j^T\boldsymbol{Z}_j(\gamma_{uj})=(\gamma_{uj})\text{diag}\left(\boldsymbol{\omega}_{\boldsymbol{Z}_j}\right)(\gamma_{uj})=\gamma_{uj}\boldsymbol{\omega}_{\boldsymbol{Z}_j}.$$

This implies that

$$\frac{1}{\sigma_\varepsilon^2}\boldsymbol{Z}_{\boldsymbol{\gamma}_{u\,j}}^T\boldsymbol{Z}_{\boldsymbol{\gamma}_{u\,j}}+\frac{b_{uj}}{\sigma_{uj}^2}\boldsymbol{I}_{K_j}=\text{diag}\left(\frac{\gamma_{uj}\boldsymbol{\omega}_{\boldsymbol{Z}_j}}{\sigma_\varepsilon^2}+\frac{b_{uj}\mathbf{1}}{\sigma_{uj}^2}\right),$$

and therefore

$$\left(\frac{1}{\sigma_\varepsilon^2}\boldsymbol{Z}_{\boldsymbol{\gamma}_{u\,j}}^T\boldsymbol{Z}_{\boldsymbol{\gamma}_{u\,j}}+\frac{b_{uj}}{\sigma_{uj}^2}\boldsymbol{I}_{K_j}\right)^{-1}=\text{diag}\left(\frac{\mathbf{1}}{\left\{\gamma_{uj}\boldsymbol{\omega}_{\boldsymbol{Z}_j}\Big/\sigma_\varepsilon^2\right\}+(b_{uj}\mathbf{1})\Big/(\sigma_{uj}^2)}\right).$$

In summary, for each $1\le j\le d_\bullet$, the full conditional distribution of $\widetilde{\boldsymbol{u}}_j$ is

$$\widetilde{\boldsymbol{u}}_j\big|\text{rest}\sim N\Bigg(\frac{\frac{1}{\sigma_\varepsilon^2}\boldsymbol{Z}_{\boldsymbol{\gamma}_{u\,j}}^T\left(\boldsymbol{y}-\mathbf{1}_n\beta_0-\boldsymbol{X}(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}})-\sum_{j'\neq j}^{d_\bullet}\boldsymbol{Z}_{j'}(\gamma_{uj'}\widetilde{\boldsymbol{u}}_{j'})\right)}{\left\{\gamma_{uj}\boldsymbol{\omega}_{\boldsymbol{Z}_j}/\sigma_\varepsilon^2\right\}+(b_{uj}\mathbf{1})/\sigma_{uj}^2},$$

$$\text{diag}\left(\frac{\mathbf{1}}{\left\{\gamma_{uj}\boldsymbol{\omega}_{\boldsymbol{Z}_j}/\sigma_\varepsilon^2\right\}+(b_{uj}\mathbf{1})/(\sigma_{uj}^2)}\right)\Bigg).$$

Recalling that $\boldsymbol{Z}_{\gamma_{uj}}^T=\boldsymbol{Z}_j(\gamma_{uj})$, we have

$$\widetilde{\boldsymbol{u}}_j\big|\text{rest}\sim N\Bigg(\text{diag}\left(\frac{\gamma_{uj}\boldsymbol{\omega}_{\boldsymbol{Z}_j}}{\sigma_\varepsilon^2}+\frac{b_{uj}\mathbf{1}}{\sigma_{uj}^2}\right)^{-1}\left(\frac{1}{\sigma_\varepsilon^2}\right)\Big[\gamma_{uj}\Big(\boldsymbol{Z}_j^T\boldsymbol{y}-\boldsymbol{Z}_j^T\mathbf{1}_n\beta_0-\boldsymbol{Z}_j^T\boldsymbol{X}$$

$$\times\left(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}}\right)-\sum_{j'\neq j}^{d_\bullet}\boldsymbol{Z}_j^T\boldsymbol{Z}_{j'}\left(\gamma_{uj'}\widetilde{\boldsymbol{u}}_{j'}\right)\Big)\Big],\text{diag}\left(\frac{\gamma_{uj}\boldsymbol{\omega}_{\boldsymbol{Z}_j}}{\sigma_\varepsilon^2}+\frac{b_{uj}\mathbf{1}}{\sigma_{uj}^2}\right)^{-1}\Bigg).\tag{2.64}$$

Recalling that (2.8) of Section 2.3.3 to be $\boldsymbol{Z}_j^T \mathbf{1}_n = \mathbf{0}$, then

$$\mathfrak{p}(\widetilde{\boldsymbol{u}}_j \big| \text{rest}) \propto \exp\left\{ \begin{bmatrix} \widetilde{\boldsymbol{u}} \\ \text{vec}(\widetilde{\boldsymbol{u}}_j \widetilde{\boldsymbol{u}}_j^T) \end{bmatrix}^T \begin{bmatrix} \frac{1}{\sigma_\varepsilon^2}\Big[\gamma_{uj}\big\{\boldsymbol{Z}_j^T\boldsymbol{y} - \boldsymbol{Z}_j^T\boldsymbol{X}\left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}\right) - \sum_{j' \neq j}^{d_\bullet} \boldsymbol{Z}_j^T \boldsymbol{Z}_{j'}\left(\gamma_{uj'}\widetilde{\boldsymbol{u}}_{j'}\right)\big\}\Big] \\ -\frac{1}{2}\text{vec}\left(\text{diag}\left(\frac{\gamma_{uj}\boldsymbol{\omega}_{\boldsymbol{Z}_j}}{\sigma_\varepsilon^2} + \frac{b_{uj}\mathbf{1}}{\sigma_{uj}^2}\right)\right) \end{bmatrix} \right\}. \tag{2.65}$$

Therefore, we obtain (2.42) in Section 2.5.3 to be

$$\begin{aligned} \widetilde{\boldsymbol{u}}_j \big| \text{rest} \;\; \sim N\Bigg(& \text{diag}\left(\frac{\gamma_{uj}\boldsymbol{\omega}_{\boldsymbol{Z}_j}}{\sigma_\varepsilon^2} + \frac{b_{uj}\mathbf{1}}{\sigma_{uj}^2}\right)^{-1}\left(\frac{1}{\sigma_\varepsilon^2}\right)\Bigg[\gamma_{uj}\Big(\boldsymbol{Z}_j^T\boldsymbol{y} - \boldsymbol{Z}_j^T\boldsymbol{X}\left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}\right) \\ & - \sum_{j' \neq j}^{d_\bullet} \boldsymbol{Z}_j^T \boldsymbol{Z}_{j'}\left(\gamma_{uj'}\widetilde{\boldsymbol{u}}_{j'}\right)\Big)\Bigg], \text{diag}\left(\frac{\gamma_{uj}\boldsymbol{\omega}_{\boldsymbol{Z}_j}}{\sigma_\varepsilon^2} + \frac{b_{uj}\mathbf{1}}{\sigma_{uj}^2}\right)^{-1}\Bigg). \end{aligned}$$

If we define

$$\boldsymbol{\omega}_6 \equiv \boldsymbol{Z}_j^T\boldsymbol{y} - \boldsymbol{Z}_j^T\boldsymbol{X}(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{j' \neq j}^{d_\bullet} \boldsymbol{Z}_j^T \boldsymbol{Z}_{j'}(\gamma_{uj'}\widetilde{\boldsymbol{u}}_{j'})$$

and

$$\boldsymbol{\omega}_7 \equiv \frac{\gamma_{uj}\boldsymbol{\omega}_{\boldsymbol{Z}_j}}{\sigma_\varepsilon^2} + \frac{b_{uj}\mathbf{1}}{\sigma_{uj}^2},$$

and generate the $(K_j \times 1)$ Multivariate Normal random vector $\boldsymbol{z} \sim N(\mathbf{0}, \boldsymbol{I}_{K_j})$ then

$$\frac{\boldsymbol{z}}{\sqrt{\boldsymbol{\omega}_7}} + \frac{\gamma_{uj}\boldsymbol{\omega}_6}{\boldsymbol{\omega}_7\sigma_\varepsilon^2} \tag{2.66}$$

has the same distribution as $\widetilde{\boldsymbol{u}}_j \big|$rest, given in Algorithm 2 of Section 2.5.3, $1 \leq j \leq d_\bullet$.

2.B.4 Full Conditional Distribution of $\gamma_{\beta j}$

First note that, for $1 \leq j \leq d_\circ + d_\bullet$,

$$\mathfrak{p}(\gamma_{\beta j} \big| \text{rest}) \propto \mathfrak{p}(\boldsymbol{y} \big| \beta_0, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u, \sigma_\varepsilon^2)\mathfrak{p}(\gamma_{\beta j}).$$

Recalling that

$$\boldsymbol{y}\big|\beta_0, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u, \sigma_\varepsilon^2 \sim N\left(\mathbf{1}_n\beta_0 + \boldsymbol{X}(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) + \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j), \sigma_\varepsilon^2 \boldsymbol{I}_n\right)$$

and

$$\gamma_{\beta j} \overset{\text{ind.}}{\sim} \text{Bernoulli}(\rho_\beta), \;\; 1 \le j \le d_\circ + d_\bullet,$$

if we define $\boldsymbol{r}_{\gamma_{\beta j}} \equiv \boldsymbol{y} - \mathbf{1}_n\beta_0 - \boldsymbol{X}_{-j}\left\{(\boldsymbol{\gamma}_\beta)_{-j} \odot (\widetilde{\boldsymbol{\beta}})_{-j}\right\} - \sum_{j'=1}^{d_\bullet} \boldsymbol{Z}_{j'}(\gamma_{uj'}\widetilde{\boldsymbol{u}}_{j'})$, as a function of $\gamma_{\beta j}$, and noting that $\gamma_{\beta j} \in \{0, 1\}$ implies $\gamma_{\beta j}^2 = \gamma_{\beta j}$, we have

$$\begin{aligned}
\mathfrak{p}(\boldsymbol{y}\big|\beta_0, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u, \sigma_\varepsilon^2) &\propto \exp(\gamma_{\beta j}\eta_{\beta j1}) \\
&\propto \exp\Bigg\{-\frac{1}{2\sigma_\varepsilon^2}\Bigg\|\boldsymbol{y} - \mathbf{1}_n\beta_0 - \boldsymbol{X}_j\gamma_{\beta j}\widetilde{\beta}_j - \boldsymbol{X}_{-j}\left\{(\boldsymbol{\gamma}_\beta)_{-j} \odot (\widetilde{\boldsymbol{\beta}})_{-j}\right\} \\
&\qquad -\sum_{j'=1}^{d_\bullet} \boldsymbol{Z}_{j'}(\gamma_{uj'}\widetilde{\boldsymbol{u}}_{j'})\Bigg\|^2\Bigg\} \\
&\propto \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\left\|\gamma_{\beta j}\widetilde{\beta}_j\boldsymbol{X}_j - \boldsymbol{r}_{\gamma_{\beta j}}\right\|^2\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\gamma_{\beta j}\left(\widetilde{\beta}_j^2\left\|\boldsymbol{X}_j\right\|^2 - 2\widetilde{\beta}_j\boldsymbol{X}_j^T\boldsymbol{r}_{\gamma_{\beta j}}\right)\right\}
\end{aligned}$$

where

$$\begin{aligned}
\eta_{\beta j1} &= -\frac{1}{2\sigma_\varepsilon^2} \\
&\times \left(\widetilde{\beta}_j^2\left\|\boldsymbol{X}_j\right\|^2 - 2\widetilde{\beta}_j\boldsymbol{X}_j^T\left[\boldsymbol{y} - \mathbf{1}_n\beta_0 - \boldsymbol{X}_{-j}\left\{(\boldsymbol{\gamma}_\beta)_{-j} \odot (\widetilde{\boldsymbol{\beta}})_{-j}\right\} - \sum_{j'=1}^{d_\bullet} \boldsymbol{Z}_{j'}(\gamma_{uj'}\widetilde{\boldsymbol{u}}_{j'})\right]\right).
\end{aligned}$$

Also

$$\mathfrak{p}(\gamma_{\beta j}) \propto \exp\left\{\gamma_{\beta j}\text{logit}(\rho_\beta)\right\}.$$

Therefore

$$\mathfrak{p}(\gamma_{\beta j}\big|\text{rest}) \propto \exp\Big[\gamma_{\beta j}\left\{\eta_{\beta j1} + \text{logit}(\rho_\beta)\right\}\Big]. \tag{2.67}$$

It follows that

$$\gamma_{\beta j}\big|\text{rest} \sim \text{Bernoulli}\left(\frac{1}{1+\exp(-\eta_{\beta j})}\right), \quad 1 \leq j \leq d_\circ + d_\bullet$$

where

$$\eta_{\beta j} = \text{logit}(\rho_\beta) - \frac{1}{2\sigma_\varepsilon^2}\left(\widetilde{\beta}_j^2\left\|\boldsymbol{X}_j\right\|^2 - 2\widetilde{\beta}_j\boldsymbol{X}_j^T\left[\boldsymbol{y} - \mathbf{1}_n\beta_0 - \boldsymbol{X}_{-j}\{(\boldsymbol{\gamma}_\beta)_{-j} \odot (\widetilde{\boldsymbol{\beta}})_{-j}\} - \sum_{j'=1}^{d_\bullet}\boldsymbol{Z}_{j'}(\gamma_{uj'}\widetilde{\boldsymbol{u}}_{j'})\right]\right). \tag{2.68}$$

Recalling that (2.7) to be $\boldsymbol{X}_j^T\mathbf{1}_n = 0$, then (2.68) becomes (2.41) of Section 2.5.3 to be

$$\eta_{\beta j} = \text{logit}(\rho_\beta) - \frac{1}{2\sigma_\varepsilon^2}\left(\widetilde{\beta}_j^2\left\|\boldsymbol{X}_j\right\|^2 - 2\widetilde{\beta}_j\boldsymbol{X}_j^T\left[\boldsymbol{y} - \boldsymbol{X}_{-j}\{(\boldsymbol{\gamma}_\beta)_{-j} \odot (\widetilde{\boldsymbol{\beta}})_{-j}\} - \sum_{j'=1}^{d_\bullet}\boldsymbol{Z}_{j'}(\gamma_{uj'}\widetilde{\boldsymbol{u}}_{j'})\right]\right).$$

If we define

$$\omega_4 \equiv \boldsymbol{X}_j^T\left[\boldsymbol{y} - \boldsymbol{X}_{-j}\{(\boldsymbol{\gamma}_\beta)_{-j} \odot (\widetilde{\boldsymbol{\beta}})_{-j}\} - \sum_{j'=1}^{d_\bullet}\boldsymbol{Z}_{j'}(\gamma_{uj'}\widetilde{\boldsymbol{u}}_{j'})\right] \quad \text{and} \quad \omega_5 \equiv \eta_{\beta_j},$$

then

$$\omega_5 = \text{logit}\,(\rho_\beta) - \frac{1}{2\sigma_\varepsilon^2}\left(\widetilde{\beta}_j^2\left\|\boldsymbol{X}_j\right\|^2 - 2\widetilde{\beta}_j\omega_4\right)$$

is given in Algorithm 2 of Section 2.5.3.

2.B.5 Full Conditional Distribution of γ_{uj}

First note that, for $1 \leq j \leq d_\bullet$,

$$\mathfrak{p}(\gamma_{uj}\big|\text{rest}) \propto \mathfrak{p}(\boldsymbol{y}\big|\beta_0, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u, \sigma_\varepsilon^2)\mathfrak{p}(\gamma_{uj}).$$

Recalling that

$$\boldsymbol{y}\big|\beta_0, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u, \sigma_\varepsilon^2 \sim N\left(\mathbf{1}_n\beta_0 + \boldsymbol{X}(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) + \sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j), \sigma_\varepsilon^2\boldsymbol{I}_n\right)$$

and

$$\gamma_{uj} \overset{\text{ind.}}{\sim} \text{Bernoulli}(\rho_u), \;\; 1 \leq j \leq d_\bullet,$$

as a function of γ_{uj}, and noting that $\gamma_{uj} \in \{0,1\}$ implies $\gamma_{uj}^2 = \gamma_{uj}$, and note that $\boldsymbol{Z}_j^T \boldsymbol{Z}_j$ is a diagonal matrix from (2.8), if we define $\boldsymbol{Z}_j^T \boldsymbol{Z}_j \equiv \text{diag}\left(\boldsymbol{\omega}_{\boldsymbol{Z}_j}\right)$, we have

$$\begin{aligned}
&\mathfrak{p}(\boldsymbol{y}\big|\beta_0, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u, \sigma_\varepsilon^2) \propto \exp\left(\gamma_{uj}\eta_{uj1}\right) \\
&\quad \propto \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\left\|\boldsymbol{y} - \boldsymbol{1}_n\beta_0 - \boldsymbol{X}(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{j'\neq j}^{d_\bullet} \boldsymbol{Z}_{j'}(\gamma_{uj'}\widetilde{\boldsymbol{u}}_{j'}) - \boldsymbol{Z}_j\left(\gamma_{uj}\widetilde{\boldsymbol{u}}_j\right)\right\|^2\right\} \\
&\quad \propto \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\left\|\gamma_{uj}\boldsymbol{Z}_j\widetilde{\boldsymbol{u}}_j - \boldsymbol{r}_{\gamma_{uj}}\right\|^2\right\} \\
&\quad \propto \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\gamma_{uj}\left(\widetilde{\boldsymbol{u}}_j^T\boldsymbol{Z}_j^T\boldsymbol{Z}_j\widetilde{\boldsymbol{u}}_j - 2\widetilde{\boldsymbol{u}}_j^T\boldsymbol{Z}_j^T\boldsymbol{r}_{\gamma_{uj}}\right)\right\} \\
&\quad \propto \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\gamma_{uj}\left(\boldsymbol{\omega}_{\boldsymbol{Z}_j}^T(\widetilde{\boldsymbol{u}}_j \odot \widetilde{\boldsymbol{u}}_j) - 2\widetilde{\boldsymbol{u}}_j^T\boldsymbol{Z}_j^T\boldsymbol{r}_{\gamma_{uj}}\right)\right\},
\end{aligned}$$

if we define

$$\boldsymbol{r}_{\gamma_{uj}} \equiv \boldsymbol{y} - \boldsymbol{1}_n\beta_0 - \boldsymbol{X}(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{j'\neq j}^{d_\bullet} \boldsymbol{Z}_{j'}(\gamma_{uj'}\widetilde{\boldsymbol{u}}_{j'})$$

and define

$$\eta_{uj1} \equiv -\frac{1}{2\sigma_\varepsilon^2}\left(\boldsymbol{\omega}_{\boldsymbol{Z}_j}^T(\widetilde{\boldsymbol{u}}_j \odot \widetilde{\boldsymbol{u}}_j) - 2\widetilde{\boldsymbol{u}}_j^T\boldsymbol{Z}_j^T\left[\boldsymbol{y} - \boldsymbol{1}_n\beta_0 - \boldsymbol{X}(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{j'\neq j}^{d_\bullet} \boldsymbol{Z}_{j'}(\gamma_{uj'}\widetilde{\boldsymbol{u}}_{j'})\right]\right).$$

Also

$$\mathfrak{p}(\gamma_{uj}) \propto \exp\left\{\gamma_{uj}\text{logit}(\rho_u)\right\}.$$

Therefore

$$\mathfrak{p}(\gamma_{uj}\big|\text{rest}) \propto \exp\left[\gamma_{uj}\left\{\eta_{uj1} + \text{logit}(\rho_u)\right\}\right].$$

It follows that

$$\gamma_{uj}\big|\text{rest} \sim \text{Bernoulli}\left(\frac{1}{1+\exp(-\eta_{uj})}\right)$$

where

$$\eta_{uj} = \text{logit}(\rho_u) - \frac{1}{2\sigma_\varepsilon^2}\Bigg(\boldsymbol{\omega}_{\boldsymbol{Z}_j}^T\left(\widetilde{\boldsymbol{u}}_j \odot \widetilde{\boldsymbol{u}}_j\right) - 2\widetilde{\boldsymbol{u}}_j^T\boldsymbol{Z}_j^T\Big[\boldsymbol{y} - \mathbf{1}_n\beta_0 \\ -\boldsymbol{X}\left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}\right) - \sum_{j'\neq j}^{d_\bullet}\boldsymbol{Z}_{j'}(\gamma_{uj'}\widetilde{\boldsymbol{u}}_{j'})\Big]\Bigg) \tag{2.69}$$

Recalling that (2.8) of Section 2.3.3 to be $\boldsymbol{Z}_j^T\mathbf{1}_n = \mathbf{0}_{K_j}$, then (2.69) becomes (2.46) of Section 2.5.3 to be

$$\eta_{uj} = \text{logit}(\rho_u) - \frac{1}{2\sigma_\varepsilon^2}\Big(\boldsymbol{\omega}_{\boldsymbol{Z}_j}^T\left(\widetilde{\boldsymbol{u}}_j \odot \widetilde{\boldsymbol{u}}_j\right) - 2\widetilde{\boldsymbol{u}}_j^T\boldsymbol{Z}_j^T\Big[\boldsymbol{y} - \boldsymbol{X}\left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}\right) - \sum_{j'\neq j}^{d_\bullet}\boldsymbol{Z}_{j'}(\gamma_{uj'}\widetilde{\boldsymbol{u}}_{j'})\Big]\Big).$$

If we define

$$\boldsymbol{\omega}_8 \equiv \boldsymbol{Z}_j^T\Big[\boldsymbol{y} - \boldsymbol{X}\left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}\right) - \sum_{j'\neq j}^{d_\bullet}\boldsymbol{Z}_{j'}(\gamma_{uj'}\widetilde{\boldsymbol{u}}_{j'})\Big] \text{ and define } \ \omega_9 \equiv \eta_{uj},$$

then

$$\omega_9 = \eta_{uj} = \text{logit}\left(\rho_{uj}\right) - \frac{1}{2\sigma_\varepsilon^2}\left(\boldsymbol{\omega}_{\boldsymbol{Z}_j}^T\left(\widetilde{\boldsymbol{u}}_j \odot \widetilde{\boldsymbol{u}}_j\right) - 2\widetilde{\boldsymbol{u}}_j^T\boldsymbol{\omega}_8\right)$$

is given in Algorithm 2 of Section 2.5.3.

2.B.6 Full Conditional Distribution of σ_ε^2

First note that

$$\mathfrak{p}(\sigma_\varepsilon^2|\text{rest}) \propto \mathfrak{p}(\boldsymbol{y}|\beta_0, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u, \sigma_\varepsilon^2)\mathfrak{p}(\sigma_\varepsilon^2|a_\varepsilon).$$

Recalling that

$$\boldsymbol{y}|\beta_0, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u, \sigma_\varepsilon^2 \sim N\left(\mathbf{1}_n\beta_0 + \boldsymbol{X}(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) + \sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j), \sigma_\varepsilon^2\boldsymbol{I}_n\right)$$

and

$$\sigma_\varepsilon^2|a_\varepsilon \sim \text{Inverse-Gamma}(\frac{1}{2}, 1/a_\varepsilon),$$

we have

$$\begin{aligned}&\mathfrak{p}(\boldsymbol{y}\big|\beta_0,\widetilde{\boldsymbol{\beta}},\boldsymbol{\gamma}_\beta,\widetilde{\boldsymbol{u}},\boldsymbol{\gamma}_u,\sigma_\varepsilon^2)\\&\quad\propto\left(\frac{1}{\sqrt{\sigma_\varepsilon^2}}\right)^n\exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\left\|\boldsymbol{y}-\mathbf{1}_n\beta_0-\boldsymbol{X}(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}})-\sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j)\right\|^2\right\}\end{aligned}$$

and

$$\mathfrak{p}(\sigma_\varepsilon^2\big|a_\varepsilon)\propto\left(\sigma_\varepsilon^2\right)^{-\frac{1}{2}-1}\exp\left(-\frac{1}{a_\varepsilon\sigma_\varepsilon^2}\right).$$

Therefore

$$\begin{aligned}&\mathfrak{p}(\sigma_\varepsilon^2\big|\text{rest})\propto\left(\sigma_\varepsilon^2\right)^{-\frac{n+1}{2}-1}\\&\quad\exp\left\{-\frac{1}{\sigma_\varepsilon^2}\left(\frac{1}{a_\varepsilon}+\frac{1}{2}\left\|\boldsymbol{y}-\mathbf{1}_n\beta_0-\boldsymbol{X}(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}})-\sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j)\right\|^2\right)\right\}.\end{aligned}$$

In summary, if the response type is Gaussian, we obtain (2.47) in Section 2.5.3 to be

$$\sigma_\varepsilon^2\big|\text{rest}\sim\text{Inverse-Gamma}\left(\frac{n+1}{2},\frac{1}{a_\varepsilon}+\frac{1}{2}\left\|\boldsymbol{y}-\mathbf{1}_n\beta_0-\boldsymbol{X}(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}})-\sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j)\right\|^2\right).$$

If we define

$$\boldsymbol{\omega}_{10}\equiv\mathbf{1}_n\beta_0+\boldsymbol{X}\left(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}}\right)+\sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j\left(\gamma_{uj}\widetilde{\boldsymbol{u}}_j\right),$$

then

$$\sigma_\varepsilon^2\big|\text{rest}\sim\text{Inverse-Gamma}\left(\frac{1}{2}\left(n+1\right),\left(1/a_\varepsilon\right)+\frac{1}{2}\left\|\boldsymbol{y}-\boldsymbol{\omega}_{10}\right\|^2\right)$$

is given in Algorithm 2 of Section 2.5.3.

If the response data is binary, note that $\mu_i=\left(\boldsymbol{\omega}_{10}\right)_i$, following (2.29) and (2.30) in Section 2.4.3, we have

$$y_i=1\text{ if and only if }c_i\geq0;$$

$$y_i\big|c_i\overset{\text{ind.}}{\sim}\text{Bernoulli}(I(c_i\geq0));$$

$$c_i\big|\beta_0,\widetilde{\boldsymbol{\beta}},\boldsymbol{\gamma}_\beta,\widetilde{\boldsymbol{u}},\boldsymbol{\gamma}_u\sim N\left(\mu_i,1\right),1\leq i\leq n.$$

Then

$$\mathfrak{p}(c_i|\text{rest}) \propto \mathfrak{p}(y_i|c_i)\mathfrak{p}(c_i|\beta_0, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u) \propto \begin{cases} \phi(c_i - \mu_i)I(c_i \geq 0); \\ \phi(c_i - \mu_i)I(c_i < 0). \end{cases}$$

Therefore

$$c_i|\text{rest} \sim \begin{cases} \text{Truncated-Normal}_+(\mu_i, 1), & y_i = 1; \\ \\ \text{Truncated-Normal}_-(\mu_i, 1), & y_i = 0. \end{cases} \tag{2.70}$$

Recalling that **Result 1.7.3**, we have

$$c_i|\text{rest} \sim \text{Truncated-Normal}_-(\mu_i, 1) \text{ equivalent to}$$
$$-c_i|\text{rest} \sim \text{Truncated-Normal}_+(-\mu_i, 1).$$

Therefore

$$c_i|\text{rest} \sim \begin{cases} \text{Truncated-Normal}_+(\mu_i, 1), & y_i = 1; \\ \\ -\text{Truncated-Normal}_+(-\mu_i, 1), & y_i = 0. \end{cases}$$

In summary, for $1 \leq i \leq n$,

$$c_i|\text{rest} \sim (2y_i - 1)\text{Truncated-Normal}_+((2y_i - 1)\mu_i, 1).$$

If define

$$\omega_{11} \sim \text{Truncated-Normal}_+ \left((2y_i - 1)\,\mu_i, 1\right),$$

then

$$c_i = (2y_i - 1)\,\omega_{11}$$

is given in Algorithm 2 of Section 2.5.3 for the binary responses.

2.B.7 Full Conditional Distribution of a_ε

First note that

$$\mathfrak{p}(a_\varepsilon|\text{rest}) \propto \mathfrak{p}(\sigma_\varepsilon^2|a_\varepsilon)\mathfrak{p}(a_\varepsilon).$$

Recalling that

$$\sigma_\varepsilon^2 \big| a_\varepsilon \sim \text{Inverse-Gamma}\left(\frac{1}{2}, \frac{1}{a_\varepsilon}\right) \text{ and } \ a_\varepsilon \sim \text{Inverse-Gamma}\left(\frac{1}{2}, \frac{1}{s_\varepsilon^2}\right),$$

we have

$$\mathfrak{p}(\sigma_\varepsilon^2 \big| a_\varepsilon) \propto \left(\frac{1}{a_\varepsilon}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{a_\varepsilon \sigma_\varepsilon^2}\right) \text{ and } \ \mathfrak{p}(a_\varepsilon) \propto (a_\varepsilon)^{-\frac{1}{2}-1} \exp\left(-\frac{1}{s_\varepsilon^2 a_\varepsilon}\right).$$

Then

$$\mathfrak{p}(a_\varepsilon \big| \text{rest}) \propto (a_\varepsilon)^{-1-1} \exp\left(-\frac{1}{a_\varepsilon}\left(\frac{1}{\sigma_\varepsilon^2} + \frac{1}{s_\varepsilon^2}\right)\right).$$

Therefore, we obtain (2.48) in Section 2.5.3 to be

$$a_\varepsilon \big| \text{rest} \sim \text{Inverse-Gamma}\left(1, \frac{1}{\sigma_\varepsilon^2} + \frac{1}{s_\varepsilon^2}\right).$$

2.B.8 Full Conditional Distribution of $b_{\beta j}$

First note that, for $1 \le j \le d_\circ + d_\bullet$,

$$\mathfrak{p}(b_{\beta j} \big| \text{rest}) \propto \mathfrak{p}(\widetilde{\beta}_j \big| b_{\beta j}, \sigma_\beta^2) \mathfrak{p}(b_{\beta j}).$$

Recalling that

$$\widetilde{\beta}_j \big| \sigma_\beta^2, b_{\beta j} \overset{\text{ind.}}{\sim} N(0, \sigma_\beta^2 / b_{\beta j}) \text{ and } \ b_{\beta j} \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(1, \tfrac{1}{2}), \ \ 1 \le j \le d_\circ + d_\bullet,$$

for $b_{\beta j} > 0$, we have

$$\begin{aligned}
\mathfrak{p}(b_{\beta j} \big| \text{rest}) &\propto b_{\beta j}^{1/2} \exp\left[-\frac{1}{2}\widetilde{\beta}_j^2 / (\sigma_\beta^2 / b_{\beta j})\right] b_{\beta j}^{-1-1} \exp\left\{-1/(2b_{\beta j})\right\} \\
&\propto (b_{\beta j})^{-\frac{3}{2}} \exp\left(-\frac{\widetilde{\beta}_j^2 b_{\beta j}}{2\sigma_\beta^2} - \frac{1}{2b_{\beta j}}\right) \propto b_{\beta j}^{-\frac{3}{2}} \exp\left\{-\frac{\left(b_{\beta j} - (\sigma_\beta / \widetilde{\beta}_j)\right)^2}{2(\sigma_\beta / \widetilde{\beta}_j)^2 b_{\beta j}}\right\}. \quad (2.71)
\end{aligned}$$

Therefore, we obtain (2.38) in Section 2.5.3 to be

$$b_{\beta j} \big| \text{rest} \sim \text{Inverse-Gaussian}\left(\frac{\sigma_\beta}{|\widetilde{\beta}_j|}, 1\right), \quad 1 \le j \le d_\circ + d_\bullet.$$

2.B.9 Full Conditional Distribution of b_{uj}

First note that, for $1 \leq j \leq d_\bullet$,

$$\mathfrak{p}(b_{uj}\big|\text{rest}) \propto \mathfrak{p}(\widetilde{\boldsymbol{u}}_j\big|b_{uj}, \sigma^2_{uj})\mathfrak{p}(b_{uj}).$$

For $1 \leq j \leq d_\bullet$, recalling that

$$\widetilde{\boldsymbol{u}}_j\big|\sigma^2_{uj}, b_{uj} \overset{\text{ind.}}{\sim} N(\mathbf{0}, (\sigma^2_{uj}/b_{uj})\boldsymbol{I}_{K_j}) \text{ and } \quad b_{uj} \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2}(K_j+1), \tfrac{1}{2}),$$

for $b_{uj} > 0$, we have

$$\begin{aligned}\mathfrak{p}(b_{uj}\big|\text{rest}) &\propto \left(\frac{1}{\sqrt{(1/b_{uj})}}\right)^{K_j} \exp\left\{-\frac{\|\widetilde{\boldsymbol{u}}_j\|^2}{2(\sigma^2_{uj}/b_{uj})\boldsymbol{I}}\right\} (b_{uj})^{-\frac{K_j+1}{2}-1} \exp\left(-\frac{1}{2b_{uj}}\right) \\ &\propto (b_{uj})^{-\frac{3}{2}} \exp\left(-\frac{b_{uj}\|\widetilde{\boldsymbol{u}}_j\|^2}{2\sigma^2_{uj}} - \frac{1}{2b_{uj}}\right) \propto (b_{uj})^{-\frac{3}{2}} \exp\left(-\frac{\left(b_{uj} - \frac{\sigma_{uj}}{\|\widetilde{\boldsymbol{u}}_j\|}\right)^2}{2\left(\frac{\sigma_{uj}}{\|\widetilde{\boldsymbol{u}}_j\|}\right)^2 b_{uj}}\right). \quad (2.72)\end{aligned}$$

Therefore, we obtain (2.43) in Section 2.5.3 to be

$$b_{uj}\big|\text{rest} \sim \text{Inverse-Gaussian}\left(\frac{\sigma_{uj}}{\|\widetilde{\boldsymbol{u}}_j\|}, 1\right).$$

2.B.10 Full Conditional Distribution of σ^2_β

First note that

$$\mathfrak{p}(\sigma^2_\beta\big|\text{rest}) \propto \mathfrak{p}(\widetilde{\boldsymbol{\beta}}\big|\sigma^2_\beta, \boldsymbol{b}_\beta)\mathfrak{p}(\sigma^2_\beta\big|a_\beta).$$

Recalling that

$$\widetilde{\beta}_j\big|\sigma^2_\beta, b_{\beta j} \sim N\left(0, \sigma^2_\beta/b_{\beta j}\right), \;\; 1 \leq j \leq d_\circ + d_\bullet \;\; \text{and} \;\; \sigma^2_\beta\big|a_\beta \sim \text{Inverse-Gamma}\left(\frac{1}{2}, \frac{1}{a_\beta}\right),$$

we have

$$\mathfrak{p}(\widetilde{\boldsymbol{\beta}}\big|\sigma^2_\beta, \boldsymbol{b}_\beta) \propto \left(\sigma^2_\beta\right)^{-\frac{d_\circ+d_\bullet}{2}} \exp\left(-\frac{\widetilde{\boldsymbol{\beta}}^T \text{diag}(\boldsymbol{b}_\beta)\widetilde{\boldsymbol{\beta}}}{2\sigma^2_\beta}\right), \;\; \mathfrak{p}(\sigma^2_\beta\big|a_\beta) \propto \left(\sigma^2_\beta\right)^{-\frac{1}{2}-1} \exp\left(-\frac{1}{a_\beta\sigma^2_\beta}\right).$$

Then

$$\mathfrak{p}(\sigma_\beta^2|\text{rest}) \propto \left(\sigma_\beta^2\right)^{-\frac{d_\circ+d_\bullet+1}{2}-1} \exp\left(-\frac{1}{\sigma_\beta^2}\left(\frac{\widetilde{\boldsymbol{\beta}}^T \text{diag}(\boldsymbol{b}_\beta)\widetilde{\boldsymbol{\beta}}}{2}+\frac{1}{a_\beta}\right)\right).$$

Therefore, we obtain (2.39) in Section 2.5.3 to be

$$\sigma_\beta^2|\text{rest} \sim \text{Inverse-Gamma}\left(\frac{d_\circ+d_\bullet+1}{2}, \frac{\widetilde{\boldsymbol{\beta}}^T \text{diag}(\boldsymbol{b}_\beta)\widetilde{\boldsymbol{\beta}}}{2}+\frac{1}{a_\beta}\right).$$

2.B.11 Full Conditional Distribution of a_β

First note that

$$\mathfrak{p}(a_\beta|\text{rest}) \propto \mathfrak{p}(\sigma_\beta^2|a_\beta)\mathfrak{p}(a_\beta).$$

Recalling that

$$\sigma_\beta^2|a_\beta \sim \text{Inverse-Gamma}\left(\frac{1}{2}, \frac{1}{a_\beta}\right) \text{ and } \ a_\beta \sim \text{Inverse-Gamma}\left(\frac{1}{2}, \frac{1}{s_\beta^2}\right),$$

following the steps in Appendix 2.B.7, we have

$$\mathfrak{p}(a_\beta|\text{rest}) \propto (a_\beta)^{-1-1} \exp\left(-\frac{1}{a_\beta}\left(\frac{1}{\sigma_\beta^2}+\frac{1}{s_\beta^2}\right)\right).$$

Therefore, we obtain (2.40) in Section 2.5.3 to be

$$a_\beta|\text{rest} \sim \text{Inverse-Gamma}\left(1, \frac{1}{\sigma_\beta^2}+\frac{1}{s_\beta^2}\right).$$

2.B.12 Full Conditional Distribution of σ_{uj}^2

First note that, for $1 \le j \le d_\bullet$,

$$\mathfrak{p}(\sigma_{uj}^2|\text{rest}) \propto \mathfrak{p}(\widetilde{\boldsymbol{u}}_j|\sigma_{uj}^2, b_{uj})\mathfrak{p}(\sigma_{uj}^2|a_{uj}).$$

Recalling that

$$\widetilde{\boldsymbol{u}}_j|\sigma_{uj}^2, b_{uj} \sim N\left(\boldsymbol{0}, \left(\sigma_{uj}^2/b_{uj}\right)\boldsymbol{I}\right) \text{ and } \ \sigma_{uj}^2|a_{uj} \sim \text{Inverse-Gamma}\left(\frac{1}{2}, \frac{1}{a_{uj}}\right), 1 \le j \le d_\bullet,$$

we have

$$\mathfrak{p}(\widetilde{\boldsymbol{u}}_j \big| \sigma_{uj}^2, b_{uj}) \propto \left(\frac{1}{\sqrt{\sigma_{uj}^2}} \right)^{K_j} \exp\left(-\frac{\left\| \widetilde{\boldsymbol{u}}_j \right\|^2}{2\left(\sigma_{uj}^2 / b_{uj}\right) \boldsymbol{I}} \right) \propto \left(\sigma_{uj}^2\right)^{-\frac{K_j}{2}} \exp\left(-\frac{b_{uj} \left\| \widetilde{\boldsymbol{u}}_j \right\|^2}{2\sigma_{uj}^2} \right)$$

and

$$\mathfrak{p}(\sigma_{uj}^2 \big| a_{uj}) \propto \left(\sigma_{uj}^2\right)^{-\frac{1}{2}-1} \exp\left(-\frac{1}{a_{uj}\sigma_{uj}^2} \right).$$

Then

$$\mathfrak{p}(\sigma_{uj}^2 \big| \text{rest}) \propto \left(\sigma_{uj}^2\right)^{-\frac{K_j+1}{2}-1} \exp\left(-\frac{1}{\sigma_{uj}^2} \left(\frac{b_{uj} \left\| \widetilde{\boldsymbol{u}}_j \right\|^2}{2} + \frac{1}{a_{uj}} \right) \right).$$

Therefore, we obtain (2.44) in Section 2.5.3 to be

$$\sigma_{uj}^2 | \text{rest} \sim \text{Inverse-Gamma}\left(\frac{K_j + 1}{2}, \frac{b_{uj} \left\| \widetilde{\boldsymbol{u}}_j \right\|^2}{2} + \frac{1}{a_{uj}} \right), 1 \leq j \leq d_\bullet.$$

2.B.13 Full Conditional Distribution of a_{uj}

First note that, for $1 \leq j \leq d_\bullet$,

$$\mathfrak{p}(a_{uj} \big| \text{rest}) \propto \mathfrak{p}(\sigma_{uj}^2 \big| a_{uj}) \mathfrak{p}(a_{uj}).$$

For $1 \leq j \leq d_\bullet$, recalling that

$$\sigma_{uj}^2 \big| a_{uj} \sim \text{Inverse-Gamma}\left(\frac{1}{2}, \frac{1}{a_{uj}} \right) \text{ and } \; a_{uj} \sim \text{Inverse-Gamma}\left(\frac{1}{2}, \frac{1}{s_u^2} \right),$$

following the steps in Appendix 2.B.7, we have

$$\mathfrak{p}(a_{uj} \big| \text{rest}) \propto \left(a_{uj}\right)^{-1-1} \exp\left(-\frac{1}{a_{uj}} \left(\frac{1}{\sigma_{uj}^2} + \frac{1}{s_u^2} \right) \right).$$

Therefore, we obtain (2.45) in Section 2.5.3 to be

$$a_{uj} \big| \text{rest} \sim \text{Inverse-Gamma}\left(1, \frac{1}{\sigma_{uj}^2} + \frac{1}{s_u^2} \right), 1 \leq j \leq d_\bullet.$$

Chapter 3

Generalized Additive Model Selection via Markov Chain Monte Carlo with Horseshoe-type Priors

Declaration The method and results presented in Section 3.2 and the parts of Section 3.3 – 3.8 for Gaussian responses, are presented in the paper *The Grouped Horseshoe Distribution and Its Statistical Properties* (He and Wand 2025). I certify that this publication was a direct result of my research towards this PhD, under the supervision of my principle supervisor Prof Matt Wand, and the reproduction in this thesis does not breach any copyright regulations.

3.1 Introduction

For parsimonious predictor selection, it is desirable to impose sparsity on coefficient vectors. This can be carried out with LASSO priors for Gaussian responses and binary responses, e.g. Laplace-Zero priors, described in Chapter 2. The Laplace-Zero distribution used in Chapter 2 is a "spike-and-slab" mixture of a point mass at zero (the spike) and an absolutely continuous Laplace density function (the slab). LASSO performs explicit selection by making some of estimates 0 and producing a sparse solution.

Alternatively, selection can be carried out by Horseshoe-type shrinkage priors (Carvalho et al., 2010). The name "horseshoe" arises from the shape of the Beta prior density of the shrinkage weight. The shrinkage weight has the same function as the posterior mean obtained under LASSO. Both can be treated as the posterior inclusion probability for recovering a sparse signal. Shrinkage methods such as Horseshoe-type prior shrink towards 0 by thresholding the shrinkage weights that behave like posterior inclusion probabilities to achieve variable selection (Bhadra et al., 2019).

The horseshoe prior is frequently used in Bayesian analysis for high-dimensional models, and it is designed for the sparse setting. The Horseshoe prior has tail robustness and super-efficient convergence properties (Bhadra et al., 2019). There are two types of Horseshoe-type priors: the Horseshoe distribution prior with 2 levels of Gamma variable conditioning and the Horseshoe$_+$ distribution prior with 4 levels of Gamma variable conditioning (Bhadra et al., 2019). In this chapter, we use the Horseshoe distribution prior with 2 levels of Gamma variable conditioning.

We have found that several properties of the univariate Horseshoe distribution are present in the grouped case. For instance, the robustness of large signals of horseshoe priors is a property that is known in the univariate case but we demonstrate that a similar notion is present for the grouped case. However, the super-efficiency property (based on risk-rates convergence), does not extend to the grouped situation.

The attraction of Horseshoe distribution prior is that each of the conditional distributions in its equivalences belongs to Normal and Gamma families, described in Section 3.4.4 for linear term coefficients and described in Section 3.4.5 for spline coefficients, with extending the Horseshoe prior to the case of grouped variables. This translates to the full conditional distributions being standard distribution and Gibbs sampling being exact for the Markov chain Monte Carlo for Gaussian responses.

For Poisson responses based on the model described in Section 3.5.3, Negative-

Binomial responses based on the model described in Section 3.5.4 and for Gaussian responses based on the model described in Section 3.5.2, we work out algorithms given in Algorithm 4, Algorithm 5 and Algorithm 6 respectively, for the Markov chain Monte Carlo approach, and this is described in Section 3.6.3. It is noted that, the auxiliary parameters $\boldsymbol{\gamma}_\beta, \boldsymbol{\gamma}_u$ and parameters $\boldsymbol{\rho}_\beta, \boldsymbol{\rho}_u$ described in Section 2.3.4 and Section 2.3.5 respectively, no longer exist in the full models using Horseshoe/Grouped Horseshoe priors, compared to the full models using Laplace-Zero/Grouped Lasso-Zero priors in Chapter 2.

In the fitting algorithm, we use slice sampling for Poisson responses and Negative-Binomial responses, described in Section 3.6.2. We are able to obtain the logarithm of the unnormalised density function of some parameters. Slice sampling is an efficient way of drawing samples based on the parameter's unnormalised density function.

In Section 3.8, for Gaussian responses, we compare the classification performance with the use of Horseshoe/Grouped Horseshoe priors described in Section 3.5.2, to the use of Laplace Zero/Grouped Lasso-Zero priors described in Chapter 2.

The classification performance comparison for Poisson responses and Negative-Binomial responses are given in Section 3.9 and Section 3.10 respectively.

Throughout Chapter 3, "rest" denotes the set of other random variables in model (3.23) for Gaussian responses, in model (3.24) for Poisson responses and in model (3.25) for Negative-Binomial responses.

In Section 3.2, $\boldsymbol{x}$ is referred to a generic random vector unrelated to the data. We aim to investigate the properties of Grouped Horseshoe distribution. This part of the thesis is not dealing with generalized additive model selection. In Section 3.3–3.10, $\boldsymbol{y}$ vector, $\boldsymbol{X}$ matrix and $\boldsymbol{Z}$ matrix are all fixed. $\boldsymbol{X}$ is referred to the data, for model estimation and to explore the application of the (Grouped) Horseshoe priors via MCMC for the Bayesian three-category GAM selection.

The additive structure of generalized additive model (GAM) described in this chapter is given in Section 3.4. The independence assumptions of GAM is given in (3.23) for Gaussian responses, (3.24) for Poisson responses and (3.25) for Negative-Binomial responses. The additive part uses splines for non-linear effects. The predictors include binary predictors and continuous predictors. The number of binary predictors is $d_\circ$. The number of continuous predictors is $d_\bullet$.

For generalized additive model described in Section 3.3–3.7, the sample size is n. The response $\boldsymbol{y}$ is a $n \times 1$ vector. The linear design matrix $\boldsymbol{X}$ is a $n \times (d_\circ + d_\bullet)$ matrix. The $\boldsymbol{Z}_j$ matrix containing the canonical Demmler-Reinsch basis is a $n \times K_j$

matrix, with $1 \leq j \leq d_\bullet$, K_j as the number of basis functions for the j-th continuous predictor with the default value of 12.

In Section 3.8.2, we conducted a simulation study for classification performance comparison in terms of accuracy for Gaussian responses, between the use of (Grouped) Horseshoe priors described in this chapter and the use of Laplace-Zero/Grouped Lasso-Zero priors described in Chapter 2. The data were generated with 30 continuous predictors, 10 having "true" linear effects, 10 having "true" non-linear effects and 10 having "true" zero effects. The sample size varied as $n \in \{500, 1000, 2000\}$. The error standard deviation varied as $\sigma_\varepsilon \in \{0.25, 0.5, 1, 2\}$. For each combination of sample size and error standard deviation, we ran simulation on 100 independently generated data sets.

In Section 3.9, we conducted a simulation study for classification performance comparison in terms of accuracy for Poisson responses, among the use of (Grouped) Horseshoe priors with fixed border of 0.5, the use of (Grouped) Horseshoe priors with k-means clustering to determine the classification border and the default version of the spikeSlabGAM described in Section 3.9. The data were generated with 15 continuous predictors, 5 having "true" linear effects, 5 having "true" non-linear effects and 5 having "true" zero effects. The sample size varied as $n \in \{1000, 2000\}$. For each sample size, we ran simulation on 15 independently generated data sets.

In Section 3.10, we conducted a simulation study for classification performance comparison in terms of accuracy for Negative-Binomial responses, between the use of (Grouped) Horseshoe priors with fixed border of 0.5 and the use of (Grouped) Horseshoe priors with k-means clustering to determine the classification border. The data were generated with 30 continuous predictors, 10 having "true" linear effects, 10 having "true" non-linear effects and 10 having "true" zero effects. The sample size varied as $n \in \{500, 1000, 2000\}$. The shape parameter varied as $\kappa \in \{1, 2, 4\}$. For each combination of sample size and shape parameter, we ran simulation on 50 independently generated data sets.

In Section 3.2.2, we derived the explicit expression of a standard grouped Horseshoe density function corresponding grouped Horseshoe variable selection. Based on this, in Section 3.2.3, we showed the grouped Horseshoe density function has a pole at the origin for any dimension. The (Grouped) Horseshoe prior's behaviour near the origin relates to its efficiency of handling sparsity. Therefore, the grouped Horseshoe distribution can be used as a prior density function for model selection. In Section 3.2.4, we explored the score function behaviour of (Grouped) Horseshoe dis-

tribution relates to its tail robustness to large signals. In Section 3.2.5, we derived the Bayes estimator's asymptotic rates of convergence, for both sparse and non-sparse situations. In Section 3.2.6, we derived the posterior mean in a Bayesian model, using Grouped Horseshoe distribution. Throughout these sections, $\boldsymbol{x}$ is denoted as a generic random vector unrelated to the data. These properties allow us to explore the application of the (Grouped) Horseshoe priors via MCMC for the Bayesian three-category GAM selection in Section 3.3–3.10.

3.2 Grouped Horseshoe Distribution and Its Statistical Properties

Carvalho et al. (2010) investigated the properties of the Horseshoe prior, which are well-established in Bayesian statistical inference. Now we investigate the properties of the horseshoe prior within its grouped situation. We use the Bayesian methodology with Grouped Horseshoe distribution arises from hierarchical structures for selection of groups of regression coefficients. We established five main results in Section 3.2. We find that most, but not all of the properties of the univariate Horseshoe distribution are present in the grouped case.

Based on the behaviour of Grouped Horseshoe distribution near the origin stated in **Result 3.2.3.1** in Section 3.2.3 and its robustness to large signals stated in **Result 3.2.4.1** in Section 3.2.4, the Grouped Horseshoe distribution can be used as a prior function for model selection for count responses and Gaussian responses, as described in Section 3.5.

3.2.1 Horseshoe Prior

Since around 2010, several types of continuous distributions have been proposed for use as prior distributions of coefficients in Bayesian regression-type models. There are seven types of priors listed in Table 1 of Bai and Ghosh (2018). They are well-known as global-local shrinkage priors. These priors are represented as scale mixtures of Normal density functions with different types of polynomial-tailed density functions. Among these seven priors, one of them is the Horseshoe prior, which is the focus of this chapter.

The Horseshoe prior is defined hierarchically for individual coefficients. This hierarchical structure makes the Horseshoe distribution (Carvalho et al., 2010) a mul-

tivariate scale mixture of normals, with the mixing distribution for variance components being related to $F(1,1)$ distribution, or with the mixing distribution for standard deviation components being related to Half-Cauchy distribution.

The Horseshoe prior achieves scale mixing by introducing heavy-tailed priors on global and local scale parameters. The local scale is unique to each coefficient and allows for individual coefficients to have a flexible level of shrinkage. The global scale ensures overall sparsity across coefficients. The mixing of global and local scales introduces flexibility and heavy tails, making it ideal for sparse and robust regression modeling.

There are many articles have discussed selecting a scalar coefficient via the Horseshoe prior, to decide a predictor is having zero or linear effect. Grouped variable selection in additive model (e.g. Scheipl et al., 2012; He and Wand, 2024) is an attractive mechanism for selecting a vector of coefficients corresponding to spline basis functions, to decide between the continuous predictor having a linear or non-linear effect, such as Group LASSO methodology. Our focus in Section 3.2 is the grouped extension of the horseshoe prior as proposed by Xu et al. (2016).

3.2.2 Density Function Explicit Form

Firstly, we determine the underlying multivariate density function corresponding to the grouped horseshoe variable selection. Section 2.2 of Xu et al. (2016) introduced the grouped horseshoe model. The underlying distribution is called the Grouped Horseshoe distribution. Equation (6) of Xu et al. (2016) uses the setting $\sigma = \tau = G = 1$ and $s_1 = d$. This leads to a $d \times 1$ random vector $\boldsymbol{x}$ having a (standard) Grouped Horseshoe distribution if and only if

$$\boldsymbol{x}|\lambda \sim N(\boldsymbol{0}, \lambda^2 \boldsymbol{I}_d) \quad \text{where} \quad \mathfrak{p}(\lambda) = \frac{2I(\lambda > 0)}{\pi(1+\lambda^2)}. \tag{3.1}$$

Let E_ν denotes *generalized exponential integral* function, given by

$$E_\nu \equiv \int_1^\infty \exp(-xt)/t^\nu dt, \quad x, \nu \in \mathbb{R}$$

(e.g.8.19.3 of Olver et al., 2023).

Result 3.2.2.1 *Let $\boldsymbol{x}$ be a $d \times 1$ random vector having a Grouped Horseshoe distribution as*

defined according to (3.1). Then the density function of $\boldsymbol{x}$*, denoted by* $\mathfrak{p}_{\mathrm{HS,d}}(\boldsymbol{x})$*, is*

$$\mathfrak{p}_{\mathrm{HS,d}}\left(\boldsymbol{x}\right)=\frac{\Gamma\left(\frac{1}{2}(d+1)\right)}{\sqrt{2\pi^{d+2}}}\exp\left(\left\|\boldsymbol{x}\right\|^{2}/2\right)E_{(d+1)/2}\left(\left\|\boldsymbol{x}\right\|^{2}/2\right)\Big/\left\|\boldsymbol{x}\right\|^{d-1},\ \ \boldsymbol{x}\in\mathbb{R}^{d}.$$

The derivation of **Result 3.2.2.1** is given in Appendix 3.A. For $d=1$, a simple consequence of **Result 3.2.2.1** is

$$\mathfrak{p}_{\mathrm{HS},1}\left(x\right)=\frac{1}{\sqrt{2\pi^{3}}}\exp\left(\frac{x^{2}}{2}\right)E_{1}\left(\frac{x^{2}}{2}\right),\ \ x\in\mathbb{R}.$$

This matches the expression given in the appendix of Carvalho et al. (2010) for the ordinary Horseshoe distribution. For $d=2$, we have

$$\mathfrak{p}_{\mathrm{HS},2}\left(x_{1},x_{2}\right)=\frac{1}{2\sqrt{2\pi^{3}}}\frac{\exp\left(\frac{x_{1}^{2}+x_{2}^{2}}{2}\right)E_{\left(\frac{3}{2}\right)}\left(\frac{x_{1}^{2}+x_{2}^{2}}{2}\right)}{\sqrt{x_{1}^{2}+x_{2}^{2}}},\ \ \left(x_{1},x_{2}\right)\in\mathbb{R}^{2}.$$

$\mathfrak{p}_{\mathrm{HS},2}$ is displayed in Figure 3.1. This perspective plot shows that $\mathfrak{p}_{\mathrm{HS},2}$ has a pole at the origin. We formalise this behaviour for general case $d\in\mathbb{N}$ in Section 3.2.3.

3.2.3 Pole at the Origin Existence

We now investigate various statistical properties of the Grouped Horseshoe distribution. We apply the Grouped Horseshoe prior in Bayesian inference to estimate a parameter vector.

In Carvalho et al. (2010), it shows $\mathfrak{p}_{\mathrm{HS},1}$ has a pole at the origin. Therefore, the ordinary horseshoe distribution can be used as a prior density function. **Result 3.2.3.1** shows that the Grouped Horseshoe density function has a pole at the origin for any dimension. The prior's behaviour near the origin relates to its efficiency of handling the sparsity (Carvalho et al., 2010). **Result 3.2.3.1** indicates the the Grouped Horseshoe distribution can be used as a prior density function and has good potential for shrinkage of noise.

Result 3.2.3.1 *For each* $d\in\mathbb{N}$, $\lim_{\boldsymbol{x}\to0}\mathfrak{p}_{\mathrm{HS,d}}\left(\boldsymbol{x}\right)=\infty$.

The derivation of **Result 3.2.3.1** is given in Appendix 3.B.

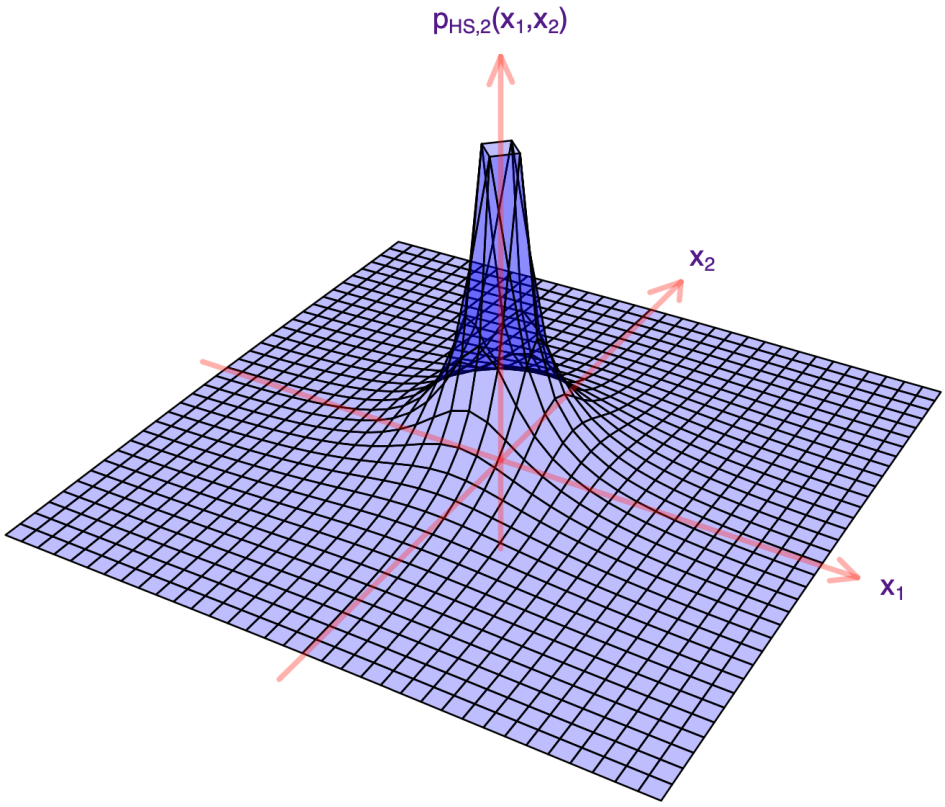


Figure 3.1: *Perspective plot of the bivariate Grouped Horseshoe density function:* $\mathfrak{p}_{\mathrm{HS},2}$.

3.2.4 Score Function and Tail Robustness

We consider the following model (3.2). Note, $\boldsymbol{y}$ here is a generic random variable, unrelated to the observed data defined in Chapter 2.

$$\boldsymbol{y}\big|\boldsymbol{\theta} \sim N(\boldsymbol{\theta}, \boldsymbol{I}_d) \quad \text{with prior} \quad \mathfrak{p}(\boldsymbol{\theta}) = \mathfrak{p}_{\mathrm{HS,d}}\left(\boldsymbol{\theta}/\tau\right)/\tau^d \tag{3.2}$$

with $\tau > 0$. τ is fixed and known. In Section 2 of Carvalho et al. (2010), the $d = 1$ case of (3.2) is considered. The score function for $d = 1$ case is shown as follows:

$$\frac{d\log\{\mathfrak{p}(y)\}}{dy} \quad \text{converges to} \quad 0 \quad \text{as} \quad \big|y\big| \to \infty. \tag{3.3}$$

Carvalho et al. (2010) stated that (3.3) implies a type of robustness for large signals which is called *tail robustness*. A prior's tail robustness is given in terms of the score function. **Result 3.2.4.1** shows that grouped horseshoe priors also possesses this property.

In **Result 3.2.4.1**, the redescending score function shows that the grouped horseshoe priors are heavy-tailed in constructing robust estimators, to prevent overshrinkage of obvious signals. We proved the score function converges to zero for large signals and the risk of the grouped horseshoe estimator is bounded. **Result 3.2.4.1** shows that the score function behaviour and robustness to large signals property of horseshoe priors extend to grouped situations. The proof of **Result 3.2.4.1** is given in Appendix 3.C.

Result 3.2.4.1 *For model (3.2), the tail behaviour of the score function is given by*

$$\nabla_{\boldsymbol{y}} \log\{\mathfrak{p}(\boldsymbol{y})\} \sim -\frac{(d+1)\,\boldsymbol{y}}{\|\boldsymbol{y}\|^2} \quad \textit{for} \quad \|\boldsymbol{y}\| \gg 1.$$

Consequently

$$\lim_{\|\boldsymbol{y}\|\to\infty} \nabla_{\boldsymbol{y}} \log\{\mathfrak{p}(\boldsymbol{y})\} = \mathbf{0} \quad \text{and} \quad E\Big\|\boldsymbol{y} - E(\boldsymbol{\theta}|\boldsymbol{y})\Big\| \le b_\tau$$

for some $b_\tau < \infty$ that depends on τ.

The first displayed equation in **Result 3.2.4.1** uses the common "$\sim$" and "$\gg$" notations. The precise result, without these notations, is

$$\lim_{\|\boldsymbol{y}\|\to\infty} \left[\nabla_{\boldsymbol{y}} \log\{\mathfrak{p}(\boldsymbol{y})\} \Big/ \left\{-\frac{(d+1)\,\boldsymbol{y}}{\|\boldsymbol{y}\|^2}\right\}\right] = 1.$$

3.2.5 Risk Convergence Rates

We are aiming to derive a general result in risk convergence rates, to be applied later to our specific data and model, stated in Section 3.3–3.10.

Consider the following model

$$\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \big| \boldsymbol{\theta} \;\; \text{independently distributed as } N\left(\boldsymbol{\theta}, \sigma^2 \boldsymbol{I}_d\right) \text{ with prior } \mathfrak{p}(\boldsymbol{\theta}) = \mathfrak{p}_{\mathrm{HS,d}}(\boldsymbol{\theta}). \tag{3.4}$$

Suppose that the true sampling distribution of the $\boldsymbol{y}_i$ is $N\left(\boldsymbol{\theta}^0, \sigma^2 \boldsymbol{I}_d\right)$. For the $d = 1$ case, Theorem 4 of Carvalho et al. (2010) states that rates of convergence results is called the Cesàro-average risk of the Bayes estimator of θ and is denoted as R_n. The risk quantity R_n is specifically referred to the estimator's asymptotic rates of convergence, measured by the Kullback-Leibler divergence between the true sampling

model and the Bayes estimator of the density function (Carvalho et al., 2010). The rates of convergence are different in the case of $\theta^0 = 0$ and $\theta^0 \neq 0$. θ^0 is the value of θ according to the sampling distribution of the $\boldsymbol{y}_i$. θ^0 is the true value of parameter. Carvalho et al. (2010) shows that the horseshoe prior leads to a super-efficient risk rate when $\theta^0 = 0$.

We derived the risk rate bounds stated in **Result 3.2.5.1** for the sparse situation ($\boldsymbol{\theta}^0 = \mathbf{0}$) and the non-sparse situations ($\boldsymbol{\theta}^0 \neq \mathbf{0}$). Theorem 4 of Carvalho et al. (2010) states the risk rate bounds for horseshoe prior ($d = 1$) in the sparse situation. Our established **Result 3.2.5.1** investigated the risk rate bounds of Grouped Horseshoe distribution, which is the extension of Theorem 4 of Carvalho et al. (2010). The definition of R_n for the Grouped Horseshoe distribution is analogous to that as given in Section 3.3 of Carvalho et al. (2010). In the derivation of **Result 3.2.5.1**, we consider the d-variate extension of the set-up as stated in Section 3.3 of Carvalho et al. (2010).

Result 3.2.5.1 *Consider model (3.4) and suppose that $\boldsymbol{y}_i$ have a sampling distribution given by $N\left(\boldsymbol{\theta}^0, \sigma^2 \boldsymbol{I}_d\right)$. Let R_n be Cesàro-average risk of the Bayes estimator of $\boldsymbol{\theta}$. When $\boldsymbol{\theta}^0 = \mathbf{0}$, we have*

$$R_n \leq \begin{cases} \dfrac{\log(n)}{2n} - \dfrac{\log\{\log(n)\}}{n} + O\left(\dfrac{1}{n}\right), & \text{if } d = 1, \\ \dfrac{\log(n)}{2n} + O\left(\dfrac{1}{n}\right), & \text{if } d \geq 2. \end{cases}$$

When $\boldsymbol{\theta}^0 \neq \mathbf{0}$, we have $R_n \leq \dfrac{d \log(n)}{2n} + O\left(\dfrac{1}{n}\right)$ for all $d \in \mathbb{N}$.

The full derivation details of **Result 3.2.5.1** is given in Appendix 3.D. For the $d = 1$ and $\theta^0 = 0$ case, the super-efficiency of risk rates corresponds to the presence of the term $-\log\{\log(n)\}/n$ in the upper bound on R_n. It implies that the horseshoe prior ($d = 1$) will ensure the Bayes estimator for the sampling density converges to the right answer at a super-efficient rate. **Result 3.2.5.1** shows that this term only arises in the $d = 1$ case for horseshoe prior. This super-efficiency property based on the risk rate of convergence does not extend to grouped situation. The Bayes estimator is not super-efficient for $d \geq 2$. The grouped horseshoe prior does not ensure the Bayes estimator for the sampling density converges to the right answer at a super-efficient rate, whether sparse or not.

The super-efficient rate terminology is used in Carvalho et al. (2010), to describe this phenomenon in risk rate bounds. In the $d = 1$ case, even though the $\log\{\log(n)\}/n$ term is negligible compared to the $\log(n)/(2n)$ term, the negative sign in front of the

$\log\{\log(n)\}/n$ term means that there is a lower asymptotic risk in the $d = 1$ case in the sparse situation.

3.2.6 Thresholding

From Section 3.2.3 – 3.2.5, we note that most properties possessed by the univariate horseshoe priors extend to the grouped situation, such as

- having a pole at the origin

  and

- the score function behavior and robustness to large signals.

However, the super efficient risk rate does not extend to the grouped situation. The Grouped Horseshoe distribution can be used as a prior density function for model selection as it has a pole at the origin and due to its robustness to large signals. To obtain the sparse fits, we use (Grouped) Horseshoe priors as sparse signals priors for count responses and Gaussian responses in Section 3.5. In this subsection, we derived the posterior mean in a Bayesian model using the Grouped Horseshoe distribution as stated in **Result 3.2.6.1**. Based on **Result 3.2.6.1**, we obtained the expression of posterior mean of γ_β and γ_u for thresholding using (Grouped) Horseshoe priors in Section 3.7.

Consider a Bayesian model as the following form

$$\boldsymbol{\theta}\big|\sigma_\theta \quad \text{has density function} \quad \mathfrak{p}_{\text{HS},d}\left(\boldsymbol{\theta}/\sigma_\theta\right)/\sigma_\theta^d \quad \text{where } \boldsymbol{\theta} \text{ is } d \times 1. \tag{3.5}$$

From results in Section 3.2.2, (3.5) is equivalent to

$$\boldsymbol{\theta}\big|\sigma_\theta, \lambda \sim N\left(\mathbf{0}, \sigma_\theta^2\lambda^2\boldsymbol{I}_d\right), \quad \mathfrak{p}(\lambda) = \frac{2I(\lambda > 0)}{\pi(1+\lambda^2)}. \tag{3.6}$$

We introduce the auxiliary variable λ, which is very important for the upcoming approach to thresholding. In the scalar case, Carvalho et al. (2010) develop a thresholding approach for deciding between

$$\theta = 0 \quad \text{and} \quad \theta \neq 0, \quad \theta \in \mathbb{R}.$$

In this subsection, we describe and evaluate the extension of the approach in scalar

case to deciding between

$$\boldsymbol{\theta} = \mathbf{0} \quad \text{and} \quad \boldsymbol{\theta} \neq \mathbf{0}, \;\; \boldsymbol{\theta} \in \mathbb{R}^d.$$

The approach stated in Carvalho et al. (2010) involves the following result concerning a "side" model:

Result 3.2.6.1 *For the Bayesian model*

$$\boldsymbol{y}\big|\boldsymbol{\psi} \sim N\left(\boldsymbol{\psi}, \tau_1^2 \boldsymbol{I}_d\right), \;\; \boldsymbol{\psi}\big|\lambda \sim N\left(\mathbf{0}, \lambda^2 \tau_2^2 \boldsymbol{I}_d\right), \;\; \mathfrak{p}\left(\lambda\right) = \frac{2I\left(\lambda > 0\right)}{\pi\left(1+\lambda^2\right)}, \;\; \tau_1, \tau_2 > 0 \;\; fixed \tag{3.7}$$

the posterior mean of $\boldsymbol{\psi}$ *is*

$$E\left(\boldsymbol{\psi}\big|\boldsymbol{y}\right) = E\left(\frac{\lambda^2\tau_2^2}{\tau_1^2 + \lambda^2\tau_2^2}\middle|\boldsymbol{y}\right)\boldsymbol{y}.$$

The full derivation of **Result 3.2.6.1** is given in Appendix 3.E.

Based on **Result 3.2.6.1**, for deciding whether or not a coefficient parameter in a Bayesian model is set to zero using (Grouped) Horseshoe prior, we have the following rule:

$$\text{decide that } \boldsymbol{\theta} = \mathbf{0} \text{ if and only if } E\left(\gamma_{\text{GHS}}\big|\boldsymbol{y}\right) < 1 - \tau \quad \text{where } \gamma_{\text{GHS}} \equiv \frac{\lambda^2\sigma_\theta^2}{\sigma_\varepsilon^2 + \lambda^2\sigma_\theta^2} \tag{3.8}$$

where τ is the threshold parameter, described in Section 2.6.

For Gaussian responses in model (3.23), we ran a simulation study for comparative performance, as described in Section 3.8, between use of the Laplace Zero/-Grouped Lasso-Zero priors described in Chapter 2 and use of the (Grouped) Horseshoe priors described in this chapter. The threshold parameter τ used for Gaussian responses is the same as that decided in Section 2.6.3 for MCMC approach, being 0.5.

Therefore, for Gaussian responses in general Bayesian models containing (3.5), or equivalently, (3.6) forms **Result 3.2.6.1** suggests the following rule

$$\text{decide that } \boldsymbol{\theta} = \mathbf{0} \text{ if and only if } E\left(\gamma_{\text{GHS}}\big|\boldsymbol{y}\right) < \frac{1}{2} \quad \text{where } \gamma_{\text{GHS}} \equiv \frac{\lambda^2\sigma_\theta^2}{\sigma_\varepsilon^2 + \lambda^2\sigma_\theta^2} \tag{3.9}$$

Bhadra et al. (2019) stated that instead of placing a prior on the model space to yield a sparse estimator, which is referred to $\boldsymbol{\gamma}_\beta$ or $\boldsymbol{\gamma}_u$ described in Chapter 2 with use

of the Laplace-Zero/Grouped Lasso-Zero prior, (Grouped) horseshoe prior models the posterior inclusion probabilities $P\left(\boldsymbol{\theta} \neq 0 \middle| \boldsymbol{y}\right)$ directly.

We use (3.8) to set up the model selection strategies on generalized additive model, with use of the (Grouped) Horseshoe priors, for count responses and Gaussian responses, described in Section 3.7. Using these model selection strategies, we ran simulation study for comparative performance on Gaussian responses, described in Section 3.8, and comparative performance on count responses, described in Section 3.9 – 3.10.

3.3 Original Input Data and Transformed Data

For Horseshoe-type priors, the original input data and transformed data follow the model described in Section 2.2.

3.3.1 The Original Input Data

The definitions of $\mathring{\boldsymbol{x}}_i^{\text{orig}}, \dot{\boldsymbol{x}}_i^{\text{orig}}, \mathring{x}_{ij}^{\text{orig}}, \dot{x}_{ij}^{\text{orig}}, d_\circ, d_\bullet, 1 \leq i \leq n$, are as in Section 2.2.1, with $1 \leq j \leq d_\circ$, $\mathring{x}_{ij}^{\text{orig}}$ referring to original predictors can only enter the model linearly, and with $1 \leq j \leq d_\bullet$, $\dot{x}_{ij}^{\text{orig}}$ referring to original predictors can enter the model either linearly or non-linearly.

3.3.2 Standardization of the Input Data for Fitting and Model Selection

The definitions of $\mathring{\boldsymbol{x}}_i, \dot{\boldsymbol{x}}_i, \mathring{x}_{ij}, \dot{x}_{ij}, a_x, b_x, c_x, d_x$ are as in Section 2.2.2. Standardization of $\mathring{\boldsymbol{x}}_{ij}$ and $\dot{\boldsymbol{x}}_{ij}$ follows (2.1) and (2.2), with $1 \leq j \leq d_\circ$, $\mathring{x}_{ij}$ referring to standardized predictors can only enter the model linearly, with $1 \leq j \leq d_\bullet$, $\dot{x}_{ij}$ referring to standardized predictors can enter the model either linearly or non-linearly. No standardization is required for the count responses data $y_i, 1 \leq i \leq n$.

3.4 Design Matrices and Auxiliary Variable Representations

3.4.1 Generalized Additive Model Form for the Full Model

In this chapter, the linear predictor given by (2.5) of Section 2.3.1 takes the following form

$$\begin{aligned}
&\beta_0 + \sum_{j=1}^{d_\circ} \left\{ (\mathbf{1}_n \beta_j) \odot \mathring{\boldsymbol{x}}_j \right\} + \sum_{j=1}^{d_\bullet} \left\{ (\mathbf{1}_n \beta_j) \odot \dot{\boldsymbol{x}}_j \right\} + \sum_{j=1}^{d_\bullet} \left\{ \sum_{k=1}^{K_j} (\mathbf{1}_n u_{jk}) \odot \dot{\boldsymbol{Z}}_{jk} \right\} \\
&= \beta_0 + \sum_{j=1}^{d_\circ + d_\bullet} \left\{ (\mathbf{1}_n \beta_j) \odot \dot{\boldsymbol{X}}_j \right\} + \sum_{j=1}^{d_\bullet} \left\{ \sum_{k=1}^{K_j} (\mathbf{1}_n u_{jk}) \odot \dot{\boldsymbol{Z}}_{jk} \right\} \\
&= \beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j \boldsymbol{u}_j. \qquad (3.10)
\end{aligned}$$

Note that (3.10) differs from (2.6) due to differences between Horseshoe and Laplace-Zero priors. The definitions of $\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{Z}_j, K_j, \boldsymbol{u}_j$ with $1 \leq j \leq d_\bullet$, the definition of $\dot{\boldsymbol{X}}_j$ with $1 \leq j \leq d_\circ + d_\bullet$, and the definition of $\dot{\boldsymbol{Z}}_{jk}$ with $1 \leq j \leq d_\bullet, 1 \leq k \leq K_j$ are as in Section 2.3.1.

3.4.2 The Linear Design Matrix

It follows Section 2.3.2 to form the Linear Design Matrix $\boldsymbol{X}$. Following (2.7), the feature of $\boldsymbol{X}$ matrix for count responses also offers simplification to be $\mathbf{1}_n^T \boldsymbol{X} = 0$.

3.4.3 The Canonical Demmler-Reinsch Basis Design Matrices

It follows Section 2.3.3 to form the canonical Demmler-Reinsch basis design matrices $\boldsymbol{Z}_j, 1 \leq j \leq d_\bullet$. Following (2.8), the feature of $\boldsymbol{Z}$ matrix for count responses also offers simplification to be $\mathbf{1}_n^T \boldsymbol{Z}_j = \mathbf{0}_{K_j}, 1 \leq j \leq d_\bullet$, and $\boldsymbol{Z}_j^T \boldsymbol{Z}_j$ is a diagonal matrix.

3.4.4 Linear Term Coefficients with Auxiliary Variable Representations

In our full model, for count responses, using Horseshoe priors, let $\boldsymbol{\beta}$ denote a $(d_\circ + d_\bullet) \times 1$ vector of linear term coefficients for the linear design matrix $\boldsymbol{X}$ described in Section 2.3.2. For each $\beta_j, 1 \leq j \leq d_\circ + d_\bullet$, based on **Result 3.2.6.1** for thresholding, it has the following Horseshoe-type prior distributions for the sparseness of coefficients:

$$\mathfrak{p}(\beta_j | \sigma_\beta) = \mathfrak{p}_{\text{HS},1}\left(\beta_j / \sigma_\beta\right) / \sigma_\beta, \;\; 1 \leq j \leq d_\circ + d_\bullet. \tag{3.11}$$

$$\sigma_\beta \sim \text{Half-Cauchy}\left(s_\beta\right). \tag{3.12}$$

$\mathfrak{p}_{\text{HS},1}\left(\beta_j / \sigma_\beta\right)$ in (3.11) is the ordinary Horseshoe distribution, described in Section 3.2.2. Following Model III in Table 1 of Neville et al. (2014), (3.11) is equivalent to

$$\begin{aligned} &\beta_j | \sigma_\beta^2, f_{2\beta_j} \overset{\text{ind.}}{\sim} N\left(0, \sigma_\beta^2 / f_{2\beta_j}\right), 1 \leq j \leq d_\circ + d_\bullet, \\ &f_{2\beta_j} | f_{1\beta_j} \overset{\text{ind.}}{\sim} \text{Gamma}\left(\tfrac{1}{2}, f_{1\beta_j}\right), f_{1\beta_j} \overset{\text{ind.}}{\sim} \text{Gamma}\left(\tfrac{1}{2}, 1\right), 1 \leq j \leq d_\circ + d_\bullet. \end{aligned} \tag{3.13}$$

(3.12) is equivalent to

$$\sigma_\beta^2 | a_\beta \sim \text{Inverse-Gamma}\left(\tfrac{1}{2}, 1/a_\beta\right), a_\beta \sim \text{Inverse-Gamma}\left(\tfrac{1}{2}, 1/s_\beta^2\right). \tag{3.14}$$

Therefore, (3.13) and (3.14) are used in our full model setting for Poisson responses described in Section 3.5.3, for Negative-Binomial responses described in Section 3.5.4 and for Gaussian responses described in Section 3.5.2.

3.4.5 Spline Coefficient Vectors with Auxiliary Variable Representations

In our full model, for count responses, let $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{d_\bullet}$ be the full set of spline coefficient vectors for the canonical Demmler-Reinsch basis matrices $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{d_\bullet}$, described in Section 2.3.3, where $\boldsymbol{u}_j$ is a $K_j \times 1$ vector, for corresponding matrix $\boldsymbol{Z}_j$, where $\boldsymbol{Z}_j$ is a $n \times K_j$ matrix with $1 \leq j \leq d_\bullet$. For each $\boldsymbol{u}_j$, based on **Result 3.2.6.1** for thresholding in Section 3.2.6, it has the following grouped LASSO extensions of the

Horseshoe-type prior distribution:

$$\mathfrak{p}(\boldsymbol{u}_j|\sigma_{uj}) = \mathfrak{p}_{\text{HS},K_j}\left(\boldsymbol{u}_j/\sigma_{uj}\right)/\sigma_{uj}^{K_j} \tag{3.15}$$

$$\sigma_{uj} \sim \text{Half-Cauchy}\left(s_u\right), 1 \leq j \leq d_\bullet. \tag{3.16}$$

$\mathfrak{p}_{\text{HS},K_j}\left(\boldsymbol{u}_j/\sigma_{uj}\right)$ in (3.15) is the Grouped Horseshoe distribution, described in Section 3.2.2. Following Model III in Table 1 of Neville et al. (2014), (3.15) is equivalent to:

$$\begin{aligned} &\boldsymbol{u}_j|\sigma_{uj}^2, f_{2uj} \overset{\text{ind.}}{\sim} N\left(0, \left(\sigma_{uj}^2/f_{2uj}\right)\boldsymbol{I}_{K_j}\right), 1 \leq j \leq d_\bullet, \\ &f_{2uj}|f_{1uj} \overset{\text{ind.}}{\sim} \text{Gamma}\left(\tfrac{1}{2}, f_{1uj}\right), f_{1uj} \overset{\text{ind.}}{\sim} \text{Gamma}\left(\tfrac{1}{2}, 1\right), 1 \leq j \leq d_\bullet. \end{aligned} \tag{3.17}$$

(3.16) is equivalent to

$$\sigma_{uj}^2|a_{uj} \sim \text{Inverse-Gamma}\left(\tfrac{1}{2}, 1/a_{uj}\right), a_{uj} \sim \text{Inverse-Gamma}\left(\tfrac{1}{2}, 1/s_u^2\right), 1 \leq j \leq d_\bullet. \tag{3.18}$$

Therefore, (3.17) and (3.18) are used in our full model setting for Poisson responses described in Section 3.5.3 , for Negative-Binomial responses described in Section 3.5.4 and for Gaussian responses described in Section 3.5.2.

3.5 Essence of the Bayesian Model

In this section, we describe, in essence, the Bayesian model for Gaussian and count responses.

3.5.1 The Likelihood of the Response

For Gaussian responses, the likelihood function of the response is

$$\boldsymbol{y}|\beta_0, \boldsymbol{\beta}, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_{d_\bullet} \sim N\left(\mathbf{1}_n\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j\boldsymbol{u}_j, \sigma_\varepsilon^2\boldsymbol{I}_n\right). \tag{3.19}$$

For Poisson responses, the likelihood function of the response is

$$y_i|\beta_0, \boldsymbol{\beta}, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_{d_\bullet} \overset{\text{ind.}}{\sim} \text{Poisson}\left(\exp\left\{\beta_0 + (\boldsymbol{X}\boldsymbol{\beta})_i + \sum_{j=1}^{d_\bullet}\left(\boldsymbol{Z}_j\boldsymbol{u}_j\right)_i\right\}\right), \quad 1 \leq i \leq n. \tag{3.20}$$

For Negative-Binomial responses, from Result 1 of Luts and Wand (2015) in Appendix A.2, also described in **Result 1.8.5** in Section 1.8 in this thesis, in our case, we have

$$\mu = \exp\left\{\beta_0 + (\boldsymbol{X}\boldsymbol{\beta})_i + \sum_{j=1}^{d_\bullet} (\boldsymbol{Z}_j \boldsymbol{u}_j)_i\right\}, \;\; a = g_i, \;\; x = y_i, \;\; 1 \le i \le n.$$

Therefore

$$\begin{aligned} &y_i \big| g_i \overset{\text{ind.}}{\sim} \text{Poisson}\left(g_i\right), \;\; 1 \le i \le n, \text{ and} \\ &g_i \big| \beta_0, \boldsymbol{\beta}, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_{d_\bullet}, \kappa \overset{\text{ind.}}{\sim} \text{Gamma}\left(\kappa, \kappa \exp\left\{-\left(\beta_0 + (\boldsymbol{X}\boldsymbol{\beta})_i + \sum_{j=1}^{d_\bullet} (\boldsymbol{Z}_j \boldsymbol{u}_j)_i\right)\right\}\right) \end{aligned} \tag{3.21}$$

is equivalent to

$$y_i \big| \beta_0, \boldsymbol{\beta}, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_{d_\bullet}, \kappa \overset{\text{ind.}}{\sim} \text{Negative-Binomial}\left(\exp\left\{\beta_0 + (\boldsymbol{X}\boldsymbol{\beta})_i + \sum_{j=1}^{d_\bullet} (\boldsymbol{Z}_j \boldsymbol{u}_j)_i\right\}, \kappa\right). \tag{3.22}$$

Therefore, (3.21) is the likelihood function of the response for Negative-Binomial responses described in Section 3.5.4. The $g_i, 1 \le i \le n$, and κ are the auxiliary parameters.

3.5.2 The Full Bayesian Gaussian Response Additive Model Selection

For Gaussian responses with Horseshoe-type priors, we set up the full model to be (3.23) in this subsection. Figure 3.2 shows the directed acyclic graph corresponding to (3.23):

$$\boldsymbol{y} \big| \beta_0, \boldsymbol{\beta}, \widetilde{\boldsymbol{u}}_1, \ldots, \widetilde{\boldsymbol{u}}_{d_\bullet}, \sigma_\varepsilon^2 \sim N\left(\boldsymbol{1}_n \beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j \boldsymbol{u}_j, \sigma_\varepsilon^2 \boldsymbol{I}_n\right), \;\; \beta_0 \sim N(0, \sigma_{\beta_0}^2),$$

$$\sigma_\varepsilon^2 \big| a_\varepsilon \sim \text{Inverse-Gamma}(\frac{1}{2}, 1/a_\varepsilon), \;\; a_\varepsilon \sim \text{Inverse-Gamma}(\frac{1}{2}, 1/s_\varepsilon^2),$$

$$\beta_j \big| \sigma_\beta^2, f_{2\beta_j} \overset{\text{ind.}}{\sim} N(0, \sigma_\beta^2 / f_{2\beta_j}), \;\; 1 \le j \le d_\circ + d_\bullet,$$

$$f_{2\beta_j} \big| f_{1\beta_j} \overset{\text{ind.}}{\sim} \text{Gamma}(\tfrac{1}{2}, f_{1\beta_j}), \;\; f_{1\beta_j} \overset{\text{ind.}}{\sim} \text{Gamma}(\tfrac{1}{2}, 1), \;\; 1 \le j \le d_\circ + d_\bullet,$$

$$\boldsymbol{u}_j \big| \sigma_{uj}^2, f_{2uj} \overset{\text{ind.}}{\sim} N(\boldsymbol{0}, (\sigma_{uj}^2 / f_{2uj}) \boldsymbol{I}_{K_j}), \;\; 1 \le j \le d_\bullet,$$

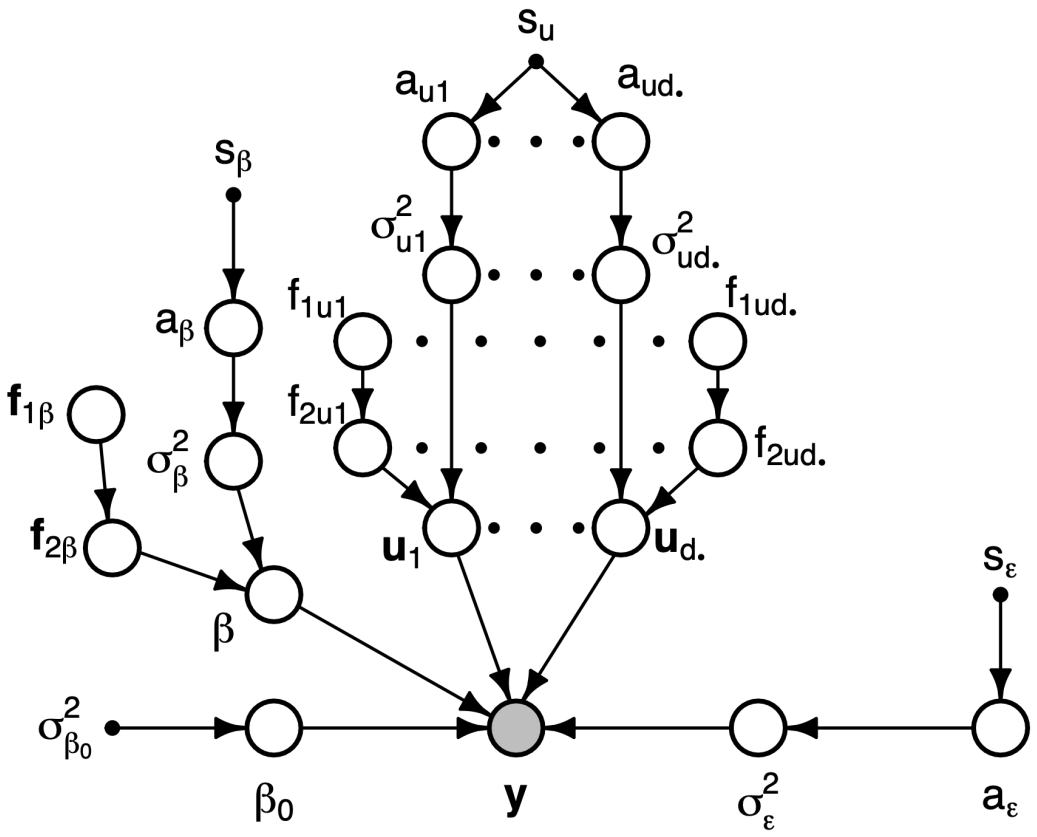


Figure 3.2: *Directed acyclic graph representation of Bayesian model (3.23) for Gaussian response using Horseshoe-type prior. Random variables and vectors are shown as larger open circles, with shading indicating to the observed response data. The small closed circles are user-specified hyperparameters.*

$$
\begin{aligned}
&f_{2uj}\big|f_{1uj} \overset{\text{ind.}}{\sim} \text{Gamma}(\tfrac{1}{2}, f_{1uj}), \;\; f_{1uj} \overset{\text{ind.}}{\sim} \text{Gamma}(\tfrac{1}{2}, 1), \;\; 1 \le j \le d_\bullet, \\
&\sigma^2_\beta\big|a_\beta \sim \text{Inverse-Gamma}(\tfrac{1}{2}, 1/a_\beta), \;\; a_\beta \sim \text{Inverse-Gamma}(\tfrac{1}{2}, 1/s^2_\beta), \\
&\sigma^2_{uj}\big|a_{uj} \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2}, 1/a_{uj}), \;\; a_{uj} \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2}, 1/s^2_u), \;\; 1 \le j \le d_\bullet.
\end{aligned}
\tag{3.23}
$$

The full set of hyperparameters in model (3.23) is:

$$\sigma_{\beta_0} = 10^5, \;\; s_\beta = s_\varepsilon = s_u = 1000.$$

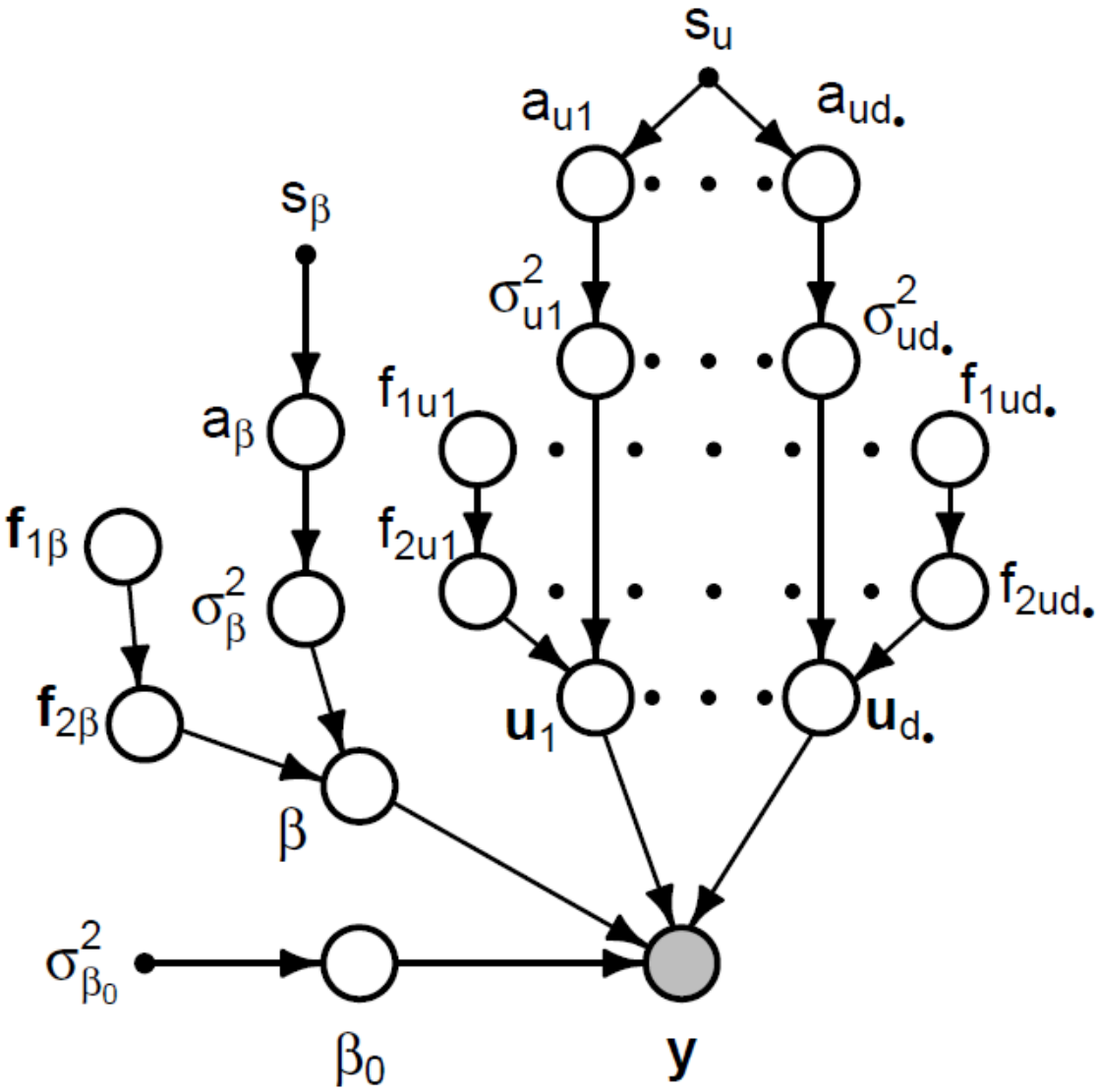


Figure 3.3: *Directed acyclic graph representation of Bayesian model (3.24) for Poisson response using Horse-type priors. Random variables and vectors are shown as larger open circles, with shading indicating to the observed response data. The small closed circles are user-specified hyperparameters.*

3.5.3 The Full Bayesian Poisson Response Additive Model Selection

For Poisson responses with Horseshoe-type priors, we set up the full model to be (3.24) in this subsection. Figure 3.3 shows the directed acyclic graph corresponding

to (3.24).

$$
\begin{aligned}
&y_i\big|\beta_0,\boldsymbol{\beta},\boldsymbol{u}_1,\ldots,\boldsymbol{u}_{d_\bullet}\sim \text{Poisson}\left(\exp\left\{\beta_0+\left(\boldsymbol{X}\boldsymbol{\beta}\right)_i+\sum_{j=1}^{d_\bullet}\left(\boldsymbol{Z}_j\boldsymbol{u}_j\right)_i\right\}\right),\ \ 1\le i\le n,\\
&\beta_0\sim N(0,\sigma^2_{\beta_0}),\ \ \sigma^2_\varepsilon\big|a_\varepsilon\sim \text{Inverse-Gamma}(\frac{1}{2},1/a_\varepsilon),\\
&a_\varepsilon\sim \text{Inverse-Gamma}(\frac{1}{2},1/s^2_\varepsilon),\ \ \beta_j\big|\sigma^2_\beta,f_{2\beta_j}\overset{\text{ind.}}{\sim}N(0,\sigma^2_\beta/f_{2\beta_j}),\ \ 1\le j\le d_\circ+d_\bullet,\\
&f_{2\beta_j}\big|f_{1\beta_j}\overset{\text{ind.}}{\sim}\text{Gamma}(\tfrac{1}{2},f_{1\beta_j}),\ \ f_{1\beta_j}\overset{\text{ind.}}{\sim}\text{Gamma}(\tfrac{1}{2},1),\ \ 1\le j\le d_\circ+d_\bullet,\\
&\boldsymbol{u}_j\big|\sigma^2_{uj},f_{2uj}\overset{\text{ind.}}{\sim}N(\boldsymbol{0},(\sigma^2_{uj}/f_{2uj})\boldsymbol{I}_{K_j}),\ \ 1\le j\le d_\bullet,\\
&f_{2uj}\big|f_{1uj}\overset{\text{ind.}}{\sim}\text{Gamma}(\tfrac{1}{2},f_{1uj}),\ \ f_{1uj}\overset{\text{ind.}}{\sim}\text{Gamma}(\tfrac{1}{2},1),\ \ 1\le j\le d_\bullet,\\
&\sigma^2_\beta\big|a_\beta\sim\text{Inverse-Gamma}(\tfrac{1}{2},1/a_\beta),\ \ a_\beta\sim\text{Inverse-Gamma}(\tfrac{1}{2},1/s^2_\beta),\\
&\sigma^2_{uj}\big|a_{uj}\overset{\text{ind.}}{\sim}\text{Inverse-Gamma}(\tfrac{1}{2},1/a_{uj}),\ \ a_{uj}\overset{\text{ind.}}{\sim}\text{Inverse-Gamma}(\tfrac{1}{2},1/s^2_u),\ \ 1\le j\le d_\bullet.
\end{aligned}
\tag{3.24}
$$

The full set of hyperparameters in model (3.24) is:

$$\sigma_{\beta_0}=10^5,\ \ s_\beta=s_u=1000.$$

3.5.4 The Full Bayesian Negative-Binomial Response Additive Model Selection

For Negative-Binomial responses with Horseshoe-type priors, we set up the full model to be (3.25) in this subsection. Figure 3.4 shows the directed acyclic graph corresponding to (3.25). In (3.25), the auxiliary parameter κ is of *Moon Rock* (McLean and Wand, 2019) exponential family distribution, to form the equivalence of likelihood for Negative-Binomial responses.

$$
\begin{aligned}
&y_i\big|g_i\overset{\text{ind.}}{\sim}\text{Poisson}\left(g_i\right),\ \ 1\le i\le n,\\
&g_i\big|\beta_0,\boldsymbol{\beta},\boldsymbol{u}_1,\ldots,\boldsymbol{u}_{d_\bullet},\kappa\overset{\text{ind.}}{\sim}\text{Gamma}\left(\kappa,\kappa\exp\left\{-\left(\beta_0+\left(\boldsymbol{X}\boldsymbol{\beta}\right)_i+\sum_{j=1}^{d_\bullet}\left(\boldsymbol{Z}_j\boldsymbol{u}_j\right)_i\right)\right\}\right),
\end{aligned}
$$

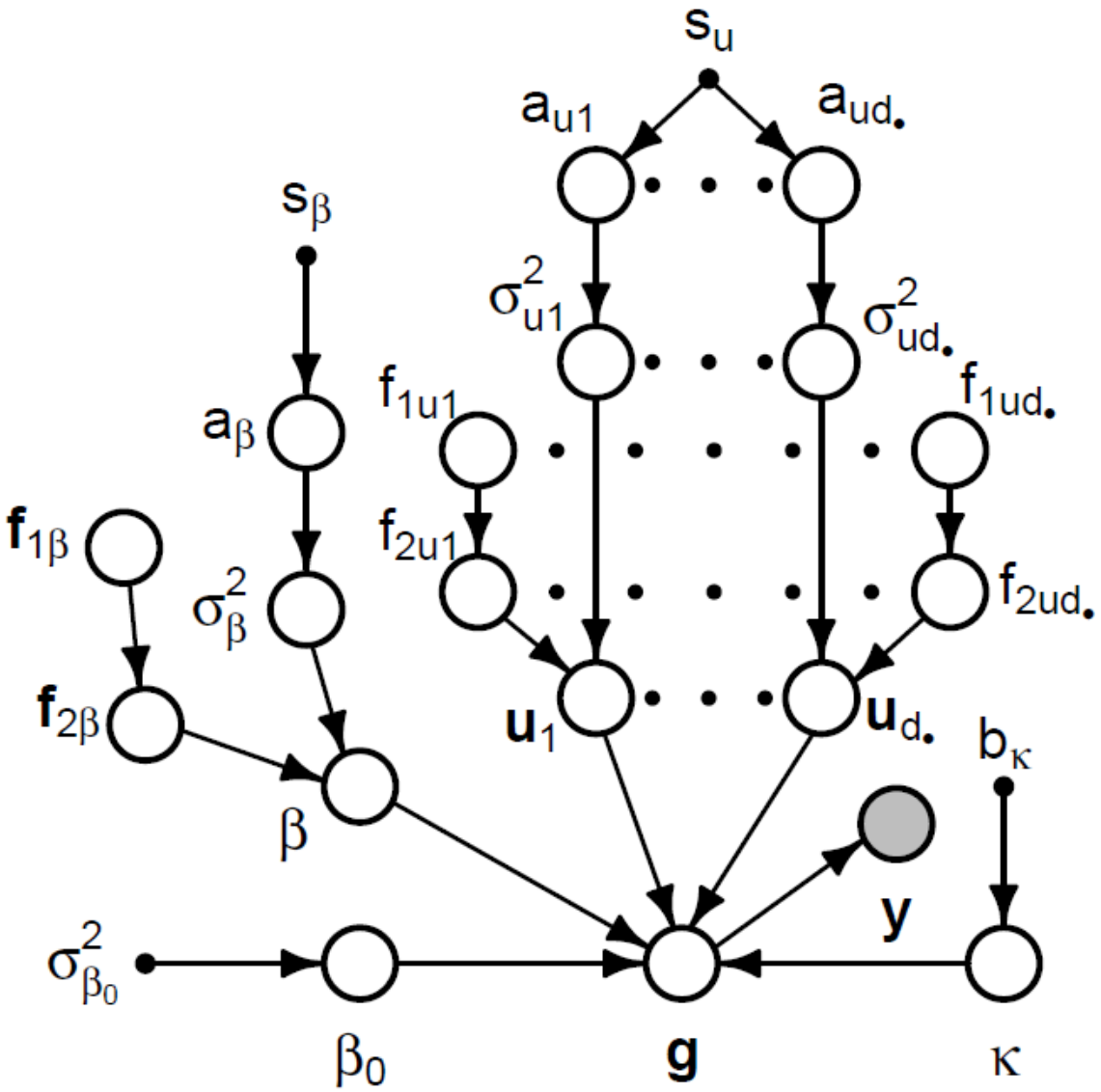


Figure 3.4: *Directed acyclic graph representation of Bayesian model (3.25) for Negative-Binomial response using Horseshoe-type priors. Random variables and vectors are shown as larger open circles, with shading indicating to the observed response data. The small closed circles are user-specified hyperparameters.*

$$1 \leq i \leq n,$$

$$\beta_0 \sim N(0, \sigma^2_{\beta_0}), \;\; \sigma^2_\varepsilon | a_\varepsilon \sim \text{Inverse-Gamma}(\frac{1}{2}, 1/a_\varepsilon),$$

$$a_\varepsilon \sim \text{Inverse-Gamma}(\frac{1}{2}, 1/s^2_\varepsilon), \;\; \beta_j \big| \sigma^2_\beta, f_{2\beta_j} \overset{\text{ind.}}{\sim} N(0, \sigma^2_\beta / f_{2\beta_j}), \;\; 1 \leq j \leq d_\circ + d_\bullet,$$

$$\kappa \sim \text{Moon-Rock}\,(0, b_\kappa)\,,\ \ f_{2\beta_j}\big|f_{1\beta_j} \overset{\text{ind.}}{\sim} \text{Gamma}(\tfrac{1}{2}, f_{1\beta_j}),\ \ 1 \le j \le d_\circ + d_\bullet,$$

$$f_{1\beta_j} \overset{\text{ind.}}{\sim} \text{Gamma}(\tfrac{1}{2}, 1),\ \ 1 \le j \le d_\circ + d_\bullet,$$

$$\boldsymbol{u}_j\big|\sigma^2_{uj}, f_{2uj} \overset{\text{ind.}}{\sim} N(\mathbf{0}, (\sigma^2_{uj}/f_{2uj})\boldsymbol{I}_{K_j}),\ \ 1 \le j \le d_\bullet,$$

$$f_{2uj}\big|f_{1uj} \overset{\text{ind.}}{\sim} \text{Gamma}(\tfrac{1}{2}, f_{1uj}),\ \ f_{1uj} \overset{\text{ind.}}{\sim} \text{Gamma}(\tfrac{1}{2}, 1),\ \ 1 \le j \le d_\bullet,$$

$$\sigma^2_\beta\big|a_\beta \sim \text{Inverse-Gamma}(\tfrac{1}{2}, 1/a_\beta),\ \ a_\beta \sim \text{Inverse-Gamma}(\tfrac{1}{2}, 1/s^2_\beta),$$

$$\sigma^2_{uj}\big|a_{uj} \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2}, 1/a_{uj}),\ \ a_{uj} \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2}, 1/s^2_u),\ \ 1 \le j \le d_\bullet. \tag{3.25}$$

The full set of hyperparameters in model (3.25) is:

$$\sigma_{\beta_0} = 10^5,\ \ s_\beta = s_u = 1000,\ \ b_\kappa = 0.001.$$

3.6 Posterior Distributions for Markov Chain Monte Carlo Practical Fitting

3.6.1 Pre-Processing and Storage of Key Matrices

The pre-processing of linear design matrix $\boldsymbol{X}$ and storage of Key Matrices $\boldsymbol{X}^T\boldsymbol{y}$, $\boldsymbol{X}^T\boldsymbol{X}$, $\boldsymbol{Z}^T\boldsymbol{y}$, $\boldsymbol{Z}^T\boldsymbol{X}$, $\boldsymbol{Z}^T\boldsymbol{Z}$ for Poisson responses and Negative-Binomial responses follow Section 2.5.1, and corresponding detailed steps of standardization follow Algorithm 1 described in Section 2.5.1. However, no pre-processing for response data $\boldsymbol{y}$ for Poisson responses and Negative-Binomial responses.

3.6.2 Slice Sampling

For Poisson response and Negative-Binomial response using Horseshoe prior, we need to use slice sampling in Markov chain Monte Carlo approach. The "stepping out" slice sampling strategy of Neal (2003) has a particularly simple implementation. For random variables $s_1 \in \mathbb{R}$, $s_2 > 0$, as well as random vectors $\boldsymbol{s}_3$ and $\boldsymbol{s}_4$, of the same dimension, let

$$x\big|s_1, s_2, \boldsymbol{s}_3, \boldsymbol{s}_4 \sim \mathcal{H}\,(s_1, s_2, \boldsymbol{s}_3, \boldsymbol{s}_4)$$

Algorithm 3: *Slice sampling generation of samples.*

Function to compute the logarithm of the unnormalised density function that arises in slice sampling for Poisson responses and Negative-Binomial responses called logUnnDens( ).

Data Inputs: $x, s_1, s_2, \boldsymbol{s}_3, \boldsymbol{s}_4,$

Outputs: $xs_1 - x^2/(2s_2) - \mathbf{1}^T(x\boldsymbol{s}_3 + \boldsymbol{s}_4)$

Function to execute slice sampling for Poisson responses and Negative-Binomial responses called sliceSampling, needs to use function logUnnDens.

Slice Sampling Data Inputs: $x, s_1, s_2, \boldsymbol{s}_3, \boldsymbol{s}_4, g,$ sumDists.

Initialize: sumDists $\longleftarrow 0$.

$\omega_1 \longleftarrow$ logUnnDens $(x, s_1, s_2, \boldsymbol{s}_3, \boldsymbol{s}_4)$.

if $(g = 2)$ $w \longleftarrow 1.0$; if $(g > 2)$ $w \longleftarrow$ sumDists$/(g-1)$

$u \longleftarrow$ a draw from the Uniform $(0, 1)$ distribution

$a \longleftarrow \omega_1 + \log(u)$; $L \longleftarrow x - wu$; $R \longleftarrow L + w$

$\omega_3 \longleftarrow$ logUnnDens $(L, s_1, s_2, \boldsymbol{s}_3, \boldsymbol{s}_4)$.

while $(a < \omega_3)$

$L \longleftarrow L - w$; $\omega_3 \longleftarrow$ logUnnDens $(L, s_1, s_2, \boldsymbol{s}_3, \boldsymbol{s}_4)$

$\omega_4 \longleftarrow$ logUnnDens $(R, s_1, s_2, \boldsymbol{s}_3, \boldsymbol{s}_4)$

while $(a < \omega_4)$

$R \longleftarrow R + w$; $\omega_4 \longleftarrow$ logUnnDens $(R, s_1, s_2, \boldsymbol{s}_3, \boldsymbol{s}_4)$

$\widetilde{L} \longleftarrow L$; $\widetilde{R} \longleftarrow R$; newPointFound $\longleftarrow$ FALSE

while (newPointFound = FALSE)

$x_{\text{New}} \longleftarrow \widetilde{L} + u\left(\widetilde{R} - \widetilde{L}\right)$

if $(a <$ logUnnDens $(x_{\text{New}}, s_1, s_2, \boldsymbol{s}_3, \boldsymbol{s}_4))$ newPointFound $\longleftarrow$ TRUE

if $(x_{\text{New}} < x)$ $\widetilde{L} \longleftarrow x_{\text{New}}$; if $(x_{\text{New}} \geq x)$ $\widetilde{R} \longleftarrow x_{\text{New}}$

Outputs: x_{New}.

denote that random variable x, conditional on $(s_1, s_2, \boldsymbol{s}_3, \boldsymbol{s}_4)$ has density function

$$\mathfrak{p}(x\big|s_1, s_2, \boldsymbol{s}_3, \boldsymbol{s}_4) \propto \exp\left\{s_1 x - x^2/(2s_2) - \mathbf{1}^T \exp\left(x\boldsymbol{s}_3 + \boldsymbol{s}_4\right)\right\}, \quad -\infty < x < \infty, \tag{3.26}$$

where $\mathbf{1}$ denotes a vector of ones having the same number of rows as $\boldsymbol{s}_3$ and $\boldsymbol{s}_4$ and $\exp\left(x\boldsymbol{s}_3 + \boldsymbol{s}_4\right)$ is evaluated element-wise.

Algorithm 3 in this subsection provides steps of function called logUnnDens, to obtain logarithm of unnormalised density function to be

$$\left\{s_1 x - x^2/(2s_2) - \mathbf{1}^T \exp\left(x\boldsymbol{s}_3 + \boldsymbol{s}_4\right)\right\}.$$

Algorithm 3 also gives detailed steps of slice sampling in the function called sliceSampling for us to draw samples for some parameters for Poisson responses and Negative-Binomial responses. Function sliceSampling needs to call function logUnnDens.

3.6.3 Markov Chain Monte Carlo for Horseshoe-type Priors

For Gaussian responses using the Markov blanket, based on the full model description in (3.23) and Figure 3.2, we obtain full conditional distributions in standard forms. The full conditional distributions of parameters listed in (3.27) – (3.39), with derivation details given in Appendix 3.F. Based on these, Algorithm 4 gives detailed steps of sample drawing for Markov chain Monte Carlo using the Horseshoe-type prior.

$$\beta_0\big|\text{rest} \sim N\left(\frac{\frac{1}{\sigma_\varepsilon^2}\mathbf{1}_n^T\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j\boldsymbol{u}_j\right)}{(n/\sigma_\varepsilon^2) + (1/\sigma_{\beta_0}^2)}, \frac{1}{(n/\sigma_\varepsilon^2) + (1/\sigma_{\beta_0}^2)}\right), \tag{3.27}$$

$$\boldsymbol{\beta}\big|\text{rest} \sim N\Bigg(\left(\frac{1}{\sigma_\varepsilon^2}\left(\boldsymbol{X}^T\boldsymbol{X}\right) + \frac{1}{\sigma_\beta^2}\text{diag}\left(\boldsymbol{f}_{2\beta}\right)\right)^{-1}\frac{1}{\sigma_\varepsilon^2}\left(\left(\boldsymbol{X}^T\boldsymbol{y}\right) - \sum_{j=1}^{d_\bullet}\left(\boldsymbol{X}^T\boldsymbol{Z}_j\right)\boldsymbol{u}_j\right), \left(\frac{1}{\sigma_\varepsilon^2}\left(\boldsymbol{X}^T\boldsymbol{X}\right) + \frac{1}{\sigma_\beta^2}\text{diag}\left(\boldsymbol{f}_{2\beta}\right)\right)^{-1}\Bigg), \tag{3.28}$$

$$f_{2\beta j}\big|\text{rest} \sim \text{Gamma}\left(1, \frac{\beta_j^2}{2\sigma_\beta^2} + f_{1\beta j}\right), \;\; 1 \leq j \leq d_\circ + d_\bullet, \tag{3.29}$$

$$f_{1\beta j}\big|\text{rest} \sim \text{Gamma}\left(1, f_{2\beta j} + 1\right), \;\; 1 \leq j \leq d_\circ + d_\bullet, \tag{3.30}$$

$$\sigma_\beta^2\big|\text{rest} \sim \text{Inverse-Gamma}\left(\frac{d_\circ + d_\bullet + 1}{2}, \frac{\boldsymbol{\beta}^T \text{diag}(\boldsymbol{f}_{2\beta})\boldsymbol{\beta}}{2} + \frac{1}{a_\beta}\right), \tag{3.31}$$

$$a_\beta\big|\text{rest} \sim \text{Inverse-Gamma}\left(1, \frac{1}{\sigma_\beta^2} + \frac{1}{s_\beta^2}\right), \tag{3.32}$$

$$\begin{aligned}\boldsymbol{u}_j\big|\text{rest} \sim N\Bigg(&\text{diag}\left(\frac{\boldsymbol{\omega}_{\boldsymbol{Z}_j}}{\sigma_\varepsilon^2} + \frac{f_{2uj}\mathbf{1}}{\sigma_{uj}^2}\right)^{-1}\left(\frac{1}{\sigma_\varepsilon^2}\right)\Bigg[\Big(\boldsymbol{Z}_j^T\boldsymbol{y} - \left(\boldsymbol{Z}_j^T\boldsymbol{X}\right)\boldsymbol{\beta} \\ &- \sum_{j' \neq j}^{d_\bullet}\left(\boldsymbol{Z}_j^T\boldsymbol{Z}_{j'}\right)\boldsymbol{u}_{j'}\Big)\Bigg], \text{diag}\left(\frac{\boldsymbol{\omega}_{\boldsymbol{Z}_j}}{\sigma_\varepsilon^2} + \frac{f_{2uj}\mathbf{1}}{\sigma_{uj}^2}\right)^{-1}\Bigg), \;\; 1 \leq j \leq d_\bullet,\end{aligned} \tag{3.33}$$

$$f_{2uj}\big|\text{rest} \sim \text{Gamma}\left(\frac{K_j + 1}{2}, \frac{\left\|\boldsymbol{u}_j\right\|^2}{2\sigma_{uj}^2} + f_{1uj}\right), \;\; 1 \leq j \leq d_\bullet, \tag{3.34}$$

$$f_{1uj}\big|\text{rest} \sim \text{Gamma}\left(1, f_{2uj} + 1\right), \;\; 1 \leq j \leq d_\bullet, \tag{3.35}$$

$$\sigma_{uj}^2\big|\text{rest} \sim \text{Inverse-Gamma}\left(\frac{K_j + 1}{2}, \frac{f_{2uj}\left\|\boldsymbol{u}_j\right\|^2}{2} + \frac{1}{a_{uj}}\right), \quad 1 \leq j \leq d_\bullet, \tag{3.36}$$

$$a_{uj}\big|\text{rest} \sim \text{Inverse-Gamma}\left(1, \frac{1}{\sigma_{uj}^2} + \frac{1}{s_u^2}\right), \quad 1 \leq j \leq d_\bullet, \tag{3.37}$$

$$\sigma_\varepsilon^2\big|\text{rest} \sim \text{Inverse-Gamma}\left(\frac{n+1}{2}, \frac{1}{a_\varepsilon} + \frac{1}{2}\left\|\boldsymbol{y} - \mathbf{1}_n\beta_0 - \boldsymbol{X}\boldsymbol{\beta} - \sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j\boldsymbol{u}_j\right\|^2\right) \tag{3.38}$$

and

$$a_\varepsilon\big|\text{rest} \sim \text{Inverse-Gamma}\left(1, \frac{1}{\sigma_\varepsilon^2} + \frac{1}{s_\varepsilon^2}\right). \tag{3.39}$$

For Poisson responses, based on full model description in (3.24) in Section 3.5.3 and Figure 3.3, for Negative-Binomial responses, based on full model description in (3.25) in Section 3.5.4 and Figure 3.4, using Markov blanket, we obtain full con-

ditional distributions of parameters or full conditional expressions of parameters in the form of (3.26) for logarithm of unnormalised density function.

For Poisson responses, full conditional distributions of parameters listed in (3.40)–(3.50), with derivation details given in Appendix 3.G. Based on these, Algorithm 5 gives detailed steps of sample drawing for Markov chain Monte Carlo, using function sliceSampling described in Algorithm 3.

$$\mathfrak{p}(\beta_0|\text{rest}) \propto \exp\left\{ n\bar{\boldsymbol{y}} - \frac{\beta_0^2}{2\sigma_{\beta_0}^2} - \mathbf{1}_n^T \exp\left(\mathbf{1}_n\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j\boldsymbol{u}_j \right) \right\}, \tag{3.40}$$

$$\mathfrak{p}(\beta_j|\text{rest}) \propto \exp\Bigg\{ \mathbf{1}_n^T \left(\boldsymbol{y} \odot \boldsymbol{X}_j\right)\beta_j - \frac{f_{2\beta_j}\beta_j^2}{2\sigma_\beta^2} - \mathbf{1}_n^T \exp\Bigg(\boldsymbol{X}_j\beta_j + \mathbf{1}_n\beta_0 + (\boldsymbol{X})_{-j}(\boldsymbol{\beta})_{-j} + \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j\boldsymbol{u}_j \Bigg) \Bigg\}, \tag{3.41}$$

$$f_{2\beta_j}|\text{rest} \overset{\text{ind.}}{\sim} \text{Gamma}\left(1, \frac{\beta_j^2}{2\sigma_\beta^2} + f_{1\beta_j}\right), \quad 1 \le j \le d_\circ + d_\bullet, \tag{3.42}$$

$$f_{1\beta_j}|\text{rest} \overset{\text{ind.}}{\sim} \text{Gamma}\left(1, 1 + f_{2\beta_j}\right), \quad 1 \le j \le d_\circ + d_\bullet, \tag{3.43}$$

$$\sigma_\beta^2|\text{rest} \sim \text{Inverse-Gamma}\left(\frac{d_\circ + d_\bullet + 1}{2}, \frac{\boldsymbol{\beta}^T \text{diag}(\boldsymbol{f}_{2\beta})\boldsymbol{\beta}}{2} + \frac{1}{a_\beta} \right), \tag{3.44}$$

$$a_\beta|\text{rest} \sim \text{Inverse-Gamma}\left(1, \frac{1}{\sigma_\beta^2} + \frac{1}{s_\beta^2} \right), \tag{3.45}$$

$$\mathfrak{p}(u_{jk}|\text{rest}) \propto \exp\Bigg\{ \mathbf{1}_n^T \left(\boldsymbol{y} \odot \boldsymbol{Z}_{jk}\right) u_{jk} - \frac{f_{2uj}u_{jk}^2}{2\sigma_{uj}^2} - \mathbf{1}_n^T \exp\Bigg(\boldsymbol{Z}_{jk}u_{jk} + \mathbf{1}_n\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \sum_{k' \ne k}^{K_j} \boldsymbol{Z}_{jk'}u_{jk'} + \sum_{j' \ne j}^{d_\bullet} \boldsymbol{Z}_{j'}\boldsymbol{u}_{j'} \Bigg) \Bigg\}, \quad 1 \le j \le d_\bullet, \ 1 \le k \le K_j, \tag{3.46}$$

$$f_{2uj}|\text{rest} \overset{\text{ind.}}{\sim} \text{Gamma}\left(\frac{K_j + 1}{2}, \frac{\left\|\boldsymbol{u}_j\right\|^2}{2\sigma_{uj}^2} + f_{1uj} \right), \quad 1 \le j \le d_\bullet, \tag{3.47}$$

$$f_{1uj}\big|\text{rest} \overset{\text{ind.}}{\sim} \text{Gamma}\left(1, 1+f_{2uj}\right), \;\; 1 \leq j \leq d_\bullet, \tag{3.48}$$

$$\sigma_{uj}^2\big|\text{rest} \sim \text{Inverse-Gamma}\left(\frac{K_j+1}{2}, \frac{f_{2uj}\left\|\widetilde{\boldsymbol{u}}_j\right\|^2}{2}+\frac{1}{a_{uj}}\right), \;\; 1 \leq j \leq d_\bullet \tag{3.49}$$

and

$$a_{uj}\big|\text{rest} \sim \text{Inverse-Gamma}\left(1, \frac{1}{\sigma_{uj}^2}+\frac{1}{s_u^2}\right), \;\; 1 \leq j \leq d_\bullet. \tag{3.50}$$

For Negative-Binomial responses, full conditional distributions of parameters listed in (3.51)–(3.63), with derivation details given in Appendix 3.H. Based on these, Algorithm 6 gives detailed steps of sample drawing for Markov chain Monte Carlo, using function sliceSampling described in Algorithm 3.

To ensure auxiliary parameter κ being positive, we use $\kappa = \exp(\lambda)$ for sample drawing with slice sampling method, using function logUnnDensLambda to obtain logarithm of unnormalised density function, described in Algorithm 6.

$$\boldsymbol{g}\big|\text{rest} \sim \text{Gamma}\left(\boldsymbol{y}+\exp\left(\lambda\right), 1+\exp\left\{\lambda-\left(\mathbf{1}_n\beta_0+\boldsymbol{X}\boldsymbol{\beta}+\sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j\boldsymbol{u}_j\right)\right\}\right), \tag{3.51}$$

$$\begin{aligned}\mathfrak{p}(\beta_0\big|\text{rest}) \propto \exp\Bigg\{&-n\kappa\beta_0-\frac{\beta_0^2}{2\sigma_{\beta_0}^2}-\mathbf{1}_n^T\exp\Big[-\mathbf{1}_n\beta_0+\mathbf{1}_n\log\left(\kappa\right)+\log\left(\boldsymbol{g}\right)\\ &-\left(\boldsymbol{X}\boldsymbol{\beta}+\sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j\boldsymbol{u}_j\right)\Big]\Bigg\},\end{aligned} \tag{3.52}$$

$$\begin{aligned}\log\left\{\mathfrak{p}(\lambda\big|\text{rest})\right\} = \lambda+n\left(\kappa\log\left(\kappa\right)-\log\left(\Gamma\left(\kappa\right)\right)\right)-\kappa\Bigg[&\sum_{i=1}^{n}\left(\beta_0+\boldsymbol{X}_i\boldsymbol{\beta}+\sum_{j=1}^{d_\bullet}\boldsymbol{Z}_{ij}\boldsymbol{u}_j\right)\\ &+b_\kappa+\sum_{i=1}^{n}\left(g_i\exp\left\{-\left(\beta_0+\boldsymbol{X}_i\boldsymbol{\beta}+\sum_{j=1}^{d_\bullet}\boldsymbol{Z}_{ij}\boldsymbol{u}_j\right)\right\}\right)-\sum_{i=1}^{n}\log\left(g_i\right)\Bigg]+\text{const},\end{aligned} \tag{3.53}$$

$$\mathfrak{p}(\beta_j\big|\text{rest}) \propto \exp\Bigg\{\left(-\kappa\sum_{i=1}^{n}X_{ij}\right)\beta_j-\frac{f_{2\beta_j}\beta_j^2}{2\sigma_\beta^2}-\mathbf{1}_n^T\exp\Big[-\boldsymbol{X}_j\beta_j+\mathbf{1}_n\log\left(\kappa\right)+\log\left(\boldsymbol{g}\right)$$

$$- \left(\mathbf{1}_n\beta_0 + (\boldsymbol{X})_{-j}(\boldsymbol{\beta})_{-j} + \sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j\boldsymbol{u}_j\right)\Bigg]\Bigg\}, \;\; 1 \leq j \leq d_\circ + d_\bullet, \quad (3.54)$$

$$f_{2\beta_j}\big|\text{rest} \overset{\text{ind.}}{\sim} \text{Gamma}\left(1, \frac{\beta_j^2}{2\sigma_\beta^2} + f_{1\beta_j}\right), \;\; 1 \leq j \leq d_\circ + d_\bullet, \quad (3.55)$$

$$f_{1\beta_j}\big|\text{rest} \overset{\text{ind.}}{\sim} \text{Gamma}\left(1, 1 + f_{2\beta_j}\right), \;\; 1 \leq j \leq d_\circ + d_\bullet, \quad (3.56)$$

$$\sigma_\beta^2\big|\text{rest} \sim \text{Inverse-Gamma}\left(\frac{d_\circ + d_\bullet + 1}{2}, \frac{\boldsymbol{\beta}^T\text{diag}(\boldsymbol{f}_{2\beta})\boldsymbol{\beta}}{2} + \frac{1}{a_\beta}\right), \quad (3.57)$$

$$a_\beta\big|\text{rest} \sim \text{Inverse-Gamma}\left(1, \frac{1}{\sigma_\beta^2} + \frac{1}{s_\beta^2}\right), \quad (3.58)$$

$$\mathfrak{p}(u_{jk}\big|\text{rest}) \propto \exp\Bigg\{\left(-\kappa\mathbf{1}_n^T\boldsymbol{Z}_{jk}\right)u_{jk} - \frac{f_{2uj}u_{jk}^2}{2\sigma_{uj}^2} - \mathbf{1}_n^T\exp\Bigg[-\boldsymbol{Z}_{jk}u_{jk} + \mathbf{1}_n\log(\kappa) + \log(\boldsymbol{g})$$

$$- \left(\mathbf{1}_n\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \sum_{k'\neq k}^{K_j}Z_{jk'}u_{jk'} + \sum_{j'\neq j}^{d_\bullet}\boldsymbol{Z}_{j'}\boldsymbol{u}_{j'}\right)\Bigg]\Bigg\}, \;\; 1 \leq j \leq d_\bullet, \;\; 1 \leq k \leq K_j, \quad (3.59)$$

$$f_{2uj}\big|\text{rest} \overset{\text{ind.}}{\sim} \text{Gamma}\left(\frac{K_j + 1}{2}, \frac{\left\|\boldsymbol{u}_j\right\|^2}{2\sigma_{uj}^2} + f_{1uj}\right), \;\; 1 \leq j \leq d_\bullet, \quad (3.60)$$

$$f_{1uj}\big|\text{rest} \overset{\text{ind.}}{\sim} \text{Gamma}\left(1, 1 + f_{2uj}\right), \;\; 1 \leq j \leq d_\bullet, \quad (3.61)$$

$$\sigma_{uj}^2\big|\text{rest} \sim \text{Inverse-Gamma}\left(\frac{K_j + 1}{2}, \frac{f_{2uj}\left\|\widetilde{\boldsymbol{u}}_j\right\|^2}{2} + \frac{1}{a_{uj}}\right), 1 \leq j \leq d_\bullet \quad (3.62)$$

and

$$a_{uj}\big|\text{rest} \sim \text{Inverse-Gamma}\left(1, \frac{1}{\sigma_{uj}^2} + \frac{1}{s_u^2}\right), \;\; 1 \leq j \leq d_\bullet. \quad (3.63)$$

3.7 Model Selection Strategies with Horseshoe-type Priors

After obtaining the full conditional distributions of parameters described in Section 3.6.3, we need to set up the model selection strategies to decide the effect type of a predictor. Specifically, whether it is zero, linear or non-linear effect.

Recalling the results obtained in Section 3.2.6, which is the derived posterior mean in a Bayesian model, using Grouped Horseshoe distribution, in this section, we obtained the expression of posterior mean $E\left(\gamma_{\text{GHS}}\middle|\boldsymbol{y}\right)$ and selection rules for Gaussian responses of model (3.23) stated in (3.74), for Poisson responses of model (3.24) and Negative-Binomial responses of model (3.25), both stated in (3.75).

3.7.1 Overview

From **Result 3.2.6.1** in Section 3.2.6, we have

$$\begin{aligned} E\left(\boldsymbol{\psi}\middle|\boldsymbol{y}\right) &= \boldsymbol{y}E\left(\frac{\lambda^2\tau_2^2}{\tau_1^2+\lambda^2\tau_2^2}\middle|\boldsymbol{y}\right) = \boldsymbol{y}E\left(\left(1-\frac{\tau_1^2}{\tau_1^2+\lambda^2\tau_2^2}\right)\middle|\boldsymbol{y}\right) \\ &= \boldsymbol{y}\left(1-E\left(\kappa\middle|\boldsymbol{y}\right)\right) = \boldsymbol{y}\left(1-\widehat{\kappa}_{\text{Bayes}}\right). \end{aligned}$$

If we define

$$\kappa \equiv \frac{\tau_1^2}{\tau_1^2+\lambda^2\tau_2^2} \quad \text{and} \quad E\left(\kappa\middle|\boldsymbol{y}\right) \equiv \widehat{\kappa}_{\text{Bayes}},$$

we then have

$$1-\widehat{\kappa}_{\text{Bayes}} = \frac{E\left(\boldsymbol{\psi}\middle|\boldsymbol{y}\right)}{\boldsymbol{y}}. \tag{3.64}$$

(3.64) matches equation (8) of Bhadra et al. (2019), for scalar coefficient case. The Bayes estimator of κ with respect to squared error loss is denoted as $\widehat{\kappa}_{\text{Bayes}}$. The shrinkage weight is denoted as $1-\widehat{\kappa}_{\text{Bayes}}$. The Bayes estimator of κ and shrinkage weight $1-\widehat{\kappa}_{\text{Bayes}}$ are as follows respectively:

$$\widehat{\kappa}_{\text{Bayes}} \equiv E\left(\kappa\middle|\boldsymbol{y}\right) = \frac{\boldsymbol{y}-E\left(\boldsymbol{\psi}\middle|\boldsymbol{y}\right)}{\boldsymbol{y}} \quad \text{and} \quad 1-\widehat{\kappa}_{\text{Bayes}} = \frac{E\left(\boldsymbol{\psi}\middle|\boldsymbol{y}\right)}{\boldsymbol{y}} = E\left(\gamma_{\text{GHS}}\middle|\boldsymbol{y}\right) \tag{3.65}$$

where γ_{GHS} is defined in (3.8). The (Grouped) Horseshoe estimator $\widehat{\kappa}_{\text{Bayes}}$ is a shrinkage estimator, with the relative amount of shrinkage that equals $E\left(\kappa\middle|\boldsymbol{y}\right)$. The greater value of the shrinkage parameter $\widehat{\kappa}_{\text{Bayes}}$, the greater shrinkage power it has. The Horseshoe shrinkage estimator $E\left(\kappa\middle|\boldsymbol{y}\right)$ for the signal is small, and the Horseshoe shrinkage estimator $E\left(\kappa\middle|\boldsymbol{y}\right)$ for the noise is big. Overall, the shrinkage weight $E\left(\boldsymbol{\psi}\middle|\boldsymbol{y}\right)/\boldsymbol{y}$ for the signal is big, and the shrinkage weight $E\left(\boldsymbol{\psi}\middle|\boldsymbol{y}\right)/\boldsymbol{y}$ for the noise is small.

The general selection strategy is based on the zeroing of a coefficient θ corresponds to the shrinkage weight $E\left(\gamma_{\text{GHS}}\middle|\boldsymbol{y}\right)$ below a classification border.

For Gaussian response, according to (3.9), with the threshold parameter τ being 0.5, we have the following rules of zeroing of θ :

$$\begin{aligned} &\text{if } E\left(\gamma_{\text{GHS}}\middle|\boldsymbol{y}\right) \leq \frac{1}{2}, \text{ then } \theta = 0, \\ &\text{if } E\left(\gamma_{\text{GHS}}\middle|\boldsymbol{y}\right) > \frac{1}{2}, \text{ then } \theta \neq 0. \end{aligned} \tag{3.66}$$

For Poisson responses in model (3.24) and Negative-Binomial responses in model (3.25), according to (3.8), with the threshold parameter τ decided by simulation study, we have the following rules of zeroing of θ :

$$\begin{aligned} &\text{if } E\left(\gamma_{\text{GHS}}\middle|\boldsymbol{y}\right) \leq 1-\tau, \text{ then } \theta = 0, \\ &\text{if } E\left(\gamma_{\text{GHS}}\middle|\boldsymbol{y}\right) > 1-\tau, \text{ then } \theta \neq 0. \end{aligned} \tag{3.67}$$

From (3.8), we then have the expressions for γ_{GHS} as follows:

$$\begin{aligned} \gamma_{\text{GHS}} &= \frac{\lambda^2\tau_2^2}{\tau_1^2+\lambda^2\tau_2^2} \text{ for Gaussian responses in model } (3.23), \text{ and} \\ \gamma_{\text{GHS}} &= \frac{\lambda^2\tau_2^2}{1+\lambda^2\tau_2^2} \text{ for Poisson responses and Negative-Binomial responses} \\ &\qquad \text{in model } (3.24) \text{ and } (3.25). \end{aligned} \tag{3.68}$$

From the first case in (3.68), for Gaussian responses using (Grouped) Horseshoe priors in model (3.23), we have the following expressions for γ_{GHS} for linear coeffi-

cients and spline coefficients vector respectively:

$$\begin{aligned}\lambda^2_{\beta_j} &= \frac{1}{f_{2\beta_j}},\ \ \gamma_{\beta_j,\text{HS}} = \frac{\lambda^2_{\beta_j}\sigma^2_\beta}{\sigma^2_\varepsilon + \lambda^2_{\beta_j}\sigma^2_\beta} = \frac{\sigma^2_\beta}{f_{2\beta_j}\sigma^2_\varepsilon + \sigma^2_\beta},\ \ 1 \le j \le d_\circ + d_\bullet, \\ \lambda^2_{uj} &= \frac{1}{f_{2uj}},\ \ \gamma_{u_j,\text{GHS}} = \frac{\lambda^2_{u_j}\sigma^2_{uj}}{\sigma^2_\varepsilon + \lambda^2_{u_j}\sigma^2_{uj}} = \frac{\sigma^2_{uj}}{f_{2u_j}\sigma^2_\varepsilon + \sigma^2_{uj}},\ \ 1 \le j \le d_\bullet.\end{aligned} \tag{3.69}$$

From the second case in (3.68), for Poisson responses in model (3.24) and Negative-Binomial responses in model (3.25), using (Grouped) Horseshoe priors, we have the following expressions for γ_{GHS} for linear coefficients and spline coefficients vector respectively:

$$\begin{aligned}\lambda^2_{\beta_j} &= \frac{1}{f_{2\beta_j}},\ \ \gamma_{\beta_j,\text{HS}} = \frac{\lambda^2_{\beta_j}\sigma^2_\beta}{1 + \lambda^2_{\beta_j}\sigma^2_\beta} = \frac{\sigma^2_\beta}{f_{2\beta_j} + \sigma^2_\beta},\ \ 1 \le j \le d_\circ + d_\bullet, \\ \lambda^2_{uj} &= \frac{1}{f_{2uj}},\ \ \gamma_{u_j,\text{GHS}} = \frac{\lambda^2_{u_j}\sigma^2_{uj}}{1 + \lambda^2_{u_j}\sigma^2_{uj}} = \frac{\sigma^2_{uj}}{f_{2u_j} + \sigma^2_{uj}},\ \ 1 \le j \le d_\bullet.\end{aligned} \tag{3.70}$$

Overall, the expression of posterior mean $E\left(\gamma_{\text{GHS}}\middle|\boldsymbol{y}\right)$ for Gaussian responses stated in (3.69) and for count responses stated in (3.70) are very similar. The only difference is that the σ^2_ε term only appears in the Gaussian response case. The expression of $E\left(\gamma_{\text{GHS}}\middle|\boldsymbol{y}\right)$ is used in Section 3.7.2–3.10 to explore the application of the (Grouped) Horseshoe priors via MCMC for the Bayesian three-category GAM selection on Gaussian responses and count responses.

3.7.2 Deciding Between an Effect Being Zero or Linear

Let $\boldsymbol{\beta}$ being a $(d_\circ + d_\bullet) \times 1$ vector of linear effects coefficients, described in Section 3.4.4. $d_\circ$ is the number of candidate predictors that may have a zero or linear effect and $d_\bullet$ is the number of candidate predictors that may have a zero, linear or non-linear effect. Based on (3.66), our "effect is zero" rule between zero effect and linear effect is

$$\begin{aligned}&\text{the effect is zero if } E\left(\gamma_{\text{GHS}}\middle|\boldsymbol{y}\right) \le 1 - \tau,\ \ 1 \le j \le d_\circ + d_\bullet, \\ &\text{otherwise the effect is linear.}\end{aligned}$$

For Gaussian responses in model (3.23), using Horseshoe priors, from (3.66) and

the first case in (3.69), our "effect is zero" rule between zero effect and linear effect is given as

$$\begin{aligned} &\gamma_{\beta_j,\text{HS}} = \frac{\sigma_\beta^2}{f_{2\beta_j}\sigma_\varepsilon^2 + \sigma_\beta^2}, \;\; 1 \leq j \leq d_\circ + d_\bullet, \\ &\text{the effect is zero if } E(\gamma_{\beta j,\text{HS}}\big|\boldsymbol{y}) \leq \frac{1}{2}, \;\; 1 \leq j \leq d_\circ + d_\bullet, \\ &\text{otherwise the effect is linear.} \end{aligned} \tag{3.71}$$

For Poisson responses in model (3.24) and Negative-Binomial responses in model (3.25), both using Horseshoe priors, from (3.67) and second case in (3.70), our "effect is zero" rule between zero effect and linear effect is given as

$$\begin{aligned} &\gamma_{\beta_j,\text{HS}} = \frac{\sigma_\beta^2}{f_{2\beta_j} + \sigma_\beta^2}, \;\; 1 \leq j \leq d_\circ + d_\bullet, \\ &\text{the effect is zero if } E(\gamma_{\beta j,\text{HS}}\big|\boldsymbol{y}) \leq 1 - \tau, \;\; 1 \leq j \leq d_\circ + d_\bullet, \\ &\text{otherwise the effect is linear.} \end{aligned} \tag{3.72}$$

3.7.3 Deciding Between an Effect Being Zero, Linear or Non-linear

Below notations are used in this subsection:

- $d_\circ \equiv$ the number of candidate predictors that may have a zero or linear effect.
- $d_\bullet \equiv$ the number of candidate predictors that may have a zero, linear or non-linear effect.
- $\boldsymbol{\beta} \equiv$ a $(d_\circ + d_\bullet) \times 1$ linear coefficient vector, as described in Section 3.4.4.
- $\boldsymbol{u}_j \equiv$ a $K_j \times 1$ spline coefficients vector $1 \leq j \leq d_\bullet$, as described in Section 3.4.5.

Based on (3.67), our "effect is zero" rule between zero, linear or non-linear effect is given as

$$\text{the effect is zero if } \begin{cases} E(\gamma_{\beta j,\text{HS}}|\boldsymbol{y}) \leq 1-\tau, \;\; 1 \leq j \leq d_\circ + d_\bullet, \text{ and} \\ E(\gamma_{u_j,\text{GHS}}|\boldsymbol{y}) \leq 1-\tau, \;\; 1 \leq j \leq d_\bullet, \end{cases}$$

$$\text{the effect is linear if } \begin{cases} E(\gamma_{\beta j,\text{HS}}|\boldsymbol{y}) > 1-\tau, \;\; 1 \leq j \leq d_\circ + d_\bullet, \text{ and} \\ E(\gamma_{u_j,\text{GHS}}|\boldsymbol{y}) \leq 1-\tau, \;\; 1 \leq j \leq d_\bullet, \end{cases} \tag{3.73}$$

otherwise the effect is non-linear,

For the Gaussian responses in model (3.23) using (Grouped) Horseshoe priors, from (3.66), (3.69) and (3.73), our strategy of deciding between an effect being zero, linear or non-linear is given as:

$$\gamma_{\beta_j,\text{HS}} = \frac{\sigma_\beta^2}{f_{2\beta_j}\sigma_\varepsilon^2 + \sigma_\beta^2}, \;\; 1 \leq j \leq d_\circ + d_\bullet,$$

$$\gamma_{uj,\text{GHS}} = \frac{\sigma_{uj}^2}{f_{2uj}\sigma_\varepsilon^2 + \sigma_{uj}^2}, \;\; 1 \leq j \leq d_\bullet,$$

$$\text{the effect is zero if } \begin{cases} E(\gamma_{\beta j,\text{HS}}|\boldsymbol{y}) \leq \dfrac{1}{2}, \;\; 1 \leq j \leq d_\circ + d_\bullet, \text{ and} \\ E(\gamma_{uj,\text{GHS}}|\boldsymbol{y}) \leq \dfrac{1}{2}, \;\; 1 \leq j \leq d_\bullet, \end{cases} \tag{3.74}$$

$$\text{the effect is linear if } \begin{cases} E(\gamma_{\beta j,\text{HS}}|\boldsymbol{y}) > \dfrac{1}{2}, \;\; 1 \leq j \leq d_\circ + d_\bullet, \text{ and} \\ E(\gamma_{uj,\text{GHS}}|\boldsymbol{y}) \leq \dfrac{1}{2}, \;\; 1 \leq j \leq d_\bullet, \end{cases}$$

otherwise the effect is non-linear.

For the Poisson responses in model (3.24) and the Negative-Binomial responses in model (3.25), both using (Grouped) Horseshoe priors, from (3.70) and (3.73), our

strategy of deciding between an effect being zero, linear or non-linear is given as:

$$\gamma_{\beta_j,\text{HS}} = \frac{\sigma_\beta^2}{f_{2\beta_j} + \sigma_\beta^2}, \;\; 1 \leq j \leq d_\circ + d_\bullet,$$

$$\gamma_{uj,\text{GHS}} = \frac{\sigma_{uj}^2}{f_{2uj} + \sigma_{uj}^2}, \;\; 1 \leq j \leq d_\bullet,$$

$$\text{the effect is zero if } \begin{cases} E(\gamma_{\beta j,\text{HS}}|\boldsymbol{y}) \leq 1-\tau, \;\; 1 \leq j \leq d_\circ + d_\bullet, \text{ and} \\ E(\gamma_{uj,\text{GHS}}|\boldsymbol{y}) \leq 1-\tau, \;\; 1 \leq j \leq d_\bullet, \end{cases} \tag{3.75}$$

$$\text{the effect is linear if } \begin{cases} E(\gamma_{\beta j,\text{HS}}|\boldsymbol{y}) > 1-\tau, \;\; 1 \leq j \leq d_\circ + d_\bullet, \text{ and} \\ E(\gamma_{uj,\text{GHS}}|\boldsymbol{y}) \leq 1-\tau, \;\; 1 \leq j \leq d_\bullet, \end{cases}$$

otherwise the effect is non-linear.

3.8 Comparative Performance of Gaussian Responses

In this subsection, we compare the classification performance on Gaussian responses in model (3.23), via the Markov chain Monte Carlo (MCMC) approach, between use of the Laplace-Zero/Grouped Lasso-Zero prior described in Chapter 2 as Algorithm 2, and use of the (Grouped) Horseshoe priors described in Section 3.6.3 as Algorithm 4.

3.8.1 Classification Between Linear Effects and Non-linear Effects

To better illustrate the effect of (3.9), we ran three simulation studies on a generalized additive model selection, with the simulation settings stated in (2.56), (2.57) and (3.76) given as follows:

$$\begin{aligned} &\text{30 predictors;} \\ &\text{10 predictors having zero effect, for } j \in \{1, 2, \ldots, 10\} \\ &\text{10 predictors having linear effect, for } j \in \{11, 12, \ldots, 20\} \\ &\text{10 predictors having non-linear effect, for } j \in \{21, 22, \ldots, 30\}\,. \end{aligned} \tag{3.76}$$

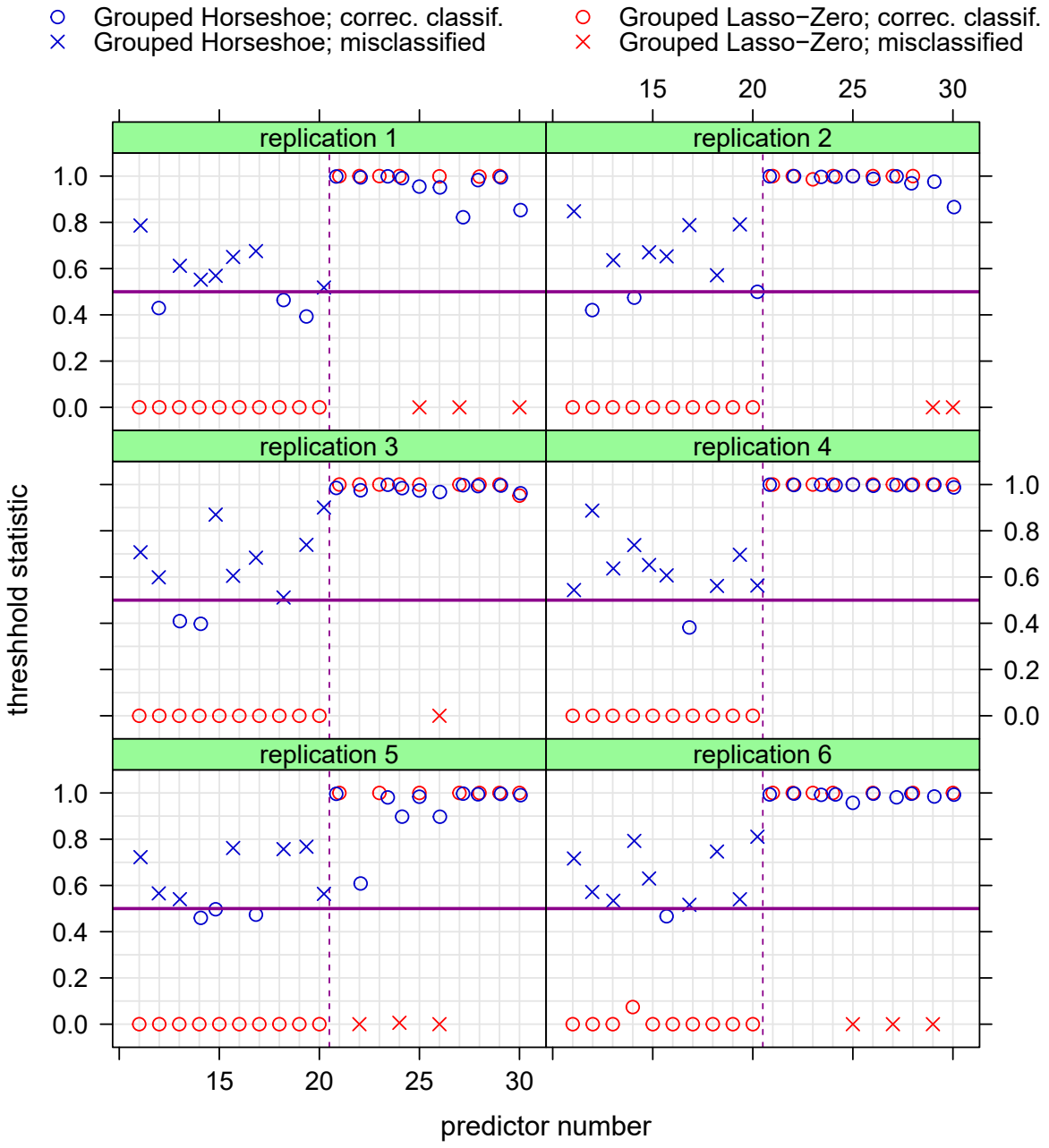


Figure 3.5: *The results of the linear versus non-linear effect classification for six replications of the simulation study, described in Section 4 of He and Wand (2024), with* $n = 500$ *and* $\sigma_\varepsilon = 2$. *The blue symbols correspond to the Markov chain Monte Carlo-approximate values* $E\left(\gamma_{uj,GHS}\middle|\boldsymbol{y}\right)$ *for* $11 \leq j \leq 30$, *which are the predictors that are simulated to have linear effects (left of the vertical dashed line) and non-linear effects (right of the vertical dashed line). A blue circle denotes the correct classification using (3.74), whilst a blue cross indicates a misclassification. The red symbols are similar, but for the* $E\left(\gamma_{uj}\middle|\boldsymbol{y}\right)$ *statistics corresponding to strategy described in (2.54). The classification border of* $\frac{1}{2}$ *is shown by the horizontal purple line.*

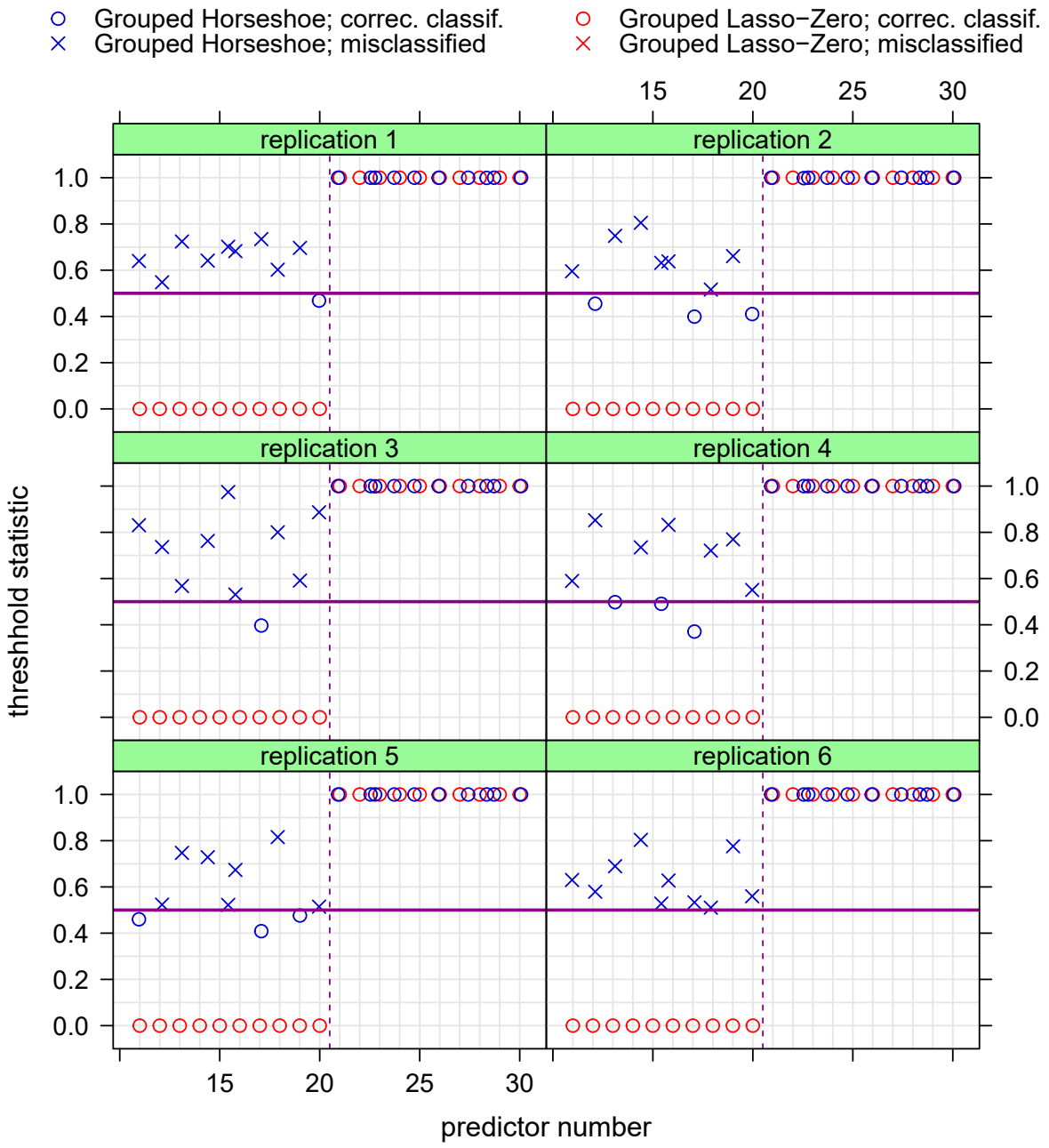


Figure 3.6: *Similar to Figure 3.5 but with a higher sample size,* $n = 2,000$*, and lower error standard deviation,* $\sigma_\varepsilon = 0.25$.

We ran three simulation study to compare the linear effect versus non-linear effect classification performance on Gaussian responses, between the use of (Grouped Horseshoe) priors in model (3.23) with selection rules on $E\left(\gamma_{u_j,\text{GHS}}\middle|\boldsymbol{y}\right)$ stated in (3.74) and the use of Grouped Lasso-Zero prior in model (2.28) with selection rules on $E\left(\gamma_{u_j}\middle|\boldsymbol{y}\right)$ stated in (2.54). We used 20 predictors generated from (3.76), 10 having

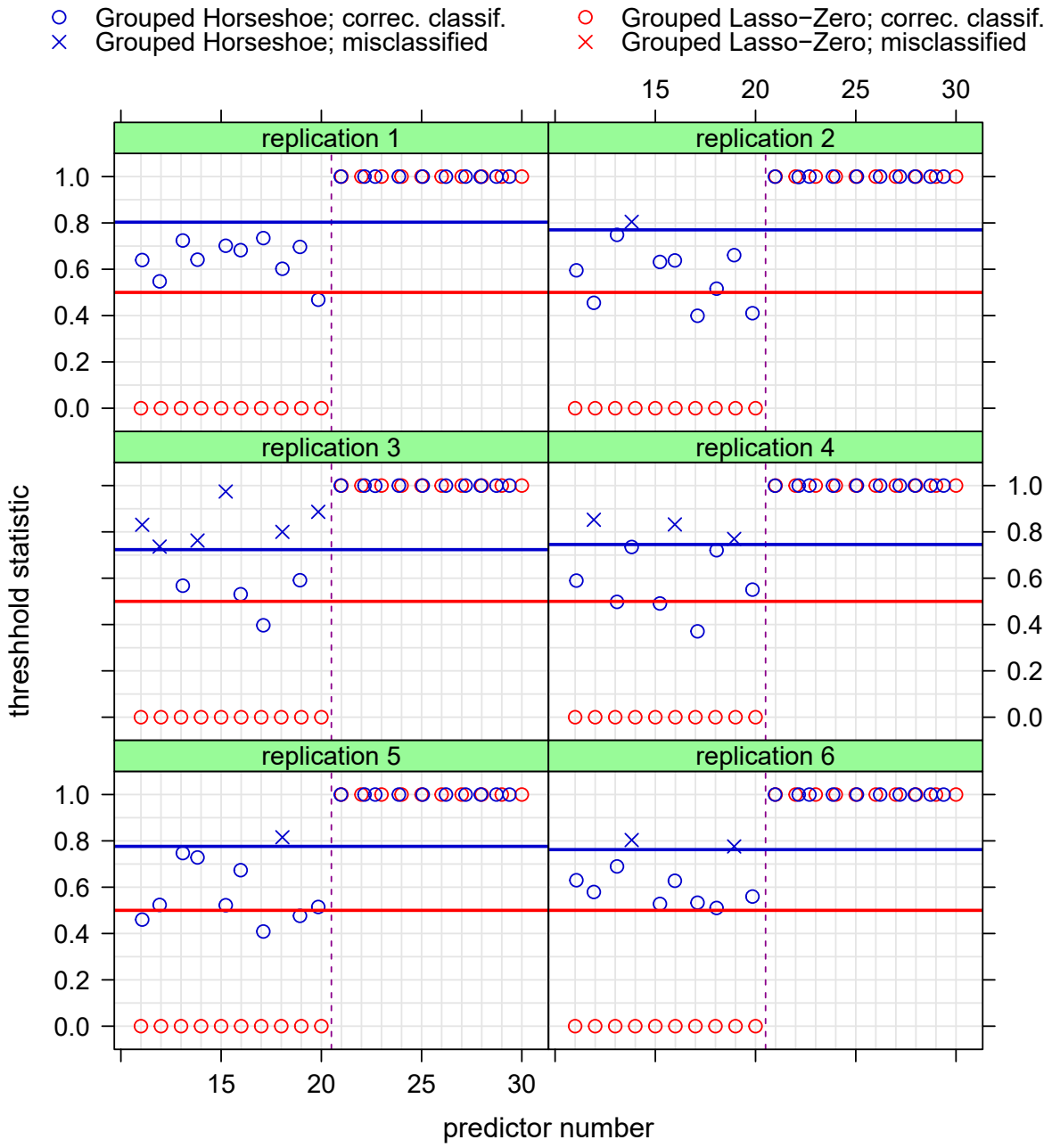


Figure 3.7: *Similar to Figure 3.6 but with the thresholding of the Grouped Horseshoe* $E\left(\gamma_{u_j,GHS}\middle|\boldsymbol{y}\right)$ *statistics based on k-means clustering. For each replication, the horizontal blue line shows the classification border arsing from* $k-means$ *clustering. The horizontal red line at* $\frac{1}{2}$ *is the threshold for the Grouped Lasso-Zero statistics.*

linear effect and 10 having non-linear effect, which implies $11 \leq j \leq 30$. The choice of the threshold parameter $\tau = \frac{1}{2}$ is on both of the rules (3.74) and (2.54). In short, for

the use of the Grouped Horseshoe priors, $11 \leq j \leq 30,$ the classification rule is:

$$\begin{aligned} &\text{the effect is linear if } E\left(\gamma_{u_j,\text{GHS}}\middle|\boldsymbol{y}\right) \leq \tfrac{1}{2}, \\ &\text{the effect is non-linear if } E\left(\gamma_{u_j,\text{GHS}}\middle|\boldsymbol{y}\right) > \tfrac{1}{2}. \end{aligned} \tag{3.77}$$

For the use of the Grouped Lasso-Zero priors, $11 \leq j \leq 30,$ the classification rule is:

$$\begin{aligned} &\text{the effect is linear if } E\left(\gamma_{u_j}\middle|\boldsymbol{y}\right) \leq \tfrac{1}{2} \\ &\text{the effect is non-linear if } E\left(\gamma_{u_j}\middle|\boldsymbol{y}\right) > \tfrac{1}{2}. \end{aligned} \tag{3.78}$$

In Figures 3.5, 3.6 and 3.7, the horizontal axis corresponds to $j = 11, 12, \ldots, 30,$ which is concerned with linear versus non-linear effect classification for six replications in the simulation studies. Among these 20 predictors, the first ten predictors are simulated to have a linear effect, corresponding to the left of the vertical dashed line. The last ten predictors are simulated to have a non-linear effect, corresponding to the right of the vertical dashed line. The classification border of 0.5 is shown by the horizontal purple line. The vertical axis corresponds to the threshold statistic. Correct classifications are shown as circles and incorrect classifications are shown as crosses. The blue circles and blue crosses correspond to the use of Grouped Horseshoe prior. The red circles and red crosses correspond to the use of Grouped Lasso-Zero prior. For each replication, we have 20 blue symbols correspond to 20 candidate predictors using the Grouped Horseshoe prior, and we have 20 red symbols correspond to 20 candidate predictors using the Grouped Lasso-Zero prior. In total, for 6 replications in each figure among Figure 3.5 – 3.7, we have 120 blue symbols and 120 red symbols. From the rules stated in (3.77) and (3.78), for the classification of the linear effect, the threshold statistics are below the classification border of 0.5, as shown in the bottom left part in each replication. For the classification of the non-linear effect, the threshold statistics are above the classification border of 0.5, as shown in the top right part in each replication.

For the simulation study in Figure 3.5, we have the sample size being $n = 500$ and the error standard deviation being $\sigma_\varepsilon = 2.$ Figure 3.5 shows that, for the use of the Grouped Lasso-Zero prior, the misclassification rate is $10\%,$ as we had 12 red crosses over 120 red symbols in six replications. For the use of the Grouped Horseshoe prior, the misclassification rate is $39.2\%,$ which is about four times worse, since we had 47 blue crosses over 120 blue symbols in six replications. Figure 3.5 shows that most of the Grouped Lasso-Zero threshold statistics are close to 1 when the true effect is

non-linear and close to 0 when the true effect is non-linear. In contrast, most of the Grouped Horseshoe threshold statistics are close to 1 when the true effect is non-linear, but scattered between 0.4 and 0.8 when the true effect is linear, which means many predictors that have a linear effect are misclassified as having a non-linear effect when the Grouped Horseshoe prior is used.

For the simulation study in Figure 3.6, we increased the sample size to $n = 2,000$ and lower the error standard deviation to $\sigma_\varepsilon = 0.25,$ compared to the simulation study of Figure 3.5. The higher sample size and the lower error standard deviation should make this linear effect versus non-linear effect classification much easier. The result shows that, for the use of Grouped Lasso-Zero prior, it leads to a perfect classification performance with no red crosses associated with misclassification in six replications. However, for the use of Grouped Horseshoe prior, the more favourable simulations conditions do not seem to help lower the misclassification rate, and most of the Grouped Horseshoe threshold statistics are still scattered between 0.4 and 0.8 when the true effect is linear, leading to a 40.8% misclassification rate, since we had 49 blue crosses over 120 blue symbols in six replications.

For the simulation study in Figure 3.7, we have the same sample size being $n = 2,000$ and the same error standard deviation being $\sigma_\varepsilon = 0.25$ as the simulation study of Figure 3.6. To improve the poor classification performance for the use of Grouped Horseshoe prior, instead of thresholding the $E\left(\gamma_{u_j,\text{GHS}}\middle|\boldsymbol{y}\right)$ at $\frac{1}{2}, 11 \leq j \leq 30,$ as stated in (3.77), we applied k-means clustering (e.g. MacQueen, 1967) to the $E\left(\gamma_{u_j,\text{GHS}}\middle|\boldsymbol{y}\right)$ observations, with the number of clusters fixed at 2, to obtain 2 clusters and corresponding classification rule. We used the function `kmeans()` within the R computing environment (R Core Team, 2024), which aims to partition each threshold points into 2 groups, such that the sum of squares from threshold points in each vector to the corresponding assigned cluster centers is minimized. The threshold for the Grouped Horseshoe statistics is based on k-means clustering. For example, for the analysis corresponding to replication 1 in Figure 3.7, the k-means threshold shown by horizontal blue line is 0.8031. The classification border of 0.5 shown by the horizontal red line is the threshold for the Grouped Lasso-Zero statistics. The result shows that with the use of this k-means alternative to Figure 3.6, the misclassification rate for the use of Grouped Horseshoe prior drops from 40.8% shown by Figure 3.6 to 10.8% shown by Figure 3.7, since we had 13 blue crosses over 120 blue symbols in Figure 3.7, compared to 49 blue crosses in Figure 3.6.

The remedy of k-means clustering approach to thresholding for generalized ad-

ditive model selection helps lower the misclassification rate for classification performance. We need to note this remedy relies on situations where there are many candidate predictors of various effect types. For example, if there are only three to six candidate predictors and most of them have strongly non-linear effects, then the k-means clustering approach to thresholding choice may not be viable.

3.8.2 Classification Among Zero, Linear and Non-linear Effects

Figure 3.5, 3.6 and 3.7 in Section 3.8.1 are based on six replications and omit the zero versus linear effect classification based on the $E\left(\gamma_{\beta_j,\text{HS}}\middle|\boldsymbol{y}\right)$ statistics with the use of Horseshoe prior and based on the $E\left(\gamma_{\beta_j}\middle|\boldsymbol{y}\right)$ statistics with the use of Laplace-Zero prior, with $1 \leq j \leq 20$ from (3.76). To get a more complete picture, we ran a bigger simulation study, relative to the one conducted in Section 3.8.1, for zero versus linear/non-linear effect classification on Gaussian responses among the use of 3 types prior distributions with different combinations of sample size and error standard deviation in this subsection. The simulation settings are stated in (3.76), (2.56), (2.57) and (3.79) as follows:

$$\begin{aligned}&\text{sample size } n \in \{500, 1000, 2000\}\,,\\ &\text{error standard deviation } \sigma_\varepsilon \in \{0.25, 0.5, 1, 2\} \text{ for Gaussian responses,}\\ &\text{threshold parameter } \tau = 0.5.\end{aligned} \tag{3.79}$$

In Figure 3.8, the performance measure was a misclassification rate for 30 candidate predictors stated in (3.76) being classified into one of three classes: zero effect, linear effect or non-linear effect. Within each panel, the blue side-by-side boxplots compare the misclassification rates among the use of three types of prior distributions given as follows:

A. the Laplace-Zero/Grouped Lasso-Zero prior approach based on model (2.28) with the classification border set to $\frac{1}{2}$ on rules (2.54),

B. the Horseshoe/Grouped Horseshoe prior approach based on model (3.23) with the classification border set to $\frac{1}{2}$ on rules (3.74),

C. the Horseshoe/Grouped Horseshoe prior approach based on model (3.23) with the classification border based on k-means clustering.

Figure 3.8 shows that the use of type A prior distribution is clearly superior to type B with a much lower misclassification rate. The use of type C prior distribution offers a big improvement compared to the use of type B. It shows that in our Bayesian generalized additive model selection setting stated in (3.76), (3.79), (2.56) and (2.57), the use of the (Grouped) Horseshoe priors does not compete very well with the use of the Laplace-Zero/Grouped Lasso-Zero priors. Perhaps this can lead to the devel-

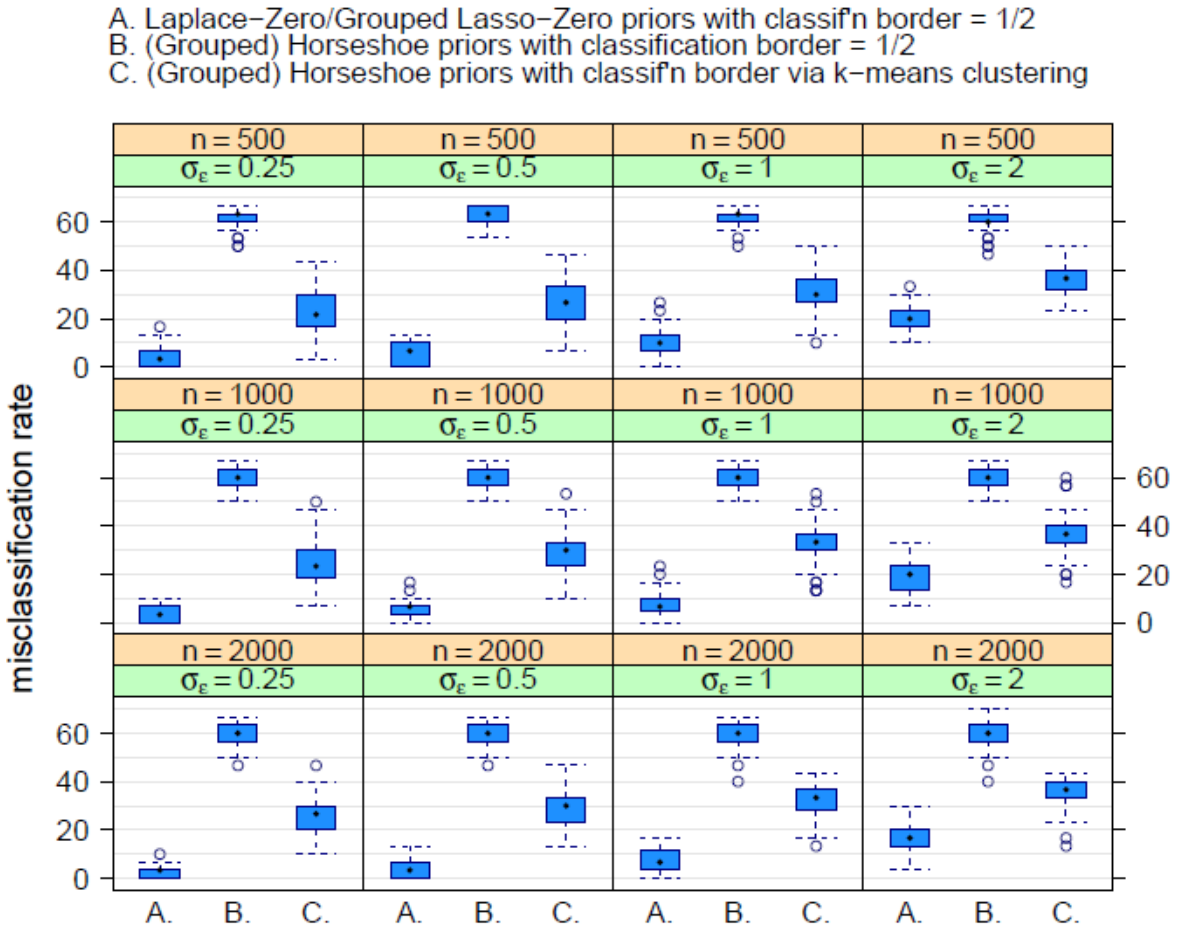


Figure 3.8: *Side-by-side boxplots of the misclassification rates for the comparative performance simulation study described in the text in the case of the response variable being Gaussian. Each panel corresponds to a different combination of sample size and error standard deviation. Within each panel, the side-by-side boxplots compare the zero, linear, non-linear effect misclassification rate across each of the three methods: A. Laplace-Zero/Grouped Lasso-Zero priors with classification border at* $\frac{1}{2}$*, B. (Grouped) Horseshoe priors with classification border at* $\frac{1}{2}$*, C. (Grouped) Horseshoe priors with classification border determined via k-means clustering.*

In summary, for Gaussian responses, using the (Grouped) Horseshoe priors with fixed border of 0.5 with model selection rules stated in (3.74) and using the (Grouped)

Horseshoe priors with k-means clustering used to determine the classification border with model selection rules stated in (3.74), both generally underperform relative to the Laplace-Zero/Grouped Lasso-Zero priors with model selection rules stated in (2.54) with threshold as 0.5. Therefore, in general, the recommendation for model selection on Gaussian responses, is using the Laplace-Zero/Grouped Lasso-Zero priors with model selection rules stated in (2.54) with threshold as 0.5.

3.9 Comparative Performance of Poisson Responses

For the three-category generalized additive model selection on Poisson response, we conducted a simulation study to evaluate the comparative performance of three approaches: A. (Grouped) Horseshoe priors described in Algorithm 5 with classification border set to $\frac{1}{2}$ in selection rule (3.75), B. (Grouped) Horseshoe priors with classification border via k-means clustering, as described in Section 3.8.1, and against C. an existing Bayesian approach spikeSlabGAM (Scheipl, 2022) as described in Scheipl et al. (2012). Details on use of the spikeSlabGAM package are given in Scheipl (2011).

In the case of spikeSlabGAM, we used the default call to its `spikeSlabGAM( )` function. The model having highest posterior probability in the `spikeSlabGAM( )` output object was selected. spikeSlabGAM uses a $N\left(0, 0.00025^2\right)$ mass as the "spike" and a Gaussian density function as the "slab". For Poisson response models, spike–SlabGAM uses Metropolis-Hastings sampling.

The data were generated with 15 candidate predictors, 5 having "true" linear effects, 5 having "true" non-linear effects and 5 having "true" zero effects. The sample size varied as $n \in \{1000, 2000\}$. Each of the predictors were generated from independent Normal distributions. The non-linear effects corresponded to quintic polynomials with random generated coefficients. We used the misclassification rate to measure the accuracy of the methodologies, defined as the proportion of times that a predictor is misclassified as zero, linear or non-linear effect based on the nature of the true effects as determined by the data generating process. For each sample size n, we ran simulation on 15 independently generated data sets. Each replication involved the generation of new coefficients.

The side-by-side boxplots in Figure 3.9 facilitate comparison of

A. the (Grouped) Horseshoe priors with a classification border at $\frac{1}{2}$,

B. the same as A., but with the classification border based on k-means clustering,

C. the "spike-and-slab" priors in spikeSlabGAM with default settings.

For A. and B. in Figure 3.9, we used $N_{\text{warm}} = N_{\text{kept}} = 1000$ for the MCMC. Figure 3.9 shows that B. outperforms A. and C. with lower misclassification rate.

It is noted that using the (Grouped) Horseshoe priors, both A. and B., for Poisson responses in the three-category GAM selection, the slice sampling method as described in Algorithm 5 was successfully drawing samples in approximately 30% to 50% of 50 replications. It shows the numerical instability of the slice sampling method in Algorithm 5 for Poisson responses. Possible reasons for numerical instability on Poisson responses via slice sampling are summarized as follows. Firstly, the Poisson response assumes the variance and the mean are equal, which may not hold in some data sets simulated. If the data is overdispersed, the model does not fit well, leading to poor convergence. Secondly, slice sampling requires evaluating the likelihood and prior densities at various points. If the likelihood or posterior distribution has very steep gradients or is poorly conditioned, some seeds may lead to numerical underflow or overflow. Moreover, slice sampling works by defining an interval around the current value and drawing samples uniformly from within that region, known as a slice. However, if the interval extends into areas where the likelihood is extremely low or effectively zero, the sampler may become stuck and fail to explore the distribution efficiently. To improve the use of (Grouped) Horseshoe priors via slice sampling for Poisson responses for the three-category GAM selection is one of future research goals.

3.10 Comparative Performance of Negative-Binomial Responses

We now repeated the same procedure as described in Section 3.9 for the three-category generalized additive model selection on Negative-Binomial response. We conducted a simulation study to evaluate the comparative performance of two approaches: A. (Grouped) Horseshoe priors described in Algorithm 6 with classification border set to $\frac{1}{2}$ in selection rule (3.75) and B. (Grouped) Horseshoe priors with classification border via k-means clustering, as described in Section 3.8.1.

The data were generated with 30 candidate predictors, 10 having "true" linear effects, 10 having "true" non-linear effects and 10 having "true" zero effects. The sample size varied as $n \in \{500, 1000, 2000\}$ and the shape parameter varied as

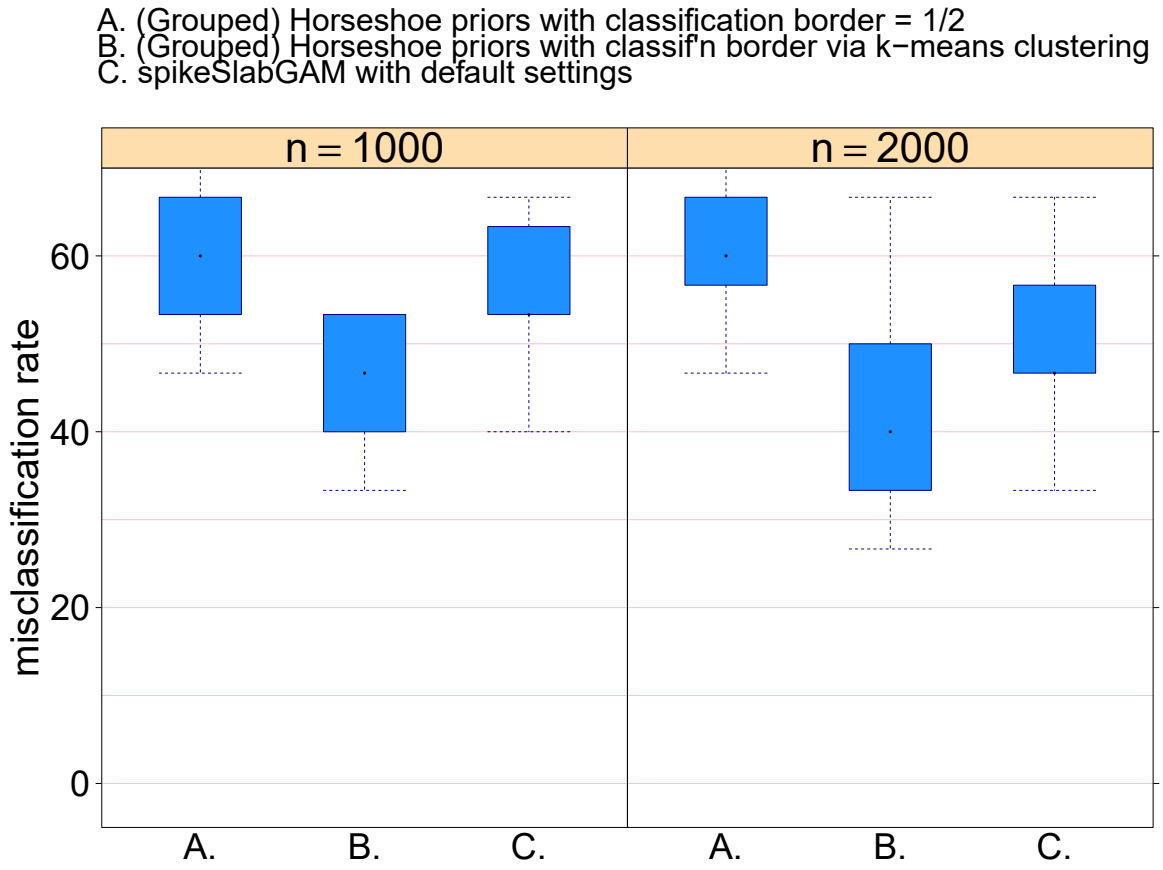


Figure 3.9: *Side-by-side boxplots of the misclassification rates for the comparative performance simulation study described in the text in the case of the response variable being Poisson. Each panel corresponds to a different sample size. Within each panel, the side-by-side boxplots compare the misclassification rate across each of three methods: A. (Grouped) Horseshoe priors with a classification border at* $\frac{1}{2}$*, B. (Grouped) Horseshoe priors with classification border determined via* k*-means clustering. C.* ***spikeSlabGAM*** *with default settings.*

$\kappa \in \{1, 2, 4\}$. For each combination of sample size n and shape parameter κ, we ran simulation on 50 independently generated data sets.

The side-by-side boxplots in Figure 3.10 allow direct comparison between the two approaches:

A. the (Grouped) Horseshoe priors with a classification border at $\frac{1}{2}$,

B. the same as A., but with the classification border based on k-means clustering.

For A. and B. in Figure 3.10, we use $N_{\text{warm}} = N_{\text{kept}} = 1000$ for the MCMC. Figure 3.10 shows that B. outperforms A. with lower misclassification rate.

Algorithm 4: *Markov chain Monte Carlo generation of samples from the posterior distributions of the parameters in (3.23) for Gaussian responses using Horseshoe-type prior.*

Data Inputs: $\boldsymbol{y}(n \times 1)$; $\boldsymbol{X}\,(n \times (d_\circ + d_\bullet))$; $\boldsymbol{Z}_j(n \times K_j), 1 \le j \le d_\bullet$.

Response Type Input: **responseType** $\in \{\text{Gaussian}\}$.

Sufficient Statistics Inputs: **XTy**, **XTX**, **ZTy**, **ZTX**, **ZTZ**.

Hyperparameter Inputs: $\sigma_{\beta_0}, s_\beta, s_\varepsilon, s_u > 0$.

Chain Length Inputs: N_{warm} and N_{kept} both positive integers.

Initialize: $\boldsymbol{\beta}^{[0]} \longleftarrow \mathbf{0}_{d_\circ + d_\bullet}$

$$\boldsymbol{u}_j^{[0]} \longleftarrow \mathbf{0}_{K_j}, 1 \le j \le d_\bullet;\ (\sigma_\varepsilon^2)^{[0]} \longleftarrow 1;\ \ a_\varepsilon^{[0]} \longleftarrow 1;\ (\sigma_\beta^2)^{[0]} \longleftarrow 1;\ a_\beta^{[0]} \longleftarrow 1$$

$$\boldsymbol{f}_{2\beta}^{[0]} \longleftarrow \mathbf{1}_{d_\circ + d_\bullet};\ \ \boldsymbol{f}_{1\beta}^{[0]} \longleftarrow \mathbf{1}_{d_\circ + d_\bullet};\ \ f_{2uj}^{[0]} \longleftarrow 1,\ \ f_{1uj}^{[0]} \longleftarrow 1,\ \ 1 \le j \le d_\bullet;$$

$$a_{uj}^{[0]} \longleftarrow 1, 1 \le j \le d_\bullet;\ (\sigma_{uj}^2)^{[0]} \longleftarrow 1, 1 \le j \le d_\bullet.$$

For $j = 1, \ldots, d_\bullet$: $\boldsymbol{\omega}_{Z_j} \longleftarrow \text{diagonal}\left(\mathbf{ZTZ}^{\langle j,j \rangle}\right)$

For $g = 1, \ldots, N_{\text{warm}} + N_{\text{kept}}$:

$$\omega_1 \longleftarrow \mathbf{yT1}$$

$$\omega_2 \longleftarrow \left\{ n/(\sigma_\varepsilon^2)^{[g-1]} + (1/\sigma_{\beta_0}^2) \right\} \quad ; \quad \beta_0^{[g]} \sim N\Big(\frac{\omega_1}{(\sigma_\varepsilon^2)^{[g-1]}\omega_2}, \frac{1}{\omega_2}\Big)$$

$$\boldsymbol{\Omega} \longleftarrow (\mathbf{XTX}) \Big/ (\sigma_\varepsilon^2)^{[g-1]} + \text{diag}\Big(\boldsymbol{f}_{2\beta}^{[g-1]}\Big) \Big/ (\sigma_\beta^2)^{[g-1]}$$

$$\boldsymbol{\omega}_3 \longleftarrow \mathbf{XTy} - \sum_{j=1}^{d_\bullet} \mathbf{ZTX}^{\langle j \rangle T} \left(\boldsymbol{u}_j^{[g-1]}\right)$$

Decompose $\boldsymbol{\Omega} = \boldsymbol{U}_{\boldsymbol{\Omega}} \text{diag}(d_{\boldsymbol{\Omega}}) \boldsymbol{U}_{\boldsymbol{\Omega}}^T$ where $\boldsymbol{U}_{\boldsymbol{\Omega}} \boldsymbol{U}_{\boldsymbol{\Omega}}^T = \boldsymbol{I}$

$$\boldsymbol{z} \sim N(\mathbf{0}, \boldsymbol{I})((d_\circ + d_\bullet) \times 1) \quad ; \quad \boldsymbol{\beta}^{[g]} \longleftarrow \boldsymbol{U}_{\boldsymbol{\Omega}} \left(\frac{\boldsymbol{U}_{\boldsymbol{\Omega}}^T \boldsymbol{z}}{\sqrt{d_{\boldsymbol{\Omega}}}} + \frac{\boldsymbol{U}_\Omega^T (\boldsymbol{\omega}_3)}{\boldsymbol{d}_{\boldsymbol{\Omega}} (\sigma_\varepsilon^2)^{[g-1]}} \right)$$

For $j = 1, \ldots, d_\circ + d_\bullet$:

$$f_{2\beta j}^{[g]} \sim \text{Gamma}\left(1, f_{1\beta j}^{[g-1]} + \frac{(\beta_j^2)^{[g]}}{2\left(\sigma_\beta^2\right)^{[g-1]}}\right)$$

continued on a subsequent page ...

Algorithm 4 continued: *This is a continuation of the description of this algorithm that commences on a preceding page.*

$$f_{1\beta j}^{[g]} \sim \text{Gamma}\left(1, f_{2\beta j}^{[g]} + 1\right)$$

$$(\sigma_\beta^2)^{[g]} \sim \text{Inverse-Gamma}\left(\frac{1}{2}(d_\circ + d_\bullet + 1), 1/a_\beta^{[g-1]} + \frac{1}{2}\boldsymbol{\beta}^{[g]T}\text{diag}\left(\boldsymbol{f}_{2\beta}^{[g]}\right)\boldsymbol{\beta}^{[g]}\right)$$

$$a_\beta^{[g]} \sim \text{Inverse-Gamma}\left(1, \left\{1/\left(\sigma_\beta^2\right)^{[g]}\right\} + \left(1/s_\beta^2\right)\right) \; ; \; \boldsymbol{\beta}^{\text{curr}} \longleftarrow \boldsymbol{\beta}^{[g]}$$

For $j = 1, \ldots, d_\bullet$: $\quad \boldsymbol{u}_j^{\text{curr}} \longleftarrow \boldsymbol{u}_j^{[g-1]}$

For $j = 1, \ldots, d_\bullet$:

$$\boldsymbol{\omega}_6 \longleftarrow \mathbf{ZTy}^{\langle j \rangle} - \mathbf{ZTX}^{\langle j \rangle}\boldsymbol{\beta}^{\text{curr}} - \sum_{j' \neq j}^{d_\bullet} \mathbf{ZTZ}^{\langle j, j' \rangle}\left(\boldsymbol{u}_{j'}^{\text{curr}}\right)$$

$$\boldsymbol{\omega}_7 \longleftarrow \left\{\boldsymbol{\omega}_{\boldsymbol{Z}_j} \Big/ \left(\sigma_\varepsilon^2\right)^{[g-1]}\right\} + \left\{f_{2uj}^{[g-1]}\mathbf{1}_{K_j} \Big/ \left(\sigma_{uj}^2\right)^{[g-1]}\right\}$$

$$\boldsymbol{z} \sim N\left(\mathbf{0}, \boldsymbol{I}_{K_j}\right) \quad ; \quad \boldsymbol{u}_j^{\text{curr}} \longleftarrow \left(\boldsymbol{z} \Big/ \sqrt{\boldsymbol{\omega}_7}\right) + \left[\boldsymbol{\omega}_6 \Big/ \left\{\boldsymbol{\omega}_7 \left(\sigma_\varepsilon^2\right)^{[g-1]}\right\}\right]$$

For $j = 1, \ldots, d_\bullet$: $\quad \boldsymbol{u}_j^{\text{curr}} \longleftarrow \boldsymbol{u}_j^{[g]}$

For $j = 1, \ldots, d_\bullet$:

$$f_{2uj}^{[g]} \sim \text{Gamma}\left(\tfrac{1}{2}\left(K_j + 1\right), f_{1uj}^{[g-1]} + \frac{\left\|\boldsymbol{u}_j^{[g]}\right\|^2}{2\left(\sigma_{uj}^2\right)^{[g-1]}}\right)$$

$$f_{1uj}^{[g]} \sim \text{Gamma}\left(1, f_{2uj}^{[g]}) + 1\right)$$

$$(\sigma_{uj}^2)^{[g]} \sim \text{Inverse-Gamma}\left(\tfrac{1}{2}(K_j + 1), \frac{1}{a_{uj}^{[g-1]}} + \tfrac{1}{2}\left\|\boldsymbol{u}_j^{[g]}\right\|^2 f_{2uj}^{[g]}\right)$$

$$a_{uj}^{[g]} \sim \text{Inverse-Gamma}\left(1, \frac{1}{\left(\sigma_{uj}^2\right)^{[g]}} + \frac{1}{s_u^2}\right)$$

$$\boldsymbol{\omega}_{10} \longleftarrow \mathbf{1}_n\beta_0^{[g]} + \boldsymbol{X}\boldsymbol{\beta}^{[g]} + \sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j\boldsymbol{u}_j^{[g]}$$

$$(\sigma_\varepsilon^2)^{[g]} \sim \text{Inverse-Gamma}\left(\tfrac{1}{2}(n+1), (1\Big/a_\varepsilon^{[g-1]}) + \tfrac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{\omega}_{10}\right\|^2\right)$$

$$a_\varepsilon^{[g]} \sim \text{Inverse-Gamma}\left(1, \left\{1\Big/(\sigma_\varepsilon^2)^{[g]}\right\} + (1\Big/s_\varepsilon^2)\right)$$

Outputs: All chains after omission of the first N_{warm} values.

Algorithm 5: *Markov chain Monte Carlo generation of samples from the posterior distributions of the parameters in (3.24) for Poisson responses using Horseshoe-type priors.*

Data Inputs: $\boldsymbol{y}(n \times 1), \boldsymbol{X}(n \times (d_\circ + d_\bullet)), \boldsymbol{Z}_j(n \times K_j), 1 \le j \le d_\bullet$.

Response Type Input: **responseType** $\in \{\text{Poisson}\}$.

Hyperparameter Inputs: $\sigma_{\beta_0}, s_\beta, s_\varepsilon, s_u, A_\beta, B_\beta, A_u, B_u > 0$.

Chain Length Inputs: N_{warm} and N_{kept} both positive integers.

Initialize: $\boldsymbol{\beta}^{[0]} \longleftarrow \mathbf{0}_{d_\circ + d_\bullet}$; $\boldsymbol{u}_j^{[0]} \longleftarrow \mathbf{0}_{K_j}$; $1 \le j \le d_\bullet$, $(\sigma_\varepsilon^2)^{[0]} \longleftarrow 1$; $a_\varepsilon^{[0]} \longleftarrow 1$;

$(\sigma_\beta^2)^{[0]} \longleftarrow 1$; $a_\beta^{[0]} \longleftarrow 1$; $\boldsymbol{f}_{2\beta}^{[0]} \longleftarrow \mathbf{1}_{d_\circ + d_\bullet}$; $\boldsymbol{f}_{1\beta}^{[0]} \longleftarrow \mathbf{1}_{d_\circ + d_\bullet}$;

$\boldsymbol{f}_{2u}^{[0]} \longleftarrow \mathbf{1}_{d_\bullet}$; $\boldsymbol{f}_{1u}^{[0]} \longleftarrow \mathbf{1}_{d_\bullet}$; $(\sigma_{uj}^2)^{[0]} \longleftarrow 1$, $a_{uj}^{[0]} \longleftarrow 1$, $1 \le j \le d_\bullet$,

$\text{sumDists}(\beta_0) \longleftarrow 0$; $\text{sumDists}(\beta_j) \longleftarrow 0$; $\text{sumDists}(u_{jk}) \longleftarrow 0$.

For $g = 1, \ldots, N_{\text{warm}} + N_{\text{kept}}$:

$s_1(\beta_0) \longleftarrow n\bar{\boldsymbol{y}}$; $s_2(\beta_0) \longleftarrow \sigma_{\beta_0}^2$; $\boldsymbol{s}_3(\beta_0) \longleftarrow \mathbf{1}_n$;

$$\boldsymbol{s}_4(\beta_0) \longleftarrow \boldsymbol{X}\boldsymbol{\beta}^{[g-1]} + \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j \boldsymbol{u}_j^{[g-1]},$$

$$\beta_0^{[g]} \longleftarrow \text{sliceSampling}\left(\beta_0^{[g-1]}, s_1(\beta_0), s_2(\beta_0), \boldsymbol{s}_3(\beta_0), \boldsymbol{s}_4(\beta_0), g, \text{sumDists}(\beta_0)\right),$$

$$\text{sumDists}(\beta_0) \longleftarrow \text{sumDists}(\beta_0) + \left|\beta_0^{[g-1]} - \beta_0^{[g]}\right|.$$

For $j = 1, \ldots, d_\circ + d_\bullet$:

$s_1(\beta_j) \longleftarrow \mathbf{1}_n^T (\boldsymbol{y} \odot \boldsymbol{X}_j)$; $s_2(\beta_j) \longleftarrow (\sigma_\beta^2)^{[g-1]} / f_{2\beta_j}^{[g-1]}$; $\boldsymbol{s}_3(\beta_j) \longleftarrow \boldsymbol{X}_j$;

$$\boldsymbol{s}_4(\beta_j) \longleftarrow \beta_0^{[g-1]} + (\boldsymbol{X})_{-j} \left(\boldsymbol{\beta}^{[g-1]}\right)_{-j} + \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j \boldsymbol{u}_j^{[g-1]};$$

$$\beta_j^{[g]} \longleftarrow \text{sliceSampling}\left(\beta_j^{[g-1]}, s_1(\beta_j), s_2(\beta_j), \boldsymbol{s}_3(\beta_j), \boldsymbol{s}_4(\beta_j), g, \text{sumDists}(\beta_j)\right),$$

$$\text{sumDists}(\beta_j) \longleftarrow \text{sumDists}(\beta_j) + \left|\beta_j^{[g-1]} - \beta_j^{[g]}\right|.$$

$$f_{2\beta_j}^{[g]} \sim \text{Gamma}\left(1, \frac{(\beta_j^2)^{[g]}}{2(\sigma_\beta^2)^{[g-1]}} + f_{1\beta_j}^{[g-1]}\right); \quad f_{1\beta_j}^{[g]} \sim \text{Gamma}\left(1, 1 + f_{2\beta_j}^{[g]}\right).$$

$$(\sigma_\beta^2)^{[g]} \sim \text{Inverse-Gamma}\left(\tfrac{1}{2}\left(d_\circ + +d_\bullet + 1\right), 1/a_\beta^{[g-1]} + \tfrac{1}{2}\boldsymbol{\beta}^{[g]T}\text{diag}(\boldsymbol{f}_{2\beta}^{[g]})\boldsymbol{\beta}^{[g]}\right)$$

continued on a subsequent page ...

Algorithm 5 continued: *This is a continuation of the description of this algorithm that commences on a preceding page.*

$a_\beta \sim \text{Inverse-Gamma}\left(1, \left\{1/\left(\sigma_\beta^2\right)^{[g]}\right\} + \left(1/s_\beta^2\right)\right)$; $\boldsymbol{s}_4(u_{jk}) \longleftarrow \beta_0^{[g]} + \boldsymbol{X}\boldsymbol{\beta}^{[g]}$,

For $j = 1, \ldots, d_\bullet$:

$s_2(u_{jk}) \longleftarrow (\sigma_{uj}^2)^{[g-1]}/f_{2uj}^{[g-1]}$; $\boldsymbol{s}_4(u_{jk}) \longleftarrow \boldsymbol{s}_4(u_{jk}) + \sum_{j' \neq j}^{d_\bullet} \boldsymbol{Z}_{j'} \boldsymbol{u}_{j'}^{[g-1]}$,

For $k = 1, \ldots, K_j$:

$s_1(u_{jk}) \longleftarrow \mathbf{1}_n^T (\boldsymbol{y} \odot \boldsymbol{Z}_{jk})$; $\boldsymbol{s}_3(u_{jk}) \longleftarrow \boldsymbol{Z}_{jk}$,

$\boldsymbol{s}_4(u_{jk}) \longleftarrow \boldsymbol{s}_4(u_{jk}) + \sum_{k' \neq k}^{K_j} \boldsymbol{Z}_{jk'} u_{jk'}^{[g-1]}$,

$u_{jk}^{[g]} \longleftarrow \text{sliceSampling}\left(u_{jk}^{[g-1]}, s_1(u_{jk}), s_2(u_{jk}), \boldsymbol{s}_3(u_{jk}), \boldsymbol{s}_4(u_{jk}), g, \text{sumDists}(u_{jk})\right)$.

$\text{sumDists}(u_{jk}) \longleftarrow \text{sumDists}(u_{jk}) + \left|u_{jk}^{[g-1]} - u_{jk}^{[g]}\right|$

$f_{2uj}^{[g]} \sim \text{Gamma}\left(\frac{1}{2}(K_j + 1), f_{1uj}^{[g-1]} + \frac{1}{2}\left\|\boldsymbol{u}_j^{[g]}\right\|^2 \Big/ (\sigma_{uj}^2)^{[g-1]}\right)$,

$f_{1uj}^{[g]} \sim \text{Gamma}\left(1, 1 + f_{2uj}^{[g]}\right)$,

$(\sigma_{uj}^2)^{[g]} \sim \text{Inverse-Gamma}\left(\frac{1}{2}(K_j + 1), \left(1/a_{uj}^{[g-1]}\right) + \frac{1}{2} f_{2uj}^{[g]} \left\|\boldsymbol{u}_j^{[g]}\right\|^2\right)$,

$a_{uj}^{[g]} \sim \text{Gamma}\left(1, \left\{1\Big/\left(\sigma_{uj}^2\right)^{[g]}\right\} + \left(1\Big/s_u^2\right)\right)$.

Outputs: All chains after omission of the first N_{warm} values.

Algorithm 6: *Markov chain Monte Carlo generation of samples from the posterior distributions of the parameters in (3.25) for Negative-Binomial responses using Horseshoe-priors.*

Function to compute the logarithm of unnormalised density function of λ for Negative-Binomial responses called logUnnDensLambda( ).

$$\boldsymbol{\eta}_\beta \longleftarrow \boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\eta}_u \longleftarrow \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j \boldsymbol{u}_j,$$

Data Inputs: $\lambda, \beta_0, n, b_\kappa, \boldsymbol{g}, \boldsymbol{\eta}_\beta, \boldsymbol{\eta}_u$.

$\kappa \longleftarrow \exp(\lambda)$.

Outputs: $\lambda + n\{\kappa \log(\kappa) - \log(\Gamma(\kappa))\}$

$$-\kappa \left[\sum_{i=1}^{n} \left(\beta_0 + \boldsymbol{\eta}_\beta + \boldsymbol{\eta}_u\right) + b_\kappa + \sum_{i=1}^{n} \left\{ g_i \exp\left(-\left(\beta_0 + \boldsymbol{\eta}_\beta + \boldsymbol{\eta}_u\right)\right)\right\} - \sum_{i=1}^{n} \log(g_i) \right]$$

Slice sampling for Negative-Binomial responses:

Data Inputs: $\boldsymbol{y}(n \times 1), \boldsymbol{X}(n \times (d_\circ + d_\bullet)), \boldsymbol{Z}_j(n \times K_j), 1 \le j \le d_\bullet$.

Response Type Input: **responseType** $\in$ {Negative-Binomial} .

Hyperparameter Inputs: $\sigma_{\beta_0}, s_\beta, s_\varepsilon, s_u, A_\beta, B_\beta, A_u, B_u, b_\kappa > 0$,

Chain Length Inputs: N_{warm} and N_{kept} both positive integers.

Initialize: $\boldsymbol{\beta}^{[0]} \longleftarrow \mathbf{0}_{d_\circ + d_\bullet}$; $\boldsymbol{u}_j^{[0]} \longleftarrow \mathbf{0}_{K_j}$; $1 \le j \le d_\bullet$, $(\sigma_\varepsilon^2)^{[0]} \longleftarrow 1$; $a_\varepsilon^{[0]} \longleftarrow 1$;

$(\sigma_\beta^2)^{[0]} \longleftarrow 1$; $a_\beta^{[0]} \longleftarrow 1$; $\boldsymbol{f}_{2\beta}^{[0]} \longleftarrow \mathbf{1}_{d_\circ + d_\bullet}$; $\boldsymbol{f}_{1\beta}^{[0]} \longleftarrow \mathbf{1}_{d_\circ + d_\bullet}$; $\lambda^{[0]} \longleftarrow 1$,

$\boldsymbol{f}_{2u}^{[0]} \longleftarrow \mathbf{1}_{d_\bullet}$; $\boldsymbol{f}_{1u}^{[0]} \longleftarrow \mathbf{1}_{d_\bullet}$; $(\sigma_{uj}^2)^{[0]} \longleftarrow 1$, $a_{uj}^{[0]} \longleftarrow 1$, $1 \le j \le d_\bullet$,

sumDists$(\beta_0) \longleftarrow 0$; sumDists$(\beta_j) \longleftarrow 0$; sumDists$(u_{jk}) \longleftarrow 0$;

sumDists$(\lambda) \longleftarrow 0$;

For $s = 1, \ldots, N_{\text{warm}} + N_{\text{kept}}$:

$$\boldsymbol{\eta}_\beta \longleftarrow \boldsymbol{X}\boldsymbol{\beta}^{[s-1]}, \boldsymbol{\eta}_u \longleftarrow \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j \boldsymbol{u}_j^{[s-1]}$$

$$\boldsymbol{g}^{[s]} \sim \text{Gamma}\left(\boldsymbol{y} + \exp(\lambda^{[s-1]}), 1 + \exp\left(\lambda^{[s-1]}\right) \exp\left(\beta_0^{[s-1]} + \boldsymbol{\eta}_\beta + \boldsymbol{\eta}_u\right)\right)$$

$s_1(\beta_0) \longleftarrow -n \exp\left(\lambda^{[s-1]}\right)$; $s_2(\beta_0) \longleftarrow \sigma_{\beta_0}^2$; $\boldsymbol{s}_3(\beta_0) \longleftarrow -\mathbf{1}_n$;

continued on a subsequent page ...

Algorithm 6 continued: *This is a continuation of the description of this algorithm that commences on a preceding page.*

$\boldsymbol{s}_4(\beta_0) \longleftarrow \mathbf{1}_n\lambda^{[s-1]} + \log\left(\boldsymbol{g}^{[s]}\right) + \boldsymbol{\eta}_\beta + \boldsymbol{\eta}_u$

$\beta_0^{[s]} \longleftarrow \mathsf{sliceSampling}\left(\beta_0^{[s-1]}, s_1(\beta_0), s_2(\beta_0), \boldsymbol{s}_3(\beta_0), \boldsymbol{s}_4(\beta_0), g, \mathrm{sumDists}(\beta_0)\right)$

$\mathrm{sumDists}(\beta_0) \longleftarrow \mathrm{sumDists}(\beta_0) + \left|\beta_0^{[s-1]} - \beta_0^{[s]}\right|$

$\omega_1 \longleftarrow \mathsf{logUnnDensLambda}\left(\lambda^{[s-1]}, \beta_0^{[s-1]}, n, b_\kappa, \boldsymbol{g}, \boldsymbol{\eta}_\beta, \boldsymbol{\eta}_u\right)$

if $(s = 2)$ $t \longleftarrow 1.0$; if $(s > 2)$ $t \longleftarrow \mathrm{sumDists}(\lambda)/(s-1)$.

$u \longleftarrow$ a draw from the Uniform$(0, 1)$ distribution

$\omega_2 \longleftarrow \omega_1 + \log(u)$; $L_\lambda \longleftarrow \lambda^{[s-1]} - tu$; $U_\lambda \longleftarrow L_\lambda + t$.

$\omega_3 \longleftarrow \mathsf{logUnnDensLambda}\left(L_\lambda, \beta_0^{[s]}, n, b_\kappa, \boldsymbol{g}^{[s]}, \boldsymbol{\eta}_\beta, \boldsymbol{\eta}_u\right)$.

while $(\omega_2 < \omega_3)$

$L_\lambda \longleftarrow L_\lambda - t$; $\omega_3 \longleftarrow \mathsf{logUnnDensLambda}\left(L_\lambda, \beta_0^{[s]}, n, b_\kappa, \boldsymbol{g}^{[s]}, \boldsymbol{\eta}_\beta, \boldsymbol{\eta}_u\right)$.

$\omega_4 \longleftarrow \mathsf{logUnnDensLambda}\left(U_\lambda, \beta_0^{[s]}, n, b_\kappa, \boldsymbol{g}^{[s]}, \boldsymbol{\eta}_\beta, \boldsymbol{\eta}_u\right)$.

while $(\omega_2 < \omega_4)$

$U_\lambda \longleftarrow U_\lambda + t$; $\omega_4 \longleftarrow \mathsf{logUnnDensLambda}\left(U_\lambda, \beta_0^{[s]}, n, b_\kappa, \boldsymbol{g}^{[s]}, \boldsymbol{\eta}_\beta, \boldsymbol{\eta}_u\right)$.

$\widetilde{L_\lambda} \longleftarrow L_\lambda$; $\widetilde{U_\lambda} \longleftarrow U_\lambda$; newPointFound $\longleftarrow$ FALSE.

while (newPointFound = FALSE)

$\lambda_{\text{new}} \longleftarrow \widetilde{L_\lambda} + u\left(\widetilde{U_\lambda} - \widetilde{L_\lambda}\right)$.

if $\left(\omega_2 < \mathsf{logUnnDensLambda}\left(\lambda_{\text{new}}, \beta_0^{[s]}, n, b_\kappa, \boldsymbol{g}^{[s]}, \boldsymbol{\eta}_\beta, \boldsymbol{\eta}_u\right)\right)$

$\lambda^{[s]} \longleftarrow \lambda_{\text{new}}$; newPointFound $\longleftarrow$ TRUE.

if $(\lambda_{\text{new}} < x)$ $\widetilde{L_\lambda} \longleftarrow \lambda_{\text{new}}$; if $(\lambda_{\text{new}} \geq x)$ $\widetilde{U_\lambda} \longleftarrow \lambda_{\text{new}}$

$\mathrm{sumDists}(\lambda) \longleftarrow \mathrm{sumDists}(\lambda) + \left|\lambda^{[s-1]} - \lambda)^{[s]}\right|$

For $j = 1, \ldots, d_\circ + d_\bullet$:

$s_1(\beta_j) \longleftarrow -\exp\left(\lambda^{[s]}\right)\mathbf{1}_n^T\boldsymbol{X}_j$; $s_2(\beta_j) \longleftarrow (\sigma_\beta^2)^{[s-1]}/f_{2\beta_j}^{[s-1]}$; $\boldsymbol{s}_3(\beta_j) \longleftarrow -\boldsymbol{X}_j$;

$\boldsymbol{s}_4(\beta_j) \longleftarrow \mathbf{1}_n\lambda^{[s]} + \log\left(\boldsymbol{g}^{[s]}\right) - \left(\mathbf{1}_n\beta_0^{[s]} + (\boldsymbol{X})_{-j}\left(\boldsymbol{\beta}^{[s-1]}\right)_{-j} + \boldsymbol{\eta}_u\right)$

$\beta_j^{[s]} \longleftarrow \mathsf{sliceSampling}\left(\beta_j^{[s-1]}, s_1(\beta_j), s_2(\beta_j), \boldsymbol{s}_3(\beta_j), \boldsymbol{s}_4(\beta_j), g, \mathrm{sumDists}(\beta_j)\right)$,

$\mathrm{sumDists}(\beta_j) \longleftarrow \mathrm{sumDists}(\beta_j) + \left|\beta_j^{[s-1]} - \beta_j^{[s]}\right|$.

continued on a subsequent page ...

Algorithm 6 continued: *This is a continuation of the description of this algorithm that commences on a preceding page.*

$$f_{2\beta_j}^{[s]} \sim \text{Gamma}\left(1, \frac{(\beta_j^2)^{[s]}}{2(\sigma_\beta^2)^{[s-1]}} + f_{1\beta_j}^{[s-1]}\right); \quad f_{1\beta_j}^{[s]} \sim \text{Gamma}\left(1, 1 + f_{2\beta_j}^{[s]}\right).$$

$$(\sigma_\beta^2)^{[s]} \sim \text{Inverse-Gamma}\left(\tfrac{1}{2}\left(d_\circ + + d_\bullet + 1\right), 1/a_\beta^{[s-1]} + \tfrac{1}{2}\boldsymbol{\beta}^{[s]T}\text{diag}(\boldsymbol{f}_{2\beta}^{[s]})\boldsymbol{\beta}^{[s]}\right)$$

$$a_\beta^{[s]} \sim \text{Inverse-Gamma}\left(1, \left\{1/\left(\sigma_\beta^2\right)^{[g]}\right\} + \left(1/s_\beta^2\right)\right);$$

$$\boldsymbol{s}_4(u_{jk}) \longleftarrow \mathbf{1}_n\lambda^{[s]} + \log\left(\boldsymbol{g}^{[s]}\right) - \left(\mathbf{1}_n\beta_0^{[s]} + \boldsymbol{X}\boldsymbol{\beta}^{[s]}\right),$$

For $j = 1, \ldots, d_\bullet$:

$$s_2(u_{jk}) \longleftarrow (\sigma_{uj}^2)^{[s-1]}/f_{2uj}^{[s-1]}; \quad \boldsymbol{s}_4(u_{jk}) \longleftarrow \boldsymbol{s}_4(u_{jk}) - \sum_{j' \neq j}^{d_\bullet} \boldsymbol{Z}_{j'}\boldsymbol{u}_{j'}^{[s-1]}$$

For $k = 1, \ldots, K_j$:

$$s_1(u_{jk}) \longleftarrow -\exp\left(\lambda^{[s]}\right)\mathbf{1}_{K_j}^T\boldsymbol{Z}_{jk}; \quad \boldsymbol{s}_3(u_{jk}) \longleftarrow -\boldsymbol{Z}_{jk},$$

$$\boldsymbol{s}_4(u_{jk}) \longleftarrow \boldsymbol{s}_4(u_{jk}) - \sum_{k' \neq k}^{K_j} \boldsymbol{Z}_{jk'}u_{jk'}^{[s-1]}$$

$$u_{jk}^{[s]} \longleftarrow \text{sliceSampling}\left(u_{jk}^{[s-1]}, s_1(u_{jk}), s_2(u_{jk}), \boldsymbol{s}_3(u_{jk}), \boldsymbol{s}_4(u_{jk}), \text{sumDists}(u_{jk})\right)$$

$$\text{sumDists}(u_{jk}) \longleftarrow \text{sumDists}(u_{jk}) + \left|u_{jk}^{[s-1]} - u_{jk}^{[s]}\right|$$

$$f_{2uj}^{[s]} \sim \text{Gamma}\left(\tfrac{1}{2}\left(K_j + 1\right), f_{1uj}^{[s-1]} + \tfrac{1}{2}\left\|\boldsymbol{u}_j^{[s]}\right\|^2 \Big/ (\sigma_{uj}^2)^{[s-1]}\right),$$

$$f_{1uj}^{[s]} \sim \text{Gamma}\left(1, 1 + f_{2uj}^{[s]}\right),$$

$$(\sigma_{uj}^2)^{[s]} \sim \text{Inverse-Gamma}\left(\tfrac{1}{2}\left(K_j + 1\right), \left(1/a_{uj}^{[s-1]}\right) + \tfrac{1}{2}f_{2uj}^{[s]}\left\|\boldsymbol{u}_j^{[s]}\right\|^2\right),$$

$$a_{uj}^{[s]} \sim \text{Gamma}\left(1, \left\{1\Big/\left(\sigma_{uj}^2\right)^{[s]}\right\} + \left(1\Big/s_u^2\right)\right).$$

Outputs: All chains after omission of the first N_{warm} values.

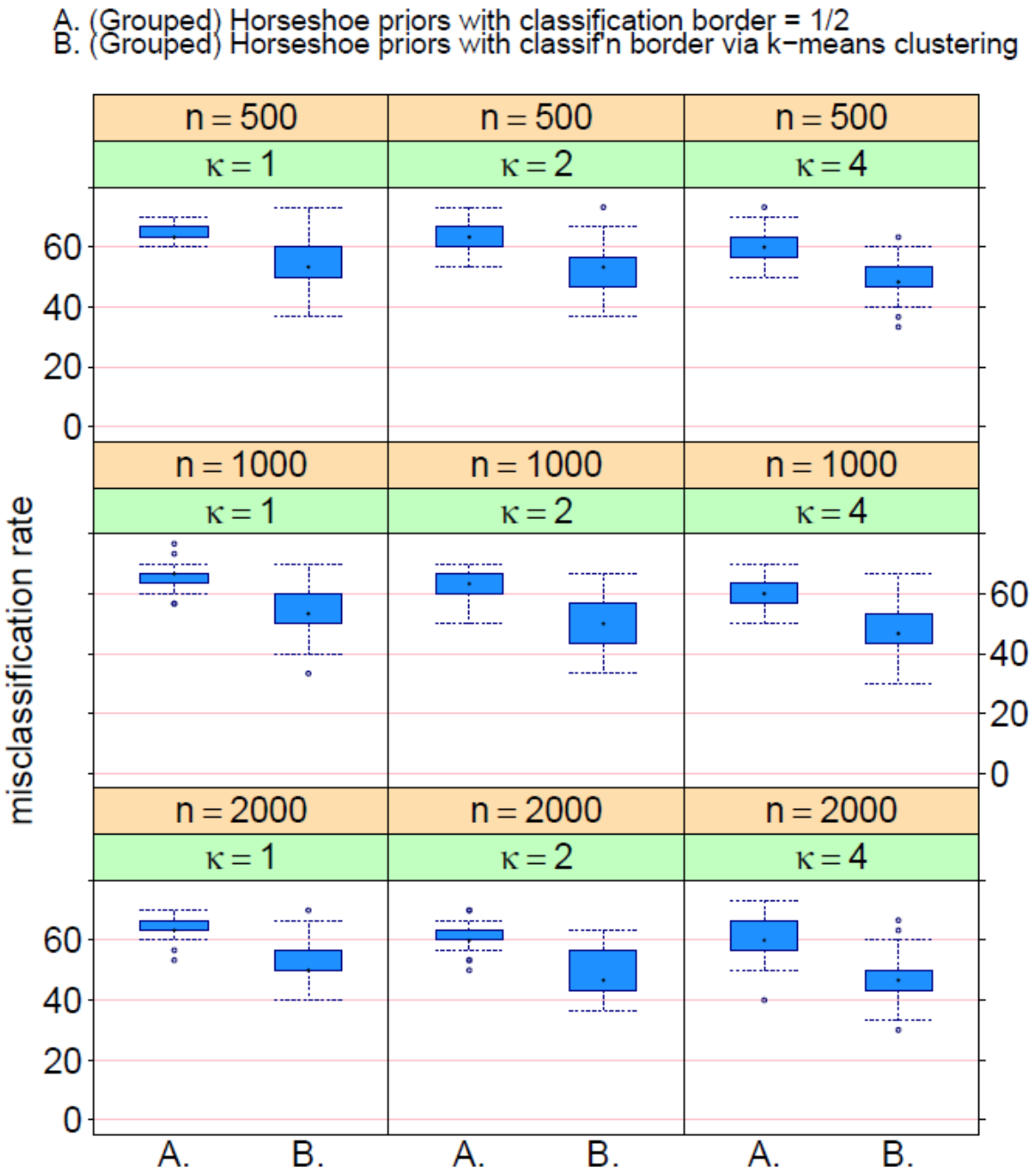


Figure 3.10: *Side-by-side boxplots of the misclassification rates for the comparative performance simulation study described in the text in the case of the response variable being Negative-Binomial. Each panel corresponds to a different combination of sample size and shape parameter. Within each panel, the side-by-side boxplots compare the misclassification rate across each of two methods: A. (Grouped) Horseshoe priors with a classification border at $\frac{1}{2}$, B. (Grouped) Horseshoe priors with classification border determined via k-means clustering.*

3.A Derivation of Result 3.2.2.1

From (3.1), for $x \in \mathbb{R}^d$, we have

$$\mathfrak{p}_{\mathrm{HS,d}}(\boldsymbol{x}) = \int_0^\infty \mathfrak{p}(\boldsymbol{x}|\lambda)\mathfrak{p}(\lambda)d\lambda = \int_0^\infty \left|2\pi\lambda^2 \boldsymbol{I}_d\right|^{-\frac{1}{2}} \exp\left(-\frac{\|\boldsymbol{x}\|^2}{2\lambda^2}\right) \frac{2}{\pi(1+\lambda^2)} d\lambda$$

$$= \left(2\pi\right)^{-\frac{d}{2}} \int_0^\infty \lambda^{-d} \exp\left(-\frac{\|\boldsymbol{x}\|^2}{2\lambda^2}\right) \frac{2}{\pi(1+\lambda^2)} d\lambda.$$

The change of variable $t \equiv \dfrac{1}{\lambda^2}$, leads to $\lambda = (t)^{-\frac{1}{2}}$, $d\lambda = -\frac{1}{2}(t)^{-\frac{3}{2}}\, dt$ and

$$\mathfrak{p}_{\mathrm{HS,d}}(\boldsymbol{x}) = \left(2\pi\right)^{-\frac{d}{2}} \int_\infty^0 \left((t)^{-\frac{1}{2}}\right)^{-d} \exp\left(-\frac{\|\boldsymbol{x}\|^2}{2}t\right) \frac{2}{\pi\left(\frac{1}{t}+1\right)} \left(-\frac{1}{2}(t)^{-\frac{3}{2}}\right) dt$$

$$= \left(2\pi\right)^{-\frac{d}{2}} \frac{1}{\pi} \int_0^\infty \left(t\right)^{\frac{d+1}{2}-1} \exp\left(-\frac{\|\boldsymbol{x}\|^2}{2}t\right) \frac{1}{t+1} dt$$

$$= \left(2\pi\right)^{-\frac{d}{2}} \frac{1}{\pi} \Gamma\left(\frac{d+1}{2}\right) \exp\left(\frac{\|\boldsymbol{x}\|^2}{2}\right) \left(\frac{\|\boldsymbol{x}\|^2}{2}\right)^{\frac{1-d}{2}}$$

$$\times \frac{\left(\frac{\|\boldsymbol{x}\|^2}{2}\right)^{\frac{d+1}{2}-1} \exp\left(-\frac{\|\boldsymbol{x}\|^2}{2}\right)}{\Gamma\left(\frac{d+1}{2}\right)} \int_0^\infty \frac{(t)^{\frac{d+1}{2}-1} \exp\left(-\frac{\|\boldsymbol{x}\|^2}{2}t\right)}{1+t} dt. \tag{3.80}$$

For the following term stated in the last step of (3.80)

$$\frac{\left(\frac{\|\boldsymbol{x}\|^2}{2}\right)^{\frac{d+1}{2}-1} \exp\left(-\frac{\|\boldsymbol{x}\|^2}{2}\right)}{\Gamma\left(\frac{d+1}{2}\right)} \int_0^\infty \frac{(t)^{\frac{d+1}{2}-1} \exp\left(-\frac{\|\boldsymbol{x}\|^2}{2}t\right)}{1+t} dt,$$

we use 8.19.4 of Olver et al. (2023) as follows:

$$E_p(z) = \frac{z^{p-1}\exp(-z)}{\Gamma(p)} \int_0^\infty \frac{t^{p-1}\exp(-zt)}{1+t} dt.$$

We then have

$$\begin{aligned}
\mathfrak{p}_{\mathrm{HS,d}}(\boldsymbol{x}) &= \left(2\pi\right)^{-\frac{d}{2}} \frac{1}{\pi} \Gamma\left(\frac{d+1}{2}\right) \exp\left(\frac{\left\|\boldsymbol{x}\right\|^2}{2}\right) \left(\frac{\left\|\boldsymbol{x}\right\|^2}{2}\right)^{\frac{1-d}{2}} E_{\left(\frac{d+1}{2}\right)}\left(\frac{\left\|\boldsymbol{x}\right\|^2}{2}\right) \\
&= \frac{\Gamma\left(\frac{d+1}{2}\right)}{\sqrt{2\pi^{d+2}}} \exp\left(\frac{\left\|\boldsymbol{x}\right\|^2}{2}\right) E_{\left(\frac{d+1}{2}\right)}\left(\frac{\left\|\boldsymbol{x}\right\|^2}{2}\right) \frac{1}{\left\|\boldsymbol{x}\right\|^{(d-1)}}.
\end{aligned}$$

3.B Derivation of Result 3.2.3.1

We will look at the derivation of **Result 3.2.3.1** by separately considering the following cases:

$$d = 1 \quad \text{and} \quad d \geq 2.$$

The $d = 1$ Case

From Theorem 1 of Carvalho et al. (2010), we have

$$\mathfrak{p}_{\mathrm{HS},1}(x) > \frac{K_1}{2} \log\left(1 + \frac{4}{x^2}\right) \quad \text{where} \quad K_1 \equiv \frac{1}{\sqrt{2\pi^3}}.$$

Then

$$\lim_{x \to 0} \mathfrak{p}_{\mathrm{HS},1}(x) > \frac{K_1}{2} \log\left(1 + 4 \lim_{x \to 0}\left(\frac{1}{x^2}\right)\right) = \infty.$$

Therefore

$$\lim_{x \to 0} \mathfrak{p}_{\mathrm{HS},1}(x) = \infty$$

and **Result 3.2.3.1** holds for $d = 1$.

The $d \geq 2$ Case

From **Result 3.2.2.1**, we have

$$\mathfrak{p}_{\mathrm{HS,d}}(\boldsymbol{x}) = K_d \exp\left(\frac{\left\|\boldsymbol{x}\right\|^2}{2}\right) E_{\left(\frac{d+1}{2}\right)}\left(\frac{\left\|\boldsymbol{x}\right\|^2}{2}\right) \frac{1}{\left\|\boldsymbol{x}\right\|^{(d-1)}} \quad \text{where} \quad K_d \equiv \frac{\Gamma\left(\frac{1}{2}(d+1)\right)}{\sqrt{2\pi^{d+2}}}.$$

Then

$$\lim_{\boldsymbol{x} \to 0} \mathfrak{p}_{\mathrm{HS,d}}(\boldsymbol{x}) = K_d \left\{\lim_{\boldsymbol{x} \to 0} \exp\left(\frac{\left\|\boldsymbol{x}\right\|^2}{2}\right)\right\} \left\{\lim_{\boldsymbol{x} \to 0} E_{\left(\frac{d+1}{2}\right)}\left(\frac{\left\|\boldsymbol{x}\right\|^2}{2}\right)\right\} \left\{\lim_{\boldsymbol{x} \to 0} \left\|\boldsymbol{x}\right\|^{1-d}\right\}.$$

Note that

$$\lim_{\boldsymbol{x} \to 0} \exp\left(\frac{\left\|\boldsymbol{x}\right\|^2}{2}\right) = 1.$$

From 8.19.6 of Olver et al. (2023) as follows:

$$E_p(0) = \frac{1}{p-1}, \quad \text{for all } p > 1,$$

we then have

$$\lim_{\boldsymbol{x}\to 0} E_{\left(\frac{d+1}{2}\right)}\left(\frac{\left\|\boldsymbol{x}\right\|^2}{2}\right) = \frac{1}{\frac{1}{2}(d+1)-1} = \frac{2}{d-1} \in [0,\infty) \quad \text{for all} \quad d \geq 2.$$

Also note that

$$\lim_{\boldsymbol{x}\to 0} \left\|\boldsymbol{x}\right\|^{1-d} = \infty \quad \text{for all} \quad d \geq 2.$$

Therefore

$$\lim_{\boldsymbol{x}\to 0} \mathfrak{p}_{\mathrm{HS,d}}(\boldsymbol{x}) = \infty \quad \text{for all} \quad d \geq 2.$$

3.C Derivation of Result 3.2.4.1

Result 3.2.4.1 implies various properties of the bivariate confluent hypergeometric function. We will start with its definition and key results. Then we show how these results link to our **Result 3.2.4.1** statement.

3.C.1 The Bivariate Confluent Hypergeometric Function

The *bivariate confluent hypergeometric function*

$$\Phi_1(\alpha,\beta,\gamma,x,y) \quad \text{for} \quad \alpha,\beta,\gamma,x,y,\in\mathbb{C}$$

is defined via a pair of partial differential equations in Section 9.262 of Gradshteyn and Ryzhik (1994). In Appendix 3.C.1, for demonstrative purposes, $\Phi_1(\alpha,\beta,\gamma,x,y)$ defined in Gradshteyn and Ryzhik (1994) is denoted as $\Phi_{1\_\text{GR}}(\alpha,\beta,\gamma,x,y)$ and $\Phi_1(\alpha,\beta,\gamma,x,y)$ defined in Gordy (1998) is denoted as $\Phi_{1\_\text{Gordy}}(\alpha,\beta,\gamma,x,y)$.

As stated in Section 9.261 of Gradshteyn and Ryzhik (1994) the following series representation applies when $|x|<1$:

$$\Phi_{1\_\text{GR}}(\alpha,\beta,\gamma,x,y)=\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\frac{(\alpha)_{m+n}(\beta)_m x^m y^n}{(\gamma)_{m+n}\,m!n!} \quad \text{where} \quad (\alpha)_k\equiv\Gamma(a+k)\,/\Gamma(\alpha). \tag{3.81}$$

Result 3.385 of Gradshteyn and Ryzhik (1994) states that

$$\int_0^{\infty}x^{\nu-1}(1-x)^{\lambda-1}(1-\beta x)^{-\rho}\exp(-\mu x)\,dx=\frac{\Gamma(\nu)\Gamma(\lambda)}{\Gamma(\nu+\lambda)}\Phi_1(\nu,\rho,\nu+\lambda,\beta,-\mu) \tag{3.82}$$

for complex numbers λ,ν,ρ,β and μ ranging over various subsets of the complex plane. If these parameters are restricted to be real numbers then the restrictions reduce to

$$\lambda,\nu>0 \quad \text{and} \quad \beta,\rho,\mu\in\mathbb{R}.$$

Equation (6) of Gordy (1998) is given as follows:

$$\Phi_{1\_\text{Gordy}}(\alpha,\beta,\gamma,x,y)=\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\frac{(\alpha)_{m+n}(\beta)_n x^m y^n}{(\gamma)_{m+n}\,m!n!}.$$

The error in equation (6) of Gordy (1998) is $(\beta)_m$ stated in (3.81) being wrongly replaced by $(\beta)_n$, even though Result 1 of Section 9.261 of an earlier edition of Grad-

shteyn and Ryzhik (1994) is given as a reference for the series representation of Φ_1. This error leads to the roles of x and y being reversed between Φ_{1_Gordy} and Φ_{1_GR}. Therefore

$$\Phi_{1\_\text{GR}}(\alpha, \beta, \gamma, x, y) = \Phi_{1\_\text{Gordy}}(\alpha, \beta, \gamma, y, x). \tag{3.83}$$

For $0 \leq x < 1$ and $0 < \alpha < \gamma$, (T4) of Gordy (1998) is the following series representation of Φ_1 in terms of univariate confluent hypergeometric function ${}_1F_1$:

$$\Phi_{1\_\text{Gordy}}(\alpha, \beta, \gamma, x, y) = \exp(x) \sum_{n=0}^{\infty} \frac{(\alpha)_n (\beta)_n y^n}{(\gamma)_n n!} {}_1F_1(\gamma - \alpha, \gamma + n, -x). \tag{3.84}$$

We refer to Result 1 of Section 9.261 of Gradshteyn and Ryzhik (1994) and use Φ_{1_GR} for Φ_1 in the rest of this thesis. According to (3.83), we have the following series representation of Φ_1 in terms of the univariate confluent hypergeometric function ${}_1F_1$:

$$\Phi_1(\alpha, \beta, \gamma, x, y) = \exp(y) \sum_{n=0}^{\infty} \frac{(\alpha)_n (\beta)_n x^n}{(\gamma)_n n!} {}_1F_1(\gamma - \alpha, \gamma + n, -y). \tag{3.85}$$

The error in equation (6) of Gordy (1998) leads to (T1) – (T4) of Gordy (1998) containing an incorrect variant of (3.85) with respect to the Φ_1, ${}_1F_1$ and ${}_2F_1$ functions as defined in Gradshteyn and Ryzhik (1994).

3.C.2 Marginal Density Function Simplification

The marginal density function of $\boldsymbol{y}$ according to model (3.2) is

$$\mathfrak{p}(\boldsymbol{y}) = \int_0^{\infty} \left\{ \int_{\mathbb{R}^d} \mathfrak{p}(\boldsymbol{y}|\boldsymbol{\theta}) \mathfrak{p}(\boldsymbol{\theta}|\lambda) d\boldsymbol{\theta} \right\} \mathfrak{p}(\lambda) d\lambda \quad \text{where} \quad \mathfrak{p}(\lambda) = \frac{2I(\lambda > 0)}{\pi(\lambda^2 + 1)}. \tag{3.86}$$

Note that model (3.2) is equivalent to the set-up as follows:

$$\boldsymbol{y}|\boldsymbol{\theta} \sim N(\boldsymbol{\theta}, \boldsymbol{I}_d), \quad \boldsymbol{\theta}|\lambda \sim N\left(\mathbf{0}, \lambda^2 \tau^2 \boldsymbol{I}_d\right), \quad \lambda \sim \text{Half-Cauchy}(1), \tag{3.87}$$

for some $\tau > 0$ that is fixed and known. If we define $\xi \equiv 1 + 1/(\lambda^2\tau^2)$ and the symmetric matrix $\boldsymbol{A} \equiv \xi \boldsymbol{I}_d$, then $\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\theta})\mathfrak{p}(\boldsymbol{\theta}|\lambda)$, stated in (3.86) for $\mathfrak{p}(\boldsymbol{y})$, becomes

$$\begin{aligned}\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\theta})\mathfrak{p}(\boldsymbol{\theta}|\lambda) &= (2\pi\tau\lambda)^{-d}\exp\left(-\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{\theta}\|^2 - \frac{\|\boldsymbol{\theta}\|}{2\lambda^2\tau^2}\right)\\ &= (2\pi\tau\lambda)^{-d}\exp\left(-\frac{1}{2}\|\boldsymbol{y}\|^2 + \boldsymbol{y}^T\boldsymbol{\theta} - \frac{1}{2}\boldsymbol{\theta}^T\left(1+\frac{1}{\lambda^2\tau^2}\boldsymbol{I}_d\boldsymbol{\theta}\right)\right)\\ &= (2\pi\tau\lambda)^{-d}\exp\left(-\frac{1}{2}\|\boldsymbol{y}\|^2\right)\exp\left(\boldsymbol{y}^T\boldsymbol{\theta} - \frac{1}{2}\boldsymbol{\theta}^T(\xi\boldsymbol{I}_d)\boldsymbol{\theta}\right)\end{aligned}$$

and

$$\boldsymbol{y}^T\boldsymbol{\theta} - \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{A}\boldsymbol{\theta} = \frac{1}{2}\boldsymbol{y}^T\boldsymbol{A}^{-1}\boldsymbol{y} - \frac{1}{2}\left(\boldsymbol{\theta}-\boldsymbol{A}^{-1}\boldsymbol{y}\right)^T\boldsymbol{A}\left(\boldsymbol{\theta}-\boldsymbol{A}^{-1}\boldsymbol{y}\right).$$

Therefore, the term $\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\theta})\mathfrak{p}(\boldsymbol{\theta}|\lambda)$, stated in (3.86) for $\mathfrak{p}(\boldsymbol{y})$, becomes

$$\begin{aligned}\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\theta})\mathfrak{p}(\boldsymbol{\theta}|\lambda) &= \Big(2\pi\tau\lambda\Big)^{-d}\exp\left(-\frac{1}{2}\|\boldsymbol{y}\|^2\right)\\ &\quad\times\exp\left(\frac{1}{2}\left(\frac{1}{\xi}\right)\|\boldsymbol{y}\|^2 - \frac{1}{2}\left(\boldsymbol{\theta}-\left(\frac{1}{\xi}\right)\boldsymbol{y}\right)^T\left(\left(\frac{1}{\xi}\right)\boldsymbol{I}_d\right)^{-1}\left(\boldsymbol{\theta}-\left(\frac{1}{\xi}\right)\boldsymbol{y}\right)\right)\\ &= \Big(2\pi\Big)^{-\frac{d}{2}}\Big(\tau\lambda\Big)^{-d}\exp\left(\frac{1}{2}\left(\frac{1}{\xi}-1\right)\|\boldsymbol{y}\|^2\right)\left|\left(\frac{1}{\xi}\right)\boldsymbol{I}_d\right|^{\frac{1}{2}}\Big(2\pi\Big)^{-\frac{d}{2}}\left|\left(\frac{1}{\xi}\right)\boldsymbol{I}_d\right|^{-\frac{1}{2}}\\ &\quad\times\exp\left(\left(\boldsymbol{\theta}-\left(\frac{1}{\xi}\right)\boldsymbol{y}\right)^T\left(\left(\frac{1}{\xi}\right)\boldsymbol{I}_d\right)^{-1}\left(\boldsymbol{\theta}-\left(\frac{1}{\xi}\right)\boldsymbol{y}\right)\right).\end{aligned} \tag{3.88}$$

Note that

$$\Big(2\pi\Big)^{-\frac{d}{2}}\left|\left(\frac{1}{\xi}\right)\boldsymbol{I}_d\right|^{-\frac{1}{2}}\exp\left(\left(\boldsymbol{\theta}-\left(\frac{1}{\xi}\right)\boldsymbol{y}\right)^T\left(\left(\frac{1}{\xi}\right)\boldsymbol{I}_d\right)^{-1}\left(\boldsymbol{\theta}-\left(\frac{1}{\xi}\right)\boldsymbol{y}\right)\right)$$

is the $N\left(\left(1/\xi\right)\boldsymbol{y}, \left(1/\xi\right)\boldsymbol{I}_d\right)$ density function in $\boldsymbol{\theta}$ and $\left|\left(1/\xi\right)\boldsymbol{I}_d\right|^{\frac{1}{2}} = \left(1/\xi\right)^{\frac{d}{2}}$. The term $\int_{\mathbb{R}^d}\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\theta})\mathfrak{p}(\boldsymbol{\theta}|\lambda)d\boldsymbol{\theta}$, stated in (3.86) for $\mathfrak{p}(\boldsymbol{y})$, becomes

$$\int_{\mathbb{R}^d}\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\theta})\mathfrak{p}(\boldsymbol{\theta}|\lambda)d\boldsymbol{\theta} = \Big(2\pi\Big)^{-\frac{d}{2}}\Big(\tau\lambda\Big)^{-d}\exp\left(\frac{1}{2}\left(\frac{1}{\xi}-1\right)\|\boldsymbol{y}\|^2\right)\left(\frac{1}{\xi}\right)^{\frac{d}{2}}.$$

Recalling that $\xi = \dfrac{\lambda^2\tau^2+1}{\lambda^2\tau^2}$, $\dfrac{1}{\xi} = \dfrac{\lambda^2\tau^2}{\lambda^2\tau^2+1}$ and $\dfrac{1}{\xi} - 1 = -\dfrac{1}{\lambda^2\tau^2+1}$, we then have

$$\int_{\mathbb{R}^d} \mathfrak{p}(\boldsymbol{y}|\boldsymbol{\theta})\mathfrak{p}(\boldsymbol{\theta}|\lambda)d\boldsymbol{\theta} = \Big(2\pi\Big)^{-\frac{d}{2}} \exp\left(-\left(\frac{\frac{\|\boldsymbol{y}\|^2}{2}}{\lambda^2\tau^2+1}\right)\right) \frac{1}{(\lambda^2\tau^2+1)^{\frac{d}{2}}}.$$

Moreover, since $\lambda \sim$ Half-Cauchy(1) and $\mathfrak{p}(\lambda) = \dfrac{2}{\pi\,(1+\lambda^2)}$, (3.86) becomes

$$\mathfrak{p}(\boldsymbol{y}) = \Big(\big(2\big)^{d-2}\big(\pi\big)^{d+2}\Big)^{-\frac{1}{2}} \int_0^\infty \exp\left(-\left(\frac{\frac{\|\boldsymbol{y}\|^2}{2}}{\lambda^2\tau^2+1}\right)\right) \frac{1}{(\lambda^2\tau^2+1)^{\frac{d}{2}}\,(1+\lambda^2)} d\lambda. \quad (3.89)$$

3.C.3 Score Function Simplification

The score function is given as the following $d \times 1$ derivative vector:

$$\nabla_{\boldsymbol{y}} \left\{\log \mathfrak{p}(\boldsymbol{y})\right\} = \begin{bmatrix} \dfrac{\partial \log\left\{\mathfrak{p}(\boldsymbol{y})\right\}}{\partial y_1} \\ \vdots \\ \dfrac{\partial \log\left\{\mathfrak{p}(\boldsymbol{y})\right\}}{\partial y_d} \end{bmatrix}. \quad (3.90)$$

The marginal density function $\mathfrak{p}(\boldsymbol{y})$ stated in (3.89) in Appendix 3.C.2 is

$$\mathfrak{p}(\boldsymbol{y}) = \Big(\big(2\big)^{d-2}\big(\pi\big)^{d+2}\Big)^{-\frac{1}{2}} \int_0^\infty \exp\left(-\left(\frac{\frac{\|\boldsymbol{y}\|^2}{2}}{\lambda^2\tau^2+1}\right)\right) \frac{1}{(\lambda^2\tau^2+1)^{\frac{d}{2}}\,(1+\lambda^2)} d\lambda.$$

Observe that

$$d_{\boldsymbol{y}}\mathfrak{p}(\boldsymbol{y}) = \Big(\big(2\big)^{d-2}\big(\pi\big)^{d+2}\Big)^{-\frac{1}{2}} \int_0^\infty d_{\boldsymbol{y}} \exp\left(-\left(\frac{\frac{\|\boldsymbol{y}\|^2}{2}}{\lambda^2\tau^2+1}\right)\right) \frac{1}{(\lambda^2\tau^2+1)^{\frac{d}{2}}\,(1+\lambda^2)} d\lambda.$$

We then have

$$
\begin{aligned}
d_{\boldsymbol{y}} \log \mathfrak{p}(\boldsymbol{y}) &= \frac{1}{\mathfrak{p}(\boldsymbol{y})} d_{\boldsymbol{y}} \mathfrak{p}(\boldsymbol{y}) \\
&= \frac{\left((2)^{d-2} (\pi)^{d+2} \right)^{-\frac{1}{2}} \int_0^{\infty} d_{\boldsymbol{y}} \exp\left(- \left(\frac{\frac{\|\boldsymbol{y}\|^2}{2}}{\lambda^2 \tau^2 + 1} \right) \right) \frac{1}{(\lambda^2 \tau^2 + 1)^{\frac{d}{2}} (1 + \lambda^2)} d\lambda}{\left((2)^{d-2} (\pi)^{d+2} \right)^{-\frac{1}{2}} \int_0^{\infty} \exp\left(- \left(\frac{\frac{\|\boldsymbol{y}\|^2}{2}}{\lambda^2 \tau^2 + 1} \right) \right) \frac{1}{(\lambda^2 \tau^2 + 1)^{\frac{d}{2}} (1 + \lambda^2)} d\lambda}.
\end{aligned}
\tag{3.91}
$$

Noting that

$$
d_{\boldsymbol{y}} \left(\|\boldsymbol{y}\|^2 \right) = d_{\boldsymbol{y}} \left(\boldsymbol{y}^T \boldsymbol{y} \right) = \left[\frac{d\left(y_1^2 + \ldots + y_n^2\right)}{dy_1}, \ldots, \frac{d\left(y_1^2 + \ldots + y_n^2\right)}{dy_n} \right] = 2\boldsymbol{y}
$$

and

$$
d_{\boldsymbol{y}} \exp\left(- \left(\frac{\frac{\|\boldsymbol{y}\|^2}{2}}{\lambda^2 \tau^2 + 1} \right) \right) = -\boldsymbol{y} \exp\left(- \left(\frac{\frac{\|\boldsymbol{y}\|^2}{2}}{\lambda^2 \tau^2 + 1} \right) \right) \frac{1}{\lambda^2 \tau^2 + 1} d\boldsymbol{y},
$$

then the integral in the numerator of $d_{\boldsymbol{y}} \log \mathfrak{p}(\boldsymbol{y})$ stated in (3.91) becomes

$$
\begin{aligned}
&\int_0^{\infty} d_{\boldsymbol{y}} \exp\left(- \left(\frac{\frac{\|\boldsymbol{y}\|^2}{2}}{\lambda^2 \tau^2 + 1} \right) \right) \frac{1}{(\lambda^2 \tau^2 + 1)^{\frac{d}{2}} (1 + \lambda^2)} d\lambda \\
&\quad = \int_0^{\infty} \exp\left(- \left(\frac{\frac{\|\boldsymbol{y}\|^2}{2}}{\lambda^2 \tau^2 + 1} \right) \right) \frac{1}{(\lambda^2 \tau^2 + 1)^{\frac{d+2}{2}} (1 + \lambda^2)} d\lambda d\boldsymbol{y} \left(-\boldsymbol{y} \right).
\end{aligned}
$$

Therefore, $d_{\boldsymbol{y}} \log \mathfrak{p}(\boldsymbol{y})$ stated in (3.91) becomes

$$d_{\boldsymbol{y}} \log \mathfrak{p}(\boldsymbol{y}) = -\boldsymbol{y} \frac{\int_0^\infty \exp\left(-\left(\frac{\frac{\|\boldsymbol{y}\|^2}{2}}{\lambda^2\tau^2+1}\right)\right) \frac{1}{(\lambda^2\tau^2+1)^{\frac{d+2}{2}}(1+\lambda^2)} d\lambda}{\int_0^\infty \exp\left(-\left(\frac{\frac{\|\boldsymbol{y}\|^2}{2}}{\lambda^2\tau^2+1}\right)\right) \frac{1}{(\lambda^2\tau^2+1)^{\frac{d}{2}}(1+\lambda^2)} d\lambda} d\boldsymbol{y}.$$

Hence, the score function stated in (3.90) becomes

$$\nabla_{\boldsymbol{y}} \{\log \mathfrak{p}(\boldsymbol{y})\} = -\boldsymbol{y} \left[\frac{\int_0^\infty \exp\left(-\left(\frac{\frac{\|\boldsymbol{y}\|^2}{2}}{\lambda^2\tau^2+1}\right)\right) \frac{1}{(\lambda^2\tau^2+1)^{\frac{d+2}{2}}(1+\lambda^2)} d\lambda}{\int_0^\infty \exp\left(-\left(\frac{\frac{\|\boldsymbol{y}\|^2}{2}}{\lambda^2\tau^2+1}\right)\right) \frac{1}{(\lambda^2\tau^2+1)^{\frac{d}{2}}(1+\lambda^2)} d\lambda} \right]. \tag{3.92}$$

3.C.4 Bivariate Confluent Hypergeometric Function Representations

In this subsection, we derive expressions for marginal density function $\mathfrak{p}(\boldsymbol{y})$ and score function $\nabla_{\boldsymbol{y}} \{\log \mathfrak{p}(\boldsymbol{y})\}$ using the bivariate confluent hypergeometric function Φ_1 as defined in Appendix 3.C.1. The marginal density function $\mathfrak{p}(\boldsymbol{y})$ is stated in (3.89) in Appendix 3.C.2. The integral stated in (3.89) is as follows:

$$\mathcal{C}(a,b) = \mathcal{C}\left(\tfrac{1}{2}\|\boldsymbol{y}\|^2, \tau^2\right)$$

where

$$\mathcal{C}(a,b) \equiv \int_0^\infty \exp\left(-\frac{a}{1+\lambda^2 b}\right) \frac{1}{(1+\lambda^2 b)^{\frac{d}{2}}(\lambda^2+1)} d\lambda. \tag{3.93}$$

The change of variable

$$x \equiv \frac{\lambda^2 b}{1+\lambda^2 b}$$

in (3.93) leads to $\lambda^2 = \dfrac{x}{b\,(1-x)}$, $d\lambda = \dfrac{1}{\sqrt{b}}\dfrac{1}{2\sqrt{-\frac{x}{x-1}}\,(1-x)^2}dx$ and

$$\mathcal{C}\,(a,b) = \int_0^1 \exp\Big(-a+ax\Big)\frac{1}{\left(\frac{1}{1-x}\right)^{\frac{d}{2}}\left(\frac{b-x(b-1)}{b(1-x)}\right)}\frac{1}{\sqrt{b}}\frac{1}{2\sqrt{-\frac{x}{x-1}}\,(1-x)^2}dx$$

$$= \frac{\exp(-a)}{2\sqrt{b}}\int_0^1 \Big(x\Big)^{\left(\frac{1}{2}-1\right)}\frac{\Big(1-x\Big)^{\left(\frac{d+1}{2}-1\right)}}{\Big(1-\left(1-\frac{1}{b}\right)x\Big)}\exp\Big(-(-ax)\Big)dx. \tag{3.94}$$

For the integral stated in the last step of (3.94), we use 3.385 from Gradshteyn and Ryzhik (1994) stated in (3.82) in Appendix 3.C.1 as follows:

$$\int_0^1 x^{\nu-1}\,(1-x)^{\lambda-1}\,(1-\beta x)^{-\rho}\,e^{-\mu x}dx = B\,(\nu,\lambda)\,\Phi_1\,(\nu,\rho,\nu+\lambda,\beta,-\mu)\,.$$

In our case, we have $\nu = 1/2$, $\lambda = (d+1)/2$, $\beta = 1-1/b$, $\rho = 1$, $\nu+\lambda = (d+2)/2$ and $\mu = -a$. Then $\mathcal{C}\,(a,b)$ stated in (3.94) becomes

$$\mathcal{C}\,(a,b) = \frac{\exp(-a)\,\sqrt{\pi}\Gamma\left(\dfrac{d}{2}+\dfrac{1}{2}\right)}{d\sqrt{b}\Gamma\left(\dfrac{d}{2}\right)}\Phi_1\left(\frac{1}{2},1,\frac{d+2}{2},1-\frac{1}{b},a\right). \tag{3.95}$$

Note that $a = \frac{1}{2}\|\boldsymbol{y}\|^2$ and $b = \tau^2$. Then $\mathcal{C}\,(a,b)$, the integral stated in (3.89), becomes

$$\mathcal{C}\,(a,b) = \int_0^\infty \exp\left(-\left(\frac{\frac{\|\boldsymbol{y}\|^2}{2}}{\lambda^2\tau^2+1}\right)\right)\frac{1}{(\lambda^2\tau^2+1)^{\frac{d}{2}}\,(1+\lambda^2)}d\lambda$$

$$= \frac{\exp\left(-\frac{1}{2}\|\boldsymbol{y}\|^2\right)\sqrt{\pi}\Gamma\left(\dfrac{d}{2}+\dfrac{1}{2}\right)}{d\tau\Gamma\left(\dfrac{d}{2}\right)}\Phi_1\left(\frac{1}{2},1,\frac{d+2}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right).$$

Therefore, the marginal density function stated in (3.89) becomes

$$\begin{aligned}
&\mathfrak{p}(\boldsymbol{y}) \\
&= \left((2)^{d-2}(\pi)^{d+2}\right)^{-\frac{1}{2}} \frac{\exp\left(-\frac{1}{2}\|\boldsymbol{y}\|^2\right)\sqrt{\pi}\Gamma\left(\frac{d}{2}+\frac{1}{2}\right)}{d\tau\Gamma\left(\frac{d}{2}\right)}\Phi_1\left(\frac{1}{2},1,\frac{d+2}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right) \\
&= \left((2)^{d-2}(\pi)^{d+1}\right)^{-\frac{1}{2}} \frac{\exp\left(-\frac{1}{2}\|\boldsymbol{y}\|^2\right)\Gamma\left(\frac{d}{2}+\frac{1}{2}\right)}{d\tau\Gamma\left(\frac{d}{2}\right)}\Phi_1\left(\frac{1}{2},1,\frac{d+2}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right).
\end{aligned} \tag{3.96}$$

For the $d=1$ case, (3.96) reduces to a similar expression, but not identical to equation (A1) of Carvalho et al. (2010). Equation (A1) of Carvalho et al. (2010) is as follows:

$$m(y) = \frac{2}{\tau\left(2\pi^3\right)^{\frac{1}{2}}}\exp\left(-\frac{y^2}{2}\right)\Phi_1\left(\frac{1}{2},1,\frac{5}{2},\frac{y^2}{2},1-\frac{1}{\tau^2}\right),$$

with $m(y)$ denoted the marginal density. The main difference is the order of the fourth and fifth arguments of the Φ_1 function. This discrepancy is attributable to an error in Gordy (1998) described in Appendix 3.C.1.

Next we seek an analogous expression for the score function $\nabla_{\boldsymbol{y}}\{\log\mathfrak{p}(\boldsymbol{y})\}$ stated in (3.92) in Appendix 3.C.3. The analogous expression is as follows:

$$\nabla_{\boldsymbol{y}}\{\log\mathfrak{p}(\boldsymbol{y})\} = -\boldsymbol{y}\frac{\mathcal{D}\left(\frac{1}{2}\|\boldsymbol{y}\|^2,\tau^2\right)}{\mathcal{C}\left(\frac{1}{2}\|\boldsymbol{y}\|^2,\tau^2\right)} \tag{3.97}$$

where

$$\mathcal{D}(a,b) \equiv \int_0^\infty \exp\left(-\left(\frac{\frac{\|\boldsymbol{y}\|^2}{2}}{\lambda^2\tau^2+1}\right)\right)\frac{1}{(\lambda^2\tau^2+1)^{\frac{d}{2}+1}(1+\lambda^2)}d\lambda.$$

Calculations of $\mathcal{D}(a,b)$ are similar to those given in the previous section for $\mathcal{C}(a,b)$. The change of variable

$$x = \frac{\lambda^2 b}{1+\lambda^2 b}$$

in $\mathcal{D}(a,b)$ leads to $\lambda^2 = \dfrac{x}{b(1-x)}$, $d\lambda = \dfrac{1}{\sqrt{b}} \dfrac{1}{2\sqrt{-\frac{x}{x-1}}(1-x)^2} dx$ and

$$\mathcal{D}(a,b) = \int_0^1 \exp\Big(-a+ax\Big) \frac{1}{\Big(\frac{1}{1-x}\Big)^{\frac{d}{2}+1}\Big(\frac{b-x(b-1)}{b(1-x)}\Big)} \frac{1}{\sqrt{b}} \frac{1}{2\sqrt{-\frac{x}{x-1}}(1-x)^2} dx$$

$$= \frac{\exp(-a)}{2\sqrt{b}} \int_0^1 \Big(x\Big)^{\left(\frac{1}{2}-1\right)} \frac{\Big(1-x\Big)^{\left(\frac{d+3}{2}-1\right)}}{\Big(1-\big(1-\frac{1}{b}\big)x\Big)} \exp\Big(-(-ax)\Big) dx. \tag{3.98}$$

For the integral stated in the last step of (3.98), we use 3.385 from Gradshteyn and Ryzhik (1994) stated in (3.82) in Appendix 3.C.1 as follows:

$$\int_0^1 x^{\nu-1}(1-x)^{\lambda-1}(1-\beta x)^{-\rho} e^{-\mu x} dx = B(\nu,\lambda)\,\Phi_1(\nu,\rho,\nu+\lambda,\beta,-\mu).$$

In our case, we have $\nu = 1/2$, $\lambda = (d+3)/2$, $\beta = 1-1/b$, $\rho = 1$, $\nu+\lambda = (d+4)/2$ and $\mu = -a$. Then $\mathcal{D}(a,b)$, the numerator of score function $\nabla_{\boldsymbol{y}}\{\log \mathfrak{p}(\boldsymbol{y})\}$, stated in (3.97), becomes

$$\mathcal{D}(a,b) = \frac{\exp(-a)}{2\sqrt{b}} \frac{\sqrt{\pi}\Gamma\left(\frac{d+3}{2}\right)}{\Gamma\left(\frac{1}{2}d+2\right)} \Phi_1\left(\frac{1}{2}, 1, \frac{d+4}{2}, 1-\frac{1}{b}, a\right)$$

$$= \frac{\exp(-a)}{2\sqrt{b}} \frac{\sqrt{\pi}\frac{1}{2}(d+1)\Gamma\left(\frac{d+1}{2}\right)}{\left(\frac{d}{2}+1\right)\Gamma\left(\frac{1}{2}d+1\right)} \Phi_1\left(\frac{1}{2}, 1, \frac{d+4}{2}, 1-\frac{1}{b}, a\right).$$

Note that $\mathcal{C}(a,b)$ stated in (3.95) in Appendix 3.C.4, which is the denominator of score function $\nabla_{\boldsymbol{y}}\{\log \mathfrak{p}(\boldsymbol{y})\}$ stated in (3.97), is as follows:

$$\mathcal{C}(a,b) = \frac{\exp(-a)\sqrt{\pi}\Gamma\left(\frac{d}{2}+\frac{1}{2}\right)}{d\sqrt{b}\Gamma\left(\frac{d}{2}\right)} \Phi_1\left(\frac{1}{2}, 1, \frac{d+2}{2}, 1-\frac{1}{b}, a\right).$$

Then the score function stated in (3.97) becomes

$$\nabla_{\boldsymbol{y}} \log\{\mathfrak{p}(\boldsymbol{y})\} = -\boldsymbol{y}\frac{\mathcal{D}(a,b)}{\mathcal{C}(a,b)} = -\boldsymbol{y}\frac{(d+1)\,\Phi_1\left(\frac{1}{2}, 1, \frac{d+4}{2}, 1-\frac{1}{\tau^2}, \frac{1}{2}\|\boldsymbol{y}\|^2\right)}{(d+2)\,\Phi_1\left(\frac{1}{2}, 1, \frac{d+2}{2}, 1-\frac{1}{\tau^2}, \frac{1}{2}\|\boldsymbol{y}\|^2\right)}. \quad (3.99)$$

For $d = 1$, the score function stated in (3.99) becomes

$$\nabla_{\boldsymbol{y}} \log\{\mathfrak{p}(\boldsymbol{y})\} = -\boldsymbol{y}\frac{2\Phi_1\left(\frac{1}{2}, 1, \frac{5}{2}, 1-\frac{1}{\tau^2}, \frac{1}{2}\|\boldsymbol{y}\|^2\right)}{3\Phi_1\left(\frac{1}{2}, 1, \frac{3}{2}, 1-\frac{1}{\tau^2}, \frac{1}{2}\|\boldsymbol{y}\|^2\right)},$$

matching equation (A2) of Carvalho et al. (2010) except for the interchanges in the fourth and the fifth arguments of the Φ_1 equation. An error in Gordy (1998), described in Appendix 3.C.1, provides an explanation for this discrepancy.

3.C.5 Large $\|\boldsymbol{y}\|^2$ Approximation of the Score Function

Result 3.2.4.1 in Section 3.2.4 concerns the limiting behaviour of the score function when $\|\boldsymbol{y}\| \to \infty$. To keep the consistency with **Result 3.2.4.1**, we assume that $\|\boldsymbol{y}\| \gg 1$. By the degree of complexity, we treat the following cases separately

$$\tau = 1, \quad \tau > 1 \quad \text{and} \quad 0 < \tau < 1.$$

In each case, we need to use the following results from Abramowitz and Stegun (1968), for the right-tail asymptotic behaviour of the univariate confluent hypergeometric function

$${}_1F_1(a,b,x) = \begin{cases} \dfrac{\Gamma(b)}{\Gamma(a)} e^x x^{a-b}\left\{1 + O\left(|x|^{-1}\right)\right\} & \text{for } x > 0 \quad \text{and } |x| \gg 1, \\ \dfrac{\Gamma(b)}{\Gamma(b-a)}(-x)^{-a}\left\{1 + O\left(|x|^{-1}\right)\right\} & \text{for } \quad x < 0 \quad \text{and } |x| \gg 1. \end{cases} \quad (3.100)$$

The $\tau = 1$ Case

Note that

$$\nabla_{\boldsymbol{y}} \log\{\mathfrak{p}(\boldsymbol{y})\} = -\boldsymbol{y}\frac{\mathcal{D}(a,b)}{\mathcal{C}(a,b)}$$

is stated in (3.97). Let us look into $\mathcal{D}(a,b)$ and $\mathcal{C}(a,b)$. $\tau = 1$ leads to $1-\frac{1}{b} = 1-\frac{1}{\tau^2} = 0$. Then $\mathcal{D}(a,b)$ stated in (3.98) in Appendix 3.C.4 becomes

$$\begin{aligned}\mathcal{D}(a,b) &= \frac{\exp(-a)}{2\sqrt{b}} \int_0^1 \Big(x\Big)^{\left(\frac{1}{2}-1\right)} \frac{\Big(1-x\Big)^{\left(\frac{d+3}{2}-1\right)}}{\left(1-\left(1-\frac{1}{b}\right)x\right)} \exp\Big(-(-ax)\Big)dx \\ &= \frac{\exp(-a)}{2\sqrt{b}} \int_0^1 \Big(x\Big)^{\left(\frac{1}{2}-1\right)} \Big(1-x\Big)^{\left(\frac{d+3}{2}-1\right)} \exp(ax)\,dx \end{aligned} \tag{3.101}$$

and $\mathcal{C}(a,b)$ stated in (3.94) in Appendix 3.C.4 becomes

$$\begin{aligned}\mathcal{C}(a,b) &= \frac{\exp(-a)}{2\sqrt{b}} \int_0^1 \Big(x\Big)^{\left(\frac{1}{2}-1\right)} \frac{\Big(1-x\Big)^{\left(\frac{d+1}{2}-1\right)}}{\left(1-\left(1-\frac{1}{b}\right)x\right)} \exp\Big(-(-ax)\Big)dx \\ &= \frac{\exp(-a)}{2\sqrt{b}} \int_0^1 \Big(x\Big)^{\left(\frac{1}{2}-1\right)} \Big(1-x\Big)^{\left(\frac{d+1}{2}-1\right)} \exp\Big(ax\Big)dx. \end{aligned} \tag{3.102}$$

For the integrals stated in the last step of (3.101) and (3.102) , we use equation (1) in Section 3.383 in Gradshteyn and Ryzhik (1994) as follows:

$$\int_0^u x^{(\nu-1)}(u-x)^{(\mu-1)} e^{\beta x}dx = \mathrm{B}(\mu,\nu)\, u^{\mu+\nu-1}{}_1F_1(\nu;\mu+\nu;\beta u) \quad [\mathrm{Re}\,\mu>0,\ \mathrm{Re}\,\nu>0].$$

In our case, for $\mathcal{D}(a,b)$, we have $\nu = 1/2,\ \ u=1,\ \ \mu=(d+3)/2,\ \ \beta = a = \left\|\boldsymbol{y}\right\|^2/2$, $\mu+\nu = (d+4)/2,\ \ \beta u = \left\|\boldsymbol{y}\right\|^2/2$. Then $\mathcal{D}(a,b)$, in terms of univariate confluent hypergeometric function, becomes

$$\mathcal{D}(a,b) = \frac{\exp(-a)}{2\sqrt{b}} \frac{\frac{1}{2}(d+1)\,\Gamma\left(\frac{1}{2}(d+1)\right)\Gamma\left(\frac{1}{2}\right)}{\frac{1}{2}(d+2)\,\frac{d}{2}\Gamma\left(\frac{d}{2}\right)} {}_1F_1\left(\frac{1}{2},\frac{1}{2}(d+4),\frac{1}{2}\left\|\boldsymbol{y}\right\|^2\right).$$

For $\mathcal{C}(a,b)$, we have $\nu = 1/2,\ \ u=1,\ \ \mu=(d+1)/2,\ \ \beta = a = \left\|\boldsymbol{y}\right\|^2/2$, $\mu+\nu = (d+2)/2,\ \ \beta u = \left\|\boldsymbol{y}\right\|^2/2$. Then $\mathcal{C}(a,b)$, in terms of univariate confluent

hypergeometric function, becomes

$$\mathcal{C}(a,b)=\frac{\exp(-a)}{2\sqrt{b}}\frac{\Gamma\left(\frac{1}{2}(d+1)\right)\Gamma\left(\frac{1}{2}\right)}{\frac{d}{2}\Gamma\left(\frac{d}{2}\right)}{}_1F_1\left(\frac{1}{2},\frac{1}{2}(d+2),\frac{1}{2}\|\boldsymbol{y}\|^2\right).$$

Therefore, the score function $\nabla_{\boldsymbol{y}}\log\{\mathfrak{p}(\boldsymbol{y})\}$, stated in (3.97), in terms of univariate confluent hypergeometric function is

$$\nabla_{\boldsymbol{y}}\log\{\mathfrak{p}(\boldsymbol{y})\}=-\boldsymbol{y}\frac{\mathcal{D}(a,b)}{\mathcal{C}(a,b)}=-\boldsymbol{y}\frac{(d+1)\,{}_1F_1\left(\frac{1}{2},\frac{1}{2}(d+4),\frac{1}{2}\|\boldsymbol{y}\|^2\right)}{(d+2)\,{}_1F_1\left(\frac{1}{2},\frac{1}{2}(d+2),\frac{1}{2}\|\boldsymbol{y}\|^2\right)}. \tag{3.103}$$

In (3.103), since the third argument of the ${}_1F_1$ function being

$$\frac{1}{2}\|\boldsymbol{y}\|^2>0 \text{ and } \left|\frac{1}{2}\|\boldsymbol{y}\|^2\right|\gg 1,$$

we use the first case stated in (3.100) in Appendix 3.C.5 as follows:

$${}_1F_1(a,b,x)=\frac{\Gamma(b)}{\Gamma(a)}e^x x^{a-b}\left\{1+O\left(|x|^{-1}\right)\right\}.$$

Therefore, the ${}_1F_1$ function in the numerator of score function $\nabla_{\boldsymbol{y}}\log\{\mathfrak{p}(\boldsymbol{y})\}$ stated in (3.103) becomes

$$\begin{aligned}&{}_1F_1\left(\frac{1}{2},\frac{1}{2}(d+4),\frac{1}{2}\|\boldsymbol{y}\|^2\right)\\&\quad=\frac{\Gamma\left(\frac{1}{2}(d+4)\right)}{\Gamma\left(\frac{1}{2}\right)}\exp\left(\frac{1}{2}\|\boldsymbol{y}\|^2\right)\left(\frac{1}{2}\|\boldsymbol{y}\|^2\right)^{-\frac{1}{2}(d+3)}\left\{1+O\left(\|\boldsymbol{y}\|^{-2}\right)\right\},\end{aligned}$$

and the ${}_1F_1$ function in the denominator of score function $\nabla_{\boldsymbol{y}} \log\{\mathfrak{p}(\boldsymbol{y})\}$ stated in (3.103) becomes

$$
\begin{aligned}
&{}_1F_1\left(\frac{1}{2}, \frac{1}{2}(d+2), \frac{1}{2}\|\boldsymbol{y}\|^2\right) \\
&\quad = \frac{\Gamma\left(\frac{1}{2}(d+2)\right)}{\Gamma\left(\frac{1}{2}\right)} \exp\left(\frac{1}{2}\|\boldsymbol{y}\|^2\right)\left(\frac{1}{2}\|\boldsymbol{y}\|^2\right)^{-\frac{1}{2}(d+1)}\left\{1+O\left(\|\boldsymbol{y}\|^{-2}\right)\right\}.
\end{aligned}
$$

Then the ${}_1F_1$ functions stated in (3.103) becomes

$$
\begin{aligned}
\frac{{}_1F_1\left(\frac{1}{2}, \frac{1}{2}(d+4), \frac{1}{2}\|\boldsymbol{y}\|^2\right)}{{}_1F_1\left(\frac{1}{2}, \frac{1}{2}(d+2), \frac{1}{2}\|\boldsymbol{y}\|^2\right)} &= 2\frac{\Gamma\left(\frac{1}{2}(d+4)\right)\|\boldsymbol{y}\|^{-2}\left\{1+O\left(\|\boldsymbol{y}\|^{-2}\right)\right\}}{\Gamma\left(\frac{1}{2}(d+2)\right)\left\{1+O\left(\|\boldsymbol{y}\|^{-2}\right)\right\}} \\
&= \frac{(d+2)\|\boldsymbol{y}\|^{-2}\left\{1+O\left(\|\boldsymbol{y}\|^{-2}\right)\right\}}{1+O\left(\|\boldsymbol{y}\|^{-2}\right)}.
\end{aligned}
$$

In summary, for $\tau = 1$, the score function stated in (3.103) becomes

$$
\nabla_{\boldsymbol{y}} \log\{\mathfrak{p}(\boldsymbol{y})\} = -\frac{(d+1)\,\boldsymbol{y}\|\boldsymbol{y}\|^{-2}\left\{1+O\left(\|\boldsymbol{y}\|^{-2}\right)\right\}}{1+O\left(\|\boldsymbol{y}\|^{-2}\right)} \quad \text{for} \quad \|\boldsymbol{y}\| \gg 1. \tag{3.104}
$$

The $\tau > 1$ Case

$\tau > 1$ leads $0 < 1 - (1/b) = 1 - (1/\tau^2) < 1$. Note that the expression for score function $\nabla_{\boldsymbol{y}} \log\{\mathfrak{p}(\boldsymbol{y})\}$, stated in (3.99) in Appendix 3.C.4, is as follows:

$$
\nabla_{\boldsymbol{y}} \log\{\mathfrak{p}(\boldsymbol{y})\} = -\boldsymbol{y}\frac{(d+1)\,\Phi_1\left(\frac{1}{2}, 1, \frac{d+4}{2}, 1-\frac{1}{\tau^2}, \frac{1}{2}\|\boldsymbol{y}\|^2\right)}{(d+2)\,\Phi_1\left(\frac{1}{2}, 1, \frac{d+2}{2}, 1-\frac{1}{\tau^2}, \frac{1}{2}\|\boldsymbol{y}\|^2\right)}.
$$

For the Φ_1 function stated in (3.99), we use (3.85) in Appendix 3.C.1 as follows:

$$
\Phi_1(\alpha, \beta, \gamma, x, y) = \exp(y)\sum_{n=0}^{\infty}\frac{(\alpha)_n(\beta)_n x^n}{(\gamma)_n n!}{}_1F_1(\gamma-\alpha, \gamma+n, -y).
$$

Then the Φ_1 function in the numerator of $\nabla_{\boldsymbol{y}} \log\{\mathfrak{p}(\boldsymbol{y})\}$ stated in (3.99) becomes

$$\begin{aligned}
&\Phi_1\left(\frac{1}{2}, 1, \frac{d+4}{2}, 1-\frac{1}{\tau^2}, \frac{1}{2}\|\boldsymbol{y}\|^2\right) \\
&\quad = \exp\left(\frac{1}{2}\|\boldsymbol{y}\|^2\right) \sum_{n=0}^{\infty} \frac{\left(\frac{1}{2}\right)_n \left(1-\frac{1}{\tau^2}\right)^n}{\left(\frac{1}{2}(d+4)\right)_n} {}_1F_1\left(\frac{1}{2}(d+3), \frac{1}{2}(d+4)+n, -\frac{1}{2}\|\boldsymbol{y}\|^2\right).
\end{aligned} \tag{3.105}$$

Since the third argument in the ${}_1F_1$ function stated in (3.105) being

$$-\frac{1}{2}\|\boldsymbol{y}\|^2 < 0 \text{ and } \left|-\frac{1}{2}\|\boldsymbol{y}\|^2\right| \gg 1,$$

we use the first case stated in (3.100) in Appendix 3.C.5 as follows:

$${}_1F_1(a, b, x) = \frac{\Gamma(b)}{\Gamma(b-a)} (-x)^{-a} \left\{1 + O\left(|x|^{-1}\right)\right\}.$$

Then the ${}_1F_1$ function stated in (3.105) becomes

$$\begin{aligned}
&{}_1F_1\left(\frac{1}{2}(d+3), \frac{1}{2}(d+4)+n, -\frac{1}{2}\|\boldsymbol{y}\|^2\right) \\
&\quad = \frac{\Gamma\left(\frac{1}{2}(d+4)+n\right)}{\Gamma\left(n+\frac{1}{2}\right)} \left(\frac{1}{2}\|\boldsymbol{y}\|^2\right)^{-\frac{1}{2}(d+3)} \left\{1 + O\left(\|\boldsymbol{y}\|^{-2}\right)\right\},
\end{aligned}$$

and the Φ_1 function stated in (3.105) becomes

$$\begin{aligned}
&\Phi_1\left(\frac{1}{2}, 1, \frac{d+4}{2}, 1-\frac{1}{\tau^2}, \frac{1}{2}\|\boldsymbol{y}\|^2\right) \\
&\quad = \frac{\tau^2 (2)^{\frac{1}{2}(d+3)} \Gamma\left(\frac{d}{2}+2\right)}{\sqrt{\pi}} \exp\left(\frac{1}{2}\|\boldsymbol{y}\|^2\right) \left(\|\boldsymbol{y}\|^{-(d+3)}\right) \left\{1 + O\left(\|\boldsymbol{y}\|^{-2}\right)\right\}.
\end{aligned} \tag{3.106}$$

Similarly, using (3.85), the Φ_1 function in the denominator of $\nabla_{\boldsymbol{y}} \log\{\mathfrak{p}(\boldsymbol{y})\}$ stated in (3.99) becomes

$$\Phi_1\left(\frac{1}{2}, 1, \frac{d+2}{2}, 1-\frac{1}{\tau^2}, \frac{1}{2}\|\boldsymbol{y}\|^2\right)$$
$$= \exp\left(\frac{1}{2}\|\boldsymbol{y}\|^2\right) \sum_{n=0}^{\infty} \frac{\left(\frac{1}{2}\right)_n \left(1-\frac{1}{\tau^2}\right)^n}{\left(\frac{1}{2}(d+2)\right)_n} {}_1F_1\left(\frac{1}{2}(d+1), \frac{1}{2}(d+2)+n, -\frac{1}{2}\|\boldsymbol{y}\|^2\right). \tag{3.107}$$

Using the first case stated in (3.100), the ${}_1F_1$ function stated in (3.107) becomes

$${}_1F_1\left(\frac{1}{2}(d+1), \frac{1}{2}(d+2)+n, -\frac{1}{2}\|\boldsymbol{y}\|^2\right)$$
$$= \frac{\Gamma\left(\frac{1}{2}(d+2)+n\right)}{\Gamma\left(n+\frac{1}{2}\right)} \left(\frac{1}{2}\|\boldsymbol{y}\|^2\right)^{-\frac{1}{2}(d+1)} \left\{1 + O\left(\|\boldsymbol{y}\|^{-2}\right)\right\},$$

and the Φ_1 function stated in (3.107) becomes

$$\Phi_1\left(\frac{1}{2}, 1, \frac{d+2}{2}, 1-\frac{1}{\tau^2}, \frac{1}{2}\|\boldsymbol{y}\|^2\right)$$
$$= \frac{\tau^2 (2)^{\frac{1}{2}(d+1)} \Gamma\left(\frac{d}{2}+1\right)}{\sqrt{\pi}} \exp\left(\frac{1}{2}\|\boldsymbol{y}\|^2\right) \left(\|\boldsymbol{y}\|^{-(d+1)}\right) \left\{1 + O\left(\|\boldsymbol{y}\|^{-2}\right)\right\}. \tag{3.108}$$

Note that Φ_1 function in the numerator of $\nabla_{\boldsymbol{y}} \log\{\mathfrak{p}(\boldsymbol{y})\}$ stated in (3.99) becomes (3.106) and the Φ_1 function in the denominator of $\nabla_{\boldsymbol{y}} \log\{\mathfrak{p}(\boldsymbol{y})\}$ stated in (3.99) becomes (3.108). In summary, for $\tau > 1$, the expression for score function $\nabla_{\boldsymbol{y}} \log\{\mathfrak{p}(\boldsymbol{y})\}$

stated in (3.99) in Appendix 3.C.4 becomes

$$\begin{aligned}\nabla_{\boldsymbol{y}} \log\{\mathfrak{p}(\boldsymbol{y})\} &= -\boldsymbol{y}\frac{(d+1)\,\Phi_1\left(\frac{1}{2},1,\frac{d+4}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right)}{(d+2)\,\Phi_1\left(\frac{1}{2},1,\frac{d+2}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right)} \\ &= -\frac{(d+1)\,\boldsymbol{y}\|\boldsymbol{y}\|^{-2}\left\{1+O\left(\|\boldsymbol{y}\|^{-2}\right)\right\}}{1+O\left(\|\boldsymbol{y}\|^{-2}\right)} \quad \text{for} \quad \|\boldsymbol{y}\| \gg 1. \end{aligned} \tag{3.109}$$

The $0<\tau<1$ Case

Note that the expression for $\nabla_{\boldsymbol{y}} \log\{\mathfrak{p}(\boldsymbol{y})\}$, stated in (3.99) in Appendix 3.C.4, is

$$\nabla_{\boldsymbol{y}} \log\{\mathfrak{p}(\boldsymbol{y})\} = -\boldsymbol{y}\frac{(d+1)\,\Phi_1\left(\frac{1}{2},1,\frac{d+4}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right)}{(d+2)\,\Phi_1\left(\frac{1}{2},1,\frac{d+2}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right)}.$$

Moreover, an analogous expression for $\nabla_{\boldsymbol{y}} \log\{\mathfrak{p}(\boldsymbol{y})\}$, stated in (3.97) in Appendix 3.C.4, is

$$\nabla_{\boldsymbol{y}} \{\log \mathfrak{p}(\boldsymbol{y})\} = -\boldsymbol{y}\frac{\mathcal{D}\left(\frac{1}{2}\|\boldsymbol{y}\|^2,\tau^2\right)}{\mathcal{C}\left(\frac{1}{2}\|\boldsymbol{y}\|^2,\tau^2\right)}$$

where $\mathcal{D}(a,b)$ stated in (3.98) and $\mathcal{C}(a,b)$ stated in (3.94) in Appendix 3.C.4 as follows:

$$\mathcal{D}(a,b) = \frac{\exp(-a)}{2\sqrt{b}}\int_0^1 \Big(x\Big)^{\left(\frac{1}{2}-1\right)} \frac{\Big(1-x\Big)^{\left(\frac{d+3}{2}-1\right)}}{\left(1-\left(1-\frac{1}{b}\right)x\right)} \exp\Big(-(-ax)\Big)dx$$

and

$$\mathcal{C}(a,b) = \frac{\exp(-a)}{2\sqrt{b}}\int_0^1 \Big(x\Big)^{\left(\frac{1}{2}-1\right)} \frac{\Big(1-x\Big)^{\left(\frac{d+1}{2}-1\right)}}{\left(1-\left(1-\frac{1}{b}\right)x\right)} \exp\Big(-(-ax)\Big)dx.$$

Note that the common term in the denominator of $\mathcal{D}(a,b)$ and $\mathcal{C}(a,b)$ is as follows:

$$1-\left(1-\frac{1}{b}\right)x=\frac{1}{b}\Big(1-(1-b)(1-x)\Big).$$

The change of variable $u=1-x$ in (3.98) and (3.94), leads to $x=1-u,\ \ dx=-du,$

$$\begin{aligned}\mathcal{D}(a,b)&=\frac{\exp(-a)}{2\sqrt{b}}\int_1^0\Big(1-u\Big)^{\left(\frac{1}{2}-1\right)}\frac{\Big(u\Big)^{\left(\frac{d+3}{2}-1\right)}}{\left(\frac{1}{b}\left(1-(1-b)u\right)\right)}\exp\Big(a(1-u)\Big)(-du)\\&=\frac{\sqrt{b}}{2}\int_0^1\Big(u\Big)^{\left(\frac{d+3}{2}-1\right)}\Big(1-u\Big)^{\left(\frac{1}{2}-1\right)}\Big(1-(1-b)u\Big)^{-1}\exp(-au)\,du,\end{aligned}\tag{3.110}$$

and

$$\begin{aligned}\mathcal{C}(a,b)&=\frac{\exp(-a)}{2\sqrt{b}}\int_1^0\Big(1-u\Big)^{\left(\frac{1}{2}-1\right)}\frac{\Big(u\Big)^{\left(\frac{d-1}{2}\right)}}{\left(\frac{1}{b}\left(1-(1-b)u\right)\right)}\exp\Big(a(1-u)\Big)(-du)\\&=\frac{\sqrt{b}}{2}\int_0^1\Big(u\Big)^{\left(\frac{d+1}{2}-1\right)}\Big(1-u\Big)^{\left(\frac{1}{2}-1\right)}\Big(1-(1-b)u\Big)^{-1}\exp(-au)\,du.\end{aligned}\tag{3.111}$$

For the integrals stated in the last step of (3.110) and (3.111), we use result 3.385 of Gradshteyn and Ryzhik (1994), stated in (3.82), as follows:

$$\int_0^\infty x^{\nu-1}(1-x)^{\lambda-1}(1-\beta x)^{-\rho}e^{-\mu x}dx=\frac{\Gamma(\nu)\Gamma(\lambda)}{\Gamma(\nu+\lambda)}\Phi_1(\nu,\rho,\nu+\lambda,\beta,-\mu).$$

In our case, for $\mathcal{D}(a,b)$, we have $\nu=(d+3)/2,\quad \lambda=1/2,\quad \beta=1-b,\quad \rho=1,$ $\nu+\lambda=(d+4)/2$ and $\mu=a$. Then $\mathcal{D}(a,b)$ stated in the last step of (3.110) becomes

$$\mathcal{D}(a,b)=\frac{\sqrt{b\pi}\Gamma\left(\frac{d+3}{2}\right)}{2\Gamma\left(\frac{d+4}{2}\right)}\Phi_1\left(\frac{d+3}{2},1,\frac{d+4}{2},1-b,-a\right).\tag{3.112}$$

For $\mathcal{C}(a,b)$, we have $\nu=(d+1)/2$, $\lambda=1/2$, $\beta=1-b$, $\rho=1$, $\nu+\lambda=(d+2)/2$ and $\mu=a$. Then $\mathcal{C}(a,b)$ stated in the last step of (3.111) becomes

$$\mathcal{C}(a,b)=\frac{\sqrt{b\pi}\Gamma\left(\frac{d+1}{2}\right)}{2\Gamma\left(\frac{d+2}{2}\right)}\Phi_1\left(\frac{d+1}{2},1,\frac{d+2}{2},1-b,-a\right). \tag{3.113}$$

Note that $a=\|\boldsymbol{y}\|^2/2$ and $b=\tau^2$. Moreover, $\mathcal{D}(a,b)$ in the numerator of $\nabla_{\boldsymbol{y}}\log\{\mathfrak{p}(\boldsymbol{y})\}$ becomes (3.112) and $\mathcal{C}(a,b)$ in the denominator of $\nabla_{\boldsymbol{y}}\log\{\mathfrak{p}(\boldsymbol{y})\}$ becomes (3.113). Therefore, for $0<\tau<1$, the score function stated in (3.99) in Appendix 3.C.4 becomes

$$\begin{aligned}\nabla_{\boldsymbol{y}}\log\{\mathfrak{p}(\boldsymbol{y})\}&=-\boldsymbol{y}\frac{\mathcal{D}(a,b)}{\mathcal{C}(a,b)}\\&=-\boldsymbol{y}\frac{(d+1)\,\Phi_1\left(\frac{d+3}{2},1,\frac{d+4}{2},1-\tau^2,-\frac{1}{2}\|\boldsymbol{y}\|^2\right)}{(d+2)\,\Phi_1\left(\frac{d+1}{2},1,\frac{d+2}{2},1-\tau^2,-\frac{1}{2}\|\boldsymbol{y}\|^2\right)}.\end{aligned} \tag{3.114}$$

For the Φ_1 function stated in the last step of (3.114), we use (3.85) in Appendix 3.C.1 as follows

$$\Phi_1(\alpha,\beta,\gamma,x,y)=\exp(y)\sum_{n=0}^{\infty}\frac{(\alpha)_n(\beta)_n x^n}{(\gamma)_n n!}{}_1F_1(\gamma-\alpha,\gamma+n,-y).$$

Then the Φ_1 function in the numerator of $\nabla_{\boldsymbol{y}}\log\{\mathfrak{p}(\boldsymbol{y})\}$ stated in the last step of (3.114) becomes

$$\begin{aligned}&\Phi_1\left(\frac{d+3}{2},1,\frac{d+4}{2},1-\tau^2,-\frac{1}{2}\|\boldsymbol{y}\|^2\right)\\&=\exp\left(-\frac{1}{2}\|\boldsymbol{y}\|^2\right)\sum_{n=0}^{\infty}\frac{\left(\frac{d+3}{2}\right)_n(1-\tau^2)^n}{\left(\frac{d+4}{2}\right)_n}{}_1F_1\left(\frac{1}{2},\left(\frac{d+4}{2}\right)+n,\frac{1}{2}\|\boldsymbol{y}\|^2\right).\end{aligned} \tag{3.115}$$

Since the third argument of the ${}_1F_1$ function stated in (3.115) being

$$\frac{1}{2}\|\boldsymbol{y}\|^2 > 0 \quad \text{and} \quad \left|\frac{1}{2}\|\boldsymbol{y}\|^2\right| \gg 1,$$

we use the first case stated in (3.100) in Appendix 3.C.5 as follows:

$${}_1F_1\left(a, b, x\right) = \frac{\Gamma\left(b\right)}{\Gamma\left(a\right)} e^x x^{a-b} \left\{1 + O\left(\left|x\right|^{-1}\right)\right\}.$$

Then the ${}_1F_1$ function stated in (3.115) becomes

$$\begin{aligned}
&{}_1F_1\left(\frac{1}{2}, \left(\frac{d+4}{2}\right) + n, \frac{1}{2}\|\boldsymbol{y}\|^2\right) \\
&\quad = \frac{\Gamma\left(\frac{1}{2}\left(d+4\right) + n\right)}{\Gamma\left(\frac{1}{2}\right)} \exp\left(\frac{1}{2}\|\boldsymbol{y}\|^2\right) \left(\frac{1}{2}\|\boldsymbol{y}\|^2\right)^{-\frac{1}{2}(d+3)-n} \left\{1 + O\left(\|\boldsymbol{y}\|^{-2}\right)\right\}.
\end{aligned}$$

Therefore, the Φ_1 function stated in (3.115) becomes

$$\begin{aligned}
&\Phi_1\left(\frac{d+3}{2}, 1, \frac{d+4}{2}, 1-\tau^2, -\frac{1}{2}\|\boldsymbol{y}\|^2\right) \\
&\quad = \frac{(2)^{\frac{d+3}{2}} \Gamma\left(\frac{d+4}{2}\right)}{\Gamma\left(\frac{d+3}{2}\right)\sqrt{\pi}} \|\boldsymbol{y}\|^{-(d+3)} \left\{1 + O\left(\|\boldsymbol{y}\|^{-2}\right)\right\} \sum_{n=0}^{\infty} \frac{\Gamma\left(\frac{d+3}{2} + n\right)\left(1-\tau^2\right)^n 2^n}{\|\boldsymbol{y}\|^{2n}} \\
&\quad = \frac{(2)^{\frac{d+3}{2}} \Gamma\left(\frac{d+4}{2}\right)}{\sqrt{\pi}} \|\boldsymbol{y}\|^{-(d+3)} \left\{1 + O\left(\|\boldsymbol{y}\|^{-2}\right)\right\}. \qquad (3.116)
\end{aligned}$$

Similarly, using (3.85), the Φ_1 function in the denominator of $\nabla_{\boldsymbol{y}} \log\left\{\mathfrak{p}(\boldsymbol{y})\right\}$ stated in

the last step of (3.114) becomes

$$\begin{aligned}
&\Phi_1\left(\frac{d+1}{2}, 1, \frac{d+2}{2}, 1-\tau^2, -\frac{1}{2}\|\boldsymbol{y}\|^2\right) \\
&\quad = \exp\left(-\frac{1}{2}\|\boldsymbol{y}\|^2\right) \sum_{n=0}^{\infty} \frac{\left(\frac{d+1}{2}\right)_n (1-\tau^2)^n}{\left(\frac{d+2}{2}\right)_n} {}_1F_1\left(\frac{1}{2}, \left(\frac{d+2}{2}\right)+n, \frac{1}{2}\|\boldsymbol{y}\|^2\right).
\end{aligned} \tag{3.117}$$

Using the first case, stated in (3.100) in Appendix 3.C.5 as follows

$${}_1F_1(a,b,x) = \frac{\Gamma(b)}{\Gamma(a)} e^x x^{a-b}\left\{1 + O\left(|x|^{-1}\right)\right\},$$

the ${}_1F_1$ function, stated in (3.117), becomes

$$\begin{aligned}
&{}_1F_1\left(\frac{1}{2}, \left(\frac{d+2}{2}\right)+n, \frac{1}{2}\|\boldsymbol{y}\|^2\right) \\
&\quad = \frac{\Gamma\left(\frac{1}{2}(d+2)+n\right)}{\Gamma\left(\frac{1}{2}\right)} \exp\left(\frac{1}{2}\|\boldsymbol{y}\|^2\right)\left(\frac{1}{2}\|\boldsymbol{y}\|^2\right)^{-\frac{1}{2}(d+1)-n}\left\{1 + O\left(\|\boldsymbol{y}\|^{-2}\right)\right\}.
\end{aligned}$$

Therefore, the Φ_1 function stated in (3.117), becomes

$$\begin{aligned}
&\Phi_1\left(\frac{d+1}{2}, 1, \frac{d+2}{2}, 1-\tau^2, -\frac{1}{2}\|\boldsymbol{y}\|^2\right) \\
&\quad = \frac{(2)^{\frac{d+1}{2}}\Gamma\left(\frac{d+2}{2}\right)}{\Gamma\left(\frac{d+1}{2}\right)\sqrt{\pi}} \|\boldsymbol{y}\|^{-(d+1)}\left\{1 + O\left(\|\boldsymbol{y}\|^{-2}\right)\right\} \sum_{n=0}^{\infty} \frac{\Gamma\left(\frac{d+1}{2}+n\right)(1-\tau^2)^n 2^n}{\|\boldsymbol{y}\|^{2n}} \\
&\quad = \frac{(2)^{\frac{d+1}{2}}\Gamma\left(\frac{d+2}{2}\right)}{\sqrt{\pi}} \|\boldsymbol{y}\|^{-(d+1)}\left\{1 + O\left(\|\boldsymbol{y}\|^{-2}\right)\right\}.
\end{aligned} \tag{3.118}$$

Note that the Φ_1 function in the numerator of $\nabla_{\boldsymbol{y}} \log\{\mathfrak{p}(\boldsymbol{y})\}$ stated in (3.114) becomes (3.116) and the Φ_1 function in the denominator of $\nabla_{\boldsymbol{y}} \log\{\mathfrak{p}(\boldsymbol{y})\}$ stated in (3.114) becomes (3.118). In summary, for $0 < \tau < 1$, the score function stated in (3.99), in

Appendix 3.C.4, becomes

$$\nabla_{\boldsymbol{y}} \log\{\mathfrak{p}(\boldsymbol{y})\} = -\boldsymbol{y}\frac{(d+1)\,\Phi_1\left(\frac{d+3}{2}, 1, \frac{d+4}{2}, 1-\tau^2, -\frac{1}{2}\|\boldsymbol{y}\|^2\right)}{(d+2)\,\Phi_1\left(\frac{d+1}{2}, 1, \frac{d+2}{2}, 1-\tau^2, -\frac{1}{2}\|\boldsymbol{y}\|^2\right)}$$

$$= -\frac{(d+1)\,\boldsymbol{y}\|\boldsymbol{y}\|^{-2}\left\{1+O\left(\|\boldsymbol{y}\|^{-2}\right)\right\}}{1+O\left(\|\boldsymbol{y}\|^{-2}\right)} \quad \text{for} \quad \|\boldsymbol{y}\| \gg 1. \tag{3.119}$$

3.C.6 Tail Limit of the Score Function

The expressions for the score function in three cases are described in Appendix 3.C.5. The expression stated in (3.104) is for the $\tau = 1$ case. The expression stated in (3.109) is for the $\tau > 1$ case. The expression stated in (3.119) is for the $0 < \tau < 1$ case. The three expressions for score function, for all cases of τ, is given as:

$$\nabla_{\boldsymbol{y}} \log\{\mathfrak{p}(\boldsymbol{y})\} = -\frac{(d+1)\,\boldsymbol{y}\|\boldsymbol{y}\|^{-2}\left\{1+O\left(\|\boldsymbol{y}\|^{-2}\right)\right\}}{1+O\left(\|\boldsymbol{y}\|^{-2}\right)} \quad \text{for} \quad \|\boldsymbol{y}\| \gg 1.$$

These results imply that, for all $\tau > 0$, the score function has the following leading term tail behaviour:

$$\nabla_{\boldsymbol{y}} \log\{\mathfrak{p}(\boldsymbol{y})\} \sim -\frac{(d+1)\,\boldsymbol{y}}{\|\boldsymbol{y}\|^2} \quad \text{for} \quad \|\boldsymbol{y}\| \gg 1. \tag{3.120}$$

We then have

$$\lim_{\|\boldsymbol{y}\|\to\infty} \nabla_{\boldsymbol{y}} \log\{\mathfrak{p}(\boldsymbol{y})\} = \mathbf{0}.$$

3.C.7 An Explicit Expression for $E\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right)$

In the first part of this subsection, we will show the derivation of the first version expression for $E\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right)$ as follows

$$E\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right) = \frac{\Phi_1\left(\frac{3}{2}, 1, \frac{1}{2}\left(d+4\right), 1-\tau^{-2}, \frac{1}{2}\|\boldsymbol{y}\|^2\right)}{(d+2)\,\Phi_1\left(\frac{1}{2}, 1, \frac{1}{2}\left(d+2\right), 1-\tau^{-2}, \frac{1}{2}\|\boldsymbol{y}\|^2\right)}. \tag{3.121}$$

Consider the model stated in (3.2):

$$\boldsymbol{y}\middle|\boldsymbol{\theta} \sim N(\boldsymbol{\theta}, \boldsymbol{I}_d) \quad \text{with prior} \quad \mathfrak{p}(\boldsymbol{\theta}) = \mathfrak{p}_{\mathrm{HS,d}}\left(\boldsymbol{\theta}/\tau\right)/\tau^d,$$

which is equivalent to (3.87):

$$\boldsymbol{y}\middle|\boldsymbol{\theta} \sim N\left(\boldsymbol{\theta}, \boldsymbol{I}_d\right), \;\; \boldsymbol{\theta}\middle|\lambda \sim N\left(\boldsymbol{0}, \lambda^2\tau^2\boldsymbol{I}_d\right), \;\; \lambda \sim \text{Half-Cauchy}\left(1\right).$$

Then note that

$$\begin{aligned} E\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right) &= \int_{\mathbb{R}^d} \boldsymbol{\theta}\mathfrak{p}(\boldsymbol{\theta}\middle|\boldsymbol{y})d\boldsymbol{\theta} = \int_{\mathbb{R}^d} \boldsymbol{\theta}\frac{\mathfrak{p}(\boldsymbol{\theta},\boldsymbol{y})}{\mathfrak{p}(\boldsymbol{y})}d\boldsymbol{\theta} = \frac{1}{\mathfrak{p}(\boldsymbol{y})}\int_{\mathbb{R}^d} \boldsymbol{\theta}\left\{\int_0^\infty \mathfrak{p}\left(\boldsymbol{y},\boldsymbol{\theta},\lambda\right)d\lambda\right\}d\boldsymbol{\theta} \\ &= \frac{1}{\mathfrak{p}(\boldsymbol{y})}\int_{\mathbb{R}^d} \boldsymbol{\theta}\left\{\int_0^\infty \mathfrak{p}\left(\boldsymbol{y}\middle|\boldsymbol{\theta}\right)\mathfrak{p}\left(\boldsymbol{\theta}\middle|\lambda\right)\mathfrak{p}\left(\lambda\right)d\lambda\right\}d\boldsymbol{\theta} \\ &= \frac{1}{\mathfrak{p}(\boldsymbol{y})}\int_0^\infty \left\{\int_{\mathbb{R}^d} \boldsymbol{\theta}\mathfrak{p}\left(\boldsymbol{y}\middle|\boldsymbol{\theta}\right)\mathfrak{p}\left(\boldsymbol{\theta}\middle|\lambda\right)d\boldsymbol{\theta}\right\}\mathfrak{p}\left(\lambda\right)d\lambda. \end{aligned} \tag{3.122}$$

Let us look into the term $\mathfrak{p}(\boldsymbol{y}\middle|\boldsymbol{\theta})\mathfrak{p}(\boldsymbol{\theta}\middle|\lambda)$ in the inner integral of $E\left(\boldsymbol{\theta}|\boldsymbol{y}\right)$, stated in the last step of (3.122). Note that the expression for $\mathfrak{p}(\boldsymbol{y}\middle|\boldsymbol{\theta})\mathfrak{p}(\boldsymbol{\theta}\middle|\lambda)$ stated in the last step of (3.88) is as follows:

$$\begin{aligned} \mathfrak{p}(\boldsymbol{y}\middle|\boldsymbol{\theta})\mathfrak{p}(\boldsymbol{\theta}\middle|\lambda) &= (2\pi)^{-\frac{d}{2}}\left(\tau\lambda\right)^{-d}\exp\left(\frac{1}{2}\left(\frac{1}{\xi}-1\right)\left\|\boldsymbol{y}\right\|^2\right)\left|\left(\frac{1}{\xi}\right)\boldsymbol{I}_d\right|^{\frac{1}{2}} \\ &\quad\times(2\pi)^{-\frac{d}{2}}\left|\left(\frac{1}{\xi}\right)\boldsymbol{I}_d\right|^{-\frac{1}{2}}\exp\left(\left(\boldsymbol{\theta}-\left(\frac{1}{\xi}\right)\boldsymbol{y}\right)^T\left(\left(\frac{1}{\xi}\right)\boldsymbol{I}_d\right)^{-1}\left(\boldsymbol{\theta}-\left(\frac{1}{\xi}\right)\boldsymbol{y}\right)\right), \end{aligned}$$

where $\xi \equiv 1 + \dfrac{1}{\lambda^2\tau^2}$. Moreover,

$$(2\pi)^{-\frac{d}{2}}\left|\left(\frac{1}{\xi}\right)\boldsymbol{I}_d\right|^{-\frac{1}{2}}\exp\left(\left(\boldsymbol{\theta}-\left(\frac{1}{\xi}\right)\boldsymbol{y}\right)^T\left(\left(\frac{1}{\xi}\right)\boldsymbol{I}_d\right)^{-1}\left(\boldsymbol{\theta}-\left(\frac{1}{\xi}\right)\boldsymbol{y}\right)\right),$$

is the $N\left(\left(1/\xi\right)\boldsymbol{y},\left(1/\xi\right)\boldsymbol{I}_d\right)$ density function in $\boldsymbol{\theta}$. The inner integral of $E\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right)$ stated in the last step of (3.122) becomes

$$\begin{aligned}
&\int_{\mathbb{R}^d}\boldsymbol{\theta}\mathfrak{p}\left(\boldsymbol{y}\middle|\boldsymbol{\theta}\right)\mathfrak{p}\left(\boldsymbol{\theta}\middle|\lambda\right)d\boldsymbol{\theta}\\
&\quad=(2\pi)^{-\frac{d}{2}}(\tau\lambda)^{-d}\exp\left(\frac{1}{2}\left(\frac{1}{\xi}-1\right)\left\|\boldsymbol{y}\right\|^2\right)\left|\left(\frac{1}{\xi}\right)\boldsymbol{I}_d\right|^{\frac{1}{2}}\left(\frac{1}{\xi}\right)\boldsymbol{y}\\
&\quad=(2\pi)^{-\frac{d}{2}}\exp\left(\frac{-\left\|\boldsymbol{y}\right\|^2/2}{\lambda^2\tau^2+1}\right)\frac{\lambda^2\tau^2\boldsymbol{y}}{\left(\lambda^2\tau^2+1\right)^{\frac{d+2}{2}}}.
\end{aligned}\tag{3.123}$$

Recalling that $\lambda\sim$ Half-Cauchy (1), then the numerator of $E\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right)$ stated in the last step of (3.122) becomes

$$\begin{aligned}
&\int_0^\infty\left\{\int_{\mathbb{R}^d}\boldsymbol{\theta}\mathfrak{p}\left(\boldsymbol{y}\middle|\boldsymbol{\theta}\right)\mathfrak{p}\left(\boldsymbol{\theta}\middle|\lambda\right)d\boldsymbol{\theta}\right\}\mathfrak{p}\left(\lambda\right)d\lambda\\
&\quad=\left((2)^{d-2}(\pi)^{d+2}\right)^{-\frac{1}{2}}\boldsymbol{y}\tau^2\int_0^\infty\exp\left(\frac{-\left\|\boldsymbol{y}\right\|^2/2}{1+\lambda^2\tau^2}\right)\frac{\lambda^2}{\left(1+\lambda^2\tau^2\right)^{\frac{d+2}{2}}\left(\lambda^2+1\right)}d\lambda\\
&\quad=\left((2)^{d-2}(\pi)^{d+2}\right)^{-\frac{1}{2}}\boldsymbol{y}\tau^2\int_0^\infty\exp\left(\frac{-a}{1+\lambda^2 b}\right)\frac{\lambda^2}{\left(1+\lambda^2 b\right)^{\frac{d+2}{2}}\left(\lambda^2+1\right)}d\lambda\\
&\quad=\left((2)^{d-2}(\pi)^{d+2}\right)^{-\frac{1}{2}}\boldsymbol{y}\tau^2\mathcal{C}^*\left(a,b\right),
\end{aligned}\tag{3.124}$$

if we define

$$\mathcal{C}^*\left(a,b\right)\equiv\int_0^\infty\exp\left(\frac{-\left\|\boldsymbol{y}\right\|^2/2}{1+\lambda^2\tau^2}\right)\frac{\lambda^2}{\left(1+\lambda^2\tau^2\right)^{\frac{d+2}{2}}\left(\lambda^2+1\right)}d\lambda,\tag{3.125}$$

where $a\equiv\dfrac{1}{2}\left\|\boldsymbol{y}\right\|^2$ and $b\equiv\tau^2$. The change of variable

$$x=\frac{\lambda^2 b}{1+\lambda^b}$$

in (3.125) leads to $\lambda^2 = \dfrac{x}{b\left(1-x\right)}$, $\dfrac{-a}{1+\lambda^2 b} = -a + ax$, $1+\lambda^2 b = \dfrac{1}{1-x}$, $\lambda^2 + 1 = \dfrac{b - x\left(b-1\right)}{b\left(1-x\right)}$, $d\lambda = \dfrac{1}{\sqrt{b}} \dfrac{1}{2\sqrt{-\frac{x}{x-1}}\left(1-x\right)^2} dx$ and

$$\begin{aligned}
\mathcal{C}^*\left(a,b\right) &= \int_0^\infty \exp\left(\frac{-a}{1+\lambda^2 b}\right) \frac{\lambda^2}{\left(1+\lambda^2 b\right)^{\frac{d+2}{2}}\left(\lambda^2+1\right)} d\lambda \\
&= \int_0^1 \exp\Big(-a+ax\Big) \frac{\frac{x}{b(1-x)}}{\left(\frac{1}{1-x}\right)^{\frac{d+2}{2}}\left(\frac{b-x(b-1)}{b(1-x)}\right)} \frac{1}{\sqrt{b}} \frac{1}{2\sqrt{-\frac{x}{x-1}}\left(1-x\right)^2} dx \\
&= \frac{\exp\left(-a\right)}{2b\sqrt{b}} \int_0^1 x^{\left(\frac{3}{2}-1\right)} \frac{\left(1-x\right)^{\left(\frac{d+1}{2}-1\right)}}{\left(1-\left(1-\frac{1}{b}\right)x\right)} \exp\Big(-\left(-ax\right)\Big) dx. \qquad (3.126)
\end{aligned}$$

For the integral stated in the last step of (3.126), we use 3.385 from Gradshteyn and Ryzhik (1994) stated in (3.82) in Appendix 3.C.1 as follows:

$$\int_0^1 x^{\nu-1}\left(1-x\right)^{\lambda-1}\left(1-\beta x\right)^{-\rho} e^{-\mu x} dx = B\left(\nu,\lambda\right)\Phi_1\left(\nu,\rho,\nu+\lambda,\beta,-\mu\right).$$

In our case, we have $\nu = 3/2$, $\lambda = \left(d+1\right)/2$, $\beta = 1 - 1/b$, $\rho = 1$, $\nu+\lambda = \left(d+4\right)/2$ and $\mu = -a$. Then $\mathcal{C}^*\left(a,b\right)$ in the numerator of $E\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right)$ stated in the last step of (3.124) becomes

$$\begin{aligned}
\mathcal{C}^*\left(a,b\right) &= \frac{\exp\left(-a\right)}{2b\sqrt{b}} B\left(\frac{3}{2},\frac{d+1}{2}\right)\Phi_1\left(\frac{3}{2},1,\frac{d+4}{2},1-\frac{1}{b},a\right) \\
&= \frac{\exp\left(-\frac{1}{2}\left\|\boldsymbol{y}\right\|^2\right)}{2\tau^2\tau} \frac{\Gamma\left(\frac{3}{2}\right)\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}+2\right)} \Phi_1\left(\frac{3}{2},1,\frac{d+4}{2},1-\frac{1}{\tau^2},\frac{1}{2}\left\|\boldsymbol{y}\right\|^2\right).
\end{aligned}$$

(3.127)

Note that the expression for $\mathfrak{p}(\boldsymbol{y})$ stated in the last step of (3.96), which is the denominator of the $E\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right)$ stated in the last step of (3.122), is as follows:

$$\begin{aligned}&\mathfrak{p}(\boldsymbol{y})\\&=\left((2)^{d-2}(\pi)^{d+1}\right)^{-\frac{1}{2}}\frac{\exp\left(-\frac{1}{2}\|\boldsymbol{y}\|^2\right)\Gamma\left(\frac{d}{2}+\frac{1}{2}\right)}{d\tau\Gamma\left(\frac{d}{2}\right)}\Phi_1\left(\frac{1}{2},1,\frac{d+2}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right).\end{aligned}$$

The numerator of $E\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right)$ stated in the last step of (3.124) is as follows:

$$\int_0^\infty\left\{\int_{\mathbb{R}^d}\boldsymbol{\theta}\mathfrak{p}\left(\boldsymbol{y}\middle|\boldsymbol{\theta}\right)\mathfrak{p}\left(\boldsymbol{\theta}\middle|\lambda\right)d\boldsymbol{\theta}\right\}\mathfrak{p}\left(\lambda\right)d\lambda=\left((2)^{d-2}(\pi)^{d+2}\right)^{-\frac{1}{2}}\boldsymbol{y}\tau^2\mathcal{C}^*\left(a,b\right),$$

with $\mathcal{C}^*\left(a,b\right)$ stated in (3.127). We then have

$$\begin{aligned}E\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right)&=\frac{1}{\mathfrak{p}(\boldsymbol{y})}\int_0^\infty\left\{\int_{\mathbb{R}^d}\boldsymbol{\theta}\mathfrak{p}\left(\boldsymbol{y}\middle|\boldsymbol{\theta}\right)\mathfrak{p}\left(\boldsymbol{\theta}\middle|\lambda\right)d\boldsymbol{\theta}\right\}\mathfrak{p}\left(\lambda\right)d\lambda=\frac{\left((2)^{d-2}(\pi)^{d+2}\right)^{-\frac{1}{2}}\boldsymbol{y}\tau^2}{\left((2)^{d-2}(\pi)^{d+1}\right)^{-\frac{1}{2}}}\\&\times\frac{\frac{\exp\left(-\frac{1}{2}\|\boldsymbol{y}\|^2\right)}{2\tau^2\tau}\frac{\Gamma\left(\frac{3}{2}\right)\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}+2\right)}\Phi_1\left(\frac{3}{2},1,\frac{d+4}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right)}{\frac{\exp\left(-\frac{1}{2}\|\boldsymbol{y}\|^2\right)\Gamma\left(\frac{d}{2}+\frac{1}{2}\right)}{d\tau\Gamma\left(\frac{d}{2}\right)}\Phi_1\left(\frac{1}{2},1,\frac{d+2}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right)}\\&=\frac{\Phi_1\left(\frac{3}{2},1,\frac{d+4}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right)}{(d+2)\,\Phi_1\left(\frac{1}{2},1,\frac{d+2}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right)}\boldsymbol{y}.\end{aligned}$$

This matches the first version expression for $E\left(\boldsymbol{\theta}|\boldsymbol{y}\right)$ to be derived, stated in (3.121). For $d=1$, $E\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right)$ stated in (3.121) becomes

$$E\left(\theta|y\right)=\frac{\Phi_1\left(\frac{3}{2},1,\frac{5}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right)}{3\Phi_1\left(\frac{1}{2},1,\frac{3}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right)}y.\tag{3.128}$$

Equation (9) in Bhadra et al. (2019) is given as follows:

$$E\left(\theta|y\right)=\left\{1-\frac{2\Phi_1\left(\frac{3}{2},1,\frac{5}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right)}{3\Phi_1\left(\frac{1}{2},1,\frac{3}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right)}\right\}y$$
$$=\left\{\frac{3\Phi_1\left(\frac{1}{2},1,\frac{3}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right)-2\Phi_1\left(\frac{3}{2},1,\frac{5}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right)}{3\Phi_1\left(\frac{1}{2},1,\frac{3}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right)}\right\}y. \tag{3.129}$$

For the $d=1$ case, it implies that

$$3\Phi_1\left(\frac{1}{2},1,\frac{3}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right)-2\Phi_1\left(\frac{3}{2},1,\frac{5}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right)$$
$$=\Phi_1\left(\frac{3}{2},1,\frac{5}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right). \tag{3.130}$$

The left-hand side of (3.130) is the numerator of $E\left(\theta\middle|y\right)$ of Bhadra et al. (2019), as stated in the last step of (3.129). The right-hand side of (3.130) is the numerator of $E\left(\theta\middle|y\right)$ of our first version expression, as stated in (3.128).

In general case, it implies

$$(d+2)\,\Phi_1\left(\frac{1}{2},1,\frac{d+2}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right)-(d+1)\,\Phi_1\left(\frac{3}{2},1,\frac{d+4}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right)$$
$$=\Phi_1\left(\frac{3}{2},1,\frac{d+4}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right). \tag{3.131}$$

The left-hand side of (3.131) is the numerator of our second version expression for $E\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right)$, which will be derived in the next part. The right-hand side of (3.131) is the numerator of our first version expression for $E\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right)$, as stated in (3.128). (3.131) implies the second version expression for $E\left(\boldsymbol{\theta}|\boldsymbol{y}\right)$ to be derived as follows:

$$E\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right)=\left\{1-\frac{(d+1)\,\Phi_1\left(\frac{1}{2},1,\frac{d+4}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right)}{(d+2)\,\Phi_1\left(\frac{1}{2},1,\frac{d+2}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right)}\right\}\boldsymbol{y}. \tag{3.132}$$

In the second part of this subsection, we will show the derivation of the second version expression for $E\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right)$ stated in (3.132). Next note that

$$\begin{aligned}
E\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right) &= \int_{\mathbb{R}^d} \boldsymbol{\theta}\mathfrak{p}(\boldsymbol{\theta}\middle|\boldsymbol{y})d\boldsymbol{\theta} = \int_{\mathbb{R}^d} \boldsymbol{\theta}\frac{\mathfrak{p}(\boldsymbol{\theta},\boldsymbol{y})}{\mathfrak{p}(\boldsymbol{y})}d\boldsymbol{\theta} = \frac{1}{\mathfrak{p}(\boldsymbol{y})}\int_{\mathbb{R}^d} \boldsymbol{\theta}\mathfrak{p}\left(\boldsymbol{y}\middle|\boldsymbol{\theta}\right)\mathfrak{p}\left(\boldsymbol{\theta}\right)d\boldsymbol{\theta} \\
&= \frac{1}{\mathfrak{p}(\boldsymbol{y})}\left(\int_{\mathbb{R}^d} (\boldsymbol{y}-(\boldsymbol{y}-\boldsymbol{\theta}))\,\mathfrak{p}\left(\boldsymbol{y}\middle|\boldsymbol{\theta}\right)\mathfrak{p}\left(\boldsymbol{\theta}\right)\right)d\boldsymbol{\theta} \\
&= \frac{1}{\mathfrak{p}(\boldsymbol{y})}\left(\boldsymbol{y}\int_{\mathbb{R}^d} \mathfrak{p}\left(\boldsymbol{\theta},\boldsymbol{y}\right)d\boldsymbol{\theta}\right) - \frac{1}{\mathfrak{p}(\boldsymbol{y})}\int_{\mathbb{R}^d} (\boldsymbol{y}-\boldsymbol{\theta})\left\{\int_0^\infty \mathfrak{p}\left(\boldsymbol{y}\middle|\boldsymbol{\theta}\right)\mathfrak{p}\left(\boldsymbol{\theta}\middle|\lambda\right)\mathfrak{p}\left(\lambda\right)d\lambda\right\}d\boldsymbol{\theta} \\
&= \boldsymbol{y} - \frac{1}{\mathfrak{p}(\boldsymbol{y})}\int_0^\infty \left\{\int_{\mathbb{R}^d} (\boldsymbol{y}-\boldsymbol{\theta})\,\mathfrak{p}\left(\boldsymbol{y}\middle|\boldsymbol{\theta}\right)\mathfrak{p}\left(\boldsymbol{\theta}\middle|\lambda\right)d\boldsymbol{\theta}\right\}\mathfrak{p}\left(\lambda\right)d\lambda. \qquad (3.133)
\end{aligned}$$

Our model is as follows:

$$\boldsymbol{y}\middle|\boldsymbol{\theta} \sim N\left(\boldsymbol{\theta},\boldsymbol{I}_d\right),\ \ \boldsymbol{\theta}\middle|\lambda \sim N\left(\boldsymbol{0},\lambda^2\tau^2\boldsymbol{I}_d\right),\ \ \lambda \sim \text{Half-Cauchy}\left(1\right).$$

Introduce the following notation defined immediately after equation (4) of Wand and Jones (1993):

$$\phi_{\boldsymbol{\Sigma}}\left(\boldsymbol{x}\right) = \left(2\pi\right)^{-d/2}\left|\boldsymbol{\Sigma}\right|^{-1/2}\exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right).$$

We then have

$$\mathfrak{p}\left(\boldsymbol{y}\middle|\boldsymbol{\theta}\right) = \phi_{\boldsymbol{I}_d}\left(\boldsymbol{y}-\boldsymbol{\theta}\right) \quad \text{and} \quad \mathfrak{p}\left(\boldsymbol{\theta}\middle|\lambda\right) = \phi_{\lambda^2\tau^2\boldsymbol{I}_d}\left(\boldsymbol{\theta}\right).$$

Using multivariate integration by parts, the differentiation results as:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}}\phi_{\boldsymbol{I}_d}\left(\boldsymbol{y}-\boldsymbol{\theta}\right) &= \left(\boldsymbol{y}-\boldsymbol{\theta}\right)\phi_{\boldsymbol{I}_d}\left(\boldsymbol{y}-\boldsymbol{\theta}\right) \quad \text{and} \\
\nabla_{\boldsymbol{\theta}}\phi_{\lambda^2\tau^2\boldsymbol{I}_d}\left(\boldsymbol{\theta}\right) &= -\left(\lambda\tau\right)^{-2}\boldsymbol{\theta}\phi_{\lambda^2\tau^2\boldsymbol{I}_d}\left(\boldsymbol{\theta}\right),
\end{aligned}$$

the inner integral in $E\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right)$ stated in the last step of (3.133) becomes

$$\begin{aligned}
&\int_{\mathbb{R}^d}(\boldsymbol{y}-\boldsymbol{\theta})\,\mathfrak{p}\left(\boldsymbol{y}\middle|\boldsymbol{\theta}\right)\mathfrak{p}\left(\boldsymbol{\theta}\middle|\lambda\right)d\boldsymbol{\theta}=\int_{\mathbb{R}^d}(\boldsymbol{y}-\boldsymbol{\theta})\,\phi_{\boldsymbol{I}_d}(\boldsymbol{y}-\boldsymbol{\theta})\,\phi_{\lambda^2\tau^2\boldsymbol{I}_d}(\boldsymbol{\theta})\,d\boldsymbol{\theta}\\
&=\int_{\mathbb{R}^d}\left\{\nabla_{\boldsymbol{\theta}}\phi_{\boldsymbol{I}_d}(\boldsymbol{y}-\boldsymbol{\theta})\right\}\phi_{\lambda^2\tau^2\boldsymbol{I}_d}(\boldsymbol{\theta})\,d\boldsymbol{\theta}=-\int_{\mathbb{R}^d}\phi_{\boldsymbol{I}_d}(\boldsymbol{y}-\boldsymbol{\theta})\left\{\nabla_{\boldsymbol{\theta}}\phi_{\lambda^2\tau^2\boldsymbol{I}_d}(\boldsymbol{\theta})\right\}d\boldsymbol{\theta}\\
&=(\lambda\tau)^{-2}\int_{\mathbb{R}^d}\phi_{\boldsymbol{I}_d}(\boldsymbol{y}-\boldsymbol{\theta})\left\{\boldsymbol{\theta}\phi_{\lambda^2\tau^2\boldsymbol{I}_d}(\boldsymbol{\theta})\right\}d\boldsymbol{\theta}=(\lambda\tau)^{-2}\int_{\mathbb{R}^d}\mathfrak{p}\left(\boldsymbol{y}\middle|\boldsymbol{\theta}\right)\boldsymbol{\theta}\mathfrak{p}\left(\boldsymbol{\theta}\middle|\lambda\right)d\boldsymbol{\theta}.
\end{aligned}$$

Then $E\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right)$ stated in the last step of (3.133) becomes

$$E(\boldsymbol{\theta}|\boldsymbol{y})=\boldsymbol{y}-\frac{1}{\mathfrak{p}(\boldsymbol{y})}\int_0^\infty(\lambda\tau)^{-2}\left\{\int_{\mathbb{R}^d}\boldsymbol{\theta}\mathfrak{p}\left(\boldsymbol{y}\middle|\boldsymbol{\theta}\right)\mathfrak{p}\left(\boldsymbol{\theta}\middle|\lambda\right)d\boldsymbol{\theta}\right\}\mathfrak{p}(\lambda)\,d\lambda. \quad (3.134)$$

The inner integral stated in (3.134), with $\xi\equiv 1+1/\left(\lambda^2\tau^2\right)$, is stated in (3.123) as follows:

$$\begin{aligned}
&\int_{\mathbb{R}^d}\boldsymbol{\theta}\mathfrak{p}\left(\boldsymbol{y}\middle|\boldsymbol{\theta}\right)\mathfrak{p}\left(\boldsymbol{\theta}\middle|\lambda\right)d\boldsymbol{\theta}\\
&=(2\pi)^{-\frac{d}{2}}(\tau\lambda)^{-d}\exp\left(\frac{1}{2}\left(\frac{1}{\xi}-1\right)\left\|\boldsymbol{y}\right\|^2\right)\left|\left(\frac{1}{\xi}\right)\boldsymbol{I}_d\right|^{\frac{1}{2}}\left(\frac{1}{\xi}\right)\boldsymbol{y}\\
&=(2\pi)^{-\frac{d}{2}}\exp\left(\frac{-\left\|\boldsymbol{y}\right\|^2/2}{\lambda^2\tau^2+1}\right)\frac{\lambda^2\tau^2\boldsymbol{y}}{\left(\lambda^2\tau^2+1\right)^{\frac{d+2}{2}}}.
\end{aligned}$$

Therefore, the double integral stated in (3.134) becomes

$$\begin{aligned}
&\int_0^\infty(\lambda\tau)^{-2}\left\{\int_{\mathbb{R}^d}\boldsymbol{\theta}\mathfrak{p}\left(\boldsymbol{y}\middle|\boldsymbol{\theta}\right)\mathfrak{p}\left(\boldsymbol{\theta}\middle|\lambda\right)\boldsymbol{\theta}\right\}\mathfrak{p}(\lambda)\,d\lambda\\
&=\int_0^\infty(\lambda\tau)^{-2}\left\{(2\pi)^{-\frac{d}{2}}\exp\left(\frac{-\left\|\boldsymbol{y}\right\|^2/2}{\lambda^2\tau^2+1}\right)\frac{\lambda^2\tau^2\boldsymbol{y}}{\left(\lambda^2\tau^2+1\right)^{\frac{d+2}{2}}}\right\}\frac{1}{1+\lambda^2}d\lambda\\
&=\left((2)^{d-2}(\pi)^{d+2}\right)^{-\frac{1}{2}}\boldsymbol{y}\int_0^\infty\exp\left(\frac{-\left\|\boldsymbol{y}\right\|^2/2}{\lambda^2\tau^2+1}\right)\frac{1}{\left(\lambda^2\tau^2+1\right)^{\frac{d+2}{2}}}\frac{1}{1+\lambda^2}d\lambda\\
&=\left((2)^{d-2}(\pi)^{d+2}\right)^{-\frac{1}{2}}\boldsymbol{y}\mathcal{C}^{**}(a,b), \quad (3.135)
\end{aligned}$$

if we define

$$\mathcal{C}^{**}(a,b) \equiv \int_0^{\infty} \exp\left(\frac{-\left\|\boldsymbol{y}\right\|^2/2}{1+\lambda^2\tau^2}\right) \frac{1}{\left(1+\lambda^2\tau^2\right)^{\frac{d+2}{2}}\left(\lambda^2+1\right)} d\lambda, \tag{3.136}$$

$a \equiv \frac{1}{2}\left\|\boldsymbol{y}\right\|^2$ and $b \equiv \tau^2$. The change of variable

$$x = \frac{\lambda^2 b}{1+\lambda^b}$$

in (3.136) leads to $\lambda^2 = \frac{x}{b(1-x)}$, $\frac{-a}{1+\lambda^2 b} = -a+ax$, $1+\lambda^2 b = \frac{1}{1-x}$, $\lambda^2+1 = \frac{b-x(b-1)}{b(1-x)}$, $d\lambda = \frac{1}{\sqrt{b}} \frac{1}{2\sqrt{-\frac{x}{x-1}}(1-x)^2} dx$ and

$$\begin{aligned}
\mathcal{C}^{**}(a,b) &= \int_0^{\infty} \exp\left(\frac{-a}{1+\lambda^2 b}\right) \frac{1}{\left(1+\lambda^2 b\right)^{\frac{d+2}{2}}\left(\lambda^2+1\right)} d\lambda \\
&= \int_0^1 \exp\Big(-a+ax\Big) \frac{1}{\left(\frac{1}{1-x}\right)^{\frac{d+2}{2}}\left(\frac{b-x(b-1)}{b(1-x)}\right)} \frac{1}{\sqrt{b}} \frac{1}{2\sqrt{-\frac{x}{x-1}}(1-x)^2} dx \\
&= \frac{\exp(-a)}{2b\sqrt{b}} \int_0^1 \Big(x\Big)^{\left(\frac{1}{2}-1\right)} \frac{(1-x)^{\left(\frac{d+3}{2}-1\right)}}{\left(1-\left(1-\frac{1}{b}\right)x\right)} \exp\Big(-(-ax)\Big) dx.
\end{aligned} \tag{3.137}$$

For the integral stated in the last step of (3.137), we use 3.385 from Gradshteyn and Ryzhik (1994) stated in (3.82) in Appendix 3.C.1 as follows:

$$\int_0^1 x^{\nu-1}(1-x)^{\lambda-1}(1-\beta x)^{-\rho} e^{-\mu x} dx = B(\nu,\lambda)\,\Phi_1(\nu,\rho,\nu+\lambda,\beta,-\mu).$$

In our case, we have $\nu = 1/2$, $\lambda = (d+3)/2$, $\beta = 1-1/b$, $\rho = 1$, $\nu+\lambda = (d+4)/2$

and $\mu = -a$. Then $\mathcal{C}^{**}(a,b)$ stated in the last step of (3.135) becomes

$$\begin{aligned}\mathcal{C}^{**}(a,b) &= \frac{\exp(-a)}{2b\sqrt{b}} B\left(\frac{1}{2}, \frac{d+3}{2}\right) \Phi_1\left(\frac{1}{2}, 1, \frac{d+4}{2}, 1-\frac{1}{b}, a\right) \\ &= \frac{\exp\left(-\frac{1}{2}\|\boldsymbol{y}\|^2\right)}{2\tau} \frac{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{d+3}{2}\right)}{\Gamma\left(\frac{d}{2}+2\right)} \Phi_1\left(\frac{1}{2}, 1, \frac{d+4}{2}, 1-\frac{1}{\tau^2}, \frac{1}{2}\|\boldsymbol{y}\|^2\right).\end{aligned} \tag{3.138}$$

Then $E\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right)$ stated in (3.134) becomes

$$\begin{aligned}E\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right) &= \boldsymbol{y} - \frac{1}{\mathfrak{p}(\boldsymbol{y})} \int_0^\infty (\lambda\tau)^{-2} \left\{ \int_{\mathbb{R}} \boldsymbol{\theta} \mathfrak{p}\left(\boldsymbol{y}\middle|\boldsymbol{\theta}\right) \mathfrak{p}\left(\boldsymbol{\theta}\middle|\lambda\right) \boldsymbol{\theta} \right\} \mathfrak{p}(\lambda)\, d\lambda \\ &= \boldsymbol{y} - \frac{1}{\mathfrak{p}(\boldsymbol{y})} \left((2)^{d-2}(\pi)^{d+2}\right)^{-\frac{1}{2}} \boldsymbol{y}\mathcal{C}^{**}(a,b),\end{aligned}$$

where $\mathcal{C}^{**}(a,b)$ stated in (3.138) and $\mathfrak{p}(\boldsymbol{y})$ stated in (3.96) as follows:

$$\begin{aligned}&\mathfrak{p}(\boldsymbol{y}) \\ &= \left((2)^{d-2}(\pi)^{d+1}\right)^{-\frac{1}{2}} \frac{\exp\left(-\frac{1}{2}\|\boldsymbol{y}\|^2\right)\Gamma\left(\frac{d}{2}+\frac{1}{2}\right)}{d\tau\Gamma\left(\frac{d}{2}\right)} \Phi_1\left(\frac{1}{2}, 1, \frac{d+2}{2}, 1-\frac{1}{\tau^2}, \frac{1}{2}\|\boldsymbol{y}\|^2\right).\end{aligned}$$

Therefore, $E\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right)$ stated in (3.134) becomes

$$\begin{aligned}
E\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right) &= \boldsymbol{y} - \frac{\left((2)^{d-2}(\pi)^{d+2}\right)^{-\frac{1}{2}}}{\left((2)^{d-2}(\pi)^{d+1}\right)^{-\frac{1}{2}}} \\
&\times \frac{\frac{\exp\left(-\frac{1}{2}\|\boldsymbol{y}\|^2\right)}{2\tau}\frac{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{d+3}{2}\right)}{\Gamma\left(\frac{d}{2}+2\right)}\Phi_1\left(\frac{1}{2},1,\frac{d+4}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right)}{\frac{\exp\left(-\frac{1}{2}\|\boldsymbol{y}\|^2\right)\Gamma\left(\frac{d}{2}+\frac{1}{2}\right)}{d\tau\Gamma\left(\frac{d}{2}\right)}\Phi_1\left(\frac{1}{2},1,\frac{d+2}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right)}\boldsymbol{y} \\
&= \left\{1 - \frac{(d+1)\,\Phi_1\left(\frac{1}{2},1,\frac{d+4}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right)}{(d+2)\,\Phi_1\left(\frac{1}{2},1,\frac{d+2}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right)}\right\}\boldsymbol{y},
\end{aligned}$$

matching the second version expression for $E\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right)$ stated in (3.132) to be derived. Comparing the second version expression for $E\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right)$ stated in (3.132) with the first version expression for $E\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right)$ stated in (3.121) as follows

$$E\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right) = \frac{\Phi_1\left(\frac{3}{2},1,\frac{1}{2}(d+4),1-\tau^{-2},\frac{1}{2}\|\boldsymbol{y}\|^2\right)}{(d+2)\,\Phi_1\left(\frac{1}{2},1,\frac{1}{2}(d+2),1-\tau^{-2},\frac{1}{2}\|\boldsymbol{y}\|^2\right)},$$

we note that, for general case, the following expression stated in (3.131) holds

$$\begin{aligned}
&(d+2)\,\Phi_1\left(\frac{1}{2},1,\frac{d+2}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right) - (d+1)\,\Phi_1\left(\frac{3}{2},1,\frac{d+4}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right) \\
&= \Phi_1\left(\frac{3}{2},1,\frac{d+4}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right).
\end{aligned}$$

In summary, the first version expression for $E\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right)$ stated in (3.121) is derived in the first part of this subsection. The second version expression for $E\left(\boldsymbol{\theta}\middle|\boldsymbol{y}\right)$ stated in (3.132) is derived in the second part of this subsection.

3.C.8 Bounding of $\left\|\boldsymbol{y}-E\left(\boldsymbol{\theta}|\boldsymbol{y}\right)\right\|$

Recalling that the second version expression for $E\left(\boldsymbol{\theta}|\boldsymbol{y}\right)$, stated in (3.132), as follows

$$E\left(\boldsymbol{\theta}|\boldsymbol{y}\right)=\left\{1-\frac{\left(d+1\right)\Phi_1\left(\frac{1}{2},1,\frac{d+4}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right)}{\left(d+2\right)\Phi_1\left(\frac{1}{2},1,\frac{d+2}{2},1-\frac{1}{\tau^2},\frac{1}{2}\|\boldsymbol{y}\|^2\right)}\right\}\boldsymbol{y},$$

and recalling that the score function has the following leading term tail behaviour stated in (3.120)

$$\nabla_{\boldsymbol{y}}\log\left\{\mathfrak{p}(\boldsymbol{y})\right\}\sim-\frac{\left(d+1\right)\boldsymbol{y}}{\|\boldsymbol{y}\|^2}\quad\text{for}\quad\|\boldsymbol{y}\|\gg1,$$

we then have

$$\left\|\boldsymbol{y}-E\left(\boldsymbol{\theta}|\boldsymbol{y}\right)\right\|=\left\|\nabla_{\boldsymbol{y}}\log\left\{\mathfrak{p}(\boldsymbol{y})\right\}\right\|.\tag{3.139}$$

From (3.120), we have

$$\nabla_{\boldsymbol{y}}\log\left\{\mathfrak{p}(\boldsymbol{y})\right\}=\boldsymbol{0}\quad\text{at}\quad\boldsymbol{y}=\boldsymbol{0}\quad\text{and}\quad\nabla_{\boldsymbol{y}}\log\left\{\mathfrak{p}(\boldsymbol{y})\right\}\sim-\frac{\left(d+1\right)\boldsymbol{y}}{\|\boldsymbol{y}\|^2}\quad\text{for}\quad\|\boldsymbol{y}\|\gg1.$$

(3.139) then provides

$$\left\|\boldsymbol{y}-E\left(\boldsymbol{\theta}|\boldsymbol{y}\right)\right\|=\boldsymbol{0}\quad\text{at}\quad\boldsymbol{y}=\boldsymbol{0}\quad\text{and}\quad\left\|\boldsymbol{y}-E\left(\boldsymbol{\theta}|\boldsymbol{y}\right)\right\|\sim-\frac{\left(d+1\right)\boldsymbol{y}}{\|\boldsymbol{y}\|^2}\quad\text{for}\quad\|\boldsymbol{y}\|\gg1.$$

This result and the fact that $\left\|\boldsymbol{y}-E\left(\boldsymbol{\theta}|\boldsymbol{y}\right)\right\|$ is a continuous in $\boldsymbol{y}$ implies that $E\left\|\boldsymbol{y}-E\left(\boldsymbol{\theta}|\boldsymbol{y}\right)\right\|$ is bounded by some $b_\tau<\infty$ that depends only on τ.

3.D Derivation of Result 3.2.5.1

Consider the Bayesian model

$$\mathfrak{p}\left(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n\middle|\boldsymbol{\theta}\right)=\prod_{i=1}^{n}\mathfrak{p}\left(\boldsymbol{y}_i\middle|\boldsymbol{\theta}\right),\ \ \boldsymbol{\theta}\text{ has prior density function }\mathfrak{p}_{\mathrm{HS,d}}\left(\boldsymbol{\theta}\right),$$

where, for $1\leq i\leq n$,

$$\mathfrak{p}\left(\boldsymbol{y}_i\middle|\boldsymbol{\theta}\right)\equiv\left(2\pi\sigma^2\right)^{-\frac{d}{2}}\exp\left\{-\frac{\left\|\boldsymbol{y}_i-\boldsymbol{\theta}\right\|^2}{2\sigma^2}\right\}.$$

Suppose $\boldsymbol{\theta}^0$ is the true value of $\boldsymbol{\theta}$. For each $\boldsymbol{\theta}\in\mathbb{R}^d$, the Kullback-Leibler divergence of $\mathfrak{p}\left(\boldsymbol{y}_i\middle|\boldsymbol{\theta}\right)$ from $\mathfrak{p}\left(\boldsymbol{y}_i\middle|\boldsymbol{\theta}^0\right)$ is

$$\mathrm{KL}\left(\mathfrak{p}\left(\boldsymbol{y}_i\middle|\boldsymbol{\theta}^0\right)\middle\|\mathfrak{p}\left(\boldsymbol{y}_i\middle|\boldsymbol{\theta}\right)\right)=\frac{\left\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\right\|^2}{2\sigma^2}.$$

For each $\varepsilon>0$, let

$$A_\epsilon\equiv\left\{\boldsymbol{\theta}:\mathrm{KL}\left(\mathfrak{p}\left(\boldsymbol{y}_i\middle|\boldsymbol{\theta}^0\right)\middle\|\mathfrak{p}\left(\boldsymbol{y}_i\middle|\boldsymbol{\theta}\right)\right)\leq\varepsilon\right\}=\left\{\left\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\right\|^2\leq 2\sigma^2\varepsilon\right\}.$$

Using Proposition 1 of Bhadra et al. (2017), which is attributed to Clarke and Barron (1990), with $\varepsilon=1/n$ leads to the risk quantity as follows

$$R_n\leq\frac{1}{n}-\frac{1}{n}\log\left(\int_{A_{1/n}}\mathfrak{p}_{\mathrm{HS,d}}\left(\boldsymbol{\theta}\right)d\boldsymbol{\theta}\right)=\frac{1}{n}-\frac{1}{n}\log\left(\int_{\left\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\right\|\leq\sigma\sqrt{2/n}}\mathfrak{p}_{\mathrm{HS,d}}\left(\boldsymbol{\theta}\right)d\boldsymbol{\theta}\right).\tag{3.140}$$

3.D.1 The $\boldsymbol{\theta}^0=\mathbf{0}$ Case

For the $\boldsymbol{\theta}^0=\mathbf{0}$ case, the risk quantity stated in (3.140) becomes

$$R_n\leq\frac{1}{n}-\frac{1}{n}\log\left(\int_{\left\|\boldsymbol{\theta}\right\|\leq\sigma\sqrt{2/n}}\mathfrak{p}_{\mathrm{HS,d}}\left(\boldsymbol{\theta}\right)d\boldsymbol{\theta}\right).\tag{3.141}$$

To work out the order of magnitude of the right-hand side of (3.141), we consider the following two cases separately: (a) $d = 1$ and (b) $d \geq 2$.

The $d = 1$ Case

For $d = 1$, we have

$$\mathfrak{p}_{\mathrm{HS},1}(\theta) = \frac{1}{\sqrt{2\pi^3}} \exp\left(\frac{\theta^2}{2}\right) E_1\left(\frac{\theta^2}{2}\right).$$

Then the risk quantity stated in (3.141) becomes

$$\begin{aligned} R_n \leq &= \frac{1}{n} - \frac{1}{n} \log\left(\int_{-\sigma\sqrt{\frac{2}{n}}}^{\sigma\sqrt{\frac{2}{n}}} \frac{1}{\sqrt{2\pi^3}} \exp\left(\frac{\theta^2}{2}\right) E_1\left(\frac{\theta^2}{2}\right) \right) d\theta \\ &= \frac{1}{n} - \frac{1}{n} \log\left(\sqrt{\frac{2}{\pi^3}} \int_0^{\sigma\sqrt{\frac{2}{n}}} \exp\left(\frac{\theta^2}{2}\right) E_1\left(\frac{\theta^2}{2}\right) \right) d\theta. \end{aligned} \tag{3.142}$$

The change of variable

$$u = \frac{\theta^2}{2},$$

in the last step of (3.142), leads to $\theta = \sqrt{2u}$, $d\theta = \frac{\sqrt{2}}{2} u^{-\frac{1}{2}} du$ and

$$R_n \leq \frac{1}{n} - \frac{1}{n} \log\left((\pi)^{-\frac{3}{2}} \int_0^{\frac{\sigma^2}{n}} (u)^{-\frac{1}{2}} \exp(u) E_1(u) \, du \right). \tag{3.143}$$

For the term $E_1(u)$ stated in (3.143), we use Section 8.214 of Gradshteyn and Ryzhik (1994) as follows

$$\begin{aligned} E_1(u) &= -\gamma - \log(u) - \sum_{k=1}^{\infty} \frac{(-u)^k}{kk!} = -\gamma - \log(u) + \sum_{k=1}^{\infty} \frac{(-1)^{k+1} (u)^k}{kk!} \\ &= -\gamma - \log(u) + u - \frac{u^2}{4} + \frac{u^3}{18} - \frac{u^4}{96} + \dots \end{aligned}$$

where $\gamma \equiv -\text{digamma}(1)$. Recalling that

$$\exp(u) = 1 + \frac{u}{1!} + \frac{u^2}{2} + \frac{u^3}{3!} + \dots,$$

then the integral in R_n, stated in (3.143), becomes

$$\begin{aligned}
&\int_0^{\frac{\sigma^2}{n}} (u)^{-\frac{1}{2}} \exp(u) E_1(u)\, du \\
&= \int_0^{\frac{\sigma^2}{n}} (u)^{-\frac{1}{2}} \left(1 + \frac{u}{1!} + \frac{u^2}{2} + \frac{u^3}{3!} + \ldots\right) \left(-\gamma - \log(u) + u - \frac{u^2}{4} + \ldots\right) du \\
&= \int_0^{\frac{\sigma^2}{n}} \left\{-\gamma (u)^{-\frac{1}{2}} - (u)^{-\frac{1}{2}} \log(u) + (1-\gamma)(u)^{\frac{1}{2}} - (u)^{\frac{1}{2}} \log(u) + \frac{1}{4}(3 - 2\gamma)(u)^{\frac{3}{2}}\right. \\
&\qquad \left. + \ldots\right\} du \\
&= -\gamma \int_0^{\frac{\sigma^2}{n}} (u)^{-\frac{1}{2}}\, du - \int_0^{\frac{\sigma^2}{n}} (u)^{-\frac{1}{2}} \log(u)\, du + (1-\gamma) \int_0^{\frac{\sigma^2}{n}} (u)^{\frac{1}{2}}\, du + \ldots \\
&= 2\sigma (n)^{-\frac{1}{2}} \log(n) + O\left(n^{-\frac{1}{2}}\right).
\end{aligned}$$

Note that the term $\log\left((\pi)^{-\frac{3}{2}} \int_0^{\frac{\sigma^2}{n}} (u)^{-\frac{1}{2}} \exp(u) E_1(u)\, du\right)$ in the second term of R_n, as stated in (3.143), becomes

$$\begin{aligned}
&\log\left((\pi)^{-\frac{3}{2}} \int_0^{\frac{\sigma^2}{n}} (u)^{-\frac{1}{2}} \exp(u) E_1(u)\, du\right) \\
&\qquad = \log\left(2(\pi)^{-\frac{3}{2}} \sigma \log(n) (n)^{-\frac{1}{2}} + O\left((n)^{-\frac{1}{2}}\right)\right).
\end{aligned}$$

Let

$$\begin{aligned}
a_n &= \log\left(2(\pi)^{-\frac{3}{2}} \sigma \log(n) (n)^{-\frac{1}{2}} + O\left((n)^{-\frac{1}{2}}\right)\right) \\
&= \log\left(C_1 \log(n) (n)^{-\frac{1}{2}} + C_2\left((n)^{-\frac{1}{2}}\right)\right) \\
&= -\frac{1}{2}\log(n) + \log(\log(n)) + \log(C_1) + \frac{C_2/C_1}{\log(n)} - \frac{1}{2}\left(\frac{C_2/C_1}{\log(n)}\right)^2 + \ldots
\end{aligned}$$

where $C_1 = 2(\pi)^{-\frac{3}{2}}\sigma$ and $C_2 = 1$. Then

$$a_n = \log\left((\pi)^{-\frac{3}{2}}\int_0^{\frac{\sigma^2}{n}}(u)^{-\frac{1}{2}}\exp(u)\,E_1(u)\,du\right)$$

$$= -\frac{1}{2}\log(n) + \log(\log(n)) + O(1).$$

Then for $\boldsymbol{\theta}^0 = \mathbf{0}$ and $d = 1$, the risk quantity R_n stated in (3.143) becomes

$$R_n \leq \frac{1}{n} - \frac{1}{n}\log\left((\pi)^{-\frac{3}{2}}\int_0^{\frac{\sigma^2}{n}}(u)^{-\frac{1}{2}}\exp(u)\,E_1(u)\,du\right)$$

$$= \frac{\log(n)}{2n} - \frac{\log\{\log(n)\}}{n} + O\left(\frac{1}{n}\right).$$

The $d \geq 2$ Case

For $d \geq 2$, to analyze

$$\mathcal{J}(d,n,\sigma) \equiv \int_{\|\boldsymbol{\theta}\| \leq \sigma\sqrt{2/n}} \mathfrak{p}_{\mathrm{HS,d}}(\boldsymbol{\theta})\,d\boldsymbol{\theta}, \tag{3.144}$$

we switch to hyper-spherical coordinates as follows:

$$\begin{aligned}
\theta_1 &= r\cos(\phi_1),\\
\theta_2 &= r\sin(\phi_1)\cos(\phi_2),\\
\theta_3 &= r\sin(\phi_1)\sin(\phi_2)\cos(\phi_3),\\
&\vdots\\
\theta_{d-1} &= r\sin(\phi_1)\ldots\sin(\phi_{d-2})\cos(\phi_{d-1}),\\
\theta_d &= r\sin(\phi_1)\ldots\sin(\phi_{d-2})\sin(\phi_{d-1})
\end{aligned}$$

where $r \geq 0$, $0 \leq \phi_1, \phi_2, \ldots, \phi_{d-2} \leq \pi$ and $0 \leq \phi_{d-1} < 2\pi$. Note that $\left\|\boldsymbol{\theta}\right\| = r$ and the determinant of the Jacobian of the transformation is

$$d\boldsymbol{\theta} = r^{d-1}\sin^{d-2}(\phi_1)\sin^{d-3}(\phi_2)\ldots\sin(\phi_{d-2})\,drd\phi_1\ldots d\phi_{d-1}.$$

Also note that $\mathfrak{p}_{\mathrm{HS,d}}(\boldsymbol{\theta})$ stated in (3.144) is derived in **Result 3.2.2.1** in Section 3.2.2 is as follows:

$$\mathfrak{p}_{\mathrm{HS,d}}(\boldsymbol{\theta}) = \frac{\Gamma\left(\frac{1}{2}(d+1)\right)}{\sqrt{2\pi^{d+2}}} \exp\left(\left\|\boldsymbol{\theta}\right\|^2/2\right) E_{(d+1)/2}\left(\left\|\boldsymbol{\theta}\right\|^2/2\right) \Big/ \left\|\boldsymbol{\theta}\right\|^{d-1}.$$

The change of variable

$$\left\|\boldsymbol{\theta}\right\|^2/2 = u$$

in (3.144) leads to $\left\|\boldsymbol{\theta}\right\| = \sqrt{2u} = r$ and $dr = \left(\sqrt{2}/2\right)(u)^{-\frac{1}{2}}\, du$. Using Wallis's Theorem as follows:

$$W_n = \int_0^{\frac{\pi}{2}} \sin^n(x)\, dx = \frac{\Gamma\left(\frac{n+1}{2}\right)\Gamma\left(\frac{1}{2}\right)}{2\Gamma\left(\frac{n}{2}+1\right)},$$

the integral stated in (3.144) becomes

$$\begin{aligned}
\mathcal{J}(d,n,\sigma) &\equiv \int_{\|\boldsymbol{\theta}\|\leq\sigma\sqrt{2/n}} \mathfrak{p}_{\mathrm{HS,d}}(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\
&= \frac{\Gamma\left(\frac{1}{2}(d+1)\right)}{\sqrt{2\pi^{d+2}}} \int_{\|\boldsymbol{\theta}\|\leq\sigma\sqrt{2/n}} \exp\left(\left\|\boldsymbol{\theta}\right\|^2/2\right) E_{(d+1)/2}\left(\left\|\boldsymbol{\theta}\right\|^2/2\right) \frac{1}{\left\|\boldsymbol{\theta}\right\|^{d-1}} r^{d-1} \\
&\quad \times \sin^{d-2}(\phi_1)\sin^{d-3}(\phi_2)\ldots\sin(\phi_{d-2})\, dr d\phi_1\ldots d\phi_{d-1} \\
&= \frac{\Gamma\left(\frac{1}{2}(d+1)\right)}{\sqrt{2\pi^{d+2}}} \int_{\|\boldsymbol{\theta}\|\leq\sigma\sqrt{2/n}} \exp\left(\left\|\boldsymbol{\theta}\right\|^2/2\right) E_{(d+1)/2}\left(\left\|\boldsymbol{\theta}\right\|^2/2\right) \left(\frac{\sqrt{2}}{2}(u)^{-\frac{1}{2}}\, du\right) \\
&\quad \times \left(2\int_0^{\frac{\pi}{2}} \sin^{d-2}(\phi_1)\, d\phi_1\right)\left(2\int_0^{\frac{\pi}{2}} \sin^{d-3}(\phi_2)\, d\phi_2\right)\times\ldots \\
&\quad \times \left(2\int_0^{\frac{\pi}{2}} \sin(\phi_{d-2})\, d\phi_{d-2}\right)\left(\int_0^{2\pi} d\phi_{d-1}\right) \\
&= \frac{\Gamma\left(\frac{1}{2}(d+1)\right)}{\sqrt{2\pi^{d+2}}}\sqrt{2}\pi \left(\int_0^{\frac{\sigma^2}{n}} (u)^{-\frac{1}{2}} \exp(u)\, E_{(d+1)/2}(u)\, du\right)(2W_{d-2})\times\ldots(2W_1) \\
&= \frac{\Gamma\left(\frac{d+1}{2}\right)}{\pi\Gamma\left(\frac{d}{2}\right)} \int_0^{\frac{\sigma^2}{n}} (u)^{-\frac{1}{2}} \exp(u)\, E_{(d+1)/2}(u)\, du.
\end{aligned}$$

In summary,

$$\mathcal{J}(d,n,\sigma) \equiv \int_{\|\boldsymbol{\theta}\|\leq\sigma\sqrt{2/n}} \mathfrak{p}_{\mathrm{HS,d}}(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \frac{\Gamma\left(\frac{d+1}{2}\right)}{\pi\Gamma\left(\frac{d}{2}\right)} \int_0^{\frac{\sigma^2}{n}} (u)^{-\frac{1}{2}} \exp(u)\, E_{(d+1)/2}(u)\, du. \tag{3.145}$$

The $d \geq 2$ Case, d Odd Case

If d is odd, then $(d+1)/2 \in \mathbb{N}$. For the term $E_{(d+1)/2}(u)$ stated in the last step of

(3.145), we use 8.19.8 of Olver et al. (2023) as follows:

$$E_n(z) = \frac{(-z)^{n-1}}{(n-1)!}(\psi(n) - \ln z) - \sum_{\substack{k=0 \\ k \neq n-1}}^{\infty} \frac{(-z)^k}{k!(1-n+k)},$$

where $\psi(n)$ is digamma () function, $n = 1, 2, 3, \ldots$. Then $E_{(d+1)/2}(u)$ stated in the last step of (3.145) becomes

$$E_{(d+1)/2}(u) = \Big(-1\Big)^{\frac{d-1}{2}} \Big(u\Big)^{\frac{d-1}{2}} \left(\frac{\text{digamma}\left(\dfrac{d+1}{2}\right) - \log(u)}{\left(\dfrac{d-1}{2}\right)!} \right) - \sum_{\substack{k=0 \\ k \neq \frac{d-1}{2}}}^{\infty} \frac{2(-1)^k (u)^k}{(2k-d+1)\,k!}.$$

Note that the term $(u)^{-\frac{1}{2}} \exp(u)$ stated in the last step of (3.145) becomes

$$(u)^{-\frac{1}{2}} \exp(u) = (u)^{-\frac{1}{2}} \left(\frac{u^0}{0!} + \frac{u^1}{1!} + \frac{u^2}{2!} + \frac{u^3}{3!} + \ldots \right) = (u)^{-\frac{1}{2}} + (u)^{\frac{1}{2}} + \frac{1}{2}(u)^{-\frac{3}{2}} + \ldots.$$

Then the integral stated in the last step of (3.145) becomes

$$\int_0^{\frac{\sigma^2}{n}} (u)^{-\frac{1}{2}} \exp(u) E_{(d+1)/2}(u)\, du$$

$$= \int_0^{\frac{\sigma^2}{n}} \frac{(-1)^{\frac{d-1}{2}} (u)^{\frac{d-1}{2}} \text{digamma}\left(\frac{d+1}{2}\right) - \log(u)}{\left(\frac{d-1}{2}\right)!} du \tag{3.146}$$

$$- \int_0^{\frac{\sigma^2}{n}} \sum_{\substack{k=0 \\ k \neq \frac{d-1}{2}}}^{\infty} \frac{2(-1)^k (u)^{k-\frac{1}{2}}}{(2k-d+1)\, k!} du$$

$$+ \int_0^{\frac{\sigma^2}{n}} \frac{(-1)^{\frac{d-1}{2}} (u)^{\frac{d}{2}} \text{digamma}\left(\frac{d+1}{2}\right) - \log(u)}{\left(\frac{d-1}{2}\right)!} du \tag{3.147}$$

$$- \int_0^{\frac{\sigma^2}{n}} \sum_{\substack{k=0 \\ k \neq \frac{d-1}{2}}}^{\infty} \frac{2(-1)^k (u)^{k+\frac{1}{2}}}{(2k-d+1)\, k!} du$$

$$+ \ldots. \tag{3.148}$$

For the integrals stated in (3.146), we use the integral results as follows:

$$\begin{aligned}
&\int_0^{\sigma^2/n} (u)^{-\frac{1}{2}}\, du = 2\sigma (n)^{-\frac{1}{2}}, \\
&\int_0^{\sigma^2/n} (u)^{\frac{1}{2}} \log(u)\, du = -\frac{2}{3}\sigma^3 \log(n) (n)^{-\frac{3}{2}} + \frac{2}{9}\{6\log(\sigma) - 2\}\sigma^3 (n)^{-\frac{3}{2}}, \\
&\int_0^{\sigma^2/n} (u)^{\frac{1}{2}}\, du = \frac{2}{3}\sigma^3 (n)^{-\frac{3}{2}}, \\
&\int_0^{\sigma^2/n} (u)^{\frac{3}{2}} \log(u)\, du = -\frac{2}{5}\sigma^5 \log(n) (n)^{-\frac{5}{2}} + \frac{2}{25}\{10\log(\sigma) - 2\}\sigma^5 (n)^{-\frac{5}{2}}, \\
&\text{and } \int_0^{\sigma^2/n} (u)^{\frac{3}{2}}\, du = \frac{2}{5}\sigma^5 (n)^{-\frac{5}{2}},
\end{aligned} \tag{3.149}$$

For $d = 3$, the integral stated in the last step of (3.145) becomes

$$\begin{aligned}
&\int_0^{\frac{\sigma^2}{n}} (u)^{-\frac{1}{2}} \exp(u)\, E_{(d+1)/2}(u)\, du \\
&\quad = \int_0^{\sigma^2/n} (u)^{-\frac{1}{2}}\, du - \text{digamma}(2) \int_0^{\sigma^2/n} (u)^{\frac{1}{2}}\, du + \int_0^{\sigma^2/n} (u)^{\frac{1}{2}} \log(u)\, du \\
&\qquad -\frac{1}{2}\int_0^{\sigma^2/n} (u)^{\frac{3}{2}}\, du + \ldots + \int_0^{\sigma^2/n} (u)^{\frac{1}{2}}\, du - \text{digamma}(2) \int_0^{\sigma^2/n} (u)^{\frac{3}{2}}\, du \\
&\qquad + \int_0^{\sigma^2/n} (u)^{\frac{3}{2}} \log(u)\, du + \ldots \\
&\quad = 2\sigma (n)^{-\frac{1}{2}} + O\left(\log(n)\,(n)^{-\frac{3}{2}}\right).
\end{aligned}$$

For $d = 5$, the integral stated in the last step of (3.145) becomes

$$\begin{aligned}
&\int_0^{\frac{\sigma^2}{n}} (u)^{-\frac{1}{2}} \exp(u)\, E_{(d+1)/2}(u)\, du \\
&\quad = \frac{1}{2}\int_0^{\sigma^2/n} (u)^{-\frac{1}{2}}\, du - \frac{1}{2}\int_0^{\sigma^2/n} (u)^{\frac{1}{2}}\, du + \left(\frac{\text{digamma}(3)}{2} - 1\right)\int_0^{\sigma^2/n} (u)^{\frac{3}{2}}\, du \\
&\qquad -\frac{1}{2}\int_0^{\sigma^2/n} (u)^{\frac{3}{2}} \log(u) + \ldots \\
&\quad = \sigma (n)^{-\frac{1}{2}} + O\left((n)^{-\frac{3}{2}}\right).
\end{aligned}$$

For $d = 7$, the integral stated in the last step of (3.145) becomes

$$\begin{aligned}
&\int_0^{\frac{\sigma^2}{n}} (u)^{-\frac{1}{2}} \exp(u)\, E_{(d+1)/2}(u)\, du \\
&\quad = \frac{1}{3}\int_0^{\sigma^2/n} (u)^{-\frac{1}{2}}\, du - \frac{1}{6}\int_0^{\sigma^2/n} (u)^{\frac{1}{2}}\, du + \left(-\frac{\text{digamma}(4)}{6} + \frac{1}{2}\right)\int_0^{\sigma^2/n} (u)^{\frac{5}{2}}\, du \\
&\qquad -\frac{\text{digamma}(4)}{6}\int_0^{\sigma^2/n} (u)^{\frac{7}{2}}\, du + \frac{1}{6}\int_0^{\sigma^2/n} (u)^{\frac{7}{2}} \log(u)\, du + \ldots \\
&\quad = \frac{2}{3}\sigma (n)^{-\frac{1}{2}} + O\left((n)^{-\frac{3}{2}}\right).
\end{aligned}$$

Application of (3.149) to the early terms in these series of integrals stated in (3.146), for $d = 3, 5, 7$, reveals that the leading term of the integral stated in the last step of (3.145) as follows:

$$-\int_0^{\frac{\sigma^2}{n}} \sum_{\substack{k=0 \\ k \neq \frac{d-1}{2}}}^{\infty} \frac{2\left(-1\right)^k \left(u\right)^{k-\frac{1}{2}}}{\left(2k-d+1\right) k!} du = \frac{2}{d-1} \int_0^{\sigma^2/n} \left(u\right)^{-\frac{1}{2}} du = \frac{4\sigma}{d-1} \left(n\right)^{-\frac{1}{2}}.$$

The second term of the integral stated in the last step of (3.145) is

$$O\left(\log\left(n\right)\left(n\right)^{-\frac{3}{2}}\right) \text{ if } d = 3 \text{ and } O\left(\left(n\right)^{-\frac{3}{2}}\right) \text{ if } d = 5, 7, 9, \ldots.$$

Therefore, for $\boldsymbol{\theta}^0 = \mathbf{0}$, the integral stated in the last step of (3.145) is

$$\int_0^{\frac{\sigma^2}{n}} \left(u\right)^{-\frac{1}{2}} \exp\left(u\right) E_{(d+1)/2}\left(u\right) du = \frac{4\sigma}{d-1} \left(n\right)^{-\frac{1}{2}} + O\left(\left(\left(n\right)^{-\frac{3}{2}}\right) \left\{\log\left(n\right)\right\}^{I(d=3)}\right), \tag{3.150}$$

for all odd integers d exceeding 1.

The $d \geq 2$ Case, d Even Case

If d is even, for the term $E_{(d+1)/2}\left(u\right)$ stated in the last step of (3.145), we use 8.19.11 of Olver et al. (2023) as follows:

$$E_p\left(z\right) = \Gamma\left(1-p\right)\left(\left(z\right)^{p-1} - \exp\left(-z\right)\sum_{k=0}^{\infty} \frac{\left(z\right)^k}{\Gamma\left(2-p+k\right)}\right).$$

Then $E_{(d+1)/2}\Big(u\Big)$ stated in the last step of (3.145) becomes

$$E_{(d+1)/2}\Big(u\Big) = \Gamma\left(\frac{1-d}{2}\right)\left(\Big(u\Big)^{\frac{d-1}{2}} - \exp\left(-u\right)\sum_{k=0}^{\infty} \frac{\left(u\right)^k}{\Gamma\left(k+\frac{3-d}{2}\right)}\right).$$

We then have

$$\left(\Gamma\left(\frac{1-d}{2}\right)\right)^{-1} \exp\Big(u\Big) E_{(d+1)/2}\Big(u\Big) = \Big(u\Big)^{\frac{d-1}{2}} \exp\Big(u\Big) - \sum_{k=0}^{\infty} \frac{\left(u\right)^k}{\Gamma\left(k+\frac{3-d}{2}\right)}$$

and

$$\begin{aligned}
&\left(\Gamma\left(\frac{1-d}{2}\right)\right)^{-1}\int_0^{\frac{\sigma^2}{n}}\left(u\right)^{-\frac{1}{2}}\exp\left(u\right)E_{(d+1)/2}\left(u\right)du \\
&=\int_0^{\frac{\sigma^2}{n}}\left(u\right)^{-\frac{1}{2}}\left(\left(u\right)^{\frac{d-1}{2}}\left(\frac{u^0}{0!}+\frac{u^1}{1!}+\frac{u^2}{2!}+\frac{u^3}{3!}+\ldots\right)-\sum_{k=0}^{\infty}\frac{\left(u\right)^k}{\Gamma\left(k+\frac{3-d}{2}\right)}\right)du \\
&=\int_0^{\frac{\sigma^2}{n}}\left(u\right)^{\frac{d-2}{2}}du+\int_0^{\frac{\sigma^2}{n}}\left(u\right)^{\frac{d}{2}}du+\int_0^{\frac{\sigma^2}{n}}\frac{1}{2}\left(u\right)^{\frac{d+2}{2}}du+\ldots \\
&\quad-\sum_{k=0}^{\infty}\frac{\frac{2}{1+2k}\left(\sigma\right)^{1+2k}\left(n\right)^{-\frac{1+2k}{2}}}{\Gamma\left(k+\frac{3-d}{2}\right)} \\
&=O\left(n\right)^{-\frac{d}{2}}-\frac{2\sigma\left(n\right)^{-\frac{1}{2}}}{\Gamma\left(\frac{3-d}{2}\right)}+O\left(n\right)^{-\frac{3}{2}} \\
&=-\frac{2\sigma\left(n\right)^{-\frac{1}{2}}}{\Gamma\left(\frac{3-d}{2}\right)}+O\left(\left(n\right)^{-\frac{\min(d,3)}{2}}\right).
\end{aligned}$$

Then, for all even positive integers d, the integral stated in the last step of (3.145) is

$$\begin{aligned}
\int_0^{\frac{\sigma^2}{n}}\left(u\right)^{-\frac{1}{2}}\exp\left(u\right)E_{(d+1)/2}\left(u\right)du&=-\frac{2\sigma\Gamma\left(\frac{1-d}{2}\right)\left(n\right)^{-\frac{1}{2}}}{\Gamma\left(1+\frac{1-d}{2}\right)}+O\left(\left(n\right)^{-\frac{\min(d,3)}{2}}\right) \\
&=\left(\frac{4\sigma}{d-1}\right)\left(n\right)^{-\frac{1}{2}}O\left(\left(n\right)^{-\frac{\min(d,3)}{2}}\right). \qquad (3.151)
\end{aligned}$$

In summary, for $\boldsymbol{\theta}^0 = \mathbf{0}$ case, the risk quantity R_n is stated in (3.140). For $d \geq 2$, and d is odd, the integral of R_n is stated in (3.150). For $d \geq 2$, and d is even case, the integral of R_n is stated in (3.151). Overall, we have the following expression for R_n

when $\boldsymbol{\theta}^0 = \mathbf{0}$, for all $d \geq 2$:

$$\begin{aligned} R_n &\leq \frac{1}{n} - \frac{1}{n} \log \left(\frac{\Gamma\left(\frac{d+1}{2}\right)}{\pi \Gamma\left(\frac{d}{2}\right)} \int_0^{\frac{\sigma^2}{n}} \left(u\right)^{-\frac{1}{2}} \exp\left(u\right) E_{(d+1)/2}\left(u\right) du \right) \\ &= \frac{1}{n} - \frac{1}{n} \log \left(\frac{\Gamma\left(\frac{d+1}{2}\right)}{\pi \Gamma\left(\frac{d}{2}\right)} \left\{ \frac{4\sigma \left(n\right)^{-\frac{1}{2}}}{d-1} + O\left(\left(n\right)^{-1}\right) \right\} \right) \\ &= O\left(\frac{1}{n}\right) + \frac{\log\left(n\right)}{2n}. \end{aligned}$$

3.D.2 The $\boldsymbol{\theta}^0 \neq \mathbf{0}$ Case

In the $\boldsymbol{\theta}^0 \neq \mathbf{0}$ case we have:

$$R_n \leq \frac{1}{n} - \frac{1}{n} \log \left(\int_{S\left(\boldsymbol{\theta}^0, \sigma, n\right)} \mathfrak{p}_{\mathrm{HS,d}}\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta} \right) \tag{3.152}$$

where

$$S\left(\boldsymbol{\theta}^0, \sigma, n\right) \equiv \left\{ \boldsymbol{\theta} : \left\| \boldsymbol{\theta} - \boldsymbol{\theta}^0 \right\| \leq \sigma \sqrt{2/n} \right\}.$$

If $C\left(\boldsymbol{\theta}^0, \sigma, n\right)$ is the largest hypercube inscribed in $S\left(\boldsymbol{\theta}^0, \sigma, n\right)$ then

$$\int_{S\left(\boldsymbol{\theta}^0, \sigma, n\right)} \mathfrak{p}_{\mathrm{HS,d}}\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta} \geq \int_{C\left(\boldsymbol{\theta}^0, \sigma, n\right)} \mathfrak{p}_{\mathrm{HS,d}}\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta}. \tag{3.153}$$

To obtain $\int_{C\left(\boldsymbol{\theta}^0, \sigma, n\right)} \mathfrak{p}_{\mathrm{HS,d}}\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta}$ for $d = 1$, $d = 2$ and $d = 3$ cases, we investigate the rates of convergence of integrals of smooth d-variate functions over hypercubes with $O\left(n^{-1/2}\right)$ rate of convergence in each direction.

Let function $f : \mathbb{R}^d \to \mathbb{R}$ be a smooth function. Then we use the following notation for partial derivatives of f :

$$f^{r_1, \ldots, r_d}\left(x_1, \ldots, x_d\right) \equiv \frac{\partial^{r_1 + \ldots + r_d}}{\partial x_1^{r_1}, \ldots, \partial x_d^{r_d}} f\left(x_1, \ldots, x_d\right).$$

The $d = 1$ Case

Consider the interval

$$C\left(x^0, 1, a, n\right) \equiv \left[x^0 - an^{-1/2}, x^0 + an^{-1/2}\right]$$

where $x^0 \in \mathbb{R}, a > 1$ and $n \in \mathbb{N}$. Suppose that the function $g : \mathbb{R} \to \mathbb{R}$ is very smooth over $C\left(x^0, 1, a, n\right)$ and the function G is such that $G^{(1)} = g$. Using Taylor expansion for one variable, we then have

$$\begin{aligned}
\int_{C(x^0,1,a,n)} g\left(x\right) dx &= \int_{x^0-an^{-1/2}}^{x^0+an^{-1/2}} g\left(x\right) dx = G\left(x^0 + an^{-1/2}\right) - G\left(x^0 - an^{-1/2}\right) \\
&= G\left(x^0\right) + an^{-1/2} g\left(x^0\right) + \frac{1}{2}a^2 n^{-1} g^{(1)}\left(x^0\right) + \frac{1}{6}a^3 n^{-3/2} g^{(2)}\left(x^0\right) + O\left(n^{-2}\right) \\
&\quad - \left\{G\left(x^0\right) - an^{-1/2} g\left(x^0\right) + \frac{1}{2}a^2 n^{-1} g^{(1)}\left(x^0\right) - \frac{1}{6}a^3 n^{-3/2} g^{(2)}\left(x^0\right) + O\left(n^{-2}\right)\right\} \\
&= 2an^{-1/2} g\left(x^0\right) + \frac{1}{3}a^3 n^{-3/2} g^{(2)}\left(x^0\right) + O\left(n^{-2}\right) \\
&= 2ag\left(x^0\right) n^{-1/2} \left\{1 + o\left(1\right)\right\}.
\end{aligned}$$

The $d = 2$ Case

Consider the square

$$\begin{aligned}
C\left(\boldsymbol{x}^0, 2, a, n\right) \equiv \big\{(x_1, x_2) : &x_1^0 - an^{-1/2} \leq x_1 \leq x_1^0 + an^{-1/2}, \\
&x_2^0 - an^{-1/2} \leq x_2 \leq x_2^0 + an^{-1/2}\big\}
\end{aligned}$$

where $x_1^0, x_2^0 \in \mathbb{R}, \boldsymbol{x}^0 \equiv \left(x_1^0, x_2^0\right), a > 1$ and $n \in \mathbb{N}$. Suppose that the function

$$g : \mathbb{R}^2 \to \mathbb{R}$$

is very smooth over $C\left(\boldsymbol{x}^0, 2, a, n\right)$ and the function G is such that $G^{(1,1)} = g$. Using

Taylor expansion for two variables, we then have

$$\int_{C(\boldsymbol{x}^0,2,a,n)} g(\boldsymbol{x})\, d\boldsymbol{x} = \int_{x_2^0-an^{-1/2}}^{x_2^0+an^{-1/2}} \left\{ \int_{x_1^0-an^{-1/2}}^{x_1^0+an^{-1/2}} g(x_1,x_2)\, dx_1 \right\} dx_2$$

$$= \int_{x_2^0-an^{-1/2}}^{x_2^0+an^{-1/2}} \left\{ G^{(0,1)}\left(x_1^0+an^{-1/2},x_2\right) - G^{(0,1)}\left(x_1^0-an^{-1/2},x_2\right) \right\} dx_2$$

$$= \int_{x_2^0-an^{-1/2}}^{x_2^0+an^{-1/2}} \left\{ G^{(0,1)}\left(x_1^0+an^{-1/2},x_2\right) \right\} dx_2 - \int_{x_2^0-an^{-1/2}}^{x_2^0+an^{-1/2}} \left\{ G^{(0,1)}\left(x_1^0-an^{-1/2},x_2\right) \right\} dx_2$$

$$= G\left(x_1^0+an^{-1/2},x_2^0+an^{-1/2}\right) - G\left(x_1^0+an^{-1/2},x_2^0-an^{-1/2}\right)$$

$$-G\left(x_1^0-an^{-1/2},x_2^0+an^{-1/2}\right) + G\left(x_1^0-an^{-1/2},x_2^0-an^{-1/2}\right).$$

Next note that for each $s_1, s_2 \in \{-1, 1)$,

$$G\left(x_1^0+s_1an^{-1/2},x_2^0+s_2an^{-1/2}\right)$$

$$= G\left(x_1^0,x_2^0\right) + s_1an^{-1/2}G^{(1,0)}\left(x_1^0,x_2^0\right) + s_2an^{-1/2}G^{(0,1)}\left(x_1^0,x_2^0\right)$$

$$+\frac{1}{2}a^2n^{-1}G^{(2,0)}\left(x_1^0,x_2^0\right) + \frac{1}{2}a^2s_1s_2n^{-1}G^{(1,1)}\left(x_1^0,x_2^0\right) + \frac{1}{2}a^2n^{-1}G^{(0,2)}\left(x_1^0,x_2^0\right)$$

$$+O\left(n^{-3/2}\right).$$

This result leads to

$$\begin{aligned}
\int_{C(\boldsymbol{x}^0,2,a,n)} g(\boldsymbol{x})\, d\boldsymbol{x} &= G\left(x_1^0, x_2^0\right) + an^{-1/2}G^{(1,0)}\left(x_1^0, x_2^0\right) + an^{-1/2}G^{(0,1)}\left(x_1^0, x_2^0\right) \\
&+\frac{1}{2}a^2n^{-1}G^{(2,0)}\left(x_1^0, x_2^0\right) + \frac{1}{2}a^2n^{-1}G^{(1,1)}\left(x_1^0, x_2^0\right) + \frac{1}{2}a^2n^{-1}G^{(0,2)}\left(x_1^0, x_2^0\right) + O\left(n^{-3/2}\right) \\
&-\left\{G\left(x_1^0, x_2^0\right) + an^{-1/2}G^{(1,0)}\left(x_1^0, x_2^0\right) - an^{-1/2}G^{(0,1)}\left(x_1^0, x_2^0\right) + \frac{1}{2}a^2n^{-1}G^{(2,0)}\left(x_1^0, x_2^0\right)\right. \\
&\left.-\frac{1}{2}a^2n^{-1}G^{(1,1)}\left(x_1^0, x_2^0\right) + \frac{1}{2}a^2n^{-1}G^{(0,2)}\left(x_1^0, x_2^0\right) + O\left(n^{-3/2}\right)\right\} \\
&-\left\{G\left(x_1^0, x_2^0\right) - an^{-1/2}G^{(1,0)}\left(x_1^0, x_2^0\right) + an^{-1/2}G^{(0,1)}\left(x_1^0, x_2^0\right) + \frac{1}{2}a^2n^{-1}G^{(2,0)}\left(x_1^0, x_2^0\right)\right. \\
&\left.-\frac{1}{2}a^2n^{-1}G^{(1,1)}\left(x_1^0, x_2^0\right) + \frac{1}{2}a^2n^{-1}G^{(0,2)}\left(x_1^0, x_2^0\right) + O\left(n^{-3/2}\right)\right\} \\
&+G\left(x_1^0, x_2^0\right) - an^{-1/2}G^{(1,0)}\left(x_1^0, x_2^0\right) - an^{-1/2}G^{(0,1)}\left(x_1^0, x_2^0\right) + \frac{1}{2}a^2n^{-1}G^{(2,0)}\left(x_1^0, x_2^0\right) \\
&+\frac{1}{2}a^2n^{-1}G^{(1,1)}\left(x_1^0, x_2^0\right) + \frac{1}{2}a^2n^{-1}G^{(0,2)}\left(x_1^0, x_2^0\right) + O\left(n^{-3/2}\right).
\end{aligned}$$

After some simplifications, we obtain

$$\begin{aligned}
\int_{C(\boldsymbol{x}^0,2,a,n)} g(\boldsymbol{x})\, d\boldsymbol{x} &= 2a^2n^{-1}G^{(1,1)}\left(x_1^0, x_2^0\right) + O\left(n^{-3/2}\right) \\
&= 2a^2n^{-1}g\left(\boldsymbol{x}^0\right)\left(x_1^0, x_2^0\right)\left\{1 + o(1)\right\}.
\end{aligned}$$

The $d = 3$ Case

Consider the cube

$$\begin{aligned}
C\left(\boldsymbol{x}^0, 3, a, n\right) \equiv \big\{&(x_1, x_2, x_3) : x_1^0 - an^{-1/2} \leq x_1 \leq x_1^0 + an^{-1/2}, \\
&x_2^0 - an^{-1/2} \leq x_2 \leq x_2^0 + an^{-1/2}, x_3^0 - an^{-1/2} \leq x_3 \leq x_3^0 + an^{-1/2}\big\}
\end{aligned}$$

where $x_1^0, x_2^0, x_3^0 \in \mathbb{R}$, $\boldsymbol{x}^0 \equiv \left(x_1^0, x_2^0, x_3^0\right)$, $a > 1$ and $n \in \mathbb{N}$. Suppose that the function $g : \mathbb{R}^3 \to \mathbb{R}$ is very smooth over $C\left(\boldsymbol{x}^0, 3, a, n\right)$ and the function G is such that $G^{(1,1,1)} = g$.

Using Taylor expansion for three variables, we have

$$\begin{aligned}
\int_{C(\boldsymbol{x}^0,3,a,n)} g(\boldsymbol{x})\, d\boldsymbol{x} &= \int_{x_3^0-an^{-1/2}}^{x_3^0+an^{-1/2}} \left[\int_{x_2^0-an^{-1/2}}^{x_2^0+an^{-1/2}} \left\{ \int_{x_1^0-an^{-1/2}}^{x_1^0+an^{-1/2}} g(x_1,x_2,x_3)\, dx_1 \right\} dx_2 \right] dx_3 \\
&= \int_{x_3^0-an^{-1/2}}^{x_3^0+an^{-1/2}} \left[\int_{x_2^0-an^{-1/2}}^{x_2^0+an^{-1/2}} \left\{ G^{(0,1,1)}\left(x_1^0+an^{-1/2},x_2,x_3\right) \right.\right. \\
&\qquad \left.\left. - G^{(0,1,1)}\left(x_1^0-an^{-1/2},x_2,x_3\right) \right\} dx_2 \right] dx_3 \\
&= \int_{x_3^0-an^{-1/2}}^{x_3^0+an^{-1/2}} \left[\int_{x_2^0-an^{-1/2}}^{x_2^0+an^{-1/2}} G^{(0,1,1)}\left(x_1^0+an^{-1/2},x_2,x_3\right) dx_2 \right] dx_3 \\
&\quad - \int_{x_3^0-an^{-1/2}}^{x_3^0+an^{-1/2}} \left[\int_{x_2^0-an^{-1/2}}^{x_2^0+an^{-1/2}} G^{(0,1,1)}\left(x_1^0-an^{-1/2},x_2,x_3\right) dx_2 \right] dx_3 \\
&= \int_{x_3^0-an^{-1/2}}^{x_3^0+an^{-1/2}} \left[G^{(0,0,1)}\left(x_1^0+an^{-1/2},x_2^0+an^{-1/2},x_3\right) \right. \\
&\qquad \left. - G^{(0,0,1)}\left(x_1^0+an^{-1/2},x_2^0-an^{-1/2},x_3\right) \right] dx_3 \\
&\quad - \int_{x_3^0-an^{-1/2}}^{x_3^0+an^{-1/2}} \left[G^{(0,0,1)}\left(x_1^0-an^{-1/2},x_2^0+an^{-1/2},x_3\right) \right. \\
&\qquad \left. - G^{(0,0,1)}\left(x_1^0-an^{-1/2},x_2^0-an^{-1/2},x_3\right) \right] dx_3 \\
&= \int_{x_3^0-an^{-1/2}}^{x_3^0+an^{-1/2}} G^{(0,0,1)}\left(x_1^0+an^{-1/2},x_2^0+an^{-1/2},x_3\right) dx_3 \\
&\quad - \int_{x_3^0-an^{-1/2}}^{x_3^0+an^{-1/2}} G^{(0,0,1)}\left(x_1^0+an^{-1/2},x_2^0-an^{-1/2},x_3\right) dx_3 \\
&\quad - \int_{x_3^0-an^{-1/2}}^{x_3^0+an^{-1/2}} G^{(0,0,1)}\left(x_1^0-an^{-1/2},x_2^0+an^{-1/2},x_3\right) dx_3 \\
&\quad + \int_{x_3^0-an^{-1/2}}^{x_3^0+an^{-1/2}} G^{(0,0,1)}\left(x_1^0-an^{-1/2},x_2^0-an^{-1/2},x_3\right) dx_3.
\end{aligned}$$

This leads to

$$\int_{C(\boldsymbol{x}^0,3,a,n)} g(\boldsymbol{x})\, d\boldsymbol{x} = \left\{G\left(x_1^0 + an^{-1/2}, x_2^0 + an^{-1/2}, x_3^0 + an^{-1/2}\right)\right. \tag{3.154}$$

$$-G\left(x_1^0 + an^{-1/2}, x_2^0 + an^{-1/2}, x_3^0 - an^{-1/2}\right)\} \tag{3.155}$$

$$-\left\{G\left(x_1^0 + an^{-1/2}, x_2^0 - an^{-1/2}, x_3^0 + an^{-1/2}\right)\right. \tag{3.156}$$

$$-G\left(x_1^0 + an^{-1/2}, x_2^0 - an^{-1/2}, x_3^0 - an^{-1/2}\right)\} \tag{3.157}$$

$$-\left\{G\left(x_1^0 - an^{-1/2}, x_2^0 + an^{-1/2}, x_3^0 + an^{-1/2}\right)\right. \tag{3.158}$$

$$-G\left(x_1^0 - an^{-1/2}, x_2^0 + an^{-1/2}, x_3^0 - an^{-1/2}\right)\} \tag{3.159}$$

$$+\left\{G\left(x_1^0 - an^{-1/2}, x_2^0 - an^{-1/2}, x_3^0 + an^{-1/2}\right)\right. \tag{3.160}$$

$$-G\left(x_1^0 - an^{-1/2}, x_2^0 - an^{-1/2}, x_3^0 - an^{-1/2}\right)\}. \tag{3.161}$$

Next note that for each $s_1, s_2, s_3 \in \{-1, 1\}$,

$$\begin{aligned}
&G\left(x_1^0 + s_1 an^{-1/2}, x_2^0 + s_2 an^{-1/2}, x_3^0 + s_3 an^{-1/2}\right) = G\left(x_1^0, x_2^0, x_3^0\right) \\
&+s_1 an^{-1/2} G^{(1,0,0)}\left(x_1^0, x_2^0, x_3^0\right) + s_2 an^{-1/2} G^{(0,1,0)}\left(x_1^0, x_2^0, x_3^0\right) + s_3 an^{-1/2} G^{(0,0,1)}\left(x_1^0, x_2^0, x_3^0\right) \\
&+\frac{1}{2}a^2 n^{-1} G^{(2,0,0)}\left(x_1^0, x_2^0, x_3^0\right) + \frac{1}{2}a^2 n^{-1} G^{(0,2,0)}\left(x_1^0, x_2^0, x_3^0\right) + \frac{1}{2}a^2 n^{-1} G^{(0,0,2)}\left(x_1^0, x_2^0, x_3^0\right) \\
&+\frac{1}{2}a^2 s_1 s_2 n^{-1} G^{(1,1,0)}\left(x_1^0, x_2^0, x_3^0\right) + \frac{1}{2}a^2 s_1 s_3 n^{-1} G^{(1,0,1)}\left(x_1^0, x_2^0, x_3^0\right) \\
&+\frac{1}{2}a^2 s_2 s_3 n^{-1} G^{(0,1,1)}\left(x_1^0, x_2^0, x_3^0\right) \\
&+\frac{1}{6}a^3 s_1^3 n^{-3/2} G^{(3,0,0)}\left(x_1^0, x_2^0, x_3^0\right) + \frac{1}{6}a^3 s_2^3 n^{-3/2} G^{(0,3,0)}\left(x_1^0, x_2^0, x_3^0\right) \\
&+\frac{1}{6}a^3 s_3^3 n^{-3/2} G^{(0,0,3)}\left(x_1^0, x_2^0, x_3^0\right) \\
&+\frac{1}{6}a^3 s_2 n^{-3/2} G^{(2,1,0)}\left(x_1^0, x_2^0, x_3^0\right) + \frac{1}{6}a^3 s_3 n^{-3/2} G^{(2,0,1)}\left(x_1^0, x_2^0, x_3^0\right) \\
&+\frac{1}{6}a^3 s_1 n^{-3/2} G^{(1,2,0)}\left(x_1^0, x_2^0, x_3^0\right) + \frac{1}{6}a^3 s_3 n^{-3/2} G^{(0,2,1)}\left(x_1^0, x_2^0, x_3^0\right) \\
&+\frac{1}{6}a^3 s_1 n^{-3/2} G^{(1,0,2)}\left(x_1^0, x_2^0, x_3^0\right) + \frac{1}{6}a^3 s_2 n^{-3/2} G^{(0,1,2)}\left(x_1^0, x_2^0, x_3^0\right) \\
&+\frac{1}{6}a^3 s_1 s_2 s_3 n^{-3/2} G^{(1,1,1)}\left(x_1^0, x_2^0, x_3^0\right) + O\left(n^{-2}\right).
\end{aligned} \tag{3.162}$$

Note that for (3.154): $\{s_1, s_2, s_3\} = \{1, 1, 1\}$; for (3.155): $\{s_1, s_2, s_3\} = \{1, 1, -1\}$; for (3.156), $\{s_1, s_2, s_3\} = \{1, -1, 1\}$; for (3.157), $\{s_1, s_2, s_3\} = \{1, -1, -1\}$; for (3.158), $\{s_1, s_2, s_3\} = \{-1, 1, 1\}$; for (3.159), $\{s_1, s_2, s_3\} = \{-1, 1, -1\}$; for (3.160), $\{s_1, s_2, s_3\} = \{-1, -1, 1\}$; for (3.161), $\{s_1, s_2, s_3\} = \{-1, -1, -1\}$.

It is clear from (3.162), the coefficient of $G\left(x_1^0, x_2^0, x_3^0\right)$ is

$$(1-1) - (1-1) - (1-1) + (1-1) = 0.$$

Hence, the $O\left(1\right)$ term corresponding to $G\left(x_1^0, x_2^0, x_3^0\right)$ cancels.

Similarly, the $O\left(n^{-1/2}\right)$ term corresponding to the following terms all cancel:

$$an^{-1/2}G^{(1,0,0)}\left(x_1^0, x_2^0, x_3^0\right), \quad an^{-1/2}G^{(0,1,0)}\left(x_1^0, x_2^0, x_3^0\right) \quad \text{and} \quad an^{-1/2}G^{(0,0,1)}\left(x_1^0, x_2^0, x_3^0\right).$$

The $O\left(n^{-1}\right)$ term corresponding to the following terms all cancel:

$$\frac{1}{2}a^2n^{-1}G^{(2,0,0)}\left(x_1^0, x_2^0, x_3^0\right), \quad \frac{1}{2}a^2n^{-1}G^{(0,2,0)}\left(x_1^0, x_2^0, x_3^0\right), \quad \frac{1}{2}a^2n^{-1}G^{(0,0,2)}\left(x_1^0, x_2^0, x_3^0\right),$$

$$\frac{1}{2}a^2n^{-1}G^{(1,1,0)}\left(x_1^0, x_2^0, x_3^0\right), \quad \frac{1}{2}a^2n^{-1}G^{(1,0,1)}\left(x_1^0, x_2^0, x_3^0\right) \quad \text{and} \quad \frac{1}{2}a^2n^{-1}G^{(0,1,1)}\left(x_1^0, x_2^0, x_3^0\right).$$

The $O\left(n^{-3/2}\right)$ term corresponding to the following terms all cancel:

$$\frac{1}{6}a^3n^{-3/2}G^{(3,0,0)}\left(x_1^0, x_2^0, x_3^0\right), \quad \frac{1}{6}a^3n^{-3/2}G^{(0,3,0)}\left(x_1^0, x_2^0, x_3^0\right), \quad \frac{1}{6}a^3n^{-3/2}G^{(0,0,3)}\left(x_1^0, x_2^0, x_3^0\right),$$

$$\frac{1}{6}a^3n^{-3/2}G^{(2,1,0)}\left(x_1^0, x_2^0, x_3^0\right), \quad \frac{1}{6}a^3n^{-3/2}G^{(2,0,1)}\left(x_1^0, x_2^0, x_3^0\right), \quad \frac{1}{6}a^3n^{-3/2}G^{(1,2,0)}\left(x_1^0, x_2^0, x_3^0\right)$$

$$\frac{1}{6}a^3n^{-3/2}G^{(0,2,1)}\left(x_1^0, x_2^0, x_3^0\right), \quad \frac{1}{6}a^3n^{-3/2}G^{(1,0,2)}\left(x_1^0, x_2^0, x_3^0\right)$$

$$\text{and} \quad \frac{1}{6}a^3n^{-3/2}G^{(0,1,2)}\left(x_1^0, x_2^0, x_3^0\right).$$

The coefficient of $\frac{1}{6}a^3n^{-3/2}G^{(1,1,1)}\left(x_1^0, x_2^0, x_3^0\right)$ is

$$(1+1) - (-1-1) - (-1-1) + (1+1) = 8.$$

Hence, the $O\left(n^{-3/2}\right)$ term corresponding to $\frac{1}{6}a^3n^{-3/2}G^{(1,1,1)}\left(x_1^0, x_2^0, x_3^0\right)$ does not cancel.

Therefore

$$\int_{C(\boldsymbol{x}^0,3,a,n)} g(\boldsymbol{x})\, d\boldsymbol{x} = \frac{4}{3}a^3 n^{-3/2} G^{(1,1,1)}\left(x_1^0, x_2^0, x_3^0\right) + O\left(n^{-2}\right)$$
$$= \frac{4}{3}a^3 n^{-3/2} g\left(x_1^0, x_2^0, x_3^0\right) + O\left(n^{-2}\right) = \frac{4}{3}a^3 n^{-3/2} g\left(\boldsymbol{x}^0\right)\{1 + o(1)\}.$$

In summary, for $\boldsymbol{\theta}^0 \neq \mathbf{0}, d = 1$,

$$\int_{C(x^0,1,a,n)} g(x)\, dx = 2ag\left(x^0\right) n^{-1/2}\{1 + o(1)\};$$

for $\boldsymbol{\theta}^0 \neq \mathbf{0}, d = 2$,

$$\int_{C(\boldsymbol{x}^0,2,a,n)} g(\boldsymbol{x})\, dx = 2a^2 n^{-1} g\left(\boldsymbol{x}^0\right)\{1 + o(1)\};$$

and, for $\boldsymbol{\theta}^0 \neq \mathbf{0}, d = 3$,

$$\int_{C(\boldsymbol{x}^0,3,a,n)} g(\boldsymbol{x})\, dx = \frac{4}{3}a^3 n^{-3/2} g\left(\boldsymbol{x}^0\right)\{1 + o(1)\}.$$

These results show that for sufficiently large n, $\mathfrak{p}_{\mathrm{HS},d}(\boldsymbol{\theta})$ is a very smooth function over $C\left(\boldsymbol{\theta}^0, \sigma, n\right)$ and Taylor series arguments can be used to establish that

$$\int_{C\left(\boldsymbol{\theta}^0,\sigma,n\right)} \mathfrak{p}_{\mathrm{HS},d}(\boldsymbol{\theta})\, d\boldsymbol{\theta} = K\left(d, \boldsymbol{\theta}^0, \sigma\right) n^{-d/2}\{1 + o(1)\}, \tag{3.163}$$

for some positive constant $K\left(d, \boldsymbol{\theta}^0, \sigma\right)$. Application of (3.153) and (3.163) to the bound in (3.152) then leads to

$$R_n \leq \frac{1}{n} - \frac{1}{n}\log\left(K\left(d, \boldsymbol{\theta}^0, \sigma\right) n^{-d/2}\{1 + o(1)\}\right) = \frac{d}{2}\frac{\log(n)}{n} + O\left(\frac{1}{n}\right).$$

Therefore, for $\boldsymbol{\theta}^0 \neq \mathbf{0}$, the bound of risk quantity R_n is as follows:

$$R_n \leq \frac{d\log(n)}{2n} + O\left(\frac{1}{n}\right), \text{ for all } d \in \mathbb{N}.$$

3.E Derivation of Result 3.2.6.1

We have the model as follows:

$$\boldsymbol{y}|\boldsymbol{\psi} \sim N\left(\boldsymbol{\psi}, \tau_1^2 \boldsymbol{I}_d\right), \;\; \boldsymbol{\psi}|\lambda \sim N\left(\boldsymbol{0}, \lambda^2 \tau_2^2 \boldsymbol{I}_d\right) \;\; \mathfrak{p}\left(\lambda\right) = \frac{2I\left(\lambda > 0\right)}{\pi\left(1+\lambda^2\right)}.$$

In this subsection, we will show the derivation of the posterior mean of $\boldsymbol{\psi}$. Firstly, we obtain the explicit expression for the posterior density function of λ. Note that

$$\begin{aligned}\mathfrak{p}\left(\lambda|\boldsymbol{y}\right) &= \frac{\mathfrak{p}\left(\boldsymbol{y},\lambda\right)}{\mathfrak{p}\left(\boldsymbol{y}\right)} = \frac{\int_{\mathbb{R}^d}\mathfrak{p}\left(\boldsymbol{y},\boldsymbol{\psi},\lambda\right)d\boldsymbol{\psi}}{\mathfrak{p}\left(\boldsymbol{y}\right)} = \frac{\int_{\mathbb{R}^d}\mathfrak{p}\left(\boldsymbol{y}|\boldsymbol{\psi}\right)\mathfrak{p}\left(\boldsymbol{\psi}|\lambda\right)\mathfrak{p}\left(\lambda\right)d\boldsymbol{\psi}}{\mathfrak{p}\left(\boldsymbol{y}\right)} \\ &= \mathfrak{p}\left(\lambda\right)\frac{\int_{\mathbb{R}^d}\mathfrak{p}\left(\boldsymbol{y}|\boldsymbol{\psi}\right)\mathfrak{p}\left(\boldsymbol{\psi}|\lambda\right)d\boldsymbol{\psi}}{\mathfrak{p}\left(\boldsymbol{y}\right)}. \end{aligned} \tag{3.164}$$

Let us look into the terms $\mathfrak{p}\left(\boldsymbol{y}|\boldsymbol{\psi}\right)$ and $\mathfrak{p}\left(\boldsymbol{\psi}|\lambda\right)$ stated in the last step of (3.164). Note that

$$\mathfrak{p}\left(\boldsymbol{y}|\boldsymbol{\psi}\right) = \left(2\pi\tau_1^2\right)^{-d/2}\exp\left(-\frac{\left\|\boldsymbol{y}\right\|^2}{2\tau_1^2}\right)\exp\left\{\left(1/\tau_1^2\right)\begin{bmatrix}\boldsymbol{\psi}\\ \\ \text{vec}\left(\boldsymbol{\psi}\boldsymbol{\psi}^T\right)\end{bmatrix}^T\begin{bmatrix}\boldsymbol{y}\\ \\ -\frac{1}{2}\text{vec}\left(\boldsymbol{I}_d\right)\end{bmatrix}\right\}$$

and

$$\mathfrak{p}\left(\boldsymbol{\psi}|\lambda\right) = \left(2\pi\lambda^2\tau_2^2\right)^{-d/2}\exp\left\{\left(1/\lambda^2\tau_2^2\right)\begin{bmatrix}\boldsymbol{\psi}\\ \\ \text{vec}\left(\boldsymbol{\psi}\boldsymbol{\psi}^T\right)\end{bmatrix}^T\begin{bmatrix}\boldsymbol{0}\\ \\ -\frac{1}{2}\text{vec}\left(\boldsymbol{I}_d\right)\end{bmatrix}\right\}.$$

Then

$$\begin{aligned}\mathfrak{p}\left(\boldsymbol{y}|\boldsymbol{\psi}\right)\mathfrak{p}\left(\boldsymbol{\psi}|\lambda\right) &= \left(2\pi\tau_1^2\right)^{-d/2}\exp\left(-\frac{\left\|\boldsymbol{y}\right\|^2}{2\tau_1^2}\right)\left(2\pi\lambda^2\tau_2^2\right)^{-d/2} \\ &\quad\times\exp\left\{\begin{bmatrix}\boldsymbol{\psi}\\ \\ \text{vec}\left(\boldsymbol{\psi}\boldsymbol{\psi}^T\right)\end{bmatrix}^T\begin{bmatrix}\left(1/\tau_1^2\right)\boldsymbol{y}\\ \\ -\frac{1}{2}\left(\frac{1}{\tau_1^2}+\frac{1}{\lambda^2\tau_2^2}\right)\text{vec}\left(\boldsymbol{I}_d\right)\end{bmatrix}\right\} \\ &= \left(2\pi\tau_1^2\right)^{-d/2}\exp\left(-\frac{\left\|\boldsymbol{y}\right\|^2}{2\tau_1^2}\right)\left(2\pi\lambda^2\tau_2^2\right)^{-d/2}\exp\left\{\begin{bmatrix}\boldsymbol{\psi}\\ \\ \text{vec}\left(\boldsymbol{\psi}\boldsymbol{\psi}^T\right)\end{bmatrix}^T\begin{bmatrix}\boldsymbol{\eta}_1\\ \\ \boldsymbol{\eta}_2\end{bmatrix}\right\},\end{aligned}$$

where

$$\boldsymbol{\eta}_1 \equiv \left(1/\tau_1^2\right)\boldsymbol{y} \text{ and } \boldsymbol{\eta}_2 \equiv -\frac{1}{2}\left(\frac{1}{\tau_1^2}+\frac{1}{\lambda^2\tau_2^2}\right)\text{vec}\left(\boldsymbol{I}_d\right).$$

Then the integral stated in the last step of (3.164) becomes

$$\begin{aligned}\int_{\mathbb{R}^d} \mathfrak{p}\left(\boldsymbol{y}\middle|\boldsymbol{\psi}\right)\mathfrak{p}\left(\boldsymbol{\psi}\middle|\lambda\right)d\boldsymbol{\psi} &= \int_{\mathbb{R}^d}\left(2\pi\tau_1^2\right)^{-d/2}\exp\left(-\frac{\left\|\boldsymbol{y}\right\|^2}{2\tau_1^2}\right)\left(2\pi\lambda^2\tau_2^2\right)^{-d/2} \\ &\times\exp\left\{\begin{bmatrix}\boldsymbol{\psi}\\ \text{vec}\left(\boldsymbol{\psi}\boldsymbol{\psi}^T\right)\end{bmatrix}^T\begin{bmatrix}\left(1/\tau_1^2\right)\boldsymbol{y}\\ -\frac{1}{2}\left(\frac{1}{\tau_1^2}+\frac{1}{\lambda^2\tau_2^2}\right)\text{vec}\left(\boldsymbol{I}_d\right)\end{bmatrix}\right\}d\boldsymbol{\psi} \\ &= \left(2\pi\tau_1^2\right)^{-d/2}\exp\left(-\frac{\left\|\boldsymbol{y}\right\|^2}{2\tau_1^2}\right)\left(2\pi\lambda^2\tau_2^2\right)^{-d/2}\int_{\mathbb{R}^d}\exp\left\{\begin{bmatrix}\boldsymbol{\psi}\\ \text{vec}\left(\boldsymbol{\psi}\boldsymbol{\psi}^T\right)\end{bmatrix}^T\begin{bmatrix}\boldsymbol{\eta}_1\\ \boldsymbol{\eta}_2\end{bmatrix}\right\}d\boldsymbol{\psi}.\end{aligned} \tag{3.165}$$

Note that the term inside the integral as stated in the last step of (3.165) is as follows:

$$\exp\left\{\begin{bmatrix}\boldsymbol{\psi}\\ \text{vec}\left(\boldsymbol{\psi}\boldsymbol{\psi}^T\right)\end{bmatrix}^T\begin{bmatrix}\boldsymbol{\eta}_1\\ \boldsymbol{\eta}_2\end{bmatrix}\right\} = \exp\left\{-\frac{1}{2}\left\{\boldsymbol{\psi}^T\left(\frac{1}{\tau_1^2}+\frac{1}{\lambda^2\tau_2^2}\right)\boldsymbol{I}_d\boldsymbol{\psi}+\boldsymbol{y}^T\frac{\boldsymbol{\psi}}{\tau_1^2}\right\}\right\}.$$

For the integral stated in the last step of (3.165), we use the following result for n-dimensional and functional generalization

$$\int_{\mathbb{R}^d}\exp\left(-\frac{1}{2}\boldsymbol{x}^TA\boldsymbol{x}+b^T\boldsymbol{x}+c\right)d^n\boldsymbol{x} = \sqrt{\det\left(2\pi A^{-1}\right)}\exp\left(\frac{1}{2}b^TA^{-1}b+c\right).$$

In our case, we have

$$\begin{aligned}&A = \left(\frac{1}{\tau_1^2}+\frac{1}{\lambda^2\tau_2^2}\right)\boldsymbol{I}_d,\ \ b=\frac{\boldsymbol{y}}{\tau_1^2}\\ &c=0 \text{ and } \sqrt{\det\left(2\pi A^{-1}\right)} = (2\pi)^{d/2}\left(\frac{1}{\tau_1^2}+\frac{1}{\lambda^2\tau_2^2}\right)^{-d/2}.\end{aligned}$$

Then the integral stated in the last step of (3.165) becomes

$$\int_{\mathbb{R}^d} \exp\left\{ \begin{bmatrix} \boldsymbol{\psi} \\ \text{vec}\left(\boldsymbol{\psi}\boldsymbol{\psi}^T\right) \end{bmatrix}^T \begin{bmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \end{bmatrix} \right\} d\boldsymbol{\psi}$$
$$= (2\pi)^{d/2} \left(\frac{1}{\tau_1^2} + \frac{1}{\lambda^2\tau_2^2}\right)^{-d/2} \exp\left\{ \frac{1}{2} \left(\frac{\boldsymbol{y}}{\tau_1^2}\right)^T \left(\frac{1}{\tau_1^2} + \frac{1}{\lambda^2\tau_2^2}\right)^{-1} \boldsymbol{I}_d \left(\frac{\boldsymbol{y}}{\tau_1^2}\right) \right\}.$$

Therefore, the integral stated in the last step of (3.164) becomes

$$\int_{\mathbb{R}^d} \mathfrak{p}\left(\boldsymbol{y}\middle|\boldsymbol{\psi}\right) \mathfrak{p}\left(\boldsymbol{\psi}\middle|\lambda\right) d\boldsymbol{\psi} = \left(2\pi\tau_1^2\right)^{-d/2} \exp\left(-\frac{\left\|\boldsymbol{y}\right\|^2}{2\tau_1^2}\right) \left(2\pi\lambda^2\tau_2^2\right)^{-d/2}$$
$$\times (2\pi)^{d/2} \left(\frac{1}{\tau_1^2} + \frac{1}{\lambda^2\tau_2^2}\right)^{-d/2} \exp\left\{ \frac{1}{2} \left(\frac{\boldsymbol{y}}{\tau_1^2}\right)^T \left(\frac{1}{\tau_1^2} + \frac{1}{\lambda^2\tau_2^2}\right)^{-1} \boldsymbol{I}_d \left(\frac{\boldsymbol{y}}{\tau_1^2}\right) \right\}$$
$$= \left(2\pi\tau_1^2\right)^{-d/2} \left(1 + \frac{\lambda^2\tau_2^2}{\tau_1^2}\right)^{-d/2} \exp\left\{ -\frac{\left\|\boldsymbol{y}\right\|^2}{2\left(\tau_1^2 + \lambda^2\tau_2^2\right)} \right\}.$$

Hence $\mathfrak{p}\left(\lambda\middle|\boldsymbol{y}\right)$ stated in (3.164) becomes

$$\mathfrak{p}\left(\lambda\middle|\boldsymbol{y}\right) = \mathfrak{p}\left(\lambda\right) \frac{\int_{\mathbb{R}^d} \mathfrak{p}\left(\boldsymbol{y}\middle|\boldsymbol{\psi}\right) \mathfrak{p}\left(\boldsymbol{\psi}\middle|\lambda\right) d\boldsymbol{\psi}}{\mathfrak{p}\left(\boldsymbol{y}\right)}$$
$$= \frac{2I\left(\lambda > 0\right)}{\pi\left(1+\lambda^2\right)\mathfrak{p}\left(\boldsymbol{y}\right)} \left(2\pi\left(\tau_1^2 + \lambda^2\tau_2^2\right)\right)^{-d/2} \exp\left\{ -\frac{\left\|\boldsymbol{y}\right\|^2}{2\left(\tau_1^2 + \lambda^2\tau_2^2\right)} \right\}. \quad (3.166)$$

The posterior density function of $\boldsymbol{\psi}$ is

$$\mathfrak{p}\left(\boldsymbol{\psi}\middle|\boldsymbol{y}\right) = \frac{\mathfrak{p}\left(\boldsymbol{\psi},\boldsymbol{y}\right)}{\mathfrak{p}\left(\boldsymbol{y}\right)} = \frac{\int_0^\infty \mathfrak{p}\left(\boldsymbol{y},\boldsymbol{\psi},\lambda\right) d\lambda}{\mathfrak{p}\left(\boldsymbol{y}\right)} = \frac{\int_0^\infty \mathfrak{p}\left(\boldsymbol{y}\middle|\boldsymbol{\psi}\right) \mathfrak{p}\left(\boldsymbol{\psi}\middle|\lambda\right) \mathfrak{p}\left(\lambda\right) d\lambda}{\mathfrak{p}\left(\boldsymbol{y}\right)}$$
$$= \frac{\mathfrak{p}\left(\boldsymbol{y}\middle|\boldsymbol{\psi}\right) \int_0^\infty \mathfrak{p}\left(\boldsymbol{\psi}\middle|\lambda\right) \mathfrak{p}\left(\lambda\right) d\lambda}{\mathfrak{p}\left(\boldsymbol{y}\right)}.$$

Then

$$\mathfrak{p}\left(\boldsymbol{y}\right) \mathfrak{p}\left(\boldsymbol{\psi}\middle|\boldsymbol{y}\right) = \mathfrak{p}\left(\boldsymbol{y}\middle|\boldsymbol{\psi}\right) \int_0^\infty \mathfrak{p}\left(\boldsymbol{\psi}\middle|\lambda\right) \mathfrak{p}\left(\lambda\right) d\lambda. \quad (3.167)$$

Let us look into the expression for $\mathfrak{p}\left(\boldsymbol{y}\middle|\boldsymbol{\psi}\right)$ and $\mathfrak{p}\left(\boldsymbol{\psi}\middle|\lambda\right)$ stated in (3.167). Introduce the following notation after equation (4) of Wand and Jones (1993):

$$\phi_{\boldsymbol{\Sigma}}\left(\boldsymbol{x}\right)=\left(2\pi\right)^{-d/2}\left|\boldsymbol{\Sigma}\right|^{-1/2}\exp\left(-\frac{1}{2}\boldsymbol{x}^{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right)\text{ is the }N\left(\boldsymbol{0},\boldsymbol{\Sigma}\right)\text{ density.}$$

In our model, we have

$$\mathfrak{p}\left(\boldsymbol{y}\middle|\boldsymbol{\psi}\right)=\phi_{\tau_1^2\boldsymbol{I}_d}\left(\boldsymbol{\psi}-\boldsymbol{y}\right)\quad\text{and}\quad\mathfrak{p}\left(\boldsymbol{\psi}\middle|\lambda\right)=\phi_{\lambda^2\tau_2^2\boldsymbol{I}_d}\left(\boldsymbol{\psi}-\boldsymbol{0}\right).$$

Then (3.167) becomes

$$\begin{aligned}\mathfrak{p}\left(\boldsymbol{y}\right)\mathfrak{p}\left(\boldsymbol{\psi}\middle|\boldsymbol{y}\right)&=\mathfrak{p}\left(\boldsymbol{y}\middle|\boldsymbol{\psi}\right)\int_0^\infty\mathfrak{p}\left(\boldsymbol{\psi}\middle|\lambda\right)\mathfrak{p}\left(\lambda\right)d\lambda\\&=\frac{2}{\pi}\int_0^\infty\phi_{\tau_1^2\boldsymbol{I}_d}\left(\boldsymbol{\psi}-\boldsymbol{y}\right)\phi_{\lambda^2\tau_2^2\boldsymbol{I}_d}\left(\boldsymbol{\psi}-\boldsymbol{0}\right)\frac{1}{1+\lambda^2}d\lambda.\end{aligned}\tag{3.168}$$

Let us look into the term $\phi_{\tau_1^2\boldsymbol{I}_d}\left(\boldsymbol{\psi}-\boldsymbol{y}\right)\phi_{\lambda^2\tau_2^2\boldsymbol{I}_d}\left(\boldsymbol{\psi}-\boldsymbol{0}\right)$ inside the integral in the last step of (3.168). From equation (A.1) of Wand and Jones (1993), we have

$$\begin{aligned}\phi_{\tau_1^2\boldsymbol{I}_d}\Big(\boldsymbol{\psi}-\boldsymbol{y}\Big)\phi_{\lambda^2\tau_2^2\boldsymbol{I}_d}\Big(\boldsymbol{\psi}-\boldsymbol{0}\Big)&=\phi_{\left(\tau_1^2+\lambda^2\tau_2^2\right)\boldsymbol{I}_d}\Big(\boldsymbol{y}\Big)\phi_{\left(\frac{\tau_1^2\lambda^2\tau_2^2}{\tau_1^2+\lambda^2\tau_2^2}\right)\boldsymbol{I}_d}\left(\boldsymbol{\psi}-\frac{\lambda^2\tau_2^2}{\tau_1^2+\lambda^2\tau_2^2}\boldsymbol{y}\right)\\&=\left(2\pi\left(\tau_1^2+\lambda^2\tau_2^2\right)\right)^{-d/2}\exp\left(-\frac{\left\|\boldsymbol{y}\right\|^2}{2\left(\tau_1^2+\lambda^2\tau_2^2\right)}\right)\phi_{\left(\frac{\tau_1^2\lambda^2\tau_2^2}{\tau_1^2+\lambda^2\tau_2^2}\right)\boldsymbol{I}_d}\left(\boldsymbol{\psi}-\frac{\lambda^2\tau_2^2}{\tau_1^2+\lambda^2\tau_2^2}\boldsymbol{y}\right).\end{aligned}$$

Then $\mathfrak{p}\left(\boldsymbol{\psi}\middle|\boldsymbol{y}\right)$ stated on the left-hand side in the first step of (3.168) becomes

$$\begin{aligned}\mathfrak{p}\left(\boldsymbol{\psi}\middle|\boldsymbol{y}\right)=\frac{2}{\pi}\left(\frac{1}{\mathfrak{p}\left(\boldsymbol{y}\right)}\right)\int_0^\infty&\left(2\pi\left(\tau_1^2+\lambda^2\tau_2^2\right)\right)^{-d/2}\exp\left(-\frac{\left\|\boldsymbol{y}\right\|^2}{2\left(\tau_1^2+\lambda^2\tau_2^2\right)}\right)\\&\times\phi_{\left(\frac{\tau_1^2\lambda^2\tau_2^2}{\tau_1^2+\lambda^2\tau_2^2}\right)\boldsymbol{I}_d}\left(\boldsymbol{\psi}-\frac{\lambda^2\tau_2^2}{\tau_1^2+\lambda^2\tau_2^2}\boldsymbol{y}\right)\frac{1}{1+\lambda^2}d\lambda.\end{aligned}$$

Recalling that $\mathfrak{p}\left(\lambda\middle|\boldsymbol{y}\right)$ stated in (3.166) as follows:

$$\mathfrak{p}\left(\lambda\middle|\boldsymbol{y}\right)=\frac{2I\left(\lambda>0\right)\mathfrak{p}\left(\boldsymbol{y}\right)}{\pi\left(1+\lambda^{2}\right)}\Big(2\pi\left(\tau_{1}^{2}+\lambda^{2}\tau_{2}^{2}\right)\Big)^{-d/2}\exp\left\{-\frac{\left\|\boldsymbol{y}\right\|^{2}}{2\left(\tau_{1}^{2}+\lambda^{2}\tau_{2}^{2}\right)}\right\},$$

we then have

$$\mathfrak{p}\left(\boldsymbol{\psi}\middle|\boldsymbol{y}\right)=\int_{0}^{\infty}\mathfrak{p}\Big(\lambda\Big|\boldsymbol{y}\Big)\phi_{\left(\frac{\tau_{1}^{2}\lambda^{2}\tau_{2}^{2}}{\tau_{1}^{2}+\lambda^{2}\tau_{2}^{2}}\right)\boldsymbol{I}_{d}}\left(\boldsymbol{\psi}-\frac{\lambda^{2}\tau_{2}^{2}}{\tau_{1}^{2}+\lambda^{2}\tau_{2}^{2}}\boldsymbol{y}\right)d\lambda.$$

Therefore

$$\begin{aligned}E\left(\boldsymbol{\psi}\middle|\boldsymbol{y}\right)&=\int_{\mathbb{R}^{d}}\boldsymbol{\psi}\mathfrak{p}\left(\boldsymbol{\psi}\middle|\boldsymbol{y}\right)d\boldsymbol{\psi}\\&=\int_{\mathbb{R}^{d}}\boldsymbol{\psi}\left\{\int_{0}^{\infty}\mathfrak{p}\Big(\lambda\Big|\boldsymbol{y}\Big)\phi_{\left(\frac{\tau_{1}^{2}\lambda^{2}\tau_{2}^{2}}{\tau_{1}^{2}+\lambda^{2}\tau_{2}^{2}}\right)\boldsymbol{I}_{d}}\left(\boldsymbol{\psi}-\frac{\lambda^{2}\tau_{2}^{2}}{\tau_{1}^{2}+\lambda^{2}\tau_{2}^{2}}\boldsymbol{y}\right)d\lambda\right\}d\boldsymbol{\psi}\\&=\int_{0}^{\infty}\mathfrak{p}\Big(\lambda\Big|\boldsymbol{y}\Big)\left\{\int_{\mathbb{R}^{d}}\boldsymbol{\psi}\phi_{\left(\frac{\tau_{1}^{2}\lambda^{2}\tau_{2}^{2}}{\tau_{1}^{2}+\lambda^{2}\tau_{2}^{2}}\right)\boldsymbol{I}_{d}}\left(\boldsymbol{\psi}-\frac{\lambda^{2}\tau_{2}^{2}}{\tau_{1}^{2}+\lambda^{2}\tau_{2}^{2}}\boldsymbol{y}\right)d\boldsymbol{\psi}\right\}d\lambda.\end{aligned}\tag{3.169}$$

The inner integral stated in the last step of (3.169) is the mean of multivariate normal

$$N\left(\left(\lambda^{2}\tau_{2}^{2}/\left(\tau_{1}^{2}+\lambda^{2}\tau_{2}^{2}\right)\right)\boldsymbol{y},\left(\tau_{1}^{2}\lambda^{2}\tau_{2}^{2}/\left(\tau_{1}^{2}+\lambda^{2}\tau_{2}^{2}\right)\right)\boldsymbol{I}_{d}\right).$$

Hence, the posterior mean of $\boldsymbol{\psi}$ is

$$\begin{aligned}E\left(\boldsymbol{\psi}\middle|\boldsymbol{y}\right)&=\int_{0}^{\infty}\mathfrak{p}\Big(\lambda\Big|\boldsymbol{y}\Big)\left\{\frac{\lambda^{2}\tau_{2}^{2}}{\tau_{1}^{2}+\lambda^{2}\tau_{2}^{2}}\boldsymbol{y}\right\}d\lambda=\left(\int_{0}^{\infty}\mathfrak{p}\Big(\lambda\Big|\boldsymbol{y}\Big)\left\{\frac{\lambda^{2}\tau_{2}^{2}}{\tau_{1}^{2}+\lambda^{2}\tau_{2}^{2}}\right\}d\lambda\right)\boldsymbol{y}\\&=E\left(\left.\frac{\lambda^{2}\tau_{2}^{2}}{\tau_{1}^{2}+\lambda^{2}\tau_{2}^{2}}\right|\boldsymbol{y}\right)\boldsymbol{y}.\end{aligned}$$

3.F Derivations of Full Conditional Distributions

3.F.1 Full Conditional Distributions of β_0 for Gaussian Responses

First note that

$$\mathfrak{p}(\beta_0 \big| \text{rest}) \propto \mathfrak{p}(\boldsymbol{y} \big| \beta_0, \boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2) \mathfrak{p}(\beta_0). \tag{3.170}$$

Then note that

$$\boldsymbol{y} \big| \beta_0, \boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2 \sim N\left(\mathbf{1}_n \beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j \boldsymbol{u}_j, \sigma_\varepsilon^2 \boldsymbol{I}_n \right) \quad \text{and} \quad \beta_0 \sim N(0, \sigma_{\beta_0}^2).$$

For $\mathfrak{p}(\boldsymbol{y} \big| \beta_0, \boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2)$ stated in (3.170), we have

$$\begin{aligned} \mathfrak{p}(\boldsymbol{y} \big| \beta_0, \boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2) &\propto \exp\left\{ -\frac{1}{2\sigma_\varepsilon^2} \left\| \boldsymbol{y} - \mathbf{1}_n \beta_0 - \boldsymbol{X}\boldsymbol{\beta} - \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j \boldsymbol{u}_j \right\|^2 \right\} \\ &\propto \exp\left\{ -\frac{1}{2\sigma_\varepsilon^2} \left\| \boldsymbol{r}_{\beta_0} - \mathbf{1}_n \beta_0 \right\|^2 \right\} \propto \exp\left\{ -\frac{1}{2\sigma_\varepsilon^2} \left(-2\beta_0^T \mathbf{1}_n^T \boldsymbol{r}_{\beta_0} + n\beta_0^2 \right) \right\}, \end{aligned}$$

if we define $\boldsymbol{r}_{\beta_0} \equiv \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j \boldsymbol{u}_j$. Therefore

$$\mathfrak{p}(\boldsymbol{y} \big| \beta_0, \boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2) \propto \exp\left\{ \begin{bmatrix} \beta_0 \\ \beta_0^2 \end{bmatrix}^T \begin{bmatrix} \mathbf{1}_n^T \left(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}} - \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j \boldsymbol{u}_j \right) \\ -n/2 \end{bmatrix} (1/\sigma_\varepsilon^2) \right\}.$$

For $\mathfrak{p}(\beta_0)$ stated in (3.170), we have

$$\mathfrak{p}(\beta_0) \propto \exp\left\{ \begin{bmatrix} \beta_0 \\ \beta_0^2 \end{bmatrix}^T \begin{bmatrix} 0 \\ -1/2 \end{bmatrix} (1/\sigma_{\beta_0}^2) \right\}.$$

Then

$$\mathfrak{p}(\beta_0|\text{rest}) \propto \exp\left\{ \begin{bmatrix} \beta_0 \\ \beta_0^2 \end{bmatrix}^T \begin{bmatrix} \left(\frac{1}{\sigma_\varepsilon^2}\right) \mathbf{1}_n^T \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j \boldsymbol{u}_j \right) \\ -\frac{1}{2}\left(\frac{n}{\sigma_\varepsilon^2} + \frac{1}{\sigma_{\beta_0}^2}\right) \end{bmatrix} \right\}. \tag{3.171}$$

Therefore $\mathfrak{p}(\beta_0|\text{rest})$ has a Normal distribution with natural parameter vector

$$\begin{bmatrix} \mathbf{1}_n^T \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j \boldsymbol{u}_j \right) \Big/ \sigma_\varepsilon^2 \\ \frac{-n}{2\sigma_\varepsilon^2} - \frac{1}{2\sigma_{\beta_0}^2} \end{bmatrix}.$$

The variance is

$$\frac{1}{(n/\sigma_\varepsilon^2) + (1/\sigma_{\beta_0}^2)}.$$

The mean is

$$\frac{\frac{1}{\sigma_\varepsilon^2}\mathbf{1}_n^T \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j \boldsymbol{u}_j \right)}{(n/\sigma_\varepsilon^2) + (1/\sigma_{\beta_0}^2)}.$$

Therefore

$$\beta_0|\text{rest} \sim N\left(\frac{\frac{1}{\sigma_\varepsilon^2}\mathbf{1}_n^T \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j \boldsymbol{u}_j \right)}{(n/\sigma_\varepsilon^2) + (1/\sigma_{\beta_0}^2)}, \frac{1}{(n/\sigma_\varepsilon^2) + (1/\sigma_{\beta_0}^2)} \right).$$

If we define

$$\omega_1 \equiv \mathbf{1}_n^T \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j \boldsymbol{u}_j \right),$$

recalling that (2.7) of Section 2.2.2 to be $\mathbf{1}_n^T \boldsymbol{X} = \mathbf{0}$ and (2.8) of Section 2.3.3 to be $\boldsymbol{Z}_j^T \mathbf{1}_n = \mathbf{0}_{K_j}, 1 \le j \le d_\bullet$, we have $\omega_1 = \mathbf{1}_n^T \boldsymbol{y}$. If define $\omega_2 \equiv (n/\sigma_\varepsilon^2) + (1/\sigma_{\beta_0}^2)$, then $\beta_0|\text{rest} \sim N\left(\frac{\omega_1}{(\sigma_\varepsilon^2)\,\omega_2}, \frac{1}{\omega_2}\right)$ is as given in Algorithm 4 in Section 3.6.3.

3.F.2 Full Conditional Distributions of $\boldsymbol{\beta}$ for Gaussian responses

First note that

$$\mathfrak{p}(\boldsymbol{\beta}\big|\text{rest}) \propto \mathfrak{p}(\boldsymbol{y}\big|\beta_0, \boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2)\mathfrak{p}(\boldsymbol{\beta}\big|\sigma_\beta^2, \boldsymbol{f}_{2\beta}). \tag{3.172}$$

Then note that

$$\boldsymbol{y}\big|\beta_0, \boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2 \sim N\left(\mathbf{1}_n\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j\boldsymbol{u}_j, \sigma_\varepsilon^2\boldsymbol{I}_n\right)$$

and

$$\beta_j\big|\sigma_\beta^2, f_{2\beta j} \sim N(0, \sigma_\beta^2/f_{2\beta j}),\ \ 1 \le j \le d_\circ + d_\bullet.$$

For $\mathfrak{p}(\boldsymbol{y}\big|\beta_0, \boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2)$, stated in (3.172), we have

$$\begin{aligned}
\mathfrak{p}(\boldsymbol{y}\big|\beta_0, \boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2) &\propto \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\left\|\boldsymbol{y} - \mathbf{1}_n\beta_0 - \boldsymbol{X}\boldsymbol{\beta} - \sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j\boldsymbol{u}_j\right\|^2\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\left\|\boldsymbol{r}_{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}\right\|^2\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\left(-2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{r}_{\boldsymbol{\beta}} + \left(\text{vec}(\widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}^T)\right)^T\text{vec}\left(\boldsymbol{X}^T\boldsymbol{X}\right)\right)\right\},
\end{aligned}$$

if we define $\boldsymbol{r}_{\boldsymbol{\beta}} \equiv \boldsymbol{y} - \mathbf{1}_n\beta_0 - \sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j\boldsymbol{u}_j$. Therefore

$$\mathfrak{p}(\boldsymbol{y}\big|\beta_0, \boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2) \propto \exp\left\{\begin{bmatrix}\boldsymbol{\beta} \\ \\ \text{vec}\left(\boldsymbol{\beta}\boldsymbol{\beta}^T\right)\end{bmatrix}^T\begin{bmatrix}\boldsymbol{X}^T\left(\boldsymbol{y} - \mathbf{1}_n\beta_0 - \sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j\boldsymbol{u}_j\right) \\ -\frac{1}{2}\text{vec}\left(\boldsymbol{X}^T\boldsymbol{X}\right)\end{bmatrix}\left(1/\sigma_\varepsilon^2\right)\right\}.$$

For $\mathfrak{p}(\boldsymbol{\beta}\big|\sigma_\beta^2, \boldsymbol{f}_{2\beta})$ stated in (3.172), we have

$$\begin{aligned}\mathfrak{p}(\boldsymbol{\beta}\big|\sigma_\beta^2, \boldsymbol{f}_{2\beta}) &\propto \exp\left(-\frac{\boldsymbol{\beta}^T \text{diag}\left(\boldsymbol{f}_{2\beta}\right)\boldsymbol{\beta}}{2\sigma_\beta^2}\right) \\ &\propto \exp\left\{\begin{bmatrix}\boldsymbol{\beta} \\ \text{vec}\left(\boldsymbol{\beta}\boldsymbol{\beta}^T\right)\end{bmatrix}^T \begin{bmatrix}\mathbf{0} \\ -\frac{1}{2}\text{vec}(\text{diag}\left(\boldsymbol{f}_{2\beta}\right)\end{bmatrix}\left(1/\sigma_\beta^2\right)\right\}.\end{aligned}$$

Therefore $\mathfrak{p}(\boldsymbol{\beta}\big|\text{rest})$ stated in (3.172) becomes

$$\mathfrak{p}(\boldsymbol{\beta}\big|\text{rest}) \propto \exp\left\{\begin{bmatrix}\boldsymbol{\beta} \\ \\ \text{vec}\left(\boldsymbol{\beta}\boldsymbol{\beta}^T\right)\end{bmatrix}^T \begin{bmatrix}\dfrac{\boldsymbol{X}^T}{\sigma_\varepsilon^2}\left(\boldsymbol{y} - \mathbf{1}_n\beta_0 - \sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j\boldsymbol{u}_j\right) \\ -\frac{1}{2}\text{vec}\left(\dfrac{\boldsymbol{X}^T\boldsymbol{X}}{\sigma_\varepsilon^2} + \dfrac{\text{diag}\left(\boldsymbol{f}_{2\beta}\right)}{\sigma_\beta^2}\right)\end{bmatrix}\right\}.$$

Then the first component of the natural parameter vector of $\mathfrak{p}(\boldsymbol{\beta}|\text{rest})$ is

$$\boldsymbol{X}^T\left(\boldsymbol{y} - \mathbf{1}_n\beta_0 - \sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j\boldsymbol{u}_j\right)\left(1/\sigma_\varepsilon^2\right).$$

The inverse vec of the second component of the natural parameter vector is

$$-\frac{1}{2}\left(\frac{1}{\sigma_\varepsilon^2}\boldsymbol{X}^T\boldsymbol{X} + \frac{1}{\sigma_\beta^2}\text{diag}\left(\boldsymbol{f}_{2\beta}\right)\right).$$

It follows that the covariance matrix of $\boldsymbol{\beta}|\text{rest}$ is

$$\left(\frac{1}{\sigma_\varepsilon^2}\boldsymbol{X}^T\boldsymbol{X} + \frac{1}{\sigma_\beta^2}\text{diag}\left(\boldsymbol{f}_{2\beta}\right)\right)^{-1}$$

and the mean vector is

$$\left(\frac{1}{\sigma_\varepsilon^2}\boldsymbol{X}^T\boldsymbol{X} + \frac{1}{\sigma_\beta^2}\text{diag}(\boldsymbol{f}_{2\beta})\right)^{-1}\frac{1}{\sigma_\varepsilon^2}\boldsymbol{X}^T\left(\boldsymbol{y} - \mathbf{1}_n\beta_0 - \sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j\boldsymbol{u}_j\right).$$

Therefore

$$\boldsymbol{\beta}|\text{rest} \sim N\Bigg(\left(\frac{1}{\sigma_\varepsilon^2}\boldsymbol{X}^T\boldsymbol{X} + \frac{1}{\sigma_\beta^2}\text{diag}\left(\boldsymbol{f}_{2\beta}\right)\right)^{-1} \frac{1}{\sigma_\varepsilon^2}\boldsymbol{X}^T\left(\boldsymbol{y} - \mathbf{1}_n\beta_0 - \sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j\boldsymbol{u}_j\right),$$

$$\left(\frac{1}{\sigma_\varepsilon^2}\boldsymbol{X}^T\boldsymbol{X} + \frac{1}{\sigma_\beta^2}\text{diag}\left(\boldsymbol{f}_{2\beta}\right)\right)^{-1}\Bigg).$$

Recalling that (2.7) in Section 3.4.2 to be $\boldsymbol{X}^T\mathbf{1}_n = \mathbf{0}$, then the full conditional distribution of $\boldsymbol{\beta}$ is then obtained to be (3.28) in Section 3.6.3:

$$\boldsymbol{\beta}|\text{rest} \sim N\Bigg(\left(\frac{1}{\sigma_\varepsilon^2}\boldsymbol{X}^T\boldsymbol{X} + \frac{1}{\sigma_\beta^2}\text{diag}\left(\boldsymbol{f}_{2\beta}\right)\right)^{-1} \frac{1}{\sigma_\varepsilon^2}\left(\boldsymbol{X}^T\boldsymbol{y} - \sum_{j=1}^{d_\bullet}\left(\boldsymbol{X}^T\boldsymbol{Z}_j\right)\boldsymbol{u}_j\right),$$

$$\left(\frac{1}{\sigma_\varepsilon^2}\boldsymbol{X}^T\boldsymbol{X} + \frac{1}{\sigma_\beta^2}\text{diag}\left(\boldsymbol{f}_{2\beta}\right)\right)^{-1}\Bigg).$$

If we define

$$\boldsymbol{\omega}_3 \equiv \boldsymbol{X}^T\boldsymbol{y} - \sum_{j=1}^{d_\bullet}\left(\boldsymbol{X}^T\boldsymbol{Z}_j\right)\boldsymbol{u}_j,$$

and define

$$\boldsymbol{\Omega} \equiv \frac{1}{\sigma_\varepsilon^2}\boldsymbol{X}^T\boldsymbol{X} + \frac{1}{\sigma_\beta^2}\text{diag}(\boldsymbol{f}_{2\beta}),$$

then decompose $\boldsymbol{\Omega} = \boldsymbol{U}_{\boldsymbol{\Omega}}\text{diag}\left(\boldsymbol{d}_{\boldsymbol{\Omega}}\right)\boldsymbol{U}_{\boldsymbol{\Omega}}^T$ where $\boldsymbol{U}_{\boldsymbol{\Omega}}^T\boldsymbol{U}_{\boldsymbol{\Omega}} = \boldsymbol{I}$. If

$$\boldsymbol{z} \sim N\left(\mathbf{0}, \boldsymbol{I}\right)\left(\left(d_\circ + d_\bullet\right) \times 1\right);$$

then

$$\boldsymbol{U}_{\boldsymbol{\Omega}}\left(\frac{\boldsymbol{U}_{\boldsymbol{\Omega}}^T\boldsymbol{z}}{\sqrt{\boldsymbol{d}_{\boldsymbol{\Omega}}}} + \frac{\boldsymbol{U}_{\boldsymbol{\Omega}}^T\boldsymbol{\omega}_3}{\boldsymbol{d}_{\boldsymbol{\Omega}}\left(\sigma_\varepsilon^2\right)}\right) \text{ has the same distribution as } \boldsymbol{\beta}|\text{rest},$$

as given in Algorithm 4 in Section 3.6.3.

3.F.3 Full Conditional Distributions of $\boldsymbol{u}_j$ for Gaussian responses

First note that, for $1 \leq j \leq d_\bullet$,

$$\mathfrak{p}(\boldsymbol{u}_j|\text{rest}) \propto \mathfrak{p}(\boldsymbol{y}|\beta_0, \boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2)\mathfrak{p}(\boldsymbol{u}_j|\sigma_{uj}^2, f_{2uj}). \tag{3.173}$$

Then note that

$$\boldsymbol{y}\big|\beta_0,\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2 \sim N\left(\mathbf{1}_n\beta_0+\boldsymbol{X}\boldsymbol{\beta}+\sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j\boldsymbol{u}_j,\sigma_\varepsilon^2\boldsymbol{I}_n\right)$$

and

$$\boldsymbol{u}_j\big|\sigma_{uj}^2,f_{2uj} \sim N\Big(0,(\sigma_{uj}^2/f_{2uj})\boldsymbol{I}\Big),\ \ 1\le j\le d_\bullet.$$

For $\mathfrak{p}(\boldsymbol{y}\big|\beta_0,\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2)$ stated in (3.173), we have

$$\begin{aligned}
&\mathfrak{p}(\boldsymbol{y}\big|\beta_0,\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2)\\
&\quad\propto \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\left\|\boldsymbol{y}-\mathbf{1}_n\beta_0-\boldsymbol{X}\boldsymbol{\beta}-\sum_{j'\neq j}^{d_\bullet}\boldsymbol{Z}_{j'}\boldsymbol{u}_{j'}-\boldsymbol{Z}_j\boldsymbol{u}_j\right\|^2\right\}\\
&\quad\propto \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\left\|\boldsymbol{r}_{\boldsymbol{u}_j}-\boldsymbol{Z}_j\boldsymbol{u}_j\right\|^2\right\}.\\
&\quad\propto \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\left(-2\boldsymbol{u}_j^T\boldsymbol{Z}_j^T\boldsymbol{r}_{\boldsymbol{u}_j}+\operatorname{vec}\left(\boldsymbol{u}_j\boldsymbol{u}_j^T\right)\operatorname{vec}\left(\boldsymbol{Z}_j^T\boldsymbol{Z}_j\right)\right)\right\},
\end{aligned}$$

if we define $\boldsymbol{r}_{\boldsymbol{u}_j}\equiv\boldsymbol{y}-\mathbf{1}_n\beta_0-\boldsymbol{X}\boldsymbol{\beta}-\sum_{j'\neq j}^{d_\bullet}\boldsymbol{Z}_{j'}\boldsymbol{u}_{j'}$. Therefore

$$\begin{aligned}
\mathfrak{p}(\boldsymbol{y}\big|\beta_0,\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2)\propto\exp\Bigg\{&\begin{bmatrix}\boldsymbol{u}_j\\ \\ \operatorname{vec}(\boldsymbol{u}_j\boldsymbol{u}_j^T)\end{bmatrix}^T\\
&\times\begin{bmatrix}\boldsymbol{Z}_j^T\left(\boldsymbol{y}-\mathbf{1}_n\beta_0-\boldsymbol{X}\boldsymbol{\beta}-\sum_{j'\neq j}^{d_\bullet}\boldsymbol{Z}_{j'}\boldsymbol{u}_{j'}\right)\\ \\ -\frac{1}{2}\operatorname{vec}\left(\boldsymbol{Z}_j^T\boldsymbol{Z}_j\right)\end{bmatrix}\left(1/\sigma_\varepsilon^2\right)\Bigg\}.
\end{aligned}$$

For $\mathfrak{p}(\boldsymbol{u}_j|\sigma^2_{uj}, f_{2uj})$ stated in (3.173), we have

$$\mathfrak{p}(\boldsymbol{u}_j|\sigma^2_{uj}, f_{2uj}) \propto \exp\left\{-\frac{1}{2\sigma^2_{uj}}\text{vec}\left(\boldsymbol{u}_j\boldsymbol{u}_j^T\right)\text{vec}\left(f_{2uj}\boldsymbol{I}_{K_j}\right)\right\}$$

$$\propto \exp\left\{\begin{bmatrix}\boldsymbol{u}_j\\ \\ \text{vec}(\boldsymbol{u}_j\boldsymbol{u}_j^T)\end{bmatrix}^T\begin{bmatrix}\boldsymbol{0}\\ \\ -\frac{1}{2}f_{2uj}\text{vec}\left(\boldsymbol{I}_{K_j}\right)\end{bmatrix}\left(1/\sigma^2_{uj}\right)\right\}.$$

Then

$$\mathfrak{p}(\boldsymbol{u}_j|\text{rest}) \sim \exp\left\{\begin{bmatrix}\boldsymbol{u}_j\\ \\ \text{vec}\left(\boldsymbol{u}_j\widetilde{\boldsymbol{u}}_j^T\right)\end{bmatrix}^T\begin{bmatrix}\frac{\boldsymbol{Z}_j^T}{\sigma^2_\varepsilon}\left(\boldsymbol{y}-\boldsymbol{1}_n\beta_0-\boldsymbol{X}\boldsymbol{\beta}-\sum_{j'\neq j}^{d\bullet}\boldsymbol{Z}_{j'}\boldsymbol{u}_{j'}\right)\\ \\ -\frac{1}{2}\text{vec}\left(\frac{\boldsymbol{Z}_j^T\boldsymbol{Z}_j}{\sigma^2_\varepsilon}+\frac{f_{2uj}}{\sigma^2_{uj}}\boldsymbol{I}_{K_j}\right)\end{bmatrix}\right\}. \tag{3.174}$$

Therefore, the first component of the natural parameter vector of $\mathfrak{p}(\boldsymbol{u}_j|\text{rest})$ is

$$\boldsymbol{Z}_j^T\left(\boldsymbol{y}-\boldsymbol{1}_n\beta_0-\boldsymbol{X}\boldsymbol{\beta}-\sum_{j'\neq j}^{d\bullet}\boldsymbol{Z}_{j'}\boldsymbol{u}_{j'}\right)\left(1/\sigma^2_\varepsilon\right).$$

The inverse vec of the second component of the natural parameter vector is

$$-\frac{1}{2}\left(\frac{1}{\sigma^2_\varepsilon}\boldsymbol{Z}_j^T\boldsymbol{Z}_j+\frac{f_{2uj}}{\sigma^2_{uj}}\boldsymbol{I}_{K_j}\right).$$

It follows that the covariance matrix of $\boldsymbol{u}_j|$rest is

$$\left(\frac{1}{\sigma^2_\varepsilon}\boldsymbol{Z}_j^T\boldsymbol{Z}_j+\frac{f_{2uj}}{\sigma^2_{uj}}\boldsymbol{I}_{K_j}\right)^{-1}$$

and the mean vector is

$$\left(\frac{1}{\sigma^2_\varepsilon}\boldsymbol{Z}_j^T\boldsymbol{Z}_j+\frac{f_{2uj}}{\sigma^2_{uj}}\boldsymbol{I}_{K_j}\right)^{-1}\left(\frac{1}{\sigma^2_\varepsilon}\right)\boldsymbol{Z}_j^T\left(\boldsymbol{y}-\boldsymbol{1}_n\beta_0-\boldsymbol{X}\boldsymbol{\beta}-\sum_{j'\neq j}^{d\bullet}\boldsymbol{Z}_{j'}\boldsymbol{u}_{j'}\right).$$

Next, we simplify $\left(\frac{1}{\sigma_\varepsilon^2}\boldsymbol{Z}_j^T\boldsymbol{Z}_j + \frac{f_{2uj}}{\sigma_{uj}^2}\boldsymbol{I}_{K_j}\right)^{-1}$. Note that $\boldsymbol{Z}_j^T\boldsymbol{Z}_j$ is a diagonal matrix from (2.8) in Section 3.4.3, if we define $\boldsymbol{Z}_j^T\boldsymbol{Z}_j \equiv \text{diag}\left(\boldsymbol{\omega}_{\boldsymbol{Z}_j}\right),\ 1 \le j \le d_\bullet$, this implies that

$$\frac{1}{\sigma_\varepsilon^2}\boldsymbol{Z}_j^T\boldsymbol{Z}_j + \frac{f_{2uj}}{\sigma_{uj}^2}\boldsymbol{I}_{K_j} = \text{diag}\left(\frac{\boldsymbol{\omega}_{\boldsymbol{Z}_j}}{\sigma_\varepsilon^2} + \frac{f_{2uj}\mathbf{1}}{\sigma_{uj}^2}\right),$$

and therefore

$$\left(\frac{1}{\sigma_\varepsilon^2}\boldsymbol{Z}_j^T\boldsymbol{Z}_j + \frac{f_{2uj}}{\sigma_{uj}^2}\boldsymbol{I}_{K_j}\right)^{-1} = \text{diag}\left(\frac{\mathbf{1}}{\left\{\boldsymbol{\omega}_{\boldsymbol{Z}_j}\Big/\sigma_\varepsilon^2\right\} + \left(f_{2uj}\mathbf{1}\right)\Big/\sigma_{uj}^2}\right).$$

In summary, for each $1 \le j \le d_\bullet$, the full conditional distribution of $\boldsymbol{u}_j$ is

$$\boldsymbol{u}_j\big|\text{rest} \sim N\Bigg(\frac{\frac{1}{\sigma_\varepsilon^2}\boldsymbol{Z}_j^T\left(\boldsymbol{y} - \mathbf{1}_n\beta_0 - \boldsymbol{X}\boldsymbol{\beta} - \sum_{j'\neq j}^{d_\bullet}\boldsymbol{Z}_{j'}\boldsymbol{u}_{j'}\right)}{\left\{\boldsymbol{\omega}_{\boldsymbol{Z}_j}/\sigma_\varepsilon^2\right\} + \left(f_{2uj}\mathbf{1}\right)/\sigma_{uj}^2},$$

$$\text{diag}\left(\frac{\mathbf{1}}{\left\{\boldsymbol{\omega}_{\boldsymbol{Z}_j}/\sigma_\varepsilon^2\right\} + \left(f_{2uj}\mathbf{1}\right)/\sigma_{uj}^2}\right)\Bigg).$$

Recalling that (2.8) of Section 3.4.3 to be $\boldsymbol{Z}_j^T\mathbf{1}_n = \mathbf{0}$, we then have

$$\boldsymbol{u}_j\big|\text{rest} \sim N\Bigg(\text{diag}\left(\frac{\boldsymbol{\omega}_{\boldsymbol{Z}_j}}{\sigma_\varepsilon^2} + \frac{f_{2uj}\mathbf{1}}{\sigma_{uj}^2}\right)^{-1}\left(\frac{1}{\sigma_\varepsilon^2}\right)\Big[\left(\boldsymbol{Z}_j^T\boldsymbol{y}\right) - \left(\boldsymbol{Z}_j^T\boldsymbol{X}\right)\boldsymbol{\beta}$$

$$-\sum_{j'\neq j}^{d_\bullet}\left(\boldsymbol{Z}_j^T\boldsymbol{Z}_{j'}\right)\boldsymbol{u}_{j'}\Big], \text{diag}\left(\frac{\boldsymbol{\omega}_{\boldsymbol{Z}_j}}{\sigma_\varepsilon^2} + \frac{f_{2uj}\mathbf{1}}{\sigma_{uj}^2}\right)^{-1}\Bigg). \tag{3.175}$$

In (3.175), we define

$$\boldsymbol{\omega}_6 \equiv \boldsymbol{Z}_j^T\boldsymbol{y} - \left(\boldsymbol{Z}_j^T\boldsymbol{X}\right)\boldsymbol{\beta} - \sum_{j'\neq j}^{d_\bullet}\left(\boldsymbol{Z}_j^T\boldsymbol{Z}_{j'}\right)\boldsymbol{u}_{j'}$$

and

$$\boldsymbol{\omega}_7 \equiv \frac{\boldsymbol{\omega}_{\boldsymbol{Z}_j}}{\sigma_\varepsilon^2} + \frac{f_{2uj}\mathbf{1}}{\sigma_{uj}^2},$$

and generate the $(K_j \times 1)$ Multivariate Normal random vector $\boldsymbol{z} \sim N(\boldsymbol{0}, \boldsymbol{I}_{K_j})$. Then

$$\frac{\boldsymbol{z}}{\sqrt{\boldsymbol{\omega}_7}} + \frac{\boldsymbol{\omega}_6}{\boldsymbol{\omega}_7 \sigma_\varepsilon^2} \tag{3.176}$$

has the same distribution as $\boldsymbol{u}_j|\text{rest}$, as given in Algorithm 4 in Section 3.6.3, for $1 \leq j \leq d_\bullet$.

3.F.4 Full Conditional Distribution of σ_ε^2 for Gaussian responses

First note that

$$\mathfrak{p}\left(\sigma_\varepsilon^2|\text{rest}\right) \propto \mathfrak{p}\left(\boldsymbol{y}|\beta_0, \boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2\right) \mathfrak{p}\left(\sigma_\varepsilon^2|a_\varepsilon\right). \tag{3.177}$$

Moreover,

$$\boldsymbol{y}|\beta_0, \boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2 \sim N\left(\boldsymbol{1}_n\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j\boldsymbol{u}_j, \sigma_\varepsilon^2 \boldsymbol{I}_n\right)$$

and

$$\sigma_\varepsilon^2|a_\varepsilon \sim \text{Inverse-Gamma}\left(\frac{1}{2}, 1/a_\varepsilon\right).$$

For $\mathfrak{p}\left(\boldsymbol{y}|\beta_0, \boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2\right)$ stated in (3.177), we have

$$\mathfrak{p}\left(\boldsymbol{y}|\beta_0, \boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2\right) \propto \left(\frac{1}{\sqrt{\sigma_\varepsilon^2}}\right)^n \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\left\|\boldsymbol{y} - \boldsymbol{1}_n\beta_0 - \boldsymbol{X}\boldsymbol{\beta} - \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j\boldsymbol{u}_j\right\|^2\right\}.$$

For $\mathfrak{p}\left(\sigma_\varepsilon^2|a_\varepsilon\right)$ stated in (3.177), we have

$$\mathfrak{p}\left(\sigma_\varepsilon^2|a_\varepsilon\right) \propto \left(\sigma_\varepsilon^2\right)^{-\frac{1}{2}-1} \exp\left(-\frac{1}{a_\varepsilon\sigma_\varepsilon^2}\right).$$

Therefore, $\mathfrak{p}\left(\sigma_\varepsilon^2|\text{rest}\right)$ stated in (3.177) becomes

$$\mathfrak{p}\left(\sigma_\varepsilon^2|\text{rest}\right) \propto \left(\sigma_\varepsilon^2\right)^{-\frac{n+1}{2}-1} \exp\left\{-\frac{1}{\sigma_\varepsilon^2}\left(\frac{1}{a_\varepsilon} + \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{1}_n\beta_0 - \boldsymbol{X}\boldsymbol{\beta} - \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j\boldsymbol{u}_j\right\|^2\right)\right\}.$$

In summary, for Gaussian responses, we obtain (3.38) in Section 3.6.3 to be

$$\sigma_\varepsilon^2|\text{rest} \sim \text{Inverse-Gamma}\left(\frac{n+1}{2}, \frac{1}{a_\varepsilon} + \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{1}_n\beta_0 - \boldsymbol{X}\boldsymbol{\beta} - \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j\boldsymbol{u}_j\right\|^2\right).$$

If we define

$$\boldsymbol{\omega}_{10} \equiv \mathbf{1}_n\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j\boldsymbol{u}_j,$$

then

$$\sigma_\varepsilon^2|\text{rest} \sim \text{Inverse-Gamma}\left(\frac{1}{2}(n+1), (1/a_\varepsilon) + \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{\omega}_{10}\right\|^2\right)$$

is as given in Algorithm 4 of Section 3.6.3.

3.F.5 Full Conditional Distribution of a_ε for Gaussian responses

Following the steps in Appendix 2.B.7, we obtain (3.39) in Section 3.6.3 to be

$$a_\varepsilon|\text{rest} \sim \text{Inverse-Gamma}\left(1, \frac{1}{\sigma_\varepsilon^2} + \frac{1}{s_\varepsilon^2}\right).$$

3.F.6 Full Conditional Distributions of $f_{2\beta j}$ for Gaussian responses

First note that, for $1 \le j \le d_\circ + d_\bullet$,

$$\mathfrak{p}\left(f_{2\beta_j}|\text{rest}\right) \propto \mathfrak{p}\left(\beta_j|\sigma_\beta^2, f_{2\beta_j}\right)\mathfrak{p}\left(f_{2\beta_j}|f_{1\beta_j}\right).$$

For $1 \le j \le d_\circ + d_\bullet$, note that

$$\beta_j|\sigma_\beta^2, f_{2\beta_j} \overset{\text{ind.}}{\sim} N\left(0, \left(\sigma_\beta^2/f_{2\beta_j}\right)\right) \text{ and } f_{2\beta_j}|f_{1\beta_j} \overset{\text{ind.}}{\sim} \text{Gamma}\left(\tfrac{1}{2}, f_{1\beta_j}\right).$$

We then have

$$\begin{aligned}\mathfrak{p}\left(f_{2\beta_j}|\text{rest}\right) &\propto \left(f_{2\beta_j}\right)^{\frac{1}{2}}\exp\left\{-\tfrac{f_{2\beta_j}\beta_j^2}{2\sigma_\beta^2}\right\}\left(f_{2\beta_j}\right)^{-\frac{1}{2}}\exp\left\{-f_{1\beta_j}f_{2\beta_j}\right\}\\ &\propto \left(f_{2\beta_j}\right)^{1-1}\exp\left\{-f_{2\beta_j}\left(\frac{\beta_j^2}{2\sigma_\beta^2} + f_{1\beta_j}\right)\right\}.\end{aligned}$$

Therefore

$$f_{2\beta_j}|\text{rest} \overset{\text{ind.}}{\sim} \text{Gamma}\left(1, \frac{\beta_j^2}{2\sigma_\beta^2} + f_{1\beta_j}\right), 1 \le j \le d_\circ + d_\bullet$$

is obtained to be (3.29) in Section 3.6.3.

3.F.7 Full Conditional Distributions of $f_{1\beta j}$ for Gaussian responses

First note that, for $1 \leq j \leq d_\circ + d_\bullet$,

$$\mathfrak{p}\left(f_{1\beta_j}\middle|\text{rest}\right) \propto \mathfrak{p}\left(f_{2\beta_j}\middle|f_{1\beta_j}\right)\mathfrak{p}\left(f_{1\beta_j}\right).$$

For $1 \leq j \leq d_\circ + d_\bullet$, note that

$$f_{2\beta_j}\middle|f_{1\beta_j} \overset{\text{ind.}}{\sim} \text{Gamma}\left(\tfrac{1}{2}, f_{1\beta_j}\right) \text{ and } \quad f_{1\beta_j} \overset{\text{ind.}}{\sim} \text{Gamma}\left(\tfrac{1}{2}, 1\right).$$

We then have

$$\begin{aligned}\mathfrak{p}\left(f_{1\beta_j}\middle|\text{rest}\right) &\propto \left(f_{1\beta_j}\right)^{\frac{1}{2}} \exp\left\{-f_{1\beta_j}f_{2\beta_j}\right\}\left(f_{1\beta_j}\right)^{-\frac{1}{2}} \exp\left\{-f_{1\beta_j}\right\} \\ &\propto \left(f_{1\beta_j}\right)^{1-1} \exp\left\{-f_{1\beta_j}\left(1+f_{2\beta_j}\right)\right\}.\end{aligned}$$

Therefore

$$f_{1\beta_j}\middle|\text{rest} \overset{\text{ind.}}{\sim} \text{Gamma}\left(1, 1+f_{2\beta_j}\right), 1 \leq j \leq d_\circ + d_\bullet$$

is obtained to be (3.30) in Section 3.6.3.

3.F.8 Full Conditional Distributions of f_{2uj} for Gaussian responses

First note that, for $1 \leq j \leq d_\bullet$,

$$\mathfrak{p}\left(f_{2uj}\middle|\text{rest}\right) \propto \mathfrak{p}\left(\boldsymbol{u}_j\middle|\sigma_{uj}^2, f_{2uj}\right)\mathfrak{p}\left(f_{2uj}\middle|f_{1uj}\right).$$

For $1 \leq j \leq d_\bullet$, note that

$$\boldsymbol{u}_j\middle|\sigma_{uj}^2, f_{2uj} \overset{\text{ind.}}{\sim} N\left(\boldsymbol{0}, \left(\sigma_{uj}^2/f_{2uj}\right)\boldsymbol{I}_{K_j}\right) \text{ and } \; f_{2uj}\middle|f_{1uj} \overset{\text{ind.}}{\sim} \text{Gamma}\left(\tfrac{1}{2}, f_{1uj}\right).$$

We then have

$$\mathfrak{p}(f_{2u_j}|\text{rest}) \propto \left(\frac{1}{\sqrt{1/f_{2uj}}}\right)^{K_j} \exp\left\{-\frac{\left\|\boldsymbol{u}_j\right\|^2}{2\left(\sigma_{uj}^2/f_{2uj}\right)\boldsymbol{I}_{K_j}}\right\}\left(f_{2uj}\right)^{-\frac{1}{2}}\exp\left\{-f_{1uj}f_{2uj}\right\}$$

$$\propto \left(f_{2uj}\right)^{\frac{K_j+1}{2}-1}\exp\left\{-f_{2uj}\left(\frac{\left\|\boldsymbol{u}_j\right\|^2}{2\sigma_{uj}^2}+f_{1uj}\right)\right\}.$$

Therefore

$$f_{2uj}|\text{rest} \overset{\text{ind.}}{\sim} \text{Gamma}\left(\frac{K_j+1}{2}, \frac{\left\|\boldsymbol{u}_j\right\|^2}{2\sigma_{uj}^2}+f_{1uj}\right), 1 \le j \le d_\bullet$$

is obtained to be (3.34) in Section 3.6.3.

3.F.9 Full Conditional Distributions of f_{1uj} for Gaussian responses

First note that, for $1 \le j \le d_\bullet$,

$$\mathfrak{p}\left(f_{1uj}|\text{rest}\right) \propto \mathfrak{p}\left(f_{2uj}|f_{1uj}\right)\mathfrak{p}(f_{1uj}).$$

For $1 \le j \le d_\bullet$, note that

$$f_{2uj}|f_{1uj} \overset{\text{ind.}}{\sim} \text{Gamma}\left(\tfrac{1}{2}, f_{1uj}\right) \text{ and } f_{1uj} \overset{\text{ind.}}{\sim} \text{Gamma}\left(\tfrac{1}{2}, 1\right).$$

We then have

$$\mathfrak{p}(f_{1uj}|\text{rest}) \propto \left(f_{1uj}\right)^{\frac{1}{2}}\exp\left\{-f_{1uj}f_{2uj}\right\}\left(f_{1uj}\right)^{-\frac{1}{2}}\exp\left\{-f_{1uj}\right\}$$

$$\propto \left(f_{1uj}\right)^{1-1}\exp\left\{-f_{1uj}\left(1+f_{2uj}\right)\right\}.$$

Therefore

$$f_{1uj}|\text{rest} \overset{\text{ind.}}{\sim} \text{Gamma}\left(1, 1+f_{2uj}\right), 1 \le j \le d_\bullet$$

is obtained to be (3.35) in Section 3.6.3.

3.F.10 Full Conditional Distribution of σ_β^2 for Gaussian responses

Following the steps in Appendix 2.B.10, we obtain (3.31) in Section 3.6.3 to be

$$\sigma_\beta^2|\text{rest} \sim \text{Inverse-Gamma}\left(\frac{d_\circ + d_\bullet + 1}{2}, \frac{\boldsymbol{\beta}^T \text{diag}(\boldsymbol{f}_{2\beta})\boldsymbol{\beta}}{2} + \frac{1}{a_\beta}\right).$$

3.F.11 Full Conditional Distribution of a_β for Gaussian responses

Following the steps in Appendix 2.B.11, we obtain (3.32) in Section 3.6.3 to be

$$a_\beta|\text{rest} \sim \text{Inverse-Gamma}\left(1, \frac{1}{\sigma_\beta^2} + \frac{1}{s_\beta^2}\right).$$

3.F.12 Full Conditional Distribution of σ_{uj}^2 for Gaussian responses

Following the steps in Appendix 2.B.12, we obtain (3.36) in Section 3.6.3 to be

$$\sigma_{uj}^2|\text{rest} \sim \text{Inverse-Gamma}\left(\frac{K_j + 1}{2}, \frac{f_{2uj}\|\boldsymbol{u}_j\|^2}{2} + \frac{1}{a_{uj}}\right), 1 \leq j \leq d_\bullet.$$

3.F.13 Full Conditional Distribution of a_{uj} for Gaussian responses

Following the steps in Appendix 2.B.13, we obtain (3.37) in Section 3.6.3 to be

$$a_{uj}|\text{rest} \sim \text{Inverse-Gamma}\left(1, \frac{1}{\sigma_{uj}^2} + \frac{1}{s_u^2}\right), 1 \leq j \leq d_\bullet.$$

3.G Full Conditional Distributions Derivation for Poisson Responses

3.G.1 Full Conditional Distributions of β_0 for Poisson Responses

First note that

$$\mathfrak{p}(\beta_0|\text{rest}) \propto \left\{\prod_{i=1}^{n} \mathfrak{p}(y_i|\beta_0, \boldsymbol{\beta}, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_{d_\bullet})\right\} \mathfrak{p}(\beta_0).$$

Recalling that

$$y_i|\beta_0, \boldsymbol{\beta}, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_{d_\bullet} \sim \text{Poisson}\left(\exp\left\{\beta_0 + (\boldsymbol{X}\boldsymbol{\beta})_i + \sum_{j=1}^{d_\bullet} (\boldsymbol{Z}_j \boldsymbol{u}_j)_i\right\}\right), \;\; 1 \leq i \leq n,$$

and

$$\beta_0 \sim N\left(0, \sigma^2_{\beta_0}\right),$$

then

$$\begin{aligned}
\mathfrak{p}(\beta_0|\text{rest}) \propto \prod_{i=1}^{n} \Bigg\{ &\left(\exp\left\{\beta_0 + (\boldsymbol{X}\boldsymbol{\beta})_i + \sum_{j=1}^{d_\bullet} (\boldsymbol{Z}_j \boldsymbol{u}_j)_i\right\}\right)^{y_i} \\
&\times \exp\left(-\exp\left\{\beta_0 + (\boldsymbol{X}\boldsymbol{\beta})_i + \sum_{j=1}^{d_\bullet} (\boldsymbol{Z}_j \boldsymbol{u}_j)_i\right\}\right)\Bigg\} \exp\left(-\frac{\beta_0^2}{2\sigma^2_{\beta_0}}\right) \\
\propto \exp&\left\{n\bar{\boldsymbol{y}}\beta_0 - \frac{\beta_0^2}{2\sigma^2_{\beta_0}} - \mathbf{1}_n^T \exp\left(\mathbf{1}_n\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j \boldsymbol{u}_j\right)\right\}
\end{aligned}$$

is obtained to be (3.40) in Section 3.6.3, if we define $\sum_{i=1}^{n} y_i \equiv n\bar{\boldsymbol{y}}$. Following the form of (3.26) for logarithm of unnormalised density function, we have

$$s_1(\beta_0) = n\bar{\boldsymbol{y}}, \;\; s_2(\beta_0) = \sigma^2_{\beta_0}, \;\; \boldsymbol{s}_3(\beta_0) = \mathbf{1}_n \;\text{ and }\; \boldsymbol{s}_4(\beta_0) = \boldsymbol{X}\boldsymbol{\beta} + \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j \boldsymbol{u}_j$$

for β_0 slice sampling for Poisson responses described in Algorithm 5.

3.G.2 Full Conditional Distributions of β_j for Poisson Responses

First note that, for $1 \leq j \leq d_\circ + d_\bullet$,

$$\mathfrak{p}(\beta_j|\text{rest}) \propto \left\{ \prod_{i=1}^{n} \mathfrak{p}(y_i|\beta_0, \boldsymbol{\beta}, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_{d_\bullet}) \right\} \mathfrak{p}(\beta_j|\sigma_\beta^2, f_{2\beta_j}).$$

Recalling that for $1 \leq i \leq n, 1 \leq j \leq d_\circ + d_\bullet$,

$$y_i|\beta_0, \boldsymbol{\beta}, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_{d_\bullet} \sim \text{Poisson}\left(\exp\left\{ \beta_0 + (\boldsymbol{X}\boldsymbol{\beta})_i + \sum_{j=1}^{d_\bullet} (\boldsymbol{Z}_j \boldsymbol{u}_j)_i \right\} \right), \;\; 1 \leq i \leq n,$$

and

$$\beta_j|\sigma_\beta^2, f_{2\beta_j} \sim N\left(0, \left(\sigma_\beta^2 / f_{2\beta_j}\right)\right), \;\; (\boldsymbol{X}\boldsymbol{\beta})_i = X_{ij}\beta_j + (\boldsymbol{X}_i)_{-j} (\boldsymbol{\beta})_{-j}, \;\; 1 \leq j \leq d_\circ + d_\bullet,$$

where

$(\boldsymbol{X}_i)_{-j}$ denotes the $1 \times (d_\circ + d_\bullet)$ vector $\boldsymbol{X}_i$ with the jth entry removed, and

$(\boldsymbol{\beta})_{-j}$ denotes the $(d_\circ + d_\bullet) \times 1$ vector $\boldsymbol{\beta}$ with the jth entry removed,

then

$$\begin{aligned}
\mathfrak{p}(\beta_j|\text{rest}) &\propto \prod_{i=1}^{n} \left[\left(\exp\left\{ \beta_0 + \left(X_{ij}\beta_j + (\boldsymbol{X}_i)_{-j} (\boldsymbol{\beta})_{-j} \right) + \sum_{j=1}^{d_\bullet} (\boldsymbol{Z}_j \boldsymbol{u}_j)_i \right\} \right)^{y_i} \right. \\
&\quad \left. \times \exp\left(- \exp\left\{ \beta_0 + \left(X_{ij}\beta_j + (\boldsymbol{X}_i)_{-j} (\boldsymbol{\beta})_{-j} \right) + \sum_{j=1}^{d_\bullet} (\boldsymbol{Z}_j \boldsymbol{u}_j)_i \right\} \right) \right] \exp\left(- \frac{\beta_j^2}{2\left(\sigma_\beta^2 / f_{2\beta_j}\right)} \right) \\
&\propto \exp\left\{ \mathbf{1}_n^T (\boldsymbol{y} \odot \boldsymbol{X}_j) \beta_j - \frac{f_{2\beta_j}\beta_j^2}{2\sigma_\beta^2} - \mathbf{1}_n^T \exp\left(\boldsymbol{X}_j\beta_j + \mathbf{1}_n\beta_0 + (\boldsymbol{X})_{-j} (\boldsymbol{\beta})_{-j} + \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j \boldsymbol{u}_j \right) \right\}
\end{aligned}$$

is obtained to be (3.41) in Section 3.6.3. Following the form of (3.26) for logarithm of unnormalised density function, we have

$$s_1 = \mathbf{1}_n^T (\boldsymbol{y} \odot \boldsymbol{X}_j), \;\; s_2 = \sigma_\beta^2 / f_{2\beta_j}, \;\; \boldsymbol{s}_3 = \boldsymbol{X}_j \;\text{ and }\; \boldsymbol{s}_4 = \mathbf{1}_n\beta_0 + (\boldsymbol{X})_{-j} (\boldsymbol{\beta})_{-j} + \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j \boldsymbol{u}_j$$

for β_j slice sampling for Poisson responses described in Algorithm 5, $1 \leq j \leq d_\circ + d_\bullet$.

3.G.3 Full Conditional Distributions of u_{jk} for Poisson Responses

First note that, for $1 \leq j \leq d_\bullet, 1 \leq k \leq K_j$,

$$\mathfrak{p}(u_{jk}|\text{rest}) \propto \left\{ \prod_{i=1}^{n} \mathfrak{p}(y_i|\beta_0, \boldsymbol{\beta}, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_{d_\bullet}) \right\} \mathfrak{p}(u_{jk}|\sigma_{uj}^2, f_{2uj}).$$

For $1 \leq i \leq n,\ \ 1 \leq j \leq d_\bullet,\ \ 1 \leq k \leq K_j$, recalling that

$$y_i|\beta_0, \boldsymbol{\beta}, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_{d_\bullet} \sim \text{Poisson}\left(\exp\left\{ \beta_0 + (\boldsymbol{X}\boldsymbol{\beta})_i + \sum_{j=1}^{d_\bullet} (\boldsymbol{Z}_j \boldsymbol{u}_j)_i \right\} \right),$$

$$u_{jk}|\sigma_{uj}^2, f_{2uj} \sim N\left(0, \left(\sigma_{uj}^2 / f_{2uj}\right)\right) \text{ and } \sum_{j=1}^{d_\bullet} (\boldsymbol{Z}_j \boldsymbol{u}_j)_i = Z_{ijk} u_{jk} + \sum_{k' \neq k}^{K_j} Z_{ijk'} u_{jk'} + \sum_{j' \neq j}^{d_\bullet} \boldsymbol{Z}_{ij'} \boldsymbol{u}_{j'},$$

then

$$\begin{aligned}
\mathfrak{p}(u_{jk}|\text{rest}) \propto \prod_{i=1}^{n} \Bigg\{ & \left(\exp\left\{ \beta_0 + (\boldsymbol{X}\boldsymbol{\beta})_i + Z_{ijk} u_{jk} + \sum_{k' \neq k}^{K_j} Z_{ijk'} u_{jk'} + \sum_{j' \neq j}^{d_\bullet} \boldsymbol{Z}_{ij'} \boldsymbol{u}_{j'} \right\} \right)^{y_i} \\
& \times \exp\left(- \exp\left\{ \beta_0 + (\boldsymbol{X}\boldsymbol{\beta})_i + Z_{ijk} u_{jk} + \sum_{k' \neq k}^{K_j} Z_{ijk'} u_{jk'} + \sum_{j' \neq j}^{d_\bullet} \boldsymbol{Z}_{ij'} \boldsymbol{u}_{j'} \right\} \right) \Bigg\} \\
& \times \exp\left(- \frac{u_{jk}^2}{2\left(\sigma_{uj}^2 / f_{2uj}\right)} \right) \\
\propto \exp\Bigg\{ & \mathbf{1}_n^T \left(\boldsymbol{y} \odot \boldsymbol{Z}_{jk} \right) u_{jk} - \frac{f_{2uj} u_{jk}^2}{2\sigma_{uj}^2} - \mathbf{1}_n^T \exp\left(\boldsymbol{Z}_{jk} u_{jk} + \mathbf{1}_n \beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \sum_{k' \neq k}^{K_j} \boldsymbol{Z}_{jk'} u_{jk'} \right. \\
& \left. + \sum_{j' \neq j}^{d_\bullet} \boldsymbol{Z}_{j'} \boldsymbol{u}_{j'} \right) \Bigg\}
\end{aligned}$$

is obtained to be (3.46) in Section 3.6.3. Following the form of (3.26) for logarithm of

unnormalised density function, we have

$$s_1 = \mathbf{1}_n^T \left(\boldsymbol{y} \odot \boldsymbol{Z}_{jk} \right), \;\; s_2 = \sigma_{uj}^2 / f_{2uj}, \;\; \boldsymbol{s}_3 = \boldsymbol{Z}_{jk},$$
$$\text{and} \quad \boldsymbol{s}_4 = \mathbf{1}_n \beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \sum_{k' \neq k}^{K_j} \boldsymbol{Z}_{jk'} u_{jk'} + \sum_{j' \neq j}^{d_\bullet} \boldsymbol{Z}_{ij'} \boldsymbol{u}_{j'}$$

for u_{jk} slice sampling for Poisson responses described in Algorithm 5, $1 \leq j \leq d_\bullet, 1 \leq k \leq K_j$.

3.G.4 Full Conditional Distributions of $f_{2\beta_j}$ for Poisson Responses

Following the steps in Appendix 3.F.6, we obtain (3.42) in Section 3.6.3 to be

$$f_{2\beta_j} \big| \text{rest} \overset{\text{ind.}}{\sim} \text{Gamma} \left(1, \frac{\beta_j^2}{2\sigma_\beta^2} + f_{1\beta_j} \right), 1 \leq j \leq d_\circ + d_\bullet.$$

3.G.5 Full Conditional Distributions of $f_{1\beta_j}$ for Poisson Responses

Following the steps in Appendix 3.F.7, we obtain (3.43) in Section 3.6.3 to be

$$f_{1\beta_j} \big| \text{rest} \overset{\text{ind.}}{\sim} \text{Gamma} \left(1, 1 + f_{2\beta_j} \right), 1 \leq j \leq d_\circ + d_\bullet.$$

3.G.6 Full Conditional Distributions of f_{2uj} for Poisson Responses

Following the steps in Appendix 3.F.8, we obtain (3.47) in Section 3.6.3 to be

$$f_{2uj} \big| \text{rest} \overset{\text{ind.}}{\sim} \text{Gamma} \left(\frac{K_j + 1}{2}, \frac{\left\| \boldsymbol{u}_j \right\|^2}{2\sigma_{uj}^2} + f_{1uj} \right), 1 \leq j \leq d_\bullet.$$

3.G.7 Full Conditional Distributions of f_{1uj} for Poisson Responses

Following the steps in Appendix 3.F.9, we obtain (3.48) in Section 3.6.3 to be

$$f_{1uj} \big| \text{rest} \overset{\text{ind.}}{\sim} \text{Gamma} \left(1, 1 + f_{2uj} \right), 1 \leq j \leq d_\bullet.$$

3.G.8 Full Conditional Distribution of σ_β^2 for Poisson Responses

Following the steps in Appendix 2.B.10, we obtain (3.44) in Section 3.6.3 to be

$$\sigma_\beta^2 \big| \text{rest} \sim \text{Inverse-Gamma}\left(\frac{d_\circ + d_\bullet + 1}{2}, \frac{\boldsymbol{\beta}^T \text{diag}(\boldsymbol{f}_{2\beta})\boldsymbol{\beta}}{2} + \frac{1}{a_\beta}\right)$$

3.G.9 Full Conditional Distributions of a_β for Poisson Responses

Following the steps in Appendix 2.B.11, we obtain (3.45) in Section 3.6.3 to be

$$a_\beta \big| \text{rest} \sim \text{Inverse-Gamma}\left(1, \frac{1}{\sigma_\beta^2} + \frac{1}{s_\beta^2}\right).$$

3.G.10 Full Conditional Distributions of σ_{uj}^2 for Poisson Responses

Following the steps in Appendix 2.B.12, we obtain (3.49) in Section 3.6.3 to be

$$\sigma_{uj}^2 \big| \text{rest} \sim \text{Inverse-Gamma}\left(\frac{K_j + 1}{2}, \frac{f_{2uj}\|\boldsymbol{u}_j\|^2}{2} + \frac{1}{a_{uj}}\right), 1 \leq j \leq d_\bullet.$$

3.G.11 Full Conditional Distributions of a_{uj} for Poisson Responses

Following the steps in Appendix 2.B.13, we obtain (3.50) in Section 3.6.3 to be

$$a_{uj} \big| \text{rest} \sim \text{Inverse-Gamma}\left(1, \frac{1}{\sigma_{uj}^2} + \frac{1}{s_u^2}\right), 1 \leq j \leq d_\bullet.$$

3.H Full Conditional Distributions Derivation for Negative-Binomial Responses

Throughout this appendix, additive constants with respect to the function argument are denoted by 'const'.

3.H.1 Full Conditional Distributions of g for Negative-Binomial Responses

First note that

$$\mathfrak{p}(\boldsymbol{g}|\text{rest}) \propto \left\{\prod_{i=1}^{n} \mathfrak{p}(y_i|g_i)\right\}\left\{\prod_{i=1}^{n} \mathfrak{p}(g_i|\beta_0, \boldsymbol{\beta}, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_{d_\bullet}, \kappa)\right\}.$$

For $1 \leq i \leq n$, recalling that

$$y_i|g_i \sim \text{Pois}\left(g_i\right), \;\; \text{Pois}(g_i) = \frac{g_i^{y_i}\exp\left(-g_i\right)}{y_i!},$$

$$\text{and} \;\; g_i|\beta_0, \boldsymbol{\beta}, \boldsymbol{u}, \kappa \sim \text{Gamma}\left(\kappa, \kappa\exp\left\{-\left(\beta_0 + \left(\boldsymbol{X}\boldsymbol{\beta}\right)_i + \sum_{j=1}^{d_\bullet}\left(\boldsymbol{Z}_j\boldsymbol{u}_j\right)_i\right)\right\}\right),$$

then we have

$$\begin{aligned}
\mathfrak{p}(\boldsymbol{g}|\text{rest}) &\propto \left\{\prod_{i=1}^{n} \mathfrak{p}(y_i|g_i)\right\}\left\{\prod_{i=1}^{n} \mathfrak{p}(g_i|\beta_0, \boldsymbol{\beta}, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_{d_\bullet}, \kappa)\right\} \\
&\propto \prod_{i=1}^{n} g_i^{y_i}\exp\left(-g_i\right)\left(g_i\right)^{\kappa-1}\exp\left[-g_i\kappa\exp\left\{-\left(\beta_0 + \left(\boldsymbol{X}\boldsymbol{\beta}\right)_i + \sum_{j=1}^{d_\bullet}\left(\boldsymbol{Z}_j\boldsymbol{u}_j\right)_i\right)\right\}\right] \\
&\propto \prod_{i=1}^{n} \left(g_i\right)^{y_i+\kappa-1}\exp\left[-g_i\left(1 + \kappa\exp\left\{-\left(\beta_0 + \left(\boldsymbol{X}\boldsymbol{\beta}\right)_i + \sum_{j=1}^{d_\bullet}\left(\boldsymbol{Z}_j\boldsymbol{u}_j\right)_i\right)\right\}\right)\right].
\end{aligned}$$

Therefore

$$g_i|\text{rest} \sim \text{Gamma}\left(y_i + \kappa, 1 + \kappa\exp\left\{-\left(\beta_0 + \left(\boldsymbol{X}\boldsymbol{\beta}\right)_i + \sum_{j=1}^{d_\bullet}\left(\boldsymbol{Z}_j\boldsymbol{u}_j\right)_i\right)\right\}\right),$$

which is equivalent to

$$\boldsymbol{g}\big|\text{rest} \sim \text{Gamma}\left(\boldsymbol{y}+\kappa, 1+\kappa\exp\left\{-\left(\mathbf{1}_n\beta_0+\boldsymbol{X}\boldsymbol{\beta}+\sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j\boldsymbol{u}_j\right)\right\}\right).$$

Recalling that $\kappa=\exp(\lambda)$, we obtain

$$\boldsymbol{g}\big|\text{rest} \sim \text{Gamma}\left(\boldsymbol{y}+\exp(\lambda), 1+\exp\left\{\lambda-\left(\mathbf{1}_n\beta_0+\boldsymbol{X}\boldsymbol{\beta}+\sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j\boldsymbol{u}_j\right)\right\}\right)$$

to be (3.51) in Section 3.6.3.

3.H.2 Full Conditional Distributions of λ for Negative-Binomial Responses

Conversion from κ to $\lambda=\log(\kappa)$ for Negative-Binomial responses is to ensure κ always being positive, we have

$$\kappa=\exp(\lambda),\ \ \frac{d(\kappa)}{d(\lambda)}=\exp(\lambda) \text{ and } \kappa\sim\text{Moon-Rock}(a_\kappa,b_\kappa).$$

Then

$$\mathfrak{p}(\kappa)\propto\left\{\frac{\kappa^\kappa}{\Gamma(\kappa)}\right\}^{a_\kappa}\exp(-b_\kappa\kappa),\kappa>0.$$

Therefore

$$\begin{aligned}\mathfrak{p}(\lambda)&\propto\left\{\frac{(\exp(\lambda))^{\exp(\lambda)}}{\Gamma(\exp(\lambda))}\right\}^{a_\kappa}\exp(-b_\kappa\exp(\lambda))\frac{d(\kappa)}{d(\lambda)}\\&\propto\left\{\frac{(\exp(\lambda))^{\exp(\lambda)}}{\Gamma(\exp(\lambda))}\right\}^{a_\kappa}\exp(\lambda-b_\kappa\exp(\lambda)).\end{aligned}$$

First note that

$$\mathfrak{p}(\lambda\big|\text{rest})\propto\left\{\prod_{i=1}^{n}\mathfrak{p}(g_i\big|\beta_0,\boldsymbol{\beta},\boldsymbol{u}_1,\ldots,\boldsymbol{u}_{d_\bullet},\lambda)\right\}\mathfrak{p}(\lambda).$$

Recalling that

$$g_i \big| \beta_0, \boldsymbol{\beta}, \boldsymbol{u}, \kappa \sim \text{Gamma}\left(\kappa, \kappa \exp\left\{-\left(\beta_0 + (\boldsymbol{X}\boldsymbol{\beta})_i + \sum_{j=1}^{d_\bullet} (\boldsymbol{Z}_j \boldsymbol{u}_j)_i\right)\right\}\right), \quad 1 \leq i \leq n,$$

and

$$\kappa = \exp(\lambda),$$

we have

$$g_i \big| \beta_0, \boldsymbol{\beta}, \boldsymbol{u}, \lambda \sim \text{Gamma}\left(\exp(\lambda), \exp\left\{\lambda - \left(\beta_0 + (\boldsymbol{X}\boldsymbol{\beta})_i + \sum_{j=1}^{d_\bullet} (\boldsymbol{Z}_j \boldsymbol{u}_j)_i\right)\right\}\right).$$

Therefore

$$\begin{aligned}
&\mathfrak{p}(g_i \big| \beta_0, \boldsymbol{\beta}, \boldsymbol{u}, \lambda) \\
&= \frac{\left[\exp\left\{\lambda - \left(\beta_0 + (\boldsymbol{X}\boldsymbol{\beta})_i + \sum_{j=1}^{d_\bullet} (\boldsymbol{Z}_j \boldsymbol{u}_j)_i\right)\right\}\right]^{\exp(\lambda)} (g_i)^{\exp(\lambda)-1}}{\Gamma(\exp(\lambda))} \\
&\qquad \times \exp\left[-g_i \exp\left\{\lambda - \left(\beta_0 + (\boldsymbol{X}\boldsymbol{\beta})_i + \sum_{j=1}^{d_\bullet} (\boldsymbol{Z}_j \boldsymbol{u}_j)_i\right)\right\}\right].
\end{aligned}$$

Then we have

$$\begin{aligned}
\mathfrak{p}(\lambda \big| \text{rest}) \propto \prod_{i=1}^{n} \Bigg\{ & \frac{\left[\exp\left\{\lambda - \left(\beta_0 + (\boldsymbol{X}\boldsymbol{\beta})_i + \sum_{j=1}^{d_\bullet} (\boldsymbol{Z}_j \boldsymbol{u}_j)_i\right)\right\}\right]^{\exp(\lambda)} (g_i)^{\exp(\lambda)-1}}{\Gamma(\exp(\lambda))} \\
& \times \exp\left[-g_i \exp\left\{\lambda - \left(\beta_0 + (\boldsymbol{X}\boldsymbol{\beta})_i + \sum_{j=1}^{d_\bullet} (\boldsymbol{Z}_j \boldsymbol{u}_j)_i\right)\right\}\right] \Bigg\} \left\{\frac{(\exp(\lambda))^{\exp(\lambda)}}{\Gamma(\exp(\lambda))}\right\}^{a_\kappa} \\
& \times \exp(\lambda - b_\kappa \exp(\lambda)).
\end{aligned}$$

After simplification, we have

$$\mathfrak{p}(\lambda|\text{rest}) \propto \frac{\left\{\frac{\{\exp(\lambda)\}^{\exp(\lambda)}}{\Gamma(\exp(\lambda))}\right\}^{n+a_\kappa} \left\{\prod_{i=1}^{n}(g_i)^{\exp(\lambda)-1}\right\}}{\prod_{i=1}^{n}\left[\left\{\exp\left(\beta_0 + (\boldsymbol{X}\boldsymbol{\beta})_i + \sum_{j=1}^{d_\bullet}(\boldsymbol{Z}_j\boldsymbol{u}_j)_i\right)\right\}^{\exp(\lambda)}\right]}$$
$$\times \exp\left[-\sum_{i=1}^{n}\left\{g_i \exp\left(\lambda - \left(\beta_0 + (\boldsymbol{X}\boldsymbol{\beta})_i + \sum_{j=1}^{d_\bullet}(\boldsymbol{Z}_j\boldsymbol{u}_j)_i\right)\right)\right\} + \lambda - b_\kappa \exp(\lambda)\right].$$

Therefore

$$\log\left\{\mathfrak{p}(\lambda|\text{rest})\right\} = (n+a_\kappa)\log\left\{\frac{\{\exp(\lambda)\}^{\exp(\lambda)}}{\Gamma(\exp(\lambda))}\right\} + \exp(\lambda)\sum_{i=1}^{n}\log(g_i) - \exp(\lambda)$$
$$\times\sum_{i=1}^{n}\left(\beta_0 + (\boldsymbol{X}\boldsymbol{\beta})_i + \sum_{j=1}^{d_\bullet}(\boldsymbol{Z}_j\boldsymbol{u}_j)_i\right)$$
$$-\sum_{i=1}^{n}\left\{g_i \exp\left(\lambda - \left(\beta_0 + (\boldsymbol{X}\boldsymbol{\beta})_i + \sum_{j=1}^{d_\bullet}(\boldsymbol{Z}_j\boldsymbol{u}_j)_i\right)\right)\right\} + \lambda - b_\kappa \exp(\lambda) + \text{const.}$$

Note that

$$\log\left\{\frac{\{\exp(\lambda)\}^{\exp(\lambda)}}{\Gamma(\exp(\lambda))}\right\} = \lambda\exp(\lambda) - \log\left\{\Gamma(\exp(\lambda))\right\}.$$

And note that

$$(n+a_\kappa)\log\left\{\frac{\{\exp(\lambda)\}^{\exp(\lambda)}}{\Gamma(\exp(\lambda))}\right\} = (n+a_\kappa)\left(\lambda\exp(\lambda) - \log\left\{\Gamma(\exp(\lambda))\right\}\right).$$

Therefore

$$\log\left\{\mathfrak{p}(\lambda\big|\text{rest})\right\} = (n+a_\kappa)\,\lambda\exp(\lambda) - (n+a_\kappa)\left(\log\left\{\Gamma\left(\exp(\lambda)\right)\right\}\right) + \exp(\lambda)$$
$$\times\sum_{i=1}^{n}\left[\log(g_i) - \left(\beta_0 + (\boldsymbol{X}\boldsymbol{\beta})_i + \sum_{j=1}^{d_\bullet}(\boldsymbol{Z}_j\boldsymbol{u}_j)_i\right)\right.$$
$$\left.-g_i\exp\left\{-\left(\beta_0 + (\boldsymbol{X}\boldsymbol{\beta})_i + \sum_{j=1}^{d_\bullet}(\boldsymbol{Z}_j\boldsymbol{u}_j)_i\right)\right\}\right] + \lambda - b_\kappa\exp(\lambda) + \text{const.}$$

If we define $a_\kappa \equiv 0$, then

$$\log\left\{\mathfrak{p}(\lambda\big|\text{rest})\right\} = -\exp(\lambda)\left[b_\kappa + \sum_{i=1}^{n}\left(g_i\exp\left\{-\left(\beta_0 + (\boldsymbol{X}\boldsymbol{\beta})_i + \sum_{j=1}^{d_\bullet}(\boldsymbol{Z}_j\boldsymbol{u}_j)_i\right)\right\}\right)\right.$$
$$\left.-\sum_{i=1}^{n}\log(g_i) - n\lambda + \sum_{i=1}^{n}\left(\beta_0 + (\boldsymbol{X}\boldsymbol{\beta})_i + \sum_{j=1}^{d_\bullet}(\boldsymbol{Z}_j\boldsymbol{u}_j)_i\right)\right]$$
$$+\lambda - n\left(\log\left\{\Gamma\left(\exp(\lambda)\right)\right\}\right) + \text{const.}$$

Recalling that

$$\exp(\lambda) = \kappa, \text{ and } \lambda = \log(\kappa),$$

then

$$\log\left\{\mathfrak{p}(\lambda\big|\text{rest})\right\} = \lambda + n\left(\kappa\log(\kappa) - \log\left(\Gamma(\kappa)\right)\right) - \kappa\left[\sum_{j=1}^{n}\left(\beta_0 + (\boldsymbol{X}\boldsymbol{\beta})_i + \sum_{j=1}^{d_\bullet}(\boldsymbol{Z}_j\boldsymbol{u}_j)_i\right)\right.$$
$$\left.+b_\kappa + \sum_{i=1}^{n}\left(g_i\exp\left\{-\left(\beta_0 + (\boldsymbol{X}\boldsymbol{\beta})_i + \sum_{j=1}^{d_\bullet}(\boldsymbol{Z}_j\boldsymbol{u}_j)_i\right)\right\}\right) - \sum_{i=1}^{n}\log(g_i)\right] + \text{const}$$

is obtained to be (3.53) in Section 3.6.3.

3.H.3 Full Conditional Distributions of β_0 for Negative-Binomial Responses

First note that

$$\mathfrak{p}(\beta_0\big|\text{rest}) \propto \left\{\prod_{i=1}^{n}\mathfrak{p}(g_i\big|\beta_0, \boldsymbol{\beta}, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_{d_\bullet}, \kappa)\right\}\mathfrak{p}(\beta_0).$$

Recalling that

$$g_i\big|\beta_0,\boldsymbol{\beta},\boldsymbol{u},\kappa \sim \text{Gamma}\left(\kappa, \kappa\exp\left\{-\left(\beta_0+(\boldsymbol{X}\boldsymbol{\beta})_i+\sum_{j=1}^{d_\bullet}(\boldsymbol{Z}_j\boldsymbol{u}_j)_i\right)\right\}\right),\ \ 1\le i\le n,$$

and

$$\beta_0 \sim N\left(0,\sigma^2_{\beta_0}\right),$$

then

$$\begin{aligned}\mathfrak{p}(\beta_0\big|\text{rest}) \propto \prod_{i=1}^{n}\Bigg[&\frac{\left(\kappa\exp\Big\{-\big(\beta_0+(\boldsymbol{X}\boldsymbol{\beta})_i+\sum_{j=1}^{d_\bullet}(\boldsymbol{Z}_j\boldsymbol{u}_j)_i\big)\Big\}\right)^{\kappa} g_i^{\kappa-1}}{\Gamma(\kappa)}\\ &\times\exp\left(-g_i\kappa\exp\Big\{-\big(\beta_0+(\boldsymbol{X}\boldsymbol{\beta})_i+\sum_{j=1}^{d_\bullet}(\boldsymbol{Z}_j\boldsymbol{u}_j)_i\big)\Big\}\right)\Bigg]\exp\left(-\frac{\beta_0^2}{2\sigma^2_{\beta_0}}\right)\\ \propto\exp\Bigg\{&-n\kappa\beta_0-\frac{\beta_0^2}{2\sigma^2_{\beta_0}}-\mathbf{1}_n^T\exp\left[-\mathbf{1}_n\beta_0+\mathbf{1}_n\log(\kappa)+\log(\boldsymbol{g})-\left(\boldsymbol{X}\boldsymbol{\beta}+\sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j\boldsymbol{u}_j\right)\right]\Bigg\}\end{aligned}$$

is obtained to be (3.52) in Section 3.6.3. Following the form of (3.26) for logarithm of unnormalised density function, we have

$$s_1=-n\kappa=-n\exp(\lambda),\ \ s_2=\sigma^2_{\beta_0},\ \ \boldsymbol{s}_3=-\mathbf{1}_n \text{ and}$$

$$\boldsymbol{s}_4=\mathbf{1}_n\log(\kappa)+\log(\boldsymbol{g})-\left(\boldsymbol{X}\boldsymbol{\beta}+\sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j\boldsymbol{u}_j\right)=\mathbf{1}_n\lambda+\log(\boldsymbol{g})-\left(\boldsymbol{X}\boldsymbol{\beta}+\sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j\boldsymbol{u}_j\right)$$

for β_0 slice sampling for Negative-Binomial responses described in Algorithm 6.

3.H.4 Full Conditional Distributions of β_j for Negative-Binomial Responses

First note that, for $1\le j\le d_\circ+d_\bullet$,

$$\mathfrak{p}(\beta_j\big|\text{rest})\propto\left\{\prod_{i=1}^{n}\mathfrak{p}(g_i\big|\beta_0,\boldsymbol{\beta},\boldsymbol{u}_1,\ldots,\boldsymbol{u}_{d_\bullet},\kappa)\right\}\mathfrak{p}(\beta_j\big|\sigma^2_\beta,f_{2\beta_j}).$$

Recalling that

$$g_i\big|\beta_0,\boldsymbol{\beta},\boldsymbol{u},\kappa \sim \text{Gamma}\left(\kappa,\kappa\exp\left\{-\left(\beta_0+(\boldsymbol{X}\boldsymbol{\beta})_i+\sum_{j=1}^{d_\bullet}(\boldsymbol{Z}_j\boldsymbol{u}_j)_i\right)\right\}\right),\ \ 1\le i\le n,$$

and

$$\beta_j\big|\sigma_\beta^2, f_{2\beta_j} \overset{\text{ind.}}{\sim} N\left(0,\left(\sigma_\beta^2/f_{2\beta_j}\right)\right),\ \ 1\le j\le d_\circ+d_\bullet,$$

then

$$\begin{aligned}\mathfrak{p}(\beta_j\big|\text{rest}) \propto \prod_{i=1}^{n}\Bigg[&\frac{\left(\kappa\exp\left\{-\left(\beta_0+(\boldsymbol{X}\boldsymbol{\beta})_i+\sum_{j=1}^{d_\bullet}(\boldsymbol{Z}_j\boldsymbol{u}_j)_i\right)\right\}\right)^{\kappa} g_i^{\kappa-1}}{\Gamma(\kappa)}\\ &\times\exp\left(-g_i\kappa\exp\left\{-\left(\beta_0+(\boldsymbol{X}\boldsymbol{\beta})_i+\sum_{j=1}^{d_\bullet}(\boldsymbol{Z}_j\boldsymbol{u}_j)_i\right)\right\}\right)\Bigg]\exp\left\{-\frac{\beta_j^2}{2\left(\sigma_\beta^2/f_{2\beta_j}\right)}\right\}\\ \propto\exp&\left\{-\kappa\sum_{i=1}^{n}(\boldsymbol{X}\boldsymbol{\beta})_i\right\}\exp\left\{-\kappa\sum_{i=1}^{n}\left[g_i\exp\left\{-\left(\beta_0+(\boldsymbol{X}\boldsymbol{\beta})_i+\sum_{j=1}^{d_\bullet}(\boldsymbol{Z}_j\boldsymbol{u}_j)_i\right)\right\}\right]\right\}\\ &\times\exp\left\{-\frac{f_{2\beta_j}\beta_j^2}{2\sigma_\beta^2}\right\}.\end{aligned}$$

Using $(\boldsymbol{X}\boldsymbol{\beta})_i = X_{ij}\beta_j + (\boldsymbol{X}_i)_{-j}(\boldsymbol{\beta})_{-j}$, we obtain (3.54) in Section 3.6.3 to be

$$\begin{aligned}\mathfrak{p}(\beta_j\big|\text{rest}) \propto \exp\Bigg\{&\left(-\kappa\mathbf{1}_n^T\boldsymbol{X}_j\right)\beta_j-\frac{f_{2\beta_j}\beta_j^2}{2\sigma_\beta^2}-\mathbf{1}_n^T\exp\Big[-\boldsymbol{X}_j\beta_j+\mathbf{1}_n\log(\kappa)+\log(\boldsymbol{g})\\ &-\left(\mathbf{1}_n\beta_0+(\boldsymbol{X})_{-j}(\boldsymbol{\beta})_{-j}+\sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j\boldsymbol{u}_j\right)\Big]\Bigg\}.\end{aligned}$$

Following the form of (3.26) for logarithm of unnormalised density function, and recalling that (2.7) of Section 2.3.2 to be $\mathbf{1}_n^T\boldsymbol{X}=\mathbf{0}$, then $\mathbf{1}_n^T\boldsymbol{X}_j=0, 1\le j\le d_\circ+d_\bullet$,

we have

$$s_1 = -\kappa \mathbf{1}_n^T \boldsymbol{X}_j = -\exp(\lambda)\, \mathbf{1}_n^T \boldsymbol{X}_j = 0, \;\; s_2 = \sigma_\beta^2 / f_{2\beta_j}, \;\; \boldsymbol{s}_3 = -\boldsymbol{X}_j \text{ and}$$

$$\begin{aligned}\boldsymbol{s}_4 &= \mathbf{1}_n \log(\kappa) + \log(\boldsymbol{g}) - \left(\mathbf{1}_n \beta_0 + (\boldsymbol{X})_{-j} (\boldsymbol{\beta})_{-j} + \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j \boldsymbol{u}_j \right) \\ &= \mathbf{1}_n \lambda + \log(\boldsymbol{g}) - \left(\mathbf{1}_n \beta_0 + (\boldsymbol{X})_{-j} (\boldsymbol{\beta})_{-j} + \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j \boldsymbol{u}_j \right)\end{aligned}$$

for β_j slice sampling for Negative-Binomial responses described in Algorithm 6, $1 \leq j \leq d_\circ + d_\bullet$.

3.H.5 Full Conditional Distributions of u_{jk} for Negative-Binomial Responses

First note that, for $1 \leq j \leq d_\bullet, 1 \leq k \leq K_j$,

$$\mathfrak{p}(u_{jk} \big| \text{rest}) \propto \left\{ \prod_{i=1}^{n} \mathfrak{p}(g_i \big| \beta_0, \boldsymbol{\beta}, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_{d_\bullet}, \kappa) \right\} \mathfrak{p}(u_{jk} \big| \sigma_{uj}^2, f_{2u_j}).$$

Recalling that

$$g_i \big| \beta_0, \boldsymbol{\beta}, \boldsymbol{u}, \kappa \sim \text{Gamma}\left(\kappa, \kappa \exp\left\{ -\left(\beta_0 + (\boldsymbol{X}\boldsymbol{\beta})_i + \sum_{j=1}^{d_\bullet} (\boldsymbol{Z}_j \boldsymbol{u}_j)_i \right) \right\} \right), \;\; 1 \leq i \leq n,$$

and

$$u_{jk} \big| \sigma_{uj}^2, f_{2u_j} \overset{\text{ind.}}{\sim} N\left(0, \left(\sigma_{uj}^2 / f_{2uj}\right)\right), \;\; 1 \leq j \leq d_\circ + d_\bullet, \;\; 1 \leq k \leq K_j,$$

then

$$\begin{aligned}\mathfrak{p}(u_{jk} \big| \text{rest}) \propto \prod_{i=1}^{n} \Bigg[& \frac{\left(\kappa \exp\left\{ -\left(\beta_0 + (\boldsymbol{X}\boldsymbol{\beta})_i + \sum_{i=1}^{d_\bullet} (\boldsymbol{Z}_j \boldsymbol{u}_j)_i \right) \right\} \right)^{\kappa} g_i^{\kappa - 1}}{\Gamma(\kappa)} \\ & \times \exp\left(-g_i \kappa \exp\left\{ -\left(\beta_0 + (\boldsymbol{X}\boldsymbol{\beta})_i + \sum_{i=1}^{d_\bullet} (\boldsymbol{Z}_j \boldsymbol{u}_j)_i \right) \right\} \right) \Bigg] \exp\left\{ -\frac{u_{jk}^2}{2\left(\sigma_{uj}^2 / f_{2uj}\right)} \right\}.\end{aligned}$$

Therefore

$$\mathfrak{p}(u_{jk}|\text{rest}) \propto \exp\left\{-\kappa\sum_{i=1}^{n}\left(\sum_{j=1}^{d_\bullet}\left(\boldsymbol{Z}_j\boldsymbol{u}_j\right)_i\right)\right\}$$
$$\times \exp\left\{-\kappa\sum_{i=1}^{n}\left[g_i\exp\left\{-\left(\beta_0+\left(\boldsymbol{X}\boldsymbol{\beta}\right)_i+\sum_{j=1}^{d_\bullet}\left(\boldsymbol{Z}_j\boldsymbol{u}_j\right)_i\right)\right\}\right]\right\}\exp\left\{-\frac{f_{2uj}u_{jk}^2}{2\sigma_{uj}^2}\right\}.$$

Recalling that $\sum_{j=1}^{d_\bullet}\left(\boldsymbol{Z}_j\boldsymbol{u}_j\right)_i = Z_{ijk}u_{jk}+\sum_{k'\neq k}^{K_j}Z_{ijk'}u_{jk'}+\sum_{j'\neq j}^{d_\bullet}\boldsymbol{Z}_{ij'}\boldsymbol{u}_{j'}$, then

$$\mathfrak{p}(u_{jk}|\text{rest}) \propto \exp\Bigg\{\left(-\kappa\mathbf{1}_n^T\boldsymbol{Z}_{jk}\right)u_{jk}-\frac{f_{2uj}u_{jk}^2}{2\sigma_{uj}^2}-\mathbf{1}_n^T\exp\Big[-\boldsymbol{Z}_{jk}u_{jk}+\mathbf{1}_n\log\left(\kappa\right)+\log\left(\boldsymbol{g}\right)$$
$$-\Big(\mathbf{1}_n\beta_0+\boldsymbol{X}\boldsymbol{\beta}+\sum_{k'\neq k}^{K_j}Z_{jk'}u_{jk'}+\sum_{j'\neq j}^{d_\bullet}\boldsymbol{Z}_{j'}\boldsymbol{u}_{j'}\Big)\Big]\Bigg\}$$

is obtained to be (3.59) in Section 3.6.3. Following the form of (3.26) for logarithm of unnormalised density function, and recalling that (2.8) of Section 2.3.3 to be $\boldsymbol{Z}_j^T\mathbf{1}_n = \mathbf{0}_{K_j}, 1\leq j\leq d_\bullet$, then $\mathbf{1}_n^T\boldsymbol{Z}_{jk}=0$, therefore

$$s_1 = -\kappa\mathbf{1}_n^T\boldsymbol{Z}_{jk} = -\exp\left(\lambda\right)\mathbf{1}_n^T\boldsymbol{Z}_{jk}=0,\ \ s_2=\sigma_{uj}^2/f_{2uj},\ \ \boldsymbol{s}_3=-\boldsymbol{Z}_{jk}\ \text{ and}$$
$$\boldsymbol{s}_4 = \mathbf{1}_n\log\left(\kappa\right)+\log\left(\boldsymbol{g}\right)-\left(\mathbf{1}_n\beta_0+\boldsymbol{X}\boldsymbol{\beta}+\sum_{k'\neq k}^{K_j}Z_{jk'}u_{jk'}+\sum_{j'\neq j}^{d_\bullet}\boldsymbol{Z}_{j'}\boldsymbol{u}_{j'}\right)$$
$$= \mathbf{1}_n\lambda+\log\left(\boldsymbol{g}\right)-\left(\mathbf{1}_n\beta_0+\boldsymbol{X}\boldsymbol{\beta}+\sum_{k'\neq k}^{K_j}Z_{jk'}u_{jk'}+\sum_{j'\neq j}^{d_\bullet}\boldsymbol{Z}_{j'}\boldsymbol{u}_{j'}\right)$$

for u_{jk} slice sampling for Negative-Binomial responses described in Algorithm 6, $1\leq j\leq d_\bullet, 1\leq k\leq K_j$.

3.H.6 Full Conditional Distributions of $f_{2\beta_j}$ for Negative-Binomial Responses

Following the steps in Appendix 3.F.6, we obtain (3.55) in Section 3.6.3 to be

$$f_{2\beta_j}|\text{rest} \overset{\text{ind.}}{\sim} \text{Gamma}\left(1, \frac{\beta_j^2}{2\sigma_\beta^2} + f_{1\beta_j}\right), 1 \leq j \leq d_\circ + d_\bullet.$$

3.H.7 Full Conditional Distributions of $f_{1\beta_j}$ for Negative-Binomial Responses

Following the steps in Appendix 3.F.7, we obtain (3.56) in Section 3.6.3 to be

$$f_{1\beta_j}|\text{rest} \overset{\text{ind.}}{\sim} \text{Gamma}\left(1, 1 + f_{2\beta_j}\right), 1 \leq j \leq d_\circ + d_\bullet.$$

3.H.8 Full Conditional Distributions of f_{2uj} for Negative-Binomial Responses

Following the steps in Appendix 3.F.8, we obtain (3.60) in Section 3.6.3 to be

$$f_{2uj}|\text{rest} \overset{\text{ind.}}{\sim} \text{Gamma}\left(\frac{K_j + 1}{2}, \frac{\left\|\boldsymbol{u}_j\right\|^2}{2\sigma_{uj}^2} + f_{1uj}\right), 1 \leq j \leq d_\bullet.$$

3.H.9 Full Conditional Distributions of f_{1uj} for Negative-Binomial Responses

Following the steps in Appendix 3.F.9, we obtain (3.61) in Section 3.6.3 to be

$$f_{1uj}|\text{rest} \overset{\text{ind.}}{\sim} \text{Gamma}\left(1, 1 + f_{2uj}\right), 1 \leq j \leq d_\bullet.$$

3.H.10 Full Conditional Distributions of σ_β^2 for Negative-Binomial Responses

Following the steps in Appendix 2.B.10, we obtain (3.57) in Section 3.6.3 to be

$$\sigma_\beta^2|\text{rest} \sim \text{Inverse-Gamma}\left(\frac{d_\circ + d_\bullet + 1}{2}, \frac{\boldsymbol{\beta}^T \text{diag}(\boldsymbol{f}_{2\beta})\boldsymbol{\beta}}{2} + \frac{1}{a_\beta}\right)$$

3.H.11 Full Conditional Distributions of a_β for Negative-Binomial Responses

Following the steps in Appendix 2.B.11, we obtain (3.58) in Section 3.6.3 to be

$$a_\beta|\text{rest} \sim \text{Inverse-Gamma}\left(1, \frac{1}{\sigma_\beta^2} + \frac{1}{s_\beta^2}\right).$$

3.H.12 Full Conditional Distributions of σ_{uj}^2 for Negative-Binomial Responses

Following the steps in Appendix 2.B.12, we obtain (3.62) in Section 3.6.3 to be

$$\sigma_{uj}^2|\text{rest} \sim \text{Inverse-Gamma}\left(\frac{K_j+1}{2}, \frac{f_{2uj}\|\boldsymbol{u}_j\|^2}{2} + \frac{1}{a_{uj}}\right), 1 \leq j \leq d_\bullet.$$

3.H.13 Full Conditional Distributions of a_{uj} for Negative-Binomial Responses

Following the steps in Appendix 2.B.13, we obtain (3.63) in Section 3.6.3 to be

$$a_{uj}|\text{rest} \sim \text{Inverse-Gamma}\left(1, \frac{1}{\sigma_{uj}^2} + \frac{1}{s_u^2}\right), 1 \leq j \leq d_\bullet.$$

Chapter 4

Mean Field Variational Bayes with Laplace-Zero Priors

Declaration The method and results presented in this chapter are presented in the paper *Bayesian Generalized Additive Model Selection Including a Fast Variational Option* (He and Wand 2024). I certify that this publication was a direct result of my research towards this PhD, under the supervision of my principal supervisor Prof Matt Wand, and the reproduction in this thesis does not breach any copyright regulations.

4.1 Introduction

Variational methods (e.g. Ormerod and Wand, 2010) have utilized in many areas such as machine learning and pattern recognition. Since the early 2000s there has been an increasing recognition of their use in mainstream Statistics (e.g. Titterington, 2004).

Consider a generic Bayesian model with observed data vector $\boldsymbol{y}$ and continuous parameter vector $\boldsymbol{\theta}$ over the parameter space $\boldsymbol{\Theta}$. The essence of variational inference is to approximate the posterior density function $\mathfrak{p}(\boldsymbol{\theta}|\boldsymbol{y})$ with an *approximating density function* $\mathfrak{q}(\boldsymbol{\theta})$. To make this approximation as close as possible, we search over $q \in \mathcal{Q}$, for some set $\mathcal{Q}$ of density functions, to find a particular density function with the minimum Kullback-Liebler (KL) distance/divergence from the actual posterior

$$\mathfrak{q}^*(\boldsymbol{\theta}) = \underset{\mathfrak{q}\in\mathcal{Q}}{\operatorname{argmin}} \ \operatorname{KL}\left\{\mathfrak{q}(\boldsymbol{\theta}|\boldsymbol{y}) \,\middle\|\, \mathfrak{p}(\boldsymbol{\theta}|\boldsymbol{y})\right\} \tag{4.1}$$

where

$$\operatorname{KL}\left\{\mathfrak{q}(\boldsymbol{\theta}|\boldsymbol{y}) \,\middle\|\, \mathfrak{p}(\boldsymbol{\theta}|\boldsymbol{y})\right\} = \int_{\boldsymbol{\Theta}} \mathfrak{q}(\boldsymbol{\theta}) \log\left\{\frac{\mathfrak{q}(\boldsymbol{\theta})}{\mathfrak{p}(\boldsymbol{\theta}|\boldsymbol{y})}\right\} d\boldsymbol{\theta}.$$

We denote d as the number of predictors, with $1 \leq j \leq d$. The parametric forms of the $\mathfrak{q}$-density functions are determined from the form of the Bayesian model and the minimum Kullback-Leibler optimisation problem, determined through (4.1). For instance, Chapter 10 of Bishop (2006) explains this aspect of mean field variational Bayes. The optimal variational factor $\mathfrak{q}_j^*(\theta_j)$ is proportional to the exponential of the expectation (under the other factors), of the the joint log-density. Therefore, an expression for $\mathfrak{q}_j^*(\theta_j)$ is:

$$\mathfrak{q}_j^*(\theta_j) \propto \exp\left[E_{\mathfrak{q}(-\boldsymbol{\theta}_j)}\left\{\log \mathfrak{p}(\boldsymbol{\theta}, \boldsymbol{y})\right\}\right], \ \ 1 \leq j \leq d.$$

If $\mathfrak{p}(\boldsymbol{\theta}, \boldsymbol{y})$ is conditionally conjugate, then $\mathfrak{q}_j(\theta_j)$ ends up in the same exponential family as the full conditional $\mathfrak{p}\left(\theta_j \middle| \boldsymbol{\theta}_{-j}, \boldsymbol{y}\right)$. For instance, if $\mathfrak{p}(\boldsymbol{\theta}, \boldsymbol{y})$ is Gaussian in θ_j, the optimal $\mathfrak{q}_j(\theta_j)$ will also be Gaussian.

A common approach to aid tractability is to impose a product density restriction on $\mathfrak{q}(\boldsymbol{\theta})$, without any restriction on the functional form of the factors (Bishop, 2006):

$$\mathcal{Q} = \left\{\mathfrak{q}(\boldsymbol{\theta}) : \mathfrak{q}(\boldsymbol{\theta}) = \prod_{j=1}^{d} \mathfrak{q}_j(\theta_j) \text{ for some partition } \{\theta_1, \ldots, \theta_d\} \text{ of } \boldsymbol{\theta}\right\} \tag{4.2}$$

The resultant $\mathfrak{q}^*(\boldsymbol{\theta})$ is known as the mean field variational Bayes approximation (MFVB) to the actual posterior, and we refer to it as the *optimal* $\mathfrak{q}$*-density.*

In this chapter, we work with the mean field variational Bayes (MFVB) approach, to overcome the heavy computational cost of the Markov chain Monte Carlo (MCMC) approach described in Chapter 2, for generalized additive model selection for Gaussian and binary responses. MFVB is a faster, deterministic, but less accurate fitting and inference alternative (e.g. Wainwright and Jordan, 2008). Same as Chapter 2, throughout Chapter 4, $\boldsymbol{y}$ vector, $\boldsymbol{X}$ matrix and $\boldsymbol{Z}$ matrix are all fixed.

We introduce the product density restrictions on approximating density functions in Section 4.2.1. Benefiting greatly on the Gibbsian full conditional distributions of the parameters, the $\mathfrak{q}$-density approximation and the implementation for MFVB involve straightforward algebraic calculations, described in Section 4.2.2. Moreover, MFVB algorithms are often very simple to implement. The MFVB approach is implemented as the R package gamselBayes, described in Algorithm 7 of Section 4.2.2. We also work out the marginal log-likelihood approximation, which is useful for monitoring convergence, described in Section 4.2.3.

In practice, either the Markov chain Monte Carlo samples or the MFVB $\mathfrak{q}$-densities can be used for approximate posterior-based decision making. Based on the approximation of relevant parameters, we set up the selection strategies for MFVB using the threshold parameter, to classify variable effect type being nonlinear, linear or zero, described in Section 4.3.1 and Section 4.3.2. Simulation studies aim at finding a good default value of sparsity threshold parameter for MFVB are described in Section 4.3.3.

We demonstrate that the MFVB algorithms achieve a good level of accuracy and improved speed with simulation studies result given in Chapter 5.

The additive structure of generalized additive model (GAM) described in this chapter is given in Section 2.3. The independence assumptions of GAM is given in (2.28) for Gaussian responses and Section 2.4.3 for binary responses. The additive part uses splines for non-linear effects. The predictors include binary predictors and continuous predictors. The number of binary predictors is $d_\circ$. The number of continuous predictors is $d_\bullet$.

For the generalized additive model described in Section 4.2, the sample size is n. The response $\boldsymbol{y}$ is a $n \times 1$ vector. The linear design matrix $\boldsymbol{X}$ is a $n \times (d_\circ + d_\bullet)$ matrix. The $\boldsymbol{Z}_j$ matrix containing the canonical Demmler-Reinsch basis is a $n \times K_j$ matrix, with $1 \leq j \leq d_\bullet$, K_j as the number of basis functions for the j-th continuous

predictor with default value of 12.

Our simulation study suggests a good default threshold value τ as 0.1 for MFVB given in Section 4.3.3. The data were generated with 30 continuous predictors, 10 having "true" linear effects, 10 having "true" non-linear effects and 10 having "true" zero effects. The sample size varied as $n \in \{500, 1000, 2000\}$. The error standard deviation varied as $\sigma_\varepsilon \in \{0.25, 0.5, 1, 2\}$. The threshold varied as $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. For each combination of sample size and error standard deviation, we ran simulation on 100 independently generated data sets.

We applied Algorithm 7 to two real-world datasets – the car auction and the Sydney real estate dataset given in Section 4.4, to demonstrate the practical applicability of the approach. The car auction dataset used in Section 4.4.1 contains $72,983$ observations and 49 predictors. The Sydney real estate dataset used in Section 4.4.2 contains $37,676$ observations and 39 predictors.

We conducted a simulation study on a prediction accuracy evaluation metric given in Figure 4.2 of Section 4.3.3. The threshold varied as $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. The sample size varied as $n \in \{500, 1000, 2000\}$. The error standard deviation varied as $\sigma_\varepsilon \in \{0.25, 0.5, 1, 2\}$. Moreover, hyperparameters sensitivity checks are given in Section 5.4.

4.2 Mean Field Variational Bayesian Inference for Laplace-Zero Priors

4.2.1 Product Density Restrictions

Consider the Bayesian inference for the parameters in (2.28) of Section 2.4.2 and in Figure 2.1 of Section 2.5.3 for Gaussian responses. The essence of our MFVB approach involves approximation of the full joint posterior density function of the form

$$\begin{aligned}&\mathfrak{p}(\beta_0, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_u, \widetilde{\boldsymbol{u}}, \boldsymbol{b}_\beta, \sigma_\beta^2, a_\beta, \boldsymbol{b}_u, \boldsymbol{\sigma}_u^2, \boldsymbol{a}_u, \sigma_\varepsilon^2, a_\varepsilon \big| \boldsymbol{y}) \\ &\quad \approx \mathfrak{q}(\beta_0, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_u, \widetilde{\boldsymbol{u}}, \boldsymbol{b}_\beta, \sigma_\beta^2, a_\beta, \boldsymbol{b}_u, \boldsymbol{\sigma}_u^2, \boldsymbol{a}_u, \sigma_\varepsilon^2, a_\varepsilon). \qquad (4.3)\end{aligned}$$

(4.3) is subject to

$$\begin{aligned}&\mathfrak{q}(\beta_0, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_u, \widetilde{\boldsymbol{u}}, \boldsymbol{b}_\beta, \sigma_\beta^2, a_\beta, \boldsymbol{b}_u, \boldsymbol{\sigma}_u^2, \boldsymbol{a}_u, \sigma_\varepsilon^2, a_\varepsilon)\\ &\quad = \mathfrak{q}(\beta_0)\mathfrak{q}(\boldsymbol{\gamma}_\beta)\mathfrak{q}(\widetilde{\boldsymbol{\beta}})\mathfrak{q}(\boldsymbol{\gamma}_u)\mathfrak{q}(\widetilde{\boldsymbol{u}})\mathfrak{q}(\boldsymbol{b}_\beta)\mathfrak{q}(\sigma_\beta^2)\mathfrak{q}(a_\beta)\mathfrak{q}(\boldsymbol{b}_u)\mathfrak{q}(\boldsymbol{\sigma}_u^2)\mathfrak{q}(\boldsymbol{a}_u)\mathfrak{q}(\sigma_\varepsilon^2)\mathfrak{q}(a_\varepsilon). \quad (4.4)\end{aligned}$$

Therefore, the mean field approximation of the joint posterior density function of the model parameters to (2.28) in Section 2.4.2 for Gaussian response is

$$\begin{aligned}&\mathfrak{p}(\beta_0, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_u, \widetilde{\boldsymbol{u}}, \boldsymbol{b}_\beta, \sigma_\beta^2, a_\beta, \boldsymbol{b}_u, \boldsymbol{\sigma}_u^2, \boldsymbol{a}_u, \sigma_\varepsilon^2, a_\varepsilon \big| \boldsymbol{y})\\ &\quad \approx \mathfrak{q}(\beta_0)\mathfrak{q}(\boldsymbol{\gamma}_\beta)\mathfrak{q}(\widetilde{\boldsymbol{\beta}})\mathfrak{q}(\boldsymbol{\gamma}_u)\mathfrak{q}(\widetilde{\boldsymbol{u}})\mathfrak{q}(\boldsymbol{b}_\beta)\mathfrak{q}(\sigma_\beta^2)\mathfrak{q}(a_\beta)\mathfrak{q}(\boldsymbol{b}_u)\mathfrak{q}(\boldsymbol{\sigma}_u^2)\mathfrak{q}(\boldsymbol{a}_u)\mathfrak{q}(\sigma_\varepsilon^2)\mathfrak{q}(a_\varepsilon)\\ &\quad \approx \mathfrak{q}(\beta_0)\left\{\prod_{j=1}^{d_\circ+d_\bullet}\mathfrak{q}(\gamma_{\beta j})\right\}\mathfrak{q}(\widetilde{\boldsymbol{\beta}})\left\{\prod_{j=1}^{d_\bullet}\mathfrak{q}(\gamma_{uj})\right\}\left\{\prod_{j=1}^{d_\bullet}\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)\right\}\left\{\prod_{j=1}^{d_\circ+d_\bullet}\mathfrak{q}(b_{\beta j})\right\}\\ &\qquad \times\mathfrak{q}(\sigma_\beta^2)\mathfrak{q}(a_\beta)\left\{\prod_{j=1}^{d_\bullet}\mathfrak{q}(b_{uj})\right\}\left\{\prod_{j=1}^{d_\bullet}\mathfrak{q}(\sigma_{uj}^2)\right\}\left\{\prod_{j=1}^{d_\bullet}\mathfrak{q}(a_{uj})\right\}\mathfrak{q}(\sigma_\varepsilon^2)\mathfrak{q}(a_\varepsilon).\end{aligned} \quad (4.5)$$

Restriction (4.5) involves the *induced product results* (Bishop, 2006, Section 10.2.5), which contains additional factorization in the $\mathfrak{q}$-density product restriction form.

A less stringent product restriction as (4.6), which will not scale well to high $d_\bullet$, but may have accuracy advantages. (4.6) keeps all of the basis function coefficients together is

$$\begin{aligned}&\mathfrak{p}(\beta_0, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_u, \widetilde{\boldsymbol{u}}, \boldsymbol{b}_\beta, \sigma_\beta^2, a_\beta, \boldsymbol{b}_u, \boldsymbol{\sigma}_u^2, \boldsymbol{a}_u, \sigma_\varepsilon^2, a_\varepsilon \big| \boldsymbol{y})\\ &\quad \approx \mathfrak{q}(\beta_0, \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{u}})\mathfrak{q}(\boldsymbol{\gamma}_\beta)\mathfrak{q}(\boldsymbol{\gamma}_u)\mathfrak{q}(\boldsymbol{b}_\beta)\mathfrak{q}(\sigma_\beta^2)\mathfrak{q}(a_\beta)\mathfrak{q}(\boldsymbol{b}_u)\mathfrak{q}(\boldsymbol{\sigma}_u^2)\mathfrak{q}(\boldsymbol{a}_u)\mathfrak{q}(\sigma_\varepsilon^2)\mathfrak{q}(a_\varepsilon).\end{aligned} \quad (4.6)$$

There are numerous stringency levels of the product restriction. The choice involves trade-offs concerning tractability, accuracy and speed. In this thesis, we will use the simpler mean field variational Bayes scheme based on (4.5).

The optimal $\mathfrak{q}$-densities, which are the mean field variational Bayes approximation to the actual posterior, are chosen to minimize Kullback-Leibler divergence between the right-hand side of (4.4) and the full joint posterior density function, which is the left-hand side of (4.3) (e.g. Wainwright and Jordan, 2008). The optimal $\mathfrak{q}$-density forms can be expressed in terms of the full conditional density functions as given by equation (6) of Ormerod and Wand (2010). Since all the full conditional distribution of parameters having standard forms described in Section 2.5.3, the optimal $\mathfrak{q}$-density functions are relatively simple and the coordinate ascent updates

have closed forms.

4.2.2 MFVB Practical Fitting for Laplace-Zero Priors

Based on full conditional distributions of parameters for Markov chain Monte Carlo given in Section 2.5.3, for the Gaussian and binary responses models, we obtain q-density approximation of parameters for mean field variational Bayes in this subsection. Some examples of the resulting optimal q-density forms are:

$$
\begin{aligned}
&\mathfrak{q}(\beta_0) \text{ has a } N(\mu_{\mathfrak{q}(\beta_0)}, \sigma^2_{\mathfrak{q}(\beta_0)}) \text{ density function,} \\
&\mathfrak{q}(\widetilde{\boldsymbol{\beta}}) \text{ has a } N(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}, \boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}) \text{ density function,} \\
&\mathfrak{q}(\widetilde{\boldsymbol{u}}_j) \text{ has a } N(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}, \boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}) \text{ density function, } \; 1 \leq j \leq d_\bullet, \\
&\mathfrak{q}(\gamma_{\beta j}) \text{ has a Bernoulli } \left(\frac{1}{1 + \exp\left\{ E_{\mathfrak{q}(\eta_{\beta j})} \right\}} \right) \text{ probability mass function,} \\
&\qquad 1 \leq j \leq d_\circ + d_\bullet, \\
&\mathfrak{q}(\gamma_{uj}) \text{ has a Bernoulli } \left(\frac{1}{1 + \exp\left\{ E_{\mathfrak{q}(\eta_{uj})} \right\}} \right) \text{ probability mass function,} \\
&\qquad 1 \leq j \leq d_\bullet, \\
&\mathfrak{q}(\boldsymbol{b}_\beta) \text{is a product of Inverse-Gaussian } \left(\mu_{\mathfrak{q}(b_{\beta j})}, 1 \right) \text{ density functions,} \\
&\qquad 1 \leq j \leq d_\circ + d_\bullet, \\
&\mathfrak{q}(\boldsymbol{b}_u) \text{ is a product of Inverse-Gaussian } \left(\mu_{\mathfrak{q}(b_{uj})}, 1 \right) \text{ denstiy functions,} \\
&\qquad 1 \leq j \leq d_\bullet, \\
&\mathfrak{q}(\sigma^2_\varepsilon) \text{ has an Inverse-Gamma}(\kappa_{\mathfrak{q}(\sigma^2_\varepsilon)}, \lambda_{\mathfrak{q}(\sigma^2_\varepsilon)}) \text{ density function,} \\
&\mathfrak{q}(\sigma^2_\beta) \text{ has an Inverse-Gamma } \left(\kappa_{\mathfrak{q}(\sigma^2_\beta)}, \lambda_{\mathfrak{q}(\sigma^2_\beta)} \right) \text{ density function,} \\
&\mathfrak{q}(\sigma^2_{uj}) \text{ has an Inverse-Gamma } \left(\kappa_{\mathfrak{q}(\sigma^2_{uj})}, \lambda_{\mathfrak{q}(\sigma^2_{uj})} \right) \text{ density function, } 1 \leq j \leq d_\bullet.
\end{aligned}
\tag{4.7}
$$

For example, the optimal Inverse Gamma shape parameter $\kappa_{\mathfrak{q}(\sigma^2_\varepsilon)}$ has explicit solution $\frac{1}{2}(n+1)$. However, the equations for the optimal values of $\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}, \boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}$ and

$\lambda_{\mathfrak{q}(\sigma_\varepsilon^2)}$ are interdependent and iteration is required to obtain their optimal values.

The Bayesian graphical model for wavelet regression described in Section 3 of Wand and Ormerod (2011) is similar in nature to the generalized additive selection model (2.28). Hence, the relevant details on the requisite mean field variational Bayes calculations for (2.28) can be gleaned from the $\mathfrak{q}$-density derivations given in Appendix D of Wand and Ormerod (2011). The $\mathfrak{q}$-density approximations of parameters for MFVB are listed in (4.8)–(4.31), with derivation details given in Appendix 4.A. The results are:

$\mathfrak{q}(\beta_0)$ has a $N\left(\mu_{\mathfrak{q}(\beta_0)}, \sigma^2_{\mathfrak{q}(\beta_0)}\right)$ density function where

$$\sigma^2_{\mathfrak{q}(\beta_0)} = \left(n\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)} + \frac{1}{\sigma^2_{\beta_0}}\right)^{-1}, \tag{4.8}$$

$$\mu_{\mathfrak{q}(\beta_0)} = \sigma^2_{\mathfrak{q}(\beta_0)}\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\mathbf{1}_n^T\boldsymbol{y}, \tag{4.9}$$

$\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)$ has a $N\left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}, \boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right)$ density function where

$$\boldsymbol{\Omega}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} = \text{diag}\left\{\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \left(\mathbf{1} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}\right)\right\} + \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}\boldsymbol{\mu}^T_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}, \tag{4.10}$$

$$\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} = \left\{\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\left[\boldsymbol{\Omega}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \left(\boldsymbol{X}^T\boldsymbol{X}\right)\right] + \mu_{\mathfrak{q}\left(1/\sigma^2_\beta\right)}\text{diag}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{b}_\beta)}\right)\right\}^{-1}, \tag{4.11}$$

$$\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} = \mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \left[\boldsymbol{X}^T\left(\boldsymbol{y} - \sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j\left(\mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right)\right)\right]\right), \tag{4.12}$$

$\mathfrak{q}\left(b_{\beta_1}, \ldots, b_{\beta_{(d_\circ + d_\bullet)}}\right)$ is a product of

Inverse-Gaussian $\left(\left\{\mu_{\mathfrak{q}\left(1/\sigma^2_\beta\right)}\left(\mu^2_{\mathfrak{q}\left(\widetilde{\beta}_j\right)} + \sigma^2_{\mathfrak{q}\left(\widetilde{\beta}_j\right)}\right)\right\}^{-1/2}, 1\right)$ density functions

over $1 \le j \le d_\circ + d_\bullet$, The $\mathfrak{q}$-density means are

$$\mu_{\mathfrak{q}\left(b_{\beta j}\right)} = \left\{\mu_{\mathfrak{q}\left(1/\sigma^2_\beta\right)}\left(\mu^2_{\mathfrak{q}\left(\widetilde{\beta}_j\right)} + \sigma^2_{\mathfrak{q}\left(\widetilde{\beta}_j\right)}\right)\right\}^{-1/2}, \tag{4.13}$$

$\mathfrak{q}\left(\sigma^2_\beta\right)$ has an Inverse-Gamma $\left(\kappa_{\mathfrak{q}\left(\sigma^2_\beta\right)}, \lambda_{\mathfrak{q}\left(\sigma^2_\beta\right)}\right)$ density function where

$$\kappa_{\mathfrak{q}\left(\sigma^2_\beta\right)} = \frac{d_\circ + d_\bullet + 1}{2}, \tag{4.14}$$

$$\lambda_{\mathfrak{q}\left(\sigma^2_\beta\right)} = \frac{1}{2}\boldsymbol{\mu}^T_{\mathfrak{q}\left(\boldsymbol{b}_\beta\right)}\left\{\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} + \text{diagonal}\left(\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right)\right\} + \mu_{\mathfrak{q}\left(1/a_\beta\right)}, \tag{4.15}$$

$\mathfrak{q}(a_\beta)$ has an Inverse-Gamma $\left(\kappa_{\mathfrak{q}(a_\beta)}, \lambda_{\mathfrak{q}(a_\beta)}\right)$ density function where

$$\kappa_{\mathfrak{q}(a_\beta)} = 1, \tag{4.16}$$

$$\lambda_{\mathfrak{q}(a_\beta)} = \mu_{\mathfrak{q}(1/\sigma_\beta^2)} + 1/s_\beta^2, \tag{4.17}$$

$\mathfrak{q}(\gamma_{\beta j})$ is the Bernoulli $\left(\dfrac{1}{1+\exp\left\{-E_{\mathfrak{q}}\left(\eta_{\beta_j}\right)\right\}}\right)$ probability mass function for each $1 \le j \le d_\circ + d_\bullet$, where

$$\begin{aligned} E_{\mathfrak{q}}\left(\eta_{\beta_j}\right) = \text{logit}\left(\rho_\beta\right) - \frac{1}{2}\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\Bigg\{ & \mathfrak{q}\left(\mu^2_{\mathfrak{q}(\widetilde{\beta}_j)} + \sigma^2_{\mathfrak{q}(\widetilde{\beta}_j)}\right)\boldsymbol{X}_j^T\boldsymbol{X}_j \\ & - 2\Bigg[\mu_{\mathfrak{q}(\widetilde{\beta}_j)}\Big\{\boldsymbol{X}_j^T\boldsymbol{y} - \sum_{j'=1}^{d_\bullet}\boldsymbol{X}_j^T\boldsymbol{Z}_{j'}\left(\mu_{\mathfrak{q}(\gamma_{uj'})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_{j'})}\right)\Big\} \\ & - \boldsymbol{X}_j^T\boldsymbol{X}_{-j}\Big\{\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}\right)_{-j} \odot \left(\left(\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right)_{-j,j} + \mu_{\mathfrak{q}(\widetilde{\beta}_j)}\left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right)_{-j}\right)\Big\}\Bigg]\Bigg\}, \end{aligned} \tag{4.18}$$

$\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)$ has a $N\left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}, \boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right)$ density function, $1 \le j \le d_\bullet$,

where $\boldsymbol{\omega}_{\boldsymbol{Z}_j} \equiv \text{diagonal}(\boldsymbol{Z}_j^T\boldsymbol{Z}_j)$,

$$\begin{aligned} \boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} &= \text{diag}\left(\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right) \\ &= \text{diag}\left[\mathbf{1}_{K_j} \Big/ \left\{\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\omega}_{\boldsymbol{Z}_j} + \mu_{\mathfrak{q}\left(1/\sigma^2_{uj}\right)}\mu_{\mathfrak{q}\left(b_{uj}\right)}\mathbf{1}_{K_j}\right\}\right], \end{aligned} \tag{4.19}$$

$$\begin{aligned} \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} = \mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\Bigg[\mu_{\mathfrak{q}(\gamma_{uj})}\Bigg\{ & \boldsymbol{Z}_j^T\boldsymbol{y} - \boldsymbol{Z}_j^T\boldsymbol{X}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right) \\ & - \sum_{j' \ne j}^{d_\bullet}\boldsymbol{Z}_j^T\boldsymbol{Z}_{j'}\left(\mu_{\mathfrak{q}(\boldsymbol{\gamma}_{uj'})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_{j'})}\right)\Bigg\}\Bigg] \odot \boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}, \end{aligned} \tag{4.20}$$

$\mathfrak{q}\left(b_{u_1}, \ldots, b_{u_{(d_\bullet)}}\right)$ is a product of Inverse-Gaussian

$\left(\left\{\mu_{\mathfrak{q}\left(1/\sigma^2_{uj}\right)}\left[\left\|\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right\|^2 + \mathbf{1}_{K_j}^T\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right]\right\}^{-1/2}, 1\right)$ density functions

over $1 \le j \le d_\bullet$. The $\mathfrak{q}$-density means are

$$\mu_{\mathfrak{q}\left(b_{uj}\right)} = \left\{\mu_{\mathfrak{q}\left(1/\sigma^2_{uj}\right)}\left[\left\|\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right\|^2 + \mathbf{1}^T_{K_j}\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right]\right\}^{-1/2}, \tag{4.21}$$

$\mathfrak{q}\left(\sigma^2_{uj}\right)$ has an Inverse-Gamma $\left(\kappa_{\mathfrak{q}\left(\sigma^2_{uj}\right)}, \lambda_{\mathfrak{q}\left(\sigma^2_{uj}\right)}\right)$ density function, $1 \leq j \leq d_\bullet$,

where $\kappa_{\mathfrak{q}\left(\sigma^2_{uj}\right)} = \dfrac{K_j + 1}{2}$, (4.22)

$$\lambda_{\mathfrak{q}\left(\sigma^2_{uj}\right)} = \frac{1}{2}\mu_{\mathfrak{q}(b_{uj})}\left(\left\|\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right\|^2 + \mathbf{1}^T_{K_j}\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right) + \mu_{\mathfrak{q}(1/a_{uj})}, \tag{4.23}$$

$\mathfrak{q}\left(a_{uj}\right) \sim$ has an Inverse-Gamma $\left(\kappa_{\mathfrak{q}(a_{uj})}, \lambda_{\mathfrak{q}(a_{uj})}\right)$ density function , $1 \leq j \leq d_\bullet$,

where $\kappa_{\mathfrak{q}\left(\sigma^2_{uj}\right)} = 1$, (4.24)

$$\lambda_{\mathfrak{q}\left(\sigma^2_{uj}\right)} = \mu_{\mathfrak{q}\left(1/\sigma^2_{uj}\right)} + 1/s^2_u, \tag{4.25}$$

and

$\mathfrak{q}(\gamma_{uj})$ is the Bernoulli $\left(\dfrac{1}{1+\exp\left(-E_{\mathfrak{q}}\left(\eta_{uj}\right)\right)}\right)$ probability mass function for each

$1 \leq j \leq d_\bullet$, where

$$E_{\mathfrak{q}}\left(\eta_{uj}\right) = \text{logit}\left(\rho_u\right) - \frac{1}{2}\mu_{\mathfrak{q}(1/\sigma^2_\varepsilon)}\left\{\boldsymbol{\omega}^T_{\boldsymbol{Z}_j}\left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} + \boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right)\right.$$
$$\left. - 2\boldsymbol{\mu}^T_{\mathfrak{q}(\widetilde{u}_j)}\boldsymbol{Z}^T_j\left[\boldsymbol{y} - \boldsymbol{X}\left(\boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{\gamma}_\beta\right)} \odot \boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\right) - \sum_{j' \neq j}^{d_\bullet}\boldsymbol{Z}_{j'}\left(\mu_{\mathfrak{q}\left(\gamma_{uj'}\right)}\boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{u}}_{j'}\right)}\right)\right]\right\}. \tag{4.26}$$

If $\boldsymbol{y}$ contains Gaussian responses, $\mathfrak{q}\left(\sigma^2_\varepsilon\right)$ has an Inverse-Gamma $\left(\kappa_{\mathfrak{q}(\sigma^2_\varepsilon)}, \lambda_{\mathfrak{q}(\sigma^2_\varepsilon)}\right)$

density function where $\kappa_{\mathfrak{q}(\sigma^2_\varepsilon)} = \dfrac{n+1}{2}$, (4.27)

$$\lambda_{\mathfrak{q}(\sigma^2_\varepsilon)} = \mu_{\mathfrak{q}(1/a_\varepsilon)} + \frac{1}{2}\left\|\boldsymbol{y} - \mathbf{1}^T_n\mu_{\mathfrak{q}(\beta_0)} - \boldsymbol{X}\left(\boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{\gamma}_\beta\right)} \odot \boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\right) - \sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j\left(\mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right)\right\|^2$$
$$+ \frac{1}{2}n\sigma^2_{\mathfrak{q}(\beta_0)} + \frac{1}{2}\text{tr}\left\{\boldsymbol{X}^T\boldsymbol{X}\left(\boldsymbol{\Omega}_{\mathfrak{q}\left(\boldsymbol{\gamma}_\beta\right)} \odot \left\{\boldsymbol{\Sigma}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)} + \boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\boldsymbol{\mu}^T_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\right\}\right)\right\}$$

$$- \frac{1}{2}\text{tr}\left\{ \boldsymbol{X}^T\boldsymbol{X}\left\{\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right\}\left\{\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right\}^T\right\}$$

$$+ \frac{1}{2}\sum_{j=1}^{d_\bullet}\boldsymbol{\omega}_{\boldsymbol{Z}_j}^T\left(\mu_{\mathfrak{q}(\gamma_{uj})}\left[\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} + \left\{1-\mu_{\mathfrak{q}(\gamma_{uj})}\right\}\left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right)\right]\right) \quad (4.28)$$

and

$\mathfrak{q}\left(a_\varepsilon\right)$ has an Inverse-Gamma $\left(\kappa_{\mathfrak{q}(a_\varepsilon)}, \lambda_{\mathfrak{q}(a_\varepsilon)}\right)$ density function where

$$\kappa_{\mathfrak{q}(a_\varepsilon)} = 1, \quad (4.29)$$

$$\lambda_{\mathfrak{q}(a_\varepsilon)} = \mu_{\mathfrak{q}(1/\sigma^2_\varepsilon)} + 1\big/s^2_\varepsilon. \quad (4.30)$$

If $\boldsymbol{y}$ contains binary responses, for $1 \leq i \leq n,\ \ y_i = 1$ or $y_i = 0$,

$\mathfrak{q}\left(\boldsymbol{c}_i\right)$ has a Truncated-Normal$_+\left(\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{c})}\right)_i, 1\right)$ density function where

$$\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{c})}\right)_i = (\boldsymbol{\mu}_{\mathfrak{q}})_i + (2y_i - 1)\,\frac{\phi\left((2y_i-1)(\boldsymbol{\mu}_{\mathfrak{q}})_i\right)}{\Phi\left((2y_i-1)(\boldsymbol{\mu}_{\mathfrak{q}})_i\right)}, \quad (4.31)$$

$$(\boldsymbol{\mu}_{\mathfrak{q}})_i = \mu_{\mathfrak{q}(\beta_0)} + \left[\boldsymbol{X}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right) + \sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j\left(\mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right)\right]_i.$$

Using the optimal $\mathfrak{q}$-density results listed in (4.8)–(4.31), we can now assemble the mean field variational Bayes algorithm, given in Algorithm 7, implemented within the R package gamselBayes, to minimize the Kullback-Leibler distance $\mathfrak{q}$-density parameters, with notation used in Section 4.2.2 for the other $\mathfrak{q}$-density parameters.

We proceed by initialising each of the $\mathfrak{q}$-density parameters and updating each parameter successively using the current estimates of the other parameters. At the end of each iteration, an updated value of the lower bound on the marginal log-likelihood is computed, and the algorithm is iterated until convergence of the lower bound on the marginal log-likelihood to its maximum. The approximate marginal log-likelihood, also known as the evidence lower bound, is given in the next subsection.

4.2.3 Variational Marginal Log-Likelihood Approximation

A final aspect of Algorithm 7 in Section 4.2.2 is determination of good stopping criteria for the coordinate ascent scheme. As is common in the mean field variational

Bayes literature, we monitor relative increases in the approximate marginal log-likelihood, which we denote by $\log \underline{p}(\boldsymbol{y}; \mathfrak{q})$. Minimizing the Kullback-Liebler divergence is equivalent to maximizing the lower bound on the marginal log-likelihood $\log \underline{p}(\boldsymbol{y}; \mathfrak{q})$.

Appendix 4.B contains the explicit expression for the approximate marginal log-likelihood for Chapter 2 models under product restriction (4.5). By computing the log-likelihood after each iteration, we obtained its corresponding relative error. Refer to R package gamselBayes, the setting of the tolerance of relative error is $\varepsilon_{\text{toler.}} = 10^{-8}$. As shown in Algorithm 7 of Section 4.2.2, we checked the relative error after each iteration. Once the relative error of log-likelihood is lower than the tolerance level, the iteration stopped. The variational marginal log-likelihood approximation is implemented as the R package gamselBayes.

4.3 Model Selection Strategies with Laplace-Zero priors for Mean Field Variational Bayes

After obtaining the $\mathfrak{q}$-density approximation in Section 4.2.2 for mean field variational Bayes, we need to set up the model selection strategies to decide whether the effect type of a predictor is zero, linear or non-linear.

4.3.1 Deciding Between an Effect Being Zero or Linear

For Gaussian and binary response generalized additive models described in Section 2.4, for the MFVB approach, our "effect is zero" rule between zero effect and linear effect is as in (2.52) of Section 2.6.1 for Markov chain Monte Carlo. For mean field variational Bayes approach, we have $E\left(\gamma_{\beta_j} \middle| \boldsymbol{y}\right) \equiv \mu_{\mathfrak{q}\left(\gamma_{\beta_j}\right)}$ in Algorithm 7, $1 \leq j \leq d_\circ + d_\bullet$. Simulation studies of deciding default values for sparsity threshold parameter τ for mean field variational Bayes described in Section 4.3.3.

4.3.2 Deciding Between an Effect Being Zero, Linear or Non-Linear

For Gaussian and binary response generalized additive models described in Section 2.4, for the MFVB approach, our "effect is zero" rule for MFVB between an effect being zero or linear or non-linear is as in (2.54) of Section 2.6.2. For mean field variational Bayes approach, we have $E\left(\gamma_{\beta_j} \middle| \boldsymbol{y}\right) \equiv \mu_{\mathfrak{q}\left(\gamma_{\beta_j}\right)}$ in Algorithm 7, $1 \leq j \leq d_\circ + d_\bullet$,

and $E\left(\gamma_{uj}\middle|\boldsymbol{y}\right) \equiv \mu_{\mathfrak{q}(\gamma_{uj})}$ in Algorithm 7, $1 \leq j \leq d_\bullet$. Simulation studies of deciding default values for sparsity threshold parameter τ for mean field variational Bayes described in Section 4.3.3.

4.3.3 Choice of Default Value of Sparsity Threshold Parameter

In this subsection, we use the misclassification rate for simulation study of default value of sparsity threshold parameter τ for MFVB. The settings of the number of continuous predictors $d_\bullet$, sample size n, error standard deviations σ_ε, and τ are as in (2.58). The generation of data sets and design of simulation are as (2.56) and (2.57) respectively in Section 2.6.3.

For mean field variational Bayes, we use $\varepsilon_{\text{toler.}} = 10^{-8}$ from Algorithm 7 in Section 4.2.2. The rules of model selection follow (2.52) of Section 2.6.1 and (2.54) of Section 2.6.2.

Figure 4.1 shows the misclassification rate as a function of τ under mean field variational Bayes. The difference of misclassification rate between $\tau < 0.5$ and $\tau \geq 0.5$ is very noticeable, recalling that $\tau = 0$ corresponding to maximum a posteriori estimation of coefficient being zero described in Section 4.3.1 and 4.3.2. Therefore, the default value of the sparsity threshold parameter τ for mean field variational Bayes is 0.1.

The range of the sparsity threshold parameter is $\tau \in (0, 1)$. Both Figure 2.3 of Section 2.6.3 and Figure 4.1 show that, tweaking the sparsity threshold parameter τ among $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ has the desired impact on misclassification rates.

In this subsection, we also use relative test error to replace effect type misclassification rate for simulation study of default value of sparsity threshold parameter τ. The settings of the number of continuous predictors $d_\bullet$, sample size n, error standard deviations σ_ε, and sparsity threshold parameter τ, are as in (2.58). The continuous predictors generation for data sets are as in Section 2.6.3.

To aid visualization, we use the $\log_{10}$ transformation to the relative test error values. Similar to the simulation study for MCMC in Section 2.6.3, Figure 4.2 shows that the relative test errors are lower for higher sample sizes under the same noise level and the relative test errors are lower for higher noise levels under the same sample size. The choice of τ does not affect the relative test errors. Therefore, it does not change our recommendation concerning the choice of τ being 0.1 for MFVB.

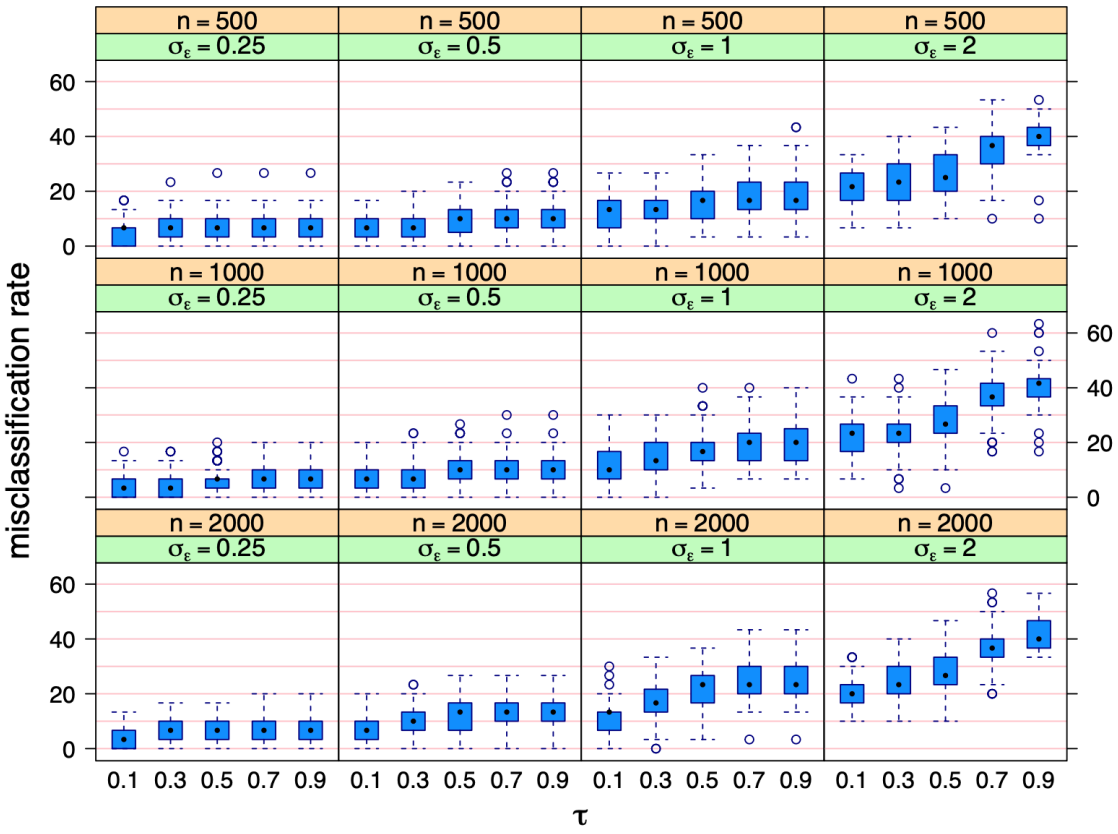


Figure 4.1: *Side-by-side boxplots of the misclassification rates for the mean field variational Bayes Algorithm 7 for the simulation study described in the text. Each panel corresponds to a different combination of sample size and error standard deviation. Within each panel, the side-by-side boxplots compare misclassification rate as a function of the sparsity threshold parameter τ.*

4.4 Data Illustration for Mean Field Variational Bayes

We implement our MFVB approach as given by Algorithm 7 in Section 4.2.2 in our R package gamselBayes. In this subsection, we apply our MFVB approach to two real data sets. The number of predictors in the first illustration for car auction data and in the second illustration for Sydney real estate data is large. This motivates the use of MFVB fitting as the fast option for model selection.

For the linear effects variables, we want to know their positive or negative impacts on the response. For the non-linear effects variables, we want to know their forms. These were able to be obtained by our R package gamselBayes.

Figure 4.2: *Side-by-side boxplots of the logarithms, to base 10, of relative test error for the mean field variational Bayes Algorithm 7 for the simulation study described in the text. Each panel corresponds to a different combination of sample size and error standard deviation. Within each panel, the side-by-side boxplots compare relative test error as a function of the sparsity threshold parameter τ.*

4.4.1 Application to Car Auction Data

The data set was obtained from the kaggle Internet platform (https://www.kaggle.com). There are 49 variables on 72,983 observations of cars purchased at car caution by automobile dealership in U.S.A.. The origin of these data is a classification competition titled "Don't Get Kicked!" that ran on the "kaggle" platform. The data set with all categorical variables converted to binary variable indicator form can be found in R package HRW Harezlak et al. (2018).

The binary response is the indicator of whether the car purchased at an auction has serious problems that prevent it being sold. We refer to such a car as a "bad buy". In the car auction data set, there are 44 binary predictors and 5 continuous predictors. We checked the unique values of the predictors and found the predictor "age at sale" only has 10 unique values. It is not conducive to spline-based estimation of non-linear effects. Therefore, we exclude age at sale variable from having non-linear

effect.

We use the fast option MFVB approach in our R package gamselBayes, with the threshold parameter τ being 0.1 as given in Section 4.3.3, with the stopping criteria being 10^{-8}, described in Section 4.2.3. Nineteen predictors were selected as having linear effects and three predictors were selected as having non-linear effects. Twenty-seven predictors out of forty-nine, which is 55% of all predictors, were discarded.

Table 4.1 shows estimation and inferential summaries for the selected linear fits. We found most of the predictor effects are intuitive. For example, with the predictor "age at sale", it shows that the older of the cars, are more likely to be a bad buy. The presence of wheel covers, lowered the probability to be a bad buy. It is interesting to note that cars purchased in some particular states, would make it more or less likely for the car to be a bad buy. For example, cars purchased in Florida and North Carolina are less likely for the car to be a bad buy. Cars purchased in Texas are more likely to be a bad buy.

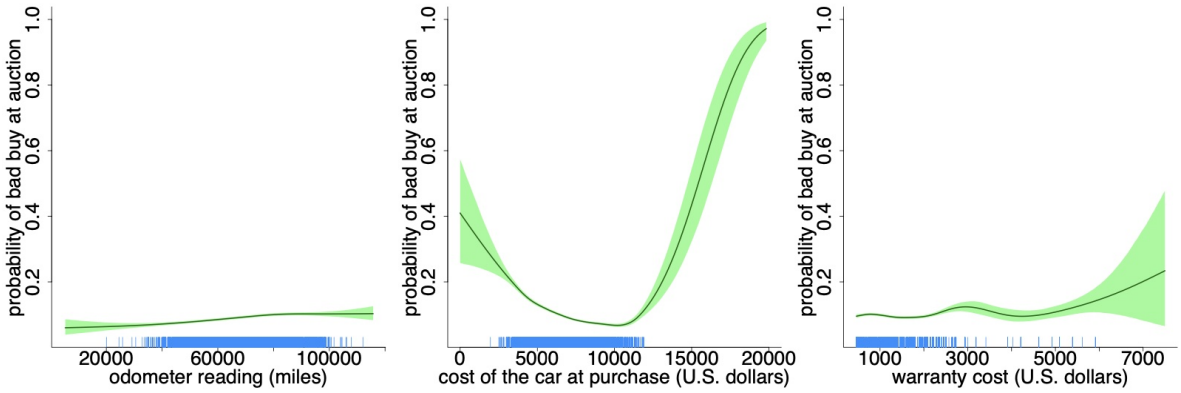


Figure 4.3: *The estimated non-linear effects for the car auction from application of Algorithm 7 and effect type estimation rules of Section 4.3.3 with $\tau = 0.1$. The curves are slices of estimated probability of bad buy as a function of the predictor, with all other selected predictors set to their median values. The shade regions correspond to pointwise approximate 95% credible intervals, but are subject to considerable mean field approximation error. The rug at the base of each plot is based on a random sample of 2,500 cars.*

Figure 4.3 shows 3 predictors having non-linear effects. They are odometer reading in miles, acquisition cost paid for the car at the time of purchase in U.S. dollars and warranty cost in U.S. dollars. The middle panel in Figure 4.3 shows that the purchase cost being around 10,000 U.S. dollars has the least probability of being a bad buy. The purchase cost below or above 10,000 U.S. dollars both increase the probability of being a bad buy. The shaded area in Figure 4.3 corresponds to pointwise approximate 95% credible intervals. However, we also need to consider considerable

mean field variational approximation error that tends to make the credible intervals overly narrow.

We also apply the MCMC approach as given by Algorithm 2 in Section 2.5.3 on this car auction data set. The categorical result shows that 14 out of the 19 predictors were selected from Table 4.1. Three predictors were not selected by the MFVB approach, but they were selected by the MCMC approach, such as indicators of the auction provider. The odometer reading predictor was considered having a non-linear effect by the MFVB approach, but was selected as having a linear effect by the MCMC approach. In summary, using the MFVB approach as given by Algorithm 7, 22 predictors were selected. Using the MCMC approach as given by Algorithm 2, 20 predictors were selected. Among those 20 predictors selected by the MCMC approach, 17 predictors were in common from the two approaches. This suggests reasonable accuracy of the MFVB approach as the fast option for this example.

4.4.2 Application of Data on House Sales in Sydney, Australia

The data set is obtained in R package HRW, which accompanies the Harezlak et al. (2018) book. There are 37,676 records and 39 variables for Sydney house sales in 2001.

The majority of the potential predictor variables are well-behaved. The histogram shows that predictor crime rate measure for the suburb in which the house is located is highly skewed and with a small fraction of the house having a large outlying value, which is not shown on histogram. Therefore, we apply the logarithm of crime rate plus one transformation to make the data of predictor crime rate measure for the suburb in which the house is located more amenable to our methodology.

Three predictors: nephelometer suspended matter measurement recorded at the air pollution monitoring station nearest to the house, particulate matter with a diameter of under 10 micrometers level recorded at the air pollution monitoring station nearest to the house, and sulphur dioxide level recorded at the air pollution monitoring station nearest to the house are quantitative but have low numbers of unique values. To avoid problems with spline fitting, these predictors are only considered as having linear or zero effects, and not considered as having a non-linear effects.

We use the fast MFVB approach in our R package gamselBayes, with the threshold parameter τ being 0.1 as given in Section 4.3.3, The categorical predictor results are shown in Table 4.2 and 4.3. Table 4.2 shows that 12 predictors were selected as having linear effects and Table 4.3 shows that 15 predictors were selected as having

non-linear effects. It implies 12 predictors were selected as having zero effects.

For the 12 predictors selected as having linear effects, we obtained the linear coefficient summary in Table 4.4. For example, among predictors financial quarter 2 in which sale took place, financial quarter 3 in which sale took place and financial quarter 4 in which sale took place, it indicates that houses sold in the 4th quarter period of October to December tend to have higher sales prices.

For the 15 predictors selected as having non-linear effects, Figure 4.4 shows that their effects on mean logarithm of house sales price. For example, the first panel shows that the effect of predictor lot size in square meters but with some imputation is quite a monotonic increasing curve, with some ramps and plateau. The fourth panel for predictor logarithm of crime rate plus one indicates that there is a sharp decrease for house sales price when predictor logarithm of crime rate plus one being around 0.32, and a sharp rise for house sales price when predictor logarithm of crime rate plus one being around 0.6 after decrease. The fifth panel shows that the predictor average weekly income of the suburb in which the house is located is mainly monotonically increasing with some ramps and plateau. The last panel for the predictor distance from house to the nearest school (kilometers) indicates that there is a sharp rise in mean house price when the distance to the nearest school is beyond 8 kilometers. This effect may due to homes in semi-rural areas being more expensive.

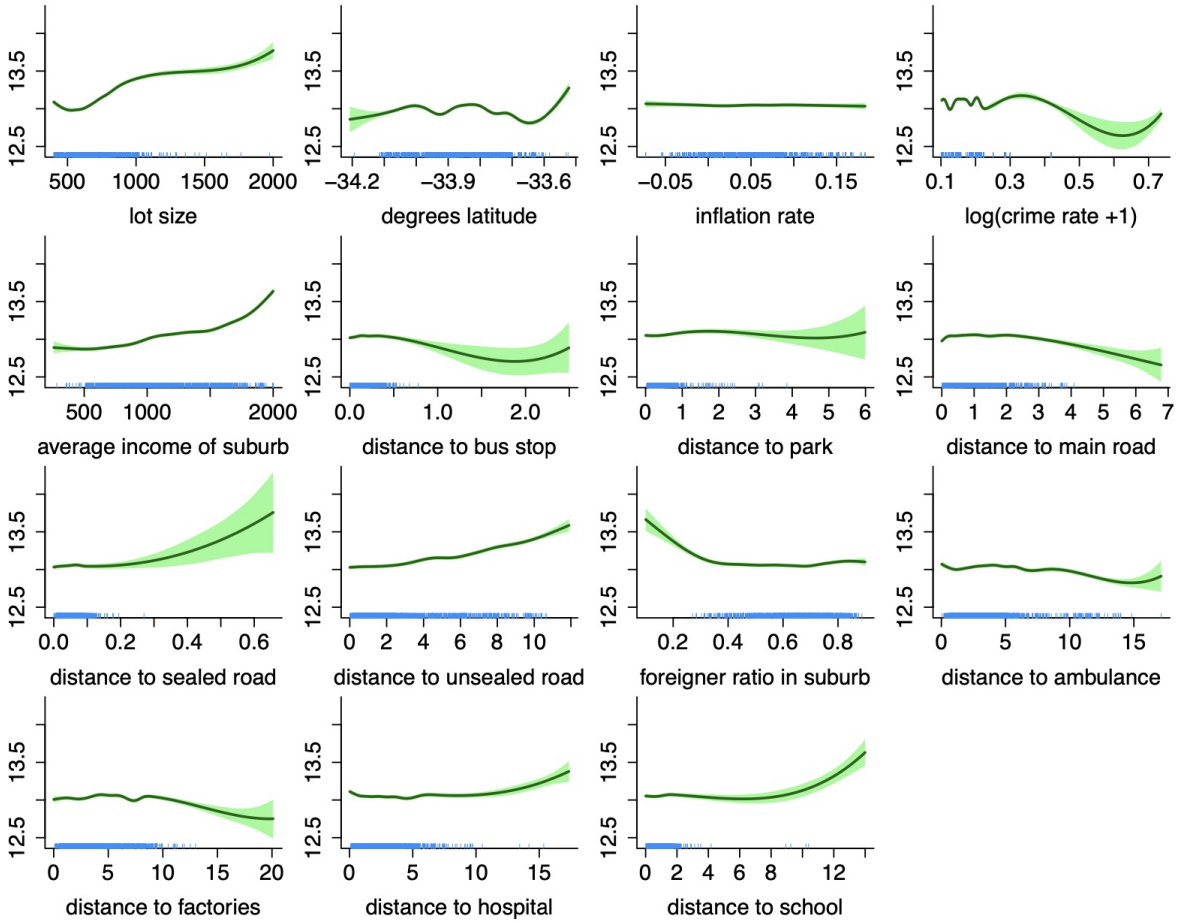


Figure 4.4: *The graph was produced by using the MFVB approach as given by Algorithm 7. The curves are vertically centred estimates of the effects of each predictor selected as having a non-linear effect. The shaded regions correspond to pointwise* 95% *credible intervals. The rugs at the base of each plot show values of each predictor. Due to the very large sample size, random subsets of size* $1,000$ *are used in the rugs.*

Algorithm 7: *Iterative determination of the optimal parameters according to a mean field variational Bayes approximation of the posterior distributions for model (2.28).*

Data Inputs: $\boldsymbol{y}(n \times 1)$; $\boldsymbol{X}(n \times (d_\circ + d_\bullet))$; $\boldsymbol{Z}_j(n \times K_j), 1 \leq j \leq d_\bullet$.

Response Type Input: **responseType** $\in$ {Gaussian,Bernoulli} .

Sufficient Statistics Inputs: **XTy, XTX, ZTy, ZTX, ZTZ**

Hyperparameters Inputs: $\sigma_{\beta_0}, s_\beta, s_\varepsilon, s_u > 0,\ \ 1 < \rho_\beta, \rho_u < 1.$

Convergence Criterion Input: $\varepsilon_{\text{toler}}$: a small positive number such as 10^{-8}.

Initialize: $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \longleftarrow \frac{1}{2}\mathbf{1}_{d_\circ + d_\bullet}$; $\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \longleftarrow \mathbf{0}_{d_\circ + d_\bullet}$; $\mu_{\mathfrak{q}(1/a_\varepsilon)} \longleftarrow 1$; $\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)} \longleftarrow 1$;

$\mu_{\mathfrak{q}(1/a_\beta)} \longleftarrow 1$; $\mu_{\mathfrak{q}(1/\sigma_\beta^2)} \longleftarrow 1$; $\kappa_{\mathfrak{q}(\sigma_\beta^2)} \longleftarrow \frac{1}{2}(d_\circ + d_\bullet + 1)$; $\kappa_{\mathfrak{q}(a_\beta)} \longleftarrow 1$

$\kappa_{\mathfrak{q}(\sigma_\varepsilon^2)} \longleftarrow \frac{1}{2}(n+1)$; $\kappa_{\mathfrak{q}(a_\varepsilon)} \longleftarrow 1$; $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{b}_\beta)} \longleftarrow \mathbf{1}_{d_\circ + d_\bullet}$;

$\mathbf{yT1}_{\text{adj}} \longleftarrow 0$; $\mathbf{XTy}_{\text{adj}} \longleftarrow \mathbf{XTy}$; $\mathbf{ZTy}_{\text{adj}} \longleftarrow \mathbf{ZTy}$

For $j = 1, \ldots, d_\bullet$:

$\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} \longleftarrow \mathbf{0}_{K_j}$; $\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} \longleftarrow \mathbf{1}_{K_j}$; $\mu_{\mathfrak{q}(\gamma_{uj})} \longleftarrow \frac{1}{2}$;

$\mu_{\mathfrak{q}(1/a_{uj})} \longleftarrow 1$; $\mu_{\mathfrak{q}(1/\sigma^2_{uj})} \longleftarrow 1$; $\mu_{\mathfrak{q}(b_{uj})} \longleftarrow 1$;

$\kappa_{\mathfrak{q}(\sigma^2_{uj})} \longleftarrow \frac{1}{2}(K_j + 1)$; $\kappa_{\mathfrak{q}(a_{uj})} \longleftarrow 1$; $\boldsymbol{\omega}_{Z_j} \longleftarrow \text{diagonal}\left(\mathbf{ZTZ}^{\langle j,j\rangle}\right)$

Cycle:

$\omega_{12} \longleftarrow \mathbf{yT1}_{adj}$

$\sigma^2_{\mathfrak{q}(\beta_0)} \longleftarrow 1/\left\{n\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)} + (1/\sigma^2_{\beta_0})\right\}$; $\mu_{\mathfrak{q}(\beta_0)} \longleftarrow \sigma^2_{\mathfrak{q}(\beta_0)}\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\omega_{12}$

$\boldsymbol{\Omega}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \longleftarrow \text{diag}\left\{\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \left(\mathbf{1} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}\right)\right\} + \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}\boldsymbol{\mu}^T_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}$

$\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \longleftarrow \left\{\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\boldsymbol{\Omega}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \left(\boldsymbol{X}^T\boldsymbol{X}\right) + \mu_{\mathfrak{q}(1/\sigma_\beta^2)}\text{diag}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{b}_\beta)}\right)\right\}^{-1}$

$\boldsymbol{\omega}_{13} \longleftarrow \mathbf{XTy}_{\text{adj}} - \sum_{j=1}^{d_\bullet}\mathbf{ZTX}^{\langle j\rangle T}\left(\mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right)$

$\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \longleftarrow \mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\omega}_{13}\right)$

$\boldsymbol{\omega}_{14} \longleftarrow \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} + \text{diagonal}\left(\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right)$; $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{b}_\beta)} \longleftarrow \left(\mu_{\mathfrak{q}(1/\sigma_\beta^2)}\boldsymbol{\omega}_{14}\right)^{-1/2}$

continued on a subsequent page...

Algorithm 7 continued: *This is a continuation of the description of this algorithm that commences on a preceding page.*

$$\lambda_{\mathfrak{q}(\sigma_\beta^2)} \longleftarrow \mu_{\mathfrak{q}(1/a_\beta)} + \frac{1}{2}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{b}_\beta)}^T \boldsymbol{\omega}_{14} \quad ; \quad \mu_{\mathfrak{q}(1/\sigma_\beta^2)} \longleftarrow \kappa_{\mathfrak{q}(\sigma_\beta^2)} \Big/ \lambda_{\mathfrak{q}(\sigma_\beta^2)}$$

$$\lambda_{\mathfrak{q}(a_\beta)} \longleftarrow \mu_{\mathfrak{q}(1/\sigma_\beta^2)} + s_\beta^{-2} \quad ; \quad \mu_{\mathfrak{q}(1/a_\beta)} \longleftarrow \kappa_{\mathfrak{q}(a_\beta)} \Big/ \lambda_{\mathfrak{q}(a_\beta)}$$

For $j = 1, \ldots, d_\bullet$: $\quad \boldsymbol{\mu}_{\mathfrak{q}(u_j)} \longleftarrow \mu_{\mathfrak{q}(\gamma_{uj})} \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}$

For $j = 1, \ldots, d_\circ + d_\bullet$:

$$\omega_{15} \longleftarrow \boldsymbol{e}_j^T \mathbf{X}\mathbf{T}\mathbf{y}_{\text{adj}} - \sum_{j=1}^{d_\bullet} \left(\mathbf{Z}\mathbf{T}\mathbf{X}^{\langle j' \rangle} \boldsymbol{e}_j \right)^T \boldsymbol{\mu}_{\mathfrak{q}(u_{j'})}$$

$$\omega_{15} \longleftarrow \mu_{\mathfrak{q}(\widetilde{\beta}_j)} \omega_{15} - \left(\mathbf{X}\mathbf{T}\mathbf{X}\boldsymbol{e}_j \right)_{-j}^T \times \left[\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \right)_{-j} \odot \left\{ \left(\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \boldsymbol{e}_j \right)_{-j} + \mu_{\mathfrak{q}(\widetilde{\beta}_j)} \left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \right)_{-j} \right\} \right]$$

$$\mu_{\mathfrak{q}(\gamma_{\beta j})} \longleftarrow \text{expit}\left(\text{logit}(\rho_\beta)) - \frac{1}{2}\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)} \left\{ \left(\mu_{\mathfrak{q}(\widetilde{\beta}_j)}^2 + \sigma_{\mathfrak{q}(\widetilde{\beta}_j)}^2 \right) \boldsymbol{e}_j^T \mathbf{X}\mathbf{T}\mathbf{X}\boldsymbol{e}_j - 2\omega_{15} \right\} \right)$$

For $j = 1, \ldots, d_\bullet$: $\quad \boldsymbol{\mu}_{\mathfrak{q}(u_j)} \longleftarrow \mu_{\mathfrak{q}(\gamma_{uj})} \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}$

For $j = 1, \ldots, d_\bullet$:

$$\boldsymbol{\omega}_{16} \longleftarrow \mathbf{Z}\mathbf{T}\mathbf{y}_{\text{adj}}^{\langle j \rangle} - \mathbf{Z}\mathbf{T}\mathbf{X}^{\langle j \rangle} \left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \right) - \sum_{j' \neq j}^{d_\bullet} \mathbf{Z}\mathbf{T}\mathbf{Z}^{\langle j, j' \rangle} \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{j'})}$$

$$\boldsymbol{\sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}^2 \longleftarrow \mathbf{1}_{K_j} \Big/ \left\{ \mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)} \mu_{\mathfrak{q}(\gamma_{uj})} \boldsymbol{\omega}_{\boldsymbol{Z}_j} + \mu_{\mathfrak{q}(1/\sigma_{uj}^2)} \mu_{\mathfrak{q}(b_{uj})} \mathbf{1}_{K_j} \right\}$$

$$\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} \longleftarrow \mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)} \mu_{\mathfrak{q}(\gamma_{uj})} \left(\boldsymbol{\omega}_{16} \odot \boldsymbol{\sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}^2 \right)$$

For $j = 1, \ldots, d_\bullet$:

$$\omega_{17} \longleftarrow \left\| \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} \right\|^2 + \mathbf{1}_{K_j}^T \boldsymbol{\sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}^2; \quad \mu_{\mathfrak{q}(b_{uj})} \longleftarrow \left(\mu_{\mathfrak{q}(1/\sigma_{uj}^2)} \omega_{17} \right)^{-1/2}$$

$$\lambda_{\mathfrak{q}(\sigma_{uj}^2)} \longleftarrow \mu_{\mathfrak{q}(1/a_{uj})} + \frac{1}{2}\mu_{\mathfrak{q}(b_{uj})} \omega_{17} \quad ; \quad \mu_{\mathfrak{q}(1/\sigma_{uj}^2)} \longleftarrow \kappa_{\mathfrak{q}(\sigma_{uj}^2)} \Big/ \lambda_{\mathfrak{q}(\sigma_{uj}^2)}$$

$$\lambda_{\mathfrak{q}(a_{uj})} \longleftarrow \mu_{\mathfrak{q}(1/\sigma_{uj}^2)} + \left(1/s_u^2\right) \quad ; \quad \mu_{\mathfrak{q}(1/a_{uj})} \longleftarrow \kappa_{\mathfrak{q}(a_{uj})} \Big/ \lambda_{\mathfrak{q}(a_{uj})}$$

continued on a subsequent page...

Algorithm 7 continued: *This is a continuation of the description of this algorithm that commences on a preceding page.*

For $j = 1, \ldots, d_\bullet$: $\quad \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_j)} \longleftarrow \mu_{\mathfrak{q}(\gamma_{uj})} \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}$

For $j = 1, \ldots, d_\bullet$:

$$\boldsymbol{\omega}_{18} \longleftarrow \mathbf{ZTy}_{\text{adj}}^{\langle j \rangle} - \mathbf{ZTX}^{\langle j \rangle} \left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \right) - \sum_{j' \neq j}^{d_\bullet} \mathbf{ZTZ}^{\langle j, j' \rangle} \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{j'})}$$

$$\omega_{19} \longleftarrow \boldsymbol{\omega}_{\boldsymbol{Z}_j}^T \left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} + \boldsymbol{\sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}^2 \right) - 2 \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}^T \boldsymbol{\omega}_{18}$$

$$\mu_{\mathfrak{q}(\gamma_{uj})} \longleftarrow \text{expit}\Big(\text{logit}(\rho_u) - \frac{1}{2} \mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)} \omega_{19} \Big)$$

$$\boldsymbol{\omega}_{20} \longleftarrow \mathbf{1}_n \mu_{\mathfrak{q}(\beta_0)} + \boldsymbol{X} \left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \right) + \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j \left(\mu_{\mathfrak{q}(\gamma_{uj})} \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} \right)$$

If **responseType** is Gaussian then

$$\boldsymbol{\Omega}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \longleftarrow \text{diag} \left\{ \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \left(\mathbf{1} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \right) \right\} + \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}^T$$

$$\begin{aligned} \lambda_{\mathfrak{q}(\sigma_\varepsilon^2)} \longleftarrow \mu_{\mathfrak{q}(1/a_\varepsilon)} &+ \frac{1}{2} \Big\| \boldsymbol{y} - \boldsymbol{\omega}_{20} \Big\|^2 + \frac{1}{2} n \sigma_{\mathfrak{q}(\beta_0)}^2 \\ &+ \frac{1}{2} \text{tr} \left[\boldsymbol{X}^T \boldsymbol{X} \left\{ \boldsymbol{\Omega}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \left(\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} + \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}^T \right) \right\} \right] \\ &- \frac{1}{2} \mathbf{tr} \left\{ \boldsymbol{X}^T \boldsymbol{X} \left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \right) \left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \right)^T \right\} \\ &+ \frac{1}{2} \sum_{j=1}^{d_\bullet} \boldsymbol{\omega}_{\boldsymbol{Z}_j}^T \left(\mu_{\mathfrak{q}(\gamma_{uj})} \left[\boldsymbol{\sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}^2 + \left\{ 1 - \mu_{\mathfrak{q}(\gamma_{uj})} \right\} \left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} \right) \right] \right) \end{aligned}$$

$$\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)} \longleftarrow \kappa_{\mathfrak{q}(\sigma_\varepsilon^2)} \Big/ \lambda_{\mathfrak{q}(\sigma_\varepsilon^2)}; \;\; \lambda_{\mathfrak{q}(a_\varepsilon)} \longleftarrow \mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)} + \left(1 \Big/ s_\varepsilon^2 \right);$$

$$\mu_{\mathfrak{q}(1/a_\varepsilon)} \longleftarrow \kappa_{\mathfrak{q}(a_\varepsilon)} \Big/ \lambda_{\mathfrak{q}(a_\varepsilon)}$$

If **responseType** is Bernoulli then

$$\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)} \longleftarrow 1 \; ; \;\; \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{c})} \longleftarrow \boldsymbol{\omega}_{20} + \left(2\boldsymbol{y} - \mathbf{1}_n \right) \odot \zeta' \left(\left(2\boldsymbol{y} - \mathbf{1}_n \right) \odot \boldsymbol{\omega}_{20} \right)$$

$$\mathbf{yT1}_{\text{adj}} \longleftarrow \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{c})}^T \mathbf{1}_n \; ; \; \mathbf{XTy}_{\text{adj}} \longleftarrow \boldsymbol{X}^T \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{c})} \; ; \; \mathbf{ZTy}_{\text{adj}} \longleftarrow \boldsymbol{Z}^T \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{c})}$$

until the relative change in the $\log(\underline{p}(\boldsymbol{y}; \mathfrak{q}))$ is below $\varepsilon_{\text{toler}}$.

Outputs: All $\mathfrak{q}$-density parameters.

predictor	posterior mean	95% credible interval
indic. made in U.S.A.	–0.04605	(–0.06925, –0.02323)
age at sale (years)	0.09344	(0.08771, 0.09896)
indic. color is red	0.0513	(0.02498, 0.07713)
indic. make is Chevrolet	–0.1103	(–0.1358, –0.08426)
indic. make is Chrysler	0.09517	(0.0691, 0.1217)
indic. make is Dodge	–0.03093	(–0.05385, –0.006961)
indic. purchased online	–0.06229	(–0.1117, –0.0000)
acquisition price (U.S. dollars)	-6.014×10^{-6}	(–9.352, –2.603)$\times 10^{-6}$
indic. purchased in 2010	0.1085	(0.0929, 0.1242)
indic. purch. in Florida	–0.1173	(–0.1391, –0.09529)
indic. purch. in North Carolina	–0.1074	(–0.1326, –0.0819)
indic. purch. in Texas	0.09706	(0.07799, 0.1162)
indic. medium-sized vehicle	–0.07368	(–0.09095, –0.05622)
indic. sports utility vehicle	0.1899	(0.1646, 0.2153)
indic. manual transmission	–0.1574	(–0.197, –0.1168)
indic. trim level is 'Bas'	0.05521	(0.0355, 0.07477)
indic. trim level is 'LS'	–0.06153	(–0.0918, –0.03176)
indic. has alloy wheels	–1.513	(–1.546, –1.48)
indic. has wheel covers	–1.585	(–1.619, –1.551)

Table 4.1: *Approximate posterior means and approximate 95% credible intervals for the coefficients of each of the selected linear fits based on the mean field variational Bayes optimal* $\mathfrak{q}$*-densities obtained from Algorithm 7 for the car auction example.*

Predictors selected as having linear effects:
nitrous oxide level recorded at the air pollution monitoring station nearest to the house,
particulate matter with a diameter of under 10 micrometers level recorded at the air pollution monitoring station nearest to the house,
sulphur dioxide level recorded at the air pollution monitoring station nearest to the house,
distance from house to the nearest highway (kilometers),
distance from house to the nearest medical services (kilometers),
distance from house to the Sydney Harbour Tunnel (kilometers),
degrees longitude of location of house,
nephelometer suspended matter measurement recorded at the air pollution monitoring station nearest to the house,
ozone level recorded at the air pollution monitoring station nearest to the house,
financial quarter 2 in which sale took place,
financial quarter 3 in which sale took place,
financial quarter 4 in which sale took place.

Table 4.2: *Predictors selected as having linear effects from Sydney real estate data set using MFVB approach as Algorithm 7.*

Predictors selected as having non-linear effects:
distance from house to the nearest ambulance station (kilometers),
distance from house to the nearest bus stop (kilometers),
distance from house to the nearest factory (kilometers),
distance from house to the nearest hospital (kilometers),
distance from house to the nearest main road (kilometers),
distance from house to the nearest park (kilometers),
distance from house to the nearest school (kilometers),
distance from house to the nearest sealed road (kilometers),
distance from house to the nearest unsealed road (kilometers),
proportion of foreigners in the suburb in which the house is located,
average weekly income of the suburb in which the house is located.
inflation rate measured as a percentage,
degrees latitude of location of house,
logarithm of crime rate measured for the suburb in which the house is located plus one,
lot size in square meters but with some imputation.

Table 4.3: *Predictors selected as having non-linear effects from Sydney real estate data set using MFVB approach as Algorithm 7.*

	posterior mean	95% credible interval
nitrous oxide level recorded at the air pollution monitoring station nearest to the house	0.01213	$(0.01081, 0.0135)$
particulate matter with a diameter of under 10 micrometers level recorded at the air pollution monitoring station nearest to the house	0.01388	$(0.01111, 0.01672)$
sulphur dioxide level recorded at the air pollution monitoring station nearest to the house	–1.526	$(-1.722, -1.326)$
distance from house to the nearest highway (kilometers)	0.001718	$(0.0006906, 0.002848)$
distance from house to the nearest medical services (kilometers)	–0.003672	$(-0.006384, -0.001)$
distance from house to the Sydney Harbour Tunnel (kilometers)	0.008914	$(0.007736, 0.01006)$
degrees longitude of location of house	2.579	$(2.501, 2.659)$
nephelometer suspended matter measurement recorded at the air pollution monitoring station nearest to the house	–0.4299	$(-0.5201, -0.3398)$

Table 4.4: *Inferential summary of the linear coefficients obtained from Sydney real estate data set using MFVB approach as Algorithm 7.*

	posterior mean	95% credible interval
ozone level recorded at the air pollution monitoring station nearest to the house	0.3396	$(0.3139, 0.3656)$
financial quarter 2 in which sale took place	0.01908	$(0.008528, 0.02939)$
financial quarter 3 in which sale took place	0.06412	$(0.05468, 0.07339)$
financial quarter 4 in which sale took place	0.1311	$(0.1219, 0.1404)$

Table 4.4 continued: *Inferential summary of the linear coefficients obtained from Sydney real estate data set using MFVB approach as Algorithm 7.*

4.A The q-Density approximation of parameters with Laplace-Zero priors

Throughout Appendix 4.A, "const" denotes a constant that does not depend on any q-density parameters.

4.A.1 The q-Density of β_0 with Laplace-Zero Priors

First note that

$$\log\{\mathfrak{q}(\beta_0)\} = E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\beta_0\middle|\text{rest}\right)\right\}\right] + \text{const.}$$

Recalling the expression for $\mathfrak{p}(\beta_0|\text{rest})$ in (2.61) of Appendix 2.B.1, we have

$$\begin{aligned}\log\left\{\mathfrak{p}\left(\beta_0\middle|\text{rest}\right)\right\} &= \begin{bmatrix}\beta_0 \\ \beta_0^2\end{bmatrix}^T \begin{bmatrix}\eta_1 \\ \eta_2\end{bmatrix} + \text{const} \\ &= \begin{bmatrix}\beta_0 \\ \beta_0^2\end{bmatrix}^T \begin{bmatrix}\frac{1}{\sigma_\varepsilon^2}\mathbf{1}_n^T\left(\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j)\right) \\ -\frac{1}{2}\left(\frac{n}{\sigma_\varepsilon^2} + \frac{1}{\sigma_{\beta_0}^2}\right)\end{bmatrix} + \text{const}\end{aligned}$$

where

$$\eta_1 \equiv \frac{1}{\sigma_\varepsilon^2}\mathbf{1}_n^T\left(\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j)\right)$$

and

$$\eta_2 \equiv -\frac{1}{2}\left(\frac{n}{\sigma_\varepsilon^2} + \frac{1}{\sigma_{\beta_0}^2}\right).$$

Moreover,

$$E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\beta_0\middle|\text{rest}\right)\right\}\right] = \begin{bmatrix}\beta_0 \\ \beta_0^2\end{bmatrix}^T E_{\mathfrak{q}}\begin{bmatrix}\eta_1 \\ \eta_2\end{bmatrix} + \text{const.}$$

We then have

$$E_{\mathfrak{q}}\left(\eta_2\right) = -\frac{1}{2}E_{\mathfrak{q}}\left(\frac{n}{\sigma_{\varepsilon}^2} + \frac{1}{\sigma_{\beta_0}^2}\right) = -\frac{1}{2}\left(n\mu_{\mathfrak{q}(1/\sigma_{\varepsilon}^2)} + \frac{1}{\sigma_{\beta_0}^2}\right) = -\frac{1}{2}\left(\sigma_{\mathfrak{q}(\beta_0)}^2\right)^{-1}$$

and therefore

$$\sigma_{\mathfrak{q}(\beta_0)}^2 = \left(n\mu_{\mathfrak{q}(1/\sigma_{\varepsilon}^2)} + \frac{1}{\sigma_{\beta_0}^2}\right)^{-1}$$

is given in (4.8) and Algorithm 7 of Section 4.2.2. And note that

$$\begin{aligned} E_{\mathfrak{q}}\left(\eta_1\right) = \frac{1}{\sigma_{\mathfrak{q}(\beta_0)}^2}\mu_{\mathfrak{q}(\beta_0)} &= E_{\mathfrak{q}}\left\{\frac{1}{\sigma_{\varepsilon}^2}\mathbf{1}_n^T\left(\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{\gamma}_{\beta}\odot\widetilde{\boldsymbol{\beta}}) - \sum_{j=1}^{d_{\bullet}}\boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j)\right)\right\} \\ &= \mu_{\mathfrak{q}(1/\sigma_{\varepsilon}^2)}\mathbf{1}_n^T\left[\boldsymbol{y} - \boldsymbol{X}\left(\boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{\gamma}_{\beta}\right)}\odot\boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\right) - \sum_{j=1}^{d_{\bullet}}\boldsymbol{Z}_j\left(\mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right)\right]. \quad (4.32) \end{aligned}$$

Recalling that (2.7) in Section 2.3.2 to be $\mathbf{1}_n^T\boldsymbol{X} = \mathbf{0}$ and (2.8) in Section 2.3.3 to be $\mathbf{1}_n^T\boldsymbol{Z}_j = \mathbf{0}, 1 \le j \le d_{\bullet}$, in (4.32), if we define

$$\omega_{12} \equiv \mathbf{1}_n^T\left(\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{\gamma}_{\beta}\right)}\odot\boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}) - \sum_{j=1}^{d_{\bullet}}\boldsymbol{Z}_j(\mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)})\right),$$

then $\omega_{12} = \mathbf{1}_n^T\boldsymbol{y}$. Therefore

$$\mu_{\mathfrak{q}(\beta_0)} = \sigma_{\mathfrak{q}(\beta_0)}^2 E_{\mathfrak{q}}\left(\eta_1\right) = \sigma_{\mathfrak{q}(\beta_0)}^2\mu_{\mathfrak{q}(1/\sigma_{\varepsilon}^2)}\omega_{12} = \sigma_{\mathfrak{q}(\beta_0)}^2\mu_{\mathfrak{q}(1/\sigma_{\varepsilon}^2)}\mathbf{1}_n^T\boldsymbol{y}$$

is given in (4.9) and Algorithm 7 of Section 4.2.2.

4.A.2 The q-Density of $\widetilde{\boldsymbol{\beta}}$ with Laplace-Zero Priors

First note that

$$\log\left\{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)\right\} = E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\widetilde{\boldsymbol{\beta}}\middle|\text{rest}\right)\right\}\right] + \text{const.}$$

Recalling the expression for $\mathfrak{p}(\widetilde{\boldsymbol{\beta}}|\text{rest})$ in (2.62) of Appendix 2.B.2, we have

$$\log\left\{\mathfrak{p}\left(\widetilde{\boldsymbol{\beta}}\middle|\text{rest}\right)\right\} = \begin{bmatrix} \widetilde{\boldsymbol{\beta}} \\ \\ \text{vec}(\widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}^T) \end{bmatrix}^T \begin{bmatrix} \boldsymbol{\eta_1} \\ \\ \boldsymbol{\eta_2} \end{bmatrix} + \text{const}$$

$$= \begin{bmatrix} \widetilde{\boldsymbol{\beta}} \\ \\ \text{vec}(\widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}^T) \end{bmatrix}^T \begin{bmatrix} \frac{1}{\sigma_\varepsilon^2}\left(\boldsymbol{\gamma}_\beta \odot \left[\boldsymbol{X}^T\left(\boldsymbol{y} - \mathbf{1}_n\beta_0 - \sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j\left(\gamma_{uj}\widetilde{\boldsymbol{u}}_j\right)\right)\right]\right) \\ -\frac{1}{2}\text{vec}\left(\frac{\left(\left(\boldsymbol{\gamma}_\beta\boldsymbol{\gamma}_\beta^T\right)\odot\left(\boldsymbol{X}^T\boldsymbol{X}\right)\right)}{\sigma_\varepsilon^2} + \frac{\text{diag}\left(\boldsymbol{b}_\beta\right)}{\sigma_\beta^2}\right) \end{bmatrix} + \text{const}$$

where

$$\boldsymbol{\eta_1} \equiv \frac{1}{\sigma_\varepsilon^2}\left(\boldsymbol{\gamma}_\beta \odot \left[\boldsymbol{X}^T\left(\boldsymbol{y} - \mathbf{1}_n\beta_0 - \sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j\left(\gamma_{uj}\widetilde{\boldsymbol{u}}_j\right)\right)\right]\right)$$

and

$$\boldsymbol{\eta_2} \equiv -\frac{1}{2}\text{vec}\left(\frac{\left(\left(\boldsymbol{\gamma}_\beta\boldsymbol{\gamma}_\beta^T\right)\odot\left(\boldsymbol{X}^T\boldsymbol{X}\right)\right)}{\sigma_\varepsilon^2} + \frac{\text{diag}\left(\boldsymbol{b}_\beta\right)}{\sigma_\beta^2}\right).$$

Moreover,

$$E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\widetilde{\boldsymbol{\beta}}\middle|\text{rest}\right)\right\}\right] = \begin{bmatrix} \widetilde{\boldsymbol{\beta}} \\ \\ \text{vec}(\widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}^T) \end{bmatrix}^T E_{\mathfrak{q}}\begin{bmatrix} \boldsymbol{\eta_1} \\ \\ \boldsymbol{\eta_2} \end{bmatrix} + \text{const.}$$

If we define $\boldsymbol{\Omega}_{\mathfrak{q}\left(\boldsymbol{\gamma}_\beta\right)} \equiv E_{\mathfrak{q}}\left(\boldsymbol{\gamma}_\beta\boldsymbol{\gamma}_\beta^T\right)$, then

$$E_{\mathfrak{q}}\left(\boldsymbol{\eta_2}\right) = -\frac{1}{2}\text{vec}\left(E_{\mathfrak{q}}\left\{\frac{\left(\left(\boldsymbol{\gamma}_\beta\boldsymbol{\gamma}_\beta^T\right)\odot\left(\boldsymbol{X}^T\boldsymbol{X}\right)\right)}{\sigma_\varepsilon^2} + \frac{\text{diag}\left(\boldsymbol{b}_\beta\right)}{\sigma_\beta^2}\right\}\right) = -\frac{1}{2}\text{vec}\left(\boldsymbol{\Sigma}^{-1}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\right)$$

$$= -\frac{1}{2}\text{vec}\left[\mu_{\mathfrak{q}\left(1/\sigma_\varepsilon^2\right)}\left\{\boldsymbol{\Omega}_{\mathfrak{q}\left(\boldsymbol{\gamma}_\beta\right)}\odot\left(\boldsymbol{X}^T\boldsymbol{X}\right)\right\} + \mu_{\mathfrak{q}\left(1/\sigma_\beta^2\right)}\text{diag}\left(\boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{b}_\beta\right)}\right)\right].$$

Therefore

$$\begin{aligned}\boldsymbol{\Omega}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} &= \mathrm{Cov}_{\mathfrak{q}}\left(\boldsymbol{\gamma}_\beta\right) + E_{\mathfrak{q}}\left(\boldsymbol{\gamma}_\beta\right) E_{\mathfrak{q}}\left(\boldsymbol{\gamma}_\beta^T\right) \\ &= \mathrm{diag}\left\{\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \left(\mathbf{1} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}\right)\right\} + \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}^T\end{aligned}$$

and

$$\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} = \left[\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\left\{\boldsymbol{\Omega}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \left(\boldsymbol{X}^T\boldsymbol{X}\right)\right\} + \mu_{\mathfrak{q}\left(1/\sigma_{\widetilde{\boldsymbol{\beta}}}^2\right)}\mathrm{diag}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{b}_\beta)}\right)\right]^{-1}$$

are obtained to be (4.10), (4.11) and given in Algorithm 7 of Section 4.2.2. Note that

$$\begin{aligned}E_{\mathfrak{q}}\left(\boldsymbol{\eta}_{\mathbf{1}}\right) &= \boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}^{-1}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \\ &= E_{\mathfrak{q}}\left\{\frac{1}{\sigma_\varepsilon^2}\left(\boldsymbol{\gamma}_\beta \odot \left[\boldsymbol{X}^T\left(\boldsymbol{y} - \mathbf{1}_n\beta_0 - \sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j)\right)\right]\right)\right\} \\ &= \mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\left(\boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{\gamma}_\beta\right)} \odot \left[\boldsymbol{X}^T\left(\boldsymbol{y} - \mathbf{1}_n\mu_{\mathfrak{q}(\beta_0)} - \sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j(\mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)})\right)\right]\right). \quad (4.33)\end{aligned}$$

Recalling that (2.7) of Section 2.3.2 to be $\mathbf{1}_n^T\boldsymbol{X} = \mathbf{0}$, in (4.33), if we define

$$\boldsymbol{\omega}_{13} \equiv \boldsymbol{X}^T\left(\boldsymbol{y} - \mathbf{1}_n\mu_{\mathfrak{q}(\beta_0)} - \sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j\left(\mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right)\right),$$

then

$$\boldsymbol{\omega}_{13} = \boldsymbol{X}^T\left(\boldsymbol{y} - \sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j(\mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)})\right)$$

and

$$\begin{aligned}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} &= \mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\left(\boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{\gamma}_\beta\right)} \odot \left[\boldsymbol{X}^T\left(\boldsymbol{y} - \sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j(\mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)})\right)\right]\right) \\ &= \mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\left(\boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{\gamma}_\beta\right)} \odot \boldsymbol{\omega}_{13}\right)\end{aligned}$$

are given in Algorithm 7 and (4.12) of Section 4.2.2.

4.A.3 The $\mathfrak{q}$-Density of $\widetilde{\boldsymbol{u}}_j$ with Laplace-Zero Priors

First note that

$$\log\left\{\mathfrak{q}\left(\widetilde{\boldsymbol{u}}_j\right)\right\} = E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\widetilde{\boldsymbol{u}}_j\middle|\text{rest}\right)\right\}\right] + \text{const}, \quad 1 \leq j \leq d_{\bullet}.$$

Recalling the expression for $\mathfrak{p}(\widetilde{\boldsymbol{u}}_j|\text{rest})$ in (2.65) of Appendix 2.B.3, we have

$$\log\left\{\mathfrak{p}\left(\widetilde{\boldsymbol{u}}_j\middle|\text{rest}\right)\right\} = \begin{bmatrix} \widetilde{\boldsymbol{u}}_j \\ \\ \text{vec}(\widetilde{\boldsymbol{u}}_j\widetilde{\boldsymbol{u}}_j^T) \end{bmatrix}^T \begin{bmatrix} \boldsymbol{\eta_1} \\ \\ \boldsymbol{\eta_2} \end{bmatrix} + \text{const} = \begin{bmatrix} \widetilde{\boldsymbol{u}}_j \\ \\ \text{vec}(\widetilde{\boldsymbol{u}}_j\widetilde{\boldsymbol{u}}_j^T) \end{bmatrix}^T$$

$$\times \begin{bmatrix} \frac{1}{\sigma_\varepsilon^2}\left(\gamma_{uj}\left[\boldsymbol{Z}_j^T\boldsymbol{y} - \boldsymbol{Z}_j^T\boldsymbol{X}\left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}\right) - \sum_{j'\neq j}^{d_\bullet}\boldsymbol{Z}_j^T\boldsymbol{Z}_{j'}\left(\gamma_{uj'}\widetilde{\boldsymbol{u}}_{j'}\right)\right]\right) \\ -\frac{1}{2}\text{vec}\left(\text{diag}\left(\frac{\gamma_{uj}\boldsymbol{\omega}_{\boldsymbol{Z}_j}}{\sigma_\varepsilon^2} + \frac{b_{uj}\mathbf{1}}{\sigma_{uj}^2}\right)\right) \end{bmatrix} + \text{const.}$$

where

$$\boldsymbol{\eta_1} \equiv \frac{1}{\sigma_\varepsilon^2}\left(\gamma_{uj}\left[\boldsymbol{Z}_j^T\boldsymbol{y} - \boldsymbol{Z}_j^T\boldsymbol{X}\left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}\right) - \sum_{j'\neq j}^{d_\bullet}\boldsymbol{Z}_j^T\boldsymbol{Z}_{j'}\left(\gamma_{uj'}\widetilde{\boldsymbol{u}}_{j'}\right)\right]\right)$$

and

$$\boldsymbol{\eta_2} \equiv -\frac{1}{2}\text{vec}\left(\text{diag}\left(\frac{\gamma_{uj}\boldsymbol{\omega}_{\boldsymbol{Z}_j}}{\sigma_\varepsilon^2} + \frac{b_{uj}\mathbf{1}}{\sigma_{uj}^2}\right)\right).$$

And note that

$$E_{\mathfrak{q}}\left\{\log\left(\mathfrak{p}\left(\widetilde{\boldsymbol{u}}_j\middle|\text{rest}\right)\right)\right\} = \begin{bmatrix} \widetilde{\boldsymbol{u}}_j \\ \\ \text{vec}(\widetilde{\boldsymbol{u}}_j\widetilde{\boldsymbol{u}}_j^T) \end{bmatrix}^T E_{\mathfrak{q}}\begin{bmatrix} \boldsymbol{\eta_1} \\ \\ \boldsymbol{\eta_2} \end{bmatrix} + \text{const.}$$

We have

$$
\begin{aligned}
E_{\mathfrak{q}}\left(\boldsymbol{\eta_2}\right) &= -\frac{1}{2}\text{vec}\left(\text{diag}\left(E_{\mathfrak{q}}\left\{\frac{\gamma_{uj}\boldsymbol{\omega}_{\boldsymbol{Z}_j}}{\sigma_{\varepsilon}^2}+\frac{b_{uj}\mathbf{1}}{\sigma_{uj}^2}\right\}\right)\right) \\
&= -\frac{1}{2}\text{vec}\Big(\text{diag}\Big(\boldsymbol{\tau}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\Big)\Big) = -\frac{1}{2}\text{vec}\left(\text{diag}\Big(\frac{\mathbf{1}_{K_j}}{\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}}\Big)\right) = -\frac{1}{2}\text{vec}\left(\boldsymbol{\Sigma}^{-1}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right) \quad (4.34)\\
&= -\frac{1}{2}\left(\mu_{\mathfrak{q}(1/\sigma_{\varepsilon}^2)}\mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\omega}_{Z_j}+\mu_{\mathfrak{q}\left(1/\sigma_{uj}^2\right)}\mu_{\mathfrak{q}\left(b_{uj}\right)}\mathbf{1}_{K_j}\right).
\end{aligned}
$$

Then

$$
\boldsymbol{\tau}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} = \mu_{\mathfrak{q}(1/\sigma_{\varepsilon}^2)}\mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\omega}_{Z_j}+\mu_{\mathfrak{q}\left(1/\sigma_{uj}^2\right)}\mu_{\mathfrak{q}\left(b_{uj}\right)}\mathbf{1}_{K_j} = \frac{\mathbf{1}_{K_j}}{\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}}.
$$

Therefore

$$
\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} = \mathbf{1}_{K_j}\Big/\left\{\mu_{\mathfrak{q}(1/\sigma_{\varepsilon}^2)}\mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\omega}_{Z_j}+\mu_{\mathfrak{q}\left(1/\sigma_{uj}^2\right)}\mu_{\mathfrak{q}\left(b_{uj}\right)}\mathbf{1}_{K_j}\right\}
$$

and

$$
\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} = \text{diag}\left(\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right) = \text{diag}\left[\mathbf{1}_{K_j}\Big/\left\{\mu_{\mathfrak{q}(1/\sigma_{\varepsilon}^2)}\mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\omega}_{Z_j}+\mu_{\mathfrak{q}\left(1/\sigma_{uj}^2\right)}\mu_{\mathfrak{q}\left(b_{uj}\right)}\mathbf{1}_{K_j}\right\}\right]
$$

are given in Algorithm 7 and (4.19) of Section 4.2.2 respectively.
If we define $\boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{u}_{j'}\right)} \equiv \mu_{\mathfrak{q}\left(\gamma_{uj'}\right)}\boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{u}}_{j'}\right)}$, then

$$
\begin{aligned}
E_{\mathfrak{q}}\left(\boldsymbol{\eta_1}\right) &= E_{\mathfrak{q}}\left\{\left(\frac{1}{\sigma_{\varepsilon}^2}\right)\left(\gamma_{uj}\left[\boldsymbol{Z}_j^T\boldsymbol{y}-\boldsymbol{Z}_j^T\boldsymbol{X}\left(\boldsymbol{\gamma}_{\beta}\odot\widetilde{\boldsymbol{\beta}}\right)-\sum_{j'\neq j}^{d_\bullet}\boldsymbol{Z}_j^T\boldsymbol{Z}_{j'}\left(\gamma_{uj'}\widetilde{\boldsymbol{u}}_{j'}\right)\right]\right)\right\} \\
&= \mu_{\mathfrak{q}(1/\sigma_{\varepsilon}^2)}\left(\mu_{\mathfrak{q}\left(\gamma_{u\,j}\right)}\left[\boldsymbol{Z}_j^T\boldsymbol{y}-\boldsymbol{Z}_j^T\boldsymbol{X}\left(\boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{\gamma}_{\beta}\right)}\odot\boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\right)-\sum_{j'\neq j}^{d_\bullet}\boldsymbol{Z}_j^T\boldsymbol{Z}_{j'}\boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{u}_{j'}\right)}\right]\right) \quad (4.35)\\
&= \boldsymbol{\Sigma}^{-1}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} = \text{diag}\Big(\frac{\mathbf{1}_{K_j}}{\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}}\Big)\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}.
\end{aligned}
$$

In (4.35), if we define

$$\boldsymbol{\omega}_{16} \equiv \boldsymbol{Z}_j^T \boldsymbol{y} - \boldsymbol{Z}_j^T \boldsymbol{X} \left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \right) - \sum_{j' \neq j}^{d_\bullet} \boldsymbol{Z}_j^T \boldsymbol{Z}_{j'} \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{j'})},$$

then

$$E_{\mathfrak{q}} \left(\boldsymbol{\eta_1} \right) = \mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)} \mu_{\mathfrak{q}(\gamma_{uj})} \boldsymbol{\omega}_{16}.$$

Therefore

$$\begin{aligned} \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} &= \operatorname{diag}\Big(\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} \Big) E_{\mathfrak{q}} \left(\boldsymbol{\eta_1} \right) = \operatorname{diag}\Big(\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} \Big) \mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)} \mu_{\mathfrak{q}(\gamma_{uj})} \boldsymbol{\omega}_{16} \\ &= \mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)} \mu_{\mathfrak{q}(\gamma_{uj})} \left(\boldsymbol{\omega}_{16} \odot \boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} \right) \end{aligned}$$

is given in (4.20) and Algorithm 7 of Section 4.2.2.

4.A.4 The q-Density of $\gamma_{\beta j}$ with Laplace-Zero Priors

First note that, for $1 \leq j \leq d_\circ + d_\bullet$,

$$\log \left\{ \mathfrak{q} \left(\gamma_{\beta j} \right) \right\} = E_{\mathfrak{q}} \left[\log \left\{ \mathfrak{p} \left(\gamma_{\beta j} \middle| \text{rest} \right) \right\} \right] + \text{const} = E_{\mathfrak{q}}(\eta_{\beta j}) \gamma_{\beta j} + \text{const}.$$

Recalling that (2.41) in Section 2.5.3 to be

$$\eta_{\beta j} = \text{logit}(\rho_\beta) - \frac{1}{2\sigma_\varepsilon^2} \left(\widetilde{\beta}_j^2 \left\| \boldsymbol{X}_j \right\|^2 - 2 \widetilde{\beta}_j \boldsymbol{X}_j^T \left[\boldsymbol{y} - \boldsymbol{X}_{-j} \left\{ (\boldsymbol{\gamma}_\beta)_{-j} \odot (\widetilde{\boldsymbol{\beta}})_{-j} \right\} - \sum_{j'=1}^{d_\bullet} \boldsymbol{Z}_{j'} (\gamma_{uj'} \widetilde{\boldsymbol{u}}_{j'}) \right] \right),$$

and note that under product density restriction (4.5) in Section 4.2.1, we have

$$\sum_{j'=1}^{d_\bullet} \boldsymbol{X}_j^T \boldsymbol{Z} j' E_{\mathfrak{q}} \left(\gamma_{uj'} \widetilde{\boldsymbol{u}}_{j'} \right) = \sum_{j'=1}^{d_\bullet} \boldsymbol{X}_j^T \boldsymbol{Z}_{j'} \left(\mu_{\mathfrak{q}(\gamma_{uj'})} \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_{j'})} \right).$$

If we define $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{j'})} \equiv \mu_{\mathfrak{q}(\gamma_{uj'})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_{j'})}$, then

$$\begin{aligned} E_{\mathfrak{q}}\left(\eta_{\beta_j}\right) &= \text{logit}\left(\rho_\beta\right) - \frac{1}{2}\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\Bigg\{\left(\mu^2_{\mathfrak{q}(\widetilde{\beta}_j)} + \sigma^2_{\mathfrak{q}(\widetilde{\beta}_j)}\right)\boldsymbol{X}_j^T\boldsymbol{X}_j \\ &\qquad -2\Bigg[\mu_{\mathfrak{q}(\widetilde{\beta}_j)}\left(\boldsymbol{X}_j^T\boldsymbol{y} - \sum_{j'=1}^{d_\bullet}\boldsymbol{X}_j^T\boldsymbol{Z}_{j'}\left(\mu_{\mathfrak{q}(\gamma_{uj'})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_{j'})}\right)\right) \\ &\qquad -\boldsymbol{X}_j^T\boldsymbol{X}_{-j}E_{\mathfrak{q}}\Big(\left(\boldsymbol{\gamma}_\beta\right)_{-j} \odot \left(\widetilde{\beta}_j(\widetilde{\boldsymbol{\beta}})_{-j}\right)\Big)\Bigg]\Bigg\} \\ &= \text{logit}\left(\rho_\beta\right) - \frac{1}{2}\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)} \\ &\quad \times\Bigg\{\left(\mu^2_{\mathfrak{q}(\widetilde{\beta}_j)} + \sigma^2_{\mathfrak{q}(\widetilde{\beta}_j)}\right)\boldsymbol{X}_j^T\boldsymbol{X}_j - 2\Bigg[\mu_{\mathfrak{q}(\widetilde{\beta}_j)}\Big(\boldsymbol{X}_j^T\boldsymbol{y} - \sum_{j'=1}^{d_\bullet}\boldsymbol{X}_j^T\boldsymbol{Z}_{j'}\big(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{j'})}\big)\Big) \\ &\quad -\boldsymbol{X}_j^T\boldsymbol{X}_{-j}\Big(\big(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}\big)_{-j} \odot \big((\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})})_{-j,j} + \mu_{\mathfrak{q}(\widetilde{\beta}_j)}(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})})_{-j}\big)\Big)\Bigg]\Bigg\} \end{aligned} \tag{4.36}$$

where $\left(\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right)_{-j,j}$ denotes the jth column of $\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}$ with the jth row removed.

In (4.36), if we define

$$\omega_{15-1} \equiv \boldsymbol{X}_j^T\boldsymbol{y} - \sum_{j'=1}^{d_\bullet}\boldsymbol{X}_j^T\boldsymbol{Z}_{j'}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{j'})}\right) \tag{4.37}$$

and further define

$$\begin{aligned} \omega_{15} \equiv \mu_{\mathfrak{q}(\widetilde{\beta}_j)}\left(\boldsymbol{X}_j^T\boldsymbol{y} - \sum_{j'=1}^{d_\bullet}\boldsymbol{X}_j^T\boldsymbol{Z}_{j'}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{j'})}\right)\right) - \boldsymbol{X}_j^T\boldsymbol{X}_{-j}\Bigg\{\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}\right)_{-j} \odot \Bigg(\left(\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right)_{-j,j} \\ +\mu_{\mathfrak{q}(\widetilde{\beta}_j)}\left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right)_{-j}\Bigg)\Bigg\}, \end{aligned}$$

then

$$\omega_{15} = \mu_{\mathfrak{q}(\widetilde{\beta}_j)}\omega_{15-1} - \boldsymbol{X}_j^T\boldsymbol{X}_{-j}\left(\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}\right)_{-j} \odot \left(\left(\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right)_{-j,j} + \mu_{\mathfrak{q}(\widetilde{\beta}_j)}\left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right)_{-j}\right)\right). \tag{4.38}$$

Therefore

$$E_{\mathfrak{q}}\left(\eta_{\beta_j}\right) = \text{logit}\left(\rho_\beta\right) - \frac{1}{2}\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\Bigg\{\left(\mu^2_{\mathfrak{q}\left(\widetilde{\beta}_j\right)} + \sigma^2_{\mathfrak{q}\left(\widetilde{\beta}_j\right)}\right)\boldsymbol{X}_j^T\boldsymbol{X}_j$$

$$-2\Bigg[\mu_{\mathfrak{q}\left(\widetilde{\beta}_j\right)}\Big\{\boldsymbol{X}_j^T\boldsymbol{y} - \sum_{j'=1}^{d_\bullet}\boldsymbol{X}_j^T\boldsymbol{Z}_{j'}\left(\mu_{\mathfrak{q}\left(\gamma_{uj'}\right)}\boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{u}}_{j'}\right)}\right)\Big\}$$

$$-\boldsymbol{X}_j^T\boldsymbol{X}_{-j}\Big\{\left(\boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{\gamma}_\beta\right)}\right)_{-j} \odot \left(\left(\boldsymbol{\Sigma}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\right)_{-j,j} + \mu_{\mathfrak{q}\left(\widetilde{\beta}_j\right)}\left(\boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\right)_{-j}\right)\Big\}\Bigg]\Bigg\} \quad (4.39)$$

$$= \text{logit}\left(\rho_\beta\right) - \frac{1}{2}\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\Bigg\{\left(\mu^2_{\mathfrak{q}\left(\widetilde{\beta}_j\right)} + \sigma^2_{\mathfrak{q}\left(\widetilde{\beta}_j\right)}\right)\boldsymbol{X}_j^T\boldsymbol{X}_j - 2\omega_{15}\Bigg\}. \quad (4.40)$$

(4.37), (4.38) and (4.40) are given in Algorithm 7. (4.39) is obtained to be (4.18) in Section 4.2.2. In summary, for $1 \leq j \leq d_\circ + d_\bullet$, $\mu_{\mathfrak{q}(\gamma_{\beta j})} = \text{expit}(E_{\mathfrak{q}}\left(\eta_{\beta_j}\right))$,

$$\mathfrak{q}\left(\gamma_{\beta j}\right) \text{ is the Bernoulli } \left(\frac{1}{1+\exp\left\{-E_{\mathfrak{q}}\left(\eta_{\beta_j}\right)\right\}}\right) \text{ probability mass function.}$$

4.A.5 The q-Density of γ_{uj} with Laplace-Zero Priors

First note that, for $1 \leq j \leq d_\bullet$,

$$\log\left\{\mathfrak{q}\left(\gamma_{uj}\right)\right\} = E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\gamma_{uj}\middle|\text{rest}\right)\right\}\right] + \text{const} = E_{\mathfrak{q}}(\eta_{uj})\gamma_{uj} + \text{const}.$$

Recalling that (2.46) in Appendix 2.B.5 to be

$$\eta_{uj} = \text{logit}(\rho_u) - \frac{1}{2\sigma_\varepsilon^2}\Big(\boldsymbol{\omega}_{\boldsymbol{Z}_j}^T\left(\widetilde{\boldsymbol{u}}_j \odot \widetilde{\boldsymbol{u}}_j\right) - 2\widetilde{\boldsymbol{u}}_j^T\boldsymbol{Z}_j^T\Big[\boldsymbol{y} - \boldsymbol{X}\left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}\right) - \sum_{j'\neq j}^{d_\bullet}\boldsymbol{Z}_{j'}(\gamma_{uj'}\widetilde{\boldsymbol{u}}_{j'})\Big]\Big),$$

and note that $\boldsymbol{Z}_j^T\boldsymbol{Z}_j$ is a diagonal matrix from (2.8), if we define $\boldsymbol{Z}_j^T\boldsymbol{Z}_j \equiv \text{diag}\left(\boldsymbol{\omega}_{\boldsymbol{Z}_j}\right)$,

$$\begin{aligned}
E_{\mathfrak{q}}\left\{\boldsymbol{\omega}_{\boldsymbol{Z}_j}^T\left(\widetilde{\boldsymbol{u}}_j \odot \widetilde{\boldsymbol{u}}_j\right)\right\} &= E_{\mathfrak{q}}\left\{\widetilde{\boldsymbol{u}}_j^T\text{diag}\left(\boldsymbol{\omega}_{\boldsymbol{Z}_j}\right)\widetilde{\boldsymbol{u}}_j\right\} \\
&= \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}^T\text{diag}\left(\boldsymbol{\omega}_{\boldsymbol{Z}_j}\right)\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} + \text{tr}\left\{\text{diag}\left(\boldsymbol{\omega}_{\boldsymbol{Z}_j}\right)\text{Cov}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right\} \\
&= \boldsymbol{\omega}_{\boldsymbol{Z}_j}^T\left(\text{diagonal}\left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}^T\right) + \text{diagonal}\left(\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right)\right) \\
&= \boldsymbol{\omega}_{\boldsymbol{Z}_j}^T\left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} + \boldsymbol{\sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}^2\right)
\end{aligned}$$

then

$$\begin{aligned}
E_{\mathfrak{q}}\left(\eta_{uj}\right) = \text{logit}\left(\rho_u\right) - \frac{1}{2}\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\Bigg\{&\boldsymbol{\omega}_{\boldsymbol{Z}_j}^T\left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} + \boldsymbol{\sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}^2\right) - 2\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{u}_j)}^T\Bigg[\boldsymbol{Z}_j^T\boldsymbol{y} \\
&-\boldsymbol{Z}_j^T\boldsymbol{X}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right) - \sum_{j'\neq j}^{d_\bullet}\boldsymbol{Z}_j^T\boldsymbol{Z}_{j'}\left(\mu_{\mathfrak{q}(\gamma_{uj'})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_{j'})}\right)\Bigg]\Bigg\}. \qquad (4.41)
\end{aligned}$$

In (4.41), if we define $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{j'})} \equiv \mu_{\mathfrak{q}(\gamma_{uj'})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_{j'})}$, and define

$$\boldsymbol{\omega}_{18} \equiv \boldsymbol{Z}_j^T\boldsymbol{y} - \boldsymbol{Z}_j^T\boldsymbol{X}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right) - \sum_{j'\neq j}^{d_\bullet}\boldsymbol{Z}_j^T\boldsymbol{Z}_{j'}\left(\mu_{\mathfrak{q}(\gamma_{uj'})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_{j'})}\right),$$

then $\boldsymbol{\omega}_{18} = \boldsymbol{Z}_j^T\boldsymbol{y} - \boldsymbol{Z}_j^T\boldsymbol{X}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right) - \sum_{j'\neq j}^{d_\bullet}\boldsymbol{Z}_j^T\boldsymbol{Z}_{j'}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{j'})}\right)$ is given in Algorithm 7 in Section 4.2.2. In (4.41), we further define

$$\begin{aligned}
\omega_{19} \equiv \boldsymbol{\omega}_{\boldsymbol{Z}_j}^T\left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} + \boldsymbol{\sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}^2\right) - 2\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{u}_j)}^T\Bigg[&\boldsymbol{Z}_j^T\boldsymbol{y} - \boldsymbol{Z}_j^T\boldsymbol{X}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right) \\
&-\sum_{j'\neq j}^{d_\bullet}\boldsymbol{Z}_j^T\boldsymbol{Z}_{j'}\left(\mu_{\mathfrak{q}(\gamma_{uj'})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_{j'})}\right)\Bigg].
\end{aligned}$$

Then

$$\omega_{19} = \boldsymbol{\omega}_{\boldsymbol{Z}_j}^T\left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} + \boldsymbol{\sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}^2\right) - 2\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{u}_j)}^T\boldsymbol{\omega}_{18}$$

and

$$E_{\mathfrak{q}}\left(\eta_{uj}\right) = \text{logit}\left(\rho_u\right) - \frac{1}{2}\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\Bigg\{\boldsymbol{\omega}_{\boldsymbol{Z}_j}^T\left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} + \boldsymbol{\sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}^2\right) - 2\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{u}_j)}^T$$
$$\times\left[\boldsymbol{Z}_j^T\boldsymbol{y} - \boldsymbol{Z}_j^T\boldsymbol{X}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right) - \sum_{j'\neq j}^{d_\bullet}\boldsymbol{Z}_j^T\boldsymbol{Z}_{j'}\left(\boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{u}_{j'}\right)}\right)\right]\Bigg\}$$
$$= \text{logit}\left(\rho_u\right) - \frac{1}{2}\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\Bigg\{\boldsymbol{\omega}_{\boldsymbol{Z}_j}^T\left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} + \boldsymbol{\sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}^2\right) - 2\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{u}_j)}^T\boldsymbol{\omega}_{18}\Bigg\}$$
$$= \text{logit}\left(\rho_u\right) - \frac{1}{2}\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\omega_{19}$$

are both given in Algorithm 7 in Section 4.2.2. In summary

$$\mathfrak{q}(\gamma_{uj}) \text{ is the Bernoulli}\left(\frac{1}{1+\exp\left(-E_{\mathfrak{q}(\eta_{uj})}\right)}\right) \text{probability mass function.}$$

$$E_{\mathfrak{q}}\left(\eta_{uj}\right) = \text{logit}\left(\rho_u\right) - \frac{1}{2}\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\Bigg\{\boldsymbol{\omega}_{\boldsymbol{Z}_j}^T\left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} + \boldsymbol{\sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}^2\right)$$
$$-2\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{u}_j)}^T\boldsymbol{Z}_j^T\left[\boldsymbol{y} - \boldsymbol{X}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right) - \sum_{j'\neq j}^{d_\bullet}\boldsymbol{Z}_{j'}\left(\mu_{\mathfrak{q}\left(\gamma_{uj'}\right)}\boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{u}}_{j'}\right)}\right)\right]\Bigg\}$$

obtained to be (4.26) of Section 4.2.2.

4.A.6 The q-Density of σ_ε^2 with Laplace-Zero Priors

For Gaussian response, recalling that (2.47) in Section 2.5.3 to be

$$\sigma_\varepsilon^2\big|\text{rest} \sim \text{Inverse-Gamma}\left(\frac{n+1}{2}, \frac{1}{a_\varepsilon} + \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{1}_n\beta_0 - \boldsymbol{X}(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j)\right\|^2\right),$$

then $\mathfrak{q}\left(\sigma_\varepsilon^2\right)$ has an Inverse-Gamma$\left(\kappa_{\mathfrak{q}(\sigma_\varepsilon^2)}, \lambda_{\mathfrak{q}(\sigma_\varepsilon^2)}\right)$density function. where

$$\kappa_{\mathfrak{q}(\sigma_\varepsilon^2)} = \frac{n+1}{2}$$

and

$$\lambda_{\mathfrak{q}(\sigma_\varepsilon^2)} = \mu_{\mathfrak{q}(1/a_\varepsilon)} + \frac{1}{2} E_{\mathfrak{q}} \left\{ \left\| \boldsymbol{y} - \mathbf{1}_n \beta_0 - \boldsymbol{X}(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j (\gamma_{uj} \widetilde{\boldsymbol{u}}_j) \right\|^2 \right\}. \tag{4.42}$$

Now note that, in (4.42), we have

$$\begin{aligned} & E_{\mathfrak{q}} \left\{ \left\| \boldsymbol{y} - \mathbf{1}_n \beta_0 - \boldsymbol{X}(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j (\gamma_{uj} \widetilde{\boldsymbol{u}}_j) \right\|^2 \right\} \\ & = \left\| \boldsymbol{y} - \mathbf{1}_n \beta_0 - \boldsymbol{X} \left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \right) - \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j \left(\mu_{\mathfrak{q}(\gamma_{uj})} \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} \right) \right\|^2 \\ & + \text{tr} \left[\text{Cov}_{\mathfrak{q}} \left\{ \boldsymbol{y} - \mathbf{1}_n \beta_0 - \boldsymbol{X} \left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}} \right) - \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j \left(\gamma_{uj} \widetilde{\boldsymbol{u}}_j \right) \right\} \right]. \end{aligned} \tag{4.43}$$

Recalling that the product density restriction in (4.5) in Section 4.2.1 to be

$$\mathfrak{q} \left(\beta_0, \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{u}}_1, \ldots, \widetilde{\boldsymbol{u}}_{d_\bullet} \right) = \mathfrak{q} \left(\beta_0 \right) \mathfrak{q} \left(\widetilde{\boldsymbol{\beta}} \right) \mathfrak{q} \left(\widetilde{\boldsymbol{u}}_1 \right) \ldots \mathfrak{q} \left(\widetilde{\boldsymbol{u}}_{d_\bullet} \right),$$

then the second term of (4.43) becomes

$$\begin{aligned} & \text{tr} \left[\text{Cov}_{\mathfrak{q}} \left\{ \boldsymbol{y} - \mathbf{1}_n \beta_0 - \boldsymbol{X} \left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}} \right) - \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j \left(\gamma_{uj} \widetilde{\boldsymbol{u}}_j \right) \right\} \right] \\ & = \text{tr} \left\{ \text{Cov}_{\mathfrak{q}} \left(\mathbf{1}_n \beta_0 \right) \right\} + \text{tr} \left\{ \text{Cov}_{\mathfrak{q}} \left\{ \boldsymbol{X} \left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}} \right) \right\} \right\} + \sum_{j=1}^{d_\bullet} \text{tr} \left\{ \text{Cov}_{\mathfrak{q}} \left\{ \boldsymbol{Z}_j \left(\gamma_{uj} \widetilde{\boldsymbol{u}}_j \right) \right\} \right\} \\ & = \text{tr} \left\{ \mathbf{1}_n \text{Cov}_{\mathfrak{q}} \left(\beta_0 \right) \mathbf{1}_n^T \right\} + \text{tr} \left\{ \boldsymbol{X} \text{Cov}_{\mathfrak{q}} \left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}} \right) \boldsymbol{X}^T \right\} + \sum_{j=1}^{d_\bullet} \text{tr} \left\{ \boldsymbol{Z}_j \text{Cov}_{\mathfrak{q}} \left(\gamma_{uj} \widetilde{\boldsymbol{u}}_j \right) \boldsymbol{Z}_j^T \right\} \\ & = n \sigma_{\mathfrak{q}(\beta_0)}^2 + \text{tr} \left\{ \boldsymbol{X}^T \boldsymbol{X} \text{Cov}_{\mathfrak{q}} \left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}} \right) \right\} + \sum_{j=1}^{d_\bullet} \text{tr} \left\{ \boldsymbol{Z}_j^T \boldsymbol{Z}_j \text{Cov}_{\mathfrak{q}} \left(\gamma_{uj} \widetilde{\boldsymbol{u}}_j \right) \right\}. \end{aligned} \tag{4.44}$$

Let us look at the second term of (4.44). Starting with $\text{Cov}_{\mathfrak{q}} \left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}} \right)$. We define

$$\boldsymbol{\Omega}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \equiv E_{\mathfrak{q}} \left(\boldsymbol{\gamma}_\beta \boldsymbol{\gamma}_\beta^T \right) = \text{Cov}_{\mathfrak{q}} \left(\boldsymbol{\gamma}_\beta \right) + E_{\mathfrak{q}} \left(\boldsymbol{\gamma}_\beta \right) E_{\mathfrak{q}} \left(\boldsymbol{\gamma}_\beta^T \right).$$

Then

$$\begin{aligned}
\mathrm{Cov}_{\mathfrak{q}}\Big(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}\Big) &= \mathrm{Cov}_{\mathfrak{q}}\left(\boldsymbol{\gamma}_\beta\right) \odot \boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} + \left\{\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}^T\right\} \odot \boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \\
&\quad + \left\{\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}^T\right\} \odot \mathrm{Cov}_{\mathfrak{q}}\left(\boldsymbol{\gamma}_\beta\right) \\
&= \left(\boldsymbol{\Omega}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}^T\right) \odot \boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} + \left\{\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}^T\right\} \odot \boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \\
&\quad + \left\{\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}^T\right\} \odot \left(\boldsymbol{\Omega}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}^T\right) \\
&= \boldsymbol{\Omega}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \left\{\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} + \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}^T\right\} - \left\{\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right\}\left\{\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right\}^T,
\end{aligned}$$

and so the second term of (4.44) becomes

$$\begin{aligned}
\mathrm{tr}\left\{\boldsymbol{X}^T\boldsymbol{X}\mathrm{Cov}_{\mathfrak{q}}\Big(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}\Big)\right\} &= \mathrm{tr}\left\{\boldsymbol{X}^T\boldsymbol{X}\left(\boldsymbol{\Omega}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \left\{\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} + \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}^T\right\}\right)\right\} \\
&\quad - \mathrm{tr}\left\{\boldsymbol{X}^T\boldsymbol{X}\left\{\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right\}\left\{\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right\}^T\right\}. \qquad (4.45)
\end{aligned}$$

Next, let us look at the third term of (4.44). Starting with $\mathrm{Cov}_{\mathfrak{q}}\Big(\gamma_{uj}\widetilde{\boldsymbol{u}}_j\Big)$.

$$\mathrm{Cov}_{\mathfrak{q}}\Big(\gamma_{uj}\widetilde{\boldsymbol{u}}_j\Big) = \mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} + \mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}^T - \left(\mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right)\left(\mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}^T\right).$$

and so the third term of (4.44) becomes

$$\begin{aligned}
&\mathrm{tr}\left\{\boldsymbol{Z}_j^T\boldsymbol{Z}_j\mathrm{Cov}_{\mathfrak{q}}\Big(\gamma_{uj}\widetilde{\boldsymbol{u}}_j\Big)\right\} \\
&\quad = \mathrm{tr}\left\{\boldsymbol{Z}_j^T\boldsymbol{Z}_j\left(\mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right)\right\} + \mathrm{tr}\left[\boldsymbol{Z}_j^T\boldsymbol{Z}_j\left\{\mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}^T\right\}\right] \\
&\qquad - \mathrm{tr}\left[\boldsymbol{Z}_j^T\boldsymbol{Z}_j\left(\mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right)\left(\mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}^T\right)\right]. \qquad (4.46)
\end{aligned}$$

Recalling from (4.35) that

$$\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} = \mathrm{diag}\left(\boldsymbol{\sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}^2\right)$$

and from (2.8) note that

$$\boldsymbol{\omega}_{\boldsymbol{Z}_j} \equiv \text{diagonal}(\boldsymbol{Z}_j^T \boldsymbol{Z}_j),$$

then the first term of (4.46) becomes

$$\begin{aligned}\text{tr}\left\{\boldsymbol{Z}_j^T\boldsymbol{Z}_j\left(\mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right)\right\} &= \text{tr}\left\{\text{diag}\left(\boldsymbol{\omega}_{\boldsymbol{Z}_j}\right)\mu_{\mathfrak{q}(\gamma_{uj})}\text{diag}\left(\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right)\right\} \\ &= \mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\omega}_{\boldsymbol{Z}_j}^T\text{diagonal}\left\{\text{diag}\left(\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right)\right\} = \mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\omega}_{\boldsymbol{Z}_j}^T\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}. \end{aligned} \tag{4.47}$$

The second term of (4.46) becomes

$$\begin{aligned}\text{tr}\left[\boldsymbol{Z}_j^T\boldsymbol{Z}_j\left\{\mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\boldsymbol{\mu}^T_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right\}\right] &= \mu_{\mathfrak{q}(\gamma_{uj})}\text{tr}\left[\text{diag}\left(\boldsymbol{\omega}_{\boldsymbol{Z}_j}\right)\left\{\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\boldsymbol{\mu}^T_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right\}\right] \\ &= \mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\omega}_{\boldsymbol{Z}_j}^T\text{diagonal}\left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\boldsymbol{\mu}^T_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right) = \mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\omega}_{\boldsymbol{Z}_j}^T\left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\odot\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right). \end{aligned} \tag{4.48}$$

The third term of (4.46) becomes

$$\text{tr}\left\{\boldsymbol{Z}_j^T\boldsymbol{Z}_j\left(\mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right)\left(\mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right)^T\right\} = \mu_{\mathfrak{q}(\gamma_{uj})}\mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\omega}_{\boldsymbol{Z}_j}^T\left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\odot\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right). \tag{4.49}$$

After obtaining (4.47)–(4.49), we have (4.46) equalling

$$\begin{aligned}\text{tr}\left\{\boldsymbol{Z}_j^T\boldsymbol{Z}_j\text{Cov}_{\mathfrak{q}}\left(\gamma_{uj}\widetilde{\boldsymbol{u}}_j\right)\right\} &= \mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\omega}_{\boldsymbol{Z}_j}^T\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} + \mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\omega}_{\boldsymbol{Z}_j}^T\left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\odot\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right) - \mu_{\mathfrak{q}(\gamma_{uj})} \\ &\quad\times\mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\omega}_{\boldsymbol{Z}_j}^T\left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\odot\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right) \\ &= \boldsymbol{\omega}_{\boldsymbol{Z}_j}^T\left[\mu_{\mathfrak{q}(\gamma_{uj})}\left\{\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} + \left(1-\mu_{\mathfrak{q}(\gamma_{uj})}\right)\left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\odot\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right)\right\}\right]. \end{aligned} \tag{4.50}$$

In summary, for Gaussian responses, the second term of (4.44) becomes (4.45), and

the third term of (4.44) becomes (4.50), then we have (4.42) being

$$\begin{aligned}\lambda_{\mathfrak{q}(\sigma_\varepsilon^2)} = \mu_{\mathfrak{q}(1/a_\varepsilon)} &+ \frac{1}{2}\left\|\boldsymbol{y} - \mathbf{1}_n^T\beta_0 - \boldsymbol{X}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right) - \sum_{d=1}^{d_\bullet}\boldsymbol{Z}_j\left(\mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right)\right\|^2 \\ &+\frac{1}{2}n\sigma^2_{\mathfrak{q}(\beta_0)} + \frac{1}{2}\text{tr}\left\{\boldsymbol{X}^T\boldsymbol{X}\left(\boldsymbol{\Omega}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \left\{\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} + \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\boldsymbol{\mu}^T_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right\}\right)\right\} \\ &-\frac{1}{2}\text{tr}\left\{\boldsymbol{X}^T\boldsymbol{X}\left\{\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right\}\left\{\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right\}^T\right\} \\ &+\frac{1}{2}\sum_{j=1}^{d_\bullet}\boldsymbol{\omega}^T_{\boldsymbol{Z}_j}\left[\mu_{\mathfrak{q}(\gamma_{uj})}\left\{\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} + \left(1-\mu_{\mathfrak{q}(\gamma_{uj})}\right)\left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right)\right\}\right],\end{aligned}$$

which is given in (4.28) and Algorithm 7 in Section 4.2.2. Then $\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)} = \kappa_{\mathfrak{q}(\sigma_\varepsilon^2)}/\lambda_{\mathfrak{q}(\sigma_\varepsilon^2)}$.

For $1 \le i \le n$, if we define

$$\boldsymbol{\omega}_{20} \equiv \boldsymbol{\mu}_{\mathfrak{q}} = \mathbf{1}_n\mu_{\mathfrak{q}(\beta_0)} + \boldsymbol{X}(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}) + \sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j(\mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}),$$

in mean field variational Bayes, we have

$$(\boldsymbol{\mu}_{\mathfrak{q}})_i = \mu_{\mathfrak{q}(\beta_0)} + \left[\boldsymbol{X}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right) + \sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j\left(\mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right)\right]_i.$$

For binary responses, $1 \le i \le n$, if we define $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{c})}$ as a $n \times 1$ vector, recalling that the full conditional distribution of auxiliary parameter c_i in (2.70) of Appendix 2.B.6 to be

$$c_i\big|\text{rest} \sim \begin{cases}\text{Truncated-Normal}_+((\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{c})})_i, 1), & y_i = 1; \\ \text{Truncated-Normal}_-((\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{c})})_i, 1), & y_i = 0\end{cases}$$

where

$$\begin{aligned}(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{c})})_i &= (\boldsymbol{\mu}_{\mathfrak{q}})_i + \frac{\phi((\boldsymbol{\mu}_{\mathfrak{q}})_i)}{\Phi((\boldsymbol{\mu}_{\mathfrak{q}})_i)}, \quad y_i = 1 \\ (\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{c})})_i &= (\boldsymbol{\mu}_{\mathfrak{q}})_i - \frac{\phi(-(\boldsymbol{\mu}_{\mathfrak{q}})_i)}{\Phi(-(\boldsymbol{\mu}_{\mathfrak{q}})_i)}, \quad y_i = 0,\end{aligned}$$

using **Result 1.7.1** and **Result 1.7.2**.

In summary, for $1 \leq i \leq n, y_i = 1$ or $y_i = 0$, if define $\zeta'(x) \equiv \dfrac{\phi(x)}{\Phi(x)}$, then we obtain (4.31) of Section 4.2.2 to be

$$\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{c})}\right)_i = (\boldsymbol{\mu}_{\mathfrak{q}})_i + (2y_i - 1)\,\frac{\phi\left((2y_i-1)(\boldsymbol{\mu}_{\mathfrak{q}})_i\right)}{\Phi\left((2y_i-1)(\boldsymbol{\mu}_{\mathfrak{q}})_i\right)} = (\boldsymbol{\omega}_{20})_i + (2y_i - 1)\zeta'((2y_i-1)(\boldsymbol{\omega}_{20})_i)$$

equivalent to

$$\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{c})} = \boldsymbol{\omega}_{20} + (2\boldsymbol{y} - \mathbf{1}_n) \odot \zeta'\left((2\boldsymbol{y} - \mathbf{1}_n) \odot \boldsymbol{\omega}_{20}\right). \tag{4.51}$$

(4.51) is given in Algorithm 7 of Section 4.2.2 for binary responses.

4.A.7 The $\mathfrak{q}$-Density of a_ε with Laplace-Zero Priors

First note that

$$\log\left\{\mathfrak{q}\left(a_\varepsilon\right)\right\} = E_{\mathfrak{q}}\left\{\log\left(\mathfrak{p}\left(a_\varepsilon\middle|\text{rest}\right)\right)\right\} + \text{const.}$$

Recalling that $a_\varepsilon|$rest in (2.48) of Section 2.5.3,

$$E_{\mathfrak{q}}\left\{\log\left(\mathfrak{p}\left(a_\varepsilon\middle|\text{rest}\right)\right)\right\} = (-1-1)\log\left(a_\varepsilon\right) + \frac{1}{a_\varepsilon}E_{\mathfrak{q}}\left(\frac{1}{\sigma_\varepsilon^2} + \frac{1}{s_\varepsilon^2}\right) + \text{const.}$$

Therefore

$$\mathfrak{q}\left(a_\varepsilon\right) \sim \text{Inverse-Gamma}\left(\kappa_{\mathfrak{q}(a_\varepsilon)}, \lambda_{\mathfrak{q}(a_\varepsilon)}\right),$$

where

$$\kappa_{\mathfrak{q}(a_\varepsilon)} = 1, \;\; \lambda_{\mathfrak{q}(a_\varepsilon)} = \mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)} + 1/s_\varepsilon^2 \;\text{ and }\; \mu_{\mathfrak{q}(1/a_\varepsilon)} = \kappa_{\mathfrak{q}(a_\varepsilon)}/\lambda_{\mathfrak{q}(a_\varepsilon)}$$

given in (6.80), (4.30) and Algorithm 7 in Section 4.2.2.

4.A.8 The $\mathfrak{q}$-Density of $b_{\beta j}$ with Laplace-Zero Priors

First note that

$$\log\left\{\mathfrak{q}\left(b_{\beta j}\right)\right\} = E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(b_{\beta j}\middle|\text{rest}\right)\right\}\right] + \text{const}, \quad 1 \leq j \leq d_\circ + d_\bullet.$$

Recalling that $\mathfrak{p}(b_{\beta j}|\text{rest})$ in (2.71) of Appendix 2.B.8, we have

$$\log\left\{\mathfrak{p}\left(b_{\beta j}\middle|\text{rest}\right)\right\} = -\frac{3}{2}\log\left(b_{\beta j}\right) - \left[\frac{1}{2}b_{\beta j}\left(\frac{\widetilde{\beta}_j}{\sigma_\beta}\right)^2 + \frac{1}{2}\frac{1}{b_{\beta j}}\right] + \text{const.}$$

Using **Result 1.6.1**, we have

$$\mathfrak{q}\left(b_{\beta j}\right) \sim \text{Inverse-Gaussian}\left(E_{\mathfrak{q}}\left(\frac{\sigma_\beta}{|\widetilde{\beta}_j|}\right), 1\right).$$

Then the $\mathfrak{q}$-density means in (4.13) of Section 4.2.2 are obtained to be

$$\mu_{\mathfrak{q}\left(b_{\beta j}\right)} = E_{\mathfrak{q}}\left(\frac{\sigma_\beta}{|\widetilde{\beta}_j|}\right) = \left[E_{\mathfrak{q}}\left\{\left(\frac{\widetilde{\beta}_j}{\sigma_\beta}\right)^2\right\}\right]^{-1/2} = \left\{\mu_{\mathfrak{q}\left(1/\sigma_\beta^2\right)}\left(\mu^2_{\mathfrak{q}\left(\widetilde{\beta}_j\right)} + \sigma^2_{\mathfrak{q}\left(\widetilde{\beta}_j\right)}\right)\right\}^{-1/2}.$$

Therefore

$$\mathfrak{q}\left(b_{\beta_1}, \ldots, b_{\beta_{(d_\circ + d_\bullet)}}\right)$$

is a product of

$$\text{Inverse-Gaussian}\left(\left\{\mu_{\mathfrak{q}\left(1/\sigma_\beta^2\right)}\left(\mu^2_{\mathfrak{q}\left(\widetilde{\beta}_j\right)} + \sigma^2_{\mathfrak{q}\left(\widetilde{\beta}_j\right)}\right)\right\}^{-1/2}, 1\right) \text{ density functions}$$

over $1 \leq j \leq d_\circ + d_\bullet$. If define $\boldsymbol{\omega}_{14} \equiv \boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)} \odot \boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)} + \text{diagnal}\left(\boldsymbol{\Sigma}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\right)$, then

$$\boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{b}_\beta\right)} = \left[\mu_{\mathfrak{q}\left(1/\sigma_\beta^2\right)}\left\{\boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)} \odot \boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)} + \text{diagnal}\left(\boldsymbol{\Sigma}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\right)\right\}\right]^{-1/2} = \left(\mu_{\mathfrak{q}\left(1/\sigma_\beta^2\right)}\boldsymbol{\omega}_{14}\right)^{-1/2}$$

as given in Algorithm 7 of Section 4.2.2.

4.A.9 The $\mathfrak{q}$-Density of b_{uj} with Laplace-Zero Priors

First note that

$$\log\left\{\mathfrak{q}\left(b_{uj}\right)\right\} = E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(b_{uj}\middle|\text{rest}\right)\right\}\right] + \text{const}, \quad 1 \leq j \leq d_\bullet.$$

Recalling that $\mathfrak{p}(b_{uj}|\text{rest})$ in (2.72) of Appendix 2.B.9, we have

$$\log\left(\mathfrak{p}\left(b_{uj}\middle|\text{rest}\right)\right) = -\frac{3}{2}\log\left(b_{uj}\right) - \left[\frac{1}{2}b_{uj}\left(\frac{\left\|\widetilde{\boldsymbol{u}}_j\right\|}{\sigma_{uj}}\right)^2 + \frac{1}{2}\frac{1}{b_{uj}}\right] + \text{const.}$$

Using **Result 1.6.1**, we have

$$\mathfrak{q}\left(b_{uj}\right) \sim \text{Inverse-Gaussian}\left(E_{\mathfrak{q}}\left(\frac{\sigma_{uj}}{\left\|\widetilde{\boldsymbol{u}}_j\right\|}\right), 1\right).$$

Recalling that (4.35) to be $\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)} = \text{diag}\left(\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right)$, we have

$$E_{\mathfrak{q}}\left\{\left(\frac{\left\|\widetilde{\boldsymbol{u}}_j\right\|}{\sigma_{uj}}\right)^2\right\} = \mu_{\mathfrak{q}\left(1/\sigma^2_{uj}\right)}\left[\left\|E_{\mathfrak{q}}\left(\widetilde{\boldsymbol{u}}_j\right)\right\|^2 + \text{tr}\Big\{\text{Cov}_{\mathfrak{q}}\left(\widetilde{\boldsymbol{u}}_j\right)\Big\}\right].$$

Note that

$$\text{tr}\left\{\text{Cov}_{\mathfrak{q}}\left(\widetilde{\boldsymbol{u}}_j\right)\right\} = \text{tr}\left\{\text{diagonal}\left(\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right)\right\} = \text{tr}\left\{\text{diagonal}\left(\text{diag}\left(\boldsymbol{\sigma}^2_{(\widetilde{\boldsymbol{u}}_j)}\right)\right)\right\} = \mathbf{1}^T_{K_j}\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}.$$

Then

$$E_{\mathfrak{q}}\left(\frac{\sigma_{uj}}{\left\|\widetilde{\boldsymbol{u}}_j\right\|}\right) = \left[E_{\mathfrak{q}}\left\{\left(\frac{\left\|\widetilde{\boldsymbol{u}}_j\right\|}{\sigma_{uj}}\right)^2\right\}\right]^{-1/2} = \left\{\mu_{\mathfrak{q}\left(1/\sigma^2_{uj}\right)}\left[\left\|\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right\|^2 + \mathbf{1}^T_{K_j}\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right]\right\}^{-1/2}.$$

If define $\omega_{17} \equiv \left\|\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right\|^2 + \mathbf{1}^T_{K_j}\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}$, then $\mu_{\mathfrak{q}(b_{uj})} = E_{\mathfrak{q}}\left(\frac{\sigma_{uj}}{\left\|\widetilde{\boldsymbol{u}}_j\right\|}\right) = \left(\mu_{\mathfrak{q}\left(1/\sigma^2_{uj}\right)}\omega_{17}\right)^{-1/2}$ is given in Algorithm 7 of Section 4.2.2. Therefore

$$\mathfrak{q}\left(b_{u_1}, \ldots, b_{u_{(d_\bullet)}}\right)$$

is a product of

$$\text{Inverse-Gaussian}\left(\left\{\mu_{\mathfrak{q}\left(1/\sigma^2_{uj}\right)}\left[\left\|\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right\|^2 + \mathbf{1}^T_{K_j}\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right]\right\}^{-1/2}, 1\right) \text{ density functions}$$

over $1 \leq j \leq d_\bullet$, obtained to be (4.21) in Section 4.2.2.

4.A.10 The q-Density of σ_β^2 with Laplace-Zero Priors

First note that

$$\log\left\{\mathfrak{q}\left(\sigma_\beta^2\right)\right\} = E_\mathfrak{q}\left[\log\left\{\mathfrak{p}\left(\sigma_\beta^2|\text{rest}\right)\right\}\right] + \text{const.}$$

Recalling the result for $\sigma_\beta^2|$rest in (2.39) of Section 2.5.3, we have

$$E_\mathfrak{q}\left\{\log\left(\mathfrak{p}\left(\sigma_\beta^2|\text{rest}\right)\right)\right\} = \left(-\frac{d_\circ + d_\bullet + 1}{2} - 1\right)\log\left(\sigma_\beta^2\right) - \frac{1}{\sigma_\beta^2}E_\mathfrak{q}\left(\frac{\widetilde{\boldsymbol{\beta}}^T \text{diag}\left(\boldsymbol{b}_\beta\right)\widetilde{\boldsymbol{\beta}}}{2} + \frac{1}{a_\beta}\right).$$

Therefore

$$\mathfrak{q}\left(\sigma_\beta^2\right) \text{ has an Inverse-Gamma } \left(\kappa_{\mathfrak{q}\left(\sigma_\beta^2\right)}, \lambda_{\mathfrak{q}\left(\sigma_\beta^2\right)}\right) \text{ density function.}$$

And note that

$$\begin{aligned} E_\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}^T \text{diag}\left(\boldsymbol{b}_\beta\right)\widetilde{\boldsymbol{\beta}}\right) &= \boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}^T \text{diag}\left(\boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{b}_\beta\right)}\right)\boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)} + \text{tr}\left\{\text{diag}\left(\boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{b}_\beta\right)}\right)\boldsymbol{\Sigma}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\right\} \\ &= \boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{b}_\beta\right)}^T \text{diagonal}\left(\boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}^T\right) + \boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{b}_\beta\right)}^T \text{diagonal}\left(\boldsymbol{\Sigma}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\right) \\ &= \boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{b}_\beta\right)}^T \left\{\boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)} \odot \boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)} + \text{diagonal}\left(\boldsymbol{\Sigma}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\right)\right\}. \end{aligned}$$

We have

$$\kappa_{\mathfrak{q}\left(\sigma_\beta^2\right)} = \frac{d_\circ + d_\bullet + 1}{2}$$

and

$$\lambda_{\mathfrak{q}\left(\sigma_\beta^2\right)} = \frac{1}{2}\boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{b}_\beta\right)}^T \left\{\boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)} \odot \boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)} + \text{diagonal}\left(\boldsymbol{\Sigma}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\right)\right\} + \mu_{\mathfrak{q}\left(1/a_\beta\right)}$$

given in (4.14) and (4.15) of Section 4.2.2. Recalling that

$$\boldsymbol{\omega}_{14} \equiv \boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)} \odot \boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)} + \text{diagnal}\left(\boldsymbol{\Sigma}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\right)$$

in Appendix 4.A.8, then

$$\lambda_{\mathfrak{q}\left(\sigma_\beta^2\right)} = \mu_{\mathfrak{q}\left(1/a_\beta\right)} + \frac{1}{2}\boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{b}_\beta\right)}^T\boldsymbol{\omega}_{14} \quad \text{and} \quad \mu_{\mathfrak{q}(1/\sigma_\beta^2)} = \kappa_{\mathfrak{q}\left(\sigma_\beta^2\right)}/\lambda_{\mathfrak{q}\left(\sigma_\beta^2\right)}$$

given in Algorithm 7 in Section 4.2.2.

4.A.11 The q-Density of a_β with Laplace-Zero Priors

First note that

$$\log\left\{\mathfrak{q}\left(a_\beta\right)\right\} = E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(a_\beta\middle|\text{rest}\right)\right\}\right] + \text{const.}$$

Following the steps in Appendix 4.A.7, we have

$$\mathfrak{q}\left(a_\beta\right) \text{ has an Inverse-Gamma } \left(\kappa_{\mathfrak{q}\left(a_\beta\right)}, \lambda_{\mathfrak{q}\left(a_\beta\right)}\right) \text{ density function}$$

where

$$\kappa_{\mathfrak{q}\left(a_\beta\right)} = 1, \;\; \lambda_{\mathfrak{q}\left(a_\beta\right)} = \mu_{\mathfrak{q}\left(1/\sigma_\beta^2\right)} + 1/s_\beta^2 \;\; \text{and} \;\; \mu_{\mathfrak{q}(1/a_\beta)} = \kappa_{\mathfrak{q}\left(a_\beta\right)}/\lambda_{\mathfrak{q}\left(a_\beta\right)}$$

are given in (4.16), (4.17) and Algorithm 7 of Section 4.2.2.

4.A.12 The q-Density of σ_{uj}^2 with Laplace-Zero Priors

First note that

$$\log\left\{\mathfrak{q}\left(\sigma_{uj}^2\right)\right\} = E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\sigma_{uj}^2\middle|\text{rest}\right)\right\}\right] + \text{const}, \quad 1 \le j \le d_\bullet.$$

Recalling that $\sigma_{uj}^2|\text{rest}$ in (2.44) of Section 2.5.3, we have

$$E_{\mathfrak{q}}\left\{\log\left(\mathfrak{p}\left(\sigma_{uj}^2\middle|\text{rest}\right)\right)\right\} = \left(-\frac{K_j+1}{2} - 1\right)\log\left(\sigma_{uj}^2\right) - \frac{1}{\sigma_{uj}^2}E_{\mathfrak{q}}\left(\frac{b_{uj}\left\|\widetilde{\boldsymbol{u}}_j\right\|^2}{2} + \frac{1}{a_{uj}}\right).$$

Note that

$$E_{\mathfrak{q}}\left(\left\|\widetilde{\boldsymbol{u}}_j\right\|^2\right) = \left\|\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right\|^2 + \text{tr}\left\{\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right\} = \left\|\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right\|^2 + \mathbf{1}_{K_j}^T\boldsymbol{\sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}^2.$$

Therefore

$$\mathfrak{q}\left(\sigma_{uj}^2\right) \text{ has an Inverse-Gamma } \left(\kappa_{\mathfrak{q}\left(\sigma_{uj}^2\right)}, \lambda_{\mathfrak{q}\left(\sigma_{uj}^2\right)}\right) \text{ density function}, \;\; 1 \le j \le d_\bullet,$$

where

$$\kappa_{\mathfrak{q}\left(\sigma_{uj}^2\right)} = \frac{K_j+1}{2} \;\; \text{and} \;\; \lambda_{\mathfrak{q}\left(\sigma_{uj}^2\right)} = \frac{1}{2}\mu_{\mathfrak{q}(b_{uj})}\left(\left\|\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right\|^2 + \mathbf{1}_{K_j}^T\boldsymbol{\sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}^2\right) + \mu_{\mathfrak{q}(1/a_{uj})}$$

are given in (4.22) and (4.23) of Section 4.2.2. Recalling that $\omega_{17} \equiv \left\|\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right\|^2 + \mathbf{1}_{K_j}^T \boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}$ in Appendix 4.A.9, we have

$$\lambda_{\mathfrak{q}\left(\sigma^2_{uj}\right)} = \mu_{\mathfrak{q}(1/a_{uj})} + \frac{1}{2}\mu_{\mathfrak{q}(b_{uj})}\omega_{17} \quad \text{and} \quad \mu_{\mathfrak{q}(1/\sigma^2_{uj})} = \kappa_{\mathfrak{q}\left(\sigma^2_{uj}\right)} \big/ \lambda_{\mathfrak{q}\left(\sigma^2_{uj}\right)}$$

given in Algorithm 7 in Section 4.2.2.

4.A.13 The $\mathfrak{q}$-Density of a_{uj} with Laplace-Zero Priors

First note that

$$\log\left\{\mathfrak{q}\left(a_{uj}\right)\right\} = E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(a_{uj}\middle|\text{rest}\right)\right\}\right] + \text{const}, \quad 1 \leq j \leq d_{\bullet}.$$

Following the steps in Appendix 4.A.7, we have

$$\mathfrak{q}\left(a_{uj}\right) \text{ has an Inverse-Gamma } \left(\kappa_{\mathfrak{q}(a_{uj})}, \lambda_{\mathfrak{q}(a_{uj})}\right) \text{ density function}$$

where

$$\kappa_{\mathfrak{q}(a_{uj})} = 1, \;\; \lambda_{\mathfrak{q}(a_{uj})} = \mu_{\mathfrak{q}\left(1/\sigma^2_{uj}\right)} + 1/s_u^2 \;\; \text{and} \;\; \mu_{\mathfrak{q}(1/a_{uj})} = \kappa_{\mathfrak{q}(a_{uj})} \big/ \lambda_{\mathfrak{q}(a_{uj})}$$

given in (4.24), (4.25) and Algorithm 7 of Section 4.2.2.

4.B Marginal Log-Likelihood Expressions

A final aspect of Algorithm 7 is determination of good stopping criteria for the coordinate ascent scheme. As is common in the mean field variational Bayes literature we monitor relative increases in the approximate marginal log-likelihood, also known as the evidence lower bound, which is denote by $\log \underline{p}(\boldsymbol{y}; \mathfrak{q})$. Appendix 6.C contains an explicit expression for the approximate marginal log-likelihood for the Section 2.4 models under product restrictions (4.5).

4.B.1 Marginal Log-Likelihood Approximation

For the Gaussian responses, the mean field variational Bayes approximate log-likelihood is

$$\begin{aligned}\log \underline{p}(\boldsymbol{y}; \mathfrak{q}) = \log \underline{p}(\boldsymbol{y}; \mathfrak{q}, \mathsf{C}) &+ E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{y}|\beta_0, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u, \sigma_\varepsilon^2\right)\right\}\right] \\ &+ E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\sigma_\varepsilon^2|a_\varepsilon\right)\right\}\right] - E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\sigma_\varepsilon^2\right)\right\}\right] \\ &+ E_{\mathfrak{q}}\left[\log \mathfrak{p}\left\{(a_\varepsilon)\right\}\right] - E_{\mathfrak{q}}\left[\log \mathfrak{q}\left\{(a_\varepsilon)\right\}\right] \end{aligned} \tag{4.52}$$

For binary responses, the mean field variational Bayes approximate log-likelihood is

$$\begin{aligned}\log \underline{p}(\boldsymbol{y}; \mathfrak{q}) = \log \underline{p}(\boldsymbol{y}; \mathfrak{q}, \mathsf{C}) &+ E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{y}|\boldsymbol{c}\right)\right\}\right] \\ &+ E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{c}|\beta_0, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_u, \widetilde{\boldsymbol{u}}\right)\right\}\right] - E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\boldsymbol{c}\right)\right\}\right]. \end{aligned} \tag{4.53}$$

where

$$\begin{aligned}\log \underline{p}(\boldsymbol{y}; \mathfrak{q}, \mathsf{C}) = {} & E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\beta_0\right)\right\}\right] - E_{\mathfrak{q}}\left[\left\{\mathfrak{q}\left(\beta_0\right)\right\}\right] + E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{\gamma}_\beta\right)\right\}\right] - E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\boldsymbol{\gamma}_\beta\right)\right\}\right] \\ & + E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\widetilde{\boldsymbol{\beta}}|\boldsymbol{b}_\beta, \sigma_\beta^2\right)\right\}\right] - E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)\right\}\right] + E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{b}_\beta\right)\right\}\right] - E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\boldsymbol{b}_\beta\right)\right\}\right] \\ & + E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\sigma_\beta^2|a_\beta\right)\right\}\right] - E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\sigma_\beta^2\right)\right\}\right] + E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(a_\beta\right)\right\}\right] - E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(a_\beta\right)\right\}\right] \\ & + E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{\gamma}_u\right)\right\}\right] - E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\boldsymbol{\gamma}_u\right)\right\}\right] + E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\widetilde{\boldsymbol{u}}|\boldsymbol{b}_u, \boldsymbol{\sigma}_u^2\right)\right\}\right] - E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\widetilde{\boldsymbol{u}}\right)\right\}\right] \\ & + E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{b}_u\right)\right\}\right] - E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\boldsymbol{b}_u\right)\right\}\right] + E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{\sigma}_u^2|\boldsymbol{a}_u\right)\right\}\right] - E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\boldsymbol{\sigma}_u^2\right)\right\}\right] \\ & + E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{a}_u\right)\right\}\right] - E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\boldsymbol{a}_u\right)\right\}\right]. \end{aligned} \tag{4.54}$$

Here "C" signifies the fact that (4.54) is common to $\log \underline{p}(\boldsymbol{y}; \mathfrak{q})$ expressions for both Gaussian response case and binary response case.

Explicit expressions for $\log \underline{p}(\boldsymbol{y}; \mathfrak{q})$ in each case can be obtained by simplifying each of the $\mathfrak{q}$-density moment expressions. For example, the first term of (4.54) is

$$\begin{aligned} E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\beta_0\right)\right\}\right] &= -\frac{1}{2}\log\left(2\pi\right) - \frac{1}{2}\log\left(\sigma_{\beta_0}^2\right) - \frac{1}{2}E_{\mathfrak{q}}\left(\beta_0^2\right)\Big/\sigma_{\beta_0}^2 \\ &= -\frac{1}{2}\log\left(2\pi\right) - \frac{1}{2}\log\left(\sigma_{\beta_0}^2\right) - \frac{1}{2}\left\{\mu_{\mathfrak{q}(\beta_0)}^2 + \sigma_{\mathfrak{q}(\beta_0)}^2\right\}\Big/\sigma_{\beta_0}^2. \end{aligned}$$

The second term of (4.54) is

$$\begin{aligned} -E_{\mathfrak{p}}\left[\log\left\{\mathfrak{q}\left(\beta_0\right)\right\}\right] &= \frac{1}{2}\log\left(2\pi\right) + \frac{1}{2}\log\left(\sigma_{\mathfrak{q}(\beta_0)}^2\right) + \frac{1}{2}E_{\mathfrak{q}}\left[\left(\beta_0 - \mu_{\mathfrak{q}(\beta_0)}\right)^2\Big/\sigma_{\beta_0}^2\right] \\ &= \frac{1}{2}\left\{\log\left(2\pi + 1\right)\right\} + \frac{1}{2}\log\left(\sigma_{\mathfrak{q}(\beta_0)}^2\right). \end{aligned}$$

The third term of (4.54) is

$$\begin{aligned} E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{\gamma}_{\beta}\right)\right\}\right] &= \sum_{j=1}^{d_\circ + d_\bullet} E_{\mathfrak{q}}\left\{\gamma_{\beta j}\log\left(\rho_\beta\right) + \left(1 - \gamma_{\beta j}\right)\log\left(1 - \rho_\beta\right)\right\} \\ &= \operatorname{logit}\left(\rho_\beta\right)\sum_{j=1}^{d_\circ + d_\bullet}\mu_{\mathfrak{q}(\gamma_{\beta j})} + \left(d_\circ + d_\bullet\right)\log\left(1 - \rho_\beta\right). \end{aligned}$$

The fourth term of (4.54) is

$$\begin{aligned} -E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\boldsymbol{\gamma}_{\beta}\right)\right\}\right] &= -\sum_{j=1}^{d_\circ + d_\bullet} E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\gamma_{\beta j}\right)\right\}\right] \\ &= -\sum_{j=1}^{d_\circ + d_\bullet} E_{\mathfrak{q}}\left[\log\left\{\left(\mu_{\mathfrak{q}(\gamma_{\beta j})}\right)^{\gamma_{\beta j}}\left(1 - \mu_{\mathfrak{q}(\gamma_{\beta j})}\right)^{1-\gamma_{\beta j}}\right\}\right] \\ &= -\sum_{j=1}^{d_\circ + d_\bullet}\left[\mu_{\mathfrak{q}(\gamma_{\beta j})}\log\left(\mu_{\mathfrak{q}(\gamma_{\beta j})}\right) + \left(1 - \mu_{\mathfrak{q}(\gamma_{\beta j})}\right)\log\left(1 - \mu_{\mathfrak{q}(\gamma_{\beta j})}\right)\right]. \end{aligned}$$

The fifth term of (4.54) is

$$
\begin{aligned}
E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\widetilde{\boldsymbol{\beta}}\middle|\boldsymbol{b}_{\beta},\sigma_{\beta}^{2}\right)\right\}\right] &= \sum_{j=1}^{d_{\circ}+d_{\bullet}} E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\widetilde{\beta}_{j}\middle|b_{\beta j},\sigma_{\beta}^{2}\right)\right\}\right] \\
&= \sum_{j=1}^{d_{\circ}+d_{\bullet}}\left[-\frac{1}{2}\log\left(2\pi\right)-\frac{1}{2}E_{\mathfrak{q}}\left\{\log\left(\sigma_{\beta}^{2}\right)\right\}+\frac{1}{2}E_{\mathfrak{q}}\left\{\log\left(b_{\beta j}\right)\right\}-\frac{1}{2}\mu_{\mathfrak{q}\left(1/\sigma_{\beta}^{2}\right)}E_{\mathfrak{q}}\left(b_{\beta j}\widetilde{\beta}_{j}^{2}\right)\right] \\
&= -\frac{1}{2}\left(d_{\circ}+d_{\bullet}\right)\log\left(2\pi\right)-\frac{1}{2}\left(d_{\circ}+d_{\bullet}\right)E_{\mathfrak{q}}\left\{\log\left(\sigma_{\beta}^{2}\right)\right\}+\frac{1}{2}\sum_{j=1}^{d_{\circ}+d_{\bullet}}E_{\mathfrak{q}}\left\{\log\left(b_{\beta j}\right)\right\} \\
&\qquad -\frac{1}{2}\mu_{\mathfrak{q}\left(1/\sigma_{\beta}^{2}\right)}\sum_{j=1}^{d_{\circ}+d_{\bullet}}\mu_{\mathfrak{q}\left(b_{\beta j}\right)}\left(\mu_{\mathfrak{q}\left(\widetilde{\beta}_{j}\right)}^{2}+\sigma_{\mathfrak{q}\left(\widetilde{\beta}_{j}\right)}^{2}\right).
\end{aligned}
$$

The sixth term of (4.54) is

$$
\begin{aligned}
-E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)\right\}\right] &= \frac{1}{2}\log\left|2\pi\boldsymbol{\Sigma}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\right|+\frac{1}{2}E_{\mathfrak{q}}\left\{\left(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\right)^{T}\boldsymbol{\Sigma}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}^{-1}\left(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\right)\right\} \\
&= \frac{1}{2}\log\left\{\left(2\pi\right)^{d_{\circ}+d_{\bullet}}\left|\boldsymbol{\Sigma}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\right|\right\}+\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}^{-1}E_{\mathfrak{q}}\left\{\left(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\right)\left(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}^{T}\right)\right\}\right] \\
&= \frac{1}{2}\log\left\{\left(2\pi\right)^{d_{\circ}+d_{\bullet}}\left|\boldsymbol{\Sigma}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\right|\right\}+\frac{1}{2}\mathrm{tr}\left(\boldsymbol{I}_{d_{\circ}+d_{\bullet}}\right) \\
&= \frac{1}{2}\log\left|\boldsymbol{\Sigma}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\right|+\frac{1}{2}\left\{\left(d_{\circ}+d_{\bullet}\right)\log\left(2\pi\right)+\left(d_{\circ}+d_{\bullet}\right)\right\}.
\end{aligned}
$$

Using **Result 1.6.2**, the seventh term of (4.54) is

$$
\begin{aligned}
E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{b}_{\beta}\right)\right\}\right] &= \sum_{j=1}^{d_{\circ}+d_{\bullet}}\left[\log\left(\frac{1}{2}\right)-2E_{\mathfrak{q}}\left\{\log\left(b_{\beta j}\right)\right\}-\frac{1}{2}E_{\mathfrak{q}}\left(1\Big/b_{\beta j}\right)\right] \\
&= -\left(d_{\circ}+d_{\bullet}\right)\log\left(2\right)-2\sum_{j=1}^{d_{\circ}+d_{\bullet}}E_{\mathfrak{q}}\left\{\log\left(b_{\beta j}\right)\right\}-\frac{1}{2}\sum_{j=1}^{d_{\circ}+d_{\bullet}}\left[\left\{1\Big/\mu_{\mathfrak{q}\left(b_{\beta j}\right)}\right\}+1\right] \\
&= -\left(d_{\circ}+d_{\bullet}\right)\left\{\log\left(2\right)+\frac{1}{2}\right\}-\frac{1}{2}\sum_{j=1}^{d_{\circ}+d_{\bullet}}\left\{1\Big/\mu_{\mathfrak{q}\left(b_{\beta j}\right)}\right\}-2\sum_{j=1}^{d_{\circ}+d_{\bullet}}E_{\mathfrak{q}}\left\{\log\left(b_{\beta j}\right)\right\}.
\end{aligned}
$$

The eighth term of (4.54) is

$$-E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\boldsymbol{b}_{\beta}\right)\right\}\right] = -\sum_{j=1}^{d_{\circ}+d_{\bullet}}\left[-\frac{1}{2}\log\left(2\pi\right) - \frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(b_{\beta j}\right)\right\} - E_{\mathfrak{q}}\left\{\frac{\left(b_{\beta j}-\mu_{\mathfrak{q}\left(b_{\beta j}\right)}\right)^{2}}{2b_{\beta j}\mu^{2}_{\mathfrak{q}\left(b_{\beta j}\right)}}\right\}\right]$$

$$= -\sum_{j=1}^{d_{\circ}+d_{\bullet}}\left[-\frac{1}{2}\log\left(2\pi\right) - \frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(b_{\beta j}\right)\right\} - \frac{1}{2}E_{\mathfrak{q}}\left(b_{\beta j}\right)\Big/\mu^{2}_{\mathfrak{q}\left(b_{\beta j}\right)}\right.$$

$$\left. +1\Big/\mu_{\mathfrak{q}\left(b_{\beta j}\right)} - \frac{1}{2}E_{\mathfrak{q}}\left(1\Big/b_{\beta j}\right)\right]$$

$$= -\sum_{j=1}^{d_{\circ}+d_{\bullet}}\left[-\frac{1}{2}\log\left(2\pi\right) - \frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(b_{\beta j}\right)\right\} - 1\Big/\left\{2\mu_{\mathfrak{q}\left(b_{\beta j}\right)}\right\} + 1\Big/\mu_{\mathfrak{q}\left(b_{\beta j}\right)}\right.$$

$$\left. -\frac{1}{2}\left\{1\Big/\mu_{\mathfrak{q}\left(b_{\beta j}\right)}+1\right\}\right]$$

$$= -\sum_{j=1}^{d_{\circ}+d_{\bullet}}\left[-\frac{1}{2}\log\left(2\pi\right) - \frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(b_{\beta j}\right)\right\} - \frac{1}{2}\right]$$

$$= \frac{1}{2}\left(d_{\circ}+d_{\bullet}\right)\left\{\log\left(2\pi\right)+1\right\} + \frac{3}{2}\sum_{j=1}^{d_{\bullet}}E_{\mathfrak{q}}\log\left(b_{\beta j}\right).$$

The ninth term of (4.54) is

$$E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\sigma^{2}_{\beta}\middle|a_{\beta}\right)\right\}\right] = E_{\mathfrak{q}}\left[\log\left\{\left(1\Big/a_{\beta}\right)^{1/2}\Big/\Gamma\left(\frac{1}{2}\right) - \frac{3}{2}\log\left(\sigma^{2}_{\beta}\right) - 1\Big/\left(\sigma^{2}_{\beta}a_{\beta}\right)\right\}\right]$$

$$= -\frac{1}{2}\log\left(\pi\right) - \frac{1}{2}E_{\mathfrak{q}}\left\{\log\left(a_{\beta}\right)\right\} - \frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(\sigma^{2}_{\beta}\right)\right\} - \mu_{\mathfrak{q}\left(1/a_{\beta}\right)}\mu_{\mathfrak{q}\left(1/\sigma^{2}_{\beta}\right)}.$$

The tenth term of (4.54) is

$$-E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\sigma^{2}_{\beta}\right)\right\}\right] = -E_{\mathfrak{q}}\left[\frac{1}{2}\left(d_{\circ}+d_{\bullet}+1\right)\log\left(\lambda_{\mathfrak{q}\left(\sigma^{2}_{\beta}\right)}\right) - \log\left\{\Gamma\left(\frac{1}{2}\left(d_{\circ}+d_{\bullet}+1\right)\right)\right\}\right.$$

$$\left. -\left\{\frac{1}{2}\left(d_{\circ}+d_{\bullet}+1\right)+1\right\}\log\left(\sigma^{2}_{\beta}\right) - \lambda_{\mathfrak{q}\left(\sigma^{2}_{\beta}\right)}\Big/\sigma^{2}_{\beta}\right]$$

$$= -\frac{1}{2}\left(d_{\circ}+d_{\bullet}+1\right)\log\left(\lambda_{\mathfrak{q}\left(\sigma^{2}_{\beta}\right)}\right) + \log\left\{\Gamma\left(\frac{1}{2}\left(d_{\circ}+d_{\bullet}+1\right)\right)\right\}$$

$$+\left\{\frac{1}{2}\left(d_{\circ}+d_{\bullet}+1\right)+1\right\}E_{\mathfrak{q}}\left\{\log\left(\sigma^{2}_{\beta}\right)\right\} + \mu_{\mathfrak{q}\left(1/\sigma^{2}_{\beta}\right)}\lambda_{\mathfrak{q}\left(\sigma^{2}_{\beta}\right)}.$$

The eleventh term of (4.54) is

$$E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(a_\beta\right)\right\}\right] = E_{\mathfrak{q}}\left(\log\left[\left\{\left(1/s_\beta\right)\Big/\Gamma\left(\frac{1}{2}\right)a_\beta^{-3/2}\right\}\exp\left(1\Big/\left(s_\beta^2 a_\beta\right)\right)\right]\right)$$
$$= -\frac{1}{2}\log\left(\pi\right) - \log\left(s_\beta\right) - \frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(a_\beta\right)\right\} - \mu_{\mathfrak{q}\left(1/a_\beta\right)}\Big/s_\beta^2.$$

The twelfth term of (4.54) is

$$-E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(a_\beta\right)\right\}\right] = -E_{\mathfrak{q}}\left[\log\left\{\lambda_{\mathfrak{q}\left(a_\beta\right)}a_\beta^{-2}\exp\left(-\lambda_{\mathfrak{q}\left(a_\beta\right)}\Big/a_\beta\right)\right\}\right]$$
$$= -\log\left\{\lambda_{\mathfrak{q}\left(a_\beta\right)}\right\} + 2E_{\mathfrak{q}}\left\{\log\left(a_\beta\right)\right\} + \lambda_{\mathfrak{q}\left(a_\beta\right)}\mu_{\mathfrak{q}\left(1/a_\beta\right)}.$$

The thirteenth term of (4.54) is

$$E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{\gamma}_u\right)\right\}\right]$$
$$= \sum_{j=1}^{d_\bullet} E_{\mathfrak{q}}\left[\gamma_{uj}\left\{\log\left(\rho_u\right)\right\} + \left(1-\gamma_{uj}\right)\log\left(1-\rho_u\right)\right]$$
$$= \text{logit}\left(\rho_u\right)\sum_{j=1}^{d_\bullet}\mu_{\mathfrak{q}\left(\gamma_{uj}\right)} + d_\bullet \text{log}\left(1-\rho_u\right).$$

The fourteenth term of (4.54) is

$$-E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\boldsymbol{\gamma}_u\right)\right\}\right] = -\sum_{j=1}^{d_\bullet} E_{\mathfrak{q}}\left[\log\left\{\mu_{\mathfrak{q}\left(\gamma_{uj}\right)}^{\gamma_{uj}}\left(1-\mu_{\mathfrak{q}\left(\gamma_{uj}\right)}\right)^{1-\gamma_{uj}}\right\}\right]$$
$$= -\sum_{j=1}^{d_\bullet}\left[\mu_{\mathfrak{q}\left(\gamma_{uj}\right)}\log\left(\mu_{\mathfrak{q}\left(\gamma_{uj}\right)}\right) + \left\{1-\mu_{\mathfrak{q}\left(\gamma_{uj}\right)}\right\}\log\left(1-\mu_{\mathfrak{q}\left(\gamma_{uj}\right)}\right)\right].$$

The fifteenth term of (4.54) is

$$E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\widetilde{\boldsymbol{u}}\middle|\boldsymbol{b}_u,\boldsymbol{\sigma}_u^2\right)\right\}\right]=\sum_{j=1}^{d_\bullet}\left[-\frac{1}{2}E_{\mathfrak{q}}\left(\log\left|\left(2\pi\right)\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right|\right)-\frac{1}{2}E_{\mathfrak{q}}\left\{\widetilde{\boldsymbol{u}}_j^T\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}^{-1}\widetilde{\boldsymbol{u}}_j\right\}\right]$$

$$=-\frac{1}{2}\log\left(2\pi\right)\left(\sum_{j=1}^{d_\bullet}K_j\right)-\frac{1}{2}\sum_{j=1}^{d_\bullet}K_jE_{\mathfrak{q}}\left\{\log\left(\sigma_{uj}^2\right)\right\}+\frac{1}{2}\sum_{j=1}^{d_\bullet}K_jE_{\mathfrak{q}}\left\{\log\left(b_{uj}\right)\right\}$$

$$-\frac{1}{2}\sum_{j=1}^{d_\bullet}\mu_{\mathfrak{q}\left(1/\sigma_{uj}^2\right)}\mu_{\mathfrak{q}\left(b_{uj}\right)}\left(\left\|\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right\|^2+\mathbf{1}_{K_j}^T\boldsymbol{\sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}^2\right).$$

The sixteenth term of (4.54) is

$$-E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\widetilde{\boldsymbol{u}}\right)\right\}\right]$$

$$=\sum_{j=1}^{d_\bullet}\left[\frac{1}{2}K_j\log\left(2\pi\right)+\frac{1}{2}\sum_{k=1}^{K_j}\log\left(\sigma_{\mathfrak{q}\left(\widetilde{u}_{jk}\right)}^2\right)+\frac{1}{2}E_{\mathfrak{q}}\left\{\frac{\left(\widetilde{\boldsymbol{u}}_j-\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right)\odot\left(\widetilde{\boldsymbol{u}}_j-\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right)}{\boldsymbol{\sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}^2}\right\}\right]$$

$$=\frac{1}{2}\left\{\log\left(2\pi\right)+1\right\}\sum_{j=1}^{d_\bullet}K_j+\frac{1}{2}\sum_{j=1}^{d_\bullet}\sum_{k=1}^{K_j}\log\left(\sigma_{\mathfrak{q}\left(\widetilde{u}_{jk}\right)}^2\right).$$

The seventeenth term of (4.54) is

$$E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{b}_u\right)\right\}\right]=\sum_{j=1}^{d_\bullet}\left[\frac{1}{2}\left(K_j+1\right)\log\left(\frac{1}{2}\right)-\log\left\{\Gamma\left(\frac{1}{2}\left(K_j+1\right)\right)\right\}\right.$$

$$\left.-\left\{\frac{1}{2}\left(K_j+1\right)+1\right\}E_{\mathfrak{q}}\left\{\log(b_{uj})\right\}-\frac{1}{2}E_{\mathfrak{q}}\left(1/b_{uj}\right)\right]$$

$$=-\frac{1}{2}\left(d_\bullet+\sum_{j=1}^{d_\bullet}K_j\right)\log\left(2\right)-\sum_{j=1}^{d_\bullet}\log\left\{\Gamma\left(\frac{1}{2}\left(K_j+1\right)\right)\right\}$$

$$-\sum_{j=1}^{d_\bullet}\left\{\frac{1}{2}\left(K_j+1\right)+1\right\}E_{\mathfrak{q}}\left\{\log\left(b_{uj}\right)\right\}-\frac{1}{2}\sum_{j=1}^{d_\bullet}\left[\left\{1/\mu_{\mathfrak{q}(b_{uj})}\right\}+1\right]$$

$$=-\frac{1}{2}d_\bullet\left\{\log\left(2\right)+1\right\}-\frac{1}{2}\log\left(2\right)\sum_{j=1}^{d_\bullet}K_j-\sum_{j=1}^{d_\bullet}\log\left\{\Gamma\left(\frac{1}{2}\left(K_j+1\right)\right)\right\}$$

$$-\sum_{j=1}^{d_\bullet}\left\{\left(\frac{1}{2}\left(K_j+1\right)\right)+1\right\}E_{\mathfrak{q}}\left\{\log\left(b_{uj}\right)\right\}-\frac{1}{2}\sum_{j=1}^{d_\bullet}\left\{1/\mu_{\mathfrak{q}(b_{uj})}\right\}.$$

The eighteenth term of (4.54) is

$$-E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\boldsymbol{b}_u\right)\right\}\right] = -\sum_{j=1}^{d_\bullet}\left[-\frac{1}{2}\log\left(2\pi\right) - \frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(b_{uj}\right)\right\} - E_{\mathfrak{q}}\left\{\frac{\left(b_{uj}-\mu_{\mathfrak{q}(b_{uj})}\right)^2}{2b_{uj}\mu^2_{\mathfrak{q}(b_{uj})}}\right\}\right]$$

$$= -\sum_{j=1}^{d_\bullet}\left[-\frac{1}{2}\log\left(2\pi\right) - \frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(b_{uj}\right)\right\} - 1/\left\{2\mu_{\mathfrak{q}(b_{uj})}\right\} + 1/\mu_{\mathfrak{q}(b_{uj})}\right.$$

$$\left. -\frac{1}{2}\left\{1/\mu_{\mathfrak{q}(b_{uj})}+1\right\}\right]$$

$$= -\sum_{j=1}^{d_\bullet}\left[-\frac{1}{2}\log\left(2\pi\right) - \frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(b_{uj}\right)\right\} - \frac{1}{2}\right] = \frac{1}{2}d_\bullet\left\{\log\left(2\pi\right)+1\right\} + \frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(b_{uj}\right)\right\}.$$

The nineteenth term of (4.54) is

$$E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{\sigma}_u^2\middle|\boldsymbol{a}_u\right)\right\}\right]$$

$$= \sum_{j=1}^{d_\bullet}E_{\mathfrak{q}}\left[\log\left\{\left(1/a_{uj}\right)^{1/2}\Big/\Gamma\left(\frac{1}{2}\right)\right\} - \frac{3}{2}\log\left(\sigma^2_{uj}\right) - 1/\left(\sigma^2_{uj}a_{uj}\right)\right]$$

$$= -\frac{1}{2}d_\bullet\log\left(\pi\right) - \sum_{j=1}^{d_\bullet}\left[\frac{1}{2}E_{\mathfrak{q}}\left\{\log\left(a_{uj}\right)\right\} + \frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(\sigma^2_{uj}\right)\right\} - \mu_{\mathfrak{q}\left(1/a_{uj}\right)}\mu_{\mathfrak{q}\left(1/\sigma^2_{uj}\right)}\right].$$

The twentieth term of (4.54) is

$$-E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\boldsymbol{\sigma}_u^2\right)\right\}\right] = -\sum_{j=1}^{d_\bullet}E_{\mathfrak{q}}\left[\frac{1}{2}\left(K_j+1\right)\log\left(\lambda_{\mathfrak{q}\left(\sigma^2_{uj}\right)}\right) - \log\left\{\Gamma\left(\frac{1}{2}\left(K_j+1\right)\right)\right\}\right.$$

$$\left. -\left\{\frac{1}{2}\left(K_j+1\right)+1\right\}E_{\mathfrak{q}}\left\{\log\left(\sigma^2_{uj}\right)\right\} - \lambda_{\mathfrak{q}\left(\sigma^2_{uj}\right)}/\sigma^2_{uj}\right]$$

$$= -\frac{1}{2}\sum_{j=1}^{d_\bullet}\left(K_j+1\right)\log\left(\lambda_{\mathfrak{q}\left(\sigma^2_{uj}\right)}\right) + \sum_{j=1}^{d_\bullet}\log\left\{\Gamma\left(\frac{1}{2}\left(K_j+1\right)\right)\right\}$$

$$+\sum_{j=1}^{d_\bullet}\left[\left\{\frac{1}{2}\left(K_j+1\right)+1\right\}E_{\mathfrak{q}}\left\{\log\left(\sigma^2_{uj}\right)\right\} + \mu_{\mathfrak{q}\left(1/\sigma^2_{uj}\right)}\lambda_{\mathfrak{q}\left(\sigma^2_{uj}\right)}\right].$$

The twenty-first term of (4.54) is

$$E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{a}_u\right)\right\}\right]=\sum_{j=1}^{d_\bullet}E_{\mathfrak{q}}\left[\log\left\{\left(1/s_u\right)\Big/\Gamma\left(\frac{1}{2}\right)\right\}-\frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(a_{uj}\right)\right\}-E_{\mathfrak{q}}\left\{1/\left(a_{uj}s_u^2\right)\right\}\right]$$

$$=-\frac{1}{2}d_\bullet\log\left(\pi\right)-d_\bullet\log\left(s_u\right)-\frac{3}{2}\sum_{j=1}^{d_\bullet}E_{\mathfrak{q}}\left\{\log\left(a_{uj}\right)\right\}-\left(1/s_u^2\right)\sum_{j=1}^{d_\bullet}\mu_{\mathfrak{q}\left(1/a_{uj}\right)}.$$

The twenty-second term of (4.54) is

$$-E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\boldsymbol{a}_u\right)\right\}\right]=\sum_{j=1}^{d_\bullet}\left[-\log\left(\lambda_{\mathfrak{q}(a_{uj})}\right)+2E_{\mathfrak{q}}\left\{\log\left(a_{uj}\right)\right\}+\lambda_{\mathfrak{q}(a_{uj})}\mu_{\mathfrak{q}\left(1/a_{uj}\right)}\right].$$

4.B.2 Collection of Logarithmic Moments of Variance Parameters

In this subsection we collect the contributions to $\log\underline{p}\left(\boldsymbol{y};\mathfrak{q},\mathrm{C}\right)$ of each of the logarithmic $\mathfrak{q}$-density moments of variance parameters.

The $E_{\mathfrak{q}}\left\{\log\left(\sigma_\beta^2\right)\right\}$ Terms

The fifth term of (4.54) contains: $-\frac{1}{2}\left(d_\circ+d_\bullet\right)E_{\mathfrak{q}}\left\{\log\left(\sigma_\beta^2\right)\right\}$.

The ninth term of (4.54) contains: $-\frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(\sigma_\beta^2\right)\right\}$.

The tenth term of (4.54) contains: $\left\{\frac{1}{2}\left(d_\circ+d_\bullet+1\right)+1\right\}E_{\mathfrak{q}}\left\{\log\left(\sigma_\beta^2\right)\right\}$.

Since

$$-\frac{1}{2}\left(d_\circ+d_\bullet\right)-\frac{3}{2}+\frac{1}{2}\left(d_\circ+d_\bullet+1\right)+1=0,$$

the $E_{\mathfrak{q}}\left\{\log\left(\sigma_\beta^2\right)\right\}$ contribution to $\log\underline{p}\left(\boldsymbol{y};\mathfrak{q},\mathrm{C}\right)$ is zero.

The $E_{\mathfrak{q}}\left\{\log\left(\sigma_{uj}^2\right)\right\}$ Terms, $1\leq j\leq d_\bullet$

The fifteenth term of (4.54) contains: $-\frac{1}{2}\sum_{j=1}^{d_\bullet}K_jE_{\mathfrak{q}}\left\{\log\left(\sigma_{uj}^2\right)\right\}$.

The nineteenth term of (4.54) contains: $-\frac{3}{2}\sum_{j=1}^{d_\bullet}E_{\mathfrak{q}}\left\{\log\left(\sigma_{uj}^2\right)\right\}$.

The twentieth term of (4.54) contains: $\sum_{j=1}^{d_\bullet} \left\{ \frac{1}{2} (K_j + 1) + 1 \right\} E_{\mathfrak{q}} \left\{ \log \left(\sigma_{uj}^2 \right) \right\}$.

Since

$$-\frac{1}{2} K_j - \frac{3}{2} + \frac{1}{2} (K_j + 1) + 1 = 0,$$

the $E_{\mathfrak{q}} \left\{ \log \left(\sigma_{uj}^2 \right) \right\}$ contributions to $\log \underline{p} (\boldsymbol{y}; \mathfrak{q}, \mathrm{C})$ are zero for each $1 \leq j \leq d_\bullet$.

4.B.3 Collection of Logarithmic Moments of Variance Auxiliary Parameters

The $E_{\mathfrak{q}} \{ \log (a_\beta) \}$ Terms

The ninth term of (4.54) contains: $-\frac{1}{2} E_{\mathfrak{q}} \{ \log (a_\beta) \}$.

The eleventh term of (4.54) contains: $-\frac{3}{2} E_{\mathfrak{q}} \{ \log (a_\beta) \}$.

The twelfth term of (4.54) contains: $2 E_{\mathfrak{q}} \{ \log (a_\beta) \}$.

Since

$$-\frac{1}{2} - \frac{3}{2} + 2 = 0,$$

the $E_{\mathfrak{q}} \{ \log (a_\beta) \}$ contribution to $\log \underline{p} (\boldsymbol{y}; \mathfrak{q}, \mathrm{C})$ is zero.

The $E_{\mathfrak{q}} \{ \log (a_{uj}) \}$ Terms, $1 \leq j \leq d_\bullet$

The nineteenth term of (4.54) contains: $-\frac{1}{2} \sum_{j=1}^{d_\bullet} E_{\mathfrak{q}} \{ \log (a_{uj}) \}$.

The twenty-first term of (4.54) contains: $-\frac{3}{2} \sum_{j=1}^{d_\bullet} E_{\mathfrak{q}} \{ \log (a_{uj}) \}$.

The twenty-second term of (4.54) contains: $2 \sum_{j=1}^{d_\bullet} E_{\mathfrak{q}} \{ \log (a_{uj}) \}$.

Since

$$-\frac{1}{2} - \frac{3}{2} + 2 = 0,$$

the $E_{\mathfrak{q}} \{ \log (a_{uj}) \}$ contributions to $\log \underline{p} (\boldsymbol{y}; \mathfrak{q}, \mathrm{C})$ are zero for each $1 \leq j \leq d_\bullet$.

4.B.4 Collection of Logarithmic Moments of Laplace Auxiliary Parameters

The $E_{\mathfrak{q}}\left\{\log\left(b_{\beta j}\right)\right\}$ Terms, $1 \leq j \leq d_{\circ} + d_{\bullet}$

The fifth term of (4.54) contains: $\frac{1}{2}\sum_{j=1}^{d_{\circ}+d_{\bullet}} E_{\mathfrak{q}}\left\{\log\left(b_{\beta j}\right)\right\}$.

The seventh term of (4.54) contains: $-2\sum_{j=1}^{d_{\circ}+d_{\bullet}} E_{\mathfrak{q}}\left\{\log\left(b_{\beta j}\right)\right\}$.

The eighth term of (4.54) contains: $\frac{3}{2}\sum_{j=1}^{d_{\bullet}} E_{\mathfrak{q}}\left\{\log\left(b_{\beta j}\right)\right\}$.

Since

$$\frac{1}{2} - 2 + \frac{3}{2} = 0,$$

the $E_{\mathfrak{q}}\left\{\log\left(b_{\beta j}\right)\right\}$ contributions to $\log \underline{p}\left(\boldsymbol{y}; \mathfrak{q}, \mathrm{C}\right)$ are zero for each $1 \leq j \leq d_{\circ} + d_{\bullet}$.

The $E_{\mathfrak{q}}\left\{\log\left(b_{uj}\right)\right\}$ Terms, $1 \leq j \leq d_{\bullet}$

The fifteenth term of (4.54) contains: $\frac{1}{2}\sum_{j=1}^{d_{\bullet}} K_j E_{\mathfrak{q}}\left\{\log\left(b_{uj}\right)\right\}$.

The seventeenth term of (4.54) contains: $-\sum_{j=1}^{d_{\bullet}} \left\{\left(\frac{1}{2}\left(K_j + 1\right)\right) + 1\right\} E_{\mathfrak{q}}\left\{\log\left(b_{uj}\right)\right\}$.

The eighteenth term of (4.54) contains: $\frac{3}{2} E_{\mathfrak{q}}\left\{\log\left(b_{uj}\right)\right\}$.

Since

$$\frac{1}{2}K_j - \left\{\left(\frac{1}{2}\left(K_j + 1\right)\right) + 1\right\} + \frac{3}{2} = 0,$$

the $E_{\mathfrak{q}}\left\{\log\left(b_{uj}\right)\right\}$ contributions to $\log \underline{p}\left(\boldsymbol{y}; \mathfrak{q}, \mathrm{C}\right)$ are zero for each $1 \leq j \leq d_{\bullet}$.

4.B.5 Collection of Constants

The constant contributions to each of the terms of $\log \underline{p}\left(\boldsymbol{y}; \mathfrak{q}, \mathrm{C}\right)$ are as follows:

first	$-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left(\sigma^2_{\beta_0}\right)$	twelfth	0
second	$\frac{1}{2}\{\log(2\pi+1)\}$	thirteenth	$d_\bullet \log(1-\rho_u)$
third	$(d_\circ + d_\bullet)\log(1-\rho_\beta)$	fourteenth	0
fourth	0	fifteenth	$-\frac{1}{2}\log(2\pi)\left(\sum_{j=1}^{d_\bullet} K_j\right)$
fifth	$-\frac{1}{2}(d_\circ + d_\bullet)\log(2\pi)$	sixteenth	$\frac{1}{2}\{\log(2\pi)+1\}\sum_{j=1}^{d_\bullet} K_j$
sixth	$\frac{1}{2}\{(d_\circ + d_\bullet)\log(2\pi) + (d_\circ + d_\bullet)\}$	seventeenth	$-\frac{1}{2}d_\bullet\{\log(2)+1\}$ $-\frac{1}{2}\log(2)\sum_{j=1}^{d_\bullet} K_j$ $-\sum_{j=1}^{d_\bullet}\log\left\{\Gamma\left(\frac{1}{2}(K_j+1)\right)\right\}$
seventh	$-(d_\circ + d_\bullet)\left\{\log(2)+\frac{1}{2}\right\}$	eighteenth	$\frac{1}{2}d_\bullet\{\log(2\pi)+1\}$
eighth	$\frac{1}{2}(d_\circ + d_\bullet)\{\log(2\pi)+1\}$	nineteenth	$-\frac{1}{2}d_\bullet\log(\pi)$
ninth	$-\frac{1}{2}\log(\pi)$	twentieth	$\sum_{j=1}^{d_\bullet}\log\left\{\Gamma\left(\frac{1}{2}(K_j+1)\right)\right\}$
tenth	$\log\left\{\Gamma\left(\frac{1}{2}(d_\circ + d_\bullet + 1)\right)\right\}$	twenty-first	$-\frac{1}{2}d_\bullet\log(\pi) - d_\bullet\log(s_u)$
eleventh	$-\frac{1}{2}\log(\pi) - \log(s_\beta)$	twenty-second	0

Table 4.5: Constants that do not depend on any $\mathfrak{q}$-density parameters of (4.54), denoted as const_1.

Therefore

$$
\begin{aligned}
&\log \underline{p}\left(\boldsymbol{y}; \mathfrak{q}, \mathrm{C}\right) \\
&= \mathrm{const}_1 - \frac{1}{2}\left\{\mu^2_{\mathfrak{q}(\beta_0)} + \sigma^2_{\mathfrak{q}(\beta_0)}\right\} \Big/ \sigma^2_{\beta_0} + \frac{1}{2}\log\left(\sigma^2_{\mathfrak{q}(\beta_0)}\right) + \mathbf{logit}(\rho_\beta) \sum_{j=1}^{d_\circ + d_\bullet} \mu_{\mathfrak{q}(\gamma_{\beta j})} \\
&\quad - \sum_{j=1}^{d_\circ + d_\bullet} \left[\mu_{\mathfrak{q}\left(\gamma_{\beta j}\right)} \log\left(\mu_{\mathfrak{q}\left(\gamma_{\beta j}\right)}\right) + \left(1 - \mu_{\mathfrak{q}\left(\gamma_{\beta j}\right)}\right) \log\left(1 - \mu_{\mathfrak{q}\left(\gamma_{\beta j}\right)}\right)\right] \\
&\quad - \frac{1}{2}\mu_{\mathfrak{q}\left(1/\sigma^2_\beta\right)} \sum_{j=1}^{d_\circ + d_\bullet} \mu_{\mathfrak{q}\left(b_{\beta j}\right)} \left(\mu^2_{\mathfrak{q}\left(\widetilde{\beta}_j\right)} + \sigma^2_{\mathfrak{q}\left(\widetilde{\beta}_j\right)}\right) + \frac{1}{2}\log\left|\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right| \\
&\quad - \frac{1}{2}\sum_{j=1}^{d_\circ + d_\bullet} \left\{1 \Big/ \mu_{\mathfrak{q}\left(b_{\beta j}\right)}\right\} - \mu_{\mathfrak{q}\left(1/a_\beta\right)} \mu_{\mathfrak{q}\left(1/\sigma^2_\beta\right)} - \frac{1}{2}\left(d_\circ + d_\bullet + 1\right) \log\left(\lambda_{\mathfrak{q}\left(\sigma^2_\beta\right)}\right) \\
&\quad + \mu_{\mathfrak{q}\left(1/\sigma^2_\beta\right)} \lambda_{\mathfrak{q}\left(\sigma^2_\beta\right)} - \mu_{\mathfrak{q}\left(1/a_\beta\right)} \Big/ s^2_\beta + \lambda_{\mathfrak{q}\left(a_\beta\right)} \mu_{\mathfrak{q}\left(1/a_\beta\right)} - \log\left\{\lambda_{\mathfrak{q}\left(a_\beta\right)}\right\} \\
&\quad - \sum_{j=1}^{d_\bullet} \left[\mu_{\mathfrak{q}\left(\gamma_{uj}\right)} \log\left(\mu_{\mathfrak{q}\left(\gamma_{uj}\right)}\right) + \left\{1 - \mu_{\mathfrak{q}\left(\gamma_{uj}\right)}\right\} \log\left(1 - \mu_{\mathfrak{q}\left(\gamma_{uj}\right)}\right)\right] \\
&\quad + \mathbf{logit}(\rho_u) \sum_{j=1}^{d_\bullet} \mu_{\mathfrak{q}(\gamma_{uj})} - \frac{1}{2}\sum_{j=1}^{d_\bullet} \mu_{\mathfrak{q}\left(1/\sigma^2_{uj}\right)} \mu_{\mathfrak{q}\left(b_{uj}\right)} \left(\left\|\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right\|^2 + \mathbf{1}^T_{K_j} \boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right) \\
&\quad + \frac{1}{2}\sum_{j=1}^{d_\bullet} \sum_{k=1}^{K_j} \log\left(\boldsymbol{\sigma}^2_{\mathfrak{q}\left(\widetilde{u}_{jk}\right)}\right) - \frac{1}{2}\sum_{j=1}^{d_\bullet} \left\{1 \Big/ \mu_{\mathfrak{q}(b_{uj})}\right\} - \sum_{j=1}^{d_\bullet} \mu_{\mathfrak{q}\left(1/a_{uj}\right)} \mu_{\mathfrak{q}\left(1/\sigma^2_{uj}\right)} \\
&\quad - \frac{1}{2}\sum_{j=1}^{d_\bullet} \left(K_j + 1\right) \log\left(\lambda_{\mathfrak{q}\left(\sigma^2_{uj}\right)}\right) + \sum_{j=1}^{d_\bullet} \mu_{\mathfrak{q}\left(1/\sigma^2_{uj}\right)} \lambda_{\mathfrak{q}\left(\sigma^2_{uj}\right)} - \left(1/s^2_u\right) \sum_{j=1}^{d_\bullet} \mu_{\mathfrak{q}\left(1/a_{uj}\right)} \\
&\quad + \sum_{j=1}^{d_\bullet} \left\{\lambda_{\mathfrak{q}(a_{uj})} \mu_{\mathfrak{q}\left(1/a_{uj}\right)} - \log\left(\lambda_{\mathfrak{q}(a_{uj})}\right)\right\}.
\end{aligned}
$$

For Gaussian response, the second term of (4.52) is

$$\begin{aligned}
&E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{y}|\beta_0,\boldsymbol{\gamma}_\beta,\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{u}},\boldsymbol{\gamma}_u,\sigma_\varepsilon^2\right)\right\}\right]\\
&\quad=-\frac{1}{2}n\log\left(2\pi\right)-\frac{1}{2}nE_{\mathfrak{q}}\left\{\log\left(\sigma_\varepsilon^2\right)\right\}\\
&\qquad-\frac{1}{2}\mu_{\mathfrak{q}\left(1/\sigma_\varepsilon^2\right)}E_{\mathfrak{q}}\left\{\left\|\boldsymbol{y}-\mathbf{1}_n^T\beta_0-\boldsymbol{X}\left(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}}\right)-\sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j\left(\gamma_{uj}\widetilde{\boldsymbol{u}}_j\right)\right\|^2\right\}.
\end{aligned}$$

The third term of (4.52) is

$$\begin{aligned}
E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\sigma_\varepsilon^2|a_\varepsilon\right)\right\}\right]&=E_{\mathfrak{q}}\left[\log\left\{1/a_\varepsilon\right)^{1/2}\Big/\Gamma\left(1\right\}-\frac{3}{2}\log\left(\sigma_\varepsilon^2\right)-1/\left(\sigma_\varepsilon^2a_\varepsilon\right)\right]\\
&=-\frac{1}{2}E_{\mathfrak{q}}\left\{\log\left(a_\varepsilon\right)\right\}-\frac{1}{2}\log\left(\pi\right)-\frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(\sigma_\varepsilon^2\right)\right\}-\mu_{\mathfrak{q}\left(1/a_\varepsilon\right)}\mu_{\mathfrak{q}\left(1/\sigma_\varepsilon^2\right)}.
\end{aligned}$$

The fourth term of (4.52) is

$$\begin{aligned}
-E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\sigma_\varepsilon^2\right)\right\}\right]&=-E_{\mathfrak{q}}\left[\frac{1}{2}\left(n+1\right)\log\left(\lambda_{\mathfrak{q}(\sigma_\varepsilon^2)}\right)-\log\left\{\Gamma\left(\frac{1}{2}\left(n+1\right)\right)\right\}\right]\\
&\qquad-\left\{\frac{1}{2}\left(n+1\right)+1\right\}E_{\mathfrak{q}}\left\{\log\left(\sigma_\varepsilon^2\right)\right\}-\lambda_{\mathfrak{q}(\sigma_\varepsilon^2)}/\sigma_\varepsilon^2\Big]\\
&=-\frac{1}{2}\left(n+1\right)\log\left(\lambda_{\mathfrak{q}(\sigma_\varepsilon^2)}\right)+\log\left\{\Gamma\left(\frac{1}{2}\left(n+1\right)\right)\right\}\\
&\qquad+\left\{\frac{1}{2}\left(n+1\right)+1\right\}E_{\mathfrak{q}}\left\{\log\left(\sigma_\varepsilon^2\right)\right\}+\mu_{\mathfrak{q}\left(1/\sigma_\varepsilon^2\right)}\Bigg\{\mu_{\mathfrak{q}(1/a_\varepsilon)}\\
&\qquad+\frac{1}{2}E_{\mathfrak{q}}\left(\left\|\boldsymbol{y}-\mathbf{1}_n\beta_0-\boldsymbol{X}(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}})+\sum_{j=1}^{d_\bullet}\boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j)\right\|^2\right)\Bigg\}.
\end{aligned}$$

The fifth term of (4.52) is

$$E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(a_\varepsilon\right)\right\}\right]=-\frac{1}{2}\log\left(\pi\right)-\log\left(s_\varepsilon\right)-\frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(a_\varepsilon\right)\right\}-\mu_{\mathfrak{q}\left(1/a_\varepsilon\right)}\Big/s_\varepsilon^2.$$

The sixth term of (4.52) is

$$-E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(a_{\varepsilon}\right)\right\}\right]=-\log\left(\lambda_{\mathfrak{q}(a_{\varepsilon})}\right)+2E_{\mathfrak{q}}\left\{\log\left(a_{\varepsilon}\right)\right\}+\lambda_{\mathfrak{q}(a_{\varepsilon})}\mu_{\mathfrak{q}\left(1/a_{\varepsilon}\right)}.$$

Therefore, for Gaussian response, we have

$$\begin{aligned}
\log\underline{p}\left(\boldsymbol{y};\mathfrak{q}\right)=&\log\underline{p}\left(\boldsymbol{y};\mathfrak{q},\mathrm{C}\right)+E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{y}\middle|\beta_0,\boldsymbol{\gamma}_{\beta},\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{u}},\boldsymbol{\gamma}_{u},\sigma_{\varepsilon}^{2}\right)\right\}\right]\\
&+E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\sigma_{\varepsilon}^{2}\middle|a_{\varepsilon}\right)\right\}\right]-E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\sigma_{\varepsilon}^{2}\right)\right\}\right]\\
&+E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(a_{\varepsilon}\right)\right\}\right]-E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(a_{\varepsilon}\right)\right\}\right]\\
=&\log\underline{p}\left(\boldsymbol{y};\mathfrak{q},\mathrm{C}\right)-\frac{1}{2}n\log\left(2\pi\right)-\frac{1}{2}nE_{\mathfrak{q}}\left\{\log\left(\sigma_{\varepsilon}^{2}\right)\right\}\\
&-\frac{1}{2}\mu_{\mathfrak{q}\left(1/\sigma_{\varepsilon}^{2}\right)}E_{\mathfrak{q}}\left\{\left\|\boldsymbol{y}-\mathbf{1}_{n}^{T}\beta_0-\boldsymbol{X}\left(\boldsymbol{\gamma}_{\beta}\odot\widetilde{\boldsymbol{\beta}}\right)-\sum_{j=1}^{d_{\bullet}}\boldsymbol{Z}_{j}\left(\gamma_{uj}\widetilde{\boldsymbol{u}}_{j}\right)\right\|^{2}\right\}\\
&-\frac{1}{2}E_{\mathfrak{q}}\left\{\log\left(a_{\varepsilon}\right)\right\}-\frac{1}{2}\log\left(\pi\right)-\frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(\sigma_{\varepsilon}^{2}\right)\right\}-\mu_{\mathfrak{q}\left(1/a_{\varepsilon}\right)}\mu_{\mathfrak{q}\left(1/\sigma_{\varepsilon}^{2}\right)}\\
&-\frac{1}{2}\left(n+1\right)\log\left(\lambda_{\mathfrak{q}(\sigma_{\varepsilon}^{2})}\right)+\log\left\{\Gamma\left(\frac{1}{2}\left(n+1\right)\right)\right\}\\
&+\left\{\frac{1}{2}\left(n+1\right)+1\right\}E_{\mathfrak{q}}\left\{\log\left(\sigma_{\varepsilon}^{2}\right)\right\}+\mu_{\mathfrak{q}\left(1/\sigma_{\varepsilon}^{2}\right)}\mu_{\mathfrak{q}(1/a_{\varepsilon})}\\
&+\frac{1}{2}\mu_{\mathfrak{q}\left(1/\sigma_{\varepsilon}^{2}\right)}E_{\mathfrak{q}}\left(\left\|\boldsymbol{y}-\mathbf{1}_{n}\beta_0-\boldsymbol{X}(\boldsymbol{\gamma}_{\beta}\odot\widetilde{\boldsymbol{\beta}})+\sum_{j=1}^{d_{\bullet}}\boldsymbol{Z}_{j}(\gamma_{uj}\widetilde{\boldsymbol{u}}_{j})\right\|^{2}\right)\\
&-\frac{1}{2}\log\left(\pi\right)-\log\left(s_{\varepsilon}\right)-\frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(a_{\varepsilon}\right)\right\}-\mu_{\mathfrak{q}\left(1/a_{\varepsilon}\right)}\Big/s_{\varepsilon}^{2}\\
&-\log\left(\lambda_{\mathfrak{q}(a_{\varepsilon})}\right)+2E_{\mathfrak{q}}\left\{\log\left(a_{\varepsilon}\right)\right\}+\lambda_{\mathfrak{q}(a_{\varepsilon})}\mu_{\mathfrak{q}\left(1/a_{\varepsilon}\right)}\\
=&\log\underline{p}\left(\boldsymbol{y};\mathfrak{q},\mathrm{C}\right)-\frac{1}{2}\left(n+1\right)\log\left(\lambda_{\mathfrak{q}(\sigma_{\varepsilon}^{2})}\right)\\
&-\mu_{\mathfrak{q}\left(1/a_{\varepsilon}\right)}\Big/s_{\varepsilon}^{2}-\log\left(\lambda_{\mathfrak{q}(a_{\varepsilon})}\right)+\lambda_{\mathfrak{q}(a_{\varepsilon})}\mu_{\mathfrak{q}\left(1/a_{\varepsilon}\right)}+\mathrm{const}_{2}.
\end{aligned}$$

where const_2 are constants that do not depend on any $\mathfrak{q}$-density parameters.

In the binary response case, recalling that (4.53) to be

$$\log \underline{p}\left(\boldsymbol{y};\mathfrak{q}\right)=\log \underline{p}\left(\boldsymbol{y};\mathfrak{q},\mathsf{C}\right)+E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{y}\middle|\boldsymbol{c}\right)\right\}\right]$$
$$+E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{c}\middle|\beta_0,\boldsymbol{\gamma}_{\beta},\widetilde{\boldsymbol{\beta}},\boldsymbol{\gamma}_{u},\widetilde{\boldsymbol{u}}\right)\right\}\right]-E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\boldsymbol{c}\right)\right\}\right],$$

and recalling that

$$\mathfrak{p}(\boldsymbol{y}|\boldsymbol{c})=\prod_{i=1}^{n}I(c_i\geq 0)^{y_i}I(c_i<0)^{1-y_i},$$

then the second term of (4.53) becomes

$$E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{y}\middle|\boldsymbol{c}\right)\right\}\right]=\sum_{i=1}^{n}\left(y_iE_{\mathfrak{q}}\left[\log\left\{I(c_i\geq 0)\right\}\right]+(1-y_i)E_{\mathfrak{q}}\left[\log\left\{I(c_i<0)\right\}\right]\right).$$

The third term of (4.53) is

$$E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{c}\middle|\beta_0,\boldsymbol{\gamma}_{\beta},\widetilde{\boldsymbol{\beta}},\boldsymbol{\gamma}_{u},\widetilde{\boldsymbol{u}}\right)\right\}\right]$$
$$=-\frac{n}{2}\log(2\pi)-\frac{1}{2}E_{\mathfrak{q}}\left(\left\|\boldsymbol{c}-\mathbf{1}_n\beta_0-\boldsymbol{X}(\boldsymbol{\gamma}_{\beta}\odot\widetilde{\boldsymbol{\beta}})-\sum_{j=1}^{d_{\bullet}}\boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j)\right\|^2\right).$$

Recalling that

$$\log\left\{\mathfrak{q}\left(\boldsymbol{c}\right)\right\}=-\frac{n}{2}\log(2\pi)+\sum_{i=1}^{n}\Bigg[y_i\log\left\{I(c_i\geq 0)\right\}-y_i\log\left\{\Phi(\boldsymbol{\eta}_i)\right\}$$
$$+(1-y_i)\log\left\{I(c_i<0)\right\}-(1-y_i)log\left\{\Phi(-\boldsymbol{\eta}_i)\right\}\Bigg]$$
$$-\frac{1}{2}\left\|\boldsymbol{c}-\mathbf{1}_n\beta_0-\boldsymbol{X}(\boldsymbol{\gamma}_{\beta}\odot\widetilde{\boldsymbol{\beta}})-\sum_{j=1}^{d_{\bullet}}\boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j)\right\|^2$$

where

$$\boldsymbol{\eta}^{\mathfrak{q}}\equiv\mathbf{1}_n\mu_{\mathfrak{q}(\beta_0))}+\boldsymbol{X}(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_{\beta})}\odot\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})])}+\sum_{j=1}^{d_{\bullet}}\boldsymbol{Z}_j(\mu_{\mathfrak{q}(\gamma_{uj})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}),$$

then the fourth term of (4.53) becomes

$$-E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\boldsymbol{c}\right)\right\}\right]=\frac{n}{2}\log(2\pi)-\sum_{i=1}^{n}\Bigg(y_iE_{\mathfrak{q}}\left[\left\{I(c_i\geq 0)\right\}\right]+(1-y_i)E_{\mathfrak{q}}\left[\left\{I(c_i<0)\right\}\right]$$
$$-y_iE_{\mathfrak{q}}\left[\log\left\{\Phi(\boldsymbol{\eta}_{\boldsymbol{i}}^{\mathfrak{q}})\right\}\right]-(1-y_i)E_{\mathfrak{q}}\left[\log\left\{\Phi(-\boldsymbol{\eta}_{\boldsymbol{i}}^{\mathfrak{q}})\right\}\right]\Bigg)$$
$$+\frac{1}{2}E_{\mathfrak{q}}\left(\left\|\boldsymbol{c}-\mathbf{1}_n\beta_0-\boldsymbol{X}(\boldsymbol{\gamma}_{\beta}\odot\widetilde{\boldsymbol{\beta}})-\sum_{j=1}^{d_{\bullet}}\boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j)\right\|^2\right).$$

Therefore, for binary response, we have

$$\log\underline{p}\left(\boldsymbol{y};\mathfrak{q}\right)=\log\underline{p}\left(\boldsymbol{y};\mathfrak{q},\mathrm{C}\right)+\sum_{i=1}^{n}\Bigg(y_iE_{\mathfrak{q}}\left[\log\left\{I(c_i\geq 0)\right\}\right]+(1-y_i)E_{\mathfrak{q}}\left[\log\left\{I(c_i<0)\right\}\right]\Bigg)$$
$$-\frac{n}{2}\log(2\pi)-\frac{1}{2}E_{\mathfrak{q}}\left(\left\|\boldsymbol{c}-\mathbf{1}_n\beta_0-\boldsymbol{X}(\boldsymbol{\gamma}_{\beta}\odot\widetilde{\boldsymbol{\beta}})-\sum_{j=1}^{d_{\bullet}}\boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j)\right\|^2\right)$$
$$+\frac{n}{2}\log(2\pi)-\sum_{i=1}^{n}\Bigg(y_iE_{\mathfrak{q}}\left[\left\{I(c_i\geq 0)\right\}\right]+(1-y_i)E_{\mathfrak{q}}\left[\left\{I(c_i<0)\right\}\right]$$
$$-y_iE_{\mathfrak{q}}\left[\log\left\{\Phi(\boldsymbol{\eta}_{\boldsymbol{i}}^{\mathfrak{q}})\right\}\right]-(1-y_i)E_{\mathfrak{q}}\left[\log\left\{\Phi(-\boldsymbol{\eta}_{\boldsymbol{i}}^{\mathfrak{q}})\right\}\right]\Bigg)$$
$$+\frac{1}{2}E_{\mathfrak{q}}\left(\left\|\boldsymbol{c}-\mathbf{1}_n\beta_0-\boldsymbol{X}(\boldsymbol{\gamma}_{\beta}\odot\widetilde{\boldsymbol{\beta}})-\sum_{j=1}^{d_{\bullet}}\boldsymbol{Z}_j(\gamma_{uj}\widetilde{\boldsymbol{u}}_j)\right\|^2\right)$$
$$=\log\underline{p}\left(\boldsymbol{y};\mathfrak{q},\mathrm{C}\right)+\sum_{i=1}^{n}\Bigg(y_iE_{\mathfrak{q}}\left[\log\left\{\Phi(\boldsymbol{\eta}_i^{\mathfrak{q}})\right\}\right]+(1-y_i)E_{\mathfrak{q}}\left[\log\left\{-\Phi(-\boldsymbol{\eta}_i^{\mathfrak{q}})\right\}\right]\Bigg)$$
$$=\log\underline{p}\left(\boldsymbol{y};\mathfrak{q},\mathrm{C}\right)+\sum_{i=1}^{n}\log\left\{\Phi\left(\left(2y_i-1\right)\boldsymbol{\eta}_i^{\mathfrak{q}}\right)\right\}$$
$$=\log\underline{p}\left(\boldsymbol{y};\mathfrak{q},\mathrm{C}\right)+\sum_{i=1}^{n}\log\Bigg\{\Phi\Bigg[\left(2y_i-1\right)\Big(\mathbf{1}_n\mu_{\mathfrak{q}(\beta_0)}+\boldsymbol{X}\Big(\boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{\gamma}_{\beta}\right)}\odot\boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\Big)$$
$$+\sum_{j=1}^{d_{\bullet}}\boldsymbol{Z}_j\left(\mu_{\mathfrak{q}\left(\gamma_{uj}\right)}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_j)}\right)\Big)_i\Bigg]\Bigg\}.$$

Chapter 5

Simulation Studies

Declaration The figures and results presented in this chapter are presented in the paper *Bayesian Generalized Additive Model Selection Including a Fast Variational Option* (He and Wand 2024). I certify that this publication was a direct result of my research towards this PhD, under the supervision of my principal supervisor Prof Matt Wand, and the reproduction in this thesis does not breach any copyright regulations.

5.1 Introduction

We have mainly focussed on the usefulness of the presented model selection tool, in the context of generalized additive models (GAM), which is a quite popular fitting procedure, and that comes with a friendly computational interface in the form of R package gamselBayes for the Laplace-Zero prior described in Chapter 2 and 4. Our goal is to arrive at a default algorithm and R package that is both scalable and possesses good three-category classification performance, whereby each predictor is determined to have either a zero effect, a linear effect or a non-linear effect.

For Gaussian and binary responses generalized additive model selection, we describe Markov chain Monte Carlo (MCMC) approach in Chapter 2 and mean field variational Bayes (MFVB) approach in Chapter 4. By imposing sparseness-inducing priors on the coefficients, the construction of the graphical model structure and the use of group absolute shrinkage and selection operator priors, ensures that the full conditionals belong to known parametric families of distributions and hence allows the implementation of MCMC in a Gibbsian setting. Our scalability goals benefit greatly from Gibbsian full conditional distributions. Exploration of the orthogonality advantages of Demmler-Reinsch spline bases and closed form expressions of the MFVB also imply that MFVB has a non-negligible computational advantage and thus it is convenient for inference in applications involving big data sets.

We demonstrate Algorithm 2 in Section 2.5.3 for the MCMC approach and Algorithm 7 in Section 4.2.2 for the MFVB approach, using these two different approaches to approximate the posterior distributions associated to the nodes of the directed acyclic graph in Figure 2.1 and 2.2 that represents the Bayesian model in (2.28). Categorization proceeds through a Bayesian graphical model whereby the effect of each predictor is categorised as being linear, or a non-linear or zero. We implement both approaches and optimise with regards to default choices of the thresholding parameter τ, in a fully documented and performant R package gamselBayes that is completely automatic, ready-to-use and publicly available on CRAN page. To improve the computational speed, our implementation used the Rcpp interface (Eddelbuettel and François, 2011) to the C++ language. We also develop a vignette (available on the CRAN page) to explain the usage of the package routines in simple settings. We evaluate both MCMC and MFVB approaches to perform meaningful "zero effect", "linear effect" and "non-linear effect" variable selection in this chapter.

There are other R packages concerned with generalized additive model analysis.

Some of them also apply LASSO-type approach. For example, such packages are BayesX (Umlauf et al., 2022b), bamlss (Umlauf et al., 2022a) and bmrs (Bürkner, 2022).

We describe the simulation study to corroborate the performance of our proposed Bayesian model selection approaches as compared to the frequentist approach of Chouldechova and Hastie (2015) accompanied with R package gamsel and a Bayesian approach of Scheipl et al. (2012) accompanied with R package spikeSlabGAM. Our main comparative performance measures are misclassification rate and computational time taken in seconds, described in Section 5.2.2 and Section 5.2.3 respectively.

We also explore how the computational time of our two proposed GAM selection approaches (MCMC and MFVB) scale with sample size and the number of predictors in Section 5.3, which shows the real benefit of the variational approach.

In (2.28), a set of default values are chose for the model hyperparameters. The sensitivity of the misclassification rate for the two proposed approaches (MCMC and MFVB) are assessed against a set of different hyperparameter values, given in Section 5.4. The simulation study in this section shows low sensitivity to the Half Cauchy distribution scale hyperparameter values.

In Section 5.2, we conducted a simulation study for classification performance comparison in terms of accuracy and speed, among our MCMC given in Algorithm 2 as described in Chapter 2, our MFVB given in Algorithm 7 as described in Chapter 4 and the default version of the spikeSlabGAM described in Section 5.2.1, for the three-category GAM selection on Gaussian and binary responses. The data were generated with 30 continuous predictors, 10 having "true" linear effects, 10 having "true" non-linear effects and 10 having "true" zero effects. The sample size varied as $n \in \{500, 1000, 2000\}$. The error standard deviation varied as $\sigma_\varepsilon \in \{0.25, 0.5, 1, 2\}$ for Gaussian responses. For each combination of sample size and error standard deviation, we ran simulation on 100 independently generated data sets.

5.2 Comparative Performance

In this section, we describe a simulation study that compares the classification performance, measured by misclassification rate and computational time taken in seconds, of our two proposed approaches MCMC and MFVB, with other alternative existing approaches in the three-category generalized additive model selection.

5.2.1 Introduction of Other R packages for Model Selection

The other approaches with R packages used in this section for comparison are:

1. spikeSlabGAM (Scheipl et al., 2012) is a Bayesian approach. According to Scheipl (2011), spikeSlabGAM uses blockwise Gibbs sampling for Gaussian response models. However, non-Gaussian responses require Metropolis-Hasting sampling.

2. gamsel (Chouldechova and Hastie, 2022) implements the frequentist approach described in Chouldechova and Hastie (2015).

For binary responses, spikeSlabGAM and gamsel approaches, the logit link function is used, in contrast to the probit link function employed by Algorithm 2 and Algorithm 7. As a result, it is necessary to apply the corresponding inverse link function to generate binary response data appropriately in this simulation study.

For spikeSlabGAM, we used the default version of the `spikeSlabGAM()` function. For the linear components, the form of the prior distributions imposed on the coefficients by spikeSlabGAM is

$$\mathfrak{p}(\beta_j|\sigma_\beta,\rho_\beta)=\frac{\rho_\beta\exp\left\{-\beta_j^2/\left(2\sigma_\beta^2\right)\right\}}{\left(2\pi\sigma_\beta^2\right)^{1/2}}+\frac{(1-\rho_\beta)\exp\left[-\beta_j^2/\left\{2\left(\upsilon_0\sigma_\beta\right)^2\right\}\right]}{\left\{2\pi\left(\upsilon_0\sigma_\beta\right)^2\right\}^{1/2}} \tag{5.1}$$

where $\upsilon_0 \ll 1, 1 \leq j \leq d_\circ + d_\bullet$, with $\upsilon_0 = 0.00025$, replaced (2.9) of Section 2.3.4 for our MCMC approach. The "slab" of (2.9) refers to a Laplace density function, compared to the "slab" of (5.1) referring to a Gaussian density function. The "spike" of (2.9) refers to the point mass at zero compared to the "spike" of (5.1) referring to a $N(0, 0.00025^2)$ mass. Therefore, (5.1) is an alternative to the "spike-and-slab" prior used in (2.9).

For the spline components, an extension of (5.1) is the alternative to (2.17) of Section 2.3.5 for our MFVB approach. Figure 1 of Scheipl (2011) demonstrates the normal mixture of Inverse Gammas with parameter expansion.

For the gamsel package, the main function is `cv.gamsel()`, which uses a regularization path approach. The user obtains a family of generalized additive model fits over a grid of regularization parameter values. To choose a single model, finding the global minima of cross-validation mean function is needed, which requires a lot of manual adjustment to the regularization parameter grids. This process is not

automated in the current version of gamsel. It should be possible to automate this process through extra coding beyond what provided in the current gamsel.

In our experiments, we found if using `cv.gamsel()` to obtain default regularization grid, the shape of the cross-validation mean function is often being monotonic rather than U-shaped. Therefore it does not lead to very good three-category classification performance.

To avoid this default grid problem of gamsel and aim for good three-category classification performance, we tried various choices provided by gamsel to minimize a k-fold cross-validation function over the grid. We found that a geometric sequence of size 50 between 0.01 and 2 often leads to U-shaped cross-validation mean functions. Therefore, this regularization grid with 10-fold cross-validation for model selection is used throughout our comparative performance in this section.

We consider two cross-validation-based choices:

1. The regularization parameter matching the absolute minimum of the mean values.

2. The largest regularization parameter value with mean minus one standard deviation below the absolute minimum.

After running the simulation study, we found the gamsel approach has a tendency to choose larger models which leads to higher misclassification rate for three-category classification than the other approaches. Therefore, we exclude the gamsel approach for the upcoming performance comparisons with graphical summaries in Section 5.2.2 and 5.2.3.

5.2.2 Performance Comparison by Misclassification Rate

Figure 5.1 in this subsection, compares misclassification rates for Gaussian responses of generalized additive model selection, for our MCMC approach described in Chapter 2 as Algorithm 2, our MFVB approach described in Chapter 4 as Algorithm 7, with the default version of spikeSlabGAM described in Section 5.2.1.

The settings of the number of continuous predictors $d_\bullet$, sample size n, error standard deviations σ_ε are as in (2.58). The threshold parameter τ is set to be 0.5 as described in Section 2.6.3. The generation of data sets and design of simulation are as (2.55), (2.56) and (2.57) in Section 2.6.3.

Figure 5.1 shows that, for the lower noise level situations with error standard deviation $\sigma_\varepsilon \in \{0.25, 0.5, 1\}$, the classification performance of these 3 methods is similar. For the higher noise level situation with error standard deviation $\sigma_\varepsilon \in \{1, 2\}$, the fast option approach of Algorithm 7 demonstrates lower accuracy with higher misclassification rate. The price of lower accuracy in classification performance needs to be mitigated against computational run time taken in seconds, described in Section 5.2.3.

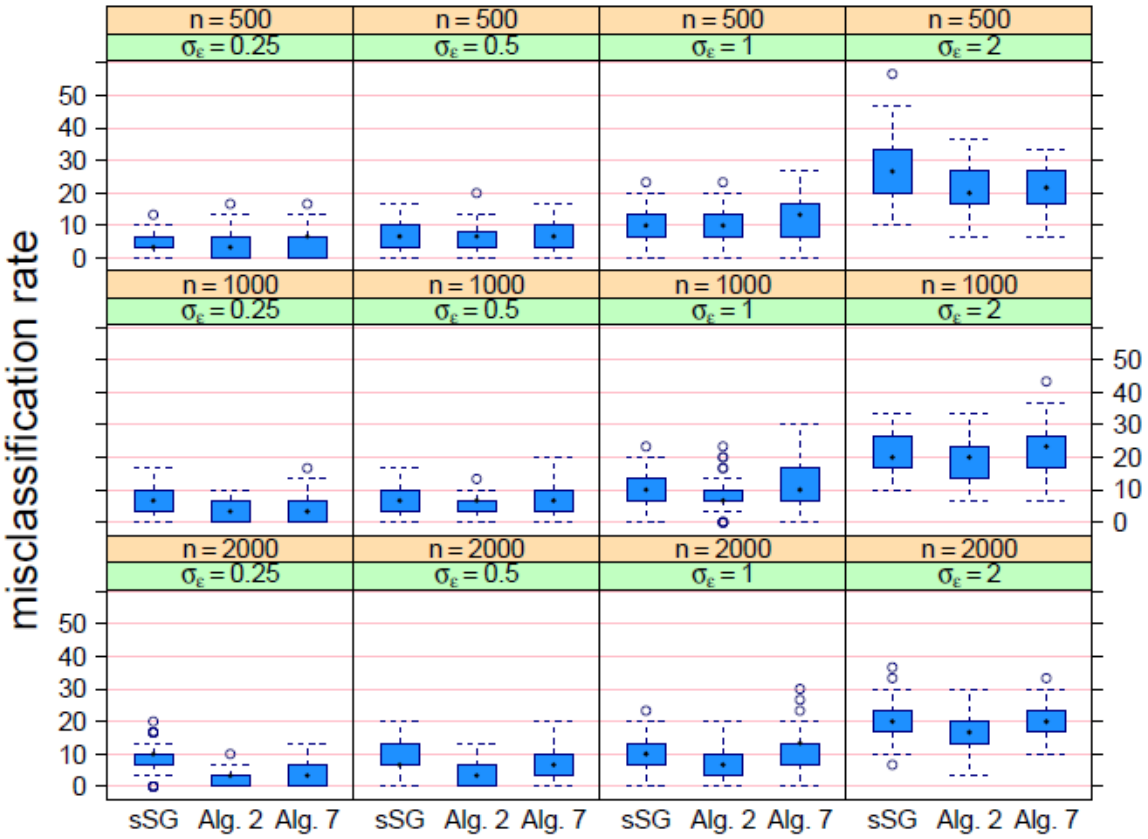


Figure 5.1: *Side-by-side boxplots of the misclassification rates for the comparative performance simulation study described in the text in the case of the response variable being Gaussian. Each panel corresponds to a different combination of sample size and error standard deviation. Within each panel, the side-by-side boxplots compare misclassification rate across each of three methods:* spikeSlabGAM *with default settings (sSG), Algorithm 2 (Alg.2) and Algorithm 7 (Alg.7).*

Figure 5.2 in this subsection compares the misclassification rates for binary responses, for our MCMC approach described in Chapter 2 as Algorithm 2, our MFVB approach described in Chapter 4 as Algorithm 7, with the default version of spikeSlab -GAM described in Section 5.2.1.

The settings of the number of continuous predictors $d_\bullet$, sample size n are as in (2.58). Threshold parameter τ is set to be 0.1 as described in Section 4.3.3. The generation of data sets and design of simulation are described by (2.55), (2.56) and (2.57) in Section 2.6.3.

Figure 5.2 shows that our Algorithm 2 and Algorithm 7 has better three-category classification performance, compared with spikeSlabGAM for the binary response simulation study.

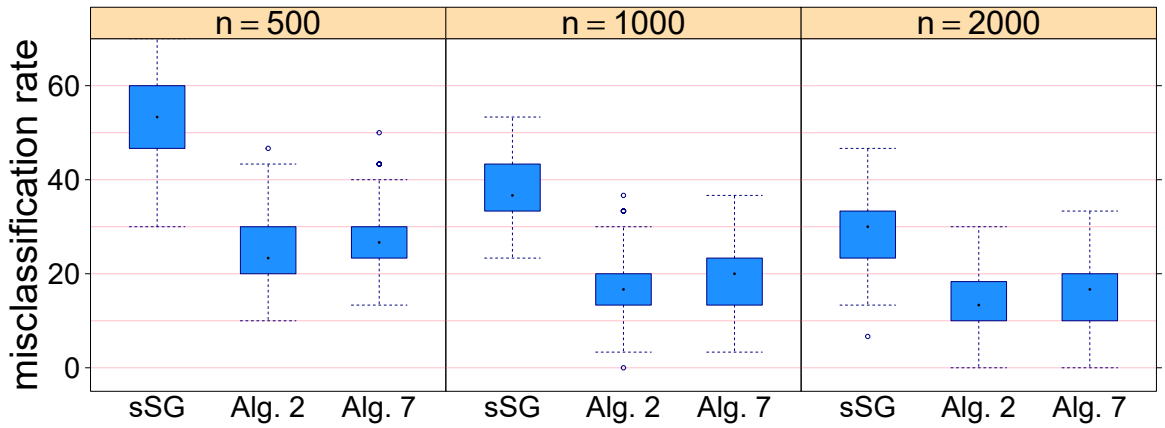


Figure 5.2: *Side-by-side boxplots of the misclassification rates for the comparative performance simulation study described in the text in the case of the response variable being binary. Each panel corresponds to a different sample size. Within each panel, the side-by-side boxplots compare misclassification rate across each of three methods:* spikeSlabGAM *with default settings (sSG), Algorithm 2 (Alg.2) and Algorithm 7 (Alg.7).*

5.2.3 Performance Comparison by Computational Time

In this subsection, we show the computing time in seconds for the four methods, our MCMC approach described by Algorithm 2, our MFVB approach described by Algorithm 7, compared with gamsel and spikeSlabGAM as described in Section 5.2.1.

Table 5.1 shows the elapsed times in seconds. The laptop that we used for simulation study is a MacBook Air laptop computer with 16 gigabytes of memory and a 3.2 gigahertz processor. Algorithm 2 and Algorithm 7 were implemented in the R package gamselBayes using the Rcpp interface (Eddelbuettel and François, 2011) to the C++ language, with the use of Rcpp package and RcppArmadillo package in R for coding, to improve the computational speed. Also, the sufficient statistics described in Section 2.5.1 offers computational speed-ups for our Algorithm 2 and Algorithm

	gamsel	spikeSlabGAM	Algorithm 2	Algorithm 7
10th percentile	3.69	84.9	1.78	0.326
50th percentile	8.12	167.0	2.14	0.466
90th percentile	17.50	339.0	3.03	0.768

Table 5.1: *10th, 50th and 90th percentiles for the number of seconds required for each generalized additive model selection approach across all settings and replications for the comparative performance simulation study.*

7. The Markov chain Monte Carlo samples size values corresponding to the default of spikeSlabGAM and gamselBayes are 1,500 and 2,000 respectively.

Table 5.1 lists the 10th, 50th and 90th percentile number of seconds took for each approach across all settings and replications. Even though spikeSlabGAM can be seen to have a very good classification performance from Figure 5.1 of Section 5.2.2, Table 5.1 in this subsection shows that spikeSlabGAM is comparatively slow and does not scale well to large data sets. Our MCMC approach as Algorithm 2 took less than 3 seconds for 90 percent of the fits in the simulation study. Our faster option MFVB as Algorithm 7, only required less than a second of computing time for most of the fits. Therefore, our new approaches with in R package gamselBayes scale well to large data sets for the generalized additive model selection problem.

The impact of sample size and number of predictors on computing times for our approaches as Algorithm 2 and Algorithm 7 will be described in Section 5.3.

5.3 Detailed Computing Time Results

The computational dimension is crucial as illustrated by the large algorithms. In this section, we present the results of simulation studies that convey an idea of the computational budget required to implement our MCMC and MFVB approaches in our R package gamselBayes.

5.3.1 Assessment of the Effect of Sample Size

We looked into the computational time versus sample size, as our first detailed computational time simulation study in this subsection, with the settings

$$\begin{aligned} &10 \text{ continuous predictors,} \\ &\text{Sample size } n = 10^k \text{ with } k = 2, 3, 4, 5, 6, \end{aligned} \tag{5.2}$$

We generated the data in a manner similar to that for the simulation studies described in (2.56) of Section 2.6.3.

Threshold parameter τ was set to be 0.5 as described in Section 2.6.3 for Algorithm 2 and τ was set to be 0.1 as described in Section 4.3.3 for Algorithm 7. To aid visualization the $\log_{10}$ transformation was applied to both the number of seconds and sample size.

Figure 5.3 shows the comparison with computational time of seconds and sample size both in logarithmic form, using the side-by-side box plots, for Gaussian and binary responses of our MCMC approach as given by Algorithm 2 and our MFVB approach as given by Algorithm 7. The relationship between mean logarithmic computational time and sample size is shown to be a linear as power law such that the power is close to 1. Therefore, mean computational time is roughly proportional to the sample size.

For example, Figure 5.3 shows that, for sample size $n = 10^5$, for binary responses, our MCMC approach as Algorithm 2 took about 100 seconds, while our MFVB approach as Algorithm 7 took about 10 seconds. Therefore, our MFVB approach has an approximately 10-fold reduction in computational time, compared to the MCMC approach.

5.3.2 Assessment of the Effect of the Number of Candidate Predictors

In this subsection we look into the computational time as a function of the number of candidate predictors. We generated the data in a manner similar to that for the simulation studies described in (2.56) of Section 2.6.3, with the settings of sample

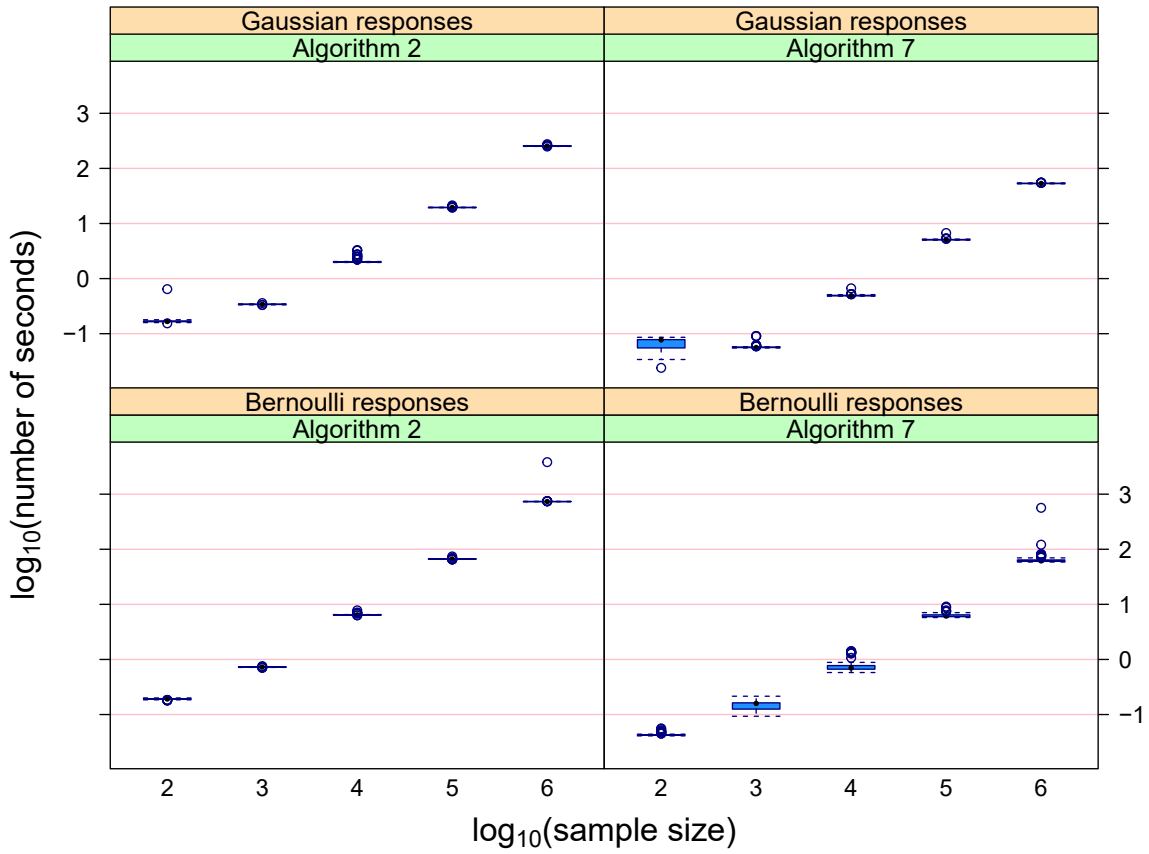


Figure 5.3: *Side-by-side boxplots of computing time in seconds versus sample size for generalized additive model selection via Algorithm 2 and Algorithm 7, for the first simulation study described in Section 5.3.1. Each axis uses a* $\log_{10}$ *scale.*

sizes and predictors to be

$$\begin{aligned} &\text{sample size } n = 5000, \\ &\text{number of continuous predictors } d_\bullet = 2^k \text{ with } k = 1, 2, 3, 4, 5, 6. \end{aligned} \tag{5.3}$$

Figure 5.4 shows results in similar way to Figure 5.3 in Section 5.3.1. To aid visualization, the $\log_{10}$ transformations is applied to the number of seconds and the $\log_2$ transformations is applied to the number of candidate predictors. Within each panel, there is an approximate linear relationship with logarithmic scales. Within each panel, the mean computing time is approximately proportional to $d_\bullet^\kappa$, with κ within the interval $(1.2, 1.5)$, dependent on the combination of response distribution and fitting algorithm.

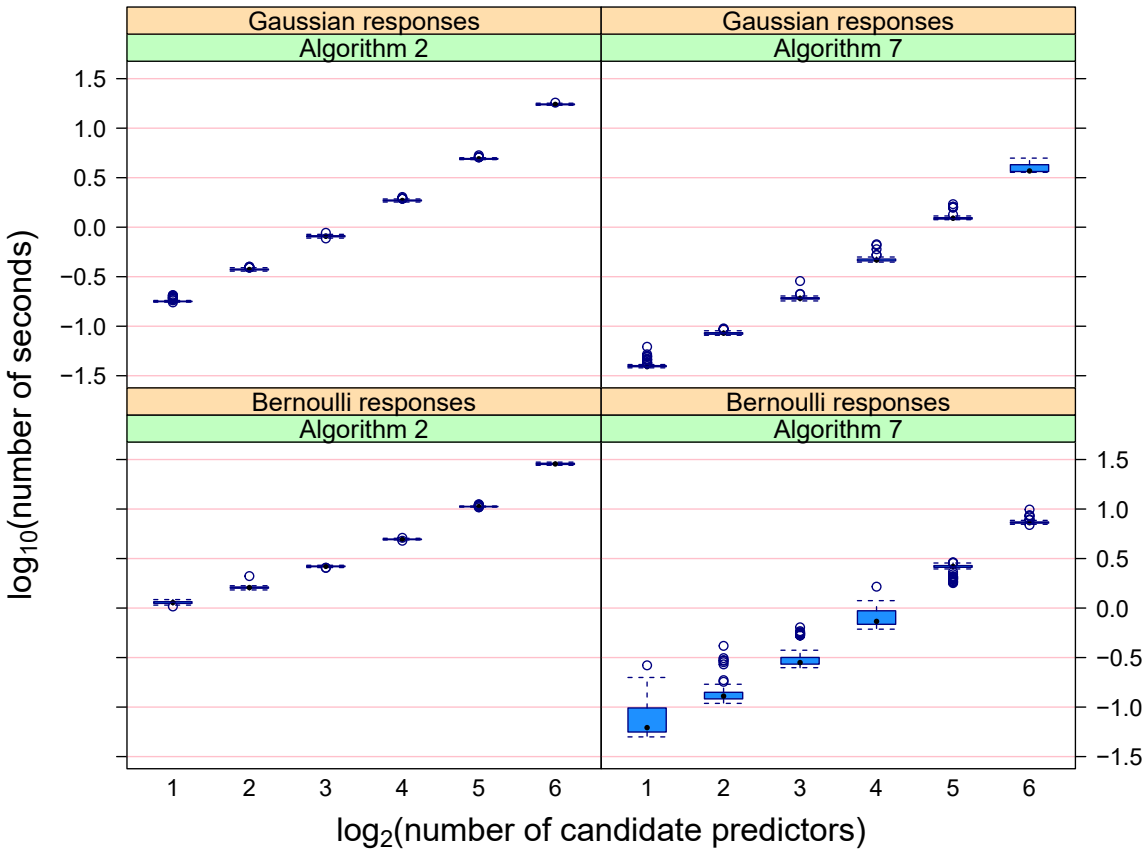


Figure 5.4: *Side-by-side boxplots of computing time in seconds versus number of candidate predictors for generalized additive model selection via Algorithm 2 and Algorithm 7, for the second simulation study described in Section 5.3.2. The horizontal axis uses a* $\log_2$ *scale and the vertical axis uses a* $\log_{10}$ *scale.*

5.4 Hyperparameter Sensitivity Checks

Figure 5.5 and Figure 5.6 show the effects of the scale hyperparameters s_β, s_ϵ and s_u on the misclassification rates. The scale hyperparameters are for the Half Cauchy prior distributions in model (2.28). The scale hyperparameters range over the set of $\left\{10^k : k = 1, 2, 3, 4\right\}$.

The setting of this simulation study is the same as the one for choice of threshold parameter τ in (2.55), (2.56) and (2.57), with τ being 0.5 as our recommended default value for MCMC described in Section 2.6.3 and τ being 0.1 for MFVB described in Section 4.3.3.

Figure 5.5 and Figure 5.6 show that Algorithm 2 and Algorithm 7 are both not

sensitive to the Half Cauchy distribution scale hyperparameter values. In summary, low sensitivity of misclassification rate to the Half Cauchy distribution scale hyperparameter values is apparent for each of our MCMC and MFVB approaches.

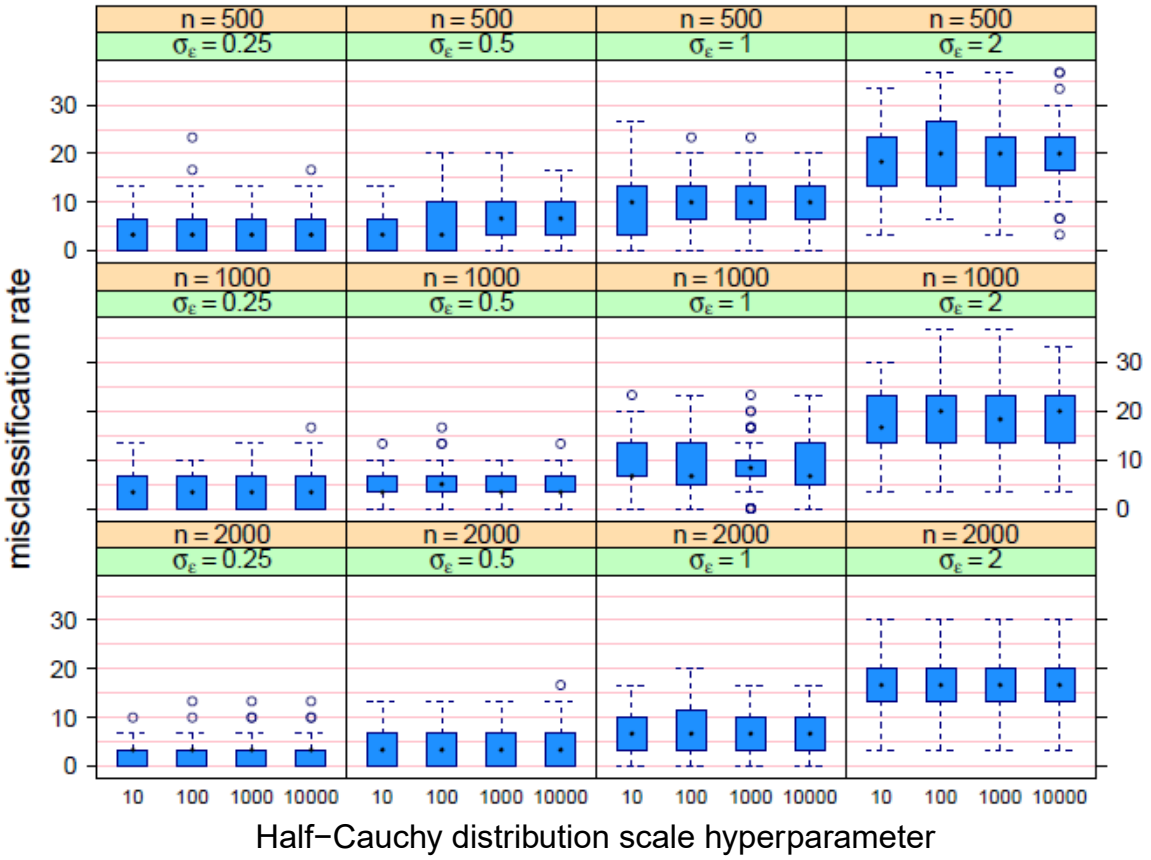


Figure 5.5: *Side-by-side boxplots for varying values of the Half Cauchy distribution scale hyperparameter for the Gaussian response version of Algorithm 2, for the first simulation study described in Section 5.4.*

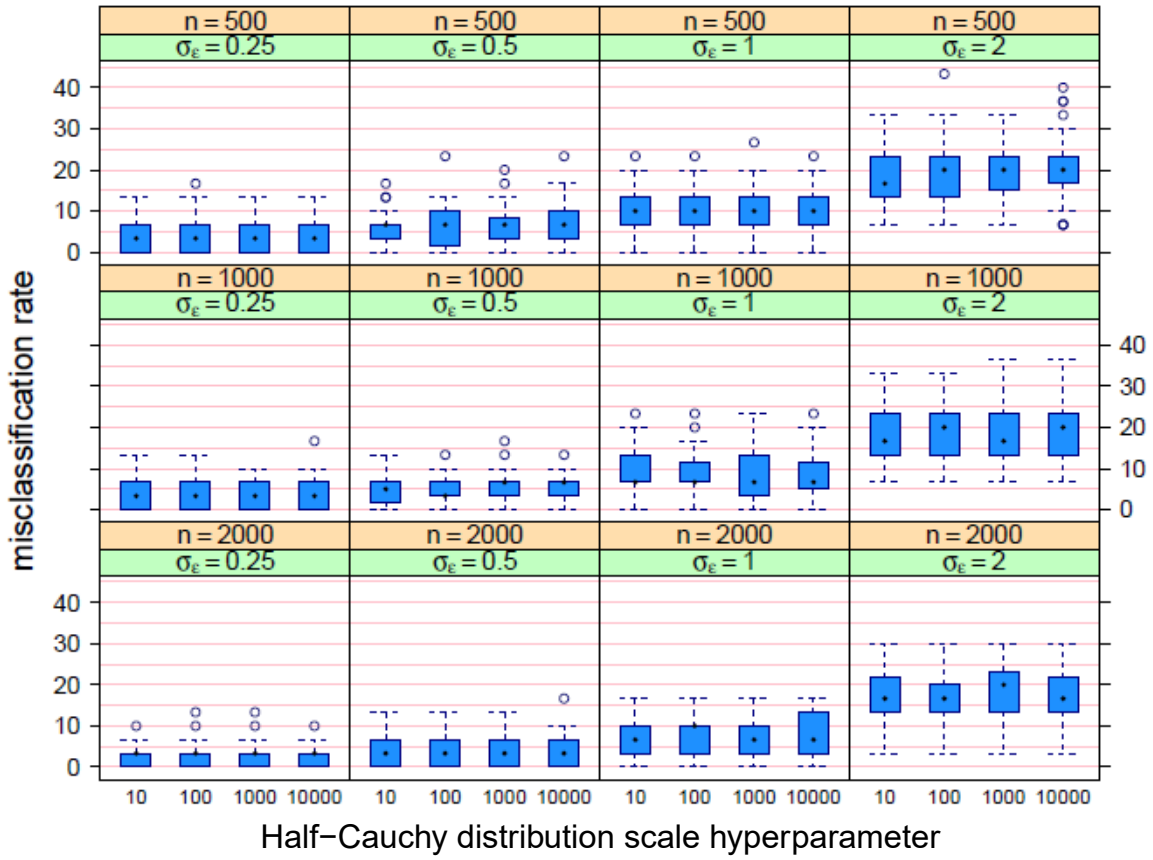


Figure 5.6: *Side-by-side boxplots for varying values of the Half Cauchy distribution scale hyperparameter for the Gaussian response version of Algorithm 7, for the first simulation study described in Section 5.4.*

Chapter 6

Extensions to Generalized Additive Mixed Models Selection

6.1 Introduction

In Chapter 2 and 4, we introduced the Laplace-Zero/Grouped Lasso-Zero priors for generalized additive model selection using the Markov chain Monte Carlo (MCMC) approach and the mean field variational Bayes (MFVB) approach as faster option. We only have the fixed effects for both of the linear coefficients and spline coefficients vector. Nowadays, many datasets have the form of grouped or hierarchical structure. Such structure may come from repeated measurements of individuals over time, such as students grouped within classrooms and schools, gene data of individuals, linked health records, house pricing grouped in suburbs and districts, or friends grouped within networks in LinkedIn or Facebook. In biomedical applications it is common for the groups to be subjects and the term longitudinal data analysis is relevant. However, many other areas have similar types of data such as multilevel data and panel data.

To meet this fundamental interest of grouped data and to best understand the effects of predictors of grouped data, we introduce a groupswise random effect into the Bayesian generalized mixed model.

In this chapter, we introduce a random intercepts vector denoted as $\boldsymbol{U}$ vector, as

described in Section 6.4.2, and obtain the MCMC and MFVB algorithms for variable selection for generalized additive mixed models (GAMM).

Data standardization for GAMM in this chapter is similar to the one for generalized additive model (GAM) in Section 2.2.2. To implement practical fitting algorithms for MCMC, similar to Section 2.4.2, we use least absolute shrinkage and selection operator (LASSO) ideas, the orthogonality advantages of Demmler-Reinsch spline bases, and tailored auxiliary variable representations to set up the full model described in Section 6.4.2 and work out the posterior distributions of each parameter in our full model described in Section 6.5.3. The posterior distributions are all in standard forms for practical implementation. Therefore, as with MCMC in Chapter 2 for GAM selection, the MCMC sampling for GAMM selection with random intercept also reduces to Gibbs sampling. The MCMC fitting algorithm listing for GAMM is given in Algorithm 9 in Section 6.5.3 for implementation into R. The sufficient statistic matrices described in Section 6.5.1, similar to those in Section 2.5.1 for GAM, offer computational speed-ups. Notation for fitting algorithms is described in Section 6.5.2, similar to the one in Section 2.5.2 for GAM. Based on posterior distributions of MCMC, the model selection strategies and simulation studies for good default value of threshold parameter for MCMC described in Section 6.7.

Since each of the full conditional distributions have standard forms for sample drawing in MCMC described in Section 6.5.3 for GAMM selection, the mean field variational Bayes q-densities are all closed forms. We introduced the product density restrictions on approximating density functions in Section 6.6.1. MFVB algorithm for GAMM is given in Algorithm 10. We work out the marginal log-likelihood approximation, which is useful for monitoring convergence, described in Section 6.6.3. The simulation study for good default value of threshold parameter for MFVB is described in Section 6.7.

Our comparative performance measures by misclassification rate described in Section 6.8.1. Application of real grouped data for GAMM selection using MCMC and MFVB is described in Chapter 7.

Throughout Chapter 6, "rest" denotes the set of other random variables in model (6.25).

The additive structure of generalized additive mixed model (GAMM) with random intercept described in this chapter is given in Section 6.3. The independence assumptions of GAMM is given in (6.25) for Gaussian responses and Section 6.4.3 for binary responses. The additive part uses splines for non-linear effects. The pre-

dictors include binary predictors and continuous predictors. The number of binary predictors is $d_\circ$. The number of continuous predictors is $d_\bullet$.

For the generalized additive mixed model with random intercept described in Section 6.2–6.6, the number of groups is m. The sample size in the i-th group is n_i, with $1 \leq i \leq m$. The response $\boldsymbol{y}$ is a $N \times 1$ vector, with $N \equiv \sum_{i=1}^{m} n_i$. The linear design matrix $\boldsymbol{X}$ is a $n \times (d_\circ + d_\bullet)$ matrix. The $\boldsymbol{Z}_j$ matrix containing the canonical Demmler-Reinsch basis is a $n \times K_j$ matrix, with $1 \leq j \leq d_\bullet$, K_j as the number of basis functions for the j-th continuous predictor with default value of 12. Throughout this chapter, $\boldsymbol{y}$ vector, $\boldsymbol{X}$ matrix and $\boldsymbol{Z}$ matrix are all fixed.

Our simulation study suggests a good default threshold value τ as 0.5 for MCMC and as 0.1 for MFVB given in Section 6.7.3. The data were generated with 30 continuous predictors, 10 having "true" linear effects, 10 having "true" non-linear effects and 10 having "true" zero effects. The number of groups varied as $m \in \{100, 200, 300\}$. The number of observations within each group is fixed at $m/10$. The error standard deviation varied as $\sigma_\varepsilon \in \{0.25, 0.5, 1, 2\}$. The threshold varied as $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. For each combination of number of groups and error standard deviation, we ran simulation on 100 independently generated data sets.

In Section 6.8, we conducted a simulation study for classification performance comparison in terms of accuracy for Gaussian and binary responses, among our MCMC (Algorithm 9), our MFVB (Algorithm 10) and the default version of the spikeSlabGAM described in Section 5.2.1. The data were generated with 30 continuous predictors, 10 having "true" linear effects, 10 having "true" non-linear effects and 10 having "true" zero effects. The number of groups varied as $m \in \{100, 200, 300\}$. The number of observations within each group is fixed at $m/10$. The error standard deviation varied as $\sigma_\varepsilon \in \{0.25, 0.5, 1, 2\}$ for Gaussian responses. For each combination of sample size and error standard deviation, we ran simulation on 100 independently generated data sets.

6.2 Bayesian Gaussian Response Additive Mixed Model Selection

6.2.1 The Original Input Data

The original input data are as follows:

$$\left(\mathring{\boldsymbol{x}}_{ij}^{\text{orig}}, \dot{\boldsymbol{x}}_{ij}^{\text{orig}}, y_{ij}^{\text{orig}}\right), \quad 1 \leq i \leq m, \quad 1 \leq j \leq n_i,$$

where, for each $1 \leq i \leq m, 1 \leq j \leq n_i$, $\mathring{\boldsymbol{x}}_{ij}^{\text{orig}}$ denotes a $d_\circ \times 1$ vector of predictors that can only enter the model linear (e.g. binary predictors) and $\dot{\boldsymbol{x}}_{ij}^{\text{orig}}$ denotes a $d_\bullet \times 1$ vector of continuous predictors that can enter the model either linearly or non-linearly.

The number of groups is m and the number of observations within the ith group is n_i. Then, for example,

$$y_{ij}^{\text{orig}} \text{ is the } j\text{th response within the } i\text{th group.}$$

6.2.2 Standardization of the Input Data for Fitting and Model Selection

Same as Section 2.2.2 for GAM, standardization of the original input data for GAMM is necessary. For fitting and model selection we replace the $\mathring{x}_{ijk}^{\text{orig}}$ by $\mathring{x}_{ijk} = \left(\mathring{x}_{ijk}^{\text{orig}} - a_x\right)/b_x$ where

$$a_x \equiv \frac{1}{n_i}\sum_{j=1}^{n_i} \mathring{x}_{ijk}^{\text{orig}} \quad \text{and} \quad b_x \equiv \sqrt{\frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(\mathring{x}_{ijk}^{\text{orig}} - a_x\right)^2} \tag{6.1}$$

with $1 \leq k \leq d_\circ$ for predictors can only enter the model linearly, $1 \leq i \leq m, 1 \leq j \leq n_i$.

And we replace the $\dot{x}_{ijk}^{\text{orig}}$ by $\dot{x}_{ijk} = \left(\dot{x}_{ijk}^{\text{orig}} - c_x\right)/d_x$ where

$$c_x \equiv \frac{1}{n_i}\sum_{j=1}^{n_i} \dot{x}_{ijk}^{\text{orig}} \quad \text{and} \quad d_x \equiv \sqrt{\frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(\dot{x}_{ijk}^{\text{orig}} - c_x\right)^2}, \tag{6.2}$$

with $1 \leq k \leq d_\bullet$ for predictors can enter the model either linearly or non-linearly, $1 \leq i \leq m, 1 \leq j \leq n_i$.

For $1 \leq i \leq m, 1 \leq j \leq n_i$, this transformation of the original predictor data implies that the transformed predictor data are centred in that

$$\sum_{j=1}^{n_i} \mathring{x}_{ijk} = 0, \;\; 1 \leq k \leq d_\circ \;\; \text{and} \;\; \sum_{j=1}^{n_i} \dot{x}_{ijk} = 0, \;\; 1 \leq k \leq d_\bullet. \tag{6.3}$$

For Gaussian response, $1 \leq i \leq m, 1 \leq j \leq n_i$, the response data for fitting and model selection are made scale-free as follows:

$$y_{ij} = y_{ij}^{\text{orig}} / s_y, \;\; s_y \equiv \sqrt{\frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left(y_{ij}^{\text{orig}} - \overline{y}^{\text{orig}} \right)^2}$$

is the sample standard deviation of the y_{ij}^{orig}s.

For binary responses, the y_{ij}s are not pre-processed and remain as values in $\{0, 1\}$.

Algorithm 8 in Section 6.5.1 provides the operational details of the standardization process.

Throughout this thesis, we define, for each $1 \leq i \leq m$,

$$\boldsymbol{y}_i \equiv \begin{bmatrix} y_{i1} \\ \vdots \\ y_{in_i} \end{bmatrix}.$$

6.3 Design Matrices and Auxiliary Variable Representations for Parameters

6.3.1 Generalized Additive Mixed Model Form for the Full Model

For Gaussian responses, we have $g(\mu) = \mu$ (referred to Section 1.10 for the choice of $g(\cdot)$). This is called the identity link. It directly models the mean of the response without transformation. Here, $g(\cdot)$ is of the mean of $\boldsymbol{y}$, with $\mathring{x}_{ijk}$ fixed, with $1 \leq i \leq m$, $1 \leq j \leq n_i$ and $1 \leq k \leq d_\circ$, and with $\dot{x}_{ijk}$ fixed, with $1 \leq i \leq m$, $1 \leq j \leq n_i$ and $1 \leq j \leq d_\bullet$.

Similar to Section 2.3.1 for generalized additive model, in this section for generalized additive mixed model with random intercept, for response data $y_{ij}, 1 \leq i \leq m, 1 \leq j \leq n_i$, the generalized additive mixed models involve linear predictors hav-

ing generic form

$$y_{ij} = \beta_0 + U_i + \sum_{k=1}^{d_\circ} \beta_k \mathring{x}_{ijk} + \sum_{k=1}^{d_\bullet} f_k\left(\dot{x}_{ijk}\right) \tag{6.4}$$

for scalar parameters $\beta_1, ..., \beta_{d_\circ}$ and the f_j are smooth real-valued functions over an interval containing the $\dot{x}_{ijk}$ data, $1 \le i \le m, 1 \le j \le n_i, 1 \le k \le d_\bullet$.
From (6.4), the detailed form of our generalized additive mixed full model is

$$\begin{aligned}
E\left(\boldsymbol{y}_i\right) &= \left(\beta_0 + U_i\right)\mathbf{1}_{n_i} + \sum_{k=1}^{d_\circ}\left\{\left(\mathbf{1}_{n_i}\beta_k\right) \odot \mathring{\boldsymbol{x}}_{ik}\right\} + \sum_{k=1}^{d_\bullet}\left\{\left(\mathbf{1}_{n_i}\beta_k\right) \odot \dot{\boldsymbol{x}}_{ik}\right\} \\
&\quad + \sum_{k=1}^{d_\bullet}\left\{\sum_{l=1}^{L_k}\left(\mathbf{1}_{n_i}u_{kl}\right) \odot \boldsymbol{Z}_{ikl}\right\} \\
&= \beta_0 + U_i + \sum_{k=1}^{d_\circ + d_\bullet}\left\{\left(\mathbf{1}_{n_i}\beta_k\right) \odot \boldsymbol{X}_{ik}\right\} + \sum_{k=1}^{d_\bullet}\left\{\sum_{l=1}^{L_k}\left(\mathbf{1}_{n_i}u_{kl}\right) \odot \boldsymbol{Z}_{ikl}\right\} \\
&= \beta_0 + U_i + \boldsymbol{X}_i\boldsymbol{\beta} + \sum_{k=1}^{d_\bullet}\boldsymbol{Z}_{ik}\boldsymbol{u}_k \\
&= \beta_0 + U_i + \boldsymbol{X}_i\left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}\right) + \sum_{k=1}^{d_\bullet}\boldsymbol{Z}_{ik}\left(\gamma_{uk}\widetilde{\boldsymbol{u}}_k\right)
\end{aligned} \tag{6.5}$$

where $\mathring{\boldsymbol{x}}_{ik}$ is the following $n_i \times 1$ vector and denotes the k-th predictor that can only enter the model linearly in the i-th group:

$$\mathring{\boldsymbol{x}}_{ik} \equiv \begin{bmatrix} \mathring{x}_{i1k} \\ \vdots \\ \mathring{x}_{in_ik} \end{bmatrix} \text{ where } 1 \le i \le m, 1 \le k \le d_\circ, \tag{6.6}$$

$\dot{\boldsymbol{x}}_{ik}$ is the following $n_i \times 1$ vector and denotes the k-th continuous predictor in the i-th group:

$$\dot{\boldsymbol{x}}_{ik} \equiv \begin{bmatrix} \dot{x}_{i1k} \\ \vdots \\ \dot{x}_{in_ik} \end{bmatrix} \text{ where } 1 \le i \le m, 1 \le k \le d_\bullet, \tag{6.7}$$

$\boldsymbol{Z}_{ik\ell}$ is the following $n_i \times 1$ vector and denotes the l-th basis function of Demmler-

Reinsch matrix $\boldsymbol{Z}_{ik}$:

$$\boldsymbol{Z}_{ik\ell} \equiv \begin{bmatrix} Z_{i1k\ell} \\ \vdots \\ Z_{in_ik\ell} \end{bmatrix} \text{ where } 1 \leq i \leq m, 1 \leq k \leq d_\bullet, 1 \leq \ell \leq L_k,$$

$\boldsymbol{Z}_{ik}$ is the following $n_i \times L_k$ matrix, and L_k is the number of columns in $\boldsymbol{Z}_{ik}$, for the k-th continuous predictor in the i-th group:

$$\boldsymbol{Z}_{ik} \equiv \begin{bmatrix} Z_{i1k1} & \dots & Z_{i1kL_K} \\ \vdots & \vdots & \vdots \\ Z_{in_ik1} & \dots & Z_{in_ikL_k} \end{bmatrix} \text{ where } 1 \leq i \leq m, 1 \leq k \leq d_\bullet, \tag{6.8}$$

$\boldsymbol{X}_{ik}$ is the following $n_i \times 1$ vector denotes the k-th binary/categorical and continuous predictors in the i-th group:

$$\boldsymbol{X}_{ik} \equiv \begin{bmatrix} X_{i1k} \\ \vdots \\ X_{in_ik} \end{bmatrix} \text{ where } 1 \leq i \leq m, 1 \leq k \leq d_\circ + d_\bullet,$$

and for the ith group, the linear design matrix $\boldsymbol{X}_i$ is the following $n_i \times (d_\circ + d_\bullet)$ matrix:

$$\boldsymbol{X}_i \equiv \begin{bmatrix} \mathring{x}_{i11} & \dots & \mathring{x}_{i1d_\circ} & \dot{x}_{i11} & \dots & \dot{x}_{i1d_\bullet} \\ \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ \mathring{x}_{in_i1} & \dots & \mathring{x}_{in_id_\circ} & \dot{x}_{in_i1} & \dots & \dot{x}_{in_id_\bullet} \end{bmatrix} \text{ where } 1 \leq i \leq m, \tag{6.9}$$

$\boldsymbol{X}$ is the following $N \times (d_\circ + d_\bullet)$ matrix, $N \equiv \sum_{i=1}^{m} n_i$:

$$\boldsymbol{X} \equiv \begin{bmatrix} \mathring{x}_{111} & \dots & \mathring{x}_{11d_\circ} & \dot{x}_{111} & \dots & \dot{x}_{11d_\bullet} \\ \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ \mathring{x}_{1n_11} & \dots & \mathring{x}_{1n_1d_\circ} & \dot{x}_{1n_11} & \dots & \dot{x}_{1n_1d_\bullet} \\ \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ \mathring{x}_{i11} & \dots & \mathring{x}_{i1d_\circ} & \dot{x}_{i11} & \dots & \dot{x}_{i1d_\bullet} \\ \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ \mathring{x}_{in_i1} & \dots & \mathring{x}_{in_id_\circ} & \dot{x}_{in_i1} & \dots & \dot{x}_{in_id_\bullet} \\ \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ \mathring{x}_{m11} & \dots & \mathring{x}_{m1d_\circ} & \dot{x}_{m11} & \dots & \dot{x}_{m1d_\bullet} \\ \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ \mathring{x}_{mn_m1} & \dots & \mathring{x}_{mn_md_\circ} & \dot{x}_{mn_m1} & \dots & \dot{x}_{mn_md_\bullet} \end{bmatrix}.$$

Let $\boldsymbol{\beta}$ be the $(d_\circ + d_\bullet) \times 1$ vector of corresponding fixed effects vector. From the standardization of original input data described in Section 6.2.2 and (6.3), it offers simplification to be

$$\mathbf{1}_N^T \boldsymbol{X} = \mathbf{0} \text{ where } N \equiv \sum_{i=1}^{m} n_i. \tag{6.10}$$

$\boldsymbol{Z}_{ik}$ defined in (6.8), partners the $\boldsymbol{y}_i$ and $\boldsymbol{X}_i$ matrices in the sense that they all have n_i rows and correspond to data for the ith group, with $1 \le i \le m$. Note

$\boldsymbol{Z}_{ik} \equiv$ matrix containing the canonical Demmler-Reinsch basis for $\dot{\boldsymbol{x}}_{ik}$, a $n_i \times L_k$ matrix, $1 \le i \le m, 1 \le k \le d_\bullet$;

$L_k \equiv$ the number of columns in $\boldsymbol{Z}_{ik}$;

$\boldsymbol{u}_k \equiv \gamma_{uk}\widetilde{\boldsymbol{u}}_k$, $\boldsymbol{u}_k$ is the spline coefficients vector for matrix $\boldsymbol{Z}_{ik}$, a $L_k \times 1$vector, $1 \le k \le d_\bullet$;

$\gamma_{uk} \equiv$ auxiliary probability parameters for $\boldsymbol{u}_k$, a scalar , $1 \le k \le d_\bullet$;

$\widetilde{\boldsymbol{u}}_k \equiv$ auxiliary parameters for $\boldsymbol{u}_k$, a $L_k \times 1$ vector , $1 \le k \le d_\bullet$.

In (6.5), the definition of linear term coefficients vector $\boldsymbol{\beta}$, auxiliary probability pa-

rameter vector $\boldsymbol{\gamma}_{\beta}$, auxiliary parameter vector $\widetilde{\boldsymbol{\beta}}$, are the same as the ones in Section 2.3.1.

Throughout Chapter 6, $\boldsymbol{y}$ vector, $\boldsymbol{X}$ matrix and $\boldsymbol{Z}$ matrix are all fixed.

6.3.2 The Fixed Effects of Linear Design Matrix

The linear design matrix $\boldsymbol{X}$ is the following $\left(\sum_{i=1}^{m} n_i\right) \times (d_{\circ} + d_{\bullet})$ matrix:

$$\boldsymbol{X} \equiv \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_m \end{bmatrix}$$

where, for each $1 \leq i \leq m$, $\boldsymbol{X}_i$ defined in (6.9).

In the linear design matrix $\boldsymbol{X}$, the kth column of matrix $\boldsymbol{X}$ is the following $N \times 1$ vector, with $N \equiv \sum_{i=1}^{m} n_i$:

$$\dot{\boldsymbol{X}}_k \equiv \begin{bmatrix} X_{11k} \\ \vdots \\ X_{1n_1k} \\ \vdots \\ X_{i1k} \\ \vdots \\ X_{in_ik} \\ \vdots \\ X_{m1k1} \\ \vdots \\ X_{mn_mk} \end{bmatrix} \text{ with } 1 \leq i \leq m, 1 \leq k \leq d_{\circ} + d_{\bullet}. \tag{6.11}$$

6.3.3 Canonical Demmler-Reinsch Basis Design Matrices

The construction of $\dot{\boldsymbol{Z}}_k$ matrix, with $1 \leq k \leq d_{\bullet}$, containing the canonical Demmler-Reinsch basis for the predictor data $\dot{\boldsymbol{x}}_k$ for generalized additive mixed model with random intercept, is the same as the construction of matrix $\boldsymbol{Z}_j, 1 \leq j \leq d_{\bullet}$ for gen-

eralized additive model selection in Section 2.3.3, described in Appendix 2.A. We define $\dot{\boldsymbol{Z}}_k$ is the following $\left(\sum_{i=1}^{m} n_i\right) \times L_k$ matrix:

$$\dot{\boldsymbol{Z}}_k \equiv \begin{bmatrix} Z_{11k1} & \dots & Z_{11kL_1} \\ \vdots & \vdots & \vdots \\ Z_{1n_1k1} & \dots & Z_{1n_1kL_1} \\ \vdots & \vdots & \vdots \\ Z_{i1k1} & \dots & Z_{i1kL_k} \\ \vdots & \vdots & \vdots \\ Z_{in_ik1} & \dots & Z_{in_ikL_k} \\ \vdots & \vdots & \vdots \\ Z_{m1k1} & \dots & Z_{m1kL_{d_\bullet}} \\ \vdots & \vdots & \vdots \\ Z_{mn_mk1} & \dots & Z_{mn_mkL_{d_\bullet}} \end{bmatrix} \text{ where } 1 \leq i \leq m, 1 \leq k \leq d_\bullet, \tag{6.12}$$

and define $\dot{\boldsymbol{x}}_k$ is the following $\left(\sum_{i=1}^{m} n_i\right) \times 1$ vector:

$$\dot{\boldsymbol{x}}_k \equiv \begin{bmatrix} \dot{x}_{11k} \\ \vdots \\ \dot{x}_{1n_1k} \\ \vdots \\ \dot{x}_{i1k} \\ \vdots \\ \dot{x}_{in_ik} \\ \vdots \\ \dot{x}_{m1k} \\ \vdots \\ \dot{x}_{mn_mk} \end{bmatrix} \text{ with } 1 \leq i \leq m, 1 \leq k \leq d_\bullet.$$

We will refer to the columns of $\dot{\boldsymbol{Z}}_k$ as the canonical Demmler-Reinsch basis for the $\dot{x}_k$s. The $\boldsymbol{Z}$ matrix we construct according to Appendix 2.A satisfies:

$$\dot{\boldsymbol{Z}}_k^T \mathbf{1}_N = \dot{\boldsymbol{Z}}_k^T \dot{\boldsymbol{x}}_k = \mathbf{0}_{L_k} \quad \text{and} \quad \dot{\boldsymbol{Z}}_k^T \dot{\boldsymbol{Z}}_k \text{ is a diagonal matrix}, \;\; 1 \leq k \leq d_\bullet, N = \sum_{i=1}^{m} n_i. \tag{6.13}$$

Spline bases satisfying (6.13) are referred to as having a Demmler-Reinsch form.

We require a version of $\dot{\boldsymbol{Z}}_k$ for each set of $\dot{x}_k$s, $1 \leq k \leq d_\bullet$. Denote the canonical Demmler-Reinsch basis matrices by

$$\dot{\boldsymbol{Z}}_1, \ldots, \dot{\boldsymbol{Z}}_{d_\bullet}.$$

with $\dot{\boldsymbol{Z}}_k$ having dimension $\left(\sum_{i=1}^{m} n_i\right) \times L_k$. Let

$$\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{d_\bullet}$$

be the corresponding coefficient vectors, where $\boldsymbol{u}_k$ has dimension $L_k \times 1$.

6.3.4 Linear Term Coefficients with Auxiliary Variable Representations

Similar to Section 2.3.4 for GAM, we have the following for GAMM

$$\boldsymbol{\beta} \equiv \boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}. \tag{6.14}$$

$$\widetilde{\beta}_k \big| b_{\beta k}, \sigma_\beta^2 \sim N\left(0, \sigma_\beta^2 / b_{\beta k}\right), b_{\beta k} \sim \text{Inverse-Gamma}\left(1, \frac{1}{2}\right), \;\; 1 \leq k \leq d_\circ + d_\bullet. \tag{6.15}$$

$$\gamma_{\beta k} \sim \text{Bernoulli}\left(\rho_\beta\right), 1 \leq k \leq d_\circ + d_\bullet. \tag{6.16}$$

equivalent to

$$\beta_k \sim \text{Laplace-Zero}\left(0, \sigma_\beta^2, \rho_\beta\right), 1 \leq k \leq d_\circ + d_\bullet. \tag{6.17}$$

The settings of scale and mixture parameters of $\beta_k, 1 \leq k \leq d_\circ + d_\bullet$ are the same as (2.16).

6.3.5 Spline Coefficient Vectors with Auxiliary Variable Representations

Similar to Section 2.3.5 for GAM, we have the following for GAMM

$$\boldsymbol{u}_k \equiv \gamma_{uk}\widetilde{\boldsymbol{u}}_k, \quad 1 \leq k \leq d_\bullet. \tag{6.18}$$

$$\widetilde{\boldsymbol{u}}_k \big| b_{uk}, \sigma^2_{uk} \overset{\text{ind.}}{\sim} N\left(\boldsymbol{0}, \left(\sigma^2_{uk}/b_{uk}\right)\boldsymbol{I}_{L_k}\right), b_{uk} \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(\frac{1}{2}\left(L_k+1\right), \frac{1}{2}\right). \tag{6.19}$$

$$\gamma_{uk} \sim \text{Bernoulli}\left(\rho_u\right), 1 \leq k \leq d_\bullet. \tag{6.20}$$

which is equivalent to

$$\boldsymbol{u}_k \sim \text{Laplace-Zero}\left(\boldsymbol{0}, \sigma^2_{uk}\boldsymbol{I}_{L_k}, \rho_u\right), 1 \leq k \leq d_\bullet. \tag{6.21}$$

The settings of scale and mixture parameters of $\boldsymbol{u}_k, 1 \leq k \leq d_\bullet$ are the similar to (2.24) as following:

$$\sigma^2_{uk} \big| a_{uk} \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(\frac{1}{2}, 1/a_{uk}\right), a_{uk} \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(\frac{1}{2}, 1/s^2_u\right), 1 \leq k \leq d_\bullet. \tag{6.22}$$

6.3.6 Standard Deviation with Auxiliary Variable Representations

As in Section 2.3.6, we have settings for the standard deviation with auxiliary variable representations the same as (2.26).

6.3.7 Scale and Mixture Parameters for Random Intercept with Auxiliary Variable Representations

In addition, in the generalized additive mixed model with random intercept for this chapter, we have the scale and mixture parameters for the random intercept $\boldsymbol{U}$ with auxiliary variable representation as follows:

$$\sigma^2_U | a_U \sim \text{Inverse-Gamma}\left(\frac{1}{2}, 1/a_U\right) \text{ and } a_U \sim \text{Inverse-Gamma}\left(\frac{1}{2}, 1/s^2_U\right). \tag{6.23}$$

6.4 Essence of the Bayesian Mixed Model

In this section we describe the Bayesian mixed model with random effect intercept.

6.4.1 The Likelihood of the Response

The likelihood of the response of our full model is:

$$\boldsymbol{y}_i\big|\beta_0, U_i, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}_1, \ldots, \widetilde{\boldsymbol{u}}_k, \gamma_{u1}, \ldots, \gamma_{ud_\bullet}, \sigma_\varepsilon^2 \overset{\text{ind.}}{\sim} N\left(\mathbf{1}_{ni}\beta_0 + \mathbf{1}_{ni}U_i + \boldsymbol{X}_i\left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}\right) + \sum_{k=1}^{d_\bullet} \boldsymbol{Z}_{ik}\left(\gamma_{uk}\widetilde{\boldsymbol{u}}_k\right), \sigma_\varepsilon^2 \boldsymbol{I}_{ni}\right), \quad 1 \le i \le m. \tag{6.24}$$

6.4.2 Gaussian Response Model Selection

For the Gaussian response case, based on the likelihood function in (6.24) given in Section 6.4.1 and model description described in Section 6.3, the full model is:

$$\boldsymbol{y}_i\big|\beta_0, U_i, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{\beta}}, \gamma_{u1}, \ldots, \gamma_{d_\bullet}, \widetilde{\boldsymbol{u}}_1, \ldots, \widetilde{\boldsymbol{u}}_{d_\bullet}, \sigma_\varepsilon^2 \overset{\text{ind.}}{\sim} N\left(\mathbf{1}_{ni}\beta_0 + \mathbf{1}_{ni}U_i + \boldsymbol{X}_i(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) + \sum_{k=1}^{d_\bullet} \boldsymbol{Z}_{ik}(\gamma_{uk}\widetilde{\boldsymbol{u}}_k), \sigma_\varepsilon^2 \boldsymbol{I}_{ni}\right), \quad 1 \le i \le m,$$

$$\beta_0 \sim N(0, \sigma_{\beta_0}^2), \;\; \sigma_\varepsilon^2\big|a_\varepsilon \sim \text{Inverse-Gamma}(\tfrac{1}{2}, 1/a_\varepsilon), \;\; a_\varepsilon \sim \text{Inverse-Gamma}(\tfrac{1}{2}, 1/s_\varepsilon^2),$$

$$U_i\big|\sigma_U^2 \sim N\left(0, \sigma_U^2\right); \;\; 1 \le i \le m,$$

$$\sigma_U^2\big|a_U \sim N\left(\tfrac{1}{2}, 1/a_U\right), a_U \sim \text{Inverse-Gamma}\left(\tfrac{1}{2}, 1/s_U^2\right),$$

$$\gamma_{\beta k} \overset{\text{ind.}}{\sim} \text{Bernoulli}(\rho_\beta), \;\; \widetilde{\beta}_k\big|\sigma_\beta^2, b_{\beta k} \overset{\text{ind.}}{\sim} N(0, \sigma_\beta^2/b_{\beta k}), \;\; 1 \le k \le d_\circ + d_\bullet,$$

$$b_{\beta k} \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(1, \tfrac{1}{2}), \quad 1 \le k \le d_\circ + d_\bullet,$$

$$\gamma_{uk} \overset{\text{ind.}}{\sim} \text{Bernoulli}(\rho_u), \;\; 1 \le k \le d_\bullet,$$

$$\widetilde{\boldsymbol{u}}_k\big|\sigma_{uk}^2, b_{uk} \overset{\text{ind.}}{\sim} N(\mathbf{0}, (\sigma_{uk}^2/b_{uk})\boldsymbol{I}_{L_k}), \;\; b_{uk} \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2}(L_k+1), \tfrac{1}{2}), \;\; 1 \le k \le d_\bullet,$$

$$\sigma_\beta^2\big|a_\beta \sim \text{Inverse-Gamma}(\tfrac{1}{2}, 1/a_\beta), \;\; a_\beta \sim \text{Inverse-Gamma}(\tfrac{1}{2}, 1/s_\beta^2),$$

$$\sigma_{uk}^2\big|a_{uk} \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2}, 1/a_{uk}), \;\; a_{uk} \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2}, 1/s_u^2), \;\; 1 \le k \le d_\bullet. \tag{6.25}$$

The full set of hyperparameters in model (6.25) is:

$$\sigma_{\beta_0} = 10^5, \;\; s_\beta = s_\varepsilon = s_U = s_u = 1000, \;\; \rho_\beta = \rho_u = \tfrac{1}{2}.$$

Figure 6.1 shows the directed acyclic graph corresponding to model (6.25). The open circle $\boldsymbol{U}$ at the bottom of Figure 6.1 is the random intercept for each group. It is a $m \times 1$ vector. The mixed model terminology is due to the fact that model (6.25) is a statistical model containing both fixed effects and random effects.

As in Section 2.4.2, the directed acyclic graph in Figure 6.1 is useful for the determination of full conditional distributions of parameters for Markov chain Monte Carlo fitting algorithms.

6.4.3 Binary Response Model Selection

For binary responses, below is the adjustment to our likelihood to model (6.25) described in Section 6.4.2, with $1 \le i \le m, 1 \le j \le n_i$:

$$\begin{aligned} y_{ij} \big| \beta_0, U_i, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{\beta}}, \gamma_{u1}, \ldots, \gamma_{u_{d_\bullet}}, \widetilde{\boldsymbol{u}}_1, \ldots, \widetilde{\boldsymbol{u}}_{d_\bullet} \\ \overset{\text{ind.}}{\sim} \text{Bernoulli} \left(\Phi \left(\beta_0 + U_i + \left(\boldsymbol{X}_i (\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) + \sum_{k=1}^{d_\bullet} \boldsymbol{Z}_{ik} (\gamma_{uk} \widetilde{\boldsymbol{u}}_k) \right)_j \right) \right). \end{aligned} \tag{6.26}$$

Similar to Section 2.4.3, we introduce an auxiliary random variables vector $\boldsymbol{c} \equiv (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_m)$, with $\boldsymbol{c}_i \equiv (c_1, \ldots, c_{n_i}), 1 \le i \le m$, such that

$$\begin{aligned} & y_{ij} = 1 \quad \text{if and only if } c_{ij} \ge 0, \;\; 1 \le i \le m, \;\; 1 \le j \le n_i, \\ & y_{ij} \big| c_{ij} \overset{\text{ind.}}{\sim} \text{Bernoulli} \left(I \left(c_{ij} \ge 0 \right) \right) \end{aligned} \tag{6.27}$$

and impose the following conditional distribution on $\boldsymbol{c} \equiv (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_m), 1 \le i \le m$:

$$\begin{aligned} \boldsymbol{c}_i \big| \beta_0, U_i, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{\beta}}, \gamma_{u1}, \ldots, \gamma_{u_{d_\bullet}}, \widetilde{\boldsymbol{u}}_1, \ldots, \widetilde{\boldsymbol{u}}_{d_\bullet} \\ \sim N \Big(\boldsymbol{1}_{n_i} \beta_0 + \boldsymbol{1}_{n_i} U_i + \boldsymbol{X}_i (\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) + \sum_{k=1}^{d_\bullet} \boldsymbol{Z}_{ik} (\gamma_{uk} \widetilde{\boldsymbol{u}}_k), \boldsymbol{I}_{n_i} \Big). \end{aligned} \tag{6.28}$$

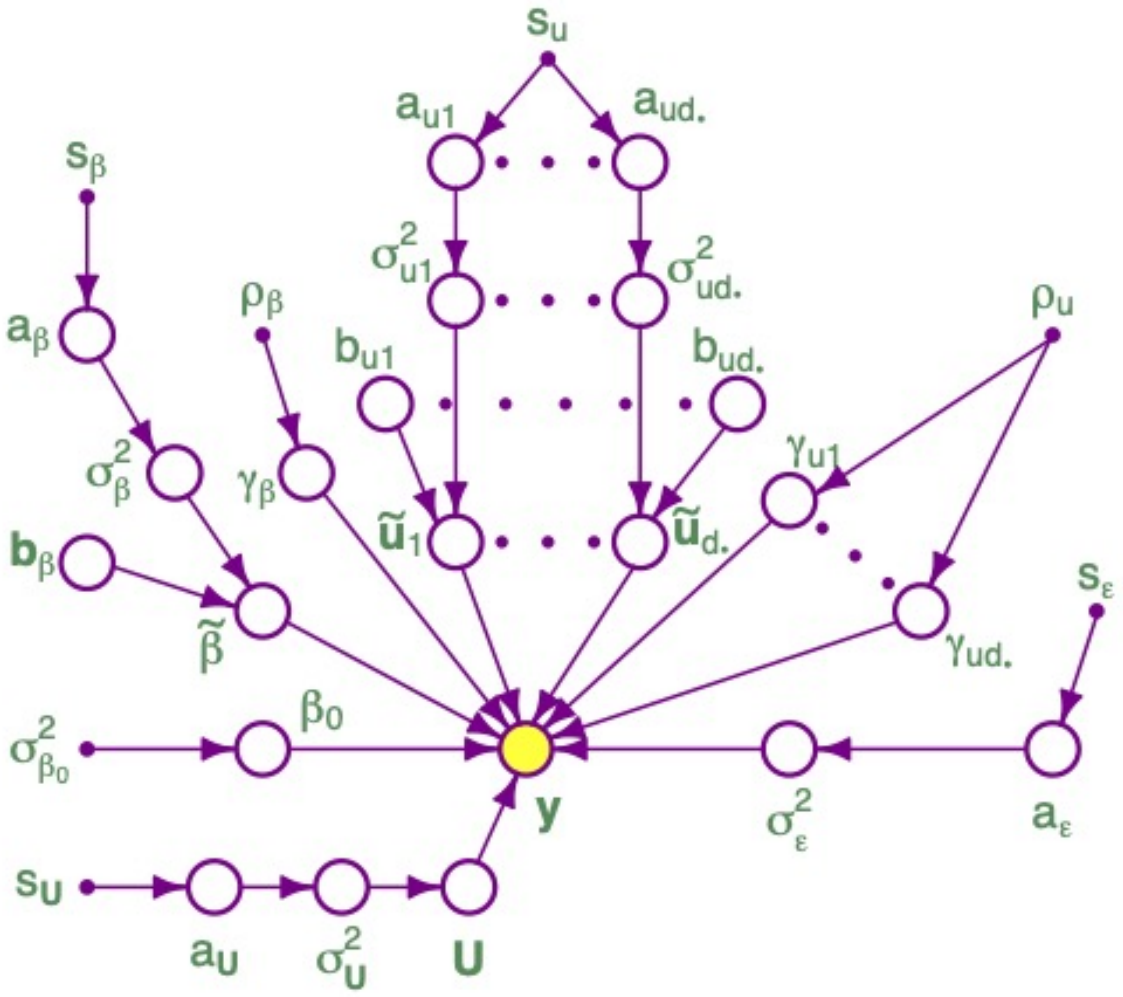


Figure 6.1: *Directed acyclic graph representation of Bayesian model (6.25) for Gaussian response. Random variables and vectors are shown as large open circles, with shading indicating to the observed response data. The vector* $\boldsymbol{y} = \left(\boldsymbol{y}_{1n_1}, \ldots, \boldsymbol{y}_{mn_m}\right)$ *is represented by a single shaded node. The small closed circles are user-specified hyperparameters.*

6.5 Posterior Distributions for Markov Chain Monte Carlo Practical Fitting

As in Section 2.5.1, we use the sufficient statistic matrices described in Section 6.5.1 for Markov chain Monte Carlo fitting algorithms for generalized additive mixed model selection. This aids computational speed-ups and scalability to large data sets.

In order to implement the practical fitting algorithm of Markov chain Monte Carlo, we need to work out the approximation of the posterior distributions of each

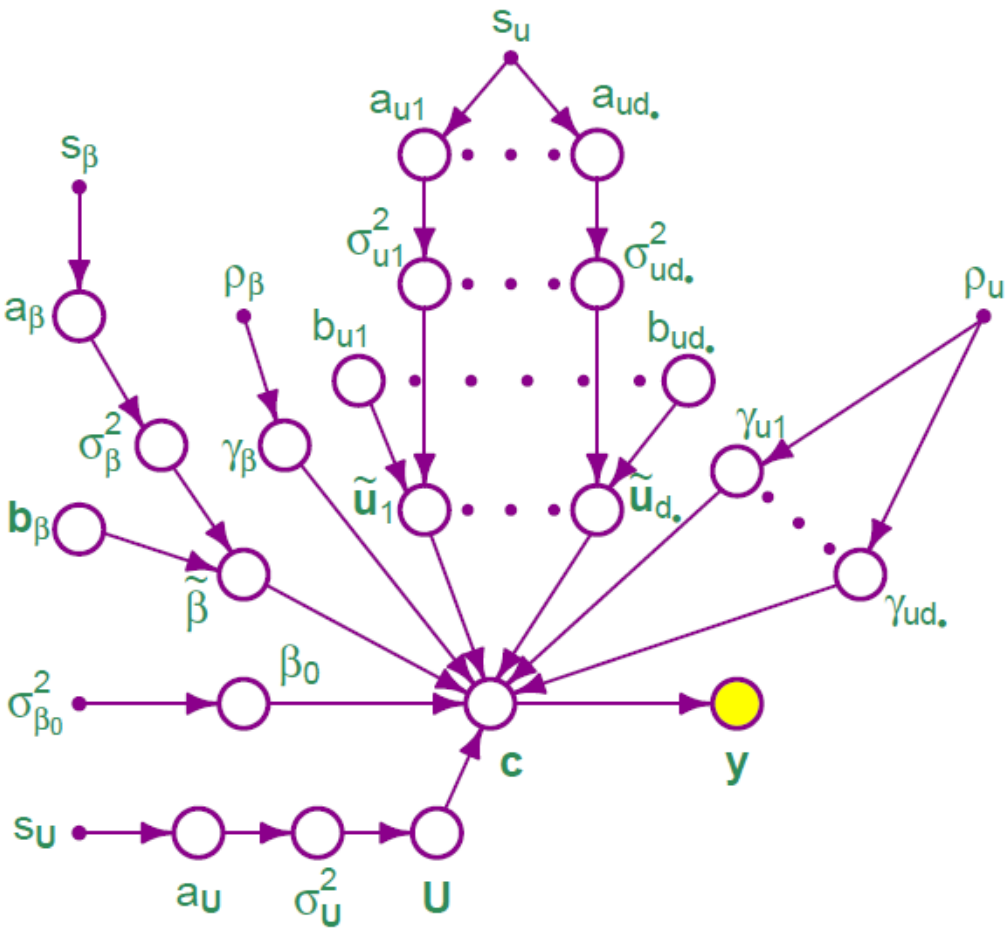


Figure 6.2: *Directed acyclic graph representation of Bayesian model (6.26) to (6.28) for binary response, adjustment to (6.25). This graph is the same as that shown in Figure 6.1 except for locations near the response variables node. The new graph has the following modifications: (1) the σ^2_ε and a_ε nodes are absent, (2) a hidden node $\boldsymbol{c}$ corresponding to the Albert-Chib auxiliary variables is added to the position held by $\boldsymbol{y}$ in the Gaussian response graph and the binary response observed data node $\boldsymbol{y}$ is a child of $\boldsymbol{c}$.*

parameter, as in Section 2.5.3. The full conditional distribution results are listed in Section 6.5.3.

6.5.1 Pre-Processing and Storage of Key Matrices

For the same purpose as described in Section 2.5.1, Algorithm 8 in this subsection provides detailed steps of standardization of the original grouped input data described in Section 6.2.2, and also the storage of sufficient statistics matrices such as $\boldsymbol{X}^T\boldsymbol{y}, \boldsymbol{X}^T\boldsymbol{X}, \boldsymbol{Z}^T\boldsymbol{y}, \boldsymbol{Z}^T\boldsymbol{X}, \boldsymbol{Z}^T\boldsymbol{Z}$ for computational speed-ups .

Algorithm 8: *Pre-processing of original data and creation of key matrices for input into Bayesian generalized additive mixed model algorithms.*

Inputs: $N = \sum_{i=1}^{m} n_i$, $\boldsymbol{y}^{\text{orig}}\,(N \times 1)$; $\mathring{\boldsymbol{x}}_k^{\text{orig}}\,(N \times 1)\,, 1 \leq k \leq d_\circ$; $\dot{\boldsymbol{x}}_k^{\text{orig}}\,(N \times 1)\,, 1 \leq k \leq d_\bullet$

mean($\boldsymbol{y}^{\text{orig}}$) $\longleftarrow$ sample mean of $\boldsymbol{y}^{\text{orig}}$; st.dev($\boldsymbol{y}^{\text{orig}}$) $\longleftarrow$ sample standard dev'n of $\boldsymbol{y}^{\text{orig}}$

If $\boldsymbol{y}^{\text{orig}}$ is continuous then $\boldsymbol{y} \longleftarrow \{\boldsymbol{y}^{\text{orig}} - \text{mean}(\boldsymbol{y}^{\text{orig}})\mathbf{1}_N\}/\text{st.dev}(\boldsymbol{y}^{\text{orig}})$

If $\boldsymbol{y}^{\text{orig}}$ is binary then $\boldsymbol{y} \longleftarrow \boldsymbol{y}^{\text{orig}}$

For $k = 1, \ldots, d_\circ$:

mean($\mathring{\boldsymbol{x}}_k^{\text{orig}}$)$\longleftarrow$ sample mean of $\mathring{\boldsymbol{x}}_k^{\text{orig}}$; st.dev($\mathring{\boldsymbol{x}}_k^{\text{orig}}$)$\longleftarrow$ sample standard dev'n of $\mathring{\boldsymbol{x}}_k^{\text{orig}}$

$\mathring{\boldsymbol{x}}_k \longleftarrow \{\mathring{\boldsymbol{x}}_k^{\text{orig}} - \text{mean}(\mathring{\boldsymbol{x}}_k^{\text{orig}})\mathbf{1}_N\}/\text{st.dev}(\mathring{\boldsymbol{x}}_k^{\text{orig}})$

For $k = 1, \ldots, d_\bullet$:

mean($\dot{\boldsymbol{x}}_k^{\text{orig}}$)$\longleftarrow$ sample mean of $\dot{\boldsymbol{x}}_k^{\text{orig}}$; st.dev($\dot{\boldsymbol{x}}_k^{\text{orig}}$)$\longleftarrow$ sample standard dev'n of $\dot{\boldsymbol{x}}_k^{\text{orig}}$

$\dot{\boldsymbol{x}}_k \longleftarrow \{\dot{\boldsymbol{x}}_k^{\text{orig}} - \text{mean}(\dot{\boldsymbol{x}}_k^{\text{orig}})\mathbf{1}_N\}/\text{st.dev}(\dot{\boldsymbol{x}}_k^{\text{orig}})$

$\boldsymbol{X} \longleftarrow \left[\mathring{\boldsymbol{x}}_1 \ldots \mathring{\boldsymbol{x}}_{d_\circ} \dot{\boldsymbol{x}}_1 \ldots \dot{\boldsymbol{x}}_{d_\bullet}\right]$

For $k = 1, \ldots, d_\bullet$:

$\dot{\boldsymbol{Z}}_k \longleftarrow N \times L_k$ matrix containing canonical Demmler-Reinsch basis for the data vector $\dot{\boldsymbol{x}}_k$, predictor using the construction described in Appendix 2.A.

$\boldsymbol{Z} \longleftarrow \left[\dot{\boldsymbol{Z}}_1 \ldots \dot{\boldsymbol{Z}}_{d_\bullet}\right]$; $\mathbf{XTy} \longleftarrow \boldsymbol{X}^T\boldsymbol{y}$; $\mathbf{XTX} \longleftarrow \boldsymbol{X}^T\boldsymbol{X}$; $\mathbf{ZTX} \longleftarrow \boldsymbol{Z}^T\boldsymbol{X}$; $\mathbf{ZTZ} \longleftarrow \boldsymbol{Z}^T\boldsymbol{Z}$;

Outputs: $\boldsymbol{y}, \boldsymbol{X}, \dot{\boldsymbol{Z}}_1, \ldots, \dot{\boldsymbol{Z}}_{d_\bullet}$, **XTy**, **XTX**, **ZTX**, **ZTZ**, mean($\boldsymbol{y}^{\text{orig}}$), st.dev($\boldsymbol{y}^{\text{orig}}$),

$\left\{(\text{mean}(\mathring{\boldsymbol{x}}_k^{\text{orig}}), \text{st.dev}(\mathring{\boldsymbol{x}}_k^{\text{orig}})) : 1 \leq k \leq d_\circ\right\}, \left\{(\text{mean}(\dot{\boldsymbol{x}}_k^{\text{orig}})), \text{st.dev}(\dot{\boldsymbol{x}}_k^{\text{orig}}) : 1 \leq k \leq d_\bullet\right\}$

6.5.2 Notation Used in the Fitting Algorithms

The notations used in Algorithm 9 and Algorithm 10 are similar to those used in Section 2.5.2. With $1 \leq k, k' \leq d_\bullet$, we have

$$\begin{aligned} \mathbf{ZTy}^{\langle k \rangle} &\equiv \dot{\boldsymbol{Z}}_k^T \boldsymbol{y} \\ \mathbf{ZTX}^{\langle k \rangle} &\equiv \dot{\boldsymbol{Z}}_k^T \boldsymbol{X} \\ \mathbf{ZTZ}^{\langle k,k' \rangle} &\equiv \dot{\boldsymbol{Z}}_k^T \dot{\boldsymbol{Z}}_k \end{aligned} \tag{6.29}$$

to locate a particular row or a particular column of the sufficient statistic matrices described in Section 6.5.1, with $\dot{\boldsymbol{Z}}_k$ described in (6.12). Their definition are similar to the definitions of $\mathbf{ZTy}^{\langle j \rangle}$, $\mathbf{ZTX}^{\langle j \rangle}$, $\mathbf{ZTZ}^{\langle j,j' \rangle}$, with $1 \leq j \leq d_\bullet$ in Section 2.5.2. We have:

$$\begin{aligned}
&\boldsymbol{L}_k \equiv \text{the number of columns in } \dot{\boldsymbol{Z}}_k, 1 \leq k \leq d_\bullet, \\
&\boldsymbol{e}_\ell^T \mathbf{ZTX}^{<k>} \equiv \text{ the } \ell\text{th row of } \mathbf{ZTX}^{<k>}, \text{ where } \boldsymbol{e}_\ell \text{ is th } L_k \times 1 \text{ vector with } \ell\text{th entry 1} \\
&\qquad\qquad \text{and 0 elsewhere, with } 1 \leq \ell \leq L_k, \\
&\mathfrak{c} \text{ is the } (d_\bullet + 1) \times 1 \text{ vector with entries } \mathfrak{c}_1 \equiv 0 \text{ and } \mathfrak{c}_{k+1} \equiv \sum_{\ell=1}^{k} L_\ell, 1 \leq k \leq d_\bullet, \\
&\mathbf{ZTy}^{\langle k \rangle} \equiv \text{the sub-block of } \mathbf{ZTy} \text{ corresponding to rows } (\mathfrak{c}_k + 1) \text{ to } \mathfrak{c}_{k+1}, 1 \leq k \leq d_\bullet, \\
&\mathbf{ZTX}^{\langle k \rangle} \equiv \text{the sub-block of } \mathbf{ZTX} \text{ corresponding to rows } (\mathfrak{c}_k + 1) \text{ to } \mathfrak{c}_{k+1}, 1 \leq k \leq d_\bullet \\
&\text{and } \mathbf{ZTZ}^{\langle k,k' \rangle} \equiv \text{the sub-block of } \mathbf{ZTZ} \text{ corresponding to rows } (\mathfrak{c}_k + 1) \text{ to } \mathfrak{c}_{k+1} \\
&\qquad\qquad \text{and columns } (\mathfrak{c}_{k'} + 1) \text{ to } (\mathfrak{c}_{k'+1}), 1 \leq k, k' \leq d_\bullet.
\end{aligned} \tag{6.30}$$

With $N \equiv \sum_{i=1}^{m} n_i$, we use notations $\mathbf{yT1}_{\text{adj}}$, $\mathbf{XTy}_{\text{adj}}$ and $\mathbf{ZTy}_{\text{adj}}$ denoted to be:

$$\begin{aligned}
\mathbf{yT1}_{\text{adj}} &\equiv \text{adjustment of } \mathbf{yT1} \equiv \begin{cases} \boldsymbol{y}^T \mathbf{1}_N & \text{for Gaussian responses,} \\ \boldsymbol{c}^T \mathbf{1}_N & \text{for binary responses,} \end{cases} \\
\mathbf{XTy}_{\text{adj}} &\equiv \text{adjustment of } \mathbf{XTy} \equiv \begin{cases} \boldsymbol{X}^T \boldsymbol{y} & \text{for Gaussian responses,} \\ \boldsymbol{X}^T \boldsymbol{c} & \text{for binary responses,} \end{cases} \\
\mathbf{ZTy}_{\text{adj}} &\equiv \text{adjustment of } \mathbf{ZTy} \equiv \begin{cases} \boldsymbol{Z}^T \boldsymbol{y} & \text{for Gaussian responses,} \\ \boldsymbol{Z}^T \boldsymbol{c} & \text{for binary responses.} \end{cases}
\end{aligned} \tag{6.31}$$

6.5.3 Markov Chain Monte Carlo

For Gaussian response, based on the model in Section 6.4.2 and Figure 6.1, we provide full algorithmic details for Markov chain Monte Carlo-based Bayesian inference. Using Markov blanket theory, we obtain the full conditional distributions of parameters, which are the posterior distributions, given in (6.32) – (6.48) in Section

6.5.3, with derivation details given in Appendix 6.A. The results are:

$$\beta_0|\text{rest} \sim N\left(\frac{\left(\frac{1}{\sigma_\varepsilon^2}\right)\left(\left(\boldsymbol{y}^T\mathbf{1}\right)-\sum_{i=1}^{m}\left(n_iU_i\right)\right)}{\left(\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{m}n_i+\frac{1}{\sigma_{\beta_0}^2}\right)},\frac{1}{\left(\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{m}n_i+\frac{1}{\sigma_{\beta_0}^2}\right)}\right), \tag{6.32}$$

$$\begin{aligned}U_i|\text{rest}&\\ &\sim N\left(\left(\frac{n_i}{\sigma_\varepsilon^2}+\frac{1}{\sigma_U^2}\right)^{-1}\left(\frac{1}{\sigma_\varepsilon^2}\right)\mathbf{1}_{ni}^T\left(\boldsymbol{y}_i-\mathbf{1}_{ni}\beta_0-\boldsymbol{X}_i(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}})-\sum_{k=1}^{d_\bullet}\boldsymbol{Z}_{ik}(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)\right),\right.\\ &\left.\left(\frac{n_i}{\sigma_\varepsilon^2}+\frac{1}{\sigma_U^2}\right)^{-1}\right),\ \ 1\le i\le m,\end{aligned} \tag{6.33}$$

$$\sigma_U^2|\text{rest} \propto \text{Inverse-Gamma}\left(\frac{m+1}{2},\frac{\left\|\boldsymbol{U}\right\|^2}{2}+\frac{1}{a_U}\right), \tag{6.34}$$

$$a_U|\text{rest} \sim \text{Inverse-Gamma}\left(1,\frac{1}{\sigma_U^2}+\frac{1}{s_U^2}\right), \tag{6.35}$$

$$\begin{aligned}\widetilde{\boldsymbol{\beta}}|\text{rest} \sim N\Bigg(&\left[\frac{1}{\sigma_\varepsilon^2}\left\{\left(\boldsymbol{\gamma}_\beta\boldsymbol{\gamma}_\beta^T\right)\odot\left(\boldsymbol{X}^T\boldsymbol{X}\right)\right\}+\frac{1}{\sigma_\beta^2}\text{diag}\left(\boldsymbol{b}_\beta\right)\right]^{-1}\left(\frac{1}{\sigma_\varepsilon^2}\right)\\ &\left[\boldsymbol{\gamma}_\beta\odot\left\{\left(\boldsymbol{X}^T\boldsymbol{y}\right)-\boldsymbol{X}^T\begin{bmatrix}U_1\mathbf{1}_{n_1}\\ \vdots\\ U_m\mathbf{1}_{n_m}\end{bmatrix}-\sum_{k=1}^{d_\bullet}\left(\dot{\boldsymbol{Z}}_k^T\boldsymbol{X}\right)^T(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)\right\}\right],\\ &\left.\left[\frac{1}{\sigma_\varepsilon^2}\left\{\left(\boldsymbol{\gamma}_\beta\boldsymbol{\gamma}_\beta^T\right)\odot\left(\boldsymbol{X}^T\boldsymbol{X}\right)\right\}+\frac{1}{\sigma_\beta^2}\text{diag}\left(\boldsymbol{b}_\beta\right)\right]^{-1}\right),\end{aligned} \tag{6.36}$$

$$\begin{aligned}\widetilde{\boldsymbol{u}}_k|\text{rest} \sim N\Bigg(&\text{diag}\left(\frac{\gamma_{uk}\boldsymbol{\omega}_{\dot{\boldsymbol{Z}}_k}}{\sigma_\varepsilon^2}+\frac{b_{uk}\mathbf{1}_{L_k}}{\sigma_{uk}^2}\right)^{-1}\\ &\times\left[\gamma_{uk}\left\{\dot{\boldsymbol{Z}}_k^T\boldsymbol{y}-\dot{\boldsymbol{Z}}_k^T\begin{bmatrix}U_1\mathbf{1}_{n_1}\\ \vdots\\ U_m\mathbf{1}_{n_m}\end{bmatrix}-\left(\dot{\boldsymbol{Z}}_k^T\boldsymbol{X}\right)\left(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}}\right)-\sum_{k'\neq k}^{d_\bullet}\left(\dot{\boldsymbol{Z}}_k^T\dot{\boldsymbol{Z}}_{k'}\right)(\gamma_{uk'}\widetilde{\boldsymbol{u}}_{k'})\right\}\right],\end{aligned}$$

$$\text{diag}\left(\frac{\gamma_{uk}\boldsymbol{\omega}_{\dot{\boldsymbol{Z}}_k}}{\sigma_\varepsilon^2}+\frac{b_{uk}\mathbf{1}_{L_k}}{\sigma_{uk}^2}\right)^{-1}\Bigg), 1\le k\le d_\bullet, \tag{6.37}$$

$$\gamma_{\beta_k}\big|\text{rest}\sim\text{Bernoulli}\left(\frac{1}{1+\exp\left(-\eta_{\beta_k}\right)}\right), 1\le k\le d_\circ+d_\bullet, \text{ where}$$

$$\begin{aligned}\eta_{\beta_k}=\text{logit}(\rho_\beta)-\frac{1}{2\sigma_\varepsilon^2}\Bigg\{&\widetilde{\beta}_k^2\left\|\dot{\boldsymbol{X}}_k\right\|^2-2\widetilde{\beta}_k\Bigg(\left(\dot{\boldsymbol{X}}_k^T\boldsymbol{y}\right)-\dot{\boldsymbol{X}}_k^T\begin{bmatrix}U_1\mathbf{1}_{n_1}\\ \vdots\\ U_m\mathbf{1}_{n_m}\end{bmatrix}\\&-\left(\dot{\boldsymbol{X}}_{(-k)}^T\dot{\boldsymbol{X}}_k\right)^T\left\{\left(\boldsymbol{\gamma}_\beta\right)_{-k}\odot\left(\widetilde{\boldsymbol{\beta}}\right)_{-k}\right\}-\sum_{k'=1}^{d_\bullet}\left(\dot{\boldsymbol{Z}}_{k'}^T\dot{\boldsymbol{X}}_k\right)^T\left(\gamma_{uk'}\widetilde{\boldsymbol{u}}_{k'}\right)\Bigg)\Bigg\}\end{aligned} \tag{6.38}$$

and

$$\gamma_{uk}\big|\text{rest}\sim\text{Bernoulli}\left(\frac{1}{1+\exp(-\gamma_{uk})}\right), 1\le k\le d_\bullet, \text{ where}$$

$$\begin{aligned}\eta_{uk}=\text{logit}(\rho_u)-\frac{1}{2\sigma_\varepsilon^2}\Bigg(&\boldsymbol{\omega}_{\dot{\boldsymbol{Z}}_k}^T\left(\widetilde{\boldsymbol{u}}_k\odot\widetilde{\boldsymbol{u}}_k\right)-2\widetilde{\boldsymbol{u}}_k^T\Bigg\{\dot{\boldsymbol{Z}}_k^T\boldsymbol{y}\\&-\dot{\boldsymbol{Z}}_k^T\begin{bmatrix}U_1\mathbf{1}_{n_1}\\ \vdots\\ U_m\mathbf{1}_{n_m}\end{bmatrix}-\dot{\boldsymbol{Z}}_k^T\boldsymbol{X}(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}})-\sum_{k'\neq k}^{d_\bullet}\dot{\boldsymbol{Z}}_k^T\dot{\boldsymbol{Z}}_{k'}^T(\gamma_{uk'}\widetilde{\boldsymbol{u}}_{k'})\Bigg\}\Bigg).\end{aligned} \tag{6.39}$$

If $\boldsymbol{y}$ contains Gaussian responses, with $N=\sum_{i=1}^m n_i$,

$$\begin{aligned}\sigma_\varepsilon^2\big|\text{rest}\sim\text{Inverse-Gamma}\Bigg(&\frac{1+N}{2},\frac{1}{a_\varepsilon}+\frac{1}{2}\Bigg\|\boldsymbol{y}-\mathbf{1}_N\beta_0-\begin{bmatrix}U_1\mathbf{1}_{n_1}\\ \vdots\\ U_m\mathbf{1}_{n_m}\end{bmatrix}\\&-\boldsymbol{X}(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}})-\sum_{k=1}^{d_\bullet}\dot{\boldsymbol{Z}}_k(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)\Bigg\|^2\Bigg)\end{aligned} \tag{6.40}$$

and

$$a_\varepsilon\big|\text{rest}\sim\text{Inverse-Gamma}\left(1,\frac{1}{\sigma_\varepsilon^2}+\frac{1}{s_\varepsilon^2}\right). \tag{6.41}$$

If $\boldsymbol{y}$ contains binary responses, then

$$y_{ij}\big|c_{ij} \overset{\text{ind.}}{\sim} \text{Bernoulli}\left(I\left(c_{ij} \geq 0\right)\right), \;\; 1 \leq i \leq m, \;\; 1 \leq j \leq n_i,$$

$$c_{ij}\big|\text{rest} \sim \left(2y_{ij}-1\right) \text{Truncated-Normal}_{+}$$
$$\left(\left(2y_{ij}-1\right) \left[\mathbf{1}_{n_i}\beta_0 + \mathbf{1}_{n_i}U_i + \boldsymbol{X}_i\left(\boldsymbol{\gamma}_{\beta} \odot \widetilde{\boldsymbol{\beta}}\right) + \sum_{k=1}^{d_\bullet} \boldsymbol{Z}_{ik}\left(\gamma_{uk}\widetilde{\boldsymbol{u}}_k\right) \right]_j , 1 \right). \tag{6.42}$$

$$b_{\beta_k}\big|\text{rest} \sim \text{Inverse-Gaussian}\left(\frac{\sigma_{\beta}}{\left|\widetilde{\beta}_k\right|}, 1\right), \;\; 1 \leq k \leq d_\circ + d_\bullet, \tag{6.43}$$

$$b_{uk}\big|\text{rest} \sim \text{Inverse-Gaussian}\left(\frac{\sigma_{uk}}{\left\|\widetilde{\boldsymbol{u}}_k\right\|}, 1\right), 1 \leq k \leq d_\bullet, \tag{6.44}$$

$$\sigma_{\beta}^2\big|\text{rest} \sim \text{Inverse-Gamma}\left(\frac{d_\circ + d_\bullet + 1}{2}, \frac{\widetilde{\boldsymbol{\beta}}^T \text{diag}(\boldsymbol{b}_{\beta})\widetilde{\boldsymbol{\beta}}}{2} + \frac{1}{a_{\beta}}\right), \tag{6.45}$$

$$a_{\beta}\big|\text{rest} \sim \text{Inverse-Gamma}\left(1, \frac{1}{\sigma_{\beta}^2} + \frac{1}{s_{\beta}^2}\right), \tag{6.46}$$

$$\sigma_{uk}^2\big|\text{rest} \sim \text{Inverse-Gamma}\left(\frac{L_k + 1}{2}, \frac{b_{uk}\|\widetilde{\boldsymbol{u}}_k\|^2}{2} + \frac{1}{a_{uk}}\right), \;\; 1 \leq k \leq d_\bullet, \tag{6.47}$$

$$a_{uk}\big|\text{rest} \sim \text{Inverse-Gamma}\left(1, \frac{1}{\sigma_{uk}^2} + \frac{1}{s_u^2}\right), \;\; 1 \leq k \leq d_\bullet. \tag{6.48}$$

The full conditional distributions all have standard forms for practical implementation of the Markov chain Monte Carlo approach. These include Multivariate Normal, Inverse Gamma, Bernoulli and Beta distributions. Some full conditional distributions are Inverse Gaussian, Inverse Gamma, Bernoulli, Multivariate Normal, Truncated Normal distributions. Michael et al. (1976) provides an algorithm for drawing from Inverse Gaussian distributions. Robert (1995) provides effective solutions for drawing from Truncated-Normal distributions. Therefore, the posterior distributions for Markov chain Monte Carlo are fairly straightforward. The Markov chain Monte Carlo sampling reduces to Gibbs sampling for our model.

Algorithm 9 in this subsection lists the full set of steps needed to draw samples from the posterior distributions of the model parameters. The fact that most of the

draws only require the sufficient statistic matrices from Algorithm 8 given in Section 6.5.1 means that the sampling can be done quite rapidly regardless of sample size.

Algorithm 9: *Markov chain Monte Carlo generation of samples from the posterior distributions of the parameters in (6.25).*

Data Inputs: $N = \sum_{i=1}^{m} n_i$, $\boldsymbol{y}\,(N \times 1)$; $\boldsymbol{X}\,(N \times (d_\circ + d_\bullet))$; $\dot{\boldsymbol{Z}}_k\,(N \times L_k)\,, 1 \le k \le d_\bullet$.

Response Type Input: **responseType** $\in$ {Gaussian, Bernoulli} .

Sufficient Statistics Inputs: **XTy**, **XTX**, **ZTy**, **ZTX**, **ZTZ**.

Hyperparameter Inputs: $\sigma_{\beta_0}, s_\beta, s_\varepsilon, s_u, s_U > 0$, $0 < \rho_\beta, \rho_u < 1$.

Chain Length Inputs: N_{warm} and N_{kept} both positive integers.

Initialize: $\boldsymbol{\gamma}_\beta^{[0]} \longleftarrow \frac{1}{2}\mathbf{1}_{d_\circ + d_\bullet}; \gamma_{uk}^{[0]} \longleftarrow \frac{1}{2}, 1 \le k \le d_\bullet$; $\widetilde{\boldsymbol{\beta}}^{[0]} \longleftarrow \mathbf{0}_{d_\circ + d_\bullet}$

$\widetilde{\boldsymbol{u}}_k^{[0]} \longleftarrow \mathbf{0}_{L_k}, 1 \le k \le d_\bullet$; $(\sigma_\varepsilon^2)^{[0]} \longleftarrow 1$; $a_\varepsilon^{[0]} \longleftarrow 1$; $(\sigma_\beta^2)^{[0]} \longleftarrow 1$; $a_\beta^{[0]} \longleftarrow 1$

$\boldsymbol{b}_\beta^{[0]} \longleftarrow \mathbf{1}_{d_\circ + d_\bullet}$; $b_{uk}^{[0]} \longleftarrow 1, 1 \le k \le d_\bullet$;

$a_{uk}^{[0]} \longleftarrow 1, 1 \le k \le d_\bullet$; $(\sigma_{uk}^2)^{[0]} \longleftarrow 1, 1 \le k \le d_\bullet$,

$U_i^{[0]} \longleftarrow 0, 1 \le i \le m$; $\boldsymbol{c} \longleftarrow \frac{1}{2}\mathbf{1}_N$

$\mathbf{yT1}_{\text{adj}} \longleftarrow 0$; $\mathbf{XTy}_{\text{adj}} \longleftarrow \mathbf{XTy}$; $\mathbf{ZTy}_{\text{adj}} \longleftarrow \mathbf{ZTy}$

If **responseType** is Gaussian then $\mathbf{y}_{\text{adj}} \longleftarrow \boldsymbol{y}$

If **responseType** is Bernoulli then $\mathbf{y}_{\text{adj}} \longleftarrow \boldsymbol{c}$

For $k = 1, \ldots, d_\bullet$: $\boldsymbol{\omega}_{\dot{\boldsymbol{Z}}_k} \longleftarrow \text{diagonal}\left(\mathbf{ZTZ}^{\langle k,k\rangle}\right)$

For $g = 1, \ldots, N_{\text{warm}} + N_{\text{kept}}$:

$$\omega_1 \longleftarrow \mathbf{yT1}_{\text{adj}} - \sum_{i=1}^{m}\left(n_i U_i^{[g-1]}\right)$$

$$\omega_2 \longleftarrow \left\{N/\left(\sigma_\varepsilon^2\right)^{[g-1]} + \left(1/\sigma_{\beta_0}^2\right)\right\} \quad ; \quad \beta_0^{[g]} \sim N\Big(\frac{\omega_1}{\left(\sigma_\varepsilon^2\right)^{[g-1]}\omega_2}, \frac{1}{\omega_2}\Big)$$

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

$$\omega_3 \longleftarrow \mathbf{1}_{n_i}^T\left(\mathbf{y}_{\text{adj}}\right)_i - \boldsymbol{n}_i\beta_0^{[g]} - \left(\boldsymbol{X}_i^T\mathbf{1}_{n_i}\right)^T\left(\boldsymbol{\gamma}_\beta^{[g-1]} \odot \widetilde{\boldsymbol{\beta}}^{[g-1]}\right)$$

$$-\sum_{k=1}^{d_\bullet}\left(\boldsymbol{Z}_{ik}^T\mathbf{1}_{n_i}\right)^T\left(\gamma_{uk}^{[g-1]}\widetilde{\boldsymbol{u}}_k^{[g-1]}\right)$$

continued on a subsequent page ...

Algorithm 9 continued: *This is a continuation of the description of this algorithm that commences on a preceding page.*

$$\omega_4 \longleftarrow \left(\boldsymbol{n}_i / \left(\sigma_\varepsilon^2\right)^{[g-1]}\right) + \left(1 / \left(\sigma_U^2\right)^{[g-1]}\right);\ U_i^{[g]} \sim N\Big(\frac{\omega_3}{\left(\sigma_\varepsilon^2\right)^{[g-1]}\omega_4}, \frac{1}{\omega_4}\Big)$$

$$\left(\sigma_U^2\right)^{[g]} \longleftarrow \text{Inverse-Gamma}\left(\frac{m+1}{2}, \frac{\left\|\boldsymbol{U}^{[g]}\right\|^2}{2} + \frac{1}{a_U^{[g-1]}}\right)$$

$$a_U^{[g]} \longleftarrow \text{Inverse-Gamma}\left(1, \frac{1}{\left(\sigma_U^2\right)^{[g]}} + \frac{1}{s_U^2}\right)$$

$$\boldsymbol{\Omega} \longleftarrow \left(\boldsymbol{\gamma}_\beta^{[g-1]}\boldsymbol{\gamma}_\beta^{[g-1]T}\right) \odot (\mathbf{XTX}) \Big/ \left(\sigma_\varepsilon^2\right)^{[g-1]} + \text{diag}\Big(\boldsymbol{b}_\beta^{[g-1]}\Big) \Big/ \left(\sigma_\beta^2\right)^{[g-1]}$$

$$\boldsymbol{\omega}_5 \longleftarrow \mathbf{XTy}_{\text{adj}} - \left(\boldsymbol{X}^T \begin{bmatrix} U_1^{[g]}\mathbf{1}_{n_1} \\ \vdots \\ U_m^{[g]}\mathbf{1}_{n_m} \end{bmatrix}\right) - \sum_{k=1}^{d_\bullet} \mathbf{ZTX}^{\langle k \rangle T}\left(\gamma_{uk}^{[g-1]}\widetilde{\boldsymbol{u}}_k^{[g-1]}\right)$$

Decompose $\boldsymbol{\Omega} = \boldsymbol{U}_{\boldsymbol{\Omega}}\text{diag}\left(d_{\boldsymbol{\Omega}}\right)\boldsymbol{U}_{\boldsymbol{\Omega}}^T$ where $\boldsymbol{U}_{\boldsymbol{\Omega}}\boldsymbol{U}_{\boldsymbol{\Omega}}^T = \boldsymbol{I}$

$$\boldsymbol{z} \sim N(\mathbf{0}, \boldsymbol{I})((d_\circ + d_\bullet) \times 1) \quad ; \quad \widetilde{\boldsymbol{\beta}}^{[g]} \longleftarrow \boldsymbol{U}_{\boldsymbol{\Omega}}\left(\frac{\boldsymbol{U}_{\boldsymbol{\Omega}}^T\boldsymbol{z}}{\sqrt{d_{\boldsymbol{\Omega}}}} + \frac{\boldsymbol{U}_\Omega^T\left(\boldsymbol{\gamma}_\beta^{[g-1]} \odot \boldsymbol{\omega}_5\right)}{\boldsymbol{d}_{\boldsymbol{\Omega}}\left(\sigma_\varepsilon^2\right)^{[g-1]}}\right)$$

$$\left(\boldsymbol{b}_\beta^{[g]}\right)_k \sim \text{Inverse-Gaussian}\left(\sigma_\beta^{[g-1]} \Big/ \left|\left(\widetilde{\boldsymbol{\beta}}^{[g]}\right)_k\right|\right),\ 1 \le k \le d_\circ + d_\bullet$$

$$\left(\sigma_\beta^2\right)^{[g]} \sim \text{Inverse-Gamma}\left(\frac{1}{2}(d_\circ + d_\bullet + 1), 1/a_\beta^{[g-1]} + \frac{1}{2}\widetilde{\boldsymbol{\beta}}^{[g]T}\text{diag}\big(\boldsymbol{b}_\beta^{[g]}\big)\widetilde{\boldsymbol{\beta}}^{[g]}\right)$$

$$a_\beta^{[g]} \sim \text{Inverse-Gamma}\left(1, \left\{1/(\sigma_\beta^2)^{[g]}\right\} + (1/s_\beta^2)\right);\ \boldsymbol{\beta}^{\text{curr}} \longleftarrow \boldsymbol{\gamma}_\beta^{[g-1]} \odot \widetilde{\boldsymbol{\beta}}^{[g]}$$

For $k = 1, \ldots, d_\bullet$: $\quad \boldsymbol{u}_k^{\text{curr}} \longleftarrow \gamma_{uk}^{[g-1]}\widetilde{\boldsymbol{u}}_k^{[g-1]}$

For $k = 1, \ldots, d_\circ + d_\bullet$:

$$\omega_6 \longleftarrow \boldsymbol{e}_k^T\mathbf{XTy}_{\text{adj}} - \dot{\boldsymbol{X}}_k^T \begin{bmatrix} U_1^{[g]}\mathbf{1}_{n_1} \\ \vdots \\ U_m^{[g]}\mathbf{1}_{n_m} \end{bmatrix} - \Big(\mathbf{XTX}\boldsymbol{e}_k\Big)_{-k}^T\Big(\boldsymbol{\beta}^{\text{curr}}\Big)_{-k}$$

$$-\sum_{k'=1}^{d_\bullet}\Big(\mathbf{ZTX}^{\langle k' \rangle}\boldsymbol{e}_k\Big)^T\boldsymbol{u}_{k'}^{\text{curr}}$$

continued on a subsequent page ...

Algorithm 9 continued: *This is a continuation of the description of this algorithm that commences on a preceding page.*

$$\omega_7 \longleftarrow \text{logit}\left(\rho_\beta\right) - \frac{1}{2}\left\{(\widetilde{\beta}_k^{[g]})^2 \boldsymbol{e}_k^T \mathbf{X}\mathbf{T}\mathbf{X}\boldsymbol{e}_k - 2\widetilde{\beta}_k^{[g]}\omega_6\right\} \Big/ \left(\sigma_\varepsilon^2\right)^{[g-1]}$$

$$\left(\boldsymbol{\gamma}_\beta^{[g]}\right)_k \sim \text{Bernoulli}\left(\text{expit}\left(\omega_7\right)\right)$$

$$\boldsymbol{\beta}^{\text{curr}} \longleftarrow \boldsymbol{\gamma}_\beta^{[g]} \odot \widetilde{\boldsymbol{\beta}}^{[g]} \quad ; \quad \text{For } k = 1, \ldots, d_\bullet: \quad \widetilde{\boldsymbol{u}}_k^{\text{curr}} \longleftarrow \widetilde{\boldsymbol{u}}_k^{[g-1]}$$

For $k = 1, \ldots, d_\bullet$:

$$\boldsymbol{\omega}_8 \longleftarrow \mathbf{ZTy}_{\text{adj}}^{\langle k \rangle} - \dot{\boldsymbol{Z}}_k^T \begin{bmatrix} U_1^{[g]}\mathbf{1}_{n_1} \\ \vdots \\ U_m^{[g]}\mathbf{1}_{n_m} \end{bmatrix} - \mathbf{ZTX}^{\langle k \rangle}\boldsymbol{\beta}^{\text{curr}} - \sum_{k' \neq k}^{d_\bullet} \mathbf{ZTZ}^{\langle k,k' \rangle}\left(\gamma_{uk'}^{[g-1]}\widetilde{\boldsymbol{u}}_{k'}^{\text{curr}}\right)$$

$$\boldsymbol{\omega}_9 \longleftarrow \left\{\gamma_{uk}^{[g-1]}\boldsymbol{\omega}_{\dot{\boldsymbol{Z}}_k}\Big/\left(\sigma_\varepsilon^2\right)^{[g-1]}\right\} + \left\{b_{uk}^{[g-1]}\mathbf{1}_{L_k}\Big/\left(\sigma_{uk}^2\right)^{[g-1]}\right\}$$

$$\boldsymbol{z} \sim N(\mathbf{0}, \boldsymbol{I}_{L_k}) \quad ; \quad \widetilde{\boldsymbol{u}}_k^{\text{curr}} \longleftarrow \left(\boldsymbol{z}\Big/\sqrt{\boldsymbol{\omega}_9}\right) + \left[\gamma_{uk}^{[g-1]}\boldsymbol{\omega}_8\Big/\left\{\boldsymbol{\omega}_9\left(\sigma_\varepsilon^2\right)^{[g-1]}\right\}\right]$$

For $k = 1, \ldots, d_\bullet$: $\quad \widetilde{\boldsymbol{u}}_k^{[g]} \longleftarrow \widetilde{\boldsymbol{u}}_k^{\text{curr}}$

For $k = 1, \ldots, d_\bullet$:

$$b_{uk}^{[g]} \sim \text{Inverse-Gaussian}\left(\sigma_{uk}^{[g-1]}\Big/\left\|\widetilde{\boldsymbol{u}}_k^{[g]}\right\|, 1\right)$$

$$\left(\sigma_{uk}^2\right)^{[g]} \sim \text{Inverse-Gamma}\left(\tfrac{1}{2}(L_k+1), \left\{1\Big/a_{uk}^{[g-1]}\right\} + \tfrac{1}{2}\left\|\widetilde{\boldsymbol{u}}_k^{[g]}\right\|^2 b_{uk}^{[g]}\right)$$

$$a_{uk}^{[g]} \sim \text{Inverse-Gamma}\Big(1, \left\{1\Big/\left(\sigma_{uk}^2\right)^{[g]}\right\} + \left(1\Big/s_u^2\right)\Big)$$

For $k = 1, \ldots, d_\bullet$: $\quad \gamma_{uk}^{\text{curr}} \longleftarrow \gamma_{uk}^{[g-1]}$

For $k = 1, \ldots, d_\bullet$:

$$\boldsymbol{\omega}_{10} \longleftarrow \mathbf{ZTy}_{\text{adj}}^{\langle k \rangle} - \dot{\boldsymbol{Z}}_k^T \begin{bmatrix} U_1^{[g]}\mathbf{1}_{n_1} \\ \vdots \\ U_m^{[g]}\mathbf{1}_{n_m} \end{bmatrix} - \mathbf{ZTX}^{\langle k \rangle}\boldsymbol{\beta}^{\text{curr}} - \sum_{k' \neq k}^{d_\bullet} \mathbf{ZTZ}^{\langle k,k' \rangle}\left(\gamma_{uk'}^{\text{curr}}\widetilde{\boldsymbol{u}}_{k'}^{[g]}\right)$$

$$\omega_{11} \longleftarrow \text{logit}(\rho_u) - \frac{1}{2}\left\{\boldsymbol{\omega}_{\dot{\boldsymbol{Z}}_k}^T\left(\widetilde{\boldsymbol{u}}_k^{[g]} \odot \widetilde{\boldsymbol{u}}_k^{[g]}\right) - 2\left(\widetilde{\boldsymbol{u}}_k^{[g]}\right)^T\boldsymbol{\omega}_{10}\right\} \Big/ \left(\sigma_\varepsilon^2\right)^{[g-1]}$$

$$\gamma_{uk}^{\text{curr}} \sim \text{Bernoulli}\left(\text{expit}\left(\omega_{11}\right)\right)$$

continued on a subsequent page ...

Algorithm 9 continued: *This is a continuation of the description of this algorithm that commences on a preceding page.*

For $k = 1, \ldots, d_\bullet$: $\quad \gamma_{uk}^{[g]} \longleftarrow \gamma_{uk}^{\text{curr}}$

$$\boldsymbol{\omega}_{12} \longleftarrow \mathbf{1}_N \beta_0^{[g]} + \begin{bmatrix} U_1^{[g]} \mathbf{1}_{n_1} \\ \vdots \\ U_m^{[g]} \mathbf{1}_{n_m} \end{bmatrix} + \boldsymbol{X}\left(\boldsymbol{\gamma}_\beta^{[g]} \odot \widetilde{\boldsymbol{\beta}}^{[g]}\right) + \sum_{k=1}^{d_\bullet} \dot{\boldsymbol{Z}}_k (\gamma_{uk}^{[g]} \widetilde{\boldsymbol{u}}_k^{[g]})$$

If **responseType** is Gaussian then

$$\left(\sigma_\varepsilon^2\right)^{[g]} \sim \text{Inverse-Gamma}\left(\tfrac{1}{2}(n+1), (1 \Big/ a_\varepsilon^{[g-1]}) + \tfrac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{\omega}_{12}\right\|^2\right)$$

$$a_\varepsilon^{[g]} \sim \text{Inverse-Gamma}\left(1, \left\{1 \Big/ (\sigma_\varepsilon^2)^{[g]}\right\} + (1 \Big/ s_\varepsilon^2)\right)$$

If **responseType** is Bernoulli then

$(\sigma_\varepsilon^2)^{[g]} \longleftarrow 1$

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

For $j = 1, \ldots, n_i$:

$\omega_{13} \longleftarrow \text{Truncated-Normal}_+\left((2y_{ij} - 1)(\boldsymbol{\omega}_{12})_{ij}, 1\right)$; $\; c_{ij} \longleftarrow (2y_{ij} - 1)\omega_{13}$

$\mathbf{yT1}_{\text{adj}} \longleftarrow \mathbf{1}^T \boldsymbol{c}$; $\; \mathbf{XTy}_{\text{adj}} \longleftarrow \boldsymbol{X}^T \boldsymbol{c}$; $\; \mathbf{ZTy}_{\text{adj}} \longleftarrow \boldsymbol{Z}^T \boldsymbol{c}$; $\; \mathbf{y}_{\text{adj}} \longleftarrow \boldsymbol{c}$

Outputs: All chains after omission of the first N_{warm} values.

6.6 Mean Field Variational Bayes Model Selection

6.6.1 Product Density Restrictions

Consider the Bayesian inference for the parameters in (6.25). Similar to Section 4.2.1, the essence of our mean field variational Bayes approach involves approximation of the full joint posterior density function of the form

$$\begin{aligned}&\mathfrak{p}(\beta_0, \boldsymbol{U}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_u, \widetilde{\boldsymbol{u}}, \boldsymbol{b}_\beta, \sigma_\beta^2, a_\beta, \boldsymbol{b}_u, \boldsymbol{\sigma}_u^2, \boldsymbol{a}_u, \sigma_\varepsilon^2, a_\varepsilon \big| \boldsymbol{y}) \\ &\quad \approx \mathfrak{q}(\beta_0, \boldsymbol{U}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_u, \widetilde{\boldsymbol{u}}, \boldsymbol{b}_\beta, \sigma_\beta^2, a_\beta, \boldsymbol{b}_u, \boldsymbol{\sigma}_u^2, \boldsymbol{a}_u, \sigma_\varepsilon^2, a_\varepsilon). \end{aligned} \tag{6.49}$$

(6.49) is subject to

$$\begin{aligned}&\mathfrak{q}(\beta_0, \boldsymbol{U}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_u, \widetilde{\boldsymbol{u}}, \boldsymbol{b}_\beta, \sigma_\beta^2, a_\beta, \boldsymbol{b}_u, \boldsymbol{\sigma}_u^2, \boldsymbol{a}_u, \sigma_\varepsilon^2, a_\varepsilon) \\ &\quad = \mathfrak{q}(\beta_0)\mathfrak{q}(\boldsymbol{U})\mathfrak{q}(\boldsymbol{\gamma}_\beta)\mathfrak{q}(\widetilde{\boldsymbol{\beta}})\mathfrak{q}(\boldsymbol{\gamma}_u)\mathfrak{q}(\widetilde{\boldsymbol{u}})\mathfrak{q}(\boldsymbol{b}_\beta)\mathfrak{q}(\sigma_\beta^2)\mathfrak{q}(a_\beta)\mathfrak{q}(\boldsymbol{b}_u)\mathfrak{q}(\boldsymbol{\sigma}_u^2)\mathfrak{q}(\boldsymbol{a}_u)\mathfrak{q}(\sigma_\varepsilon^2)\mathfrak{q}(a_\varepsilon). \end{aligned} \tag{6.50}$$

Therefore, the mean field approximation of the joint posterior density function of the model parameters to (6.25) for Gaussian response is

$$\begin{aligned}&\mathfrak{p}(\beta_0, \boldsymbol{U}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_u, \widetilde{\boldsymbol{u}}, \boldsymbol{b}_\beta, \sigma_\beta^2, a_\beta, \boldsymbol{b}_u, \boldsymbol{\sigma}_u^2, \boldsymbol{a}_u, \sigma_\varepsilon^2, a_\varepsilon \big| \boldsymbol{y}) \\ &\quad \approx \mathfrak{q}(\beta_0)\mathfrak{q}(\boldsymbol{U})\mathfrak{q}(\boldsymbol{\gamma}_\beta)\mathfrak{q}(\widetilde{\boldsymbol{\beta}})\mathfrak{q}(\boldsymbol{\gamma}_u)\mathfrak{q}(\widetilde{\boldsymbol{u}})\mathfrak{q}(\boldsymbol{b}_\beta)\mathfrak{q}(\sigma_\beta^2)\mathfrak{q}(a_\beta)\mathfrak{q}(\boldsymbol{b}_u)\mathfrak{q}(\boldsymbol{\sigma}_u^2)\mathfrak{q}(\boldsymbol{a}_u)\mathfrak{q}(\sigma_\varepsilon^2)\mathfrak{q}(a_\varepsilon) \\ &\quad \approx \mathfrak{q}(\beta_0)\left\{\textstyle\prod_{i=1}^{m}\mathfrak{q}(U_i)\right\}\left\{\textstyle\prod_{k=1}^{d_\circ+d_\bullet}\mathfrak{q}(\gamma_{\beta k})\right\}\mathfrak{q}(\widetilde{\boldsymbol{\beta}})\left\{\textstyle\prod_{k=1}^{d_\bullet}\mathfrak{q}(\gamma_{uk})\right\}\left\{\textstyle\prod_{k=1}^{d_\bullet}\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)\right\} \\ &\qquad \times\left\{\textstyle\prod_{k=1}^{d_\circ+d_\bullet}\mathfrak{q}(b_{\beta k})\right\}\mathfrak{q}(\sigma_\beta^2)\mathfrak{q}(a_\beta)\left\{\textstyle\prod_{k=1}^{d_\bullet}\mathfrak{q}(b_{\beta k})\right\}\left\{\textstyle\prod_{k=1}^{d_\bullet}\mathfrak{q}(\sigma_{uk}^2)\right\} \\ &\qquad \times\left\{\textstyle\prod_{k=1}^{d_\bullet}\mathfrak{q}(a_{uk})\right\}\mathfrak{q}(\sigma_\varepsilon^2)\mathfrak{q}(a_\varepsilon). \end{aligned} \tag{6.51}$$

6.6.2 $\mathfrak{q}$-Densities Forms

Based on full conditional distributions of parameters for Markov chain Monte Carlo given in Section 6.5.3, for the Gaussian and binary responses models, we obtain $\mathfrak{q}$-density approximation of parameters for mean field variational Bayes in this subsection. Some examples of the resulting optimal $\mathfrak{q}$-density forms are:

$\mathfrak{q}(\beta_0)$ has a $N(\mu_{\mathfrak{q}(\beta_0)}, \sigma^2_{\mathfrak{q}(\beta_0)})$ density function,

$\mathfrak{q}(U_i)$ has a $N\left(\mu_{\mathfrak{q}(U_i)}, \sigma^2_{\mathfrak{q}(U_i)}\right)$ density function, $1 \leq i \leq m$,

σ^2_U has an Inverse-Gamma $\left(\kappa_{\mathfrak{q}(\sigma^2_U)}, \lambda_{\mathfrak{q}(\sigma^2_U)}\right)$ density function,

$\mathfrak{q}(\widetilde{\boldsymbol{\beta}})$ has a $N(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}, \boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})})$ density function,

$\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)$ has a $N(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}, \boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)})$ density function, $1 \leq k \leq d_\bullet$,

$\mathfrak{q}(\gamma_{\beta k})$ has a Bernoulli $\left(\dfrac{1}{1+\exp\left\{E_{\mathfrak{q}(\eta_{\beta k})}\right\}}\right)$ probability mass function,

$1 \leq k \leq d_\circ + d_\bullet$,

$\mathfrak{q}(\gamma_{uk})$ has a Bernoulli $\left(\dfrac{1}{1+\exp\left\{E_{\mathfrak{q}(\eta_{uk})}\right\}}\right)$ probability mass function,

$1 \leq k \leq d_\bullet$,

$\mathfrak{q}(\boldsymbol{b}_\beta)$is a product of Inverse-Gaussian $\left(\mu_{\mathfrak{q}(b_{\beta k})}, 1\right)$ density functions,

$1 \leq k \leq d_\circ + d_\bullet$,

$\mathfrak{q}(\boldsymbol{b}_u)$ is a product of Inverse-Gaussian $\left(\mu_{\mathfrak{q}(b_{uk})}, 1\right)$ denstiy functions,

$1 \leq k \leq d_\bullet$,

$\mathfrak{q}(\sigma^2_\varepsilon)$ has an Inverse-Gamma$(\kappa_{\mathfrak{q}(\sigma^2_\varepsilon)}, \lambda_{\mathfrak{q}(\sigma^2_\varepsilon)})$ density function,

$\mathfrak{q}(\sigma^2_\beta)$ has an Inverse-Gamma $\left(\kappa_{\mathfrak{q}(\sigma^2_\beta)}, \lambda_{\mathfrak{q}(\sigma^2_\beta)}\right)$ density function,

$\mathfrak{q}(\sigma^2_{uk})$ has an Inverse-Gamma $\left(\kappa_{\mathfrak{q}(\sigma^2_{uk})}, \lambda_{\mathfrak{q}(\sigma^2_{uk})}\right)$ density function, $1 \leq k \leq d_\bullet$.

(6.52)

The q-density approximations of parameters for mean field variational Bayes are listed in (6.53)–(6.82), with derivation details given in Appendix 6.B. The results are:

$\mathfrak{q}(\beta_0)$ has a $N\left(\mu_{\mathfrak{q}(\beta_0)}, \sigma^2_{\mathfrak{q}(\beta_0)}\right)$ density function, where

$$\sigma^2_{\mathfrak{q}(\beta_0)} = \left(\left(\sum_{i=1}^m n_i\right)\mu_{\mathfrak{q}(1/\sigma^2_\varepsilon)} + \frac{1}{\sigma^2_{\beta_0}}\right)^{-1}, \tag{6.53}$$

$$\mu_{\mathfrak{q}(\beta_0)} = \sigma^2_{\mathfrak{q}(\beta_0)}\mu_{\mathfrak{q}(1/\sigma^2_\varepsilon)}\left(\mathbf{1}_n^T\boldsymbol{y} - \left(\sum_{i=1}^m n_i U_i\right)\right), \tag{6.54}$$

$\mathfrak{q}(U_i)$ has a $N\left(\mu_{\mathfrak{q}(U_i)}, \sigma^2_{\mathfrak{q}(U_i)}\right)$ density function , $1 \le i \le m,$ where

$$\sigma^2_{\mathfrak{q}(U_i)} = \left(n_i\mu_{\mathfrak{q}(1/\sigma^2_\varepsilon)} + \mu_{\mathfrak{q}\left(1/\sigma^2_U\right)}\right)^{-1}, \tag{6.55}$$

$$\mu_{\mathfrak{q}(U_i)} = \sigma^2_{\mathfrak{q}(U_i)}\mu_{\mathfrak{q}(1/\sigma^2_\varepsilon)}\Bigg(\boldsymbol{y}_i^T\mathbf{1}_{n_i} - n_i\mu_{\mathfrak{q}(\beta_0)} - \left(\boldsymbol{X}_i^T\mathbf{1}_{n_i}\right)^T\left(\boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{\gamma}_\beta\right)} \odot \boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\right)$$

$$- \sum_{k=1}^{d_\bullet}\left(\boldsymbol{Z}_{ik}^T\mathbf{1}_{n_i}\right)^T\left(\mu_{\mathfrak{q}(\gamma_{uk})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right)\Bigg), \tag{6.56}$$

$\mathfrak{q}\left(\sigma^2_U\right)$ has an Inverse-Gamma $\left(\kappa_{\mathfrak{q}\left(\sigma^2_U\right)}, \lambda_{\mathfrak{q}\left(\sigma^2_U\right)}\right)$ density function where

$$\kappa_{\mathfrak{q}\left(\sigma^2_U\right)} = \frac{m+1}{2}, \tag{6.57}$$

$$\lambda_{\mathfrak{q}\left(\sigma^2_U\right)} = \left\|\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{U})}\right\|^2 + \sum_{i=1}^m \mu_{\mathfrak{q}\left(\sigma^2_{U_i}\right)}, \tag{6.58}$$

$\mathfrak{q}(a_U)$ has an Inverse-Gamma $\left(\kappa_{\mathfrak{q}(a_U)}, \lambda_{\mathfrak{q}(a_U)}\right)$ density function where

$$\kappa_{\mathfrak{q}(a_U)} = 1, \tag{6.59}$$

$$\lambda_{\mathfrak{q}(a_U)} = \mu_{\mathfrak{q}\left(1/\sigma^2_U\right)} + 1/s^2_U, \tag{6.60}$$

$\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)$ has a $N\left(\boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}, \boldsymbol{\Sigma}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\right)$ density function where

$$\boldsymbol{\Omega}_{\mathfrak{q}\left(\boldsymbol{\gamma}_\beta\right)} = \text{diag}\left\{\boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{\gamma}_\beta\right)} \odot \left(\mathbf{1} - \boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{\gamma}_\beta\right)}\right)\right\} + \boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{\gamma}_\beta\right)}\boldsymbol{\mu}^T_{\mathfrak{q}\left(\boldsymbol{\gamma}_\beta\right)}, \tag{6.61}$$

$$\boldsymbol{\Sigma}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)} = \left\{\mu_{\mathfrak{q}(1/\sigma^2_\varepsilon)}\left[\boldsymbol{\Omega}_{\mathfrak{q}\left(\boldsymbol{\gamma}_\beta\right)} \odot \left(\boldsymbol{X}^T\boldsymbol{X}\right)\right] + \mu_{\mathfrak{q}\left(1/\sigma^2_\beta\right)}\text{diag}\left(\boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{b}_\beta\right)}\right)\right\}^{-1}, \tag{6.62}$$

$$\boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)} = \mu_{\mathfrak{q}(1/\sigma^2_\varepsilon)}\boldsymbol{\Sigma}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\Bigg(\boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{\gamma}_\beta\right)} \odot \Bigg(\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T\begin{bmatrix}\mu_{\mathfrak{q}(U_1)}\mathbf{1}_{n_1}\\ \vdots \\ \mu_{\mathfrak{q}(U_m)}\mathbf{1}_{n_m}\end{bmatrix}$$

$$- \sum_{k=1}^{d_\bullet}\left(\dot{\boldsymbol{Z}}_k\boldsymbol{X}\right)^T\left(\mu_{\mathfrak{q}(\gamma_{uk})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right)\Bigg)\Bigg), \tag{6.63}$$

$\mathfrak{q}\left(b_{\beta_1},\ldots,b_{\beta_{(d_\circ+d_\bullet)}}\right)$ is a product of

Inverse-Gaussian $\left(\left\{\mu_{\mathfrak{q}\left(1/\sigma_\beta^2\right)}\left(\mu^2_{\mathfrak{q}\left(\widetilde{\beta}_k\right)}+\sigma^2_{\mathfrak{q}\left(\widetilde{\beta}_k\right)}\right)\right\}^{-1/2},1\right)$ density functions

over $1\leq k\leq d_\circ+d_\bullet$, The q-density means are

$$\mu_{\mathfrak{q}\left(b_{\beta k}\right)}=\left\{\mu_{\mathfrak{q}\left(1/\sigma_\beta^2\right)}\left(\mu^2_{\mathfrak{q}\left(\widetilde{\beta}_k\right)}+\sigma^2_{\mathfrak{q}\left(\widetilde{\beta}_k\right)}\right)\right\}^{-1/2}, \tag{6.64}$$

$\mathfrak{q}\left(\sigma_\beta^2\right)$ has an Inverse-Gamma $\left(\kappa_{\mathfrak{q}\left(\sigma_\beta^2\right)},\lambda_{\mathfrak{q}\left(\sigma_\beta^2\right)}\right)$ density function where

$$\kappa_{\mathfrak{q}\left(\sigma_\beta^2\right)}=\frac{d_\circ+d_\bullet+1}{2}, \tag{6.65}$$

$$\lambda_{\mathfrak{q}\left(\sigma_\beta^2\right)}=\frac{1}{2}\boldsymbol{\mu}^T_{\mathfrak{q}\left(\boldsymbol{b}_\beta\right)}\left\{\boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\odot\boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}+\text{diagonal}\left(\boldsymbol{\Sigma}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\right)\right\}+\mu_{\mathfrak{q}\left(1/a_\beta\right)}, \tag{6.66}$$

$\mathfrak{q}\left(a_\beta\right)$ has an Inverse-Gamma $\left(\kappa_{\mathfrak{q}\left(a_\beta\right)},\lambda_{\mathfrak{q}\left(a_\beta\right)}\right)$ density function where

$$\kappa_{\mathfrak{q}\left(a_\beta\right)}=1, \tag{6.67}$$

$$\lambda_{\mathfrak{q}\left(a_\beta\right)}=\mu_{\mathfrak{q}\left(1/\sigma_\beta^2\right)}+1/s_\beta^2, \tag{6.68}$$

$\mathfrak{q}\left(\gamma_{\beta k}\right)$ is the Bernoulli $\left(\frac{1}{1+\exp\left\{-E_{\mathfrak{q}}\left(\eta_{\beta_k}\right)\right\}}\right)$ probability mass function for each

$1\leq k\leq d_\circ+d_\bullet$, where

$$\begin{aligned}E_{\mathfrak{q}}\left(\eta_{\beta_k}\right)=\text{logit}\left(\rho_\beta\right)-\frac{1}{2}\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\Bigg\{&\left(\mu^2_{\mathfrak{q}\left(\widetilde{\beta}_k\right)}+\sigma^2_{\mathfrak{q}\left(\widetilde{\beta}_k\right)}\right)\dot{\boldsymbol{X}}_k^T\dot{\boldsymbol{X}}_k\\&-2\Bigg[\mu_{\mathfrak{q}\left(\widetilde{\beta}_k\right)}\Big\{\dot{\boldsymbol{X}}_k^T\boldsymbol{y}-\dot{\boldsymbol{X}}_k^T\begin{bmatrix}\mu_{\mathfrak{q}(U_1)}\mathbf{1}_{n_1}\\\vdots\\\mu_{\mathfrak{q}(U_m)}\mathbf{1}_{n_m}\end{bmatrix}-\sum_{k'=1}^{d_\bullet}\left(\dot{\boldsymbol{Z}}_{k'}^T\dot{\boldsymbol{X}}_k\right)^T\left(\mu_{\mathfrak{q}(\gamma_{uk'})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_{k'})}\right)\Big\}\\&-\left(\dot{\boldsymbol{X}}_{-k}^T\dot{\boldsymbol{X}}_k\right)^T\left\{\left(\boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{\gamma}_\beta\right)}\right)_{-k}\odot\left(\left(\boldsymbol{\Sigma}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\right)_{-k,k}+\mu_{\mathfrak{q}\left(\widetilde{\beta}_k\right)}\left(\boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\right)_{-k}\right)\right\}\Bigg]\Bigg\},\end{aligned} \tag{6.69}$$

$\mathfrak{q}\left(\widetilde{\boldsymbol{u}}_k\right)$ has a $N\left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)},\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right)$ density function where $1\leq k\leq d_\bullet$,

$\boldsymbol{\omega}_{\dot{\boldsymbol{Z}}_k}\equiv\text{diagonal}(\dot{\boldsymbol{Z}}_k^T\dot{\boldsymbol{Z}}_k)$,

$\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}=\text{diag}\left(\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right)$

$$= \text{diag}\left[\mathbf{1}_{L_k} \Big/ \left\{\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\mu_{\mathfrak{q}(\gamma_{uk})}\boldsymbol{\omega}_{\dot{\boldsymbol{Z}}_k} + \mu_{\mathfrak{q}\left(1/\sigma_{uk}^2\right)}\mu_{\mathfrak{q}(b_{uk})}\mathbf{1}_{L_k}\right\}\right], \tag{6.70}$$

$$\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)} = \mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\Bigg((\mu_{\mathfrak{q}(\gamma_{uk})}\Big\{\dot{\boldsymbol{Z}}_k^T\boldsymbol{y} - \dot{\boldsymbol{Z}}_k^T\begin{bmatrix}\mu_{\mathfrak{q}(U_1)}\mathbf{1}_{n_1}\\ \vdots \\ \mu_{\mathfrak{q}(U_m)}\mathbf{1}_{n_m}\end{bmatrix} - \left(\dot{\boldsymbol{Z}}_k^T\boldsymbol{X}\right)\left(\boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{\gamma}_\beta\right)} \odot \boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\right)$$

$$- \sum_{k' \neq k}^{d_\bullet}\dot{\boldsymbol{Z}}_k^T\dot{\boldsymbol{Z}}_{k'}\left(\mu_{\mathfrak{q}(\gamma_{uk'})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_{k'})}\right)\Big\}\Bigg) \odot \boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}, \tag{6.71}$$

$\mathfrak{q}\left(b_{u_1}, \ldots, b_{u_{(d_\bullet)}}\right)$ is a product of

Inverse-Gaussian $\left(\left\{\mu_{\mathfrak{q}\left(1/\sigma_{uk}^2\right)}\left[\left\|\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right\|^2 + \mathbf{1}_{L_k}^T\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right]\right\}^{-1/2}, 1\right)$ density functions

over $1 \leq k \leq d_\bullet$. The q-density means are

$$\mu_{\mathfrak{q}(b_{uk})} = \left\{\mu_{\mathfrak{q}\left(1/\sigma_{uk}^2\right)}\left[\left\|\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right\|^2 + \mathbf{1}_{L_k}^T\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right]\right\}^{-1/2}, \tag{6.72}$$

$\mathfrak{q}\left(\sigma_{uk}^2\right)$ has an Inverse-Gamma $\left(\kappa_{\mathfrak{q}\left(\sigma_{uk}^2\right)}, \lambda_{\mathfrak{q}\left(\sigma_{uk}^2\right)}\right)$ density function, $1 \leq k \leq d_\bullet$, where

$$\kappa_{\mathfrak{q}\left(\sigma_{uk}^2\right)} = \frac{L_k + 1}{2}, \tag{6.73}$$

$$\lambda_{\mathfrak{q}\left(\sigma_{uk}^2\right)} = \frac{1}{2}\mu_{\mathfrak{q}(b_{uk})}\left(\left\|\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right\|^2 + \mathbf{1}_{L_k}^T\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right) + \mu_{\mathfrak{q}(1/a_{uk})}, \tag{6.74}$$

$\mathfrak{q}\left(a_{uk}\right)$ has an Inverse-Gamma $\left(\kappa_{\mathfrak{q}(a_{uk})}, \lambda_{\mathfrak{q}(a_{uk})}\right)$ density function, $1 \leq k \leq d_\bullet$, where

$$\kappa_{\mathfrak{q}\left(\sigma_{uk}^2\right)} = 1, \tag{6.75}$$

$$\lambda_{\mathfrak{q}\left(\sigma_{uk}^2\right)} = \mu_{\mathfrak{q}\left(1/\sigma_{uk}^2\right)} + 1/s_u^2, \tag{6.76}$$

and

$\mathfrak{q}(\gamma_{uk})$ is the Bernoulli $\left(\dfrac{1}{1 + \exp\left(-E_{\mathfrak{q}}\left(\eta_{uk}\right)\right)}\right)$ probability mass function for each

$1 \leq k \leq d_\bullet$, where

$$E_{\mathfrak{q}}\left(\eta_{uk}\right) = \text{logit}\left(\rho_u\right) - \frac{1}{2}\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\Big\{\boldsymbol{\omega}_{\dot{\boldsymbol{Z}}_k}^T\left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)} + \boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right) - 2\boldsymbol{\mu}^T_{\mathfrak{q}(\widetilde{u}_k)}$$

$$\times \dot{\boldsymbol{Z}}_k^T \Bigg(\boldsymbol{y} - \begin{bmatrix} \mu_{\mathfrak{q}(U_1)} \mathbf{1}_{n_1} \\ \vdots \\ \mu_{\mathfrak{q}(U_m)} \mathbf{1}_{n_m} \end{bmatrix} - \boldsymbol{X} \left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \right) - \sum_{k' \neq k}^{d_\bullet} \dot{\boldsymbol{Z}}_{k'} \left(\mu_{\mathfrak{q}(\gamma_{uk'})} \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_{k'})} \right) \Bigg) \Bigg\}. \tag{6.77}$$

If $\boldsymbol{y}$ contains Gaussian responses, $\mathfrak{q}\left(\sigma_\varepsilon^2\right)$ has an Inverse-Gamma$\left(\kappa_{\mathfrak{q}(\sigma_\varepsilon^2)}, \lambda_{\mathfrak{q}(\sigma_\varepsilon^2)}\right)$

density function where

$$\kappa_{\mathfrak{q}(\sigma_\varepsilon^2)} = \frac{1}{2}(N+1), \quad N \equiv \sum_{i=1}^{m} n_i, \tag{6.78}$$

$$\begin{aligned} \lambda_{\mathfrak{q}(\sigma_\varepsilon^2)} = \mu_{\mathfrak{q}(1/a_\varepsilon)} &+ \frac{1}{2} \Bigg\| \boldsymbol{y} - \mathbf{1}_N^T \mu_{\mathfrak{q}(\beta_0)} - \begin{bmatrix} \mu_{\mathfrak{q}(U_1)} \mathbf{1}_{n_1} \\ \vdots \\ \mu_{\mathfrak{q}(U_m)} \mathbf{1}_{n_m} \end{bmatrix} - \boldsymbol{X} \left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \right) \\ &- \sum_{k=1}^{d_\bullet} \dot{\boldsymbol{Z}}_k \left(\mu_{\mathfrak{q}(\gamma_{uk})} \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)} \right) \Bigg\|^2 + \frac{1}{2} N \sigma^2_{\mathfrak{q}(\beta_0)} + \frac{1}{2} \sum_{i=1}^{m} \left(n_i \sigma^2_{\mathfrak{q}(U_i)} \right) \\ &+ \frac{1}{2} \mathrm{tr} \left\{ \boldsymbol{X}^T \boldsymbol{X} \left(\boldsymbol{\Omega}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \left\{ \boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} + \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \boldsymbol{\mu}^T_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \right\} \right) \right\} \\ &- \frac{1}{2} \mathrm{tr} \left\{ \boldsymbol{X}^T \boldsymbol{X} \left\{ \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \right\} \left\{ \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \right\}^T \right\} \\ &+ \frac{1}{2} \sum_{k=1}^{d_\bullet} \boldsymbol{\omega}^T_{\dot{\boldsymbol{Z}}_k} \left(\mu_{\mathfrak{q}(\gamma_{uk})} \left[\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)} + \left\{ 1 - \mu_{\mathfrak{q}(\gamma_{uk})} \right\} \left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)} \right) \right] \right) \end{aligned} \tag{6.79}$$

and

$\mathfrak{q}(a_\varepsilon)$ has an Inverse-Gamma $\left(\kappa_{\mathfrak{q}(a_\varepsilon)}, \lambda_{\mathfrak{q}(a_\varepsilon)}\right)$ density function where

$$\kappa_{\mathfrak{q}(a_\varepsilon)} = 1, \tag{6.80}$$

$$\lambda_{\mathfrak{q}(a_\varepsilon)} = \mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)} + 1 \big/ s_\varepsilon^2. \tag{6.81}$$

If $\boldsymbol{y}$ contains binary responses, then for $1 \leq i \leq m, \;\; 1 \leq j \leq n_i, \;\; y_{ij} = 1$ or $y_{ij} = 0,$

$\mathfrak{q}(\boldsymbol{c}_{ij})$ has a Truncated-Normal$_+$ $\left(\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{c})} \right)_{ij}, 1 \right)$ density function where

$$\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{c})}\right)_{ij} = (\boldsymbol{\mu}_{\mathfrak{q}})_{ij} + (2y_{ij}-1)\,\frac{\phi\left((2y_{ij}-1)\left(\boldsymbol{\mu}_{\mathfrak{q}}\right)_{ij}\right)}{\Phi\left((2y_{ij}-1)\left(\boldsymbol{\mu}_{\mathfrak{q}}\right)_{ij}\right)}, \tag{6.82}$$

$$(\boldsymbol{\mu}_{\mathfrak{q}})_{ij} = \mu_{\mathfrak{q}(\beta_0)} + U_i + \left[\boldsymbol{X}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right) + \sum_{k=1}^{d_\bullet} \dot{\boldsymbol{Z}}_k \left(\mu_{\mathfrak{q}(\gamma_{uk})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right)\right]_{ij}.$$

Using the optimal $\mathfrak{q}$-density results listed in (6.53)–(6.82), we can now assemble the mean field variational Bayes algorithm, given in Algorithm 10.

We proceed by initialising each of the $\mathfrak{q}$-density parameters and updating each parameter successively using the current estimates of the other parameters. At the end of each iteration, an updated value of the lower bound on the marginal log-likelihood is computed, and the algorithm is iterated until convergence of the lower bound on the marginal log-likelihood to its maximum. The approximate marginal log-likelihood, also known as the evidence lower bound, is given in the next subsection.

6.6.3 Variational Marginal Log-Likelihood Approximation

A final aspect of Algorithm 10 in Section 6.6.2 is determination of good stopping criteria for the coordinate ascent scheme. As is common in the mean field variational Bayes literature, we monitor relative increases in the approximate marginal log-likelihood, which we denote by $\log \underline{p}(\boldsymbol{y}; \mathfrak{q})$. Minimizing the Kullback-Liebler divergence is equivalent to maximizing the lower bound on the marginal log-likelihood $\log \underline{p}(\boldsymbol{y}; \mathfrak{q})$.

Appendix 6.C contains the explicit expression for the approximate marginal log-likelihood for Chapter 6 models under product restriction (6.51). By computing the log-likelihood after each iteration, we obtained its corresponding relative error. The setting of the tolerance of relative error is $\varepsilon_{\text{toler.}} = 10^{-8}$. As shown in Algorithm 10 in Section 6.6.2, we checked the relative error after each iteration. Once the relative error of log-likelihood is lower than the tolerance level, the iteration stopped.

6.7 Model Selection Strategies

After obtaining the full conditional distributions of parameters for Markov chain Monte Carlo described in Section 6.5.3 or obtaining the $\mathfrak{q}$-densities approximation

described in Section 6.6.2 for mean field variational Bayes, for generalized additive mixed model selection with random intercept, we need to set up the model selection strategies to decide between the effect type of a predictor being zero, linear or non-linear. In practice, either the Markov chain Monte Carlo samples or mean field variational Bayes $\mathfrak{q}$-densities are used for approximate posterior-based decision making. However, we will demonstrate our selection strategies in terms of exact posterior distributions – starting with the zero effect versus linear effect decision.

6.7.1 Deciding Between an Effect Being Zero or Linear

For predictors that can only enter the model linearly (e.g. binary predictors), as in (2.52) in Section 2.6.1 for generalized additive model selection, our rule for deciding between a zero effect and a linear effect is

$$\begin{aligned} &\text{the effect is zero if } E(\gamma_\beta|\boldsymbol{y}) \leq 1-\tau, \\ &\text{otherwise the effect is linear.} \end{aligned} \tag{6.83}$$

For Markov chain Monte Carlo approach, $E(\gamma_\beta|\boldsymbol{y})$ is approximate the sample mean of $\gamma_\beta^{[g]}$ in Algorithm 9, $1 \leq g \leq N_{\text{kept}}$. For mean field variational Bayes approach, we have $E\left(\gamma_\beta|\boldsymbol{y}\right) \equiv \mu_{\mathfrak{q}(\gamma_\beta)}$ in Algorithm 10.

6.7.2 Deciding Between an Effect Being Zero, Linear or Non-Linear

For predictors that can enter the model either linearly or non-linearly, as in (2.54) in Section 2.6.2 for generalized additive model selection, our strategy of deciding between an effect being zero, linear or non-linear is:

$$\begin{aligned} &\text{the effect is zero if } \begin{cases} E(\gamma_\beta|\boldsymbol{y}) \leq 1-\tau, \text{ and} \\ E(\gamma_u|\boldsymbol{y}) \leq 1-\tau, \end{cases} \\ &\text{the effect is linear if } \begin{cases} E(\gamma_\beta|\boldsymbol{y}) > 1-\tau, \text{ and} \\ E(\gamma_u|\boldsymbol{y}) \leq 1-\tau, \end{cases} \\ &\text{otherwise the effect is non-linear.} \end{aligned} \tag{6.84}$$

For Markov chain Monte Carlo approach, $E(\gamma_\beta|\boldsymbol{y})$ and $E\left(\gamma_u|\boldsymbol{y}\right)$ are approximate the sample mean of $\gamma_\beta^{[g]}$ and $\gamma_u^{[g]}$ in Algorithm 9, $1 \leq g \leq N_{\text{kept}}$. For mean field varia-

tional Bayes approach, we have $E\left(\gamma_\beta \middle| \boldsymbol{y}\right) \equiv \mu_{\mathfrak{q}\left(\gamma_\beta\right)}$ and $E\left(\gamma_u \middle| \boldsymbol{y}\right) \equiv \mu_{\mathfrak{q}(\gamma_u)}$ in Algorithm 10.

6.7.3 Choice of Default Value of Sparsity Threshold Parameter

To assess the effect of the sparsity threshold parameter τ for selection of a generalized additive mixed model with random intercept, we ran a simulation study similar to the one described in Section 2.6.3 for generalized additive model. We use the settings as stated in (2.55), (2.56), (2.57) and the settings of the number of continuous predictors $d_\bullet$, number of groups m, total number of observations N across all groups, error standard deviations σ_ε, and sparsity threshold parameter τ are to be

$$\begin{aligned}
&d_\bullet = 30, \text{ number of groups } m = \{100, 200, 300\}, \\
&\text{the value of } n_i,\ 1 \leq i \leq m, \text{ being the number of observations within each group,} \\
&\qquad \text{is fixed at } m/10, \\
&\text{error standard deviation } \sigma_\varepsilon \in \{0.25, 0.5, 1, 2\} \text{ for Gaussian response,} \\
&\text{standard deviation for random intercept: } \sigma_U = 0.5, \\
&\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}.
\end{aligned} \tag{6.85}$$

For each combination of number of groups m and error standard deviation σ_ε, we simulated 100 replicated data sets generated. For the Markov chain Monte Carlo approach, we use $N_{\text{warm}} = N_{\text{kept}} = 1000$ from Algorithm 9 in Section 6.5.3. The rules of model selection follow (6.83) of Section 6.7.1 and (6.84) of Section 6.7.2.

Both Figure 6.3 and Figure 6.4 show the misclassification rate as a function of sparsity threshold parameter τ under Markov chain Monte Carlo in generalized additive mixed model with random intercept. The performance measure was misclassification rate for 30 candidate predictors being classified into one of three classes: zero effect, linear effect or non-linear effect. Each panel corresponds to a different combination of number of groups m and error standard deviation σ_ε. Within each panel, the side-by-side boxplots compare the misclassification rates under different values of sparsity threshold parameter τ. Figure 6.3, for Gaussian response, low noise level with $\sigma_\varepsilon < 1$, there is not much of a difference in misclassification rates for $\tau \in \{0.1, 0.3, 0.5, 0.7\}$. However, for Gaussian response, high noise level with $\sigma_\varepsilon \geq 1$, it has the advantages of having τ being 0.5 for lower misclassification rate. Figure 6.4 shows that having τ being 0.5 is the best choice for threshold parameter

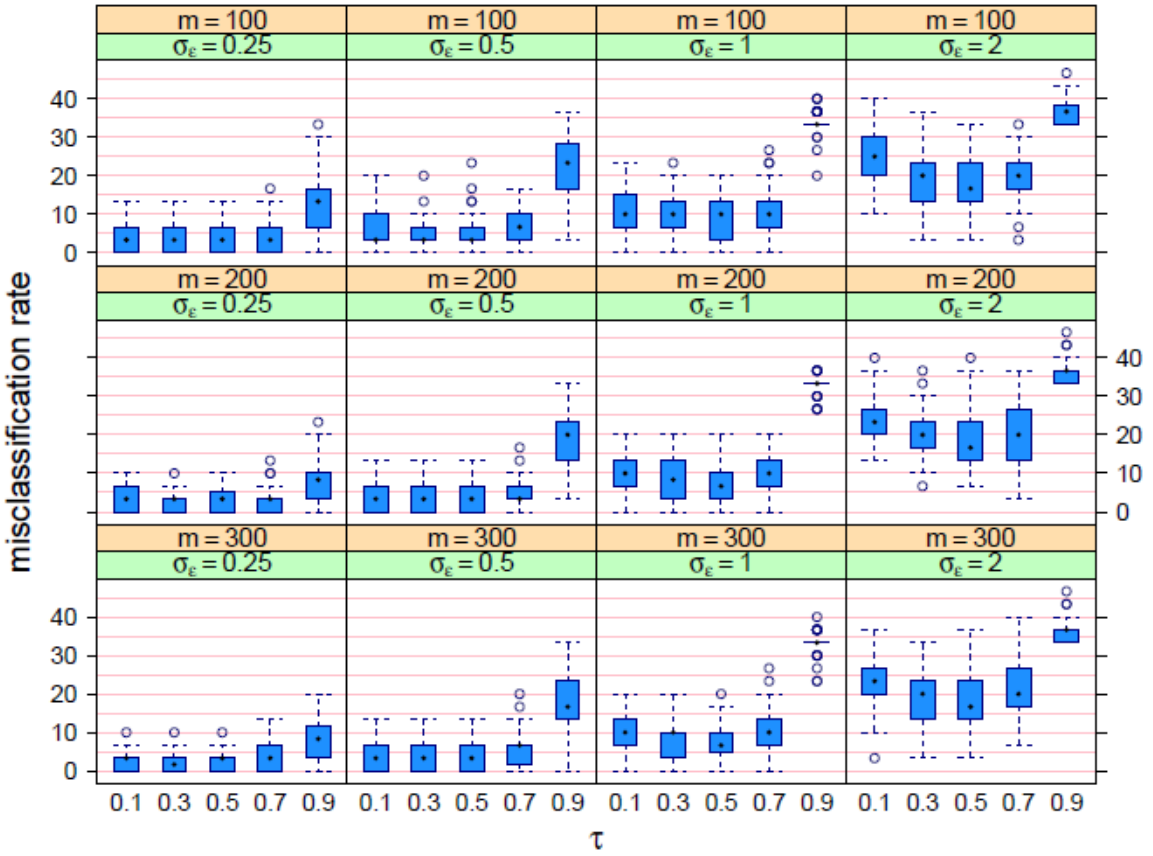


Figure 6.3: *Side-by-side boxplots of the misclassification rates for the Markov chain Monte Carlo Algorithm 9 for Gaussian response for the simulation study described in the text. Each panel corresponds to a different combination of group size and error standard deviation. Within each panel, the side-by-side boxplots compare misclassification rate as a function of the sparsity threshold parameter τ. The value of m is the number of groups. The number of observations within each group is fixed at $m/10$.*

for binary response. Therefore, the default value of sparsity threshold parameter τ for Markov chain Monte Carlo in generalized additive mixed model with random intercept is 0.5.

For MFVB, we have $E(\gamma_{\beta k}) \equiv \mu_{\mathfrak{q}(\gamma_{\beta k})}, 1 \leq k \leq d_\circ + d_\bullet$, and $E(\gamma_{uk}) \equiv \mu_{\mathfrak{q}(\gamma_{uk})}, 1 \leq k \leq d_\bullet$. For binary or categorical predictors, the selection rule is the same as (6.83) for MCMC. For continuous predictors, the selection rule is the same as (6.84) for MCMC.

To make our methodology completely automatic, a good default value of τ as threshold is desirable. The simulation study settings of the best sparsity threshold parameter τ for MFVB, are the same as the one for MCMC in Section 6.7.3, stated in (2.55), (2.56), (2.57) and (6.85).

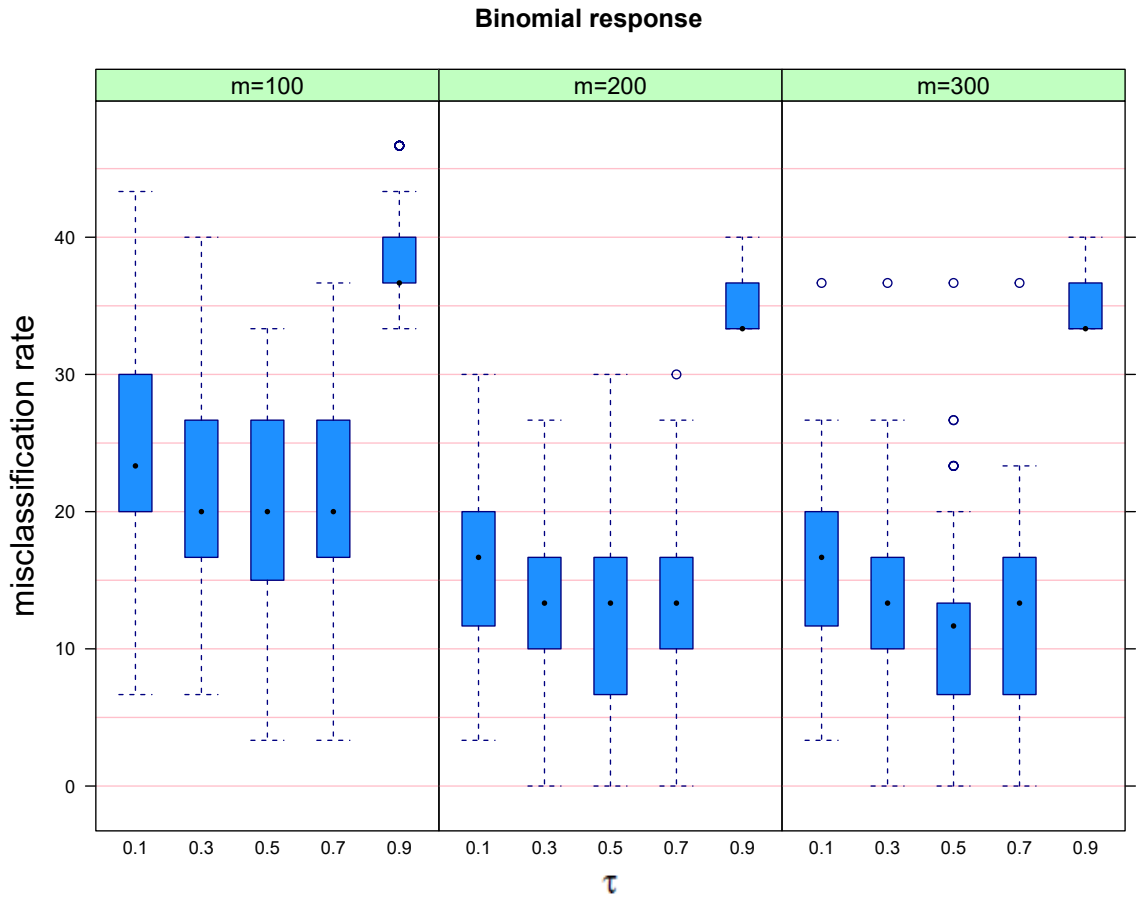


Figure 6.4: *Side-by-side boxplots of the misclassification rates for the Markov chain Monte Carlo Algorithm 9 for binary response for the simulation study described in the text. Each panel corresponds to a different number of group size. Within each panel, the side-by-side boxplots compare misclassification rate as a function of the sparsity threshold parameter τ. The value of m is the number of groups. The number of observations within each group is fixed at $m/10$.*

Both Figure 6.5 and Figure 6.6 show the misclassification rate as a function of sparsity threshold parameter τ under mean field variational Bayes in generalized additive mixed model with random intercept. The performance measure was misclassification rate for 30 candidate predictors being classified into one of three classes: zero effect, linear effect or non-linear effect. Each panel corresponds to a different combination of number of groups m and error standard deviation σ_ε. Within each panel, the side-by-side boxplots compare the misclassification rates under different values of sparsity threshold parameter τ. Figure 6.5 for Gaussian response shows that it has the advantages of having τ being 0.1 for lower misclassification rate. Figure 6.6 shows that having τ being 0.1 is the best choice for threshold parameter for binary response. Therefore, the default value of sparsity threshold parameter τ for

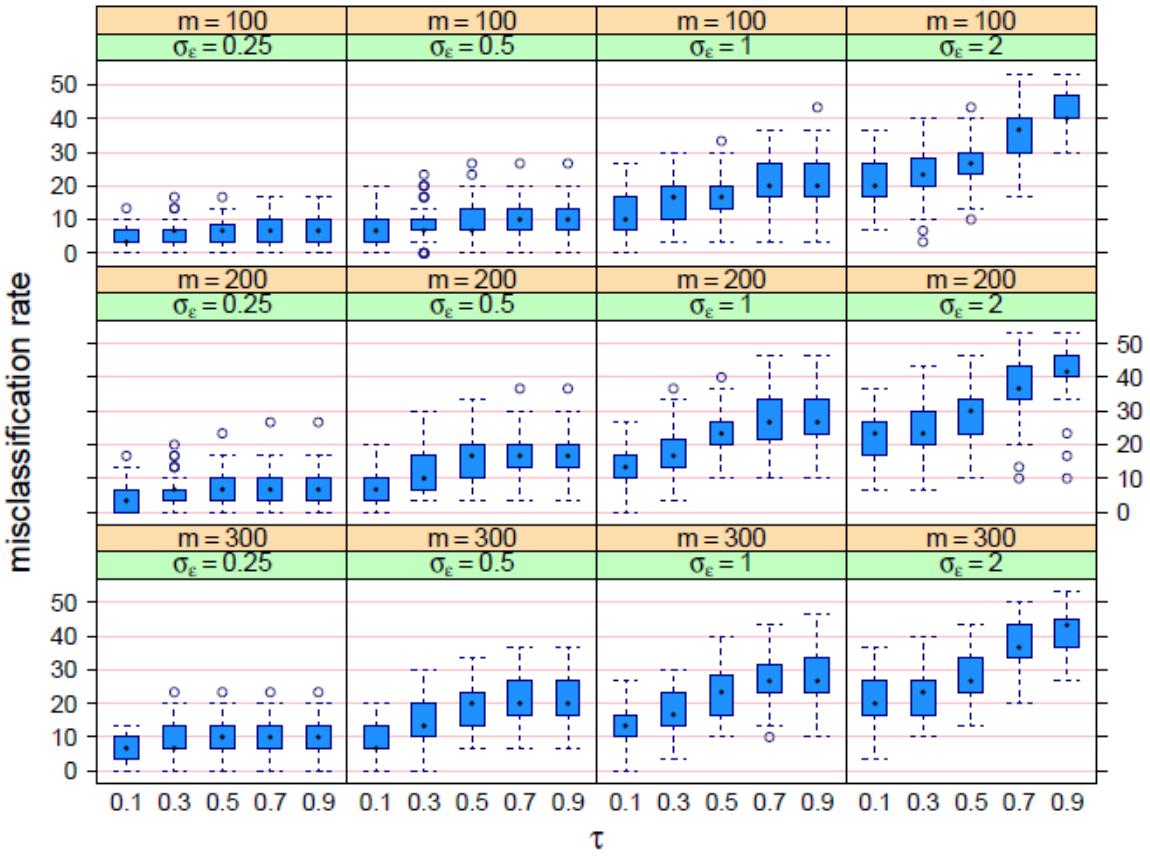


Figure 6.5: *Side-by-side boxplots of the misclassification rates for the mean field variational Bayes Algorithm 10 for Gaussian response for the simulation study described in the text. Each panel corresponds to a different combination of group size and error standard deviation. Within each panel, the side-by-side boxplots compare misclassification rate as a function of the sparsity threshold parameter τ. The value of m is the number of groups. The number of observations within each group is fixed at $m/10$.*

mean field variational Bayes in generalized additive mixed model with random intercept is 0.1.

6.8 Comparative Performance

In this section, we describe a simulation study comparing the classification performance, measured by misclassification rate and computational time in seconds, of the default version of the spikeSlabGAM approach, our MCMC approach described in Algorithm 9 and our MFVB approach described in Algorithm 10, in the context of the three-category generalized additive mixed model selection with random intercept in

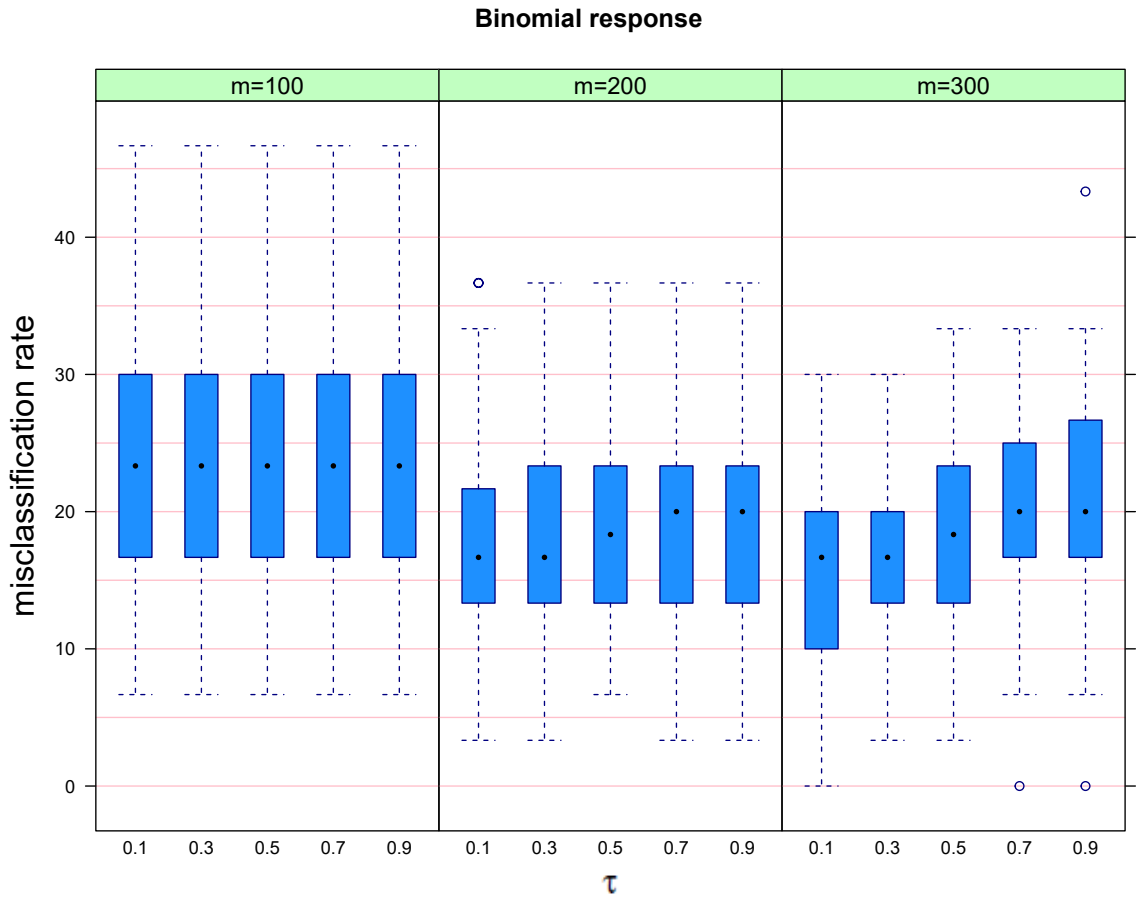


Figure 6.6: *Side-by-side boxplots of the misclassification rates for the mean field variational Bayes Algorithm 10 for binary response for the simulation study described in the text. Each panel corresponds to a different number of group size. Within each panel, the side-by-side boxplots compare misclassification rate as a function of the sparsity threshold parameter τ. The value of m is the number of groups. The number of observations within each group is fixed at $m/10$.*

model (6.25).

6.8.1 Performance Comparison by Misclassification Rate

Figure 6.7 compares the misclassification rates for Gaussian responses in generalized additive mixed model selection with random intercept, among the default version of the spikeSlabGAM approach as described in Section 5.2.1, our MCMC approach described in Section 6.5.3 (Algorithm 9) and our MFVB approach described in Section 6.6.2 (Algorithm 10). The performance measure is the misclassification rate when 30 candidate predictors are classified into one of three classes: zero effect, linear effect or non-linear effect.

The number of continuous predictors $d_\bullet$, the number of groups m, the number of observations within each group $n_i, 1 \leq i \leq m$, and the error standard deviations σ_ε are as in setting (6.85). The threshold parameter τ is set to 0.5 as described in Section 6.7.3. The generation of data sets and simulation design are as (2.55), (2.56) and (2.57) in Section 2.6.3.

In Figure 6.7 for Gaussian responses, 25 replicated data sets were generated, for each combination of number of groups m and error standard deviation σ_ε. The results show that when the number of groups m is $100,$ these three methods perform similarly across all error standard deviations. However, when the number of groups increases to $m \in \{200, 300\}$, our MCMC and MFVB approaches demonstrate better classification performance with lower misclassification rate.

Figure 6.8 shows the misclassification rates for binary responses, based on the same simulation settings used for Gaussian case in Figure 6.7. That is, 25 replicated dataset for each group. The comparison includes the default version of the spikeSlabGAM approach (Section 5.2.1), our MCMC approach (Algorithm 9) and our MFVB approach (Algorithm 10). Threshold parameter τ is set to 0.1 as described in Section 6.7.

For the binary responses simulation study, Figure 6.8 shows that our MCMC and MFVB approaches have better classification performance when the number of groups m is 100. For the number of groups increases to $m \in \{200, 300\}$, these three methods all have similar classification performance.

6.8.2 Performance Comparison by Computational Time

In this subsection, we show the computing time in seconds among the three methods, the default version of the spikeSlabGAM approach (Section 5.2.1), our MCMC approach (Algorithm 9) and our MFVB approach (Algorithm 10).

Table 6.1 reports the elapsed computing times in seconds. The laptop used for the simulation study was a MacBook Air with 16 gigabytes of memory and a 3.2 gigahertz processor. Algorithm 9 and Algorithm 10 were implemented in the R package using the Rcpp interface (Eddelbuettel and François, 2011) to the C++ language, specifically with the Rcpp package and RcppArmadillo package in R, to improve the computational speed. Additionally, the sufficient statistics described in Section 6.5.1 offered computational speed-ups for Algorithm 9 and Algorithm 10.

Table 6.1 presents the 10th, 50th and 90th percentile computing time in seconds for each approach across all settings and replications as described in Section 6.8.1.

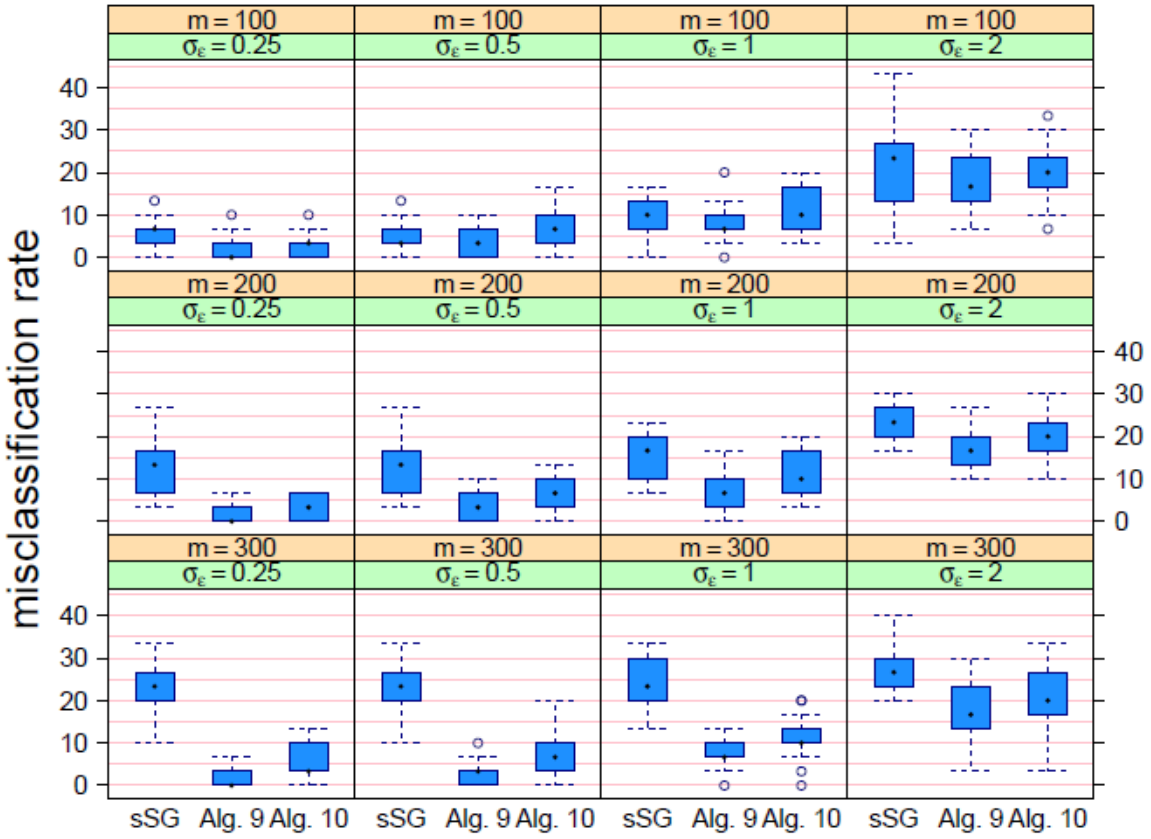


Figure 6.7: *Side-by-side boxplots of the misclassification rates for the comparative performance simulation study described in the text in the case of the response variable being Gaussian. Each panel corresponds to a different combination of group size and error standard deviation. Within each panel, the side-by-side boxplots compare misclassification rate across each of three methods:* spikeSlabGAM *with default settings (sSG), Algorithm 9 (Alg. 9) and Algorithm 10 (Alg. 10). The value of m is the number of groups. The number observations within each group is fixed at $m/10$.*

Despite exhibiting very good classification performance, spikeSlabGAM is comparatively slow and does not scale well to large problems. Our MCMC approach (Algorithm 9) took less than 42 seconds for 90 percent of the fits in the simulation study. The faster alternative MFVB (Algorithm 10), only required less than 10 seconds of computing time for most fits. Therefore, both the MCMC and MFVB approaches scale well with large data sets for the generalized additive mixed model selection problem with random intercept, with MFVB approach being the faster option.

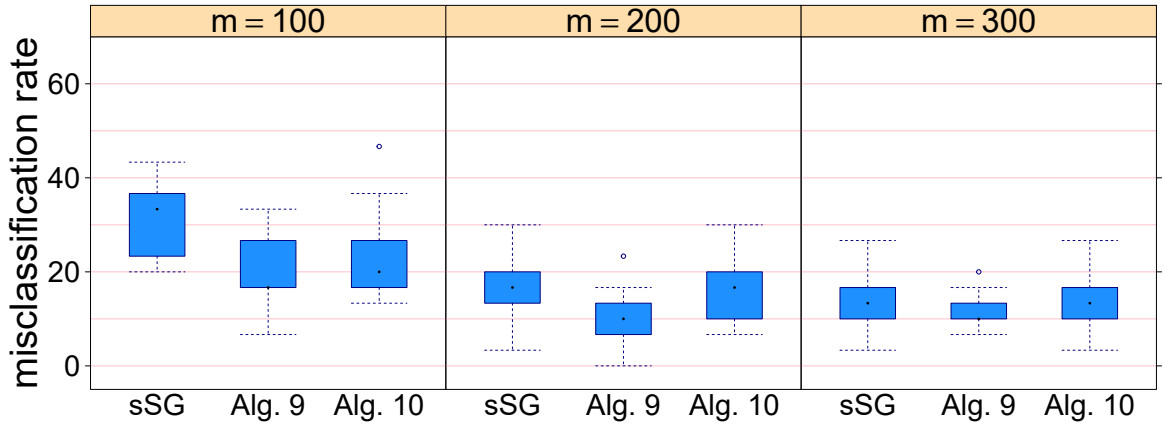


Figure 6.8: *Side-by-side boxplots of the misclassification rates for the comparative performance simulation study described in the text in the case of the response variable being binary. Each panel corresponds to a different group size. Within each panel, the side-by-side boxplots compare misclassification rate across each of three methods:* spikeSlabGAM *with default settings (sSG), Algorithm 9 (Alg. 9) and Algorithm 10 (Alg. 10). The value of m is the number of groups. The number observations within each group is fixed at $m/10$.*

Algorithm 10: *Iterative determination of the optimal parameters according to a mean field variational Bayes approximation of the posterior distributions for model (6.25).*

Data Inputs: $N = \sum_{i=1}^{m} n_i$, $\boldsymbol{y}(N \times 1)$; $\boldsymbol{X}(N \times (d_\circ + d_\bullet))$; $\dot{\boldsymbol{Z}}_k(N \times L_k), 1 \le k \le d_\bullet$.

Response Type Input: **responseType** $\in$ {Gaussian,Bernoulli} .

Sufficient Statistics Inputs: **XTy, XTX, ZTy, ZTX, ZTZ**

Hyperparameters Inputs: $\sigma_{\beta_0}, s_\beta, s_\varepsilon, s_u > 0, s_U > 0$, $1 < \rho_\beta, \rho_u < 1$.

Convergence Criterion Input: $\varepsilon_{\text{toler}}$: a small positive number such as 10^{-8}.

Initialize: $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \longleftarrow \frac{1}{2}\mathbf{1}_{d_\circ+d_\bullet}$; $\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \longleftarrow \mathbf{0}_{d_\circ+d_\bullet}$; $\mu_{\mathfrak{q}(1/a_\varepsilon)} \longleftarrow 1$; $\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)} \longleftarrow 1$;

$\mu_{\mathfrak{q}(1/a_\beta)} \longleftarrow 1$; $\mu_{\mathfrak{q}(1/\sigma_\beta^2)} \longleftarrow 1$; $\kappa_{\mathfrak{q}(\sigma_\beta^2)} \longleftarrow \frac{1}{2}(d_\circ + d_\bullet + 1)$; $\kappa_{\mathfrak{q}(a_\beta)} \longleftarrow 1$;

$\kappa_{\mathfrak{q}(\sigma_\varepsilon^2)} \longleftarrow \frac{1}{2}(N+1)$; $\kappa_{\mathfrak{q}(a_\varepsilon)} \longleftarrow 1$; $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{b}_\beta)} \longleftarrow \mathbf{1}_{d_\circ+d_\bullet}$;

$\kappa_{\mathfrak{q}(\sigma_U^2)} \longleftarrow \frac{m+1}{2}$; $\kappa_{\mathfrak{q}(a_U)} \longleftarrow 1$; $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{c})} \longleftarrow \frac{1}{2}\mathbf{1}_N$

$\mathbf{yT1}_{\text{adj}} \longleftarrow 0$; $\mathbf{XTy}_{\text{adj}} \longleftarrow \mathbf{XTy}$; $\mathbf{ZTy}_{\text{adj}} \longleftarrow \mathbf{ZTy}$;

If **responseType** is Gaussian then $\mathbf{y}_{\text{adj}} \longleftarrow \boldsymbol{y}$

If **responseType** is Bernoulli then $\mathbf{y}_{\text{adj}} \longleftarrow \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{c})}$

For $k = 1, \ldots, d_\bullet$:

$\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)} \longleftarrow \mathbf{0}_{L_k}$; $\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)} \longleftarrow \mathbf{1}_{L_k}$; $\mu_{\mathfrak{q}(\gamma_{uk})} \longleftarrow \frac{1}{2}$;
$\mu_{\mathfrak{q}(1/a_{uk})} \longleftarrow 1$; $\mu_{\mathfrak{q}(1/\sigma_{uk}^2)} \longleftarrow 1$; $\mu_{\mathfrak{q}(b_{uk})} \longleftarrow 1$;
$\kappa_{\mathfrak{q}(\sigma_{uk}^2)} \longleftarrow \frac{1}{2}(L_k+1)$; $\kappa_{\mathfrak{q}(a_{uk})} \longleftarrow 1$; $\boldsymbol{\omega}_{\dot{\boldsymbol{Z}}_k} \longleftarrow \text{diagonal}\left(\mathbf{ZTZ}^{\langle k,k\rangle}\right)$

Cycle:

$\omega_{12} \longleftarrow \mathbf{yT1}_{adj} - \sum_{i=1}^{m} n_i U_i$

$\sigma^2_{\mathfrak{q}(\beta_0)} \longleftarrow 1/\left\{N\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)} + (1/\sigma^2_{\beta_0})\right\}$; $\mu_{\mathfrak{q}(\beta_0)} \longleftarrow \sigma^2_{\mathfrak{q}(\beta_0)}\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\omega_{12}$

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

$\sigma^2_{\mathfrak{q}(U_i)} \longleftarrow 1\Big/\left\{n_i\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)} + \mu_{\mathfrak{q}\left(1/\sigma_U^2\right)}\right\}$

continued on a subsequent page...

Algorithm 10 continued: *This is a continuation of the description of this algorithm that commences on a preceding page.*

$$\mu_{\mathfrak{q}(U_i)} \longleftarrow \sigma^2_{\mathfrak{q}(U_i)}\mu_{\mathfrak{q}(1/\sigma^2_\varepsilon)}\mathbf{1}_{n_i}^T\left(\left(\mathbf{y}_{\text{adj}}\right)_i - \boldsymbol{X}_i\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right) - \sum_{k=1}^{d_\bullet}\boldsymbol{Z}_{ik}\left(\mu_{\mathfrak{q}(\gamma_{uk})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right)\right)$$

$$\kappa_{\mathfrak{q}\left(1/\sigma^2_U\right)} \longleftarrow \left\|\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{U})}\right\|^2 + \sum_{i=1}^{m}\mu_{\mathfrak{q}\left(\sigma^2_{U_i}\right)}; \quad \mu_{\mathfrak{q}\left(1/\sigma^2_U\right)} \longleftarrow \kappa_{\mathfrak{q}\left(\sigma^2_U\right)}\big/\lambda_{\mathfrak{q}\left(\sigma^2_U\right)}$$

$$\lambda_{\mathfrak{q}((a_U)} \longleftarrow \mu_{\mathfrak{q}\left(1/\sigma^2_U\right)} + 1/s^2_U; \quad \mu_{\mathfrak{q}(1/a_U)} \longleftarrow \kappa_{\mathfrak{q}(a_U)}\big/\lambda_{\mathfrak{q}(a_U)}$$

$$\boldsymbol{\Omega}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \longleftarrow \text{diag}\left\{\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \left(\mathbf{1} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}\right)\right\} + \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}\boldsymbol{\mu}^T_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}$$

$$\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \longleftarrow \left\{\mu_{\mathfrak{q}(1/\sigma^2_\varepsilon)}\boldsymbol{\Omega}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \left(\boldsymbol{X}^T\boldsymbol{X}\right) + \mu_{\mathfrak{q}(1/\sigma^2_\beta)}\text{diag}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{b}_\beta)}\right)\right\}^{-1}$$

$$\boldsymbol{\omega}_{13} \longleftarrow \mathbf{XTy}_{\text{adj}} - \boldsymbol{X}^T\begin{bmatrix}\mu_{\mathfrak{q}(U_1)}\mathbf{1}_{n_1}\\ \vdots \\ \mu_{\mathfrak{q}(U_m)}\mathbf{1}_{n_m}\end{bmatrix} - \sum_{k=1}^{d_\bullet}\mathbf{ZTX}^{\langle k\rangle T}\left(\mu_{\mathfrak{q}(\gamma_{uk})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right)$$

$$\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \longleftarrow \mu_{\mathfrak{q}(1/\sigma^2_\varepsilon)}\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\omega}_{13}\right)$$

$$\boldsymbol{\omega}_{14} \longleftarrow \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} + \text{diagonal}\left(\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right) \; ; \; \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{b}_\beta)} \longleftarrow \left(\mu_{\mathfrak{q}(1/\sigma^2_\beta)}\boldsymbol{\omega}_{14}\right)^{-1/2}$$

$$\lambda_{\mathfrak{q}(\sigma^2_\beta)} \longleftarrow \mu_{\mathfrak{q}(1/a_\beta)} + \frac{1}{2}\boldsymbol{\mu}^T_{\mathfrak{q}(\boldsymbol{b}_\beta)}\boldsymbol{\omega}_{14} \quad ; \quad \mu_{\mathfrak{q}(1/\sigma^2_\beta)} \longleftarrow \kappa_{\mathfrak{q}(\sigma^2_\beta)}\Big/\lambda_{\mathfrak{q}(\sigma^2_\beta)}$$

$$\lambda_{\mathfrak{q}(a_\beta)} \longleftarrow \mu_{\mathfrak{q}(1/\sigma^2_\beta)} + s_\beta^{-2} \quad ; \quad \mu_{\mathfrak{q}(1/a_\beta)} \longleftarrow \kappa_{\mathfrak{q}(a_\beta)}\Big/\lambda_{\mathfrak{q}(a_\beta)}$$

$$\mathbf{XTU}_{\mathfrak{q}} \longleftarrow \boldsymbol{X}^T\begin{bmatrix}\mu_{\mathfrak{q}(U_1)}\mathbf{1}_{n_1}\\ \vdots \\ \mu_{\mathfrak{q}(U_m)}\mathbf{1}_{n_m}\end{bmatrix}; \quad \mathbf{ZTU}_{\mathfrak{q}} \longleftarrow \boldsymbol{Z}^T\begin{bmatrix}\mu_{\mathfrak{q}(U_1)}\mathbf{1}_{n_1}\\ \vdots \\ \mu_{\mathfrak{q}(U_m)}\mathbf{1}_{n_m}\end{bmatrix};$$

For $k = 1, \ldots, d_\bullet$: $\quad \boldsymbol{\mu}_{\mathfrak{q}(u_k)} \longleftarrow \mu_{\mathfrak{q}(\gamma_{uk})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}$

For $k = 1, \ldots, d_\circ + d_\bullet$:

$$\omega_{15} \longleftarrow \boldsymbol{e}_k^T\mathbf{XTy}_{\text{adj}} - \boldsymbol{e}_k^T\mathbf{XTU}_{\mathfrak{q}} - \boldsymbol{e}_k^T\mathbf{X}\sum_{k=1}^{d_\bullet}\left(\mathbf{ZTX}^{\langle k'\rangle}\boldsymbol{e}_k\right)^T\boldsymbol{\mu}_{\mathfrak{q}(u_{k'})}$$

$$\omega_{15} \longleftarrow \mu_{\mathfrak{q}(\widetilde{\beta}_k)}\omega_{15} - \left(\mathbf{XTX}\boldsymbol{e}_k\right)^T_{-k} \times\left[\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}\right)_{-k} \odot \left\{\left(\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\boldsymbol{e}_k\right)_{-k} + \mu_{\mathfrak{q}(\widetilde{\beta}_k)}\left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right)_{-k}\right\}\right]$$

continued on a subsequent page...

Algorithm 10 continued: *This is a continuation of the description of this algorithm that commences on a preceding page.*

$$\mu_{\mathfrak{q}(\gamma_{\beta j})} \longleftarrow \text{expit}\left(\text{logit}(\rho_\beta)) - \frac{1}{2}\mu_{\mathfrak{q}(1/\sigma^2_{\tilde{\varepsilon}})}\left\{\left(\mu^2_{\mathfrak{q}(\widetilde{\beta}_k)} + \sigma^2_{\mathfrak{q}(\widetilde{\beta}_k)}\right)\boldsymbol{e}_k^T\mathbf{XTX}\boldsymbol{e}_k \right.\right.$$

$$\left.\left. -2\omega_{15}\right\}\right)$$

For $k = 1, \ldots, d_\bullet$: $\quad \boldsymbol{\mu}_{\mathfrak{q}(u_k)} \longleftarrow \mu_{\mathfrak{q}(\gamma_{uk})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}$

For $k = 1, \ldots, d_\bullet$:

$$\boldsymbol{\omega}_{16} \longleftarrow \mathbf{ZTy}^{\langle k\rangle}_{\text{adj}} - \mathbf{ZTU}^{\langle k\rangle}_{\mathfrak{q}} - \mathbf{ZTX}^{\langle k\rangle}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right) - \sum_{k'\neq k}^{d_\bullet}\mathbf{ZTZ}^{\langle k,k'\rangle}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{k'})}$$

$$\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)} \longleftarrow \mathbf{1}_{L_k}\Big/\left\{\mu_{\mathfrak{q}(1/\sigma^2_{\tilde{\varepsilon}})}\mu_{\mathfrak{q}(\gamma_{uk})}\boldsymbol{\omega}_{\dot{\boldsymbol{Z}}_k} + \mu_{\mathfrak{q}(1/\sigma^2_{uk})}\mu_{\mathfrak{q}(b_{uk})}\mathbf{1}_{L_k}\right\}$$

$$\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)} \longleftarrow \mu_{\mathfrak{q}(1/\sigma^2_{\tilde{\varepsilon}})}\mu_{\mathfrak{q}(\gamma_{uk})}\left(\boldsymbol{\omega}_{16} \odot \boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right)$$

For $k = 1, \ldots, d_\bullet$:

$$\omega_{17} \longleftarrow \left\|\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right\|^2 + \mathbf{1}^T_{L_k}\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}; \quad \mu_{\mathfrak{q}(b_uk)} \longleftarrow \left(\mu_{\mathfrak{q}(1/\sigma^2_{uj})}\omega_{17}\right)^{-1/2}$$

$$\lambda_{\mathfrak{q}(\sigma^2_{uk})} \longleftarrow \mu_{\mathfrak{q}(1/a_{uk})} + \frac{1}{2}\mu_{\mathfrak{q}(b_{uk})}\omega_{17} \quad ; \quad \mu_{\mathfrak{q}(1/\sigma^2_{uk})} \longleftarrow \kappa_{\mathfrak{q}(\sigma^2_{uk})}\Big/\lambda_{\mathfrak{q}(\sigma^2_{uk})}$$

$$\lambda_{\mathfrak{q}(a_{uk})} \longleftarrow \mu_{\mathfrak{q}(1/\sigma^2_{uk})} + \left(1/s^2_u\right) \quad ; \quad \mu_{\mathfrak{q}(1/a_{uk})} \longleftarrow \kappa_{\mathfrak{q}(a_{uk})}\Big/\lambda_{\mathfrak{q}(a_{uk})}$$

For $k = 1, \ldots, d_\bullet$: $\quad \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_k)} \longleftarrow \mu_{\mathfrak{q}(\gamma_{uk})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}$

For $k = 1, \ldots, d_\bullet$:

$$\boldsymbol{\omega}_{18} \longleftarrow \mathbf{ZTy}^{\langle k\rangle}_{\text{adj}} - \mathbf{ZTU}^{\langle k\rangle}_{\mathfrak{q}} - \mathbf{ZTX}^{\langle k\rangle}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right) - \sum_{k'\neq k}^{d_\bullet}\mathbf{ZTZ}^{\langle k,k'\rangle}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{k'})}$$

$$\omega_{19} \longleftarrow \boldsymbol{\omega}^T_{\dot{\boldsymbol{Z}}_k}\left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)} + \boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right) - 2\boldsymbol{\mu}^T_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\boldsymbol{\omega}_{18}$$

$$\mu_{\mathfrak{q}(\gamma_{uk})} \longleftarrow \text{expit}\Big(\text{logit}(\rho_u) - \frac{1}{2}\mu_{\mathfrak{q}(1/\sigma^2_{\tilde{\varepsilon}})}\omega_{19}\Big)$$

$$\boldsymbol{\omega}_{20} \longleftarrow \mathbf{1}_N\mu_{\mathfrak{q}(\beta_0)} + \begin{bmatrix} \mu_{\mathfrak{q}(U_1)}\mathbf{1}_{n_1} \\ \vdots \\ \mu_{\mathfrak{q}(U_m)}\mathbf{1}_{n_m} \end{bmatrix} + \boldsymbol{X}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right) + \sum_{k=1}^{d_\bullet}\dot{\boldsymbol{Z}}_k\left(\mu_{\mathfrak{q}(\gamma_{uk})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right)$$

continued on a subsequent page...

Algorithm 10 continued: *This is a continuation of the description of this algorithm that commences on a preceding page.*

If **responseType** is Gaussian then

$$\boldsymbol{\Omega}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \longleftarrow \text{diag}\left\{\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \left(\mathbf{1} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}\right)\right\} + \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}\boldsymbol{\mu}^T_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}$$

$$\begin{aligned}\lambda_{\mathfrak{q}(\sigma^2_\varepsilon)} \longleftarrow\ & \mu_{\mathfrak{q}(1/a_\varepsilon)} + \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{\omega}_{20}\right\|^2 + \frac{1}{2}N\sigma^2_{\mathfrak{q}(\beta_0)} + \frac{1}{2}\sum_{i=1}^{m}\left(n_i\sigma^2_{\mathfrak{q}(U_i)}\right) \\ & + \frac{1}{2}\text{tr}\left[\boldsymbol{X}^T\boldsymbol{X}\left\{\boldsymbol{\Omega}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \left(\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} + \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\boldsymbol{\mu}^T_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right)\right\}\right] \\ & - \frac{1}{2}\textbf{tr}\left\{\boldsymbol{X}^T\boldsymbol{X}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right)\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right)^T\right\} \\ & + \frac{1}{2}\sum_{k=1}^{d_\bullet}\boldsymbol{\omega}^T_{\bullet_k}\boldsymbol{Z}_k\left(\mu_{\mathfrak{q}(\gamma_{uk})}\left[\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)} + \left\{1 - \mu_{\mathfrak{q}(\gamma_{uk})}\right\}\left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right)\right]\right)\end{aligned}$$

$$\mu_{\mathfrak{q}(1/\sigma^2_\varepsilon)} \longleftarrow \kappa_{\mathfrak{q}(\sigma^2_\varepsilon)}\Big/\lambda_{\mathfrak{q}(\sigma^2_\varepsilon)};\ \ \lambda_{\mathfrak{q}(a_\varepsilon)} \longleftarrow \mu_{\mathfrak{q}(1/\sigma^2_\varepsilon)} + \left(1\Big/s^2_\varepsilon\right);$$

$$\mu_{\mathfrak{q}(1/a_\varepsilon)} \longleftarrow \kappa_{\mathfrak{q}(a_\varepsilon)}\Big/\lambda_{\mathfrak{q}(a_\varepsilon)}$$

If **responseType** is Bernoulli then

$$\mu_{\mathfrak{q}(1/\sigma^2_\varepsilon)} \longleftarrow 1\ ;\ \ \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{c})} \longleftarrow \boldsymbol{\omega}_{20} + \left(2\boldsymbol{y} - \mathbf{1}_N\right) \odot \zeta'\left(\left(2\boldsymbol{y} - \mathbf{1}_N\right) \odot \boldsymbol{\omega}_{20}\right)$$

$$\mathbf{yT1}_{\text{adj}} \longleftarrow \boldsymbol{\mu}^T_{\mathfrak{q}(\boldsymbol{c})}\mathbf{1}_N\ ;\ \mathbf{XTy}_{\text{adj}} \longleftarrow \boldsymbol{X}^T\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{c})}\ ;\ \mathbf{ZTy}_{\text{adj}} \longleftarrow \boldsymbol{Z}^T\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{c})};\ \ \mathbf{y}_{\text{adj}} \longleftarrow \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{c})};$$

until the relative change in the $\log(\underline{p}(\boldsymbol{y};\mathfrak{q}))$ is below $\varepsilon_{\text{toler}}$.

Outputs: All $\mathfrak{q}$-density parameters.

	spikeSlabGAM	Algorithm 9	Algorithm 10
10th percentile	218.1	5.66	1.12
50th percentile	1384.7	15.90	3.35
90th percentile	5261.8	41.42	9.86

Table 6.1: *10th, 50th and 90th percentiles for the number of seconds required for each generalized additive mixed model selection approach with random intercept across all settings and replications for the comparative performance simulation study.*

6.A Derivations of Full Conditional Distributions

The const in Appendix 6.A is a constant that does not depend on any parameters.

6.A.1 Full Conditional Distribution of β_0

First note that

$$\mathfrak{p}(\beta_0|\text{rest}) \propto \prod_{i=1}^{m} \left\{ \mathfrak{p}\left(\boldsymbol{y}_i | \beta_0, U_i, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u, \sigma_\varepsilon^2 \right) \right\} \mathfrak{p}\left(\beta_0\right).$$

We have

$$\begin{aligned}
&\prod_{i=1}^{m} \left\{ \mathfrak{p}\left(\boldsymbol{y}_i | \beta_0, U_i, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u, \sigma_\varepsilon^2 \right) \right\} \\
&\quad \propto \exp\left\{ -\frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^{m} \left\| \boldsymbol{y}_i - \mathbf{1}_{ni}\beta_0 - \mathbf{1}_{ni}U_i - \boldsymbol{X}_i(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{k=1}^{d_\bullet} \boldsymbol{Z}_{ik}(\gamma_{uk}\widetilde{\boldsymbol{u}}_k) \right\|^2 \right\} \\
&\quad \propto \exp\left\{ -\frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^{m} \left\| \boldsymbol{r}_{i\beta_0} - \mathbf{1}_{ni}\beta_0 \right\|^2 \right\} \\
&\quad \propto \exp\left\{ -\frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^{m} \left(-2\beta_0 \mathbf{1}_{ni}^T \boldsymbol{r}_{i\beta_0} + n_i\beta_0^2 \right) \right\}
\end{aligned}$$

if we define

$$\boldsymbol{r}_{i\beta_0} \equiv \boldsymbol{y}_i - \mathbf{1}_{ni}U_i - \boldsymbol{X}_i(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{k=1}^{d_\bullet} \boldsymbol{Z}_{ik}(\gamma_{uk}\widetilde{\boldsymbol{u}}_k),$$

and note that

$$\begin{aligned}
\sum_{i=1}^{m} \left(\mathbf{1}_{ni}^T \boldsymbol{r}_{i\beta_0} \right) &= \sum_{i=1}^{m} \left(\mathbf{1}_{ni}^T \left(\boldsymbol{y}_i - \mathbf{1}_{ni}U_i - \boldsymbol{X}_i(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{k=1}^{d_\bullet} \boldsymbol{Z}_{ik}(\gamma_{uk}\widetilde{\boldsymbol{u}}_k) \right) \right) \\
&= \boldsymbol{y}^T\mathbf{1} - \sum_{i=1}^{m} \left(n_i U_i \right) - \left(\boldsymbol{X}^T\mathbf{1} \right)^T \left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}} \right) - \sum_{k=1}^{d_\bullet} \left(\dot{\boldsymbol{Z}}_k^T\mathbf{1} \right)^T (\gamma_{uk}\widetilde{\boldsymbol{u}}_k)
\end{aligned}$$

where $\dot{\boldsymbol{Z}}_k$ defined in (6.12). Then

$$\prod_{i=1}^{m} \mathfrak{p}\left(\boldsymbol{y}_i \middle| \beta_0, U_i, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u, \sigma_\varepsilon^2\right)$$

$$\propto \exp\left\{ \begin{bmatrix} \beta_0 \\ \beta_0^2 \end{bmatrix}^T \begin{bmatrix} \dfrac{\boldsymbol{y}^T\mathbf{1} - \sum_{i=1}^{m}(n_i U_i) - \left(\boldsymbol{X}^T\mathbf{1}\right)^T\left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}\right) - \sum_{k=1}^{d_\bullet}\left(\dot{\boldsymbol{Z}}_k^T\mathbf{1}\right)^T(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)}{\sigma_\varepsilon^2} \\ -\dfrac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{m} n_i \end{bmatrix} \right\}.$$

Since $\mathfrak{p}(\beta_0)$ is the $N(0, \sigma_{\beta_0}^2)$ density function, we have

$$\mathfrak{p}(\beta_0) \propto \exp\left\{ \begin{bmatrix} \beta_0 \\ \beta_0^2 \end{bmatrix}^T \begin{bmatrix} 0 \\ -1/2 \end{bmatrix} (1/\sigma_{\beta_0}^2) \right\}.$$

Therefore

$$\mathfrak{p}(\beta_0 \big| \text{rest})$$

$$\propto \exp\left\{ \begin{bmatrix} \beta_0 \\ \beta_0^2 \end{bmatrix}^T \begin{bmatrix} \dfrac{\boldsymbol{y}^T\mathbf{1} - \sum_{i=1}^{m}(n_i U_i) - \left(\boldsymbol{X}^T\mathbf{1}\right)^T\left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}\right) - \sum_{k=1}^{d_\bullet}\left(\dot{\boldsymbol{Z}}_k^T\mathbf{1}\right)^T(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)}{\sigma_\varepsilon^2} \\ -\dfrac{1}{2}\left(\dfrac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{m} n_i + \dfrac{1}{\sigma_{\beta_0}^2}\right) \end{bmatrix} \right\}. \tag{6.86}$$

Therefore $\mathfrak{p}(\beta_0 \big| \text{rest})$ has a Normal distribution with natural parameter vector

$$\begin{bmatrix} \left(\boldsymbol{y}^T\mathbf{1} - \sum_{i=1}^{m}(n_i U_i) - \left(\boldsymbol{X}^T\mathbf{1}\right)^T\left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}\right) - \sum_{k=1}^{d_\bullet}\left(\dot{\boldsymbol{Z}}_k^T\mathbf{1}\right)^T(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)\right) \Big/ \sigma_\varepsilon^2 \\ -\dfrac{1}{2}\left(\dfrac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{m} n_i + \dfrac{1}{\sigma_{\beta_0}^2}\right) \end{bmatrix}.$$

The variance is

$$1 \Big/ \left(\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{m} n_i + \frac{1}{\sigma_{\beta_0}^2}\right).$$

The mean is

$$\frac{\frac{1}{\sigma_\varepsilon^2}\left(\boldsymbol{y}^T\mathbf{1}-\sum_{i=1}^{m}\left(n_iU_i\right)-\left(\boldsymbol{X}^T\mathbf{1}\right)^T\left(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}}\right)-\sum_{k=1}^{d_\bullet}\left(\dot{\boldsymbol{Z}}_k^T\mathbf{1}\right)^T\left(\gamma_{uk}\widetilde{\boldsymbol{u}}_k\right)\right)}{\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{m}n_i+\frac{1}{\sigma_{\beta_0}^2}}.$$

Therefore

$$\beta_0\big|\text{rest}\sim N\left(\frac{\frac{1}{\sigma_\varepsilon^2}\left(\boldsymbol{y}^T\mathbf{1}-\sum_{i=1}^{m}\left(n_iU_i\right)-\left(\boldsymbol{X}^T\mathbf{1}\right)^T\left(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}}\right)-\sum_{k=1}^{d_\bullet}\left(\dot{\boldsymbol{Z}}_k^T\mathbf{1}\right)^T\left(\gamma_{uk}\widetilde{\boldsymbol{u}}_k\right)\right)}{\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{m}n_i+\frac{1}{\sigma_{\beta_0}^2}},\ \frac{1}{\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{m}n_i+\frac{1}{\sigma_{\beta_0}^2}}\right).$$

We define

$$\begin{aligned}\omega_1&\equiv\boldsymbol{y}^T\mathbf{1}-\sum_{i=1}^{m}\left(n_iU_i\right)-\left(\boldsymbol{X}^T\mathbf{1}\right)^T\left(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}}\right)-\sum_{k=1}^{d_\bullet}\left(\dot{\boldsymbol{Z}}_k^T\mathbf{1}\right)^T\left(\gamma_{uk}\widetilde{\boldsymbol{u}}_k\right)\\&=\left(\boldsymbol{y}^T\mathbf{1}\right)-\sum_{i=1}^{m}\left(n_iU_i\right).\end{aligned}$$

The last step above is using feature of $\boldsymbol{X}$ matrix in (6.10) to be $\mathbf{1}_N^T\boldsymbol{X}=\mathbf{0}, N=\sum_{i=1}^{m}n_i$ and $\boldsymbol{Z}$ matrix in (6.13) to be $\dot{\boldsymbol{Z}}_k^T\mathbf{1}=\mathbf{0}_{L_k}, 1\le k\le d_\bullet$. We also define

$$\omega_2\equiv\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{m}n_i+\frac{1}{\sigma_{\beta_0}^2}.$$

Then $\beta_0\big|\text{rest}\sim N\left(\frac{\omega_1}{(\sigma_\varepsilon^2)\,\omega_2},\frac{1}{\omega_2}\right)$ is as given in Algorithm 9 in Section 6.5.3.

6.A.2 Full Conditional Distribution of U_i

First note that, for $1 \leq i \leq m$,

$$\mathfrak{p}(U_i \big| \text{rest}) \propto \mathfrak{p}(\boldsymbol{y}_i \big| \beta_0, U_i, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u, \sigma_\varepsilon^2) \mathfrak{p}(U_i \big| \sigma_U^2).$$

We have

$$\begin{aligned}
&\mathfrak{p}(\boldsymbol{y}_i \big| \beta_0, U_i, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u, \sigma_\varepsilon^2) \\
&\quad \propto \exp\left\{ -\frac{1}{2\sigma_\varepsilon^2} \left\| \boldsymbol{y}_i - \mathbf{1}_{ni}\beta_0 - \mathbf{1}_{ni} U_i - \boldsymbol{X}_i(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{k=1}^{d_\bullet} \boldsymbol{Z}_{ik}(\gamma_{uk} \widetilde{\boldsymbol{u}}_k) \right\|^2 \right\} \\
&\quad \propto \exp\left\{ -\frac{1}{2\sigma_\varepsilon^2} \left\| \boldsymbol{r}_{U_i} - \mathbf{1}_{ni} U_i \right\|^2 \right\} \propto \exp\left\{ -\frac{1}{2\sigma_\varepsilon^2} \left(-2U_i \mathbf{1}_{ni}^T \boldsymbol{r}_{U_i} + n_i U_i^2 \right) \right\}.
\end{aligned}$$

If we define

$$\boldsymbol{r}_{U_i} \equiv \boldsymbol{y}_i - \mathbf{1}_{ni}\beta_0 - \boldsymbol{X}_i(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{k=1}^{d_\bullet} \boldsymbol{Z}_{ik}(\gamma_{uk} \widetilde{\boldsymbol{u}}_k)$$

then

$$\begin{aligned}
&\mathfrak{p}\left(\boldsymbol{y}_i \big| \beta_0, U_i, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u, \sigma_\varepsilon^2 \right) \\
&\quad \propto \exp\left\{ \begin{bmatrix} U_i \\ U_i^2 \end{bmatrix}^T \begin{bmatrix} \mathbf{1}_{ni}^T \left(\boldsymbol{y}_i - \mathbf{1}_{ni}\beta_0 - \boldsymbol{X}_i(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{k=1}^{d_\bullet} \boldsymbol{Z}_{ik}(\gamma_{uk} \widetilde{\boldsymbol{u}}_k) \right) \\ -n_i/2 \end{bmatrix} (1/\sigma_\varepsilon^2) \right\}.
\end{aligned}$$

From $U_i \big| \sigma_U^2 \sim N\left(0, \sigma_U^2\right), 1 \leq i \leq m$, we have

$$\mathfrak{p}(U_i \big| \sigma_U^2) \propto \exp\left\{ \begin{bmatrix} U_i \\ U_i^2 \end{bmatrix}^T \begin{bmatrix} 0 \\ -\frac{1}{2} \end{bmatrix} \left(1/\sigma_U^2\right) \right\}.$$

Then

$$\begin{aligned}\mathfrak{p}(U_i\big|\text{rest}) &\propto \mathfrak{p}(\boldsymbol{y}_i\big|\beta_0, U_i, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u, \sigma_\varepsilon^2)\mathfrak{p}(U_i\big|\sigma_U^2) \\ &\propto \exp\left\{ \begin{bmatrix} U_i \\ U_i^2 \end{bmatrix}^T \begin{bmatrix} \frac{1}{\sigma_\varepsilon^2}\mathbf{1}_{ni}^T\left(\boldsymbol{y}_i - \mathbf{1}_{ni}\beta_0 - \boldsymbol{X}_i(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{k=1}^{d_\bullet}\boldsymbol{Z}_{ik}(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)\right) \\ -\frac{1}{2}\left(\frac{n_i}{\sigma_\varepsilon^2} + \frac{1}{\sigma_U^2}\right) \end{bmatrix} \right\}.\end{aligned} \tag{6.87}$$

Therefore $\mathfrak{p}(U_i\big|\text{rest})$ has a Normal distribution with natural parameter vector

$$\begin{bmatrix} \mathbf{1}_{ni}^T\left(\boldsymbol{y}_i - \mathbf{1}_{ni}\beta_0 - \boldsymbol{X}_i(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{k=1}^{d_\bullet}\boldsymbol{Z}_{ik}(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)\right) \Big/ \sigma_\varepsilon^2 \\ -\frac{1}{2}\left(\frac{n_i}{\sigma_\varepsilon^2} + \frac{1}{\sigma_U^2}\right) \end{bmatrix}.$$

The variance is

$$\frac{1}{\frac{n_i}{\sigma_\varepsilon^2} + \frac{1}{\sigma_U^2}}.$$

The mean is

$$\frac{\frac{1}{\sigma_\varepsilon^2}\mathbf{1}_{ni}^T\left(\boldsymbol{y}_i - \mathbf{1}_{ni}\beta_0 - \boldsymbol{X}_i(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{k=1}^{d_\bullet}\boldsymbol{Z}_{ik}(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)\right)}{\frac{n_i}{\sigma_\varepsilon^2} + \frac{1}{\sigma_U^2}}.$$

In summary, for $1 \le i \le m$,

$$\begin{aligned}U_i\big|\text{rest} \sim N\Bigg(&\left(\frac{n_i}{\sigma_\varepsilon^2} + \frac{1}{\sigma_U^2}\right)^{-1}\frac{1}{\sigma_\varepsilon^2}\mathbf{1}_{ni}^T\Big\{\boldsymbol{y}_i - \mathbf{1}_{ni}\beta_0 - \boldsymbol{X}_i(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) \\ &-\sum_{k=1}^{d_\bullet}\boldsymbol{Z}_{ik}(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)\Big\}, \left(\frac{n_i}{\sigma_\varepsilon^2} + \frac{1}{\sigma_U^2}\right)^{-1}\Bigg).\end{aligned}$$

If we define

$$\omega_3 \equiv \mathbf{1}_{ni}^T \left(\boldsymbol{y}_i - \mathbf{1}_{ni}\beta_0 - \boldsymbol{X}_i(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{k=1}^{d_\bullet} \boldsymbol{Z}_{ik}(\gamma_{uk}\widetilde{\boldsymbol{u}}_k) \right)$$

$$= \left(\mathbf{1}_{ni}^T \boldsymbol{y}_i\right) - n_i\beta_0 - \left(\boldsymbol{X}_i^T \mathbf{1}_{ni}\right)^T \left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}\right) - \sum_{k=1}^{d_\bullet} \left(\boldsymbol{Z}_{ik}^T \mathbf{1}_{ni}\right)^T (\gamma_{uk}\widetilde{\boldsymbol{u}}_k)$$

and

$$\omega_4 \equiv \frac{n_i}{\sigma_\varepsilon^2} + \frac{1}{\sigma_U^2},$$

then $U_i \big| \text{rest} \sim N\left(\frac{\omega_3}{\sigma_\varepsilon^2 \omega_4}, \frac{1}{\omega_4}\right), \ \ 1 \le i \le m, \ \ $ as in Algorithm 9 in Section 6.5.3.

6.A.3 Full Conditional Distribution of σ_U^2

First note that

$$\mathfrak{p}(\sigma_U^2 \big| \text{rest}) \propto \left\{ \prod_{i=1}^m \mathfrak{p}(U_i \big| \sigma_U^2) \right\} \mathfrak{p}(\sigma_U^2 \big| a_U).$$

From

$$U_i \big| a_U \overset{\text{ind.}}{\sim} N\left(0, \sigma_U^2\right), 1 \le i \le m,$$

we have

$$\prod_{i=1}^m \mathfrak{p}(U_i \big| \sigma_U^2) \propto \left(\sigma_U^2\right)^{-\frac{m}{2}} \exp\left(-\frac{\left\|\boldsymbol{U}\right\|^2}{2\sigma_U^2} \right).$$

From

$$\sigma_U^2 \big| a_U \sim \text{Inverse-Gamma}\left(\frac{1}{2}, 1/a_U\right),$$

we have

$$\mathfrak{p}(\sigma_U^2 \big| a_U) \sim \left(\sigma_U^2\right)^{-\frac{1}{2}-1} \exp\left(-\frac{1}{a_U \sigma_U^2} \right).$$

Then we obtain

$$\mathfrak{p}(\sigma_U^2 \big| \text{rest}) \propto \left(\sigma_U^2\right)^{-\frac{m+1}{2}-1} \exp\left(-\frac{1}{\sigma_U^2} \left(\frac{\left\|\boldsymbol{U}\right\|^2}{2} + \frac{1}{a_U} \right) \right).$$

Therefore

$$\sigma_U^2 \big| \text{rest} \sim \text{Inverse-Gamma}\left(\frac{m+1}{2}, \frac{\left\|\boldsymbol{U}\right\|^2}{2} + \frac{1}{a_U}\right).$$

6.A.4 Full Conditional Distribution of a_U

First note that

$$\mathfrak{p}(a_U \big| \text{rest}) \propto \mathfrak{p}(\sigma_U^2 \big| a_U)\mathfrak{p}(a_U).$$

In our full model, we have

$$\sigma_U^2 \big| a_U \sim \text{Inverse-Gamma}\left(\frac{1}{2}, 1/a_U\right) \text{ and } a_U \sim \text{Inverse-Gamma}\left(\frac{1}{2}, 1/s_U^2\right).$$

Then

$$\begin{aligned}\mathfrak{p}(a_U \big| \text{rest}) &\propto (a_U)^{-\frac{1}{2}} \exp\left(-\tfrac{1}{a_U\sigma_U^2}\right) (a_U)^{-\frac{1}{2}-1} \exp\left(-\frac{1}{a_U s_U^2}\right) \\ &\propto (a_U)^{-1-1} \exp\left(-\frac{1}{a_U}\left(\frac{1}{\sigma_U^2} + \frac{1}{s_U^2}\right)\right).\end{aligned}$$

Therefore

$$a_U \big| \text{rest} \sim \text{Inverse-Gamma}\left(1, \frac{1}{\sigma_U^2} + \frac{1}{s_U^2}\right).$$

6.A.5 Full Conditional Distribution of $\widetilde{\boldsymbol{\beta}}$

First note that

$$\mathfrak{p}(\widetilde{\boldsymbol{\beta}} \big| \text{rest}) \propto \prod_{i=1}^{m} \left\{\mathfrak{p}(\boldsymbol{y}_i \big| \beta_0, U_i, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u, \sigma_\varepsilon^2)\right\} \mathfrak{p}(\widetilde{\boldsymbol{\beta}} \big| \sigma_\beta^2, \boldsymbol{b}_\beta).$$

If we define

$$\boldsymbol{r}_{i\widetilde{\boldsymbol{\beta}}} \equiv \boldsymbol{y}_i - \mathbf{1}_{ni}\beta_0 - \mathbf{1}_{ni}U_i - \sum_{k=1}^{d_\bullet} \boldsymbol{Z}_{ik}(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)$$

and

$$\boldsymbol{X}_{i\boldsymbol{\gamma}_\beta} \equiv \boldsymbol{X}_i \text{diag}\left(\boldsymbol{\gamma}_\beta\right)$$

then we have

$$\begin{aligned}
&\prod_{i=1}^{m}\left\{\mathfrak{p}(\boldsymbol{y}_i\big|\beta_0,U_i,\widetilde{\boldsymbol{\beta}},\boldsymbol{\gamma}_\beta,\widetilde{\boldsymbol{u}},\boldsymbol{\gamma}_u,\sigma_\varepsilon^2)\right\}\\
&\quad\propto\exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{m}\left\|\boldsymbol{y}_i-\mathbf{1}_{ni}\beta_0-\mathbf{1}_{ni}U_i-\boldsymbol{X}_i\mathrm{diag}\left(\boldsymbol{\gamma}_\beta\right)\widetilde{\boldsymbol{\beta}}-\sum_{k=1}^{d_\bullet}\boldsymbol{Z}_{ik}(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)\right\|^2\right\}\\
&\quad\propto\exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{m}\left\|\boldsymbol{r}_{i\widetilde{\boldsymbol{\beta}}}-\boldsymbol{X}_{i\boldsymbol{\gamma}_\beta}\widetilde{\boldsymbol{\beta}}\right\|^2\right\}\\
&\quad\propto\exp\left[-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{m}\left\{-2\widetilde{\boldsymbol{\beta}}^T\boldsymbol{X}_{i\boldsymbol{\gamma}_\beta}^T\boldsymbol{r}_{i\widetilde{\boldsymbol{\beta}}}+\left(\mathrm{vec}\left(\widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}^T\right)\right)^T\mathrm{vec}\left(\boldsymbol{X}_{i\boldsymbol{\gamma}_\beta}^T\boldsymbol{X}_{i\boldsymbol{\gamma}_\beta}\right)\right\}\right].
\end{aligned}$$

Then

$$\begin{aligned}
&\prod_{i=1}^{m}\mathfrak{p}(\boldsymbol{y}_i\big|\beta_0,U_i,\widetilde{\boldsymbol{\beta}},\boldsymbol{\gamma}_\beta,\widetilde{\boldsymbol{u}},\boldsymbol{\gamma}_u,\sigma_\varepsilon^2)\\
&\propto\exp\left\{\begin{bmatrix}\widetilde{\boldsymbol{\beta}}\\ \mathrm{vec}\left(\widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}^T\right)\end{bmatrix}^T\begin{bmatrix}\sum_{i=1}^{m}\left\{\frac{\boldsymbol{X}_{i\boldsymbol{\gamma}_\beta}^T}{\sigma_\varepsilon^2}\left(\boldsymbol{y}_i-\mathbf{1}_{ni}\beta_0-\mathbf{1}_{ni}U_i-\sum_{k=1}^{d_\bullet}\boldsymbol{Z}_{ik}(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)\right)\right\}\\ -\frac{1}{2}\sum_{i=1}^{m}\mathrm{vec}\left(\frac{\boldsymbol{X}_{i\boldsymbol{\gamma}_\beta}^T\boldsymbol{X}_{i\boldsymbol{\gamma}_\beta}}{\sigma_\varepsilon^2}\right)\end{bmatrix}\right\}
\end{aligned}$$

as well as

$$\mathfrak{p}(\widetilde{\boldsymbol{\beta}}\big|\sigma_\beta^2,\boldsymbol{b}_\beta)\propto\exp\left\{\begin{bmatrix}\widetilde{\boldsymbol{\beta}}\\ \mathrm{vec}\left(\widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}^T\right)\end{bmatrix}^T\begin{bmatrix}\mathbf{0}\\ -\frac{1}{2}\mathrm{vec}\left(\mathrm{diag}\left(\boldsymbol{b}_\beta\right)\right).\end{bmatrix}(1/\sigma_\beta^2)\right\}.$$

Therefore

$$\begin{aligned}
&\mathfrak{p}(\widetilde{\boldsymbol{\beta}}\big|\text{rest})\\
&\propto\exp\left\{\begin{bmatrix}\widetilde{\boldsymbol{\beta}}\\ \mathrm{vec}\left(\widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}^T\right)\end{bmatrix}^T\begin{bmatrix}\sum_{i=1}^{m}\left\{\frac{\boldsymbol{X}_{i\boldsymbol{\gamma}_\beta}^T}{\sigma_\varepsilon^2}\left(\boldsymbol{y}_i-\mathbf{1}_{ni}\beta_0-\mathbf{1}_{ni}U_i-\sum_{k=1}^{d_\bullet}\boldsymbol{Z}_{ik}(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)\right)\right\}\\ -\frac{1}{2}\mathrm{vec}\left\{\frac{1}{\sigma_\varepsilon^2}\left(\sum_{i=1}^{m}\boldsymbol{X}_{i\boldsymbol{\gamma}_\beta}^T\boldsymbol{X}_{i\boldsymbol{\gamma}_\beta}\right)+\frac{\mathrm{diag}\left(\boldsymbol{b}_\beta\right)}{\sigma_\beta^2}\right\}\end{bmatrix}\right\}.
\end{aligned}$$

The first component of the natural parameter vector of $\widetilde{\boldsymbol{\beta}}|$rest is:

$$\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{m}\left[\boldsymbol{X}_{i\boldsymbol{\gamma}_\beta}^T\left\{\boldsymbol{y}_i-\mathbf{1}_{ni}\beta_0-\mathbf{1}_{ni}U_i-\sum_{k=1}^{d_\bullet}\boldsymbol{Z}_{ik}(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)\right\}\right].$$

The inverse vec of the second component of natural parameter vector is

$$-\frac{1}{2}\left\{\frac{1}{\sigma_\varepsilon^2}\left(\sum_{i=1}^{m}\boldsymbol{X}_{i\boldsymbol{\gamma}_\beta}^T\boldsymbol{X}_{i\boldsymbol{\gamma}_\beta}\right)+\frac{\text{diag}\,(\boldsymbol{b}_\beta)}{\sigma_\beta^2}\right\}.$$

It follows that the covariance matrix of $\widetilde{\boldsymbol{\beta}}|$rest is

$$\left\{\frac{1}{\sigma_\varepsilon^2}\left(\sum_{i=1}^{m}\boldsymbol{X}_{i\boldsymbol{\gamma}_\beta}^T\boldsymbol{X}_{i\boldsymbol{\gamma}_\beta}\right)+\frac{\text{diag}\,(\boldsymbol{b}_\beta)}{\sigma_\beta^2}\right\}^{-1}$$

and the mean vector is

$$\left\{\frac{1}{\sigma_\varepsilon^2}\left(\sum_{i=1}^{m}\boldsymbol{X}_{i\boldsymbol{\gamma}_\beta}^T\boldsymbol{X}_{i\boldsymbol{\gamma}_\beta}\right)+\frac{\text{diag}\,(\boldsymbol{b}_\beta)}{\sigma_\beta^2}\right\}^{-1}\left(\frac{1}{\sigma_\varepsilon^2}\right)$$
$$\times\sum_{i=1}^{m}\left[\boldsymbol{X}_{i\boldsymbol{\gamma}_\beta}^T\left\{\boldsymbol{y}_i-\mathbf{1}_{ni}\beta_0-\mathbf{1}_{ni}U_i-\sum_{k=1}^{d_\bullet}\boldsymbol{Z}_{ik}(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)\right\}\right].$$

This leads to

$$\widetilde{\boldsymbol{\beta}}|\text{rest}\sim N\left(\left\{\frac{1}{\sigma_\varepsilon^2}\left(\sum_{i=1}^{m}\boldsymbol{X}_{i\boldsymbol{\gamma}_\beta}^T\boldsymbol{X}_{i\boldsymbol{\gamma}_\beta}\right)+\frac{\text{diag}\,(\boldsymbol{b}_\beta)}{\sigma_\beta^2}\right\}^{-1}\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{m}\left[\boldsymbol{X}_{i\boldsymbol{\gamma}_\beta}^T\left\{\boldsymbol{y}_i-\mathbf{1}_{ni}\beta_0\right.\right.\right.$$
$$\left.\left.\left.-\mathbf{1}_{ni}U_i-\sum_{k=1}^{d_\bullet}\boldsymbol{Z}_{ik}(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)\right\}\right],\left\{\frac{1}{\sigma_\varepsilon^2}\left(\sum_{i=1}^{m}\boldsymbol{X}_{i\boldsymbol{\gamma}_\beta}^T\boldsymbol{X}_{i\boldsymbol{\gamma}_\beta}\right)+\frac{\text{diag}\,(\boldsymbol{b}_\beta)}{\sigma_\beta^2}\right\}^{-1}\right).$$

Note that

$$\sum_{i=1}^{m}\boldsymbol{X}_{i\boldsymbol{\gamma}_\beta}^T\boldsymbol{X}_{i\boldsymbol{\gamma}_\beta}=\sum_{i=1}^{m}\left\{\left(\boldsymbol{X}_i\text{diag}(\boldsymbol{\gamma}_\beta)\right)^T\boldsymbol{X}_i\text{diag}(\boldsymbol{\gamma}_\beta)\right\}$$
$$=\text{diag}(\boldsymbol{\gamma}_\beta)\left(\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{X}_i\right)\text{diag}(\boldsymbol{\gamma}_\beta)=\left(\boldsymbol{\gamma}_\beta\boldsymbol{\gamma}_\beta^T\right)\odot\left(\boldsymbol{X}^T\boldsymbol{X}\right)$$

and note that

$$\sum_{i=1}^{m}\left[\boldsymbol{X}_{i\boldsymbol{\gamma}_\beta}^{T}\left\{\boldsymbol{y}_i-\mathbf{1}_{ni}\beta_0-\mathbf{1}_{ni}U_i-\sum_{k=1}^{d_\bullet}\boldsymbol{Z}_{ik}(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)\right\}\right]$$

$$=\sum_{i=1}^{m}\left[\boldsymbol{\gamma}_\beta\odot\left\{\boldsymbol{X}_i^{T}\left(\boldsymbol{y}_i-\mathbf{1}_{ni}\beta_0-\mathbf{1}_{ni}U_i-\sum_{k=1}^{d_\bullet}\boldsymbol{Z}_{ik}(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)\right)\right\}\right]$$

$$=\boldsymbol{\gamma}_\beta\odot\left\{\left(\boldsymbol{X}^{T}\boldsymbol{y}\right)-\beta_0\left(\boldsymbol{X}^{T}\mathbf{1}\right)-\boldsymbol{X}^{T}\begin{bmatrix}U_1\mathbf{1}_{n_1}\\ \vdots\\ U_m\mathbf{1}_{n_m}\end{bmatrix}-\left(\sum_{k=1}^{d_\bullet}\left(\dot{\boldsymbol{Z}}_k^{T}\boldsymbol{X}\right)^{T}(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)\right)\right\}$$

where $\dot{\boldsymbol{Z}}_k$ defined in (6.12). If we define

$$\boldsymbol{\omega}_5\equiv\left(\boldsymbol{X}^{T}\boldsymbol{y}\right)-\beta_0\left(\boldsymbol{X}^{T}\mathbf{1}\right)-\boldsymbol{X}^{T}\begin{bmatrix}U_1\mathbf{1}_{n_1}\\ \vdots\\ U_m\mathbf{1}_{n_m}\end{bmatrix}-\left(\sum_{k=1}^{d_\bullet}\left(\dot{\boldsymbol{Z}}_k^{T}\boldsymbol{X}\right)^{T}(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)\right)$$

$$=\left(\boldsymbol{X}^{T}\boldsymbol{y}\right)-\boldsymbol{X}^{T}\begin{bmatrix}U_1\mathbf{1}_{n_1}\\ \vdots\\ U_m\mathbf{1}_{n_m}\end{bmatrix}-\left(\sum_{k=1}^{d_\bullet}\left(\dot{\boldsymbol{Z}}_k^{T}\boldsymbol{X}\right)^{T}(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)\right),$$

the last step above is using the feature of the $\boldsymbol{X}$ matrix in (6.10) that $\boldsymbol{X}^{T}\mathbf{1}=\mathbf{0}$. Therefore, we obtain

$$\widetilde{\boldsymbol{\beta}}|\text{rest}\sim N\left(\left\{\frac{\left(\boldsymbol{\gamma}_\beta\boldsymbol{\gamma}_\beta^{T}\right)\odot\left(\boldsymbol{X}^{T}\boldsymbol{X}\right)}{\sigma_\varepsilon^2}+\frac{\text{diag}\left(\boldsymbol{b}_\beta\right)}{\sigma_\beta^2}\right\}^{-1}\frac{1}{\sigma_\varepsilon^2}\left[\boldsymbol{\gamma}_\beta\odot\left\{\left(\boldsymbol{X}^{T}\boldsymbol{y}\right)-\boldsymbol{X}^{T}\right.\right.\right.$$

$$\left.\left.\left.\times\begin{bmatrix}U_1\mathbf{1}_{n_1}\\ \vdots\\ U_m\mathbf{1}_{n_m}\end{bmatrix}-\left(\sum_{k=1}^{d_\bullet}\left(\dot{\boldsymbol{Z}}_k^{T}\boldsymbol{X}\right)^{T}(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)\right)\right\}\right],\left\{\frac{\left(\boldsymbol{\gamma}_\beta\boldsymbol{\gamma}_\beta^{T}\right)\odot\left(\boldsymbol{X}^{T}\boldsymbol{X}\right)}{\sigma_\varepsilon^2}+\frac{\text{diag}\left(\boldsymbol{b}_\beta\right)}{\sigma_\beta^2}\right\}^{-1}\right).$$

We also define

$$\boldsymbol{\Omega}\equiv\frac{1}{\sigma_\varepsilon^2}(\boldsymbol{X}^{T}\boldsymbol{X})\odot(\boldsymbol{\gamma}_\beta\boldsymbol{\gamma}_\beta^{T})+\frac{1}{\sigma_\beta^2}\text{diag}(\boldsymbol{b}_\beta),$$

then decompose $\boldsymbol{\Omega}=\boldsymbol{U}_{\boldsymbol{\Omega}}\text{diag}\left(\boldsymbol{d}_{\boldsymbol{\Omega}}\right)\boldsymbol{U}_{\boldsymbol{\Omega}}^{T}$ where $\boldsymbol{U}_{\boldsymbol{\Omega}}^{T}\boldsymbol{U}_{\boldsymbol{\Omega}}=\boldsymbol{I}$. If

$$\boldsymbol{z}\sim N\left(\mathbf{0},\boldsymbol{I}\right)\left((d_\circ+d_\bullet)\times 1\right),$$

then

$$\boldsymbol{U}_{\boldsymbol{\Omega}}\left(\frac{\boldsymbol{U}_{\boldsymbol{\Omega}}^T\boldsymbol{z}}{\sqrt{\boldsymbol{d}_{\boldsymbol{\Omega}}}}+\frac{\boldsymbol{U}_{\boldsymbol{\Omega}}^T\left(\boldsymbol{\gamma}_\beta\odot\boldsymbol{\omega}_5\right)}{\boldsymbol{d}_{\boldsymbol{\Omega}}\left(\sigma_\varepsilon^2\right)}\right) \text{ has the same distribution as } \widetilde{\boldsymbol{\beta}}\big|\text{rest},$$

which is as given in Algorithm 9 in Section 6.5.3.

6.A.6 Full Conditional Distribution of $\widetilde{\boldsymbol{u}}_k$

First note that, for $1\le k\le d_\bullet$,

$$\mathfrak{p}\left(\widetilde{\boldsymbol{u}}_k\big|\text{rest}\right)\propto\left\{\prod_{i=1}^m\mathfrak{p}\left(\boldsymbol{y}_i\big|\beta_0,U_i,\widetilde{\boldsymbol{\beta}},\boldsymbol{\gamma}_\beta,\widetilde{\boldsymbol{u}},\boldsymbol{\gamma}_u,\sigma_\varepsilon^2\right)\right\}\mathfrak{p}\left(\widetilde{\boldsymbol{u}}\big|\sigma_{uk}^2,b_{uk}\right).$$

If we define

$$\boldsymbol{r}_{i\widetilde{\boldsymbol{u}}_k}\equiv\boldsymbol{y}_i-\mathbf{1}_{ni}\beta_0-\mathbf{1}_{ni}U_i-\boldsymbol{X}_i(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}})-\sum_{k'\neq k}^{d_\bullet}\boldsymbol{Z}_{ik'}(\gamma_{uk'}\widetilde{\boldsymbol{u}}_{k'})$$

and define

$$\boldsymbol{Z}_{i\boldsymbol{\gamma}_{uk}}\equiv\boldsymbol{Z}_{ik}(\gamma_{uk}),$$

we have

$$\begin{aligned}\prod_{i=1}^m\mathfrak{p}(\boldsymbol{y}_i\big|\beta_0,U_i,\widetilde{\boldsymbol{\beta}},\boldsymbol{\gamma}_\beta,\widetilde{\boldsymbol{u}},\boldsymbol{\gamma}_u,\sigma_\varepsilon^2)&\propto\exp\Bigg\{-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^m\Bigg\|\boldsymbol{y}_i-\mathbf{1}_{ni}\beta_0-\mathbf{1}_{ni}U_i-\boldsymbol{X}_i(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}})\\&\qquad-\sum_{k'\neq k}^{d_\bullet}\boldsymbol{Z}_{ik'}(\gamma_{uk'}\widetilde{\boldsymbol{u}}_{k'})-\boldsymbol{Z}_{ik}(\gamma_{uk})\widetilde{\boldsymbol{u}}_k\Bigg\|^2\Bigg\}\\&\propto\exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^m\left\|\boldsymbol{r}_{i\widetilde{\boldsymbol{u}}_k}-\boldsymbol{Z}_{i\boldsymbol{\gamma}_{uk}}\right\|^2\right\}\\&\propto\exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^m\left(-2\widetilde{\boldsymbol{u}}_k^T\boldsymbol{Z}_{i\boldsymbol{\gamma}_{uk}}^T\boldsymbol{r}_{i\widetilde{\boldsymbol{u}}_k}+\left(\text{vec}\left(\widetilde{\boldsymbol{u}}_k\widetilde{\boldsymbol{u}}_k^T\right)\right)^T\text{vec}\left(\boldsymbol{Z}_{i\boldsymbol{\gamma}_{uk}}^T\boldsymbol{Z}_{i\boldsymbol{\gamma}_{uk}}\right)\right)\right\}.\end{aligned}$$

Then

$$\prod_{i=1}^{m} \mathfrak{p}(\boldsymbol{y}_i \big| \beta_0, U_i, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u, \sigma_\varepsilon^2) \propto \exp\left\{ \begin{bmatrix} \widetilde{\boldsymbol{u}}_k \\ \\ \text{vec}\left(\widetilde{\boldsymbol{u}}_k \widetilde{\boldsymbol{u}}_k^T\right) \end{bmatrix}^T \right.$$

$$\left. \times \begin{bmatrix} \sum_{i=1}^{m} \left\{ \dfrac{\boldsymbol{Z}_{i\boldsymbol{\gamma}_{uk}}^T}{\sigma_\varepsilon^2} \right\} \left(\boldsymbol{y}_i - \mathbf{1}_{ni}\beta_0 - \mathbf{1}_{ni}U_i - \boldsymbol{X}_i(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{k' \neq k}^{d\bullet} \boldsymbol{Z}_{ik'}(\gamma_{uk'}\widetilde{\boldsymbol{u}}_{k'}) \right) \\ -\dfrac{1}{2}\sum_{i=1}^{m} \text{vec}\left(\dfrac{\boldsymbol{Z}_{i\boldsymbol{\gamma}_{uk}}^T \boldsymbol{Z}_{i\boldsymbol{\gamma}_{uk}}}{\sigma_\varepsilon^2} \right) \end{bmatrix} \right\},$$

as well as

$$\mathfrak{p}(\widetilde{\boldsymbol{u}}_k \big| \sigma_{uk}^2, b_{uk}) \propto \exp\left\{ \begin{bmatrix} \widetilde{\boldsymbol{u}}_k \\ \\ \text{vec}\left(\widetilde{\boldsymbol{u}}_k \widetilde{\boldsymbol{u}}_k^T\right) \end{bmatrix}^T \begin{bmatrix} \mathbf{0} \\ \\ -\dfrac{1}{2}\text{vec}\left(\text{diag}\left(b_{uk}\boldsymbol{I}_{L_k}\right)\right). \end{bmatrix} (1/\sigma_{uk}^2) \right\}.$$

Therefore

$$\mathfrak{p}(\widetilde{\boldsymbol{u}}_k \big| \text{rest}) \propto \exp\left\{ \begin{bmatrix} \widetilde{\boldsymbol{u}}_k \\ \\ \text{vec}\left(\widetilde{\boldsymbol{u}}_k \widetilde{\boldsymbol{u}}_k^T\right) \end{bmatrix}^T \right.$$

$$\left. \times \begin{bmatrix} \sum_{i=1}^{m} \left\{ \dfrac{\boldsymbol{Z}_{i\boldsymbol{\gamma}_{uk}}^T}{\sigma_\varepsilon^2} \left(\boldsymbol{y}_i - \mathbf{1}_{ni}\beta_0 - \mathbf{1}_{ni}U_i - \boldsymbol{X}_i(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{k' \neq k}^{d\bullet} \boldsymbol{Z}_{ik'}(\gamma_{uk'}\widetilde{\boldsymbol{u}}_{k'}) \right) \right\} \\ -\dfrac{1}{2}\text{vec}\left\{ \dfrac{1}{\sigma_\varepsilon^2} \left(\sum_{i=1}^{m} \boldsymbol{Z}_{i\boldsymbol{\gamma}_{uk}}^T \boldsymbol{Z}_{i\boldsymbol{\gamma}_{uk}} \right) + \dfrac{b_{uk}}{\sigma_{uk}^2} \boldsymbol{I}_{L_k} \right\} \end{bmatrix} \right\}.$$

The first component of the natural parameter vector of $\widetilde{\boldsymbol{u}}_k \big| \text{rest}$ is:

$$\frac{1}{\sigma_\varepsilon^2} \sum_{i=1}^{m} \left[\boldsymbol{Z}_{i\boldsymbol{\gamma}_{uk}}^T \left(\boldsymbol{y}_i - \mathbf{1}_{ni}\beta_0 - \mathbf{1}_{ni}U_i - \boldsymbol{X}_i(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{k' \neq k}^{d\bullet} \boldsymbol{Z}_{ik'}(\gamma_{uk'}\widetilde{\boldsymbol{u}}_{k'}) \right) \right].$$

The inverse vec of the second component of natural parameter vector is

$$-\frac{1}{2}\left\{\frac{1}{\sigma_\varepsilon^2}\left(\sum_{i=1}^{m}\boldsymbol{Z}_{i\boldsymbol{\gamma}_{uk}}^T\boldsymbol{Z}_{i\boldsymbol{\gamma}_{uk}}\right)+\frac{b_{uk}}{\sigma_{uk}^2}\boldsymbol{I}_{L_k}\right\}.$$

It follows that the covariance matrix of $\widetilde{\boldsymbol{u}}_k|\text{rest}, 1\leq k\leq d_\bullet$, is

$$\left\{\frac{1}{\sigma_\varepsilon^2}\left(\sum_{i=1}^{m}\boldsymbol{Z}_{i\boldsymbol{\gamma}_{uk}}^T\boldsymbol{Z}_{i\boldsymbol{\gamma}_{uk}}\right)+\frac{b_{uk}}{\sigma_{uk}^2}\boldsymbol{I}_{L_k}\right\}^{-1}$$

and the mean vector is

$$\left\{\frac{1}{\sigma_\varepsilon^2}\left(\sum_{i=1}^{m}\boldsymbol{Z}_{i\boldsymbol{\gamma}_{uk}}^T\boldsymbol{Z}_{i\boldsymbol{\gamma}_{uk}}\right)+\frac{b_{uk}}{\sigma_{uk}^2}\boldsymbol{I}_{L_k}\right\}^{-1}\left(\frac{1}{\sigma_\varepsilon^2}\right)$$
$$\times\sum_{i=1}^{m}\left[\boldsymbol{Z}_{i\boldsymbol{\gamma}_{uk}}^T\left(\boldsymbol{y}_i-\mathbf{1}_{ni}\beta_0-\mathbf{1}_{ni}U_i-\boldsymbol{X}_i(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}})-\sum_{k'\neq k}^{d_\bullet}\boldsymbol{Z}_{ik'}(\gamma_{uk'}\widetilde{\boldsymbol{u}}_{k'})\right)\right].$$

Next, we simplify the covariance matrix. Recalling that

$$\boldsymbol{Z}_{i\boldsymbol{\gamma}_{uk}}=\boldsymbol{Z}_{ik}\left(\gamma_{uk}\right),$$

then

$$\sum_{i=1}^{m}\boldsymbol{Z}_{i\boldsymbol{\gamma}_{uk}}^T\boldsymbol{Z}_{i\boldsymbol{\gamma}_{uk}}=\sum_{i=1}^{m}\left(\boldsymbol{Z}_{ik}\left(\gamma_{uk}\right)\right)^T\left(\boldsymbol{Z}_{ik}\left(\gamma_{uk}\right)\right)=\left(\gamma_{uk}\right)\dot{\boldsymbol{Z}}_k^T\dot{\boldsymbol{Z}}_k\left(\gamma_{uk}\right)=\gamma_{uk}\text{diag}\left(\boldsymbol{\omega}_{\dot{\boldsymbol{Z}}_k}\right)$$

where

$$\text{diag}\left(\boldsymbol{\omega}_{\dot{\boldsymbol{Z}}_k}\right)\equiv\left(\dot{\boldsymbol{Z}}_k^T\dot{\boldsymbol{Z}}_k\right)\text{ and }\dot{\boldsymbol{Z}}_k\text{ is defined in (6.12).}$$

Then

$$\frac{1}{\sigma_\varepsilon^2}\left(\sum_{i=1}^{m}\boldsymbol{Z}_{i\boldsymbol{\gamma}_{uk}}^T\boldsymbol{Z}_{i\boldsymbol{\gamma}_{uk}}\right)+\frac{b_{uk}}{\sigma_{uk}^2}\boldsymbol{I}_{L_k}=\text{diag}\left(\frac{\gamma_{uk}\boldsymbol{\omega}_{\dot{\boldsymbol{Z}}_k}}{\sigma_\varepsilon^2}+\frac{b_{uk}\mathbf{1}_{L_k}}{\sigma_{uk}^2}\right).$$

Therefore

$$\left\{\frac{1}{\sigma_\varepsilon^2}\left(\sum_{i=1}^{m}\boldsymbol{Z}_{i\boldsymbol{\gamma}_{uk}}^T\boldsymbol{Z}_{i\boldsymbol{\gamma}_{uk}}\right)+\frac{b_{uk}}{\sigma_{uk}^2}\boldsymbol{I}_{L_k}\right\}^{-1}=\text{diag}\left(\frac{\gamma_{uk}\boldsymbol{\omega}_{\dot{\boldsymbol{Z}}_k}}{\sigma_\varepsilon^2}+\frac{b_{uk}\mathbf{1}_{L_k}}{\sigma_{uk}^2}\right)^{-1}.$$

Note that

$$\sum_{i=1}^{m}\left[\boldsymbol{Z}_{i\boldsymbol{\gamma}_{uk}}^{T}\left(\boldsymbol{y}_i-\mathbf{1}_{ni}\beta_0-\mathbf{1}_{ni}U_i-\boldsymbol{X}_i(\boldsymbol{\gamma}_{\beta}\odot\widetilde{\boldsymbol{\beta}})-\sum_{k'\neq k}^{d_\bullet}\boldsymbol{Z}_{ik'}(\gamma_{uk'}\widetilde{\boldsymbol{u}}_{k'})\right)\right]$$

$$=\gamma_{uk}\Bigg\{\dot{\boldsymbol{Z}}_k^T\boldsymbol{y}-\beta_0\left(\dot{\boldsymbol{Z}}_k^T\mathbf{1}\right)-\dot{\boldsymbol{Z}}_k^T\begin{bmatrix}U_1\mathbf{1}_{n_1}\\ \vdots\\ U_m\mathbf{1}_{n_m}\end{bmatrix}-\left(\dot{\boldsymbol{Z}}_k^T\boldsymbol{X}\right)\left(\boldsymbol{\gamma}_{\beta}\odot\widetilde{\boldsymbol{\beta}}\right)$$

$$-\sum_{k'\neq k}^{d_\bullet}\left(\dot{\boldsymbol{Z}}_k^T\dot{\boldsymbol{Z}}_{k'}\right)(\gamma_{uk'}\widetilde{\boldsymbol{u}}_{k'})\Bigg\}.$$

The required result follows if we define

$$\boldsymbol{\omega}_8\equiv\dot{\boldsymbol{Z}}_k^T\boldsymbol{y}-\beta_0\left(\dot{\boldsymbol{Z}}_k^T\mathbf{1}\right)-\dot{\boldsymbol{Z}}_k^T\begin{bmatrix}U_1\mathbf{1}_{n_1}\\ \vdots\\ U_m\mathbf{1}_{n_m}\end{bmatrix}-\left(\dot{\boldsymbol{Z}}_k^T\boldsymbol{X}\right)\left(\boldsymbol{\gamma}_{\beta}\odot\widetilde{\boldsymbol{\beta}}\right)-\sum_{k'\neq k}^{d_\bullet}\left(\dot{\boldsymbol{Z}}_k^T\dot{\boldsymbol{Z}}_{k'}\right)(\gamma_{uk'}\widetilde{\boldsymbol{u}}_{k'}).$$

Using the feature of the $\boldsymbol{Z}$ matrix in (6.13) that $\dot{\boldsymbol{Z}}_k^T\mathbf{1}=\mathbf{0}, 1\leq k\leq d_\bullet$, we obtained $\boldsymbol{\omega}_8$ in Algorithm 9 in Section 6.5.3 to be

$$\boldsymbol{\omega}_8=\dot{\boldsymbol{Z}}_k^T\boldsymbol{y}-\dot{\boldsymbol{Z}}_k^T\begin{bmatrix}U_1\mathbf{1}_{n_1}\\ \vdots\\ U_m\mathbf{1}_{n_m}\end{bmatrix}-\left(\dot{\boldsymbol{Z}}_k^T\boldsymbol{X}\right)\left(\boldsymbol{\gamma}_{\beta}\odot\widetilde{\boldsymbol{\beta}}\right)-\sum_{k'\neq k}^{d_\bullet}\left(\dot{\boldsymbol{Z}}_k^T\dot{\boldsymbol{Z}}_{k'}\right)(\gamma_{uk'}\widetilde{\boldsymbol{u}}_{k'}).$$

Therefore, we obtain (6.37) in Section 6.5.3 to be:

$$\widetilde{\boldsymbol{u}}_k\big|\text{rest}\sim N\Bigg(\text{diag}\left(\frac{\gamma_{uk}\boldsymbol{\omega}_{\dot{\boldsymbol{Z}}_k}}{\sigma_\varepsilon^2}+\frac{b_{uk}\mathbf{1}_{L_k}}{\sigma_{uk}^2}\right)^{-1}\Bigg[\gamma_{uk}\Big\{\dot{\boldsymbol{Z}}_k^T\boldsymbol{y}-\dot{\boldsymbol{Z}}_k^T\begin{bmatrix}U_1\mathbf{1}_{n_1}\\ \vdots\\ U_m\mathbf{1}_{n_m}\end{bmatrix}$$

$$-\left(\dot{\boldsymbol{Z}}_k^T\boldsymbol{X}\right)\left(\boldsymbol{\gamma}_{\beta}\odot\widetilde{\boldsymbol{\beta}}\right)-\sum_{k'\neq k}^{d_\bullet}\left(\dot{\boldsymbol{Z}}_k^T\dot{\boldsymbol{Z}}_{k'}\right)(\gamma_{uk'}\widetilde{\boldsymbol{u}}_{k'})\Big\}\Bigg],\text{diag}\left(\frac{\gamma_{uk}\boldsymbol{\omega}_{\dot{\boldsymbol{Z}}_k}}{\sigma_\varepsilon^2}+\frac{b_{uk}\mathbf{1}_{L_k}}{\sigma_{uk}^2}\right)^{-1}\Bigg).$$

We also define

$$\boldsymbol{\omega}_9\equiv\frac{\gamma_{uk}\boldsymbol{\omega}_{\dot{\boldsymbol{Z}}_k}}{\sigma_\varepsilon^2}+\frac{b_{uk}\mathbf{1}_{L_k}}{\sigma_{uk}^2}$$

and generate the $L_k \times 1$ Multivariate Normal random vector $\boldsymbol{z} \sim N\left(\boldsymbol{0}, \boldsymbol{I}_{L_k}\right)$. Then

$$\frac{\boldsymbol{z}}{\sqrt{\boldsymbol{\omega}_9}} + \frac{\gamma_{uk}\boldsymbol{\omega}_8}{\boldsymbol{\omega}_9\sigma_\varepsilon^2}$$

has the same distribution as $\widetilde{\boldsymbol{u}}_k\big|$rest, which is given in Algorithm 9 in Section 6.5.3, $1 \leq k \leq d_\bullet$.

6.A.7 Full Conditional Distribution of γ_{β_k}

First note that, for $1 \leq k \leq d_\circ + d_\bullet$,

$$\mathfrak{p}(\gamma_{\beta_k}\big|\text{rest}) \propto \left\{\prod_{i=1}^{m} \mathfrak{p}(\boldsymbol{y}_i\big|\beta_0, U_i, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u, \sigma_\varepsilon^2)\right\} \mathfrak{p}(\gamma_{\beta_k}).$$

If we define

$$\boldsymbol{r}_{i\gamma_{\beta_k}} \equiv \boldsymbol{y}_i - \boldsymbol{1}_{ni}\beta_0 - \boldsymbol{1}_{ni}U_i - (\boldsymbol{X}_i)_{-k}\left\{\left(\boldsymbol{\gamma}_\beta\right)_{-k} \odot \left(\widetilde{\boldsymbol{\beta}}\right)_{-k}\right\} - \sum_{k'=1}^{d_\bullet} \boldsymbol{Z}_{ik'}(\gamma_{uk'}\widetilde{\boldsymbol{u}}_{k'}),$$

and note that as a function of γ_{β_k}, using $\gamma_{\beta_k} \in \{0, 1\}$ implies that $\gamma_{\beta_k}^2 = \gamma_{\beta_k}$, we have

$$\begin{aligned}
&\prod_{i=1}^{m} \mathfrak{p}(\boldsymbol{y}_i\big|\beta_0, U_i, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u, \sigma_\varepsilon^2) \propto \exp\left(\gamma_{\beta_k}\eta_{\beta_{k1}}\right) \\
&\quad \propto \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{m}\left\|\boldsymbol{y}_i - \boldsymbol{1}_{ni}\beta_0 - \boldsymbol{1}_{ni}U_i - (\boldsymbol{X}_i)_{-k}\left\{\left(\boldsymbol{\gamma}_\beta\right)_{-k} \odot \left(\widetilde{\boldsymbol{\beta}}\right)_{-k}\right\} - \boldsymbol{X}_{ik}\gamma_{\beta_k}\widetilde{\beta}_k \right.\right. \\
&\qquad\qquad \left.\left. - \sum_{k'=1}^{d_\bullet} \boldsymbol{Z}_{ik'}(\gamma_{uk'}\widetilde{\boldsymbol{u}}_{k'})\right\|^2\right\} \\
&\quad \propto \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{m}\left\|\gamma_{\beta_k}\widetilde{\beta}_k\boldsymbol{X}_{ik} - \boldsymbol{r}_{i\gamma_{\beta_k}}\right\|^2\right\} \\
&\quad \propto \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\gamma_{\beta_k}\left(\widetilde{\beta}_k^2\sum_{i=1}^{m}\left\|\boldsymbol{X}_{ik}\right\|^2 - 2\widetilde{\beta}_k\sum_{i=1}^{m}\boldsymbol{X}_{ik}^T\boldsymbol{r}_{i\gamma_{\beta_k}}\right)\right\}.
\end{aligned}$$

This implies that

$$\prod_{i=1}^{m} \mathfrak{p}(\boldsymbol{y}_i \big| \beta_0, U_i, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u, \sigma_\varepsilon^2) \propto \exp\left[-\frac{1}{2\sigma_\varepsilon^2} \gamma_{\beta_k} \left\{ \widetilde{\beta}_k^2 \sum_{i=1}^{m} \left\| \boldsymbol{X}_{ik} \right\|^2 - 2\widetilde{\beta}_k \right.\right.$$
$$\left.\left. \times \sum_{i=1}^{m} \boldsymbol{X}_{ik}^T \left(\boldsymbol{y}_i - \mathbf{1}_{ni}\beta_0 - \mathbf{1}_{ni} U_i - (\boldsymbol{X}_i)_{-k} \left\{ \left(\boldsymbol{\gamma}_\beta\right)_{-k} \odot \left(\widetilde{\boldsymbol{\beta}}\right)_{-k} \right\} - \sum_{k'=1}^{d_\bullet} \boldsymbol{Z}_{ik'} (\gamma_{uk'} \widetilde{\boldsymbol{u}}_{k'}) \right) \right\} \right]$$

where

$$\eta_{\beta_{k1}} = -\frac{1}{2\sigma_\varepsilon^2} \left\{ \widetilde{\beta}_k^2 \sum_{i=1}^{m} \left\| \boldsymbol{X}_{ik} \right\|^2 - 2\widetilde{\beta}_k \sum_{i=1}^{m} \boldsymbol{X}_{ik}^T \left(\boldsymbol{y}_i - \mathbf{1}_{ni}\beta_0 - \mathbf{1}_{ni} U_i \right.\right.$$
$$\left.\left. - (\boldsymbol{X}_i)_{-k} \left\{ \left(\boldsymbol{\gamma}_\beta\right)_{-k} \odot \left(\widetilde{\boldsymbol{\beta}}\right)_{-k} \right\} - \sum_{k'=1}^{d_\bullet} \boldsymbol{Z}_{ik'} \left(\gamma_{uk'} \widetilde{\boldsymbol{u}}_{k'}\right) \right) \right\}.$$

Also, for each $1 \leq k \leq d_\circ + d_\bullet$, we have $\mathfrak{p}(\gamma_{\beta_k}) \propto \exp\left\{\gamma_{\beta_k} \text{logit}(\rho_\beta)\right\}$. Therefore, $\mathfrak{p}(\gamma_{\beta_k} \big| \text{rest}) \propto \exp\left[\gamma_{\beta_k} \left\{\eta_{\beta_{k1}} + \text{logit}(\rho_\beta)\right\}\right]$. It follows that

$$\gamma_{\beta_k} \big| \text{rest} \sim \text{Bernoulli}\left(\frac{1}{1 + \exp(-\gamma_{\beta_k})} \right), 1 \leq k \leq d_\circ + d_\bullet$$

where

$$\eta_{\beta_k} = \text{logit}(\rho_\beta) - \frac{1}{2\sigma_\varepsilon^2} \left[\widetilde{\beta}_k^2 \sum_{i=1}^{m} \left\| \boldsymbol{X}_{ik} \right\|^2 - 2\widetilde{\beta}_k \sum_{i=1}^{m} \boldsymbol{X}_{ik}^T \left\{ \boldsymbol{y}_i - \mathbf{1}_{ni}\beta_0 - \mathbf{1}_{ni} U_i - (\boldsymbol{X}_i)_{-k} \right.\right.$$
$$\left.\left. \times \left\{ \left(\boldsymbol{\gamma}_\beta\right)_{-k} \odot \left(\widetilde{\boldsymbol{\beta}}\right)_{-k} \right\} - \sum_{k'=1}^{d_\bullet} \boldsymbol{Z}_{ik'} \left(\gamma_{uk'} \widetilde{\boldsymbol{u}}_{k'}\right) \right\} \right].$$

If we define

$$\omega_6 \equiv \sum_{i=1}^{m} \boldsymbol{X}_{ik}^T \left(\boldsymbol{y}_i - \mathbf{1}_{ni}\beta_0 - \mathbf{1}_{ni} U_i - (\boldsymbol{X}_i)_{-k} \left\{ \left(\boldsymbol{\gamma}_\beta\right)_{-k} \odot \left(\widetilde{\boldsymbol{\beta}}\right)_{-k} \right\} - \sum_{k'=1}^{d_\bullet} \boldsymbol{Z}_{ik'} \left(\gamma_{uk'} \widetilde{\boldsymbol{u}}_{k'}\right) \right),$$

using the feature of the $\boldsymbol{X}$ matrix in (6.10) that $\dot{\boldsymbol{X}}_k^T \mathbf{1} = 0, 1 \leq k \leq d_\circ + d_\bullet$, with $\dot{\boldsymbol{X}}_k$ defined in (6.11) and $\dot{\boldsymbol{Z}}_k$ defined in (6.12), we obtain ω_6 in Algorithm 9 in Section

6.5.3 to be:

$$\omega_6 = \left(\dot{\boldsymbol{X}}_k^T \boldsymbol{y}\right) - \beta_0 \left(\dot{\boldsymbol{X}}_k^T \mathbf{1}\right) - \dot{\boldsymbol{X}}_k^T \begin{bmatrix} U_1 \mathbf{1}_{n_1} \\ \vdots \\ U_m \mathbf{1}_{n_m} \end{bmatrix} - \left(\dot{\boldsymbol{X}}_{(-k)}^T \dot{\boldsymbol{X}}_k\right)^T \left\{ \left(\boldsymbol{\gamma}_\beta\right)_{-k} \odot \left(\widetilde{\boldsymbol{\beta}}\right)_{-k} \right\}$$

$$-\sum_{k'=1}^{d_\bullet} \left(\dot{\boldsymbol{Z}}_{k'}^T \dot{\boldsymbol{X}}_k\right)^T \left(\gamma_{uk'} \widetilde{\boldsymbol{u}}_{k'}\right)$$

$$= \left(\boldsymbol{X}^T \boldsymbol{y}\right)^{\langle k \rangle} - \left(\boldsymbol{X}^T \begin{bmatrix} U_1 \mathbf{1}_{n_1} \\ \vdots \\ U_m \mathbf{1}_{n_m} \end{bmatrix} \right)^{\langle k \rangle} - \left\{ \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{\langle -k,k \rangle} \right\}^T$$

$$\times \left\{ \left(\boldsymbol{\gamma}_\beta\right)_{-k} \odot \left(\widetilde{\boldsymbol{\beta}}\right)_{-k} \right\} - \sum_{k'=1}^{d_\bullet} \left\{ \left(\boldsymbol{Z}^T \boldsymbol{X}\right)^{\langle k',k \rangle} \right\}^T \left(\gamma_{uk'} \widetilde{\boldsymbol{u}}_{k'}\right).$$

Note that

$$\sum_{i=1}^{m} \left\| \boldsymbol{X}_{ik} \right\|^2 = \left\| \dot{\boldsymbol{X}}_k \right\|^2 = \left(\dot{\boldsymbol{X}}_k\right)^T \dot{\boldsymbol{X}}_k = \boldsymbol{e}_k^T \left(\boldsymbol{X}^T \boldsymbol{X}\right) \boldsymbol{e}_k,$$

and we define

$$\omega_7 \equiv \text{logit}(\rho_\beta) - \frac{1}{2\sigma_\varepsilon^2} \left[\widetilde{\beta}_k^2 \sum_{i=1}^{m} \left\| \boldsymbol{X}_{ik} \right\|^2 - 2\widetilde{\beta}_k \sum_{i=1}^{m} \boldsymbol{X}_{ik}^T \left\{ \boldsymbol{y}_i - \mathbf{1}_{ni} \beta_0 - \mathbf{1}_{ni} U_i \right.\right.$$

$$\left.\left. - \left(\boldsymbol{X}_i\right)_{-k} \left\{ \left(\boldsymbol{\gamma}_\beta\right)_{-k} \odot \left(\widetilde{\boldsymbol{\beta}}\right)_{-k} \right\} - \sum_{k'=1}^{d_\bullet} \boldsymbol{Z}_{ik'} \left(\gamma_{uk'} \widetilde{\boldsymbol{u}}_{k'}\right) \right\} \right].$$

Then we obtain ω_7 in Algorithm 9 in Section 6.5.3 to be:

$$\omega_7 = \text{logit}(\rho_\beta) - \frac{1}{2\sigma_\varepsilon^2} \left\{ \widetilde{\beta}_k^2 \boldsymbol{e}_k^T \left(\boldsymbol{X}^T \boldsymbol{X}\right) \boldsymbol{e}_k - 2\widetilde{\beta}_k \omega_6 \right\}.$$

Then we obtain (6.38) in Section 6.5.3 to be:

$$\eta_{\beta_k} = \text{logit}(\rho_\beta) - \frac{1}{2\sigma_\varepsilon^2}\left\{\widetilde{\beta}_k^2\left\|\dot{\boldsymbol{X}}_k\right\|^2 - 2\widetilde{\beta}_k\left(\left(\dot{\boldsymbol{X}}_k^T\boldsymbol{y}\right) - \dot{\boldsymbol{X}}_k^T\begin{bmatrix} U_1\mathbf{1}_{n_1} \\ \vdots \\ U_m\mathbf{1}_{n_m}\end{bmatrix}\right.\right.$$
$$\left.\left. - \left(\dot{\boldsymbol{X}}_{(-k)}^T\dot{\boldsymbol{X}}_k\right)^T\left\{\left(\boldsymbol{\gamma}_\beta\right)_{-k}\odot\left(\widetilde{\boldsymbol{\beta}}\right)_{-k}\right\} - \sum_{k'=1}^{d_\bullet}\left(\dot{\boldsymbol{Z}}_{k'}^T\dot{\boldsymbol{X}}_k\right)^T(\gamma_{uk'}\widetilde{\boldsymbol{u}}_{k'})\right)\right\}.$$

6.A.8 Full Conditional Distribution of γ_{uk}

First note that, for $1 \le k \le d_\bullet$,

$$\mathfrak{p}(\gamma_{uk}|\text{rest}) \propto \left\{\prod_{i=1}^m \mathfrak{p}(\boldsymbol{y}_i|\beta_0, U_i, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u, \sigma_\varepsilon^2)\right\}\mathfrak{p}(\gamma_{uk}).$$

If we define

$$\boldsymbol{r}_{i\gamma_{uk}} \equiv \boldsymbol{y}_i - \mathbf{1}_{ni}\beta_0 - \mathbf{1}_{ni}U_i - \boldsymbol{X}_i(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}}) - \sum_{k'\neq k}^{d_\bullet}\boldsymbol{Z}_{ik'}(\gamma_{uk'}\widetilde{\boldsymbol{u}}_{k'}),$$

and note that

$$\sum_{i=1}^m\left(\boldsymbol{Z}_{ik}^T\boldsymbol{Z}_{ik}\right) = \dot{\boldsymbol{Z}}_k^T\dot{\boldsymbol{Z}}_k.$$

As a function of γ_{uk} and noting that $\gamma_{uk} \in \{0, 1\}$ implies that $\gamma_{uk}^2 = \gamma_{uk}$, we have

$$\begin{aligned}
&\prod_{i=1}^{m} \mathfrak{p}(\boldsymbol{y}_i \big| \beta_0, U_i, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u, \sigma_\varepsilon^2) \propto \exp\left(\gamma_{uk}\eta_{uk1}\right) \\
&\quad \propto \exp\Bigg\{ -\frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^{m} \Bigg\| \boldsymbol{y}_i - \mathbf{1}_{ni}\beta_0 - \mathbf{1}_{ni}U_i - \boldsymbol{X}_i(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{k' \neq k}^{d_\bullet} \boldsymbol{Z}_{ik'}(\gamma_{uk'}\widetilde{\boldsymbol{u}}_{k'}) \\
&\qquad\qquad - \boldsymbol{Z}_{ik}\gamma_{uk}\widetilde{\boldsymbol{u}}_k \Bigg\|^2 \Bigg\} \\
&\quad \propto \exp\left\{ -\frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^{m} \left\| \boldsymbol{r}_{i\gamma_{uk}} - \gamma_{uk}\boldsymbol{Z}_{ik}\widetilde{\boldsymbol{u}}_k \right\|^2 \right\} \\
&\quad \propto \exp\left[-\frac{1}{2\sigma_\varepsilon^2} \left\{ \gamma_{uk} \left(\boldsymbol{\omega}_{\dot{\boldsymbol{Z}}_k}^T (\widetilde{\boldsymbol{u}}_k \odot \widetilde{\boldsymbol{u}}_k) - 2\widetilde{\boldsymbol{u}}_k^T \sum_{i}^{m} \boldsymbol{Z}_{ik}^T \boldsymbol{r}_{i\gamma_{uk}} \right) \right\} \right]
\end{aligned}$$

with $\dot{\boldsymbol{Z}}_k$ defined as in (6.12). Using the feature of the $\boldsymbol{Z}$ matrix in (6.13) that $\dot{\boldsymbol{Z}}_k^T \mathbf{1} = \mathbf{0}, 1 \leq k \leq d_\bullet$, we have

$$\begin{aligned}
\sum_{i=1}^{m} (\boldsymbol{Z}_{ik}^T \boldsymbol{r}_{i\gamma_{uk}}) &= \sum_{i=1}^{m} \left(\boldsymbol{Z}_{ik}^T \left(\boldsymbol{y}_i - \mathbf{1}_{ni}\beta_0 - \mathbf{1}_{ni}U_i - \boldsymbol{X}_i(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{k' \neq k}^{d_\bullet} \boldsymbol{Z}_{ik'}(\gamma_{uk'}\widetilde{\boldsymbol{u}}_{k'}) \right) \right) \\
&= \dot{\boldsymbol{Z}}_k^T \boldsymbol{y} - \dot{\boldsymbol{Z}}_k^T \mathbf{1}\beta_0 - \dot{\boldsymbol{Z}}_k^T \begin{bmatrix} U_1 \mathbf{1}_{n_1} \\ \vdots \\ U_m \mathbf{1}_{n_m} \end{bmatrix} - \dot{\boldsymbol{Z}}_k^T \boldsymbol{X} \left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}} \right) - \sum_{k \neq k'}^{d_\bullet} \dot{\boldsymbol{Z}}_k^T \dot{\boldsymbol{Z}}_{k'} \left(\gamma_{uk'} \widetilde{\boldsymbol{u}}_{k'} \right) \\
&= \dot{\boldsymbol{Z}}_k^T \boldsymbol{y} - \dot{\boldsymbol{Z}}_k^T \begin{bmatrix} U_1 \mathbf{1}_{n_1} \\ \vdots \\ U_m \mathbf{1}_{n_m} \end{bmatrix} - \dot{\boldsymbol{Z}}_k^T \boldsymbol{X} \left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}} \right) - \sum_{k \neq k'}^{d_\bullet} \dot{\boldsymbol{Z}}_k^T \dot{\boldsymbol{Z}}_{k'} \left(\gamma_{uk'} \widetilde{\boldsymbol{u}}_{k'} \right).
\end{aligned}$$

This implies that

$$\prod_{i=1}^{m} \mathfrak{p}(\boldsymbol{y}_i \big| \beta_0, U_i, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u, \sigma_\varepsilon^2) \propto \exp\Bigg[-\frac{1}{2\sigma_\varepsilon^2} \gamma_{uk} \Bigg\{ \boldsymbol{\omega}_{\dot{\boldsymbol{Z}}_k}^T \left(\widetilde{\boldsymbol{u}}_k \odot \widetilde{\boldsymbol{u}}_k \right) - 2\widetilde{\boldsymbol{u}}_k^T$$

$$\times \Big(\dot{\boldsymbol{Z}}_k^T \boldsymbol{y} - \dot{\boldsymbol{Z}}_k^T \begin{bmatrix} U_1 \mathbf{1}_{n_1} \\ \vdots \\ U_m \mathbf{1}_{n_m} \end{bmatrix} - \dot{\boldsymbol{Z}}_k^T \boldsymbol{X} \left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}} \right) - \sum_{k \neq k'}^{d_\bullet} \dot{\boldsymbol{Z}}_k^T \dot{\boldsymbol{Z}}_{k'} \left(\gamma_{uk'} \widetilde{\boldsymbol{u}}_{k'} \right) \Big) \Bigg\} \Bigg]$$

$$\propto \exp\left(\gamma_{uk} \eta_{uk1} \right)$$

where

$$\eta_{uk1} = -\frac{1}{2\sigma_\varepsilon^2} \Bigg\{ \boldsymbol{\omega}_{\dot{\boldsymbol{Z}}_k}^T \left(\widetilde{\boldsymbol{u}}_k \odot \widetilde{\boldsymbol{u}}_k \right) - 2\widetilde{\boldsymbol{u}}_k^T \Bigg(\dot{\boldsymbol{Z}}_k^T \boldsymbol{y} - \dot{\boldsymbol{Z}}_k^T \begin{bmatrix} U_1 \mathbf{1}_{n_1} \\ \vdots \\ U_m \mathbf{1}_{n_m} \end{bmatrix}$$

$$- \dot{\boldsymbol{Z}}_k^T \boldsymbol{X} \Big(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}} \Big) - \sum_{k \neq k'}^{d_\bullet} \dot{\boldsymbol{Z}}_k^T \dot{\boldsymbol{Z}}_{k'} \left(\gamma_{uk'} \widetilde{\boldsymbol{u}}_{k'} \right) \Bigg) \Bigg\}.$$

Also, for each $1 \leq k \leq d_\bullet$,

$$\mathfrak{p}(\gamma_{uk}) \propto \exp\left\{ \gamma_{uk} \text{logit}(\rho_u) \right\}.$$

Therefore

$$\mathfrak{p}(\gamma_{uk} \big| \text{rest}) \propto \exp\left[\gamma_{uk} \left\{ \eta_{uk1} + \text{logit}(\rho_u) \right\} \right].$$

Hence, (6.39) is obtained in Section 6.5.3 to be:

$$\gamma_{uk} \big| \text{rest} \sim \text{Bernoulli} \left(\frac{1}{1 + \exp(-\gamma_{uk})} \right), 1 \leq k \leq d_\bullet$$

where

$$\eta_{uk} = \text{logit}(\rho_u) - \frac{1}{2\sigma_\varepsilon^2}\Bigg[\boldsymbol{\omega}_{\dot{\boldsymbol{Z}}_k}^T\left(\widetilde{\boldsymbol{u}}_k \odot \widetilde{\boldsymbol{u}}_k\right) - 2\widetilde{\boldsymbol{u}}_k^T\Bigg\{\dot{\boldsymbol{Z}}_k^T\boldsymbol{y} - \dot{\boldsymbol{Z}}_k^T\begin{bmatrix} U_1\mathbf{1}_{n_1} \\ \vdots \\ U_m\mathbf{1}_{n_m}\end{bmatrix} - \dot{\boldsymbol{Z}}_k^T\boldsymbol{X}\left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}\right) - \sum_{k\neq k'}^{d_\bullet}\dot{\boldsymbol{Z}}_k^T\dot{\boldsymbol{Z}}_{k'}\left(\gamma_{uk'}\widetilde{\boldsymbol{u}}_{k'}\right)\Bigg\}\Bigg].$$

Therefore, $\boldsymbol{\omega}_{10}$ and ω_{11} are defined and obtained in Algorithm 9 in Section 6.5.3 to be:

$$\boldsymbol{\omega}_{10} \equiv \dot{\boldsymbol{Z}}_k^T\boldsymbol{y} - \dot{\boldsymbol{Z}}_k^T\begin{bmatrix} U_1\mathbf{1}_{n_1} \\ \vdots \\ U_m\mathbf{1}_{n_m}\end{bmatrix} - \dot{\boldsymbol{Z}}_k^T\boldsymbol{X}\left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}\right) - \sum_{k\neq k'}^{d_\bullet}\dot{\boldsymbol{Z}}_k^T\dot{\boldsymbol{Z}}_{k'}\left(\gamma_{uk'}\widetilde{\boldsymbol{u}}_{k'}\right)$$

and

$$\begin{aligned}\omega_{11} \equiv \eta_{uk} &= \text{logit}(\rho_u) - \frac{1}{2\sigma_\varepsilon^2}\Bigg[\boldsymbol{\omega}_{\dot{\boldsymbol{Z}}_k}^T\left(\widetilde{\boldsymbol{u}}_k \odot \widetilde{\boldsymbol{u}}_k\right) - 2\widetilde{\boldsymbol{u}}_k^T\Bigg\{\dot{\boldsymbol{Z}}_k^T\boldsymbol{y} - \dot{\boldsymbol{Z}}_k^T\begin{bmatrix} U_1\mathbf{1}_{n_1} \\ \vdots \\ U_m\mathbf{1}_{n_m}\end{bmatrix} - \dot{\boldsymbol{Z}}_k^T\boldsymbol{X}\left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}\right) - \sum_{k\neq k'}^{d_\bullet}\dot{\boldsymbol{Z}}_k^T\dot{\boldsymbol{Z}}_{k'}\left(\gamma_{uk'}\widetilde{\boldsymbol{u}}_{k'}\right)\Bigg\}\Bigg] \\ &= \text{logit}(\rho_u) - \frac{1}{2\sigma_\varepsilon^2}\left\{\boldsymbol{\omega}_{\dot{\boldsymbol{Z}}_k}^T\left(\widetilde{\boldsymbol{u}}_k \odot \widetilde{\boldsymbol{u}}_k\right) - 2\widetilde{\boldsymbol{u}}_k^T\boldsymbol{\omega}_{10}\right\}.\end{aligned}$$

6.A.9 Full Conditional Distribution of σ_ε^2

First note that, for $1 \leq i \leq m$,

$$\mathfrak{p}(\sigma_\varepsilon^2|\text{rest}) \propto \left\{\prod_{i=1}^{m}\mathfrak{p}(\boldsymbol{y}_i|\beta_0, U_i, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u, \sigma_\varepsilon^2)\right\}\mathfrak{p}(\sigma_\varepsilon^2|a_\varepsilon)$$

with $N \equiv \sum_{i=1}^{m} n_i$, where

$$\prod_{i=1}^{m} \mathfrak{p}(\boldsymbol{y}_i|\beta_0, U_i, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u, \sigma_\varepsilon^2) \propto \left(\frac{1}{\sqrt{\sigma_\varepsilon^2}}\right)^N \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{m}(\boldsymbol{y}_i - \mathbf{1}_{n_i}\beta_0 - \mathbf{1}_{n_i}U_i - \boldsymbol{X}_i(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{k=1}^{d_\bullet}\boldsymbol{Z}_{ik}(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)\right)\right\}$$

$$\propto \left(\frac{1}{\sqrt{\sigma_\varepsilon^2}}\right)^N \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\left(\boldsymbol{y} - \mathbf{1}_N\beta_0 - \begin{bmatrix} U_1\mathbf{1}_{n_1} \\ \vdots \\ U_m\mathbf{1}_{n_m}\end{bmatrix} - \boldsymbol{X}(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{k=1}^{d_\bullet}\dot{\boldsymbol{Z}}_k(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)\right)\right\}$$

and

$$\mathfrak{p}(\sigma_\varepsilon^2|a_\varepsilon) \propto \left(\sigma_\varepsilon^2\right)^{-\frac{1}{2}-1}\exp\left(-\frac{1}{a_\varepsilon\sigma_\varepsilon^2}\right).$$

Therefore

$$\mathfrak{p}\left(\sigma_\varepsilon^2|\text{rest}\right) \propto \left(\sigma_\varepsilon^2\right)^{-\frac{1+N}{2}-1} \exp\left\{-\frac{1}{\sigma_\varepsilon^2}\left(\frac{1}{a_\varepsilon} + \frac{1}{2}\left\|\boldsymbol{y} - \mathbf{1}_N\beta_0 - \begin{bmatrix} U_1\mathbf{1}_{n_1} \\ \vdots \\ U_m\mathbf{1}_{n_m}\end{bmatrix} - \boldsymbol{X}(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{k=1}^{d_\bullet}\dot{\boldsymbol{Z}}_k(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)\right\|^2\right)\right\}.$$

In summary, if the response type is Gaussian, we obtain (6.40) in Section 6.5.3 to be:

$$\sigma_\varepsilon^2|\text{rest} \sim \text{Inverse-Gamma}\left(\frac{1+N}{2}, \frac{1}{a_\varepsilon} + \frac{1}{2}\left\|\boldsymbol{y} - \mathbf{1}_N\beta_0 - \begin{bmatrix} U_1\mathbf{1}_{n_1} \\ \vdots \\ U_m\mathbf{1}_{n_m}\end{bmatrix} - \boldsymbol{X}\left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}\right) - \sum_{k=1}^{d_\bullet}\dot{\boldsymbol{Z}}_k\left(\gamma_{uk}\widetilde{\boldsymbol{u}}_k\right)\right\|^2\right).$$

If we define

$$\boldsymbol{\omega}_{12} = \mathbf{1}_N \beta_0 + \begin{bmatrix} U_1 \mathbf{1}_{n_1} \\ \vdots \\ U_m \mathbf{1}_{n_m} \end{bmatrix} + \boldsymbol{X}\left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}\right) + \sum_{k=1}^{d_\bullet} \dot{\boldsymbol{Z}}_k \left(\gamma_{uk} \widetilde{\boldsymbol{u}}_k\right)$$

then

$$\sigma_\varepsilon^2 \big| \text{rest} \sim \text{Inverse-Gamma}\left(\frac{1+N}{2}, \left(1/a_\varepsilon\right) + \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{\omega}_{12}\right\|^2\right).$$

If the response data is binary, note that $\mu_{ij} = \left(\boldsymbol{\omega}_{12}\right)_{ij}, 1 \le i \le m, 1 \le j \le n_i$, following (6.26) and (6.27) in Section 6.4.3, we have

$$y_{ij} = 1 \text{ if and only if } c_{ij} \ge 0;$$

$$y_{ij} \big| c_{ij} \overset{\text{ind.}}{\sim} \text{Bernoulli}(I(c_{ij} \ge 0));$$

$$c_{ij} \big| \beta_0, U_i, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u \sim N\left(\mu_{ij}, 1\right), \;\; 1 \le i \le m, \;\; 1 \le j \le n_i.$$

Then

$$\mathfrak{p}(c_{ij} \big| \text{rest}) \propto \mathfrak{p}(y_{ij} \big| c_{ij}) \mathfrak{p}(c_{ij} \big| \beta_0, U_i, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u) \propto \begin{cases} \phi(c_{ij} - \mu_{ij}) I(c_{ij} \ge 0); \\ \phi(c_{ij} - \mu_{ij}) I(c_{ij} < 0). \end{cases}$$

Therefore

$$c_{ij} \big| \text{rest} \sim \begin{cases} \text{Truncated-Normal}_+(\mu_{ij}, 1), & y_{ij} = 1; \\ \\ \text{Truncated-Normal}_-(\mu_{ij}, 1), & y_{ij} = 0. \end{cases}$$

Using **Result 1.7.3**, we have

$$c_{ij} \big| \text{rest} \sim \text{Truncated-Normal}_-(\mu_{ij}, 1) \text{ equivalent to}$$

$$-c_{ij} \big| \text{rest} \sim \text{Truncated-Normal}_+(-\mu_{ij}, 1).$$

Therefore

$$c_{ij} \big| \text{rest} \sim \begin{cases} \text{Truncated-Normal}_+(\mu_{ij}, 1), & y_{ij} = 1; \\ \\ -\text{Truncated-Normal}_+(-\mu_{ij}, 1), & y_{ij} = 0. \end{cases}$$

In summary, for $1 \leq i \leq m, \;\; 1 \leq j \leq n_i$,

$$c_{ij}\big|\text{rest} \sim (2y_{ij}-1)\text{Truncated-Normal}_{+}((2y_{ij}-1)\mu_{ij}, 1)$$

is as given in Algorithm 9 in Section 6.5.3 for the binary responses. Then we obtain (6.42) in Section 6.5.3 to be

$$\begin{aligned}
&y_{ij}\big|c_{ij} \overset{\text{ind.}}{\sim} \text{Bernoulli}\left(I\left(c_{ij} \geq 0\right)\right), \;\; 1 \leq i \leq m, \;\; 1 \leq j \leq n_i, \\
&c_{ij}\big|\text{rest} \sim (2y_{ij}-1)\,\text{Truncated-Normal}_{+} \\
&\left((2y_{ij}-1)\left[\mathbf{1}_{n_i}\beta_0 + \mathbf{1}_{n_i}U_i + \boldsymbol{X}_i\left(\boldsymbol{\gamma}_{\beta} \odot \widetilde{\boldsymbol{\beta}}\right) + \sum_{k=1}^{d_{\bullet}} \boldsymbol{Z}_{ik}\left(\gamma_{uk}\widetilde{\boldsymbol{u}}_k\right)\right]_j, 1\right), \\
&y_{ij} = 1 \quad \text{or} \quad y_{ij} = 0.
\end{aligned}$$

Recalling that

$$\begin{aligned}
\mu_{ij} = \left(\boldsymbol{\omega}_{12}\right)_{ij} &= \left\{\mathbf{1}_N\beta_0 + \begin{bmatrix} U_1\mathbf{1}_{n_1} \\ \vdots \\ U_m\mathbf{1}_{n_m} \end{bmatrix} + \boldsymbol{X}\left(\boldsymbol{\gamma}_{\beta} \odot \widetilde{\boldsymbol{\beta}}\right) + \sum_{k=1}^{d_{\bullet}} \dot{\boldsymbol{Z}}_k\left(\gamma_{uk}\widetilde{\boldsymbol{u}}_k\right)\right\}_{ij} \\
&= \left[\mathbf{1}_{n_i}\beta_0 + \mathbf{1}_{n_i}U_i + \boldsymbol{X}_i\left(\boldsymbol{\gamma}_{\beta} \odot \widetilde{\boldsymbol{\beta}}\right) + \sum_{k=1}^{d_{\bullet}} \boldsymbol{Z}_{ik}\left(\gamma_{uk}\widetilde{\boldsymbol{u}}_k\right)\right]_j
\end{aligned}$$

then

$$c_{ij}\big|\text{rest} \sim (2y_{ij}-1)\,\text{Truncated-Normal}_{+}\left((2y_{ij}-1)\,\mu_{ij}, 1\right), y_{ij} = 1 \text{ or } y_{ij} = 0.$$

If we define

$$\left(\boldsymbol{\omega}_{13}\right)_{ij} \equiv \text{Truncated-Normal}_{+}((2y_{ij}-1)\mu_{ij}, 1),$$

then

$$c_{ij}\big|\text{rest} \sim (2y_{ij}-1)\left(\boldsymbol{\omega}_{13}\right)_{ij}$$

is as given in Algorithm 9 in Section 6.5.3 for the binary responses.

6.A.10 Full Conditional Distribution of a_ε

Following the steps in Appendix 2.B.7, we obtain (6.41) in Section 6.5.3 to be

$$a_\varepsilon \big| \text{rest} \sim \text{Inverse-Gamma}\left(1, \frac{1}{\sigma_\varepsilon^2} + \frac{1}{s_\varepsilon^2}\right).$$

6.A.11 Full Conditional Distribution of b_{β_k}

Following the steps in Appendix 2.B.8, we obtain (6.43) in Section 6.5.3 to be:

$$b_{\beta_k} \big| \text{rest} \sim \text{Inverse-Gaussian}\left(\frac{\sigma_\beta}{|\widetilde{\beta}_k|}, 1\right), \quad 1 \leq k \leq d_\circ + d_\bullet.$$

6.A.12 Full Conditional Distribution of b_{u_k}

Following the steps in Appendix 2.B.9, we obtain (6.44) in Section 6.5.3 to be:

$$b_{uk} \big| \text{rest} \sim \text{Inverse-Gaussian}\left(\frac{\sigma_{uk}}{\|\widetilde{\boldsymbol{u}}_k\|}, 1\right), 1 \leq k \leq d_\bullet.$$

6.A.13 Full Conditional Distribution of σ_β^2

Following the steps in Appendix 2.B.10, we obtain (6.45) in Section 6.5.3 to be:

$$\sigma_\beta^2 \big| \text{rest} \sim \text{Inverse-Gamma}\left(\frac{d_\circ + d_\bullet + 1}{2}, \frac{\widetilde{\boldsymbol{\beta}}^T \text{diag}(\boldsymbol{b}_\beta)\widetilde{\boldsymbol{\beta}}}{2} + \frac{1}{a_\beta}\right).$$

6.A.14 Full Conditional Distribution of a_β

Following the steps in Appendix 2.B.11, we obtain (6.46) in Section 6.5.3 to be:

$$a_\beta \big| \text{rest} \sim \text{Inverse-Gamma}\left(1, \frac{1}{\sigma_\beta^2} + \frac{1}{s_\beta^2}\right).$$

6.A.15 Full Conditional Distribution of σ^2_{uk}

Following the steps in Appendix 2.B.12, we obtain (6.47) in Section 6.5.3 to be:

$$\sigma^2_{uk}\big|\text{rest} \sim \text{Inverse-Gamma}\left(\frac{L_k+1}{2}, \frac{b_{uk}\|\widetilde{\boldsymbol{u}}_k\|^2}{2} + \frac{1}{a_{uk}}\right), 1 \leq k \leq d_\bullet.$$

6.A.16 Full Conditional Distribution of a_{uk}

Following the steps in Appendix 2.B.13, we obtain (6.48) in Section 6.5.3 to be:

$$a_{uk}\big|\text{rest} \sim \text{Inverse-Gamma}\left(1, \frac{1}{\sigma^2_{uk}} + \frac{1}{s^2_u}\right), 1 \leq k \leq d_\bullet.$$

6.B The q-Density approximation of parameters

Throughout Appendix 6.B, "const" denotes a constant that does not depend on any q-density parameters.

6.B.1 The q-Density of β_0

Similar to the approach of used to obtain the q-density of β_0 in Appendix 4.A.1, and recalling that $\mathfrak{p}\left(\beta_0|\text{rest}\right)$ in (6.86), we have

$$\mathfrak{q}\left(\beta_0\right) \text{ having a } N\left(\mu_{\mathfrak{q}(\beta_0)}, \sigma^2_{\mathfrak{q}(\beta_0)}\right) \text{ density function}$$

where the q-density parameters are determined from

$$E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\beta_0\middle|\text{rest}\right)\right\}\right] = \begin{bmatrix}\beta_0 \\ \\ \beta_0^2\end{bmatrix}^T$$

$$\times E_{\mathfrak{q}}\left[\begin{array}{c}\dfrac{1}{\sigma_\varepsilon^2}\left(\boldsymbol{y}^T\mathbf{1} - \displaystyle\sum_{i=1}^{m}\left(n_i U_i\right) - \left(\boldsymbol{X}^T\mathbf{1}\right)^T\left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}\right) - \sum_{k=1}^{d_\bullet}\left(\dot{\boldsymbol{Z}}_k^T\mathbf{1}\right)^T\left(\gamma_{uk}\widetilde{\boldsymbol{u}}_k\right)\right) \\ \\ -\dfrac{1}{2}\left(\dfrac{1}{\sigma_\varepsilon^2}\displaystyle\sum_{i=1}^{m} n_i + \dfrac{1}{\sigma_{\beta_0}^2}\right)\end{array}\right].$$

If we continue to define $N \equiv \displaystyle\sum_{i=1}^{m} n_i$, note that

$$-\frac{1}{2}E_{\mathfrak{q}}\left(\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{m} n_i + \frac{1}{\sigma_{\beta_0}^2}\right) = -\frac{1}{2}\left(N\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)} + \frac{1}{\sigma_{\beta_0}^2}\right) = -\frac{1}{2}\left(\sigma^2_{\mathfrak{q}(\beta_0)}\right)^{-1}$$

and therefore

$$\sigma^2_{\mathfrak{q}(\beta_0)} = \left(N\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)} + \frac{1}{\sigma_{\beta_0}^2}\right)^{-1}$$

is given in (6.53) and Algorithm 10 of Section 6.6.2. The other q-density natural parameter is

$$\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\Bigg[\boldsymbol{y}^T\mathbf{1}-\sum_{i=1}^{m}\left(n_i\mu_{\mathfrak{q}(U_i)}\right)-\left(\boldsymbol{X}^T\mathbf{1}\right)^T\left(\boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{\gamma}_\beta\right)}\odot\boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\right)$$
$$-\sum_{k=1}^{d_\bullet}\left(\dot{\boldsymbol{Z}}_k^T\mathbf{1}\right)^T\left(\mu_{\mathfrak{q}(\gamma_{uk})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right)\Bigg]. \tag{6.88}$$

Recalling that (6.10) has $\mathbf{1}_N^T\boldsymbol{X}=\mathbf{0}$ and (6.13) in Section 6.3.3 has $\dot{\boldsymbol{Z}}_k^T\mathbf{1}_N=\mathbf{0}, 1\le k\le d_\bullet$, in (6.88), if we define

$$\omega_{12}\equiv\boldsymbol{y}^T\mathbf{1}-\sum_{i=1}^{m}\left(n_i\mu_{\mathfrak{q}(U_i)}\right)-\left(\boldsymbol{X}^T\mathbf{1}\right)^T\left(\boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{\gamma}_\beta\right)}\odot\boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\right)-\sum_{k=1}^{d_\bullet}\left(\dot{\boldsymbol{Z}}_k^T\mathbf{1}\right)^T\left(\mu_{\mathfrak{q}(\gamma_{uk})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right),$$

then $\omega_{12}=\boldsymbol{y}^T\mathbf{1}-\sum_{i=1}^{m}\left(n_i\mu_{\mathfrak{q}(U_i)}\right)$. Therefore

$$\mu_{\mathfrak{q}(\beta_0)}=\sigma^2_{\mathfrak{q}(\beta_0)}\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\omega_{12}=\sigma^2_{\mathfrak{q}(\beta_0)}\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\left(\boldsymbol{y}^T\mathbf{1}-\sum_{i=1}^{m}\left(n_i\mu_{\mathfrak{q}(U_i)}\right)\right)$$

matches the expression given in (6.54) and Algorithm 10 of Section 6.6.2.

6.B.2 The q-Density of U_i

Noting the expression for $\mathfrak{p}\left(U_i\middle|\text{rest}\right)$ in (6.87), $1\le i\le m$, we have

$$\mathfrak{q}\left(U_i\right)\text{ having a }N\left(\mu_{\mathfrak{q}(U_i)},\sigma^2_{\mathfrak{q}(U_i)}\right)\text{ density function,}$$

where the q-density parameters are determined from

$$E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(U_i\middle|\text{rest}\right)\right\}\right]$$
$$=\begin{bmatrix}U_i\\U_i^2\end{bmatrix}^T E_{\mathfrak{q}}\begin{bmatrix}\frac{1}{\sigma_\varepsilon^2}\mathbf{1}_{n_i}^T\left(\boldsymbol{y}_i-\mathbf{1}_{n_i}\beta_0-\boldsymbol{X}_i\left(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}}\right)-\sum_{k=1}^{d_\bullet}\boldsymbol{Z}_{ik}\left(\gamma_{uk}\widetilde{\boldsymbol{u}}_k\right)\right)\\-\frac{1}{2}\left(\frac{n_i}{\sigma_\varepsilon^2}+\frac{1}{\sigma_U^2}\right)\end{bmatrix}.$$

Note that

$$-\frac{1}{2}E_{\mathfrak{q}}\left(\frac{n_i}{\sigma_\varepsilon^2}+\frac{1}{\sigma_U^2}\right)=-\frac{1}{2}\left(n_i\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}+\mu_{\mathfrak{q}\left(1/\sigma_U^2\right)}\right)=-\frac{1}{2}\left(\sigma^2_{\mathfrak{q}(U_i)}\right)^{-1}$$

and therefore

$$\sigma^2_{\mathfrak{q}(U_i)}=\left(n_i\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}+\mu_{\mathfrak{q}\left(1/\sigma_U^2\right)}\right)^{-1}$$

is given in (6.55) and Algorithm 10 of Section 6.6.2. The other q-density natural parameter is

$$\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\Bigg(\boldsymbol{y}_i^T\mathbf{1}_{n_i}-n_i\mu_{\mathfrak{q}(\beta_0)}-\left(\boldsymbol{X}_i^T\mathbf{1}_{n_i}\right)^T\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}\odot\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right)$$
$$-\sum_{k=1}^{d_\bullet}\left(\boldsymbol{Z}_{ik}\mathbf{1}_{n_i}\right)^T\left(\mu_{\mathfrak{q}(\gamma_{uk})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right)\Bigg). \tag{6.89}$$

Then

$$\mu_{\mathfrak{q}(U_i)}=\sigma^2_{\mathfrak{q}(U_i)}\mu_{\mathfrak{q}\left(1/\sigma_\varepsilon^2\right)}\Bigg(\boldsymbol{y}_i^T\mathbf{1}_{n_i}-n_i\mu_{\mathfrak{q}(\beta_0)}$$
$$-\left(\boldsymbol{X}_i^T\mathbf{1}_{n_i}\right)^T\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}\odot\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right)-\sum_{k=1}^{d_\bullet}\left(\boldsymbol{Z}_{ik}\mathbf{1}_{n_i}\right)^T\left(\mu_{\mathfrak{q}(\gamma_{uk})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right)\Bigg)$$

which matches the expression given in (6.56) and Algorithm 10 of Section 6.6.2.

6.B.3 The q-Density of σ_U^2

Recalling the expression of $\sigma_U^2\big|$rest in (6.34), we note that

q-density of σ_U^2 is the Inverse-Gamma $\left(\frac{m+1}{2},\lambda_{\mathfrak{q}\left(\sigma_U^2\right)}\right)$ density function where

$$\lambda_{\mathfrak{q}\left(\sigma_U^2\right)}=\mu_{\mathfrak{q}(1/a_U)}+\frac{1}{2}\left(\left\|\boldsymbol{\mu}_{\mathfrak{q}(U)}\right\|^2+\sum_{i=1}^{m}\sigma^2_{\mathfrak{q}(U_i)}\right).$$

6.B.4 The q-Density of a_U

Recalling the expression of $a_U|\text{rest}$ in (6.35), we have

$$\mathfrak{q}\left(a_U\right) \text{ has an Inverse-Gamma } \left(1, \mu_{\mathfrak{q}\left(1/\sigma_U^2\right)} + s_U^2\right) \text{ density function.}$$

6.B.5 The q-Density of $\widetilde{\boldsymbol{\beta}}$

Recalling the expression for $\widetilde{\boldsymbol{\beta}}|\text{rest}$ in (6.36), the q-density of $\widetilde{\boldsymbol{\beta}}$ is the

$$N\left(\boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}, \boldsymbol{\Sigma}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\right) \text{ density function}$$

where the q-density parameters are determined from

$$E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\widetilde{\boldsymbol{\beta}}|\text{rest}\right)\right\}\right] = \begin{bmatrix} \widetilde{\boldsymbol{\beta}} \\ \text{vec}\left(\widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}^T\right) \end{bmatrix}^T$$

$$\times E_{\mathfrak{q}}\left[\begin{array}{c} \frac{1}{\sigma_\varepsilon^2}\left[\boldsymbol{\gamma}_\beta \odot \left\{\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T\begin{bmatrix} U_1\mathbf{1}_{n_1} \\ \vdots \\ U_m\mathbf{1}_{n_m} \end{bmatrix} - \sum_{k=1}^{d_\bullet}\left(\dot{\boldsymbol{Z}}_k^T\boldsymbol{X}\right)^T\left(\gamma_{uk}\widetilde{\boldsymbol{u}}_k\right)\right\}\right] \\ -\frac{1}{2}\text{vec}\left\{\left(\frac{1}{\sigma_\varepsilon^2}\right)\left(\boldsymbol{\gamma}_\beta\boldsymbol{\gamma}_\beta^T\right) \odot \left(\boldsymbol{X}^T\boldsymbol{X}\right) + \frac{\text{diag}\left(\boldsymbol{b}_\beta\right)}{\sigma_\beta^2}\right\} \end{array}\right]$$

As in Appendix 4.A.2, we define

$$\begin{aligned} \boldsymbol{\Omega}_{\mathfrak{q}\left(\boldsymbol{\gamma}_\beta\right)} &\equiv E_{\mathfrak{q}}\left(\boldsymbol{\gamma}_\beta\boldsymbol{\gamma}_\beta^T\right) = \text{Cov}_{\mathfrak{q}}\left(\boldsymbol{\gamma}_\beta\right) + E_{\mathfrak{q}}\left(\boldsymbol{\gamma}_\beta\right)E_{\mathfrak{q}}\left(\boldsymbol{\gamma}_\beta^T\right) \\ &= \text{diag}\left\{\boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{\gamma}_\beta\right)} \odot \left(\mathbf{1} - \boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{\gamma}_\beta\right)}\right)\right\} + \boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{\gamma}_\beta\right)}\boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{\gamma}_\beta\right)}^T. \end{aligned}$$

Note that

$$-\frac{1}{2}E_{\mathfrak{q}}\left\{\frac{1}{\sigma_\varepsilon^2}\left(\boldsymbol{\gamma}_\beta\boldsymbol{\gamma}_\beta^T\right) \odot \left(\boldsymbol{X}^T\boldsymbol{X}\right) + \frac{\text{diag}\left(\boldsymbol{b}_\beta\right)}{\sigma_\beta^2}\right\} = -\frac{1}{2}\boldsymbol{\Sigma}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}^{-1}.$$

Then

$$\boldsymbol{\Sigma}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)} = \left[\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\left\{\boldsymbol{\Omega}_{\mathfrak{q}\left(\boldsymbol{\gamma}_\beta\right)} \odot \left(\boldsymbol{X}^T\boldsymbol{X}\right)\right\} + \mu_{\mathfrak{q}\left(1/\sigma_\beta^2\right)}\text{diag}\left(\boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{b}_\beta\right)}\right)\right]^{-1}.$$

The other q-density natural parameter is

$$\boldsymbol{\Sigma}^{-1}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)} = \mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\Bigg(\boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{\gamma}_\beta\right)} \odot \Bigg(\boldsymbol{X}^T\boldsymbol{y} - \mu_{\mathfrak{q}(\beta_0)}\left(\boldsymbol{X}^T\mathbf{1}\right) - \boldsymbol{X}^T\begin{bmatrix}\mu_{\mathfrak{q}(U_1)}\mathbf{1}_{n_1}\\ \vdots \\ \mu_{\mathfrak{q}(U_m)}\mathbf{1}_{n_m}\end{bmatrix}$$

$$-\sum_{k=1}^{d_\bullet}\left(\dot{\boldsymbol{Z}}_k^T\boldsymbol{X}\right)^T\left(\mu_{\mathfrak{q}(\gamma_{uk})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right)\Bigg)\Bigg). \tag{6.90}$$

Recalling that (6.10) has $\mathbf{1}_N^T\boldsymbol{X} = \mathbf{0}$, we have

$$\boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)} = \mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\boldsymbol{\Sigma}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\Bigg(\boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{\gamma}_\beta\right)} \odot \Bigg(\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T\begin{bmatrix}\mu_{\mathfrak{q}(U_1)}\mathbf{1}_{n_1}\\ \vdots \\ \mu_{\mathfrak{q}(U_m)}\mathbf{1}_{n_m}\end{bmatrix}$$

$$-\sum_{k=1}^{d_\bullet}\left(\dot{\boldsymbol{Z}}_k^T\boldsymbol{X}\right)^T\left(\mu_{\mathfrak{q}(\gamma_{uk})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right)\Bigg)\Bigg).$$

6.B.6 The q-Density of $\widetilde{\boldsymbol{u}}_k$

For the q-density derivation given here, we have the induced product restriction

$$\mathfrak{q}\left(\widetilde{\boldsymbol{u}}_1, \ldots, \widetilde{\boldsymbol{u}}_{d_\bullet}\right) = \mathfrak{q}\left(\widetilde{\boldsymbol{u}}_1\right)\ldots\mathfrak{q}\left(\widetilde{\boldsymbol{u}}_{d_\bullet}\right).$$

Recalling the result for $\widetilde{\boldsymbol{u}}_k|$rest in (6.37), $1 \le k \le d_\bullet$, the q-density of $(\widetilde{\boldsymbol{u}}_1, \ldots, \widetilde{\boldsymbol{u}}_{d_\bullet})$ is the product of

$$N\left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}, \boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right)$$

density functions over $1 \leq k \leq d_\bullet$ where the q-density parameters are determined from

$$E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\widetilde{\boldsymbol{u}}_k\middle|\text{rest}\right)\right\}\right] = \begin{bmatrix} \widetilde{\boldsymbol{u}}_k \\ \\ \text{vec}\left(\widetilde{\boldsymbol{u}}_k\widetilde{\boldsymbol{u}}_k^T\right) \end{bmatrix}^T$$

$$\times E_{\mathfrak{q}}\left[\begin{array}{c} \left(\frac{\gamma_{uk}}{\sigma_\varepsilon^2}\right)\left\{\dot{\boldsymbol{Z}}_k^T\boldsymbol{y} - \dot{\boldsymbol{Z}}_k^T\begin{bmatrix} U_1\mathbf{1}_{n_1} \\ \vdots \\ U_m\mathbf{1}_{n_m}\end{bmatrix} - \left(\dot{\boldsymbol{Z}}_k^T\boldsymbol{X}\right)\left(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}\right) - \sum_{k\neq k'}^{d_\bullet}\left(\dot{\boldsymbol{Z}}_k^T\dot{\boldsymbol{Z}}_{k'}\right)\left(\gamma_{uk'}\widetilde{\boldsymbol{u}}_{k'}\right)\right\} \\ \\ -\frac{1}{2}\text{vec}\left\{\text{diag}\left(\frac{\gamma_{uk}\boldsymbol{\omega}_{\dot{\boldsymbol{Z}}_k}}{\sigma_\varepsilon^2} + \frac{b_{uk}\mathbf{1}_{L_k}}{\sigma_{uk}^2}\right)\right\} \end{array}\right].$$

Note that

$$-\frac{1}{2}\text{vec}\left(\text{diag}\left(E_{\mathfrak{q}}\left(\frac{\gamma_{uk}\boldsymbol{\omega}_{\dot{\boldsymbol{Z}}_k}}{\sigma_\varepsilon^2} + \frac{b_{uk}\mathbf{1}_{L_k}}{\sigma_{uk}^2}\right)\right)\right) = -\frac{1}{2}\text{vec}\left(\text{diag}\left(\frac{\mathbf{1}_{L_k}}{\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}}\right)\right).$$

Therefore

$$\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)} = \mathbf{1}_{L_k} \Big/ \left\{\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\mu_{\mathfrak{q}(\gamma_{uk})}\boldsymbol{\omega}_{\dot{\boldsymbol{Z}}_k} + \mu_{\mathfrak{q}\left(1/\sigma_{uk}^2\right)}\mu_{\mathfrak{q}(b_{uk})}\mathbf{1}_{L_k}\right\}.$$

If we define $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{k'})} \equiv \mu_{\mathfrak{q}(\gamma_{uk'})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_{k'})}$, then the other natural q-density parameter is

$$\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\left(\mu_{\mathfrak{q}(\gamma_{uk})}\left\{\dot{\boldsymbol{Z}}_k^T\boldsymbol{y} - \dot{\boldsymbol{Z}}_k^T\begin{bmatrix} \mu_{\mathfrak{q}(U_1)}\mathbf{1}_{n_1} \\ \vdots \\ \mu_{\mathfrak{q}(U_m)}\mathbf{1}_{n_m}\end{bmatrix} - \dot{\boldsymbol{Z}}_k^T\boldsymbol{X}\left(\boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{\gamma}_\beta\right)} \odot \boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\right)\right.\right.$$
$$\left.\left. - \sum_{k'\neq k}^{d_\bullet}\dot{\boldsymbol{Z}}_k^T\dot{\boldsymbol{Z}}_{k'}^T\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{k'})}\right\}\right) = \text{diag}\left(\frac{\mathbf{1}_{L_k}}{\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}}\right)\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}.$$

We define

$$\boldsymbol{\omega}_{16} \equiv \dot{\boldsymbol{Z}}_k^T \boldsymbol{y} - \dot{\boldsymbol{Z}}_k^T \begin{bmatrix} \mu_{\mathfrak{q}(U_1)} \mathbf{1}_{n_1} \\ \vdots \\ \mu_{\mathfrak{q}(U_m)} \mathbf{1}_{n_m} \end{bmatrix} - \dot{\boldsymbol{Z}}_k^T \boldsymbol{X} \left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \right) - \sum_{k' \neq k}^{d_\bullet} \dot{\boldsymbol{Z}}_k^T \dot{\boldsymbol{Z}}_{k'}^T \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{k'})}.$$

Then

$$\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)} = \operatorname{diag}\left(\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)} \right) \mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)} \mu_{\mathfrak{q}(\gamma_{uk})} \boldsymbol{\omega}_{16} = \mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)} \mu_{\mathfrak{q}(\gamma_{uk})} \left(\boldsymbol{\omega}_{16} \odot \boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)} \right).$$

6.B.7 The q-Density of γ_{β_k}

Recalling the expression of η_{β_k} in (6.38), for $1 \leq k \leq d_\circ + d_\bullet$, we have

$$\log \left\{ \mathfrak{q} \left(\gamma_{\beta_k} \right) \right\} = E_{\mathfrak{q}} \left[\log \left\{ \mathfrak{p} \left(\gamma_{\beta_k} \middle| \text{rest} \right) \right\} \right] + \text{const} = E_{\mathfrak{q}}(\eta_{\beta_k}) \gamma_{\beta_k} + \text{const}.$$

Hence,

$$\mathfrak{q} \left(\gamma_{\beta_k} \right) \text{ is the Bernoulli } \left(\frac{1}{1 + \exp\left(-E_{\mathfrak{q}\left(\eta_{\beta_k} \right)} \right)} \right) \text{ probability mass function.}$$

As in Appendix 4.A.4, if we define

$$\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{k'})} \equiv \mu_{\mathfrak{q}(\gamma_{uk'})} \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_{k'})}, \;\; 1 \leq k' \leq d_\bullet,$$

we have

$$\begin{aligned} E_{\mathfrak{q}} \left(\eta_{\beta_k} \right) = \text{logit}(\rho_\beta) - \frac{1}{2} \mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)} \Bigg\{ & \left(\mu^2_{\mathfrak{q}\left(\widetilde{\beta}_k \right)} + \sigma^2_{\mathfrak{q}\left(\widetilde{\beta}_k \right)} \right) \left\| \dot{\boldsymbol{X}}_k \right\|^2 \\ & - 2 \Bigg[\mu_{\mathfrak{q}\left(\widetilde{\beta}_k \right)} \Bigg(\dot{\boldsymbol{X}}_k^T \boldsymbol{y} - \dot{\boldsymbol{X}}_k^T \begin{bmatrix} \mu_{\mathfrak{q}(U_1)} \mathbf{1}_{n_1} \\ \vdots \\ \mu_{\mathfrak{q}(U_m)} \mathbf{1}_{n_m} \end{bmatrix} - \sum_{k'=1}^{d_\bullet} \left(\dot{\boldsymbol{Z}}_{k'}^T \dot{\boldsymbol{X}}_k \right)^T \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{k'})} \Bigg) \\ & - \left(\dot{\boldsymbol{X}}_{(-k)}^T \dot{\boldsymbol{X}}_k \right)^T E_{\mathfrak{q}} \Big\{ \widetilde{\beta}_k \left(\boldsymbol{\gamma}_\beta \right)_{-k} \odot \left(\widetilde{\boldsymbol{\beta}} \right)_{-k} \Big\} \Bigg] \Bigg\}. \end{aligned}$$

We define

$$\omega_{15} \equiv \dot{\boldsymbol{X}}_k^T \boldsymbol{y} - \dot{\boldsymbol{X}}_k^T \begin{bmatrix} \mu_{\mathfrak{q}(U_1)} \mathbf{1}_{n_1} \\ \vdots \\ \mu_{\mathfrak{q}(U_m)} \mathbf{1}_{n_m} \end{bmatrix} - \sum_{k'=1}^{d_\bullet} \left(\dot{\boldsymbol{Z}}_{k'}^T \dot{\boldsymbol{X}}_k \right)^T \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{k'})}$$

and note that

$$E_{\mathfrak{q}} \left\{ \widetilde{\beta}_k \left(\boldsymbol{\gamma}_\beta \right)_{-k} \odot \left(\widetilde{\boldsymbol{\beta}} \right)_{-k} \right\} = \left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \right)_{-k} \odot \left(\left(\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \right)_{-k,k} + \mu_{\mathfrak{q}(\widetilde{\beta}_k)} \left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \right)_{-k} \right).$$

Therefore

$$E_{\mathfrak{q}} \left(\eta_{\beta_k} \right) = \text{logit}(\rho_\beta) - \frac{1}{2} \mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)} \Bigg\{ \left(\mu^2_{\mathfrak{q}(\widetilde{\beta}_k)} + \sigma^2_{\mathfrak{q}(\widetilde{\beta}_k)} \right) \left\| \dot{\boldsymbol{X}}_k \right\|^2 - 2 \Bigg[\mu_{\mathfrak{q}(\widetilde{\beta}_k)} \omega_{15}$$

$$- \left(\dot{\boldsymbol{X}}_{(-k)}^T \dot{\boldsymbol{X}}_k \right)^T \left\{ \left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \right)_{-k} \odot \left(\left(\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \right)_{-k,k} + \mu_{\mathfrak{q}(\widetilde{\beta}_k)} \left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \right)_{-k} \right) \right\} \Bigg] \Bigg\}.$$

6.B.8 The q-Density of γ_{uk}

Recalling that η_{uk} in (6.39), for $1 \leq k \leq d_\bullet$, we have

$$\log \left\{ \mathfrak{q} \left(\gamma_{uk} \right) \right\} = E_{\mathfrak{q}} \left[\log \left\{ \mathfrak{p} \left(\gamma_{uk} \middle| \text{rest} \right) \right\} \right] + \text{const} = E_{\mathfrak{q}}(\eta_{uk}) \gamma_{uk} + \text{const}.$$

Hence,

$$\mathfrak{q} \left(\gamma_{uk} \right) \text{ is the Bernoulli } \left(\frac{1}{1 + \exp\left(-E_{\mathfrak{q}(\eta_{uk})} \right)} \right) \text{ probability mass function.}$$

As in Appendix 4.A.5, if we define

$$\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{k'})} \equiv \mu_{\mathfrak{q}(\gamma_{uk'})} \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_{k'})}, \;\; 1 \leq k', k \leq d_\bullet, \;\; k' \neq k,$$

we have

$$E_{\mathfrak{q}}\left(\eta_{uk}\right) = \text{logit}\left(\rho_u\right) - \frac{1}{2}\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)}\Bigg\{\boldsymbol{\omega}_{\dot{\boldsymbol{Z}}_k}^T\left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right) - 2\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}^T\Bigg[\dot{\boldsymbol{Z}}_k^T\boldsymbol{y}$$

$$-\dot{\boldsymbol{Z}}_k^T\begin{bmatrix}\mu_{\mathfrak{q}\left(U_1\mathbf{1}_{n_1}\right)}\\ \vdots \\ \boldsymbol{\mu}_{\mathfrak{q}(U_m\mathbf{1}_{n_m})}\end{bmatrix} - \dot{\boldsymbol{Z}}_k^T\boldsymbol{X}\left(\boldsymbol{\mu}_{\mathfrak{q}\left(\boldsymbol{\gamma}_\beta\right)} \odot \boldsymbol{\mu}_{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)}\right) - \sum_{k'\neq k}^{d_\bullet}\left(\dot{\boldsymbol{Z}}_k^T\dot{\boldsymbol{Z}}_{k'}\right)\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{k'})}\Bigg]\Bigg\}.$$

6.B.9 The q-Density of σ_ε^2

Noting the result for $\sigma_\varepsilon^2\big|$rest in (6.40), we have

$$\mathfrak{q}\left(\sigma_\varepsilon^2\right) \text{ has an Inverse-Gamma } \left(\kappa_{\mathfrak{q}(\sigma_\varepsilon^2)}, \lambda_{\mathfrak{q}(\sigma_\varepsilon^2)}\right) \text{density function,}$$

where

$$\kappa_{\mathfrak{q}(\sigma_\varepsilon^2)} = \frac{N+1}{2}, \text{ with } N \equiv \sum_{i=1}^{m} n_i$$

and

$$\lambda_{\mathfrak{q}(\sigma_\varepsilon^2)} = \mu_{\mathfrak{q}(1/a_\varepsilon)} + \frac{1}{2}E_{\mathfrak{q}}\left\{\left\|\boldsymbol{y} - \mathbf{1}_N\beta_0 - \begin{bmatrix}U_1\mathbf{1}_{n_1}\\ \vdots \\ U_m\mathbf{1}_{n_m}\end{bmatrix} - \boldsymbol{X}(\boldsymbol{\gamma}_\beta \odot \widetilde{\boldsymbol{\beta}}) - \sum_{k=1}^{d_\bullet}\dot{\boldsymbol{Z}}_k(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)\right\|^2\right\}. \tag{6.91}$$

As in (4.43) obtained in Appendix 4.A.6, now note that, in (6.91), we have

$$E_{\mathfrak{q}}\left\{\left\|\boldsymbol{y}-\mathbf{1}_N\beta_0-\begin{bmatrix}U_1\mathbf{1}_{n_1}\\ \vdots \\ U_m\mathbf{1}_{n_m}\end{bmatrix}-\boldsymbol{X}(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}})-\sum_{k=1}^{d_\bullet}\dot{\boldsymbol{Z}}_k(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)\right\|^2\right\}$$

$$=\left\|\boldsymbol{y}-\mathbf{1}_N\beta_0-\begin{bmatrix}\mu_{\mathfrak{q}(U_1)}\mathbf{1}_{n_1}\\ \vdots \\ \mu_{\mathfrak{q}(U_m)}\mathbf{1}_{n_m}\end{bmatrix}-\boldsymbol{X}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}\odot\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right)-\sum_{k=1}^{d_\bullet}\dot{\boldsymbol{Z}}_k\left(\mu_{\mathfrak{q}(\gamma_{uk})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right)\right\|^2$$

$$+\text{tr}\left[\text{Cov}_{\mathfrak{q}}\left\{\boldsymbol{y}-\mathbf{1}_N\beta_0-\begin{bmatrix}U_1\mathbf{1}_{n_1}\\ \vdots \\ U_m\mathbf{1}_{n_m}\end{bmatrix}-\boldsymbol{X}\left(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}}\right)-\sum_{k=1}^{d_\bullet}\dot{\boldsymbol{Z}}_k\left(\gamma_{uk}\widetilde{\boldsymbol{u}}_k\right)\right\}\right]. \quad (6.92)$$

Recalling that the product density restriction as follows:

$$\mathfrak{q}\left(\beta_0,\boldsymbol{U},\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{u}}_1,\ldots,\widetilde{\boldsymbol{u}}_{d_\bullet}\right)=\mathfrak{q}\left(\beta_0\right)\mathfrak{q}\left(\boldsymbol{U}\right)\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)\mathfrak{q}\left(\widetilde{\boldsymbol{u}}_1\right)\ldots\mathfrak{q}\left(\widetilde{\boldsymbol{u}}_{d_\bullet}\right),$$

the second term of (6.92) becomes

$$\begin{aligned}
&\mathrm{tr}\left[\mathrm{Cov}_{\mathfrak{q}}\left\{\boldsymbol{y}-\mathbf{1}_N\beta_0-\mathrm{blockdiag}\left(\mathbf{1}_{n_i}\right)\begin{bmatrix}U_1\\ \vdots\\ U_m\end{bmatrix}-\boldsymbol{X}\left(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}}\right)-\sum_{k=1}^{d_\bullet}\dot{\boldsymbol{Z}}_k\left(\gamma_{uk}\widetilde{\boldsymbol{u}}_k\right)\right\}\right]\\
&=\mathrm{tr}\left\{\mathrm{Cov}_{\mathfrak{q}}\left(\mathbf{1}_N\beta_0\right)\right\}+\mathrm{tr}\left\{\mathrm{Cov}_{\mathfrak{q}}\left(\mathrm{blockdiag}\left(\mathbf{1}_{n_i}\right)\boldsymbol{U}\right)\right\}+\mathrm{tr}\left\{\mathrm{Cov}_{\mathfrak{q}}\left\{\boldsymbol{X}\left(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}}\right)\right\}\right\}\\
&\quad+\sum_{k=1}^{d_\bullet}\mathrm{tr}\left\{\mathrm{Cov}_{\mathfrak{q}}\left\{\dot{\boldsymbol{Z}}_k\left(\gamma_{uk}\widetilde{\boldsymbol{u}}_k\right)\right\}\right\}\\
&=\mathrm{tr}\left\{\mathbf{1}_N\mathrm{Cov}_{\mathfrak{q}}\left(\beta_0\right)\mathbf{1}_N^T\right\}+\mathrm{tr}\left\{\mathrm{blockdiag}\left(\mathbf{1}_{n_i}\right)\mathrm{Cov}_{\mathfrak{q}(\boldsymbol{U})}\mathrm{blockdiag}\left(\mathbf{1}_{n_i}^T\right)\right\}\\
&\quad+\mathrm{tr}\left\{\boldsymbol{X}\mathrm{Cov}_{\mathfrak{q}}\left(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}}\right)\boldsymbol{X}^T\right\}+\sum_{k=1}^{d_\bullet}\mathrm{tr}\left\{\dot{\boldsymbol{Z}}_k\mathrm{Cov}_{\mathfrak{q}}\left(\gamma_{uk}\widetilde{\boldsymbol{u}}_k\right)\dot{\boldsymbol{Z}}_k^T\right\}\\
&=N\sigma^2_{\mathfrak{q}(\beta_0)}+\sum_{i=1}^{m}\left(n_i\sigma^2_{\mathfrak{q}(U_i)}\right)+\mathrm{tr}\left\{\boldsymbol{X}^T\boldsymbol{X}\mathrm{Cov}_{\mathfrak{q}}\left(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}}\right)\right\}\\
&\quad+\sum_{k=1}^{d_\bullet}\mathrm{tr}\left\{\dot{\boldsymbol{Z}}_k^T\dot{\boldsymbol{Z}}_k\mathrm{Cov}_{\mathfrak{q}}\left(\gamma_{uk}\widetilde{\boldsymbol{u}}_k\right)\right\}. \qquad (6.93)
\end{aligned}$$

As in (4.45) obtained in Appendix 4.A.6, the third term of (6.93) becomes

$$\begin{aligned}
\mathrm{tr}\left\{\boldsymbol{X}^T\boldsymbol{X}\mathrm{Cov}_{\mathfrak{q}}\left(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}}\right)\right\}&=\mathrm{tr}\left\{\boldsymbol{X}^T\boldsymbol{X}\left(\boldsymbol{\Omega}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}\odot\left\{\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}+\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\boldsymbol{\mu}^T_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right\}\right)\right\}\\
&\quad-\mathrm{tr}\left\{\boldsymbol{X}^T\boldsymbol{X}\left\{\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}\odot\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right\}\left\{\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}\odot\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right\}^T\right\}. \qquad (6.94)
\end{aligned}$$

As in (4.50) obtained in Appendix 4.A.6, the fourth term of (6.93) becomes

$$\mathrm{tr}\left\{\dot{\boldsymbol{Z}}_k^T\dot{\boldsymbol{Z}}_k\mathrm{Cov}_{\mathfrak{q}}\left(\gamma_{uk}\widetilde{\boldsymbol{u}}_k\right)\right\}=\boldsymbol{\omega}^T_{\dot{\boldsymbol{Z}}_k}\left[\mu_{\mathfrak{q}(\gamma_{uk})}\left\{\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}+\left(1-\mu_{\mathfrak{q}(\gamma_{uk})}\right)\left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\odot\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right)\right\}\right]. \qquad (6.95)$$

In summary, for Gaussian responses, (6.91) becomes

$$\lambda_{\mathfrak{q}(\sigma_\varepsilon^2)} = \mu_{\mathfrak{q}(1/a_\varepsilon)} + \frac{1}{2}\left\| \boldsymbol{y} - \mathbf{1}_N\beta_0 - \begin{bmatrix} \mu_{\mathfrak{q}(U_1)}\mathbf{1}_{n_1} \\ \vdots \\ \mu_{\mathfrak{q}(U_m)}\mathbf{1}_{n_m} \end{bmatrix} - \boldsymbol{X}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right) \right.$$
$$\left. - \sum_{k=1}^{d_\bullet} \dot{\boldsymbol{Z}}_k \left(\mu_{\mathfrak{q}(\gamma_{uk})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right) \right\|^2 + \frac{1}{2}N\sigma^2_{\mathfrak{q}(\beta_0)} + \frac{1}{2}\sum_{i=1}^{m}\left(n_i\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)} + \mu_{\mathfrak{q}\left(1/\sigma_U^2\right)}\right)$$
$$+\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{X}^T\boldsymbol{X}\left(\boldsymbol{\Omega}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \left\{\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} + \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\boldsymbol{\mu}^T_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right\}\right)\right\}$$
$$-\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{X}^T\boldsymbol{X}\left\{\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right\}\left\{\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right\}^T\right\}$$
$$+\frac{1}{2}\sum_{k=1}^{d_\bullet}\boldsymbol{\omega}^T_{\dot{\boldsymbol{Z}}_k}\left[\mu_{\mathfrak{q}(\gamma_{uk})}\left\{\boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)} + \left(1-\mu_{\mathfrak{q}(\gamma_{uk})}\right)\left(\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right)\right\}\right].$$

Then

$$\mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)} = \kappa_{\mathfrak{q}(\sigma_\varepsilon^2)}/\lambda_{\mathfrak{q}(\sigma_\varepsilon^2)}.$$

If we define

$$\boldsymbol{\omega}_{20} \equiv \mu_{\mathfrak{q}(\beta_0)}\mathbf{1}_N + \begin{bmatrix} \mu_{\mathfrak{q}(U_1)}\mathbf{1}_{n_1} \\ \vdots \\ \mu_{\mathfrak{q}(U_m)}\mathbf{1}_{n_m} \end{bmatrix} + \boldsymbol{X}(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}) + \sum_{k=1}^{d_\bullet}\dot{\boldsymbol{Z}}_k(\mu_{\mathfrak{q}(\gamma_{uk})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}),$$

we have, for $1 \le i \le m, 1 \le j \le n_i$,

$$(\boldsymbol{\omega}_{20})_{ij} = \mu_{\mathfrak{q}(\beta_0)} + \mu_{\mathfrak{q}(U_i)} + \left[\boldsymbol{X}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right) + \sum_{k=1}^{d_\bullet}\dot{\boldsymbol{Z}}_k\left(\mu_{\mathfrak{q}(\gamma_{uk})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right)\right]_{ij}.$$

For binary responses, we define $\boldsymbol{c}$ and $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{c})}$ both as $N \times 1$ vector. Recalling that the full conditional distribution of the auxiliary parameter c_{ij} in (6.42), we have

$$\text{the q-density of } c_{ij} \text{ is the } \begin{cases} \text{Truncated-Normal}_+(\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{c})}\right)_{ij}, 1), & y_{ij} = 1; \\ \text{Truncated-Normal}_-(\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{c})}\right)_{ij}, 1), & y_{ij} = 0 \end{cases}$$

where

$$\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{c})}\right)_{ij} = (\boldsymbol{\omega}_{20})_{ij} + \frac{\phi(\boldsymbol{\omega}_{20})_{ij}}{\Phi(\boldsymbol{\omega}_{20})_{ij}}, \quad y_{ij} = 1,$$

$$\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{c})}\right)_{ij} = (\boldsymbol{\omega}_{20})_{ij} - \frac{\phi(-(\boldsymbol{\omega}_{20})_{ij})}{\Phi(-(\boldsymbol{\omega}_{20})_{ij})}, \quad y_{ij} = 0,$$

using **Result 1.7.1** and **Result 1.7.2**.

In summary, for $1 \leq i \leq m, \;\; 1 \leq j \leq n_i, \;\; y_{ij} = 1$ or $y_{ij} = 0$, if define $\zeta'(x) \equiv \log\{2\Phi(x)\}$, then

$$\begin{aligned}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{c})}\right)_{ij} &= (\boldsymbol{\omega}_{20})_{ij} + (2y_{ij}-1)\,\frac{\phi\left((2y_{ij}-1)(\boldsymbol{\omega}_{20})_{ij}\right)}{\Phi\left((2y_{ij}-1)(\boldsymbol{\omega}_{20})_{ij}\right)} \\ &= (\boldsymbol{\omega}_{20})_{ij} + (2y_{ij}-1)\zeta'((2y_{ij}-1)(\boldsymbol{\omega}_{20})_{ij}),\end{aligned}$$

which is

$$\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{c})} = \boldsymbol{\omega}_{20} + (2\boldsymbol{y} - \mathbf{1}_N) \odot \zeta'\left((2\boldsymbol{y} - \mathbf{1}_N) \odot \boldsymbol{\omega}_{20}\right).$$

6.B.10 The q-Density of a_ε

As in Appendix 4.A.7,

$$\mathfrak{q}(a_\varepsilon) \text{ has an Inverse-Gamma } \left(\kappa_{\mathfrak{q}(a_\varepsilon)}, \lambda_{\mathfrak{q}(a_\varepsilon)}\right) \text{ density function,}$$

where

$$\kappa_{\mathfrak{q}(a_\varepsilon)} = 1, \;\; \lambda_{\mathfrak{q}(a_\varepsilon)} = \mu_{\mathfrak{q}(1/\sigma_\varepsilon^2)} + 1/s_\varepsilon^2 \;\text{ and }\; \mu_{\mathfrak{q}(1/a_\varepsilon)} = \kappa_{\mathfrak{q}(a_\varepsilon)}/\lambda_{\mathfrak{q}(a_\varepsilon)}.$$

6.B.11 The q-Density of b_{β_k}

As in Appendix 4.A.8, we have

$$\mathfrak{q}\left(b_{\beta_1}, \ldots, b_{\beta_{(d_\circ + d_\bullet)}}\right)$$

is a product of

$$\text{Inverse-Gaussian}\left(\left\{\mu_{\mathfrak{q}\left(1/\sigma_\beta^2\right)}\left(\mu^2_{\mathfrak{q}\left(\widetilde{\beta}_k\right)} + \sigma^2_{\mathfrak{q}\left(\widetilde{\beta}_k\right)}\right)\right\}^{-1/2}, 1\right) \text{ density functions}$$

over $1 \leq k \leq d_\circ + d_\bullet$. If define $\boldsymbol{\omega}_{14} \equiv \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} + \text{diagnal}\left(\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right)$, then

$$\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{b}_\beta)} = \left[\mu_{\mathfrak{q}\left(1/\sigma_\beta^2\right)}\left\{\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} + \text{diagnal}\left(\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right)\right\}\right]^{-1/2} = \left(\mu_{\mathfrak{q}\left(1/\sigma_\beta^2\right)}\boldsymbol{\omega}_{14}\right)^{-1/2}.$$

6.B.12 The $\mathfrak{q}$-Density of b_{u_k}

As in Appendix 4.A.9, we have

$$\mathfrak{q}\left(b_{u_1}, \ldots, b_{u_{(d_\bullet)}}\right)$$

is a product of

$$\text{Inverse-Gaussian}\left(\left\{\mu_{\mathfrak{q}\left(1/\sigma_{uk}^2\right)}\left[\left\|\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right\|^2 + \mathbf{1}_{L_k}^T\boldsymbol{\sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}^2\right]\right\}^{-1/2}, 1\right) \text{ density functions}$$

over $1 \leq k \leq d_\bullet$. If define $\omega_{17} \equiv \left\|\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right\|^2 + \mathbf{1}_{L_k}^T\boldsymbol{\sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}^2$ then

$$\mu_{\mathfrak{q}(b_{uk})} = E_{\mathfrak{q}}\left(\frac{\sigma_{uk}}{\left\|\widetilde{\boldsymbol{u}}_k\right\|}\right) = \left(\mu_{\mathfrak{q}\left(1/\sigma_{uk}^2\right)}\omega_{17}\right)^{-1/2}.$$

6.B.13 The $\mathfrak{q}$-Density of σ_β^2

As in Appendix 4.A.10, we have

$$\mathfrak{q}\left(\sigma_\beta^2\right) \text{ has an Inverse-Gamma}\left(\kappa_{\mathfrak{q}\left(\sigma_\beta^2\right)}, \lambda_{\mathfrak{q}\left(\sigma_\beta^2\right)}\right) \text{ density function}$$

where

$$\kappa_{\mathfrak{q}\left(\sigma_\beta^2\right)} = \frac{d_\circ + d_\bullet + 1}{2}$$

and

$$\lambda_{\mathfrak{q}\left(\sigma_\beta^2\right)} = \frac{1}{2}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{b}_\beta)}^T\left\{\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} + \text{diagonal}\left(\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right)\right\} + \mu_{\mathfrak{q}\left(1/a_\beta\right)}.$$

Recalling that

$$\boldsymbol{\omega}_{14} \equiv \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} \odot \boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})} + \text{diagnal}\left(\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right)$$

in Appendix 6.B.11, then

$$\lambda_{\mathfrak{q}\left(\sigma_\beta^2\right)} = \mu_{\mathfrak{q}\left(1/a_\beta\right)} + \frac{1}{2}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{b}_\beta)}^T\boldsymbol{\omega}_{14} \quad \text{and} \quad \mu_{\mathfrak{q}(1/\sigma_\beta^2)} = \kappa_{\mathfrak{q}\left(\sigma_\beta^2\right)}/\lambda_{\mathfrak{q}\left(\sigma_\beta^2\right)}.$$

6.B.14 The $\mathfrak{q}$-Density of a_β

As in Appendix 4.A.11,

$$\mathfrak{q}(a_\beta) \text{ has an Inverse-Gamma } \left(\kappa_{\mathfrak{q}(a_\beta)}, \lambda_{\mathfrak{q}(a_\beta)}\right) \text{ density function}$$

where

$$\kappa_{\mathfrak{q}(a_\beta)} = 1, \;\; \lambda_{\mathfrak{q}(a_\beta)} = \mu_{\mathfrak{q}\left(1/\sigma_\beta^2\right)} + 1/s_\beta^2 \;\; \text{and} \;\; \mu_{\mathfrak{q}(1/a_\beta)} = \kappa_{\mathfrak{q}(a_\beta)}/\lambda_{\mathfrak{q}(a_\beta)}.$$

6.B.15 The $\mathfrak{q}$-Density of σ_{uk}^2

As in Appendix 4.A.12, we have

$$\mathfrak{q}\left(\sigma_{uk}^2\right) \text{ is the Inverse-Gamma } \left(\kappa_{\mathfrak{q}\left(\sigma_{uk}^2\right)}, \lambda_{\mathfrak{q}\left(\sigma_{uk}^2\right)}\right) \text{ density function, } 1 \leq k \leq d_\bullet,$$

where

$$\kappa_{\mathfrak{q}\left(\sigma_{uk}^2\right)} = \frac{L_k + 1}{2} \;\; \text{and} \;\; \lambda_{\mathfrak{q}\left(\sigma_{uk}^2\right)} = \frac{1}{2}\mu_{\mathfrak{q}(b_{uk})}\left(\left\|\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right\|^2 + \mathbf{1}_{L_k}^T \boldsymbol{\sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}^2\right) + \mu_{\mathfrak{q}(1/a_{uk})}.$$

Recalling that $\omega_{17} \equiv \left\|\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right\|^2 + \mathbf{1}_{L_k}^T \boldsymbol{\sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}^2$ in Appendix 6.B.12, we have

$$\lambda_{\mathfrak{q}\left(\sigma_{uk}^2\right)} = \mu_{\mathfrak{q}(1/a_{uk})} + \frac{1}{2}\mu_{\mathfrak{q}(b_{uk})}\omega_{17} \;\; \text{and} \;\; \mu_{\mathfrak{q}\left(1/\sigma_{uk}^2\right)} = \kappa_{\mathfrak{q}\left(\sigma_{uk}^2\right)}/\lambda_{\mathfrak{q}\left(\sigma_{uk}^2\right)}.$$

6.B.16 The $\mathfrak{q}$-Density of a_{uk}

As in Appendix 4.A.13, we have

$$\mathfrak{q}(a_{uk}) \text{ is the Inverse-Gamma } \left(\kappa_{\mathfrak{q}(a_{uk})}, \lambda_{\mathfrak{q}(a_{uk})}\right) \text{ density function, } 1 \leq k \leq d_\bullet$$

where

$$\kappa_{\mathfrak{q}(a_{uk})} = 1, \;\; \lambda_{\mathfrak{q}(a_{uk})} = \mu_{\mathfrak{q}\left(1/\sigma_{uk}^2\right)} + 1/s_u^2 \;\; \text{and} \;\; \mu_{\mathfrak{q}(1/a_{uk})} = \kappa_{\mathfrak{q}(a_{uk})}/\lambda_{\mathfrak{q}(a_{uk})}.$$

6.C Marginal Log-Likelihood Expressions

A final aspect of Algorithm 10 is determination of good stopping criteria for the coordinate ascent scheme. As is common in the mean field variational Bayes literature we monitor relative increases in the approximate marginal log-likelihood, also known as the evidence lower bound, which is denote by $\log \underline{p}\left(\boldsymbol{y}; \mathfrak{q}\right)$. Appendix 6.C contains an explicit expression for the approximate marginal log-likelihood for the Section 6.4 models under product restrictions (6.51).

6.C.1 Marginal Log-Likelihood Approximation

For the generalized additive mixed model with random intercept and Gaussian responses, the mean field variational Bayes approximate log-likelihood, similar to (4.52) for generalized additive models, becomes

$$\begin{aligned}\log \underline{p}\left(\boldsymbol{y}; \mathfrak{q}\right) = \log \underline{p}\left(\boldsymbol{y}; \mathfrak{q}, \mathrm{C}\right) &+ E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{y}\middle|\beta_0, \boldsymbol{U}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{u}}, \boldsymbol{\gamma}_u, \sigma_\varepsilon^2\right)\right\}\right] \\ &+ E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\sigma_\varepsilon^2\middle|a_\varepsilon\right)\right\}\right] - E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\sigma_\varepsilon^2\right)\right\}\right] \\ &+ E_{\mathfrak{q}}\left[\log \mathfrak{p}\left\{\left(a_\varepsilon\right)\right\}\right] - E_{\mathfrak{q}}\left[\log \mathfrak{q}\left\{\left(a_\varepsilon\right)\right\}\right]. \end{aligned} \tag{6.96}$$

For the binary responses, the mean field variational Bayes approximate log-likelihood, similar to (4.53) for generalized additive models, becomes

$$\begin{aligned}\log \underline{p}\left(\boldsymbol{y}; \mathfrak{q}\right) = \log \underline{p}\left(\boldsymbol{y}; \mathfrak{q}, \mathrm{C}\right) &+ E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{y}\middle|\boldsymbol{c}\right)\right\}\right] \\ &+ E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{c}\middle|\beta_0, \boldsymbol{U}, \boldsymbol{\gamma}_\beta, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_u, \widetilde{\boldsymbol{u}}\right)\right\}\right] - E_{\mathfrak{q}}\Big[\log\big\{\mathfrak{q}\left(\boldsymbol{c}\right)\big\}\Big]. \end{aligned} \tag{6.97}$$

The expression of $\log \underline{p}(\boldsymbol{y}; \mathfrak{q}, \mathrm{C})$, similar to (4.54) for generalized additive model, becomes

$$\begin{aligned}
\log \underline{p}(\boldsymbol{y}; \mathfrak{q}, \mathrm{C}) = {} & E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\beta_0\right)\right\}\right] - E_{\mathfrak{q}}\left[\left\{\mathfrak{q}\left(\beta_0\right)\right\}\right] + E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{\gamma}_{\beta}\right)\right\}\right] - E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\boldsymbol{\gamma}_{\beta}\right)\right\}\right] \\
& + E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\widetilde{\boldsymbol{\beta}} \middle| \boldsymbol{b}_{\beta}, \sigma_{\beta}^2\right)\right\}\right] - E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)\right\}\right] + E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{b}_{\beta}\right)\right\}\right] - E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\boldsymbol{b}_{\beta}\right)\right\}\right] \\
& + E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\sigma_{\beta}^2 \middle| a_{\beta}\right)\right\}\right] - E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\sigma_{\beta}^2\right)\right\}\right] + E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(a_{\beta}\right)\right\}\right] - E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(a_{\beta}\right)\right\}\right] \\
& + E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{\gamma}_{u}\right)\right\}\right] - E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\boldsymbol{\gamma}_{u}\right)\right\}\right] + E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\widetilde{\boldsymbol{u}} \middle| \boldsymbol{b}_{u}, \boldsymbol{\sigma}_{u}^2\right)\right\}\right] - E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\widetilde{\boldsymbol{u}}\right)\right\}\right] \\
& + E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{b}_{u}\right)\right\}\right] - E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\boldsymbol{b}_{u}\right)\right\}\right] + E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{\sigma}_{u}^2 \middle| \boldsymbol{a}_{u}\right)\right\}\right] - E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\boldsymbol{\sigma}_{u}^2\right)\right\}\right] \\
& + E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{a}_{u}\right)\right\}\right] - E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\boldsymbol{a}_{u}\right)\right\}\right] \\
& + E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{U} \middle| \sigma_{U}^2\right)\right\}\right] - E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\boldsymbol{U}\right)\right\}\right] \\
& + E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\sigma_{U}^2 \middle| a_{U}\right)\right\}\right] - E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\sigma_{U}^2\right)\right\}\right] \\
& + E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(a_{U}\right)\right\}\right] - E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(a_{U}\right)\right\}\right]. \qquad (6.98)
\end{aligned}$$

Here "C" signifies the fact that (6.98) is common to $\log \underline{p}(\boldsymbol{y}; \mathfrak{q})$ expressions for both Gaussian response case and binary response case.

Explicit expressions for $\log \underline{p}(\boldsymbol{y}; \mathfrak{q})$ in each case can be obtained by simplifying each of the $\mathfrak{q}$-density moment expressions. For example, the first term of (6.98) is

$$\begin{aligned}
E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\beta_0\right)\right\}\right] &= -\frac{1}{2}\log\left(2\pi\right) - \frac{1}{2}\log\left(\sigma_{\beta_0}^2\right) - \frac{1}{2}E_{\mathfrak{q}}\left(\beta_0^2\right) \Big/ \sigma_{\beta_0}^2 \\
&= -\frac{1}{2}\log\left(2\pi\right) - \frac{1}{2}\log\left(\sigma_{\beta_0}^2\right) - \frac{1}{2}\left\{\mu_{\mathfrak{q}(\beta_0)}^2 + \sigma_{\mathfrak{q}(\beta_0)}^2\right\} \Big/ \sigma_{\beta_0}^2.
\end{aligned}$$

The second term of (6.98) is

$$\begin{aligned}
-E_{\mathfrak{p}}\left[\log\left\{\mathfrak{q}\left(\beta_0\right)\right\}\right] &= \frac{1}{2}\log\left(2\pi\right) + \frac{1}{2}\log\left(\sigma_{\mathfrak{q}(\beta_0)}^2\right) + \frac{1}{2}E_{\mathfrak{q}}\left(\beta_0 - \mu_{\mathfrak{q}(\beta_0)}\right)^2 \Big/ \sigma_{\beta_0}^2 \\
&= \frac{1}{2}\left\{\log\left(2\pi + 1\right)\right\} + \frac{1}{2}\log\left(\sigma_{\mathfrak{q}(\beta_0)}^2\right).
\end{aligned}$$

The third term of (6.98) is

$$E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{\gamma}_{\beta}\right)\right\}\right] = \sum_{k=1}^{d_{\circ}+d_{\bullet}} E_{\mathfrak{q}}\left\{\gamma_{\beta k}\log\left(\rho_{\beta}\right) + \left(1-\gamma_{\beta k}\right)\log\left(1-\rho_{\beta}\right)\right\}$$

$$= \operatorname{logit}\left(\rho_{\beta}\right)\sum_{k=1}^{d_{\circ}+d_{\bullet}} \mu_{\mathfrak{q}\left(\gamma_{\beta k}\right)} + \left(d_{\circ}+d_{\bullet}\right)\log\left(1-\rho_{\beta}\right).$$

The fourth term of (6.98) is

$$-E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\boldsymbol{\gamma}_{\beta}\right)\right\}\right] = -\sum_{k=1}^{d_{\circ}+d_{\bullet}} E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\gamma_{\beta k}\right)\right\}\right]$$

$$= -\sum_{k=1}^{d_{\circ}+d_{\bullet}} E_{\mathfrak{q}}\left[\log\left\{\left(\mu_{\mathfrak{q}\left(\gamma_{\beta k}\right)}\right)^{\gamma_{\beta k}}\left(1-\mu_{\mathfrak{q}\left(\gamma_{\beta k}\right)}\right)^{1-\gamma_{\beta k}}\right\}\right]$$

$$= -\sum_{k=1}^{d_{\circ}+d_{\bullet}} \left[\mu_{\mathfrak{q}\left(\gamma_{\beta k}\right)}\log\left(\mu_{\mathfrak{q}\left(\gamma_{\beta k}\right)}\right) + \left(1-\mu_{\mathfrak{q}\left(\gamma_{\beta k}\right)}\right)\log\left(1-\mu_{\mathfrak{q}\left(\gamma_{\beta k}\right)}\right)\right].$$

The fifth term of (6.98) is

$$E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\widetilde{\boldsymbol{\beta}}\middle|\boldsymbol{b}_{\beta},\sigma_{\beta}^{2}\right)\right\}\right] = \sum_{k=1}^{d_{\circ}+d_{\bullet}} E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\widetilde{\beta}_{k}\middle|b_{\beta k},\sigma_{\beta}^{2}\right)\right\}\right]$$

$$= \sum_{k=1}^{d_{\circ}+d_{\bullet}} \left[-\frac{1}{2}\log\left(2\pi\right) - \frac{1}{2}E_{\mathfrak{q}}\left\{\log\left(\sigma_{\beta}^{2}\right)\right\} + \frac{1}{2}E_{\mathfrak{q}}\left\{\log\left(b_{\beta k}\right)\right\} - \frac{1}{2}\mu_{\mathfrak{q}\left(1/\sigma_{\beta}^{2}\right)}E_{\mathfrak{q}}\left(b_{\beta k}\widetilde{\beta}_{k}^{2}\right)\right]$$

$$= -\frac{1}{2}\left(d_{\circ}+d_{\bullet}\right)\log\left(2\pi\right) - \frac{1}{2}\left(d_{\circ}+d_{\bullet}\right)E_{\mathfrak{q}}\left\{\log\left(\sigma_{\beta}^{2}\right)\right\} + \frac{1}{2}\sum_{k=1}^{d_{\circ}+d_{\bullet}} E_{\mathfrak{q}}\left\{\log\left(b_{\beta k}\right)\right\}$$

$$-\frac{1}{2}\mu_{\mathfrak{q}\left(1/\sigma_{\beta}^{2}\right)}\sum_{k=1}^{d_{\circ}+d_{\bullet}} \mu_{\mathfrak{q}\left(b_{\beta k}\right)}\left(\mu_{\mathfrak{q}\left(\widetilde{\beta}_{k}\right)}^{2} + \sigma_{\mathfrak{q}\left(\widetilde{\beta}_{k}\right)}^{2}\right).$$

The sixth term of (6.98) is

$$
\begin{aligned}
-E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\widetilde{\boldsymbol{\beta}}\right)\right\}\right] &= \frac{1}{2}\log\left|2\pi\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right| + \frac{1}{2}E_{\mathfrak{q}}\left\{\left(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right)^{T}\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}^{-1}\left(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right)\right\} \\
&= \frac{1}{2}\log\left\{(2\pi)^{d_{\circ}+d_{\bullet}}\left|\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right|\right\} + \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}^{-1}E_{\mathfrak{q}}\left\{\left(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right)\left(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}^{T}\right)\right\}\right] \\
&= \frac{1}{2}\log\left\{(2\pi)^{d_{\circ}+d_{\bullet}}\left|\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right|\right\} + \frac{1}{2}\mathrm{tr}\left(\boldsymbol{I}_{d_{\circ}+d_{\bullet}}\right) \\
&= \frac{1}{2}\log\left|\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right| + \frac{1}{2}\left\{(d_{\circ}+d_{\bullet})\log(2\pi) + (d_{\circ}+d_{\bullet})\right\}.
\end{aligned}
$$

Using **Result 1.6.2**, the seventh term of (6.98) is

$$
\begin{aligned}
E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{b}_{\beta}\right)\right\}\right] &= \sum_{k=1}^{d_{\circ}+d_{\bullet}}\left[\log\left(\frac{1}{2}\right) - 2E_{\mathfrak{q}}\left\{\log\left(b_{\beta k}\right)\right\} - \frac{1}{2}E_{\mathfrak{q}}\left(1\Big/b_{\beta k}\right)\right] \\
&= -\left(d_{\circ}+d_{\bullet}\right)\log(2) - 2\sum_{k=1}^{d_{\circ}+d_{\bullet}}E_{\mathfrak{q}}\left\{\log\left(b_{\beta k}\right)\right\} - \frac{1}{2}\sum_{k=1}^{d_{\circ}+d_{\bullet}}\left[\left\{1\Big/\mu_{\mathfrak{q}\left(b_{\beta k}\right)}\right\} + 1\right] \\
&= -\left(d_{\circ}+d_{\bullet}\right)\left\{\log(2) + \frac{1}{2}\right\} - \frac{1}{2}\sum_{k=1}^{d_{\circ}+d_{\bullet}}\left\{1\Big/\mu_{\mathfrak{q}\left(b_{\beta k}\right)}\right\} - 2\sum_{k=1}^{d_{\circ}+d_{\bullet}}E_{\mathfrak{q}}\left\{\log\left(b_{\beta k}\right)\right\}.
\end{aligned}
$$

The eighth term of (6.98) is

$$-E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\boldsymbol{b}_{\beta}\right)\right\}\right]=-\sum_{k=1}^{d_{\circ}+d_{\bullet}}\left[-\frac{1}{2}\log\left(2\pi\right)-\frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(b_{\beta k}\right)\right\}-E_{\mathfrak{q}}\left\{\frac{\left(b_{\beta k}-\mu_{\mathfrak{q}\left(b_{\beta k}\right)}\right)^{2}}{2b_{\beta k}\mu_{\mathfrak{q}\left(b_{\beta k}\right)}^{2}}\right\}\right]$$

$$\begin{aligned}&=-\sum_{k=1}^{d_{\circ}+d_{\bullet}}\left[-\frac{1}{2}\log\left(2\pi\right)-\frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(b_{\beta k}\right)\right\}-\frac{1}{2}E_{\mathfrak{q}}\left(b_{\beta k}\right)\Big/\mu_{\mathfrak{q}\left(b_{\beta k}\right)}^{2}\right.\\&\qquad\left.+1\Big/\mu_{\mathfrak{q}\left(b_{\beta k}\right)}-\frac{1}{2}E_{\mathfrak{q}}\left(1\Big/b_{\beta k}\right)\right]\\&=-\sum_{k=1}^{d_{\circ}+d_{\bullet}}\left[-\frac{1}{2}\log\left(2\pi\right)-\frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(b_{\beta k}\right)\right\}-1\Big/\left\{2\mu_{\mathfrak{q}\left(b_{\beta k}\right)}\right\}+1\Big/\mu_{\mathfrak{q}\left(b_{\beta k}\right)}\right.\\&\qquad\left.-\frac{1}{2}\left\{1\Big/\mu_{\mathfrak{q}\left(b_{\beta k}\right)}+1\right\}\right]\\&=-\sum_{k=1}^{d_{\circ}+d_{\bullet}}\left[-\frac{1}{2}\log\left(2\pi\right)-\frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(b_{\beta k}\right)\right\}-\frac{1}{2}\right]\\&=\frac{1}{2}\left(d_{\circ}+d_{\bullet}\right)\left\{\log\left(2\pi\right)+1\right\}+\frac{3}{2}\sum_{k=1}^{d_{\bullet}}E_{\mathfrak{q}}\log\left(b_{\beta k}\right).\end{aligned}$$

The ninth term of (6.98) is

$$\begin{aligned}E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\sigma_{\beta}^{2}|a_{\beta}\right)\right\}\right]&=E_{\mathfrak{q}}\left[\log\left\{\left(1/a_{\beta}\right)^{1/2}\Big/\Gamma\left(\frac{1}{2}\right)-\frac{3}{2}\log\left(\sigma_{\beta}^{2}\right)-1\Big/\left(\sigma_{\beta}^{2}a_{\beta}\right)\right\}\right]\\&=-\frac{1}{2}\log\left(\pi\right)-\frac{1}{2}E_{\mathfrak{q}}\left\{\log\left(a_{\beta}\right)\right\}-\frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(\sigma_{\beta}^{2}\right)\right\}-\mu_{\mathfrak{q}\left(1/a_{\beta}\right)}\mu_{\mathfrak{q}\left(1/\sigma_{\beta}^{2}\right)}.\end{aligned}$$

The tenth term of (6.98) is

$$\begin{aligned}-E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\sigma_{\beta}^{2}\right)\right\}\right]&=-E_{\mathfrak{q}}\left[\frac{1}{2}\left(d_{\circ}+d_{\bullet}+1\right)\log\left(\lambda_{\mathfrak{q}\left(\sigma_{\beta}^{2}\right)}\right)-\log\left\{\Gamma\left(\frac{1}{2}\left(d_{\circ}+d_{\bullet}+1\right)\right)\right\}\right.\\&\qquad\left.-\left\{\frac{1}{2}\left(d_{\circ}+d_{\bullet}+1\right)+1\right\}\log\left(\sigma_{\beta}^{2}\right)-\lambda_{\mathfrak{q}\left(\sigma_{\beta}^{2}\right)}\Big/\sigma_{\beta}^{2}\right]\\&=-\frac{1}{2}\left(d_{\circ}+d_{\bullet}+1\right)\log\left(\lambda_{\mathfrak{q}\left(\sigma_{\beta}^{2}\right)}\right)+\log\left\{\Gamma\left(\frac{1}{2}\left(d_{\circ}+d_{\bullet}+1\right)\right)\right\}\\&\qquad+\left\{\frac{1}{2}\left(d_{\circ}+d_{\bullet}+1\right)+1\right\}E_{\mathfrak{q}}\left\{\log\left(\sigma_{\beta}^{2}\right)\right\}+\mu_{\mathfrak{q}\left(1/\sigma_{\beta}^{2}\right)}\lambda_{\mathfrak{q}\left(\sigma_{\beta}^{2}\right)}.\end{aligned}$$

The eleventh term of (6.98) is

$$E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(a_\beta\right)\right\}\right] = E_{\mathfrak{q}}\left(\log\left[\left\{\left(1/s_\beta\right)\Big/\Gamma\left(\frac{1}{2}\right)a_\beta^{-3/2}\right\}\exp\left(1\Big/\left(s_\beta^2 a_\beta\right)\right)\right]\right)$$
$$= -\frac{1}{2}\log\left(\pi\right) - \log\left(s_\beta\right) - \frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(a_\beta\right)\right\} - \mu_{\mathfrak{q}\left(1/a_\beta\right)}\Big/s_\beta^2.$$

The twelfth term of (6.98) is

$$-E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(a_\beta\right)\right\}\right] = -E_{\mathfrak{q}}\left[\log\left\{\lambda_{\mathfrak{q}\left(a_\beta\right)}a_\beta^{-2}\exp\left(-\lambda_{\mathfrak{q}\left(a_\beta\right)}\Big/a_\beta\right)\right\}\right]$$
$$= -\log\left\{\lambda_{\mathfrak{q}\left(a_\beta\right)}\right\} + 2E_{\mathfrak{q}}\left\{\log\left(a_\beta\right)\right\} + \lambda_{\mathfrak{q}\left(a_\beta\right)}\mu_{\mathfrak{q}\left(1/a_\beta\right)}.$$

The thirteenth term of (6.98) is

$$E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{\gamma}_u\right)\right\}\right]$$
$$= \sum_{k=1}^{d_\bullet} E_{\mathfrak{q}}\left[\gamma_{uk}\left\{\log\left(\rho_u\right)\right\} + \left(1-\gamma_{uk}\right)\log\left(1-\rho_u\right)\right]$$
$$= \text{logit}\left(\rho_u\right)\sum_{k=1}^{d_\bullet}\mu_{\mathfrak{q}(\gamma_{uk})} + d_\bullet \text{log}\left(1-\rho_u\right).$$

The fourteenth term of (6.98) is

$$-E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\boldsymbol{\gamma}_u\right)\right\}\right] = -\sum_{k=1}^{d_\bullet} E_{\mathfrak{q}}\left[\log\left\{\left(\mu_{\mathfrak{q}\left(\gamma_{\beta k}\right)}\right)^{\gamma_{\beta k}}\left(1-\mu_{\mathfrak{q}\left(\gamma_{\beta k}\right)}\right)^{1-\gamma_{\beta k}}\right\}\right]$$
$$= -\sum_{k=1}^{d_\bullet}\left[\mu_{\mathfrak{q}\left(\gamma_{\beta k}\right)}\log\left(\mu_{\mathfrak{q}\left(\gamma_{\beta k}\right)}\right) + \left\{1-\mu_{\mathfrak{q}\left(\gamma_{\beta k}\right)}\right\}\log\left(1-\mu_{\mathfrak{q}\left(\gamma_{\beta k}\right)}\right)\right].$$

The fifteenth term of (6.98) is

$$E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\widetilde{\boldsymbol{u}}\middle|\boldsymbol{b}_u,\boldsymbol{\sigma}_u^2\right)\right\}\right]=\sum_{k=1}^{d_\bullet}\left[-\frac{1}{2}E_{\mathfrak{q}}\left(\log\left|\left(2\pi\right)\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right|\right)-\frac{1}{2}E_{\mathfrak{q}}\left\{\widetilde{\boldsymbol{u}}_k^T\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}^{-1}\widetilde{\boldsymbol{u}}_k\right\}\right]$$

$$=-\frac{1}{2}\log\left(2\pi\right)\left(\sum_{k=1}^{d_\bullet}L_k\right)-\frac{1}{2}\sum_{k=1}^{d_\bullet}L_kE_{\mathfrak{q}}\left\{\log\left(\sigma_{uk}^2\right)\right\}+\frac{1}{2}\sum_{k=1}^{d_\bullet}L_kE_{\mathfrak{q}}\left\{\log\left(b_{uk}\right)\right\}$$

$$-\frac{1}{2}\sum_{k=1}^{d_\bullet}\mu_{\mathfrak{q}\left(1/\sigma_{uk}^2\right)}\mu_{\mathfrak{q}(b_{uk})}\left(\left\|\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right\|^2+\mathbf{1}_{L_k}^T\boldsymbol{\sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}^2\right).$$

The sixteenth term of (6.98) is

$$-E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\widetilde{\boldsymbol{u}}\right)\right\}\right]$$

$$=\sum_{k=1}^{d_\bullet}\left[\frac{1}{2}L_k\log\left(2\pi\right)+\frac{1}{2}\sum_{l=1}^{L_k}\log\left(\sigma_{\mathfrak{q}(\widetilde{u}_{kl})}^2\right)+\frac{1}{2}E_{\mathfrak{q}}\left\{\frac{\left(\widetilde{\boldsymbol{u}}_k-\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right)\odot\left(\widetilde{\boldsymbol{u}}_k-\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right)}{\boldsymbol{\sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}^2}\right\}\right]$$

$$=\frac{1}{2}\left\{\log\left(2\pi\right)+1\right\}\sum_{k=1}^{d_\bullet}L_k+\frac{1}{2}\sum_{k=1}^{d_\bullet}\sum_{l=1}^{L_k}\log\left(\sigma_{\mathfrak{q}(\widetilde{u}_{kl})}^2\right).$$

The seventeenth term of (6.98) is

$$E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{b}_u\right)\right\}\right]=\sum_{k=1}^{d_\bullet}\left[\frac{1}{2}\left(L_k+1\right)\log\left(\frac{1}{2}\right)-\log\left\{\Gamma\left(\frac{1}{2}\left(L_k+1\right)\right)\right\}\right.$$

$$\left.-\left\{\frac{1}{2}\left(L_k+1\right)+1\right\}E_{\mathfrak{q}}\left\{\log(b_{uk})\right\}-\frac{1}{2}E_{\mathfrak{q}}\left(1/b_{uk}\right)\right]$$

$$=-\frac{1}{2}\left(d_\bullet+\sum_{k=1}^{d_\bullet}L_k\right)\log\left(2\right)-\sum_{k=1}^{d_\bullet}\log\left\{\Gamma\left(\frac{1}{2}\left(L_k+1\right)\right)\right\}$$

$$-\sum_{k=1}^{d_\bullet}\left\{\frac{1}{2}\left(L_k+1\right)+1\right\}E_{\mathfrak{q}}\left\{\log\left(b_{uk}\right)\right\}-\frac{1}{2}\sum_{k=1}^{d_\bullet}\left[\left\{1/\mu_{\mathfrak{q}(b_{uk})}\right\}+1\right]$$

$$=-\frac{1}{2}d_\bullet\left\{\log\left(2\right)+1\right\}-\frac{1}{2}\log\left(2\right)\sum_{k=1}^{d_\bullet}L_k-\sum_{k=1}^{d_\bullet}\log\left\{\Gamma\left(\frac{1}{2}\left(L_k+1\right)\right)\right\}$$

$$-\sum_{k=1}^{d_\bullet}\left\{\left(\frac{1}{2}\left(L_k+1\right)\right)+1\right\}E_{\mathfrak{q}}\left\{\log\left(b_{uk}\right)\right\}-\frac{1}{2}\sum_{k=1}^{d_\bullet}\left\{1/\mu_{\mathfrak{q}(b_{uk})}\right\}.$$

The eighteenth term of (6.98) is

$$-E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\boldsymbol{b}_u\right)\right\}\right] = -\sum_{k=1}^{d_\bullet}\left[-\frac{1}{2}\log\left(2\pi\right) - \frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(b_{uk}\right)\right\} - E_{\mathfrak{q}}\left\{\frac{\left(b_{uk}-\mu_{\mathfrak{q}(b_{uk})}\right)^2}{2b_{uk}\mu^2_{\mathfrak{q}(b_{uk})}}\right\}\right]$$

$$= -\sum_{k=1}^{d_\bullet}\left[-\frac{1}{2}\log\left(2\pi\right) - \frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(b_{uk}\right)\right\} - 1/\left\{2\mu_{\mathfrak{q}(b_{uk})}\right\} + 1/\mu_{\mathfrak{q}(b_{uk})}\right.$$

$$\left. -\frac{1}{2}\left\{1/\mu_{\mathfrak{q}(b_{uk})}+1\right\}\right]$$

$$= -\sum_{k=1}^{d_\bullet}\left[-\frac{1}{2}\log\left(2\pi\right) - \frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(b_{uk}\right)\right\} - \frac{1}{2}\right] = \frac{1}{2}d_\bullet\left\{\log\left(2\pi\right)+1\right\} + \frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(b_{uk}\right)\right\}.$$

The nineteenth term of (6.98) is

$$E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{\sigma}_u^2\middle|\boldsymbol{a}_u\right)\right\}\right]$$

$$= \sum_{k=1}^{d_\bullet}E_{\mathfrak{q}}\left[\log\left\{\left(1/a_{uk}\right)^{1/2}\Big/\Gamma\left(\frac{1}{2}\right)\right\} - \frac{3}{2}\log\left(\sigma^2_{uk}\right) - 1/\left(\sigma^2_{uk}a_{uk}\right)\right]$$

$$= -\frac{1}{2}d_\bullet\log\left(\pi\right) - \sum_{k=1}^{d_\bullet}\left[\frac{1}{2}E_{\mathfrak{q}}\left\{\log\left(a_{uk}\right)\right\} + \frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(\sigma^2_{uk}\right)\right\} - \mu_{\mathfrak{q}\left(1/a_{uk}\right)}\mu_{\mathfrak{q}\left(1/\sigma^2_{uk}\right)}\right].$$

The twentieth term of (6.98) is

$$-E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\boldsymbol{\sigma}_u^2\right)\right\}\right] = -\sum_{k=1}^{d_\bullet}E_{\mathfrak{q}}\left[\frac{1}{2}\left(L_k+1\right)\log\left(\lambda_{\mathfrak{q}\left(\sigma^2_{uk}\right)}\right) - \log\left\{\Gamma\left(\frac{1}{2}\left(L_k+1\right)\right)\right\}\right.$$

$$\left. -\left\{\frac{1}{2}\left(L_k+1\right)+1\right\}E_{\mathfrak{q}}\left\{\log\left(\sigma^2_{uk}\right)\right\} - \lambda_{\mathfrak{q}\left(\sigma^2_{uk}\right)}/\sigma^2_{uk}\right]$$

$$= -\frac{1}{2}\sum_{k=1}^{d_\bullet}\left(L_k+1\right)\log\left(\lambda_{\mathfrak{q}\left(\sigma^2_{uk}\right)}\right) + \sum_{k=1}^{d_\bullet}\log\left\{\Gamma\left(\frac{1}{2}\left(L_k+1\right)\right)\right\}$$

$$+\sum_{k=1}^{d_\bullet}\left[\left\{\frac{1}{2}\left(L_k+1\right)+1\right\}E_{\mathfrak{q}}\left\{\log\left(\sigma^2_{uk}\right)\right\} + \mu_{\mathfrak{q}\left(1/\sigma^2_{uk}\right)}\lambda_{\mathfrak{q}\left(\sigma^2_{uk}\right)}\right].$$

The twenty-first term of (6.98) is

$$E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{a}_u\right)\right\}\right]=\sum_{k=1}^{d_\bullet}E_{\mathfrak{q}}\left[\log\left\{\left(1/s_u\right)\Big/\Gamma\left(\frac{1}{2}\right)\right\}-\frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(a_{uk}\right)\right\}-E_{\mathfrak{q}}\left\{1/\left(a_{uk}s_u^2\right)\right\}\right]$$

$$=-\frac{1}{2}d_\bullet\log\left(\pi\right)-d_\bullet\log\left(s_u\right)-\frac{3}{2}\sum_{k=1}^{d_\bullet}E_{\mathfrak{q}}\left\{\log\left(a_{uk}\right)\right\}-\left(1/s_u^2\right)\sum_{k=1}^{d_\bullet}\mu_{\mathfrak{q}\left(1/a_{uk}\right)}.$$

The twenty-second term of (6.98) is

$$-E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\boldsymbol{a}_u\right)\right\}\right]=\sum_{k=1}^{d_\bullet}\left[-\log\left(\lambda_{\mathfrak{q}(a_{uk})}\right)+2E_{\mathfrak{q}}\left\{\log\left(a_{uk}\right)\right\}+\lambda_{\mathfrak{q}(a_{uk})}\mu_{\mathfrak{q}\left(1/a_{uk}\right)}\right].$$

The twenty-third term of (6.98) is

$$E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{U}\middle|\sigma_U^2\right)\right\}\right]=\sum_{i=1}^{m}E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(U_i|\sigma_U^2\right)\right\}\right]$$

$$=\sum_{i=1}^{m}\left[-\frac{1}{2}\log\left(2\pi\right)-\frac{1}{2}E_{\mathfrak{q}}\left\{\log\left(\sigma_U^2\right)\right\}-\frac{1}{2}\mu_{\mathfrak{q}\left(1/\sigma_U^2\right)}E_{\mathfrak{q}}\left(U_i^2\right)\right]$$

$$=-\frac{m}{2}\log\left(2\pi\right)-\frac{m}{2}E_{\mathfrak{q}}\left\{\log\left(\sigma_U^2\right)\right\}-\frac{1}{2}\mu_{\mathfrak{q}\left(1/\sigma_U^2\right)}\sum_{i=1}^{m}\left(\mu_{\mathfrak{q}(U_i)}^2+\sigma_{\mathfrak{q}(U_i)}^2\right).$$

The twenty-fourth term of (6.98) is

$$-E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\boldsymbol{U}\right)\right\}\right]=\sum_{i=1}^{m}\left(\frac{1}{2}\log\left(2\pi\right)+\frac{1}{2}\log\left(\sigma_{\mathfrak{q}(U_i)}^2\right)+\frac{1}{2}E_{\mathfrak{q}}\left[\left(U_i-\mu_{\mathfrak{q}(U_i)}\right)^2\Big/\sigma_{U_i}^2\right]\right)$$

$$=\frac{m}{2}\left\{\log\left(2\pi+1\right)\right\}+\frac{1}{2}\sum_{i=1}^{m}\left(\log\left(\sigma_{\mathfrak{q}(U_i)}^2\right)\right).$$

The twenty-fifth term of (6.98) is

$$E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\sigma_U^2\middle|a_U\right)\right\}\right]=E_{\mathfrak{q}}\left[\log\left\{\left(1/a_U\right)^{1/2}\Big/\Gamma\left(\frac{1}{2}\right)-\frac{3}{2}\log\left(\sigma_U^2\right)-1\Big/\left(\sigma_U^2a_U\right)\right\}\right]$$

$$=-\frac{1}{2}\log\left(\pi\right)-\frac{1}{2}E_{\mathfrak{q}}\left\{\log\left(a_U\right)\right\}-\frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(\sigma_U^2\right)\right\}-\mu_{\mathfrak{q}(1/a_U)}\mu_{\mathfrak{q}\left(1/\sigma_U^2\right)}.$$

The twenty-sixth term of (6.98) is

$$\begin{aligned}
&-E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\sigma_U^2\right)\right\}\right] \\
&\quad= -E_{\mathfrak{q}}\left[\frac{1}{2}(m+1)\log\left(\lambda_{\mathfrak{q}\left(\sigma_U^2\right)}\right) - \log\left\{\Gamma\left(\frac{1}{2}(m+1)\right)\right\}\right. \\
&\qquad\left. - \left\{\frac{1}{2}(m+1)+1\right\}\log\left(\sigma_U^2\right) - \lambda_{\mathfrak{q}\left(\sigma_U^2\right)}\Big/\sigma_U^2\right] \\
&\quad= -\frac{1}{2}(m+1)\log\left(\lambda_{\mathfrak{q}\left(\sigma_U^2\right)}\right) + \log\left\{\Gamma\left(\frac{1}{2}(m+1)\right)\right\} \\
&\qquad+ \left\{\frac{1}{2}(m+1)+1\right\}E_{\mathfrak{q}}\left\{\log\left(\sigma_U^2\right)\right\} + \mu_{\mathfrak{q}\left(1/\sigma_U^2\right)}\lambda_{\mathfrak{q}\left(\sigma_U^2\right)}.
\end{aligned}$$

The twenty-seventh term of (6.98) is

$$\begin{aligned}
&E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(a_U\right)\right\}\right] \\
&\quad= E_{\mathfrak{q}}\left(\log\left[\left\{\left(1/s_U\right)\Big/\Gamma\left(\frac{1}{2}\right)a_U^{-3/2}\right\}\exp\left(1\Big/\left(s_U^2 a_U\right)\right)\right]\right) \\
&\quad= -\frac{1}{2}\log(\pi) - \log\left(s_U\right) - \frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(a_U\right)\right\} - \mu_{\mathfrak{q}(1/a_U)}\Big/s_U^2.
\end{aligned}$$

The twenty-eighth term of (6.98) is

$$\begin{aligned}
&-E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(a_U\right)\right\}\right] \\
&\quad= -E_{\mathfrak{q}}\left[\log\left\{\lambda_{\mathfrak{q}(a_U)}a_U^{-2}\exp\left(-\lambda_{\mathfrak{q}(a_U)}\Big/a_U\right)\right\}\right] \\
&\quad= -\log\left\{\lambda_{\mathfrak{q}(a_U)}\right\} + 2E_{\mathfrak{q}}\left\{\log\left(a_U\right)\right\} + \lambda_{\mathfrak{q}(a_U)}\mu_{\mathfrak{q}(1/a_U)}.
\end{aligned}$$

6.C.2 Collection of Logarithmic Moments of Variance Parameters

In this subsection we collect the contributions to $\log \underline{p}\left(\boldsymbol{y}; \mathfrak{q}, \mathrm{C}\right)$ of each of the logarithmic $\mathfrak{q}$-density moments of variance parameters.

The $E_{\mathfrak{q}}\left\{\log\left(\sigma_\beta^2\right)\right\}$ Terms

The fifth term of (6.98) contains: $-\frac{1}{2}\left(d_\circ + d_\bullet\right)E_{\mathfrak{q}}\left\{\log\left(\sigma_\beta^2\right)\right\}$.

The ninth term of (6.98) contains: $-\frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(\sigma_{\beta}^{2}\right)\right\}$.

The tenth term of (6.98) contains: $\left\{\frac{1}{2}\left(d_{\circ}+d_{\bullet}+1\right)+1\right\}E_{\mathfrak{q}}\left\{\log\left(\sigma_{\beta}^{2}\right)\right\}$.

Since

$$-\frac{1}{2}\left(d_{\circ}+d_{\bullet}\right)-\frac{3}{2}+\frac{1}{2}\left(d_{\circ}+d_{\bullet}+1\right)+1=0,$$

the $E_{\mathfrak{q}}\left\{\log\left(\sigma_{\beta}^{2}\right)\right\}$ contribution to $\log\underline{p}\left(\boldsymbol{y};\mathfrak{q},\mathrm{C}\right)$ is zero.

The $E_{\mathfrak{q}}\left\{\log\left(\sigma_{uk}^{2}\right)\right\}$ Terms, $1\leq k\leq d_{\bullet}$

The fifteenth term of (6.98) contains: $-\frac{1}{2}\sum_{k=1}^{d_{\bullet}}L_{k}E_{\mathfrak{q}}\left\{\log\left(\sigma_{uk}^{2}\right)\right\}$.

The nineteenth term of (6.98) contains: $-\frac{3}{2}\sum_{k=1}^{d_{\bullet}}E_{\mathfrak{q}}\left\{\log\left(\sigma_{uk}^{2}\right)\right\}$.

The twentieth term of (6.98) contains: $\sum_{k=1}^{d_{\bullet}}\left\{\frac{1}{2}\left(L_{k}+1\right)+1\right\}E_{\mathfrak{q}}\left\{\log\left(\sigma_{uk}^{2}\right)\right\}$.

Since

$$-\frac{1}{2}L_{k}-\frac{3}{2}+\frac{1}{2}\left(L_{k}+1\right)+1=0,$$

the $E_{\mathfrak{q}}\left\{\log\left(\sigma_{uk}^{2}\right)\right\}$ contributions to $\log\underline{p}\left(\boldsymbol{y};\mathfrak{q},\mathrm{C}\right)$ are zero for each $1\leq k\leq d_{\bullet}$.

The $E_{\mathfrak{q}}\left\{\log\left(\sigma_{U}^{2}\right)\right\}$ Terms

The twenty-third term of (6.98) contains: $-\frac{m}{2}E_{\mathfrak{q}}\left\{\log\left(\sigma_{U}^{2}\right)\right\}$.

The twenty-fifth term of (6.98) contains: $-\frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(\sigma_{U}^{2}\right)\right\}$.

The twenty-sixth term of (6.98) contains: $\left(\frac{m+1}{2}+1\right)E_{\mathfrak{q}}\left\{\log\left(\sigma_{U}^{2}\right)\right\}$,

Since

$$-\frac{m}{2}-\frac{3}{2}+\left(\frac{m+1}{2}+1\right)=0,$$

the $E_{\mathfrak{q}}\left\{\log\left(\sigma_{U}^{2}\right)\right\}$ contributions to $\log\underline{p}\left(\boldsymbol{y};\mathfrak{q},\mathrm{C}\right)$ is zero.

6.C.3 Collection of Logarithmic Moments of Variance Auxiliary Parameters

The $E_{\mathfrak{q}}\left\{\log\left(a_{\beta}\right)\right\}$ Terms

The ninth term of (6.98) contains: $-\frac{1}{2}E_{\mathfrak{q}}\left\{\log\left(a_{\beta}\right)\right\}$.

The eleventh term of (6.98) contains: $-\frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(a_{\beta}\right)\right\}$.

The twelfth term of (6.98) contains: $2E_{\mathfrak{q}}\left\{\log\left(a_{\beta}\right)\right\}$.

Since

$$-\frac{1}{2}-\frac{3}{2}+2=0,$$

the $E_{\mathfrak{q}}\left\{\log\left(a_{\beta}\right)\right\}$ contribution to $\log\underline{p}\left(\boldsymbol{y};\mathfrak{q},\mathrm{C}\right)$ is zero.

The $E_{\mathfrak{q}}\left\{\log\left(a_{uk}\right)\right\}$ Terms, $1\leq k\leq d_{\bullet}$

The nineteenth term of (6.98) contains: $-\frac{1}{2}\sum_{k=1}^{d_{\bullet}}E_{\mathfrak{q}}\left\{\log\left(a_{uk}\right)\right\}$.

The twenty-first term of (6.98) contains: $-\frac{3}{2}\sum_{k=1}^{d_{\bullet}}E_{\mathfrak{q}}\left\{\log\left(a_{uk}\right)\right\}$.

The twenty-second term of (6.98) contains: $2\sum_{k=1}^{d_{\bullet}}E_{\mathfrak{q}}\left\{\log\left(a_{uk}\right)\right\}$.

Since

$$-\frac{1}{2}-\frac{3}{2}+2=0,$$

the $E_{\mathfrak{q}}\left\{\log\left(a_{uk}\right)\right\}$ contributions to $\log\underline{p}\left(\boldsymbol{y};\mathfrak{q},\mathrm{C}\right)$ are zero for each $1\leq k\leq d_{\bullet}$.

The $E_{\mathfrak{q}}\left\{\log\left(a_{U}\right)\right\}$ Terms

The twenty-fifth term of (6.98) contains: $-\frac{1}{2}E_{\mathfrak{q}}\left\{\log\left(a_{U}\right)\right\}$.

The twenty-seventh term of (6.98) contains: $-\frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(a_{U}\right)\right\}$.

The twenty-eighth term of (6.98) contains: $2E_{\mathfrak{q}}\left\{\log\left(a_U\right)\right\}$.

Since

$$-\frac{1}{2}-\frac{3}{2}+2=0,$$

the $E_{\mathfrak{q}}\left\{\log\left(a_U\right)\right\}$ contribution to $\log \underline{p}\left(\boldsymbol{y};\mathfrak{q},\mathrm{C}\right)$ is zero.

6.C.4 Collection of Constants

The constant contributions to each of the terms of $\log \underline{p}\left(\boldsymbol{y};\mathfrak{q},\mathrm{C}\right)$ are as follows:

first	$-\frac{1}{2}\log(2\pi)$ $-\frac{1}{2}\log\left(\sigma_{\beta_0}^2\right)$	fifteenth	$-\frac{1}{2}\log(2\pi)\sum_{k=1}^{d_\bullet} L_k$
second	$\frac{1}{2}\{\log(2\pi+1)\}$	sixteenth	$\frac{1}{2}\{\log(2\pi)+1\}\sum_{k=1}^{d_\bullet} L_k$
third	$(d_\circ + d_\bullet)$ $\times \log(1-\rho_\beta)$	seventeenth	$-\frac{1}{2}d_\bullet\{\log(2)+1\}$ $-\frac{1}{2}\log(2)\sum_{k=1}^{d_\bullet} L_k$ $-\sum_{k=1}^{d_\bullet}\log\left\{\Gamma\left(\frac{1}{2}(L_k+1)\right)\right\}$
fourth	0	eighteenth	$\frac{1}{2}d_\bullet\{\log(2\pi)+1\}$
fifth	$-\frac{1}{2}(d_\circ + d_\bullet)\log(2\pi)$	nineteenth	$-\frac{1}{2}d_\bullet\log(\pi)$
sixth	$\frac{1}{2}\{(d_\circ + d_\bullet)\log(2\pi)$ $+(d_\circ + d_\bullet)\}$	twentieth	$\sum_{k=1}^{d_\bullet}\log\left\{\Gamma\left(\frac{1}{2}(L_k+1)\right)\right\}$
seventh	$-(d_\circ + d_\bullet)$ $\times\left\{\log(2)+\frac{1}{2}\right\}$	twenty-first	$-\frac{1}{2}d_\bullet\log(\pi) - d_\bullet\log(s_u)$
eighth	$\frac{1}{2}(d_\circ + d_\bullet)$ $\times\{\log(2\pi)+1\}$	twenty-second	0
ninth	$-\frac{1}{2}\log(\pi)$	twenty-third	$-\frac{m}{2}\log(2\pi)$
tenth	$\log\left\{\Gamma\left(\frac{1}{2}\right.\right.$ $\times\ (d_\circ + d_\bullet + 1))\}$	twenty-fourth	$\frac{m}{2}\{\log(2\pi)+1\}$
eleventh	$-\frac{1}{2}\log(\pi) - \log(s_\beta)$	twenty-fifth	$-\frac{1}{2}\log(\pi)$
twelfth	0	twenty-sixth	$\log\left\{\Gamma\left(\frac{m+1}{2}\right)\right\}$
thirteenth	$d_\bullet\log(1-\rho_u)$	twenty-seventh	$-\frac{1}{2}\log(\pi) - \log(s_U)$
fourteenth	0	twenty-eighth	0

Table 6.2: Constants that do not depend on any q-density parameters of (6.98), denoted as const_1.

Therefore

$$
\begin{aligned}
&\log \underline{p}\left(\boldsymbol{y}; \mathfrak{q}, \mathrm{C}\right) \\
&= \mathrm{const}_1 - \frac{1}{2}\left\{\mu^2_{\mathfrak{q}(\beta_0)} + \sigma^2_{\mathfrak{q}(\beta_0)}\right\} \Big/ \sigma^2_{\beta_0} + \frac{1}{2}\log\left(\sigma^2_{\mathfrak{q}(\beta_0)}\right) + \mathrm{logit}(\rho_\beta) \sum_{k=1}^{d_\circ + d_\bullet} \mu_{\mathfrak{q}(\gamma_{\beta k})} \\
&\quad - \sum_{k=1}^{d_\circ + d_\bullet} \left[\mu_{\mathfrak{q}\left(\gamma_{\beta k}\right)} \log\left(\mu_{\mathfrak{q}\left(\gamma_{\beta k}\right)}\right) + \left(1 - \mu_{\mathfrak{q}\left(\gamma_{\beta k}\right)}\right) \log\left(1 - \mu_{\mathfrak{q}\left(\gamma_{\beta k}\right)}\right)\right] \\
&\quad - \frac{1}{2}\mu_{\mathfrak{q}\left(1/\sigma^2_\beta\right)} \sum_{k=1}^{d_\circ + d_\bullet} \mu_{\mathfrak{q}\left(b_{\beta k}\right)} \left(\mu^2_{\mathfrak{q}\left(\widetilde{\beta}_k\right)} + \sigma^2_{\mathfrak{q}\left(\widetilde{\beta}_k\right)}\right) + \frac{1}{2}\log\left|\boldsymbol{\Sigma}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right| \\
&\quad - \frac{1}{2}\sum_{k=1}^{d_\circ + d_\bullet} \left\{1 \Big/ \mu_{\mathfrak{q}\left(b_{\beta k}\right)}\right\} - \mu_{\mathfrak{q}\left(1/a_\beta\right)} \mu_{\mathfrak{q}\left(1/\sigma^2_\beta\right)} - \frac{1}{2}\left(d_\circ + d_\bullet + 1\right) \log\left(\lambda_{\mathfrak{q}\left(\sigma^2_\beta\right)}\right) \\
&\quad + \mu_{\mathfrak{q}\left(1/\sigma^2_\beta\right)} \lambda_{\mathfrak{q}\left(\sigma^2_\beta\right)} - \mu_{\mathfrak{q}\left(1/a_\beta\right)} \Big/ s^2_\beta + \lambda_{\mathfrak{q}\left(a_\beta\right)} \mu_{\mathfrak{q}\left(1/a_\beta\right)} - \log\left\{\lambda_{\mathfrak{q}\left(a_\beta\right)}\right\} \\
&\quad - \sum_{k=1}^{d_\bullet} \left[\mu_{\mathfrak{q}(\gamma_{uk})} \log\left(\mu_{\mathfrak{q}(\gamma_{uk})}\right) + \left\{1 - \mu_{\mathfrak{q}(\gamma_{uk})}\right\} \log\left(1 - \mu_{\mathfrak{q}(\gamma_{uk})}\right)\right] \\
&\quad + \mathrm{logit}(\rho_u) \sum_{k=1}^{d_\bullet} \mu_{\mathfrak{q}(\gamma_{uk})} - \frac{1}{2}\sum_{k=1}^{d_\bullet} \mu_{\mathfrak{q}\left(1/\sigma^2_{uk}\right)} \mu_{\mathfrak{q}(b_{uk})} \left(\left\|\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right\|^2 + \mathbf{1}^T_{L_k} \boldsymbol{\sigma}^2_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right) \\
&\quad + \frac{1}{2}\sum_{k=1}^{d_\bullet} \sum_{l=1}^{L_k} \log\left(\sigma^2_{\mathfrak{q}(\widetilde{u}_{kl})}\right) - \frac{1}{2}\sum_{k=1}^{d_\bullet} \left\{1 \Big/ \mu_{\mathfrak{q}(b_{uk})}\right\} - \sum_{k=1}^{d_\bullet} \mu_{\mathfrak{q}\left(1/a_{uk}\right)} \mu_{\mathfrak{q}\left(1/\sigma^2_{uk}\right)} \\
&\quad - \frac{1}{2}\sum_{k=1}^{d_\bullet} \left(L_k + 1\right) \log\left(\lambda_{\mathfrak{q}\left(\sigma^2_{uk}\right)}\right) + \sum_{k=1}^{d_\bullet} \mu_{\mathfrak{q}\left(1/\sigma^2_{uk}\right)} \lambda_{\mathfrak{q}\left(\sigma^2_{uk}\right)} - \left(1/s^2_u\right) \sum_{k=1}^{d_\bullet} \mu_{\mathfrak{q}\left(1/a_{uk}\right)} \\
&\quad + \sum_{k=1}^{d_\bullet} \left\{\lambda_{\mathfrak{q}(a_{uk})} \mu_{\mathfrak{q}\left(1/a_{uk}\right)} - \log\left(\lambda_{\mathfrak{q}(a_{uk})}\right)\right\} \\
&\quad - \frac{1}{2}\mu_{\mathfrak{q}\left(1/\sigma^2_U\right)} \sum_{i=1}^{m} \left(\mu^2_{\mathfrak{q}(U_i)} + \sigma^2_{\mathfrak{q}(U_i)}\right) + \frac{1}{2}\sum_{i=1}^{m} \log\left(\sigma^2_{\mathfrak{q}(U_i)}\right) \\
&\quad - \mu_{\mathfrak{q}\left(1/a_U\right)} \mu_{\mathfrak{q}\left(1/\sigma^2_U\right)} - \frac{m+1}{2}\log\left(\lambda_{\mathfrak{q}\left(\sigma^2_U\right)}\right) + \mu_{\mathfrak{q}\left(1/\sigma^2_U\right)} \lambda_{\mathfrak{q}\left(\sigma^2_U\right)} \\
&\quad - \mu_{\mathfrak{q}(1/a_U)} \big/ S^2_U - \log\left\{\lambda_{\mathfrak{q}(a_U)}\right\} + \lambda_{\mathfrak{q}(a_U)} \mu_{\mathfrak{q}\left(1/a_U\right)}.
\end{aligned}
$$

For the Gaussian responses, the second term of (6.96) is

$$
\begin{aligned}
&E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{y}\middle|\beta_0,\boldsymbol{U},\boldsymbol{\gamma}_{\beta},\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{u}},\boldsymbol{\gamma}_u,\sigma_{\varepsilon}^2\right)\right\}\right] \\
&\quad=-\frac{N}{2}\log\left(\pi\right)-\frac{N}{2}E_{\mathfrak{q}}\left\{\log\left(\sigma_{\varepsilon}^2\right)\right\} \\
&\qquad-\frac{1}{2}\mu_{\mathfrak{q}\left(1/\sigma_{\varepsilon}^2\right)}E_{\mathfrak{q}}\left(\sum_{i=1}^{m}\left\|\boldsymbol{y}_i-\mathbf{1}_{n_i}\beta_0-\mathbf{1}_{n_i}U_i-\boldsymbol{X}_i\left(\boldsymbol{\gamma}_{\beta}\odot\widetilde{\boldsymbol{\beta}}\right)-\sum_{k=1}^{d_{\bullet}}\boldsymbol{Z}_{ik}\left(\gamma_{uk}\widetilde{\boldsymbol{u}}_k\right)\right\|^2\right).
\end{aligned}
$$

The third term of (6.96) is

$$
\begin{aligned}
E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\sigma_{\varepsilon}^2\middle|a_{\varepsilon}\right)\right\}\right]&=E_{\mathfrak{q}}\left[\log\left\{1/a_{\varepsilon}\right)^{1/2}\Big/\Gamma\left(1\right\}-\frac{3}{2}\log\left(\sigma_{\varepsilon}^2\right)-1/\left(\sigma_{\varepsilon}^2a_{\varepsilon}\right)\right] \\
&=-\frac{1}{2}E_{\mathfrak{q}}\left\{\log\left(a_{\varepsilon}\right)\right\}-\frac{1}{2}\log\left(\pi\right)-\frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(\sigma_{\varepsilon}^2\right)\right\}-\mu_{\mathfrak{q}\left(1/a_{\varepsilon}\right)}\mu_{\mathfrak{q}\left(1/\sigma_{\varepsilon}^2\right)}.
\end{aligned}
$$

The fourth term of (6.96) is

$$
\begin{aligned}
-E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\sigma_{\varepsilon}^2\right)\right\}\right]&=-E_{\mathfrak{q}}\left[\frac{1}{2}\left(N+1\right)\log\left(\lambda_{\mathfrak{q}(\sigma_{\varepsilon}^2)}\right)-\log\left\{\Gamma\left(\frac{1}{2}\left(N+1\right)\right)\right\}\right. \\
&\qquad\left.-\left\{\frac{1}{2}\left(N+1\right)+1\right\}E_{\mathfrak{q}}\left\{\log\left(\sigma_{\varepsilon}^2\right)\right\}-\lambda_{\mathfrak{q}(\sigma_{\varepsilon}^2)}/\sigma_{\varepsilon}^2\right] \\
&=-\frac{1}{2}\left(N+1\right)\log\left(\lambda_{\mathfrak{q}(\sigma_{\varepsilon}^2)}\right)+\log\left\{\Gamma\left(\frac{1}{2}\left(N+1\right)\right)\right\} \\
&\quad+\left\{\frac{1}{2}\left(N+1\right)+1\right\}E_{\mathfrak{q}}\left\{\log\left(\sigma_{\varepsilon}^2\right)\right\}+\mu_{\mathfrak{q}\left(1/\sigma_{\varepsilon}^2\right)}\left\{\mu_{\mathfrak{q}(1/a_{\varepsilon})}+\right. \\
&\quad\left.+\frac{1}{2}E_{\mathfrak{q}}\left(\sum_{i=1}^{m}\left\|\boldsymbol{y}_i-\mathbf{1}_{n_i}\beta_0-\mathbf{1}_{n_i}U_i-\boldsymbol{X}_i(\boldsymbol{\gamma}_{\beta}\odot\widetilde{\boldsymbol{\beta}})+\sum_{k=1}^{d_{\bullet}}\boldsymbol{Z}_{ik}(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)\right\|^2\right)\right\}.
\end{aligned}
$$

The fifth term of (6.96) is

$$
E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(a_{\varepsilon}\right)\right\}\right]=-\frac{1}{2}\log\left(\pi\right)-\log\left(s_{\varepsilon}\right)-\frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(a_{\varepsilon}\right)\right\}-\mu_{\mathfrak{q}\left(1/a_{\varepsilon}\right)}\Big/s_{\varepsilon}^2.
$$

The sixth term of (6.96) is

$$-E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(a_{\varepsilon}\right)\right\}\right]=-\log\left(\lambda_{\mathfrak{q}(a_{\varepsilon})}\right)+2E_{\mathfrak{q}}\left\{\log\left(a_{\varepsilon}\right)\right\}+\lambda_{\mathfrak{q}(a_{\varepsilon})}\mu_{\mathfrak{q}\left(1/a_{\varepsilon}\right)}.$$

Therefore, for the Gaussian responses, we have

$$\begin{aligned}
\log\underline{p}\left(\boldsymbol{y};\mathfrak{q}\right)&=\log\underline{p}\left(\boldsymbol{y};\mathfrak{q},\mathsf{C}\right)+E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{y}\middle|\beta_0,\boldsymbol{U},\boldsymbol{\gamma}_{\beta},\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{u}},\boldsymbol{\gamma}_{u},\sigma_{\varepsilon}^{2}\right)\right\}\right]\\
&\quad+E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\sigma_{\varepsilon}^{2}\middle|a_{\varepsilon}\right)\right\}\right]-E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\sigma_{\varepsilon}^{2}\right)\right\}\right]\\
&\quad+E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(a_{\varepsilon}\right)\right\}\right]-E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(a_{\varepsilon}\right)\right\}\right]\\
&=\log\underline{p}\left(\boldsymbol{y};\mathfrak{q},\mathsf{C}\right)-\frac{N}{2}\log\left(2\pi\right)-\frac{N}{2}E_{\mathfrak{q}}\left\{\log\left(\sigma_{\varepsilon}^{2}\right)\right\}\\
&\quad-\frac{1}{2}\mu_{\mathfrak{q}\left(1/\sigma_{\varepsilon}^{2}\right)}E_{\mathfrak{q}}\left\{\sum_{i=1}^{m}\left\|\boldsymbol{y}_i-\mathbf{1}_{n_i}\beta_0-\mathbf{1}_{n_i}U_i-\boldsymbol{X}_i\left(\boldsymbol{\gamma}_{\beta}\odot\widetilde{\boldsymbol{\beta}}\right)-\sum_{k=1}^{d_{\bullet}}\boldsymbol{Z}_{ik}\left(\gamma_{uk}\widetilde{\boldsymbol{u}}_k\right)\right\|^2\right\}\\
&\quad-\frac{1}{2}E_{\mathfrak{q}}\left\{\log\left(a_{\varepsilon}\right)\right\}-\frac{1}{2}\log\left(\pi\right)-\frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(\sigma_{\varepsilon}^{2}\right)\right\}-\mu_{\mathfrak{q}\left(1/a_{\varepsilon}\right)}\mu_{\mathfrak{q}\left(1/\sigma_{\varepsilon}^{2}\right)}\\
&\quad-\frac{1}{2}\left(N+1\right)\log\left(\lambda_{\mathfrak{q}(\sigma_{\varepsilon}^{2})}\right)+\log\left\{\Gamma\left(\frac{1}{2}\left(N+1\right)\right)\right\}\\
&\quad+\left\{\frac{1}{2}\left(N+1\right)+1\right\}E_{\mathfrak{q}}\left\{\log\left(\sigma_{\varepsilon}^{2}\right)\right\}+\mu_{\mathfrak{q}\left(1/\sigma_{\varepsilon}^{2}\right)}\mu_{\mathfrak{q}(1/a_{\varepsilon})}\\
&\quad+\frac{1}{2}\mu_{\mathfrak{q}\left(1/\sigma_{\varepsilon}^{2}\right)}E_{\mathfrak{q}}\left(\sum_{i=1}^{m}\left\|\boldsymbol{y}_i-\mathbf{1}_{n_i}\beta_0-\mathbf{1}_{n_i}U_i-\boldsymbol{X}_i(\boldsymbol{\gamma}_{\beta}\odot\widetilde{\boldsymbol{\beta}})+\sum_{k=1}^{d_{\bullet}}\boldsymbol{Z}_{ik}(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)\right\|^2\right)\\
&\quad-\frac{1}{2}\log\left(\pi\right)-\log\left(s_{\varepsilon}\right)-\frac{3}{2}E_{\mathfrak{q}}\left\{\log\left(a_{\varepsilon}\right)\right\}-\mu_{\mathfrak{q}\left(1/a_{\varepsilon}\right)}\Big/s_{\varepsilon}^{2}\\
&\quad-\log\left(\lambda_{\mathfrak{q}(a_{\varepsilon})}\right)+2E_{\mathfrak{q}}\left\{\log\left(a_{\varepsilon}\right)\right\}+\lambda_{\mathfrak{q}(a_{\varepsilon})}\mu_{\mathfrak{q}\left(1/a_{\varepsilon}\right)}\\
&=\log\underline{p}\left(\boldsymbol{y};\mathfrak{q},\mathsf{C}\right)-\frac{1}{2}\left(N+1\right)\log\left(\lambda_{\mathfrak{q}(\sigma_{\varepsilon}^{2})}\right)\\
&\quad-\mu_{\mathfrak{q}\left(1/a_{\varepsilon}\right)}\Big/s_{\varepsilon}^{2}-\log\left(\lambda_{\mathfrak{q}(a_{\varepsilon})}\right)+\lambda_{\mathfrak{q}(a_{\varepsilon})}\mu_{\mathfrak{q}\left(1/a_{\varepsilon}\right)}+\text{const}_2.
\end{aligned}$$

where const_2 denotes constants that do not depend on any $\mathfrak{q}$-density parameters.

In the binary response case, recalling that (6.97) is

$$\log \underline{p}\left(\boldsymbol{y};\mathfrak{q}\right)=\log \underline{p}\left(\boldsymbol{y};\mathfrak{q},\mathrm{C}\right)+E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{y}\middle|\boldsymbol{c}\right)\right\}\right]$$
$$+E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{c}\middle|\beta_0,\boldsymbol{U},\boldsymbol{\gamma}_{\beta},\widetilde{\boldsymbol{\beta}},\boldsymbol{\gamma}_u,\widetilde{\boldsymbol{u}}\right)\right\}\right]-E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\boldsymbol{c}\right)\right\}\right],$$

and recalling that

$$\mathfrak{p}(\boldsymbol{y}|\boldsymbol{c})=\prod_{i=1}^{m}\prod_{j=1}^{n_i}I(c_{ij}\geq 0)^{y_{ij}}I(c_{ij}<0)^{1-y_{ij}},$$

then the second term of (6.97) becomes

$$E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{y}\middle|\boldsymbol{c}\right)\right\}\right]=\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(y_{ij}E_{\mathfrak{q}}\left[\log\left\{I(c_{ij}\geq 0)\right\}\right]+(1-y_{ij})E_{\mathfrak{q}}\left[\log\left\{I(c_{ij}<0)\right\}\right]\right).$$

The third term of (6.97) is

$$E_{\mathfrak{q}}\left[\log\left\{\mathfrak{p}\left(\boldsymbol{c}\middle|\beta_0,\boldsymbol{U},\boldsymbol{\gamma}_{\beta},\widetilde{\boldsymbol{\beta}},\boldsymbol{\gamma}_u,\widetilde{\boldsymbol{u}}\right)\right\}\right]$$
$$=-\frac{N}{2}\log(2\pi)-\frac{1}{2}E_{\mathfrak{q}}\left(\sum_{i=1}^{m}\left\|\boldsymbol{c}_i-\mathbf{1}_{n_i}\beta_0-\mathbf{1}_{n_i}U_i-\boldsymbol{X}_i(\boldsymbol{\gamma}_{\beta}\odot\widetilde{\boldsymbol{\beta}})-\sum_{k=1}^{d_{\bullet}}\boldsymbol{Z}_{ik}(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)\right\|^2\right).$$

Recalling that

$$\log\left\{\mathfrak{q}\left(\boldsymbol{c}\right)\right\}=-\frac{N}{2}\log(2\pi)+\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left[y_{ij}\log\left\{I(c_{ij}\geq 0)\right\}-y_{ij}\log\left\{\Phi(\boldsymbol{\eta}_{ij})\right\}\right.$$
$$\left.+(1-y_{ij})\log\left\{I(c_{ij}<0)\right\}-(1-y_{ij})log\left\{\Phi(-\boldsymbol{\eta}_{ij})\right\}\right]$$
$$-\frac{1}{2}\sum_{i=1}^{m}\left\|\boldsymbol{c}_i-\mathbf{1}_{n_i}\beta_0-\mathbf{1}_{n_i}U_i-\boldsymbol{X}_i(\boldsymbol{\gamma}_{\beta}\odot\widetilde{\boldsymbol{\beta}})-\sum_{k=1}^{d_{\bullet}}\boldsymbol{Z}_{ik}(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)\right\|^2$$

where

$$\boldsymbol{\eta}^{\mathfrak{q}}\equiv\mu_{\mathfrak{q}(\beta_0)}\mathbf{1}_N+\begin{bmatrix}\mu_{\mathfrak{q}(U_1)}\mathbf{1}_{n_1}\\ \vdots\\ \mu_{\mathfrak{q}(U_m)}\mathbf{1}_{n_M}\end{bmatrix}+\boldsymbol{X}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_{\beta})}\odot\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right)+\sum_{k=1}^{d_{\bullet}}\dot{\boldsymbol{Z}}_k\left(\mu_{\mathfrak{q}(\gamma_{uK})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_K)}\right),$$

then the fourth term of (6.97) becomes

$$
\begin{aligned}
-E_{\mathfrak{q}}\left[\log\left\{\mathfrak{q}\left(\boldsymbol{c}\right)\right\}\right] = \frac{N}{2}\log(2\pi) - \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(y_{ij}E_{\mathfrak{q}}\left[\{I(c_{ij}\geq 0)\}\right] + (1-y_{ij})E_{\mathfrak{q}}\left[\{I(c_{ij}<0)\}\right]\right. \\
-y_{ij}E_{\mathfrak{q}}\left[\log\left\{\Phi(\boldsymbol{\eta}_{\boldsymbol{ij}}^{\mathfrak{q}})\right\}\right] - (1-y_{ij})E_{\mathfrak{q}}\left[\log\left\{\Phi(-\boldsymbol{\eta}_{\boldsymbol{ij}}^{\mathfrak{q}})\right\}\right]\Big) \\
+\frac{1}{2}E_{\mathfrak{q}}\left(\sum_{i=1}^{m}\left\|\boldsymbol{c}_i - \mathbf{1}_{n_i}\beta_0 - \mathbf{1}_{n_i}U_i - \boldsymbol{X}_i(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}}) - \sum_{k=1}^{d_\bullet}\boldsymbol{Z}_{ik}(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)\right\|^2\right).
\end{aligned}
$$

Therefore, for binary response, we have

$$
\begin{aligned}
&\log\underline{p}\left(\boldsymbol{y};\mathfrak{q}\right) = \log\underline{p}\left(\boldsymbol{y};\mathfrak{q},\mathsf{C}\right) \\
&+\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(y_{ij}E_{\mathfrak{q}}\left[\log\left\{I(c_{ij}\geq 0)\right\}\right] + (1-y_{ij})E_{\mathfrak{q}}\left[\log\left\{I(c_{ij}<0)\right\}\right]\right) - \frac{N}{2}\log(2\pi) \\
&-\frac{1}{2}E_{\mathfrak{q}}\left(\sum_{i=1}^{m}\left\|\boldsymbol{c}_i - \mathbf{1}_{n_i}\beta_0 - \mathbf{1}_{n_i}U_i - \boldsymbol{X}_i(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}}) - \sum_{k=1}^{d_\bullet}\boldsymbol{Z}_{ik}(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)\right\|^2\right) \\
&+\frac{N}{2}\log(2\pi) - \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(y_{ij}E_{\mathfrak{q}}\left[\{I(c_{ij}\geq 0)\}\right] + (1-y_{ij})E_{\mathfrak{q}}\left[\{I(c_{ij}<0)\}\right]\right. \\
&-y_{ij}E_{\mathfrak{q}}\left[\log\left\{\Phi(\boldsymbol{\eta}_{\boldsymbol{ij}}^{\mathfrak{q}})\right\}\right] - (1-y_{ij})E_{\mathfrak{q}}\left[\log\left\{\Phi(-\boldsymbol{\eta}_{\boldsymbol{ij}}^{\mathfrak{q}})\right\}\right]\Big) \\
&+\frac{1}{2}E_{\mathfrak{q}}\left(\sum_{i=1}^{m}\left\|\boldsymbol{c}_i - \mathbf{1}_{n_i}\beta_0 - \mathbf{1}_{n_i}U_i - \boldsymbol{X}_i(\boldsymbol{\gamma}_\beta\odot\widetilde{\boldsymbol{\beta}}) - \sum_{k=1}^{d_\bullet}\boldsymbol{Z}_{ik}(\gamma_{uk}\widetilde{\boldsymbol{u}}_k)\right\|^2\right) \\
&= \log\underline{p}\left(\boldsymbol{y};\mathfrak{q},\mathsf{C}\right) + \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(y_{ij}E_{\mathfrak{q}}\left[\log\left\{\Phi(\boldsymbol{\eta}_{ij}^{\mathfrak{q}})\right\}\right] + (1-y_{ij})E_{\mathfrak{q}}\left[\log\left\{-\Phi(-\boldsymbol{\eta}_{ij}^{\mathfrak{q}})\right\}\right]\right) \\
&= \log\underline{p}\left(\boldsymbol{y};\mathfrak{q},\mathsf{C}\right) + \sum_{i=1}^{m}\sum_{j=1}^{n_j}\log\left\{\Phi\left((2y_{ij}-1)\,\boldsymbol{\eta}_{ij}^{\mathfrak{q}}\right)\right\} \\
&= \log\underline{p}\left(\boldsymbol{y};\mathfrak{q},\mathsf{C}\right) + \sum_{i=1}^{m}\sum_{j=1}^{n_i}\log\Bigg\{\Phi\Bigg((2y_{ij}-1)\left[\mu_{\mathfrak{q}(\beta_0)} + \mu_{\mathfrak{q}(U_i)}\right. \\
&\quad\left.+\left(\boldsymbol{X}_i\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\gamma}_\beta)}\odot\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{\beta}})}\right) + \sum_{k=1}^{d_\bullet}\boldsymbol{Z}_{ik}\left(\mu_{\mathfrak{q}(\gamma_{uk})}\boldsymbol{\mu}_{\mathfrak{q}(\widetilde{\boldsymbol{u}}_k)}\right)\right)_j\right]\Bigg)\Bigg\}.
\end{aligned}
$$

Chapter 7

Application to the Australian Red Cross Blood Service Datasets

7.1 Introduction

Blood donation is currently one of the few methods to collect blood for therapeutic purposes and is essential for healthcare worldwide. Due to the typically voluntary nature of blood donations, donor attendance rates can vary significantly. As a result, these variations affect both the financial and human resources of the healthcare system (Boksmati et al., 2016). Consequently, blood collection services aim to understand the factors influencing donor attendance, such as Charbonneau et al. (2015), Bagot et al. (2013) and Gemelli et al. (2017). However, most studies rely on small datasets and lack the use of modern statistical tools for analysis.

Our dataset, provided by the Australian Red Cross Blood Service, has over 3 million donation records, collected in Queensland between 2015 and 2017. Table 7.1 provides the descriptions of 37 relevant recorded variables. Our goal is to identify the key factors influencing blood donors' appointment outcomes, such as whether they attend their scheduled donation, while incorporating more variables and a larger sample size than previous studies. Computationally efficient methods are necessary due to the large size of the dataset.

The Bayesian model selection paradigm is used to facilitate the GAMM selection with random intercept described in Chapter 6, with accuracy and tremendous flexibility to achieve the scalability when handling large data sets. We apply the methodology described in Chapter 6 for the GAMM model selection with random intercept to the Australian Red Cross Blood Service dataset in this chapter.

We aim to categorize each candidate predictor's effect type to be linear, non-linear or zero on the mean response, to identify the contribution of variables to the response data, via the MCMC approach and the MFVB approach as the faster option. For the linear effect predictors, we need to know they are having positive or negative impacts on the mean response by the posterior mean and the 95% credible intervals of all the selected linear coefficients. For the non-linear effect predictors, we display relevant slices of the additive model fit, for which each of the other functions are evaluated at the median value of their predictor. For the zero effect predictors, they can be discarded, making the selected model parsimonious.

The data cleaning process of the Australian Red Cross Blood Service data is described in Section 7.2.1. We simplify the data for the three-category GAMM selection with random intercept in grouped data. The selection result is given in Section 7.2.2, using the MCMC approach, as described in Algorithm 9 in Section 6.5.3. Further, the selection result is given in Section 7.2.3, using the MFVB approach, as the faster option as described in Algorithm 10 in Section 6.6.2.

7.2 Analysis of the Australian Red Cross Blood Service Dataset

Table 7.1 provides a description of each of the 37 relevant variables recorded. These 37 variables will be used in Section 7.2.2–7.2.3 for GAMM selection with random intercept.

Table 7.1: *A table that shows each of the variables in the analysis, explains what they mean and lists the type of data they are.*

Variable name	**Description**
ID	A multilevel factor variable with each donor's ID number.
attendance	A two–level factor whether the donors either *attended* or were *absent* in their appointment. Didn't attend is the reference.
age	A continuous variable with the donor's age in years.
sex	A two–level factor variable with the donor's gender. Female is the reference.
bloodType	A nine–level factor variable with donor's blood type. It can be $A+$, $A-$, $AB-$, $AB+$, $B-$, $B+$, $O-$, $O+$ or *Unknown*.
pltDon	A count variable with the number of platelet donations the donor has had.
WBDon	A count variable with the number of whole blood donations the donor has had.
plsDon	A count variable with the number of plasma donations the donor has had.
apptTime	A continuous positive variable with time of the appointment, converted to proportion of a day measured from midnight in hours.
apptWkday	A six–level factor variable with weekday of the appointment from 2 to 7, where 2,3,4,5,6,7 are weekdays Monday, Tuesday, Wednesday, Thursday, Friday, Saturday respectively.
apptMth	A 12–level factor variable with the month of the appointment.
apptSeas	A four–level factor variable with season of the appointment.
apptYear	A three–level factor variable with the year of the appointment.

Continued on next page

Table 7.1: *A table that shows each of the variables in the analysis, explains what they mean and lists the type of data they are.* (Continued)

Variable name	Description
apptPubHolRad	A five–level factor variable with the number of days either before or after the appointment to the nearest public holiday. If a public holiday is not within seven days of an appointment, this variable takes the value 0. If an appointment is one day from a public holiday, this variable takes the value 1. If an appointment is two days from a public holiday, this variable takes the value 2, etc.
apptCreDelta	A continuous variable with the difference between the date of creating the booking and the actual donation date.
apptBookType	A seven–level factor with the type of booking made. *In centre* bookings are those made after the previous donation. *NCC inbound* is where the donor calls the centre. *NCC outbound* is where the centre calls the donor. *NCC unknown* is a booking made over the phone, but it is unsure who called who. *Community relations* are bookings made through community outreach. *Portal* are bookings made using the portal software and *Webbooking* are those made over the web.
colTypeSched	A four–level factor variable with the type of blood collection procedure scheduled to be conducted. It can be either *plasmapheresis*, *plateletpheresis*, *whole blood* or *no blood*.
colTypeTaken	A four–level factor variable with the type of blood collection procedure conducted. It can be either *plasmapheresis*, *plateletpheresis*, *whole blood*, *no blood* taken during appointment.
centreType	A two–level factor with the type of centre where the collection of blood was scheduled, i.e. whether it was a *static* collection room or a *mobile* blood van. Here *mobile* is the reference category.
centreLoc	A two–level factor whether the scheduled collection centre was *regional* or *metro*. Here *regional* is the reference category.

Continued on next page

Table 7.1: *A table that shows each of the variables in the analysis, explains what they mean and lists the type of data they are.* (Continued)

Variable name	**Description**
lastDefLen	A continuous variable given as the length of time in weeks since the donor last deferred an appointment. This is set to 0 for when the donor has not had a previous deferral.
lastDefEndLen	A continuous variable with length of time in weeks since the donor last deferral ended. This is set to 0 for when the donor has not had a previous deferral.
inbndSMS	A count variable with the number of messages sent to the Red Cross from the donor.
inbndPhone	A count variable with the number of phone calls made to the Red Cross by the donor.
inbndEmail	A count variable with the number of emails sent to the Red Cross by the donor.
inbndUnknown	A count variable with the number of unknown communications sent to the Red Cross by the donor.
inbndInternet	A count variable with the number of internet communications sent to the Red Cross by the donor.
outbndPhone	A count variable with the number of phone calls made to the donor by the Red Cross.
outbndSMS	A count variable with the number of messages sent to the donor by the Red Cross.
outbndEmail	A count variable with the number of emails sent to the donor by the Red Cross.
outbndUnknown	A count variable with the number of unknown communications sent to the donor by the Red Cross.
outbndInternet	A count variable with the number of internet communications sent to the donor by the Red Cross.

Continued on next page

Table 7.1: *A table that shows each of the variables in the analysis, explains what they mean and lists the type of data they are.* (Continued)

Variable name	Description
lastTTC	A continuous variable with the time it took in the last appointment before the donor was on the couch ready for the phlebotomist to take blood.
lastBleedTime	A continuous variable with the time it took in the last appointment from needle in to needle out. This is linked to collection type.
lastDonMultiArm	A two–level factor variable which indicates whether or not the donor was punctured in one or both arms. *True* if one arm and *False* if both.
lastDAEVVR	A two–level factor variable indicating whether or not the donor had a vasovagal reaction in their last visit. *True* if one did happen and *False* if one did not.
lastDAEnonVVR	A two–level factor variable indicating whether or not the donor has a non-vassalage reaction in their last visit. *True* if one did happen and *False* if one did not.

7.2.1 Data Cleaning

The Australian Red Cross Blood Service dataset was obtained from various sources and required data cleaning. Donors had the freedom to visit any room at any location in Queensland as often as they wished, resulting in many donors attending different rooms for their appointments. The data has a complex structure, with donations nested within donors and crossed with rooms, where rooms are further nested within areas (see Figure 7.1). Additionally, the true structure of the data includes a time effect. However, due to the methodological limitations discussed in this thesis, we simplify the data to a single level of nesting, with donations nested by donor, by grouping all donations under each donor (see Figure 7.2).

For the application to the Australian Red Cross Blood Service dataset, 37 relevant candidate predictors with descriptions listed in Table 7.1 enter the mode for the GAMM selection with random intercept, as described in Section 7.2.2 using the MCMC approach and as described in Section 7.2.3 using the MFVB approach.

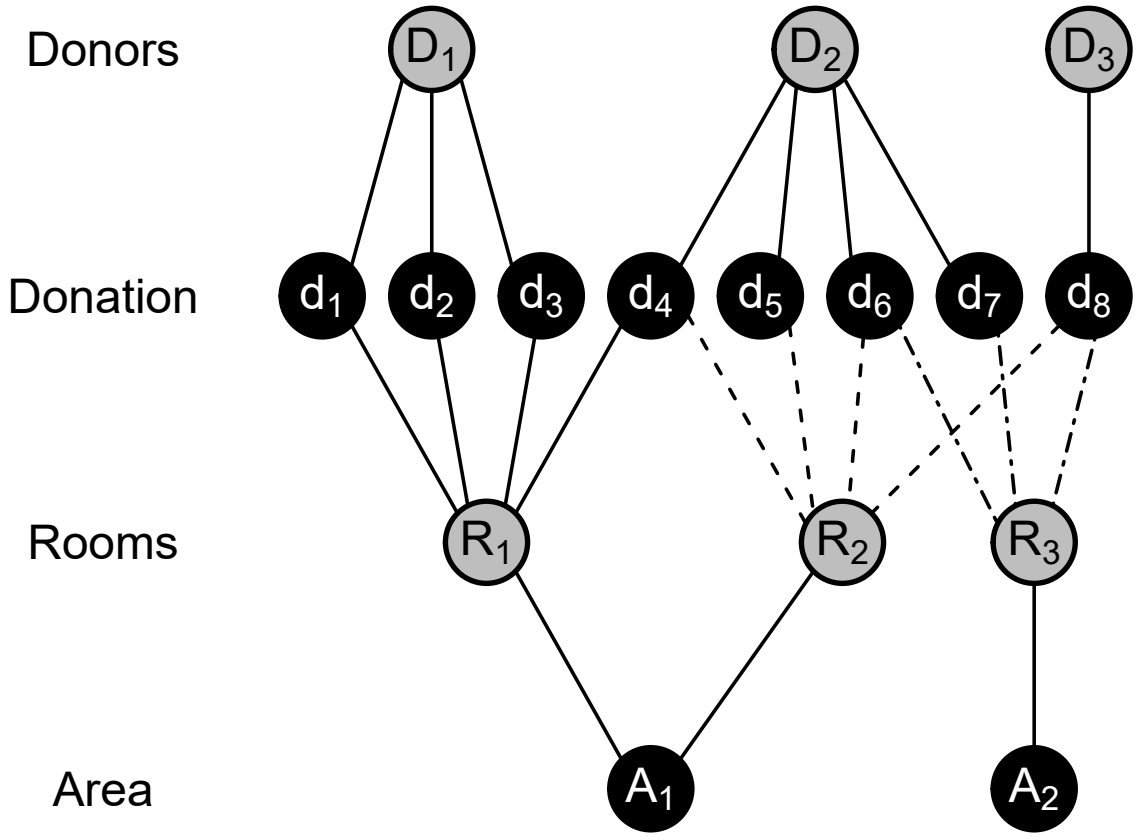


Figure 7.1: *A plot shows the two level multiple membership structure of the data, where donations are nested under donors crossed with rooms, which are nested under area. The first row represents each donor, the second row shows each donation of each donor, the third row shows which room each donation was done in, and the fourth row shows the area of each room. Note the time aspect of the data is not included in this graph.*

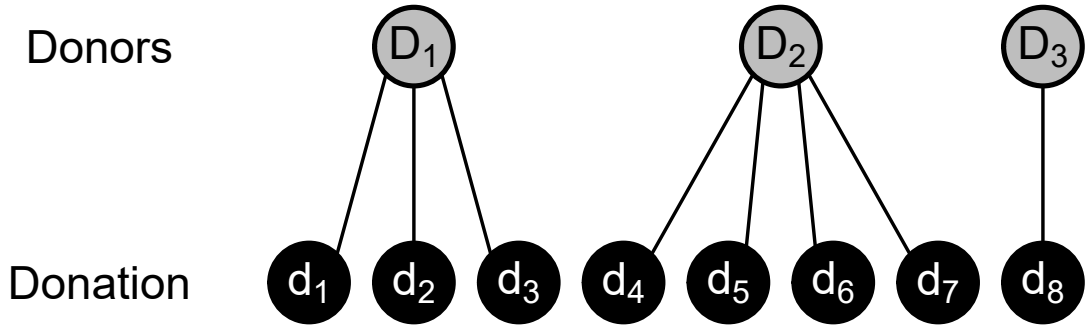


Figure 7.2: *A plot shows the one level structure of the data after the simplification for the GAMM selection with random intercept. Donations are nested under each donor. The first row shows each donor. The second row shows each donation of each donor.*

Some variables were recorded as they were used in the collection rooms and hence were not in a format suitable for statistical analysis. The predictor apptTime listed in Table 7.1, originally in a date-time format, was converted to hours since midnight, with the date component separated into years.

The response variable attendance listed in Table 7.1, is the patient appointment result. It was originally recorded as a six-level factor with the following outcomes being: attended, blood was taken, cancelled, rescheduled, active or unknown. We simplified it to a two-level factor called attendance, where the donors either attended or were absent, with *absent* as the reference category. The response variable attendance is the indicator of whether the donor attended the donation.

The predictor apptCreDelta listed in Table 7.1, is referred to the date gap between creating the booking and actual donation date. The dataset contained negative values for apptCreDelta, representing instances where patients attended without booking in advance. These values occurred since the appointment time was recorded after the donor's blood was taken by staff. These negative values were not useful for predicting donor attendance rates, therefore, we removed these negative values from the dataset.

The dataset also contains several missing values. For first time donors, there are missing values for predictors WBDON, pltDon, plsDon and lastDefLen listed in Table 7.1. These predictors represent the number of whole blood donations the donor has had, the number of platelet donations the donor has had, the number of plasma donations the donor has had and length of time in weeks since the donor last deferral, which are not available for first time donors. Since the primary goal of this analysis is to demonstrate the methodology, we do not investigate this further and instead perform a complete case analysis. Moreover, a spurious observation, where a donor's age made their last deferral length impossible, was removed from the dataset.

The counts of each communication method between the Red Cross and the donor were converted from count variables, indicating the number of times each form of communication occurred, into a binary variable indicating whether there was no communication or at least one communication.

The dataset included several donors with an exceptionally high number of observations (the highest having $2,247$ bookings). It is speculated that these were automated bookings triggered by various information flags, which were either not attended or cancelled. As donors with a high number of bookings were relatively common, we chose not to remove them from the dataset.

The predictors plsDon, pltDon, lastDefLen, lastDefEndLen and lastTTC are mildly or highly skewed. Therefore, we implemented a transform of $\log(x+1)$.

After the conversion of each of the categorical variables to indicator form, the dataset contained 80 candidate predictors. We used a 10% subset of the original data. There are $9,053$ groups with a total of $78,727$ observations for the GAMM model selection with random intercept via MCMC and MFVB, descried in Section 7.2.2 – 7.2.3 respectively.

7.2.2 Data Illustration for Markov Chain Monte Carlo

Using the MCMC approach described in Algorithm 9 in Section 6.5.3, we obtained the categorization results shown in Table 7.2. The model selection process via MCMC takes 237.39 seconds. Among 80 candidate predictors, 41 were selected: 37 linear effects, 4 non-linear effects and 39 predictors discarded.

We obtained the inferential summary for all the selected linear fits using the MCMC approach as shown in Table 7.3 . The descriptions of these predictors are listed in Table 7.1. The results show that apptYear of *2017*, outbndSMS, inbndPhone and outbndPhone, each have positive impacts on the probability of donor attendance. Conversely, centreType, inbndSMS, apptBookType of *NCC outbound*, colTypeSched of *Plateletpheresis Collection* and apptBookType of *Web Booking* each have negative impacts on the probability of donor attendance.

There are four predictors that exhibit non-linear effects on the probability of donor attendance, which are apptCreDelta, WBDon, $\log\left(\text{plsDon}+1\right)$ and apptTime. The descriptions of these predictors are listed in Table 7.1. We obtained Figure 7.3 to visualize the non-linear effects. The first panel for predictor apptCreDelta indicates that the probability of donor attendance sharply decreases from 0.8 to 0.3 as apptCreDelta increases from 0 to 75, and then continues to decrease moderately from 0.3 to

Predictors selected as having linear effects:

colTypeTaken of *no blood taken*, colTypeTaken of *plasmapheresis taken*,
colTypeTaken of *Platelepheresis taken*, colTypeTaken of *whole blood taken*,
age, apptYear of *2016*, apptYear of *2017*, bloodType of $B+$,
apptSeas of *Summer*, apptBookType of *NCC unknown*, apptBookType of *NCC outbound*,
centreLoc, centreType, apptMth of *December*, apptBookType of *NCC inbound*,
lastDAEVVR, apptMth of *February*, apptWkday of *Friday*, outbndUnknown,
apptMth of *January*, apptMth of *June*, $\log\Big(\mathsf{lastDefLen}+1\Big)$, $\log\Big(\mathsf{lastDefEndLen}+1\Big)$,
$\log\Big(\mathsf{pltDon}+1\Big)$, apptMth of *March*, lastDAEnonVVR, bloodType of $O-$,
apptBookType of *other*, inbndPhone, outbndPhone,
colTypeSched of *Plateletpheresis Collection*, sex, inbndSMS, outbndSMS,
bloodType of *unknown*, apptBookType of *Web Booking*, apptWkday of *Wednesday*.

Predictors selected as having non-linear effects:

apptCreDelta, WBDon, $\log\Big(\mathsf{plsDon}+1\Big)$, apptTime.

Table 7.2: *Categorization results obtained from Red Cross Blood Service application dataset using the MCMC approach as given by Algorithm 9.*

0.2 in the range from 100 to 200. The second panel for predictor WBDon shows that the probability of donor attendance increases steadily from 0.4 to 0.6 with particularly high values further increasing the probability of donor attendance. Similarly, the third panel for predictor $\log\Big(\mathsf{plsDon}+1\Big)$ shows a similar steady increase from 0.4 to 0.6 with higher values resulting in higher probability of donor attendance. The fourth panel for predictor apptTime shows that the probability of donor attendance peaks around 0.5 in the range from 0.3 to 0.4 and from 0.7 to 0.8.

7.2.3 Data Illustration for Mean Field Variational Bayes

Using the MFVB approach described in Algorithm 10 in Section 6.6.2, we obtained the categorization results shown in Table 7.4. The model selection process via MFVB takes 45.23 seconds. Among 80 candidate predictors, 36 candidate predictors were selected: 31 linear effects, 5 non-linear effects and 44 discarded. Among the 31 linear

	posterior mean	95% credible interval
colTypeTaken of *no blood taken*	–0.09168	$(-0.1875, 0.0000)$
colTypeTaken of *plasmapheresis taken*	0.05964	$(-0.005898, 0.1768)$
colTypeTaken of *Platelepheresis taken*	–0.06558	$(-0.2024, 0.0000)$
colTypeTaken of *whole blood taken*	0.02673	$(-0.06051, 0.1404)$
age	0.01038	$(0.009372, 0.01148)$
apptYear of *2016*	–0.05056	$(-0.3601, 0.005011)$
apptYear of *2017*	0.3107	$(0.0000, 0.3873)$
bloodType of $B+$	0.06942	$(0.0000, 0.1461)$
apptSeas of *Summer*	–0.08098	$(-0.186, 0.08692)$
apptBookType of *NCC unknown*	0.4996	$(0.4021, 0.6462)$
apptBookType of *NCC outbound*	–0.7836	$(-0.8889, -0.635)$
centreLoc	–0.03121	$(-0.06249, 0.0000)$
centreType	–0.1653	$(-0.2091, -0.1252)$
apptMth of *December*	0.161	$(0.0000, 0.2716)$
apptBookType of *NCC inbound*	–0.4067	$(-0.5105, -0.2535)$
lastDAEVVR	–0.1913	$(-0.2781, -0.09297)$
apptMth of *February*	–0.06906	$(-0.2517, 0.0119)$
apptWkday of *Friday*	–0.07165	$(-0.1102, -0.02197)$
outbndUnknown	0.597	$(0.0000, 1.141)$
apptMth of *January*	–0.05212	$(-0.2215, 0.05397)$
apptMth of *June*	–0.04275	$(-0.09215, 0.0000)$
$\log\left(\text{lastDefLen} + 1\right)$	0.01786	$(0.0000, 0.02852)$
$\log\left(\text{lastDefEndLen} + 1\right)$	–0.01404	$(-0.02346, 0.0000)$
$\log\left(\text{pltDon} + 1\right)$	0.03502	$(0.0000, 0.06323)$

Table 7.3: *Inferential summary of the linear coefficients for the Australian Red Cross Blood Service dataset, using the MCMC approach as given by Algorithm 9.*

effect predictors, 28 are in common with MCMC approach (Algorithm 9). Among the 5 non-linear effect predictors, 4 are in common with MCMC approach. Predictor

	posterior mean	95% credible interval
apptMth of *March*	–0.0543	$(-0.1090, 0.0000)$
lastDAEnonVVR	–0.0949	$(-0.2063, 0.0000)$
bloodType of $O-$	–0.0856	$(-0.1380, 0.0000)$
apptBookType of *other*	–0.1091	$(-0.2170, 0.0000)$
inbndPhone	0.1136	$(0.0839, 0.1455)$
outbndPhone	0.1293	$(0.0970, 0.1605)$
colTypeSched of *Plateletpheresis Collection*	–0.1568	$(-0.2512, -0.0784)$
sex	0.0646	$(0.0366, 0.0942)$
inbndSMS	–0.3547	$(-0.4008, -0.3051)$
outbndSMS	0.3037	$(0.2804, 0.3277)$
bloodType of *unknown*	–1.0179	$(-1.2695, -0.7499)$
apptBookType of *Web Booking*	–1.8091	$(-2.1848, -1.4384)$
apptWkday of *Wednesday*	–0.0321	$(-0.0792, 0.0000)$

Table 7.3 continued: *Inferential summary of the linear coefficients for the Australian Red Cross Blood Service dataset, using the MCMC approach as given by Algorithm 9.*

$\log\Big(\text{lastDefLen} + 1\Big)$ is selected to have linear effect in MCMC, but it is selected to have non-linear effect in MFVB.

We obtained the inferential summary for all the selected linear fits using the MFVB approach as shown in Table 7.5. The descriptions of these predictors are listed in Table 7.1. Among the 31 selected linear predictors, 28 predictors are also selected by the MCMC approach shown in Table 7.3.

The results indicate that apptYear of *2017*, outbndSMS, apptBookType of *In centre*, inbndPhone and outbndPhone each have positive impacts on the probability of donor attendance. Conversely, colTypeTaken of *no blood taken*, centreType, apptBookType of *NCC inbound*, lastDAEVVR, apptMth of *February*, colTypeSched of *Plateletpheresis Collection*, inbndSMS and apptBookType of *Web Booking* each have negative impacts on the probability of donor attendance.

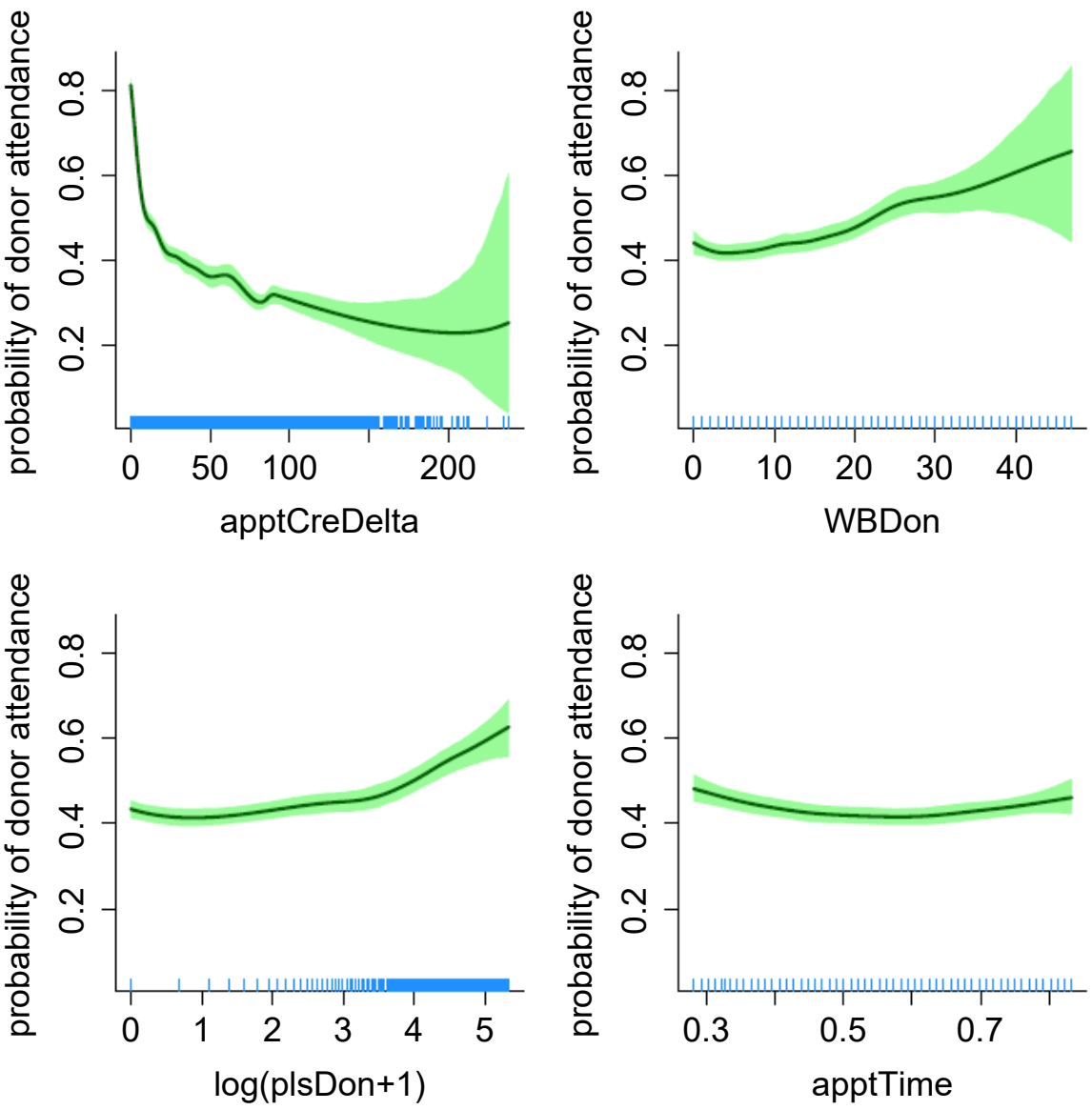


Figure 7.3: *The estimated non-linear effects of* ***apptCreDelta*** *(dates between the booking and donation),* ***WBDon*** *(the number of whole blood donations),* $\log\left(\textbf{plsDon}+1\right)$ *(the logarithm of the number plasma donations plus one) and* ***apptTime*** *(proportion of a day since midnight in hours) for the Australian Red Cross Blood Service dataset from application of Algorithm 9 and effect type estimation rules of Section 6.7.2 with* $\tau = 0.5$*, with* $1,000$ *samples as warm up and* $1,000$ *samples as retained. Each curve is the slices of estimated probability of donor attendance as a function of the predictor, with all other selected predictors set to their median values. The shade region corresponds to pointwise approximate* 95% *credible intervals. The rug at the base of the plot show values of each predictor.*

There are five predictors having non-linear effects on the probability of donor at-

Predictors selected as having linear effects:
colTypeTaken of *no blood taken*, colTypeTaken of *plateletpheresis taken*, age, apptYear of *2017*, bloodType of $B+$, apptBookType of *NCC unknown*, apptBookType of *NCC outbound*, centreLoc, centreType, apptMth of *December*, apptBookType of *NCC inbound*, lastDAEVVR, apptMth of *February*, apptWkday of *Friday*, apptBookType of *In centre*, outbndUnknown, apptMth of *January*, apptMth of *June*, apptMth of *March*, apptWkday of *Monday*, lastDAEnonVVR, bloodType of $O-$, inbndPhone, outbndPhone, colTypeSched of *Plateletpheresis Collection*, sex, inbndSMS, outbndSMS, apptWkday of *Tuesday*, bloodType of *unknown*, apptBookType of *Web Booking*.
Predictors selected as having non-linear effects:
apptCreDelta, WBDon, $\log\Big(\mathsf{plsDon}+1\Big)$, $\log\Big(\mathsf{lastDefLen}+1\Big)$, apptTime.

Table 7.4: *Categorization results obtained from Red Cross Blood Service application dataset using the MFVB approach as given by Algorithm 10.*

tendance, which are apptCreDelta, WBDon, $\log\Big(\mathsf{plsDon}+1\Big)$, $\log\Big(\mathsf{lastDefLen}+1\Big)$ and apptTime. The descriptions of these predictors are listed in Table 7.1. We obtained Figure 7.4 to visualize these non-linear effects. The first panel for predictor apptCreDelta, the second panel for predictor WBDon, the third panel for predictor $\log\Big(\mathsf{plsDon}+1\Big)$ and the fifth panel for predictor apptTime closely resemble the ones in Figure 7.3 using MCMC approach. The fourth panel for predictor $\log\Big(\mathsf{lastDefLen}+1\Big)$ shows that the probability of donor attendance remains steady at approximately 0.4 in the range from 0 and 3 and then gradually decreases from 0.4 to 0.3 between 3 and 7.

In summary, using the MFVB approach (Algorithm 10) takes approximately 20% of the model selection processing time as using the MCMC approach (Algorithm 9). MCMC method selected 36 variables and MFVB selected 41, with 33 variables in

	posterior mean	95% credible interval
colTypeTaken of *no blood taken*	–0.1265	$(-0.153, -0.09987)$
colTypeTaken of *plateletpheresis taken*	–0.07279	$(-0.1208, -0.02504)$
age	0.01074	$(0.01025, 0.01123)$
apptYear of *2017*	0.3651	$(0.3491, 0.3814)$
bloodType of $B+$	0.06408	$(0.03548, 0.09304)$
apptBookType of *NCC unknown*	0.5562	$(0.5046, 0.6075)$
apptBookType of *NCC outbound*	–0.703	$(-0.7587, -0.648)$
centreLoc	–0.02594	$(-0.04108, -0.01049)$
centreType	–0.1528	$(-0.1761, -0.1296)$
apptMth of *December*	0.0787	$(0.05456, 0.1027)$
apptBookType of *NCC inbound*	–0.3214	$(-0.3745, -0.2681)$
lastDAEVVR	–0.1954	$(-0.25, -0.1416)$
apptMth of *February*	–0.163	$(-0.1921, -0.1345)$
apptWkday of *Friday*	–0.05132	$(-0.07235, -0.03068)$
apptBookType of *In centre*	0.1206	$(0.07229, 0.1688)$
outbndUnknown	0.5979	$(0.2858, 0.9004)$
apptMth of *January*	–0.1403	$(-0.1706, -0.1103)$
apptMth of *June*	–0.05898	$(-0.08532, -0.03349)$
apptMth of *March*	–0.06999	$(-0.09966, -0.04098)$
apptWkday of *Monday*	0.03939	$(0.01962, 0.05923)$
lastDAEnonVVR	–0.1277	$(-0.1928, -0.06271)$
bloodType of $O-$	–0.07107	$(-0.09282, -0.04896)$
inbndPhone	0.1088	$(0.08761, 0.1298)$
outbndPhone	0.1197	$(0.09738, 0.1417)$
colTypeSched of *Plateletpheresis Collection*	–0.1131	$(-0.1592, -0.06768)$

Table 7.5: *Inferential summary of the linear coefficients for the Australian Red Cross Blood Service dataset, using the MFVB approach as given by Algorithm 10.*

	posterior mean	95% credible interval
sex	0.06275	$(0.04838, 0.07768)$
inbndSMS	–0.3536	$(-0.3853, -0.3213)$
outbndSMS	0.285	$(0.2687, 0.3017)$
apptWkday of *Tuesday*	0.03824	$(0.01902, 0.05762)$
bloodType of *unknown*	–0.9712	$(-1.107, -0.8337)$
apptBookType of *Web Booking*	–1.626	$(-1.792, -1.466)$

Table 7.5 continued: *Inferential summary of the linear coefficients for the Australian Red Cross Blood Service dataset, using the MFVB approach as given by Algorithm 10.*

common between these two approaches.

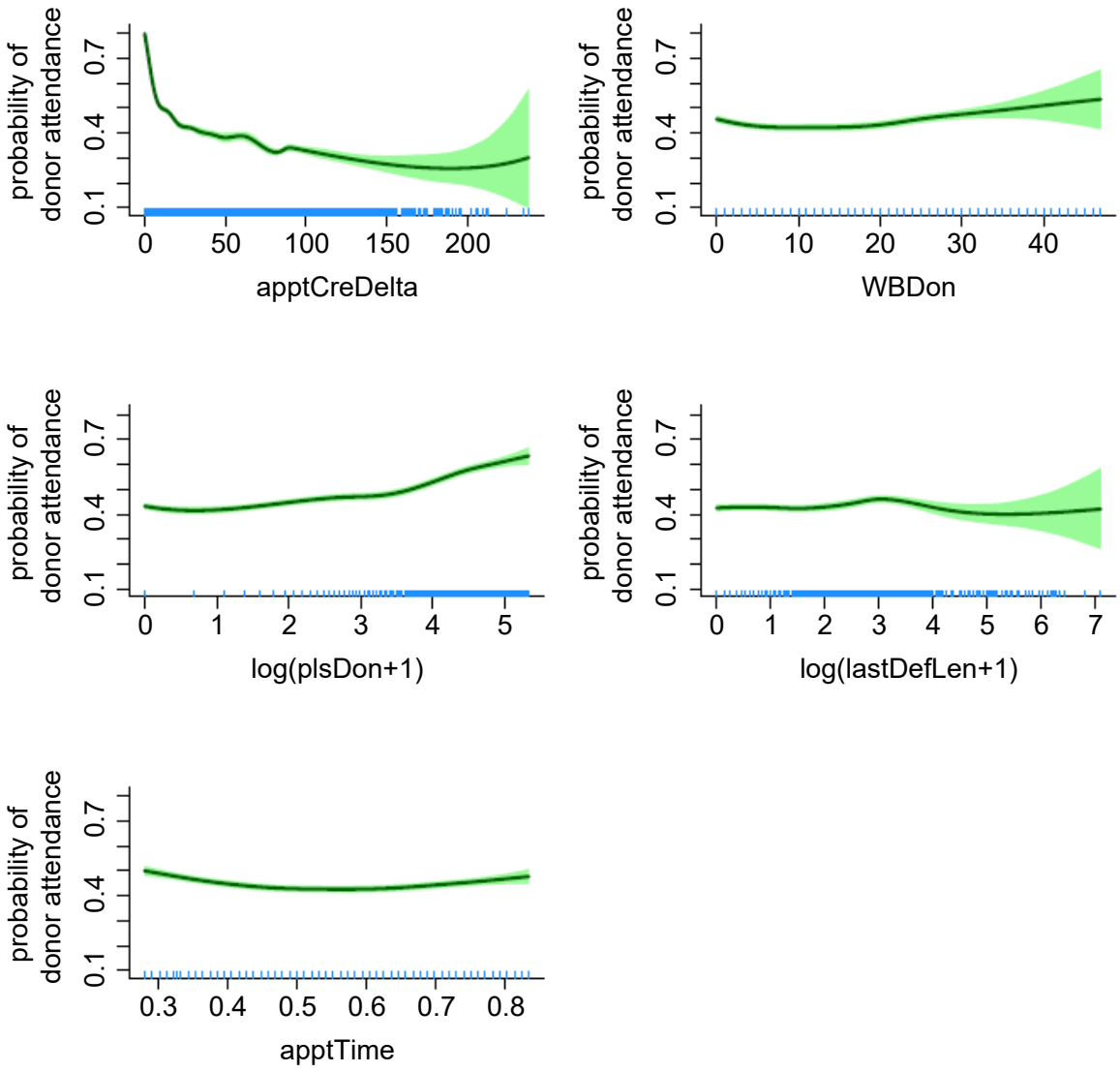


Figure 7.4: *The estimated non-linear effects of* ***apptCreDelta*** *(dates between the booking and donation),* ***WBDon*** *(the number of whole blood donations),* $\log\left(\textbf{\textit{plsDon}}+1\right)$ *(the logarithm of the number plasma donations plus one),* $\log\left(\textbf{\textit{lastDefLen}}+1\right)$ *(the logarithm of the number of weeks since the donor last deferred an appointment plus one) and* ***apptTime*** *(proportion of a day since midnight in hours) for the Australian Red for the Australian Red Cross Blood Service dataset from application of Algorithm 10 and effect type estimation rules of Section 6.7.2 with* $\tau = 0.1$*. The curve is the slice of estimated probability of donor attendance as a function of the predictor, with all other selected predictors set to their median values. The shaded regions correspond to pointwise approximate* 95% *credible intervals, but are subject to considerable mean field approximation error.*

Chapter 8

Conclusion

This thesis presents fully automatic and accurate Bayesian three-category GAM selection methodology with improved speed and scalability, to categorize each candidate predictor's effect to be zero, linear or non-linear on the mean response. The framework supports both Gaussian and binary response cases. Both the MCMC and the faster MFVB approach are implemented in our R package gamselBayes. Our approaches are inspired by the regularization path approach (Chouldechova and Hastie, 2015), which is a frequentist approach, based on the group LASSO and the Demmler-Reinsch spline bases. Instead of relying on the regularization path and cross-validation, we incorporate their framework into a Bayesian graphical model and apply Bayesian principles to perform model selection, allowing us to avoid the practical difficulties of finding the cross-validation minima and to explore a bigger model space compared with the regularization path approach. We use the misclassification rate as the primary performance metric, representing the proportion of times that a predictor is misclassified as zero, linear or non-linear effect based on its true simulated effect.

In Chapter 2, we introduced the use of Laplace-Zero/Grouped Lasso-Zero priors and tailored auxiliary variable representations, enabling Gibbs sampling for the MCMC and implemented Algorithm 2 in our R package gamselBayes, for the Bayesian three-category GAM selection on Gaussian and binary responses. Using the model selection rules stated in (2.54), our simulation study suggests a good default threshold value of 0.5 for the MCMC (see Figure 2.3). We applied Algorithm 2 to two real-world datasets – the mortgage dataset and the educational dataset, to demonstrate the practical applicability of the approach.

In Chapter 3, we investigated the properties of Grouped Horseshoe distribution and explored the application of the (Grouped) Horseshoe priors via MCMC for the Bayesian three-category GAM selection on Gaussian responses (Algorithm 4) using Gibbs sampling, as well as on Poisson responses (Algorithm 5) and Negative-Binomial responses (Algorithm 6), both using slice sampling. The score function behaviour and robustness to large signals of the univariate Horseshoe distribution are seen to transfer to the grouped case (see **Result 3.2.3.1** and **Result 3.2.4.1**). However, the super-efficiency based on risk rates of convergence does not extend to grouped case (see **Result 3.2.5.1**). Further, we conducted a simulation study for classification performance comparison in terms of accuracy. For Gaussian responses, using the model selection rules stated in (3.74) for the (Grouped) Horseshoe priors and the rules stated in (2.54) for the Laplace-Zero/Grouped Lasso-Zero priors, Figures 3.5–3.8 indicate that the (Grouped) Horseshoe priors generally underperform relative to the Laplace-Zero/Grouped Lasso-Zero priors described in Chapter 2. For count responses, we use the model selection rules stated in (3.75). For Poisson responses, Figures 3.9 shows that the (Grouped) Horseshoe priors with k-means clustering used to determine the classification border performs better than the (Grouped) Horseshoe priors with fixed border of 0.5 and the default version of the spikeSlabGAM approach. For Negative-Binomial responses, Figure 3.10 shows that the (Grouped) Horseshoe priors with k-means clustering used to determine the classification border performs better than the use of a fixed border of 0.5.

In Chapter 4, following the work of Chapter 2 via the MCMC, with use of the Laplace-Zero/Grouped Lasso-Zero priors and tailored auxiliary variable representations, we derived the closed forms updates for the MFVB algorithm and implemented Algorithm 7 as the faster option in our R package gamselBayes, for the Bayesian three-category GAM selection on Gaussian and binary responses. Using the model selection rules stated in (2.54), our simulation study suggests a good default threshold value of 0.1 for the MFVB (see Figure 4.1). We applied Algorithm 7 to the car auction data and to the Sydney real estate data.

In Chapter 5, we conducted a simulation study for classification performance comparison in terms of accuracy and speed, among our MCMC given in Algorithm 2 as described in Chapter 2, our MFVB given in Algorithm 7 as described in Chap-

ter 4 and the default version of the spikeSlabGAM approach as described in Section 5.2.1, for the three-category GAM selection on Gaussian and binary responses. Figure 5.1 shows both MCMC and MFVB perform similarly to spikeSlabGAM for low noise level on Gaussian responses. For high noise level on Gaussian responses, our MFVB exhibits slightly reduced accuracy but improved speed and scalability. Figure 5.2 shows that both MCMC and MFVB outperform spikeSlabGAM on binary responses. Table 5.1 shows although spikeSlabGAM can yield strong classification result, it is comparatively slow and lacks scalability to large data sets, relative to our MCMC and MFVB methods.

In Chapter 6, we extended the Bayesian three-category GAM selection via the MCMC and the MFVB, as described in Chapter 2 and Chapter 4 respectively, to GAMM selection with random intercept. Using the Laplace-Zero/Grouped Lasso-Zero priors and tailored auxiliary variable representations, we derived Gibbs sampling for MCMC (Algorithm 9) and closed forms updates for the MFVB (Algorithm 10). Using the model selection rules stated in (6.84), our simulation study suggests a good default threshold value of 0.5 for the MCMC (see Figure 6.3) and 0.1 for the MFVB (see Figure 6.5). We also conducted a simulation study for classification performance comparison in terms of accuracy and speed, among our MCMC (Algorithm 9), our MFVB (Algorithm 10) and the default version of the spikeSlabGAM approach as described in Section 5.2.1. For Gaussian responses, our methods match the performance of the existing spikeSlabGAM approach at smaller sample size and outperform it as sample size increases (see Figure 6.7). For binary responses (see Figure 6.8), for smaller sample size, both MCMC and MFVB approaches show better classification performance. For larger sample size, these three methods all have similar performance. A comparison of computational time in seconds (see Table 6.1) shows both MCMC and MFVB are faster and scale well across simulation settings for the GAMM selection with random intercept problem.

In Chapter 7, we demonstrated the application of our Bayesian three-category GAMM selection with random intercept described in Chapter 6, to the Australian Red Cross Blood Service data via the MCMC and the MFVB. This is a very large and complex dataset with repeated measure. Both of our MCMC (Algorithm 9) and MFVB (Algorithm 10) for the Bayesian three-category GAMM selection with random intercept scale efficiently to the very large dataset and offer fast computational

performance.

In conclusion, this thesis contributes to the development of robust and scalable methodologies for Bayesian three-category GAM selection. Specifically, we have implemented both MCMC and MFVB algorithms in the public available R package gamselBayes, supporting Gaussian and binary responses. In addition, we investigated the Grouped Horseshoe distribution properties and its use in Bayesian model selection, and extended the framework to Bayesian GAMM selection with random intercept.

For future research work in the Bayesian three-category GAM selection with extension to GAMM selection, we aim to introduce random effects into coefficients and correlations on Gaussian and binary responses. Moreover, we aim to explore robust slice sampling method for Poisson responses under the Bayesian framework.

References

Abramowitz, M. and Stegun, I. A., editors (1968). *Handbook of Mathematical Functions*. New York: Dover.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, **88**:669–679.

Azzalini, A. (2021). The R package 'sn': The Skew-Normal and related distributions such as the Skew-t and the Unified Skew-Normal (version 2.0.0). http://azzalini.stat.unipd.it/SN.

Bagot, K. L., Bove, L. L., Masser, B. M., Bednall, T. C., and Buzza, M. (2013). Perceived deterrents to being a plasmapheresis donor in a voluntary, nonremunerated environment. *Transfusion*, **53**:1108–1119.

Bai, R. and Ghosh, M. (2018). High dimensional multivariate posterior consistency under global-local shrinkage priors. *Journal of Multivariate Analysis*, **167**:157–170.

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2017). The horseshoe+ estimator for ultra-sparse signals. *Bayesian Analysis*, **12**:1105–1131.

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2019). Lasso meets horseshoe: a survey. *Statistical Science*, **34**:405–427.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

Boksmati, N., Butler-Henderson, K., Anderson, K., and Sahama, T. (2016). The effectiveness of SMS reminders on appointment attendance: a meta-analysis. *Journal of Medical Systems*, **40**:article number 90.

Bürkner, P. C. (2022). bmrs 2.18.0: Bayesian regression models using Stan. R package. https://r-project.org.

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

Charbonneau, J., Cloutier, M. S., and Carrier, É. (2015). Why do blood donors lapse or reduce their donation's frequency? *Transfusion Medicine Reviews*, **30**:1–5.

Chouldechova, A. and Hastie, T. (2015). Generalized additive model selection. https://arXiv.org/abs/1506.03850v2.

Chouldechova, A. and Hastie, T. (2022). gamsel 1.8: Fit regularization path for generalized additive models. R package. https://r-project.org.

Clarke, B. and Barron, A. R. (1990). Information-theoretic asymptotics of Bayes methods. *IEEE Transactions of Information Theory*, **36**:453–471.

Dechter, R. and Pearl, J. (1988). *Network-Based Heuristics for Constraint-Satisfaction Problems*. New York: Springer.

Eddelbuettel, D. and François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, **40**:1–18.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**:398–409.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**:721–741.

Gemelli, C. N., Hayman, J., and Waller, D. (2017). Frequent whole blood donors: understanding this population and predictors of lapse. *Transfusion*, **57**:108–114.

Gordy, M. B. (1998). A generalization of generalized beta distributions. In *Finance and Economics Discussion Series.* Board of Governors of the Federal Reserve System, United States of America.

Gradshteyn, I. S. and Ryzhik, I. M. (1994). *Tables of Integrals Series, and Products, 5th Edition*. New York: Academic Press.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. London: Chapman and Hall.

Harezlak, J., Ruppert, D., and Wand, M. P. (2018). *Semiparametric Regression with R*. New York: Springer.

Hastie, T., Tibshirani, R., and Tibshirani, R. (2020). Best subset, forward stepwise or lasso? Analysis and recommendations based on extensive comparisons. *Statistical Science*, **35**:579–592.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. New York: Chapman and Hall.

He, V. X. and Wand, M. P. (2024). Bayesian generalized additive model selection including a fast variational option. *Advances in Statistical Analysis*, **108**:639–668.

He, V. X. and Wand, M. P. (2025). The grouped horseshoe distribution and its statistical properties. *Stat*, **14**:e70050, 1–6.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *The Annals of Statistics*, **33**:730–733.

Jordan, M. I. (2004). Graphical models. *Statistical Science*, **19**:140–155.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, **37**:183–233.

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, **5**:369–412.

Ligges, U., Thomas, A., Spiegelhalter, D., Best, N., Lunn, D., Rice, K., and Sturtz, S. (2009). BRugs 0.5. *OpenBUGS* and its *R/S-PLUS* interface *BRugs*.

Luts, J. and Wand, M. P. (2015). Variational inference for count response semiparametric regression. *Bayesian Analysis*, **10**:991–1023.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, **1**:281–297.

McGrory, C. A. and Titterington, D. (2007). Variational approximations in Bayesian model selection for finite mixture distributions. *Computational Statistics and Data Analysis*, **51**:5352–5367.

McLean, M. W. and Wand, M. P. (2019). Variational message passing for elaborate response regression models. *Bayesian Analysis*, **14**:371–398.

Michael, J. R., Schucany, W. R., and Haas, R. W. (1976). Generating random variates using transformations with multiple roots. *The American Statistician*, **30**:88–90.

Neal, R. (2003). Slice sampling (with discussion). *The Annals of Statistics*, **31**:705–767.

Neville, S. E., Ormerod, J., and Wand, M. P. (2014). Mean field variational Bayes for continuous sparse signal shrinkage: pitfalls and remedies. *Electronic Journal of Statistics*, **8**:1113–1151.

Ngo, L. and Wand, M. P. (2004). Smoothing with mixed model software. *Journal of Statistical Software*, **9**:1–54.

Olver, F. W., Olde Daalhuis, D., Lozier, D. W., Schneider, B. I., Boisvert, C. W., Clark, C. W., Miller, B. R., Saunders, B. V., Cohl, H. S., and McClain, M. A., editors (2023). *U.S. National Institute of Standards and Technology Digital Library of Mathematical Functions.* Release 1.1.11., https://dlmf.nist.gov.

Ormerod, J. T. and Wand, M. P. (2010). Explaining variational approximations. *The American Statistician*, **64**:140–153.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statistical Science*, **1**:502–527.

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

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. San Mateo, California: Morgan Kaufmann.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society, Series B*, **71**:1009–1030.

Reich, B. J., Sorlie, C. B., and Bondell, H. D. (2009). Variable selection in smoothing spline ANOVA: application to deterministic computer codes. *Technometrics*, **51**:110–120.

Robert, C. P. (1995). Simulation of truncated normal variates. *Statistics and Computing*, **5**:121–125.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. New York: Springer.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. New York: Cambridge University Press.

Scheipl, F. (2011). spikeSlabGAM: Bayesian variable selection, model choice and regularization for generalized additive mixed models in R. *Journal of Statistical Software*, **43**:1–24.

Scheipl, F. (2022). spikeSlabGAM 1.1: Bayesian variable selection and model choice for generalized additive mixed models. R package. https://github.com/fabian-s/spikeSlabGAM.

Scheipl, F., Fahrmeir, L., and Kenib, T. (2012). Spike-and-slab priors for function selection in structured additive regression models. *Journal of the American Statistical Association*, **107**:1518–1532.

Searle, S. R. (1982). *Matrix Algebra Useful for Statistics*. New York: Wiley.

Shively, T. S., Kohn, R., and Wood, S. (1999). Variable selection and function estimation in additive nonparametric regression using a data-based prior. *Journal of the American Statistical Association*, **94**:777–794.

Stan Development Team (2015). Stan: A C++ Library for probability and sampling, version 2.6.

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

Titterington, D. M. (2004). Bayesian methods for neural networks and related methods. *Statistical Science*, **19**:128–139.

Umlauf, N., Klein, N., Zeileis, A., and Simon, T. (2022a). bamlss 1.1: Bayesian additive models for location, scale, and shape (and beyond). R package. https://www.bamlss.org.

Umlauf, N., Kneib, T., and Klein, N. (2022b). BayesX 0.3: R utilities accompanying the software package BayesX. R package. https://www.BayesX.org.

Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia: SIAM.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families and variational inference. *Foundations and Trends in Machine Learning*, **1**:1–305.

Wand, M. P. and Jones, M. C. (1993). Comparison of smoothing parameterizations in bivariate density estimation. *Journal of the American Statistical Association*, **88**:520–528.

Wand, M. P. and Ormerod, J. T. (2008). On semiparametric regression with O'Sullivan penalized splines. *Australian and New Zealand Journal of Statistics*, **50**:179–198.

Wand, M. P. and Ormerod, J. T. (2011). Penalized wavelets: embedding wavelets into semiparametric regression. *Electronic Journal of Statistics*, **5**:1654–1717.

Wand, M. P. and Ormerod, J. T. (2012). Continued fraction enhancement of Bayesian computing. *Stat*, **1**:31–41.

Wand, M. P., Ormerod, J. T., Padoan, S. A., and Fruhwirth, R. (2012). Mean field variational Bayes for elaborate distributions. *Bayesian Analysis*, **7**:847–900.

Wang, B. and Titterington, D. (2006). Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. *Bayesian Analysis*, **1**:625–650.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R, Second Edition*. Boca Raton, Florida: CRC Press.

Xu, Z., Schmidt, D. F., Enes, M., Qian, G., and Hopper, J. L. (2016). Bayesian grouped horseshoe regression with application to additive models. In B.H. Kang and Q.Bai (eds.). *AI 2016: Advances in Artificial Intelligence*, pp. 229–240, Cham, Switzerland: Springer.

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

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, **68**:49–67.

Zhang, H. H. and Lin, C.-Y. (2013). cosso: Fit regularized nonparametric regression models using COSSO. R package version 2.1. https://www.r-project.org.