University of Wollongong

Research Online

University of Wollongong Thesis Collection 1954-2016

University of Wollongong Thesis Collections

2015

Variational inference machines for semiparametric regression

Shen Wang
University of Wollongong

Follow this and additional works at: https://ro.uow.edu.au/theses

University of Wollongong

Copyright Warning

You may print or download ONE copy of this document for the purpose of your own research or study. The University does not authorise you to copy, communicate or otherwise make available electronically to any other person any copyright material contained on this site.

You are reminded of the following: This work is copyright. Apart from any use permitted under the Copyright Act 1968, no part of this work may be reproduced by any process, nor may any other exclusive right be exercised, without the permission of the author. Copyright owners are entitled to take legal action against persons who infringe their copyright. A reproduction of material that is protected by copyright may be a copyright infringement. A court may impose penalties and award damages in relation to offences and infringements relating to copyright material. Higher penalties may apply, and higher damages may be awarded, for offences and infringements involving the conversion of material into digital or electronic form.

Unless otherwise indicated, the views expressed in this thesis are those of the author and do not necessarily represent the views of the University of Wollongong.

Recommended Citation

Wang, Shen, Variational inference machines for semiparametric regression, Doctor of Philosophy thesis, School of Mathematics and Applied Statistics, University of Wollongong, 2015. https://ro.uow.edu.au/theses/4666

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au

UNIVERSITY OF WOLLONGONG

Variational Inference Machines for Semiparametric Regression

Shen Wang

M Stat (Distinction), University of Wollongong

B Math, Beijing University of Technology

Supervisor:

Prof. Matt Wand

This thesis is presented as part of the requirements for the conferral of the degree:

Doctor of Philosophy

The University of Wollongong

School of Mathematics and Applied Statistics

December 2015

Declaration

I, Shen Wang, declare that this thesis submitted in partial fulfilment of the requirements for the conferral of the degree Doctor of Philosophy, from the University of Wollongong, is wholly my own work unless otherwise referenced or acknowledged. This document has not been submitted for qualifications at any other academic institution.

Shen Wang

May 2, 2016

Abstract

Variational approximation methods are enjoying an increasing amount of development and use in statistical problems. In the Bayesian field, we develop mean field variational Bayes (MFVB) algorithms that perform variable selection and fit complicated regression models. We also produce a new Bayesian inference software, `InferMachine()`, which can perform the MFVB inference using `BRugs` model code. Finally, a new computational framework, Infer.NET, for approximate Bayesian inference in hierarchical Bayesian models is demonstrated. We assess the accuracy of MFVB via comparison with a Markov chain Monte Carlo (MCMC) baseline. The simulation results show that the results of the MFVB inference agree with those of the MCMC approach. In the non-Bayesian field, the precise asymptotic distributional behaviour of Gaussian variational approximate estimators in a single predictor Poisson mixed model is derived. A simulation study shows that the Gaussian variational approximate confidence intervals possess good to excellent coverage properties.

Acknowledgments

I would like to sincerely thank the University of Wollongong for its generous grants. I truly acknowledge the valuable time, patience and support of my supervisory, Professor Matt Wand. I am deeply grateful for the continued guidance and supervision. He taught me the fundamentals of academic life and how to be a good researcher. Professor Wand guided me to enhance the quality of this work and present it in the best possible way. Professor David Steel, must also be thanked for his valuable advice. I also thank Professor Kenneth Russell for proofreading of hundreds of thesis drafts. Many thanks go to the staff and research students of the University of Wollongong, with whom I interacted and who made this research journey enjoyable, despite the difficulties I experienced.

Contents

[1]This chapter is based on: Wang, S.S.J. and Wand, M.P. Using Infer.NET for statistical analyses. *The American Statistician*, 65, 2 (2011), 115-126.

[2]This chapter is based on: Hall, P., Pham, T., Wand, M.P. and Wang, S.S.J. Asymptotic Normality and Valid Inference for Gaussian Variational Approximation. *The Annals of Statistics*, (2011), 39, 2502-2532.

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Literature Review

Variational approximation methods have recently enjoyed increasing use and development in statistical problems (Jordan, Ghahramani, Jaakkola and Saul, 1999). The mean field variational Bayes (MFVB) method uses variational approximation methods for inference in a hierarchical Bayesian model, and is a fast, deterministic alternative to Markov chain Monte Carlo (MCMC). In frequentist fields, the variational approximation achieves satisfactory results for generalized linear mixed model analysis (Ormerod and Wand, 2012). Gaussian variational approximation (Hall, Ormerod and Wand, 2011) is a relatively simple, fast, natural alternative to Laplace approximation for maximum likelihood estimation. This literature review firstly explains the origins of variational approximation and introduces developments in the computer science field. Secondly, the use of MFVB is reviewed and suggestions of future work in this field are presented. Finally, I summarise the latest progress in variational approximation for maximum likelihood estimation.

1.1.1 Variational approximations

Variational methods have a long history of use in physics, mathematics and control theory (Jaakkola and Jordan, 2000). The variational methods for approximating intractable computations have roots in the calculus of variations and include a wide range of tools for evaluating integrals and functionals. Generally, the calculus of variations involves optimizing a functional over a given class of functions.

Variational approximations were first explored and used in the field of computer science in the 1990s. Jordan, Ghahramani, Jaakkola and Saul (1999) introduced the use of variational methods for inference and learning in graphical models. They showed the relationship between variational approximations and graphical models and demonstrated how variational algorithms can be formulated using different models. Finally, they described a general framework for generating variational transformations based on convex duality.

Bishop (2006) included two chapters about graphical models and approximate inference which established a preliminary systematic theoretical system of variational approximations inference for graphical models, including the relationship between hierarchical Bayesian models and graphical models. This system also introduced the method of variational approximations and a general framework for the derivation of the variational Bayes approximation model based on the Kullback-Leibler divergence between the true and approximating distributions. Although variational approximations are commonly used in machine learning models, Bishop (2006) indicated that variational approximations are efficient alternatives to MCMC for Bayesian inference.

Based on variational inference methods, some variational inference engines (Wand, 2009) were developed for conducting inference in graphical models. These included Variational Inference for BayESian networks (VIBES) (Bishop *et al.*, 2003) and Infer.NET (Minka *et al.*, 2014).

1.1.2 Mean field variational Bayes

In recent years, statisticians have started to explore the use of variational approximations for Bayesian inference, and to establish the connections between variational approximate inference and Bayesian models. The concept of mean field variational Bayes (MFVB) originated in statistical physics (Parisi, 1988), where it was called mean field theory. In MFVB, the posterior density function of the parameter vector is factorized into a particular product structure. The approximate posterior density function is obtained by maximizing a lower bound on the marginal likelihood over the restricted space. Ormerod and Wand (2010) explained the use of the MFVB method in Bayesian statistics, and Faes, Ormerod and Wand (2011) established a method to assess the accuracy of MFVB.

More recent works have been focused on two aspects of MFVB: to extend the MFVB method to handle a variety of models; and to assess the accuracy of MFVB in different statistical models. McGrory and Titterington (2007, 2009) performed model selection in finite mixture distributions and fit hidden Markov models. Pham, Ormerod and Wand (2013) fitted nonparametric regression with measurement error. Faes, Ormerod and Wand (2011) discussed MFVB for elaborate distributions. Based on the MFVB method, Ormerod (2011) introduced grid-based variational approximations method for Bayesian inference.

There still remains several statistical models that are largely mysterious and unexplored in MFVB's fields. As an important problem in statistical analysis, variable selection is the choice of an optimal model from a set of a priori plausible models (O'Hara and Sillanpää, 2009). Variable selection for semiparametric regression models consists of two components: model selection for nonparametric components and selection of significant variables for the parametric portion. This area have not been explored in the context of variational approximations. This forms the basis of the thesis. Moreover, we will extend current MFVB methodolo-

gy for Bayesian Lasso regression and linear variable selection models with indicator variables.

1.1.3 Gaussian variational approximation

In non-Bayesian fields, the maximum likelihood method is frequently used to estimate the parameters. However, intractable issues for likelihoods are encountered in a wide range of complex models. Ormerod and Wand (2012) incorporated the variational approximation method into maximum likelihood estimation and introduced the Gaussian variational approximation (GVA) method for fitting generalized linear mixed models (GLMMs). Ormerod and Wand (2012) proposed the point estimation method and showed that the estimation of Gaussian variational approximations for the grouped GLMM model is very accurate. Hall, Ormerod and Wand (2011) proved consistency for Poisson mixed models. However, there is almost no literature on asymptotic validity of variational inference methods. Hall, Pham, Wand and Wang (2011) derived the precise asymptotic distributional behavior of Gaussian variational approximate estimators of the parameters in a single-predictor Poisson mixed model.

1.2 Basics of Mean Field Variational Bayes

The process of MFVB is summarised in Chapter 10 of Bishop (2006) and Ormerod and Wand (2010). Consider a generic Bayesian model with parameter $\boldsymbol{\theta}$ and observed data $\boldsymbol{y}$. We suppose that $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_M)$ is continuous and is in the set Θ. The posterior density function of the parameter $\boldsymbol{\theta}$, $p(\boldsymbol{\theta}|\boldsymbol{y})$, can be obtained by

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

where $p(\boldsymbol{y})$ is the marginal likelihood. Except in some simple models, the posterior density function for each $\boldsymbol{\theta}_m$, $1 \leq m \leq M$, is difficult to obtain.

Let q be an arbitrary density function over Θ, We can obtain the q-dependent lower bound on the marginal likelihood as

$$\underline{p}(\boldsymbol{y}; q) \equiv \exp \int q(\boldsymbol{\theta}) \log \left\{ \frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right\} d\boldsymbol{\theta}$$

and

$$p(\boldsymbol{y}) \geq \underline{p}(\boldsymbol{y}; q)$$

with equality if and only if $q(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{y})$ almost everywhere.

The MFVB approximation assumes that the posterior density function $p(\boldsymbol{\theta}|\boldsymbol{y})$ can be factorized for some partition $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_M)$, i.e.

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \approx \prod_{m=1}^{M} q(\boldsymbol{\theta}_m)$$

.

The lower bound for this factorization is given by

$$\log \underline{p}(\boldsymbol{y}; q) = \int \prod_{m=1}^{M} q(\boldsymbol{\theta}_m) \left\{ \log(p(\boldsymbol{y}, \boldsymbol{\theta})) - \sum_{m=1}^{M} \log(q(\boldsymbol{\theta}_m)) \right\} d\boldsymbol{\theta}_1, ..., d\boldsymbol{\theta}_M$$

Maximizing $\log \underline{p}(\boldsymbol{y}; q)$ over each of $q_1, ..., q_M$ leads to the optimal posterior density function. Algorithm updates can be derived from the expression

$$q^*(\boldsymbol{\theta}_m) \propto \exp \left\{ E_{q(\boldsymbol{\theta}_{-m})} \log p(\boldsymbol{\theta}_m | \boldsymbol{y}, \boldsymbol{\theta}_{-m}) \right\}, \quad 1 \leq m \leq M, \tag{1.1}$$

where $\boldsymbol{\theta}_{-m}$ denotes the set $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_M\}$ with $\boldsymbol{\theta}_m$ excluded. Each iteration results in an increase in $\log \underline{p}(\boldsymbol{y}; q)$, and this quantity can be used to assess convergence. In this thesis, we use a high fixed number of iterations to ensure conver-

gence. Convexity properties can be used to show that convergence to at least local optima is guaranteed (Boyd and Vandenberghe, 2004), and each iteration results in an increase in $\log \underline{p}(\boldsymbol{y}; q)$.

1.3 Accuracy Measure

An accuracy measure for MFVB was defined in Faes *et al.* (2011). Suppose that $p(\theta|\boldsymbol{y})$ is the true posterior density function of θ, and $q^*(\theta)$ is the posterior distribution function given by MFVB. The accuracy of MFVB approximate posterior density functions was measured via L_1 distance. The L_1 error, or integrated absolute error (IAE), of $q^*(\theta)$ is given by

$$\text{IAE}(q^*) = \int_{-\infty}^{\infty} |q^*(\theta) - p(\theta|\boldsymbol{y})| d\theta.$$

Because the L_1 error is a scale-independent number between 0 and 2 and is invariant to monotone transformations on the parameter θ, the accuracy of $q^*(\theta)$ is defined to be

$$\text{accuracy}(q^*) = 1 - \frac{1}{2}\,\text{IAE}(q^*).$$

For most Bayesian models, the real posterior density function $p(\theta|\boldsymbol{y})$ is difficult to obtain. In practice, we can obtain an accurate approximate value of $p(\theta|\boldsymbol{y})$ with MCMC sampling by using `BRugs` (Thomas *et al.*, 2006). All examples in this thesis use a burn-in of size 10000, thinning factor of 5 and sampling of size 50000. This is overly large for some models but adequate for all models considered. Density estimates were obtained using the binned kernel density estimate `bkde()` function in the R package `KernSmooth` (Wand, 2015). The bandwidth was chosen using a direct plug-in rule, corresponding to the default version of the `dpik()` function in `KernSmooth`.

1.4 Graphical Models and Factorized Approximation

Our approach to variational inference is based on a factorized approximation for the true posterior distribution. This factorized form of variational inference corresponds to an approximation framework developed in physics called mean field theory (Parisi, 1988). Graphical models allow us to better understand the structure of a Bayesian model and a factorized approximation.

1.4.1 Graphical models

A graph comprises nodes connected by links. In a probabilistic graphical model, each node represents a random variable, and the links express probabilistic relationships between these variables.

Definition 1.1. *A directed cycle is a closed path within the graph such that we can move from node to node along links following the direction of the arrows and end up back at the starting node.*

Definition 1.2. *A graph is called a directed acyclic graph (DAG), if there are no directed cycles in the graph.*

Definition 1.3. *In a DAG, the directed links (arrows) from the nodes correspond to the variables on which the distribution is conditioned.*

Definition 1.4. *Node A is node B's parent, if there are directed links from node A to node B.*

Definition 1.5. *Node A is node B's child, if there are directed links from node B to node A.*

Definition 1.6. *Two nodes are co-parents if they share a common child node.*

Definition 1.7. *The Markov blanket of a node is the set of children, parents and co-parents of that node.*

Figure 1.1 is a DAG showing the Markov blanket of X, where nodes C and D are parent nodes of X, nodes E and F are child nodes of X, and nodes A and B are co-parents with X. The corresponding conditional distributions include $p(X|C, D)$, $p(E|X, A)$ and $p(F|X, B)$.

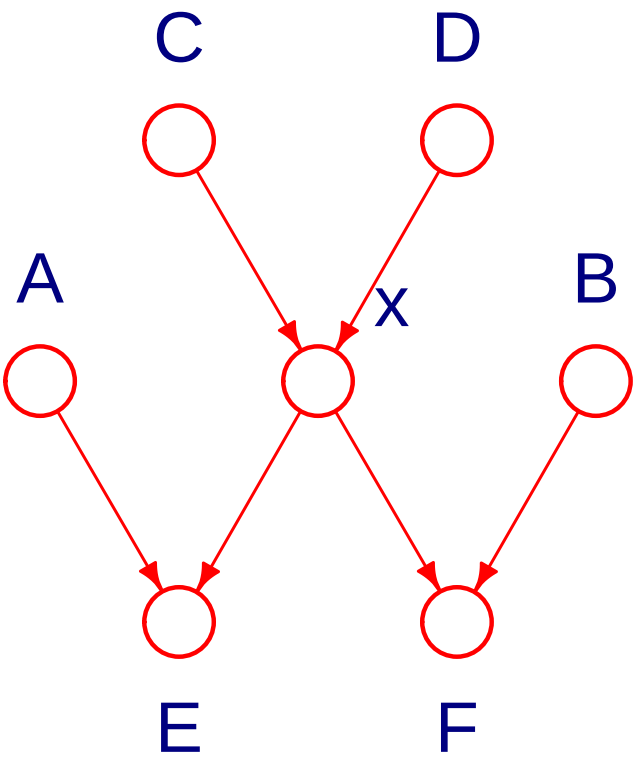

Figure 1.1: The Markov blanket of a node X comprises the set of parents, children and co-parents of the node.

Result 1.1. *The full conditionals for node X involves localized calculations on the Markov blanket. i.e.*

$$p(x|rest) = p(x|Markov\ blanket\ of\ X).$$

Result 1.2. *Using Result 1.1 in the optimal equation (1.1), we can get*

$$q^*(\boldsymbol{\theta}_m) \propto \exp\left\{E_{q(\boldsymbol{\theta}_{-m})} \log p(\boldsymbol{\theta}_m|Markov\ blanket\ of\ \boldsymbol{\theta}_m)\right\}. \tag{1.2}$$

1.4.2 Induced factorizations

Induced factorizations arise from an interaction between the factorization assumed in the variational posterior distribution and the conditional independence properties of the joint distribution of the random variables.

Definition 1.8. *Consider three variables A, B, and C, and suppose that the conditional distribution of A, given B and C, does not depend on the value of B, so that*

$$p(A|B, C) = p(A|C).$$

We say that A is conditionally independent of B given C.

Definition 1.9. *We use the notation $A \perp\!\!\!\perp C \mid B$ to denote that is conditionally independent of B given C.*

The *induced factorizations* arise from an interaction between the factorization assumed in the variational posterior distribution and the conditional independence properties of the true joint distribution. Consider three variables A, B and C. Firstly, we assume a factorization between C and the remaining variables A and B, so that:

$$q(A, B, C) = q(A, B)q(C)$$

Secondly, if A is conditionally independent of B given C, we can factorise this posterior distribution between A and B, so that

$$q(A, B, C) = q(A)q(B)q(C).$$

The ability to recognise induced factorisations helps to streamline derivation of MFVB methodology (Bishop, 2006).

1.5 Notation, Definitions and Results

The following notation is used throughout the thesis.

1.5.1 Vector notation

If $\boldsymbol{a}$ and $\boldsymbol{b}$ are n-dimensional vectors, we write $\boldsymbol{a}$ and $\boldsymbol{b} \in \mathbb{R}^n$, where

$$\boldsymbol{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \text{ and } \boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}.$$

The component-wise product of two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ is:

$$\boldsymbol{a} \odot \boldsymbol{b} = \begin{bmatrix} a_1 b_1 \\ a_2 b_2 \\ \vdots \\ a_n b_n \end{bmatrix}.$$

The norm of a vector $\boldsymbol{a}$ is $\sqrt{\boldsymbol{a}^T\boldsymbol{a}}$. $\text{diag}(\boldsymbol{a})$ is a diagonal matrix with diagonal entries corresponding to those of $\boldsymbol{a}$, i.e.

$$\text{diag}(\boldsymbol{a}) = \begin{pmatrix} a_1 & 0 & \cdots & 0 \\ 0 & a_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_n \end{pmatrix}.$$

1.5.2 Matrix

Let $\boldsymbol{M}$ be a square matrix. Then $\text{tr}(\boldsymbol{M})$ is the trace of $\boldsymbol{M}$ and $|\boldsymbol{M}|$ is the determinant of $\boldsymbol{M}$.

1.5.3 Random variable

Let x and y be random variable. Then $p(x)$ is the density function of x, $E(x)$ is the expected value of x, and $\text{Var}(x)$ is the variance of x. The conditional density of x given y is denoted by $p(x|y)$.

If $\boldsymbol{x}$ and $\boldsymbol{y}$ are random vectors, then $p(\boldsymbol{x})$ is the density function of $\boldsymbol{x}$, $E(\boldsymbol{x})$ is the expected vector of $\boldsymbol{x}$, and $\text{Cov}(\boldsymbol{x})$ is the covariance matrix of $\boldsymbol{x}$.

1.5.4 Mean field variational Bayes notation

We use q to denote a approximating density function that arises from MFVB inference. The optimal density function is denoted by q^*.

If z is a random variable with density function $q(z)$ and $f(z)$ is any function of z, then

$$\mu_{q(f(z))} \equiv E_q[f(z)], \text{ and } \sigma^2_{q(f(z))} \equiv \text{Var}_q[f(z)].$$

If $\boldsymbol{\theta}$ is a random vector generated from the density function $q(\boldsymbol{\theta})$, and $f(\boldsymbol{\theta})$ is any function of $\boldsymbol{\theta}$, then

$$\mu_{q(f(\boldsymbol{\theta}))} \equiv E_q[f(\boldsymbol{\theta})], \text{ and } \Sigma_{q(f(\boldsymbol{\theta}))} \equiv \mathrm{Cov}_q[f(\boldsymbol{\theta})].$$

1.5.5 Function

Definition 1.10. *We define the integral* $\mathcal{J}(\cdot,\cdot,\cdot,\cdot)$ *by*

$$\mathcal{J}(p,q,r,s) \equiv \int_{-\infty}^{\infty} x^p \exp\{qx - rx^2 - s\boldsymbol{e}^{-x}\}dx,$$

where $p \geq 0$, $-\infty < q < \infty$, $r > 0$ *and* $s > 0$.

Definition 1.11. *We use the notation* $\Gamma(\cdot)$ *to denote the Gamma function, which is defined by*

$$\Gamma(x) = \int_0^{\infty} u^{x-1}e^{-x}\,du.$$

Definition 1.12. *We use the notation* $\psi(\cdot)$ *to denote the digamma function, which is the derivative of the logarithm Gamma function*

Definition 1.13. *We use the notation* $B(\alpha,\beta)$ *to denote the Beta function, which is defined by*

$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,dt.$$

1.5.6 Distribution

Definition 1.14. *We use the notation* $x \sim N(\mu,\sigma^2)$ *to denote that* x *follows a Normal (Gaussian) distribution with mean* μ *and variance* $\sigma^2 > 0$. *The corresponding density function is*

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty.$$

Definition 1.15. *We use the notation $\phi(\cdot)$ to denote the probability density function of the standard normal distribution. i.e. if $x \sim N(0,1)$, then the corresponding density function is*

$$p(x) = \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad -\infty < x < \infty.$$

Definition 1.16. *We use the notation $\phi_{\sigma^2}(\cdot)$ to denote the probability density function of the Normal distribution with mean 0 and variance σ^2.*

Result 1.3. *Let $x \sim N(\mu, \sigma^2)$. Then*

$$E[x] = \mu, \ \mathrm{Var}[x] = \sigma^2.$$

Definition 1.17. *We use the notation $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote that $\boldsymbol{x}$ follows a Multivariate Normal (Gaussian) distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. The corresponding density function is*

$$p(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}|}} \exp\left\{ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right\}, \quad \boldsymbol{x} \in \mathbb{R}^k.$$

Definition 1.18. *The notation $x \sim$ Inverse-Gaussian(μ, λ) is used to denote that x follows an Inverse-Gaussian distribution with mean parameter $\mu > 0$ and shape parameter $\lambda > 0$. The corresponding density function is*

$$p(x) = \left[\frac{\lambda}{2\pi x^3} \right]^{1/2} \exp\left\{ \frac{-\lambda (x-\mu)^2}{2\mu^2 x} \right\}, \quad -\infty < x < \infty.$$

Result 1.4. *Let $x \sim$ Inverse-Gaussian(μ, λ). Then*

$$E[x] = \mu, \ E[1/x] = 1/\mu + 1/\lambda.$$

Definition 1.19. *The notation $x \sim$ Log-Normal(μ, σ^2) is used to denote that x*

follows a Log-Normal distribution with parameters μ and σ^2. The corresponding density function is

$$p(x) = \frac{1}{x\sigma\sqrt{2\pi}}\, e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}, \quad x > 0.$$

Definition 1.20. *The notation $x \sim$ Bernoulli(ρ), $x \in \{0, 1\}$ means that x has a Bernoulli distribution with probability parameter ρ. The corresponding probability mass function is*

$$p(x) = \rho^x (1-\rho)^{1-x}, \;\; x = 0, 1.$$

Definition 1.21. *The notation $x \sim$ Half-Cauchy(A) means that x has a Half-Cauchy distribution with scale parameter $A > 0$. The corresponding density function is*

$$p(x) = \frac{2A}{\pi\{A^2 + x^2\}}, \;\; x > 0.$$

Result 1.5. *Let σ and a be random variables such that*

$$\sigma^2 \sim \text{Inverse-Gamma}(1/2, 1/a) \;\; \text{and} \;\; a \sim \text{Inverse-Gamma}(1/2, 1/A^2),$$

where $A > 0$. Then $\sigma \sim$ Half-Cauchy(A).

Definition 1.22. *We use the notation $x \sim Beta(\alpha, \beta)$ to denote that x follows a Beta distribution with parameters $\alpha > 0$ and $\beta > 0$. The corresponding density function is*

$$p(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, \;\; x \in [0, 1],$$

where $B(\alpha, \beta)$ is a Beta function.

Result 1.6. *Let $x \sim Beta(\alpha, \beta)$. Then*

$$E[x] = \frac{\alpha}{\alpha+\beta}, \;\; E[\ln x] = \psi(\alpha) - \psi(\alpha+\beta),$$

where ψ is the digamma function.

Definition 1.23. *The notation $x \sim$ Inverse-Gamma(A, B) is used to denote that x has an Inverse-Gamma distribution with shape parameter $\alpha > 0$ and rate parameter $\beta > 0$. The corresponding density function is*

$$p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp\left\{-\frac{\beta}{x}\right\}, \ x > 0.$$

Result 1.7. *Let $x \sim$ Inverse-Gamma(A, B). Then*

$$E[1/x] = \frac{\alpha}{\beta}, \ E[\ln x] = \ln(\beta) - \psi(\alpha),$$

where ψ is the digamma function.

Definition 1.24. *The notation $x \sim$ Gamma(A, B) is used to denote that x has a Gamma distribution with shape parameter $\alpha > 0$ and rate parameter $\beta > 0$. The corresponding density function is*

$$p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp\{-\beta x\}, \ x > 0.$$

Result 1.8. *Let $x \sim$ Gamma(A, B), then*

$$E[x] = \frac{\alpha}{\beta}, \ E[\ln x] = -\ln(\beta) + \psi(\alpha),$$

where ψ is the digamma function.

Definition 1.25. *We use the notation $x \sim$ Laplace(λ) to denote that x follows a Laplace distribution with parameter $\lambda > 0$. The corresponding density function is*

$$p(x) = \frac{\lambda}{2} \exp\left\{-\lambda|x|\right\}, \ -\infty < x < \infty.$$

Result 1.9. *Let x and a be random variables such that*

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

Then $x \sim \text{Laplace}(\lambda)$.

Definition 1.26. *We use the notation $x \sim t(x, \sigma^2, \nu)$ to denote that x follows a Student's t-distribution with parameters σ^2 and $\nu > 0$. The corresponding density function is*

$$p(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1 + \frac{x^2}{\sigma^2\nu}\right)^{-\frac{\nu+1}{2}}, \quad -\infty < x < \infty.$$

Result 1.10. *Let x and a be random variables such that*

$$x|a \sim N(0, a\sigma^2) \text{ and } a \sim \text{Inverse-Gamma}(\frac{\nu}{2}, \frac{\nu}{2}).$$

Then $x \sim t(x, \sigma^2, \nu)$.

Definition 1.27. *Suppose that $x \sim N(\mu, \sigma^2)$ has a normal distribution and lies within the interval $x \in (a, b)$, $-\infty \leq a < b \leq \infty$. Then x conditional on $a < x < b$ has a truncated normal distribution. Its probability density function for $a < x < b$, is given by*

$$p(x; a, b) = \frac{\frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})},$$

where $\phi(\cdot)$ is the probability density function of the standard normal distribution and $\Phi(\cdot)$ is its cumulative distribution function.

Result 1.11. *Suppose x, $a < x < b$, has a truncated normal distribution. Then*

$$E[x] = \mu + \frac{\phi(\frac{a-\mu}{\sigma}) - \phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}\sigma.$$

Definition 1.28. *We use the notation* $\boldsymbol{x} \sim$ *Dirichlet*$(\boldsymbol{\alpha})$ *to denote that* $\boldsymbol{x} = (x_1, ..., x_K)$ *follows a Dirichlet distribution of order* $K \leq 2$ *with parameters* $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_K)$. *The corresponding density function is*

$$Dirichlet(\boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}.$$

Definition 1.29. *The probability density function of a Normal Mixture distribution with mean* μ_1,...,μ_K *and variance* σ_1^2,...,σ_K^2 *is given by*

$$p(x; \mu_1, ..., \mu_n, \sigma_1^2, ..., \sigma_n^2) = \sum_{i=1}^{K} w_i \frac{1}{\sigma_i} \phi(\frac{x - \mu_i}{\sigma_i}),$$

where $w_i \geq 0$ *and* $\sum_{i=1}^{K} w_i = 1$.

Result 1.12. *Suppose that* x *has a Normal Mixture distribution with density function*

$$p(x; \mu_1, ..., \mu_n, \sigma_1^2, ..., \sigma_n^2) = \sum_{i=1}^{n} w_i \frac{1}{\sigma_i} \phi(\frac{x - \mu_i}{\sigma_i}).$$

Then

$$E[x] = \mu = \sum_{i=1}^{n} w_i \mu_i, \mathrm{Var}[x] = \sigma^2 = \sum_{i=1}^{n} w_i [(\mu_i - \mu)^2 + \sigma_i^2].$$

Definition 1.30. *The Dirac delta function is denoted by* $\delta_0(\cdot)$ *and is defined by*

$$\delta_0(x) = \begin{cases} 1 & \text{if } x = 0, \\ 0 & \text{if } x \neq 0. \end{cases}$$

Definition 1.31. *The notation* $x \sim$ *Normal-Zero*(σ^2, ρ) *denotes that* x *has a Normal-Zero distribution. The corresponding density function is*

$$p(x|\sigma^2, \rho) = \rho \phi_{\sigma^2}(x) + (1 - \rho)\delta_0(x), \;\; x \in \mathbb{R},$$

where ρ is a random variable over [0,1].

Result 1.13. *Let γ and x be random variables such that*

$$p(x|\gamma) = \gamma\phi_{\sigma^2}(x) + (1-\gamma)\delta_0(x), \text{ and } \gamma|\rho \sim \textit{Bernoulli}(\rho).$$

Then $x|\sigma^2, \rho \sim$ Gaussian-Zero(σ^2, ρ).

Result 1.14. *Let γ and θ be random variables such that*

$$\gamma \sim \textit{Bernoulli}(\rho), \text{ and } \theta|\sigma^2 \sim N(0, \sigma^2),$$

and suppose that $x = \gamma\theta$. Then $x|\sigma^2, \rho \sim$ Gaussian-Zero(σ^2, ρ).

Result 1.15. *Let y and a be random variables such that*

$$y = \begin{cases} 1 & \textit{if } a \geq 0 \\ 0 & \textit{otherwise} \end{cases}$$
$$\textit{and } a \sim N(\mu, 1).$$

Then $y \sim$ Bernoulli$(\Phi(\mu))$.

Result 1.16. *Consider a Bayesian linear mixed model:*

$$\begin{aligned} &\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2 \sim N(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \sigma_\varepsilon^2\boldsymbol{I}), \\ &\boldsymbol{u}|\sigma_u^2 \sim N(0, \sigma_u^2\boldsymbol{I}), \\ &\boldsymbol{\beta} \sim N(0, \sigma_\beta^2\boldsymbol{I}), \\ &\sigma_\varepsilon^2 \sim \textit{Inverse-Gamma}(A_\varepsilon, B_\varepsilon), \\ &\sigma_u^2 \sim \textit{Inverse-Gamma}(A_u, B_u). \end{aligned} \tag{1.3}$$

Then, the posterior density function of $\boldsymbol{\beta}$ *and* $\boldsymbol{u}$, $p(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{y})$ *is given by*

$$p(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{y}) \propto \int_0^\infty \int_0^\infty \exp\left\{-\frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}\|^2}{2\sigma_\varepsilon^2} - \frac{\|\boldsymbol{\beta}\|^2}{2\sigma_\beta^2} - \frac{\|\boldsymbol{u}\|^2}{2\sigma_u^2}\right\} p(\sigma_\varepsilon^2)p(\sigma_u^2)d\sigma_\varepsilon^2 d\sigma_u^2.$$

An alternative model, which employs an auxiliary data vector $\boldsymbol{a}$, *is*

$$\begin{aligned}
&\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2 \sim N(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \sigma_\varepsilon^2 \boldsymbol{I}).\\
&\boldsymbol{a}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma_u^2 \sim N\left(\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix}, \begin{bmatrix} \sigma_\beta^2 \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_u^2 \boldsymbol{I} \end{bmatrix}\right).\\
&\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} \sim N(\boldsymbol{0}, b\boldsymbol{I}),\\
&\sigma_\varepsilon^2 \sim \textit{ Inverse-Gamma}(A_\varepsilon, B_\varepsilon),\\
&\sigma_u^2 \sim \textit{ Inverse-Gamma}(A_u, B_u).
\end{aligned} \tag{1.4}$$

Then the posterior density function of $\boldsymbol{\beta}$ *and* $\boldsymbol{u}$ *satisfies:*

$$p_b(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{y}, \boldsymbol{a} = 0) \propto \exp\left\{-\frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}\|^2}{2\sigma_\varepsilon^2} - \frac{1 + \sigma_\beta^2/b}{2\sigma_\beta^2} \| \boldsymbol{\beta} \|^2 - \frac{1 + \sigma_u^2/b}{2\sigma_u^2} \| \boldsymbol{u} \|^2 \right\} p(\sigma_\varepsilon^2)p(\sigma_u^2)d\sigma_\varepsilon^2 d\sigma_u^2.$$

It is apparent from this that

$$\lim_{b\to\infty} p_b(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{y}, \boldsymbol{a} = 0) = p(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{y}).$$

Similar results hold for the other posterior density functions. Hence, using the model (1.4) with b *set to be a very large number and with the auxiliary vector set to have an observed value 0 leads to essentially the same results for posterior*

density functions of model (1.3). This follows from Wang and Wand (2011).

Result 1.17. *Let x, a and b be random variables such that*

$$x|a \sim N(0,a),\ a|b \sim \textit{Inverse-Gamma}(M, M/b),\ b|\lambda \sim \textit{Inverse-Gamma}(1, \lambda^2/2),$$

where $M > 0$. Then the density function $p(x; M, \lambda)$ leads to a good approximation to the Laplace(λ) distribution when M is large. This follows Luts et al. (2015).

Result 1.18. *The expression of $-log(1+e^x)$ is the maxima of a family of parabolas:*

$$-log(1+e^x) = \max_{\xi \in \mathbb{R}} \left\{ A(\xi)x^2 - \frac{1}{2}x + C(\xi) \right\} \ \textit{for all } \xi \in \mathbb{R}$$

where

$$A(\xi) \equiv -\frac{\tanh(\xi/2)}{4\xi} \ \textit{and}\ C(\xi) \equiv \frac{\xi}{2} - log(1+e^{\xi}) + \frac{\xi \tanh(\xi/2)}{4}.$$

This follows Jaakkola and Jordan (2000).

1.6 Overview

The aim of this PhD research is to explore **how to use variational approximations methods to deal with the linear and semiparametric regression model and perform variable selection**. Traditional approximations methods, such as the MCMC and Laplace approximations method, result in a model fit that is too slow. Therefore, we attempt to use the variational approximations method, including mean field variational Bayes (MFVB) and Gaussian variational approximation (GVA), to deal with the models mentioned.

This thesis has 8 chapters: Chapter 1 reviews variational approximations and gives required notation and results. Chapter 2 develops MFVB inference for

Bayesian variable selection based on the posterior probabilities of the model. Chapter 3 presents a MFVB linear variable selection method based on indicator variables. Chapter 4 develops MFVB inference for the Bayesian lasso model. Chapter 5 introduces how to use a new approximate Bayesian inference, Infer.NET, for statistical analyses. Chapter 6 presents asymptotic theory for Gaussian variational approximations, which is an non-Bayesian variational approximations method. Chapter 7 introduces a new Bayesian inference software, InferMachine(), which can perform the MFVB inference by using BRugs model code. Some discussion is given in Chapter 8.

Chapter 2

Mean Field Variational Bayes Variable Selection

2.1 Introduction

The selection of variables in regression problems has occupied the minds of many statisticians. An important problem in statistical analysis is the choice of an optimal model from a set of *a priori* alternative models. In most instances, we consider how to select a subset of variables that should be included in the model. An extensive literature (e.g. George, 2000; Robert & Casella, 2004; Broman & Speed, 2002; Liang *et al.*, 2008) introduces a variety of algorithms for searching the model space and selection criteria for choosing between competing models. Given a set of potential predictor variables x_1,...,x_p, the models M_k, $k = 1, ..., K$, are alternative models containing the subsets X_k of total potential predictor variables and unknown parameters $\boldsymbol{\theta}_k$. For a given prior distribution of the unknown parameter $\boldsymbol{\theta}_k$, we can obtain the posterior model probabilities under the data:

$$p(Y|M_k) = \int p(Y|M_k, \theta_k)p(\theta_k)d\theta_k,$$

known as the marginal likelihood of the data under the model. For given model prior probabilities $p(M_k)$, we can obtain the posterior probabilities of each model:

$$p(M_k|Y) = \frac{p(M_k)p(Y|M_k)}{\sum_{k=1}^{K} p(M_k)p(Y|M_k)}.$$

For a simple or specific model, the approximate value of posterior model probabilities can be obtained by integrating the likelihood function directly (Liang *et al.*, 2008). In the general case, several MCMC methods have been proposed for estimating probabilities of models in the presence of model uncertainty (Kass & Raftery, 1995; West, 2003; Drummond & Rambaut, 2007).

In the case of the mean field variational Bayes (MFVB) approximation, the lower bound on the marginal likelihood, $\underline{p}(Y|M_k)$, can be derived more directly. For MFVB model selection, $\underline{p}(Y|M_k)$ will be used to replace the marginal likelihood, $p(Y|M_k)$, to obtain the posterior probabilities of each model:

$$p(M_k|Y) = \frac{p(M_k)\underline{p}(Y|M_k)}{\sum_{k=1}^{K} p(M_k)\underline{p}(Y|M_k)}.$$

Next, I will describe the general framework of variable selection for linear and non-linear cases.

2.1.1 Linear variable selection

We consider the Gaussian linear regression model with response $\boldsymbol{y}$ and a set of potential predictor variables $\boldsymbol{x}_1$,...,$\boldsymbol{x}_p$. Variable selection will select a subset $\boldsymbol{X}_k = (\boldsymbol{x}_{k1}, \boldsymbol{x}_{k2}, ..., \boldsymbol{x}_{kp_k})$ from the predictor variables and generate the alternative M_k as:

$$M_k : \boldsymbol{y} = 1\beta_0 + \boldsymbol{X}_k\boldsymbol{\beta}_k + \boldsymbol{\varepsilon},$$

where $\boldsymbol{y} = (y_1, y_2, ..., y_n)^T$ is the response vector and β_0 is an intercept that will be included in each alternative model. The error vector, $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)^T$, contains terms that are independent and identically distributed from a normal distribution with mean 0 and variance σ_ε^2. $\boldsymbol{\beta}_k$ is a p_k-dimensional vector of nonzero regression coefficients.

A Bayesian linear regression model is:

$$\begin{aligned} &\boldsymbol{y}|\beta_0, \boldsymbol{\beta}, \sigma_\varepsilon^2 \overset{\text{ind.}}{\sim} N(\mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta}, \sigma_\varepsilon^2 \boldsymbol{I}), \\ &\sigma_\varepsilon^2 \sim \text{ Inverse-Gamma}(A_\varepsilon, B_\varepsilon), \\ &\beta_0 \sim N(0, \sigma_{\beta_0}^2), \\ &\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma_{\boldsymbol{\beta}}^2 \boldsymbol{I}), \end{aligned} \tag{2.1}$$

where A_ε, B_ε, $\sigma_{\beta_0}^2$ and $\sigma_{\boldsymbol{\beta}}^2 > 0$ are hyperparameters. Bishop (2006) shown a conjugate prior structure exists that gives a closed form for the posterior of model (2.1). Ormerod and Wand (2010) gave a pedagogical derivation of MFVB and obtained the $\log \underline{p}(\boldsymbol{y}; q)$ expression. Assuming the prior probabilities $p(M_k)$ are equal and using the lower bound $\underline{p}(y; q)$ instead of the marginal likelihood, the variational approximate posterior probabilities of models are:

$$p(M_k|\boldsymbol{y}) = \frac{\underline{p}(\boldsymbol{y}; q, M_k)}{\sum_{k=1}^{K} \underline{p}(\boldsymbol{y}; q, M_k)}, \;\; 1 \leq k \leq K,$$

and the alternative model with the largest posterior probability is defined to be the estimated optimal model.

2.1.2 Non-linear variable selection

In general, we are not only interested in whether a variable can be selected into the optimal model with a linear structure, but also we are interested in whether the selected variable is linear or non-linear. We use additive models with mixed

model-based penalised splines to deal with non-linear effects. Similarly to the linear case, we consider a set of potential predictor variables $\boldsymbol{x}_1$,...,$\boldsymbol{x}_p$ and generate the alternative M_k as:

$$M_k : y_i = \beta_0 + \sum_{j=1}^{k_1} \beta_j x_{i,j} + \sum_{r=1}^{k_2} f_r(x_{i,r}) + \varepsilon_i, \;\; 1 \leq i \leq n,$$

where y_i is the ith response and β_0 is an intercept that will be included in each alternative model. The error, ε_i, is independent and identically distributed from a normal distribution with mean 0 and variance σ^2_ε. The variables $x_{i,j}$, $1 \leq j \leq k_1$, are linear effect variables and $x_{i,r}$, $1 \leq r \leq k_2$, are non-linear effect variables. The quantities k_1 and k_2 are the numbers of linear and non-linear variables selected into the alternative M_k. We will model each of the $f_r(\cdot)$ using low-rank smoothing splines with a mixed model representation:

$$f_r(x) = \beta_r x + \sum_{k=1}^{K} u_{r,k} z_{r,k}(x)$$

$$u_{r,k} \sim N(0, \sigma^2_{u_r})$$

where $z_{r,k}$, $1 \leq k \leq K_r$, are O'Sullivan Penalised Splines (Wand & Ormerod, 2008) functions over x. $u_{r,1}, ..., u_{r,K}$ are random effect terms with the normal distribution independently.

Our Bayesian additive model is

$$\begin{aligned}
&\boldsymbol{y}|\beta_0, \boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2 \overset{\text{ind.}}{\sim} N(\mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \sigma_\varepsilon^2 I_n),\\
&\beta_0 \sim N(0, \sigma_{\beta_0}^2),\\
&\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma_{\boldsymbol{\beta}}^2 \boldsymbol{I}),\\
&\boldsymbol{u}_\ell|\sigma_{\boldsymbol{u}_\ell} \sim N(0, \sigma_{\boldsymbol{u}_\ell}^2 I_{K_\ell}) \quad 1 \le \ell \le r,\\
&\sigma_{\boldsymbol{u}_\ell}^2 \sim \text{Inverse-Gamma}(A_{\boldsymbol{u}}, B_{\boldsymbol{u}}) \quad 1 \le \ell \le r,\\
&\sigma_\varepsilon^2 \sim \text{Inverse-Gamma}(A_\varepsilon, B_\varepsilon).
\end{aligned} \tag{2.2}$$

Here A_ε, B_ε, $\sigma_{\beta_0}^2$, $A_{\boldsymbol{u}}$, $B_{\boldsymbol{u}}$ and $\sigma_{\boldsymbol{\beta}}^2 > 0$ are hyperparameters and $\boldsymbol{Z} = [\boldsymbol{Z}_1, ..., \boldsymbol{Z}_r]$ where

$$\boldsymbol{Z}_\ell = \begin{bmatrix} z_{\ell,1}(x_{1,\ell}) & \cdots & z_{\ell,K}(x_{1,\ell}) \\ \vdots & \ddots & \vdots \\ z_{\ell,1}(x_{n,\ell}) & \cdots & z_{\ell,K}(x_{n,\ell}) \end{bmatrix}, \quad 1 \le \ell \le r,$$

is the spline basis design matrix and z_k, $1 \le k \le K$, is an O'Sullivan spline basis (Wand & Ormerod, 2008). Using the MFVB inference method, we can obtain the $\log \underline{p}(\boldsymbol{y}; q)$ expression (Ormerod & Wand, 2010). Similarly to the linear case, the variational approximate posterior probabilities of the models can be obtained using $\underline{p}(y; q)$, and the alternative model with the largest posterior probability is the estimated optimal model.

In this chapter, we will consider the different model structures for the alternative model sets M_k and examine the linear and non-linear variable selection models. The results of variable selection under different prior distributions will be compared. To make the variable selection fast, a stepwise method will be used for variable selection.

2.2 Prior Distribution for Fixed Effects

In general, we set the fixed effect parameters β_i, $1 \leq i \leq p$, to be generated from a normal prior distribution with mean 0 and variance $\sigma^2_{\boldsymbol{\beta}}$, where $\sigma^2_{\boldsymbol{\beta}}$ is a large number, such as 10^8. This is a non-informative prior distribution and suitable for general models. Liang *et al.* (2008) studied mixtures of g priors as an alternative to default prior distribution for fixed effects for use in Bayesian variable selection. In this chapter, a ridge penalty (Hoerl & Kennard, 1970) for the fixed effect parameters and a corresponding Bayesian hierarchical model will be considered.

Ridge regression shrinks the regression coefficients by imposing a penalty on their size. The ridge coefficients minimize a penalized residual sum of squares:

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) = \arg\min_{\beta_0, \boldsymbol{\beta}} \left\{ (\boldsymbol{y} - 1_n\beta_0 - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - 1_n\beta_0 - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \tag{2.3}$$

where the intercept β_0 has been left out of the penalty term. The ridge estimate can be derived as a Bayes maximum *a posteriori* estimate under independent Normal priors for each regression coefficient (Hastie, Tibshirani & Friedman, 2009). For model (2.1) with given σ^2_ε, the posterior distribution of β_0 and $\boldsymbol{\beta}$ is given:

$$\begin{aligned} p(\beta_0, \boldsymbol{\beta} | \boldsymbol{y}, \sigma^2_\varepsilon) =& (2\pi\sigma^2_\varepsilon)^{-\frac{n}{2}} \exp\left\{ \frac{(\boldsymbol{y} - 1_n\beta_0 - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - 1_n\beta_0 - \boldsymbol{X}\boldsymbol{\beta})}{\sigma^2_\varepsilon} \right\} \\ & \times (2\pi\sigma^2_{\boldsymbol{\beta}})^{-\frac{p}{2}} \exp\left\{ -\frac{\sum_{j=1}^{p} \beta_j^2}{\sigma^2_{\boldsymbol{\beta}}} \right\} \\ & \propto \exp\left\{ -\frac{(\boldsymbol{y} - 1_n\beta_0 - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - 1_n\beta_0 - \boldsymbol{X}\boldsymbol{\beta})}{\sigma^2_\varepsilon} - \frac{\sum_{j=1}^{p} \beta_j^2}{\sigma^2_{\boldsymbol{\beta}}} \right\}. \end{aligned} \tag{2.4}$$

So the Bayes maximum *a posteriori* estimate for $(\beta_0, \boldsymbol{\beta})$ is given by:

$$\arg\min_{\beta_0, \boldsymbol{\beta}} \left\{ (\boldsymbol{y} - 1_n\beta_0 - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - 1_n\beta_0 - \boldsymbol{X}\boldsymbol{\beta}) + \frac{\sigma^2_\varepsilon}{\sigma^2_{\boldsymbol{\beta}}} \sum_{j=1}^{p} \beta_j^2 \right\}.$$

It is similar to (2.3) when we set $\lambda = \sigma_\varepsilon^2/\sigma_{\boldsymbol{\beta}}^2$. If we set a non-informative prior for β_j, e.g. $\sigma_{\boldsymbol{\beta}}^2 = 10^{10}$, $\sigma_\varepsilon^2/\sigma_{\boldsymbol{\beta}}^2$ will be close to 0. Then the Bayes maximum *a posteriori* estimates for the regression coefficients are ridge regression least squares estimators.

Linear regression with a ridge penalty in the fixed effect

The model for a Bayesian linear regression with a ridge penalty for the fixed effects is:

$$\begin{aligned}
&\boldsymbol{y}|\beta_0, \boldsymbol{\beta}, \sigma_\varepsilon^2 \overset{\text{ind.}}{\sim} N(\mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta}, \sigma_\varepsilon^2 \boldsymbol{I}),\\
&\sigma_\varepsilon^2 \sim \text{ Inverse-Gamma}(A_\varepsilon, B_\varepsilon),\\
&\beta_0 \sim N(0, \sigma_{\beta_0}^2),\\
&\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma_{\boldsymbol{\beta}}^2 \boldsymbol{I}),\\
&\sigma_{\boldsymbol{\beta}}^2 \sim \text{ Inverse-Gamma}(A_{\boldsymbol{\beta}}, B_{\boldsymbol{\beta}}),
\end{aligned} \tag{2.5}$$

where A_ε, B_ε, $\sigma_{\beta_0}^2$, A_β and $B_{\boldsymbol{\beta}} > 0$ are hyperparameters. We seek an approximate inference to the posterior distribution corresponding to the model given in (2.5). A tractable solution arises if we impose the product restriction:

$$q(\beta_0, \boldsymbol{\beta}, \sigma_\varepsilon^2, \sigma_{\boldsymbol{\beta}}^2) = q(\beta_0, \boldsymbol{\beta}) q(\sigma_\varepsilon^2, \sigma_{\boldsymbol{\beta}}^2).$$

The theory of induced factorizations (e.g., Bishop, 2006, Section 10.2.5) leads to a solution with the additional product structure:

$$q(\beta_0, \boldsymbol{\beta}, \sigma_\varepsilon^2, \sigma_{\boldsymbol{\beta}}^2) = q(\beta_0, \boldsymbol{\beta}) q(\sigma_\varepsilon^2) q(\sigma_{\boldsymbol{\beta}}^2).$$

Then, using the equation (1.1), the optimal q^* densities for the parameters in model (2.5) take the form:

$$\begin{aligned} &q^*(\beta_0, \boldsymbol{\beta}) \text{ is a Multivariate Normal density function,} \\ &q^*(\sigma^2_\varepsilon) \text{ is an Inverse Gamma density function, and} \\ &q^*(\sigma^2_{\boldsymbol{\beta}}) \text{ is an Inverse Gamma density function.} \end{aligned} \tag{2.6}$$

Let $\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})}$ and $\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}$ denote the mean vector and covariance matrix for the Multivariate Normal density function $q^*(\beta_0, \boldsymbol{\beta})$, and $A_{q(\sigma^2_\varepsilon)}$ and $B_{q(\sigma^2_\varepsilon)}$ denote the shape and rate parameters for $q^*(\sigma^2_\varepsilon)$. A similar definition is used for the parameters in $q^*(\sigma^2_{\boldsymbol{\beta}})$. Let $\boldsymbol{C} = [1, \boldsymbol{X}]$. Derivations for the optimal densities are deferred to Appendix 2.A.

The convergence of algorithm (2.2.1)[1] can be monitored using the following expression for the lower bound on the marginal log-likelihood:

$$\begin{aligned} \log \underline{p}(\boldsymbol{y}; q) \;=\; & -\frac{n}{2}\log(2\boldsymbol{\pi}) + \frac{1+p}{2} - \frac{1}{2}\log(\sigma^2_{\beta_0}) - \frac{1}{2}\frac{\mu^2_{q(\beta_0)} + \sigma^2_{q(\beta_0)}}{\sigma^2_{\beta_0}} \\ & + A_\varepsilon \log(B_\varepsilon) - \log\Gamma(A_\varepsilon) \\ & - (A_\varepsilon + \frac{n}{2})\log(B_{q(\sigma^2_\varepsilon)}) + \log\Gamma(A_\varepsilon + \frac{n}{2}) \\ & + A_{\boldsymbol{\beta}}\log(B_{\boldsymbol{\beta}}) - \log\Gamma(A_{\boldsymbol{\beta}}) \\ & - A_{q(\sigma^2_{\boldsymbol{\beta}})}\log(B_{q(\sigma^2_{\boldsymbol{\beta}})}) + \log\Gamma(A_{q(\sigma^2_{\boldsymbol{\beta}})}) \\ & + \frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})}|. \end{aligned}$$

[1]For many semiparametric regression model, the order of update does not matter. When the regression model includes indictor variables (Chapter 3), the order my matter.

Algorithm 2.2.1: MFVB iterative scheme to obtain the parameters of the optimal densities $q^*(\beta_0, \boldsymbol{\beta})$, $q^*(\sigma^2_{\boldsymbol{\beta}})$ and $q^*(\sigma^2_\varepsilon)$ for Bayesian linear regression with the ridge penalized method in the fixed effect model (2.5).

Initialize $\mu_{q(1/\sigma^2_\varepsilon)}$ and $\mu_{q(1/\sigma^2_{\boldsymbol{\beta}})}$;

Cycle

$$\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})} \leftarrow \left\{\mu_{q(1/\sigma^2_\varepsilon)}\boldsymbol{C}^T\boldsymbol{C} + \text{blockdiag}\left[(\sigma^2_{\beta_0})^{-1}, \mu_{q(1/\sigma^2_{\boldsymbol{\beta}})}\boldsymbol{I}_p\right]\right\}^{-1}$$

$$\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})} \leftarrow \mu_{q(1/\sigma^2_\varepsilon)}\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}\boldsymbol{C}^T\boldsymbol{y}$$

$$A_{q(\sigma^2_\varepsilon)} \leftarrow \frac{n}{2} + A_\varepsilon$$

$$B_{q(\sigma^2_\varepsilon)} \leftarrow B_\varepsilon + \tfrac{1}{2}\left\{\| \boldsymbol{y} - C\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})} \|^2 + \text{tr}(\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})})\right\}$$

$$\mu_{q(1/\sigma^2_\varepsilon)} \leftarrow \frac{A_{q(\sigma^2_\varepsilon)}}{B_{q(\sigma^2_\varepsilon)}}$$

$$A_{q(\sigma^2_{\boldsymbol{\beta}})} \leftarrow \frac{p}{2} + A_{\boldsymbol{\beta}}$$

$$B_{q(\sigma^2_{\boldsymbol{\beta}})} \leftarrow B_{\boldsymbol{\beta}} + \tfrac{1}{2}\left\{\| \boldsymbol{\mu}_{q(\boldsymbol{\beta})} \|^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\right\}$$

$$\mu_{q(1/\sigma^2_{\boldsymbol{\beta}})} \leftarrow \frac{A_{q(\sigma^2_{\boldsymbol{\beta}})}}{B_{q(\sigma^2_{\boldsymbol{\beta}})}}$$

until the increase in $\log \underline{p}(\boldsymbol{y}; q)$ is negligible.

2.3 Prior Distribution for the Variance

The Inverse-Gamma(A, B) prior distribution is an attempt at non-informativeness with A and B set to low values, such as 0.01 or 0.001. As a conjugate for the variance in the Normal response model, the Inverse-Gamma prior distribution is used in a large number of Bayesian hierarchical models. In the previous section, we used the Inverse-Gamma as the prior distribution to build a variable selection model. However, alternative variance (scale) parameter priors are often discussed and considered in the hierarchical model. Gelman (2006) discussed non-informative prior

distributions for scale parameters, such as Half-t prior distributions. As a special case, a Half-Cauchy prior distribution is used for scale parameters (square roots of variances), which are estimated from a small number of groups and have generated good results. Polson and Scott (2012) used the Half-Cauchy prior for a global scale parameter. Marley and Wand (2010) used the Half-Cauchy prior in robust non-parametric regression via the t-distribution. In Bayesian variable selection, Cottet, Kohn and Nott (2008) used a log-normal prior distribution for the variance of the spline coefficient. I will begin by giving a brief introduction to those distributions and the corresponding Bayesian model structure.

2.3.1 Half-Cauchy prior

The Half-Cauchy is a special case of the Half-t prior distributions. Figure 2.1 shows the Half-Cauchy density function with a scale parameter A. The density function has a broad peak at zero and will be close to a Uniform distribution when $A \to \infty$.

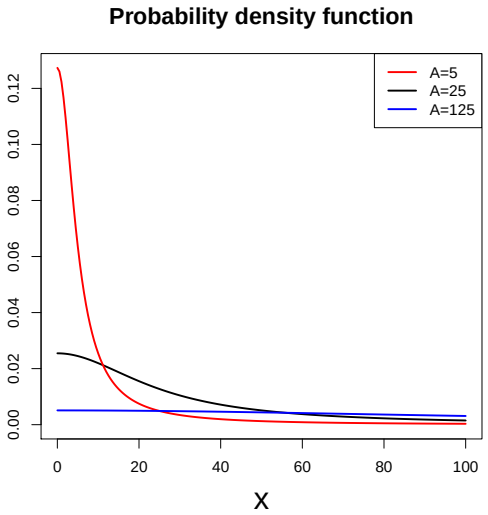

Figure 2.1: The Half-Cauchy density function with different values of the scale parameter A.

To consider a scale parameter $\sigma \sim$ Half-Cauchy(A), we introduce the auxiliary variables corresponding to Result 1.5. The model with auxiliary variables is

$$\sigma^2 \sim \text{ Inverse-Gamma}(1/2, 1/a) \text{ and } a \sim \text{ Inverse-Gamma}(1/2, 1/A^2)$$

Linear regression model with a Half-Cauchy prior for σ_ε^2

A Bayesian linear regression model with a Half-Cauchy prior is:

$$\begin{aligned}
&\boldsymbol{y}|\beta_0, \boldsymbol{\beta}, \sigma_\varepsilon^2 \overset{\text{ind.}}{\sim} N(\mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta}, \sigma_\varepsilon^2 \boldsymbol{I}),\\
&\beta_0 \sim N(0, \sigma_{\beta_0}^2),\\
&\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma_{\boldsymbol{\beta}}^2 \boldsymbol{I}),\\
&\sigma_\varepsilon^2|a \sim \text{ Inverse-Gamma}(1/2, 1/a),\\
&a \sim \text{ Inverse-Gamma}(1/2, 1/A^2).
\end{aligned} \tag{2.7}$$

Here $\sigma_{\beta_0}^2$, $\sigma_{\boldsymbol{\beta}}^2$ and $A > 0$ are hyperparameters. The product restriction that we impose here is

$$q(\beta_0, \boldsymbol{\beta}, \sigma_\varepsilon^2, a) = q(\beta_0, \boldsymbol{\beta}) q(\sigma_\varepsilon^2) q(a).$$

Then, as shown in Appendix 2.B, the optimal q^* densities for the parameters in model (2.7) take the form:

$$\begin{aligned}
&q^*(\beta_0, \boldsymbol{\beta}) \text{ is a Multivariate Normal density function,}\\
&q^*(\sigma_\varepsilon^2) \text{ is an Inverse Gamma density function, and}\\
&q^*(a) \text{ is an Inverse Gamma density function.}
\end{aligned} \tag{2.8}$$

Let $\boldsymbol{\mu}_{q(\beta_0, \boldsymbol{\beta})}$ and $\boldsymbol{\Sigma}_{q(\beta_0, \boldsymbol{\beta})}$ denote the mean vector and covariance matrix for the Multivariate Normal density function $q^*(\beta_0, \boldsymbol{\beta})$, and $A_{q(\sigma_\varepsilon^2)}$ and $B_{q(\sigma_\varepsilon^2)}$ denote the shape and rate parameters for $q^*(\sigma_\varepsilon^2)$. The 1 and $B_{q(a)}$ denote the shape and rate

parameters for $q^*(a)$. Let $\boldsymbol{C} = [1, \boldsymbol{X}]$. The convergence of algorithm (2.3.1) can be monitored using the following expression for the lower bound on the marginal log-likelihood:

$$\begin{aligned}\log \underline{p}(\boldsymbol{y}; q) \;=\; & \frac{1+p}{2} - \frac{n}{2}\log(2\boldsymbol{\pi}) - \frac{1}{2}\log(\sigma^2_{\beta_0}) - \frac{\mu^2_{q(\beta_0)} + \sigma^2_{q(\beta_0)}}{2\sigma^2_{\beta_0}} \\ & -\frac{p}{2}\log(\sigma^2_{\boldsymbol{\beta}}) - \frac{1}{2\sigma^2_{\boldsymbol{\beta}}}\{\|\mu_{q(\boldsymbol{\beta})}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\} \\ & +\frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})}| \\ & -\log(\pi) - \log(A) - \log(B_{q(a)}) + \mu_{q(1/a)}\mu_{1(1/\sigma^2_\varepsilon)} \\ & -A_{q(\sigma^2_\varepsilon)}\log(B_{q(\sigma^2_\varepsilon)}) + \log\Gamma(A_{q(\sigma^2_\varepsilon)}).\end{aligned}$$

Algorithm 2.3.1: MFVB iterative scheme for obtaining the parameters of the optimal densities $q^*(\beta_0, \boldsymbol{\beta})$, $q^*(a)$ and $q^*(\sigma^2_\varepsilon)$ for model (2.7): a linear regression model with Half Cauchy prior in σ^2_ε.

Initialize $\mu_{q(1/\sigma^2_\varepsilon)}$;
Cycle

$$\begin{aligned}&\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})} \leftarrow \left\{\mu_{q(1/\sigma^2_\varepsilon)}\boldsymbol{C}^T\boldsymbol{C} + \mathrm{blockdiag}\left[(\sigma^2_{\beta_0})^{-1}, (\sigma^2_{\boldsymbol{\beta}})^{-1}\boldsymbol{I}_p\right]\right\}^{-1} \\ &\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})} \leftarrow \mu_{q(1/\sigma^2_\varepsilon)}\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}\boldsymbol{C}^T\boldsymbol{y} \\ &A_{q(\sigma^2_\varepsilon)} \leftarrow \frac{n+1}{2} \\ &B_{q(\sigma^2_\varepsilon)} \leftarrow \mu_{q(1/a)} + \tfrac{1}{2}\left\{\|\,\boldsymbol{y} - C\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})}\,\|^2 + \mathrm{tr}(\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})})\right\} \\ &\mu_{q(1/\sigma^2_\varepsilon)} \leftarrow \frac{A_{q(\sigma^2_\varepsilon)}}{B_{q(\sigma^2_\varepsilon)}} \\ &B_{q(a)} \leftarrow \frac{1}{A^2} + \mu_{q(1/\sigma^2_\varepsilon)} \\ &\mu_{q(1/a)} \leftarrow \frac{1}{B_{q(a)}}\end{aligned}$$

until the increase in $\log \underline{p}\,(\boldsymbol{y}; q)$ is negligible.

2.3.2 Log-Normal prior

Cottet, Kohn and Nott (2008) used a Log-Normal prior distribution for the variance. The prior distribution for the variance σ^2 is:

$$\begin{aligned} \sigma^2 &\sim \text{Log-Normal}(a, b), \\ a &\sim N(0, \sigma_a^2), \\ b &\sim \text{Inverse-Gamma}(A_b, B_b), \end{aligned} \tag{2.9}$$

where the σ_a^2, A_b and $B_b > 0$ are hyperparameters. In our study, I will follow the result of Cottet *et al.* (2008) and set the hyperparameters values to $\sigma_a^2 = 100$, $A_b = 101$ and $B_b = 10100$. Figure 2.3.2 shows the density functions of the Inverse Gamma, Log-Normal and Half-Cauchy distributions. The density functions are similar in being flat-tailed but are significantly different around zero.

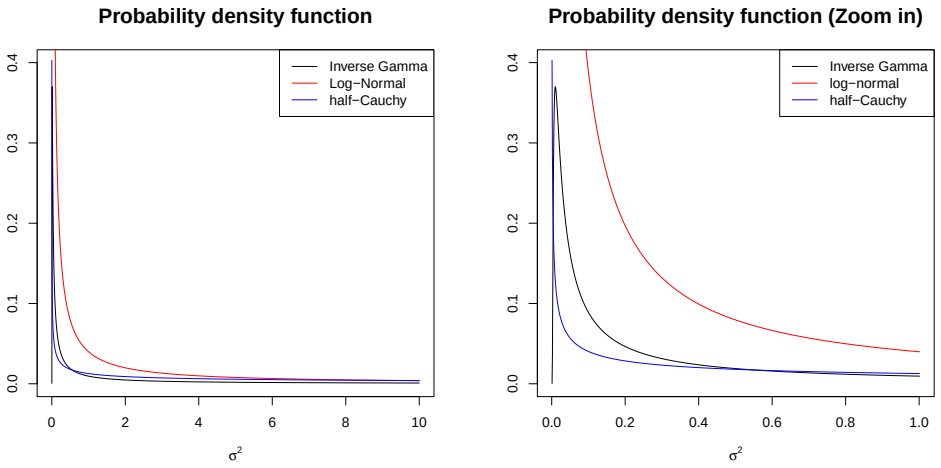

Figure 2.2: The density functions of the Inverse-Gamma, Log-Normal and Half-Cauchy distributions. The right figure is a enlarged portion of the left figure.

Linear regression model with a Log-Normal prior for σ_ε^2

A Bayesian linear regression model with a log-normal prior for σ_ε^2 is:

$$\begin{aligned}
&\boldsymbol{y}|\beta_0, \boldsymbol{\beta}, \sigma_\varepsilon^2 \overset{\text{ind.}}{\sim} N(\mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta}, \sigma_\varepsilon^2),\\
&\sigma_\varepsilon^2|a_\varepsilon, b_\varepsilon \sim \text{Log-Normal}(a_\varepsilon, b_\varepsilon),\\
&a_\varepsilon \sim N(0, \sigma_a^2),\\
&b_\varepsilon \sim \text{ Inverse-Gamma}(A_b, B_b),\\
&\beta_0 \sim N(0, \sigma_{\beta_0}^2),\\
&\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma_{\boldsymbol{\beta}}^2 \boldsymbol{I}),
\end{aligned} \tag{2.10}$$

where $\sigma_{\beta_0}^2$, $\sigma_{\boldsymbol{\beta}}^2$ A_b, B_b and $\sigma_a^2 > 0$ are hyperparameters. The product restriction that we impose here is:

$$q(\beta_0, \boldsymbol{\beta}, \sigma_\varepsilon^2, a_\varepsilon, b_\varepsilon) = q(\beta_0, \boldsymbol{\beta}) q(\sigma_\varepsilon^2) q(a_\varepsilon) q(b_\varepsilon).$$

Then, as shown in Appendix 2.C, the optimal densities of the parameters in model (2.10) take the form:

$$\begin{aligned}
&q^*(\beta_0, \boldsymbol{\beta}) \text{ is a Multivariate Normal density function,}\\
&q^*(a_\varepsilon) \text{ is a Normal density function,}\\
&q^*(b_\varepsilon) \text{ is an Inverse Gamma density function, and}\\
&q^*(\sigma_\varepsilon^2) = \frac{(\sigma_\varepsilon^2)^{C_{l1}^* - 1} \exp\{-\frac{C_{l2}^*}{\sigma_\varepsilon^2} - C_{l3}^* [\log \sigma_\varepsilon^2]^2\}}{\mathcal{J}(0, C_{l1}^*, C_{l3}^*, C_{l2}^*)}, \;\; \sigma_\varepsilon > 0.
\end{aligned} \tag{2.11}$$

Here $\mathcal{J}$ is defined as in Definition 1.10 and

$$\begin{aligned}
C^*_{l1} &= -\frac{n}{2} + \mu_{q(a_\varepsilon)}\mu_{q(1/b_\varepsilon)}, \\
C^*_{l2} &= \frac{1}{2}[\|\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})}\|^2 + \mathrm{tr}(\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})})], \\
C^*_{l3} &= \frac{1}{2}\mu_{q(1/b_\varepsilon)}.
\end{aligned}$$

Let $\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})}$ and $\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}$ denote the mean vector and covariance matrix for the Multivariate Normal density function $q^*(\beta_0, \boldsymbol{\beta})$, $\mu_{q(a_\varepsilon)}$ and $\sigma^2_{q(a_\varepsilon)}$ denote the mean and variance for the Normal density function $q^*(a_\varepsilon)$, and $A_{q(b_\varepsilon)}$ and $B_{q(b_\varepsilon)}$ denote the shape and rate parameters for $q^*(b_\varepsilon)$.

Let $\boldsymbol{C} = [1, \boldsymbol{X}]$. The convergence of algorithm (2.3.2) can be monitored using the following expression for the lower bound on the marginal log-likelihood:

$$\begin{aligned}
\log \underline{p}(\boldsymbol{y}; q) \quad = \quad & \frac{2+p}{2} - \frac{n+1}{2}\log(2\boldsymbol{\pi}) - \frac{1}{2}\log(\sigma^2_{\beta_0}) - \frac{\mu^2_{q(\beta_0)} + \sigma^2_{q(\beta_0)}}{2\sigma^2_{\beta_0}} \\
& -\frac{p}{2}\log(\sigma^2_{\boldsymbol{\beta}}) - \frac{1}{2\sigma^2_{\boldsymbol{\beta}}}\{\|\mu_{q(\boldsymbol{\beta})}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\} \\
& +\frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}| \\
& -\frac{1}{2}\log(\sigma^2_a) - \frac{1}{2}\frac{\sigma^2_{q(a_\varepsilon)} + \mu^2_{q(a_\varepsilon)}}{\sigma^2_a} + \frac{1}{2}\log(\sigma^2_{q(a_\varepsilon)}) \\
& +A_b\log(B_b) - \log\Gamma(A_b) \\
& -A_{q(b_\varepsilon)}\log(B_{q(b_\varepsilon)}) + \log\Gamma(A_{q(b_\varepsilon)}) \\
& +\log\mathcal{J}(0, C^*_{l1}, C^*_{l3}, C^*_{l2}) - \mu_{q(a_\varepsilon)}\mu_{q(1/b_\varepsilon)}\mu_{(\log\sigma^2_\varepsilon)} \\
& +\frac{1}{2}\mu_{q(1/b_\varepsilon)}\mu_{q([\log\sigma^2_\varepsilon]^2)}.
\end{aligned}$$

Algorithm 2.3.2: MFVB iterative scheme for obtaining the parameters of the optimal densities $q^*(\beta_0, \boldsymbol{\beta})$, $q^*(a_\varepsilon)$, $q^*(b_\varepsilon)$ and $q^*(\sigma_\varepsilon^2)$ for linear regression with a Log-Normal prior in model (2.10) for σ_ε^2.

Initialize $\mu_{q(1/\sigma_\varepsilon^2)}$, $\mu_{q(1/b_\varepsilon)}$, $\mu_{q(a_\varepsilon)}$, $\mu_{q(\log\sigma_\varepsilon^2)}$ and $\mu_{q([\log\sigma_\varepsilon^2]^2)}$;
Cycle

$$\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})} \leftarrow \left\{\mu_{q(1/\sigma_\varepsilon^2)}\boldsymbol{C}^T\boldsymbol{C} + \text{blockdiag}\left[(\sigma_{\beta_0}^2)^{-1}, (\sigma_{\boldsymbol{\beta}}^2)^{-1}\boldsymbol{I}_p\right]\right\}^{-1}$$

$$\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})} \leftarrow \mu_{q(1/\sigma_\varepsilon^2)}\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}\boldsymbol{C}^T\boldsymbol{y}$$

$$A_{q(b_\varepsilon)} \leftarrow A_b + \frac{1}{2}$$

$$B_{q(b_\varepsilon)} \leftarrow B_b + \frac{1}{2}\{\mu_{q(a_\varepsilon)}^2 + \sigma_{q(a_\varepsilon)}^2 - 2\mu_{q(a_\varepsilon)}\mu_{q(\log\sigma_\varepsilon^2)} + \mu_{q([\log\sigma_\varepsilon^2]^2)}\}$$

$$\mu_{q(1/b_\varepsilon)} \leftarrow \frac{A_{q(b_\varepsilon)}}{B_{q(b_\varepsilon)}}$$

$$\sigma_{q(a_\varepsilon)}^2 \leftarrow [\frac{1}{\sigma_a^2} + \mu_{q(1/b_\varepsilon)}]^{-1}$$

$$\mu_{q(a_\varepsilon)} \leftarrow \sigma_{q(a_\varepsilon)}^2\mu_{q(1/b_\varepsilon)}\mu_{q(\log\sigma_\varepsilon^2)}$$

$$C_{l1}^* \leftarrow -\frac{n}{2} + \mu_{q(a_\varepsilon)}\mu_{q(1/b_\varepsilon)}$$

$$C_{l2}^* \leftarrow \frac{1}{2}[(\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})})^2 + \text{tr}(\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})})]$$

$$C_{l3}^* \leftarrow \frac{1}{2}\mu_{q(1/b_\varepsilon)}$$

$$\mu_{q(1/\sigma_\varepsilon^2)} \leftarrow \frac{\mathcal{J}(0, C_{l1}^* - 1, C_{l3}^*, C_{l2}^*)}{\mathcal{J}(0, C_{l1}^*, C_{l3}^*, C_{l2}^*)}$$

$$\mu_{q(\log\sigma_\varepsilon^2)} \leftarrow \frac{\mathcal{J}(1, C_{l1}^*, C_{l3}^*, C_{l2}^*)}{\mathcal{J}(0, C_{l1}^*, C_{l3}^*, C_{l2}^*)}$$

$$\mu_{q([\log\sigma_\varepsilon^2]^2)} \leftarrow \frac{\mathcal{J}(2, C_{l1}^*, C_{l3}^*, C_{l2}^*)}{\mathcal{J}(0, C_{l1}^*, C_{l3}^*, C_{l2}^*)}$$

until the increase in $\log \underline{p}(\boldsymbol{y}; q)$ is negligible.

2.4 Stepwise Variable Selection

When we use the highest posterior probability as a criterion for the model choice, all potential regression models need to be considered. If the number of variables of interest in the regression model is p, we need to fit 2^p regression models for linear variable selection. We have to spend a lot of time dealing with the variable selection if p is large, if we consider all possible models. Therefore a stepwise procedure is used for variable selection.

2.4.1 Introduction to stepwise variable selection

In stepwise variable selection, we build models by adding new terms or dropping a term and seeing how much this action improve the regression model fitting. Because each candidate variable is added or dropped iteratively, the computation can be time-consuming.

Generally, the stepwise variable selection process includes the following three steps:

1. Start variable selection for initial candidate model.

2. Add or drop each term within the candidate model and calculate the value of the criterion for variable selection. The "best" model will be used as a new candidate model.

3. Stop when no new candidate model can increase the criterion.

In the `gam` package (Hastie, 2015), the AIC statistic is used as the criterion for variable selection in the function `step.gam()`. In our MFVB variable selection, we use the model posterior probability as the criterion value in the variable selection. If we are just adding or improving a new term in the candidate model, this stepwise search process is named "forward"; if we are only removing or reducing a term in

the candidate model, this stepwise search process is named "backward". If we are adding or dropping a term in each step, this process of step-wise search is named "both". For the "forward" stepwise search, the NULL model (without any variables) is always selected as the initial candidate model; for the "backward" and "both" stepwise searches, the FULL model (including all of the variables) should be considered as the initial candidate model. For a linear regression model with p candidate variables, if we start with a NULL model by using the "forward" search or a FULL model by using the "backward" search, the maximal number of fitted model is $p(p+1)/2$. It is always less than 2^p.

2.4.2 Stepwise linear variable selection

We have given an overview of stepwise variable selection. Now we focus on stepwise linear variable selection. In the linear variable selection, we only need to consider whether a variable should be in or out of the candidate model. Algorithm 2.4.1 is a stepwise algorithm for linear variable selection.

Algorithm 2.4.1: Algorithm for stepwise linear variable selection

1. Initialize the candidate model (for the "forward" step-wise search, the NULL model is selected as the initial candidate model; for the "backward" and "both" stepwise searches, the Full model is selected as the initial candidate model).
2. Add each candidate variable into the candidate model one at a time or remove each variable from the candidate model one at a time. Calculate the the posterior probability of each new model. The model with largest posterior probability is set as the new candidate model.
3. Stop if the posterior probability model cannot be increased.

Example of stepwise linear variable selection

We use a example to describe the stepwise linear variable selection. The data were generated from the model

$$y_i = \beta_0 + x_{1i} + 2 \times x_{2i} + 0 \times x_{3i}\varepsilon_i, \;\; 1 \leq t \leq n,$$

The number of observation is $n = 400$, and the number of predictor variables of interest is $p = 3$. The predictors x_1, x_2 and x_3 are generated from uniform distributions on $(0, 1)$, where x_1 and x_2 are linear terms and x_3 is noise. The errors ε_i are independent and identically distributed from a normal distribution with mean 0 and variance 0.09. We use a "forward" stepwise search and set the NULL model to be the candidate model. We let M_{x_1} denote the model including the variable x_1 and $\log \underline{p}(y; M)$ denote the lower bound on the marginal log-likelihood, which is the criterion value used to evaluate the variable selection.

initialization

- $\underline{p}(y; M_{NULL}) = -434.9$

Step 1 Candidate model M_{NULL}.

- $\underline{p}(y; M_{x_1}) = -402.4$
- $\underline{p}(y; M_{x_2}) = -239.0$
- $\underline{p}(y; M_{x_3}) = -438.7$
- M_{x_2} is new candidate model.

Step 2 Candidate model M_{x_2}

- $\underline{p}(y; M_{x_2,x_1}) = -85.9$
- $\underline{p}(y; M_{x_2,x_3}) = -241.2$

- M_{x_1,x_2} is the new candidate model.

Step 3 Candidate model M_{x_1,x_2}

- $\underline{p}(y; M_{x_1,x_2,x_3}) = -88.9$
- Stop. Model M_{x_1,x_2} is the estimated optimal model.

2.5 Simulation Study

Earlier, we listed a basic model framework that can be used in Bayesian variable selection. In this section, a simulation study will be used to analyze the effect of alternative parameter priors on variable selection.

2.5.1 Simulation for linear variable selection

The basic model for Bayesian linear regression has been given in Section 2.1. Combining the alternative variance (scale) parameter and alternative fixed effect parameter priors, we can get three derivative models: linear regression with the ridge penalized method for the fixed effects, linear regression with Half-Cauchy prior in σ_ε^2 and linear regression with log-normal prior in σ_ε^2.

Setup of the numerical simulation, and the result

We generated a series of simulations with different correlations between candidate predictor variables and different signal-to-noise ratios (Hastie *et al.*, 2009, page 649) to evaluate the performance of the linear variable selection. We consider the linear model form:

$$y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + \varepsilon_i, \;\; 1 \leq i \leq n.$$

The number of observations is $n = 400$, and the number of predictors of interest is $p = 3$. The $\boldsymbol{X}_i$, $1 \leq i \leq n$, are generated from a Multivariate Normal distribution,

$N(\mathbf{0}, \mathbf{\Sigma})$, where

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho_x & \cdots & \rho_x \\ \rho_x & 1 & \cdots & \rho_x \\ \vdots & \vdots & \ddots & \vdots \\ \rho_x & \rho_x & \cdots & 1 \end{pmatrix}.$$

Then, $\rho_x = 0.2, 0.5, 0.8$ correspond to low correlation, medium correlation and high correlation. The ε_i is generated from the $N(0, \sigma^2)$ distribution. Following Hastie *et al.* (2009), the standard deviation, σ, was chosen in each case so that the signal-to-noise ratio (SNR) is equal to a fixed value. We set the SNR equal to 1, 5 and 25 to represent low, medium and high values. The true value of $\boldsymbol{\beta}$ is $(0, 0.5, 0.3)^T$. In this Thesis, all data have be standardised.

SNR	1	1	1	5	5	5	25	25	25
ρ	0.2	0.5	0.8	0.2	0.5	0.8	0.2	0.5	0.8
OM	0.02	0	0	0.91	0.53	0.01	1.00	1.00	0.98
LN	0.02	0	0	0.90	0.53	0.01	1.00	1.00	0.98
HC	0.02	0	0	0.90	0.53	0.01	1.00	1.00	0.98
RP	0.90	0.73	0.32	0.97	0.92	0.83	0.98	0.98	0.97

Table 2.1: The correct rate of variable selection for the original linear model (OM),the ridge penalized method in the fixeds effect model (RP), the Log-Normal prior in σ_ε^2 (LN) and the Half-Cauchy prior in σ_ε^2 (HC).

Table 2.1 shows that the correct rate (number of estimated optimal models that are the same as the true model / number of simulations) of variable selection has not significantly improved by updating the prior distribution of σ_ε^2 from Inverse Gamma to Log-Normal prior or Half-Cauchy prior. Correspondingly, the correct rate has been significantly increased by using a ridge penalized method for the fixed effect term.

2.5.2 Simulation for non-linear variable selection

The simulation has shown that the correct rate of variable selection cannot be improved by using a Log-Normal prior or a Half-Cauchy prior as the prior distribution for the error variance σ_ε^2. In this section, we will evaluate the performance of non-linear variable selection by using various prior distributions on the random variance term σ_u^2. The following model structure will be considered: (1) original mixed model, (2) mixed model with the ridge penalized method for the fixed effects and a Inverse Gamma prior for the random variance, (3) mixed model with the ridge penalized method for the fixed effects and a Log-Normal prior for the random variance, and (4) mixed model with the ridge penalized method for the fixed effects and a Half-Cauchy prior for the random variance. (The corresponding lower bound of the marginal likelihood, Algorithm and derivation are in Appendices 2.D, 2.E and 2.F)

We consider the non-linear model form

$$y_i = x_{1,i} + \frac{1}{2} m_j(x_{2,i}) + 0 \times x_{3,i} + \varepsilon_i, \;\; 1 \leq i \leq n,$$

where function $m_j = \sqrt{x(1-x)} + \sin(\frac{2\pi(1+2^{1.8-0.8j})}{x+2^{1.8-0.8j}})$, $j = 1$ and 3 and x_1 is a linear predictor variable, x_2 is non-linear predictor variable and x_3 is a noise variable. The number of observations is $n = 400$. The vector $(x_{1,i}, x_{2,i}, x_{3,i})^T$ was generated from a Uniform(0,1) distribution with fixed correlation ρ between the variables of a triplet. Then, $\rho = 0.2$ and 0.8 correspond to low correlation and high correlation. The ε_i, $i = 1, 2, ..., n$, are independent and identically distributed from the Normal distribution observations with mean 0 and variance σ_ε^2, where $\sigma_\varepsilon = 0.2$ and 0.8.

j		1	1	1	1	3	3	3	3
ρ		0.8	0.2	0.8	0.2	0.8	0.2	0.8	0.2
σ		0.8	0.8	0.2	0.2	0.8	0.8	0.2	0.2
Model (1)	x_1	0.67	1.00	1.00	1.00	0.46	0.99	1.00	1.00
	x_2	0.23	1.00	1.00	1.00	0.80	0.81	1.00	1.00
	x_3	0.92	0.99	0.98	1.00	0.93	0.93	1.00	1.00
Model (2)	x_1	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	x_2	0.89	1.00	1.00	1.00	0.99	0.99	1.00	1.00
	x_3	0.87	0.99	0.98	0.99	0.91	0.92	0.99	0.99
Model (3)	x_1	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	x_2	0.89	1.00	1.00	1.00	0.99	0.99	1.00	1.00
	x_3	0.87	0.98	0.98	0.98	0.92	0.91	0.99	0.99
Model (4)	x_1	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	x_2	0.89	1.00	1.00	1.00	1.00	0.99	1.00	1.00
	x_3	0.86	0.99	1.00	0.99	0.92	0.93	1.00	0.99

Table 2.2: The correct rate of non-linear variable selection when using four model structures: model (1) is the original mixed model, model (2) is the mixed model with ridge penalized method for the in fixed effects, model (3) is the mixed model with ridge penalized method for the fixed effects and a Log-Normal prior for random variance, and model (4) is the mixed model with ridge penalized method for the fixed effects and a Half-Cauchy prior for the random variance.

Table 2.2 shows that the correct rate has not been significantly improved by updating the prior distribution of σ_u^2 from the Inverse Gamma distribution to the Log-Normal distribution or to the Half-Cauchy distribution. For the fixed effect term, the correct rate can be significantly increased by adding a ridge penalized method for the the fixed effects if the predictor variables have a strong correlation. Therefore, we suggest that the model which includes the ridge penalized method for the fixed effects and a Inverse Gamma prior for the random variance term is best for MFVB non-linear variable selection.

2.5.3 Simulation for stepwise variable selection

Similarly to simulation for linear variable selection, we consider the linear model form

$$y_i = \boldsymbol{X}_i\boldsymbol{\beta} + \varepsilon_i, \;\; 1 \leq i \leq n.$$

The number of observations is $n = 200$, and the number of predictors of interest is $p = 10$. The $\boldsymbol{X}_i$, $1 \leq i \leq n$, are generated from a Multivariate Normal distribution, $N(\mathbf{0}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$$

Then $\rho = 0$, 0.2, 0.5, 0.8 correspond to non correlation, low correlation, medium correlation and high correlation. The ε_i is generated from the $N(0, \sigma^2)$ distribution. Following Hastie *et al.* (2009), the standard deviation, σ, was chosen in each case so that the signal-to-noise ratio (SNR) is equal to a fixed value. We set the SNR equal to 1, 5 and 25 to represent low, medium and high values. The true value of $\boldsymbol{\beta}$ is $(1, 2, 0, 0, 0, 1, 2, 0, 0, 0)^T$. The following two performance measures were calculated:

1. The concordance rate (#P1): consider the stepwise model selection method and MFVB full model variable selection method. then
   $\#P1 = \dfrac{\text{number of predictors } x \text{ that are selected/dropped in both method}}{\text{number of simulations}}$;

2. The accuracy rate (#P2): if potential predictor x is included in the true model, then
   $\#P2 = \dfrac{\text{number of estimated optimal models that include } x}{\text{number of simulations}}$;
   otherwise,
   $\#P2 = \dfrac{\text{number of estimated optimal models that exclude } x}{\text{number of simulations}}$.

ρ	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
#P1	0.99	1.00	0.99	0.99	0.99	1.00	0.99	0.99	0.99	1.00
#P2	0.94	.98	0.98	0.95	0.96	0.95	0.98	0.96	0.97	0.95
True coeff.	1	2	0	0	0	1	2	0	0	0

Table 2.3: The concordance rate (#P1) and accuracy rate (#P2) over 400 simulations for $\rho = 0$, 0.2, 0.5, 0.8 and SNR = 1, 5, 25, where $p = 10$ and $n = 200$.

The result is shown in Table 2.3. There is good agreement between the results from the MFVB full model variable selection and the stepwise variable selection. The effects of the correlation among the predictors and the signal-to-noise ratio for variable selection are not significant. Therefore, we can use the stepwise variable selection method instead of the full model method.

2.6 Variable Selection for a Binary Response

Binary response data can be fitted by using a probit regression model. Similarly to variable selection for a Gaussian response, we obtain the MFVB algorithm for Bayesian linear probit regression and non-linear probit regression model. Nextly, the lower bound on the marginal log-likelihood, $\underline{p}(\boldsymbol{y}|M_k)$, is used to obtain the posterior probabilities of each model and to select the variables based on the best posterior probabilities.

2.6.1 Linear variable selection for a binary response

We have shown that the results of variable selection can be improved by adding the ridge penalized method for the fixed effects in the Gaussian response case. We also use the ridge penalized method for the fixed coefficient of the probit regression

model. The Bayesian linear probit regression is:

$$\begin{aligned}
&p(y_i|\beta_0, \boldsymbol{\beta}) \overset{\text{ind.}}{\sim} \text{Bernoulli}(\Phi((\mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta})_i)),\\
&\beta_0 \sim N(0, \sigma^2_{\beta_0}),\\
&\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma^2_{\boldsymbol{\beta}}\boldsymbol{I}),\\
&\sigma^2_{\boldsymbol{\beta}} \sim \text{Inverse-Gamma}(A_{\boldsymbol{\beta}}, B_{\boldsymbol{\beta}}),
\end{aligned} \tag{2.12}$$

where A_ε, B_ε, $\sigma^2_{\beta_0}$, > 0 are hyperparameters. Then, with the auxiliary variables $\boldsymbol{a}$ as in Result 1.15, we can re-write model (2.12) as

$$\begin{aligned}
&p(y_i|a_i) = I(a_i \geq 0)^{y_i} I(a_i < 0)^{1-y_i},\\
&\boldsymbol{a}|\beta_0, \boldsymbol{\beta} \overset{\text{ind.}}{\sim} N(\mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{I}),\\
&\beta_0 \sim N(0, \sigma^2_{\beta_0}),\\
&\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma^2_{\boldsymbol{\beta}}\boldsymbol{I}),\\
&\sigma^2_{\boldsymbol{\beta}} \sim \text{Inverse-Gamma}(A_{\boldsymbol{\beta}}, B_{\boldsymbol{\beta}}).
\end{aligned} \tag{2.13}$$

The product restriction that we impose here is:

$$q(\beta_0, \boldsymbol{\beta}, \sigma^2_{\boldsymbol{\beta}}, \boldsymbol{a}) = q(\beta_0, \boldsymbol{\beta}) q(\sigma^2_{\boldsymbol{\beta}}) q(\boldsymbol{a}).$$

Then, as shown in Appendix 2.G, the optimal q^* densities for the parameters in model (2.10) take the form:

$q^*(\beta_0, \boldsymbol{\beta})$ is a Multivariate Normal density function,

$q^*(\sigma^2_{\boldsymbol{\beta}})$ is an Inverse Gamma density function, and

$$q^*(a_i) = \left\{ \left[\frac{I(\boldsymbol{a}_i \geq 0)}{\Phi((\boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})})_i)} \right]^{y_i} \left[\frac{I(\boldsymbol{a}_i < 0)}{1 - \Phi((\boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})})_i)} \right]^{1-y_i} \right\} \phi\left((\boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})})_i \right).$$

Let $\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})}$ and $\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}$ denote the mean vector and covariance matrix for the Multivariate Normal density function $q^*(\beta_0, \boldsymbol{\beta})$, and $A_{q(\sigma^2_{\boldsymbol{\beta}})}$ and $B_{q(\sigma^2_{\boldsymbol{\beta}})}$ denote the shape and rate parameters for $q^*(\sigma^2_{\boldsymbol{\beta}})$. Let $\boldsymbol{C} = [1, \boldsymbol{X}]$. The convergence of algorithm (2.6.1) can be monitored using the following expression for the lower bound on the marginal log-likelihood:

$$\begin{aligned}
\log \underline{p}(\boldsymbol{y}; q) \;=\; & \frac{1+p}{2} - \frac{1}{2}\log(\sigma^2_{\beta_0}) - \frac{1}{2}\frac{\mu^2_{q(\beta_0)} + \sigma^2_{q(\beta_0)}}{\sigma^2_{\beta_0}} \\
& + \boldsymbol{y}^T \log[\Phi(\boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})})] + (\mathbf{1}_n - \boldsymbol{y})^T \log[1_n - \Phi(\boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})})] \\
& + A_{\boldsymbol{\beta}} \log(B_{\boldsymbol{\beta}}) - \log\Gamma(A_{\boldsymbol{\beta}}) - A_{q(\sigma^2_{\boldsymbol{\beta}})} \log(B_{q(\sigma^2_{\boldsymbol{\beta}})}) + \log\Gamma(A_{q(\sigma^2_{\boldsymbol{\beta}})}) \\
& + \frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}| - \tfrac{1}{2}\mathrm{tr}(\boldsymbol{C}^T\boldsymbol{C})\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}.
\end{aligned}$$

Algorithm 2.6.1: MFVB iterative scheme for obtaining the parameters of the optimal densities $q^*(\beta_0, \boldsymbol{\beta})$, $q^*(\sigma^2_{\boldsymbol{\beta}})$ and $q^*(a_i)$ for model (2.13)

Initialize $\mu_{q(1/\sigma^2_{\boldsymbol{\beta}})}$;

Cycle

$$\begin{aligned}
\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})} &\leftarrow \left\{ \boldsymbol{C}^T\boldsymbol{C} + \mathrm{blockdiag}\left[(\sigma^2_{\beta_0})^{-1}, \mu_{q(1/\sigma^2_{\boldsymbol{\beta}})} \boldsymbol{I}_p \right] \right\}^{-1} \\
\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})} &\leftarrow \boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})} \boldsymbol{C}^T \mu_{q(a)} \\
A_{q(\sigma^2_{\boldsymbol{\beta}})} &\leftarrow \frac{p}{2} + A_{\boldsymbol{\beta}} \\
B_{q(\sigma^2_{\boldsymbol{\beta}})} &\leftarrow B_{\boldsymbol{\beta}} + \tfrac{1}{2}\left\{ \| \boldsymbol{\mu}_{q(\boldsymbol{\beta})} \|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}) \right\} \\
\mu_{q(1/\sigma^2_{\boldsymbol{\beta}})} &\leftarrow \frac{A_{q(\sigma^2_{\boldsymbol{\beta}})}}{B_{q(\sigma^2_{\boldsymbol{\beta}})}} \\
\mu_{q(a_i)} &\leftarrow (\boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})})_i + \frac{\phi((\boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})})_i)}{\Phi((\boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})})_i)^{y_i}[1 - \Phi((\boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})})_i)]^{1-y_i}}
\end{aligned}$$

until the increase in $\log \underline{p}\,(\boldsymbol{y}; q)$ is negligible.

2.6.2 Non-linear variable selection for a binary response

The Bayesian non-linear probit regression is:

$$
\begin{aligned}
&p(y_i|\beta_0, \boldsymbol{\beta}, \boldsymbol{u}) \overset{\text{ind.}}{\sim} \text{Bernoulli}\{\Phi((\mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u})_i)\},\\
&\beta_0 \sim N(0, \sigma^2_{\beta_0}),\\
&\boldsymbol{\beta}|\sigma^2_{\boldsymbol{\beta}} \sim N(\mathbf{0}, \sigma^2_{\boldsymbol{\beta}}\boldsymbol{I}),\\
&\sigma^2_{\boldsymbol{\beta}} \sim \text{IG}(A_{\boldsymbol{\beta}}, B_{\boldsymbol{\beta}}),\\
&\boldsymbol{u}_\ell|\sigma_{\boldsymbol{u}_\ell} \sim N(0, \sigma^2_{\boldsymbol{u}_\ell} I_{K_\ell}) \;\; 1 \leq \ell \leq r,\\
&\sigma_{\boldsymbol{u}_\ell} \sim \text{Half-Cauchy}(A),
\end{aligned}
\tag{2.14}
$$

where we use a Half-Cauchy prior on the standard deviation of random effect. Then, with auxiliary variables $\boldsymbol{a}$ as in Result 1.15 and b as in Result 1.5, we can re-write model (2.14) as:

$$
\begin{aligned}
&p(y_i|a_i) = I(a_i \geq 0)^{y_i} I(a_i < 0)^{1-y_i}, \;\; 1 \leq i \leq n,\\
&\boldsymbol{a}|\beta_0, \boldsymbol{\beta}, \boldsymbol{u} \sim N(\mathbf{1}\beta_0 + X\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, I_n),\\
&\beta_0 \sim N(0, \sigma^2_{\beta_0}),\\
&\boldsymbol{\beta}|\sigma^2_{\boldsymbol{\beta}} \sim N(\mathbf{0}, \sigma^2_{\boldsymbol{\beta}}\boldsymbol{I}),\\
&\sigma^2_{\boldsymbol{\beta}} \sim \text{IG}(A_{\boldsymbol{\beta}}, B_{\boldsymbol{\beta}}),\\
&\boldsymbol{u}_\ell|\sigma_{\boldsymbol{u}_\ell} \sim N(0, \sigma^2_{\boldsymbol{u}_\ell} I_{K_\ell}) \;\; 1 \leq \ell \leq r,\\
&\sigma^2_{\boldsymbol{u}_\ell}|b \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(1/2, 1/b),\\
&b \sim \text{Inverse-Gamma}(1/2, 1/A^2).
\end{aligned}
\tag{2.15}
$$

Here $\sigma^2_{\beta_0}$, A, $A_{\boldsymbol{\beta}}$, $B_{\boldsymbol{\beta}} > 0$ are hyperparameters. The product restriction that we impose here is:

$$
q(\beta_0, \boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{a}, \sigma^2_{\boldsymbol{\beta}}, b, \sigma^2_{\boldsymbol{u}_1}, ..., \sigma^2_{\boldsymbol{u}_r}) = q(\beta_0, \boldsymbol{\beta}, \boldsymbol{u}) q(\boldsymbol{a}) q(\sigma^2_{\boldsymbol{\beta}}) q(b) q(\sigma^2_{\boldsymbol{u}_1}) ... q(\sigma^2_{\boldsymbol{u}_r}).
$$

Then, the optimal q^* densities for the parameters in model (2.15) take the form:

$$
\begin{aligned}
&q^*(\beta_0, \boldsymbol{\beta}, \boldsymbol{u}) \text{ is a Multivariate Normal density function,} \\
&q^*(\sigma^2_{\boldsymbol{\beta}}) \text{ is an Inverse Gamma density function,} \\
&q^*(\sigma^2_{\boldsymbol{u}_\ell}) \text{ is an Inverse Gamma density function,} \\
&q^*(b) \text{ is an Inverse Gamma density function, and} \\
&q^*(a_i) = \left\{ \left[\frac{I(\boldsymbol{a}_i \geq 0)}{\Phi((\boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{n})})_i)} \right]^{y_i} \left[\frac{I(\boldsymbol{a}_i < 0)}{1 - \Phi((\boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})})_i)} \right]^{1-y_i} \right\} \\
&\qquad \times \phi\left((\boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})})_i \right).
\end{aligned}
\tag{2.16}
$$

Let $\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})}$ and $\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})}$ denote the mean vector and covariance matrix for the Multivariate Normal density function $q^*(\beta_0, \boldsymbol{\beta}, \boldsymbol{u})$, and $A_{q(\sigma^2_{\boldsymbol{\beta}})}$ and $B_{q(\sigma^2_{\boldsymbol{\beta}})}$ denote the shape and rate parameters for $q^*(\sigma^2_{\boldsymbol{\beta}})$. A similar definition was used for the parameters in $q^*(\sigma^2_{\boldsymbol{u}_\ell})$ and $q^*(b)$. Let $\boldsymbol{C} = [1, \boldsymbol{X}, \boldsymbol{Z}]$. The convergence of algorithm (2.6.2) can be monitored using the following expression for the lower bound on the marginal log-likelihood:

$$
\begin{aligned}
\log \underline{p}(\boldsymbol{y}; q) \;=\; & \frac{1 + p + K_\ell}{2} - \frac{1}{2}\log(\sigma^2_{\beta_0}) - \frac{\mu^2_{q(\beta_0)} + \sigma^2_{q(\beta_0)}}{2\sigma^2_{\beta_0}} \\
& + \frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})}| + \sum_{\ell=1}^{r} \log\Gamma(A_{q(\sigma^2_{\boldsymbol{u}_\ell})}) \\
& + A_{\boldsymbol{\beta}}\log(B_{\boldsymbol{\beta}}) - \log\Gamma(A_{\boldsymbol{\beta}}) \\
& - A_{q(\sigma^2_{\boldsymbol{\beta}})}\log(B_{q(\sigma^2_{\boldsymbol{\beta}})}) + \log\Gamma(A_{q(\sigma^2_{\boldsymbol{\beta}})}) \\
& - \log(A) - \frac{1 + r}{2}\log(\pi) - A_{q(b)}\log(B_{q(b)}) + \log\Gamma(A_{q(b)}) \\
& - \sum_{\ell=1}^{r} A_{q(\sigma^2_{\boldsymbol{u}_\ell})}\log(B_{q(\sigma^2_{\boldsymbol{u}_\ell})}) \\
& - \tfrac{1}{2}\mathrm{tr}(\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})}) + \boldsymbol{y}^T\log[\Phi(C\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})})] \\
& + (\mathbf{1}_n - \boldsymbol{y})^T\log[1_n - \Phi(\boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})})].
\end{aligned}
$$

Algorithm 2.6.2: MFVB iterative scheme for obtaining the parameters of the optimal densities $q^*(\beta_0, \boldsymbol{\beta}, \boldsymbol{u})$, $q^*(\sigma^2_{\boldsymbol{\beta}})$, $q^*(\sigma^2_{\boldsymbol{u}_\ell})$, $q^*(b)$ and $q^*(a_i)$ for probit non-linear variable selection model (2.15)

Initialize $\mu_{q(1/\sigma^2_{\boldsymbol{\beta}})}$, $\mu_{q(1/\sigma^2_{\boldsymbol{u}_1})}$,...,$\mu_{q(1/\sigma^2_{\boldsymbol{u}_r})}$;
Cycle

$$\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \big\{\boldsymbol{C}^T\boldsymbol{C} + \text{blockdiag}\big[(\sigma^2_{\beta_0})^{-1}, \mu_{q(1/\sigma^2_{\boldsymbol{\beta}})}\boldsymbol{I}_p, \mu_{q(1/\sigma^2_{\boldsymbol{u}_1})}\boldsymbol{I}_{K_1}, ..., \mu_{q(1/\sigma^2_{\boldsymbol{u}_r})}\boldsymbol{I}_{K_r}\big]\big\}^{-1}$$

$$\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T\mu_{q(\boldsymbol{a})}$$

$$A_{q(\sigma^2_{\boldsymbol{\beta}})} \leftarrow \frac{p}{2} + A_{\boldsymbol{\beta}}$$

$$B_{q(\sigma^2_{\boldsymbol{\beta}})} \leftarrow B_{\boldsymbol{\beta}} + \tfrac{1}{2}\left\{\| \boldsymbol{\mu}_{q(\boldsymbol{\beta})} \|^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\right\}$$

$$\mu_{q(1/\sigma^2_{\boldsymbol{\beta}})} \leftarrow \frac{A_{q(\sigma^2_{\boldsymbol{\beta}})}}{B_{q(\sigma^2_{\boldsymbol{\beta}})}}$$

$$A_{q(\sigma^2_{\boldsymbol{u}_\ell})} \leftarrow \frac{K_{\boldsymbol{u}_\ell}}{2} + \tfrac{1}{2}$$

$$B_{q(\sigma^2_{\boldsymbol{u}_\ell})} \leftarrow \mu_{q(1/b)} + \tfrac{1}{2}\left\{\| \boldsymbol{\mu}_{q(\boldsymbol{u}_\ell)} \|^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_\ell)})\right\}$$

$$\mu_{q(1/\sigma^2_{\boldsymbol{u}_\ell})} \leftarrow \frac{A_{q(\sigma^2_{\boldsymbol{u}_\ell})}}{B_{q(\sigma^2_{\boldsymbol{u}_\ell})}}$$

$$A_{q(b)} \leftarrow \frac{1+r}{2}$$

$$B_{q(b)} \leftarrow \frac{1}{A^2} + \sum_{\ell=1}^{r} \mu_{q(1/\sigma^2_{\boldsymbol{u}_\ell})}$$

$$\mu_{q(1/b)} \leftarrow \frac{A_{q(b)}}{B_{q(b)}}$$

$$\mu_{q(a_i)} \leftarrow (\boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})})_i + \frac{\phi((\boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})})_i)}{\Phi((\boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})})_i)^{y_i}[1-\Phi((\boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})})_i)]^{1-y_i}}$$

until the increase in $\log \underline{p}(y;q)$ is negligible.

2.6.3 Simulation study

The simulation will be carried out to evaluate the performance of variable selection for a binary response.

Linear variable selection for binary response

Similarly to Hu and Johnson (2009), we consider the model:

$$y_i \overset{\text{ind.}}{\sim} \text{Bernoulli}\{\Phi(\boldsymbol{X}\boldsymbol{\beta})_i\}, 1 \leq i \leq n.$$

We set the number of observations $n = 100$, and the number of predictors of interest $p = 10$. The $\boldsymbol{X}_i$, $1 \leq i \leq n$, are generated from a Multivariate Normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho_x & \cdots & \rho_x \\ \rho_x & 1 & \cdots & \rho_x \\ \vdots & \vdots & \ddots & \vdots \\ \rho_x & \rho_x & \cdots & 1 \end{pmatrix}.$$

The true value of $\boldsymbol{\beta}$ is equal to $(0.5, 1, 1.5, 2, 2.5, 0, 0, 0, 0, 0)^T$, $\rho_x = 0$, 0.2, 0.5 and 0.8 correspond to non-correlation, low correlation, medium correlation and high correlation.

ρ	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
0	0.34	0.9	1	1	1	0	0.02	0.03	0.01	0
0.2	0.21	0.72	0.99	1	1	0	0.01	0.01	0.02	0.01
0.5	0.14	0.44	0.88	0.97	0.99	0.01	0.01	0.03	0.02	0.01
0.8	0.08	0.28	0.52	0.73	0.86	0.06	0.04	0.08	0.04	0.04
True coeff.	0.5	1	1.5	2	2.5	0	0	0	0	0

Table 2.4: Marginal probabilities that variables are selected for various ρ_x

The results for a simulation size of 500 are in Table 2.4, which lists the estimated marginal posterior probabilities that variables were included in a sampled model. For the low correlation cases (i.e., $\rho_x = 0$ and 0.2), the variables x_2, x_3, x_4 and x_5 were selected with high probability, and the variable x_1 was classified as the noise variable and ignored. For the medium and high correlation cases (i.e., $\rho_x = 0.5$ and 0.8), the variables x_3, x_4 and x_5 were selected with high probability, and the variable x_1 and x_2 were classified as the noise variable and ignored. This means that the variables with large value of their coefficients were easily selected and that the correct rate of variable selection was decreased by increasing the correlation between predictor variables. At the same time, the marginal probabilities of variables x_6 to x_{10} shown the noise variable could always be ignored for any values of the correlation. It means that the selected model contained fewer false positive results.

Non-linear variable selection for binary Response

The data were generated from the model

$$y_i \overset{\text{ind.}}{\sim} \text{Bernoulli}\{\Phi(x_{1i} + \sin(2\pi x_{2i}))\}, \;\; 1 \leq i \leq n.$$

The number of observation was $n = 400$, and the number of predictors of interest was $p = 3$. The predictors x_1, x_2 and x_3 were generated from the uniform distribution on $(0, 1)$, where x_1 is the linear term, x_2 is a non-linear variable and x_3 is noise. The accuracy rates, obtained using a simulation size of 500, for x_1, x_2 and x_3 were 1.00, 0.80 and 0.99. This means that 100% of the selected models included that the x_1 is a linear predictor, 80% of the selected models included x_2 as the non-linear term, and 99% of the selected models correctly omitted x_3. It can be seen that the result of variable selection was quite good.

In linear variable selection for binary response models, we find that the result

of variable selection will be affected by the values of the predictor's coefficients. We also generated another data set from the model

$$y_i \overset{\text{ind.}}{\sim} \text{Bernoulli}\{\Phi(0.5x_{1i} + \sin(0.5\pi x_{2i}))\}, \;\; 1 \leq i \leq n,$$

where the coefficients for each predictor has been reduced by a factor of 0.5. The accuracy rate for x_1, x_2 and x_3 were decreased to 0.57, 0.35 and 0.95. The performance of variable selection was significantly reduced.

2.7 Discussion

In this chapter, based on the MFVB inference method, a framework of MFVB variable selection was introduced, which used the lower bound of the marginal likelihood to estimate the posterior probabilities of candidate models and facilitate models selection.

For Gaussian response models, the accuracy of variable selection is significantly increased by adding a ridge penalized method for the fixed effects. The simulation result shows that MFVB inference can perform variable selection quite well. Moreover, the stepwise method can be used to speed up the variable selection when the number of predictors is large. The simulations show that stepwise MFVB variable selection performs similarly to MFVB variable selection with all possible subsets.

For the binary response case, the probit regression model can be used to select the variables, but the simulation result shows that result will always be affected by the value of a predictor's coefficients. This phenomenon also appear in Bayesian variable selection by MCMC (Hu & Johnson, 2009).

2.A Appendix: Derivation of Algorithm 2.2.1

2.A.1 Full conditionals

Full conditional for β_0 and $\boldsymbol{\beta}$

$$\log p(\beta_0, \boldsymbol{\beta}|\text{rest}) \;=\; -\frac{1}{2}\frac{\|\boldsymbol{y} - \boldsymbol{C}\tilde{\boldsymbol{\beta}}\|^2}{\sigma_\varepsilon^2} - \frac{1}{2}\tilde{\boldsymbol{\beta}}^T \boldsymbol{F}^{-1}\tilde{\boldsymbol{\beta}}$$

where

$$\tilde{\boldsymbol{\beta}} = \begin{bmatrix}\beta_0 \\ \boldsymbol{\beta}\end{bmatrix}, \;\; \boldsymbol{C} = [\mathbf{1}, \boldsymbol{X}]$$

$$\text{and } \boldsymbol{F}^{-1} = \text{diag}(\sigma_{\beta_0}^{-2}, \sigma_\beta^{-2}, ..., \sigma_{\boldsymbol{\beta}}^{-2}).$$

Derivation:

$$\begin{aligned} p(\beta_0, \boldsymbol{\beta}|\text{rest}) \;&\propto\; p(\boldsymbol{y}|\beta_0, \boldsymbol{\beta}, \sigma_\varepsilon^2)p(\boldsymbol{\beta}|\sigma_{\boldsymbol{\beta}}^2)p(\beta_0) \\ &=\; (2\pi)^{-n/2}\sigma_\varepsilon^{-n}\exp\left\{-\frac{\|\boldsymbol{y} - \mathbf{1}\beta_0 - \boldsymbol{X}\boldsymbol{\beta}\|^2}{2\sigma_\varepsilon^2}\right\} \\ &\quad \times(2\pi)^{-p/2}\sigma_{\boldsymbol{\beta}}^{-p}\exp\left\{-\frac{\|\boldsymbol{\beta}\|^2}{2\sigma_{\boldsymbol{\beta}}^2}\right\} \\ &\quad \times(2\pi)^{-1/2}\sigma_{\beta_0}^{-1}\exp\left\{-\frac{\beta_0^2}{2\sigma_{\beta_0}^2}\right\} \end{aligned}$$

Taking logarithms, we get

$$\begin{aligned} \log p(\beta_0, \boldsymbol{\beta}|\text{rest}) \;&=\; -\frac{\|\boldsymbol{y} - \mathbf{1}\beta_0 - \boldsymbol{X}\boldsymbol{\beta}\|^2}{2\sigma_\varepsilon^2} - \frac{\|\boldsymbol{\beta}\|^2}{2\sigma_{\boldsymbol{\beta}}^2} - \frac{\beta_0^2}{2\sigma_{\beta_0}^2} + \text{const.} \\ &=\; -\frac{\|\boldsymbol{y} - \boldsymbol{C}\tilde{\boldsymbol{\beta}}\|^2}{2\sigma_\varepsilon^2} - \frac{1}{2}\tilde{\boldsymbol{\beta}}^T\boldsymbol{F}^{-1}\tilde{\boldsymbol{\beta}} + \text{const.} \end{aligned}$$

Full conditional for σ_ε^2

$$\log p(\sigma_\varepsilon^2|\text{rest}) = (-1 - A_\varepsilon - \frac{n}{2})\log(\sigma_\varepsilon^2) - \frac{2B_\varepsilon + \|\boldsymbol{y} - \boldsymbol{C}\tilde{\boldsymbol{\beta}}\|^2}{2\sigma_\varepsilon^2} + \text{const.}$$

Derivation:

$$\begin{aligned} p(\sigma_\varepsilon^2|\text{rest}) &\propto p(\boldsymbol{y}|\beta_0, \boldsymbol{\beta}, \sigma_\varepsilon^2)p(\sigma_\varepsilon^2) \\ &= (2\pi)^{-n/2}\sigma_\varepsilon^{-n}\exp\left\{-\frac{\|\boldsymbol{y} - \mathbf{1}\beta_0 - \boldsymbol{X}\boldsymbol{\beta}\|^2}{2\sigma_\varepsilon^2}\right\} \\ &\quad \times \frac{B_\varepsilon^{A_\varepsilon}}{\Gamma(A_\varepsilon)}(\sigma_\varepsilon^2)^{-1-A_\varepsilon}\exp\left\{\frac{-B_\varepsilon}{\sigma_\varepsilon^2}\right\}. \end{aligned}$$

Taking logarithms, we get:

$$\log p(\sigma_\varepsilon^2|\text{rest}) = (-1 - A_\varepsilon - \frac{n}{2})\log(\sigma_\varepsilon^2) - \frac{2B_\varepsilon + \|\boldsymbol{y} - \boldsymbol{C}\tilde{\boldsymbol{\beta}}\|^2}{2\sigma_\varepsilon^2} + \text{const.}$$

Full conditional for $\sigma_{\boldsymbol{\beta}}^2$

$$\log p(\sigma_{\boldsymbol{\beta}}^2|\text{rest}) = (-1 - A_{\boldsymbol{\beta}} - \frac{p}{2})\log(\sigma_{\boldsymbol{\beta}}^2) - \frac{2B_{\boldsymbol{\beta}} + \|\boldsymbol{\beta}\|^2}{2\sigma_{\boldsymbol{\beta}}^2} + \text{const.}$$

Derivation:

$$\begin{aligned} p(\sigma_{\boldsymbol{\beta}}^2|\text{rest}) &\propto p(\boldsymbol{\beta}|\sigma_{\boldsymbol{\beta}}^2)p(\sigma_{\boldsymbol{\beta}}^2) \\ &= (2\pi)^{-p/2}\sigma_{\boldsymbol{\beta}}^{-p}\exp\left\{-\frac{\|\boldsymbol{\beta}\|^2}{2\sigma_{\boldsymbol{\beta}}^2}\right\} \\ &\quad \times \frac{B_{\boldsymbol{\beta}}^{A_{\boldsymbol{\beta}}}}{\Gamma(A_{\boldsymbol{\beta}})}(\sigma_{\boldsymbol{\beta}}^2)^{-1-A_{\boldsymbol{\beta}}}\exp\left\{\frac{-B_{\boldsymbol{\beta}}}{\sigma_{\boldsymbol{\beta}}^2}\right\}. \end{aligned}$$

Taking logarithms, we get:

$$\log p(\sigma_{\boldsymbol{\beta}}^2|\text{rest}) = (-1 - A_{\boldsymbol{\beta}} - \frac{n}{2})\log(\sigma_\varepsilon^2) - \frac{2B_{\boldsymbol{\beta}} + \|\boldsymbol{y} - \boldsymbol{C}\tilde{\boldsymbol{\beta}}\|^2}{2\sigma_{\boldsymbol{\beta}}^2} + \text{const.}$$

2.A.2 Optimal q^* densities

Expressions for $q^*(\beta_0, \boldsymbol{\beta})$, $\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})}$ and $\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}$

$$q^*(\beta_0, \boldsymbol{\beta}) \sim N(\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})}, \boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}),$$

$$\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})} = \left\{ \mu_{q(1/\sigma_\varepsilon^2)} \boldsymbol{C}^T \boldsymbol{C} + \text{blockdiag} \left[(\sigma_{\beta_0}^2)^{-1}, \mu_{q(1/\sigma_{\boldsymbol{\beta}}^2)} \boldsymbol{I}_p \right] \right\}^{-1}$$

and

$$\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})} = \mu_{q(1/\sigma_\varepsilon^2)} \boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})} \boldsymbol{C}^T \boldsymbol{y}$$

Derivation:

$$\begin{aligned} \log q^*(\beta_0, \boldsymbol{\beta}) &= E_q \left[\log p(\beta_0, \boldsymbol{\beta} | \text{rest}) \right] \\ &= -\frac{1}{2} \tilde{\boldsymbol{\beta}}^T \text{blockdiag} \left[(\sigma_{\beta_0}^2)^{-1}, \mu_{q(1/\sigma_{\boldsymbol{\beta}}^2)} \boldsymbol{I}_p \right] \tilde{\boldsymbol{\beta}} \\ &\quad -\frac{1}{2} \mu_{q(1/\sigma_\varepsilon^2)} \| \boldsymbol{y} - \boldsymbol{C} \tilde{\boldsymbol{\beta}} \|^2 + \text{const.} \\ &= \left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})} \right)^T \boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}^{-1} \left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})} \right) + \text{const.} \end{aligned}$$

Therefore,

$$q^*(\beta_0, \boldsymbol{\beta}) = \exp \left\{ \left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})} \right)^T \boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}^{-1} \left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})} \right) + \text{const} \right\}.$$

The results then follow from the definition 1.17 of the Multivariate Normal distribution.

Expressions for $q^*(\sigma_\varepsilon^2)$

$$q^*(\sigma_\varepsilon^2) \sim \text{ Inverse-Gamma}(A_{q(\sigma_\varepsilon^2)}, B_{q(\sigma_\varepsilon^2)}),$$

where

$$\begin{aligned} A_{q(\sigma_\varepsilon^2)} &= \frac{n}{2} + A_\varepsilon, \\ B_{q(\sigma_\varepsilon^2)} &= B_\varepsilon + \tfrac{1}{2}\left\{\| \boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})} \|^2 + \mathrm{tr}(\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})})\right\} \end{aligned}$$

and

$$\mu_{q(1/\sigma_\varepsilon^2)} = \frac{A_{q(\sigma_\varepsilon^2)}}{B_{q(\sigma_\varepsilon^2)}}.$$

Derivation:

$$\begin{aligned} \log q^*(\sigma_\varepsilon^2) &= E_q\left[\log p(\sigma_\varepsilon^2|\text{rest})\right] \\ &= (-1 - A_\varepsilon - \frac{n}{2})\log(\sigma_\varepsilon^2) - \frac{2B_\varepsilon + E_q[\|\boldsymbol{y} - \boldsymbol{C}\tilde{\boldsymbol{\beta}}\|^2]}{2\sigma_\varepsilon^2} + \text{const.} \\ &= (-1 - A_\varepsilon - \frac{n}{2})\log(\sigma_\varepsilon^2) \\ &\quad - \frac{2B_\varepsilon + \| \boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})} \|^2 + \mathrm{tr}(\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})})}{2\sigma_\varepsilon^2} + \text{const.} \end{aligned}$$

The results then follow from the definition (1.23) and results (1.7) for the Inverse-Gamma distribution.

Expressions for $q^*(\sigma_{\boldsymbol{\beta}}^2)$

$$q^*(\sigma_{\boldsymbol{\beta}}^2) \sim \text{ Inverse-Gamma}(A_{q(\sigma_{\boldsymbol{\beta}}^2)}, B_{q(\sigma_{\boldsymbol{\beta}}^2)}),$$

where

$$A_{q(\sigma_{\boldsymbol{\beta}}^2)} = \frac{p}{2} + A_{\boldsymbol{\beta}},\ B_{q(\sigma_{\boldsymbol{\beta}}^2)} = B_{\boldsymbol{\beta}} + \tfrac{1}{2}\left\{\| \boldsymbol{\mu}_{q(\boldsymbol{\beta})} \|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\right\}$$

and

$$\mu_{q(1/\sigma_{\boldsymbol{\beta}}^2)} = \frac{A_{q(\sigma_{\boldsymbol{\beta}}^2)}}{B_{q(\sigma_{\boldsymbol{\beta}}^2)}}.$$

Derivation:

$$\begin{aligned}
\log q^*(\sigma^2_{\boldsymbol{\beta}}) &= E_q\left[\log p(\sigma^2_{\boldsymbol{\beta}}|\text{rest})\right] \\
&= (-1 - A_{\boldsymbol{\beta}} - \frac{p}{2})\log(\sigma^2_{\boldsymbol{\beta}}) - \frac{2B_{\boldsymbol{\beta}} + E_q[\|\boldsymbol{\beta}\|^2]}{2\sigma^2_{\boldsymbol{\beta}}} + \text{const.} \\
&= (-1 - A_{\boldsymbol{\beta}} - \frac{1}{2})\log(\sigma^2_{\boldsymbol{\beta}}) - \frac{2B_{\boldsymbol{\beta}} + \| \boldsymbol{\mu}_{q(\boldsymbol{\beta})} \|^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})}{2\sigma^2_{\boldsymbol{\beta}}} + \text{const.}
\end{aligned}$$

The results then follow from the definition (1.23) and results (1.7) for the Inverse-Gamma distribution.

2.A.3 Derivation of lower bound

We note that

$$\begin{aligned}
\log \underline{p}(y; q) &= E_q\{\log\, p(\boldsymbol{y}, \beta_0, \boldsymbol{\beta}, \sigma^2_\varepsilon, \sigma^2_{\boldsymbol{\beta}}) - \log\, q(\beta_0, \boldsymbol{\beta}, \sigma^2_\varepsilon, \sigma^2_{\boldsymbol{\beta}})\} \\
&= E_q\{\log p(\boldsymbol{y}|\boldsymbol{\beta}_0, \boldsymbol{\beta}, \sigma^2_\varepsilon) + \log p(\sigma^2_\varepsilon) \\
&\quad + \log p(\beta_0) + \log p(\boldsymbol{\beta}|\sigma^2_{\boldsymbol{\beta}}) + \log p(\sigma^2_{\boldsymbol{\beta}}) \\
&\quad - \log q(\beta_0, \boldsymbol{\beta}) - \log q(\boldsymbol{\sigma}^2_\varepsilon) - \log q(\sigma^2_{\boldsymbol{\beta}})\}
\end{aligned}$$

Firstly,

$$\begin{aligned}
&E_q\{\log\, p(\boldsymbol{y}|\beta_0, \boldsymbol{\beta}, \sigma^2_\varepsilon)\} \\
&\quad = -\frac{n}{2}\log(2\boldsymbol{\pi}) - \frac{n}{2}\mu_{q(\log \sigma^2_\varepsilon)} \\
&\qquad -\frac{1}{2}\mu_{q(1/\sigma^2_\varepsilon)}\{\|\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})}\|^2 + \text{tr}(\boldsymbol{C}^T\boldsymbol{C})\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}\}
\end{aligned}$$

Secondly,

$$
\begin{aligned}
&E_q\{\log\, p(\sigma_\varepsilon^2) - \log\, q(\sigma_\varepsilon^2)\} \\
&\quad = A_\varepsilon \log(B_\varepsilon) - \log\Gamma(A_\varepsilon) \\
&\qquad -(A_\varepsilon + \frac{n}{2})\log(B_{q(\sigma_\varepsilon^2)}) + \log\Gamma(A_\varepsilon + \frac{n}{2}) \\
&\qquad +\frac{n}{2}\mu_{q(\log\sigma_\varepsilon^2)} + \frac{1}{2}\mu_{q(1/\sigma_\varepsilon^2)}\{\|\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})}\|^2 + \mathrm{tr}(\boldsymbol{C}^T\boldsymbol{C})\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}\}
\end{aligned}
$$

Thirdly,

$$
\begin{aligned}
&E_q\{\log p(\sigma_{\boldsymbol{\beta}}^2) - \log q(\sigma_{\boldsymbol{\beta}}^2)\} \\
&\quad = A_{\boldsymbol{\beta}}\log(B_{\boldsymbol{\beta}}) - \log\Gamma(A_{\boldsymbol{\beta}}) \\
&\qquad -A_{q(\sigma_{\boldsymbol{\beta}}^2)}\log(B_{q(\sigma_{\boldsymbol{\beta}}^2)}) + \log\Gamma(A_{q(\sigma_{\boldsymbol{\beta}}^2)}) \\
&\qquad +\frac{p}{2}\mu_{q(\log\sigma_{\boldsymbol{\beta}}^2)} + \frac{1}{2}\mu_{q(1/\sigma_{\boldsymbol{\beta}}^2)}\{\|\mu_{q(\boldsymbol{\beta})}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\}
\end{aligned}
$$

Fourthly,

$$
\begin{aligned}
&E_q\{\log p(\beta_0) + \log p(\boldsymbol{\beta}|\sigma_{\boldsymbol{\beta}}^2) - \log q(\beta_0, \boldsymbol{\beta})\} \\
&\quad = \frac{1+p}{2} - \frac{1}{2}\log(\sigma_{\beta_0}^2) - \frac{1}{2}\frac{\mu_{q(\beta_0)}^2 + \sigma_{q(\beta_0)}^2}{\sigma_{\beta_0}^2} \\
&\qquad -\frac{p}{2}\mu_{q(\log\sigma_{\boldsymbol{\beta}}^2)} - \frac{1}{2}\mu_{q(1/\sigma_{\boldsymbol{\beta}}^2)}\{\|\mu_{q(\boldsymbol{\beta})}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\} \\
&\qquad +\frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})}|
\end{aligned}
$$

Substitution of these gives the lower bound:

$$\begin{aligned}\log \underline{p}(y;q) &= -\frac{n}{2}\log(2\boldsymbol{\pi}) + \frac{1+p}{2} - \frac{1}{2}\log(\sigma^2_{\beta_0}) - \frac{\mu^2_{q(\beta_0)} + \sigma^2_{q(\beta_0)}}{2\sigma^2_{\beta_0}} \\ &\quad + A_\varepsilon \log(B_\varepsilon) - \log\Gamma(A_\varepsilon) \\ &\quad - (A_\varepsilon + \frac{n}{2})\log(B_{q(\sigma^2_\varepsilon)}) + \log\Gamma(A_\varepsilon + \frac{n}{2}) \\ &\quad + A_{\boldsymbol{\beta}}\log(B_{\boldsymbol{\beta}}) - \log\Gamma(A_{\boldsymbol{\beta}}) \\ &\quad - A_{q(\sigma^2_{\boldsymbol{\beta}})}\log(B_{q(\sigma^2_{\boldsymbol{\beta}})}) + \log\Gamma(A_{q(\sigma^2_{\boldsymbol{\beta}})}) \\ &\quad + \frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})}|\end{aligned}$$

2.B Appendix: Derivation of Algorithm 2.3.1

In Algorithm 2.3.1, the MFVB calculations for β_0, $\boldsymbol{\beta}$ and σ^2_ε are similar to the Algorithm 2.2.1. We obtain them by using

$$(\sigma^2_{\boldsymbol{\beta}})^{-1} \text{ to replace } \mu_{q(1/\sigma^2_{\boldsymbol{\beta}})},$$

$$\tfrac{1}{2} \text{ to replace } A_\varepsilon,$$

$$\text{and } \mu_{q(1/a)} \text{ to replace } B_\varepsilon.$$

Therefore, I only show the derivation for a.

2.B.1 Full conditionals

Full conditional for a

$$\log p(a|\text{rest}) = -2\log(a) - \frac{1}{a}\left\{\frac{1}{\sigma^2_\varepsilon} + \frac{1}{A^2}\right\} + \text{const.}$$

Derivation:

$$\begin{aligned} p(a|\text{rest}) &\propto p(\sigma_\varepsilon^2|a)p(a) \\ &= \frac{a^{-1/2}}{\Gamma(1/2)}(\sigma_\varepsilon^2)^{-3/2}\exp\left\{-\frac{1}{a\sigma_\varepsilon^2}\right\} \\ &\quad \times\frac{A^{-1}}{\Gamma(1/2)}a^{-3/2}\exp\left\{-\frac{1}{A^2a}\right\}. \end{aligned}$$

Taking logarithms, we get:

$$\log p(a|\text{rest}) = -2\log(a) - \frac{1}{a}\left\{\frac{1}{\sigma_\varepsilon^2} + \frac{1}{A^2}\right\} + \text{const.}$$

2.B.2 Optimal q^* densities

Expressions for $q^*(a)$

$$q^*(a) \sim \text{Inverse-Gamma}(1, B_{q(a)}),$$

where

$$B_{q(a)} = \frac{1}{A^2} + \mu_{q(1/\sigma_\varepsilon^2)} \text{ and } \mu_{q(1/a)} = \frac{1}{B_{q(a)}}.$$

Derivation:

$$\begin{aligned} \log q^*(a) &= E_q\left[\log p(a|\text{rest})\right] \\ &= -2\log(a) - \frac{1}{a}\left\{E_q\left[\frac{1}{\sigma_\varepsilon^2}\right] + \frac{1}{A^2}\right\} + \text{const.} \\ &= -2\log(a) - \frac{1}{a}\left\{\mu_{q(1/\sigma_\varepsilon^2)} + \frac{1}{A^2}\right\} + \text{const.} \end{aligned}$$

The results then follow from the definition (1.23) and results (1.7) for the Inverse-Gamma distribution.

2.B.3 Derivation of lower bound

We note that

$$\begin{aligned}\log \underline{p}(y;q) &= E_q\{\log p(\boldsymbol{y},\beta_0,\boldsymbol{\beta},\sigma_\varepsilon^2,a) - \log q(\beta_0,\boldsymbol{\beta},\sigma_\varepsilon^2,a)\} \\ &= E_q\{\log p(\boldsymbol{y}|\boldsymbol{\beta}_0,\boldsymbol{\beta},\sigma_\varepsilon^2) + \log p(\sigma_\varepsilon^2|a) + \log p(a) \\ &\quad +\log p(\beta_0) + \log p(\boldsymbol{\beta}) \\ &\quad -\log q(\beta_0,\boldsymbol{\beta}) - \log q(\boldsymbol{\sigma}_\varepsilon^2) - \log q(a)\}\end{aligned}$$

Firstly,

$$\begin{aligned}&E_q\{\log p(\boldsymbol{y}|\beta_0,\boldsymbol{\beta},\sigma_\varepsilon^2)\} \\ &\quad = -\frac{n}{2}\log(2\boldsymbol{\pi}) - \frac{n}{2}\mu_{(\log\sigma_\varepsilon^2)} \\ &\qquad -\frac{1}{2}\mu_{q(1/\sigma_\varepsilon^2)}\{\|\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})}\|^2 + \mathrm{tr}(\boldsymbol{C}^T\boldsymbol{C})\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}\}\end{aligned}$$

Secondly,

$$\begin{aligned}&E_q\{\log p(\sigma_\varepsilon^2|a) - \log q(\sigma_\varepsilon^2)\} \\ &\quad = -\frac{1}{2}\mu_{q(\log a)} - \frac{1}{2}\log(\pi) \\ &\qquad -A_{q(\sigma_\varepsilon^2)}\log(B_{q(\sigma_\varepsilon^2)}) + \log\Gamma(A_{q(\sigma_\varepsilon^2)}) \\ &\qquad +\frac{n}{2}\mu_{q(\log\sigma_\varepsilon^2)} + \frac{1}{2}\mu_{q(1/\sigma_\varepsilon^2)}\{\|\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})}\|^2 + \mathrm{tr}(\boldsymbol{C}^T\boldsymbol{C})\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}\}\end{aligned}$$

Thirdly,

$$\begin{aligned}&E_q\{\log p(a) - \log q(a)\} \\ &\quad = -\log A - \frac{1}{2}\log(\pi) - \log(B_{q(a)}) + \frac{1}{2}\mu_{q(\log a)} + \mu_{q(1/a)}\mu_{q(1/\sigma_\varepsilon^2)}\end{aligned}$$

Fourthly,

$$\begin{aligned}
&E_q\{\log p(\beta_0) + \log p(\boldsymbol{\beta}) - \log q(\beta_0, \boldsymbol{\beta})\} \\
&\quad = \frac{1+p}{2} - \frac{1}{2}\log(\sigma^2_{\beta_0}) - \frac{\mu^2_{q(\beta_0)} + \sigma^2_{q(\beta_0)}}{2\sigma^2_{\beta_0}} \\
&\qquad -\frac{p}{2}\log(\sigma^2_{\beta}) - \frac{1}{2\sigma^2_{\boldsymbol{\beta}}}\{\|\mu_{q(\boldsymbol{\beta})}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\} \\
&\qquad +\frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})}|
\end{aligned}$$

Substitution of these gives the lower bound:

$$\begin{aligned}
\log \underline{p}(y;q) \quad = \quad & \frac{1+p}{2} - \frac{n}{2}\log(2\boldsymbol{\pi}) - \frac{1}{2}\log(\sigma^2_{\beta_0}) - \frac{\mu^2_{q(\beta_0)} + \sigma^2_{q(\beta_0)}}{2\sigma^2_{\beta_0}} \\
& -\frac{p}{2}\log(\sigma^2_{\boldsymbol{\beta}}) - \frac{1}{2\sigma^2_{\boldsymbol{\beta}}}\{\|\mu_{q(\boldsymbol{\beta})}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\} \\
& +\frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})}| \\
& -\log(\pi) - \log A - \log(B_{q(a)}) + \mu_{q(1/a)}\mu_{1(1/\sigma^2_\varepsilon)} \\
& -A_{q(\sigma^2_\varepsilon)}\log(B_{q(\sigma^2_\varepsilon)}) + \log\Gamma(A_{q(\sigma^2_\varepsilon)}
\end{aligned}$$

2.C Appendix: Derivation of Algorithm 2.3.2

In Algorithm 2.3.2, the MFVB calculations for β_0 and $\boldsymbol{\beta}$ are the same as those of Algorithm 2.3.1. Therefore, I only show the derivation for σ^2_ε, a_ε and b_ε.

2.C.1 Full conditionals

Full conditional for σ^2_ε

$$\log p(\sigma^2_\varepsilon|\text{rest}) = -\left[\log\sigma^2_\varepsilon\right]^2\left(\frac{1}{2b_\varepsilon}\right) + \left(-1 - \frac{n}{2} + \frac{a_\varepsilon}{b_\varepsilon}\right)\log(\sigma^2_\varepsilon) - \frac{\|\boldsymbol{y} - \boldsymbol{C}\tilde{\boldsymbol{\beta}}\|^2}{2\sigma^2_\varepsilon} + \text{const.}$$

where

$$\tilde{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \boldsymbol{\beta} \end{bmatrix}, \quad \boldsymbol{C} = [\mathbf{1}, \boldsymbol{X}] .$$

Derivation:

$$\begin{aligned} p(\sigma_\varepsilon^2|\text{rest}) &\propto p(\boldsymbol{y}|\beta_0, \boldsymbol{\beta}, \sigma_\varepsilon^2) p(\sigma_\varepsilon^2) \\ &= (2\pi)^{-n/2} \sigma_\varepsilon^{-n} \exp\left\{ -\frac{\|\boldsymbol{y} - \mathbf{1}\beta_0 - \boldsymbol{X}\boldsymbol{\beta}\|^2}{2\sigma_\varepsilon^2} \right\} \\ &\quad \times \frac{1}{\sigma_\varepsilon^2 b_\varepsilon^{1/2} \sqrt{2\pi}} \exp\left\{ -\frac{\left(\log\sigma_\varepsilon^2 - a_\varepsilon\right)^2}{2b_\varepsilon} \right\} . \end{aligned}$$

Taking logarithms, we get:

$$\begin{aligned} \log p(\sigma_\varepsilon^2|\text{rest}) &= (-1 - \frac{n}{2}) \log(\sigma_\varepsilon^2) - \frac{\|\boldsymbol{y} - \boldsymbol{C}\tilde{\boldsymbol{\beta}}\|^2}{2\sigma_\varepsilon^2} - \frac{\left(\log(\sigma_\varepsilon^2) - a_\varepsilon\right)^2}{2b_\varepsilon} + \text{const.} \\ &= -\left[\log(\sigma_\varepsilon^2)\right]^2 \left(\frac{1}{2b_\varepsilon}\right) + \left(-1 - \frac{n}{2} + \frac{a_\varepsilon}{b_\varepsilon}\right) \log(\sigma_\varepsilon^2) - \frac{\|\boldsymbol{y} - \boldsymbol{C}\tilde{\boldsymbol{\beta}}\|^2}{2\sigma_\varepsilon^2} + \text{const.} \end{aligned}$$

Full conditional for a_ε

$$\log p(a_\varepsilon|\text{rest}) = -a_\varepsilon^2 \left(\frac{1}{2b_\varepsilon} + \frac{1}{2\sigma_a^2}\right) + a_\varepsilon \frac{\log(\sigma_\varepsilon^2)}{b_\varepsilon} + \text{const.}$$

Derivation:

$$\begin{aligned} p(a_\varepsilon|\text{rest}) &\propto p(\sigma_\varepsilon^2|a_\varepsilon, b_\varepsilon) p(a_\varepsilon) \\ &= \frac{1}{\sigma_\varepsilon^2 b_\varepsilon^{1/2} \sqrt{2\pi}} \exp\left\{ -\frac{\left(\log(\sigma_\varepsilon^2) - a_\varepsilon\right)^2}{2b_\varepsilon} \right\} \\ &\quad \times \frac{1}{\sqrt{2\pi}\sigma_a} \exp\left\{ -\frac{a_\varepsilon^2}{2\sigma_a^2} \right\} . \end{aligned}$$

Taking logarithms, we get:

$$\begin{aligned}\log p(a_\varepsilon|\text{rest}) &= -\frac{\left(\log(\sigma_\varepsilon^2)+a_\varepsilon\right)^2}{2b_\varepsilon} - \frac{a_\varepsilon^2}{2\sigma_a^2} + \text{const.} \\ &= -a_\varepsilon^2\left(\frac{1}{2b_\varepsilon}+\frac{1}{2\sigma_a^2}\right) + a_\varepsilon\frac{\log(\sigma_\varepsilon^2)}{b_\varepsilon} + \text{const.}\end{aligned}$$

Full conditional for b_ε

$$\log p(b_\varepsilon|\text{rest}) = -\frac{1}{2b_\varepsilon}\left\{2B_b + \left(\log(\sigma_\varepsilon^2)-a_\varepsilon\right)^2\right\} - (\frac{3}{2}+A_b)\log(b_\varepsilon) + \text{const.}$$

Derivation:

$$\begin{aligned}p(b_\varepsilon|\text{rest}) &\propto p(\sigma_\varepsilon^2|a_\varepsilon,b_\varepsilon)p(b_\varepsilon) \\ &= \frac{1}{\sigma_\varepsilon^2 b_\varepsilon^{1/2}\sqrt{2\pi}}\exp\left\{-\frac{\left(\log(\sigma_\varepsilon^2)-a_\varepsilon\right)^2}{2b_\varepsilon}\right\} \\ &\quad \times\frac{B_b^{A_b}}{\Gamma(A_b)}(b_\varepsilon)^{-1-A_b}\exp\left\{-\frac{B_b}{b_\varepsilon}\right\}.\end{aligned}$$

Taking logarithms, we get:

$$\begin{aligned}\log p(b_\varepsilon|\text{rest}) &= -\frac{\left(\log(\sigma_\varepsilon^2)-a_\varepsilon\right)^2}{2b_\varepsilon} - \frac{B_b}{b_\varepsilon} - (\frac{3}{2}+A_b)\log(b_\varepsilon) + \text{const.} \\ &= -\frac{1}{2b_\varepsilon}\left\{2B_b + \left(\log(\sigma_\varepsilon^2)-a_\varepsilon\right)^2\right\} - (\frac{3}{2}+A_b)\log(b_\varepsilon) + \text{const.}\end{aligned}$$

2.C.2 Optimal q^* densities

Expressions for b_ε

$$q^*(b_\varepsilon) \sim \text{ Inverse-Gamma}(A_{q(b_\varepsilon)}, B_{q(b_\varepsilon)}),$$

where

$$A_{q(b_\varepsilon)} = A_b + \frac{1}{2},$$

$$B_{q(b_\varepsilon)} = B_b + \frac{1}{2}\{\mu^2_{q(a_\varepsilon)} + \sigma^2_{q(a_\varepsilon)} - 2\mu_{q(a_\varepsilon)}\mu_{q(\log\sigma^2_\varepsilon)} + \mu_{q([\log(\sigma^2_\varepsilon)]^2)}\}$$

and

$$\mu_{q(1/b_\varepsilon)} = \frac{A_{q(b_\varepsilon)}}{B_{q(b_\varepsilon)}}$$

Derivation:

$$\begin{aligned}
\log q^*(b_\varepsilon) &= E_q\left[\log p(b_\varepsilon|\text{rest})\right] \\
&= -\frac{1}{2b_\varepsilon}\left\{2B_b + E_q\left[\left(\log(\sigma^2_\varepsilon) - a_\varepsilon\right)^2\right]\right\} - (\frac{3}{2} + A_b)\log(b_\varepsilon) + \text{const.} \\
&= -\frac{1}{2b_\varepsilon}\left\{2B_b + \mu^2_{q(a_\varepsilon)} + \sigma^2_{q(a_\varepsilon)} - 2\mu_{q(a_\varepsilon)}\mu_{q(\log\sigma^2_\varepsilon)} + \mu_{q([\log(\sigma^2_\varepsilon)]^2)}\right\} \\
&\quad -(\frac{3}{2} + A_b)\log(b_\varepsilon) + \text{const.}
\end{aligned}$$

The results then follow from the definition (1.23) and results (1.7) for the Inverse-Gamma distribution.

Expressions for a_ε

$$q^*(a_\varepsilon) \sim N(\mu_{q(a_\varepsilon)}, \sigma^2_{q(a_\varepsilon)}),$$

where

$$\sigma^2_{q(a_\varepsilon)} = \left[\frac{1}{\sigma^2_a} + \mu_{q(1/b_\varepsilon)}\right]^{-1} \text{ and } \mu_{q(a_\varepsilon)} = \sigma^2_{q(a_\varepsilon)}\mu_{q(1/b_\varepsilon)}\mu_{q(\log\sigma^2_\varepsilon)}.$$

Derivation:

$$\begin{aligned}
\log q^*(a_\varepsilon) &= E_q\left[\log p(a_\varepsilon|\text{rest})\right] \\
&= -a^2_\varepsilon E_q\left[\frac{1}{2b_\varepsilon} + \frac{1}{2\sigma^2_a}\right] + a_\varepsilon E_q\left[\frac{\log(\sigma^2_\varepsilon)}{b_\varepsilon}\right] + \text{const.} \\
&= -\frac{a^2_\varepsilon}{2}\left[\frac{1}{\sigma^2_a} + \mu_{q(1/b_\varepsilon)}\right] + a_\varepsilon\left[\mu_{q(1/b_\varepsilon)}\mu_{q(\log\sigma^2_\varepsilon)}\right].
\end{aligned}$$

The results then from follow the definition 1.14 of the Normal distribution.

Expressions for σ_ε^2

$$q^*(\sigma_\varepsilon^2) = \frac{(\sigma_\varepsilon^2)^{C_{l1}^*-1}\exp\{-\frac{C_{l2}^*}{\sigma_\varepsilon^2} - C_{l3}^*[\log(\sigma_\varepsilon^2)]^2\}}{\mathcal{J}(0, C_{l1}^*, C_{l3}^*, C_{l2}^*)},$$

$$\mu_{q(1/\sigma_\varepsilon^2)} = \frac{\mathcal{J}(0, C_{l1}^* - 1, C_{l3}^*, C_{l2}^*)}{\mathcal{J}(0, C_{l1}^*, C_{l3}^*, C_{l2}^*)},$$

$$\mu_{q(\log\sigma_\varepsilon^2)} = \frac{\mathcal{J}(1, C_{l1}^*, C_{l3}^*, C_{l2}^*)}{\mathcal{J}(0, C_{l1}^*, C_{l3}^*, C_{l2}^*)},$$

and

$$\mu_{q([\log(\sigma_\varepsilon^2)]^2)} = \frac{\mathcal{J}(2, C_{l1}^*, C_{l3}^*, C_{l2}^*)}{\mathcal{J}(0, C_{l1}^*, C_{l3}^*, C_{l2}^*)}$$

where

$$\begin{aligned}
C_{l1}^* &= -\frac{n}{2} + \mu_{q(a_\varepsilon)}\mu_{q(1/b_\varepsilon)}, \\
C_{l2}^* &= \frac{1}{2}[\|\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})}\|^2 + \mathrm{tr}(\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})})], \\
C_{l3}^* &= \frac{1}{2}\mu_{q(1/b_\varepsilon)}.
\end{aligned}$$

Derivation:

$$\begin{aligned}
\log q^*(\sigma_\varepsilon^2) &= E_q\left[\log p(\sigma_\varepsilon^2|\text{rest})\right] \\
&= -\left[\log(\sigma_\varepsilon^2)\right]^2 E_q\left[\frac{1}{2b_\varepsilon}\right] + \left(1 - \frac{n}{2} + E_q\left[\frac{a_\varepsilon}{b_\varepsilon}\right]\right)\log\sigma_\varepsilon^2 \\
&\quad -\frac{E_q\left[\|\boldsymbol{y} - \boldsymbol{C}\tilde{\boldsymbol{\beta}}\|^2\right]}{2\sigma_\varepsilon^2} + \text{const.} \\
&= -C_{l3}^*\left[\log(\sigma_\varepsilon^2)\right]^2 + (C_{l2}^* - 1)\log(\sigma_\varepsilon^2) - \frac{C_{l2}^*}{\sigma_\varepsilon^2} + \text{const.}
\end{aligned}$$

Taking exponents, we get:

$$\begin{aligned} q^*(\sigma_\varepsilon^2) &= \frac{(\sigma_\varepsilon^2)^{C_{l1}^*-1}\exp\{-\frac{C_{l2}^*}{\sigma_\varepsilon^2} - C_{l3}^*[\log(\sigma_\varepsilon^2)]^2\}}{\int (\sigma_\varepsilon^2)^{C_{l1}^*-1}\exp\{-\frac{C_{l2}^*}{\sigma_\varepsilon^2} - C_{l3}^*[\log(\sigma_\varepsilon^2)]^2\}\, d\sigma_\varepsilon^2} \\ &= \frac{(\sigma_\varepsilon^2)^{C_{l1}^*-1}\exp\{-\frac{C_{l2}^*}{\sigma_\varepsilon^2} - C_{l3}^*[\log(\sigma_\varepsilon^2)]^2\}}{\mathcal{J}(0, C_{l1}^*, C_{l3}^*, C_{l2}^*)}. \end{aligned}$$

Next, we get:

$$\begin{aligned} \mu_{q(1/\sigma_\varepsilon^2)} &= \int \frac{1}{\sigma_\varepsilon^2} q^*(\sigma_\varepsilon^2)\, d\sigma_\varepsilon^2 \\ &= \frac{\int (\sigma_\varepsilon^2)^{C_{l1}^*-2}\exp\{-\frac{C_{l2}^*}{\sigma_\varepsilon^2} - C_{l3}^*[\log(\sigma_\varepsilon^2)]^2\}\, d\sigma_\varepsilon^2}{\mathcal{J}(0, C_{l1}^*, C_{l3}^*, C_{l2}^*)} \\ &= \frac{\mathcal{J}(0, C_{l1}^* - 1, C_{l3}^*, C_{l2}^*)}{\mathcal{J}(0, C_{l1}^*, C_{l3}^*, C_{l2}^*)}. \end{aligned}$$

Similarly, we can get $\mu_{q(\log\sigma_\varepsilon^2)}$ and $\mu_{q([\log(\sigma_\varepsilon^2)]^2)}$.

2.C.3 Derivation of lower bound

We note that

$$\begin{aligned} \log \underline{p}(y; q) &= E_q\{\log p(\boldsymbol{y}, \beta_0, \boldsymbol{\beta}, \sigma_\varepsilon^2, a_\varepsilon, b_\varepsilon) - \log q(\beta_0, \boldsymbol{\beta}, \sigma_\varepsilon^2, a_\varepsilon, b_\varepsilon)\} \\ &= E_q\{\log p(\boldsymbol{y}|\boldsymbol{\beta}_0, \boldsymbol{\beta}, \sigma_\varepsilon^2) + \log p(\sigma_\varepsilon^2|a_\varepsilon, b_\varepsilon) \\ &\quad + \log p(a_\varepsilon) + \log p(b_\varepsilon) + \log p(\beta_0) + \log p(\boldsymbol{\beta}) \\ &\quad - \log q(\beta_0, \boldsymbol{\beta}) - \log q(\sigma_\varepsilon^2) - \log q(a_\varepsilon) - \log q(b_\varepsilon)\} \end{aligned}$$

Firstly,

$$\begin{aligned}
&E_q\{\log p(\boldsymbol{y}|\beta_0, \boldsymbol{\beta}, \sigma_\varepsilon^2)\} \\
&\quad = -\frac{n}{2}\log(2\boldsymbol{\pi}) - \frac{n}{2}\mu_{q(\log\sigma_\varepsilon^2)} \\
&\qquad -\frac{1}{2}\mu_{q(1/\sigma_\varepsilon^2)}\{\|\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})}\|^2 + \mathrm{tr}(\boldsymbol{C}^T\boldsymbol{C})\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}\}
\end{aligned}$$

Secondly,

$$\begin{aligned}
&E_q\{\log p(a_\varepsilon) - \log q(a_\varepsilon)\} \\
&\quad = \frac{1}{2} - \frac{1}{2}\log(\sigma_a^2) - \frac{1}{2}\frac{\sigma_{q(a_\varepsilon)}^2 + \mu_{q(a_\varepsilon)}^2}{\sigma_a^2} + \frac{1}{2}\log(\sigma_{q(a_\varepsilon)}^2)
\end{aligned}$$

Thirdly,

$$\begin{aligned}
&E_q\{\log p(b_\varepsilon) - \log q(b_\varepsilon)\} \\
&\quad = A_b\log(B_b) - \log\Gamma(A_b) \\
&\qquad -A_{q(b_\varepsilon)}\log(B_{q(b_\varepsilon)}) + \log\,\Gamma(A_{q(b_\varepsilon)}) + \frac{1}{2}\mu_{q(\log(b_\varepsilon))} \\
&\qquad +\frac{1}{2}\mu_{q(1/b_\varepsilon)}\{\mu_{q(a_\varepsilon)}^2 + \sigma_{q(a_\varepsilon)}^2 - 2\mu_{q(a_\varepsilon)}^2\mu_{q(\log\sigma_\varepsilon^2)} + \mu_{q([\log(\sigma_\varepsilon^2)]^2)}\}
\end{aligned}$$

Fourthly,

$$
\begin{aligned}
&E_q\{\log p(\sigma_\varepsilon^2) - \log q(\sigma_\varepsilon^2)\} \\
&\quad = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\mu_{q(\log b_\varepsilon)} \\
&\quad -\frac{1}{2}\mu_{q(1/b_\varepsilon)}\{\mu^2_{q(a_\varepsilon)} + \sigma^2_{q(a_\varepsilon)} - 2\mu^2_{q(a_\varepsilon)}\mu_{q(\log \sigma^2_\varepsilon)} + \mu_{q([\log(\sigma^2_\varepsilon)]^2)}\} \\
&\quad\ +\log \mathcal{J}(0, C^*_{l1}, C^*_{l3}, C^*_{l2}) - [\mu_{q(a_\varepsilon)}\mu_{q(1/b_\varepsilon)} - \frac{n}{2}]\mu_{(\log\, \sigma^2_\varepsilon)} \\
&\quad\ +\frac{1}{2}\mu_{q(1/b_\varepsilon)}\mu_{q([\log\, \sigma^2_\varepsilon]^2)} \\
&\quad\ +\frac{1}{2}\mu_{(1/\sigma^2_\varepsilon)}\{\|\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})}\|^2 + \mathrm{tr}(\boldsymbol{C}^T\boldsymbol{C})\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}\}
\end{aligned}
$$

Fifthly,

$$
\begin{aligned}
&E_q\{\log p(\beta_0) + \log p(\boldsymbol{\beta}) - \log q(\beta_0, \boldsymbol{\beta})\} \\
&\quad = \frac{1+p}{2} - \frac{1}{2}\log(\sigma^2_{\beta_0}) - \frac{1}{2}\frac{\mu^2_{q(\beta_0)} + \sigma^2_{q(\beta_0)}}{\sigma^2_{\beta_0}} \\
&\quad\ -\frac{p}{2}\log(\sigma^2_{\boldsymbol{\beta}}) - \frac{1}{2\sigma^2_{\boldsymbol{\beta}}}\{\|\mu_{q(\boldsymbol{\beta})}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\} \\
&\quad\ +\frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})}|
\end{aligned}
$$

Substitution of these gives the lower bound:

$$\begin{aligned}
\log \underline{p}(y;q) \;=\; & \frac{2+p}{2} - \frac{n+1}{2}\log(2\boldsymbol{\pi}) - \frac{1}{2}\log(\sigma^2_{\beta_0}) - \frac{\mu^2_{q(\beta_0)} + \sigma^2_{q(\beta_0)}}{2\sigma^2_{\beta_0}} \\
& -\frac{p}{2}\log(\sigma^2_{\boldsymbol{\beta}}) - \frac{1}{2\sigma^2_{\boldsymbol{\beta}}}\{\|\mu_{q(\boldsymbol{\beta})}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\} \\
& +\frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}| \\
& -\frac{1}{2}\log(\sigma^2_a) - \frac{1}{2}\frac{\sigma^2_{q(a_\varepsilon)} + \mu^2_{q(a_\varepsilon)}}{\sigma^2_a} + \frac{1}{2}\log(\sigma^2_{q(a_\varepsilon)}) \\
& +A_b\log(B_b) - \log\,\Gamma(A_b) \\
& -A_{q(b_\varepsilon)}\log(B_{q(b_\varepsilon)}) + \log\,\Gamma(A_{q(b_\varepsilon)}) \\
& +\log\mathcal{J}(0, C^*_{l1}, C^*_{l3}, C^*_{l2}) - \mu_{q(a_\varepsilon)}\mu_{q(1/b_\varepsilon)}\mu_{(\log\sigma^2_\varepsilon)} \\
& +\frac{1}{2}\mu_{q(1/b_\varepsilon)}\mu_{q([\log(\sigma^2_\varepsilon)]^2)}
\end{aligned}$$

2.D Appendix: MFVB Algorithm of model (2)

The mixed model with the ridge penalized method for the fixed effects and a Normal prior for the random variance is given by:

$$\begin{aligned}
& \boldsymbol{y}|\beta_0, \boldsymbol{\beta}, \boldsymbol{u}, \sigma^2_\varepsilon \overset{\text{ind.}}{\sim} N(\mathbf{1}\beta_0 + X\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \sigma^2_\varepsilon I_n) \\
& \beta_0 \sim N(0, \sigma^2_{\beta_0}) \\
& \boldsymbol{\beta}|\sigma^2_{\boldsymbol{\beta}} \sim N(\mathbf{0}, \sigma^2_{\boldsymbol{\beta}}\boldsymbol{I}) \\
& \sigma^2_{\boldsymbol{\beta}} \sim \text{ Inverse-Gamma}(A_{\boldsymbol{\beta}}, B_{\boldsymbol{\beta}}) \\
& \boldsymbol{u}_\ell|\sigma_{\boldsymbol{u}_\ell} \sim N(0, \sigma^2_{\boldsymbol{u}_\ell} I_{K_\ell}) \;\; 1 \le \ell \le r \\
& \sigma^2_{\boldsymbol{u}_\ell} \sim \text{ Inverse-Gamma}(A_{\boldsymbol{u}}, B_{\boldsymbol{u}}) \;\; 1 \le \ell \le r \\
& \sigma^2_\varepsilon \sim \text{ Inverse-Gamma}(A_\varepsilon, B_\varepsilon)
\end{aligned} \tag{2.17}$$

where A_ε, B_ε, $\sigma^2_{\beta_0}$, $A_{\boldsymbol{u}}$, $B_{\boldsymbol{u}}$,$A_{\boldsymbol{\beta}}$, $B_{\boldsymbol{\beta}} > 0$ are hyperparameters. The product restriction that we impose here is:

$$q(\beta_0, \boldsymbol{\beta}, \boldsymbol{u}, \sigma^2_\varepsilon, \sigma^2_{\boldsymbol{\beta}}, \sigma^2_{\boldsymbol{u}_1}, ..., \sigma^2_{\boldsymbol{u}_r}) = q(\beta_0, \boldsymbol{\beta}, \boldsymbol{u})q(\sigma^2_\varepsilon)q(\sigma^2_{\boldsymbol{\beta}})q(\sigma^2_{\boldsymbol{u}_1}, ..., \sigma^2_{\boldsymbol{u}_r}).$$

Then, the optimal q^* densities for the parameters in model (2.17) take the form:

$$\begin{aligned} &q^*(\beta_0, \boldsymbol{\beta}, \boldsymbol{u}) \text{ is a Multivariate Normal density function,} \\ &q^*(\sigma^2_\varepsilon) \text{ is an Inverse Gamma density function,} \\ &q^*(\sigma^2_{\boldsymbol{\beta}}) \text{ is an Inverse Gamma density function,} \\ &q^*(\sigma^2_{\boldsymbol{u}_\ell}) \text{ is an Inverse Gamma density function,} \end{aligned} \tag{2.18}$$

Let $\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})}$ and $\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})}$ denote the mean vector and covariance matrix for the Multivariate Normal density function $q^*(\beta_0, \boldsymbol{\beta}, \boldsymbol{u})$, and $A_{q(\sigma^2_{\boldsymbol{\beta}})}$ and $B_{q(\sigma^2_{\boldsymbol{\beta}})}$ denote the shape and rate parameters for $q^*(\sigma^2_{\boldsymbol{\beta}})$. A similar definition is used for the parameters in $q^*(\sigma^2_{\boldsymbol{u}_\ell})$ and $q^*(\sigma^2_\varepsilon)$. Let $\boldsymbol{C} = [1, \boldsymbol{X}, \boldsymbol{Z}]$. The convergence of Algorithm 2.D.1 can be monitored using the following expression for the lower bound on the

marginal log-likelihood:

$$\begin{aligned}
\log \underline{p}(\boldsymbol{y}; q) &= -\frac{n}{2}\log(2\boldsymbol{\pi}) + \frac{1+p+K_\ell}{2} - \frac{1}{2}\log(\sigma^2_{\beta_0}) - \frac{\mu^2_{q(\beta_0)} + \sigma^2_{q(\beta_0)}}{2\sigma^2_{\beta_0}} \\
&\quad + A_\varepsilon \log(B_\varepsilon) - \log\,\Gamma(A_\varepsilon) \\
&\quad - (A_\varepsilon + \frac{n}{2})\log(B_{q(\sigma^2_\varepsilon)}) + \log\,\Gamma(A_\varepsilon + \frac{n}{2}) \\
&\quad + A_{\boldsymbol{\beta}} \log(B_{\boldsymbol{\beta}}) - \log\,\Gamma(A_{\boldsymbol{\beta}}) \\
&\quad - A_{q(\sigma^2_{\boldsymbol{\beta}})}\log(B_{q(\sigma^2_{\boldsymbol{\beta}})}) + \log\,\Gamma(A_{q(\sigma^2_{\boldsymbol{\beta}})}) \\
&\quad + r(A_{\boldsymbol{u}}\log(B_{\boldsymbol{u}}) - \log\,\Gamma(A_{\boldsymbol{u}})) \\
&\quad - \sum_{\ell=1}^{r}[A_{q(\sigma^2_{\boldsymbol{u}_\ell})}\log(B_{q(\sigma^2_{\boldsymbol{u}_\ell})}) - \log\,\Gamma(A_{q(\sigma^2_{\boldsymbol{u}_\ell})})] \\
&\quad + \frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})}|
\end{aligned}$$

The derivation of MFVB Algorithms is similar to the linear case. Therefore, I only show the derivation of the lower bound.

Derivation of the lower bound

We note that

$$\begin{aligned}
\log \underline{p}(y; q) &= E_q\{\log p(\boldsymbol{y}, \beta_0, \boldsymbol{\beta}, \boldsymbol{u}, \sigma^2_\varepsilon, \sigma^2_{\boldsymbol{\beta}}, \sigma^2_{\boldsymbol{u}_1}, ..., \sigma^2_{\boldsymbol{u}_r}) \\
&\quad - \log q(\beta_0, \boldsymbol{\beta}, \boldsymbol{u}, \sigma^2_\varepsilon, \sigma^2_{\boldsymbol{\beta}}, \sigma^2_{\boldsymbol{u}_1}, ..., \sigma^2_{\boldsymbol{u}_r})\} \\
&= E_q\{\log p(\boldsymbol{y}|\boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{u}, \sigma^2_\varepsilon) + \log p(\sigma^2_\varepsilon) + \log p(\beta_0) \\
&\quad + \log p(\boldsymbol{\beta}|\sigma^2_{\boldsymbol{\beta}}) + \log p(\sigma^2_{\boldsymbol{\beta}}) + \sum_{\ell=1}^{r} \log p(\boldsymbol{u}_\ell|\sigma^2_{\boldsymbol{u}_\ell}) + \sum_{\ell=1}^{r} \log p(\sigma^2_{\boldsymbol{u}_\ell}) \\
&\quad - \log q(\beta_0, \boldsymbol{\beta}, \boldsymbol{u}) - \log q(\boldsymbol{\sigma}^2_\varepsilon) - \log q(\sigma^2_{\boldsymbol{\beta}}) - \sum_{\ell=1}^{r} \log q(\sigma^2_{\boldsymbol{u}_\ell})\}
\end{aligned}$$

Algorithm 2.D.1: MFVB iterative scheme for obtaining the parameters of the optimal densities $q^*(\beta_0, \boldsymbol{\beta}, \boldsymbol{u})$, $q^*(\sigma^2_{\boldsymbol{\beta}})$, $q^*(\sigma^2_{\boldsymbol{u}_\ell})$ and $q^*(\sigma^2_\varepsilon)$ for the mixed model with the ridge penalized method for the fixed effects and Normal prior for the random variance.

Initialize $\mu_{q(1/\sigma^2_{\boldsymbol{\beta}})}$, $\mu_{q(1/\sigma^2_{\boldsymbol{u}_1})}$,...,$\mu_{q(1/\sigma^2_{\boldsymbol{u}_r})}$;
Cycle

$$\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \Big\{\mu_{q(1/\sigma^2_\varepsilon)}\boldsymbol{C}^T\boldsymbol{C} + blockdiag\big[(\sigma^2_{\beta_0})^{-1}, \mu_{q(1/\sigma^2_{\boldsymbol{\beta}})}\boldsymbol{I}_p, \mu_{q(1/\sigma^2_{\boldsymbol{u}_1})}\boldsymbol{I}_{K_1}, ..., \mu_{q(1/\sigma^2_{\boldsymbol{u}_1})}\boldsymbol{I}_{K_r}\big]\Big\}^{-1}$$

$$\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T\boldsymbol{y}\mu_{q(1/\sigma^2_\varepsilon)}$$

$$A_{q(\sigma^2_\varepsilon)} \leftarrow \frac{n}{2} + A_\varepsilon$$

$$B_{q(\sigma^2_\varepsilon)} \leftarrow B_\varepsilon + \tfrac{1}{2}\left\{\| \boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})} \|^2 + \mathrm{tr}(\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})})\right\}$$

$$\mu_{q(1/\sigma^2_\varepsilon)} \leftarrow \frac{A_{q(\sigma^2_\varepsilon)}}{B_{q(\sigma^2_\varepsilon)}}$$

$$A_{q(\sigma^2_{\boldsymbol{\beta}})} \leftarrow \frac{p}{2} + A_{\boldsymbol{\beta}}$$

$$B_{q(\sigma^2_{\boldsymbol{\beta}})} \leftarrow B_{\boldsymbol{\beta}} + \tfrac{1}{2}\left\{\| \boldsymbol{\mu}_{q(\boldsymbol{\beta})} \|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\right\}$$

$$\mu_{q(1/\sigma^2_{\boldsymbol{\beta}})} \leftarrow \frac{A_{q(\sigma^2_{\boldsymbol{\beta}})}}{B_{q(\sigma^2_{\boldsymbol{\beta}})}}$$

$$A_{q(\sigma^2_{\boldsymbol{u}_\ell})} \leftarrow \frac{K_\ell}{2} + A_{\boldsymbol{u}}$$

$$B_{q(\sigma^2_{\boldsymbol{u}_\ell})} \leftarrow B_{\boldsymbol{u}} + \tfrac{1}{2}\left\{\| \boldsymbol{\mu}_{q(\boldsymbol{u}_\ell)} \|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_\ell)})\right\}$$

$$\mu_{q(1/\sigma^2_{\boldsymbol{u}_\ell})} \leftarrow \frac{A_{q(\sigma^2_{\boldsymbol{u}_\ell})}}{B_{q(\sigma^2_{\boldsymbol{u}_\ell})}}$$

until the increase in $\log \underline{p}(y; q)$ is negligible.

Firstly,

$$
\begin{aligned}
&E_q\{\log p(\boldsymbol{y}|\beta_0, \boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2)\} \\
&\quad = -\frac{n}{2}\log(2\boldsymbol{\pi}) - \frac{n}{2}\mu_{q(\log\sigma_\varepsilon^2)} \\
&\qquad -\frac{1}{2}\mu_{q(1/\sigma_\varepsilon^2)}\{\|\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})}\|^2 + \mathrm{tr}(\boldsymbol{C}^T\boldsymbol{C})\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})}\}
\end{aligned}
$$

Secondly,

$$
\begin{aligned}
&E_q\{\log p(\sigma_\varepsilon^2) - \log q(\sigma_\varepsilon^2)\} \\
&\quad = A_\varepsilon\log(B_\varepsilon) - \log\,\Gamma(A_\varepsilon) \\
&\qquad -(A_\varepsilon + \frac{n}{2})\log(B_{q(\sigma_\varepsilon^2)}) + \log\,\Gamma(A_\varepsilon + \frac{n}{2}) \\
&\qquad +\frac{n}{2}\mu_{q(\log\sigma_\varepsilon^2)} + \frac{1}{2}\mu_{q(1/\sigma_\varepsilon^2)}\{\|\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})}\|^2 + \mathrm{tr}(\boldsymbol{C}^T\boldsymbol{C})\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})}\}
\end{aligned}
$$

Thirdly,

$$
\begin{aligned}
&E_q\{\log p(\sigma_{\boldsymbol{\beta}}^2) - \log q(\sigma_{\boldsymbol{\beta}}^2)\} \\
&\quad = A_{\boldsymbol{\beta}}\log(B_{\boldsymbol{\beta}}) - \log\,\Gamma(A_{\boldsymbol{\beta}}) \\
&\qquad -A_{q(\sigma_{\boldsymbol{\beta}}^2)}\log(B_{q(\sigma_{\boldsymbol{\beta}}^2)}) + \log\,\Gamma(A_{q(\sigma_{\boldsymbol{\beta}}^2)}) \\
&\qquad +\frac{p}{2}\mu_{q(\log\sigma_{\boldsymbol{\beta}}^2)} + \frac{1}{2}\mu_{q(1/\sigma_{\boldsymbol{\beta}}^2)}\{\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\}
\end{aligned}
$$

Fourthly,

$$
\begin{aligned}
&E_q\{\log p(\sigma_{\boldsymbol{u}_\ell}^2) - \log q(\sigma_{\boldsymbol{u}_\ell}^2)\} \\
&\quad A_{\boldsymbol{u}}\log(B_{\boldsymbol{u}}) - \log\,\Gamma(A_{\boldsymbol{u}}) \\
&\qquad -A_{q(\sigma_{\boldsymbol{u}_\ell}^2)}\log(B_{q(\sigma_{\boldsymbol{u}_\ell}^2)}) + \log\,\Gamma(A_{q(\sigma_{\boldsymbol{u}_\ell}^2)}) \\
&\qquad +\frac{K_\ell}{2}\mu_{q(\log(\sigma_{\boldsymbol{u}_\ell}^2))} + \frac{1}{2}\mu_{q(1/\sigma_{\boldsymbol{u}_\ell}^2)}\{\|\boldsymbol{\mu}_{q(\boldsymbol{u}_\ell)}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_\ell)})\}
\end{aligned}
$$

Fifthly,

$$
\begin{aligned}
&E_q\{\log p(\sigma_\varepsilon^2|a) - \log q(\sigma_\varepsilon^2)\} \\
&\quad -\frac{1}{2}\mu_{q(\log a)} - \frac{1}{2}\log(\pi) \\
&\quad -A_{q(\sigma_\varepsilon^2)}\log(B_{q(\sigma_\varepsilon^2)}) + \log\,\Gamma(A_{q(\sigma_\varepsilon^2)} \\
&\quad +\frac{n}{2}\mu_{q(\log\sigma_\varepsilon^2)} + \frac{1}{2}\mu_{q(1/\sigma_\varepsilon^2)}\{\|\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})}\|^2 + \mathrm{tr}(\boldsymbol{C}^T\boldsymbol{C})\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}\}
\end{aligned}
$$

Sixthly,

$$
\begin{aligned}
&E_q\{\log p(a) - \log q(a)\} \\
&\quad -\log A - \frac{1}{2}\log(\pi) - \log(B_{q(a)}) + \frac{1}{2}\mu_{q(\log a)} + \mu_{q(1/a)}\mu_{1(1/\sigma_\varepsilon^2)}
\end{aligned}
$$

Seventhly,

$$
\begin{aligned}
&E_q\{\log p(\beta_0) + \log p(\boldsymbol{\beta}|\sigma_{\boldsymbol{\beta}}^2) + \sum_{\ell=1}^{r}\log p(\boldsymbol{u}_\ell|\sigma_{\boldsymbol{u}_\ell}^2) - \log q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})\} \\
&\quad \frac{1+p+K_\ell}{2} - \frac{1}{2}\log(\sigma_{\beta_0}^2) - \frac{1}{2}\frac{\mu_{q(\beta_0)}^2 + \sigma_{q(\beta_0)}^2}{\sigma_{\beta_0}^2} \\
&\quad -\frac{p}{2}\mu_{q(\log(\sigma_{\boldsymbol{\beta}}^2))} - \frac{1}{2}\mu_{q(1/\sigma_{\boldsymbol{\beta}}^2)}\{\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\} \\
&\quad -\frac{K_\ell}{2}\mu_{q(\log(\sigma_{\boldsymbol{u}_\ell}^2))} - \frac{1}{2}\mu_{q(1/\sigma_{\boldsymbol{u}_\ell}^2)}\{\|\boldsymbol{\mu}_{q(\boldsymbol{u}_\ell)}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_\ell)})\} \\
&\quad +\frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})}|
\end{aligned}
$$

Substitution of these gives the lower bound:

$$\begin{aligned}
\log \underline{p}(y;q) \;=\; & -\frac{n}{2}\log(2\boldsymbol{\pi}) + \frac{1+p+K_\ell}{2} - \frac{1}{2}\log(\sigma^2_{\beta_0}) - \frac{1}{2}\frac{\mu^2_{q(\beta_0)} + \sigma^2_{q(\beta_0)}}{\sigma^2_{\beta_0}} \\
& + A_\varepsilon \log(B_\varepsilon) - \log\,\Gamma(A_\varepsilon) \\
& - (A_\varepsilon + \frac{n}{2})\log(B_{q(\sigma^2_\varepsilon)}) + \log\,\Gamma(A_\varepsilon + \frac{n}{2}) \\
& + A_{\boldsymbol{\beta}} \log(B_{\boldsymbol{\beta}}) - \log\,\Gamma(A_{\boldsymbol{\beta}}) \\
& - A_{q(\sigma^2_{\boldsymbol{\beta}})}\log(B_{q(\sigma^2_{\boldsymbol{\beta}})}) + \log\,\Gamma(A_{q(\sigma^2_{\boldsymbol{\beta}})}) \\
& + r(A_{\boldsymbol{u}}\log(B_{\boldsymbol{u}}) - \log\,\Gamma(A_{\boldsymbol{u}})) \\
& - \sum_{\ell=1}^{r}[A_{q(\sigma^2_{\boldsymbol{u}_\ell})}\log(B_{q(\sigma^2_{\boldsymbol{u}_\ell})}) - \log\,\Gamma(A_{q(\sigma^2_{\boldsymbol{u}_\ell})})] \\
& + \frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})}|
\end{aligned}$$

2.E Appendix: MFVB Algorithm of model (3)

The mixed model with the ridge penalized method for the fixed effects and a Log-Normal prior for the random variance is given by:

$$\begin{aligned}
& \boldsymbol{y}|\beta_0, \boldsymbol{\beta}, \boldsymbol{u}, \sigma^2_\varepsilon \overset{\text{ind.}}{\sim} N(\mathbf{1}\beta_0 + X\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \sigma^2_\varepsilon I_n) \\
& \sigma^2_\varepsilon \sim \text{ Inverse-Gamma}(A_\varepsilon, B_\varepsilon) \\
& \beta_0 \sim N(0, \sigma^2_{\beta_0}) \\
& \boldsymbol{\beta}|\sigma^2_{\boldsymbol{\beta}} \sim N(\mathbf{0}, \sigma^2_{\boldsymbol{\beta}}\boldsymbol{I}) \\
& \sigma^2_{\boldsymbol{\beta}} \sim \text{ Inverse-Gamma}(A_{\boldsymbol{\beta}}, B_{\boldsymbol{\beta}}) \\
& \boldsymbol{u}_\ell|\sigma^2_{\boldsymbol{u}_\ell} \sim N(0, \sigma^2_{\boldsymbol{u}_\ell} I_{K_\ell}) \;\; 1 \le \ell \le r \\
& \sigma^2_{\boldsymbol{u}_\ell}|a_{\boldsymbol{u}}, b_{\boldsymbol{u}} \sim \text{log-normal}(a_{\boldsymbol{u}}, b_{\boldsymbol{u}}) \;\; 1 \le \ell \le r \\
& a_{\boldsymbol{u}} \sim N(0, \sigma^2_{a_{\boldsymbol{u}}}) \\
& b_{\boldsymbol{u}} \sim \text{ Inverse-Gamma}(A_{\boldsymbol{u}}, B_{\boldsymbol{u}})
\end{aligned} \tag{2.19}$$

where $\sigma^2_{\beta_0}$, A_ε, B_ε, $A_{\boldsymbol{\beta}}$, $B_{\boldsymbol{\beta}}$, $\sigma^2_{a_{\boldsymbol{u}}}$, $A_{\boldsymbol{u}}$ and $B_{\boldsymbol{u}}$ are hyperparameters. The product restriction that we impose here is:

$$q(\beta_0, \boldsymbol{\beta}, \boldsymbol{u}, \sigma^2_\varepsilon, \sigma^2_{\boldsymbol{\beta}}, a_{\boldsymbol{u}}, b_{\boldsymbol{u}}, \sigma^2_{\boldsymbol{u}_1}, ..., \sigma^2_{\boldsymbol{u}_r}) = q(\beta_0, \boldsymbol{\beta}, \boldsymbol{u}) q(\sigma^2_\varepsilon) q(\sigma^2_{\boldsymbol{\beta}}) q(a_{\boldsymbol{u}}) q(b_{\boldsymbol{u}}) q(\sigma^2_{\boldsymbol{u}_1}, ..., \sigma^2_{\boldsymbol{u}_r}).$$

Then, the optimal q^* densities for the parameters in model (2.19) take the form:

$$\begin{aligned}
&q^*(\beta_0, \boldsymbol{\beta}, \boldsymbol{u}) \text{ is a Multivariate Normal density function,} \\
&q^*(\sigma^2_\varepsilon) \text{ is an Inverse Gamma density function,} \\
&q^*(\sigma^2_{\boldsymbol{\beta}}) \text{ is an Inverse Gamma density function,} \\
&q^*(a_{\boldsymbol{u}}) \text{ is a Normal density functions,} \\
&q^*(b_{\boldsymbol{u}}) \text{ is an Inverse Gamma density function, and} \\
&q^*(\sigma^2_{\boldsymbol{u}_\ell}) = \frac{(\sigma^2_\varepsilon)^{C^*_{\ell,1}-1} \exp\{-\frac{C^*_{\ell,2}}{\sigma^2_\varepsilon} - C^*_{\ell,3}[\log(\sigma^2_\varepsilon)]^2\}}{\mathcal{J}(0, C^*_{\ell,1}, C^*_{\ell,3}, C^*_{\ell,2})}.
\end{aligned} \tag{2.20}$$

Here $\mathcal{J}$ is defined in Definition 1.10 and

$$\begin{aligned}
C^*_{\ell,1} &= -\frac{K_\ell}{2} + \mu_{q(a_{\boldsymbol{u}})} \mu_{q(1/b_{\boldsymbol{u}})}, \\
C^*_{\ell,2} &= \frac{1}{2}[\|\boldsymbol{\mu}_{q(\boldsymbol{u}_\ell)}\|^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_\ell)})], \\
C^*_{\ell,3} &= \frac{1}{2} \mu_{q(1/b_{\boldsymbol{u}})}.
\end{aligned}$$

Let $\boldsymbol{\mu}_{q(\beta_0, \boldsymbol{\beta}, \boldsymbol{u})}$ and $\boldsymbol{\Sigma}_{q(\beta_0, \boldsymbol{\beta}, \boldsymbol{u})}$ denote the mean vector and covariance matrix for the Multivariate Normal density function $q^*(\beta_0, \boldsymbol{\beta}, \boldsymbol{u})$, $\mu_{q(a_{\boldsymbol{u}})}$ and $\sigma^2_{q(a_{\boldsymbol{u}})}$ denote the mean and variance for the Normal density function $q^*(a_{\boldsymbol{u}})$, and $A_{q(\sigma^2_{\boldsymbol{\beta}})}$ and $B_{q(\sigma^2_{\boldsymbol{\beta}})}$ denote the shape and rate parameters for $q^*(\sigma^2_{\boldsymbol{\beta}})$. A similar definition is used for the parameters in $q^*(b_{\boldsymbol{u}})$ and $q^*(\sigma^2_\varepsilon)$. Let $\boldsymbol{C} = [1, \boldsymbol{X}, \boldsymbol{Z}]$. The convergence of Algorithm 2.E.1 can be monitored using the following expression for the lower

bound on the marginal log-likelihood:

$$
\begin{aligned}
\log \underline{p}(y;q) \;=\; & \frac{2+p+\sum_{\ell=1}^{r} K_\ell}{2} - \frac{n+r}{2}\log(2\pi) \\
& + A_\varepsilon \log(B_\varepsilon) - \log\,\Gamma(A_\varepsilon) - \frac{1}{2}\log(\boldsymbol{\sigma}^2_{\boldsymbol{\beta}_0}) \\
& + A_{\boldsymbol{\beta}} \log(B_{\boldsymbol{\beta}}) - \log\,\Gamma(A_{\boldsymbol{\beta}}) - \frac{1}{2}\log(\boldsymbol{\sigma}^2_{a_{\boldsymbol{u}}}) \\
& + A_{\boldsymbol{u}} \log(B_{\boldsymbol{u}}) - \log\,\Gamma(A_{\boldsymbol{u}}) \\
& - (A_\varepsilon + \frac{n}{2})\log(B_{q(\sigma^2\varepsilon)}) + \log\,\Gamma(A_\varepsilon + \frac{n}{2}) \\
& - (A_{\boldsymbol{u}} + \frac{r}{2})\log(B_{q(\sigma^2_{\boldsymbol{u}})}) + \log\,\Gamma(A_{\boldsymbol{u}} + \frac{r}{2}) \\
& - (A_{\boldsymbol{\beta}} + \frac{p}{2})\log(B_{q(\sigma^2_{\boldsymbol{\beta}})}) + \log \mathcal{J}(0, C^*_{\ell,1}, C^*_{\ell,3}, C^*_{\ell,2}) \\
& + \frac{1}{2}\log|\boldsymbol{\sigma}^2_{q(a_{\boldsymbol{u}})}| + \frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})}| \\
& - \frac{\mu^2_{q(\beta_0)} + \sigma^2_{q(\beta_0)}}{2\sigma^2_{\boldsymbol{\beta}_0}} - \frac{\mu^2_{q(a_{\boldsymbol{u}})} + \sigma^2_{q(a_{\boldsymbol{u}})}}{2\sigma^2_{a_{\boldsymbol{u}}}} \\
& + \sum_{\ell=1}^{r} \left\{ \frac{1}{2}\mu_{q(1/b_{\boldsymbol{u}})}\mu_{q([\log\sigma^2_{\boldsymbol{u}_\ell}]^2)} - \nu_{q(a_{\boldsymbol{u}})}\mu_{q(1/b_{\boldsymbol{u}})}\mu_{q(\log\sigma^2_{\boldsymbol{u}_\ell})} \right\}
\end{aligned}
$$

Algorithm 2.E.1: MFVB iterative scheme for obtaining the parameters of the optimal densities $q^*(\beta_0, \boldsymbol{\beta}, \boldsymbol{u})$, $q^*(\sigma^2_{\boldsymbol{\beta}})$, $q^*(\sigma^2_{\boldsymbol{u}_\ell})$, $q^*(\sigma^2_\varepsilon)$, $q^*(a_{\boldsymbol{u}})$ and $q^*(b_{\boldsymbol{u}})$ for the mixed model with the ridge penalized method for the fixed effects and a Log-Normal prior for the random variance.

Initialize $\mu_{q(1/\sigma^2_\varepsilon)}$, $\mu_{q(1/b_\varepsilon)}$, $\mu_{q(a_\varepsilon)}$, $\mu_{q(\log\sigma^2_\ell)}$ and $\mu_{q([\log\sigma^2_\ell]^2)}$;
Cycle

$$\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \Big\{\boldsymbol{C}^T\boldsymbol{C}\mu_{q(1/\sigma^2_\varepsilon)} + \text{blockdiag}\big[(\sigma^2_{\beta_0})^{-1}, \mu_{q(1/\sigma^2_{\boldsymbol{\beta}})}\boldsymbol{I}_p, \mu_{q(1/\sigma^2_{\boldsymbol{u}_1})}\boldsymbol{I}_{K_1}, ..., \mu_{q(1/\sigma^2_{\boldsymbol{u}_1})}\boldsymbol{I}_{K_r}\big]\Big\}^{-1}$$

$$\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T\boldsymbol{y}\mu_{q(1/\sigma^2_\varepsilon)}$$

$$A_{q(\sigma^2_\varepsilon)} \leftarrow \frac{n}{2} + A_\varepsilon$$

$$B_{q(\sigma^2_\varepsilon)} \leftarrow B_\varepsilon + \tfrac{1}{2}\left\{\| \boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})} \|^2 + \text{tr}(\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})})\right\}$$

$$\mu_{q(1/\sigma^2_\varepsilon)} \leftarrow \frac{A_{q(\sigma^2_\varepsilon)}}{B_{q(\sigma^2_\varepsilon)}}$$

$$A_{q(\sigma^2_{\boldsymbol{\beta}})} \leftarrow \frac{p}{2} + A_{\boldsymbol{\beta}}$$

$$B_{q(\sigma^2_{\boldsymbol{\beta}})} \leftarrow B_{\boldsymbol{\beta}} + \tfrac{1}{2}\left\{\| \boldsymbol{\mu}_{q(\boldsymbol{\beta})} \|^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\right\}$$

$$\mu_{q(1/\sigma^2_{\boldsymbol{\beta}})} \leftarrow \frac{A_{q(\sigma^2_{\boldsymbol{\beta}})}}{B_{q(\sigma^2_{\boldsymbol{\beta}})}}$$

$$A_{q(b_{\boldsymbol{u}})} \leftarrow A_b + \frac{1}{2}$$

$$B_{q(b_{\boldsymbol{u}})} \leftarrow B_b + \frac{1}{2}\sum_{\ell=1}^{r}\{\mu^2_{q(a_{\boldsymbol{u}})} + \sigma^2_{q(a_{\boldsymbol{u}})} - 2\mu_{q(a_{\boldsymbol{u}})}\mu_{q(\log\sigma^2_{\boldsymbol{u}_\ell})} + \mu_{q([\log\sigma^2_{\boldsymbol{u}_\ell}]^2)}\}$$

$$\mu_{q(1/b_{\boldsymbol{u}})} \leftarrow \frac{A_{q(b_{\boldsymbol{u}})}}{B_{q(b_{\boldsymbol{u}})}};\ \sigma^2_{q(a_{\boldsymbol{u}})} \leftarrow [\frac{1}{\sigma^2_{a_{\boldsymbol{u}}}} + \mu_{q(1/b_{\boldsymbol{u}})}]^{-1}$$

$$\mu_{q(a_{\boldsymbol{u}})} \leftarrow \sigma^2_{q(a_{\boldsymbol{u}})}\mu_{q(1/b_{\boldsymbol{u}})}\sum_{\ell=1}^{r}\mu_{q(\log\sigma^2_{\boldsymbol{u}_\ell})}$$

$$C^*_{\ell,1} \leftarrow -\frac{K_\ell}{2} + \mu_{q(a_{\boldsymbol{u}})}\mu_{q(1/b_{\boldsymbol{u}})};\ C^*_{\ell,2} \leftarrow \frac{1}{2}[\|\boldsymbol{\mu}_{q(\boldsymbol{u}_\ell)}\|^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_\ell)})]$$

$$C^*_{\ell,3} \leftarrow \frac{1}{2}\mu_{q(1/b_{\boldsymbol{u}})}$$

$$\mu_{q(1/\sigma^2_{\boldsymbol{u}_\ell})} \leftarrow \frac{\mathcal{J}(0, C^*_{\ell,1} - 1, C^*_{\ell,3}, C^*_{\ell,2})}{\mathcal{J}(0, C^*_{\ell,1}, C^*_{\ell,3}, C^*_{\ell,2})}$$

$$\mu_{q(\log\sigma^2_{\boldsymbol{u}_\ell})} \leftarrow \frac{\mathcal{J}(1, C^*_{\ell,1}, C^*_{\ell,3}, C^*_{\ell,2})}{\mathcal{J}(0, C^*_{\ell,1}, C^*_{\ell,3}, C^*_{\ell,2})};\ \mu_{q([\log\sigma^2_{\boldsymbol{u}_\ell}]^2)} \leftarrow \frac{\mathcal{J}(2, C^*_{\ell,1}, C^*_{\ell,3}, C^*_{\ell,2})}{\mathcal{J}(0, C^*_{\ell,1}, C^*_{\ell,3}, C^*_{\ell,2})}$$

until the increase in $\log \underline{p}\,(\boldsymbol{y}; q)$ is negligible.

2.F Appendix: MFVB Algorithm of model (4)

The mixed model with the ridge penalized method for the fixed effects and a Half-Cauchy prior for the random variance

$$
\begin{aligned}
&\boldsymbol{y}|\beta_0, \boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2 \overset{\text{ind.}}{\sim} N(\mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \sigma_\varepsilon^2 I_n) \\
&\beta_0 \sim N(0, \sigma_{\beta_0}^2) \\
&\boldsymbol{\beta}|\sigma_{\boldsymbol{\beta}}^2 \sim N(\mathbf{0}, \sigma_{\boldsymbol{\beta}}^2 \boldsymbol{I}) \\
&\sigma_{\boldsymbol{\beta}}^2 \sim \text{ Inverse-Gamma}(A_{\boldsymbol{\beta}}, B_{\boldsymbol{\beta}}) \\
&\boldsymbol{u}_\ell|\sigma_{\boldsymbol{u}_\ell} \sim N(0, \sigma_{\boldsymbol{u}_\ell}^2 I_{K_\ell}) \;\; 1 \leq \ell \leq r \\
&\sigma_{\boldsymbol{u}_\ell}^2|a \sim \text{ Inverse-Gamma}(1/2, 1/a) \;\; 1 \leq \ell \leq r \\
&a \sim \text{ Inverse-Gamma}(1/2, 1/A^2) \\
&\sigma_\varepsilon^2 \sim \text{ Inverse-Gamma}(A_\varepsilon, B_\varepsilon)
\end{aligned}
\tag{2.21}
$$

where A_ε, B_ε, $\sigma_{\beta_0}^2$, A, $A_{\boldsymbol{\beta}}$, $B_{\boldsymbol{\beta}} > 0$ are hyperparameters. The product restriction that we impose here is:

$$
q(\beta_0, \boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2, \sigma_{\boldsymbol{\beta}}^2, a, \sigma_{\boldsymbol{u}_1}^2, ..., \sigma_{\boldsymbol{u}_r}^2) = q(\beta_0, \boldsymbol{\beta}, \boldsymbol{u}) q(\sigma_\varepsilon^2) q(\sigma_{\boldsymbol{\beta}}^2) q(a) q(\sigma_{\boldsymbol{u}_1}^2, ..., \sigma_{\boldsymbol{u}_r}^2).
$$

Then, the optimal q^* densities for the parameters in model (2.21) take the form:

$$
\begin{aligned}
&q^*(\beta_0, \boldsymbol{\beta}, \boldsymbol{u}) \text{ is a Multivariate Normal density function,} \\
&q^*(\sigma_\varepsilon^2) \text{ is an Inverse Gamma density function,} \\
&q^*(\sigma_{\boldsymbol{\beta}}^2) \text{ is an Inverse Gamma density function,} \\
&q^*(\sigma_{\boldsymbol{u}_\ell}^2) \text{ is an Inverse Gamma density function,} \\
&q^*(a) \text{ is an Inverse Gamma density function.}
\end{aligned}
\tag{2.22}
$$

Let $\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})}$ and $\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})}$ denote the mean vector and covariance matrix for the Multivariate Normal density function $q^*(\beta_0, \boldsymbol{\beta}, \boldsymbol{u})$, and $A_{q(\sigma^2_{\boldsymbol{\beta}})}$ and $B_{q(\sigma^2_{\boldsymbol{\beta}})}$ denote the shape and rate parameters for $q^*(\sigma^2_{\boldsymbol{\beta}})$. A similar definition is used for the parameters in $q^*(a)$, $q^*(\sigma^2_{\boldsymbol{u}_\ell})$ and $q^*(\sigma^2_\varepsilon)$. Let $\boldsymbol{C} = [1, \boldsymbol{X}, \boldsymbol{Z}]$. The convergence of Algorithm 2.F.1 can be monitored using the following expression for the lower bound on the marginal log-likelihood:

$$
\begin{aligned}
\log \underline{p}(y; q) \;=\; & \frac{1+p+K_\ell}{2} - \frac{n}{2}\log(2\boldsymbol{\pi}) - \frac{1}{2}\log(\sigma^2_{\beta_0}) \\
& - \frac{\mu^2_{q(\beta_0)} + \sigma^2_{q(\beta_0)}}{2\sigma^2_{\beta_0}} + \frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})}| \\
& + A_\varepsilon \log(B_\varepsilon) - \log\,\Gamma(A_\varepsilon) \\
& - (A_\varepsilon + \frac{n}{2})\log(B_{q(\sigma^2_\varepsilon)}) + \log\,\Gamma(A_\varepsilon + \frac{n}{2}) \\
& + A_{\boldsymbol{\beta}} \log(B_{\boldsymbol{\beta}}) - \log\,\Gamma(A_{\boldsymbol{\beta}}) \\
& - A_{q(\sigma^2_{\boldsymbol{\beta}})}\log(B_{q(\sigma^2_{\boldsymbol{\beta}})}) + \log\,\Gamma(A_{q(\sigma^2_{\boldsymbol{\beta}})}) \\
& - \log A - \frac{1+r}{2}\log(\pi) - A_{q(a)}\log(B_{q(a)}) + \log\,\Gamma(A_{q(a)}) \\
& + \sum_{\ell=1}^{r} \mu_{q(1/a)}\mu_{1(1/\sigma^2_{\boldsymbol{u}_\ell})} - \sum_{\ell=1}^{r} A_{q(\sigma^2_{\boldsymbol{u}_\ell})}\log(B_{q(\sigma^2_{\boldsymbol{u}_\ell})}) + \sum_{\ell=1}^{r} \log\,\Gamma(A_{q(\sigma^2_{\boldsymbol{u}_\ell})})
\end{aligned}
$$

Algorithm 2.F.1: MFVB iterative scheme for obtaining the parameters of the optimal densities $q^*(\beta_0, \boldsymbol{\beta}, \boldsymbol{u})$, $q^*(\sigma^2_{\boldsymbol{\beta}})$, $q^*(\sigma^2_{\boldsymbol{u}_\ell})$, $q^*(\sigma^2_\varepsilon)$ and $q^*(a)$ for the mixed model with the ridge penalized method for the fixed effects and a Half-Cauchy prior for the random variance.

Initialize $\mu_{q(1/\sigma^2_{\boldsymbol{\beta}})}$, $\mu_{q(1/\sigma^2_{\boldsymbol{u}_1})}$,...,$\mu_{q(1/\sigma^2_{\boldsymbol{u}_r})}$;

Cycle

$$\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \Big\{\mu_{q(1/\sigma^2_\varepsilon)}\boldsymbol{C}^T\boldsymbol{C} + \text{blockdiag}\big[(\sigma^2_{\beta_0})^{-1}, \mu_{q(1/\sigma^2_{\boldsymbol{\beta}})}\boldsymbol{I}_p, \mu_{q(1/\sigma^2_{\boldsymbol{u}_1})}\boldsymbol{I}_{K_1}, ..., \mu_{q(1/\sigma^2_{\boldsymbol{u}_1})}\boldsymbol{I}_{K_r}\big]\Big\}^{-1}$$

$$\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T\boldsymbol{y}\mu_{q(1/\sigma^2_\varepsilon)}$$

$$A_{q(\sigma^2_\varepsilon)} \leftarrow \frac{n}{2} + A_\varepsilon$$

$$B_{q(\sigma^2_\varepsilon)} \leftarrow B_\varepsilon + \tfrac{1}{2}\left\{\| \boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})} \|^2 + \text{tr}(\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})})\right\}$$

$$\mu_{q(1/\sigma^2_\varepsilon)} \leftarrow \frac{A_{q(\sigma^2_\varepsilon)}}{B_{q(\sigma^2_\varepsilon)}}$$

$$A_{q(\sigma^2_{\boldsymbol{\beta}})} \leftarrow \frac{p}{2} + A_{\boldsymbol{\beta}}$$

$$B_{q(\sigma^2_{\boldsymbol{\beta}})} \leftarrow B_{\boldsymbol{\beta}} + \tfrac{1}{2}\left\{\| \boldsymbol{\mu}_{q(\boldsymbol{\beta})} \|^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\right\}$$

$$\mu_{q(1/\sigma^2_{\boldsymbol{\beta}})} \leftarrow \frac{A_{q(\sigma^2_{\boldsymbol{\beta}})}}{B_{q(\sigma^2_{\boldsymbol{\beta}})}}$$

$$A_{q(\sigma^2_{\boldsymbol{u}_\ell})} \leftarrow \frac{K_{\boldsymbol{u}_\ell}}{2} + \tfrac{1}{2}$$

$$B_{q(\sigma^2_{\boldsymbol{u}_\ell})} \leftarrow \mu_{q(1/a)} + \tfrac{1}{2}\left\{\| \boldsymbol{\mu}_{q(\boldsymbol{u}_\ell)} \|^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_\ell)})\right\}$$

$$\mu_{q(1/\sigma^2_{\boldsymbol{u}_\ell})} \leftarrow \frac{A_{q(\sigma^2_{\boldsymbol{u}_\ell})}}{B_{q(\sigma^2_{\boldsymbol{u}_\ell})}}$$

$$A_{q(a)} \leftarrow \frac{1+r}{2}$$

$$B_{q(a)} \leftarrow \frac{1}{A^2} + \sum_{\ell=1}^{r}\mu_{q(1/\sigma^2_{\boldsymbol{u}_\ell})}$$

$$\mu_{q(1/a)} \leftarrow \frac{A_{q(a)}}{B_{q(a)}}$$

until the increase in $\log \underline{p}(y; q)$ is negligible.

2.G Appendix: Derivation of Algorithm 2.6.1

In Algorithm 2.6.1, the MFVB calculations for β_0, $\boldsymbol{\beta}$ and $\sigma^2_{\boldsymbol{\beta}}$ are similar to the Gaussian case (Algorithm 2.2.1). We obtain them by using

$$1 \text{ to replace } \mu_{1/\sigma^2_\varepsilon}$$

$$\text{and } \boldsymbol{\mu}_{q(\boldsymbol{a})} \text{ to replace } \boldsymbol{y}.$$

Therefore, I only show the derivation for $\boldsymbol{a}$.

2.G.1 Full conditional for $\boldsymbol{a}$

$$\begin{aligned}\log p(a_i|\text{rest}) &= -\tfrac{1}{2}[a_i - (\boldsymbol{C}\boldsymbol{\beta})_i]^2 \\ &\quad +\log\left[(\boldsymbol{I}(a_i \geq 0))^{y_i}(\boldsymbol{I}(a_i < 0))^{1-y_i}\right] + \text{const.}\end{aligned}$$

where

$$\tilde{\boldsymbol{\beta}} = \begin{bmatrix}\beta_0 \\ \boldsymbol{\beta}\end{bmatrix}, \quad \boldsymbol{C} = [\mathbf{1}, \boldsymbol{X}]\,.$$

Derivation:

$$\begin{aligned}p(a_i|\text{rest}) &\propto p(a_i|\beta_0, \boldsymbol{\theta}, \boldsymbol{\gamma})p(y_i|a_i) \\ &= (2\pi)^{-1/2}\exp\left\{-\frac{[a_i - \beta_0 - \boldsymbol{X}_i\boldsymbol{\beta}]^2}{2}\right\} \\ &\quad \times(\boldsymbol{I}(a_i \geq 0))^{y_i}(\boldsymbol{I}(a_i < 0))^{1-y_i}\end{aligned}$$

Taking logarithms, we get

$$\begin{aligned}\log p(a_i|\text{rest}) &= -\frac{[a_i - \beta_0 - \boldsymbol{X}_i\boldsymbol{\beta}]^2}{2} \\ &\quad +\log\left[(\boldsymbol{I}(a_i \geq 0))^{y_i}(\boldsymbol{I}(a_i < 0))^{1-y_i}\right] + \text{const.} \\ &= -\tfrac{1}{2}[a_i - (\boldsymbol{C}\tilde{\boldsymbol{\beta}})_i]^2 \\ &\quad +\log\left[(\boldsymbol{I}(a_i \geq 0))^{y_i}(\boldsymbol{I}(a_i < 0))^{1-y_i}\right] + \text{const.}\end{aligned}$$

2.G.2 Expressions for $q^*(\boldsymbol{a})$ and $\boldsymbol{\mu}_{q(\boldsymbol{a})}$

If $y_i = 1$

$$q(a_i) = \frac{\phi(a_i - (\boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})})_i)}{\Phi((\boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})})_i)}, \quad a_i \geq 0,$$

which is a truncated normal density function on $(0, \infty)$; and if $y_i = 0$,

$$q(a_i) = \frac{\phi(a_i - (\boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})})_i)}{1 - \Phi((\boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})})_i)}, \quad a_i < 0,$$

which is a truncated normal density function on $(-\infty, 0)$.

Derivation:

If $y_i = 1$, then $a_i \geq 0$ and

$$\begin{aligned}\log q^*(a_i) &= E_q[p(a_i|\text{rest})] \\ &= -\tfrac{1}{2}E_q\left\{[a_i - (\boldsymbol{C}\tilde{\boldsymbol{\beta}})_i]^2\right\} + const. \\ &= -\tfrac{1}{2}[a_i - (\mu_{q(\boldsymbol{C}\tilde{\boldsymbol{\beta}})})_i]^2 + const. \\ &= -\tfrac{1}{2}[a_i - (\boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})})_i]^2 + const.\end{aligned}$$

The results for $y_i \geq 0$ then follow from the definition (1.27) for the truncated Gaussian distribution. Similarly, we can obtain the result for $y_i < 0$. Using Result

(1.11) of the truncated Gaussian distribution, we obtain the expression

$$\mu_{q(\boldsymbol{a})} = \boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})} + \frac{\phi(\boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})})}{\Phi(\boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})})^{\boldsymbol{y}} \odot [\boldsymbol{1} - \Phi(\boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})})]^{\boldsymbol{1}-\boldsymbol{y}}}.$$

2.G.3 Derivation of lower bound

We note that

$$\begin{aligned}
\log \underline{p}(y;q) &= E_q\{\log p(\boldsymbol{y},\beta_0,\boldsymbol{\beta},\boldsymbol{a}) - \log q(\beta_0,\boldsymbol{\beta},\boldsymbol{a}\} \\
&= E_q\{\log p(\boldsymbol{y}|\boldsymbol{a}) + \log p(\boldsymbol{a}|\boldsymbol{\beta}_0,\boldsymbol{\beta}) + \log p(\beta_0) + \log p(\boldsymbol{\beta}) \\
&\quad -\log q(\beta_0,\boldsymbol{\beta}) - \log q(\boldsymbol{a})\}.
\end{aligned}$$

Firstly,

$$\begin{aligned}
& E_q\{\log p(\beta_0) + \log p(\boldsymbol{\beta}) - \log q(\beta_0,\boldsymbol{\beta})\} \\
= \; & \frac{1+p}{2} - \frac{1}{2}\log(\sigma^2_{\beta_0}) - \frac{1}{2}\frac{\mu^2_{q(\beta_0)} + \sigma^2_{q(\beta_0)}}{\sigma^2_{\beta_0}} \\
& -\frac{p}{2}\log(\sigma^2_{\boldsymbol{\beta}}) - \frac{1}{2\sigma^2_{\boldsymbol{\beta}}}\{\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\} \\
& +\frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})}|.
\end{aligned}$$

Secondly,

$$\begin{aligned}
& E_q\{\log(p(\boldsymbol{y}|\boldsymbol{a})) + \log(p(\boldsymbol{a}|\beta_0,\boldsymbol{\beta})) - \log q(\boldsymbol{a})\} \\
= \; & -\tfrac{1}{2}\mathrm{tr}(\boldsymbol{C}^T\boldsymbol{C})\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})} \\
& +\boldsymbol{y}^T\log[\Phi(\boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})}))] + (\boldsymbol{1}_n - \boldsymbol{y})^T\log[\boldsymbol{1}_n - \Phi(\boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})}))].
\end{aligned}$$

Substitution of these gives the lower bound expression:

$$\begin{aligned}
\log \underline{p}(y;q) \;=\; & \frac{1+p}{2} - \frac{1}{2}\log(\sigma^2_{\beta_0}) - \frac{\mu^2_{q(\beta_0)} + \sigma^2_{q(\beta_0)}}{2\sigma^2_{\beta_0}} \\
& -\frac{p}{2}\log(\sigma^2_{\boldsymbol{\beta}}) - \frac{1}{2\sigma^2_{\boldsymbol{\beta}}}\{\|\mu_{q(\boldsymbol{\beta})}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\} \\
& +\frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta},\boldsymbol{u})}| \\
& -\tfrac{1}{2}\mathrm{tr}(\boldsymbol{C}^T\boldsymbol{C})\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})} + \boldsymbol{y}^T\log[\Phi(\boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})})] \\
& +(\mathbf{1}_n - \boldsymbol{y})^T\log[\mathbf{1}_n - \Phi(\boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})})].
\end{aligned}$$

Chapter 3

Mean Field Variational Bayes Indicator Variable Selection

3.1 Introduction

In Chapter 2, we used the mean field variational Bayes method to obtain the posterior probability of the candidate models M_k, $1 \leq k \leq 2^p$, where p is the number of candidate variables, and to choose the best model based on the maximum posterior model probability. However, we need to consider all of the possible subsets of candidate variables and to compute the posterior probabilities of all candidate models. In a regression model with p candidate variables, the limitation of this method is that computation of the posterior probabilities for all 2^p linear models or 3^p non-linear models would be required. In this chapter, instead of using global searching, we conduct model selection using indictor variables.

The most direct approach to indictor variable selection is setting a spike-and-slab prior. A spike-and-slab prior is a mixture of two distributions, a spike distribution and a slab distribution, where the spike is a distribution with its mass concentrated around zero, and the slab is a flat distribution spread over the pa-

rameter space. The generic form of a spike-and-slab prior is:

$$p(\beta|\gamma) = \gamma p_{\text{slab}}(\beta) + (1-\gamma)p_{\text{spike}}(\beta),$$

where the random variable γ is the indicator variable. If $\gamma = 1$ (i.e., the corresponding variable β is selected), the prior distribution $p(\beta)$ is set to the slab distribution $p_{\text{slab}}(\beta)$; alternatively, if $\gamma = 0$ (i.e., the corresponding variable β is not selected), the prior distribution $p(\beta)$ is set to the spike distribution $p_{\text{spike}}(\beta)$. This prior and model section method was introduced by George and McCulloch (1996) and named stochastic search variable selection (SSVS). This method has been extended to the multivariate case and to various distributions for the spike-and-slab prior (George & McCulloch, 1996; Chipman, Hamada & Wu, 1997; Brown, Vannucci & Fearn, 1998; Yi, George & Allison, 2003; Yi, 2004) and applied to the results of real-world problems, such as gene research (Theo & Mike, 2004).

Dellaportas *et al.* (1995) and Kuo and Mallick (1998) used a point mass at zero instead of a distribution concentrated around zero as their spike distribution. The generic form is:

$$p(\beta|\gamma) = \gamma p_{\text{slab}}(\beta) + (1-\gamma)\delta_0(\beta),$$

where δ_0 is the Dirac delta function.

Several research studies (Dellaportas *et al.*, 1995; Kuo & Mallick, 1998; Liang *et al.*, 2008) have chosen a Gaussian density function as the slab distribution density function (i.e., used the Gaussian-Zero density function as the prior of the candidate variable). This approach to variable selection sets the slab, $p(\beta|\gamma = 1)$, equal to the Gaussian density function and the spike, $p(\beta|\gamma = 0)$, equal to 0. In this chapter, I use the same Gaussian-Zero prior in the candidate variable to build the indictor variable selection models and derive a MFVB approximate inference algorithm.

The Gaussian response linear model and the binary response generalized linear model will be considered in the following two sections.

3.2 Gaussian Response Linear Models

Firstly, I will consider Gaussian response linear models. Chapter 2 showed that the ridge-type, i.e. ℓ_2, penalization for linear coefficients can be used to improve the results of model selection. I will preserve this penalization in the following models.

3.2.1 Models

The ridge penalized linear model with the Gaussian-Zero prior takes the form:

$$\begin{aligned}
&\boldsymbol{y}|\beta_0, \boldsymbol{\beta}, \sigma_\varepsilon^2 \sim N(\mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta}, \sigma_\varepsilon^2 \boldsymbol{I}),\\
&\sigma_\varepsilon^2 \sim \text{ Inverse-Gamma}(A_\varepsilon, B_\varepsilon),\\
&\beta_0 \sim N(0, \sigma_{\beta_0}^2),\\
&p(\beta_j|\sigma_\beta^2, \gamma_j) = \gamma_j \phi_{\sigma_\beta^2}(\beta_j) + (1-\gamma_j)\delta_0(\beta_j),\ 1 \le j \le p,\\
&\sigma_\beta^2 \sim \text{ Inverse-Gamma}(A_\beta, B_\beta),\\
&\gamma_j|\rho \overset{\text{ind.}}{\sim} \text{Bernoulli}(\rho),\\
&\rho \sim \text{ Uniform}(0,1).
\end{aligned} \tag{3.1}$$

We introduce specially tailored auxiliary variables, θ_j, to facilitate more efficient MFVB derivations for model (3.1). Suppose that $\beta_j = \gamma_j \theta_j$, where

$$\theta_j|\sigma_\beta^2 \overset{\text{ind.}}{\sim} N(0, \sigma_\beta^2).$$

Through application of Result 1.14, we have the equivalent model:

$$
\begin{aligned}
&\boldsymbol{y}|\beta_0, \boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma_\varepsilon^2 \sim N(\mathbf{1}\beta_0 + \boldsymbol{X}(\boldsymbol{\gamma} \odot \boldsymbol{\theta}), \sigma_\varepsilon^2 \boldsymbol{I}),\\
&\sigma_\varepsilon^2 \sim \text{ Inverse-Gamma}(A_\varepsilon, B_\varepsilon),\\
&\beta_0 \sim N(0, \sigma_{\beta_0}^2),\\
&\theta_j|\sigma_\beta^2 \overset{\text{ind.}}{\sim} N(0, \sigma_\beta^2),\\
&\sigma_\beta^2 \sim \text{ Inverse-Gamma}(A_\beta, B_\beta),\\
&\gamma_j|\rho \overset{\text{ind.}}{\sim} \text{Bernoulli}(\rho),\\
&\rho \sim \text{ Uniform}(0, 1),
\end{aligned}
\tag{3.2}
$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, ..., \gamma_p)^T$ is the vector of indicator variables; $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_p)^T$ are auxiliary variables corresponding to Result 1.14; $A_\varepsilon, B_\varepsilon, A_\beta, B_\beta, \sigma_{\beta_0}^2 > 0$ are constant hyperparameters. Figure 3.1 shows the directed acyclic graph corresponding to (3.2).

The full conditionals for the Markov chain Monte Carlo can be shown to be:

$$
\begin{aligned}
&p(\beta_0, \boldsymbol{\theta}, \boldsymbol{\gamma}|\text{rest}) = \phi_{\sigma_\varepsilon^2 \boldsymbol{I}_n}(\boldsymbol{y} - \boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\boldsymbol{\theta}})\phi_{\boldsymbol{F}}(\tilde{\boldsymbol{\theta}})\rho^{\gamma_\bullet}(1-\rho)^{p-\gamma_\bullet},\\
&\rho|\text{rest} \sim \text{Beta}(1 + \gamma_\bullet, 1 + p - \gamma_\bullet),\\
&\sigma_\varepsilon^2|\text{rest} \sim \text{IG}(A_\varepsilon + \frac{n}{2}, B_\varepsilon + \frac{1}{2} \parallel \boldsymbol{y} - \boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\boldsymbol{\theta}} \parallel^2),\\
&\sigma_\beta^2|\text{rest} \sim \text{IG}(A_\beta + \frac{p}{2}, B_\beta + \frac{1}{2} \parallel \boldsymbol{\theta} \parallel^2),
\end{aligned}
$$

where

$$
\tilde{\boldsymbol{\theta}} = (\beta_0, \theta_1, ..., \theta_p)^T, \ \gamma_\bullet = \sum_{j=1}^{p} \gamma_j,
$$

$$
\boldsymbol{X}_{\boldsymbol{\gamma}} = [\mathbf{1}, \boldsymbol{X}]\text{diag}(1, \gamma_1, ..., \gamma_p) \ \text{ and } \ \boldsymbol{F} = \text{diag}(\sigma_{\beta_0}^2, \sigma_\beta^2, ..., \sigma_\beta^2).
$$

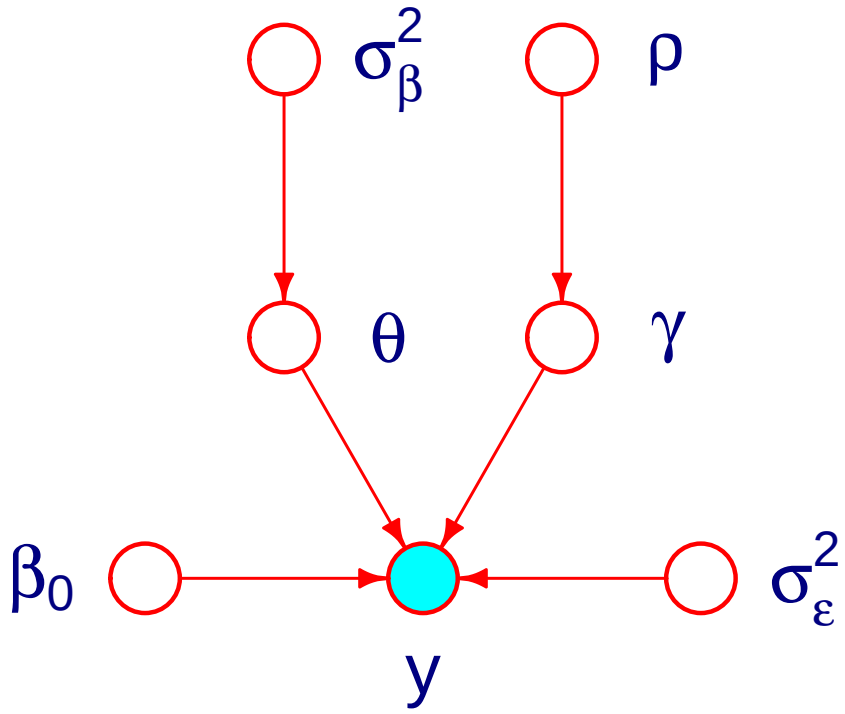

Figure 3.1: Directed acyclic graph for model (3.2).

3.2.2 Mean field variational Bayes

We now seek a quick deterministic approximate inference approach for (3.2) based on MFVB. A tractable solution arises if we impose the product restriction:

$$q(\beta_0, \boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma_\varepsilon^2, \sigma_\beta^2, \rho) = q(\beta_0, \boldsymbol{\theta}, \boldsymbol{\gamma}) q(\sigma_\varepsilon^2, \sigma_\beta^2, \rho).$$

The theory of induced factorizations (e.g., Bishop, 2006, Section 10.2.5) leads to a solution with the additional product structure:

$$q(\beta_0, \boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma_\varepsilon^2, \sigma_\beta^2, \rho) = q(\beta_0, \boldsymbol{\theta}, \boldsymbol{\gamma}) q(\sigma_\varepsilon^2) q(\sigma_\beta^2) q(\rho). \tag{3.3}$$

Then, as shown in Appendix 3.A, the optimal q^* densities for the parameters in Model (3.2) take the form:

$$\begin{aligned}
&q^*(\rho) \text{ is a Beta density function,} \\
&q^*(\sigma_\varepsilon^2) \text{ is an Inverse Gamma density function,} \\
&q^*(\sigma_\beta^2) \text{ is an Inverse Gamma density function,} \\
&q^*(\beta_0, \boldsymbol{\theta}|\boldsymbol{\gamma}) \text{ is a conditional multivariate Gaussian density function,} \\
&q^*(\beta_0, \boldsymbol{\theta}) \text{ is a mixture of multivariate Gaussian density functions, and} \\
&q^*(\gamma_j), 1 \leq j \leq p, \text{ are Bernoulli probability mass functions.}
\end{aligned} \tag{3.4}$$

Let $\boldsymbol{m}_{\boldsymbol{\gamma} q(\beta_0, \boldsymbol{\theta})}$ and $\boldsymbol{V}_{\boldsymbol{\gamma} q(\beta_0, \boldsymbol{\theta})}$ denote the mean vector and covariance matrix for the conditional multivariate Gaussian density function $q^*(\beta_0, \boldsymbol{\theta}|\boldsymbol{\gamma})$, and $A_{q(\sigma_\varepsilon^2)}$ and $B_{q(\sigma_\varepsilon^2)}$ denote the shape and rate parameters for $q^*(\sigma_\varepsilon^2)$. A similar definition was used for the parameters in $q^*(\sigma_\beta^2)$. The $\mu_{q(\omega_{\boldsymbol{\gamma}})}$ denote the weight parameters for $q^*(\beta_0, \boldsymbol{\theta})$, which was formed as

$$q^*(\beta_0, \boldsymbol{\theta}) = \sum_{\boldsymbol{\gamma} \in \mathbb{B}} \mu_{q(\omega_{\boldsymbol{\gamma}})} \phi_{\boldsymbol{V}_{\boldsymbol{\gamma} q(\beta_0, \boldsymbol{\theta})}} \left(\begin{bmatrix} \beta_0 \\ \boldsymbol{\theta} \end{bmatrix} - \boldsymbol{m}_{\boldsymbol{\gamma} q(\beta_0, \boldsymbol{\theta})} \right),$$

where $\mathbb{B}$ denotes a p-dimensional space of $\{0, 1\}$. The convergence of Algorithm (3.2.1) can be monitored using the following expression for the lower bound on the marginal log-likelihood:

$$
\begin{aligned}
\log \underline{p}(y;q) = & -\frac{n}{2}\log(2\pi) + \frac{1+p}{2} - \frac{1}{2}\log(\sigma^2_{\beta_0}) \\
& + A_\varepsilon \log B_\varepsilon - \log\,\Gamma(A_\varepsilon) \\
& - (A_\varepsilon + \frac{n}{2})\log B_{q(\sigma^2_\varepsilon)} + \log\,\Gamma(A_\varepsilon + \frac{n}{2}) \\
& + A_\beta \log B_\beta - \log\,\Gamma(A_\beta) \\
& - (A_\beta + \frac{p}{2})\log B_{q(\sigma^2_\beta)} + \log\,\Gamma(A_\beta + \frac{p}{2}) \\
& + \frac{1}{2}\sum_{\boldsymbol{\gamma}\in\mathbb{B}} \mu_{q(\omega_{\boldsymbol{\gamma}})} \left\{ \log|\boldsymbol{V}_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})}| - \frac{\mu^2_{\boldsymbol{\gamma}q(\beta_0)} + \sigma^2_{\boldsymbol{\gamma}q(\beta_0)}}{\sigma^2_{\beta_0}} - 2\log\mu_{q(\omega_{\boldsymbol{\gamma}})} \right\}.
\end{aligned}
\tag{3.5}
$$

3.2.3 Accuracy assessment

We now investigate the accuracy of the MFVB approximate inference scheme in this context. I give three examples, including one simulation example and two real data examples.

Simulated data

In this example, the data are generated from a simple linear model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, \;\; 1 \leq i \leq n,$$

where the sample size is $n = 100$ and the true parameter values are $\beta_0 = 1$, $\beta_1 = 2$ and $\beta_2 = 0$. This means that x_1 is in the model, and x_2 is a superfluous noise variable. The ε_i, x_{1i} and x_{2i}, $1 \leq i \leq n$, are generated independently from the standard Gaussian distribution. Figure 3.2 summarizes the MCMC inference for the linear indicator variable selection model (3.2) for fitting this simulation data set. Columns two to four indicate that the MCMC convergence is quite

Algorithm 3.2.1: Iterative scheme for obtaining the parameters in the optimal densities $q^*(\beta_0, \boldsymbol{\theta}|\boldsymbol{\gamma})$, $q^*(\beta_0, \boldsymbol{\theta})$, $q^2(\sigma_\varepsilon^2)$, $q^*(\sigma_\beta^2)$, $q^*(\rho)$ and $q^*(\gamma_j)$ for the linear indicator variable selection model (3.2).

Initialize;
Cycle

$$\boldsymbol{V}_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})} \leftarrow \left\{ \boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X}_{\boldsymbol{\gamma}} \mu_{q(1/\sigma_\varepsilon^2)} + \text{blockdiag}\left[(\sigma_{\beta_0}^2)^{-1}, \mu_{q(1/\sigma_\beta^2)} \boldsymbol{I}_p \right] \right\}^{-1}$$

$$\boldsymbol{m}_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})} \leftarrow \boldsymbol{V}_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})} \boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{y} \mu_{q(1/\sigma_\varepsilon^2)}$$

$$\eta_{\boldsymbol{\gamma}} \leftarrow \exp\Big\{ \tfrac{1}{2}\log|\boldsymbol{V}_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})}| + \gamma_\bullet \mu_{q(\log(\rho))} + (p-\gamma_\bullet)\mu_{q(\log(\rho))} + \tfrac{1}{2}\mu_{q(1/\sigma_\varepsilon^2)}^2 \boldsymbol{y}^T \boldsymbol{X}_{\boldsymbol{\gamma}} \boldsymbol{V}_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})} \boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{y} \Big\}$$

$$\mu_{q(\omega_{\boldsymbol{\gamma}})} \leftarrow \frac{\eta_{\boldsymbol{\gamma}}}{\sum_{\boldsymbol{\gamma}\in\mathbb{B}} \eta_{\boldsymbol{\gamma}}}$$

$$\mu_{q(\gamma_j)} \leftarrow \sum_{\boldsymbol{\gamma} \text{ if } \gamma_j=1} \mu_{q(\omega_{\boldsymbol{\gamma}})}$$

$$\mu_{q(\log(\rho))} \leftarrow \psi(1+\sum_{j=1}^{p} \mu_{q(\gamma_j)}) - \psi(2+p)$$

$$\mu_{q(\log(1-\rho))} \leftarrow \psi(1-\sum_{j=1}^{p} \mu_{q(\gamma_j)}) - \psi(2+p)$$

$$A_{q(\sigma_\varepsilon^2)} \leftarrow \frac{n}{2} + A_\varepsilon$$

$$B_{q(\sigma_\varepsilon^2)} \leftarrow B_\varepsilon + \tfrac{1}{2} \sum_{\boldsymbol{\gamma}\in\mathbb{B}} \mu_{q(\omega_{\boldsymbol{\gamma}})} \left\{ \| \boldsymbol{y} - \boldsymbol{X}_{\boldsymbol{\gamma}} \boldsymbol{m}_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})} \|^2 + \text{tr}(\boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X}_{\boldsymbol{\gamma}} \boldsymbol{V}_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})}) \right\}$$

$$\mu_{q(1/\sigma_\varepsilon^2)} \leftarrow \frac{A_{q(\sigma_\varepsilon^2)}}{B_{q(\sigma_\varepsilon^2)}}$$

$$\mu_{q(\boldsymbol{\theta})} \leftarrow \sum_{\boldsymbol{\gamma}\in\mathbb{B}} \mu_{q(\omega_{\boldsymbol{\gamma}})} \boldsymbol{m}_{\boldsymbol{\gamma}q(\boldsymbol{\theta})}$$

$$\Sigma_{q(\boldsymbol{\theta})} \leftarrow \sum_{\boldsymbol{\gamma}\in\mathbb{B}} \mu_{q(\omega_{\boldsymbol{\gamma}})} \boldsymbol{V}_{\boldsymbol{\gamma}q(\boldsymbol{\theta})} - \mu_{q(\boldsymbol{\theta})} \mu_{q(\boldsymbol{\theta})}^T$$

$$A_{q(\sigma_\beta^2)} \leftarrow \frac{p}{2} + A_\beta$$

$$B_{q(\sigma_\beta^2)} \leftarrow B_\beta + \tfrac{1}{2} \left\{ \| \mu_{q(\boldsymbol{\theta})} \|^2 + \text{tr}(\Sigma_{q(\boldsymbol{\theta})}) \right\}$$

$$\mu_{q(1/\sigma_\beta^2)} \leftarrow \frac{A_{q(\sigma_\beta^2)}}{B_{q(\sigma_\beta^2)}}$$

until the increase in $\log \underline{p}(y;q)$ is negligible.

good. Figure 3.3 shows the approximate posterior density functions of the model parameters obtained from both the MFVB inference and the MCMC approach. We consider a predictor to be selected when the posterior probability of its indicator variable γ is bigger than 0.5. We see that the correspondence of the density functions for the parameters is very good. The lower bound on the marginal log-likelihood $\underline{p}(y;q)$ in the upper right panel of Figure 3.3 illustrates that the speed of convergence of the MFVB inference is very fast.

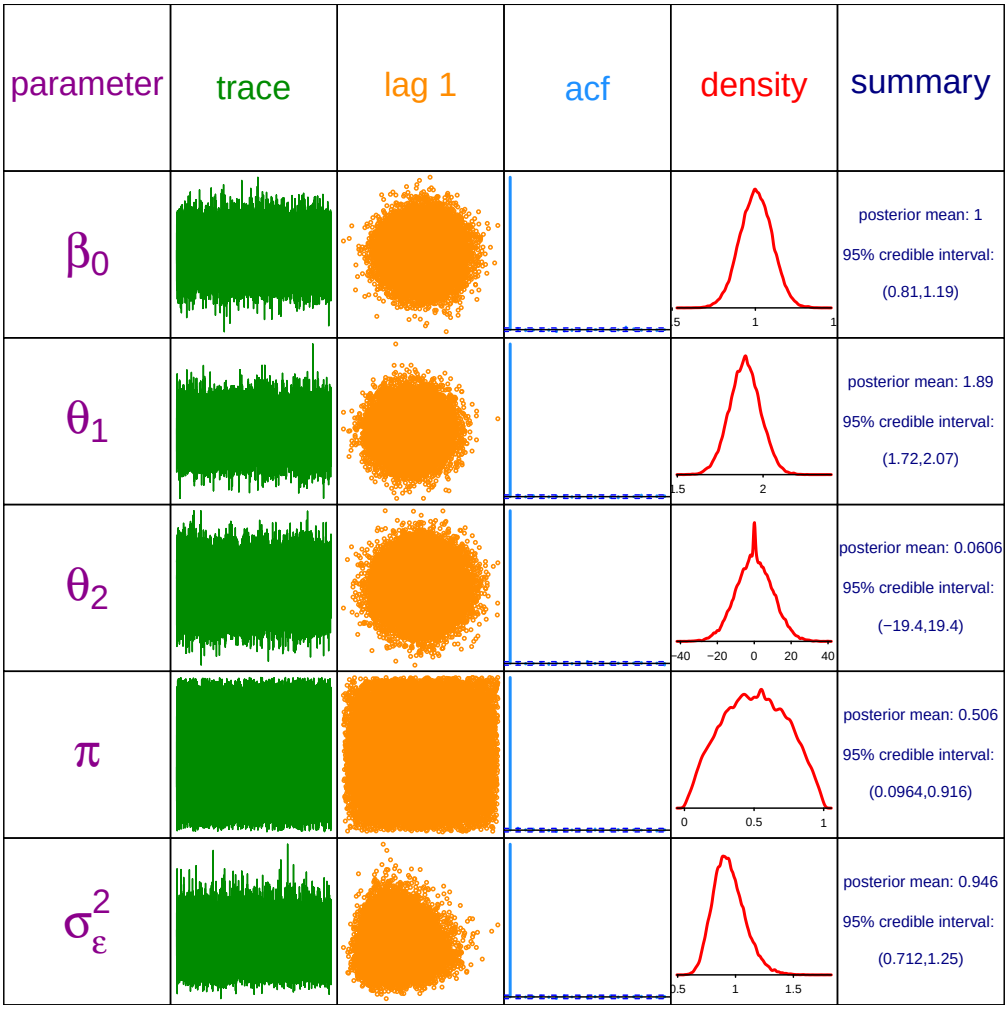

Figure 3.2: Summary of MCMC inference for linear indicator variable selection model (3.2).

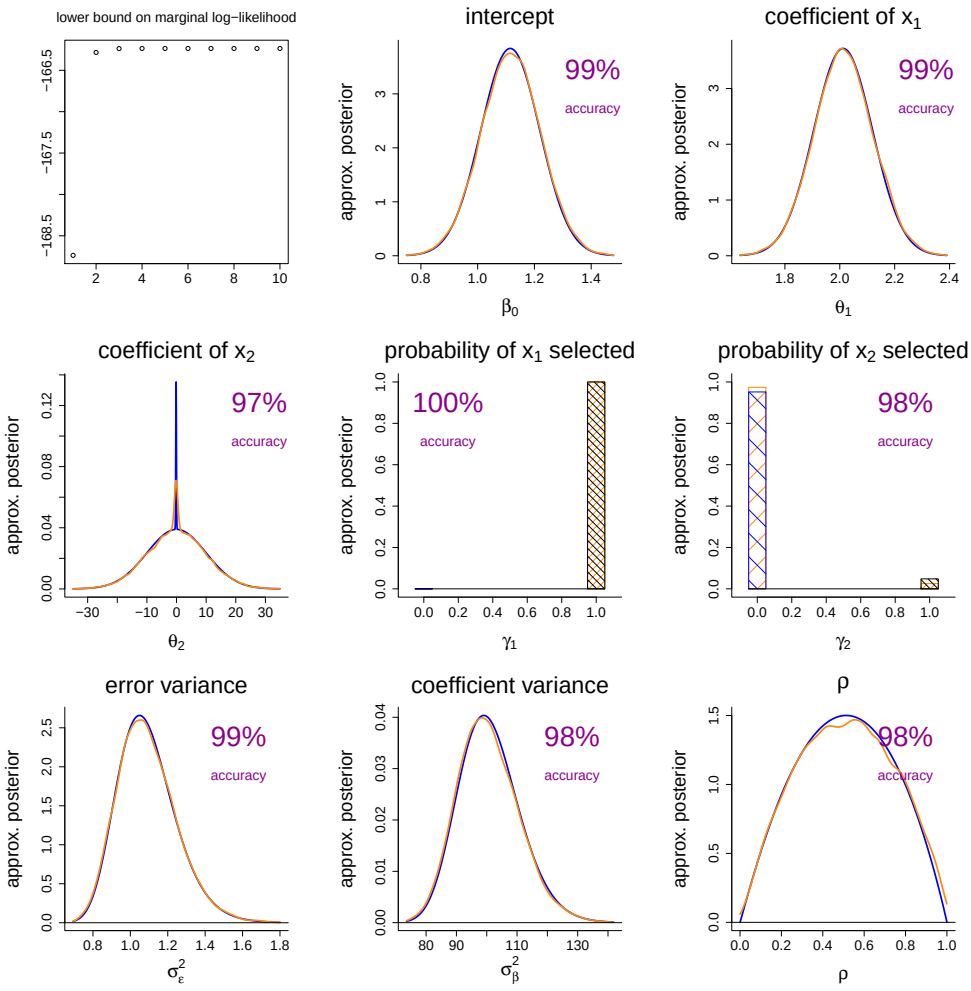

Figure 3.3: Upper right panel: successive values of lower bound on marginal log-likelihood to monitor convergence of the MFVB algorithm. Other panels: MFVB (blue) and MCMC (orange) approximate posterior densities for fitting (3.2) to a simulation data. The percentage are the accuracies of the MFVB fit compared with the MCMC fit.

Cheese data

In a study of cheddar cheese from the La Trobe Valley of Victoria, Australia, samples of cheese were analyzed to determine the amounts of lactic acid, acetic acid and hydrogen sulfide they contained. Then, overall taste scores for each cheese were obtained by combining the scores from several tasters. The goal was

to predict the taste score (**taste**) based on the lactic acid (**lactic**), acetic acid (**acetic**) and hydrogen sulfide (**H2S**) contents.

Figure 3.4 shows the approximate posterior density function of the model parameters for fitting model (3.2) to the cheese data. There is good agreement between the results of the MFVB inference and those of the MCMC approach. The posterior probability of γ_{H2S} and γ_{lactic} strongly suggest that both predictors (**lactic** and **H2S**) should be selected. The predictor **acetic** will also be included in the model for predicting the taste score, because the posterior probability of γ_{acetic} is slightly larger than 0.5. Moreover, it can be seen that the approximate posterior marginal density functions of the intercept, and acetic and lactic predictors, are slightly skewed (i.e., they do not have Gaussian distributions). The approximate posterior marginal density produced by the MFVB inference can handle this skewed distribution, because the marginal posterior distribution for β_0 and $\boldsymbol{\theta}$ is a mixture of multivariate Gaussian distributions, which can be used to fit any continuous density function approximatively.

Prostate cancer data

The prostate cancer data come from a study that examined the correlation between the level of prostate specific antigen and a number of clinical measures in men who were about to receive a radical prostatectomy. It is a data frame with 97 rows and 9 columns (Stamey *et al.*, 1989). This data set can be obtained in the R package `lasso2` (Lokhorst, Venables and Turlach, 2014). The data has the following components: log(cancer volume) (**lcavol**), log(prostate weight) (**lweight**), age (**age**), log(benign prostatic hyperplasia amount) (**lbph**), seminal vesicle invasion (**svi**), log(capsular penetration) (**lcp**), Gleason score (**gleason**), percentage Gleason scores 4 or 5 (**pgg45**) and log(prostate specific antigen) (**lpsa**). The response variable is **lcavol**, and the predictors include **lweight**, **age**, **lbph**, **svi**, **lcp**,

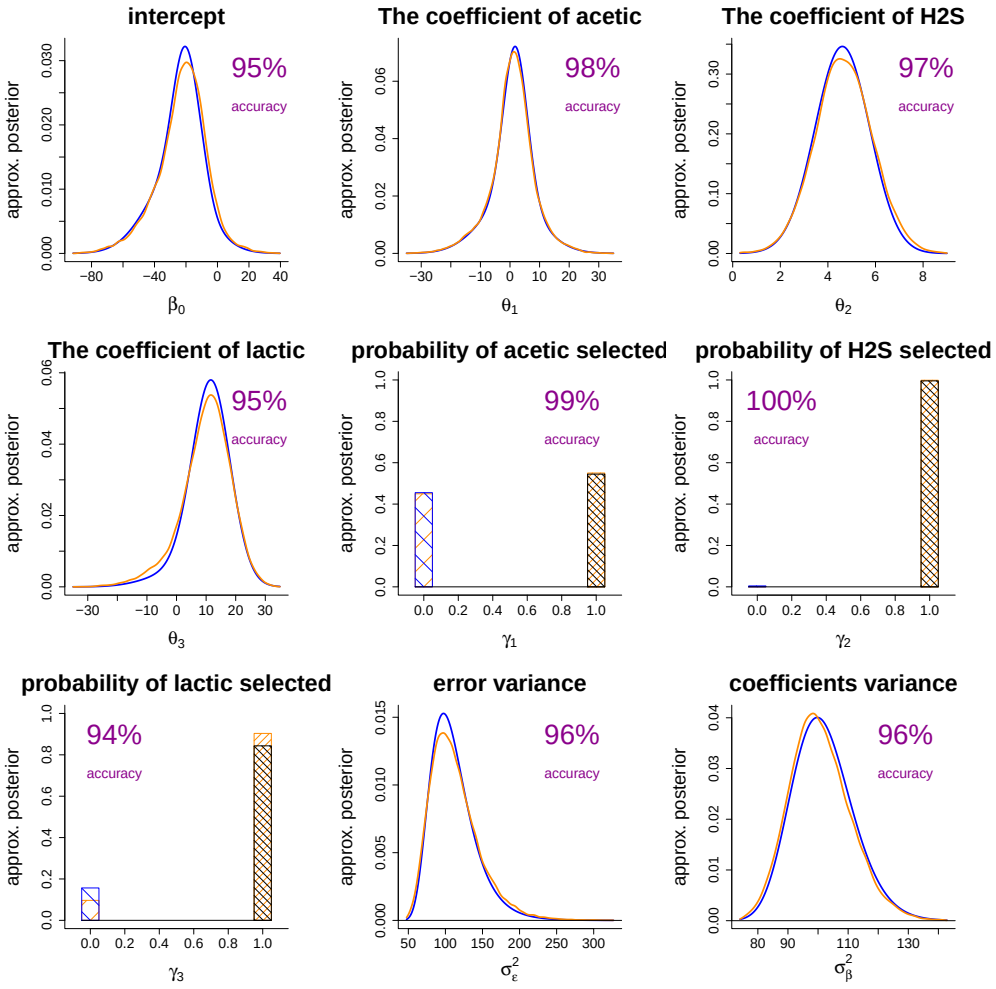

Figure 3.4: MFVB (blue) and MCMC (orange) approximate posterior density functions of model parameters in fitting (3.2) to the cheese data. The percentages are the accuracies of the MFVB fit compared with the MCMC fit.

gleason, **pgg45** and **lpsa**.

Figures 3.5 and 3.6 present the approximate posterior density function of each coefficient and the approximate probability of selecting each variable. There is good agreement between the results from the MFVB inference and the MCMC approach. Figure 3.6 suggests that the variables **svi** and **pgg45** should be selected.

These three examples show that the accuracy of the MFVB approximate pos-

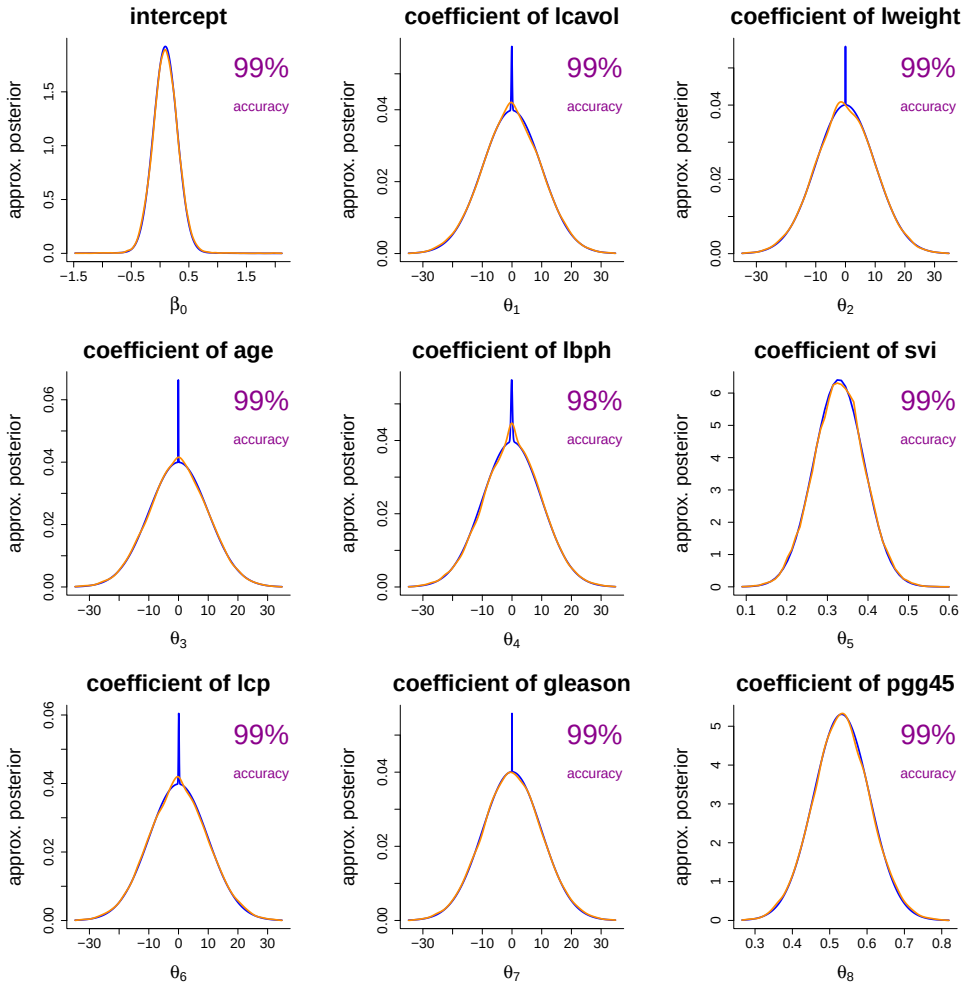

Figure 3.5: MFVB (blue) and MCMC (orange) approximate posterior density functions of coefficients in fitting (3.2) to prostate cancer data. The percentages are the accuracies of the MFVB fit compared with the MCMC fit.

terior density function is very high. Next, a simulation is carried out to evaluate the performance of variable selection.

3.2.4 Results for model selection

In Chapter 2, a series of simulations with different correlation between candidate predictor and different signal-to-noise ratio (Hastie *et al.*, 2009, page 649) were

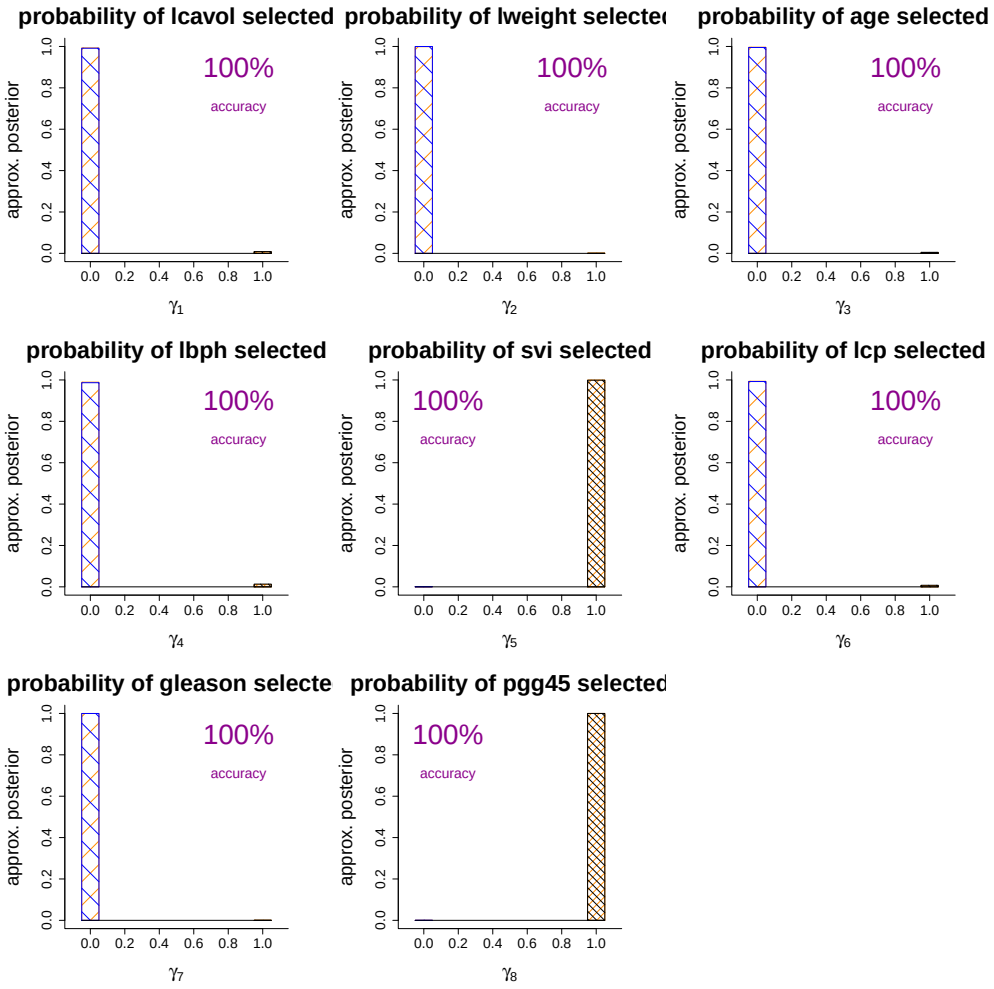

Figure 3.6: MFVB (blue) and MCMC (orange) approximate posterior probabilities of selecting each variable in fitting (3.2) to prostate cancer data. The percentages are the accuracies of the MFVB fit compared with the MCMC fit.

used to evaluate the performance of linear and non-linear model selection method. In this chapter, we also use simulation with similar settings to evaluate the performance of the MFVB indicator variable selection. We consider the linear model form

$$y_i = \boldsymbol{X}_i\boldsymbol{\beta} + \varepsilon_i, \quad 1 \leq i \leq n.$$

The number of observations is $n = 100$, and the number of predictors of interest is p=10. The $\boldsymbol{X}_i$, $1 \leq i \leq n$, are generated from a Multivariate Gaussian distribution, $N(\mathbf{0}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = . \begin{pmatrix} 1 & \rho_x & \cdots & \rho_x \\ \rho_x & 1 & \cdots & \rho_x \\ \vdots & \vdots & \ddots & \vdots \\ \rho_x & \rho_x & \cdots & 1 \end{pmatrix}$$

Then, $\rho_x = 0$, 0.2, 0.5, 0.8 correspond to non correlation, low correlation, medium correlation and high correlation. The ε_i, $1 \leq i \leq n$, are generated from a Gaussian distribution $N(0, \sigma^2)$. Following Hastie *et al.* (2009), the standard deviation, σ, was chosen in each case so that the signal-to-noise ratio (SNR) is equal to a fixed value. We also set the SNR equal to 1, 5 and 25 to represent low, medium and high values. The true value of $\boldsymbol{\beta}$ is $(2, 2, 0, 0, 0, -2, -2, 0, 0, 0)^T$. We consider a predictor to be selected when the posterior probability of its indicator variable γ is bigger than 0.5. Table 3.1 shows the simulation results of the marginal probabilities that variables are selected for various ρ_x and SNR for a simulation size of 200. The accuracy results for the variable selection show that model (3.2) and the corresponding MFVB inference algorithm (3.2.1) can perform the linear variable selection effectively.

SNR	ρ	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1	0	0.96	0.97	0	0	0.01	0.99	0.91	0	0.02	0
5	0	1	1	0	0	0.01	1	1	0	0.01	0
25	0	1	1	0.01	0	0	1	1	0	0	0
1	0.2	0.99	1	0.01	0.03	0	1	1	0	0	0
5	0.2	1	1	0.01	0	0	1	1	0.01	0.01	0
25	0.2	1	1	0.01	0	0	1	1	0	0	0
1	0.5	0.99	1	0	0.04	0.01	1	1	0.01	0	0.02
5	0.5	1	1	0	0	0	1	1	0	0	0
25	0.5	1	1	0	0	0	1	1	0.01	0	0
1	0.8	1	0.99	0.01	0.01	0.01	1	0.98	0.01	0	0
5	0.8	1	1	0.01	0	0	1	1	0	0.01	0
25	0.8	1	1	0.01	0	0.01	1	1	0	0.01	0
True	coef.	2	2	0	0	0	-2	-2	0	0	0

Table 3.1: Marginal probabilities that variables are selected.

3.3 Non-Gaussian Response Linear Models

In this section we extend the Gaussian response linear model to binary response models.

3.3.1 Models

The indicator variable selection model for a binary response is:

$$
\begin{aligned}
& y_i|\beta_0, \boldsymbol{\beta} \overset{\text{ind.}}{\sim} \text{Bernoulli}\{\Phi((\mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta})_i)\}, \\
& \beta_0 \sim N(0, \sigma^2_{\beta_0}), \\
& p(\beta_j|\sigma^2_\beta, \gamma_j) = \gamma_j N(0, \sigma^2_\beta) + (1-\gamma_j)\delta(0),\ 1 \le j \le p, \\
& \sigma^2_\beta \sim \text{Inverse-Gamma}(A_\beta, B_\beta), \\
& \gamma_j|\rho \overset{\text{ind.}}{\sim} \text{Bernoulli}(\rho), \\
& \rho \sim \text{Uniform}(0,1).
\end{aligned}
\tag{3.6}
$$

We introduce the same auxiliary variables $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_p)^T$ as in Result 1.15, such that:

$$\theta_j | \sigma_\beta^2 \overset{\text{ind.}}{\sim} N(0, \sigma_\beta^2),$$

and a vector of auxiliary variables $\boldsymbol{a} = (a_1, a_2, ..., a_n)^T$ as in Result 1.14, such that:

$$\boldsymbol{a} \sim N(\mathbf{1}\beta_0 + \boldsymbol{X}(\boldsymbol{\gamma} \odot \boldsymbol{\theta}), \boldsymbol{I}).$$

The full model with auxiliary variables is then:

$$\begin{aligned}
&y_i | a_i \overset{\text{ind.}}{\sim} \text{Bernoulli}\{\boldsymbol{I}(a_i \geq 0)\}, \\
&\boldsymbol{a} \sim N(\mathbf{1}\beta_0 + \boldsymbol{X}(\boldsymbol{\gamma} \odot \boldsymbol{\theta}), \boldsymbol{I}), \\
&\beta_0 \sim N(0, \sigma_{\beta_0}^2), \\
&\theta_j | \sigma_\beta^2 \overset{\text{ind.}}{\sim} N(0, \sigma_\beta^2), \\
&\sigma_\beta^2 \sim \text{Inverse-Gamma}(A_\beta, B_\beta), \\
&\gamma_j | \rho \overset{\text{ind.}}{\sim} \text{Bernoulli}(\rho), \\
&\rho \sim \text{Uniform}(0, 1),
\end{aligned} \tag{3.7}$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, ..., \gamma_p)^T$ is the vector of indicator variables, and $A_\beta, B_\beta, \sigma_{\beta_0}^2 > 0$ are constants hyperparameters. Figure 3.7 shows the directed acyclic graph corresponding to (3.7).

The full conditional distributions for the Markov chain Monte Carlo can be

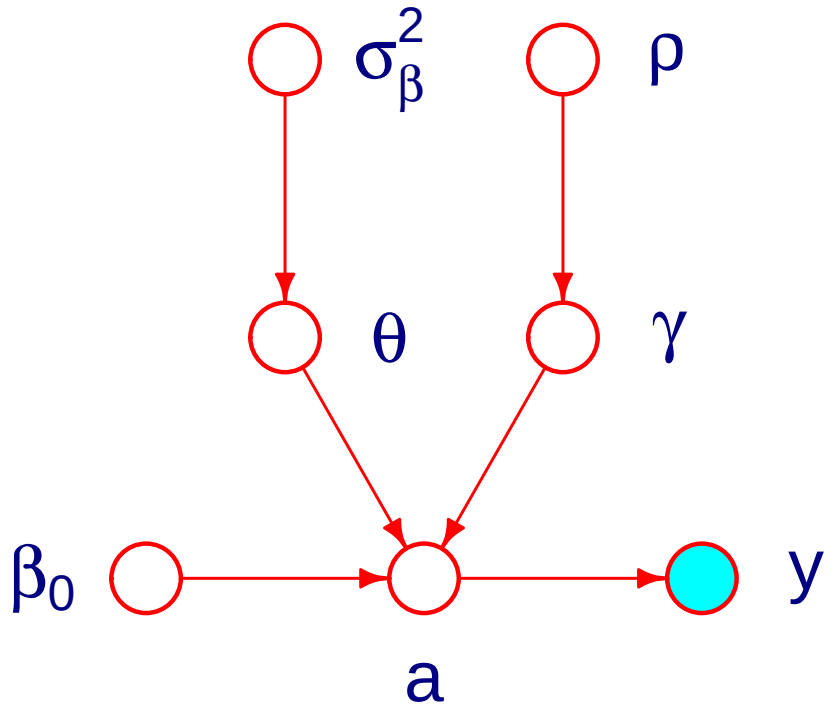

Figure 3.7: Directed acyclic graph for model (3.7).

shown to be:

$$p(\beta_0, \boldsymbol{\theta}, \boldsymbol{\gamma}|\text{rest}) \sim \phi_{\boldsymbol{I}_n}(\boldsymbol{a} - \boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\boldsymbol{\theta}})\phi_F(\tilde{\boldsymbol{\theta}})\rho^{\gamma_\bullet}(1-\rho)^{p-\gamma_\bullet},$$

$$\rho|\text{rest} \sim \text{Beta}(1+\gamma_\bullet, 1+p-\gamma_\bullet),$$

$$\sigma_\beta^2|\text{rest} \sim \text{IG}(A_\beta + \frac{p}{2}, B_\beta + \frac{1}{2} \parallel \boldsymbol{\theta} \parallel^2),$$

$$a_i|\text{rest} \overset{\text{ind.}}{\sim} (\boldsymbol{I}(a_i \geq 0))^{y_i}(\boldsymbol{I}(a_i < 0))^{1-y_i} N(\beta_0 + (\boldsymbol{X}(\boldsymbol{\gamma} \odot \boldsymbol{\theta}))_i, 1).$$

Here

$$\tilde{\boldsymbol{\theta}} = (\beta_0, \theta_1, ..., \theta_p)^T,\ \gamma_\bullet = \sum_{j=1}^{p} \gamma_j,$$

$$\boldsymbol{X}_{\boldsymbol{\gamma}} = [\mathbf{1}, \boldsymbol{X}]\text{diag}(1, \gamma_1, ..., \gamma_p) \text{ and } \boldsymbol{F} = \text{diag}(\sigma^2_{\beta_0}, \sigma^2_\beta, ..., \sigma^2_\beta).$$

3.3.2 Mean field variational Bayes scheme

We now seek a quick deterministic approximate inference procedure for (3.7) based on the MFVB. A tractable solution arises if we impose the product restriction:

$$q(\beta_0, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{a}, \sigma^2_\beta, \rho) = q(\beta_0, \boldsymbol{\theta}, \boldsymbol{\gamma})q(\boldsymbol{a}, \sigma^2_\beta, \rho).$$

The induced factorizations theory (e.g., Bishop, 2006, Section 10.2.5) leads to the solution having the additional product structure:

$$q(\beta_0, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{a}, \sigma^2_\beta, \rho) = q(\beta_0, \boldsymbol{\theta}, \boldsymbol{\gamma})q(\boldsymbol{a})q(\sigma^2_\beta)q(\rho). \tag{3.8}$$

Then, as shown in Appendix 3.B:

$$\begin{aligned}
&q^*(\rho) \text{ is a Beta density function,}\\
&q^*(a_i) \text{ is a truncated Gaussian density function,}\\
&q^*(\sigma^2_\beta) \text{ is an Inverse Gamma density functions,}\\
&q^*(\beta_0, \boldsymbol{\theta}|\boldsymbol{\gamma}) \text{ is a multivariate Gaussian density function,}\\
&q^*(\beta_0, \boldsymbol{\theta}) \text{ is a mixture of multivariate Gaussian density functions, and}\\
&q^*(\gamma_j), 1 \leq j \leq p, \text{ are Bernoulli probability mass functions.}
\end{aligned} \tag{3.9}$$

Similarly to the Gaussian response case, let $\boldsymbol{m}_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})}$ and $\boldsymbol{V}_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})}$ denote the mean vector and covariance matrix for the conditional multivariate Gaussian density function $q^*(\beta_0, \boldsymbol{\theta}|\boldsymbol{\gamma})$, and $A_{q(\sigma^2_\beta)}$ and $B_{q(\sigma^2_\beta)}$ denote the shape and rate parameters

for $q^*(\sigma_\beta^2)$. The $\mu_{q(\omega_{\boldsymbol{\gamma}})}$ denote the weight parameters for $q^*(\beta_0, \boldsymbol{\theta})$, which is formed as

$$q^*(\beta_0, \boldsymbol{\theta}) = \sum_{\boldsymbol{\gamma} \in \mathbb{B}} \mu_{q(\omega_{\boldsymbol{\gamma}})} \phi_{\boldsymbol{V}_{\boldsymbol{\gamma} q(\beta_0, \boldsymbol{\theta})}} \left(\begin{bmatrix} \beta_0 \\ \boldsymbol{\theta} \end{bmatrix} - \boldsymbol{m}_{\boldsymbol{\gamma} q(\beta_0, \boldsymbol{\theta})} \right).$$

The convergence of algorithm (3.3.1) can be monitored using the following expression for the lower bound on the marginal log-likelihood:

$$\begin{aligned}
\log \underline{p}(y; q) \;=\; & \frac{1+p}{2} - \frac{1}{2}\log(\sigma_{\beta_0}^2) \\
& + A_\beta \log B_\beta - \log\,\Gamma(A_\beta) \\
& - (A_\beta + \frac{p}{2})\log(B_{q(\sigma_\beta)}) + \log\,\Gamma(A_\beta + \frac{p}{2}) \\
& + y^T \log[\Phi(\sum_{\boldsymbol{\gamma} \in \mathbb{B}} \mu_{q(\omega_{\boldsymbol{\gamma}})} \boldsymbol{X}_{\boldsymbol{\gamma}} \boldsymbol{m}_{\boldsymbol{\gamma} q(\beta_0, \boldsymbol{\theta})})] \\
& + (\mathbf{1} - y)^T \log[\mathbf{1} - \Phi(\sum_{\boldsymbol{\gamma} \in \mathbb{B}} \mu_{q(\omega_{\boldsymbol{\gamma}})} \boldsymbol{X}_{\boldsymbol{\gamma}} \boldsymbol{m}_{\boldsymbol{\gamma} q(\beta_0, \boldsymbol{\theta})})] \\
& - \frac{1}{2} \sum_{\boldsymbol{\gamma} \in \mathbb{B}} \mu_{q(\omega_{\boldsymbol{\gamma}})} \left\{ \mathrm{tr}(\boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X}_{\boldsymbol{\gamma}} V_{q(\beta_0, \boldsymbol{\theta})}) - \log|\boldsymbol{V}_{\boldsymbol{\gamma} q(\beta_0, \boldsymbol{\theta})}| \right\} \\
& - \frac{1}{2} \sum_{\boldsymbol{\gamma} \in \mathbb{B}} \mu_{q(\omega_{\boldsymbol{\gamma}})} \left\{ \frac{\mu_{\boldsymbol{\gamma} q(\beta_0)}^2 + \sigma_{\boldsymbol{\gamma} q(\beta_0)}^2}{\sigma_{\beta_0}^2} + 2\log \mu_{q(\omega_{\boldsymbol{\gamma}})} \right\}.
\end{aligned}$$

3.3.3 Accuracy assessment

We now investigate the accuracy of a mean field variational Bayes approximation inference scheme in this context. Firstly, I will give one example of fitting model (3.7) to simulated data. After the example, a simulation will be carried out to compare the accuracies of the MFVB inference and the MCMC inference.

Algorithm 3.3.1: Iterative scheme for obtaining the parameters in the optimal densities $q^*(\beta_0, \boldsymbol{\theta}|\boldsymbol{\gamma})$, $q^*(\beta_0, \boldsymbol{\theta})$, $q^2(\boldsymbol{a})$, $q^*(\sigma_\beta^2)$, $q^*(\rho)$ and $q^*(\gamma_j)$ for the probit indicator variable selection model (3.7)

Initialize;
Cycle

$$\boldsymbol{V}_{\boldsymbol{\gamma} q(\beta_0,\boldsymbol{\theta})} \leftarrow \left\{ \boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X}_{\boldsymbol{\gamma}} + \text{blockdiag}\left[(\sigma_{\beta_0}^2)^{-1}, \mu_{q(1/\sigma_\beta^2)} \boldsymbol{I}_p \right] \right\}^{-1}$$

$$\boldsymbol{m}_{\boldsymbol{\gamma} q(\beta_0,\boldsymbol{\theta})} \leftarrow \boldsymbol{V}_{\boldsymbol{\gamma} q(\beta_0,\boldsymbol{\theta})} \boldsymbol{X}_{\boldsymbol{\gamma}}^T \mu_{q(\boldsymbol{a})}$$

$$\eta_{\boldsymbol{\gamma}} \leftarrow \exp\Big\{ \tfrac{1}{2}\log|\boldsymbol{V}_{\boldsymbol{\gamma} q(\beta_0,\boldsymbol{\theta})}| + \gamma_\bullet \mu_{q(\log(\rho))} + (p-\gamma_\bullet)\mu_{q(\log(\rho))} + \tfrac{1}{2}\mu_{q(a)}^T \boldsymbol{X}_{\boldsymbol{\gamma}} \boldsymbol{V}_{\boldsymbol{\gamma} q(\beta_0,\boldsymbol{\theta})} \boldsymbol{X}_{\boldsymbol{\gamma}}^T \mu_{q(a)} \Big\}$$

$$\mu_{q(\omega_{\boldsymbol{\gamma}})} \leftarrow \frac{\eta_{\boldsymbol{\gamma}}}{\sum_{\boldsymbol{\gamma}\in\mathbb{B}} \eta_{\boldsymbol{\gamma}}}$$

$$\mu_{q(\boldsymbol{\gamma}_j)} \leftarrow \sum_{\boldsymbol{\gamma} \text{ if } \gamma_j=1} \mu_{q(\omega_{\boldsymbol{\gamma}})}$$

$$\mu_{q(\log(\rho))} \leftarrow \psi(1 + \sum_{j=1}^{p} \mu_{q(\gamma_j)}) - \psi(2+p)$$

$$\mu_{q(\log(1-\rho))} \leftarrow \psi(1 - \sum_{j=1}^{p} \mu_{q(\gamma_j)}) - \psi(2+p)$$

$$\mu_{q(\boldsymbol{\theta})} \leftarrow \sum_{\boldsymbol{\gamma}\in\mathbb{B}} \mu_{q(\omega_{\boldsymbol{\gamma}})} \boldsymbol{m}_{\boldsymbol{\gamma} q(\boldsymbol{\theta})}$$

$$\Sigma_{q(\boldsymbol{\theta})} \leftarrow \sum_{\boldsymbol{\gamma}\in\mathbb{B}} \mu_{q(\omega_{\boldsymbol{\gamma}})} \boldsymbol{V}_{\boldsymbol{\gamma} q(\boldsymbol{\theta})} - \mu_{q(\boldsymbol{\theta})} \mu_{q(\boldsymbol{\theta})}^T$$

$$A_{q(\sigma_\beta^2)} \leftarrow \frac{p}{2} + A_\beta$$

$$B_{q(\sigma_\beta^2)} \leftarrow B_\beta + \tfrac{1}{2}\left\{ \| \mu_{q(\boldsymbol{\theta})} \|^2 + \text{tr}(\Sigma_{q(\boldsymbol{\theta})}) \right\}$$

$$\mu_{q(1/\sigma_\beta^2)} \leftarrow \frac{A_{q(\sigma_\beta^2)}}{B_{q(\sigma_\beta^2)}}$$

$$\mu_{q(\boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\theta})} \leftarrow \sum_{\boldsymbol{\gamma}\in\mathbb{B}} \mu_{q(\omega_{\boldsymbol{\gamma}})} \boldsymbol{X}_{\boldsymbol{\gamma}} \boldsymbol{m}_{\boldsymbol{\gamma} q(\beta_0,\boldsymbol{\theta})}$$

$$\mu_{q(\boldsymbol{a})} \leftarrow \mu_{q(\boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\theta})} + \frac{\phi(\mu_{q(\boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\theta})})}{\Phi(\mu_{q(\boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\theta})})^{\boldsymbol{y}} \odot [\mathbf{1} - \Phi(\mu_{q(\boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\theta})})]^{\mathbf{1}-\boldsymbol{y}}}$$

until the increase in $\log \underline{p}(y; q)$ is negligible.

Simulated data

The data were generated according to:

$$\text{logit}\{P(y=1)\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$$

where the sample size is $n = 100$ and the true values are $\beta_0 = -1$, $\beta_1 = 0$, $\beta_2 = 2$, $\beta_3 = 0$, $\beta_4 = -1$ and $\beta_5 = 0$. This means that x_2 and x_4 are in the model, and x_1, x_3 and x_5 are superfluous noise variables. The x_j, $1 \leq j \leq 5$, are independently generated from the standard Gaussian distribution. Figure 3.8 summarizes the MCMC results for the binary response linear indicator variable selection model (3.7) to fit using these simulated data. Columns two to four indicate that the MCMC convergence is quite good. The lower bound on the marginal log-likelihood $\underline{p}(y; q)$ in Figure 3.9 shows that the convergence of the MFVB inference is very fast. Figure 3.11 shows the approximate posterior density functions of the coefficients. The strong correlation between the coefficient variable $\boldsymbol{\theta}$ and the auxiliary variable $\boldsymbol{a}$ results in decreased accuracy. Figure 3.11 shows the selection probability for each variable obtained from both the MFVB inference and the MCMC inference. We see that the correspondence is very good.

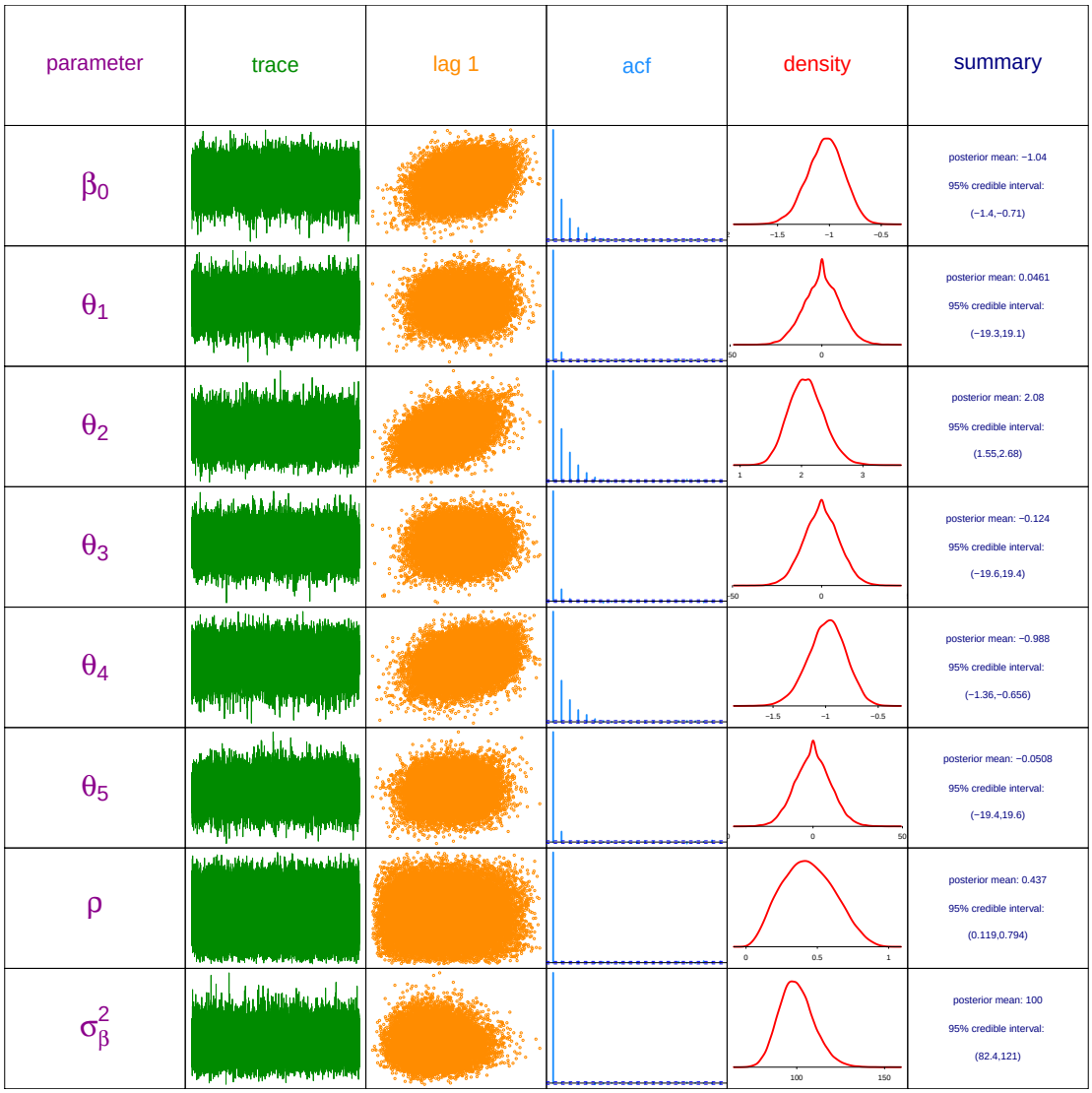

Figure 3.8: Summary of MCMC inference for linear indicator variable selection model (3.7).

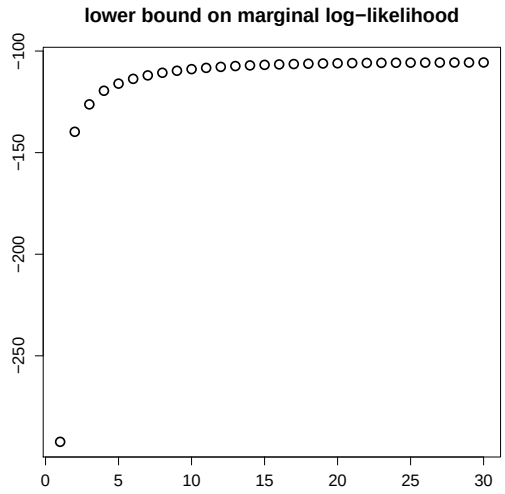

Figure 3.9: Successive values of the lower bound on marginal log-likelihood to monitor convergence of the MFVB algorithm for fit model (3.7).

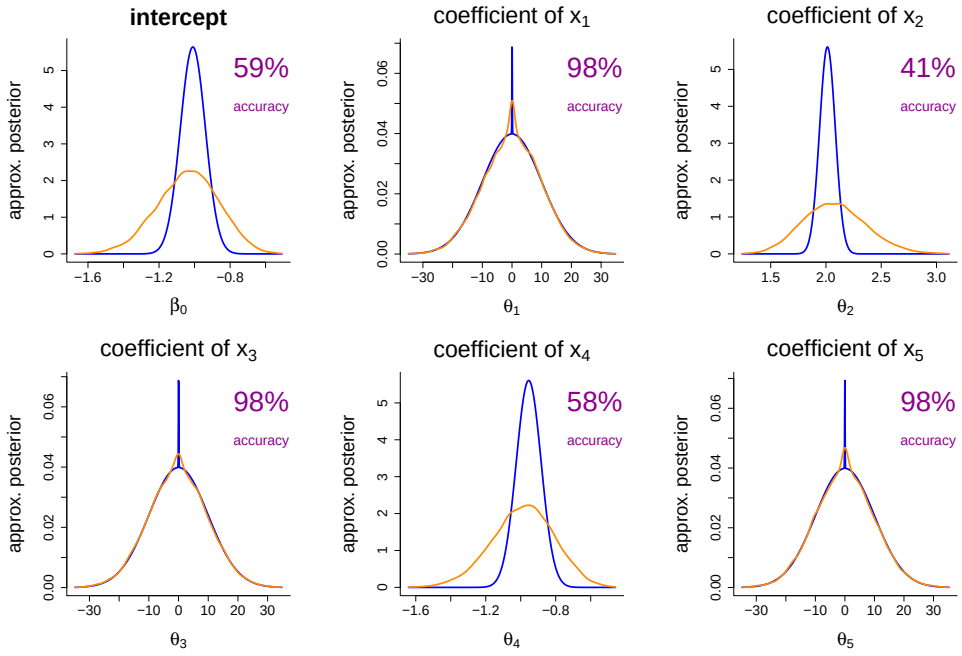

Figure 3.10: MFVB (blue) and MCMC (orange) approximate posterior density functions of coefficients obtained by fitting (3.7) to simulated data. The percentages are the accuracies of the MFVB fit compared with the MCMC fit.

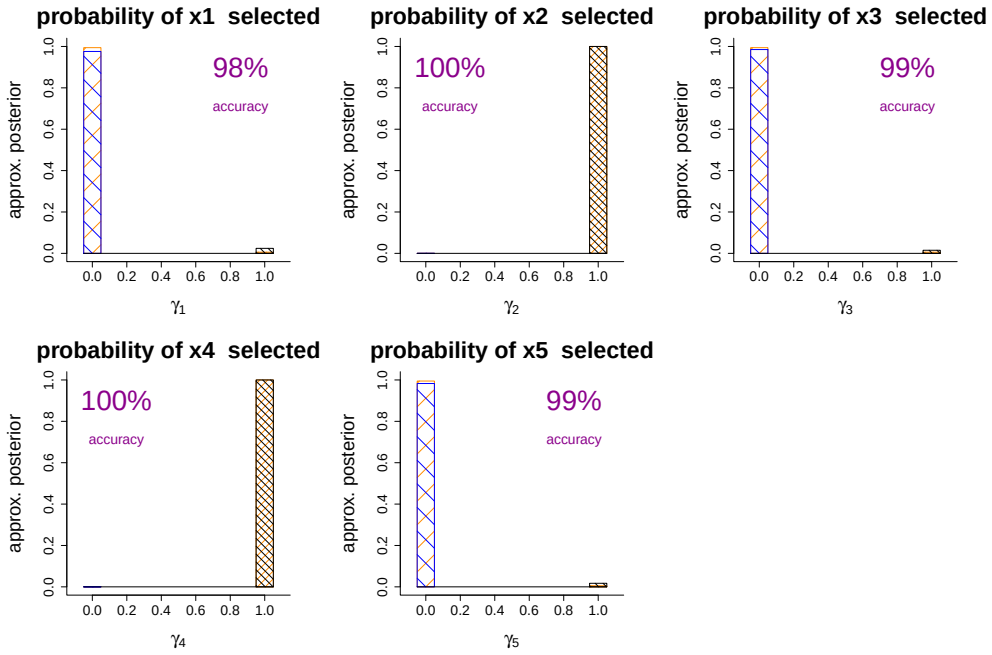

Figure 3.11: MFVB (blue) and MCMC (orange) approximate posterior probabilities of selection of each variable by using model (3.7) to simulate data. The percentages are the accuracies of the MFVB fit compared with the MCMC fit.

Simulation study for accuracy

The results using simulated data show that the approximate posterior density functions of the coefficients have insufficient accuracy. In this part, I carry out a simulation study to evaluate the accuracy of fits in fitting model (3.7) by using the MFVB inference. The data were also generated according to:

$$\text{logit}\{P(y=1)\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5,$$

where we set the sample size to be $n = 100$ and the true parameter values to be $\beta_0 = -1$, $\beta_1 = 0$, $\beta_2 = 2$, $\beta_3 = 3$, $\beta_4 = 0$ and $\beta_5 = 0$. The x_j, $1 \leq j \leq 5$, are independently generated from the standard Gaussian distribution. I simulate 100 times, and the accuracy of the fits is shown in Figure 3.12. In spite of low

accuracies in the posterior density functions of the coefficients, we can keep using this model to perform the variable selection because of the accurate results of the posterior probability of the indicator variable.

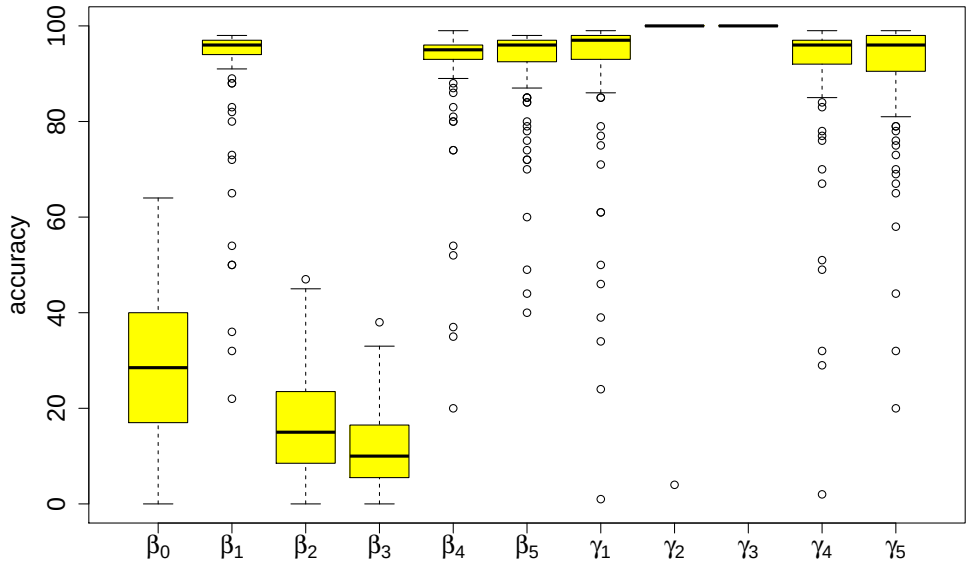

Figure 3.12: The boxplots of accuracy values for the Probit linear indicator model (3.7) study described in section 3.3.3.

3.3.4 Results for model selection

Similarly to the Gaussian response case, we consider the model:

$$y_i \overset{\text{ind.}}{\sim} \text{Bernoulli}\{\Phi(\boldsymbol{X}\boldsymbol{\beta})_i\}, 1 \leq i \leq n.$$

We set the number of observations $n = 100$, and the number of predictors of interest $p = 10$. The $\boldsymbol{X}_i$, $1 \leq i \leq n$, are generated from a multivariate normal

distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho_x & \cdots & \rho_x \\ \rho_x & 1 & \cdots & \rho_x \\ \vdots & \vdots & \ddots & \vdots \\ \rho_x & \rho_x & \cdots & 1 \end{pmatrix}$$

Then, $\rho_x = 0$, 0.2, 0.5, 0.8 correspond to no correlation, low correlation, medium correlation and high correlation. The true value of $\boldsymbol{\beta}$ is equal to $(0.5, 1, 1.5, 2, 2.5, 0, 0, 0, 0, 0)^T$. The results for a simulation size of 500 are in Table 3.2, which lists the mean posterior probability of each predictor's indicator variable. For the case of low correlation (i.e., $\rho_x = 0$ and 0.2), the variables x_3, x_4 and x_5 can be selected into the sample model, and the variable x_1 and x_2 will be classified into the noise variable and ignored. The obtained result is similar to that obtain by Hu and Johnson (2009), which used the MCMC for variable selection using similar models. The right side of model (3.7) and the corresponding MFVB algorithm (3.3.1) are such that the selected model has fewer false positive results. Unfortunately, the accuracy of the variable selection decreases further when the correlation among the predictors is stronger.

ρ	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
0	0.02	0.09	0.60	0.92	1	0.00	0.00	0	0.01	0.01
0.2	0.02	0.08	0.35	0.85	0.98	0.01	0.00	0.01	0.00	0.00
0.5	0.03	0.09	0.23	0.57	0.86	0.01	0.01	0.03	0.01	0.01
0.8	0.03	0.09	0.16	0.24	0.52	0.03	0.05	0.01	0.01	0.02
True coeff.	0.5	1	1.5	2	2.5	0	0	0	0	0

Table 3.2: Mean posterior probability of each predictor's indicator variable.

3.4 Discussion

In this chapter we used indicator variables to perform the linear model selection for the Gaussian response model and binary response model. The corresponding models (3.7) and (3.2) and MFVB Algorithms 3.3.1 and 3.2.1 were carried out. The accuracy of the MFVB inference and the MCMC inference and the variable selection results were presented and compared.

For the Gaussian response, the MFVB inference can perform model selection accurately and efficiently. However, using the indicator model to perform the model selection for a binary response model is challenging. The MFVB indicator selection procedure generated an eclectic model, in which there was a trade-off between selecting the correct variable and generating false positive results. The reason for this has two aspects: (1) as in the simulation in Figure 3.12, the accuracy of the MFVB inference for binary response model is lower than that for the Gaussian case; and (2), the MCMC results of Hu and Johnson (2009) showed that model (3.3.1) cannot perform the model selection faultlessly and that predictors with small coefficients are always ignored. Other prior distributions, such as the Laplace-zero (Johnstone & Silverman, 2005), can be considered instead of the Gaussian-Zero in future research.

3.A Appendix: Derivation of Algorithm 3.2.1

3.A.1 Full conditionals

Full conditional for $\beta_0, \boldsymbol{\theta}, \boldsymbol{\gamma}$

$$\begin{aligned}\log p(\beta_0, \boldsymbol{\theta}, \boldsymbol{\gamma}|\text{rest}) &= -\tfrac{1}{2}\frac{\|\boldsymbol{y} - \boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\boldsymbol{\theta}}\|^2}{\sigma_\varepsilon^2} - \frac{1}{2}\tilde{\boldsymbol{\theta}}^T \boldsymbol{F}^{-1}\tilde{\boldsymbol{\theta}} \\ &\quad +\gamma_\bullet \log(\rho) + (p - \gamma_\bullet)\log(1-\rho) + \text{const.}\end{aligned}$$

where

$$\tilde{\boldsymbol{\theta}} = (\beta_0, \theta_1, ..., \theta_p)^T,\ \gamma_\bullet = \sum_{j=1}^{p} \gamma_j,$$

$$\boldsymbol{X}_{\boldsymbol{\gamma}} = [\mathbf{1}, \boldsymbol{X}]\text{diag}(1, \gamma_1, ..., \gamma_p) \text{ and } \boldsymbol{F}^{-1} = \text{diag}(\sigma_{\beta_0}^{-2}, \sigma_\beta^{-2}, ..., \sigma_\beta^{-2}).$$

Derivation:

$$\begin{aligned}p(\beta_0, \boldsymbol{\theta}, \boldsymbol{\gamma}|\text{rest}) &\propto p(\boldsymbol{y}|\beta_0, \boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma_\varepsilon^2) p(\beta_0) \prod_{j=1}^{p} p(\theta_j|\sigma_\beta^2) p(\gamma_j|\rho) \\ &= (2\pi)^{-n/2}\sigma_\varepsilon^{-n}\exp\left\{-\frac{\|\boldsymbol{y} - \mathbf{1}\beta_0 - \boldsymbol{X}(\boldsymbol{\gamma}\odot\boldsymbol{\theta})\|^2}{2\sigma_\varepsilon^2}\right\} \\ &\quad \times (2\pi)^{-1/2}\sigma_{\beta_0}^{-1}\exp\left\{-\frac{\beta_0^2}{2\sigma_{\beta_0}^2}\right\} \\ &\quad \times \prod_{j=1}^{p}(2\pi)^{-1/2}\sigma_\beta^{-1}\exp\left\{-\frac{\theta_j^2}{2\sigma_\beta^2}\right\}\rho^{\gamma_j}(1-\rho)^{1-\gamma_j}.\end{aligned}$$

Taking logarithms, we get

$$\begin{aligned}\log p(\beta_0, \boldsymbol{\theta}, \boldsymbol{\gamma}|\text{rest}) &= -\frac{\|\boldsymbol{y} - \mathbf{1}\beta_0 - \boldsymbol{X}(\boldsymbol{\gamma} \odot \boldsymbol{\theta})\|^2}{2\sigma_\varepsilon^2} - \frac{\beta_0^2}{2\sigma_{\beta_0}^2} - \sum_{j=1}^{p} \frac{\theta_j^2}{2\sigma_\beta^2} \\ &\quad + \sum_{j=1}^{p} \gamma_j \log(\rho) + (1-\gamma_j)\log(1-\rho) + \text{const.} \\ &= -\frac{\|\boldsymbol{y} - \boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\boldsymbol{\theta}}\|^2}{2\sigma_\varepsilon^2} - \tfrac{1}{2}\tilde{\boldsymbol{\theta}}^T \boldsymbol{F}^{-1}\tilde{\boldsymbol{\theta}} \\ &\quad + \gamma_\bullet \log(\rho) + (p - \gamma_\bullet)\log(1-\rho) + \text{const.}\end{aligned}$$

Full conditional for ρ

$$\log p(\rho|\text{rest}) = \gamma_\bullet \log(\rho) + (p - \gamma_\bullet)\log(1-\rho).$$

Derivation:

$$\begin{aligned}p(\rho|\text{rest}) &\propto \prod_{j=1}^{p} p(\gamma_j|\rho)p(\rho) \\ &= \prod_{j=1}^{p} \rho^{\gamma_j}(1-\rho)^{1-\gamma_j}.\end{aligned}$$

Taking logarithms, we get:

$$\log p(\rho|\text{rest}) = \sum_{j=1}^{p} \gamma_j \log(\rho) + (1-\gamma_j)\log(1-\rho).$$

Full conditional for σ_ε^2

$$\log p(\sigma_\varepsilon^2|\text{rest}) = (-1 - A_\varepsilon - \frac{n}{2})\log(\sigma_\varepsilon^2) - \frac{2B_\varepsilon + \|\boldsymbol{y} - \boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\boldsymbol{\theta}}\|^2}{2\sigma_\varepsilon^2} + \text{const.}$$

Derivation:

$$\begin{aligned} p(\sigma_\varepsilon^2|\text{rest}) &\propto p(\boldsymbol{y}|\beta_0, \boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma_\varepsilon^2) p(\sigma_\varepsilon^2) \\ &= (2\pi)^{-n/2} \sigma_\varepsilon^{-n} \exp\left\{ -\frac{\|\boldsymbol{y} - \mathbf{1}\beta_0 - \boldsymbol{X}(\boldsymbol{\gamma} \odot \boldsymbol{\theta})\|^2}{2\sigma_\varepsilon^2} \right\} \\ &\quad \times \frac{B_\varepsilon^{A_\varepsilon}}{\Gamma(A_\varepsilon)} (\sigma_\varepsilon^2)^{-1-A_\varepsilon} \exp\left\{ \frac{-B_\varepsilon}{\sigma_\varepsilon^2} \right\}. \end{aligned}$$

Taking logarithms, we get:

$$\begin{aligned} \log p(\sigma_\varepsilon^2|\text{rest}) &= (-1 - A_\varepsilon - \frac{n}{2}) \log(\sigma_\varepsilon^2) \\ &\quad - \frac{2B_\varepsilon + \|\boldsymbol{y} - \mathbf{1}\beta_0 - \boldsymbol{X}(\boldsymbol{\gamma} \odot \boldsymbol{\theta})\|^2}{2\sigma_\varepsilon^2} + \text{const.} \end{aligned}$$

Full conditional for σ_β^2

$$\log p(\sigma_\beta^2|\text{rest}) = (-1 - A_\beta - \frac{p}{2}) \log(\sigma_\beta^2) - \frac{2B_\beta + \|\boldsymbol{\theta}\|^2}{2\sigma_\beta^2} + \text{const.}$$

Derivation:

$$\begin{aligned} p(\sigma_\beta^2|\text{rest}) &\propto p(\boldsymbol{\theta}|\sigma_\beta^2) p(\sigma_\beta^2) \\ &= (2\pi)^{-p/2} \sigma_\varepsilon^{-p} \exp\left\{ -\frac{\|\boldsymbol{\theta}\|^2}{2\sigma_\beta^2} \right\} \times \frac{B_\beta^{A_\beta}}{\Gamma(A_\beta)} (\sigma_\beta^2)^{-1-A_\beta} \exp\left\{ \frac{-B_\beta}{\sigma_\beta^2} \right\}. \end{aligned}$$

Taking logarithms, we get:

$$\log p(\sigma_\beta^2|\text{rest}) = (-1 - A_\beta - \frac{p}{2}) \log(\sigma_\beta^2) - \frac{2B_\beta + \|\boldsymbol{\theta}\|^2}{2\sigma_\beta^2} + \text{const.}$$

3.A.2 Optimal q^* densities

Expressions for $q^*(\beta_0, \boldsymbol{\beta})$

$$q^*(\beta_0, \boldsymbol{\theta}, \boldsymbol{\gamma}) = \mu_{q(\omega_{\boldsymbol{\gamma}})} \phi_{\boldsymbol{V}_{\boldsymbol{\gamma} q(\beta_0, \boldsymbol{\theta})}} (\tilde{\boldsymbol{\theta}} - \boldsymbol{m}_{\boldsymbol{\gamma} q(\beta_0, \boldsymbol{\theta})}),$$

where

$$\begin{aligned}
\boldsymbol{V}_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})} &= \left\{\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}}\mu_{q(1/\sigma_\varepsilon^2)} + \text{blockdiag}\left[(\sigma_{\beta_0}^2)^{-1}, \mu_{q(1/\sigma_\beta^2)}\boldsymbol{I}_p\right]\right\}^{-1}, \\
\boldsymbol{m}_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})} &= V_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})}\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{y}\mu_{q(1/\sigma_\varepsilon^2)}, \\
\eta_{\boldsymbol{\gamma}} &= \exp\Big\{\tfrac{1}{2}\log|V_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})}| + \gamma_\bullet\mu_{q(\log(\rho))} + (p-\gamma_\bullet)\mu_{q(\log(\rho))} \\
&\quad + \tfrac{1}{2}\mu_{q(1/\sigma_\varepsilon^2)}^2\boldsymbol{y}^T\boldsymbol{X}_{\boldsymbol{\gamma}}V_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})}\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{y}\Big\}
\end{aligned}$$

and

$$\mu_{q(\omega_{\boldsymbol{\gamma}})} = \frac{\eta_{\boldsymbol{\gamma}}}{\sum_{\boldsymbol{\gamma}\in\mathbb{B}}\eta_{\boldsymbol{\gamma}}}.$$

Derivation:

$$\begin{aligned}
\log q^*(\beta_0,\boldsymbol{\theta},\boldsymbol{\gamma}) &= E_q\left[\log p(\beta_0,\boldsymbol{\theta},\boldsymbol{\gamma}|\text{rest})\right] \\
&= E_q[-\tfrac{1}{2}\frac{\|\boldsymbol{y}-\boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\boldsymbol{\theta}}\|^2}{\sigma_\varepsilon^2} - \frac{1}{2}\tilde{\boldsymbol{\theta}}^T\boldsymbol{F}^{-1}\tilde{\boldsymbol{\theta}} \\
&\quad +\gamma_\bullet\log(\rho) + (p-\gamma_\bullet)\log(1-\rho) + \text{const}] \\
&= -\tfrac{1}{2}\mu_{q(1/\sigma_\varepsilon^2)}\|\boldsymbol{y}-\boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\boldsymbol{\theta}}\|^2 - \frac{1}{2}\tilde{\boldsymbol{\theta}}^T\boldsymbol{F}_q^{-1}\tilde{\boldsymbol{\theta}} \\
&\quad +\gamma_\bullet\mu_{q(\log(\rho))} + (p-\gamma_\bullet)\mu_{q(\log(1-\rho))},
\end{aligned}$$

where

$$\boldsymbol{F}_q^{-1} = \begin{bmatrix} 1/\sigma_{\beta_0}^2 & \boldsymbol{0} \\ \boldsymbol{0} & \mu_{q(1/\sigma_\beta^2)}\boldsymbol{I}_p \end{bmatrix}.$$

Note:

$$\begin{aligned}(\boldsymbol{y}-\boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\boldsymbol{\theta}})^T(1/\mu_{q(1/\sigma_\varepsilon^2)}\boldsymbol{I})^{-1}(\boldsymbol{y}-\boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\boldsymbol{\theta}})+\tilde{\boldsymbol{\theta}}^T\boldsymbol{F}_q^{-1}\tilde{\boldsymbol{\theta}}&\\ =\left\{\tilde{\boldsymbol{\theta}}-\left[\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}}\mu_{q(1/\sigma_\varepsilon^2)}+\boldsymbol{F}_q^{-1}\right]^{-1}\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{y}\mu_{q(1/\sigma_\varepsilon^2)}\right\}^T&\\ \times\left\{\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}}\mu_{q(1/\sigma_\varepsilon^2)}+\boldsymbol{F}_q^{-1}\right\}&\\ \times\left\{\tilde{\boldsymbol{\theta}}-\left[\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}}\mu_{q(1/\sigma_\varepsilon^2)}+\boldsymbol{F}_q^{-1}\right]^{-1}\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{y}\mu_{q(1/\sigma_\varepsilon^2)}\right\}&\\ -\mu_{q(1/\sigma_\varepsilon^2)}^2\boldsymbol{y}^T\boldsymbol{X}_{\boldsymbol{\gamma}}\left\{\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}}\mu_{q(1/\sigma_\varepsilon^2)}+\boldsymbol{F}_q^{-1}\right\}^{-1}\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{y}.&\end{aligned}$$

Then:

$$\begin{aligned}\log q^*(\beta_0,\boldsymbol{\theta},\boldsymbol{\gamma}) \;=\; & -\tfrac{1}{2}\left\{\tilde{\boldsymbol{\theta}}-\left[\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}}\mu_{q(1/\sigma_\varepsilon^2)}+\boldsymbol{F}_q^{-1}\right]^{-1}\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{y}\mu_{q(1/\sigma_\varepsilon^2)}\right\}^T\\ & \times\left\{\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}}\mu_{q(1/\sigma_\varepsilon^2)}+\boldsymbol{F}_q^{-1}\right\}\\ & \times\left\{\tilde{\boldsymbol{\theta}}-\left[\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}}\mu_{q(1/\sigma_\varepsilon^2)}+\boldsymbol{F}_q^{-1}\right]^{-1}\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{y}\mu_{q(1/\sigma_\varepsilon^2)}\right\}\\ & +\tfrac{1}{2}\mu_{q(1/\sigma_\varepsilon^2)}^2\boldsymbol{y}^T\boldsymbol{X}_{\boldsymbol{\gamma}}\left\{\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}}\mu_{q(1/\sigma_\varepsilon^2)}+\boldsymbol{F}_q^{-1}\right\}^{-1}\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{y}\\ & +\gamma_\bullet\mu_{q(\log(\rho))}+(p-\gamma_\bullet)\mu_{q(\log(1-\rho))}+\text{const.}\end{aligned}$$

This means that

$$q^*(\beta_0,\boldsymbol{\theta},\boldsymbol{\gamma})\propto\eta_{\boldsymbol{\gamma}}\phi_{\boldsymbol{V}_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})}}(\tilde{\boldsymbol{\theta}}-\boldsymbol{m}_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})}),$$

where

$$\begin{aligned}\boldsymbol{V}_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})} &= \left\{\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}}\mu_{q(1/\sigma_\varepsilon^2)}+\text{blockdiag}\left[(\sigma_{\beta_0}^2)^{-1},\mu_{q(1/\sigma_\beta^2)}\boldsymbol{I}_p\right]\right\}^{-1},\\ \boldsymbol{m}_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})} &= V_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})}\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{y}\mu_{q(1/\sigma_\varepsilon^2)}\end{aligned}$$

and

$$\begin{aligned}\eta_{\boldsymbol{\gamma}} &= \exp\Big\{\tfrac{1}{2}\log|V_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})}| + \gamma_\bullet \mu_{q(\log(\rho))} + (p-\gamma_\bullet)\mu_{q(\log(\rho))} \\ &\quad + \tfrac{1}{2}\mu^2_{q(1/\sigma^2_\varepsilon)} \boldsymbol{y}^T \boldsymbol{X}_{\boldsymbol{\gamma}} V_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})} \boldsymbol{X}^T_{\boldsymbol{\gamma}} \boldsymbol{y}\Big\}.\end{aligned}$$

Because

$$\int\int \eta_{\boldsymbol{\gamma}} \phi_{\boldsymbol{V}_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})}}(\tilde{\boldsymbol{\theta}} - \boldsymbol{m}_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})}) d\tilde{\boldsymbol{\theta}} d\boldsymbol{\gamma} = \sum_{\boldsymbol{\gamma}\in\mathbb{B}} \eta_{\boldsymbol{\gamma}},$$

hence

$$q^*(\beta_0, \boldsymbol{\theta}, \boldsymbol{\gamma}) = \mu_{q(\omega_{\boldsymbol{\gamma}})} \phi_{\boldsymbol{V}_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})}}(\tilde{\boldsymbol{\theta}} - \boldsymbol{m}_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})}),$$

where

$$\mu_{q(\omega_{\boldsymbol{\gamma}})} = \frac{\eta_{\boldsymbol{\gamma}}}{\sum_{\boldsymbol{\gamma}\in\mathbb{B}} \eta_{\boldsymbol{\gamma}}}.$$

Expressions for $q^*(\beta_0, \boldsymbol{\theta}|\boldsymbol{\gamma})$

From the expression of $q^*(\beta_0, \boldsymbol{\theta}, \boldsymbol{\gamma})$, we can obtain

$$q^*(\beta_0, \boldsymbol{\theta}|\boldsymbol{\gamma}) \sim N(\boldsymbol{m}_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})}, \boldsymbol{V}_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})}),$$

where

$$\boldsymbol{V}_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})} = \Big\{\boldsymbol{X}^T_{\boldsymbol{\gamma}} \boldsymbol{X}_{\boldsymbol{\gamma}} \mu_{q(1/\sigma^2_\varepsilon)} + \text{blockdiag}\left[(\sigma^2_{\beta_0})^{-1}, \mu_{q(1/\sigma^2_\beta)} \boldsymbol{I}_p\right]\Big\}^{-1}$$

and

$$\boldsymbol{m}_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})} = V_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})} \boldsymbol{X}^T_{\boldsymbol{\gamma}} \boldsymbol{y} \mu_{q(1/\sigma^2_\varepsilon)}.$$

Expressions for $q^*(\beta_0, \boldsymbol{\theta})$, $\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\theta})}$ and $\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\theta})}$

From the expression for $q^*(\beta_0, \boldsymbol{\theta}, \boldsymbol{\gamma})$, we can obtain

$$q^*(\beta_0, \boldsymbol{\theta}) = \sum_{\boldsymbol{\gamma} \in \mathbb{B}} \mu_{q(\omega_{\boldsymbol{\gamma}})} \phi_{\boldsymbol{V}_{\boldsymbol{\gamma} q(\beta_0,\boldsymbol{\theta})}}(\tilde{\boldsymbol{\theta}} - \boldsymbol{m}_{\boldsymbol{\gamma} q(\beta_0,\boldsymbol{\theta})}),$$

which is a mixture of Multivariate Gaussian density functions by definition (1.29). Using the results (1.12) for a mixture of multivariate Gaussian distribution,

$$\mu_{q(\beta_0,\boldsymbol{\theta})} = \sum_{\boldsymbol{\gamma} \in \mathbb{B}} \mu_{q(\omega_{\boldsymbol{\gamma}})} \boldsymbol{m}_{\boldsymbol{\gamma} q(\beta_0,\boldsymbol{\theta})}$$

and

$$\Sigma_{q(\beta_0,\boldsymbol{\theta})} = \sum_{\boldsymbol{\gamma} \in \mathbb{B}} \mu_{q(\omega_{\boldsymbol{\gamma}})} V_{\boldsymbol{\gamma} q(\beta_0,\boldsymbol{\theta})} - \mu_{q(\beta_0,\boldsymbol{\theta})} \mu_{q(\beta_0,\boldsymbol{\theta})}^T.$$

Expressions for $q^*(\rho)$, $\mu_{q(\mathbf{log}(\rho))}$ and $\mu_{q(\mathbf{log}(1-\rho))}$

$$q^*(\rho) \sim \text{Beta}(1 + \sum_{j=1}^{p} \mu_{q(\gamma_j)}, 1 + p - \sum_{j=1}^{p} \mu_{q(\gamma_j)}),$$

$$\mu_{q(\log(\rho))} = \psi(1 + \sum_{j=1}^{j=p} \mu_{q(\gamma_j)}) - \psi(2 + p)$$

and

$$\mu_{q(\log(1-\rho))} = \psi(1 - \sum_{j=1}^{j=p} \mu_{q(\gamma_j)}) - \psi(2 + p).$$

Derivation:

$$\begin{aligned} \log q^*(\rho) &= E_q \left[\log p(\rho|\text{rest})\right] \\ &= \sum_{j=1}^{p} \mu_{q(\gamma_j)} \log(\rho) + (p - \sum_{j=1}^{p} \mu_{q(\gamma_j)}) \log(1-\rho) + \text{const.} \end{aligned}$$

Therefore,

$$q^*(\rho) \;=\; \exp\left\{\sum_{j=1}^{p}\mu_{q(\gamma_j)}\log(\rho)+(p-\sum_{j=1}^{p}\mu_{q(\gamma_j)})\log(1-\rho)+\text{const.}\right\}$$

The results then follow from definition (1.22) and result (1.6) for the Beta distribution.

Expressions for $q^*(\sigma_\varepsilon^2)$

$$q^*(\sigma_\varepsilon^2) \sim \text{ Inverse-Gamma}(A_{q(\sigma_\varepsilon^2)}, B_{q(\sigma_\varepsilon^2)}),$$

where

$$\begin{aligned} A_{q(\sigma_\varepsilon^2)} &= \frac{n}{2}+A_\varepsilon, \\ B_{q(\sigma_\varepsilon^2)} &= B_\varepsilon+\tfrac{1}{2}\sum_{\boldsymbol{\gamma}\in\mathbb{B}}\mu_{q(\omega_{\boldsymbol{\gamma}})}\left\{\parallel \boldsymbol{y}-\boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{m}_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})}\parallel^2+tr(\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}}V_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})})\right\} \end{aligned}$$

and

$$\mu_{q(1/\sigma_\varepsilon^2)}=\frac{A_{q(\sigma_\varepsilon^2)}}{B_{q(\sigma_\varepsilon^2)}}.$$

Derivation:

$$\begin{aligned} \log q^*(\sigma_\varepsilon^2) &= E_q\left[\log p(\sigma_\varepsilon^2|\text{rest})\right] \\ &= (-1-A_\varepsilon-\frac{n}{2})\log(\sigma_\varepsilon^2)-\frac{2B_\varepsilon+E_q\left[\|\boldsymbol{y}-\boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\boldsymbol{\theta}}\|^2\right]}{2\sigma_\varepsilon^2}. \end{aligned}$$

Using the iterated expectation method,

$$\begin{aligned} & E_q\left[\|\boldsymbol{y}-\boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\boldsymbol{\theta}}\|^2\right] \\ & \qquad = E_{q(\boldsymbol{\gamma})}\left[E_{q(\tilde{\boldsymbol{\theta}})}\left[\|\boldsymbol{y}-\boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\boldsymbol{\theta}}\|^2|\boldsymbol{\gamma}\right]\right] \\ & \qquad = \sum_{\boldsymbol{\gamma}\in\mathbb{B}} \mu_{q(\omega_{\boldsymbol{\gamma}})}\left\{\| \boldsymbol{y}-\boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{m}_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})} \|^2 + tr(\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}}V_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})})\right\}. \end{aligned}$$

The results then follow from definition (1.23) and results (1.7) for the Inverse-Gamma distribution.

Expressions for $q^*(\sigma_\beta^2)$

$$q^*(\sigma_\beta^2) \sim \text{ Inverse-Gamma}(A_{q(\sigma_\beta^2)}, B_{q(\sigma_\beta^2)}),$$

where

$$\begin{aligned} A_{q(\sigma_\beta^2)} &= \frac{p}{2} + A_\beta, \\ B_{q(\sigma_\beta^2)} &= B_\beta + \tfrac{1}{2}\left\{\| \mu_{q(\boldsymbol{\theta})} \|^2 + \text{tr}(\Sigma_{q(\boldsymbol{\theta})})\right\} \end{aligned}$$

and

$$\mu_{q(1/\sigma_\beta^2)} \leftarrow \frac{A_{q(\sigma_\beta^2)}}{B_{q(\sigma_\beta^2)}}.$$

Derivation:

$$\begin{aligned} \log q^*(\sigma_\beta^2) &= E_q\left[\log p(\sigma_\beta^2|\text{rest})\right] \\ &= (-1-A_\beta-\frac{1}{2})\log(\sigma_\beta^2) - \frac{2B_\beta + E_q[\|\boldsymbol{\theta}\|^2]}{2\sigma_\beta^2} + const. \\ &= (-1-A_\beta-\frac{1}{2})\log(\sigma_\beta^2) - \frac{2B_\beta + \|\boldsymbol{\mu}_{\boldsymbol{\theta}}\|^2 + \text{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}})}{2\sigma_\beta^2} + \text{const.} \end{aligned}$$

The results then follow from definition (1.23) and result (1.7) for the Inverse-Gamma distribution.

3.A.3 Derivation of lower bound

We note that

$$\begin{aligned}
\log \underline{p}(y;q) &= E_q\{\log\, p(\beta_0,\boldsymbol{\theta},\sigma_\beta^2,\sigma_\varepsilon^2,\boldsymbol{\gamma},\rho) - \log\, q(\beta_0,\boldsymbol{\theta},\boldsymbol{\gamma},\rho,\sigma_\beta^2,\sigma_\varepsilon^2)\} \\
&= E_q\log\, p(y|\beta_0,\boldsymbol{\theta},\boldsymbol{\gamma},\sigma_\varepsilon^2) + E_q\log\, p(\sigma_\varepsilon^2) + E_q\log\, p(\sigma_\beta^2) \\
&\quad + E_q\log\, p(\boldsymbol{\theta}|\sigma_\beta^2) + E_q\log\, p(\beta_0) + E_q\log\, p(\boldsymbol{\gamma}|\rho) + E_q\log\, p(\rho) \\
&\quad - E_q\log\, q(\beta_0,\boldsymbol{\theta},\boldsymbol{\gamma}) - E_q\log\, q(\sigma_\beta^2) - E_q\log\, q(\sigma_\varepsilon^2) - E_q\log\, q(\rho).
\end{aligned}$$

Firstly,

$$\begin{aligned}
&E_q[\log\, p(y|\beta_0,\boldsymbol{\theta},\boldsymbol{\gamma},\sigma_\varepsilon^2)] \\
&\quad = -\frac{n}{2}\mu_{q(\log(\sigma_\varepsilon^2))} - \frac{n}{2}\log(2\pi) - \frac{1}{2}\mu_{q(1/\sigma_\varepsilon^2)} E_q\left[\parallel \boldsymbol{y} - \boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\boldsymbol{\theta}} \parallel^2\right] \\
&\quad = -\frac{n}{2}\mu_{q(\log(\sigma_\varepsilon^2))} - \frac{n}{2}\log(2\pi) \\
&\qquad -\frac{1}{2}\mu_{q(1/\sigma_\varepsilon^2)} \sum_{\boldsymbol{\gamma}\in\mathbb{B}} \mu_{q(\omega_{\boldsymbol{\gamma}})} \left\{\parallel \boldsymbol{y} - \boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{m}_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})} \parallel^2 + \mathrm{tr}(\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{V}_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})})\right\}.
\end{aligned}$$

Secondly,

$$\begin{aligned}
&E_q[\log\, p(\sigma_\varepsilon^2) - \log\, q(\sigma_\epsilon^2)] \\
&\quad = A_\varepsilon \log B_\varepsilon - \log\Gamma(A_\varepsilon) \\
&\qquad -(A_\varepsilon + \frac{n}{2})\log(B_{q(\sigma_\varepsilon^2)}) + \log\Gamma(A_\varepsilon + \frac{n}{2}) + \frac{n}{2}\mu_{q(\log(\sigma_\varepsilon^2))} \\
&\qquad +\frac{1}{2}\mu_{q(1/\sigma_\varepsilon^2)} \sum_{\boldsymbol{\gamma}\in\mathbb{B}} \mu_{q(\omega_{\boldsymbol{\gamma}})} \left\{\parallel \boldsymbol{y} - \boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{m}_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})} \parallel^2 + \mathrm{tr}(\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{V}_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})})\right\}.
\end{aligned}$$

Thirdly,

$$
\begin{aligned}
&E_q[\log\, p(\sigma_\beta^2) - \log\, q(\sigma_\beta^2)] \\
&\quad = A_\beta \log B_\beta - \log\Gamma(A_\beta) \\
&\qquad -(A_\beta + \frac{p}{2})\log(B_{q(\sigma_\beta^2)}) + \log\Gamma(A_\beta + \frac{p}{2}) \\
&\qquad +\frac{p}{2}\mu_{q(\log(\sigma_\beta^2))} + \frac{p}{2}\mu_{q(1/\sigma_\beta^2)} \left\{\parallel \mu_{q(\boldsymbol{\theta})} \parallel^2 + \mathrm{tr}(\Sigma_{q(\boldsymbol{\theta})})\right\}.
\end{aligned}
$$

Fourthly,

$$
\begin{aligned}
&E_q[\log\, p(\boldsymbol{\gamma}|\rho) + \log\, p(\rho) - \log\, q(\rho)] \\
&\quad = \sum_{j=1}^{p} \left\{\mu_{q(\gamma_j)} E_q[\log(\rho)] + (1 - \mu_{q(\gamma_j)}) E_q[\log(1-\rho)]\right\} \\
&\qquad - \left\{\sum_{j=1}^{p} \mu_{q(\gamma_j)} E_q[\log(\rho)] + (p - \sum_{j=1}^{p} \mu_{q(\gamma_j)}) E_q[\log(1-\rho)]\right\} \\
&\quad = 0.
\end{aligned}
$$

Fifthly,

$$
\begin{aligned}
&E_q[\log\, p(\beta_0) + \log\, p(\boldsymbol{\theta}|\sigma_\beta^2) - \log\, q(\beta_0, \boldsymbol{\theta}, \boldsymbol{\gamma})] \\
&\quad = \frac{1+p}{2} - \frac{1}{2}\log(\sigma_{\beta_0}^2) - \frac{p}{2}\mu_{q(\log(\sigma_\beta^2))} \\
&\qquad -\frac{1}{2}\frac{\mu_{\boldsymbol{\gamma} q(\beta_0)}^2 + \sigma_{\boldsymbol{\gamma} q(\beta_0)}^2}{\sigma_{\beta_0}^2} + \frac{1}{2}\sum_{\boldsymbol{\gamma}\in\mathbb{B}} \mu_{q(\omega_{\boldsymbol{\gamma}})} \log|\boldsymbol{V}_{\boldsymbol{\gamma} q(\beta_0, \boldsymbol{\theta})}| \\
&\qquad -\frac{p}{2}\mu_{q(1/\sigma_\beta^2)} \left\{\parallel \mu_{q(\boldsymbol{\theta})} \parallel^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})})\right\} - \frac{1}{2}\sum_{\boldsymbol{\gamma}\in\mathbb{B}} \mu_{q(\omega_{\boldsymbol{\gamma}})} \log \mu_{q(\omega_{\boldsymbol{\gamma}})}.
\end{aligned}
$$

Substitution of these gives the lower bound (3.5):

$$\begin{aligned}\log \underline{p}(y;q) &= -\frac{n}{2}\log(2\pi)+\frac{1+p}{2}-\frac{1}{2}\log(\sigma^2_{\beta_0})\\ &\quad +A_\varepsilon \log B_\varepsilon - \log\,\Gamma(A_\varepsilon)\\ &\quad -(A_\varepsilon+\frac{n}{2})\log(B_{q(\sigma^2_\varepsilon)})+\log\,\Gamma(A_\varepsilon+\frac{n}{2})\\ &\quad +A_\beta \log B_\beta - \log\,\Gamma(A_\beta)\\ &\quad -(A_\beta+\frac{p}{2})\log(B_{q(\sigma^2_\beta)})+\log\,\Gamma(A_\beta+\frac{p}{2})\\ &\quad +\frac{1}{2}\sum_{\boldsymbol{\gamma}\in\mathbb{B}}\mu_{q(\omega_{\boldsymbol{\gamma}})}\left\{\log|\boldsymbol{V}_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})}|-\frac{\mu^2_{\boldsymbol{\gamma}q(\beta_0)}+\sigma^2_{\boldsymbol{\gamma}q(\beta_0)}}{\sigma^2_{\beta_0}}-2\log(\mu_{q(\omega_{\boldsymbol{\gamma}})})\right\}.\end{aligned}$$

3.B Appendix: Derivation of Algorithm 3.3.1

In Algorithm 3.3.1, the MFVB calculations for β_0, $\boldsymbol{\theta}$, σ^2_β, ρ and $\boldsymbol{\gamma}$ are similar to the Gaussian case (Algorithm 3.2.1). We obtain them by using

$$1 \text{ to replace } \mu_{1/\sigma^2_\varepsilon}$$

$$\text{and } \boldsymbol{\mu}_{q(\boldsymbol{a})} \text{ to replace } \boldsymbol{y}.$$

Therefore, I only show the derivation for $\boldsymbol{a}$.

3.B.1 Full conditional for $\boldsymbol{a}$

$$\begin{aligned}\log p(a_i|\text{rest}) &= -\tfrac{1}{2}[a_i-(\boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\boldsymbol{\theta}})_i]^2\\ &\quad +\log\left[(\boldsymbol{I}(a_i\geq 0))^{y_i}(\boldsymbol{I}(a_i<0))^{1-y_i}\right]+\text{const.}\end{aligned}$$

Derivation:

$$\begin{aligned} p(a_i|\text{rest}) &\propto p(a_i|\beta_0, \boldsymbol{\theta}, \boldsymbol{\gamma}) p(y_i|a_i) \\ &= (2\pi)^{-1/2} \exp\left\{ -\frac{[a_i - \beta_0 - \boldsymbol{X}_i(\boldsymbol{\gamma} \odot \boldsymbol{\theta})]^2}{2} \right\} \\ &\quad \times (\boldsymbol{I}(a_i \geq 0))^{y_i} (\boldsymbol{I}(a_i < 0))^{1-y_i} \end{aligned}$$

Taking logarithms, we get

$$\begin{aligned} \log p(a_i|\text{rest}) &= -\frac{[a_i - \beta_0 - \boldsymbol{X}_i(\boldsymbol{\gamma} \odot \boldsymbol{\theta})]^2}{2} \\ &\quad + \log\left[(\boldsymbol{I}(a_i \geq 0))^{y_i} (\boldsymbol{I}(a_i < 0))^{1-y_i} \right] + \text{const.} \\ &= -\tfrac{1}{2}[a_i - (\boldsymbol{X}_{\boldsymbol{\gamma}} \tilde{\boldsymbol{\theta}})_i]^2 \\ &\quad + \log\left[(\boldsymbol{I}(a_i \geq 0))^{y_i} (\boldsymbol{I}(a_i < 0))^{1-y_i} \right] + \text{const.} \end{aligned}$$

3.B.2 Expressions for $q^*(\boldsymbol{a})$ and $\boldsymbol{\mu}_{q(\boldsymbol{a})}$

If $y_i = 1$,

$$q(a_i) = \frac{\phi(a_i - (\mu_{q(\boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\theta})})_i)}{\Phi((\mu_{q(\boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\theta})})_i)}, \quad a_i \geq 0,$$

which is a truncated normal density function on $(0, \infty)$; and if $y_i = 0$,

$$q(a_i) = \frac{\phi(a_i - (\mu_{q(\boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\theta})})_i)}{1 - \Phi((\mu_{q(\boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\theta})})_i)}, \quad a_i < 0$$

which is a truncated normal density function on $(-\infty, 0)$, where

$$\mu_{q(\boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\theta})} = \sum_{\boldsymbol{\gamma} \in \mathbb{B}} \mu_{q(\omega_{\boldsymbol{\gamma}})} \boldsymbol{X}_{\boldsymbol{\gamma}} \boldsymbol{m}_{\boldsymbol{\gamma} q(\beta_0, \boldsymbol{\theta})}.$$

Derivation:

If $y_i = 1$, then $a_i \geq 0$ and

$$\begin{aligned}\log q^*(a_i) &= E_q[p(a_i|\text{rest})] \\ &= -\tfrac{1}{2}E_q\left\{[a_i - (\boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\boldsymbol{\theta}})_i]^2\right\} + const. \\ &= -\tfrac{1}{2}[a_i - (\mu_{q(\boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\theta})})_i]^2 + const.\end{aligned}$$

where

$$\mu_{q(\boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\theta})} = E_q[\boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\boldsymbol{\theta}}].$$

Using the iterated expectation method,

$$\begin{aligned}\mu_{q(\boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\theta})} &= E_{q(\boldsymbol{\gamma})}\left[E_{q(\tilde{\boldsymbol{\theta}})}\left[\boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\theta}|\boldsymbol{\gamma}\right]\right] \\ &= \sum_{\boldsymbol{\gamma}\in\mathbb{B}} \mu_{q(\omega_{\boldsymbol{\gamma}})}\boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{m}_{\boldsymbol{\gamma}q(\beta_0,\boldsymbol{\theta})}.\end{aligned}$$

The results for $y_i \geq 0$ then follow definition (1.27) for the truncated Gaussian distribution. Similarly, we can obtain the result for $y_i < 0$. Using the Result (1.11) of the truncated Gaussian distribution, we obtain the expression

$$\mu_{q(\boldsymbol{a})} = \mu_{q(\boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\theta})} + \frac{\phi(\mu_{q(\boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\theta})})}{\Phi(\mu_{q(\boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\theta})})^{\boldsymbol{y}} \odot [\mathbf{1} - \Phi(\mu_{q(\boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\theta})})]^{\mathbf{1}-\boldsymbol{y}}}$$

3.B.3 Derivation of lower bound

We note that

$$\begin{aligned}\log \underline{p}(y;q) &= E_q\{\log\, p(\beta_0,\boldsymbol{\theta},\sigma_\beta^2,\boldsymbol{\gamma},\rho,\boldsymbol{a}) - \log q(\beta_0,\boldsymbol{\theta},\boldsymbol{\gamma},\rho,\sigma_\beta^2,\boldsymbol{a})\} \\ &= E_q[\log\, p(y|\boldsymbol{a}) + \log\, p(\boldsymbol{a}|\beta_0,\boldsymbol{\theta},\boldsymbol{\gamma}) + \log\, p(\sigma_\beta^2) \\ &\quad +\log\, p(\boldsymbol{\theta}|\sigma_\beta^2) + \log\, p(\beta_0) + \log\, p(\boldsymbol{\gamma}|\rho) + \log\, p(\rho) \\ &\quad -\log\, q(\beta_0,\boldsymbol{\theta},\boldsymbol{\gamma}) - \log\, q(\sigma_\beta^2) - \log\, q(\boldsymbol{a}) - \log\, q(\rho)].\end{aligned}$$

Firstly,

$$\begin{aligned}
&E_q[\log\, p(y|\boldsymbol{a}) + p(\boldsymbol{a}|\beta_0, \boldsymbol{\theta}, \boldsymbol{\gamma}) - q(\beta_0, \boldsymbol{\theta}, \boldsymbol{\gamma})] \\
&\quad = y^T \log[\Phi(\sum_{\boldsymbol{\gamma}\in\mathbb{B}} \mu_{q(\omega_{\boldsymbol{\gamma}})} \boldsymbol{X}_{\boldsymbol{\gamma}} \boldsymbol{m}_{\boldsymbol{\gamma} q(\beta_0, \boldsymbol{\theta})})] \\
&\qquad + (\mathbf{1} - y)^T \log[\mathbf{1} - \Phi(\sum_{\boldsymbol{\gamma}\in\mathbb{B}} \mu_{q(\omega_{\boldsymbol{\gamma}})} \boldsymbol{X}_{\boldsymbol{\gamma}} \boldsymbol{m}_{\boldsymbol{\gamma} q(\beta_0, \boldsymbol{\theta})})] \\
&\qquad - \frac{1}{2} \sum_{\boldsymbol{\gamma}\in\mathbb{B}} \mu_{q(\omega_{\boldsymbol{\gamma}})} \left\{ \mathrm{tr}(\boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X}_{\boldsymbol{\gamma}} \boldsymbol{V}_{q(\beta_0, \boldsymbol{\theta})}) \right\}.
\end{aligned}$$

Secondly,

$$\begin{aligned}
&E_q[\log\, p(\sigma_\beta^2) - \log\, q(\sigma_\beta^2)] \\
&\quad = A_\beta \log B_\beta - \log \Gamma(A_\beta) \\
&\qquad - (A_\beta + \frac{p}{2}) \log(B_{q(\sigma_\beta^2)}) + \log \Gamma(A_\beta + \frac{p}{2}) \\
&\qquad + \frac{p}{2} \mu_{q(\log(\sigma_\beta^2))} + \frac{p}{2} \mu_{q(1/\sigma_\beta^2)} \left\{ \| \mu_{q(\boldsymbol{\theta})} \|^2 + \mathrm{tr}(\Sigma_{q(\boldsymbol{\theta})}) \right\}.
\end{aligned}$$

Thirdly,

$$\begin{aligned}
&E_q[\log\, p(\boldsymbol{\gamma}|\rho) + \log\, p(\rho) - \log\, q(\rho)] \\
&\quad = \sum_{j=1}^{p} \left\{ \mu_{q(\gamma_j)} E_q[\log(\rho)] + (1 - \mu_{q(\gamma_j)}) E_q[\log(1-\rho)] \right\} \\
&\qquad - \left\{ \sum_{j=1}^{p} \mu_{q(\gamma_j)} E_q[\log(\rho)] + (p - \sum_{j=1}^{p} \mu_{q(\gamma_j)}) E_q[\log(1-\rho)] \right\} \\
&\quad = 0.
\end{aligned}$$

Fourthly,

$$
\begin{aligned}
&E_q[\log\, p(\beta_0) + \log\, p(\boldsymbol{\theta}|\sigma_\beta^2) - \log\, q(\beta_0, \boldsymbol{\theta}, \boldsymbol{\gamma})] \\
&\quad = \frac{1+p}{2} - \frac{1}{2}\log(\sigma_{\beta_0}^2) - \frac{p}{2}\mu_{q(\log(\sigma_\beta^2))} \\
&\qquad -\frac{1}{2}\frac{\mu_{\boldsymbol{\gamma} q(\beta_0)}^2 + \sigma_{\boldsymbol{\gamma} q(\beta_0)}^2}{\sigma_{\beta_0}^2} + \frac{1}{2}\sum_{\boldsymbol{\gamma}\in\mathbb{B}} \mu_{q(\omega_{\boldsymbol{\gamma}})} \log|\boldsymbol{V}_{\boldsymbol{\gamma} q(\beta_0,\boldsymbol{\theta})}| \\
&\qquad -\frac{p}{2}\mu_{q(1/\sigma_\beta^2)} \left\{ \| \boldsymbol{\mu}_{q(\boldsymbol{\theta})} \|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}) \right\} - \frac{1}{2}\sum_{\boldsymbol{\gamma}\in\mathbb{B}} \mu_{q(\omega_{\boldsymbol{\gamma}})} \log(\mu_{q(\omega_{\boldsymbol{\gamma}})}).
\end{aligned}
$$

Substitution of these gives the lower bound expression:

$$
\begin{aligned}
\log \underline{p}(y; q) \;=\; & \frac{1+p}{2} - \frac{1}{2}\log(\sigma_{\beta_0}^2) \\
& + A_\beta \log B_\beta - \log\, \Gamma(A_\beta) \\
& - (A_\beta + \frac{p}{2})\log(B_{q(\sigma_\beta)}) + \log\, \Gamma(A_\beta + \frac{p}{2}) \\
& + y^T \log[\Phi(\sum_{\boldsymbol{\gamma}\in\mathbb{B}} \mu_{q(\omega_{\boldsymbol{\gamma}})} \boldsymbol{X}_{\boldsymbol{\gamma}} \boldsymbol{m}_{\boldsymbol{\gamma} q(\beta_0,\boldsymbol{\theta})})] \\
& + (\mathbf{1} - y)^T \log[\mathbf{1} - \Phi(\sum_{\boldsymbol{\gamma}\in\mathbb{B}} \mu_{q(\omega_{\boldsymbol{\gamma}})} \boldsymbol{X}_{\boldsymbol{\gamma}} \boldsymbol{m}_{\boldsymbol{\gamma} q(\beta_0,\boldsymbol{\theta})})] \\
& - \frac{1}{2}\sum_{\boldsymbol{\gamma}\in\mathbb{B}} \mu_{q(\omega_{\boldsymbol{\gamma}})} \left\{ \mathrm{tr}(\boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X}_{\boldsymbol{\gamma}} \boldsymbol{V}_{q(\beta_0,\boldsymbol{\theta})}) - \log|\boldsymbol{V}_{\boldsymbol{\gamma} q(\beta_0,\boldsymbol{\theta})}| \right\} \\
& - \frac{1}{2}\sum_{\boldsymbol{\gamma}\in\mathbb{B}} \mu_{q(\omega_{\boldsymbol{\gamma}})} \left\{ \frac{\mu_{\boldsymbol{\gamma} q(\beta_0)}^2 + \sigma_{\boldsymbol{\gamma} q(\beta_0)}^2}{\sigma_{\beta_0}^2} + 2\log \mu_{q(\omega_{\boldsymbol{\gamma}})} \right\}.
\end{aligned}
$$

Chapter 4

Variational Bayesian Lasso

4.1 Introduction

The Least Absolute Shrinkage and Selection Operator (Lasso), attributed to Tibshirani (1996), is a regression method that involves penalizing the absolute size of the regression coefficients. Consider a linear regression model:

$$\boldsymbol{y} = \mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$ is an n-dimensional vector of response variables, $\boldsymbol{X}$ is an $n \times p$ matrix, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$ is an error vector of independent and identically distributed normal random variables with mean 0 and unknown variance, σ_ε^2. β_0 is an intercept coefficient and $\boldsymbol{\beta}$ is a p-dimensional coefficient vector. Lasso estimation is a form of penalized least squares that minimizes the residual sum of squares when penalizing the ℓ_1-norm of $\boldsymbol{\beta}$:

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) = \underset{\beta_0, \boldsymbol{\beta}}{\arg\min} \left(\boldsymbol{y} - 1_n\beta_0 - \boldsymbol{X}\boldsymbol{\beta}\right)^T \left(\boldsymbol{y} - 1_n\beta_0 - \boldsymbol{X}\boldsymbol{\beta}\right) + \lambda \sum_{j=1}^{p} |\beta_j|, \qquad (4.1)$$

where λ is nonnegative. If $\lambda = 0$, Lasso estimation is ordinary least-squares estimation. The modification Least Angle Regression (LARS) developed by Efron *et al.* (2004) can be used to compute the Lasso estimations. Zhao and Yu (2007) proposed a boosted Lasso algorithm to produce the complete regularization path for general Lasso problems. Lastly, `lars` (Hastie & Efron, 2013) is a convenient R package that fits an entire Lasso sequence at the cost of a single least squares fit.

Shrinkage of the vector of regression coefficients toward zero via the ℓ_1 norm allows some coefficients to be set to identically equal zero. This key feature allows Lasso to be used for variable selection. The Lasso method builds a sequence of candidate models from which the final model is chosen. The sequence is controlled by the value of λ. Zhao and Yu (2006) proposed the Irrepresentable Condition for the model selection consistency of Lasso.

Tibshirani (1996) found a connection between the Lasso and Bayesian inference. The Lasso estimate can be derived as the Bayes maximum *a posteriori* under independent Laplace priors for each regression coefficient, β_j, as:

$$f(\beta_j) = \frac{\lambda}{2}\exp\left(-\lambda|\beta_j|\right), \quad j = 1, 2, \ldots, p. \tag{4.2}$$

Fernandez and Steel (2000) used this Bayesian Lasso model to deal with linear regression under independent sampling from general scale mixtures of normals. Balakrishnan and Madigan (2010) obtained a demi-Bayesian Lasso method to deal with the sparse Bayesian learning problem (Tipping, 2001). Armagan (2009) developed a variational approximation method for Tibshirani's Lasso model.

Park and Casella (2008) indicated that using the unconditional prior (4.2) leads to a bimodal joint density for β and σ^2, and the marginal posterior density of β is also bimodal for the simplest linear regression model (one predictor). For MCMC or Gibbs samplers, lack of unimodality results in slow convergence and leads to

less meaningful point estimation. For MFVB, assumed independence between β and σ_ε^2 will result in a unimodal posterior distribution for each parameter estimate. Park and Casella (2008) introduced a conditional Laplace prior specification of the form:

$$f(\beta_j|\sigma_\varepsilon^2) = \frac{\lambda}{2\sqrt{\sigma_\varepsilon^2}}\exp\left(\frac{-\lambda|\beta_j|}{\sqrt{\sigma_\varepsilon^2}}\right),\ j = 1, 2, \ldots, p. \tag{4.3}$$

The regression parameter depends on the variance of the model error, σ_ε^2. The joint posterior distribution of β and σ_ε^2 is unimodal, under typical prior distribution of σ_ε^2 and any $\lambda \geq 0$. Park and Casella (2008) also proposed a Bayesian regression model with latent variables via Gibbs samplers. Based on this Bayesian Lasso model, Hans (2009) recommended a direct characterization of the posterior using two new Gibbs samplers that do not require the use of latent variables.

The Bayesian Lasso path of regression coefficients is smooth for shrinkage parameters λ, and it is unable to achieve a sparsity in the regression coefficients vector. Therefore, the Bayesian Lasso model cannot deal with variable selection directly. Yuan and Lin (2005) used a method of combining stochastic model selection and Bayesian Lasso for variable selection and coefficient estimation in linear regression models. Hans (2010) discussed the uncertainty of variable selection for Bayesian Lasso regression.

A MFVB inference based on Park and Casella's Bayesian Lasso model and methods for obtaining hyperparameter will be presented in this chapter, and the corresponding results will be compared with a MCMC method. The development of MFVB methodology for high-dimensional linear regression model will also be addressed in this chapter.

4.2 Basic Bayesian Lasso Model

We firstly consider Park and Casella's Bayesian Lasso for a linear regression model with given λ. Next, two methods, the variational expectation-maximization (VEM) algorithm and adding a prior distribution in λ, will be presented to choose the hyperparameter λ.

4.2.1 Models

The Bayesian Lasso model with Laplace prior takes the form:

$$\begin{aligned}
&\boldsymbol{y} \sim N(1_n\beta_0 + \boldsymbol{X}\boldsymbol{\beta}, \sigma_\varepsilon^2 \boldsymbol{I}_n),\\
&\beta_0 \sim N(0, \sigma_{\beta_0}^2),\\
&\beta_j|\sigma_\varepsilon^2 \overset{\text{ind.}}{\sim} \text{Laplace}(0, \sigma_\varepsilon^2, \lambda), \quad j = 1, 2, \ldots, p,\\
&\sigma_\varepsilon^2 \sim \text{ Inverse-Gamma}(A_\varepsilon, B_\varepsilon),
\end{aligned} \tag{4.4}$$

where $\sigma_{\beta_0}^2$, A_ε, $B_\varepsilon > 0$ and $\lambda \geq 0$ are hyperparameters. We introduce the auxiliary variables $\boldsymbol{a} = (a_1, a_2, ..., a_p)^T$ corresponding to Result 1.9 to avoid the irreducible integrals in derivation of MFVB. The full model with auxiliary variables is then:

$$\begin{aligned}
&\boldsymbol{y} \sim N(1_n\beta_0 + \boldsymbol{X}\boldsymbol{\beta}, \sigma_\varepsilon^2 \boldsymbol{I}_n),\\
&\beta_0 \sim N(0, \sigma_{\beta_0}^2),\\
&\beta_j|a_j \overset{\text{ind.}}{\sim} N(0, \frac{\sigma_\varepsilon^2}{a_j}), \quad j = 1, 2, \ldots, p,\\
&a_j|\lambda^2 \overset{\text{ind.}}{\sim} \text{ Inverse-Gamma}(1, \frac{\lambda^2}{2}),\\
&\sigma_\varepsilon^2 \sim \text{ Inverse-Gamma}(A_\varepsilon, B_\varepsilon).
\end{aligned} \tag{4.5}$$

Figure 4.1 shows the directed acyclic graph corresponding to (4.5).

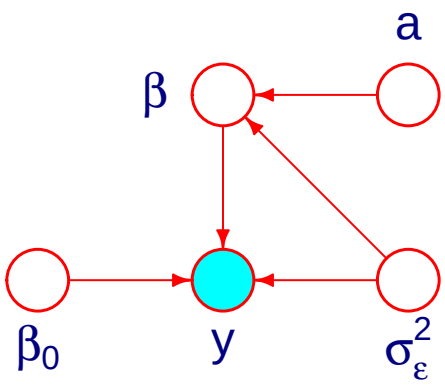

Figure 4.1: Directed acyclic graph for model (4.5).

4.2.2 Mean field variational Bayes scheme

We now seek a quick deterministic approximate inference procedure for (4.5) based on the MFVB method. A tractable solution arises if we impose the product restriction:

$$q(\beta_0, \boldsymbol{\beta}, \boldsymbol{a}, \sigma_\varepsilon^2) = q(\beta_0, \boldsymbol{\beta}) q(\boldsymbol{a}, \sigma_\varepsilon^2).$$

The theory of induced factorizations (e.g., Bishop, 2006, Section 10.2.5) leads to a solution with the additional product structure:

$$q(\beta_0, \boldsymbol{\beta}, \boldsymbol{a}, \sigma_\varepsilon^2) = q(\beta_0, \boldsymbol{\beta}) \prod_{j=1}^{p} q^*(a_j) q(\sigma_\varepsilon^2).$$

Algorithm 4.2.1: MFVB iterative scheme for obtaining the parameters of the optimal densities $q^*(\beta_0, \boldsymbol{\beta})$, $q^*(a_i)$ and $q^*(\sigma_\varepsilon^2)$ for the Bayesian Lasso model (4.5).

Initialize $\mu_{q(1/\sigma_\varepsilon^2)}$ and $\mu_{q(a_j)}$, $j = 1, \ldots, p$;

Cycle

$$
\begin{aligned}
&\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})} \leftarrow \left\{\mu_{q(1/\sigma_\varepsilon^2)}\boldsymbol{C}^T\boldsymbol{C} + \text{blockdiag}\left[(\sigma_{\beta_0}^2)^{-1}, \mu_{q(1/\sigma_\varepsilon^2)}\boldsymbol{D}_{\mu_q(\boldsymbol{a})}\right]\right\}\\
&\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})} \leftarrow \boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}\boldsymbol{C}^T\boldsymbol{y}\mu_{q(1/\sigma_\varepsilon^2)}\\
&A_{q(\sigma_\varepsilon^2)} \leftarrow \frac{p+n}{2} + A_\varepsilon\\
&B_{q(\sigma_\varepsilon^2)} \leftarrow B_\varepsilon + \tfrac{1}{2}\big\{ \parallel \boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})} \parallel^2 + \text{tr}(\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})})\\
&\qquad\qquad + \text{tr}(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}^T\boldsymbol{D}_{\mu_q(\boldsymbol{a})} + \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\boldsymbol{D}_{\mu_q(\boldsymbol{a})})\big\}\\
&\mu_{q(1/\sigma_\varepsilon^2)} \leftarrow \frac{A_{q(\sigma_\varepsilon^2)}}{B_{q(\sigma_\varepsilon^2)}}\\
&\mu_{q(a_j)} \leftarrow \sqrt{\frac{\lambda^2}{(\mu_{q(\beta_i)}^2 + \sigma_{q(\beta_i)}^2)\mu_{q(1/\sigma_\varepsilon^2)}}} \quad j = 1, 2, \ldots, p\\
&\boldsymbol{D}_{\mu_q(\boldsymbol{a})} \leftarrow \text{diag}(\mu_{q(a_1)}, \mu_{q(a_2)}, \ldots, \mu_{q(a_p)})
\end{aligned}
$$

until the increase in $\log \underline{p}(y; q)$ is negligible.

Then, as shown in Appendix 4.A, the optimal q^* densities for the parameters in model (4.5) take the form:

$$
\begin{aligned}
&q^*(\beta_0, \boldsymbol{\beta}) \text{ is a Multivariate Gaussian density function,}\\
&q^*(a_j) \text{ is an Inverse Gaussian density function,}\\
&q^*(\sigma_\varepsilon^2) \text{ is an Inverse Gamma density functions.}
\end{aligned}
\tag{4.6}
$$

Let $\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})}$ and $\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}$ denote the mean vector and covariance matrix for the multivariate Gaussian density function $q^*(\beta_0, \boldsymbol{\beta})$, and $A_{q(\sigma_\varepsilon^2)}$ and $B_{q(\sigma_\varepsilon^2)}$ denote the shape and rate parameters for $q^*(\sigma_\varepsilon^2)$. The $\mu_{q(a_j)}$ denotes the mean parameter for

$q^*(a_i)$ and the shape parameter for $q^*(a_i)$ is equal to λ^2. Let $\boldsymbol{C} = [1, \boldsymbol{X}]$. Then convergence of Algorithm 4.2.1 can be monitored using the following expression for the lower bound on the marginal log-likelihood:

$$\begin{aligned}\log \underline{p}(y;q) &= \frac{p+1}{2} - p\,\log(2) + \frac{p-n}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2_{\beta_0}) \\ &\quad - \frac{\mu^2_{q(\beta_0)} + \sigma^2_{q(\beta_0)}}{2\sigma^2_{\beta_0}} + \frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\beta_0\beta)}| + \frac{p}{2}\log(\lambda^2) \\ &\quad + A_\varepsilon \log(B_\varepsilon) - \log\Gamma(A_\varepsilon) - \sum_{j=1}^{p} \frac{\lambda^2}{2\mu_{q(a_j)}} \\ &\quad - \left(A_\varepsilon + \frac{n+p}{2}\right)\log(B_{q(\sigma^2_\varepsilon)}) + \log\Gamma\left(A_\varepsilon + \frac{n+p}{2}\right).\end{aligned}$$

4.2.3 Choosing the Hyperparameter λ

The Bayesian Lasso method includes four hyperparameters: $\sigma^2_{\beta_0}$, A_ε, B_ε and λ. Setting the values of the hyperparameters is an important part of Bayesian inference. As general hyperparameters for the Normal prior and Inverse-Gamma prior, $\sigma^2_{\beta_0}$, A_ε, B_ε were set to be non-informative hyperparameters. As a special hyperparameter, λ appears in the Lasso definition and is used to penalize the regression coefficient $\boldsymbol{\beta}$. Selecting the value of λ is an important part of the algorithm for using the Lasso method. In the original Lasso, Tibshirani (1996) suggested the use of k-fold cross-validation to choose λ. For the Bayesian Lasso, Park and Casella (2008) offered two uniquely Bayesian alternatives: empirical Bayesian via marginal maximum likelihood, or use of an appropriate hyperprior. In this section, we develop two MFVB approaches to choose λ: variational EM for empirical Bayesian via marginal a maximum likelihood, and an MFVB inference for the Bayesian Lasso model with a specific prior on λ.

Empirical Bayesian via Variational EM

Compared with standard Bayesian methods, in which the hyperparameter is set before the data are observed, the empirical Bayesian method estimates the parameters of the prior distribution based on the observation. In empirical Bayesian analysis, the hyperparameters of the Bayesian hierarchical model are estimated by some estimation procedure. Usually, estimation is done via maximum likelihood, and then one proceeds by obtaining the posterior distribution as if this hyperparameter is fixed. Casella (2001) proposed a Monte Carlo EM algorithm that complements the Gibbs sampler and provides marginal maximum likelihood estimates of hyperparameters. Park and Casella (2008) used a Monte Carlo EM algorithm to deal with the Bayesian Lasso model. In this section, a variational EM is developed instead of the Monte Carlo EM algorithm to obtain the optimal hyperparameter λ.

Algorithm 4.2.2: Iterative scheme for a variational expectation-maximization algorithm for estimating the hyperparameter, λ.

Initialize λ^0;

Cycle: For $k = 1, ..., K$.

Use Algorithm (4.2.1) to obtain $q^*(\beta_0, \boldsymbol{\beta}, \boldsymbol{a}, \sigma_\varepsilon^2; \lambda^{(k-1)})$

E-step : $Q(\lambda|\lambda^{(k-1)}) = E_{q^*(\beta_0,\boldsymbol{\beta},\boldsymbol{a},\sigma_\varepsilon^2;\lambda^{(k-1)})}\{\ell(y; \beta_0, \boldsymbol{\beta}, \sigma_\varepsilon^2, \boldsymbol{a})\}$

M-step : $\lambda^{(k)} = \arg\max_{\lambda} Q(\lambda|\lambda^{(k-1)})$

The EM algorithm is an iterative method. It alternates between performing an expectation (E) step, which computes the expectation of the log-marginal-likelihood evaluated by using the current estimates of variables, and a maximization (M) step, which computes parameters by maximizing the expected log

marginal likelihood found on the E step.

The Variational Expectation-Maximization (VEM) algorithm 4.2.2 for estimating hyperparameters is a modified EM algorithm in the E-step. The variational optimal posterior distribution is used to obtain the expectation of the log likelihood using MFVB inference. In the M-step, the optimal value of the hyperparameter is computed by maximizing the expected log likelihood. Using the new hyperparameter, a new posterior distribution can be obtained by using MFVB inference again.

The Bayesian Lasso log-likelihood is:

$$\begin{aligned}
\ell(y;\beta_0,\boldsymbol{\beta},\sigma_\varepsilon^2,\boldsymbol{a}) &= \log\left\{\pi(y|\beta_0,\boldsymbol{\beta},\sigma_\varepsilon^2)\pi(\beta_0)\pi(\sigma_\varepsilon^2)\prod_{j=1}^{p}\pi(\beta_j|a_j)\pi(a_j|\lambda)\right\} \\
&= \log(\prod_{j=1}^{p}\pi(a_j|\lambda)) + \text{ term not depending on } \lambda \\
&= \sum_{j=1}^{p}\left\{\log(\lambda^2) - \frac{\lambda^2}{2a_i}\right\} + \text{ term not depending on } \lambda.
\end{aligned}$$

This leads to:

E-step:

$$\begin{aligned}
Q(\lambda|\lambda^{(k-1)}) &= E_{\boldsymbol{q}^*(\beta_0,\boldsymbol{\beta},\boldsymbol{a},\sigma_\varepsilon^2;\lambda^{(k-1)})}[\ell(y;\beta_0,\boldsymbol{\beta},\sigma_\varepsilon^2,\boldsymbol{a})] \\
&= p\,\log(\lambda^2) - \tfrac{1}{2}\lambda^2\sum_{j=1}^{p}\mu_{q(1/a_j)} + \text{constant}.
\end{aligned}$$

M-step:

$$\begin{aligned}
(\lambda^2)^{(k)} &= \arg\max_{\lambda} Q(\lambda|\lambda^{(k-1)}) \\
&= \frac{2p}{\sum_{j=1}^{p}\mu_{q(1/a_j)}}.
\end{aligned}$$

If λ is initialized at a small number such as 0.01, then the optimal posterior distribution in the first cycle is close to the optimal posterior distribution of ordinary least squares. The convergence of the EM algorithm is fast. However, if a large number (10^8) is used as the initial value of λ, then all the coefficients ($\beta_j,\ 1 \leq j \leq p$) will be close to 0 and convergence of the EM algorithms will be slow.

Prior Distribution in λ

Another method is to choose a prior distribution $\pi(\lambda)$ for λ with a uninformative hyperparameters. The adjusted Bayesian Lasso model takes the form:

$$\begin{aligned}
&\boldsymbol{y} \sim N(1_n\beta_0 + \boldsymbol{X}\boldsymbol{\beta}, \sigma_\varepsilon^2 \boldsymbol{I}_n),\\
&\beta_0 \sim N(0, \sigma_{\beta_0}^2),\\
&\beta_j | a_j \overset{\text{ind.}}{\sim} N(0, \frac{\sigma_\varepsilon^2}{a_j}), \quad j = 1, 2, \ldots, p,\\
&a_j | \gamma \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(1, \frac{1}{2\gamma}),\\
&\gamma \sim \text{Inverse-Gamma}(A_\gamma, B_\gamma),\\
&\sigma_\varepsilon^2 \sim \text{Inverse-Gamma}(A_\varepsilon, B_\varepsilon),
\end{aligned} \tag{4.7}$$

where $\gamma \equiv 1/\lambda^2$ has the Inverse Gamma distribution, and the A_γ, B_γ, $\sigma_{\beta_0}^2$, A_ε and $B_\varepsilon > 0$ are fixed non-informative hyperparameters.

We now seek a quick deterministic approximate inference procedure for (4.7) based on the MFVB. A tractable solution arises if we impose the product restriction:

$$q(\beta_0, \boldsymbol{\beta}, \boldsymbol{a}, \sigma_\varepsilon^2, \gamma) = q(\beta_0, \boldsymbol{\beta}) \prod_{j=1}^{p} q(a_j) q(\sigma_\varepsilon^2) q(\gamma).$$

Then, as shown in Appendix 4.B, the optimal q^* densities for the parameters in

model (4.7) take the form:

$$\begin{aligned} &q^*(\beta_0, \boldsymbol{\beta}) \text{ is a multivariate Gaussian density function,} \\ &q^*(a_j) \text{ is an Inverse Gaussian density function,} \\ &q^*(\gamma) \text{ is an Inverse Gamma density functions,} \\ &q^*(\sigma_\varepsilon^2) \text{ is an Inverse Gamma density functions.} \end{aligned} \tag{4.8}$$

Similarly, let $\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})}$ and $\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}$ denote the mean vector and covariance matrix for the multivariate Gaussian density function $q^*(\beta_0, \boldsymbol{\beta})$, $A_{q(\sigma_\varepsilon^2)}$ and $B_{q(\sigma_\varepsilon^2)}$ denote the shape and rate parameters for $q^*(\sigma_\varepsilon^2)$, a similar notation is used for the parameters in $q^*(\gamma)$, and let $\mu_{q(a_j)}$ and $\mu_{q(1/\gamma)}$ denote the mean and shape parameter for the Inverse Gaussian $q^*(a_i)$. Let $\boldsymbol{C} = [\mathbf{1}, \boldsymbol{X}]$ and

$$\boldsymbol{D}_{\mu_q(\boldsymbol{a})} = \begin{pmatrix} \mu_q(a_1) & 0 & \cdots & 0 \\ 0 & \mu_q(a_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mu_q(a_p) \end{pmatrix}.$$

Algorithm 4.2.3: Iterative scheme for obtaining the parameters of the optimal densities $q^*(\beta_0, \boldsymbol{\beta})$, $q^*(\gamma)$, $q^*(a_i)$ and $q^*(\sigma_\varepsilon^2)$ for the Bayesian Lasso model (4.7).

Initialize $\mu_{q(1/\sigma_\varepsilon^2)}$, $\mu_{q(a_j)}$ and $\mu_{q(1/a_j)}$, $j = 1, \ldots, p$;

Cycle

$$\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})} \leftarrow \left\{\mu_{q(1/\sigma_\varepsilon^2)}\boldsymbol{C}^T\boldsymbol{C} + \text{blockdiag}\left[(\sigma_{\beta_0}^2)^{-1}, \mu_{q(1/\sigma_\varepsilon^2)}\boldsymbol{D}_{\mu_q(\boldsymbol{a})}\right]\right\}$$

$$\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})} \leftarrow \mu_{q(1/\sigma_\varepsilon^2)}\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}\boldsymbol{C}^T\boldsymbol{y}$$

$$A_{q(\sigma_\varepsilon^2)} \leftarrow \frac{p+n}{2} + A_\varepsilon$$

$$B_{q(\sigma_\varepsilon^2)} \leftarrow B_\varepsilon + \tfrac{1}{2}\Big\{ \parallel \boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})} \parallel^2 + \text{tr}(\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}) + \text{tr}(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}^T\boldsymbol{D}_{\mu_q(\boldsymbol{a})} + \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\boldsymbol{D}_{\mu_q(\boldsymbol{a})})\Big\}$$

$$\mu_{q(1/\sigma_\varepsilon^2)} \leftarrow \frac{A_{q(\sigma_\varepsilon^2)}}{B_{q(\sigma_\varepsilon^2)}}$$

$$A_{q(\gamma)} \leftarrow A_\gamma + p$$

$$B_{q(\gamma)} \leftarrow B_\gamma + \tfrac{1}{2}\sum_{j=1}^{p}\mu_{q(1/a_i)}$$

$$\mu_{q(1/\gamma)} \leftarrow A_{q(\gamma)}/B_{q(\gamma)}$$

$$\mu_{q(a_j)} \leftarrow \sqrt{\frac{\mu_{q(1/\gamma)}}{(\mu_{q(\beta_j)}^2 + \sigma_{q(\beta_j)}^2)\mu_{q(1/\sigma_\varepsilon^2)}}} \quad j = 1, 2, \ldots, p$$

$$\boldsymbol{D}_{\mu_q(\boldsymbol{a})} \leftarrow \text{diag}(\mu_{q(a_1)}, \mu_{q(a_2)}, \ldots, \mu_{q(a_p)})$$

$$\mu_{q(1/a_j)} \leftarrow \frac{1}{\mu_{q(a_j)}} + \frac{1}{\mu_{q(1/\gamma)}}$$

until the increase in $\log \underline{p}(y; q)$ is negligible.

Then convergence of Algorithm 4.2.3 can be monitored using the following expression for the lower bound on the marginal log-likelihood:

$$\begin{aligned}
\log \underline{p}(y;q) \;=\; & \frac{p+1}{2} - p\,\log(2) + \frac{p-n}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2_{\beta_0}) \\
& -\frac{1}{2\sigma^2_{\beta_0}}(\mu^2_{q(\beta_0)} + \sigma^2_{q(\beta_0)}) \\
& +\frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}| - \frac{p}{2}\log(\mu_{q(1/\gamma)}) \\
& +A_\varepsilon \log(B_\varepsilon) - \log\,\Gamma(A_\varepsilon) \\
& +A_\gamma \log(B_\gamma) - \log\,\Gamma(A_\gamma) \\
& -(A_\gamma + p)\log(B_{q(\gamma)}) + \log\,\Gamma(A_\gamma + p) \\
& -\left(A_\varepsilon + \frac{n+p}{2}\right)\log\,B_{q(\sigma^2_\gamma)} + \log\,\Gamma\left(A_\varepsilon + \frac{n+p}{2}\right).
\end{aligned}$$

4.2.4 Diabetes Data

The diabetes data is a classical example used in Lasso research (Efron *et al.*, 2004; Park & Casella, 2008). The sample size is $n = 442$ and the number of predictor variables is $p = 10$. The response variable is a continuous index of disease progression one year after baseline, and the predictor variables include age (**age**), sex (**sex**), body mass index (**bmi**), average blood pressure (**map**) and six blood serum measurements (**tc**, **ldl**, **hdl**, **tch**, **ltg** and **glu**).

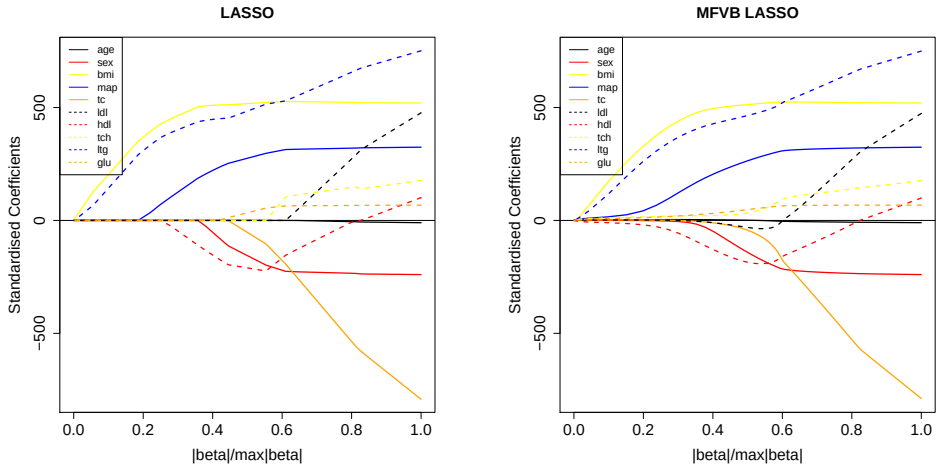

Figure 4.2: The trace plots of Lasso and MFVB Lasso for estimates of the diabetes data regression parameters.

Figure 4.2 compares the MFVB estimates with the ordinary Lasso estimates. The left panel shows the paths of the estimates as a function of their ℓ_1 norm relative to the ℓ_1 norm of the corresponding least squares estimate. The right panel shows the paths of posterior mean estimates using Algorithm 4.2.1. The paths of the MFVB Lasso estimates are similar in shape to the path of the ordinary Lasso estimates, but the paths of the MFVB Lasso estimates are smoother.

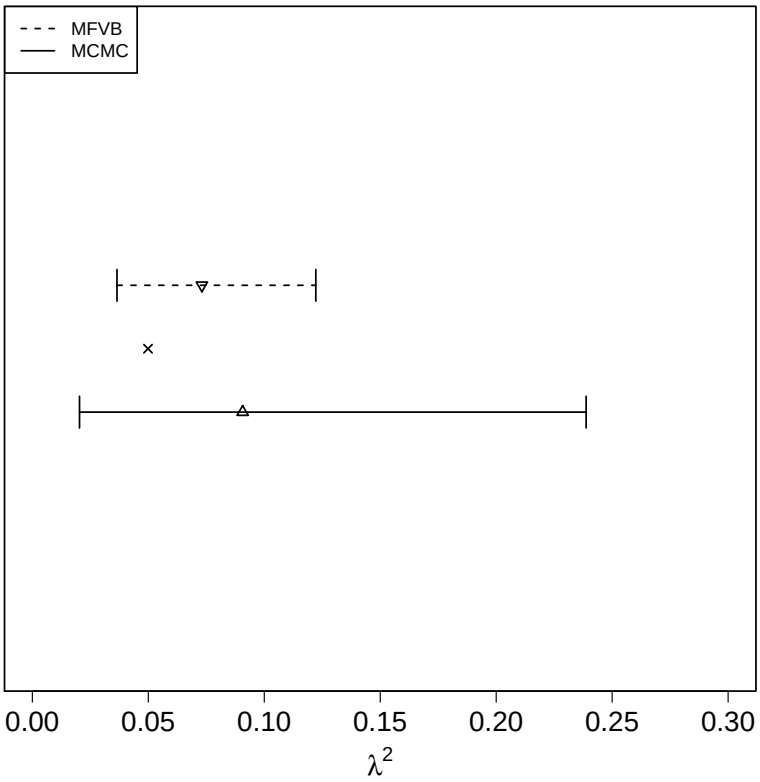

Figure 4.3: Estimates of the hyperparameter λ using empirical Bayesian via Variational EM ($\times$), Inverse-Gamma prior via MCMC with mean ($\triangle$) and Inverse-Gamma prior via MFVB inference with mean ($\bigtriangledown$), and corresponding 95% credible intervals for MCMC and MFVB.

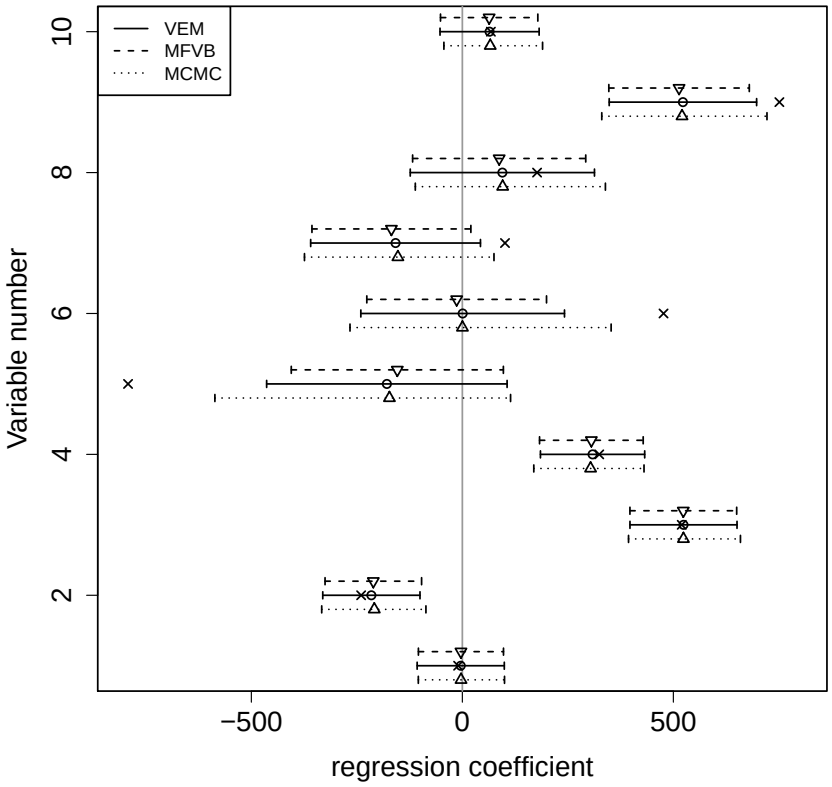

Figure 4.4: The posterior distribution of the Lasso regression estimates using an Inverse-Gamma prior via MCMC ($\triangle$), an Inverse-Gamma prior via MFVB ($\bigtriangledown$), an estimate of λ via VEM ($\bigcirc$), and corresponding 95% credible intervals. ($\times$) is the ordinary least squares estimate.

Figure 4.3 compares MCMC and MFVB for the estimation of λ^2. The empirical Bayesian method via the MFVB EM yields an optimal λ^2 of approximately 0.05. When λ^2 is given an Inverse Gamma prior distribution, the posterior distributions are obtained by MCMC and MFVB inference. In the MCMC method, the posterior mean for λ^2 is approximately 0.086, and a 95% posterior credible interval for λ^2 is approximately (0.020,0.226), and the corresponding values of MFVB are 0.073 and (0.036,0.122). The posterior medians and 95% credible intervals of coefficients are shown in Figure 4.4. The approximate posterior density function computed in

model (4.7) by using MFVB inference and MCMC for the diabetes data is shown in Figure 4.5. Good to excellent accuracy of MFVB inference is apparent for all posterior densities.

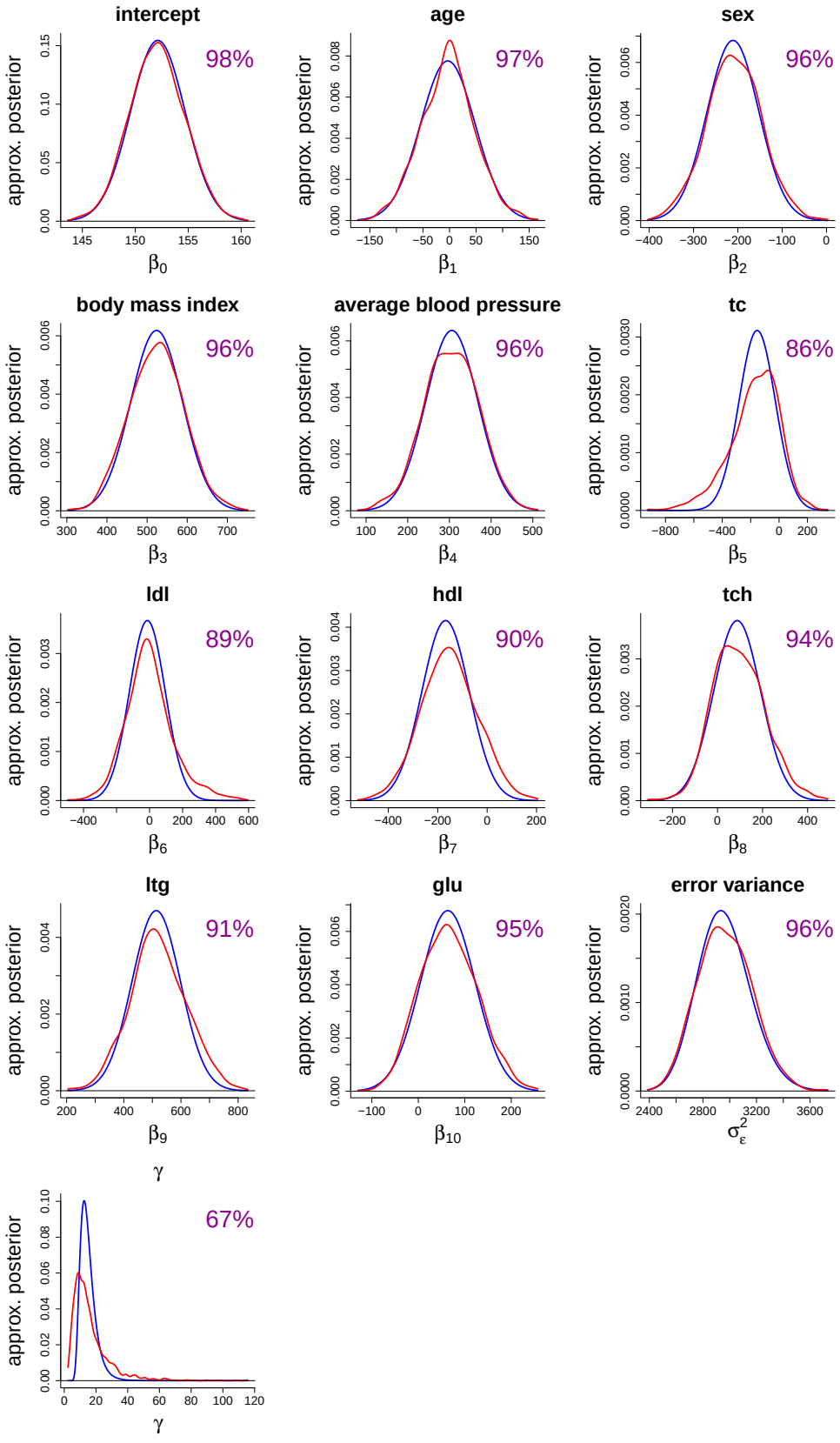

Figure 4.5: The approximate posterior density functions produced by MFVB (blue) and MCMC (orange) for fitting model (4.7) to the diabetes data. The percentages are the accuracies of the MFVB fit compared with the MCMC fit.

4.3 Bayesian Lasso in high-dimensional data

We have developed the MFVB algorithm for the Lasso linear regression model and to choose the value of the hyperparameter λ. In this section, I will modify Lasso regression model to fit the high-dimensional data and perform variable selection. High-dimensional data analysis has become increasingly frequent and important in diverse fields. In classical statistical analysis and model selection, the number of predictors (p) is a small value and always smaller than the number of observations (n). High-dimensional data analysis is used to deal with data with dimensions larger than those considered in classical multivariate analysis. Recently, researchers have become more interested in even larger dimension case, i.e. $p \gg n$. The computational costs of MFVB should be increased significantly to adapt to high-dimensional data. The correlation matrix with p dimensions need to be calculated during each iteration until the algorithms converge, which makes the increment of time or computational cost $O(p^2)$. To avoid a high-dimensional correlation matrix, we can factorize the variable $\boldsymbol{\beta}$ as:

$$q(\boldsymbol{\beta}) = q(\beta_1)q(\beta_2)...q(\beta_p).$$

In the MFVB algorithm, the variance of each variable β_j will be calculated respectively. This factorization is based on the "incorrect" assumption that β_1,...,β_p are independent, but MCMC shows that there is correlation between β_1,...,β_p. This creates a trade-off between accuracy and the computational costs of time and memory to fit the models for high-dimensional data.

Mean field variational Bayes inference

Consider the Bayesian Lasso model (4.7). For high-dimensional MFVB inference, we impose the product restriction

$$q(\beta_0, \boldsymbol{\beta}, \boldsymbol{a}, \sigma_\varepsilon^2, \gamma) = q(\beta_0)q(\beta_1)..q(\beta_p)q(a_1)...q(a_p)q(\sigma_\varepsilon^2)q(\gamma). \tag{4.9}$$

The optimal densities take the forms:

$$\begin{aligned}
&q^*(\beta_0) \text{ is a Gaussian density function,} \\
&q^*(\beta_j),\ 1 \leq j \leq p, \text{ are Gaussian density functions,} \\
&q^*(a_j),\ 1 \leq j \leq p, \text{ are Inverse Gaussian density functions,} \\
&q^*(\sigma_\varepsilon^2) \text{ is an Inverse Gamma density function,} \\
&q^*(\gamma) \text{ is an Inverse Gamma density function.}
\end{aligned} \tag{4.10}$$

Let $\mu_{q(\beta_0)}$ and $\sigma^2_{q(\beta_0)}$ denote the mean and variance for the Gaussian density function $q^*(\beta_0)$. A similar definition is used for the parameters in $q^*(\beta_j)$. Let $A_{q(\sigma^2_\varepsilon)}$ and $B_{q(\sigma^2_\varepsilon)}$ denote the shape and rate parameters for $q^*(\sigma^2_\varepsilon)$. A similar definition is used for the parameters in $q^*(\gamma)$. Let $\mu_{q(a_j)}$ and $\mu_{q(1/\gamma)}$ denote the mean and shape parameter for the Inverse Gaussian density function $q^*(a_i)$.

Algorithm 4.3.1: MFVB iterative scheme to obtain the parameters of the optimal densities $q^*(\beta_0)$, $q^*(\beta_j)$, $q^*(a_j)$, $q^*(\sigma_\varepsilon^2)$ and $q^*(\gamma)$ for the Bayesian Lasso model in the high dimensional case

Initialize $\mu_{q(1/\sigma_\varepsilon^2)}$, $\mu_{q(a_j)}$ and $\mu_{q(1/a_j)}$, $j = 1, \ldots, p$;

Cycle

$$
\begin{aligned}
\sigma^2_{q(\beta_0)} &\leftarrow \frac{1}{\mu_{q(1/\sigma_\varepsilon^2)} + \sigma^2_{\beta_0}} \\
\mu_{q(\beta_0)} &\leftarrow \sigma^2_{q(\beta_0)} \mu_{q(1/\sigma_\varepsilon^2)} \mathbf{1}^T (\boldsymbol{y} - \boldsymbol{X}\mu_{q(\beta)}) \\
\sigma^2_{q(\beta_j)} &\leftarrow \frac{1}{\mu_{q(1/\sigma_\varepsilon^2)} \|x_j\|^2 + \mu_{q(1/\sigma_\varepsilon^2)} \mu_{q(a_j)}} \\
\mu_{q(\beta_j)} &\leftarrow \sigma^2_{q(\beta_j)} \mu_{q(1/\sigma_\varepsilon^2)} \boldsymbol{X}_j^T \left(\boldsymbol{y} - \mathbf{1}\mu_{q(\beta_0)} - \boldsymbol{X}_{-j} \boldsymbol{\mu}_{q(\beta_{-j})} \right) \\
B_{q(\sigma_\varepsilon^2)} &\leftarrow B_\varepsilon + \frac{1}{2} \{ \|\boldsymbol{y} - \mathbf{1}\mu_{q(\beta_0)} - \sum_{j=1}^{P} \boldsymbol{X}_j \mu_{q(\beta_j)}\|^2 \\
&\qquad + n\sigma^2_{q(\beta_0)} + \sum_{j=1}^{p} \boldsymbol{X}_j^T \boldsymbol{X}_j \sigma^2_{q(\beta_j)} + \sum_{j=1}^{p} \mu_{q(a_j)} [\mu^2_{q(\beta_j)} + \sigma^2_{q(\beta_j)}] \} \\
A_{q(\sigma_\varepsilon^2)} &\leftarrow A_\varepsilon + \frac{n+p}{2} \\
\mu_{q(1/\sigma_\varepsilon^2)} &\leftarrow \frac{A_{q(\sigma_\varepsilon^2)}}{B_{q(\sigma_\varepsilon^2)}} \\
B_{q(\gamma)} &\leftarrow B_\gamma + \frac{1}{2} \sum_{j=1}^{p} \mu_{q(1/a_i)} \\
A_{q(\gamma)} &\leftarrow A_\gamma + p \\
\mu_{q(1/\gamma)} &\leftarrow \frac{A_{q(\gamma)}}{B_{q(\gamma)}} \\
\mu_{q(a_j)} &\leftarrow \sqrt{\frac{\mu_{q(1/\gamma)}}{\mu^2_{q(1/\sigma_\varepsilon^2)} [\mu^2_{q(\beta_j)} + \sigma^2_{q(\beta_j)}]}}
\end{aligned}
$$

until the increase in $\log \underline{p}(y; q)$ is negligible.

Convergence of Algorithm 4.3.1 can be monitored using the following expression for the lower bound on the marginal log-likelihood:

$$\begin{aligned}
\log p(y;q) &= \frac{p-n}{2}\log(2\pi) - p\,\log(2) + \frac{p+1}{2} \\
&\quad -\frac{1}{2}\log\,\sigma^2_{\beta_0} - \frac{1}{2\sigma^2_{\beta_0}}(\mu^2_{q(\beta_0)} + \sigma^2_{q(\beta_0)}) \\
&\quad +A_\varepsilon\log(B_\varepsilon) - \log\,\Gamma(A_\varepsilon) + A_\gamma\log(B_\gamma) - \log\,\Gamma(A_\gamma) \\
&\quad -\frac{p}{2}\log(\mu_{1/\gamma}) + \frac{1}{2}\log(\sigma^2_{q(\beta_0)}) + \frac{1}{2}\sum_{i=1}^{p}\log(\sigma^2_{q(\beta_j)}) \\
&\quad -(A_\gamma + p)\log(B_{q(\gamma)}) + \log\,\Gamma(A_\gamma + p) \\
&\quad -\left(A_\varepsilon + \frac{n+p}{2}\right)\log(B_{q(\gamma)}) + \log\,\Gamma\left(A_\varepsilon + \frac{n+p}{2}\right).
\end{aligned}$$

Figure 4.6 shows comparisons of posterior density functions from the MFVB high-dimensional Lasso algorithm 4.3.1 and the MCMC method to fit diabetes data. The different variances of posterior distributions obtain from the MFVB and MCMC methods. However, the means of the posterior distributions from the MFVB and MCMC methods are very close. Therefore, we can use Algorithm 4.3.1 to fit the Lasso model for high dimensional data when we more interested in the mean values of the coefficients.

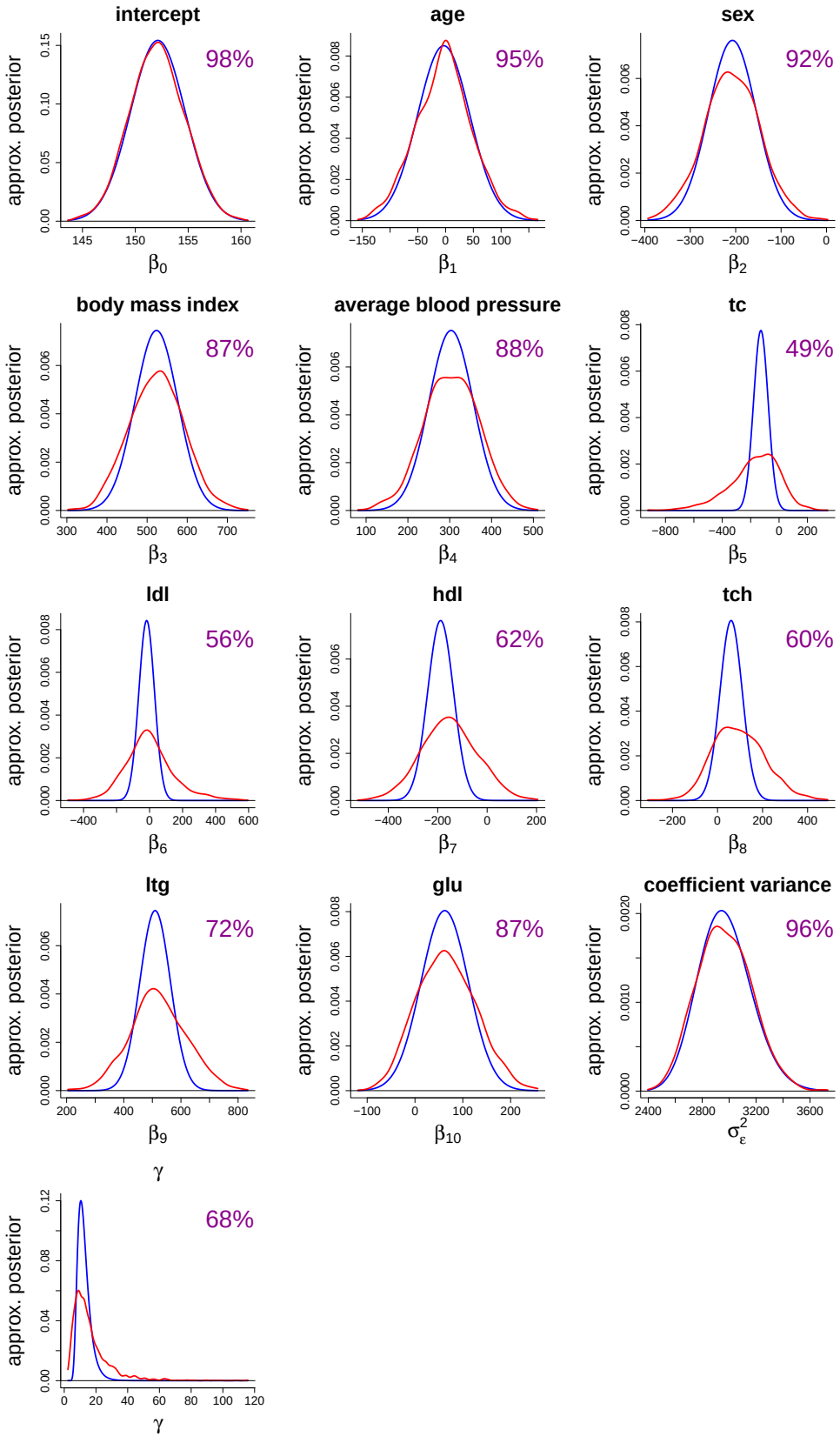

Figure 4.6: The approximate posterior density functions produced by MFVB (blue) using high-dimensional Lasso algorithm (4.3.1) and MCMC (orange) to the diabetes data. The percentages are the accuracies of the MFVB fit compared with the MCMC fit.

High-Dimensional Data Variable Selection

This study attempts to use the Bayesian Lasso method to select the variables for high-dimensional data. As is well known, the Bayesian Lasso method cannot give a coefficient that is exactly zero, so the a Z-value will be used to select the variables for the high-dimensional data, where the Z-value is defined as:

$$z = \frac{x - \mu}{\sigma},$$

where the μ is the mean of the population and σ is the standard deviation of the population.

Because the approximate posterior density function of coefficient of predictor, $q(\beta_j)$, is $N(\mu_{q(\beta_j)}, \sigma^2_{q(\beta_j)})$, the $Z_i = |\mu_{q(\beta_j)}|/\sigma_{q(\beta_j)}$ can be obtained and used to evaluate the corresponding predictor. A predictor in the model should be selected when $Z_i \geq 2$.

Similarly to the low-dimensional cases in Chapter 2 and 3, we consider the linear model form

$$y_i = \boldsymbol{X}_i\boldsymbol{\beta} + \varepsilon_i, \;\; 1 \leq i \leq n.$$

The high-dimensional data are simulated using the number of observation, n= 400, and the number of predictors of interest is p=1500. The $\boldsymbol{X}_i$, $1 \leq i \leq n$, are generated from a Multivariate Normal distribution, $N(\mathbf{0}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = . \begin{pmatrix} 1 & \rho_x & \cdots & \rho_x \\ \rho_x & 1 & \cdots & \rho_x \\ \vdots & \vdots & \ddots & \vdots \\ \rho_x & \rho_x & \cdots & 1 \end{pmatrix}.$$

Then, $\rho_x = 0$, 0.2, 0.5, 0.8 correspond to no correlation, low correlation, medium correlation and high correlation. The ε_i are generated from a Gaussian distribution

$N(\mathbf{0}, \sigma^2)$. Following Hastie *et al.* (2009), the standard deviation, σ, is chosen in each case so that the signal-to-noise ratio is equal to a fixed value. We also set the SNR equal to 1, 5 and 25 to represent the low, medium and high values. The true value of β_j is 3 for $j = 1, 301, 601, 901, 1201$ and 0 for remaining values of j.

Signal to noise ratio	ρ	x_1	x_{301}	x_{601}	x_{901}	x_{1201}
1	0	1	1	1	1	1
5	0	1	1	1	1	1
25	0	1	1	1	1	1
1	0.2	0.94	0.96	0.95	0.96	0.96
5	0.2	1	1	1	1	1
25	0.2	1	1	1	1	1
1	0.5	0.22	0.16	0.13	0.09	0.19
5	0.5	1	1	1	0.99	0.99
25	0.5	1	1	1	1	1
1	0.8	0	0	0	0	0
5	0.8	0.01	0	0	0.01	0.01
25	0.8	0.89	0.94	0.90	0.89	0.89

Table 4.1: Marginal probabilities that variables are selected for various ρ_x and SNRs by using MFVB inference for the Bayesian high-dimensional Lasso model.

Table 4.1 shows the simulation results of the marginal probabilities that variables, x_1, x_{301}, x_{601}, x_{901} and x_{1201} are selected for various ρ_x and SNR for a simulation size of 200. In the uncorrelated and low correlation cases, the performance of the variable selection is good for high-dimensional data with $p > n$ when using Algorithm 4.3.1. Unfortunately, the accuracy of the variable selection will decrease when the correlation among the predictors is stronger and/or the signal-to-noise ratio is lower.

4.4 Discussion

The Bayesian Lasso estimate for a linear regression parameter can be interpreted as a Bayesian posterior mode estimate when the regression parameters have Laplace priors. This chapter presents a successful development of MFVB methodology for the Bayesian Lasso model. Comparison between MFVB inference and MCMC inference reveals that the posterior density functions of MFVB inference are accurate. The VEM algorithm, which uses the MFVB method to replace MCMC in the E-step, can be used to choose λ. We have also developed the MFVB methodology for the Bayesian Lasso model with a prior distribution for λ, which also can estimate the value of λ.

A extended application for the Bayesian Lasso model are presented at the end of this chapter. The MFVB inference for a high-dimensional Bayesian Lasso linear regression model reduces the cost of computational time and memory but sacrifices accuracy to a certain degree. The simulation shows that the MFVB inference's ability to select variables is good for high-dimensional data with $p > n$ when there are not high correlation between predictors.

4.A Appendix: Derivation of Algorithm 4.2.1

4.A.1 Full conditionals

Full conditional for β_0 and $\boldsymbol{\beta}$

$$\log p(\beta_0, \boldsymbol{\beta}|\text{rest}) = -\frac{\|\boldsymbol{y} - \boldsymbol{C}\tilde{\boldsymbol{\beta}}\|^2}{2\sigma_\varepsilon^2} - \frac{1}{2}\tilde{\boldsymbol{\beta}}^T \text{blockdiag}\left[\sigma_{\beta_0}, \sigma_\varepsilon^2 \boldsymbol{D}_a^{-1}\right]\tilde{\boldsymbol{\beta}} + \text{const.}$$

where

$$\tilde{\boldsymbol{\beta}} = \begin{bmatrix}\beta_0 \\ \boldsymbol{\beta}\end{bmatrix}, \quad \boldsymbol{C} = [\mathbf{1}, \boldsymbol{X}] \quad \text{and} \quad \boldsymbol{D}_a = \begin{pmatrix} a_1 & 0 & \cdots & 0 \\ 0 & a_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_p \end{pmatrix},$$

Derivation:

$$\begin{aligned} p(\beta_0, \boldsymbol{\beta}|\text{rest}) &\propto p(\boldsymbol{y}|\beta_0, \boldsymbol{\beta}, \sigma_\varepsilon^2) p(\boldsymbol{\beta}|\boldsymbol{a}, \sigma_\varepsilon^2) p(\beta_0) \\ &= (2\pi)^{-n/2}\sigma_\varepsilon^{-n}\exp\left\{-\frac{\|\boldsymbol{y} - \mathbf{1}\beta_0 - \boldsymbol{X}\boldsymbol{\beta}\|^2}{2\sigma_\varepsilon^2}\right\} \\ &\quad \times (2\pi)^{-1/2}\sigma_{\beta_0}^{-1}\exp\left\{-\frac{\beta_0^2}{2\sigma_{\beta_0}^2}\right\} \\ &\quad \times \prod_{j=1}^{p}(2\pi)^{-p/2}\sigma_\varepsilon^{-1}a_j^{1/2}\exp\left\{-\frac{a_j\beta_j^2}{2\sigma_\varepsilon^2}\right\} \end{aligned}$$

Taking logarithms, we get

$$\begin{aligned} \log p(\beta_0, \boldsymbol{\beta}|\text{rest}) &= -\frac{\|\boldsymbol{y} - \mathbf{1}\beta_0 - \boldsymbol{X}\boldsymbol{\beta}\|^2}{2\sigma_\varepsilon^2} - \frac{\beta_0^2}{2\sigma_{\beta_0}^2} - \sum_{j=1}^{p}\frac{a_j\beta_j^2}{2\sigma_\varepsilon^2} + \text{const.} \\ &= -\frac{\|\boldsymbol{y} - \boldsymbol{C}\tilde{\boldsymbol{\beta}}\|^2}{2\sigma_\varepsilon^2} - \frac{1}{2}\tilde{\boldsymbol{\beta}}^T \text{blockdiag}\left[\sigma_{\beta_0}, \sigma_\varepsilon^2 \boldsymbol{D}_a^{-1}\right]\tilde{\boldsymbol{\beta}} + \text{const.} \end{aligned}$$

Full conditional for σ_ε^2

$$\begin{aligned}\log p(\sigma_\varepsilon^2|\text{rest}) &= \left(-1 - A_\varepsilon - \frac{p+n}{2}\right)\log\sigma_\varepsilon^2 \\ &\quad -\frac{1}{2\sigma_\varepsilon^2}\left(\|\boldsymbol{y} - \boldsymbol{C}\tilde{\boldsymbol{\beta}}\|^2 - \boldsymbol{\beta}^T\boldsymbol{D}_a\boldsymbol{\beta}\right) + \text{const.}\end{aligned}$$

Derivation:

$$\begin{aligned}p(\sigma_\varepsilon^2|\text{rest}) &\propto p(\boldsymbol{y}|\beta_0, \boldsymbol{\beta}, \sigma_\varepsilon^2)p(\boldsymbol{\beta}|\boldsymbol{a}, \sigma_\varepsilon^2)p(\sigma_\varepsilon^2) \\ &= (2\pi)^{-n/2}\sigma_\varepsilon^{-n}\exp\left\{-\frac{\|\boldsymbol{y} - \mathbf{1}\beta_0 - \boldsymbol{X}\boldsymbol{\beta}\|^2}{2\sigma_\varepsilon^2}\right\} \\ &\quad \times\prod_{j=1}^{p}(2\pi)^{-p/2}\sigma_\varepsilon^{-1}a_j^{1/2}\exp\left\{-\frac{a_j\beta_j^2}{2\sigma_\varepsilon^2}\right\} \\ &\quad \times\frac{B_\varepsilon^{A_\varepsilon}}{\Gamma(A_\varepsilon)}(\sigma_\varepsilon^2)^{-1-A_\varepsilon}\exp\left\{\frac{-B_\varepsilon}{\sigma_\varepsilon^2}\right\}.\end{aligned}$$

Taking logarithms, we get

$$\begin{aligned}\log p(\sigma_\varepsilon^2|\text{rest}) &= \left(-1 - A_\varepsilon - \frac{p+n}{2}\right)\log\sigma_\varepsilon^2 \\ &\quad -\frac{1}{2\sigma_\varepsilon^2}\left(\|\boldsymbol{y} - \boldsymbol{C}\tilde{\boldsymbol{\beta}}\|^2 - \boldsymbol{\beta}^T\boldsymbol{D}_a\boldsymbol{\beta}\right) + \text{const.}\end{aligned}$$

Full conditional for a_j, $1 \leq j \leq p$

Derivation:

$$\begin{aligned}p(a_j|\text{rest}) &\propto p(\beta_j|a_j, \sigma_\varepsilon^2)p(a_j) \\ &= (2\pi)^{-p/2}\sigma_\varepsilon^{-1}a_j^{1/2}\exp\left\{-\frac{a_j\beta_j^2}{2\sigma_\varepsilon^2}\right\} \\ &\quad \times\frac{\lambda^2}{2}{a_j}^{-2}\exp\left\{-\frac{\lambda^2}{2a_j}\right\}\end{aligned}$$

Taking logarithms, we get

$$\log p(a_j|\text{rest}) \;=\; -\frac{3}{2}\log(a_j) - \frac{a_j\beta_j^2}{2\sigma_\varepsilon^2} - \frac{\lambda^2}{2a_j} + \text{const.}$$

4.A.2 Optimal q^* densities

Expressions for $q^*(\beta_0, \boldsymbol{\beta})$, $\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})}$ and $\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}$

$$q^*(\beta_0, \boldsymbol{\beta}) \sim N(\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})}, \boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}),$$

$$\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})} = \left\{\boldsymbol{C}^T\boldsymbol{C}\mu_{q(1/\sigma_\varepsilon^2)} + \text{blockdiag}\left[(\sigma_{\beta_0}^2)^{-1}, \mu_{q(1/\sigma_\varepsilon^2)}\boldsymbol{D}_{\mu_q(\boldsymbol{a})}\right]\right\},$$

and

$$\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})} = \mu_{q(1/\sigma_\varepsilon^2)}\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}\boldsymbol{C}^T\boldsymbol{y},$$

where

$$\boldsymbol{D}_{\mu_q(\boldsymbol{a})} = \begin{pmatrix} \mu_q(a_1) & 0 & \cdots & 0 \\ 0 & \mu_q(a_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mu_q(a_p) \end{pmatrix}.$$

Derivation:

$$\begin{aligned} \log q^*(\beta_0, \boldsymbol{\beta}) &= E_q\left[\log p(\beta_0, \boldsymbol{\beta}|\text{rest})\right] \\ &= -\frac{1}{2}\tilde{\boldsymbol{\beta}}^T \text{blockdiag}\left[(\sigma_{\beta_0}^2)^{-1}, \mu_{q(1/\sigma_{\boldsymbol{\beta}}^2)}\boldsymbol{D}_{\mu_q(\boldsymbol{a})}\right]\tilde{\boldsymbol{\beta}} \\ &\quad -\frac{1}{2}\mu_{q(1/\sigma_\varepsilon^2)}\|\boldsymbol{y} - \boldsymbol{C}\tilde{\boldsymbol{\beta}}\|^2 + \text{const.} \\ &= \left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})}\right)^T \boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}^{-1}\left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})}\right) + \text{const.} \end{aligned}$$

Therefore,

$$q^*(\beta_0, \boldsymbol{\beta}) \;=\; \exp\left\{\left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})}\right)^T \boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}^{-1}\left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})}\right) + \text{const}\right\}.$$

The results then follow from definition 1.17 of the Multivariate Normal distribution.

Expressions for $q^*(\sigma_\varepsilon^2)$ and $\mu_{q(1/\sigma_\varepsilon^2)}$

$$q^*(\sigma_\varepsilon^2) \sim \text{Inverse-Gamma}(A_{q(\sigma_\varepsilon^2)}, B_{q(\sigma_\varepsilon^2)}),$$

and

$$\mu_{q(1/\sigma_\varepsilon^2)} = \frac{A_{q(\sigma_\varepsilon^2)}}{B_{q(\sigma_\varepsilon^2)}},$$

where

$$\begin{aligned} A_{q(\sigma_\varepsilon^2)} &= \frac{p+n}{2} + A_\varepsilon, \\ B_{q(\sigma_\varepsilon^2)} &= B_\varepsilon + \frac{1}{2}\Big\{ \parallel \boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})} \parallel^2 + \text{tr}(\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}) \\ &\quad + \text{tr}(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}^T\boldsymbol{D}_{\mu_q(\boldsymbol{a})} + \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\boldsymbol{D}_{\mu_q(\boldsymbol{a})})\Big\}. \end{aligned}$$

Derivation:

$$\begin{aligned} \log q^*(\sigma_\varepsilon^2) &= E_q\left[\log p(\sigma_\varepsilon^2|\text{rest})\right] \\ &= (-1 - A_\varepsilon - \frac{p+n}{2})\log(\sigma_\varepsilon^2) \\ &\quad - \frac{2B_\varepsilon + E_q[\|\boldsymbol{y} - \boldsymbol{C}\tilde{\boldsymbol{\beta}}\|^2 + \boldsymbol{\beta}^T\boldsymbol{D}_a\boldsymbol{\beta}}{2\sigma_\varepsilon^2} + \text{const.} \\ &= (-1 - A_\varepsilon - \frac{p+n}{2})\log(\sigma_\varepsilon^2) \\ &\quad - \frac{2B_\varepsilon + \parallel \boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})} \parallel^2 + \text{tr}(\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})})}{2\sigma_\varepsilon^2} \\ &\quad - \frac{\text{tr}(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}^T\boldsymbol{D}_{\mu_q(\boldsymbol{a})} + \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\boldsymbol{D}_{\mu_q(\boldsymbol{a})})}{2\sigma_\varepsilon^2} \end{aligned}$$

The results then follow from definition (1.23) and result (1.7) for the Inverse-

Gamma distribution.

Expressions for $q^*(a_j)$

$$q^*(a_j) \sim \text{Inverse-Gaussian}(\mu_{q(a_j)}, \lambda^2)$$

and

$$\mu_{q(a_j)} = \sqrt{\frac{\lambda^2}{(\mu^2_{q(\beta_j)} + \sigma^2_{q(\beta_j)})\mu_{q(1/\sigma^2_\varepsilon)}}}.$$

Derivation:

$$\begin{aligned}
\log q^*(a_j) &= E_q\left[\log p(a_j|\text{rest})\right] \\
&= -\frac{3}{2}\log(a_j) - \frac{\lambda^2}{2a_j} - a_j E_q\left[\frac{\beta_j^2}{2\sigma_\varepsilon^2}\right] + \text{const.} \\
&= -\frac{3}{2}\log(a_j) - \frac{\lambda^2}{2a_j} - a_j\frac{(\mu^2_{q(\beta_j)} + \sigma^2_{q(\beta_j)})\mu_{q(1/\sigma^2_\varepsilon)}}{2} + \text{const.}
\end{aligned}$$

Then,

$$q^*(a_j) \propto a^{-3/2}\exp\left\{-\frac{\lambda^2}{2a_j} - a_j\frac{(\mu^2_{q(\beta_j)} + \sigma^2_{q(\beta_j)})\mu_{q(1/\sigma^2_\varepsilon)}}{2}\right\}$$

The results then follow from definition 1.18 for the Inverse-Gaussian distribution.

4.A.3 Derivation of lower bound

We note that

$$\begin{aligned}\log \underline{p}(y;q) &= E[\log\, p(y,\beta_0,\boldsymbol{\beta},\boldsymbol{a},\sigma_\varepsilon^2) - q(\beta_0,\boldsymbol{\beta},\boldsymbol{a},\sigma_\varepsilon^2)] \\ &= E[\log\, p(y|\beta_0,\boldsymbol{\beta},\sigma_\varepsilon^2) + \log\, p(\beta_0) \\ &\quad +\log\, p(\boldsymbol{\beta}|\boldsymbol{a},\sigma_\varepsilon^2) + \log\, p(\boldsymbol{a}) + \log\, p(\sigma_\varepsilon^2) \\ &\quad -\log\, q(\beta_0,\boldsymbol{\beta}) - \log\, q(\sigma_\varepsilon) - \log\, q(\boldsymbol{a})].\end{aligned}$$

Firstly,

$$\begin{aligned}&E_q\{\log\, p(y|\beta_0,\boldsymbol{\beta},\ \sigma_\varepsilon^2)\} \\ &\quad = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\mu_{q(\log\sigma_\varepsilon^2)} \\ &\qquad -\frac{1}{2}\mu_{q(1/\sigma_\varepsilon^2)}\{\parallel \boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})} \parallel^2 + \mathrm{tr}(\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})})\}.\end{aligned}$$

Secondly,

$$\begin{aligned}&E_q\{\log\, p(\sigma_\varepsilon^2) - \log\, q(\sigma_\varepsilon^2)\} \\ &\quad = A_\varepsilon\log(B_\varepsilon) - \log\,\Gamma(A_\varepsilon) \\ &\qquad -(A_\varepsilon + \frac{n+p}{2})\log(B_{q(\sigma_\varepsilon^2)}) + \log\,\Gamma(A_\varepsilon + \frac{n+p}{2}) \\ &\qquad +\frac{n+p}{2}\mu_{q(\log\sigma_\varepsilon^2)} + \frac{1}{2}\mu_{q(1/\sigma_\varepsilon^2)}\big\{\parallel \boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})} \parallel^2 + \mathrm{tr}(\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}) \\ &\qquad\qquad +\mathrm{tr}(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}^T\boldsymbol{D}_{\mu_q(\boldsymbol{a})} + \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\boldsymbol{D}_{\mu_q(\boldsymbol{a})})\big\}\end{aligned}$$

Thirdly,

$$\begin{aligned}
&E_q\{\log\, p(\beta_0) + \log\, p(\boldsymbol{\beta}|a, \sigma_\varepsilon^2) - \log\, q(\beta_0\beta)\} \\
&\quad = -\frac{1}{2}\log(\sigma_{\beta_0}^2) - \frac{1}{2\sigma_{\beta_0}^2}(\mu_{q(\beta_0)}^2 + \sigma_{q(\beta_0)}^2) \\
&\qquad + \frac{p+1}{2} + \frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}| \\
&\qquad - \frac{p}{2}\mu_{q(\log\sigma_\varepsilon^2)} + \frac{1}{2}\sum_{j=1}^{P}\mu_{q(\log a_j)} \\
&\qquad - \frac{1}{2}\mu_{q(1/\sigma_\varepsilon^2)}\left\{\mathrm{tr}(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}^T\boldsymbol{D}_{\mu_q(\boldsymbol{a})} + \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\boldsymbol{D}_{\mu_q(\boldsymbol{a})})\right\}.
\end{aligned}$$

Fourthly,

$$\begin{aligned}
&E_q\{\log\, p(\boldsymbol{a}) - \log\, q(\boldsymbol{a})\} \\
&\quad = \sum_{j=1}^{p}\left\{-\log(2) + \frac{1}{2}\log(\lambda^2) - \frac{1}{2}\mu_{q(\log a_j)} + \frac{1}{2}\log(2\pi) - \frac{\lambda^2}{2\mu_{q(a_j)}}\right\}
\end{aligned}$$

Substitution of these gives the lower bound:

$$\begin{aligned}
\log \underline{p}(y;q) \;=\; & \frac{p+1}{2} - p\,\log(2) + \frac{p-n}{2}\log(2\pi) - \frac{1}{2}\log(\sigma_{\beta_0}^2) \\
& - \frac{\mu_{q(\beta_0)}^2 + \sigma_{q(\beta_0)}^2}{2\sigma_{\beta_0}^2} + \frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\beta_0\beta)}| + \frac{p}{2}\log(\lambda^2) \\
& + A_\varepsilon\log(B_\varepsilon) - \log\,\Gamma(A_\varepsilon) - \sum_{j=1}^{p}\frac{\lambda^2}{2\mu_{q(a_j)}} \\
& - \left(A_\varepsilon + \frac{n+p}{2}\right)\log(B_{q(\sigma_\varepsilon^2)}) + \log\,\Gamma\left(A_\varepsilon + \frac{n+p}{2}\right).
\end{aligned}$$

4.B Appendix: Derivation of Algorithm 4.2.3

In Algorithm 4.2.3, the MFVB calculations for β_0, $\boldsymbol{\beta}$, $\boldsymbol{a}$ and σ_ε^2 are similar to those in Algorithm 4.2.1. We obtain them by using

$$\mu_{q(1/\gamma)} \text{ to replace } \lambda^2.$$

Therefore, I only show the derivation for $\boldsymbol{a}$.

4.B.1 Full conditionals

Full conditional for γ

Derivation:

$$\begin{aligned} p(\gamma|\text{rest}) &\propto p(\gamma)\prod_{j=1}^{p} p(a_j|\gamma) \\ &= \prod_{j=1}^{p} \frac{1}{2\gamma} {a_j}^{-2} \exp\left\{-\frac{1}{2a_j\gamma}\right\} \\ &\quad \times \frac{B_\gamma^{A_\gamma}}{\Gamma(A_\gamma)} \gamma^{-A_\gamma-1} \exp\{-\frac{B_\gamma}{\gamma}\} \end{aligned}$$

Taking logarithms, we get:

$$\log p(\boldsymbol{a}|\text{rest}) = (-A_\gamma - p - 1)\log\gamma - \frac{1}{\gamma}\left(B_\gamma + \sum_{j=1}^{p}\frac{1}{2a_j}\right) + \text{const.}$$

4.B.2 Optimal q^* densities

Expressions for $q^*(\gamma)$ and $\mu_{q(1/\gamma)}$

$$q^*(\gamma) \sim \text{Inverse-Gamma}(A_{q(\gamma)}, B_{q(\gamma)}) \text{ and } \mu_{q(1/\gamma)} = \frac{A_{q(\gamma)}}{B_{q(\gamma)}}$$

where

$$A_{q(\gamma)} = A_\gamma + p \text{ and } B_{q(\gamma)} = B_\gamma + \tfrac{1}{2}\sum_{j=1}^{p} \mu_{q(1/a_j)}$$

Derivation:

$$\begin{aligned} \log q^*(\gamma) &= E_q\left[\log p(\gamma|\text{rest})\right] \\ &= (-A_\gamma - p - 1)\log\gamma - \frac{1}{\gamma}\left(B_\gamma + \sum_{j=1}^{p}\frac{\mu_{q(1/a_j)}}{2}\right) + \text{const.} \end{aligned}$$

The results then follow from definition (1.23) and result (1.7) for the Inverse-Gamma distribution.

4.B.3 Derivation of lower bound

We note that

$$\begin{aligned} \log \underline{p}(y;q) &= p(\boldsymbol{y}, \beta_0, \boldsymbol{\beta}, \boldsymbol{a}, \gamma, \sigma_\varepsilon^2) - \log q(\beta_0, \boldsymbol{\beta}, \boldsymbol{a}, \gamma, \sigma_\varepsilon^2)\} \\ &= E[\log\, p(\boldsymbol{y}|\beta_0, \boldsymbol{\beta}, \sigma_\varepsilon^2) + \log\, p(\beta_0) + \log\, p(\boldsymbol{\beta}|a, \sigma_\varepsilon^2) \\ &\quad +\log(\boldsymbol{a}|\gamma) + \log(\sigma_\varepsilon^2) + \log(\gamma) \\ &\quad -\log\, q(\beta_0, \boldsymbol{\beta}) - \log\, q(\boldsymbol{a}) - \log\, q(\gamma) - \log\, q(\sigma_\varepsilon^2)]. \end{aligned}$$

Firstly,

$$\begin{aligned} &E_q\{\log\, p(y|\beta_0, \boldsymbol{\beta},\ \sigma_\varepsilon^2)\} \\ &\quad = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\mu_{q(\log\sigma_\varepsilon^2)} \\ &\quad\quad -\frac{1}{2}\mu_{q(1/\sigma_\varepsilon^2)}\{\parallel \boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})} \parallel^2 + \text{tr}(\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})})\}. \end{aligned}$$

Secondly,

$$
\begin{aligned}
&E_q\{\log\, p(\sigma_\varepsilon^2) - \log\, q(\sigma_\varepsilon^2)\} \\
&\quad = A_\varepsilon \log(B_\varepsilon) - \log\,\Gamma(A_\varepsilon) \\
&\qquad -(A_\varepsilon + \frac{n+p}{2})\log(B_{q(\sigma_\varepsilon^2)}) + \log\,\Gamma(A_\varepsilon + \frac{n+p}{2}) \\
&\qquad +\frac{n+p}{2}\mu_{q(\log\sigma_\varepsilon^2)} + \frac{1}{2}\mu_{(1/\sigma_\varepsilon^2)}\big\{\,\|\,\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\beta_0,\boldsymbol{\beta})}\,\|^2 + \mathrm{tr}(\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}) \\
&\qquad\qquad +\mathrm{tr}(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}^T\boldsymbol{D}_{\mu_q(\boldsymbol{a})} + \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\boldsymbol{D}_{\mu_q(\boldsymbol{a})})\big\}.
\end{aligned}
$$

Thirdly,

$$
\begin{aligned}
&E_q\{\log\, p(\beta_0) + \log\, p(\boldsymbol{\beta}|a,\sigma_\varepsilon^2) - \log\, q(\beta_0\beta)\} \\
&\quad = -\frac{1}{2}\log(\sigma_{\beta_0}^2) - \frac{1}{2\sigma_{\beta_0}^2}(\mu_{q(\beta_0)}^2 + \sigma_{q(\beta_0)}^2) \\
&\qquad +\frac{p+1}{2} + \frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}| \\
&\qquad -\frac{p}{2}\mu_{q(\log\sigma_\varepsilon^2)} + \frac{1}{2}\sum_{j=1}^{P}\mu_{q(\log a_j)} \\
&\qquad -\frac{1}{2}\mu_{(1/\sigma_\varepsilon^2)}\big\{\mathrm{tr}(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}^T\boldsymbol{D}_{\mu_q(\boldsymbol{a})} + \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\boldsymbol{D}_{\mu_q(\boldsymbol{a})})\big\}.
\end{aligned}
$$

Fourthly,

$$
\begin{aligned}
&E_q\{\log\, p(\boldsymbol{a}) - \log\, q(\boldsymbol{a})\} \\
&\quad = \sum_{j=1}^{p}\big\{ -\log(2) - \mu_{q(\log\gamma)} - \mu_{q(\log a_j)} + \frac{1}{2}\log(2\pi) \\
&\qquad\qquad -\frac{1}{2}\log(\mu_{q(1/\gamma)}) - \frac{\mu_{q(1/\gamma)}}{2\mu_{q(1/a_j)}}\big\}.
\end{aligned}
$$

Fifthly,

$$
\begin{aligned}
&E_q\{\log\, p(\gamma) - \log\, q(\gamma)\} \\
&\quad = A_\gamma \log(B_\gamma) - \log\,\Gamma(A_\gamma) + \mu_{q(\log\,\gamma)} \\
&\qquad -(A_\gamma + P)\log(B_{q(\gamma)}) + \log\,\Gamma(A_\gamma + P) + \frac{1}{2}\sum_{j=1}^{p} \mu_{q(1/a_j)}.
\end{aligned}
$$

Substitution of those gives the lower bound expression:

$$
\begin{aligned}
\log \underline{p}(y;q) \;=\; & \frac{p+1}{2} - p\,\log(2) + \frac{p-n}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2_{\beta_0}) \\
& -\frac{1}{2\sigma^2_{\beta_0}}(\mu^2_{q(\beta_0)} + \sigma^2_{q(\beta_0)}) \\
& +\frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\beta_0,\boldsymbol{\beta})}| - \frac{p}{2}\log\mu_{q(1/\gamma)} \\
& +A_\varepsilon \log(B_\varepsilon) - \log\,\Gamma(A_\varepsilon) \\
& +A_\gamma \log(B_\gamma) - \log\,\Gamma(A_\gamma) \\
& -(A_\gamma + p)\log(B_{q(\gamma)}) + \log\,\Gamma(A_\gamma + p) \\
& -\left(A_\varepsilon + \frac{n+p}{2}\right)\log(B_{q(\sigma^2_\gamma)}) + \log\,\Gamma\left(A_\varepsilon + \frac{n+p}{2}\right).
\end{aligned}
$$

Chapter 5

Using Infer.NET for Statistical Analyses[1]

5.1 Introduction

`Infer.NET` (Minka *et al.*, 2014) is a new computational framework for approximate Bayesian inference in hierarchical Bayesian models. The first beta version of `Infer.NET` was released in December 2008. `Infer.NET` can be downloaded from www.research.microsoft.com/infernet. At the time of this writing, the current version of `Infer.NET` is 2.6 and all advice given in this article is based on that version. Since `Infer.NET` is in its infancy, it is anticipated that new and improved versions will be released quite regularly in the coming years.

Over the past 20 years, the `BUGS` (Bayesian inference Using Gibbs Sampling) package has been the most popular software for the Bayesian analysis of complex statistical models using Markov chain Monte Carlo (MCMC) methods. `Infer.NET` is similar to `BUGS` in that both facilitate the fitting of hierarchical Bayesian models.

[1]This chapter is based on: Wang, S.S.J. and Wand, M.P. Using Infer.NET for statistical analyses. *The American Statistician*, 65, 2 (2011), 115-126.

They differ in their methods for approximate inference. `BUGS` uses Markov chain Monte Carlo (MCMC) samples from the posterior distributions of parameters of interest. `Infer.NET` instead uses deterministic approximation methods, known as variational message passing (VMP) (Winn & Bishop, 2005) and expectation propagation (EP) (Minka, 2001; Kim & Wand 2016) to approximate posterior distributions. Deterministic approximate inference methods have the advantage of being quite fast in comparison with the MCMC method, and they do not require laborious convergence checks. However, they can be considerably less accurate than MCMC with the latter having the advantage of improved accuracy through larger samples. `Infer.NET` has a Gibbs sampling option, which means that it can also perform MCMC-based approximate inference. However, this is for a much narrower class of models compared with `BUGS`.

Variational message passing (VMP) is a special case of MFVB. VMP sends messages between nodes in the Graphical model and updates posterior beliefs using local operations at each node. Each such update increases a lower bound on the log evidence (marginal likelihood).

Expectation propagation (EP) is a different class of deterministic approximation methods. An early reference is Minka (2001), although similar approaches such as assumed density filtering and moment matching have a longer history. For certain models, EP has been seen to achieve greater accuracy than VMP (e.g. Bishop, 2006, Section 10.7.1). There are fewer models that admit EP analytic solutions compared to VMP.

`Infer.NET` can be used from any of the so-called .NET languages, a family that includes C#, C++, Visual Basic, and Iron Python. Unfortunately, however, there are no R packages, which can provide an interface to `Infer.NET` and allow users to analyse Bayesian models by using `Infer.NET` in R. Hence I wrote a tool, named *InferNETSupport*, to perform `Infer.NET` within the R environmen-

t. In this chapter, all the model fittings were done in the R environment using *InferNETSupport*.

In this chapter, we introduce the use of `Infer.NET` for statistical analyses via four simple examples in Section 2. Five advanced examples are described in Section 3. Section 4 compares `Infer.NET` with `BUGS`. Section 5 presents a summary.

5.2 Simple Examples

We start with four examples involving simple Bayesian models. The first of these is Bayesian simple linear regression. We then describe extensions to binary responses, and to random effects. Our last example in this section is concerned with the classical finite normal mixture fitting problem. The simplicity of the examples allows the essential aspects of `Infer.NET` to be delineated more clearly.

All continuous variables are first transformed to the unit interval and weakly informative hyperparameter choices are used. The resulting approximate posterior densities are then back-transformed to the original units.

5.2.1 Simple linear regression

The first example is the Bayesian simple linear regression model

$$\begin{aligned} &y_i|\beta_0,\beta_1,\tau_\epsilon \sim N(\beta_0+\beta_1 X_i,\tau_\epsilon^{-1}), \quad 1\le i\le n,\\ &\beta_0,\beta_1 \sim N(0,\sigma_\beta^2),\\ &\tau_\epsilon \sim \text{Gamma}(A,B), \end{aligned} \tag{5.1}$$

where A, B and σ_β^2 are hyperparameters to be specified by the analyst. The joint posterior density of the model parameters $q(\beta_0,\beta_1,\tau_\epsilon|y)$ does not have a closed

form expression and `Infer.NET` will fit the product density approximation

$$q(\beta_0, \beta_1, \tau_\epsilon|y) = q_{\beta_0}(\beta_0)q_{\beta_1}(\beta_1)q_{\tau_\epsilon}(\tau_\epsilon) \tag{5.2}$$

or

$$q(\beta_0, \beta_1, \tau_\epsilon|y) = q_{\beta_0,\beta_1}(\beta_0, \beta_1)q_{\tau_\epsilon}(\tau_\epsilon). \tag{5.3}$$

Factorization (5.2) assumes the regression coefficients β_0 and β_1 have independent posterior densities. For the product restriction (5.2), the prior distribution for β_0 and β_1 are specified via:

```
Variable<double> beta0 = Variable.GaussianFromMeanAndVariance
                         (0.0,sigsqBeta).Named("beta0");
Variable<double> beta1 = Variable.GaussianFromMeanAndVariance
                         (0.0,sigsqBeta).Named("beta1");
```

Let $\boldsymbol{\beta} = [\beta_0, \beta_1]^T$, where the alternative expression of model (5.1) that matches the product restriction (5.3) is given by

$$\begin{aligned} &y_i|\boldsymbol{\beta}, \tau_\epsilon \sim N(\boldsymbol{\beta}^T\boldsymbol{X}_i, \tau_\epsilon^{-1}), \quad 1 \leq i \leq n, \\ &\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma_\beta^2\boldsymbol{I}), \\ &\tau_\epsilon \sim \text{Gamma}(A, B). \end{aligned} \tag{5.4}$$

The prior for $\boldsymbol{\beta}$ is specified via:

```
PositiveDefiniteMatrix SigmaBeta =
        PositiveDefiniteMatrix.IdentityScaledBy(2,sigsqBeta);
Variable<Vector> beta =
        Variable.VectorGaussianFromMeanAndVariance(
```

```
        new Vector(new double[]{0.0,0.0}),
        SigmaBeta).Named("beta");
```

We set $\sigma_\beta^2 = 10^8$ and $A = B = 0.01$ and fit the simple linear regression model for data on the age and price of $n = 39$ Mitsubishi cars (Smith, 1998). Figure 5.2 shows the approximate posterior densities obtained from `Infer.NET`. The fitted regression line, point-wise credible intervals and Bayesian prediction intervals are shown in Figure 5.1.

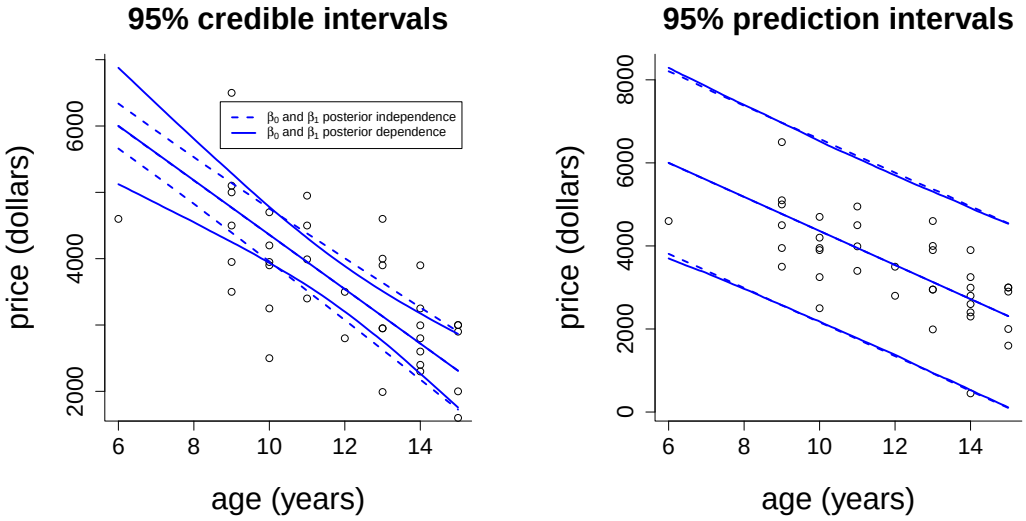

Figure 5.1: Fitted regression line, pointwise 95% credible intervals and pointwise 95% Bayesian prediction intervals for data on the age and price of 39 Mitsubishi cars. (source: Smith, 1998)

The error variance posterior approximation is unaffected by the type of variational Bayes restriction, but this is far from the case for the regression coefficients. The comparisons with the accurate MCMC-based posterior approximations, given in Figure 5.2, demonstrate that variational Bayes approximation (5.3) is quite accurate, but that variational Bayes approximation (5.2) is poor. The good performance of (5.3) is to be expected for diffuse independent priors because of the

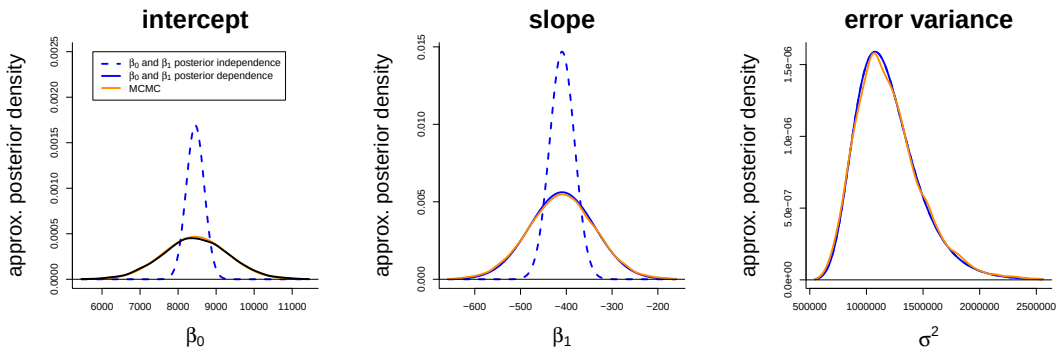

Figure 5.2: MFVB approximate posterior density functions produced by `Infer.NET` for the simple linear regression fit to the Mitsubishi car price/age data.

orthogonality between β and τ in likelihood-based inference. However, β_0 and β_1 are far from orthogonal, and this affects the accuracy of (5.2). In particular, both variational Bayes approximation (5.3) and the MCMC method lead to (for the pre-transformed data)

$$\mathrm{Cov}(\beta_0, \beta_1) \approx 0.92,$$

but variational Bayes approximation 5.2 forces this value to zero, leading to posterior density functions with incorrect amounts of spread and strange behaviour in the 95% pointwise credible intervals. The prediction intervals are less affected by the differences between (5.2) and (5.3) since the error variance posterior contribution dominates. In the following example, the regression coefficient should make a block in the approximation factorization of the product density.

5.2.2 Binary response regression

When the response variable $y_i \in \{0, 1\}$, then the appropriate regression models take the form:

$$P(y = 1|\boldsymbol{\beta}) = F(\boldsymbol{X}\boldsymbol{\beta}), \quad \boldsymbol{\beta} \sim N(0, \tau_\beta^{-1} I)$$

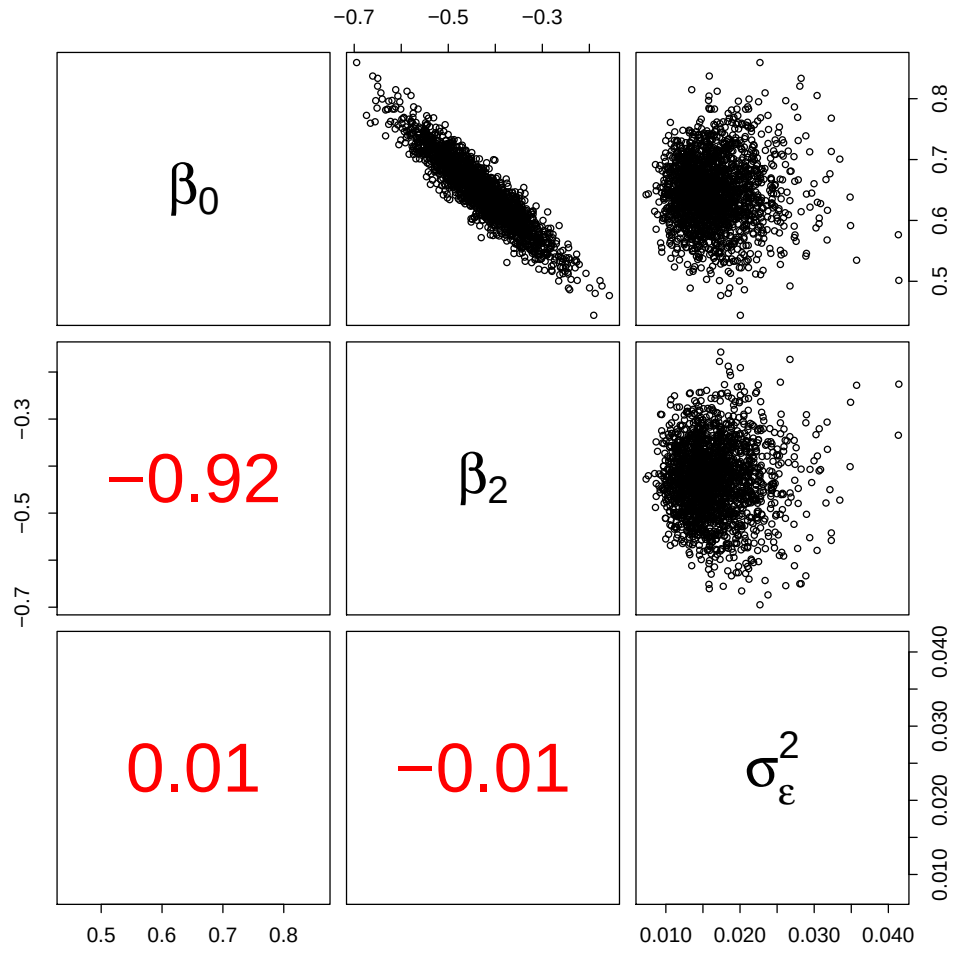

Figure 5.3: Pairwise scatter plots of MCMC samples and sample correlations between β_0, β_1 and σ_ϵ^2.

where F:$R \to (0, 1)$ is an inverse link function. We can obtain the logistic regression model and probit regression model by choosing F to be the logistic function and probit function respectively. `Infer.NET` can accommodate both types of binary regression models by using different approximation methods.

For the logistic regression model, `Infer.NET` can perform logistic regression by using the Jaakkola and Jordan (2000) trick (Result 1.18). Ormerod and Wand (2010) also use this trick for variational approximations logistic regression. The

logistic regression model works on the VMP method. The likelihood specification for the logistic regression model is:

```
Range index = new Range(n).Named("index");
VariableArray<bool> y = Variable.Array<bool>(index).Named("y");
VariableArray<Vector> xvec = Variable.Array<Vector>(index).
                             Named("xvec");
y[index] = Variable.BernoulliFromLogOdds(
           Variable.InnerProduct(beta,xvec[index]));
```

For the probit regression model, we introduce auxiliary variables corresponding to Result 1.15 of Albert and Chib (1993) to fit the probit regression. The probit repression model uses the EP method. The likelihood specification for the probit regression model is:

```
y[index] = Variable.IsPositive(
           Variable.GaussianFromMeanAndVariance(
           Variable.InnerProduct(beta,xvec[index]),1));
```

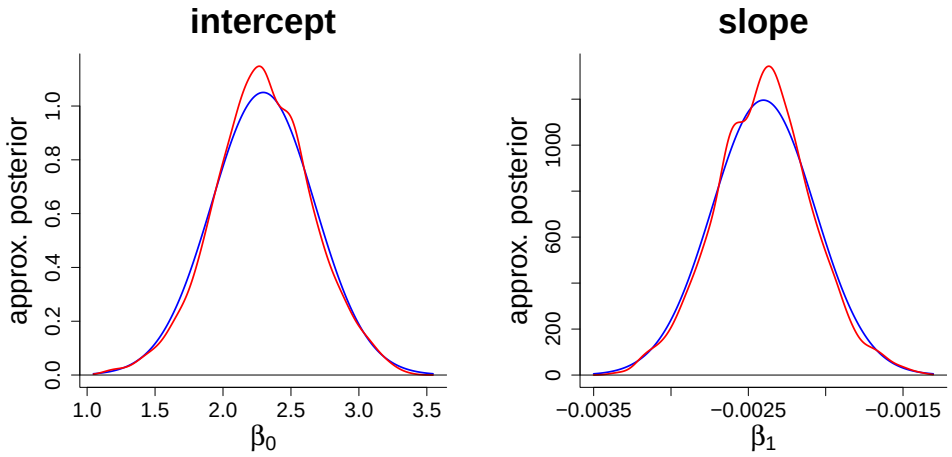

Figure 5.4: Variational Bayes approximate posterior density functions produced by `Infer.NET` (blue) and MCMC (orange) for the probit regression model fit to the BPD data.

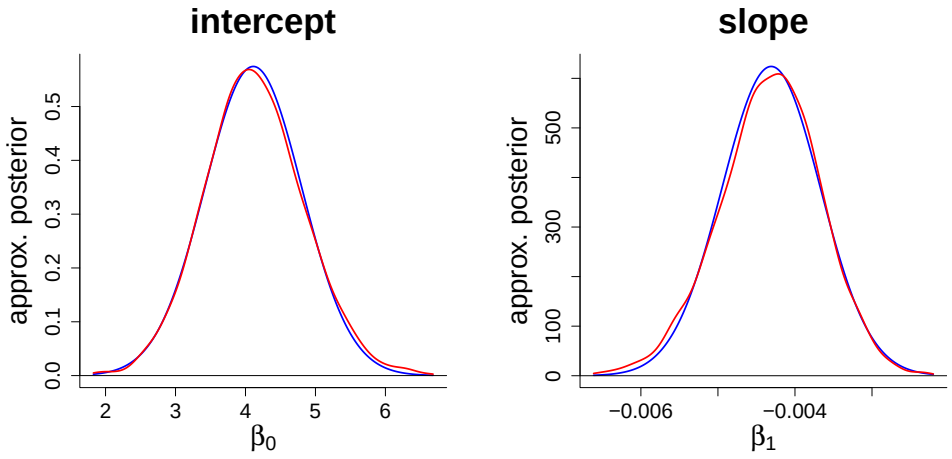

Figure 5.5: Variational Bayes approximate posterior density functions produced by `Infer.NET` (blue) and MCMC (orange) for the logistic regression model fit to the BPD data.

We fit the binary response regression to the bronchopulmonary dysplasia (BPD) data set (source: Pagano & Gauvreau, 2000). The predictor and response variable are birthweight (grammes) and the indicator variable respectively. The hyperpa-

rameter is set at $\sigma_\beta^2 = 10^8$. Figures 5.4 and 5.5 show the approximate posterior density obtained from `Infer.NET` and MCMC for probit regression and logistic regression. The fitted probability curves and pointwise 95% credible sets are shown in Figure 5.6.

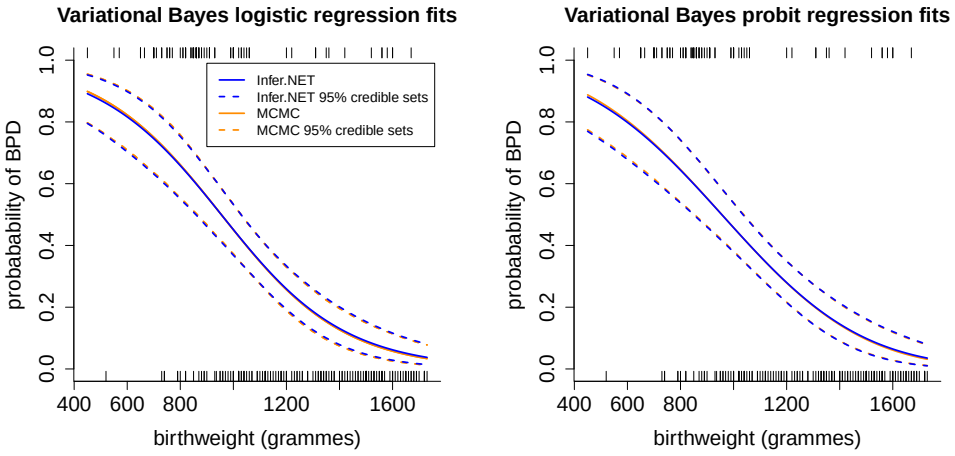

Figure 5.6: Binary response regression fits to the bronchopulmonary dysplasia (BPD) data using `Infer.NET`. The solid line is the posterior probability of BPD for a given birthweight. The dashed lines are point-wise 95% credible sets.

5.2.3 Random intercept model

The random intercept model is a simple linear mixed model. We consider the model:

$$
\begin{aligned}
&y_{ij}|\boldsymbol{\beta}, u_i, \tau_\epsilon \overset{\text{ind.}}{\sim} N(\boldsymbol{\beta}^T \boldsymbol{x}_{ij} + u_i, \tau_\epsilon^{-1}), \;\; 1 \le i \le m, \;\; 1 \le j \le n_i, \\
&u_i \overset{\text{ind.}}{\sim} N(0, \tau_u^{-1}), \\
&\boldsymbol{\beta} \sim N(\mathbf{0}, \tau_\beta^{-1} I), \\
&\tau_\epsilon \sim \text{Gamma}(A_\epsilon, B_\epsilon), \\
&\tau_u \sim \text{Gamma}(A_u, B_u),
\end{aligned}
\tag{5.5}
$$

where y_{ij} is the jth response measurement in the ith group and m is the number of groups. The quantify n_i is the number of observations in the ith group and u_i is a random intercept specific to the ith group.

To avoid the effect of correlations between regression coefficients, we use the following factorization for approximate posterior density,

$$q_{\boldsymbol{\beta},\boldsymbol{u},\tau_\epsilon,\tau_u}(\boldsymbol{\beta},\boldsymbol{u},\tau_\epsilon,\tau_u) = q_{\boldsymbol{\beta},\boldsymbol{u}}(\boldsymbol{\beta},\boldsymbol{u})q_{\tau_\epsilon}(\tau_\epsilon)q_{\tau_\epsilon}(\tau_\epsilon), \tag{5.6}$$

where $\boldsymbol{u} = (u_1, ..., u_m)^T$ is a vector of random effect coefficients. The random intercept model has the form:

$$\begin{aligned}
&y|\boldsymbol{\beta},\boldsymbol{u},\tau_\epsilon \sim N(X\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \tau_\epsilon^{-1}) \\
&\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} |\tau_u \sim N\left(0, \begin{bmatrix} \sigma_\beta^2 I & 0 \\ 0 & \tau_u^{-1} I \end{bmatrix}\right) \\
&\tau_u \sim \text{Gamma}(A_u, B_u) \\
&\tau_\epsilon \sim \text{Gamma}(A_\epsilon, B_\epsilon),
\end{aligned} \tag{5.7}$$

where $\boldsymbol{X}$ contains the x_{ij} and $Z = \boldsymbol{I}_m \otimes \boldsymbol{1}_n$ is the indicator matrix for matching the x_{ij}s with their corresponding u_i ($1 \leq i \leq m$, $1 \leq j \leq n$).

In `Infer.NET`, we want to form $\boldsymbol{\beta}$ and $\boldsymbol{u}$ as a block to avoid the effect of correlations between regression coefficients. We introduce an auxiliary variable, $\boldsymbol{a}$,

corresponding to Result 1.16. Then the model is given by:

$$\begin{aligned}
&\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \tau_\epsilon \sim N(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \tau_\epsilon^{-1}), \\
&\boldsymbol{a}|\boldsymbol{\beta}, \boldsymbol{u}, \tau_u \sim N\left(\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix}, \begin{bmatrix} \sigma_\beta^2 I & 0 \\ 0 & \tau_u^{-1} I \end{bmatrix}\right), \\
&\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} \sim N(0, \kappa^{-1}\boldsymbol{I}), \\
&\tau_u \sim \text{Gamma}(A_u, B_u), \\
&\tau_\epsilon \sim \text{Gamma}(A_\epsilon, B_\epsilon),
\end{aligned} \tag{5.8}$$

where κ is a hyperparameter with a very small value, and $\boldsymbol{a}$ is set to have an observed value $\underline{0}$.

The input data correspond to four longitudinal orthodontic measurements on each of $m = 27$ children (source: Pinheiro & Bates, 2000). The data are available in the R computing environment (R Core Team, 2015) via the package `nlme` (Pinheiro *et al.*, 2015), in the object `Orthodont`. The y_{ij} corresponds to distances from the pituitary to the pterygomaxillary fissure (mm) and

$$\boldsymbol{\beta}^T x_{ij} = \beta_0 + \beta_1 age_{ij} + \beta_2 male_i$$

where age_{ij} is the age of the child when y_{ij} was recorded and $male_i$ is an indicator variable for the child being male.

Figure 5.7 shows the approximate posterior density obtained from `Infer.NET`. We make a transformation using $\sigma_\epsilon^2 = \tau_\epsilon^{-1}$ and $\sigma_u^2 = \tau_u^{-1}$.

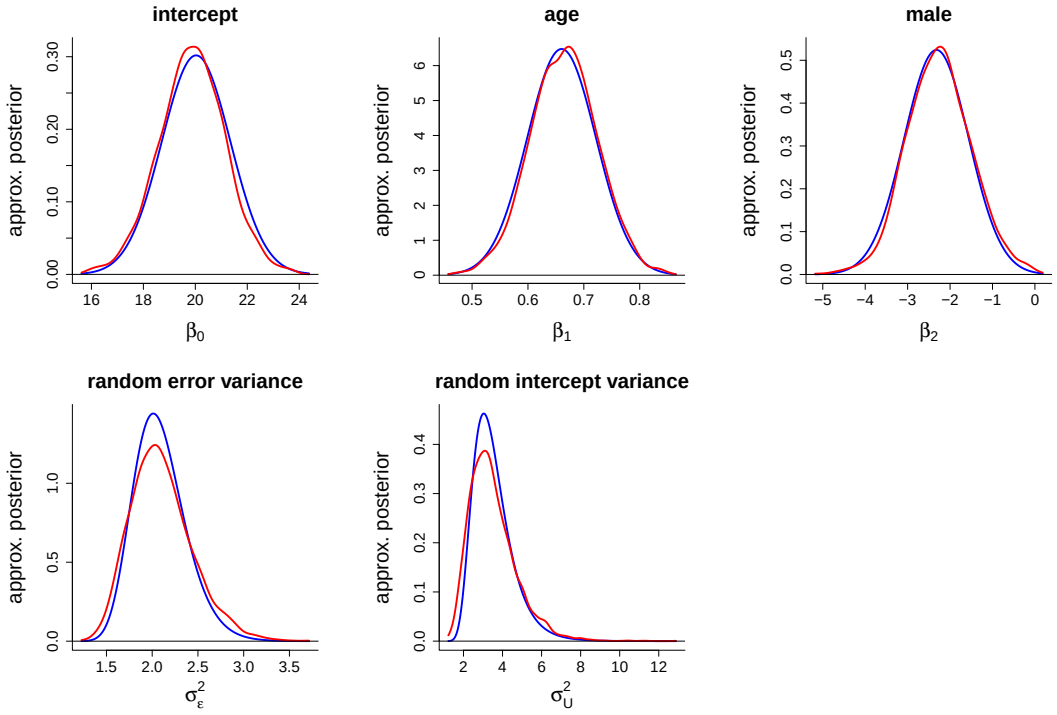

Figure 5.7: Variational Bayes approximate posterior density functions produced by `Infer.NET` (blue) and MCMC (orange) for the simple linear mixed model fit to the orthodontic data.

5.2.4 Normal mixture model

Consider the Normal mixture density model:

$$\begin{aligned} x_i &\sim \sum_{k=1}^{K} \frac{\omega_k}{\tau_k^{-1}} \phi\left(\frac{x_i - \mu_k}{\tau_k^{-1}}\right), \\ \tau_k^{-1} &\sim \text{Gamma}(A, B), \\ \mu_k &\sim N(0, \sigma_\mu^2), \\ \boldsymbol{\omega} &\sim \text{Dirichlet}(\boldsymbol{\alpha}), \end{aligned} \tag{5.9}$$

where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_K)$, $\boldsymbol{\omega} = (\omega_1, ..., \omega_K)$, A, B and σ_μ^2 are hyperparameters, and x_i, $1 \leq i \leq n$ is a univariate sample. This is the classic finite normal mixture

problem and is the topic of an enormous amount of literature (e.g., McLachlan & Peel, 2000). Ormerod and Wand (2010) described a MFVB algorithm for fitting (5.9). We use `Infer.NET` to fit this model and choose the value of K, Section 10.2.4 of Bishop (2006), to introduce and choose K by maximizing

$$\log \underline{p}(\boldsymbol{x}; q) + log(K!).$$

Here $\log \underline{p}(x; q)$ is the MFVB approximation to the marginal log-likelihood. The $\log(K!)$ term accounts for the $K!$ configurations of (w_k, μ_k, τ_k) that give rise to the same normal mixture density function. We choose $K \in \Xi = \{1, 2, ..., K_{max}\}$.

We use `Infer.NET` to fit a finite normal mixture to the data on the eruption durations of a geyser. The geyser data are available in R via the `MASS` package (Venables & Ripley, 2002), in the data frame entitled `geyser`. The hyperparameter are set at $A = B = 0.01$, $\boldsymbol{\alpha} = (1, 1, ..., 1)$ and $\sigma_\mu^2 = 10^8$.

`Infer.NET` can compute $\log \underline{p}(x; q)$ by creating a mixture of the current model with an empty model. The learnt mixing weight is then the marginal log-likelihood. Further details on this trick are given in the `Infer.NET` user guide, where the term *model evidence* is used for $\underline{p}(x; q)$. We first need to set up an auxiliary Bernoulli variable as follows:

```
Variable<bool> auxML = Variable.Bernoulli(0.5).Named("auxML");
```

The code for the normal mixture fitting is then given by:

```
IfBlock model = Variable.If(auxML);
```

and

```
model.CloseBlock();
```

The quantify $\log \underline{p}(x; q)$ is then obtained from:

```
double marginalLogLikelihood =
                        engine.Infer<Bernoulli>(auxML).LogOdds;
```

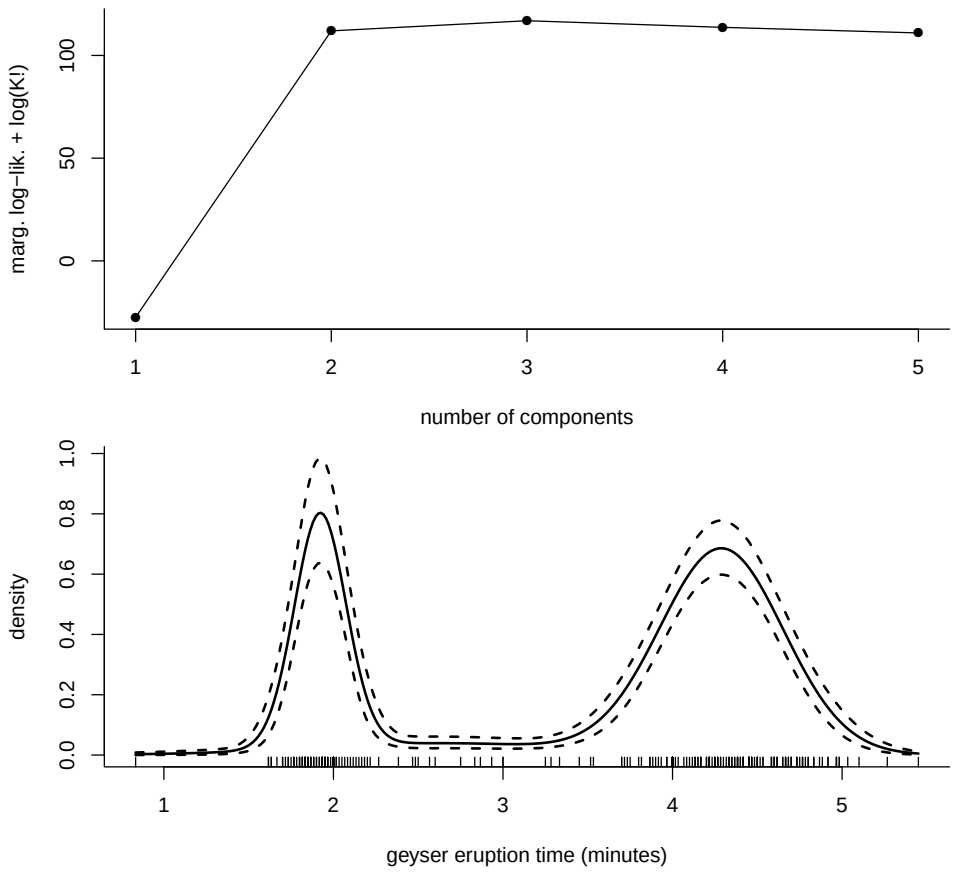

Figure 5.8: Upper panel: log $p(\boldsymbol{x}; q) + \log(K!)$ versus K, where K is the number of components in the normal mixture fit to the transformed geyser duration data. Lower panel: fitted normal mixture density for $K = 3$ (the K that maximizes the criterion of the upper panel plot). The dashed curves correspond to pointwise 95% credible sets.

Note that K = 3 maximizes $\log \underline{p}(\boldsymbol{x}; q) + \log(K!)$, as shown in the upper panel of Figure 5.8. The lower panel shows the K = 3 `Infer.NET` fit. Also shown are 95% pointwise credible sets based on Monte Carlo samples of size 10000 from the approximate posterior distributions.

5.3 Advanced Examples

In this section, we describe some advanced examples that illustrate the capabilities of `Infer.NET` for more challenging data analyses. Some complex models with non-conjugate priors will be considered.

5.3.1 Normal additive model with Half-Cauchy prior

A three-predictor Normal additive model is

$$\begin{aligned} y_i &= \beta_0 + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}) + \varepsilon_i, \\ \varepsilon_i &\sim N(0, \tau_\varepsilon^{-1}), \end{aligned} \tag{5.10}$$

where, for $1 \leq i \leq n$, the y_i are measurements on a continuous response variable and (x_{1i}, x_{2i}, x_{3i}) are triples containing measurements on three continuous predictor variables. We will model each of the $f_j(\cdot)$ using low-rank smoothing splines with a mixed model representation:

$$f_j(x) = \beta_j x + \sum_{k=1}^{K} u_{jk} z_{jk}(x),$$

$$u_{jk} \sim N(0, \tau_{u_j}^{-1}),$$

where z_{jk}, $1 \leq k \leq K$, are O'Sullivan penalised spline basis functions (Wand & Ormerod, 2008) over x_j. The vector $\boldsymbol{u}_j = (u_{1,j}, ..., u_{K,j})^T$ is a random effect vector with its elements having independent normal distributions. In the last section, we use an Inverse Gamma distribution as a prior density for the random effects variance. In this section, we use Half-Cauchy distribution (Gelman, 2006) as the prior distribution for the standard deviation of the random effects. Then, the Half

Cauchy distribution is given by

$$\tau_{u_j}^{-1} \sim \text{Half-Cauchy}(A),$$

where $A > 0$ is a hyperparameter. We introduce an auxiliary vector $\boldsymbol{a} = (a_1, a_2, a_3)^T$ corresponding to Result 1.5, such that:

$$\tau_{u_j}|a_j \sim \text{Gamma}(1/2, a_j) \text{ and } a_j \sim \text{Gamma}(1/2, 1/A^2),$$

where $1 \leq j \leq 3$. The full model with auxiliary variables is then:

$$\begin{aligned}
&\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \tau_\varepsilon^{-1} \sim N(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \tau_\varepsilon^{-1}\boldsymbol{I}),\\
&\boldsymbol{u}|\tau_1, \tau_2, \tau_3 \sim N\left(0, \begin{bmatrix} \tau_1^{-1}\boldsymbol{I} & 0 & 0 \\ 0 & \tau_2^{-1}\boldsymbol{I} & 0 \\ 0 & 0 & \tau_3^{-1}\boldsymbol{I} \end{bmatrix}\right),\\
&\boldsymbol{\beta} \sim (0, \sigma_\beta^2 I),\\
&\tau_\varepsilon \sim \text{Gamma}(A_\varepsilon, B_\varepsilon),\\
&\tau_{uj}|a_j \sim \text{Gamma}(1/2, a_j), \;\; 1 \leq j \leq 3,\\
&a_j \sim \text{Gamma}(1/2, 1/A^2),
\end{aligned} \tag{5.11}$$

where

$$\boldsymbol{X} = [1\, x_{1i}\, x_{2i}\, x_{3i}]_{1 \leq i \leq n}$$

and

$$\boldsymbol{Z} = [z_{1k}(x_{1i}) \underset{1 \leq k \leq K_1}{|}\, z_{2k}(x_{2i}) \underset{1 \leq k \leq K_2}{|}\, z_{3k}(x_{3i})]_{1 \leq i \leq n} \underset{1 \leq k \leq K_3}{}$$

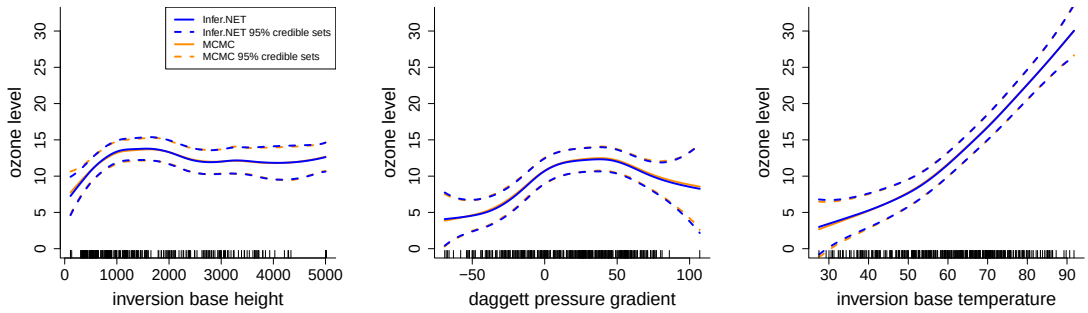

Figure 5.9: Variational Bayes additive model fits as produced by `Infer.NET` for the California ozone data. The dashed curves correspond to point-wise 95% credible sets.

We fit this model to variables in the Ozone data frame in the R package `mlbench` (Leisch & Dimitriadou, 2009). The response variable is daily maximum ozone level and the predictor variables are the inversion base height (feet), pressure gradient to the town of Daggett (mm Hg), and inversion base temperature (degrees Fahrenheit) at Los Angeles International Airport. All variables were transformed to the unit interval for `Infer.NET` fitting and the hyperparameters were fixed at $\sigma_\beta^2 = 10^8$, $A_\varepsilon = 0.01$, $B_\varepsilon = 0.01$, $A = 25$.

The fitted curves in Figure 5.9 shows interesting non-linear effects. Note that the convergence speed of the Normal additive model with the half-Cauchy prior is remarkably slower than with the Inverse-Gamma prior using `Infer.NET`.

5.3.2 Generalized logistic additive model

`Infer.NET` can fit the logistic regression model and the Normal additive model. In this section, we illustrates binary response regression through the model:

$$y_i \sim \text{Bernoulli}\{\text{logistic}(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + f(x_{4i}) + f(x_{5i}) + f(x_{6i})),\}$$

where, for $1 \leq i \leq n$, the y_i are measurements on a binary response variable, and x_{ji}, $1 \leq j \leq 6$, are six predictor variables. The effect of the variables x_1, x_2 and x_3 are linear; the effect of variables x_4, x_5 and x_6 are non-linear. We will model each of the $f_\ell(\cdot)$, $1 \leq \ell \leq 3$, using low-rank smoothing splines with mixed model representation. The binary response logistic additive model is given by

$$
\begin{aligned}
&y|\boldsymbol{\beta}, \boldsymbol{u}, \sim \text{Bernoulli}\left(\text{logit}^{-1}\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}\right)\right), \\
&\boldsymbol{u}|\tau_{u_1}, \tau_{u_2}, \tau_{u_3} \sim N\left(0, \begin{bmatrix} \tau_{u_1}^{-1}\boldsymbol{I} & 0 & 0 \\ 0 & \tau_{u_2}^{-1}\boldsymbol{I} & 0 \\ 0 & 0 & \tau_{u_3}^{-1}\boldsymbol{I} \end{bmatrix}\right), \\
&\boldsymbol{\beta} \sim N(0, \sigma_\beta^2 I), \\
&\tau_{u_j}^{-1/2} \sim \text{Half-Cauchy}(A), \;\; 1 \leq j \leq 3,
\end{aligned}
$$

where the $\boldsymbol{X}$ and $\boldsymbol{Z}$ are defined as in the previous section. A and σ_β^2 are hyperparameters. We introduce an auxiliary vector $\boldsymbol{a} = (a_1, a_2, a_3)^T$ corresponding to Result 1.5 and an auxiliary vector $\boldsymbol{\alpha}$ corresponding to Result 1.16. Then, the actual model implemented in `Infer.NET` is:

$$
\begin{aligned}
&\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sim \text{Bernoulli}\left(\text{logistic}\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}\right)\right), \\
&\boldsymbol{\alpha}|\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\tau}_u \sim N\left(\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\tau}_\beta^{-1}\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \text{diag}(\boldsymbol{\tau}_u^{-1}) \end{bmatrix}\right), \\
&\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} \sim N(0, \kappa^{-1}\boldsymbol{I}), \\
&\tau_{u_j}|a_j \sim \text{Gamma}(1/2, a_j), \;\; 1 \leq j \leq 3, \\
&a_j \sim \text{Gamma}(1/2, 1/A^2),
\end{aligned}
$$

where $\boldsymbol{\alpha}$ is set as an observed value $\mathbf{0}$, and $\boldsymbol{\tau}_u^{-1} = (\boldsymbol{\tau}_{u_1}^{-1}, \boldsymbol{\tau}_{u_2}^{-1}, \boldsymbol{\tau}_{u_3}^{-1})^T$. We use the union membership data set (Ruppert, Wand & Carroll, 2003).

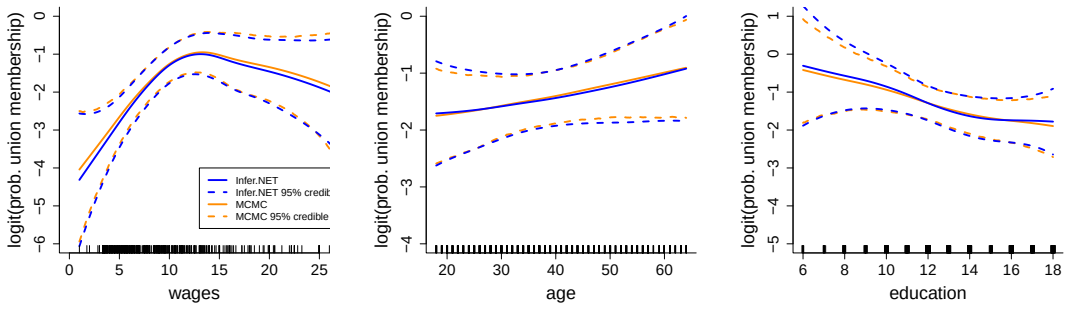

Figure 5.10: Variational Bayesian logistic additive model fits by `Infer.NET` for the union membership data. The dashed curves correspond to point-wise 95% credible sets.

The response is the indicator of union membership (union). The predictor variables are: race (indicator of white race); gender (indicator of female); south (indicator of living in southern region of the U.S.); wages (wages in dollars/ hour); age (age in years); ed (number of years of education). The first three variables (race, gender, and south) are all indicators, so they were modeled with linear effects. The effects of the last three (wages, age, and ed) were modeled nonparametrically using splines. The hyperparameters were set at $\tau_\beta = 10^{-10}$, $A = 25$, and $\kappa = 10^{-10}$, while the number of MFVB iterations was fixed at 100. Figure 5.10 illustrates that there is good agreement between the results from `Infer.NET` and the MCMC method.

5.3.3 Bayesian Lasso regression

In Chapter 4, we fit the Bayesian Lasso mode using the MFVB method with an auxiliary variables $\boldsymbol{a} = (a_1, a_2, ..., a_p)^T$ corresponding to Result 1.9. The full model

with auxiliary variables is:

$$
\begin{aligned}
&\boldsymbol{y} \sim N(\mathbf{1}_n\beta_0 + \boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{I}_n\sigma_\varepsilon^2),\\
&\beta_0 \sim N(0, \sigma_{\beta_0}^2),\\
&\beta_j|a_j \sim N(0, \frac{\sigma_\varepsilon^2}{a_j}) \quad j = 1, 2, \ldots, p,\\
&a_j \sim \text{IG}(1, \frac{\lambda^2}{2}),\\
&\sigma_\varepsilon^2 \sim \text{IG}(A_\varepsilon, B_\varepsilon),
\end{aligned}
\tag{5.12}
$$

where the prior and posterior distributions for the auxiliary variable a_i are both Inverse-Gaussian distributions. Therefore, there are not conjugate distributions for variable a_i, and `Infer.NET` is not able to handle those models. We introduce the auxiliary variables $\boldsymbol{b} = (b_1, b_2, ..., b_p)^T$ corresponding to Result 1.17. The full model with auxiliary variables is then:

$$
\begin{aligned}
&\boldsymbol{y} \sim N(\mathbf{1}_n\beta_0 + \boldsymbol{X}\boldsymbol{\beta}, \tau_\varepsilon^{-1}\boldsymbol{I}_n),\\
&\beta_0 \sim N(0, \tau_{\beta_0}^{-1}),\\
&\beta_j|a_j \overset{\text{ind.}}{\sim} N(0, \tau_\varepsilon^{-1}a_j^{-1}), \quad j = 1, 2, \ldots, p,\\
&a_j|b_j \sim \text{Gamma}(M, Mb_j),\\
&b_j \sim \text{Gamma}(1, \frac{\lambda^2}{2}),\\
&\tau_\varepsilon \sim \text{Gamma}(A_\varepsilon, B_\varepsilon).
\end{aligned}
\tag{5.13}
$$

After introducing an auxiliary vector, $\boldsymbol{\alpha}$, corresponding to Result 1.16, the actual model fitted in `Infer.NET` is:

$$
\begin{aligned}
&\boldsymbol{y} \sim N(\mathbf{1}_n\beta_0 + \boldsymbol{X}\boldsymbol{\beta}, \tau_\varepsilon^{-1}\boldsymbol{I}_n),\\
&\boldsymbol{\alpha}|\beta_0, \boldsymbol{\beta}, \boldsymbol{a}, \tau_\varepsilon^{-1} \sim N\left(\begin{bmatrix} \beta_0 \\ \boldsymbol{\beta} \end{bmatrix}, \begin{bmatrix} \tau_{\beta_0}^{-1} & \mathbf{0} \\ \mathbf{0} & \tau_\varepsilon^{-1}\text{diag}(\boldsymbol{a}) \end{bmatrix}\right),\\
&\begin{bmatrix} \beta_0 \\ \boldsymbol{\beta} \end{bmatrix} \sim N(\mathbf{0}, \kappa^{-1}\boldsymbol{I}),\\
&a_j|b_j \sim \text{Gamma}(M, Mb_j), \quad j = 1, 2, \ldots, p,\\
&b_j \sim \text{Gamma}(1, \frac{\lambda^2}{2}),\\
&\tau_\varepsilon \sim \text{Gamma}(A_\varepsilon, B_\varepsilon),
\end{aligned}
\tag{5.14}
$$

where $\boldsymbol{\alpha}$ is set to have observed value $\mathbf{0}$. We use $M = 100$, $\kappa = 10^{-10}$, $A_\varepsilon = B_\varepsilon = 0.01$ and $\tau_{\beta_0}^{-1} = 10^{-10}$.

We illustrate Lasso model fitting in `Infer.NET` using same diabetes data (Efron *et al.*, 2004; Park & Casella, 2008). The sample size is $n = 442$ and the number of predictor variables is $p = 10$. The response variable is a continuous index of disease progression one year after baseline. and the predictor variables include age (**age**), sex (**sex**), body mass index (**bmi**), average blood pressure (**map**) and six blood serum measurements (**tc**, **ldl**, **hdl**, **tch**, **ltg** and **glu**).

Figure 5.11 compares `Infer.NET` Lasso estimates with estimates from the ordinary Lasso. The left panel shows the paths of estimates as a function of their ℓ_1 norm relative to the ℓ_1 norm of the least squares estimate. The right panel shows the paths of posterior mean estimates using `Infer.NET`. The paths of the `Infer.NET` Lasso estimates are similar in shape to the ordinary Lasso paths, but the paths of the `Infer.NET` Lasso estimates are smoother.

Similarly to work in chapter 4, we can choose a prior distribution $\pi(\lambda)$ for λ

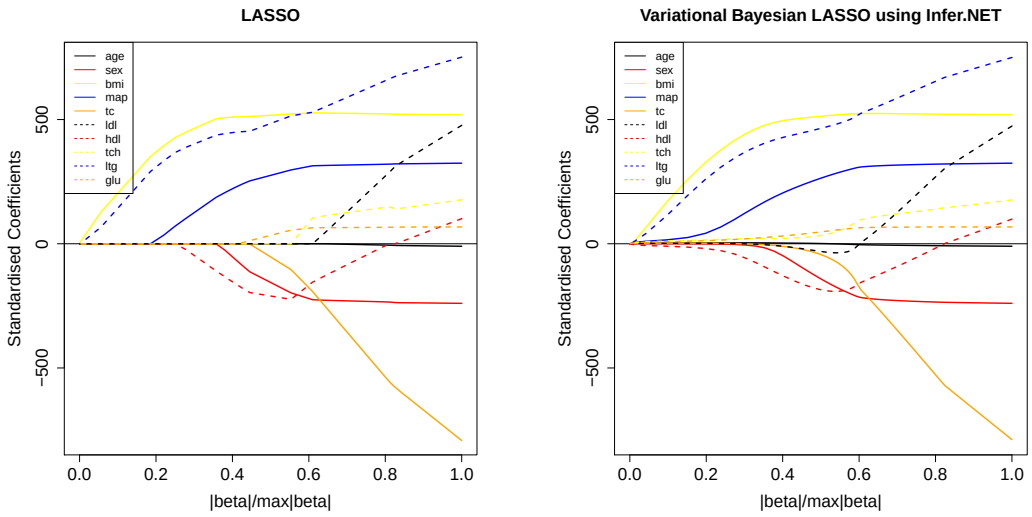

Figure 5.11: The trace plots of the original Lasso and `Infer.NET` Lasso for estimates of the diabetes data regression parameters.

with a uninformative hyperparameter. The actual adjusted Bayesian Lasso model in `Infer.NET` is:

$$
\begin{aligned}
&\boldsymbol{y} \sim N(1_n\beta_0 + \boldsymbol{X}\boldsymbol{\beta}, \tau_\varepsilon^{-1}\boldsymbol{I}_n),\\
&\boldsymbol{\alpha}|\beta_0, \boldsymbol{\beta}, \boldsymbol{a}, \tau_\varepsilon^{-1} \sim N\left(\begin{bmatrix} \beta_0 \\ \boldsymbol{\beta} \end{bmatrix}, \begin{bmatrix} \tau_{\beta_0}^{-1} & \mathbf{0} \\ \mathbf{0} & \tau_\varepsilon^{-1}\mathrm{diag}(\boldsymbol{a}) \end{bmatrix}\right),\\
&\begin{bmatrix} \beta_0 \\ \boldsymbol{\beta} \end{bmatrix} \sim N(\mathbf{0}, \kappa^{-1}\boldsymbol{I}),\\
&a_j|b_j \sim \mathrm{Gamma}(M, Mb_j), \quad j = 1, 2, \ldots, p,\\
&b_j|\lambda^2 \sim \mathrm{Gamma}(1, \frac{\lambda^2}{2}),\\
&\tau_\varepsilon \sim \mathrm{Gamma}(A_\varepsilon, B_\varepsilon),\\
&\lambda^2 \sim \mathrm{Gamma}(A_\lambda, B_\lambda).
\end{aligned}
\tag{5.15}
$$

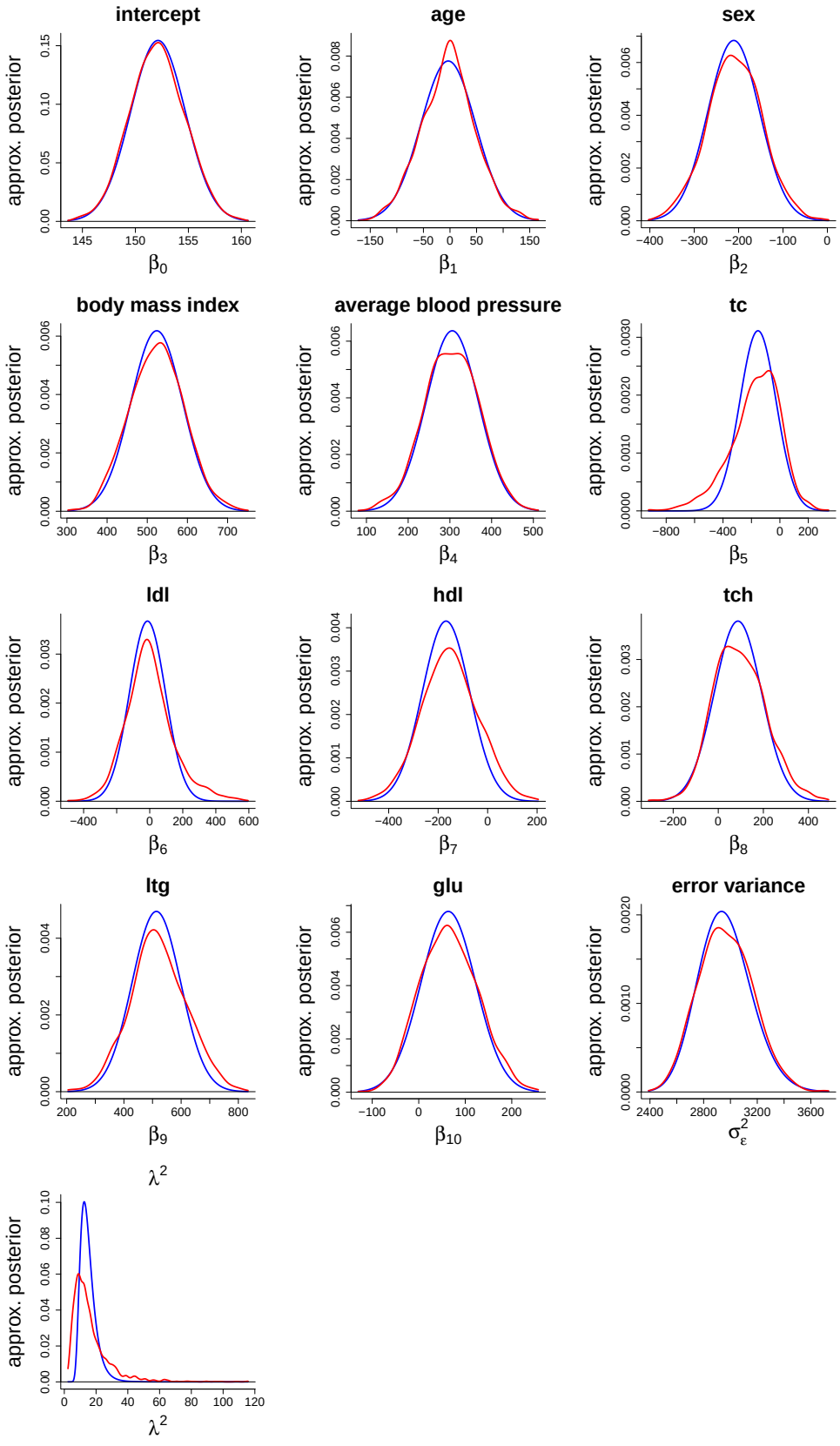

Figure 5.12: Bayes approximate posterior density functions produced by MFVB inference using `Infer.NET` (blue) and MCMC (orange) for fitting model (5.15) to the diabetes data set.

We use $M = 100$, $\kappa = 10^{-10}$, $A_\varepsilon = B_\varepsilon = 0.01$, $A_\lambda = B_\lambda = 0.01$ and $\tau_{\beta_0}^{-1} = 10^{-10}$. The approximate posterior density functions computed in model (5.15) by using `Infer.NET` and MCMC for diabetes data are shown in Figure 5.12. Good to excellent accuracy of MFVB inference is apparent for all posterior densities.

5.3.4 Robust nonparametric regression based on the t-distribution

Assuming that the response variable has a t-distribution is a popular model-based approach for robust regression. Wand *et al.* (2011) presented an MFVB algorithm for fitting Bayesian t-distribution model to a univariate random sample. In this section, we illustrates robust nonparametric regression based on the t-distribution. We consider a penalized spline mixed model approach to nonparametric regression using the t-distribution (Staudenmayer *et al.*, 2009):

$$\begin{aligned} &y_i \sim t(\beta_0 + \beta_1 x_i + f(x_i), \tau_\varepsilon^{-1}, \nu), \\ &p(\nu) \text{ discrete on a finite set } \Xi, \end{aligned} \tag{5.16}$$

where the $f(\cdot)$ is a low-rank smoothing spline with mixed model representation as in previous examples, and ν is a discrete distribution. Wand *et al.* (2011) extended an ordinary MFVB method to discrete distribution. Since variational message passing is a special case of MFVB, `Infer.NET` can fit Model 5.16 using structured MFVB (Saul & Jordan, 1996). First, Model 5.16 can be fitted by using `Infer.NET` for each $\nu \in \Xi$. Next, the results of each of these fits are combined.

Because the t-distribution is not supported by `Infer.NET`, we introduce an auxiliary vector $\boldsymbol{a} = (a_1, a_2, ..., a_n)^T$ corresponding to Result 1.10 to deal with the

t-distribution, such that:

$$y|a \sim N(0, a^{-1}\tau^{-1}) \text{ and } a \sim \text{Gamma}(\frac{\nu}{2}, \frac{\nu}{2}),$$

implies $y \sim t(x, \tau^{-1}, \nu)$.

Similarly to the previous example, we introduce an auxiliary vector, $\boldsymbol{\alpha}$, corresponding to Result 1.16 to handle the $\boldsymbol{\beta}$ and $\boldsymbol{u}$. The prior distribution for ν was set to be a discrete uniform distribution between 0.5 to 10 on the set Ξ, and the interval was set to be 0.1. For each $\nu \in \Xi$, the actual model fitted in `Infer.NET` is:

$$\begin{aligned}
&\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{a}, \tau_\varepsilon^{-1} \sim N(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \tau_\varepsilon^{-1}\text{diag}(\boldsymbol{a}^{-1})),\\
&a_i \sim \text{Gamma}(\nu/2, \nu/2), \;\; 1 \leq i \leq n,\\
&\boldsymbol{\alpha}|\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\tau}_u \sim N\left(\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\tau}_\beta^{-1}\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \tau_u^{-1}\boldsymbol{I} \end{bmatrix}\right),\\
&\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} \sim N(0, \kappa^{-1}\boldsymbol{I}),\\
&\tau_u \sim \text{Gamma}(A_u, B_u),\\
&\tau_\varepsilon \sim \text{Gamma}(A_\varepsilon, B_\varepsilon),
\end{aligned} \tag{5.17}$$

where $\boldsymbol{Z}$ and $\boldsymbol{X}$ are defined as in the previous examples. We set $\boldsymbol{\alpha}$ as an observed value $\boldsymbol{0}$, and set $\kappa = 10^{-10}$, $A_\varepsilon = B_\varepsilon = 0.01$, $A_u = B_u = 0.01$ and $\tau_{\beta_0}^{-1} = 10^{-10}$ as the values of the hyperparameters.

Next, for $\nu \in \Xi$, the approximate posterior densities are obtained from:

$$
\begin{aligned}
q(\nu) &= \frac{p(\nu)\underline{p}(\boldsymbol{y}|\nu)}{\sum_{\nu' \in \Xi} p(\nu')\underline{p}(\boldsymbol{y}|\nu')}, \\
q(\boldsymbol{\beta}, \boldsymbol{u}) &= \textstyle\sum_{\nu \in \Xi} q(\nu) q(\boldsymbol{\beta}, \boldsymbol{u}|\nu), \\
q(\tau_u) &= \textstyle\sum_{\nu \in \Xi} q(\nu) q(\tau_u|\nu), \\
q(\tau_\varepsilon) &= \textstyle\sum_{\nu \in \Xi} q(\nu) q(\tau_\varepsilon|\nu).
\end{aligned}
$$

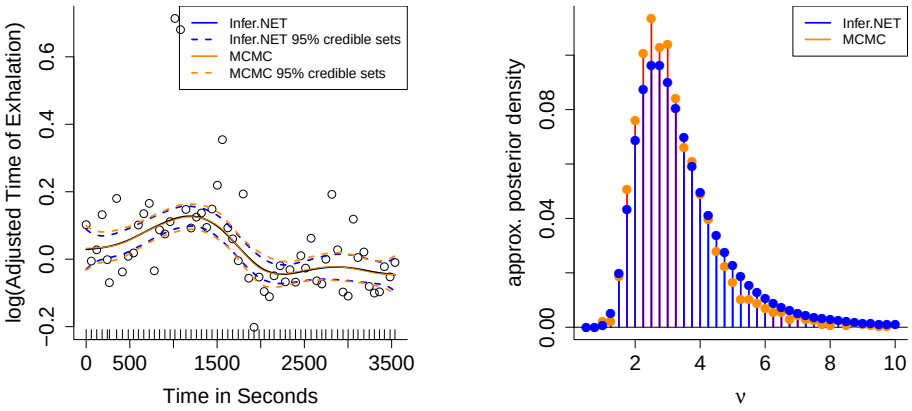

Figure 5.13: The robust nonparametric regression model (5.17) fits to the respiratory experiment data using structured mean field variational Bayesian, based on `Infer.NET`, and MCMC. Left: Posterior mean and pointwise 95% credible sets for the regression function. Right: Approximate posterior mass for the degrees of freedom parameter ν.

Figure 5.13 shows the results of the structured MFVB analysis from `Infer.NET` fitting of 5.17 to a data set on a respiratory experiment (Staudenmayer *et al.*, 2009). The log-adjusted response, $\boldsymbol{y}$, is the log of a subject's response for the experiment minus her mean response at baseline, and $\boldsymbol{x}$ is equal to the time in seconds. The right panel of Figure 5.13 shows the fitted curve and pointwise 95% credible sets by using the MCMC method and `Infer.NET`. The `Infer.NET` fits and pointwise

95% credible sets are quite close to those obtained by using the MCMC method. The approximate posterior function of ν is in the right-hand panel of Figure 5.13. There is good agreement between the results of the Infer.net and those of the MCMC approach.

5.4 Timing Comparison

Table 5.1 shows the relative computing times for the examples in this chapter using `Infer.NET` and the MCMC method. In this table, we report the time elapsed (and standard error) of the computing times over 100 runs with the number of `Infer.NET` iterations set to 100 and the MCMC sample with 10,000 burn-in and 10,000 regular iterations. This was sufficient for convergence in these particular examples. Table 5.1 reveals that `Infer.NET` is considerably faster than BUGS for most of the examples. This is particularly the case for some advanced examples, where the computing times are reduced from minutes to seconds when going from BUGS to `Infer.NET`. The slowness of the robust nonparametric regression fit is mainly explained by the multiple calls to `Infer.NET`, corresponding to the degrees of freedom grid.

5.5 Discussion

In this chapter, eight examples were used to illustrate various types of statistical regression model analysis via `Infer.NET`. Most example show that the inference by `Infer.NET` is quite accurate. The first two panels of Figure 5.2 show that the posterior density functions produced by `Infer.NET` can be overly narrow for stringent product density restrictions. By adding the auxiliary variable, t-distribution and Laplace distribution can also be fitted by `Infer.NET`, and there is good agreement

regression model	time in seconds for `Infer.NET`	time in seconds for MCMC
Simple linear	0.05 (0.01)	0.18 (0.01)
Simple logistic	0.06 (0.01)	2.98 (0.03)
Simple probit	0.03 (0.01)	1.56 (0.02)
Random intercept	2.15 (0.01)	1.34 (0.03)
Normal mixture	11.2 (0.01)	25.5 (0.04)
Normal additive	7.59 (0.01)	2226 (1.03)
Logistic additive	8.56 (0.01)	5682 (1.74)
Bayesian Lasso	0.11 (0.02)	583 (1.02)
Robust nonparametric	90.8 (0.03)	23.7 (0.02)

Table 5.1: Average (standard errors) run times in seconds over 100 runs of the methods for each of the examples in Chapter 5.

between `Infer.NET` and the MCMC method for those models.

`Infer.NET` is inherently inaccurate since it relies on deterministic approximation methods. Our examples demonstrate that `Infer.NET` can be use to fit some part of current popular statistical models. Some useful distributions and models cannot be defined in the current version of `Infer.NET`. We hope that this chapter will lead to useful discourse on the confluence between `Infer.NET` and statistical analyses.

Chapter 6

Asymptotic Normality and Valid Inference for Gaussian Variational Approximation[1]

6.1 Introduction

The generalized linear mixed models (GLMMs) are an extension to the generalized linear model in which the linear predictor contains random effects in addition to the usual fixed effects (Williams, 1982; Breslow, 1984; Zeger, Liang and Albert, 1988). Maximum likelihood estimation for GLMMs always involves complicated problems, including irreducible high-dimensional integrals. Bayesian inference can be used to avoid the numerical integration by obtaining the posterior distribution using the MCMC method (Zeger and Karim, 1991). However, there are significant drawbacks involving time and computational costs. Breslow and Clayton (1993) presented an approximation method based on the marginal quasi-likelihood using

[1]This chapter is based on: Hall, P., Pham, T., Wand, M.P. and Wang, S.S.J. Asymptotic Normality and Valid Inference for Gaussian Variational Approximation. *The Annals of Statistics*, (2011), 39, 2502-2532.

Laplace's method.

Recently, Hall, Ormerod and Wand (2011) and Ormerod and Wand (2012) extended variational approximation technology to statistical settings, and introduced a frequentist, rather than Bayesian, inference for GLMMs. This approach was named as Gaussian variational approximation (GVA), which involves minimum Kullback-Liebler divergence from a family of Gaussian densities. Ormerod and Wand (2012) presented the algorithm of GVA for GLMMs and showed the GVA to be quite accurate for $n \simeq 5$, where n is the number of repeated measures in each group. Hall *et al.* (2011) proved consistency of the variational approximate estimators and established the rate of consistency of GVA for a simple Poisson mixed model, i.e. the Poisson mixed model with a single predictor variable and random intercept. Those theories pointed out that there exits the bounds for GVA estimators and a consistency rate of $m^{-1/2} + n^{-1}$, where m is the number of groups and n is number of observations in each group.

In this chapter, we improve upon Hall *et al.* (2011)'s results for the simplest Poisson mixed model, i.e. the Poisson mixed model with a constant and a random intercept, and obtain the asymptotic distributions of the estimators. The results show that the estimators are asymptotically normal, have negligible bias and that their constant parameter and variances decay at least as fast as m^{-1}, where m is the number of groups. Ormerod and Wand (2012) give the details about point estimators for each parameter. Using the asymptotic normality result, confidence intervals for all model parameters can be obtained directly, without any numerical integration or MCMC simulation.

Section 2 describes the simplest Poisson mixed model and GVA. Section 3 introduces an asymptotic normality theorem. In Section 4, we discuss the implications for valid inference and perform some numerical evaluations. The proof is attached in the Appendix.

6.2 Gaussian Variational Approximation for the Simple Poisson Mixed Model

The simplest Poisson mixed model is a simple GLMM where the fixed effect is a constant and the random effects correspond to a random intercept. The responses conditional on the random effects, are assumed to be Poisson. The simplest Poisson mixed model is:

$$\begin{aligned} &y_{ij}|U_i \text{ independent Poisson with mean } \exp(\beta_0 + U_i) \\ &U_i \text{ independent } N(0, \sigma^2) \end{aligned} \tag{6.1}$$

The observed data y_{ij}, $1 \leq i \leq m$, $1 \leq j \leq n$, are non-negative integers, where typically $m \gg n$. The quantify U_i is unobserved latent variables for ith group. The log likelihood function for the simplest Poisson mixed model (6.1) is given by:

$$\begin{aligned} \ell(\beta_0, \sigma^2) &= \log\left(\int p(\boldsymbol{y}|\beta_0, \boldsymbol{U}) p(\boldsymbol{U}) d\boldsymbol{U}\right) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n}\left\{\beta_0 y_{ij} - \log(y_{ij}) - \frac{1}{2}\log(2\pi\sigma^2)\right\} \\ &\quad + \sum_{i=1}^{m}\log\int \exp\left\{\sum_{j=1}^{n}(y_{ij}U_i) - e^{\beta_0+U_i} - \frac{U_i^2}{\sigma^2}\right\} dU_i, \end{aligned} \tag{6.2}$$

where $\boldsymbol{y} = (y_{11}, ..., y_{1n}, ..., y_{mn})^T$ and $\boldsymbol{U} = (U_1, ..., U_m)^T$. The maximum likelihood estimates of β_0 and σ^2 are:

$$(\underline{\widehat{\beta}}_0, \underline{\widehat{\sigma}}^2) = \arg\max_{\beta_0, \sigma^2} \ell(\beta_0, \sigma^2).$$

The intractable integrals in (6.2) impede maximum likelihood estimation. The GVA specifies that the density function of U_i is a Gaussian density function with

mean μ_i and variance λ_i (Ormerod & Wand, 2012). The integral term of 6.2 for each i, $1 \leq i \leq m$, can be re-written as:

$$\begin{aligned}
&\log \int \exp\left\{\sum_{j=1}^{n}(y_{ij}U_i) - e^{\beta_0+U_i} - \frac{U_i^2}{\sigma^2}\right\} dU_i \\
=&\log \int \exp\left\{\sum_{j=1}^{n}(y_{ij}U_i) - e^{\beta_0+U_i} - \frac{U_i^2}{\sigma^2}\right\} \frac{(2\pi\lambda_i)^{-1/2}\exp\left\{\frac{(U_i-\mu_i)^2}{(2\lambda_i)}\right\}}{(2\pi\lambda_i)^{-1/2}\exp\left\{\frac{(U_i-\mu_i)^2}{(2\lambda_i)}\right\}} dU_i \\
=&-\frac{1}{2}\log(2\pi\lambda_i) + \log E_{U_i}\left[\exp\left\{\sum_{j=1}^{n}(y_{ij}U_i) - e^{\beta_0+U_i} - \frac{U_i^2}{\sigma^2} - \frac{(U_i-\mu_i)^2}{2\lambda_i}\right\}\right].
\end{aligned} \tag{6.3}$$

The lower bound of the expected term in (6.3) can be obtained using Jensen's inequality. It is given by

$$\begin{aligned}
&\log E_{U_i}\left[\exp\left\{\sum_{j=1}^{n}(y_{ij}U_i) - e^{\beta_0+U_i} - \frac{U_i^2}{\sigma^2} - \frac{(U_i-\mu_i)^2}{2\lambda_i}\right\}\right] \\
&\geq E_{U_i}\left[\sum_{j=1}^{n}(y_{ij}U_i) - e^{\beta_0+U_i} - \frac{U_i^2}{\sigma^2} - \frac{(U_i-\mu_i)^2}{2\lambda_i}\right].
\end{aligned} \tag{6.4}$$

Then the lower bound of the likelihood function (6.2) is given by:

$$\ell(\beta_0, \sigma^2) \geq \underline{\ell}(\beta_0, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\lambda}),$$

where $\boldsymbol{\mu} = (\mu_1, ..., \mu_m)$, $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_m)$, and

$$\begin{aligned}
\underline{\ell}(\beta_0, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\lambda}) =& \sum_{i=1}^{m}\sum_{j=1}^{n}\left\{y_{ij}(\beta_0+\mu_i) - e^{\beta_0+\mu_i+\lambda_i/2} - \log(y_{ij})\right\} \\
&- \frac{1}{2\sigma^2}\sum_{i=1}^{m}(\mu_i^2+\lambda_i) - \frac{m}{2}\log(\sigma^2) + \frac{1}{2}\sum_{i=1}^{m}\log(\lambda_i) + \frac{m}{2}.
\end{aligned} \tag{6.5}$$

In the GVA, the lower bound of the likelihood function, $\underline{\ell}(\beta_0, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\lambda})$, is used

to approximate the likelihood function $\ell(\beta_0, \sigma^2)$. The GVA of β_0 and σ^2 are:

$$(\underline{\widehat{\beta}}_0, \underline{\widehat{\sigma}}^2) = (\beta_0, \sigma^2) \text{ component of } \underset{\beta_0,\sigma^2,\boldsymbol{\mu},\boldsymbol{\lambda}}{\arg\max}\, \underline{\ell}(\beta_0, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\lambda}).$$

The algorithm for GVA estimation is presented in Ormerod and Wand (2012).

6.3 Asymptotic Normality Results.

Our upcoming theorem relies on the following assumptions:

(1) $m = m(n)$ diverges to infinity with n, such that $n/m \to 0$ as $n \to \infty$;

(2) for a constant $C > 0$, $m = O(n^C)$ as m and n diverge.

The precise asymptotic behavior of $\underline{\widehat{\beta}}_0$ and $\underline{\widehat{\sigma}}^2$ is conveyed by:

Theorem 6.1. *Assume that conditions (1) and (2) hold. Then:*

$$\underline{\widehat{\beta}}_0 - \beta_0^0 = m^{-1/2} N_0 + o_p\big(n^{-1} + m^{-1/2}\big), \tag{6.6}$$

where the random variable N_0 is normal $N(0, (\sigma^2)^0)$;

$$\underline{\widehat{\sigma}}^2 - (\sigma^2)^0 = m^{-1/2}\, N_1 + o_p(n^{-1} + m^{-1/2}), \tag{6.7}$$

where the random variable N_1 is normal $N(0, 2\{(\sigma^2)^0\}^2)$.

Theorem 6.1 implies:

- GVA estimators, $\underline{\widehat{\beta}}_0$ and $\underline{\widehat{\sigma}}^2$, have asymptotically normal distributions;
- The bias of GVA estimators is $o_p(n^{-1} + m^{-1/2})$, which can be ignored as m and n diverge;

- GVA estimators, $\widehat{\underline{\beta}}_0$ and $\widehat{\underline{\sigma}}^2$, have variances of size $m^{-1/2}$, as m and n diverge;
- the variance of $\widehat{\underline{\beta}}_0$ and $\widehat{\underline{\sigma}}^2$ can be decreased by increasing the group size m.

6.4 Asymptotically Valid Inference

Theorem 6.1 implies that $\widehat{\underline{\beta}}_0$ and $\widehat{\underline{\sigma}}^2$ follow asymptotically normal distributions and converge to the true parameter values, β_0^0 and $(\sigma^2)^0$. Therefore, we can obtain the following approximate $100(1-\alpha)\%$ confidence intervals for β_0^0 and $(\sigma^2)^0$:

$$\begin{aligned} \widehat{\underline{\beta}}_0 &\pm \Phi^{-1}(1-\frac{1}{2}\alpha)\sqrt{\frac{\widehat{\underline{\sigma}}^2}{m}}, \\ \widehat{\underline{\sigma}}^2 &\pm \Phi^{-1}(1-\frac{1}{2}\alpha)\widehat{\underline{\sigma}}^2\sqrt{\frac{2}{m}}. \end{aligned} \tag{6.8}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. These confidence intervals are asymptotically valid since they involve studentization based on consistent estimators of all unknown quantities.

We ran a simulation study to evaluate the coverage properties of the Gaussian variational approximate confidence intervals (6.8). The true parameter vector $(\beta_0^0, (\sigma^2)^0)^T$ was taken to vary over the four members of the following set:

$$\{(2,1),\ (-0.02,0.5),\ (0.5,2),\ (1,0.09)\}$$

The number of observations in each group, n, varied over 10, 20,..., 100 with the number of groups, m, fixed at n^2 throughout the study. For each simulation scenario, we generated 1000 samples and computed 95% confidence intervals based on (6.8).

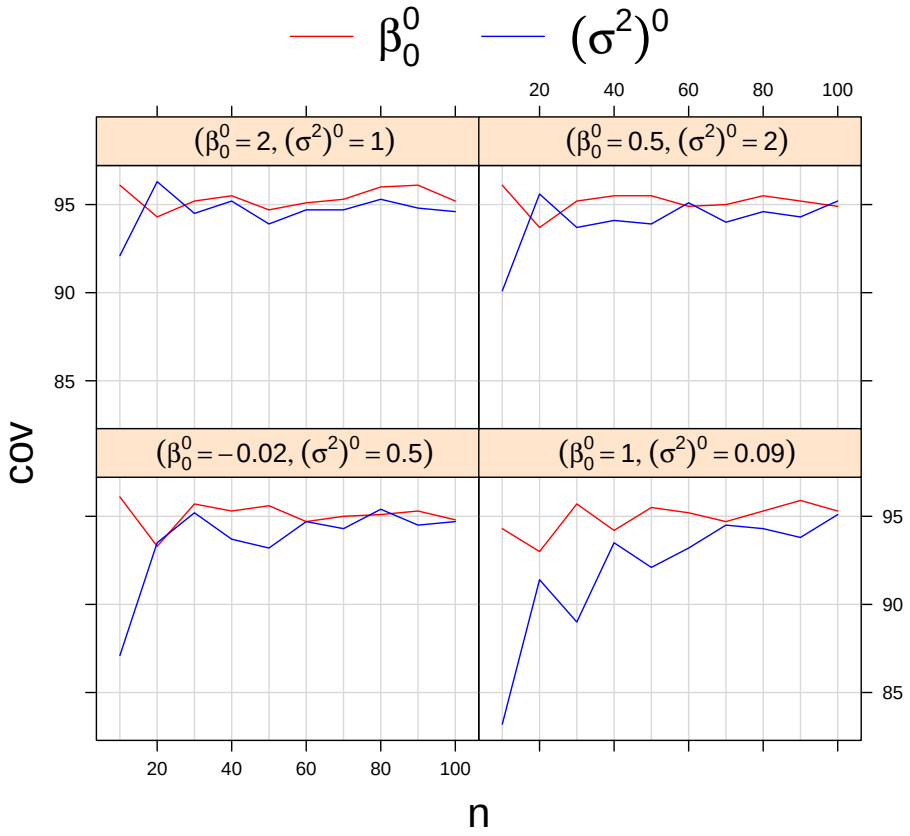

Figure 6.1: Actual coverage percentage of nominally 95% Gaussian variational approximate confidence intervals for the parameters in the simplest Poisson mixed model. The percentages are based on 500 replications. The values of n are 10, 20,...,100. The value of m is fixed at $m = n^2$.

Figure 6.1 shows the actual coverage percentages for the nominally 95% confidence intervals. In the case of β_0^0 for each simulation, the actual and nominal percentages are seen to have very good agreement, even for $(m,\ n) = (100,\ 10)$. The actual coverage is close to 95% within 5% of the nominal level. For $(\sigma^2)^0$, we get very good agreement between the actual and nominal percentages for the first three simulations. For the fourth simulation, we see that $n \geq 30$ is required to get the actual coverage above 90%, that is, within 5% of the nominal level.

The similarity of the actual coverage of β_0^0 and $(\sigma^2)^0$ is in keeping with the same convergence rate apparent from Theorem 6.1.

6.5 Discussion

We derived the asymptotic distributional behavior of GVA estimators of the parameters for the simplest Poisson model. The simulation result showed that GVA confidence intervals possess a excellent coverage properties. Nextly, Hall, Pham, Wand and Wang (2011) have derived the precise asymptotic distributional behavior of Gaussian variational approximate estimators of the parameters in a single-predictor Poisson mixed model. The main barriers is doing further detailed theory of GVA, if we want to use the GVA methods to deal with a complex regression model.

6.A Appendix: Proof

Notation

Write β_0^0 and $(\sigma^2)^0$ to denote the true values of the parameters in model (6.1), and $\underline{\widehat{\beta}}_0$ and $\underline{\widehat{\sigma}}^2$ to denote their respective Gaussian variational approximate estimators. The definitions of the $O_{(k)},\ k = 1, 2, ..., 6$, are in Table 6.1.

Notation	Meaning
$O_{(1)}$	$O_p(m^{-1/2} + n^{-1})$
$O_{(2)}$	$O_p(n^{\varepsilon-1/2})$, uniformly in $1 \leq i \leq m$, for each $\varepsilon > 0$
$O_{(3)}$	$O_p(n^{\varepsilon-1})$, uniformly in $1 \leq i \leq m$, for each $\varepsilon > 0$
$O_{(4)}$	$O_p(n^{\varepsilon-3/2})$, uniformly in $1 \leq i \leq m$, for each $\varepsilon > 0$
$O_{(5)}$	$O_p(m^{-1} + n^{\varepsilon-3/2})$, for each $\varepsilon > 0$
$O_{(6)}$	$O_p\{(m^{-1} + n^{-1/2})^3 n^{\varepsilon}\}$, for each $\varepsilon > 0$

Table 6.1: Definitions of the $O_{(k)}$ notation used in the proofs.

Expression of estimators

Firstly, we obtain the GVA estimators. We derive the lower bound of the likelihood function (6.5), and get:

$$\begin{aligned}
\frac{\partial \underline{\ell}(\beta_0, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\lambda})}{\partial \beta_0} &= \sum_{i=1}^{m}\sum_{j=1}^{n}\left\{y_{ij} - e^{\beta_0+\mu_i+\lambda_i/2}\right\}, \\
\frac{\partial \underline{\ell}(\beta_0, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\lambda})}{\partial \sigma^2} &= \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{m}\left\{\mu_i^2 + \lambda_i\right\} - \frac{m}{2\sigma^2}, \\
\frac{\partial \underline{\ell}(\beta_0, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\lambda})}{\partial \mu_i} &= \sum_{j=1}^{n} y_{ij} - ne^{\beta_0+\mu_i+\lambda_i/2} - \frac{\mu_i}{\sigma^2}, \\
\frac{\partial \underline{\ell}(\beta_0, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\lambda})}{\partial \lambda_i} &= \frac{n}{2}e^{\beta_0+\mu_i+\lambda_i/2} - \frac{1}{2\sigma^2} + \frac{1}{2\lambda_i}.
\end{aligned}$$

Then, the GVA estimators are $\underline{\widehat{\beta}}_0$ and $\underline{\widehat{\sigma}}^2$ subject to:

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\left\{y_{ij}-\exp\left(\underline{\widehat{\beta}}_0+\underline{\widehat{\mu}}_i+\underline{\widehat{\lambda}}_i/2\right)\right\}=0 \tag{6.9}$$

$$\sum_{i=1}^{m}\left\{\underline{\widehat{\mu}}_i^2+\underline{\widehat{\lambda}}_i\right\}-m\underline{\widehat{\sigma}}^2=0 \tag{6.10}$$

$$\sum_{j=1}^{n}y_{ij}-n\exp\left(\underline{\widehat{\beta}}_0+\underline{\widehat{\mu}}_i+\underline{\widehat{\lambda}}_i/2\right)-\frac{\underline{\widehat{\mu}}_i}{\underline{\widehat{\sigma}}^2}=0 \tag{6.11}$$

$$n\exp\left(\underline{\widehat{\beta}}_0+\underline{\widehat{\mu}}_i+\underline{\widehat{\lambda}}_i/2\right)-\frac{1}{\underline{\widehat{\sigma}}^2}+\frac{1}{\underline{\widehat{\lambda}}_i}=0,\ \text{ for } 1\le i\le m. \tag{6.12}$$

Summing (6.11) for $1\le i\le m$ and subtracting (6.9), we get:

$$-mn\frac{\underline{\widehat{\mu}}_i}{\underline{\widehat{\sigma}}^2}=0.$$

Therefore,

$$\underline{\widehat{\mu}}_i=0.$$

We define:

$$\exp(\zeta_i)=\exp(-\beta_0^0-U_i)\left(\frac{1}{n}\sum_{j=1}^{n}y_{ij}-\frac{\underline{\widehat{\mu}}_i}{n\underline{\widehat{\sigma}}^2}\right).$$

and write (6.11) as:

$$\exp\left(\underline{\widehat{\beta}}_0+\underline{\widehat{\mu}}_i+\underline{\widehat{\lambda}}_i/2\right)=\exp\left(\beta_0^0+U_i+\zeta_i\right).$$

Approximate formulae for U_i and λ_i

The convergence in probability of n implies

$$\frac{1}{n}\sum_{j=1}^{n}y_{ij}-E\left[\frac{1}{n}\sum_{j=1}^{n}y_{ij}\right]=O_p\left(\sqrt{\mathrm{Var}\left(\frac{1}{n}\sum_{j=1}^{n}y_{ij}\right)}\right),\ \text{as } n\to\infty.$$

Here we use the Theorem 14.4-1 in Bishop, Fienberg and Holland (2007).

Let $U_i \sim N(0, (\sigma^2)^0)$ and $m = O(n^C)$. From the properties of extrema of Gaussian variables, we get

$$\max_{1\leq i\leq m}|U_i| = O_p\left\{(\log n)^{1/2}\right\}.$$

Then $\max_{1\leq i\leq m}|\exp(U_i)| = O(n^{\varepsilon})$ for any $\varepsilon > 0$. Therefore,

$$\frac{1}{n}\sum_{j=1}^{n} y_{ij} - \exp\left(\beta_0^0 + U_i\right) = O_{(2)}.$$

Then (6.11) implies

$$(1 + O_{(2)})\exp\left(\beta_0^0 + U_i\right) = \exp\left(\underline{\widehat{\beta}}_0 + \underline{\widehat{\mu}}_i + \underline{\widehat{\lambda}}_i/2\right) + \frac{\underline{\widehat{\mu}}_i}{n\underline{\widehat{\sigma}}^2}.$$

Using Theorem 4 of Hall, Ormerod and Wand (2011),

$$\beta_0^0 - \underline{\widehat{\beta}}_0 = O_{(1)},$$

and so we get

$$(1 + O_{(2)})\exp\left(U_i\right) = \exp\left(\underline{\widehat{\mu}}_i + \underline{\widehat{\lambda}}_i/2\right) + \frac{\underline{\widehat{\mu}}_i}{n\,\exp(\beta_0^0)\underline{\widehat{\sigma}}^2}. \tag{6.13}$$

Theorem 2 of Hall, Ormerod and Wand (2011) indicate that

$$\underline{\widehat{\mu}}_i < \underline{\widehat{\sigma}}^2 \sum_{j=1}^{n} y_{ij}.$$

Then, (6.13) implies

$$(1 + O_{(2)})\exp\left(U_i\right) = \exp\left(\underline{\widehat{\mu}}_i + \underline{\widehat{\lambda}}_i/2\right),$$

and, taking logarithms,

$$U_i = \widehat{\underline{\mu}}_i + \frac{1}{2}\widehat{\underline{\lambda}}_i + O_{(2)}. \tag{6.14}$$

Next, we also use Theorem 4 of Hall, Ormerod and Wand (2011) and substitute (6.14) into (6.12), and get

$$\widehat{\underline{\lambda}}_i = (1 + O_{(2)}) \left\{ n \, \exp(\beta_0^0 + U_i) \right\}^{-1}.$$

Therefore,

$$\widehat{\underline{\lambda}}_i = \left\{ n \, \exp(\beta_0^0 + U_i) \right\}^{-1} + O_{(5)}. \tag{6.15}$$

Initial approximations to $\widehat{\underline{\beta}}_0 - \beta_0^0$

Note, taking logarithms of both sides of

$$\exp\left(\widehat{\underline{\beta}}_0 + \widehat{\underline{\mu}}_i + \widehat{\underline{\lambda}}_i/2\right) = \exp\left(\beta_0^0 + U_i + \zeta_i\right),$$

we obtain

$$\widehat{\underline{\beta}}_0 + \widehat{\underline{\mu}}_i + \widehat{\underline{\lambda}}_i/2 = \beta_0^0 + U_i + \zeta_i. \tag{6.16}$$

Combining (6.16) for $1 \leq i \leq m$, we obtain

$$\widehat{\underline{\beta}}_0 - \beta_0^0 \;=\; \frac{1}{m}\sum_{i=1}^{m}\left\{U_i + \zeta_i - \widehat{\underline{\mu}}_i - \widehat{\underline{\lambda}}_i/2\right\}. \tag{6.17}$$

Using (6.15), we obtain

$$\widehat{\underline{\beta}}_0 - \beta_0^0 \;=\; \frac{1}{m}\sum_{i=1}^{m}\{U_i + \zeta_i\} - \frac{1}{2m}\sum_{i=1}^{m}\left\{n \, \exp(\beta_0^0 + U_i)\right\}^{-1} + O_{(5)}. \tag{6.18}$$

Using $E[\exp(-U)] = \exp\left\{\frac{1}{2}(\sigma^2)^2\right\}$, we obtain

$$\widehat{\underline{\beta}}_0 - \beta_0^0 = \frac{1}{m}\sum_{i=1}^{m}(U_i + \zeta_i) - \left\{2n\,\exp(\beta_0^0 - \frac{1}{2}(\sigma^2)^0)\right\}^{-1} + O_{(5)}.$$

Approximation to ζ_i

From the 2nd order Taylor's Formula for $\exp(\widehat{\underline{\beta}}_0)$, we obtain,

$$\exp\widehat{\underline{\beta}}_0 = \left[1 + (\widehat{\underline{\beta}}_0 - \beta_0^0) + \frac{1}{2}(\widehat{\underline{\beta}}_0 - \beta_0^0)\right]\exp(\beta_0^0) + O_{(6)}.$$

Define

$$\delta_i = -U_i + \widehat{\underline{\mu}}_i + \frac{1}{2}\widehat{\underline{\lambda}}_i,$$

where $\delta_i = O_{(2)}$ from (6.14). Then, the left-hand side of (6.11) is equal to

$$\begin{aligned}
\sum_{j=1}^{n} y_{ij} - n\,\exp\left(\widehat{\underline{\beta}}_0 + \widehat{\underline{\mu}}_i + \widehat{\underline{\lambda}}_i/2\right) - \frac{\widehat{\underline{\mu}}_i}{\widehat{\underline{\sigma}}^2} \\
&= \sum_{j=1}^{n} y_{ij} - \frac{U_i + \delta_i - \frac{1}{2}\widehat{\underline{\lambda}}_i}{\widehat{\underline{\sigma}}^2} \\
&\quad -n\,\exp(\widehat{\underline{\beta}}_0)\exp(U_i + \delta_i) \\
&= \sum_{j=1}^{n} y_{ij} - \frac{U_i + \delta_i - \frac{1}{2}\widehat{\underline{\lambda}}_i}{\widehat{\underline{\sigma}}^2} \\
&\quad -n\,\exp(\widehat{\underline{\beta}}_0)\exp(U_i)(1 + \delta_i + \frac{1}{2}\delta_i^2 + O_{(4)}) \\
&= \sum_{j=1}^{n} y_{ij} - \frac{U_i + \delta_i - \frac{1}{2}\widehat{\underline{\lambda}}_i}{\widehat{\underline{\sigma}}^2} \\
&\quad -n\left[1 + (\widehat{\underline{\beta}}_0 - \beta_0^0) + \frac{1}{2}(\widehat{\underline{\beta}}_0 - \beta_0^0)\right]\exp(\beta_0^0) \\
&\quad \times\exp(U_i)(1 + \delta_i + \frac{1}{2}\delta_i^2 + O_{(4)}) + nO_{(6)}.
\end{aligned}$$

Next define

$$\Delta_i = \sum_{j=1}^{n} y_{ij} - n \exp(\beta_0^0 + U_i),$$

and

$$\chi_i = n[(\widehat{\underline{\beta}}_0 - \beta_0^0) + \frac{1}{2}(\widehat{\underline{\beta}}_0 - \beta_0^0)^2]\exp(\beta_0^0).$$

Then, (6.11) implies

$$\begin{aligned} & \Delta_i - \frac{U_i + \delta_i - \frac{1}{2}\widehat{\underline{\lambda}}_i}{\widehat{\underline{\sigma}}^2} - n(\delta_i + \frac{1}{2}\delta_i^2 + O_{(4)})\exp(\beta_0^0)\exp(U_i) \\ = \; & \chi_i(1 + \delta_i + \frac{1}{2}\delta_i^2 + O_{(4)})\exp(U_i) + nO_{(6)}. \end{aligned}$$

Therefore,

$$\begin{aligned} & \delta_i + \frac{1}{2}\delta_i^2 \frac{[n \exp(\beta_0^0) + \chi_i]\exp(U_i)}{[n \exp(\beta_0^0) + \chi_i]\exp(U_i) + (\widehat{\underline{\sigma}}^2)^{-1}} \\ = \; & \frac{\Delta_i - \chi_i \exp(U_i) - (\widehat{\underline{\sigma}}^2)^{-1}(U_i - \widehat{\underline{\lambda}}_i)}{[n \exp(\beta_0^0) + \chi_i]\exp(U_i) + (\widehat{\underline{\sigma}}^2)^{-1}} + O_{(6)} + O_{(4)}. \end{aligned}$$

This implies

$$\begin{aligned} \delta_i &= \frac{\Delta_i - \chi_i \exp(U_i)}{[n \exp(\beta_0^0) + \chi_i]\exp(U_i)} + O_{(3)} \\ &= \left\{n \exp(\beta_0^0)\right\}^{-1} \exp(-U_i)[\Delta_i - \chi_i \exp(U_i)] - (\widehat{\underline{\beta}}_0 - \beta_0^0) + O_{(3)}. \end{aligned}$$

Therefore, we deduce that

$$\begin{aligned}
\underline{\widehat{\mu}}_i &= U_i + \delta_i - \frac{1}{2}\underline{\widehat{\lambda}}_i \\
&= U_i - \frac{1}{2}\left\{n\,\exp(\beta_0^0 + U_i)\right\}^{-1} \\
&\quad + \left\{n\,\exp(\beta_0^0)\right\}^{-1}\exp(-U_i)[\Delta_i - \chi_i\exp(U_i)] - (\underline{\widehat{\beta}}_0 - \beta_0^0) + O_{(3)} \\
&= U_i + \left\{n\,\exp(\beta_0^0)\right\}^{-1}\exp(-U_i)[\Delta_i - \chi_i\exp(U_i)] \\
&\quad -(\underline{\widehat{\beta}}_0 - \beta_0^0) + O_{(3)}.
\end{aligned}$$

Defining $\bar{U} = \underline{\widehat{\beta}}_0 - \beta_0^0$, we obtain

$$\underline{\widehat{\mu}}_i = U_i + \bar{U} + \left\{n\,\exp(\beta_0^0)\right\}^{-1}\exp(-U_i)[\Delta_i - \chi_i\exp(U_i)] + O_{(3)}. \tag{6.19}$$

Combing the definition of ζ_i and (6.19), we obtain

$$\begin{aligned}
1 + \zeta_i + \zeta_i^2 &= \exp\{-\beta_0^0 - U_i\}\left(\frac{1}{n}\sum_{j=1}^{n} y_{ij} - \frac{\underline{\widehat{\mu}}_i}{n\underline{\widehat{\sigma}}^2}\right) + O_{(4)} \\
&= \exp\{-\beta_0^0 - U_i\}\left(\frac{1}{n}\sum_{j=1}^{n} y_{ij} - \exp\{\beta_0^0 + U_i\}\right) \\
&\quad +\exp\{-\beta_0^0 - U_i\}(n\underline{\widehat{\sigma}}^2)^{-1}\left\{\bar{U} + \frac{\exp(-U_i)[\Delta_i - \chi_i\exp(U_i)]}{n\,\exp(\beta_0^0)}\right\} \\
&\quad +\exp\{-\beta_0^0 - U_i\}(n\underline{\widehat{\sigma}}^2)^{-1}U_i + 1 + O_{(4)}.
\end{aligned}$$

Therefore,

$$\zeta_i + \zeta_i^2 = n^{-1}\exp\{-\beta_0^0 - U_i\}[\Delta_i - (\underline{\widehat{\sigma}}^2)^{-1}U_i] + O_{(4)}$$

and

$$\begin{aligned}
\zeta_i &= n^{-1}\exp\{-\beta_0^0 - U_i\}[\Delta_i - (\underline{\widehat{\sigma}}^2)^{-1}U_i] \\
&\quad -\frac{1}{2}n^{-2}\exp\{-2\beta_0^0 - 2U_i\}\Delta_i^2 + O_{(4)}.
\end{aligned}$$

Final approximations to $\widehat{\underline{\beta}}_0 - \beta_0^0$

Note,

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}\zeta_i &= n^{-1}\exp(-\beta_0^0)E\left\{\exp(-U_i)[\Delta_i - (\widehat{\underline{\sigma}}^2)^{-1}U_i]\right\} \\ &\quad - \frac{1}{2n^2}\exp(-2\beta_0^0)E\left[\exp(-2U_i^2)\Delta_i^2\right] + O_p(n^{-1}) \qquad (6.20)\\ &= \left\{2n\,\exp(\beta_0^0 - \frac{1}{2}(\sigma^2)^0)\right\}^{-1} + O_p(n^{-1}).\end{aligned}$$

Combining (6.20) and (6.18), we obtain,

$$\widehat{\underline{\beta}}_0 - \beta_0^0 = \frac{1}{m}\sum_{i=1}^{m}U_i + O_p(m^{-1/2} + n^{-1}). \qquad (6.21)$$

Result (6.6) of Theorem 6.1 is a direct consequence of (6.21).

Final approximations to $\widehat{\underline{\sigma}}^2 - (\sigma^2)_0^0$

Using (6.16), we get:

$$\begin{aligned}\widehat{\underline{\mu}}_i &= \beta_0^0 - \widehat{\underline{\beta}}_0 + U_i + \zeta_i - \widehat{\underline{\lambda}}_i/2 \\ &= U_i + \zeta_i - \left\{2n\,\exp(\beta_0^0 + U_i)\right\}^{-1} \\ &\quad -\frac{1}{m}\sum_{i=1}^{m}\{U_i + \zeta_i\} + \frac{1}{m}\sum_{i=1}^{m}\left\{n\,\exp(\beta_0^0 - \frac{1}{2}(\sigma^2)^0)\right\}^{-1} + O_{(5)}.\end{aligned}$$

Taking the squares and adding them together, we get.

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m}\underline{\widehat{\mu}}_i^2 \\
&\quad = \frac{1}{m}\sum_{i=1}^{m}\left(U_i+\zeta_i-\bar{U}-\bar{\zeta}\right)^2 \\
&\qquad -\frac{1}{m}\sum_{i=1}^{m}\left\{\left[n\,\exp(\beta_0^0+U_i)\right]^{-1}\left(U_i+\zeta_i-\bar{U}-\bar{\zeta}\right)\right\}+O_{(5)}.
\end{aligned} \tag{6.22}$$

Combining (6.10), (6.15) and (6.22), we obtain

$$\begin{aligned}
\underline{\widehat{\sigma}}^2 &= \frac{1}{m}\sum_{i=1}^{m}(\underline{\widehat{\lambda}}_i+\underline{\widehat{\mu}}_i^2) \\
&= \frac{1}{m}\sum_{i=1}^{m}\left(U_i+\zeta_i-\bar{U}-\bar{\zeta}\right)^2+\frac{1}{m}\sum_{i=1}^{m}\left\{n\,\exp(\beta_0^0+U_i)\right\}^{-1} \\
&\quad -\frac{1}{m}\sum_{i=1}^{m}\left\{\left[n\,\exp(\beta_0^0+U_i)\right]^{-1}\left(U_i+\zeta_i-\bar{U}-\bar{\zeta}\right)\right\}+O_{(5)} \\
&= \frac{1}{m}\sum_{i=1}^{m}\left(U_i+\zeta_i-\bar{U}-\bar{\zeta}\right)^2 \\
&\quad -\left\{n\,\exp\left(\beta_0^0-\frac{1}{2}(\sigma^2)^0\right)\right\}^{-1}\left(1+(\sigma^2)^0\right)+O_{(5)} \\
&= \frac{1}{m}\sum_{i=1}^{m}U_i^2+\frac{1}{m}\sum_{i=1}^{m}\zeta_i^2+\frac{2}{m}\sum_{i=1}^{m}U_i\zeta_i-c_o n^{-1}\left(1+(\sigma^2)^0\right)+O_{(5)},
\end{aligned} \tag{6.23}$$

where

$$c_0=\left\{\exp\left(\beta_0^0-\frac{1}{2}(\sigma^2)^0\right)\right\}^{-1}.$$

Note,

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m}\zeta_i^2 \\
&\quad= \frac{1}{mn^2}\sum_{i=1}^{m}\exp(-2\beta_0^0 - 2U_i)\Delta_i^2 + o_p(n^{-1}) \\
&\quad= \frac{1}{n^2}E\left[\exp(-2\beta_0^0 - 2U_i)\Delta_i^2\right] + o_p(n^{-1}) \\
&\quad= \frac{1}{n}E\left[\exp(-\beta_0^0 - U_i)\right] + o_p(n^{-1}) \\
&\quad= c_0 n^{-1} + o_p(n^{-1}),
\end{aligned} \tag{6.24}$$

and

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m}U_i\zeta_i \\
&\quad= -\frac{1}{mn(\sigma^2)^0}\sum_{i=1}^{m}U_i^2\exp(-\beta_0^0 - U_i) \\
&\qquad- \frac{1}{2mn^2}\sum_{i=1}^{m}U_i\exp(-\beta_0^0 - U_i)\Delta_i^2 + o_p(n^{-1}) \\
&\quad= -\frac{\exp(-\beta_0^0)}{n(\sigma^2)^0}E\left[U_i^2\exp(-U_i)\right] + o_p(n^{-1}) \\
&\quad= \left\{n\,\exp\left(\beta_0^0 - \frac{1}{2}(\sigma^2)^0\right)\right\}^{-1}\left(1 + \frac{(\sigma^2)^0}{2}\right) + o_p(n^{-1}) \\
&\quad= c_o n^{-1}\left(1 + \frac{(\sigma^2)^0}{2}\right) + o_p(n^{-1}).
\end{aligned} \tag{6.25}$$

Combining (6.23), (6.24) and (6.25), we obtain

$$\underline{\widehat{\sigma}}^2 = \frac{1}{m}\sum_{i=1}^{m}U_i^2 + O_p(m^{-1/2} + n^{-1}),$$

which implies

$$\widehat{\underline{\sigma}}^2 - (\sigma^2)^0 = \frac{1}{m}\sum_{i=1}^{m}\left\{U_i^2 - (\sigma^2)^0\right\} + O_p(m^{-1/2} + n^{-1}). \tag{6.26}$$

Result (6.7) of Theorem 6.1 is a direct consequence of (6.26).

Chapter 7

A New Mean Field Variational Bayes Inference Machine

7.1 Introduction

BUGS (Bayesian inference Using Gibbs Sampling) is a well-known Bayesian inference software package. The BUGS user specifies a statistical model by simply stating the distributional relationships between variables. The software then determines an appropriate MCMC scheme for the specified model. BUGS includes two main versions, namely WinBUGS (Lunn, Thomas, Best & Spiegelhalter, 2000) and OpenBUGS. WinBUGS controls the Bayesian analysis by standard 'point-and-click' operations. Many researchers from the statistical community are familiar with `OpenBUGS` because it can run in other statistical environments, such as R and SAS. The R package `BRugs` provides an interface to `OpenBUGS` and allows fitting and inference for Bayesian models using MCMC techniques in R.

In the wake of developments in Bayesian inference for graphical models, a new computational framework for approximate Bayesian inference, `Infer.NET`, was launched (Minka *et al.*, 2014). `Infer.NET` performs Bayesian inference by using

deterministic approximations methods: MFVB and expectation propagation. We showed how to using `Infer.NET` to deal with statistical models in Chapter 5. Although a link tools, *InferNETSupport*, was coded to perform `Infer.NET` within the R environment, there is an difficulty for users in the statistical community in that they have to learn new programing languages, the so-called .NET languages.

In this chapter, I introduce a new R function, `InferMachine()`, which can perform MFVB by using the same model specification syntax that `BUGS` uses. Currently, `InferMachine()` only support the Gaussian and binary response models with specific distributional forms. However, the principle applies to a much larger class of models.

7.2 User Manual for InferMachine()

The R function `InferMachine()` reads `BUGS` model files and input data and performs the MFVB inference. In this section, I will introduce the function's usage, arguments and output value. The manual page for `InferMachine()` is:

Usage

```
InferMachine(modelFile, data, numIter)
```

Arguments

```
modelFile:
    a character string giving the pathname of the model file that
    specifies the BUGS model structure.

data:
    a list including name-value pairs.  The name specifies the name
    of any hyperparameters or input data in the BUGS model
```

```
    structure file, and the value specifies the corresponding
    value of any hyperparameters or input data.

numIter:
    the number of mean field variational Bayes iterations.
```

Value

```
The InferMachine() return a list including variable names and para
-meters of its posterior density function.
```

7.3 Illustration for Gaussian Response Models

In this section, we will give some examples involving Gaussian response models, including the simple linear model, the ridge penalized linear model, the nonparametric regression model and the semiparametric mixed model. We also make some accuracy comparisons between `InferMachine()` and `BUGS`.

7.3.1 Simple linear model

Consider the Bayesian simple linear model:

$$\begin{aligned} &y_i|\beta_0,\beta_1,\tau_\varepsilon \overset{\text{ind.}}{\sim} N(\beta_0+\beta_1 x_i,\tau_\varepsilon^{-1}), \quad 1\le i\le n,\\ &\beta_0,\beta_1 \overset{\text{ind.}}{\sim} N(0,\sigma_\beta^2),\ \ \tau_\varepsilon\sim \text{Gamma}(A,B), \end{aligned} \tag{7.1}$$

where A, B and $\sigma_\beta^2>0$ are hyperparameters. The `BUGS` model code file, `linearModel.txt`, for specifying (7.1) is

```
linearModel
   {
      for (i in 1:n)
```

```
    {
       mu[i] <- beta0+beta1*x[i]
       y[i] ~ dnorm(mu[i],tau)
    }
    beta0 ~ dnorm(0,tauBeta) ; beta1 ~ dnorm(0,tauBeta)
    tau ~ dgamma(A,B)
}
```

The R code that processes the `BUGS` file, including the data and hyperparameters specification, is as fellows:

```
# Set hyperparameters:
   sigsqBeta <- 1e08 ; A <- 0.01 ; B <- 0.01

# Set up input data and hyperparameters:
   allData <- list(n=n,x=age,y=price,
                   tauBeta=1/sigsqBeta,A=A,B=B)
```

The list `allData` stores the input data and hyperparameters. The variables `sigsqBeta`, `A` and `B` correspond to the hyperparameters σ^2_β, A and B, which are set at $\sigma^2_\beta = 10^8$ and $A = B = 0.01$. We fit (7.1) for data on the age and price of $n = 39$ Mitsubishi cars (Smith, 1998). The `x` and `y` are vectors containing the observed values of age and price, and `n` is the number of observations. Fitting is then performed by using

```
   fit <- InferMachine("linearModel.txt",allData,200)
```

For model (7.1), the function `InferMachine()` does the MFVB inference by imposing the product restriction

$$q(\beta_0, \beta_1, \tau_\varepsilon) = q(\beta_0, \beta_1)q(\tau_\varepsilon).$$

The output is a list with the variable name and values of the approximate posterior density function parameters. The names of the output list includes `beta0`, `beta1`, `tau` and `muR`. The function $q(\beta_0, \beta_1)$ is the Multivariate Normal density function with mean vector `fit$muR[[1]]` and covariance matrix `fit$muR[[2]]`. The posterior density function $q(\tau_\varepsilon)$ is a Gamma density function, where the shape parameter is the list entry `fit$tau[[1]]` and the rate parameter is the list entry `fit$tau[[2]]`. The marginal posterior density function $q(\beta_0)$ is a Normal density function with mean `fit$beta0[[1]]` and variance `fit$beta0[[2]]`. The marginal posterior density function $q(\beta_1)$ is a Normal density function with mean `fit$beta1[[1]]` and variance `fit$beta1[[2]]`

Figure 7.1 shows that the result for `InferMachine()` is the same as that from `Infer.NET`. Compared with the MCMC, the function `InferMachine()` can obtain highly accurate results for posterior density functions. Note that we use $\sigma^2 \equiv 1/\tau$, instead of τ, for the error variance.

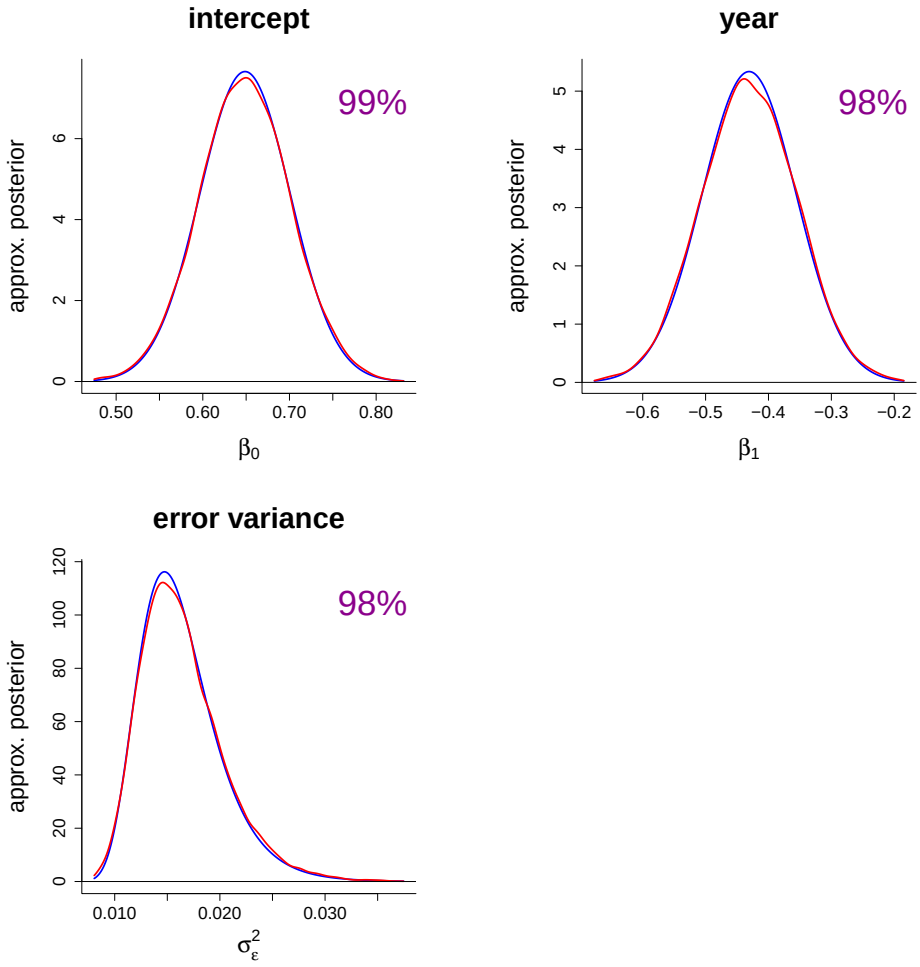

Figure 7.1: The posterior density functions produced by `InferMachine()` (blue), `Infer.NET` (yellow) and MCMC (orange) for a simple linear regression fit to the Mitsubishi car price/age data. The percentages are the accuracies of the `InferMachine()` fit compared with the MCMC fit.

7.3.2 Ridge penalized linear model

The ridge penalized linear model was defined in (2.5). Consider a Bayesian ridge penalized linear model

$$
\begin{aligned}
&\boldsymbol{y}|\beta_0, \boldsymbol{\beta}, \tau_\varepsilon \sim N(\mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta}, \tau_\varepsilon^{-1}\boldsymbol{I}),\\
&\beta_0 \sim N(0, \sigma^2_{\beta_0}),\\
&\boldsymbol{\beta}|\tau_\beta \sim N(\mathbf{0}, \tau_\beta^{-1}\boldsymbol{I}),\\
&\tau_\varepsilon \sim \text{Gamma}(A_\varepsilon, B_\varepsilon),\\
&\tau_\beta \sim \text{Gamma}(A_\beta, B_\beta),
\end{aligned}
\tag{7.2}
$$

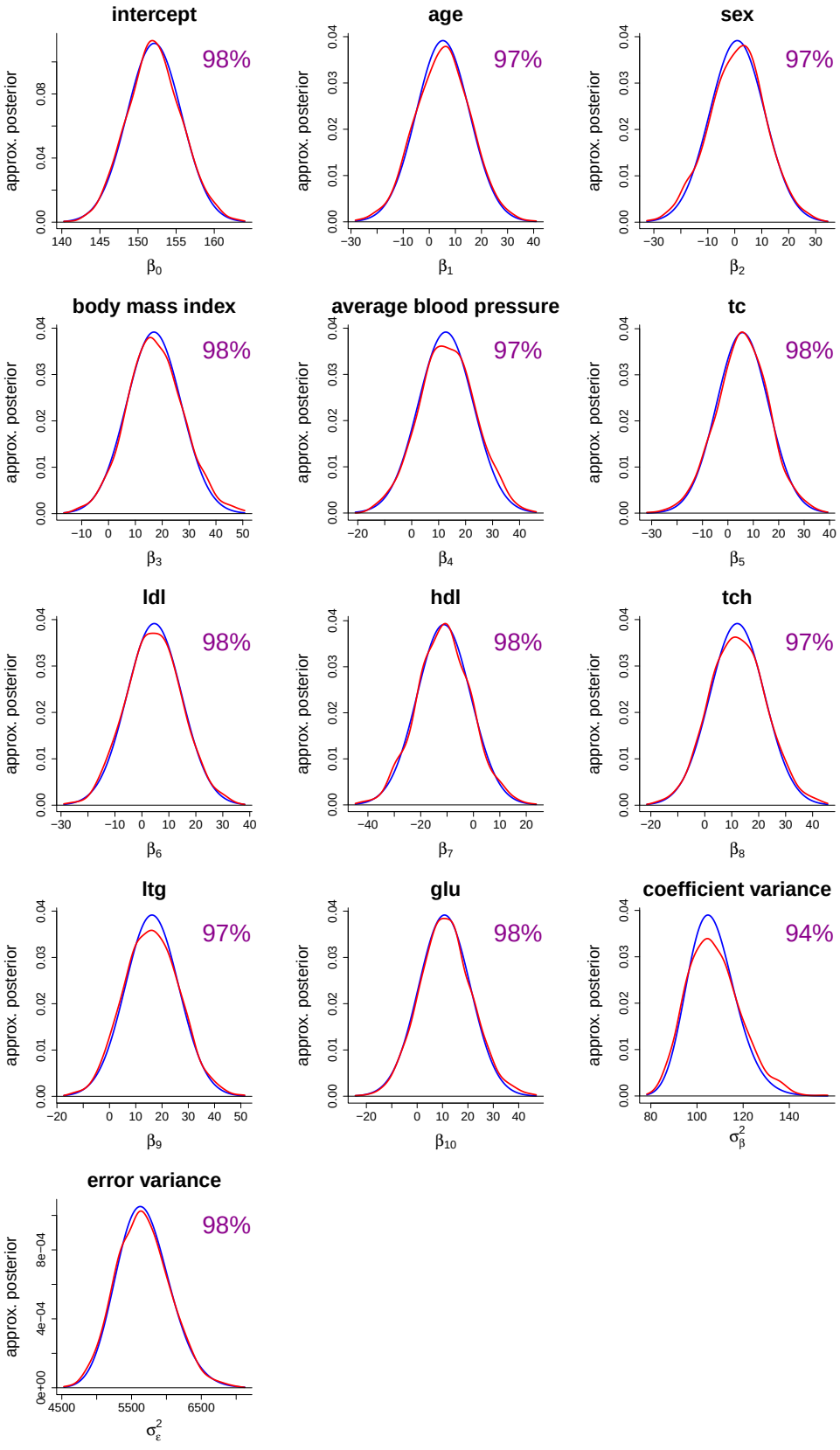

Figure 7.2: The approximate posterior density functions produced by `InferMachine()` (blue) and the MCMC method (orange) for a ridge penalized linear model to fit diabetes data. The percentages are the accuracies of the `InferMachine()` fit compared with the MCMC fit.

where A_ε, B_ε, $\sigma^2_{\beta_0}$, A_β and $B_\beta > 0$ are hyperparameters. The `BUGS` model code file, `RidgeLinearModel.txt`, for specifying (7.2) is

```
RidgeLinearModel
{
   for (i in 1:n)
   {
      mu[i] <- beta0 + inprod(beta[],X[i,])
      y[i] ~ dnorm(mu[i],tauEps)
   }
   for (j in 1:p)
   {
      beta[j] ~ dnorm(0,tauBeta)
   }
   beta0 ~ dnorm(0,tauBeta0)
   tauEps ~ dgamma(Aeps,Beps)
   tauBeta ~ dgamma(Abeta,Bbeta)
}
```

The R code data and hyperparameters specification is

```
allData <- list(n=n,X=X,y=y,p=p,
                tauBeta0=1/sigsqBeta0,
                Abeta=Abeta,Bbeta=Bbeta,
                Aeps=Aeps,Beps=Beps)
```

The variables `Aeps`, `Beps`, `sigsqBeta0`, `ABeta` and `Bbeta` correspond to hyperparameters A_ε, B_ε, $\sigma^2_{\beta_0}$, A_β and B_β. The quantify `n` is the number of observations, and `p` is the number of predictors. For the model (7.2), the MFVB inference

imposes the product restriction

$$q(\beta_0, \boldsymbol{\beta}, \tau_\varepsilon, \tau_\beta) = q(\beta_0, \boldsymbol{\beta}) q(\tau_\varepsilon) q(\tau_\beta).$$

We fit the ridge penalized linear model for the diabetes data, which have been used in Chapter 4, and compare the results of `InferMachine()` and MCMC. The hyperparameters were set at $A_\varepsilon = 0.01$, $B_\varepsilon = 0.01$, $\sigma^2_{\beta_0} = 10^8$, $A_\beta = 0.01$ and $B_\beta = 0.01$. Fitting is then performed using:

```
fit <- InferMachine("RidgeLinearModel.txt",allData,200)
```

The `BUGS` code for model (7.1) sets the β_0 and β_1 separately, and the `BUGS` code uses a vector $\boldsymbol{\beta}$ for model (7.2) and also sets β_0 and $\boldsymbol{\beta}$ separately. In `InferMachine()`, the (β_0, β_1) or $(\beta_0, \boldsymbol{\beta})$ were blocked in the product restriction. Similarly to the simple linear example, the parameters of the posterior density functions $q(\beta_0, \boldsymbol{\beta})$ were stored in `fit$muR`, and the parameters of the marginal posterior density functions $q(\beta_0)$ and $q(\boldsymbol{\beta})$ were stored in `fit$beta0` and `fit$beta`. Figure 7.2 shows that the function `InferMachine()` can get a high accuracy result in the posterior density functions. Note that we use $\sigma^2 \equiv 1/\tau$, instead of τ, for the error variance and coefficients variance.

7.3.3 Simple nonparametric regression

Consider a one-predictor Gaussian nonparametric regression model:

$$y_i = f(x_i) + \varepsilon_i,$$
$$\varepsilon_i \sim N(0, \sigma^2_\varepsilon),$$

where, for $1 \leq i \leq n$, the y_i are measurements on a continuous response variable and the x_i are continuous predictor variables. By using low-rank smoothing splines

with mixed model representation, we can get

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} u_k \boldsymbol{z}_k(x),$$

$$u_k \overset{\text{ind.}}{\sim} N(0, \sigma_u^2),$$

where $\boldsymbol{z}_k$, $1 \leq k \leq K$ is an O'Sullivan basis spline (Wand & Ormerod, 2008). Then, the Bayesian nonparametric regression model is

$$\begin{aligned}
&\boldsymbol{y}|\beta_0, \beta_1, \tau_\varepsilon \sim N(\mathbf{1}\beta_0 + \boldsymbol{x}\beta_1 + \boldsymbol{Z}\boldsymbol{u}, \tau_\varepsilon^{-1}\boldsymbol{I}),\\
&\beta_0, \beta_1 \overset{\text{ind.}}{\sim} N(0, \sigma_\beta^2),\\
&\boldsymbol{u} \sim N(\mathbf{0}, \tau_u^{-1}\boldsymbol{I}),\\
&\tau_\varepsilon \sim \text{Gamma}(A_\varepsilon, B_\varepsilon),\\
&\tau_u \sim \text{Gamma}(A_u, B_u),
\end{aligned} \tag{7.3}$$

where A_ε, B_ε, $\sigma_{\beta_0}^2$, A_u and $B_u > 0$ are hyperparameters, and

$$\boldsymbol{Z} = \begin{bmatrix} z_1(x_1) & \cdots & z_k(x_1) \\ \vdots & \ddots & \vdots \\ z_1(x_n) & \cdots & z_k(x_n) \end{bmatrix}$$

is the spline basis design matrix. The `BUGS` model code file, `fossilModel.txt`, for specifying (7.3) is

```
fossilModel
{
   for(i in 1:n)
   {
      mu[i] <- beta0 + beta1*x[i] + inprod(u[],Z[i,])
      y[i] ~ dnorm(mu[i],tauEps)
```

```
   }
   for (k in 1:numKnots)
   {
      u[k] ~  dnorm(0,tauU)
   }
   beta0 ~ dnorm(0,tauBeta)
   beta1 ~ dnorm(0,tauBeta)
   tauU ~ dgamma(AU,BU)
   tauEps ~ dgamma(AEps,BEps)
}
```

Similarly, we store the `BUGS` data file, including the data and hyperparameters specification, in `allData` as follows:

```
allData <- list(n=n,numKnots=K,
                x=age,y=strontium.ratio,Z=Z,
                tauBeta=1/sigsqBeta,
                AU=AU,BU=BU,AEps=AEps,BEps=BEps)
```

We fit model (7.3) for data on the age and the strontium ratio of fossil (Ruppert *et al.*, 2003). The quantifies `Aeps`, `Beps`, `sigsqBeta`, `Au` and `Bu` correspond to hyperparameters A_ε, B_ε, σ_β^2, A_u and B_u. The quantify `n` is the number of observations, and `K` is the number of spline basis functions. The quantifies `x` and `y` are vectors containing the observed values of age and strontium ratio, and `Z` is the spline basis design matrix $\boldsymbol{Z}$. For model (7.3), the MFVB inference imposes the product restriction

$$q(\beta_0, \beta_1, \boldsymbol{u}, \tau_\varepsilon, \tau_u) = q(\beta_0, \beta_1, \boldsymbol{u})q(\tau_\varepsilon)q(\tau_u).$$

The hyperparameters are set at $A_\varepsilon = 0.01$, $B_\varepsilon = 0.01$, $\sigma^2_{\beta_0} = 10^8$, $A_u = 0.01$ and $B_u = 0.01$. A fitting is performed by using:

```
fit <- InferMachine("fossilModel.txt",allData,200)
```

Figure 7.3 shows the fitted curves and pointwise 95% credible sets for the fossil data, and Figure 7.4 shows approximate posterior density functions for σ^2_u and σ^2_ϵ (Note that we also use $\sigma^2 \equiv 1/\tau$, instead of τ, for the error variance and random variance). It is seen that both the `InferMachine()` and MCMC fits are quite similar in terms of point estimation, interval estimation and posterior density functions.

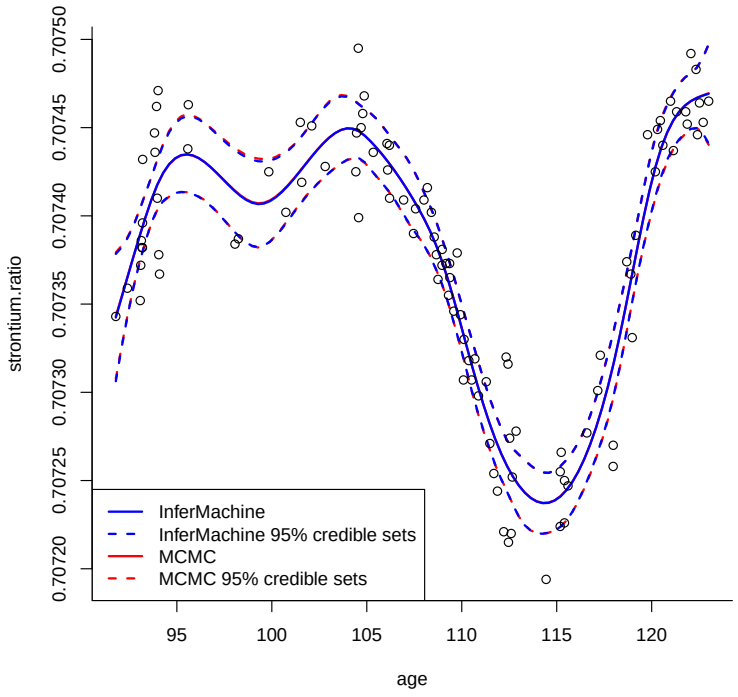

Figure 7.3: Fossil data. Fitted regression line and pointwise 95% credible intervals from `InferMachine()` and the MCMC method for a simple nonparametric regression.

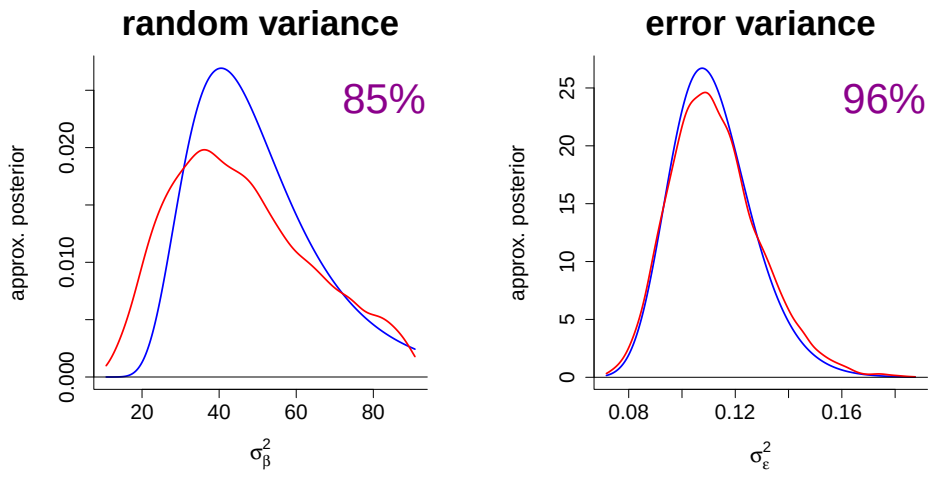

Figure 7.4: The approximate posterior density functions produced by `InferMachine()` (blue) and the MCMC method (orange) for a simple nonparametric regression to fit fossil data. The percentages are the accuracies of the `InferMachine()` fit compared with the MCMC fit.

7.3.4 Semiparametric mixed model

The nonparametric model based on penalized splines is in a mixed model framework and hence can be performed by `InferMachine()`. In this section, I illustrate the use of `InferMachine()` for fitting semiparametric mixed models, which include a splines component and a random effect. It can be used to fit some longitudinal data sets. The model has the form

$$y_{ij}|\boldsymbol{\beta}, \boldsymbol{u}, U_i, \sigma^2_\epsilon \overset{\text{ind.}}{\sim} N\left(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_{ij} + U_i + \sum_{k=1}^{K} u_k z_k(s_{ij}), \sigma^2_\epsilon\right),$$
$$U_i|\sigma^2_U \overset{\text{ind.}}{\sim} N(0, \sigma^2_U), \;\; 1 \leq j \leq n_i, \;\; 1 \leq i \leq m,$$

where x_{ij} is a vector of predictors that enter the model linearly, and s_{ij} is another predictor that enters the model non-linearly via penalized splines. For each $1 \leq i \leq m$, U_i denotes the random intercept for the ith subject. The corresponding

Bayesian semiparametric mixed model is

$$\begin{aligned}
&\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \tau_\varepsilon \sim N(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \tau_\varepsilon^{-1}\boldsymbol{I}),\\
&\boldsymbol{u}|\tau_U, \tau_u \sim N\left(\mathbf{0}, \begin{bmatrix} \tau_U^{-1}\boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & \tau_u^{-1}\boldsymbol{I} \end{bmatrix}\right),\\
&\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma_\beta^2\boldsymbol{I}),\\
&\tau_U \sim \text{Gamma}(A_U, B_U),\\
&\tau_u \sim \text{Gamma}(A_u, B_u),\\
&\tau_\epsilon \sim \text{Gamma}(A_\varepsilon, B_\varepsilon),
\end{aligned} \tag{7.4}$$

where σ_β^2, A_U, B_U, A_u, B_u A_ϵ and $B_\epsilon > 0$ are user-specified hyperparameters. Introduction of the random intercepts results in

$$\boldsymbol{u} = \begin{bmatrix} U_1 \\ \vdots \\ U_m \\ u_1 \\ \vdots \\ u_K \end{bmatrix}, \ \boldsymbol{Z} = \begin{bmatrix} 1 & \cdots & 0 & z_1(s_{11}) & \cdots & z_K(s_{11}) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & \cdots & 0 & z_1(s_{1n_1}) & \cdots & z_K(s_{1n_1}) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & z_1(s_{m1}) & \cdots & z_K(s_{m1}) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & z_1(s_{mn_m}) & \cdots & z_K(s_{mn_m}) \end{bmatrix}.$$

I fitted this semiparametric mixed model for data from a study on spinal bone mineral density (Bachrach *et al.*, 1999). A population of 230 female subjects aged between 8 and 27 was followed over time and each subject contributed either one, two, three or four spinal bone mineral density measurements. Data on ethnicity are available, and the entries for x_{ij} correspond to the indicator variables for Black (x_{1ij}), Hispanic (x_{2ij}) and White (x_{3ij}), with Asian ethnicity corresponding to the baseline. Age enters the model non-linearly and corresponds to s_{ij} and x_{4ij}. We

respectively specify the random intercept term and smoothing term in the BUGS code as

$$\boldsymbol{U}_{obj} = \begin{bmatrix} U_1 \\ \vdots \\ U_m \end{bmatrix}, \ \boldsymbol{Z}_{obj} = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}.$$

and

$$\boldsymbol{u}_{spl} = \begin{bmatrix} u_1 \\ \vdots \\ u_K \end{bmatrix}, \ \boldsymbol{Z}_{spl} = \begin{bmatrix} z_1(s_{11}) & \cdots & z_K(s_{11}) \\ \vdots & \ddots & \vdots \\ z_1(s_{1n_1}) & \cdots & z_K(s_{1n_1}) \\ \vdots & \ddots & \vdots \\ z_1(s_{m1}) & \cdots & z_K(s_{m1}) \\ \vdots & \ddots & \vdots \\ z_1(s_{mn_m}) & \cdots & z_K(s_{mn_m}) \end{bmatrix}.$$

Note that an identification number can be used in BUGS to increase the speed of Gibbs sampling. Unfortunately, `InferMachine()` does not support this approach. I will use BUGS code without indictors for `InferMachine()` inference. The BUGS model code file, `SBMDfemModel.txt`, for specifying (7.4) is

```
SBMDfemModel
{
   for(i in 1:numObs)
   {
      mu[i] <- beta0 + beta1*black[i] + beta2*hispanic[i]
                + beta3*white[i] + betaAge*age[i]
```

```
                + inprod(Uobj[],Zobj[i,])
                + inprod(uSpl[],Zspl[i,])
      spnbmd[i] ~ dnorm(mu[i],tauEps)
   }
   for (iSbj in 1:numSbj)
   {
      Uobj[iSbj] ~ dnorm(0,tauU)
   }
   for (iSpl in 1:numSpl)
   {
      uSpl[iSpl] ~ dnorm(0,tauu)
   }

   beta0  ~ dnorm(0,tauBeta)
   beta1  ~ dnorm(0,tauBeta)
   beta2  ~ dnorm(0,tauBeta)
   beta3  ~ dnorm(0,tauBeta)
   betaAge ~ dnorm(0,tauBeta)

   tauu ~ dgamma(Au,Bu);
   tauU ~ dgamma(AU,BU);
   tauEps ~ dgamma(Atau,Btau);
}
```

The R code that processes the `BUGS` file, including the specification of the data and hyperparameters, is

```
allData <- list(numObs=numObs,numSpl=numSpl,numSbj=numSbj,
```

```
                black=black,hispanic=hispanic,white=white,
                age=age,spnbmd=spnbmd,Zspl=Zspl,Zobj=Zobj,
                Atau=Atau,Btau=Btau,
                Au=Au,Bu=Bu,
                AU=Au,BU=BU,
                tauBeta=1/sigsqBeta)
```

and the MFVB inference imposes the product restriction:

$$q(\boldsymbol{\beta}, \boldsymbol{u}, \tau_\varepsilon, \tau_u, \tau_U) = q(\boldsymbol{\beta}, \boldsymbol{u})q(\tau_\varepsilon)q(\tau_u)q(\tau_U)$$

After setting the hyperparameters at $A_\varepsilon = 0.01$, $B_\varepsilon = 0.01$, $\sigma_\beta^2 = 10^8$, $A_u = B_u = 0.01$ and $A_U = B_U = 0.01$, fitting is then performed by using

```
fit <- InferMachine("SBMDfemModel.txt",allData,200)
```

Figure 7.5 shows fitted regression lines and pointwise 95% credible intervals produced by `InferMachine()` and the MCMC method. The estimated regression lines and credible intervals from the `InferMachine()` and MCMC fittings are highly similar. There exists a statistically significant difference in mean spinal bone mineral density between Asian and Black subjects. The approximate posterior density functions shown in Figure 7.6 can confirm this difference. It is also seen that no statistically significant difference is found between Asian and Hispanic subjects and between Asian and White subjects. The accuracy shown in Figure 7.6 also indicates that `InferMachine()` achieves high accuracy in the semiparametric mixed model.

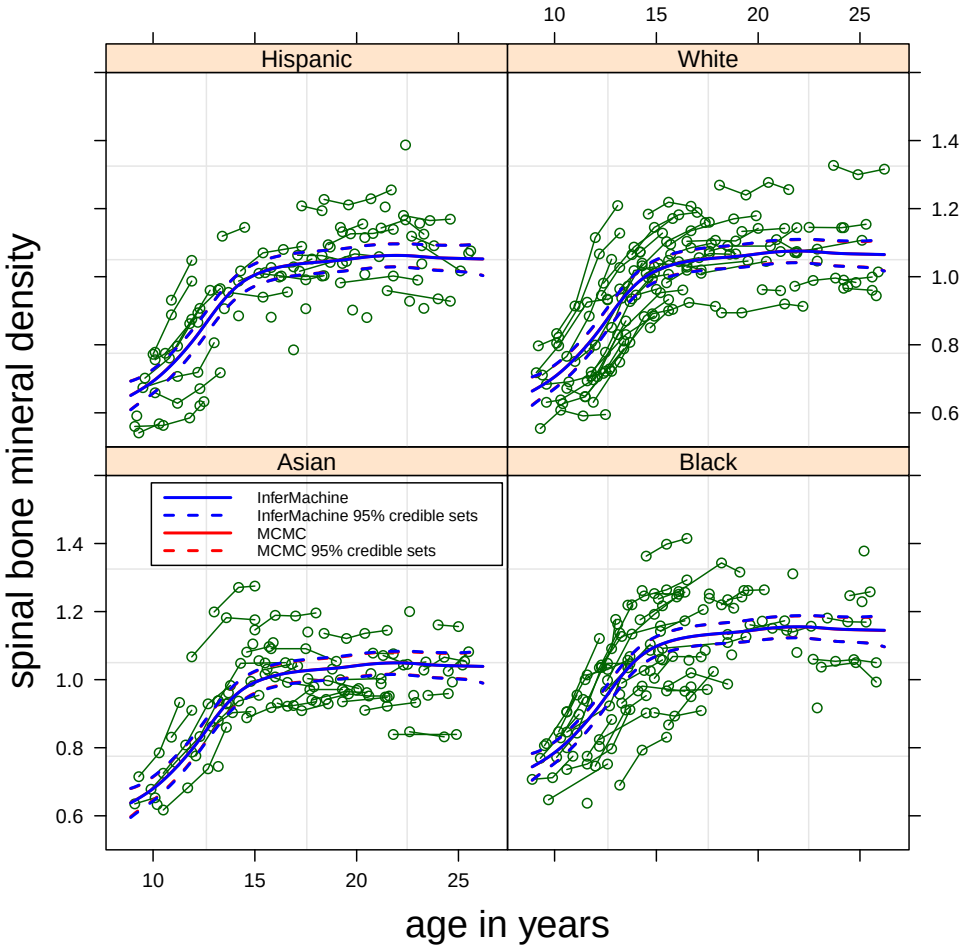

Figure 7.5: Fitted regression line and pointwise 95% credible intervals produced by `InferMachine()` and the MCMC method to spinal bone mineral density data.

7.4 Illustration for Binary Response Model

MFVB for the Gaussian response model can be extended to the binary response probit models easily by using the "trick" introduced by Albert and Chib (1993). By using the Result 1.15 to add auxiliary variables, we perform the MFVB inference for the Bayesian probit regression model with different mean structures. For this reason, I only give a one simple example to show How to use the `InferMachine()`

to handle a binary response model.

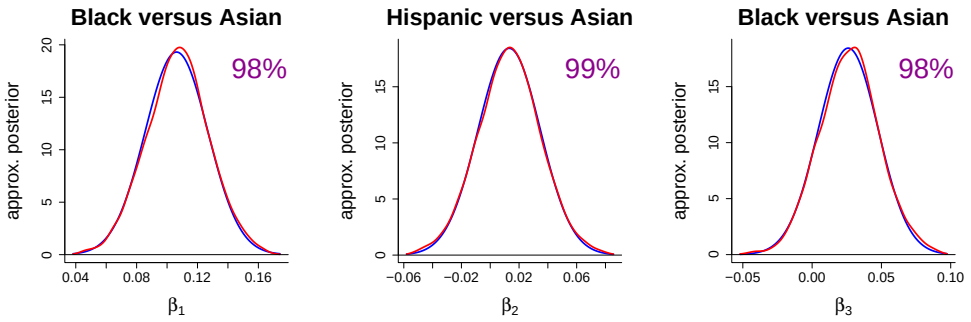

Figure 7.6: The approximate posterior density functions produced by `InferMachine()` (blue) and MCMC (orange) estimation of ethnic group parameters to fit a simple semiparametric mixed model to the spinal bone mineral density data set. The percentages are the accuracies of the `InferMachine()` fit compared with the MCMC fit.

how to build a probit model by using `InferMachine()`.

7.4.1 Probit regression with ridge penalized

Consider the probit regression with the ridge penalized model

$$\begin{aligned} y_i|\beta_0, \boldsymbol{\beta} &\overset{\text{ind.}}{\sim} \text{Bernoulli}\{\Phi((\beta_0 + \boldsymbol{X}\boldsymbol{\beta})_i)\} \\ \beta_0 &\sim N(0, \sigma^2_{\beta_0}) \\ \beta_j|\tau_\beta &\overset{\text{ind.}}{\sim} N(0, \tau_\beta^{-1}),\ 1 \le j \le p, \\ \tau_\beta &\sim \text{Gamma}(A_\beta, B_\beta), \end{aligned} \tag{7.5}$$

where $\sigma^2_{\beta_0}$, A_β, $B_\beta > 0$ are hyperparameters. The `BUGS` model code file, `RidgeBinaryModel.txt`, for specifying (7.5) is

```
RidgeBinaryModel
{
```

```
   for (i in 1:n)
   {
      mu[i] <- phi(beta0 + inprod(beta[],X[i,]))
      y[i] ~ dbern(mu[i])
   }
   for (j in 1:p)
   {
      beta[j] ~ dnorm(0,tauBeta)
   }
   beta0 ~ dnorm(0,tauBeta0)
   tauBeta ~ dgamma(Abeta,Bbeta)
}
```

The R code that processes the `BUGS` file, including the specifications of the data and hyperparameters, is

```
allData <- list(n=n,X=X,y=y,p=p,tauBeta=1/sigsqBeta0,
                    Abeta=Abeta,Bbeta=Bbeta)
```

where `sigsqBeta0`, `ABeta` and `Bbeta` correspond to the hyperparameters $\sigma^2_{\beta_0}$, A_β and B_β. The hyperparameters are set equal to $\sigma^2_{\beta_0} = 10^8$, $A_\beta = 0.01$ and $B_\beta = 0.01$. Fitting is then performed by using:

```
   fit <- InferMachine("RidgeBinaryModel.txt",allData,200)
```

Note that the current version only perform the `BUGS` code for the probit regression model, i.e. the link function should be $\Phi(.)$ (the corresponding `BUGS` code is `phi(.)`). Adding an auxiliary vector $\boldsymbol{a}$, the probit regression with the ridge

penalized model for MFVB inference is

$$\begin{aligned}
&p(y_i|a_i) = I(a_i \geq 0)^{y_i} I(a_i < 0)^{1-y_i}, \;\; 1 \leq i \leq n, \\
&\boldsymbol{a}|\beta_0, \boldsymbol{\beta} \overset{\text{ind.}}{\sim} N(\mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{I}), \\
&\beta_0 \sim N(0, \sigma^2_{\beta_0}), \\
&\beta_j|\tau_\beta \overset{\text{ind.}}{\sim} N(0, \tau_\beta^{-1}), \;\; 1 \leq j \leq p, \\
&\tau_\beta \sim \text{Gamma}(A_\beta, B_\beta).
\end{aligned} \tag{7.6}$$

For model (7.6), the MFVB inference imposes the product restriction:

$$q(\beta_0, \boldsymbol{\beta}, \tau_\beta, \boldsymbol{a}) = q(\beta_0, \boldsymbol{\beta}) q(\tau_\beta) q(\boldsymbol{a}).$$

Figure 7.7 shows the approximate posterior density functions produced by `InferMachine()` to fit a probit regression model to the ICU data set. The data correspond to a study on the survival of patients following admission to an adult intensive care unit (ICU). The binary response variable is **died** (0 = Lived, 1 = Died). A subset of predictors include: **age:** patient's age in years; **cancer:** is cancer part of the present problem? (0 = No, 1 = Yes); **SBP:** systolic blood pressure at ICU admission; **emergency:** type of admission (0 = Elective, 1 = Emergency); **hiPH:** pH from initial blood gases (**hiPH**=1 if pH is bigger than 7.25; otherwise, **hiPH**=0) and **hiPCO2:** PCO2 from initial blood gases (**hiPCO2**=1 if PCO2 is bigger than 45; otherwise, **hiPCO2**=0). Figure 7.7 suggests that the direct relationship between age and death rate is remarkable.

Compared with the Gaussian response case, the accuracy of the approximate posterior density of the probit regression model is considerably reduced. There exists a strong correlation between auxiliary variable $\boldsymbol{a}$ and the linear effect variable $\boldsymbol{\beta}$, so the factorized approximations result in reduced accuracy.

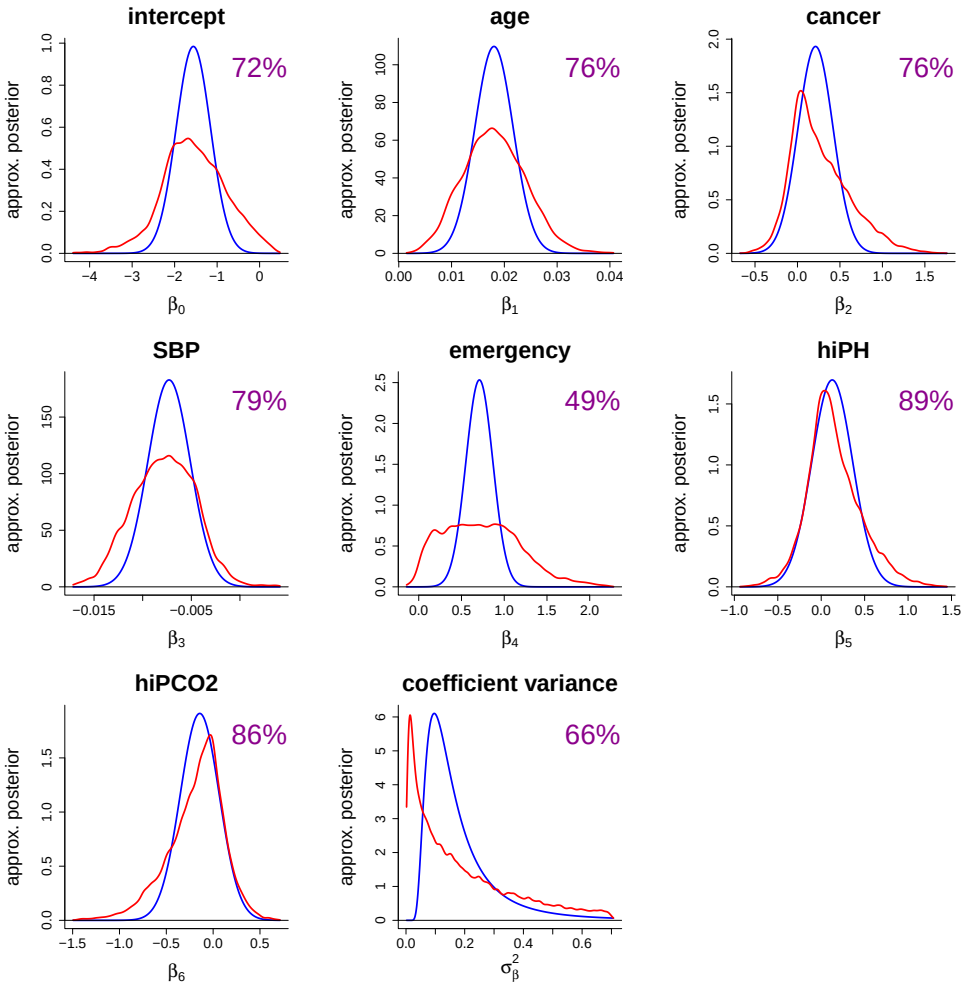

Figure 7.7: The approximate posterior density functions produced by `InferMachine()` (blue) and the MCMC method (orange) to fit a probit regression to the ICU data. The percentages are the accuracies of the `InferMachine()` fit compared with the MCMC fit.

7.5 Timing Comparison

Table 7.1 shows the relative computing times for the examples in this chapter by using `InferMachine()`, `Infer.NET` and the MCMC method. In this table, we report the average time elapsed (and standard error) of the computing times

over 100 runs with the number of `InferMachine()` and `Infer.NET` iterations set to 100 and the MCMC sample with 10000 burn-in and 10000 iterations. This was sufficient for convergence in these particular examples. Table 7.1 reveals that `InferMachine()` is considerably faster compared with `BUGS` and is also faster than `Infer.NET`.

regression model	time in seconds for `InferMachine`	time in seconds for `Infer.NET`	time in seconds for MCMC
simple linear	0.03 (0.01)	0.08 (0.01)	0.51 (0.03)
ridge penalized	0.06 (0.01)	0.11 (0.02)	583.13 (1.02)
nonpar. reg'n	0.08 (0.01)	1.12 (0.15)	58.71 (0.51)
semipar. mixed	10.08 (0.05)	581.23 (1.24)	1710.46 (1.58)
probit reg'n	0.11 (0.01)	0.15 (0.02)	99.52 (0.78)

Table 7.1: Average run times (standard errors) in seconds over 100 runs of the methods for each of the examples in Chapter 7.

7.6 Discussion

`InferMachine()` can perform MFVB inference for Gaussian and binary response model by using `BUGS` model code files. We obtain high accuracy for the Gaussian response model. The accuracy of the binary response model was influenced by the factorized restriction and is lower than the accuracy for the Gaussian response model.

There are positive results in the timing comparison for the semiparametric mixed model. The time taken by `Infer.NET` to build this model and run 100 iterations is about 10 minutes; however the `InferMachine()` only spends 10 seconds building the same model and running 100 iterations. Therefore, the gains in computational speed offered by `InferMachine()` make it a viable choice for larger models and/or sample sizes.

In the future, by adding more algorithms based on the MFVB, `InferMachine()` can be extended to include more responses and more prior distribution structures.

Chapter 8

Conclusion

The study explored variational approximation methods in semiparametric regression applications. In the Bayesian field, the study confirmed that the MFVB can perform Bayesian inference quickly and accurately. We also extended variational approximation methods into non-Bayesian fields and derived the asymptotic distributional behaviour of Gaussian variational approximate methods. We learned many things about the deriving MFVB algorithm. Asymptotic for variational inference was also visited.

In Chapter 2, MFVB methodology for Bayesian variable selection was established, based on the posterior probabilities of the model. For Gaussian response models, adding a ridge penalized technique for the fixed effect can improve the performance of variable selection. The simulation also demonstrated that stepwise MFVB variable selection performs similarly to MFVB variable selection with all possible subsets. However, for the binary response case, if the value of a predictor's coefficient is small, the accuracy of variable selection decreases.

In Chapter 3, MFVB methodology for linear variable selection was established using the indicator variable. We found that Gaussian-Zero models performed well in linear variable selection. For the binary response case, the value of a predictor's

coefficient also affects the performance of variable selection.

In Chapter 4, we provided MFVB methodology for the Bayesian Lasso regression model. We focused on choosing the Hyperparameter λ and considered two MFVB approaches to choose λ: variational EM for empirical Bayesian via marginal maximum likelihood, and a MFVB inference for the Bayesian LASSO model with a specific prior on λ. We also extended the Lasso regression model to a high-dimensional Lasso model and the ℓ_1 penalization smoothing model. Generally, we found that the MFVB estimation of Lasso was of high quality.

In Chapter 5, the new Bayesian inference software, `Infer.NET`, was demonstrated and critiqued in terms of its capacity for statistical analyses.

In Chapter 6, the application of variational approximation methods in the non-Bayesian field was studied. Hall, Ormerod and Wand (2011) introduced the Gaussian variational approximation (GVA) method to fit the generalized linear mixed models (GLMM). We derived the precise asymptotic distributional behavior of the Gaussian variational approximate estimators of the parameters in the simplest Poisson mixed model. The simulation study indicated that the coverage properties of Gaussian variational approximate confidence intervals are very good.

In Chapter 7, a new R function, `InferMachine()`, was provided to perform MFVB using the same model specification syntax that `BUGS` uses. This work is quite promising for the viability of MFVB, because the end-user can use the MFVB to fit the statistical model directly.

Overall, variational approximation methods, including MFVB and GVA, provide a fast, deterministic alternative for avoiding intractable integration in statistical inference. There was a trade-off between accuracy and computational cost. Even the accuracy of MFVB varies according to the various regression models and applications, though they still allow the possibility of using Bayesian methods to deal with big data. The GVA provides a new alternative for non-Bayesian

inference. I am the first to prove asymptotically valid inference for a variational approximation method. This can be used to properly understand variational approximation methods and to conduct hypothesis testing..

Bibliography

Albert, J. H., and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association, 88* (422), 669-679.

Armagan, A. (2009). Variational bridge regression. In *International Conference on Artificial Intelligence and Statistics,* 17-24.

Bachrach, L. K., Hastie, T., Wang, M. C., Narasimhan, B., and Marcus, R. (1999). Bone mineral acquisition in healthy Asian, Hispanic, Black, and Caucasian youth: a longitudinal study 1. *The Journal of Clinical Endocrinology & Metabolism, 84* 12, 4702-4712.

Balakrishnan, S., and Madigan, D. (2010). Priors on the variance in sparse Bayesian learning: the demi-Bayesian lasso. In Chen, M. H., Müller, P., Sun, D., Ye, K., and Dey, D. (Eds.), *Frontiers of Statistical Decision Making and Bayesian Analysis: In Honor of James O. Berger* (1st ed., pp. 346-359). New York: Springer.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67* 1, 1-48.

Bishop, C. M. (2006). *Pattern recognition and machine learning.* New York:

Springer.

Bishop, C. M., and Winn, J. (2003). Structured variational distributions in VIBES. In *Proceedings Artificial Intelligence and Statistics* (pp. 3-6). Key West, FL: Society for Artificial Intelligence and Statistics.

Bishop, Y. M., Fienberg, S. E., and Holland, P. W. (2007). *Discrete multivariate analysis: theory and practice*. New York: Springer.

Boyd, S., and Vandenberghe, L. (2004). *Convex optimization.* Cambridge: Cambridge University Press.

Breslow, N. E. (1984). Extra-Poisson variation in log-linear models. *Journal of the Royal Statistical Society, Series C (Applied Statistics), 33* (1), 38-44.

Breslow, N. E., and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association, 88* (421), 9-25.

Broman, K. W., and Speed, T. P. (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society, Series B (Statistical Methodology), 64* (4), 641-656.

Brown, P. J., Vannucci, M., and Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society, Series B (Statistical Methodology), 60* (3), 627-641.

Casella, G. (2001). Empirical Bayes Gibbs sampling. *Biostatistics, 2* (4), 485-500.

Celeux, G., Forbes, F., Robert, C. P., and Titterington, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis, 1* (4), 651-

673.

Chambers, J. M., and Hastie, T. J. (1991). *Statistical models in S*. London: Chapman & Hall.

Chipman, H., Hamada, M., and Wu, C. F. J. (1997). A Bayesian variable selection approach for analyzing designed experiments with complex aliasing. *Technometrics, 39* (4), 372-381.

Clyde, M., and George, E. I. (2004). Model uncertainty. *Statistical Science, 19* (1), 81-94.

Cottet, R., Kohn, R. J., and Nott, D. J. (2008). Variable selection and model averaging in semiparametric overdispersed generalized linear models. *Journal of the American Statistical Association, 103* (482), 661-671.

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2002). On Bayesian model and variable selection using MCMC. *Statistics and Computing, 12* (1), 27-36.

Dellaportas, P., and Stephens, D. A. (1995). Bayesian analysis of errors in variables regression models. *Biometrics, 51* (3), 1085-1095.

Drummond, A. J., and Rambaut, A. (2007). Beast: Bayesian evolutionary anal- ysis by sampling trees. *BMC Evolutionary Biology, 7*, 214. URL *http://www.biomedcentral.com/1471-2148/7/214*

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics, 32* (2), 407-499.

Faes, C., Ormerod, J. T., and Wand, M. P. (2011). Variational Bayesian inference for parametric and nonparametric regression with missing data. *Journal of*

the American Statistical Association, 106 (495), 959-971.

Fernandez, C., and Steel, M. F. (2000). Bayesian regression analysis with scale mixtures of normals. *Econometric Theory, 16* (1), 80-101.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis, 1* (3), 515-534.

George, E. I. (2000). The variable selection problem. *Journal of the American Statistical Association, 95* (452), 1304-1308.

George, E. I., and McCulloch, R. E. (1996). Stochastic search variable selection. In Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (Eds.), *Markov Chain Monte Carlo in Practice* (pp. 203-214). London: Chapman & Hall.

Hall, P., Ormerod, J. T., and Wand, M. P. (2011). Theory of Gaussian variational approximation for a Poisson mixed model. *Statistica Sinica, 21* (1), 369-389.

Hall, P., Pham, T., Wand, M. P., and Wang, S. S. J. (2011). Asymptotic normality and valid inference for gaussian variational approximation. *The Annals of Statistics, 39* (5), 2502-2532.

Hans, C. (2009). Bayesian lasso regression. *Biometrika, 96* (4), 4, 835-845.

Hans, C. (2010). Model uncertainty and variable selection in Bayesian lasso regression. *Statistics and Computing, 20* (2), 221-229.

Hastie, T. (2015). *gam: Generalized Additive Models*. R package version 1.12. URL http://CRAN.R-project.org/package=gam

Hastie, T., and Efron, B. (2013). *lars: Least Angle Regression, Lasso and Forward Stagewise*. R package version 1.2.
URL http://CRAN.R-project.org/package=lars

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2nd ed.). New York: Springer.

Hoerl, A. E., and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12* (1), 55-67.

Hu, J., and Johnson, V. E. (2009). Bayesian model selection using test statistics. *Journal of the Royal Statistical Society, Series B (Statistical Methodology), 71* (1), 143-158.

Jaakkola, T. S., and Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing, 10* (1), 25-37.

Johnstone, I. M., and Silverman, B. W. (2005). Empirical Bayes selection of wavelet thresholds. *The Annals of Statistics, 33* (4), 1700-1752.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning, 37* (2), 183-233.

Kass, R. E., and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90* (430), 773-795.

Kauermann, G., Ormerod, J. T., and Wand, M. P. (2010). Parsimonious classification via generalized linear mixed models. *Journal of Classification, 27* (1), 89-110.

Kim, A.S.I. and Wand, M.P. (2016). The Explicit Form of Expectation Propagation for a Simple Statistical Model. *Electronic Journal of Statistics, 10*, 550-581.

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

Leisch, F., and Dimitriadou, E. (2009). *`mlbench`: A Collection of Artificial and Real-World Machine Learning Problems*. R package version 2.1.1.
URL http://CRAN.R-project.org/package=mlbench

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association, 103* (481), 410-423.

Liddle, A. R. (2007). Information criteria for astrophysical model selection. *Monthly Notices of the Royal Astronomical Society: Letters, 377* (1), 74-78.

Lokhorst, J., Venables, B., and Turlach, B. (2014). *`lasso2`: L_1 constrained estimation aka 'lasso'*. R package version 1.2-19.
URL http://CRAN.R-project.org/package=lasso2

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). `WinBUGS`: a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Somputing, 10* (4), 325-337.

Luts, J., Wang, S. S. J., Ormerod, J., and Wand, M. P. (2016). Semiparametric regression analysis via `Infer.NET`. *Journal of Statistical Software.* Accepted.

Marley, J. K., and Wand, M. P. (2010). Non-standard semiparametric regression via `BRugs`. *Journal of Statistical Software, 37* (5), 1-30.

McGrory, C. A., and Titterington, D. M. (2007). Variational approximations in Bayesian model selection for finite mixture distributions. *Computational Statistics & Data Analysis, 51* (11), 5352-5367.

McGrory, C. A., and Titterington, D. M. (2009). Variational Bayesian analysis for hidden Markov models. *Australian & New Zealand Journal of Statistics, 51* (2), 227-244.

McLachlan, G. J., and Peel, D. (2004). *Finite mixture models.* Hoboken, NJ: John Wiley & Sons.

Minka, T., Winn, J., Guiver, J., Webster, S., Zaykov, Y., Yangel, B., Spengler, A., and Bronskill, J. (2014). `Infer.NET 2.6`, Microsoft Research Cambridge. URL http://research.microsoft.com/infernet.

Minka, T. P. (2001). Expectation propagation for approximate Bayesian inference. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence* (pp. 362-369). San Francisco: Morgan Kaufmann.

Ngo, L., and Wand, M. P. (2004). Smoothing with mixed model software. *Journal of Statistical Software, 9* (1), 1-54.

O'Hara, R. B., and Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian analysis, 4* (1), 85-117.

Ormerod, J. T. (2011). Grid based variational approximations. *Computational Statistics & Data Analysis, 55* (1), 45-56.

Ormerod, J. T., and Wand, M. P. (2010). Explaining variational approximations. *The American Statistician, 64* (2), 140-153.

Ormerod, J. T., and Wand, M. P. (2012). Gaussian variational approximate inference for generalized linear mixed models. *Journal of Computational and Graphical Statistics, 21* (1), 2-17.

Pagano, M., and Gauvreau, K. (2000). *Principles of biostatistics* (2nd ed.). Pacific Grove, CA: Duxbury.

Parisi, G. (1988). *Statistical field theory.* Boston, MA: Addison-Wesley.

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

Pham, T. H., Ormerod, J. T., and Wand, M. P. (2013). Mean field variational Bayesian inference for nonparametric regression with measurement error. *Computational Statistics & Data Analysis, 68*, 375-387.

Pinheiro, J. C., and Bates, D. M. (2006). *Mixed-effects models in S and S-PLUS.* Berlin: Springer Science & Business Media.

Pinheiro, J. C., Bates, D. M., DebRoy, S., Sarkar, D., and R Core Team, T. (2015). *`nlme`: Linear and Nonlinear Mixed Effects Models.* R package version 3.1-120.
URL http://CRAN.R-project.org/package=nlme

Polson, N. G., and Scott, J. G. (2012). On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis, 7* (4), 887-902.

R Core Team. (2015). *`R`: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.
URL http://www.R-project.org/.

Robert, C., and Casella, G. (2004). *Monte Carlo Statistical Methods.* New York: Springer.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression.* Cambridge: Cambridge University Press.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2009). Semiparametric regression during 2003-2007. *Electronic Journal of Statistics 3*, 1193-1256.

Saul, L., and Jordan, M. I. (1996). Exploiting tractable substructures in intractable networks. In *Advances in Neural Information Processing Systems 8* (pp. 486-492). Cambridge, MA: MIT Press.

Shafer, G. (1982). Lindley's paradox. *Journal of the American Statistical Association, 77* (378), 325-334.

Smith, P. J. (1998). *Into statistics: a guide to understanding statistical concepts in engineering and the sciences.* New York: Springer.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B (Statistical Methodology), 64* (4), 583-639.

Stamey, T. A., Kabalin, J. N., McNeal, J. E., Johnstone, I. M., Freiha, F., Redwine, E. A., and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. radical prostatectomy treated patients. *The Journal of Urology, 141* (5), 1076- 1083.

Staudenmayer, J., Lake, E. E., and Wand, M. P. (2009). Robustness for general design mixed models using the t-distribution. *Statistical Modelling, 9* (3), 235-255.

Theo, H. E., and Mike, E. G. (2004). Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. *Genetics Selection Evolution, 36* (3), 261-279.

Thomas, A., O'Hara, B., Ligges, U., and Sturtz, S. (2006). Making `BUGS` open. *R news, 6* (1), 12-17.

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

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *The journal of Machine Learning Research, 1*, 211-244.

Venables, W. N., and Ripley, B. D. (2002). *Modern Applied Statistics with S* (4th ed.). New York: Springer.

Wand, M. P. (2009). Semiparametric regression and graphical models. *Australian & New Zealand Journal of Statistics, 51* (1), 9-41.

Wand, M. P. (2014). Fully simplified Multivariate normal updates in non-conjugate variational message passing. *The Journal of Machine Learning Research, 15* (1), 1351-1369.

Wand, M. P. (2015). *`KernSmooth`: Functions for Kernel Smoothing Supporting Wand & Jones (1995).* R package version 2.23-14.
URL http://CRAN.R-project.org/package=KernSmooth

Wand, M. P., and Ormerod, J. T. (2008). On semiparametric regression with OSullivan penalized splines. *Australian & New Zealand Journal of Statistics, 50* (2), 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., Ormerod, J. T., Padoan, S. A., and Fuhrwirth, R. (2011). Mean field variational Bayes for elaborate distributions. *Bayesian Analysis, 6* (4), 847-900.

Wang, S. S. J., and Wand, M. P. (2011). Using `Infer.NET` for statistical analyses. *The American Statistician, 65* (2), 115-126.

West M. (2003). Bayesian factor egression models in the "large p, small n" paradigm. In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M. (Eds.), *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting* (pp. 733-742). Oxford: Oxford University Press.

Williams, D. A. (1982). Extra-binomial variation in logistic linear models. *Applied Statistics, 31* (2), 144-148.

Winn, J., and Bishop, C. M. (2005). Variational message passing. *Journal of Machine Learning Research, 6*, 661-694.

Wu, B., McGrory, C. A., and Pettitt, A. N. (2012). The variational Bayesian approach to fitting mixture models to circular wave direction data. *Journal of Applied Meteorology and Climatology, 51* (10), 1750-1762.

Yi, N. (2004). A unified Markov chain Monte Carlo framework for mapping multiple quantitative trait loci. *Genetics, 167* (2), 967-975.

Yi, N., George, V., and Allison, D. B. (2003). Stochastic search variable selection for identifying multiple quantitative trait loci. *Genetics, 164* (3), 1129-1138.

Yuan, M., and Lin, Y. (2005). Efficient empirical Bayes variable selection and estimation in linear models. *Journal of the American Statistical Association, 100* (472), 1215-1225.

Zeger, S. L., and Karim, M. R. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American statistical association, 86* (413), 79-86.

Zeger, S. L., Liang, K. Y., and Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics, 44* (4), 1049-1060.

Zhao, P., and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research, 7*, 2541-2563.

Zhao, P., and Yu, B. (2007). Stagewise Lasso. *Journal of Machine Learning Research*, *8*, 2701-2726.