Fast Approximate Inference for Longitudinal and Multilevel Data Analysis

Cathy Yuen Yi Lee

GStat MBiostat BMathAdv(Hons)

A PhD thesis submitted in fulfilment of the requirements for
Doctor of Philosophy in Mathematics
in the School of Mathematical and Physical Sciences,
University of Technology Sydney

2016

Certificate of Original Authorship

I, Cathy Yuen Yi Lee, certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of student:

Date:

Acknowledgements

Fear protects us in many ways, but what has served me is diving into my own obliviousness. Being more confident than I should be trying things that you never would have tried. Your inexperience is an asset in that it will make you think in original, unconventional ways. Accept your lack of knowledge and use it as your asset. Natalie Portman

I never thought I would go to university, let alone complete a PhD degree. I am always and forever grateful to my lovely parents for instilling the value of education in me and for making countless sacrifice to always provide for their children. I wouldn't be who I am today if it wasn't for the undying love and support my parents have unconditionally given to me.

Words are simply not enough to express my sincere gratitude to my academic parents, Matt and Louise, for believing in me and seeing my good qualities way before I ever could. Through watching them, I learned dedication and perseverance to succeed. I am no longer afraid of following my dream I had set since I was 16. Thank you for guiding me to reach a stage where I can be proud of who I am and what I have achieved for the past three years.

I must express my gratitude to Jason, my companion, for his enthusiasm and continuous support and encouragement. I was amazed by his willingness and determination to support me in every possible way throughout this journey. He continually challenged me to dig deeper and learn more, and I greatly appreciate everything he has done to make this possible.

"Friends are the bold essence of life". Completing this work would have been more difficult were it not for the support and friendship provided by my colleagues at the School of Mathematical and Physical Sciences. A special thanks to Chris, who has proofread countless pages of my work. Also a special thanks to Marianne, who has always been there for me and experienced all of the laugher and tears from my special journey.

I finish here with Hon's favourite quote "Stay Hungry, Stay Foolish". I believe this quote will continue to motivate me in life for many years to come.

Publications

Lee, C. Y . Y. and Wand, M. P. (2016). Streamlined mean field variational Bayes for longitudinal and multilevel data analysis. *Biometrical Journal*. DOI: 10.1002/bimj.201500007

Bentley, J. and Lee, C. Y. Y. (2016). Bayesian Inference for Gaussian Semiparametric Multilevel Models. *SAS Global Forum*, April 18-21, Las Vegas, United States. `http://support.sas.com/resources/papers/proceedings16/7200-2016.pdf`

Lee, C. Y. Y. and Wand, M. P. (2015). Variational methods for fitting complex Bayesian mixed effects models to health data. *Statistics in Medicine*. DOI: 10.1002/sim.6737.

Lee, C. Y. Y., Homer, C., Bisits, A. and Ryan, L. (2015). Increasing variation in hospital caesarean section rates among low-risk nulliparous women in Australia, from 1994 to 2010. *Birth in revision*.

Chen, J., Lee, C. Y. Y., Coull, B., Valeri, L., Gruskin, S., Beckfield, J., Singh, N. and Krieger, N. (2015). 50-year trends and state variation in socioeconomic and racial/ethnic inequalities in US infant death rates, 1960 to 2010. *Under preparation*.

The last listed publication related to my 3-week academic visit at the Harvard School of Public Health during March-April 2014. Throughout my visit, I collaborated with a research team of world-renowned social epidemiologists, biostatisticians and research scientists on a study under an overarching project entitled *The Public Health Disparities Geocoding Project Monograph*. For more information, please refer to https://www.hsph.harvard.edu/thegeocodingproject/.

Presentations

Lee, C. Y. Y and Bentley, J. P. Fast variational Bayesian inference for Gaussian semi-parametric multilevel models in SAS/IML®, Las Vegas, United States, 2016. (Accepted for a 20-min workshop presentation)

Lee, C. Y.Y. Fast approximate inference for longitudinal and multilevel data analysis. *The 16th J.B. Douglas Awards, Statistical Society of Australia*, Sydney, Australia, 2015. (Awarded Equal First Price for oral presentation)

Lee, C. Y. Y., Wand, M. P. and Ryan, L. Fast approximate inference for longitudinal and multilevel data analysis. *The 16th J.B. Douglas Awards, Statistical Society of Australia*, Sydney, Australia, 2015. (Awarded Equal First Price for oral presentation)

Lee, C. Y. Y., Chen, J., Coull, B., Valeri, L., Gruskin, S., Beckfield, J., Singh, N. and Krieger, N. 50-year trends and state variation in socioeconomic and racial/ethnic inequalities in US infant death rates, 1960 to 2010. *Joint Statistical Meeting*, Seattle, United States, 2015.

Lee, C. Y. Y. and Wand, M. P. Fast approximate variational inference for Bayesian multilevel models. *Young Statisticians Conference*, Adelaide, Australia, 2015. (Awarded 3rd Price for oral presentation)

Ryan, L. and Lee, C. Y. Y. Spatio-temporal variation in cesarean rates in New South Wales, Australia, from 1994 to 2010. *Joint Statistical Meeting*, Boston, United States, 2014.

Lee, C. Y. Y. and Wand, M. P. Fast approximate variational inference for Bayesian multilevel models. *Australian Statistical Conference in conjunction with the Institute of Mathematical Statistics Annual Meeting*, Sydney, Australia, 2014.

The fourth listed presentation related to my 3-week academic visit at the Harvard School of Public Health during March-April 2014. Throughout my visit, I collaborated with a research team of world-renowned social epidemiologists, biostatisticians and research scientists on a study under an overarching project entitled *The Public Health Disparities Geocoding Project Monograph.* For more information, please refer to https://www.hsph.harvard.edu/thegeocodingproject/.

Lee, C. Y. Y. and Wand, M. P. Fast approximate variational inference for Bayesian longitudinal and multilevel semiparametric regression models. *University of Queensland*, Brisbane, Australia, 2014.

Contents

List of Figures

List of Tables

Abstract

Generalised linear mixed models are the cornerstone of longitudinal and multilevel data analysis. However, exact inference for Bayesian mixed models with semiparametric extensions is typically intractable, requiring approximate inference methods for use in practice. Markov chain Monte Carlo or MCMC is one of the most commonly used approximate inference methods in this setting, but can be computationally intensive and often suffers from poor convergence in complex models. A faster, deterministic alternative to MCMC is variational approximations, a class of deterministic algorithms that is based on reformulating the problem of computing the posterior distribution as an optimisation problem, simplifying that problem and finding solutions to the perturbed problem. In this thesis, we work with a particular class of variational approximations, known as the mean field variational Bayes (MFVB). In essence, MFVB approximations are based upon optimising the Kullback-Leibler divergence with respect to the so-called approximating distribution. We derive MFVB algorithms for a wide variety of Bayesian semiparametric mixed models with Gaussian, Student-t, Bernoulli and Poisson responses. In order to overcome the computational cost of the direct naïve approach to the underlying MFVB calculations for models, we introduce a novel, streamlined approach that involves matrix permutation and block decomposition. Through a series of numerical studies, we demonstrate that the MFVB algorithms achieve a good level of accuracy compared to a MCMC benchmark (our gold standard). Furthermore, our developed streamlined algorithms are shown to have a complexity that is linear in the number of groups at each level, representing a two orders of magnitude improvement over the naïve approach. More importantly, the modularity of MFVB allows relatively simple extensions to more complicated scenarios, including higher-level random effects, measurement error and/or missing data problems, models with group-specific curves and real-time or online data processing. Illustrations from various real data examples are provided.

Chapter 1

Introduction

The world is being re-shaped by the convergence of social, mobile, cloud, big data, community and other powerful forces. The combination of these technologies unlocks an incredible opportunity to connect everything together in a new way and is dramatically transforming the way we live and work.

Marc Benioff

1.1 Aim and motivation

In today's digital age, we are surrounded by a vast amount of data, so called "Big Data". In a recent talk at Google, Kenneth Cukier, Data Editor of *The Economist* stated that Big Data is a revolution that will transform how we live, work and think about the economy, health and society in the years to come. "The amount of data in the world is increasing exponentially, and we don't necessarily have the tools to handle it.", Kenneth says. So the question is, how can we take this opportunity and turn this new wealth of information into big insights? The answer is *fast data analysis*. The key role of data analysis is to translate the raw, intimidating numbers into plain english, scientific findings. "The doubling of computing power every 18 months (Moore's Law) is nothing compared to a big algorithm - a set of rules that can be used to solve a problem a thousand times faster than conventional computational methods could", says Albert J. Weatherhead III University Professor Gary King at Harvard University. Inspired by these people, I am motivated to take on the Big Data challenges that show a presence within this thesis.

Most datasets nowadays in science have some form of grouped or hierarchical structure. Such structure may arise from repeated measurements of individuals over time, students grouped within classrooms and schools, or friends grouped within networks (much like LinkedIn or Facebook). The reality of scientific research is that our fundamental units of interest are usually intertwined, and quite often, of primary interest from a theoretical

perspective. Insightful information can be gleaned from connecting gene expression data, linked administrative health records or environmental sequences, but how can we best extract signals from these correlated data? From a methodological perspective, the complex structure of these data complicates the conventional statistical methods and threatens our ability to fitting models and performing statistical inference efficiently.

Bayesian generalised linear mixed models (e.g. Gelman and Hill, 2007) continue to grow in popularity in response to the demands of large grouped/clustered datasets and have proved powerful in many real-world applications. These models are based on the deceptively simple Bayes's Theorem. However, the computations often involve intractable multivariate integrals in likelihood and posterior density expressions, requiring approximate inference methods for use in practice. *Markov chain Monte Carlo* or MCMC is one of the most prominent methods for estimating Bayesian statistical models and has proven useful in a wide array of problems. Unfortunately, in many situations a MCMC type of sampling strategy is undesirably slow and suffers from poor mixing. In parallel to these challenges in statistics, the computer science community has been developing approximate solutions to inferential problems using variational algorithms (e.g. Bishop, 2006; Ormerod and Wand, 2010). These are deterministic algorithms that facilitate analytical calculations for the posterior distributions, an iterative approach resembling an Expectation-Maximisation method. In contrast to these algorithms, MCMC methods yield their results in the form of a set of samples from the posteriors. Practical implementations of variational algorithms make use of factorised approximating posteriors and priors that belong to the conjugate-exponential family, making the required integrals tractable. The resultant variational approximations therefore sacrifice some accuracy of MCMC, albeit minimal, but offer vast improvements in terms of speed and memory efficiency.

What is currently lacking in the statistical literature is research on the accuracy of the variational approximations for Bayesian generalised linear mixed models with various extensions. All of these bring the overarching question of my thesis:

> *Are variational approximations a viable set of tools for Bayesian longitudinal and multilevel data analysis?*

To address this we: first, identify models for which standard inference is impaired by time constraints; second, develop variational algorithms; and third, empirically assess the performance of variational algorithms in terms of inferential accuracy and computational speed. In what follows, we commence our discussion by presenting the literature review on approximate Bayesian inference.

1.2 Structure of the thesis

Each chapter in this thesis is dedicated to the development of variational algorithms catered to a particular longitudinal or multilevel model setting. In Chapter 2, we consider the generalised linear mixed models for longitudinal and multilevel data analysis. Four common response distributions, including Gaussian, Student-t, Bernoulli and Poisson, are explored. We provide a step-by-step, notationally friendly guide on the development of fast variational algorithms for fitting and inference in mixed models, where all algorithmic updates are available in closed form. The advantages of variational methods as compared to the gold standard MCMC methods in terms of speed, coverage and accuracy are illustrated via a series of comprehensive numerical studies. A computational challenge in the directly-implemented variational algorithms arises from the update expression of the effects covariance matrix, which involves inversion and storage of a large sparse matrix, a cost that grows cubically in the number of groups. The centrepieces for Chapter 3 are our developed novel, streamlined variational algorithms for fast and memory efficient fitting and inference in large Bayesian generalised linear mixed models with semiparametric extensions - currently represent the state-of-the-art algorithms in this area. The number of operations is linear in the number of groups at each level, which constitutes a major improvement over the naïve/direct approach. Storage requirements are also lessened considerably.

The beauty of the Bayesian paradigm combined with variational algorithms is its tremendous flexibility. In Chapter 4, we extend our variational algorithms to cater for three interesting and challenging scenarios, including data subject to measurement error problems, data subject to missing data problems, and data that are processed in real-time as they arrive. Accuracy and time comparisons are considered as an effective assessment of the developed algorithms. Chapter 5 is motivated by a study that examines trends in caesarean section rates in the largest state of Australia, New South Wales, between 1994 and 2010. We propose a group-specific curve model that encapsulates the complex non-linear features of the overall and hospital-specific trends in caesarean section rates, while taking into account hospital variability over time. We use penalised spline based smooth functions that represent trends and implement a fully variational Bayes approach to model fitting. In Chapter 6, we present a discussion on a variant of the standard variational Bayes approach that is based on natural parameters and sufficient statistics and allows an easier extension to arbitrarily large models via factor graph fragments. We demonstrate this approach for the two models considered in Chapters 2 and 3.

The majority of the materials written in this thesis have been published (one is under review at the time of writing) and presented at five conferences. Main results in Chapters 2, 3 and 5 have been published in Lee and Wand (2015a), Lee *et al.* (2015) and Lee and Wand (2015b) respectively.

1.3 Semiparametric regression

Parametric regression methods for longitudinal and multilevel data analysis have been well developed in the past two decades. Such methods can be broadly classified into estimating equations based methods (e.g. Breslow and Clayton, 1993; Zeger and Liang, 1986) and mixed models (e.g. Gelman and Hill, 2007; Goldstein, 2011; Hastie and Tibshirani, 1990). Parametric regression assumes a functional form (typically linear) in the relationship between the mean of a response and predictors. Although such parametric assumption offers simplicity, it is inappropriate for situations when the relationship between the mean response and predictors is unknown.

Semiparametric regression is a seamless fusion between parametric and nonparametric regression analysis (Ruppert *et al.*, 2009). It is a prominent field synthesising research across several branches of statistics including parametric and non-parametric regression, longitudinal and multilevel data analysis, and Bayesian hierarchical models. Early comprehensive reviews on semiparametric regression analysis are Ruppert *et al.* (2003) and Ruppert *et al.* (2009). Semiparametric regression models extends classical generalised linear mixed models to accomodate non-linear predictor effects flexibly using penalised basis functions such as B-splines and Daubechies wavelets. Such penalisation can be achieved through fitting random effects that have the same formulation as those used traditionally in longitudinal and multilevel data analysis. Throughout this thesis, we work with a direct generalisation of smoothing splines, namely the *O'Sullivan splines* (O'Sullivan, 1986, Section 3). O'Sullivan splines are a class of penalised splines based on B-spline basis functions. They possess the attractive feature of requiring considerably fewer basis functions and their smoothness, numerical stability and natural boundary conditions make them the most widely used class of penalised splines in standard statistical software, for example, the `smooth.spline()` function in R (R Development Core Team, 2015) and the `MIXED` procedure in SAS (SAS Institute Inc., 2013). Wand and Ormerod (2008) give a detailed description of O'Sullivan penalised splines and their mixed model representation. A brief sketch of their description is provided here for reference.

Consider the simplest non-parametric regression setting

$$y_i = f(x_i) + \varepsilon_i, \quad 1 \leq i \leq n,$$

where $(x_i, y_i) \in \mathbb{R}\times\mathbb{R}$. Suppose we are interested in estimating f over the interval $[a, b]$ containing the x_is via a set of cubic B-spline basis functions $\boldsymbol{B}_x \equiv [B_1(x), \ldots B_{K+4}(x)]$ for $K \leq n$. The corresponding knot sequence is defined by $a = \kappa_1 = \kappa_2 = \kappa_3 = \kappa_4 < \kappa_5 < \cdots \kappa_{K+4} < \kappa_{K+5} = \kappa_{K+6} = \kappa_{K+7} = \kappa_{K+8} = b$ (Hastie *et al.*, 2005). The coefficients may be estimated in a number of ways. The simplest is the *penalised residual sum of squares*

or PRSS, which involves choosing the coefficients to minimise

$$PRSS(f,\lambda) = \sum_{i=1}^{n}\{y_i - f(x_i)\}^2 + \lambda\int_a^b f''(x)\,dx. \tag{1.1}$$

The expression $\lambda\int_a^b f''(x)dx$ is the so-called *penalty term* because it penalises fits that are too rough, thus yielding a smoother result. The amount of smoothing is controlled by $\lambda > 0$, and is usually referred to as *a smoothing parameter*. The solution to (1.1) is the O'Sullivan penalised spline $f(\boldsymbol{x}) = \boldsymbol{B\nu}$ and thus (1.1) can be rewritten as

$$PRSS(\boldsymbol{\nu},\lambda) = (\boldsymbol{y} - \boldsymbol{B\nu})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{B\nu}) + \lambda\boldsymbol{\nu}^{\mathsf{T}}\boldsymbol{\Omega\nu}, \tag{1.2}$$

where $B_{ik} = B_k(x_i)$ and $\Omega_{kk'} = \int_a^b B_k''(x)B_{k'}''(x)dx$. Straightforward algebraic manipulation leads to the following fitted O'Sullivan penalised spline with a solution to (1.2)

$$\hat{f}(\boldsymbol{x}) = \boldsymbol{B}\hat{\boldsymbol{\nu}}, \quad \hat{\boldsymbol{\nu}} = (\boldsymbol{B}^{\mathsf{T}}\boldsymbol{B} + \lambda\boldsymbol{\Omega})^{-1}\boldsymbol{B}^{\mathsf{T}}\boldsymbol{y}. \tag{1.3}$$

Computation of the design matrix $\boldsymbol{B}$ is straightforward and is readily available in the R environment. However, computation of the penalty matrix $\boldsymbol{\Omega}$ requires some effort. In Section 6 of Wand and Ormerod (2008), an exact matrix algebraic expression for $\boldsymbol{\Omega}$ is derived by applying the Simpson's rule over each of the inter-knot differences, given as

$$\boldsymbol{\Omega} = (\widetilde{\boldsymbol{B}}'')^{\mathsf{T}}\mathrm{diag}(\boldsymbol{w})\widetilde{\boldsymbol{B}}'',$$

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

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

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

$$\begin{aligned}\boldsymbol{w} \;=\; &\left\{\tfrac{1}{6}(\triangle\boldsymbol{\kappa})_1, \tfrac{4}{6}(\triangle\boldsymbol{\kappa})_1, \tfrac{1}{6}(\triangle\boldsymbol{\kappa})_1, \tfrac{1}{6}(\triangle\boldsymbol{\kappa})_2, \tfrac{4}{6}(\triangle\boldsymbol{\kappa})_2,\right.\\ &\left.\tfrac{1}{6}(\triangle\boldsymbol{\kappa})_2, \ldots, \tfrac{1}{6}(\triangle\boldsymbol{\kappa})_{K+7}, \tfrac{4}{6}(\triangle\boldsymbol{\kappa})_{K+7}, \tfrac{1}{6}(\triangle\boldsymbol{\kappa})_{K+7}\right\},\end{aligned}$$

where $(\triangle\boldsymbol{\kappa})_k \equiv \kappa_{K+1} - \kappa_K$, $1 \le k \le K+7$. A common default choice for the number of knots is $K = \min(n_U/4, 35)$, where n_U is the number of unique x_is, and the distribution of knots can either be quantile-based or equally spaced (e.g. Ruppert *et al.*, 2003). In the next section we show how the O'Sullian penalised splines can be expressed in the mixed model and Bayesian hierarchical model framework.

1.3.1 Mixed model representation

Ruppert *et al.* (2003) presents the general framework of semiparametric regression using the inferential equivalence between penalised likelihood models and mixed models. In this subsection we explain why one can view penalised splines as mixed models. Consider the regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon}, \quad \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \sigma_u^2 \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_\varepsilon^2 \boldsymbol{I} \end{bmatrix} \right), \tag{1.4}$$

for the general design matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$. Let $\boldsymbol{C} \equiv [\boldsymbol{X}\ \boldsymbol{Z}]$, the least squares estimator to (1.4) can be expressed as a best linear unbiased predictor given as

$$\hat{\boldsymbol{\nu}} = \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{u}} \end{bmatrix} = \left(\boldsymbol{C}^\mathsf{T}\boldsymbol{C} + \lambda \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \right)^{-1} \boldsymbol{C}^\mathsf{T}\boldsymbol{y}, \tag{1.5}$$

where the smoothing parameter is $\lambda = \sigma_u^2/\sigma_\varepsilon^2$. The equivalence of (1.3) can be achieved if a $(K+4)\times(K+4)$ linear transformation matrix $\boldsymbol{L}$ can be found such that

$$\boldsymbol{C} = \boldsymbol{B}\boldsymbol{L} \quad \text{and} \quad \boldsymbol{L}^\mathsf{T}\boldsymbol{\Omega}\boldsymbol{L} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix}.$$

According to Wand and Ormerod (2008), the usual method for obtaining $\boldsymbol{L}$ and $\boldsymbol{\Omega}$ is through spectral decomposition, and their resultant forms are respectively

$$\boldsymbol{L} = [\boldsymbol{U}_X \mid \boldsymbol{U}_Z \operatorname{diag}(\boldsymbol{d}_Z^{-1/2})] \quad \text{and} \quad \boldsymbol{\Omega} = \boldsymbol{U}\operatorname{diag}(d)\boldsymbol{U}^\mathsf{T},$$

where $\boldsymbol{U}^\mathsf{T}\boldsymbol{U} = \boldsymbol{I}$ and $\boldsymbol{d}$ is a $(K+4)\times 1$ vector with exactly two zero entries and $K+2$ positive entries. The term $\boldsymbol{d}_Z$ is the $(K+2)\times 1$ is a subvector of $\boldsymbol{d}$ containing these positive entries and $\boldsymbol{U}_Z$ is the $(K+4)\times(K+2)$ submatrix of $\boldsymbol{U}$ with columns corresponding to the positive entries of $\boldsymbol{d}$. Following this, model (1.4) with the O'Sullivan penalised splines can be fitted with the following design matrices:

$$\boldsymbol{X} = \boldsymbol{B}\boldsymbol{U}_X \quad \text{and} \quad \boldsymbol{Z} = \boldsymbol{B}\boldsymbol{U}_Z\operatorname{diag}(\boldsymbol{d}_Z^{-1/2}).$$

1.4 Graphical models

Graphical models, originated from the field of machine learning and pattern recognition, are a branch of mathematics that combines ideas from graph theory and probability (e.g. Bishop, 2006; Jordan, 2004). Graphical models, also known as the *probabilistic graphical models*, are diagrammatic representations of probability distributions. A graph comprises

a set of *nodes* connected by *edges*. Adapting the conventions of Bishop (2006), nodes are shown as circles and edges are shown as line segments. Each node represents a random variable (or a set of random variables), and the edges express probabilistic relationships between these variables. There are two major types of graphical models: *directed acyclic graphs* (DAGs), also known as the *Bayesian networks*, and *undirected graphs*, also known as the *Markov random fields*. DAGs are distinguished from the undirected graphs by their directed edges, each of which has a single-headed arrow (double-headed arrows are not allowed). In both cases the graph expresses the way in which the joint density function over all of the random variables can be decomposed into a product of factors each depending on a subset of nodes, but the underlying relationship between the nodes and the factorisation is different for the two types of graphs. Of the two types, DAGs are more relevant to the theme of this thesis, and therefore we restrict our attention to DAGs.

Figure 1.1 is an elementary example of a DAG. In the graphical models literature, the variables x_1, x_2, x_3, x_5, x_6 are random ("*hidden*") nodes denoted by $\mathcal{H}$, corresponding to each of the open circles. The variable x_4 is an observed ("*evidence*") node denoted by $\mathcal{E}$, corresponding to the shaded circle. A DAG can be viewed as a "family tree", where each directed edge signifies a "parent-child" relationship between the corresponding nodes. For example, the node x_2 meets by the arrow-head is a *child* of the node x_4 (*parent*). The nodes x_2 and x_3 have a common child x_4 are called *co-parents* of each other. The *ancestors* of a node, say x_4, is the set consisting of the node's parents, parents of the node's parents, and so on, i.e. $\{x_1, x_2, x_3, x_4\}$. The *Markov blanket* of a node, say x_4, is the set consisting of the node's parents, co-parents and children, i.e. $\{x_2, x_3, x_5, x_6\}$, as indicated by the blue dashed line polygon in Figure 1.1. The Markov blanket of a node separates that node from the remainder of the graph probabilistically and provides an important *locality property of probabilistic graphs* as we shall elaborate later on.

The DAG structure can easily be used to convey conditional independence properties of a model using the notions of *smallest ancestral subgraphs* and *moral graphs*. Let $S = \{x_1, x_2, x_3, x_4\}$ be a subset of nodes, then an ancestral subgraph is a subgraph of a DAG, which, for any node in S, each of the node's ancestors are also in the subgraph. The smallest ancestral subgraph containing S is the ancestral subgraph that has the fewest number of nodes (see Figure 1.2 left).The moral graph of a DAG is obtained by linking any pair of unconnected parents of a common child node with an undirected edge and changing all directed edges to undirected edges (referred to as *moralisation*, see Figure 1.2 right). Using the moral (undirected) graph in Figure 1.2, and suppose $A = \{x_1\}$, $B = \{x_4\}$ and $C = \{x_2, x_3\}$, then the conditional independence structure for these three sets of nodes is

$$A \perp\!\!\!\perp B \mid C \quad \text{i.e.} \quad x_1 \perp\!\!\!\perp x_4 \mid \{x_2, x_3\},$$

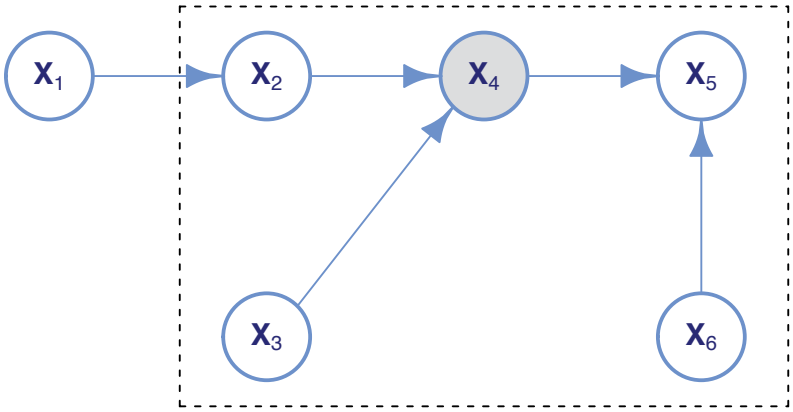

Figure 1.1: Directed acyclic graph involving six random variables: $x_1, \ldots, x_6$. The open circles represent the random ("hidden") nodes. The shaded circle represents the observed ("evidence") node. The grey dashed line polygon represents the Markov blanket of the node x_4.

given that all the paths connecting nodes in A and nodes in B would pass through one or more nodes in C. The set of nodes in C *blocks* the paths from A to B and therefore the conditional independence holds. This conditional independence can also be derived using the *d-separation theorem* for a DAG.

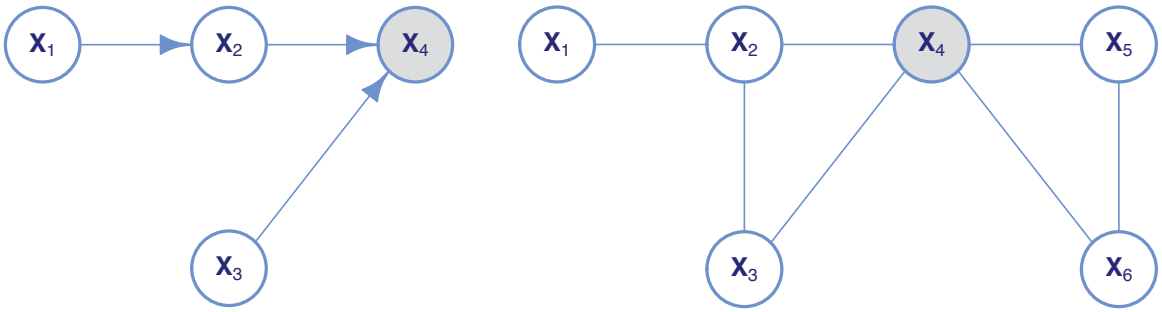

Figure 1.2: Left: The smallest ancestral subgraph corresponding to Figure 1.1. Right: The moral graph corresponding to Figure 1.1.

Through exploiting the graph-theoretic representation, the joint density function over all of the random variables can be factorised into a product of conditional density functions,

$$\begin{aligned} p(x_1, x_2, x_3, x_4, x_5, x_6) &= \prod_{i=1}^{6} p(x_i \,|\, \text{parents of } x_i) \\ &= p(x_1)\, p(x_2|x_1)\, p(x_3)\, p(x_4|x_2, x_3)\, p(x_5|x_4, x_6)\, p(x_6), \end{aligned}$$

where parenthood is determined by the directed edges of the DAG. As demonstrated, conditional independence properties are useful in simplifying the structure of probabilis-

tic graphical models and thus more efficient algorithms can be devised for performing statistical inference. In a Bayesian setting, the model parameters are treated as random variables and represented as nodes in the graph. Computing the posterior distribution of those parameters is simply yet another inference problem!

In the next section, we describe probability and statistics from a Bayesian viewpoint, which provides a principled formalism through which all sources of uncertainty can be addressed consistently. Graphical models offers several useful properties in such a viewpoint including: (i) they provide a simple way to visualise the hierarchical structure of a probabilistic model; (ii) insights into conditional properties of the model can be obtained via inspection of the graph; and (iii) complex computations for performing inference can be expressed in terms of graphical manipulations and therefore the underlying algebraic expressions are carried along implicitly (Bishop, 2008).

1.5 Bayesian inference

Bayesian inference involves finding the joint posterior distribution of parameters of interest given the observed data. In many instances, exact inference is infeasible due to the posterior distributions being intractable, requiring approximate inference methods for use in practice. In recent years, Bayesian inference engines have emerged for approximate inference for general classes of mixed models. Examples include BUGS (Ligges *et al.*, 2009) and Stan (Stan Development Team, 2015), which are based on MCMC methods. In Subsection 1.5.2, we describe MCMC being one of the current standard methods for estimating Bayesian statistical models and has proven useful in a wide range of problems. However, for large models with complex posteriors, MCMC can be computationally intensive and suffers from poor mixing that leads to slow convergence. We therefore opt for an alternative approach, namely *variational approximations*, as we shall elaborate on in Subsection 1.5.3.

1.5.1 A Bayesian viewpoint of probability and statistics

Consider a generic Bayesian model with parameter vector $\boldsymbol{\theta} = \{\theta_1, \dots \theta_n\} \in \Theta$ and observed data vector $\boldsymbol{y}$. The goal of Bayesian inference is to infer the conditional density function of posterior $\boldsymbol{\theta}$ given $\boldsymbol{y}$ (known as the *posterior density function*),

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

The numerator $p(\boldsymbol{y}, \boldsymbol{\theta})$ is known as the *joint density function* of $\boldsymbol{y}$ and $\boldsymbol{\theta}$, which can be written as a product of the sampling density function $p(\boldsymbol{y}|\boldsymbol{\theta})$ and the prior density function $p(\boldsymbol{\theta})$. The denominator $p(\boldsymbol{y})$ is known as the *marginal likelihood*, where $p(\boldsymbol{y}) =$

$\int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta})\, p(\boldsymbol{y}|\boldsymbol{\theta})\, d\boldsymbol{\theta}$ in the case of continuous $\boldsymbol{\theta}$, or replace $\int_{\boldsymbol{\Theta}}$ with $\sum_{\boldsymbol{\Theta}}$ in the case of discrete $\boldsymbol{\theta}$.

For many models of practical interest the posterior density functions are intractable; they cannot be used to directly calculate the marginal density functions of parameters or other quantities of interest. For a long time the only generally applicable method was to use MCMC sampling techniques. MCMC is a sampling-based simulation algorithm which circumvents the intractability problem by generating a Markov Chain of samples of $\boldsymbol{\theta}^{(t)}$, $0 \leq t \leq T$, to approximate the target posterior density function $p(\boldsymbol{\theta}|\boldsymbol{y})$. A *Markov Chain* is a sequence of random variables $\boldsymbol{\theta}^{(0)}, \ldots \boldsymbol{\theta}^{(T)}$ for which

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

so that the current state depends only on the preceding state.

1.5.2 Markov chain Monte Carlo

Gibbs sampling is the simplest component-wise algorithm of the MCMC methods (Gelfand and Smith, 1990; Geman and Geman, 1984; Robert and Casella, 2004). The Markov chain of samples of $\boldsymbol{\theta}^{(t)}$, $0 \leq t \leq T$, is obtained by directly sampling from the *full conditional density functions* $p\left(\theta_i^{(t)}|\theta_{[i]}^{(t-1,t)}\right)$, specifying the density of θ_i conditional on other parameters $\theta_{[i]}$. Each subvector $\theta_i^{(t)}$ is updated conditional on the latest value of the other parameters $\theta_{[i]}$, which are the iteration t values for the parameters already updated and the iteration $t-1$ values for the others. This is illustrated in Algorithm 1.

Initialise: $\theta_1^{(0)}, \ldots, \theta_n^{(T)}$

Cycle:

For $t = 0, \ldots, T$:

$$\begin{aligned}
\theta_1^{(t)} &\sim p(\theta_1|\theta_2^{(t-1)}, \ldots, \theta_n^{(t-1)}) \\
\theta_2^{(t)} &\sim p(\theta_2|\theta_1^{(t)}, \ldots, \theta_n^{(t-1)}) \\
&\vdots \\
\theta_{n-1}^{(t)} &\sim p(\theta_{n-1}|\theta_1^{(t)}, \ldots, \theta_n^{(t-1)}) \\
\theta_n^{(t)} &\sim p(\theta_n|\theta_1^{(t)}, \ldots, \theta_{n-1}^{(t)})
\end{aligned}$$

until convergence reaches.

Algorithm 1: The Gibbs sampler algorithm in generic form.

Unfortunately, this approach tends to be computationally costly for complex models

that leads to unacceptably slow convergence and does not scale well to real-world applications involving large datasets.

To address this problem, a research group from Columbia University has created Stan: new, high-performance open-source software for Bayesian inference on multilevel models (Stan Development Team, 2015). Rather than the conventional Gibbs sampler, Stan uses a variant of Hamiltonian Monte Carlo algorithm to speed up convergence of generalised multilevel linear models. Instead of the typical random-walk behaviours, Hamiltonian Monte Carlo is a MCMC method that simulates a physical system governed by Hamiltonian dynamics (a way that physicists use to describe an object's motion in terms of its location and momentum at some time point) to propose future states in the Markov chain. This allows the Markov chain to explore the posterior distribution much more efficiently, resulting in faster convergence.

1.5.3 A brief introduction to variational approximations

Variational approximations are a fast, versatile and deterministic class of approximations with origins in the statistical physics and computer science literature (Bishop, 2006; Jordan *et al.*, 1999; Titterington, 2004; Wainwright and Jordan, 2008). Since about 2005, variational methods have been increasingly explored in the statistical literature. For example, McGrory and Titterington (2007), Wand *et al.* (2012) and Ormerod and Wand (2010) present variational methodology for a diversified range of applications, from finite mixture models, complex models with elaborate distributions (such as asymmetric Laplace and skew normal) to spline and wavelet regression models. In addition, Wang and Titterington (2006) prove convergence of variational algorithms for normal mixture models and You *et al.* (2014) propose several information criteria that are useful for model selection. We provide here a brief overview of variational approximations and highlight the key concepts. Comprehensive summaries can be found in Bishop (2006) and Ormerod and Wand (2010).

1.5.3.1 Mean field variational Bayes

Consider a generic Bayesian model with observed data vector $\boldsymbol{y}$ and parameter vector $\boldsymbol{\theta}$ that is continuous over the parameter space $\boldsymbol{\Theta}$. The essence of variational inference is to approximate the posterior density function $p(\boldsymbol{\theta}|\boldsymbol{y})$ with a so-called *approximating density function* $q(\boldsymbol{\theta})$. To make this approximation as close as possible, we search over $q \in \mathcal{Q}$, for some set $\mathcal{Q}$ of density functions, to find a particular density function with the minimum Kullback-Liebler (KL) distance/divergence with the actual posterior

$$q^*(\boldsymbol{\theta}) = \underset{q \in \mathcal{Q}}{\operatorname{argmin}} \, \mathrm{KL}\left\{q(\boldsymbol{\theta}) \,\|\, p(\boldsymbol{\theta}|\boldsymbol{y})\right\}, \tag{1.6}$$

where

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

The approximating distribution can be chosen to have a simple analytical form. For example, it could be a Gaussian distribution, whose mean and covariance are optimised to minimise the Kullback-Liebler distance with respect to the actual posterior. A more flexible framework is to impose some form of factorisation or product density restriction on $q(\boldsymbol{\theta})$, without any restriction on the functional form of the factors (Bishop, 2006):

$$\mathcal{Q} = \left\{ q(\boldsymbol{\theta}) \,:\, q(\boldsymbol{\theta}) = \prod_{i=1}^{M} q_i(\boldsymbol{\theta}_i) \text{ for some partition } \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\} \text{ of } \boldsymbol{\theta} \right\}, \tag{1.7}$$

known as the *mean field restriction*. The resultant $q^*(\boldsymbol{\theta})$ is known as the *mean field variational Bayes approximation* (MFVB) to the actual posterior, and from now on we refer to it as *optimal q-density*. Essentially, it assumes posterior independence among parameters that may not be present in the actual posterior. Depending on the chosen partition of $\boldsymbol{\theta}$, this independence assumption may be rather unrealistic in settings where there exists high posterior correlations among parameters, thus leading to poor approximations. The particular parametric families that constitute each of the approximating q-density factors $q_i(\boldsymbol{\theta}_i)$ are derived through the variational methodology. Indeed, minimising the Kullback-Liebler divergence in (1.6) is equivalent to maximising the lower bound $\underline{p}(\boldsymbol{y}; q)$ since the marginal log-likelihood can be expressed as (Ormerod and Wand, 2010)

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

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

The optimal q-density functions under product restriction (1.7) can be obtained via an *iterative coordinate ascent algorithm* that is analogous to the Expectation-Maximisation algorithm (Bishop, 2006; Ormerod and Wand, 2010). Define

$$E_{q(-\boldsymbol{\theta}_i)}\{\log p(\boldsymbol{y},\boldsymbol{\theta})\} = \int \log p(\boldsymbol{y},\boldsymbol{\theta}) \prod_{j \neq i} q(\boldsymbol{\theta}_j) d\boldsymbol{\theta}_j$$

to be the log posterior averaged over the current estimates of the approximating density functions for all but the ith parameter vector. The term $E_{q(-\boldsymbol{\theta}_i)}$ denotes expectation with respect to the q-densities of all parameters except $\boldsymbol{\theta}_i$. Variational approximations is now reduced to solving an optimisation problem in the form of (1.6). We proceed by initialising each of the q-density factors $q_i(\boldsymbol{\theta}_i)$ and updating each factor successively using the current estimates of the other factors. At the end of each iteration, an updated value of the lower bound is computed,

$$\log \underline{p}(\boldsymbol{y}; q) \quad = \quad E_q\{\log p(\boldsymbol{y}, \boldsymbol{\theta}) - \log q(\boldsymbol{\theta})\}, \tag{1.8}$$

and the algorithm is iterated until convergence of the lower bound to its maximum, see Algorithm 2. Convergence of such simple iterative updates is guaranteed in most cases since the algorithm is fundamentally a generalisation of Expectation-Maximisation (Chappell *et al.*, 2009). The second term on the right hand side of (1.8) corresponds to the *entropy* of a density function. If x is a random vector with a density function p then the corresponding entropy is given by

$$\text{Entropy}(p) \equiv E_p\{-\log p(x)\}.$$

For many common distribution families, the entropy can be expressed algebraically in terms of the distribution's parameters. For example, if $p(x; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate normal density function of dimension d with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ then

$$\text{Entropy}(p; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \tfrac{1}{2}\, d\{1 + \log(2\pi)\} + \tfrac{1}{2} \log|\boldsymbol{\Sigma}|.$$

Another common entropy expression that arises throughout this thesis is the inverse-Gamma family of density functions. The entropy corresponds to $p(x; \kappa, \lambda)$ with parameters $\kappa, \lambda > 0$ is

$$\text{Entropy}(p; \kappa, \lambda) = \log(\lambda) + \kappa + \log\{\Gamma(\kappa)\} - (\kappa + 1)\, \text{digamma}(\kappa),$$

where $\text{digamma}(\kappa) \equiv (d/dx) \log \Gamma(\kappa)$ is the digamma function.

While the distributional forms of the approximating density functions $q_i(\boldsymbol{\theta}_i)$ are unspecified, the structure of the statistical model itself lends to a solution that lies in a particular parametric family for each of the $q_i(\boldsymbol{\theta}_i)$. For example, when all the parameters in a model are conditionally conjugate, the optimal q-density functions in Algorithm 2 are available in closed form. If $q_i(\boldsymbol{\theta}_i)$ can not be recognised as a standard distribution, then numerical integration methods are required to estimate the marginal likelihood, which is computationally more demanding.

Initialise: $q_1(\boldsymbol{\theta}_1), \ldots, q_M(\boldsymbol{\theta}_M)$

Cycle:

$$q_1(\boldsymbol{\theta}_1) \leftarrow \frac{\exp[E_{q(-\boldsymbol{\theta}_1)}\{\log p(\boldsymbol{y}, \boldsymbol{\theta})\}]}{\int \exp[E_{q(-\boldsymbol{\theta}_1)}\{\log p(\boldsymbol{y}, \boldsymbol{\theta})\}] d\boldsymbol{\theta}_1}$$

$$\vdots$$

$$q_M(\boldsymbol{\theta}_M) \leftarrow \frac{\exp[E_{q(-\boldsymbol{\theta}_M)}\{\log p(\boldsymbol{y}, \boldsymbol{\theta})\}]}{\int \exp[E_{q(-\boldsymbol{\theta}_M)}\{\log p(\boldsymbol{y}, \boldsymbol{\theta})\}] d\boldsymbol{\theta}_M}$$

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

Algorithm 2: Iterative scheme for obtaining the optimal q-density functions under product restriction (1.7).

Upon convergence, Algorithm 2 shows that the optimal occurs when

$$q_i^*(\boldsymbol{\theta}_i) \propto \exp\left[E_{(-\boldsymbol{\theta}_i)}\{\ln\, p(\boldsymbol{\theta}, \boldsymbol{y})\}\right], \quad 1 \leq i \leq M.$$

This means that the log of the optimal solution for each factor is simply obtained by considering the log of the joint density function over all of the random and observed variables and taking the expectation with respect to all of the other factors $q_j(\boldsymbol{\theta}_j)$, $j \neq i$ (Ormerod and Wand, 2010). An alternative expression for $q_i^*(\boldsymbol{\theta}_i)$ is

$$q_i^*(\boldsymbol{\theta}_i) \quad \propto \quad \exp\left[E_{q(-\boldsymbol{\theta}_i)}\{\ln\, p(\boldsymbol{\theta}_i|\text{rest})\}\right], \quad 1 \leq i \leq M, \tag{1.9}$$

where "rest" denotes all of the random variables excluding $\boldsymbol{\theta}_i$, i.e. $\{\boldsymbol{y}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1}, \boldsymbol{\theta}_{i+1}, \ldots \boldsymbol{\theta}_M\}$, and the distribution $p(\boldsymbol{\theta}_i|\text{rest})$ is known as the *full conditional density function*. The formulation of (1.9) greatly benefits from a DAG-related concept known as the *Markov blanket theory* - Theorem 1.5.1. The Markov blanket of a node on a DAG is the set of children, parents and coparents of that node. Two nodes are coparents if they have at least one child node in common.

Theorem 1.5.1. For each node on a probabilistic DAG, the conditional distribution of the node given the rest of the nodes is the same as the conditional distribution of the node given its Markov blanket (Dechter and Pearl, 1988).

For our generic Bayesian example, this implies that

$$p(\boldsymbol{\theta}_i|\text{rest}) = p(\boldsymbol{\theta}_i|\text{Markov blanket of } \boldsymbol{\theta}_i).$$

It immediately follows that

$$q_i^*(\boldsymbol{\theta}_i) \quad \propto \quad \exp\left[E_{q(-\boldsymbol{\theta}_i)}\{\log p(\boldsymbol{\theta}_i|\text{Markov blanket of } \boldsymbol{\theta}_i)\}\right], \quad 1 \leq i \leq M. \quad (1.10)$$

This is known as the *locality property of DAGs*. For large DAGs, this property yields considerable algebraic benefits. In particular, it shows that the $q_i^*(\boldsymbol{\theta}_i)$ require only local calculations on the model's DAG.

1.5.3.2 Variational message passing

Variational message passing or VMP(e.g. Winn and Bishop, 2005) is a message passing algorithmic implementation of the mean field approximation, limited to conjugate-exponential models. The VMP works with *messages* passed between neighbouring nodes (i.e. parent and child) on a *factor graph* of the model. Figure 1.3 shows an example of a factor graph corresponding to the DAG in Figure 1.1. Examples of the messages are:

$$m_{p(x_1) \to x_1}(x_1), \quad m_{x_1 \to p(x_1)}(x_1), \quad m_{x_1 \to p(x_2|x_1)}(x_1), \quad m_{p(x_2|x_1) \to x_1}(x_2).$$

The VMP messages are equivalent to the q-density factors $q_i(\boldsymbol{\theta}_i)$ in MFVB, and their conditional density functions are members of the exponential family including Gaussian, Poisson, Gamma etc. Thus, calculating $q_i(\boldsymbol{\theta}_i)$ simply involves summing sufficient statistics, and similarly, the variational optimisation can be decomposed into a set of local computations that depend only on messages from neighbouring nodes in the DAG.

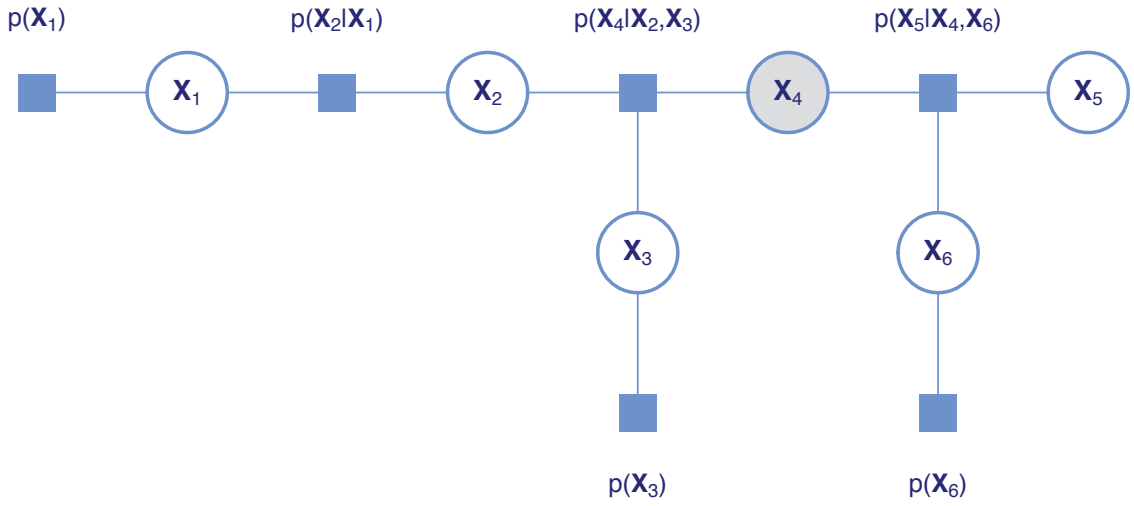

Figure 1.3: Factor graph corresponding to the DAG in Figure 1.1.

An in-depth explanation as well as a discussion on the factor graphs and VMP is deferred to Chapter 6.

1.5.3.3 Non-conjugate variational message passing

Knowles and Minka (2011) introduce *non-conjugate variational message passing* (NCVMP) which widens the scope of tractable models for VMP and MFVB in general. Specifically, it extends variational inference to non-conjugate exponential models. The modularity of NCVMP allows modifications only to those model parameters that involve previously intractable solutions, whilst keeping VMP or MFVB solutions of the other model parameters the same. Let us reconsider the generic Bayesian model with parameter vector $\boldsymbol{\theta}$ and observed data vector $\boldsymbol{y}$, but this time include an additional parameter vector $\boldsymbol{\phi}$. From (1.7), the MFVB approximates the joint posterior density function $p(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{y})$ by the factorised form

$$q_{\boldsymbol{\theta}_1}(\boldsymbol{\theta}_1), \ldots, q_{\boldsymbol{\theta}_M}(\boldsymbol{\theta}_M)\, q_{\boldsymbol{\phi}}(\boldsymbol{\phi}), \tag{1.11}$$

and suppose that $q_{\boldsymbol{\theta}_1}(\boldsymbol{\theta}_1), \ldots, q_{\boldsymbol{\theta}_M}(\boldsymbol{\theta}_M)$ are in the exponential family and $q_{\boldsymbol{\phi}}(\boldsymbol{\phi})$ is not. The optimal q-density functions satisfy

$$\begin{aligned} q^*_{\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) &\propto \exp\left[E_{q(-\boldsymbol{\theta}_i)}\left\{\log p(\boldsymbol{\theta}_i|\text{rest})\right\}\right], \quad 1 \leq i \leq M, \\ \text{and} \quad q^*_{\boldsymbol{\phi}}(\boldsymbol{\phi}) &\propto \exp\left[E_{q(-\boldsymbol{\phi})}\left\{\log p(\boldsymbol{\phi}|\text{rest})\right\}\right]. \end{aligned}$$

The term $E_{q(-\boldsymbol{\phi})}\left\{\log p(\boldsymbol{\phi}|\text{rest})\right\}$ is intractable due to difficulties arising from non-conjugacy. Non-conjugate variational message passing offers a remedy by postulating $q_{\boldsymbol{\phi}}(\boldsymbol{\phi})$ to be an exponential family density function with the *natural parameter vector* $\boldsymbol{\eta}$, *natural statistic vector* $\boldsymbol{T}(\boldsymbol{\phi})$ and the *log-partition function* $A(\boldsymbol{\eta})$:

$$q(\boldsymbol{\theta}; \boldsymbol{\eta}) = \exp\left\{\boldsymbol{T}(\boldsymbol{\phi})^{\mathsf{T}}\boldsymbol{\eta} - A(\boldsymbol{\eta}) + \text{const}\right\}.$$

where "const" denotes additive constants with respect to the function argument. Hence, (1.11) becomes

$$q_{\boldsymbol{\theta}_1}(\boldsymbol{\theta}_1), \ldots, q_{\boldsymbol{\theta}_M}(\boldsymbol{\theta}_M)\, q_{\boldsymbol{\phi}}(\boldsymbol{\phi}; \boldsymbol{\eta}).$$

From Theorem 1 of Knowles and Minka (2011), the optimal q-density functions can then be found using

$$\begin{aligned} q^*_{\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) &\propto \exp\left[E_{q(-\boldsymbol{\theta}_i)}\left\{\log p(\boldsymbol{\theta}_i|\text{rest})\right\}\right], \quad 1 \leq i \leq M, \\ \boldsymbol{\eta} &\leftarrow \left[\text{var}\left\{\boldsymbol{T}(\boldsymbol{\phi})\right\}\right]^{-1}\left\{\mathsf{D}_{\boldsymbol{\eta}} E_{q(\boldsymbol{\theta},\boldsymbol{\phi})}\left\{\log p(\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\phi})\right\}\right\}, \end{aligned} \tag{1.12}$$

where $\mathsf{D}_{\boldsymbol{x}} f$ is defined in Section 1.10. This thesis only concerns a special case where $q(\boldsymbol{\theta}; \boldsymbol{\eta})$ corresponds to a d-dimensional multivariate normal density function. The advantage of

exponential family density functions is that expectations of their logarithms are tractable to compute and their state can be summarised completely by the natural parameter vector. A recent example of utilising non-conjugate VMP for approximate Bayesian inference in generalised linear mixed models is given in Tan and Nott (2013).

1.6 Terminological and notational conventions

Lower-case Roman and Greek letters in boldface are used to denote random vectors with entries consisting of subscripts. For example, $\boldsymbol{x}$ denotes a $n\times 1$ vector containing $x_1, \ldots, x_n$. Upper-case Roman and Greek letters in boldface are used to denote random matrices. For example, $\boldsymbol{X}$ denotes a $m \times n$ matrix containing n vectors of dimension $m \times 1$, i.e. $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$. Scalar functions applied to vectors are evaluated element-wise. For example,

$$\exp(a_1, a_2, a_3) \equiv (\exp(a_1), \exp(a_2), \exp(a_3)).$$

For two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of the same dimensions, the element-wise product denoted $\boldsymbol{A} \odot \boldsymbol{B}$ is a matrix with elements $(\boldsymbol{A} \odot \boldsymbol{B})_{ij} = (\boldsymbol{A})_{ij}\,(\boldsymbol{B})_{ij}$. For two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of any different dimension $m \times n$ and $p \times q$ respectively, the Kronecker product denoted $\boldsymbol{A} \otimes \boldsymbol{B}$ is a matrix with dimension $mn \times pq$ with elements $(\boldsymbol{A} \otimes \boldsymbol{B})_{ij} = (\boldsymbol{A})_{ij}\boldsymbol{B}$. If $\boldsymbol{A}$ and $\boldsymbol{B}$ have the same number of rows, then the notation $(\boldsymbol{A}|\boldsymbol{B})$ denotes concatenation (by columns) of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. We use $\boldsymbol{1}_d$ to denote the $d \times 1$ vector with all entries equal to 1 and $\boldsymbol{I}_d$ to denote the $d \times d$ identity matrix. The norm of a vector denoted $\|\boldsymbol{v}\|$ is defined to be $\sqrt{\boldsymbol{v}^\mathsf{T}\boldsymbol{v}}$. For a $n \times n$ symmetric matrix $\boldsymbol{M}$, the trace and determinant are denoted by $\mathrm{tr}(\boldsymbol{M})$ and $|\boldsymbol{M}|$ respectively and adopt the usual definitions.

For a $d \times 1$ vector $\boldsymbol{a}$, we use $\mathrm{diag}(\boldsymbol{a})$ to denote the $d \times d$ diagonal matrix containing entries of $\boldsymbol{a}$ along the main diagonal. For a $d \times d$ square matrix $\boldsymbol{A}$, we use $\mathrm{diagonal}(\boldsymbol{A})$ to denote the $d \times 1$ vector containing the main diagonal entries of $\boldsymbol{A}$. For square matrices $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_r$, we use $\mathrm{blockdiag}(\boldsymbol{A}_1, \ldots, \boldsymbol{A}_r)$ to denote the block diagonal matrix with the i^{th} block equal to $\boldsymbol{A}_i$. The transpose of $\boldsymbol{A}$ is denoted by $\boldsymbol{A}^\mathsf{T}$. We let $\boldsymbol{A}^{-1}$ denote the inverse of a matrix provided $\boldsymbol{A}$ is an invertible matrix.

For a $d \times d$ matrix $\boldsymbol{A}$, we use $\mathrm{vec}(\boldsymbol{A})$ to denote the $d^2 \times 1$ vector obtaining by concatenating the columns of $\boldsymbol{A}$ underneath each other from left to right. The term $\mathrm{vech}(\boldsymbol{A})$ denotes the $\frac{1}{2}\,d(d+1) \times 1$ vector obtained from $\mathrm{vec}(\boldsymbol{A})$ by eliminating the above-diagonal entries of $\boldsymbol{A}$. For example,

$$\mathrm{vec}\left(\begin{bmatrix} 3 & 8 \\ 4 & 9 \end{bmatrix}\right) = \begin{bmatrix} 3 \\ 4 \\ 8 \\ 9 \end{bmatrix} \quad \text{and} \quad \mathrm{vech}\left(\begin{bmatrix} 3 & 8 \\ 4 & 9 \end{bmatrix}\right) = \begin{bmatrix} 3 \\ 4 \\ 9 \end{bmatrix}.$$

Let u and v be two continuous or discrete random scalar variables, we denote $p(u)$ as the density function (probability mass function for discrete and probability density function for continuous). This density function notation extends, analogously, to joint and conditional density functions. For example, $p(u, v)$ denotes the joint density function of u and v, and $p(u|v)$ denotes the conditional density function of u given by v. The expectation and covariance matrix of u are denoted by $E(u)$ and $\text{Var}(u)$ respectively. Analogous terminology applies to random vectors and matrices.

The q denotes the density functions that arise from the MFVB approximation. For a generic random scalar variable θ and density function q we define

$$\mu_{q(\theta)} \equiv E_q(\theta) \quad \text{and} \quad \sigma_{q(\theta)} \equiv \text{Var}_q(\theta).$$

For a generic random vector $\boldsymbol{\theta}$ and density function we define

$$\boldsymbol{\mu}_{q(\boldsymbol{\theta})} \equiv E_q(\boldsymbol{\theta}) \quad \text{and} \quad \boldsymbol{\Sigma}_{q(\boldsymbol{\theta})} \equiv \text{Cov}_q(\boldsymbol{\theta}).$$

1.7 Common distributions

If y_i has a distribution D_i for each $1 \leq i \leq n$ and the y_i are independent, then we write $y_i \overset{\text{ind.}}{\sim} D_i$. Table 1.1 lists all of the distributions used in this thesis. In particular, the parametrisation of the corresponding density functions is provided.

1.8 Distributional results

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

$$x|\, a \sim \text{Inverse-Gamma}\,(\tfrac{1}{2}, 1/a) \quad \text{and} \quad a \sim \text{Inverse-Gamma}\,(\tfrac{1}{2}, 1/A^2).$$

Then $\sqrt{x} \sim \text{Half-Cauchy}(A)$.

The above hierarchical representation enables MCMC and variational methods easier to carry out because of the conditional conjugacy properties of the Inverse-Gamma distribution. Extending Result 1.8.1 leads to the following:

$$\begin{aligned} \boldsymbol{\Sigma}|a_1, \ldots, a_p &\sim \text{Inverse-Wishart}\,(\nu + p - 1, 2\,\nu\,\text{diag}(1/a_1, \ldots, 1/a_p)) \\ \text{and} \quad a_k &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2}, 1/A_k^2), \quad k = 1, \ldots, p, \end{aligned}$$

where $\boldsymbol{\Sigma}$ is a $p \times p$ random matrix and $\boldsymbol{a}$ is a $p \times 1$ random vector. According to Huang and Wand (2013) simulation results, the scale-mixture $\boldsymbol{\Sigma}$ offers an overall increase in practical performance over the classical inverse-Wishart prior for sparse covariance matrix

Distribution	Probability density or mass function in x	Abbreviation
Bernoulli	$p^x\,(1-p)^{1-x}; \quad x=0,1;\, p\in(0,1)$	Bernoulli(p)
Poisson	$\lambda^x\, e^{-\lambda}/x!; \quad x=0,1,\ldots;\, \lambda>0$	Poisson(λ)
Uniform	$1/(b-a); \quad a<x<b$	Uniform(a,b)
Normal	$\frac{1}{\sqrt{2\pi\sigma}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}; \quad x,\mu\in\mathbb{R};\, \sigma^2>0$	$N(\mu,\sigma^2)$
Student-t	$\frac{\Gamma\{(\nu+1)/2\}}{\sqrt{\pi\nu\sigma^2}\,\Gamma(\frac{\nu}{2})\left\{1+\frac{(x-\mu)^2}{\nu\sigma^2}\right\}^{(\nu+1)/2}}; \quad \sigma,\nu>0$	$t(\mu,\sigma^2,\nu)$
Gamma	$\frac{B^A\,x^{A-1}e^{-B\,x}}{\Gamma(A)}; \quad x>0;\; A,B>0$	Gamma(A,B)
Inverse-Gamma	$\frac{B^A\,x^{-A-1}e^{-B/x}}{\Gamma(A)}; \quad x>0;\; A,B>0$	Inverse-Gamma(A,B)
Half-Cauchy	$\frac{2\sigma}{\pi(x^2+\sigma^2)}; \quad x>0;\; \sigma>0$	Half-Cauchy(σ)
Multivariate Normal	$\lvert 2\pi\boldsymbol{\Sigma}\rvert^{-1/2}\exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}$	$N(\boldsymbol{\mu},\boldsymbol{\Sigma})$
	$\boldsymbol{x}\in\mathbb{R}^d$; $\boldsymbol{\Sigma}$ is symmetric and positive definite	
Inverse-Wishart	$(\mathcal{C}_{d,A}^{-1}\lvert\boldsymbol{B}\rvert^{A/2}\lvert\boldsymbol{X}\rvert^{(A+d+1)/2}\exp\left\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{B}\boldsymbol{X})\right\}$	Wishart($A,\boldsymbol{B}$)
	where $(\mathcal{C}_{d,A}\equiv 2^{Ad/2}\pi^{d(d-1)/4}\prod_{j=1}^{d}\Gamma\left(\frac{A+1-j}{2}\right)$.	
	$A>0$; $\boldsymbol{B}$ is symmetric and positive definite	

Table 1.1: Distributions used in this thesis and their corresponding probability density or mass functions.

estimation. In addition, a compelling feature of the scale-mixture $\boldsymbol{\Sigma}$ is that, its marginal distribution is invariant even when the matrix itself is expanded or collapsed over any set of variables in a self-consistent manner.

Result 1.8.2. Let x and b be random variables such that

$$x|\,b\sim \mathrm{N}(\mu, b\,\sigma^2) \quad \text{and} \quad b\sim \text{Inverse-Gamma}(\tfrac{\nu}{2},\nu/2).$$

Then $\sqrt{x}\sim \mathrm{t}(\mu,\sigma,\nu)$.

Result 1.8.3. If $\nu\sim$ Inverse-Gamma(A,B) then

$$E(1/\nu)=A/B \quad \text{and} \quad E\left\{\log(\nu)\right\}=\log(B)-\text{digamma}(A).$$

1.9 Vector differential calculus

Definition 1.9.1. The derivative factor of a scalar-valued function f with respect to $x \in \mathbb{R}^p$ is denoted by $\mathsf{D}_x f$, with the ith entry equals to

$$\frac{\partial f(\boldsymbol{x})_i}{\partial x_i}.$$

Definition 1.9.2. The Hessian matrix of a scalar-valued function f with respect to $x \in \mathbb{R}^p$ is denoted by $\mathsf{H}_x f$, with the (i,j)th entry equals to

$$\frac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j}.$$

Definition 1.9.3. Let $\boldsymbol{U}$ and $\boldsymbol{V}$ are matrix functions, then

$$\begin{aligned}
&\mathrm{vec}(d\boldsymbol{U}) = d\mathrm{vec}(\boldsymbol{U}) \\
&d|\boldsymbol{U}| = |\boldsymbol{U}|\mathrm{tr}(\boldsymbol{U}^{-1}d\boldsymbol{U}) \\
&d\boldsymbol{U}^{-1} = -\boldsymbol{U}^{-1}(d\boldsymbol{U})\boldsymbol{U}^{-1} \\
&d(\mathrm{tr}(\boldsymbol{U})) = \mathrm{tr}(d\boldsymbol{U}) \\
&d(\boldsymbol{U} \odot \boldsymbol{V}) = (d\boldsymbol{U}) \odot \boldsymbol{V} + \boldsymbol{U} \odot (d\boldsymbol{V}) \\
&d(\log|\boldsymbol{U}|) = \mathrm{tr}(\boldsymbol{U}^{-1})d\boldsymbol{U} \\
&d(\boldsymbol{U}\boldsymbol{V}) = (d\boldsymbol{U})\boldsymbol{V} + \boldsymbol{U}(d\boldsymbol{V})
\end{aligned}$$

1.10 Special functions

Definition 1.10.1. The integral $\mathcal{F}$ is a non-analytic integral family, defined as

$$\mathcal{F}(p,q,r,s,t) \equiv \int_s^t x^p \exp\left[q\left\{\log(x/2) - \log\Gamma(x/2)\right\} - \tfrac{1}{2}rx\right] dx, \quad p \geq 0,\ q,r,s,t > 0.$$

Definition 1.10.2. The logit$(\cdot)$ function is defined by

$$\mathrm{logit}(p) = \log\left(\frac{p}{1-p}\right) = \log(p) - \log(1-p), \quad 0 \leq p \leq 1.$$

Definition 1.10.3. The digamma function is denoted by $\psi(\cdot)$ and is the derivative of the logarithm of the gamma function

$$\psi(x) = \frac{d}{dx}\ln\Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}, \quad x > 0.$$

Definition 1.10.4. The duplication matrix $\boldsymbol{D}_d$ is an unique $d^2 \times \frac{1}{2}d(d+1)$ matrix of

zeros and one that transforms any $d \times d$ symmetric matrix $\boldsymbol{A}$ from $\text{vech}(\boldsymbol{A})$ into $\text{vec}(\boldsymbol{A})$:

$$\boldsymbol{D}_d \,\text{vech}(\boldsymbol{A}) = \text{vec}(\boldsymbol{A}) \quad \text{for} \quad A = A^{\mathsf{T}}.$$

Definition 1.10.5. The Moore-Penrose Inverse of $\boldsymbol{D}_d$, defined as

$$\boldsymbol{D}_d^{+} \equiv (\boldsymbol{D}_d^T \, \boldsymbol{D}_d)^{-1} \boldsymbol{D}_d^T,$$

is an unique $\frac{1}{2}\, d(d+1) \times d^2$ matrix that transforms any $d \times d$ symmetric matrix $\boldsymbol{A}$ from $\text{vec}(\boldsymbol{A})$ into $\text{vech}(\boldsymbol{A})$:

$$\boldsymbol{D}_d^{+} \,\text{vec}(\boldsymbol{A}) = \text{vech}(\boldsymbol{A}).$$

Definition 1.10.6. The two identities involving vec and vech are:

$$\text{vec}^{-1}(\text{vec}(\boldsymbol{A})) = \boldsymbol{A} \quad \text{and} \quad \text{vec}(\text{vec}^{-1}(\boldsymbol{a})) = \boldsymbol{a},$$

where $\boldsymbol{A}$ is a $d \times d$ matrix, $\boldsymbol{a}$ is a $d^2 \times 1$ vector and $\text{vec}^{-1}(\boldsymbol{a})$ is a $d \times d$ matrix obtained from unstacking the entries of a in a column-wise fashion.

1.11 Matrix results

Result 1.11.1. If $\boldsymbol{A}$ is a square matrix, then $\text{tr}(\boldsymbol{A}^{\mathsf{T}}) = \text{tr}(\boldsymbol{A})$.

Result 1.11.2. If $\boldsymbol{A}$ is a $m \times n$ matrix and $\boldsymbol{B}$ is a $n \times m$ matrix, then $\text{tr}(\boldsymbol{AB}) = \text{tr}(\boldsymbol{BA})$.

Result 1.11.3. If $\boldsymbol{A}$ and $\boldsymbol{B}$ are $m \times n$ matrices, then $\text{tr}(\boldsymbol{A}^{\mathsf{T}}\boldsymbol{B}) = \text{vec}(\boldsymbol{A})^{\mathsf{T}}\text{vec}(\boldsymbol{B})$.

Result 1.11.4. If $\boldsymbol{a}$ and $\boldsymbol{b}$ are vectors of the same length, then

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

Result 1.11.5. Let $\boldsymbol{A}$ be a symmetric invertible matrix and $\boldsymbol{x}$ and $\boldsymbol{b}$ be column vectors with the same number of rows as $\boldsymbol{A}$. Then

$$\tfrac{1}{2}\boldsymbol{x}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^{\mathsf{T}}\boldsymbol{x} = -\tfrac{1}{2}\,(\boldsymbol{x} - \boldsymbol{A}^{-1}\boldsymbol{b})^{\mathsf{T}}\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{A}^{-1}\boldsymbol{b}) + \tfrac{1}{2}\,\boldsymbol{b}^{\mathsf{T}}\boldsymbol{A}^{-1}\boldsymbol{b}.$$

Result 1.11.6. Let $\boldsymbol{v}$ be a random vector. Then

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

Result 1.11.7. Let $\boldsymbol{v}$ be a random vector and let $\boldsymbol{A}$ be a fixed matrix with the same

number of rows as $\boldsymbol{v}$. Then

$$E(\boldsymbol{v}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{v}) = E(\boldsymbol{v})\boldsymbol{A}E(\boldsymbol{v})^{\mathsf{T}} + \operatorname{tr}\{\boldsymbol{A}\operatorname{Cov}(\boldsymbol{v})\}.$$

Result 1.11.8. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be $n \times 1$ random vectors such that $\boldsymbol{x}$ is conditioned on $\boldsymbol{y}$. Then,

$$\operatorname{Cov}(\boldsymbol{y}) = E\{\operatorname{Cov}(\boldsymbol{y}|\boldsymbol{x})\} + \operatorname{Cov}\{E(\boldsymbol{y}|\boldsymbol{x})\}.$$

1.12 Statistical software

Many longitudinal and multilevel models can be couched in the framework of Bayesian hierarchical models and thus have natural graphical model representations. However, the posterior density functions arise in these types of models are often difficult to obtain analytically. Hence one relies on taking advantage of the growing body of software for approximate inference in general graphical models.

In recent years, graphical model-based Bayesian inference engines have emerged for facilitating such approximate inference for general classes of parametric and nonparametric regression models (e.g. Luts, 2015; Marley and Wand, 2010). Examples include BUGS (Bayesian inference using Gibbs Sampling) (Ligges *et al.*, 2009), Stan (Stan Development Team, 2015), Infer.NET (Minka *et al.*, 2009) and VIBES (Variational Inference for Bayesian Networks)(Bishop *et al.*, 2002). Each engine offers a different form of the approximate inference. BUGS assumes the model is specified in a directed acyclic graph form and uses relatively slow but highly accurate MCMC Gibbs sampling for inference. Stan is a probabilistic programming language written in C++ and uses the Hamiltonian Monte Carlo No U-turn sampling for implementing full Bayesian statistical inference. Both engines are easy to use, especially since they can be directly called from the R computing environment via the wrapper packages BRugs (Ligges *et al.*, 2009) and rstan (Stan Development Team, 2015) respectively. The other two inference engines are designed for performing variational inference using graphical models, but suffer from versatility limitations. Infer.NET uses faster, but less accurate, deterministic approximations such as mean field variational Bayes and expectation propagation. VIBES is open-source Java and is similar to BUGS in terms of functionality. It also uses variational algorithms for inference, but is limited to the conjugate exponential family. Throughout this thesis, we use `rstan` for MCMC sampling.

Chapter 2

Mean Field Variational Bayes Approximations for Longitudinal and Multilevel Data Analysis

An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.

John Tukey

2.1 Introduction

The past decades have seen the emergence of machine learning and data mining tools designed to help capture and understand patterns from large datasets with complex structure. The most common data types are the longitudinal and multilevel data, which have a grouped or hierarchical structure and are frequently seen in many applied areas such as education, epidemiology, medicine and social science. Examples include human growth studies in which individuals' body measurements are collected at multiple follow-up times, medical studies in which patients are grouped within hospitals and hospital-specific disease rates are examined, and sample surveys in which respondents to questionnaires are grouped within households and geographical districts. These data structures give rise to correlations among observations within groups, therefore sophisticated statistical models are required to analyse these data taking into account correlations induced in the data. A

The main content of this chapter is published as: Lee, C. Y. Y. and Wand, M. P. (2015). Streamlined mean field variational Bayes for longitudinal and multilevel data analysis. *Biometrical Journal article in press.* This research has been accepted for presentation at four conferences.

popular model is the *generalised linear mixed model* (GLMM), which directly acknowledges multiple levels of dependency and accounts for different response types.

GLMMs extend generalised linear models by adding normal random effects on the linear predictor scale, to account for correlated observations within groups. However, this flexibility is traded with analytical tractability and may lead to negative implications on computational complexity and efficiency. In the frequentist paradigm, estimation of GLMMs using maximum likelihood is challenging since the integral over the random effects is intractable. Various approximate methods to such problems have been developed in the past two decades, including a seminal approach first introduced by Breslow and Clayton (1993), namely penalised quasi-likelihood, Laplace approximation and its extension (e.g. Raudenbush *et al.*, 2000), and Gaussian variational approximation (e.g. Ormerod and Wand, 2012). In the Bayesian paradigm, an early work by Zeger and Karim (1991) describe approximate Gibbs sampling for GLMMs, using normal distributions to approximate the nonstandard conditional distributions. Fong *et al.* (2009) propose the integrated nested Laplace approximation as a computationally convenient alternative to MCMC. Nonetheless, many of the aforementioned methods involving numerical quadrature or MCMC techniques to approximate intractable integrals suffer from computational intensity.

In this chapter, we examine the use of variational approximations, in particular MFVB, for fitting and inference in Bayesian GLMMs, with an emphasis on the Gaussian, Student-t, Bernoulli and Poisson responses. Variational methods (e.g. Bishop, 2006; Ormerod and Wand, 2010) have evolved mainly in areas such as machine learning and pattern recognition. However, since the early 2000s there has been an increasing recognition of their usefulness in mainstream statistics (e.g. Titterington, 2004). An early contribution is Teschendorff *et al.* (2005), who apply MFVB to mixture modelling for gene expression data. Longitudinal and multilevel models (Diggle *et al.*, 2002; Fitzmaurice *et al.*, 2012; Gelman and Hill, 2007; Goldstein, 2011) represent a major branch of statistics which, to date, has had a relatively little overlap with the graphical models viewpoint and variational methods. Recall from (1.7) that MFVB approximations arise upon restricting $q(\boldsymbol{\theta})$ to some class of density functions as follows:

$$\mathcal{Q} = \left\{ q(\boldsymbol{\theta}) \,:\, q(\boldsymbol{\theta}) = \prod_{i=1}^{M} q_i(\boldsymbol{\theta}_i) \text{ for some partition } \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\} \text{ of } \boldsymbol{\theta} \right\}.$$

The class of density functions $\mathcal{Q}$ can either be parametric, semiparametric or nonparametric. Parametric MFVB involves setting $\mathcal{Q}$ to be a parametric family of density functions. A recent contribution of this type is Opper and Archambeau (2009) and Challis and Barber (2013) who use the Gaussian families. Semiparametric MFVB involves the relaxation of the ordinary mean-field restriction in which some of the density functions in the postu-

lated product density form are pre-specified to be particular parametric density functions (Rhode and Wand, 2015). Nonparametric MFVB means that there is no parametric specification at all on $\mathcal{Q}$ and the product restriction is the only pre-specification being made. The parametric and semiparametric approaches to restricting $q(\boldsymbol{\theta})$ are the focus of this chapter.

This chapter makes three major contributions. First, we show, step by step, how to implement MFVB algorithms for Bayesian GLMMs with four common response distributions and their semiparametric extensions. Second, we empirically assess the performance of MFVB algorithms, benchmarking against the standard MCMC methods, through a series of comprehensive simulations. Finally, we derive a variational lower bound for each model that can be useful for model selection. Ultimately, we demonstrate that variational approximations make feasible inferences and estimation of models that would be difficult or impossible to estimate using standard MCMC methods.

Section 2.2 introduces longitudinal and multilevel models and shares a graph theoretical viewpoint of these models, focusing on random intercepts models, random intercepts and slopes models, and their semiparametric regression extensions. Details on the direct implementation of variational algorithms and inference for the four response models are given in Sections 2.5-2.9. In Section 2.10 we provide numerical evidence of the efficacy of variational methods, in terms of both inferential accuracy, credible interval coverage and computational speed. The concluding remarks are given in Section 2.11.

2.2 A brief introduction to longitudinal and multilevel models

Longitudinal and multilevel data are common in many applied areas such as education, medicine, epidemiology and social science. Such data have a complex grouped structure. An example of this is depicted in Figure 2.1. The data consist of mathematics test scores of 728 students from 48 schools in inner London, United Kingdom. Details of the data are described in Goldstein (2010).

In any complex structure we can identify atomic units that are at the lowest level of the system. For multilevel data, these atomic units are often individuals. Individuals are then grouped into *higher-level* units, commonly referred to as "groups" or "clusters". By convention we then say that the individuals are at Level 1 and the higher-level units are at Level 2 in our structure. For longitudinal data, repeat measurements on different occasions are at Level 1 and the individuals are at Level 2.

When individuals form groups or clusters, we might expect that two randomly selected individuals from the same group will be more alike than two individuals selected from different groups. For example, patients in the same hospital share common risk profiles

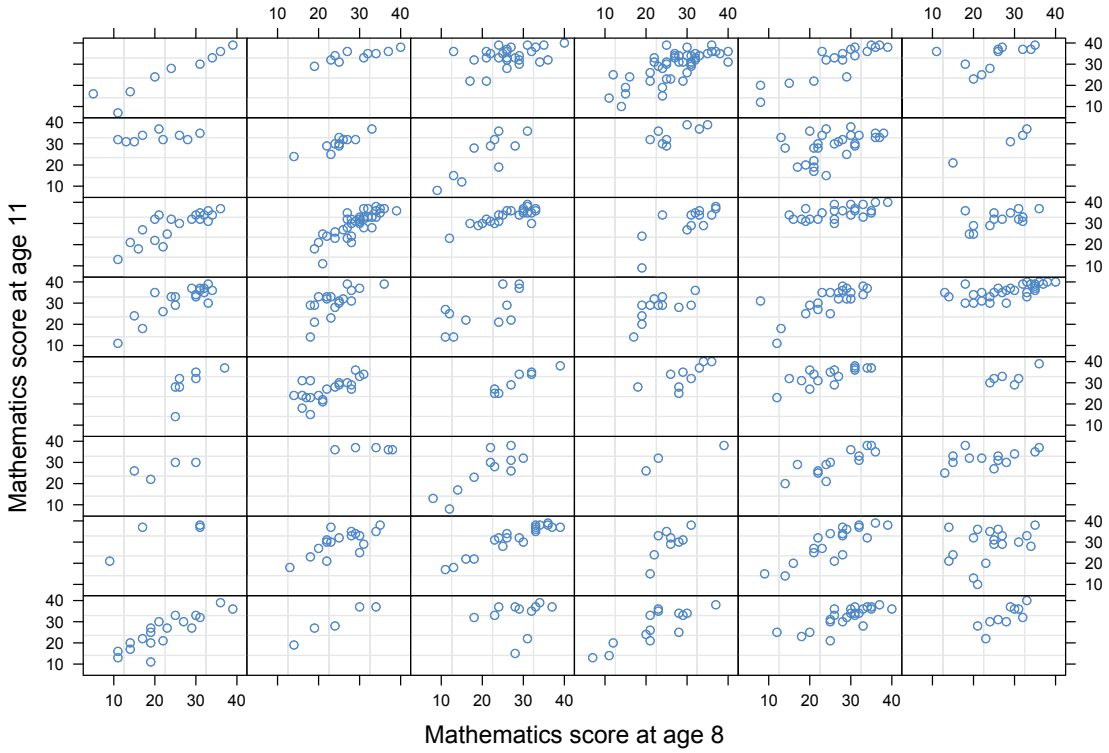

Figure 2.1: Panel of scatterplots of the 11-year old mathematics test scores versus the eight-year old test scores. Each panel corresponds to a school. Details of the data are described in Goldstein (2010).

and hospital characteristics, such as clinical management and practice, are likely to have a similar clinical outcome. By similar argument, measurements taken on the same individual over time will be more highly correlated than two measurements from different individuals. Such correlation violates the independence assumption of the classic regression model and therefore requires a more sophisticated model to appropriately take the induced correlation into account, namely *longitudinal or multilevel models*. Longitudinal or multilevel models are also known as hierarchical models, mixed models and variance components models. For the sake of consistency, we use the term *mixed models* throughout this thesis to represent any general classes of longitudinal and multilevel models, unless stated otherwise. In addition, we restrict our attention to data with a *nested hierarchical* structure, where lower-level unit nests in one and only one higher-level unit. However, it would be relatively straightforward to extend our methods below to include non-nested random effects models.

2.2.1 Graph theoretical viewpoint of longitudinal and multilevel models

Graphical models have become increasingly viewed as a general Bayesian inference engine. In many instances, sophisticated probabilistic models are embedded into the framework of graphical models in order to simplify statistical inference and improve computational efficiency. Graphical models are a great way to represent conditional independence assumptions by using graphs. Specifically, nodes represent random variables and lack of edges represent conditional independencies. In addition, the graph is a useful visual representation for complex stochastic systems. The graphical structure is also the basis of efficient inference algorithms.

2.2.2 Random intercepts models

Figure 2.1 is a frequently cited example in education, where students are grouped or clustered within schools. A natural starting point to analysing such a data structure is to assume that the overall mean and deviation of the ith school from that overall mean are simply straight lines. This leads to the standard random intercept model (Laird and Ware, 1982). Consider the Bayesian version of a two-level Gaussian mixed model with a random intercept for each school, and with simplicity in mind, we confine the model to the case of a single predictor, say $\boldsymbol{x}$,

$$\begin{aligned} y_{ij}|\beta_0, \beta_x, u_i^{\mathrm{R}}, \sigma_\varepsilon^2 &\overset{\text{ind.}}{\sim} N(\beta_0 + \beta_x\, x_{ij} + u_i^{\mathrm{R}},\, \sigma_\varepsilon^2) \\ \beta_0 &\sim N(\mu_{\beta_0},\, \sigma_{\beta_0}^2), \quad \beta_x \sim N(\mu_{\beta_x},\, \sigma_{\beta_x}^2), \\ u_i^{\mathrm{R}}|\sigma_{\mathrm{R}}^2 &\overset{\text{ind.}}{\sim} N(0,\, \sigma_{\mathrm{R}}^2), \\ \sigma_{\mathrm{R}}^2 &\sim \text{Inverse-Gamma}(A_{\mathrm{R}},\, B_{\mathrm{R}}) \\ \text{and} \quad \sigma_\varepsilon^2 &\sim \text{Inverse-Gamma}(A,\, B), \quad 1 \le j \le n_i,\, 1 \le i \le m. \end{aligned} \tag{2.1}$$

The letters i, j denote the *nested* indexing for the group level (Level 2) and unit level (Level 1) respectively. Further, y_{ij} denotes the jth response at Level 1 within the ith group at Level 2, x_{ij} denotes the jth predictor within the ith group, β_0 and β_x are the so-called *fixed* intercept and slope respectively, and u_i^{R} is the so-called *random* intercept for each group. The *mixed model* terminology is due to the fact that model (2.1) is a statistical model containing both *fixed effects* and *random effects*. Here we use the random intercepts to model the dependence of the ys within higher levels. It is also reasonable to assume that the level parameters are independent of each other and each of them is independent of the error terms ε_{ij}.

Recall from Section 1.4 that DAGs provide a useful "road map" of the model's structure and aid the algebra required for variational approximations. A DAG representation of model (2.1) is depicted in Figure 2.2. In graphical model terminology, $\mathcal{E} = \{y_{11}, \ldots, y_{mn_m}\}$

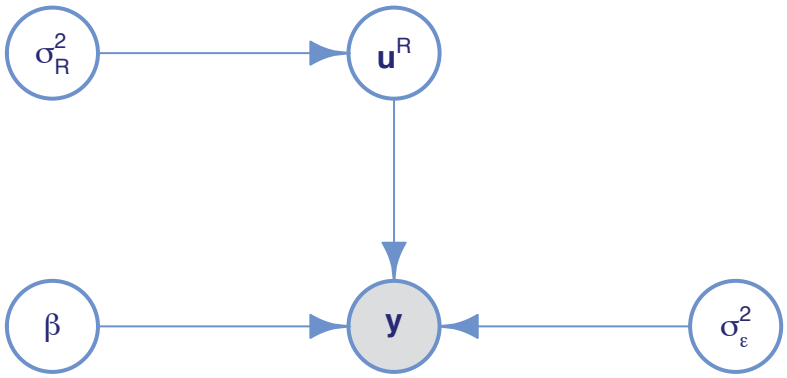

Figure 2.2: Directed acyclic graph corresponding to model (2.1). The open circles represent the random nodes. The shaded circle represents the observed node. The vector $\boldsymbol{y} = (y_{11}, \ldots, y_{mn_m})$ is represented by a single node, and β_0 and β_x are represented by the $\boldsymbol{\beta}$ node. Similarly, $\boldsymbol{u}^{\mathrm{R}} = (u_1^{\mathrm{R}} \ \ldots, u_m^{\mathrm{R}})$.

is the evidence node and $\mathcal{H} = \left\{\beta_0, \beta_x, u_1^{\mathrm{R}} \ \ldots, u_m^{\mathrm{R}}, \sigma^2_{\beta_0}, \sigma^2_{\beta_x}, \sigma^2_{\mathrm{R}}, \sigma^2_{\varepsilon}\right\}$ is the set of hidden nodes. Bayesian inference relies upon

$$\begin{aligned} p(\mathcal{H}|\mathcal{E}) &= p(\beta_0, \beta_x, u_1^{\mathrm{R}} \ \ldots, u_m^{\mathrm{R}}, \sigma^2_{\beta_0}, \sigma^2_{\beta_0}, \sigma^2_{\mathrm{R}}, \sigma^2_{\varepsilon}|y_{11}, \ldots, y_{mn_m}) \\ &= \frac{p(\mathcal{H}, \mathcal{E})}{\mathcal{E}} = \frac{p(\beta_0, \beta_x, u_1^{\mathrm{R}} \ \ldots, u_m^{\mathrm{R}}, \sigma^2_{\beta_0}, \sigma^2_{\beta_x}, \sigma^2_{\mathrm{R}}, \sigma^2_{\varepsilon}, y_{11}, \ldots, y_{mn_m})}{p(y_{11}, \ldots, y_{mn_m})}, \end{aligned}$$

the posterior density function of all of the random nodes given the observed node. Quite often we are stuck with intractable integrals arising from integrating over the variance components, and therefore obtaining the posterior density function via direct calculation becomes infeasible.

2.2.3 Random intercepts and slopes models

Model (2.1) can be easily extended to include both a random intercept and random slope for each school as follows:

$$\begin{aligned} y_{ij}|\beta_0, \beta_x, u_{0i}^{\mathrm{R}}, u_{1i}^{\mathrm{R}}, \sigma^2_{\varepsilon} &\overset{\text{ind.}}{\sim} N(\beta_0 + \beta_x\, x_{ij} + u_{0i}^{\mathrm{R}} + u_{1i}^{\mathrm{R}}\, x_{ij},\, \sigma^2_{\varepsilon}), \\ \beta_0 &\sim N(\mu_{\beta_0},\, \sigma^2_{\beta_0}), \quad \beta_x \sim N(\mu_{\beta_x},\, \sigma^2_{\beta_x}), \\ \begin{bmatrix} u_{0i}^{\mathrm{R}} & u_{1i}^{\mathrm{R}} \end{bmatrix}^{\mathsf{T}} |\boldsymbol{\Sigma}^{\mathrm{R}} &\overset{\text{ind.}}{\sim} N\left(\boldsymbol{0}, \boldsymbol{\Sigma}^{\mathrm{R}} \equiv \begin{bmatrix} \sigma^2_{R_0} & \sigma_{R_0,R_1} \\ \sigma_{R_0,R_1} & \sigma^2_{R_1} \end{bmatrix}\right), \\ \boldsymbol{\Sigma}^{\mathrm{R}} &\sim \text{Inverse-Wishart}(A_{\Sigma}, \boldsymbol{B}_{\Sigma}) \\ \text{and} \quad \sigma^2_{\varepsilon} &\sim \text{Inverse-Gamma}(A,\, B), \quad 1 \leq j \leq n_i,\, 1 \leq i \leq m. \end{aligned} \tag{2.2}$$

The majority of the parameters are defined analogously in Subsection 2.2.2. The terms u^{R}_{0i} and u^{R}_{1i} are the so-called random intercepts and random slopes, being treated as a random sample from a bivariate normal distribution with an unstructured 2×2 covariance matrix $\boldsymbol{\Sigma}^{\mathrm{R}}$. This model accounts for possible variability in the group intercepts $\sigma^2_{R_0}$ and slopes $\sigma^2_{R_1}$ and allows for an intercept-slope correlation ρ_{R_0,R_1}.

2.2.4 Extension to semiparametric regression

Semiparametric regression models extend generalised linear mixed models to accommodate non-linear predictor effects. The essence of semiparametric regression is to include penalised spline basis functions through random effects formulation that have the same form as those used traditionally in longitudinal and multilevel data analysis (e.g. Ruppert *et al.*, 2003).

Here we extend model (2.2) by simply adding a smooth, but otherwise unspecified, function for the ℓth predictor, say s_ℓ, $1 \leq \ell \leq L$,

$$f_\ell(s_\ell) = \beta_\ell s_\ell + \sum_{k=1}^{q^{\mathrm{G}}_\ell} u^{\mathrm{G}}_{\ell k}\, z_{\ell k}(s_\ell), \quad 1 \leq \ell \leq L, \tag{2.3}$$

where $z_{\ell 1}(\cdot), \ldots, z_{\ell q^{\mathrm{G}}_\ell}(\cdot)$ is a set of spline functions appropriate for the linear component being unpenalised. Our default choice for the $z_\ell(\cdot)$ is the B-spline basis and penalty set-up of O'Sullivan (1986), since this leads to approximate smoothing splines which have good boundary and extrapolation properties. The quantity of spline basis functions has a minimal effect on the adequacy of (2.3) and $q^{\mathrm{G}}_\ell = 25$ is recommended for models of practical interest (Li and Ruppert, 2008). The coefficients $u^{\mathrm{G}}_{\ell k}$ can be considered as a measure of the basis amplitude since they regulate the contribution from the ℓth curve.

From a computational point of view, penalised splines are equivalent to the random effects, in that we penalise the spline basis function coefficients by treating them as a random sample from a multivariate normal distribution to avoid overfitting of the data. That is,

$$u^{\mathrm{G}}_{11}, \ldots, u^{\mathrm{G}}_{Lq^{\mathrm{G}}_L} | \sigma^2_{u1}, \ldots, \sigma^2_{uL} \sim N\left(\mathbf{0}, \underset{1\leq\ell\leq L}{\mathrm{blockdiag}}(\sigma^2_{u\ell} \boldsymbol{I}_{q^{\mathrm{G}}_\ell})\right).$$

2.3 Mixed model representation

Classical mixed models can be viewed as an intermediate step between frequentist and Bayesian models since the grouped effects are treated as random. Bayesian analysis treats all model parameters as random, assigns prior distributions to characterise our prior knowledge about these parameters prior to data collection, and computes the joint posterior

density function of all of the random parameters given the observed data as the basis of inference.

In common practice, we generalise the fixed and random components of the mixed models to arbitrary general design matrices (Zhao *et al.*, 2006). This allows one to take advantage of the ever-expanding methods and software for inference in these models. Henceforth, we combine models (2.2) and (2.3) in matrix notation as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon}, \tag{2.4}$$

where

$$\begin{aligned}
\boldsymbol{y} &\equiv \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_m \end{bmatrix}, \quad \boldsymbol{X}_i \equiv \begin{bmatrix} \mathbf{1} & \boldsymbol{x}_i \end{bmatrix}, \quad \boldsymbol{X} \equiv \begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{s}_1 \\ \vdots & \vdots \\ \boldsymbol{X}_m & \boldsymbol{s}_m \end{bmatrix}, \quad \boldsymbol{\varepsilon} \equiv \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_m \end{bmatrix}, \\
\boldsymbol{Z} &\equiv \begin{bmatrix} z_{11}(\boldsymbol{s}_{11}) & \dots & z_{1q_1^{\mathrm{G}}}(\boldsymbol{s}_{11}) & \dots & z_{L1}(\boldsymbol{s}_{L1}) & \dots & z_{Lq_L^{\mathrm{G}}}(\boldsymbol{s}_{L1}) & \boldsymbol{X}_1 & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots & \dots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ z_{11}(\boldsymbol{s}_{1m}) & \dots & z_{1q_1^{\mathrm{G}}}(\boldsymbol{s}_{1m}) & \dots & z_{L1}(\boldsymbol{s}_{Lm}) & \dots & z_{Lq_L^{\mathrm{G}}}(\boldsymbol{s}_{Lm}) & \mathbf{0} & \dots & \boldsymbol{X}_m \end{bmatrix}, \\
\boldsymbol{\beta} &\equiv \begin{bmatrix} \beta_0 & \beta_x & \beta_{s1} & \cdots & \beta_{sL} \end{bmatrix}^{\mathsf{T}} \quad \text{and} \quad \boldsymbol{u} \equiv \begin{bmatrix} \boldsymbol{u}^{\mathrm{G}} \\ \boldsymbol{u}_1^{\mathrm{R}} \\ \vdots \\ \boldsymbol{u}_m^{\mathrm{R}} \end{bmatrix}.
\end{aligned}$$

For the ith group, $\boldsymbol{y}_i$ is an $n_i \times 1$ vector of response variables, $\boldsymbol{x}_i$ and $\boldsymbol{s}_i$ are $n_i \times 1$ vectors of predictor variables, $\boldsymbol{\varepsilon}_i$ is an $n_i \times 1$ vector of errors and $\boldsymbol{u}_i^{\mathrm{R}}$ is a $q^{\mathrm{R}} \times 1$ vector comprising the random intercept and random slope of the ith group. The $\sum_{\ell=1}^{L} q_\ell^{\mathrm{G}} \times 1$ vector of spline basis coefficients is denoted by $\boldsymbol{u}^{\mathrm{G}} \equiv [\boldsymbol{u}_1^{\mathrm{G}} \ \dots \ \boldsymbol{u}_L^{\mathrm{G}}]$. The matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the respective $\sum_{i=1}^{m} n_i \times p$ *fixed effects design matrix* and $\sum_{i=1}^{m} n_i \times q$ *random effects design matrix*, $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are the $p \times 1$ fixed effects vector and $q \times 1$ random effects vector. The random effects have a multivariate normal distribution with zero mean and covariance matrix $\boldsymbol{G}$, given as

$$\boldsymbol{G} = \begin{bmatrix} \underset{1 \le \ell \le L}{\mathrm{blockdiag}}(\sigma_{u\ell}^2 \boldsymbol{I}_{q_\ell^{\mathrm{G}}}) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_m \otimes (\boldsymbol{\Sigma}^{\mathrm{R}}) \end{bmatrix}.$$

Putting all of these components together gives the following *linear mixed model* form

$$\boldsymbol{y} | \, \boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2 \sim N(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \sigma_\varepsilon^2 \boldsymbol{I}), \quad \boldsymbol{u} | \, \boldsymbol{G} \sim N(\mathbf{0}, \boldsymbol{G}). \tag{2.5}$$

A common extension of (2.5) involves replacement of $\sigma_\varepsilon^2 \boldsymbol{I}$ by a general covariance matrix $\boldsymbol{R}$, but is not treated here.

A very wide range of models can be obtained through various choices of $\boldsymbol{Z}$ and $\boldsymbol{G}$. In this chapter, we focus on those corresponding to two-level hierarchical structures, which are often treated separately within the broad branches of longitudinal data analysis (Diggle *et al.*, 2002; Fitzmaurice *et al.*, 2012) and multilevel modelling (Gelman and Hill, 2007; Goldstein, 2011). Since around 2000, there has been a major interplay between longitudinal data analysis and semiparametric regression. An overview is given in Fitzmaurice *et al.* (2008). As we shall see in subsequent chapters, the general model (2.4) is extremely rich in that it encompasses a wide range of models that are useful in practice.

The class of generalised linear mixed models is much more general than (2.5), as explained by Zhao *et al.* (2006). Staying within the one-parameter exponential family, a more general class of models is

$$\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u} \sim \exp\left\{\boldsymbol{y}^{\mathsf{T}}(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u}) - \boldsymbol{1}^{\mathsf{T}}b(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u}) + \boldsymbol{1}^{\mathsf{T}}c(\boldsymbol{y})\right\}, \quad \boldsymbol{u}|\boldsymbol{G} \sim N(\boldsymbol{0},\boldsymbol{G}), \quad (2.6)$$

where the design matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ and covariance matrix $\boldsymbol{G}$ are quite general and described above. The functions b and c are specific to members of the family. The most common examples are Bernoulli for which $b(x) \equiv \log(1+e^x)$ and $c(x) \equiv 0$ and Poisson for which $b(x) \equiv e^x$ and $c(x) \equiv -\log(x!)$.

2.3.1 Sample size and subscript notation

Many of the models used in longitudinal and multilevel data analysis are mathematically equivalent, but use different sample size and subscript notation. It is prudent to delineate the various notational conventions commonly in use so that methodology developed using one set of notation can be transferred to models that use other notational systems. We achieve that in this Subsection by way of a concrete example of (2.5).

The special case of (2.5) with

$$\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0}, \quad \boldsymbol{Z} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{G} = \sigma^2_{\mathrm{R}}\boldsymbol{I} \quad (2.7)$$

imposes the following covariance matrix on the response vector:

$$\mathrm{Cov}(\boldsymbol{y}) = \begin{bmatrix} \sigma^2_{\mathrm{R}}+\sigma^2_{\varepsilon} & \sigma^2_{\mathrm{R}} & 0 & 0 & 0 & 0 & 0 \\ \sigma^2_{\mathrm{R}} & \sigma^2_{\mathrm{R}}+\sigma^2_{\varepsilon} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2_{\mathrm{R}}+\sigma^2_{\varepsilon} & \sigma^2_{\mathrm{R}} & \sigma^2_{\mathrm{R}} & 0 & 0 \\ 0 & 0 & \sigma^2_{\mathrm{R}} & \sigma^2_{\mathrm{R}}+\sigma^2_{\varepsilon} & \sigma^2_{\mathrm{R}} & 0 & 0 \\ 0 & 0 & \sigma^2_{\mathrm{R}} & \sigma^2_{\mathrm{R}} & \sigma^2_{\mathrm{R}}+\sigma^2_{\varepsilon} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2_{\mathrm{R}}+\sigma^2_{\varepsilon} & \sigma^2_{\mathrm{R}} \\ 0 & 0 & 0 & 0 & 0 & \sigma^2_{\mathrm{R}} & \sigma^2_{\mathrm{R}}+\sigma^2_{\varepsilon} \end{bmatrix}.$$

This block covariance structure is in keeping with two hierarchical levels with 3 groups at the first level and sample sizes of 2, 3 and 2 observations within each group.

There are at least two established notations for conveying the hierarchical structure imposed by (2.7). The first is

$$E(y_{ij}|\, u^{\mathrm{R}}_i) = u^{\mathrm{R}}_i,$$

where

$$\begin{aligned} y_{ij} &\equiv j\text{th measurement in the } i\text{th group}, \\ u^{\mathrm{R}}_i &\equiv \text{random effect in the } i\text{th group}, \\ &\qquad \text{for } 1 \le i \le m,\ 1 \le j \le n_i. \end{aligned} \tag{2.8}$$

In the current example, $m = 3$ and $n_1 = 2, n_2 = 3, n_3 = 2$. This notation is used by prominent longitudinal data analysis textbooks, Diggle *et al.* (2002) and Fitzmaurice *et al.* (2012), as well as the semiparametric regression book, Ruppert *et al.* (2003).

An alternative notation is

$$E(y_i|u^{\mathrm{R}}_{j[i]}) = u^{\mathrm{R}}_{j[i]}, \quad 1 \le i \le 7, \quad 1 \le j \le J,$$

where for this example, $J = 3$ and the $j[i]$ mapping is

$$1[1] \equiv 1,\ 1[2] \equiv 1,\ 1[3] \equiv 1,\ 2[3] \equiv 2,\ 2[4] \equiv 2,\ 2[5] \equiv 2,\ 3[6] \equiv 3,\ 3[7] \equiv 3$$

and

$$y_i = i\text{th entry of } \boldsymbol{y}.$$

This is used in the multilevel model textbook by Gelman and Hill (2007). Goldstein (2011) uses

$$y_{ij} \equiv i\text{th measurement in the } j\text{th group}, \quad 1 \le j \le m$$

for two-level models with m groups.

For the remainder of this thesis we will use the sample size and subscript notation adopted by the above-mentioned longitudinal data analysis textbooks. This involves sub-

script notation corresponding to (2.8) and

$$m \equiv \text{number of groups} \quad \text{and} \quad n_i \equiv \text{number of response measurements in the } i\text{th group.}$$

The mathematical equivalence between longitudinal and multilevel models means that our methodology applies equally to both areas. However, as illustrated above, some notational adjustment is required to match common multilevel model notations.

2.4 Prior distributions for fixed and random effects

Throughout this thesis we take the prior distribution for the fixed effects vector $\boldsymbol{\beta}$, standard deviation parameters for the spline bases $\sigma_{u\ell}$ and random effects covariance matrix $\boldsymbol{\Sigma}^{\mathrm{R}}$ of the form:

$$\begin{aligned} \boldsymbol{\beta} &\sim N(\mathbf{0}, \sigma_\beta^2 \boldsymbol{I}_p), \quad \sigma_{u\ell} \overset{\text{ind.}}{\sim} \text{Half-Cauchy}(A_{u\ell}) \\ \text{and} \quad & \boldsymbol{\Sigma}^{\mathrm{R}} \sim \text{Inverse-Wishart}(A_\Sigma, \boldsymbol{B}_\Sigma), \end{aligned} \tag{2.9}$$

where the hyperparameters σ_β^2, $A_{u\ell} > 0$ are to be specified by the user. The standard deviation parameters have independent half-Cauchy priors. As explained in Gelman (2006), (2.9) is an attractive means by which weakly-informative priors can be imposed on the $\sigma_{u\ell}$s. Result 5 of Wand *et al.* (2012) leads to the following equivalent distributional statement:

$$\sigma_{u\ell}^2 \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2}, 1/a_{u\ell}), \quad a_{u\ell} \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2}, 1/A_{u_\ell}^2). \tag{2.10}$$

We use representation (2.10) rather than (2.9) since it is more conductive to MFVB.

For $\boldsymbol{\Sigma}^{\mathrm{R}}$ we use the following extension of (2.10) (Huang and Wand, 2013):

$$\begin{aligned} \boldsymbol{\Sigma}^{\mathrm{R}} | a_1^{\mathrm{R}}, \ldots, a_{q^{\mathrm{R}}}^{\mathrm{R}} &\sim \text{Inverse-Wishart}\left(\nu + q^{\mathrm{R}} - 1,\ 2\,\nu\,\text{diag}(1/a_1^{\mathrm{R}}, \ldots, 1/a_{q^{\mathrm{R}}}^{\mathrm{R}})\right), \\ a_r^{\mathrm{R}} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(\tfrac{1}{2}, 1/A_{\mathrm{R}r}^2\right), \quad r = 1, \ldots, q^{\mathrm{R}}, \end{aligned}$$

where $\nu, A_{\mathrm{R}r} > 0$. The choice $\nu = 2$ corresponds to the correlation parameters having uniform distributions over $(-1, 1)$ and the standard deviation parameters having Half-t distributions with 2 degrees of freedom.

For notation convenience, define the following parameter and matrices throughout this chapter as follows:

$$N \equiv \sum_{i=1}^{m} n_i, \quad \boldsymbol{C} \equiv \begin{bmatrix} \boldsymbol{X} & \boldsymbol{Z} \end{bmatrix}, \quad \boldsymbol{C}_{12} \equiv \underset{1 \le i \le N}{\text{diag}}\,(b_i) \quad \text{and} \quad \boldsymbol{v} \equiv \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix}.$$

The effects vector $\boldsymbol{v}$ has a multivariate normal distribution with zero mean and covariance

matrix $\boldsymbol{\Omega}$, given as

$$\boldsymbol{v}|\sigma_\beta^2, \sigma_{u1}^2, \ldots, \sigma_{uL}^2, \boldsymbol{\Sigma}^{\mathrm{R}} \sim N\left(\boldsymbol{0}, \begin{bmatrix} \sigma_\beta^{-2}\boldsymbol{I}_p & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \underset{1\le\ell\le L}{\mathrm{blockdiag}}(\sigma_{u\ell}^{-2}\boldsymbol{I}_{q_\ell^{\mathrm{G}}}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_m \otimes (\boldsymbol{\Sigma}^{\mathrm{R}})^{-1} \end{bmatrix}\right).$$

In what follows, we systematically present approximate Bayesian analysis of four common GLMMs in scientific research.

2.5 Gaussian semiparametric mixed models

We first consider the case where the distribution of response vector $\boldsymbol{y}$ is Gaussian, the two-level Bayesian Gaussian semiparametric mixed model stated in full is

$$\begin{aligned}
\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2 &\sim N(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \sigma_\varepsilon^2\boldsymbol{I}), \quad \boldsymbol{\beta} \sim N(0, \sigma_\beta^2\boldsymbol{I}_p), \\
\boldsymbol{u}|\sigma_{u1}^2, \ldots, \sigma_{uL}^2, \boldsymbol{\Sigma}^{\mathrm{R}} &\sim N\left(\boldsymbol{0}, \begin{bmatrix} \underset{1\le\ell\le L}{\mathrm{blockdiag}}(\sigma_{u\ell}^2\boldsymbol{I}_{q_\ell^{\mathrm{G}}}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_m \otimes \boldsymbol{\Sigma}^{\mathrm{R}} \end{bmatrix}\right), \\
\boldsymbol{\Sigma}^{\mathrm{R}}|a_1^{\mathrm{R}}, \ldots, a_{q^{\mathrm{R}}}^{\mathrm{R}} &\sim \text{Inverse-Wishart}\left(\nu + q^{\mathrm{R}} - 1,\ 2\,\nu\,\mathrm{diag}(1/a_1^{\mathrm{R}}, \ldots, 1/a_{q^{\mathrm{R}}}^{\mathrm{R}})\right), \\
a_r^{\mathrm{R}} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(\tfrac{1}{2}, 1/A_{\mathrm{R}r}^2\right), \quad r = 1, \ldots, q^{\mathrm{R}}, \\
\sigma_{u\ell}^2|a_{u\ell} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2},\ 1/a_{u\ell}), \quad a_{u\ell} \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2},\ 1/A_{u_\ell}^2), \\
\sigma_\varepsilon^2|a_\varepsilon &\sim \text{Inverse-Gamma}(\tfrac{1}{2},\ 1/a_\varepsilon), \quad a_\varepsilon \sim \text{Inverse-Gamma}(\tfrac{1}{2},\ 1/A_\varepsilon^2).
\end{aligned} \tag{2.11}$$

Figure 2.3 shows the graph-theoretic representation of model (2.11). The advantages to such a graphical representation are two-fold: first, it provides a visualisation of the hierarchical structure of its corresponding Bayesian model; and second, graph-theoretic results can be used to determine probabilistic relationships between nodes. The node $\boldsymbol{a}^{\mathrm{R}}$ corresponds to the random vector $[a_1^{\mathrm{R}} \ldots a_{q^{\mathrm{R}}}^{\mathrm{R}}]^{\mathsf{T}}$. The nodes $\boldsymbol{\sigma}_u^2$ and $\boldsymbol{a}_u$ are defined in a similar vein, i.e. $\boldsymbol{\sigma}_u^2 \equiv [\sigma_{u1}^2 \ldots \sigma_{uL}^2]^{\mathsf{T}}$ and $\boldsymbol{a}_u \equiv [a_{u1}^2, \ldots, a_{uL}^2]^{\mathsf{T}}$ respectively. The node $\boldsymbol{u}$ is separated into two nodes $\boldsymbol{u}^{\mathrm{R}}$ and $\boldsymbol{u}^{\mathrm{G}}$. Note that the random effects have been partitioned into spline coefficients (superscript G) and group effects (superscript R).

2.5.1 Approximate Bayesian inference via mean field variational Bayes

Consider the problem of Bayesian inference for the random parameters in model (2.11) as depicted in Figure 2.3. The process of MFVB approximation is best illustrated in Figure 2.5 using "life cycle of nodes" as a metaphor. We begin with the DAG that can be viewed as a family tree, where each directed edge signifies a "parent-child" relationship. Moralisation then follows, which involves "marrying" any unconnected parents of a common child. The resultant relationships are complex and often lead to an intractable problem. This is where

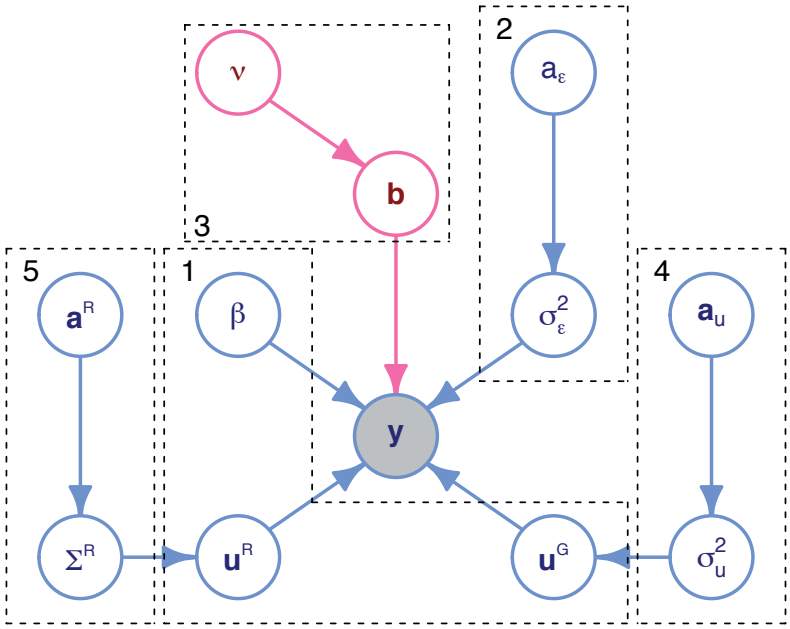

Figure 2.3: Directed acyclic graph for the two-level Bayesian semiparametric mixed models with the Gaussian and Student-t responses. The shaded node corresponds to the observed data vector and the open nodes correspond to the random or auxiliary variables. The pink open nodes $\boldsymbol{b}$ and ν are additional parameters for the Student-t response model. The numbered grey dashed polygons indicate the corresponding update expressions in Algorithm 3.

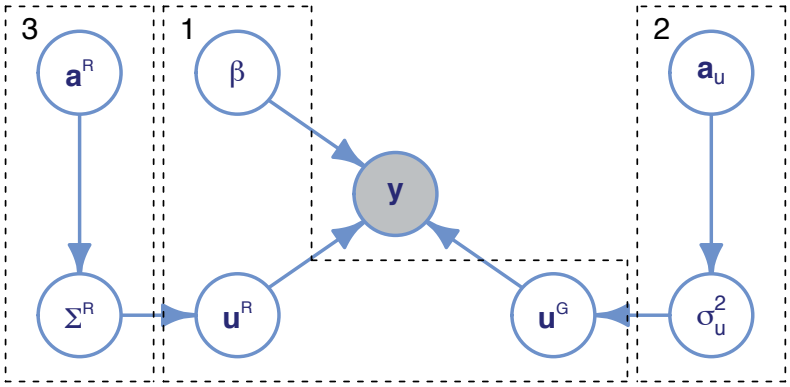

Figure 2.4: Directed acyclic graph for the two-level Bayesian semiparametric mixed models with the Bernoulli and Poisson responses. The shaded node corresponds to the observed data vector and the open node corresponds to the random or auxiliary variables. The numbered grey dashed polygons indicate the corresponding update expressions in Algorithms 4 and 6.

the MFVB approximation comes into play by "getting rid of some kids" and assumes posterior independence among parameters in order to achieve tractability.

To elaborate mathematically, the essence of this approach involves approximation of the full joint posterior density function of the form

$$p(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{a}^{\mathrm{R}}, \boldsymbol{a}_u, a_\varepsilon, \boldsymbol{\Sigma}^{\mathrm{R}}, \boldsymbol{\sigma}_u^2, \sigma_\varepsilon^2 | \boldsymbol{y}) \approx q(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{a}^{\mathrm{R}}, \boldsymbol{a}_u, a_\varepsilon, \boldsymbol{\Sigma}^{\mathrm{R}}, \boldsymbol{\sigma}_u^2, \sigma_\varepsilon^2)$$

subject to the q-density product restriction

$$q(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{a}^{\mathrm{R}}, \boldsymbol{a}_u, a_\varepsilon, \boldsymbol{\Sigma}^{\mathrm{R}}, \boldsymbol{\sigma}_u^2, \sigma_\varepsilon^2) = q(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{a}^{\mathrm{R}}, \boldsymbol{a}_u, a_\varepsilon) \times q(\boldsymbol{\Sigma}^{\mathrm{R}}, \boldsymbol{\sigma}_u^2, \sigma_\varepsilon^2). \tag{2.12}$$

Restriction (2.12) is the minimal factorisation in the MFVB approximation. Menictas & Wand (2013) provides some heuristic arguments for the restriction in (2.12). Essentially, such a restriction is underpinned by the well-known parameter orthogonality result arising from the block-diagonal form of the Fisher information matrix, which implies that the maximum likelihood estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ in linear regression models are asymptotically independent. Applying this result to the approximate Bayesian inference context leads to (2.12) and so we would expect this posterior independence assumption to be reasonable, and therefore, should incur little loss in accuracy.

The q-densities are chosen to minimise the Kullback-Leibler divergence between the full joint posterior density function and the right hand side of (2.12) (Wainwright and Jordan, 2008). Although the mild restriction (2.12) is sufficient to produce closed-form q-densities, *induced product results* (Bishop, 2006) lead to further factorisations in the product q-density form that simplify calculations. Such induced factorisations arise from an interaction between the factorisation assumed in the approximating posterior density function and the conditional independence properties of the actual joint density function, and can be easily detected using simple graphical tests such as moralisation. As an illustration, Figure 2.5b is the moral graph of model (2.11). Let $A = \sigma_\varepsilon^2$, $B = \boldsymbol{\Sigma}^{\mathrm{R}}$ and $C = \{\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{y}\}$, then the smallest ancestral set containing $A \cup B \cup C$ is the entire DAG. It is evident that all paths connecting nodes in A and B must pass through at least one of the nodes in C. In other words, the set of nodes in C blocks the paths from A to B and therefore the following conditional independence statement holds under the global Markov property:

$$A \perp\!\!\!\perp B \mid C, \quad \text{i.e.} \quad \sigma_\varepsilon^2 \perp\!\!\!\perp \boldsymbol{\Sigma}^{\mathrm{R}} \mid \{\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{y}\}.$$

Similarly, we can deduce from the global Markov property the conditional independences

$$\{\boldsymbol{\beta}, \boldsymbol{u}\} \perp\!\!\!\perp \boldsymbol{a}^{\mathrm{R}} \mid \{\boldsymbol{y}, \boldsymbol{\Sigma}^{\mathrm{R}}, \sigma_\varepsilon^2\} \quad \text{and} \quad \boldsymbol{a}^{\mathrm{R}} \perp\!\!\!\perp \boldsymbol{a}_u \mid \{\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\Sigma}^{\mathrm{R}}, \sigma_\varepsilon^2, \sigma_u^2\}.$$

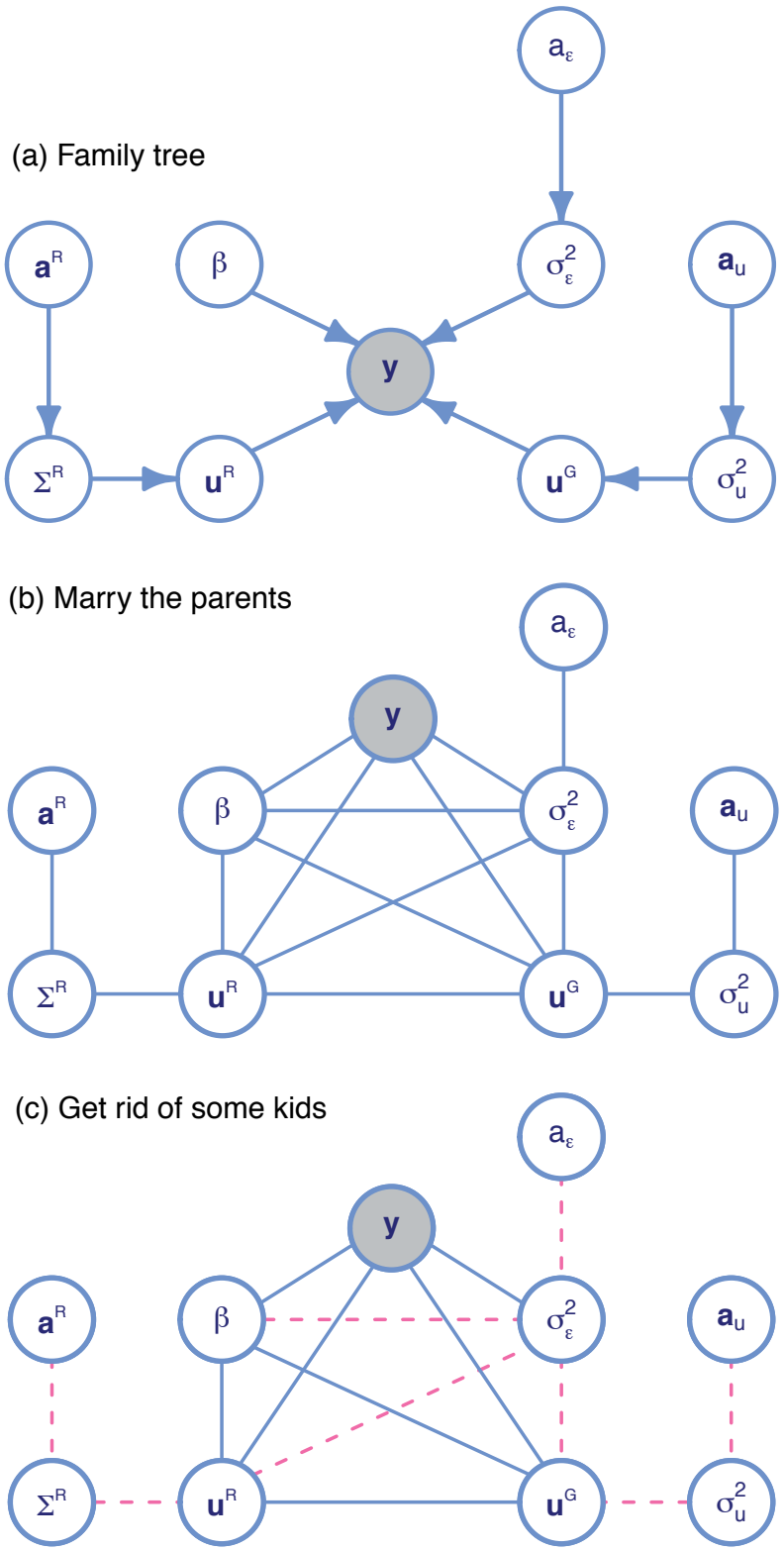

Figure 2.5: Life cycle of nodes for mean field variational Bayes: (a) directed acyclic graph for model (2.11), (b) moral graph for model (2.11), (c) modification of Figure 2.5b with eight edges removed to impose q-density product restriction. In each graph, shading is used to signify the observed data vector.

Similar arguments lead to the q-densities having the product form:

$$\begin{aligned} & q(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{a}^{\mathrm{R}}, \boldsymbol{a}_u, a_\varepsilon, \boldsymbol{\Sigma}^{\mathrm{R}}, \boldsymbol{\sigma}_u^2, \sigma_\varepsilon^2) \\ & = q(\boldsymbol{\beta}, \boldsymbol{u})\, q(a_\varepsilon)\, q(\boldsymbol{\Sigma}^{\mathrm{R}})\, q(\sigma_\varepsilon^2) \left\{ \prod_{r=1}^{q^{\mathrm{R}}} q(a_r^{\mathrm{R}}) \right\} \left\{ \prod_{\ell=1}^{L} q(\sigma_{u\ell}^2) \right\} \left\{ \prod_{\ell=1}^{L} q(a_{u\ell}) \right\}. \end{aligned} \tag{2.13}$$

Visually, this is analogous to removing edges between relevant nodes in the moralised DAG, represented by the pink dotted lines in Figure 2.5c. Under the product restriction (2.13), all parameters of model (2.11) have closed-form expressions for their full conditional density function:

$$\begin{aligned} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} \Bigg| \text{rest} &\sim N\left((\sigma_\varepsilon^{-2} \boldsymbol{C}^{\mathsf{T}} \boldsymbol{C} + \boldsymbol{\Omega}^{-1})^{-1} \sigma_\varepsilon^{-2} \boldsymbol{C}^{\mathsf{T}} \boldsymbol{y},\, (\sigma_\varepsilon^{-2} \boldsymbol{C}^{\mathsf{T}} \boldsymbol{C} + \boldsymbol{\Omega}^{-1})^{-1} \right), \\ \sigma_\varepsilon^2 | \text{rest} &\sim \text{Inverse-Gamma}\left(\tfrac{1}{2} \left(\textstyle\sum_{i=1}^m n_i + 1 \right),\, \tfrac{1}{2} \|\boldsymbol{y} - \boldsymbol{C}\boldsymbol{v}\|^2 + a_\varepsilon^{-1} \right), \\ a_\varepsilon | \text{rest} &\sim \text{Inverse-Gamma}\left(1,\, \sigma_\varepsilon^{-2} + A_\varepsilon^{-2} \right), \\ \sigma_{u\ell}^2 | \text{rest} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(\tfrac{1}{2} \left(q_\ell^{\mathrm{G}} + 1 \right),\, \tfrac{1}{2} \|\boldsymbol{u}_\ell^{\mathrm{G}}\|^2 + a_\ell^{-1} \right), \\ a_{u\ell} | \text{rest} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(1,\, \sigma_{u\ell}^{-2} + A_{u_\ell}^{-2} \right), \\ \boldsymbol{\Sigma}^{\mathrm{R}} | \text{rest} &\sim \text{Inverse-Wishart}\left(\nu + m + 1,\, \sum_{i=1}^m \{ u_i^{\mathrm{R}} (u_i^{\mathrm{R}})^{\mathsf{T}} \} + 2\,\nu\, \mathrm{diag}(1/a_1^{\mathrm{R}}, \ldots, 1/a_{q^{\mathrm{R}}}^{\mathrm{R}}) \right), \\ a_r^{\mathrm{R}} | \text{rest} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(\tfrac{1}{2}(\nu + q^{\mathrm{R}}),\, \nu (\boldsymbol{\Sigma}^{\mathrm{R}})_{rr}^{-1} + A_{\mathrm{R}r}^{-2} \right), \quad r = 1, \ldots, q^{\mathrm{R}}. \end{aligned}$$

Here "rest" denotes the set of other random variables in model (2.11) and $(\boldsymbol{\Sigma}^{\mathrm{R}})_{rr}^{-1}$ denotes the (r, r) entry of $(\boldsymbol{\Sigma}^{\mathrm{R}})^{-1}$. The full conditionals are of standard form, so MCMC methods can be used to construct a Markov chain that targets from the full conditionals that leads to samples from their joint posterior. However, we demonstrate fitting via MFVB approximation since it provides much faster computation than MCMC.

Recall from Section 1.5.3.1 that the optimal q-density for a parameter vector $\boldsymbol{\theta}$, denoted by $q^*(\boldsymbol{\theta})$, were shown to satisfy:

$$q_i^*(\boldsymbol{\theta}_i) \propto \exp\{ E_{q(-\boldsymbol{\theta}_i)} \log p(\boldsymbol{\theta}_i | \text{Markov blanket of } \boldsymbol{\theta}_i) \},$$

where $E_{q(-\boldsymbol{\theta}_i)}$ denotes expectation with respect to the q-densities of all parameters except $\boldsymbol{\theta}_i$. As shown through the derivations in Appendix 2.A, the optimal q-densities admit the following forms:

$$\begin{aligned} & q^*(\boldsymbol{\beta}, \boldsymbol{u}) \text{ is the } N(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})},\, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}) \text{ density function}, \\ & q^*(\sigma_\varepsilon^2) \text{ is the Inverse-Gamma}\left(\tfrac{1}{2} \left(\textstyle\sum_{i=1}^m n_i + 1 \right),\, B_{q(\sigma_\varepsilon^2)} \right) \text{ density function}, \\ & q^*(a_\varepsilon) \text{ is the Inverse-Gamma}(1,\, B_{q(a_\varepsilon)}) \text{ density function}, \end{aligned} \tag{2.14}$$

$q^*(\sigma^2_{u\ell})$ is the Inverse-Gamma $\left(\frac{1}{2}\left(q^{\mathrm{G}}_\ell + 1\right),\, B_{q(\sigma^2_{u\ell})}\right)$ density function,
$q^*(a_{u\ell})$ is the Inverse-Gamma$(1,\, B_{q(a_{u\ell})})$ density function,
$q^*(\boldsymbol{\Sigma}^{\mathrm{R}})$ is the Inverse-Wishart $\left(\nu + m + q^{\mathrm{R}} + 1,\, \boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\mathrm{R}})}\right)$ density function,
$q^*(a^{\mathrm{R}}_r)$ is the Inverse-Gamma $\left(\frac{1}{2}(\nu + q^{\mathrm{R}}),\, B_{q(a^{\mathrm{R}}_r)}\right)$ density function,

for parameters $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$, the mean vector and covariance matrix of $q^*(\boldsymbol{\beta},\boldsymbol{u})$, $B_{q(\sigma^2_\varepsilon)}$, the scale parameter of $q^*(\sigma^2_\varepsilon)$, $B_{q(a_\varepsilon)}$, the scale parameter of $q^*(a_\varepsilon)$, $B_{q(\sigma^2_{u\ell})}$, the scale parameter of $q^*(\sigma^2_{u\ell})$, $B_{q(a_{u\ell})}$, the scale parameter of $q^*(a_{u\ell})$, $B_{q(a^{\mathrm{R}})}$, the scale parameter of $q^*(a^{\mathrm{R}})$ and $\boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\mathrm{R}})}$, the scale parameter of $q^*(\boldsymbol{\Sigma}^{\mathrm{R}})$. The MFVB approximation involves standard density functions and requires only simple closed-form coordinate ascent updates, see Algorithm 3. Convergence of Algorithm 3 for the Gaussian response model can be monitored using successive values of the lower bound on the marginal log-likelihood:

$$
\begin{aligned}
\log \underline{p}(\boldsymbol{y}; q) \;=\;& \tfrac{1}{2}\, q^{\mathrm{R}}(\nu + q^{\mathrm{R}} - 1)\log(2\nu) - \tfrac{1}{2}\log(2\pi)\textstyle\sum_{i=1}^m n_i - \left(\tfrac{1}{2}q^{\mathrm{R}} + L + 1\right)\log(\pi) \\
& - \tfrac{1}{2}p\log(\sigma^2_\beta) - \tfrac{1}{2}\sigma^{-2}_\beta\left(\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\right) + \tfrac{1}{2}\left(\textstyle\sum_{\ell=1}^L q^{\mathrm{G}}_\ell + p + m\right) \\
& + \tfrac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}| - \log(\mathcal{C}_{q^{\mathrm{R}},\,\nu+q^{\mathrm{R}}-1}) + \log(\mathcal{C}_{q^{\mathrm{R}},\,\nu+m+q^{\mathrm{R}}-1}) \\
& - \tfrac{1}{2}\left(\nu + m + q^{\mathrm{R}} - 1\right)\log|\boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\mathrm{R}})}| \\
& + \textstyle\sum_{\ell=1}^L \log\Gamma\left\{\frac{1}{2}\left(q^{\mathrm{G}}_\ell + 1\right)\right\} - \frac{1}{2}\sum_{\ell=1}^L\left(q^{\mathrm{G}}_\ell + 1\right)\log(B_{q(\sigma^2_{u\ell})}) \\
& + \log\Gamma\left\{\tfrac{1}{2}(\textstyle\sum_{i=1}^m n_i + 1)\right\} - \tfrac{1}{2}\left(\textstyle\sum_{i=1}^m n_i + 1\right)\log\left(B_{q(\sigma^2_\varepsilon)}\right) \\
& - \textstyle\sum_{r=1}^{q^{\mathrm{R}}}\log(A_{\mathrm{R}r}) + q^{\mathrm{R}}\log\Gamma\left\{\frac{1}{2}(\nu + q^{\mathrm{R}})\right\} \\
& - \textstyle\sum_{\ell=1}^L\log(A_{u\ell}) - \sum_{\ell=1}^L\log\left(B_{q(a_{u\ell})}\right) + \sum_{\ell=1}^L \mu_{q(1/a_{u\ell})}\mu_{q(1/\sigma^2_{u\ell})} \\
& + \textstyle\sum_{r=1}^{q^{\mathrm{R}}}\nu(\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})})_{rr}\,\mu_{q(1/a^{\mathrm{R}}_r)} - \frac{1}{2}\left(\nu + q^{\mathrm{R}}\right)\sum_{r=1}^{q^{\mathrm{R}}}\log(B_{q(a^{\mathrm{R}}_r)}) \\
& - \log(A_\varepsilon) - \log\left(B_{q(a_\varepsilon)}\right) + \mu_{q(1/a_\varepsilon)}\mu_{q(1/\sigma^2_\varepsilon)},
\end{aligned}
$$

where $\boldsymbol{\mu}_{q(\boldsymbol{\beta})}$ is defined to be the subvector of $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ corresponding to $\boldsymbol{\beta}$ and similarly, $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}$ is the sub-matrix of $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ corresponding to $\boldsymbol{\beta}$. The terms $\mathcal{C}_{q^{\mathrm{R}},\nu+q^{\mathrm{R}}-1}$ and $\mathcal{C}_{q^{\mathrm{R}},\nu+m+q^{\mathrm{R}}-1}$ are defined in Table 1.1. Details on the derivation for the optimal q-densities and lower bound expression are deferred to Appendix 2.A.

2.5.2 Approximating density functions of entries of $\boldsymbol{\Sigma}^{\mathrm{R}}$

In Section 2.10, we present the results from a series of comprehensive simulation studies that compare the MFVB methodology against MCMC. We do this comparison on the regression coefficients and the entries of the effects covariance matrix $\boldsymbol{\Sigma}^{\mathrm{R}}_{11}, \boldsymbol{\Sigma}^{\mathrm{R}}_{12}, \boldsymbol{\Sigma}^{\mathrm{R}}_{21}$ and $\boldsymbol{\Sigma}^{\mathrm{R}}_{22}$. The approximating density functions of regression coefficients are straightforward to attain, however, the ones related to the entries of the effects covariance matrix require additional effort. The derivations of the approximating density functions of these entries

are given in this subsection.

2.5.2.1 Main diagonal entries

The distributions of the main diagonal entries of $q^*(\boldsymbol{\Sigma}^{\mathrm{R}})$ are obtained using the following theorem (Menictas, 2015):

Theorem 2.5.1. If $\boldsymbol{X}$ is an $n \times n$ positive definite matrix that has an inverse-Wishart distribution with degrees of freedom a and scale matrix $\boldsymbol{B}$, then the diagonal entries $\boldsymbol{X}_{jj}$ are such that:

$$\boldsymbol{X}_{jj} \sim \text{Inverse-Gamma}\left(\tfrac{1}{2}\left(a-n+1\right), \tfrac{1}{2}\boldsymbol{B}_{jj}\right), \quad 1 \leq j \leq n.$$

From Algorithm 3 we see that the optimal q-density of $\boldsymbol{\Sigma}^{\mathrm{R}}$ has the form

$$\begin{aligned} q^*(\boldsymbol{\Sigma}^{\mathrm{R}}) &\sim \text{Inverse-Wishart}\left(A_{q(\boldsymbol{\Sigma}^{\mathrm{R}})},\, \boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\mathrm{R}})}\right), \\ \text{where} \quad A_{q(\boldsymbol{\Sigma}^{\mathrm{R}})} &= \nu + m + q^{\mathrm{R}} - 1 \\ \boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\mathrm{R}})} &= \sum_{i=1}^{m} (\boldsymbol{\mu}_{q(\boldsymbol{u}_i^{\mathrm{R}})} \boldsymbol{\mu}_{q(\boldsymbol{u}_i^{\mathrm{R}})}^{\mathsf{T}} + \boldsymbol{\Sigma}_{q(\boldsymbol{u}_i^{\mathrm{R}})}) + 2\,\nu\, \mathrm{diag}(\mu_{q(1/a_1^{\mathrm{R}})}, \ldots, \mu_{q(1/a_{q^{\mathrm{R}}}^{\mathrm{R}})}). \end{aligned}$$

Using Theorem 2.5.1, the main diagonal entries take the form

$$q^*(\boldsymbol{\Sigma}_{rr}^{\mathrm{R}}) \sim \text{Inverse-Gamma}\left(\tfrac{1}{2}(\nu+m),\, \tfrac{1}{2}(\boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\mathrm{R}})})_{jj}\right), \quad 1 \leq r \leq q^{\mathrm{R}}.$$

2.5.2.2 Off-diagonal entries

The distributions of the off diagonal entries of an inverse-Wishart random matrix $\boldsymbol{X}$ can be obtained analytically via the fast Fourier transform inversion of the characteristic function (Menictas, 2015):

Theorem 2.5.2. If $\boldsymbol{X}$ is an $n \times n$ positive definite matrix that has an inverse-Wishart distribution with degrees of freedom a and scale matrix $\boldsymbol{B}$ then the characteristic function of the off-diagonal entries $\boldsymbol{X}_{jj'}$ take the form:

$$E\left(e^{itX_{jj'}}\right) = \exp\left\{-\frac{a}{2}\sum_{\ell=1}^{n}\log(\lambda_\ell)\right\},$$

where λ_ℓ are the eigenvalues of the matrix $\boldsymbol{I} - 2i\boldsymbol{T}\boldsymbol{B}^{-1}$ and $\boldsymbol{T}$ is a matrix with the (j, j') and (j', j) entries equal to $t/2$ and all other entries equal to 0.

Unfortunately, there is no straightforward way of obtaining the above characteristic function. We therefore use a Monte Carlo sampling approach, subject to Monte Carlo errors, to find the distributions of the off diagonal elements in $q^*(\boldsymbol{\Sigma}^{\mathrm{R}})$.

2.6 Student-t semiparametric mixed models

We now consider the case where the distribution of response vector $\boldsymbol{y}$ is Student-t

$$\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \nu, \sigma_\varepsilon^2 \;\sim\; t\left((\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}),\, \sigma_\varepsilon^2,\, \nu\right), \quad \nu \sim \text{Uniform}(\nu_{\min},\, \nu_{\max}), \tag{2.15}$$

where $\nu_{\max},\ \nu_{\min} > 0$. The low degrees of freedom value $\nu > 0$ corresponds to the Student-t distribution with very heavy tails, and it is commonly used to model data containing gross outliers.

The notational infrastructure described in Section 2.5 for the Gaussian response model can be directly transferred to the Student-t response case. The only differences are the response distribution specification and inclusion of additional parameters $\boldsymbol{b}$ and ν. Using Result 1.8.2, we can rewrite (2.15) and thus the full two-level Bayesian Student-t semiparametric mixed model takes the following form:

$$\begin{aligned}
\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{b}, \sigma_\varepsilon^2 &\sim N\left((\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}), \sigma_\varepsilon^2 \boldsymbol{C}_{12}\right), \quad \boldsymbol{\beta} \sim N(0, \sigma_\beta^2 \boldsymbol{I}_p), \\
\boldsymbol{u}|\sigma_{u1}^2, \ldots, \sigma_{uL}^2, \boldsymbol{\Sigma}^{\text{R}} &\sim N\left(\boldsymbol{0}, \begin{bmatrix} \underset{1\le\ell\le L}{\text{blockdiag}}(\sigma_{u\ell}^2 \boldsymbol{I}_{q_\ell^{\text{G}}}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_m \otimes \boldsymbol{\Sigma}^{\text{R}} \end{bmatrix}\right), \\
\boldsymbol{\Sigma}^{\text{R}}|a_1^{\text{R}}, \ldots, a_{q^{\text{R}}}^{\text{R}} &\sim \text{Inverse-Wishart}\left(\nu + q^{\text{R}} - 1,\ 2\,\nu\,\text{diag}(1/a_1^{\text{R}}, \ldots, 1/a_{q^{\text{R}}}^{\text{R}})\right), \\
a_r^{\text{R}} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(\tfrac{1}{2}, 1/A_{\text{R}r}^2\right), \quad r = 1, \ldots, q^{\text{R}}, \\
\sigma_{u\ell}^2|a_{u\ell} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2},\, 1/a_{u\ell}), \quad a_{u\ell} \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2},\, 1/A_{u_\ell}^2), \\
\sigma_\varepsilon^2|a_\varepsilon &\sim \text{Inverse-Gamma}(\tfrac{1}{2},\, 1/a_\varepsilon), \quad a_\varepsilon \sim \text{Inverse-Gamma}(\tfrac{1}{2},\, 1/A_\varepsilon^2), \\
b_i|\nu &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{\nu}{2}, \nu/2) \quad \text{and} \quad \nu \sim \text{Uniform}(\nu_{\min},\, \nu_{\max}).
\end{aligned} \tag{2.16}$$

Recall that N denotes the total number of observations, i.e. $\sum_{i=1}^m n_i$ and $\boldsymbol{C}_{12} \equiv \underset{1\le i\le N}{\text{diag}}\,(b_i)$.

2.6.1 Approximate Bayesian inference via mean field variational Bayes

A good starting point is to look at the DAG for model (2.16) as illustrated in Figure 2.3. It shares many similarities with the DAG for model (2.11), with the exception of additional parameters $\boldsymbol{b}$ and ν. This is where the locality property of MFVB approximation comes to the fore through the factorisation afforded by DAG models. As set out in Section 1.4, the distribution of a node within a DAG depends only on the nodes within its Markov blanket. But what impact does this locality property have on the derivation of optimal q-densities for model (2.16)? Inspecting Figure 2.3 we can see similarities between the DAGs for the Gaussian and Student-t response models. The parameters $\boldsymbol{\Sigma}^{\text{R}}, \boldsymbol{a}^{\text{R}}, \sigma_{u\ell}^2$ and $a_{u\ell}$ play the same role as in the Gaussian response model and therefore we expect to see their optimal q-density remain unchanged. However, the optimal q-densities of $(\boldsymbol{\beta}, \boldsymbol{u})$ and σ_ε^2

have dependencies on the parameter $\boldsymbol{b}$ and therefore we expect to see discrepancies of these optimal q-densities between the Gaussian and Student-t response models. Ultimately, the locality property of MFVB not only affords considerable simplification for the model at hand, but also allows MFVB results for one model to be transferred to another.

Standard manipulations lead to the following full conditionals for parameters $\boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2, b_i$ and ν:

$$\begin{aligned}
\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix} \,\Big|\, \text{rest} &\sim N\left((\sigma_\varepsilon^{-2}\boldsymbol{C}^\mathsf{T}\boldsymbol{C}_{12}\boldsymbol{C}+\boldsymbol{\Omega}^{-1})^{-1}\sigma_\varepsilon^{-2}\boldsymbol{C}^\mathsf{T}\boldsymbol{C}_{12}\boldsymbol{y},\, (\sigma_\varepsilon^{-2}\boldsymbol{C}^\mathsf{T}\boldsymbol{C}_{12}\boldsymbol{C}+\boldsymbol{\Omega}^{-1})^{-1}\right),\\
\sigma_\varepsilon^2|\text{rest} &\sim \text{Inverse-Gamma}\left(\tfrac{1}{2}\,(\textstyle\sum_{i=1}^m n_i+1)\,,\, \tfrac{1}{2}\,(\boldsymbol{y}-\boldsymbol{C}\boldsymbol{v})^\mathsf{T}\boldsymbol{C}_{12}(\boldsymbol{y}-\boldsymbol{C}\boldsymbol{v})+a_\varepsilon^{-1}\right),\\
b_i|\text{rest} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(\tfrac{1}{2}(\nu+1),\, \tfrac{1}{2}\sigma_\varepsilon^{-2}\left\{(\boldsymbol{y}-\boldsymbol{C}\boldsymbol{v})_i^\mathsf{T}(\boldsymbol{y}-\boldsymbol{C}\boldsymbol{v})_i+\nu\right\}\right),\\
p(\nu|\text{rest}) &= \frac{1}{(\nu_{\max}-\nu_{\min})}\prod_{i=1}^N\left\{\frac{(0.5\,\nu)^{0.5\nu}}{\Gamma(0.5\,\nu)}(b_i)^{-\frac{1}{2}(\nu+2)}\exp\left(-\frac{\nu}{2b_i}\right)\right\}.
\end{aligned}$$

Note that the full conditional of ν is of non-standard form.

Similar to Subsection 2.5.1, we impose the following product restriction on the approximating q-densities:

$$\begin{aligned}
q(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{a}^\text{R},\boldsymbol{a}_u,a_\varepsilon,\boldsymbol{\Sigma}^\text{R},\boldsymbol{\sigma}_u^2,\nu,\boldsymbol{b}) &= q(\boldsymbol{\beta},\boldsymbol{u})\,q(a_\varepsilon)\,q(\boldsymbol{\Sigma}^\text{R})\,q(\sigma_\varepsilon^2)\,q(\nu) \\
&\quad\times\left\{\prod_{r=1}^{q^\text{R}}q(a_r^\text{R})\right\}\left\{\prod_{\ell=1}^{L}q(\sigma_{u\ell}^2)\right\}\left\{\prod_{\ell=1}^{L}q(a_{u\ell})\right\}\left\{\sum_{i=1}^{N}q(b_i)\right\}.
\end{aligned} \tag{2.17}$$

As shown through the derivations in Appendix 2.B, the optimal q-densities admit the following forms:

$$\begin{aligned}
&q^*(\boldsymbol{\beta},\boldsymbol{u}) \text{ is the } N\left(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})},\, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\right) \text{density function,}\\
&q^*(\sigma_\varepsilon^2) \text{ is the Inverse-Gamma } \left(\tfrac{1}{2}\,(\textstyle\sum_{i=1}^m n_i+1)\,,\, B_{q(\sigma_\varepsilon^2)}\right) \text{density function,}\\
&q^*(a_\varepsilon) \text{ is the Inverse-Gamma}(1,\, B_{q(a_\varepsilon)}) \text{ density function,}\\
&q^*(\sigma_{u\ell}^2) \text{ is the Inverse-Gamma } \left(\tfrac{1}{2}\,(q_\ell^\text{G}+1),\, B_{q(\sigma_{u\ell}^2)}\right) \text{density function,}\\
&q^*(a_{u\ell}) \text{ is the Inverse-Gamma}(1,\, B_{q(a_{u\ell})}) \text{ density function,}\\
&q^*(\boldsymbol{\Sigma}^\text{R}) \text{ is the Inverse-Wishart } \left(\nu+m+q^\text{R}+1,\, \boldsymbol{B}_{q(\boldsymbol{\Sigma}^\text{R})}\right) \text{density function,}\\
&q^*(a_r^\text{R}) \text{ is the Inverse-Gamma } \left(\tfrac{1}{2}(\nu+q^\text{R}),\, B_{q(a_r^\text{R})}\right) \text{density function,}\\
&q^*(b_i) \overset{\text{ind.}}{\sim} \text{Inverse-Gamma } \left(\tfrac{1}{2}(\mu_{q(\nu)}+1), B_{q(b_i)}\right) \text{density function and}\\
&q^*(\nu) = \frac{\exp\left[N\left\{\frac{1}{2}\nu\log(\nu/2)-\log\Gamma(\nu/2)\right\}-\frac{1}{2}\nu\,C_1\right]}{\mathcal{F}(0,N,C_1,\nu_{\min},\nu_{\max})}, \quad \nu_{\min}\le\nu\le\nu_{\max},
\end{aligned}$$

for parameters $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$, the mean vector and covariance matrix of $q^*(\boldsymbol{\beta},\boldsymbol{u})$,

$B_{q(\sigma_\varepsilon^2)}$, the scale parameter of $q^*(\sigma_\varepsilon^2)$, $B_{q(a_\varepsilon)}$, the scale parameter of $q^*(a_\varepsilon)$, $B_{q(\sigma_{u\ell}^2)}$, the scale parameter of $q^*(\sigma_{u\ell}^2)$, f$B_{q(a_{u\ell})}$, the scale parameter of $q^*(a_{u\ell})$, $B_{q(a_r^{\mathrm{R}})}$, the scale parameter of $q^*(a_r^{\mathrm{R}})$ and $\boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\mathrm{R}})}$, the scale parameter of $q^*(\boldsymbol{\Sigma}^{\mathrm{R}})$. $B_{q(b_i)}$, the scale parameter of $q^*(b_i)$. The last density uses the definition: $C_1 \equiv \sum_{i=1}^N \{\mu_{q(\log(b_i))} + \mu_{q(1/b_i)}\}$.

The MFVB approximation involves standard density functions except for the parameter ν. The optimal q-density $q^*(\nu)$ involves a non-analytic integral (Wand *et al.*, 2012)

$$\mathcal{F}(p,q,r,s,t) \equiv \int_s^t x^p \exp\left[q\left\{\tfrac{1}{2}x\log(x/2) - \log\Gamma(x/2)\right\} - \tfrac{1}{2}\,r\,x\right] dx,$$

where $p \geq 0$ and $q, r, s, t > 0$, and requires univariate quadrature to solve numerically. A simple and effective form of univariate quadrature is the trapezoidal rule, which can be made arbitrarily accurate as the number of trapezoidal elements increases.

Algorithm 3 gives details of an iterative scheme for finding optimal q-densities for all key model parameters under product restriction (2.17). Convergence of Algorithm 3 can be monitored using successive values of the lower bound on the marginal log-likelihood $\log \underline{p}(\boldsymbol{y}; q)$ and is given as:

$$\begin{aligned}
\log \underline{p}(\boldsymbol{y}; q) \;=\;& \tfrac{1}{2}\,q^{\mathrm{R}}(\nu + q^{\mathrm{R}} - 1)\log(2\nu) - \tfrac{1}{2}\log(2\pi)\textstyle\sum_{i=1}^m n_i - \left(\tfrac{1}{2}q^{\mathrm{R}} + L + 1\right)\log(\pi) \\
& - \tfrac{1}{2}p\log(\sigma_\beta^2) - \tfrac{1}{2}\sigma_\beta^{-2}\left(\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\right) + \tfrac{1}{2}\left(\textstyle\sum_{\ell=1}^L q_\ell^{\mathrm{G}} + p + m\right) \\
& + \tfrac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}| - \log(\mathcal{C}_{q^{\mathrm{R}},\,\nu+q^{\mathrm{R}}-1}) + \log(\mathcal{C}_{q^{\mathrm{R}},\,\nu+m+q^{\mathrm{R}}-1}) \\
& - \tfrac{1}{2}\,(\nu + m + q^{\mathrm{R}} - 1)\log|\boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\mathrm{R}})}| \\
& + \textstyle\sum_{\ell=1}^L \log\Gamma\left\{\tfrac{1}{2}\,(q_\ell^{\mathrm{G}} + 1)\right\} - \tfrac{1}{2}\,\sum_{\ell=1}^L (q_\ell^{\mathrm{G}} + 1)\,\log(B_{q(\sigma_{u\ell}^2)}) \\
& + \log\Gamma\left\{\tfrac{1}{2}(\textstyle\sum_{i=1}^m n_i + 1)\right\} - \tfrac{1}{2}\,(\textstyle\sum_{i=1}^m n_i + 1)\log\left(B_{q(\sigma_\varepsilon^2)}\right) \\
& - \textstyle\sum_{r=1}^{q^{\mathrm{R}}} \log(A_{\mathrm{R}r}) + q^{\mathrm{R}}\log\Gamma\left\{\tfrac{1}{2}(\nu + q^{\mathrm{R}})\right\} \\
& - \textstyle\sum_{\ell=1}^L \log(A_{u\ell}) - \sum_{\ell=1}^L \log\left(B_{q(a_{u\ell})}\right) + \sum_{\ell=1}^L \mu_{q(1/a_{u\ell})}\mu_{q(1/\sigma_{u\ell}^2)} \\
& + \textstyle\sum_{r=1}^{q^{\mathrm{R}}} \nu(\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})})_{rr}\,\mu_{q(1/a_r^{\mathrm{R}})} - \tfrac{1}{2}\,(\nu + q^{\mathrm{R}})\,\sum_{r=1}^{q^{\mathrm{R}}}\log(B_{q(a_r^{\mathrm{R}})}) \\
& - \log(A_\varepsilon) - \log\left(B_{q(a_\varepsilon)}\right) + \mu_{q(1/a_\varepsilon)}\mu_{q(1/\sigma_\varepsilon^2)} \\
& - \tfrac{1}{2}\mu_{q(\nu+1)}\textstyle\sum_{i=1}^N \log B_{q(b_i)} + \log\Gamma\left\{\tfrac{1}{2}(\mu_{q(\nu)} + 1)\right\} + \sum_{i=1}^N B_{q(b_i)}\mu_{q(1/b_i)} \\
& - \log(\nu_{\max} - \nu_{\min}) - \tfrac{1}{2}\mu_{q(\nu)}\textstyle\sum_{i=1}^N \mu_{q(\log(b_i))} + \log\mathcal{F}(0, \sum_{i=1}^m n_i, C_1, \nu_{\max}, \nu_{\min}),
\end{aligned}$$

where $\boldsymbol{\mu}_{q(\boldsymbol{\beta})}$ is defined to be the subvector of $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ corresponding to $\boldsymbol{\beta}$ and similarly, $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}$ is the sub-matrix of $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ corresponding to $\boldsymbol{\beta}$. The terms $\mathcal{C}_{q^{\mathrm{R}},\nu+q^{\mathrm{R}}-1}$ and $\mathcal{C}_{q^{\mathrm{R}},\nu+m+q^{\mathrm{R}}-1}$ are defined in Table 1.1. Details on the derivation for the optimal q-densities and lower bound expression are deferred to Appendix 2.B.

Initialise: $\nu = 2$, $\mu_{q(1/\sigma_\varepsilon^2)} > 0$, $\mu_{q(1/a_\varepsilon)} > 0$, $\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})}$, a $q^{\mathrm{R}} \times q^{\mathrm{R}}$ positive definite matrix, $\mu_{q(1/a_r^{\mathrm{R}})} > 0$, $1 \leq r \leq q^{\mathrm{R}}$, $\mu_{q(1/\sigma_{u\ell}^2)} > 0$, $1 \leq \ell \leq L$, $\mu_{q(1/b_i)} > 0, 1 \leq i \leq N$ and $\mu_{q(\nu)} > 0$.

Cycle through updates:

$$\textbf{Define: } \boldsymbol{M}_{q(\boldsymbol{\Omega}^{-1})} \equiv \begin{bmatrix} \sigma_\beta^{-2}\boldsymbol{I}_p & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \underset{1\leq\ell\leq L}{\text{blockdiag}}\left(\mu_{q(1/\sigma_{u\ell}^2)}\boldsymbol{I}_{q_\ell^{\mathrm{G}}}\right) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_m \otimes \boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})} \end{bmatrix}$$

$N \leftarrow \sum_{i=1}^m n_i$; Gaussian: $\boldsymbol{\Psi} = \boldsymbol{I}$; Student-t: $\boldsymbol{\Psi} \leftarrow \underset{1\leq i\leq N}{\text{diag}}\left(\mu_{q(1/b_i)}\right)$

1. Update multivariate normal $q^*(\boldsymbol{\beta}, \boldsymbol{u})$ parameters:

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \left(\mu_{q(1/\sigma_\varepsilon^2)}\boldsymbol{C}^{\mathsf{T}}\boldsymbol{\Psi}\boldsymbol{C} + \boldsymbol{M}_{q(\boldsymbol{\Omega})^{-1}}\right)^{-1}$$

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

2. Update inverse-Gamma $q^*(\sigma_\varepsilon^2)$ and inverse-Gamma $q^*(a_\varepsilon)$ parameters:

$$B_{q(\sigma_\varepsilon^2)} \leftarrow \mu_{q(1/a_\varepsilon)} + \tfrac{1}{2}\left\{(\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})})^{\mathsf{T}}\boldsymbol{\Psi}(\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}) + \text{tr}(\boldsymbol{C}^{\mathsf{T}}\boldsymbol{\Psi}\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})\right\}$$

$$\mu_{q(1/\sigma_\varepsilon^2)} \leftarrow \tfrac{1}{2}\left(\textstyle\sum_{i=1}^m n_i + 1\right)/B_{q(\sigma_\varepsilon^2)} \quad ; \quad \mu_{q(1/a_\varepsilon)} \leftarrow 1/\left(\mu_{q(1/\sigma_\varepsilon^2)} + A_\varepsilon^{-2}\right)$$

3. For Student-t response model only, update inverse-Gamma $q^*(b_i)$ and $q^*(\nu)$ parameters:

For $1 \leq i \leq N$:

$$B_{q(b_i)} \leftarrow \tfrac{1}{2}\left[\mu_{q(\nu)} + \mu_{q(1/\sigma_\varepsilon^2)}\left\{(\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})})_i^2 + (\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^{\mathsf{T}})_{ii}\right\}\right]$$
$$\mu_{q(1/b_i)} \leftarrow \tfrac{1}{2}(\mu_{q(\nu)} + 1)/B_{q(b_i)}$$
$$\mu_{q(\log b_i)} \leftarrow \log(B_{q(b_i)}) - \text{digamma}\left\{\tfrac{1}{2}(\mu_{q(\nu)} + 1)\right\}$$

$$C_1 \leftarrow \textstyle\sum_{i=1}^N \left\{\mu_{q(\log(b_i))} + \mu_{q(1/b_i)}\right\}$$

$$\mu_{q(\nu)} \leftarrow \exp\left\{\log \mathcal{F}(1, N, C_1, \nu_{\min}, \nu_{\max}) - \log \mathcal{F}(0, N, C_1, \nu_{\min}, \nu_{\max})\right\}$$

4. Update inverse-Gamma $q^*(a_{u\ell})$ and inverse-Gamma $q^*(\sigma_{u\ell}^2)$ parameters:

For $\ell = 1, \ldots, L$:

$$\mu_{q(1/a_{u\ell})} \leftarrow 1/(\mu_{q(1/\sigma_{u\ell}^2)} + A_{u\ell}^{-2})$$
$$\mu_{q(\sigma_{u\ell}^2)} \leftarrow \frac{q_\ell^{\mathrm{G}} + 1}{||\boldsymbol{\mu}_{q(\boldsymbol{u}_\ell^{\mathrm{G}})}||^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_\ell^{\mathrm{G}})}) + 2\,\mu_{q(1/a_{u\ell})}}$$

5. Update inverse-Gamma $q^*(a_r^{\mathrm{R}})$ and inverse-Wishart $q^*(\boldsymbol{\Sigma}^{\mathrm{R}})$parameters:

For $r = 1, \ldots, q^{\mathrm{R}}$:

$$B_{q(a_r^{\mathrm{R}})} \leftarrow \nu(\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})})_{rr} + A_{\mathrm{R}r}^{-2} \quad ; \quad \mu_{q(a_r^{\mathrm{R}})} \leftarrow \tfrac{1}{2}(\nu + q^{\mathrm{R}})/B_{q(a_r^{\mathrm{R}})}$$

$$\boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\mathrm{R}})} \leftarrow \sum_{i=1}^{m} (\boldsymbol{\mu}_{q(\boldsymbol{u}_i^{\mathrm{R}})} \boldsymbol{\mu}_{q(\boldsymbol{u}_i^{\mathrm{R}})}^{\mathsf{T}} + \boldsymbol{\Sigma}_{q(\boldsymbol{u}_i^{\mathrm{R}})}) + 2\,\nu\,\mathrm{diag}(\mu_{q(1/a_1^{\mathrm{R}})}, \ldots, \mu_{q(1/a_{q^{\mathrm{R}}}^{\mathrm{R}})})$$

$$\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})} \leftarrow (\nu + m + q^{\mathrm{R}} - 1)\ \boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\mathrm{R}})}^{-1}$$

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

Algorithm 3: Mean field variational Bayes algorithm for obtaining model parameters in the optimal q-density functions for the two-level Bayesian semiparametric mixed models with the Gaussian and Student-t responses.

2.7 Bernoulli semiparametric mixed models

We now consider the case where the response distribution of $\boldsymbol{y}$ is Bernoulli distributed, the logistic mixed model

$$\begin{aligned} \boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u} &\sim \text{Bernoulli}\left(\text{logit}^{-1}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u})\right), \\ \text{where} \qquad &\text{logit}^{-1}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}) = \frac{e^{\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u}}}{\boldsymbol{1} + e^{\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u}}} \end{aligned} \tag{2.18}$$

is appropriate. The notation $v \sim \text{Bernoulli}(p)$ used in (2.18) is shorthand for the entries of the random vector v having independent Bernoulli distributions with parameters corresponding to the entries of p.

The notational infrastructure described in Section 2.5 for the Gaussian response model can be directly transferred to the Bernoulli response case. The only change is the response distribution specification and omission of parameters σ_ε^2 and a_ε. The two-level Bayesian Bernoulli semiparametric mixed model stated in full is:

$$\begin{aligned} \boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2 &\sim \text{Bernoulli}\left(\text{logit}^{-1}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u})\right), \quad \boldsymbol{\beta} \sim N(0, \sigma_\beta^2 \boldsymbol{I}_p), \\ \boldsymbol{u}|\sigma_{u1}^2, \ldots, \sigma_{uL}^2, \boldsymbol{\Sigma}^{\mathrm{R}} &\sim N\left(\boldsymbol{0}, \begin{bmatrix} \underset{1\le\ell\le L}{\text{blockdiag}}(\sigma_{u\ell}^2 \boldsymbol{I}_{q_\ell^{\mathrm{G}}}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_m \otimes \boldsymbol{\Sigma}^{\mathrm{R}} \end{bmatrix}\right), \\ \boldsymbol{\Sigma}^{\mathrm{R}}|a_1^{\mathrm{R}}, \ldots, a_{q^{\mathrm{R}}}^{\mathrm{R}} &\sim \text{Inverse-Wishart}\left(\nu + q^{\mathrm{R}} - 1,\ 2\,\nu\,\mathrm{diag}(1/a_1^{\mathrm{R}}, \ldots, 1/a_{q^{\mathrm{R}}}^{\mathrm{R}})\right), \\ a_r^{\mathrm{R}} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(\tfrac{1}{2}, 1/A_{\mathrm{R}r}^2\right), \quad r = 1, \ldots, q^{\mathrm{R}}, \\ \sigma_{u\ell}^2|a_{u\ell} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2},\ 1/a_{u\ell}), \quad a_{u\ell} \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2},\ 1/A_{u\ell}^2). \end{aligned} \tag{2.19}$$

2.7.1 Approximate Bayesian inference via mean field variational Bayes

We once again utilise the locality property of MFVB approximation to derive the necessary optimal q-densities for model (2.19). Inspection of Figure 2.4 reviews great similarities between the DAGs for the Gaussian and Bernoulli response models. The only change that is required would be for $(\boldsymbol{\beta}, \boldsymbol{u})$ since the response vector $\boldsymbol{y}$ is now Bernoulli distributed and hence the optimal q-density $q^*(\boldsymbol{\beta}, \boldsymbol{u})$ is the focus of this subsection.

The marginal likelihood is

$$p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}) \;=\; \exp\left[\boldsymbol{y}^{\mathsf{T}}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}) - \mathbf{1}^{\mathsf{T}} \log\left\{\mathbf{1} + \exp(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u})\right\}\right].$$

The joint density function of $(\boldsymbol{\beta}, \boldsymbol{u})$ satisfies

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{u}) \;\propto\; & \exp\left[\boldsymbol{y}^{\mathsf{T}}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}) - \mathbf{1}^{\mathsf{T}}\log\left\{\mathbf{1} + \exp(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u})\right\}\right] \\ & \times \prod_{i=1}^{m} \exp\left[-\tfrac{1}{2}(\boldsymbol{u}^{\mathrm{R}})^{\mathsf{T}}\left\{\boldsymbol{I}_m \otimes (\boldsymbol{\Sigma}^{\mathrm{R}})^{-1}\right\}\boldsymbol{u}^{\mathrm{R}}\right] \\ & \times \prod_{\ell=1}^{L} \exp\left\{-\tfrac{1}{2}(\boldsymbol{u}_\ell^{\mathrm{G}})^{\mathsf{T}}(\sigma_{u\ell}^2 \boldsymbol{I}_{q_\ell^{\mathrm{G}}})^{-1}\boldsymbol{u}_\ell^{\mathrm{G}}\right\} \exp\left\{-\tfrac{1}{2}\,\boldsymbol{\beta}^{\mathsf{T}}(\sigma_\beta^2 \boldsymbol{I}_p)^{-1}\boldsymbol{\beta}\right\}. \end{aligned}$$

This is a non-standard form and poses tractability problems with regard to approximate inference for $(\boldsymbol{\beta}, \boldsymbol{u})$. The term $-\log(1 + e^x)$ is a convex function of x and we propose to transform this convex term to a simple quadratic function $f(x; \boldsymbol{\xi})$, a trick first introduced in Jaakkola and Jordan (1997). Different values of $\boldsymbol{\xi}$ correspond to different parabolas, all of which is smaller than $-\log(1 + e^x)$, and thus we have the following representation of $-\log(1 + e^x)$ as the maxima of a family of parabolas (Ormerod and Wand, 2010):

$$\begin{aligned} -\log(1 + e^x) \;&=\; \max_{\xi \in \mathbb{R}} \left\{\lambda(\xi)x^2 - \tfrac{1}{2}x + \psi(\xi)\right\} \text{ for all } x \in \mathbb{R}, \qquad (2.20) \\ \text{where} \quad \lambda(\xi) \;&\equiv\; -\tanh(\xi/2)/(4\xi) \\ \text{and} \quad \psi(\xi) \;&\equiv\; \xi/2 - \log(1 + e^{\xi}) + \xi\tanh(\xi/2)/4. \end{aligned}$$

Figure 2.6 is the graphical representation of (2.20), in which the function $-\log(1 + e^x)$ is seen to be the maximum of a family of parabolas. It follows from (2.20) that

$$\begin{aligned} \mathbf{1}^{\mathsf{T}}\log\left\{\mathbf{1} + \exp(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u})\right\} \;&\geq\; \mathbf{1}^{\mathsf{T}}\left\{\lambda(\boldsymbol{\xi}) \odot (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u})^2 - \tfrac{1}{2}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}) + \psi(\boldsymbol{\xi})\right\}, \\ &=\; (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u})^{\mathsf{T}}\,\mathrm{diag}\{\lambda(\boldsymbol{\xi})\}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}) \\ &\quad -\tfrac{1}{2}\mathbf{1}^{\mathsf{T}}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}) + \mathbf{1}^{\mathsf{T}}\psi(\boldsymbol{\xi}), \end{aligned}$$

where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N)$ is an $(\sum_{i=1}^m n_i) \times 1$ vector of variational parameters. Recall that

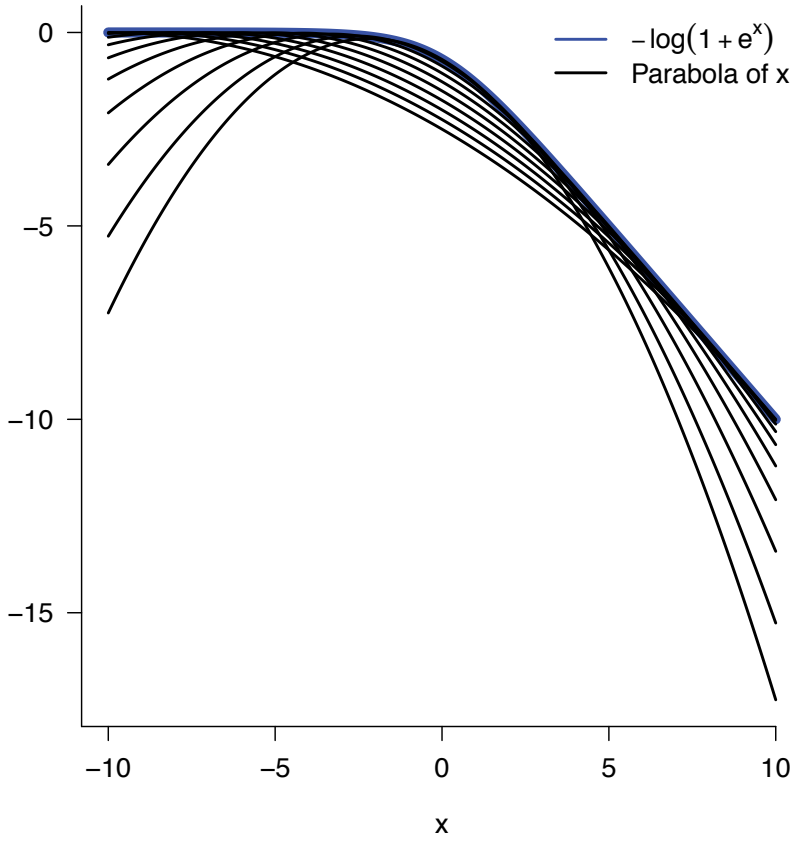

Figure 2.6: Variational representation of the function $-\log(1+e^x)$ as the maximum of a family of parabolas, corresponding to (2.20).

$\boldsymbol{C} \equiv \left[\boldsymbol{X}\ \boldsymbol{Z}\right]$ and $\boldsymbol{v} = \left[\boldsymbol{\beta}^{\mathsf{T}}\ \boldsymbol{u}^{\mathsf{T}}\right]^{\mathsf{T}}$, the lower bound on $p(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\xi})$ is

$$\underline{p}\,(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\xi}) \quad \propto \quad \exp\left(\boldsymbol{y}^{\mathsf{T}}\boldsymbol{C}\boldsymbol{v} - \left[\boldsymbol{v}^{\mathsf{T}}\boldsymbol{C}^{\mathsf{T}}\,\mathrm{diag}\{\lambda(\boldsymbol{\xi})\}\,\boldsymbol{C}\boldsymbol{v} - \tfrac{1}{2}\mathbf{1}^{\mathsf{T}}\,\boldsymbol{C}\boldsymbol{v} + \mathbf{1}^{\mathsf{T}}\,\psi(\boldsymbol{\xi})\right] - \tfrac{1}{2}\boldsymbol{v}^{\mathsf{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{v}\right),$$

where $\boldsymbol{\Omega}$ is as defined in Section 2.4. Noting that since the right hand side of the above equation is a quadratic form and, by completing the square in the usual way to identify the mean and covariance, we approximate the full conditional of $(\boldsymbol{\beta}, \boldsymbol{u})$ by a multivariate normal distribution

$$\begin{aligned}\boldsymbol{\beta}, \boldsymbol{u} | \boldsymbol{y}; \boldsymbol{\xi} \quad \sim \quad N\Big(& (2\boldsymbol{C}^{\mathsf{T}}\mathrm{diag}\{\lambda(\boldsymbol{\xi})\}\boldsymbol{C} + \boldsymbol{\Omega}^{-1})^{-1}\boldsymbol{C}^{\mathsf{T}}\left(\boldsymbol{y} - \tfrac{1}{2}\mathbf{1}\right), \\ & (2\boldsymbol{C}^{\mathsf{T}}\mathrm{diag}\{\lambda(\boldsymbol{\xi})\}\boldsymbol{C} + \boldsymbol{\Omega}^{-1})^{-1}\Big).\end{aligned}$$

The Bernoulli response model entails a variant of variational approximation, namely the *tangent transform approach* of Jaakkola and Jordan (1997). As described above, we succeed in simplifying the convex function of x with a simple quadratic function. The only trade-off is to include the additional variational parameter vector $\boldsymbol{\xi}$ and to obtain

the lower bound with respect to $\boldsymbol{\xi}$. The Jaakkola and Jordan (1997) trick is one of a few approaches that has been proposed in the variational approximation literature for handling binary response regression models. Others include the use of probit auxiliary variable trick of Albert and Chib (1993) and Gaussian variational approximation (e.g. Knowles and Minka, 2011; Ormerod and Wand, 2012). The Gaussian variational approximation approach emerges as a fast, deterministic alternative to Laplace approximation for fitting of grouped data generalised linear mixed models. The approach is conceptually simple and its derivation requires application of Jensen's inequality to the log-likelihood to obtain a variational lower bound. Various trade-offs are involved with the choice amongst these options. For example, Gaussian variational approximation requires univariate numerical integration whereas Algorithm 4 has only purely algebraic updates.

Similar to Subsection 2.5.1, the q-densities are subject to the following product form:

$$q(\boldsymbol{\beta},\boldsymbol{u})\,q(\boldsymbol{\Sigma}^{\rm R})\left\{\prod_{r=1}^{q^{\rm R}}q(a_r^{\rm R})\right\}\left\{\prod_{\ell=1}^{L}q(\sigma_{u\ell}^2)\right\}\left\{\prod_{\ell=1}^{L}q(a_{u\ell})\right\}. \tag{2.21}$$

As shown through the derivations in Appendix 2.C, the optimal q-densities admit the following forms:

$$\begin{aligned}
&\boldsymbol{\xi}\leftarrow\sqrt{\text{diagonal}\left\{\boldsymbol{C}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}+\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}\,\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}^{\mathsf{T}}\right)\right\}\boldsymbol{C}^{\mathsf{T}}},\\
&q^*(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})\,\text{is the } N(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})},\,\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})})\,\text{density function},\\
&q^*(\sigma_{u\ell}^2)\,\text{is the Inverse-Gamma}\left(\tfrac{1}{2}\,(q_\ell^{\rm G}+1),\,B_{q(\sigma_{u\ell}^2)}\right)\,\text{density function},\\
&q^*(a_{u\ell})\,\text{is the Inverse-Gamma}(1,\,B_{q(a_{u\ell})})\,\text{density function},\\
&q^*(\boldsymbol{\Sigma}^{\rm R})\,\text{is the Inverse-Wishart}\left(\nu+m+q^{\rm R}+1,\,\boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\rm R})}\right)\,\text{density function},\\
&q^*(a_r^{\rm R})\,\text{is the Inverse-Gamma}\left(\tfrac{1}{2}(\nu+q^{\rm R}),\,B_{q(a_r^{\rm R})}\right)\,\text{density function},
\end{aligned}$$

for parameters $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$, the mean vector and covariance matrix of $q^*(\boldsymbol{\beta},\boldsymbol{u})$, $B_{q(\sigma_{u\ell}^2)}$, the scale parameter of $q^*(\sigma_{u\ell}^2)$, $B_{q(a_{u\ell})}$, the scale parameter of $q^*(a_{u\ell})$, $B_{q(a_r^{\rm R})}$, the scale parameter of $q^*(a_r^{\rm R})$ and $\boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\rm R})}$, the scale parameter of $q^*(\boldsymbol{\Sigma}^{\rm R})$.

Algorithm 4 gives details of an iterative scheme for finding optimal q-densities for all key model parameters under product restriction (2.21). Convergence of Algorithm 4 can be monitored using successive values of the lower bound on the marginal log-likelihood and is given as:

$$\begin{aligned}
\log\underline{p}(\boldsymbol{y};q) \;=\;& \tfrac{1}{2}\,q^{\rm R}(\nu+q^{\rm R}-1)\log(2\nu)-\left(\tfrac{1}{2}q^{\rm R}+L\right)\log(\pi)-\lambda(\boldsymbol{\xi})^{\mathsf{T}}(\boldsymbol{\xi}^2)\\
&+\boldsymbol{1}^{\mathsf{T}}\psi(\boldsymbol{\xi})+\left(\boldsymbol{y}-\tfrac{1}{2}\boldsymbol{1}\right)^{\mathsf{T}}\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}+\tfrac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}|\\
&-\tfrac{1}{2}p\log(\sigma_\beta^2)-\tfrac{1}{2}\sigma_\beta^{-2}\left(\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2+\text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\right)+\tfrac{1}{2}\left(\textstyle\sum_{\ell=1}^{L}q_\ell^{\rm G}+p+m\right)
\end{aligned}$$

$$
\begin{aligned}
&- \log(\mathcal{C}_{q^{\mathrm{R}},\nu+q^{\mathrm{R}}-1}) + \log(\mathcal{C}_{q^{\mathrm{R}},\nu+m+q^{\mathrm{R}}-1}) - \tfrac{1}{2}\,(\nu+m+q^{\mathrm{R}}-1)\log|B_{q(\boldsymbol{\Sigma}^{\mathrm{R}})}| \\
&+ \textstyle\sum_{\ell=1}^{L} \log\Gamma\left\{\tfrac{1}{2}\,(q_\ell^{\mathrm{G}}+1)\right\} - \tfrac{1}{2}\,\sum_{\ell=1}^{L}(q_\ell^{\mathrm{G}}+1)\,\log(B_{q(\sigma^2_{u\ell})}) \\
&- \textstyle\sum_{r=1}^{q^{\mathrm{R}}} \log(A_{\mathrm{R}r}) + q^{\mathrm{R}}\log\Gamma\left\{\tfrac{1}{2}(\nu+q^{\mathrm{R}})\right\} \\
&+ \textstyle\sum_{r=1}^{q^{\mathrm{R}}} \nu(\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})})_{rr}\,\mu_{q(1/a_r^{\mathrm{R}})} \\
&-\tfrac{1}{2}\,(\nu+q^{\mathrm{R}})\,\textstyle\sum_{r=1}^{q^{\mathrm{R}}}\log(B_{q(a_r^{\mathrm{R}})}) - \sum_{\ell=1}^{L}\log(A_{u\ell}) \\
&- \textstyle\sum_{\ell=1}^{L}\log\left(B_{q(a_{u\ell})}\right) + \sum_{\ell=1}^{L}\mu_{q(1/a_{u\ell})}\mu_{q(1/\sigma^2_{u\ell})},
\end{aligned}
$$

where $\boldsymbol{\mu}_{q(\boldsymbol{\beta})}$ is defined to be the subvector of $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ corresponding to $\boldsymbol{\beta}$ and similarly, $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}$ is the sub-matrix of $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ corresponding to $\boldsymbol{\beta}$. The terms $\mathcal{C}_{q^{\mathrm{R}},\nu+q^{\mathrm{R}}-1}$ and $\mathcal{C}_{q^{\mathrm{R}},\nu+m+q^{\mathrm{R}}-1}$ are defined in Table 1.1. Details on the derivation for the optimal q-densities and lower bound expression are deferred to Appendix 2.C.

2.8 Poisson semiparametric mixed models

Finally, we consider the distribution of response vector $\boldsymbol{y}$ as Poisson, the full two-level Bayesian Poisson semiparametric mixed model is

$$
\begin{aligned}
\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u} &\sim \text{Poisson}\left(\exp(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u})\right), \quad \boldsymbol{\beta}\sim N(0,\sigma_\beta^2\boldsymbol{I}_p), \\
\boldsymbol{u}|\sigma^2_{u1},\ldots,\sigma^2_{uL},\boldsymbol{\Sigma}^{\mathrm{R}} &\sim N\left(\boldsymbol{0},\begin{bmatrix}\underset{1\le\ell\le L}{\text{blockdiag}}(\sigma^2_{u\ell}\boldsymbol{I}_{q_\ell^{\mathrm{G}}}) & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}_m\otimes\boldsymbol{\Sigma}^{\mathrm{R}}\end{bmatrix}\right), \\
\boldsymbol{\Sigma}^{\mathrm{R}}|a_1^{\mathrm{R}},\ldots,a_{q^{\mathrm{R}}}^{\mathrm{R}} &\sim \text{Inverse-Wishart}\left(\nu+q^{\mathrm{R}}-1,\ 2\,\nu\,\text{diag}(1/a_1^{\mathrm{R}},\ldots,1/a_{q^{\mathrm{R}}}^{\mathrm{R}})\right), \\
a_r^{\mathrm{R}} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(\tfrac{1}{2},1/A^2_{\mathrm{R}r}\right), \quad r=1,\ldots,q^{\mathrm{R}}, \\
\sigma^2_{u\ell}|a_{u\ell} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2},\,1/a_{u\ell}), \quad a_{u\ell}\overset{\text{ind.}}{\sim}\text{Inverse-Gamma}(\tfrac{1}{2},\,1/A^2_{u_\ell}).
\end{aligned}
\tag{2.22}
$$

Figure 2.4 provides a DAG representation of (2.19) and (2.22). Observed data are indicated by the shaded node and random parameters are represented by the open nodes. The visual representation shows that the Bernoulli and Poisson response models have all parts of their graphs in common. The locality property of MFVB approximation suggests that the algorithms for the two models share some of the updates.

2.8.1 Approximate Bayesian inference via non-conjugate variational message passing

In generalised response regression, the Poisson distribution is often bracketed with the Bernoulli distribution since both belong to the one-parameter exponential family. However, variational approximations for the Poisson response models are not the same as those with the Bernoulli response models. For the Bayesian logistic mixed models, Jaakkola and

Initialise: $\nu = 2$, $\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})}$, a $q^{\mathrm{R}} \times q^{\mathrm{R}}$ positive definite matrix, $\mu_{q(1/a_r^{\mathrm{R}})} > 0$, $1 \leq r \leq q^{\mathrm{R}}$, $\mu_{q(1/\sigma_{u\ell}^2)} > 0$, $1 \leq \ell \leq L$ and $\boldsymbol{\xi}$, a $(\sum_{i=1}^m n_i) \times 1$ vector of positive entries.

Cycle through updates:

$$\textbf{Define: } \boldsymbol{M}_{q(\boldsymbol{\Omega}^{-1})} \equiv \begin{bmatrix} \sigma_\beta^{-2}\boldsymbol{I}_p & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \underset{1\leq\ell\leq L}{\text{blockdiag}}\left(\mu_{q(1/\sigma_{u\ell}^2)}\boldsymbol{I}_{q_\ell^{\mathrm{G}}}\right) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{I}_m \otimes \boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})} \end{bmatrix}$$

1. Update multivariate normal $q^*(\boldsymbol{\beta}, \boldsymbol{u})$ and $\boldsymbol{\xi}$ parameters:

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \left[2\,\boldsymbol{C}^{\mathsf{T}}\text{diag}\{\lambda(\boldsymbol{\xi})\}\boldsymbol{C} + \boldsymbol{M}_{q(\boldsymbol{\Omega}^{-1})}\right]^{-1}$$

$$\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^{\mathsf{T}}(\boldsymbol{y} - \tfrac{1}{2}\mathbf{1})$$

$$\boldsymbol{\xi} \leftarrow \sqrt{\text{diagonal}\left\{\boldsymbol{C}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}^{\mathsf{T}}\right)\boldsymbol{C}^{\mathsf{T}}\right\}}$$

2. Update inverse-Gamma $q^*(a_{u\ell})$ and inverse-Gamma $q^*(\sigma_{u\ell}^2)$ parameters:

For $\ell = 1, \ldots, L$:

$$\mu_{q(1/a_{u\ell})} \leftarrow 1/(\mu_{q(1/\sigma_{u\ell}^2)} + A_{u\ell}^{-2})$$

$$\mu_{q(\sigma_{u\ell}^2)} \leftarrow \frac{q_\ell^{\mathrm{G}} + 1}{||\boldsymbol{\mu}_{q(\boldsymbol{u}_\ell^{\mathrm{G}})}||^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_\ell^{\mathrm{G}})}) + 2\,\mu_{q(1/a_{u\ell})}}$$

3. Update inverse-Gamma $q^*(a_r^{\mathbf{R}})$ and inverse-Wishart $q^*(\boldsymbol{\Sigma}^{\mathbf{R}})$parameters:

For $r = 1, \ldots, q^{\mathrm{R}}$:

$$B_{q(a_r^{\mathrm{R}})} \leftarrow \nu(\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})})_{rr} + A_{\mathrm{R}r}^{-2} \quad ; \quad \mu_{q(a_r^{\mathrm{R}})} \leftarrow \tfrac{1}{2}(\nu + q^{\mathrm{R}})/B_{q(a_r^{\mathrm{R}})}$$

$$\boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\mathrm{R}})} \leftarrow \sum_{i=1}^{m}(\boldsymbol{\mu}_{q(\boldsymbol{u}_i^{\mathrm{R}})}\boldsymbol{\mu}_{q(\boldsymbol{u}_i^{\mathrm{R}})}^{\mathsf{T}} + \boldsymbol{\Sigma}_{q(\boldsymbol{u}_i^{\mathrm{R}})}) + 2\,\nu\,\text{diag}(\mu_{q(1/a_1^{\mathrm{R}})}, \ldots, \mu_{q(1/a_{q^{\mathrm{R}}}^{\mathrm{R}})})$$

$$\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})} \leftarrow (\nu + m + q^{\mathrm{R}} - 1)\;\boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\mathrm{R}})}^{-1}$$

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

Algorithm 4: Mean field variational Bayes algorithm for obtaining model parameters in the optimal q-density functions for the two-level Bayesian semiparametric mixed model with Bernoulli response.

Jordan (1997) derive a lower bound on the marginal likelihood that leads to tractable approximate inference. The Albert and Chib (1993) auxiliary variable representation of Bayesian probit regression leads to a different type of variational approximation for the binary outcomes. There do not appear to be analogues of these approaches for the Poisson regression and therefore a different technique is explored.

We begin with the MFVB approximation such that the actual joint posterior density function is approximated by the product restricted form

$$p(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\Sigma}^{\mathrm{R}}, \boldsymbol{a}^{\mathrm{R}}, \boldsymbol{\sigma}_u^2, \boldsymbol{a}_u|\boldsymbol{y}) \approx q(\boldsymbol{\beta}, \boldsymbol{u})\, q(\boldsymbol{\Sigma}^{\mathrm{R}}, \boldsymbol{\sigma}_u^2)\, q(\boldsymbol{a}^{\mathrm{R}}, \boldsymbol{a}_u). \tag{2.23}$$

However, the optimal q-density function $q^*(\boldsymbol{\beta}, \boldsymbol{u})$ under (2.23) satisfies

$$q^*(\boldsymbol{\beta}, \boldsymbol{u}) \propto \exp\left[E_{q(-(\boldsymbol{\beta},\boldsymbol{u}))}\left\{\log p(\boldsymbol{\beta}, \boldsymbol{u}|\text{Markov blanket of } (\boldsymbol{\beta}, \boldsymbol{u}))\right\}\right],$$

which involves non-closed form multivariate integrals and therefore is not tractable. A remedy in this instance is to postulate $q(\boldsymbol{\beta}, \boldsymbol{u})$ to be a special case of the exponential family density function, in this case multivariate normal

$$p(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\Sigma}^{\mathrm{R}}, \boldsymbol{a}^{\mathrm{R}}, \boldsymbol{\sigma}_u^2, \boldsymbol{a}_u|\boldsymbol{y}) \approx q(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})\, q(\boldsymbol{\Sigma}^{\mathrm{R}}, \boldsymbol{\sigma}_u^2)\, q(\boldsymbol{a}^{\mathrm{R}}, \boldsymbol{a}_u),$$

where

$$q(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}) \text{ is the } N(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}) \text{ density function.} \tag{2.24}$$

Recall that $\boldsymbol{\upsilon} = \begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}$ and express (2.24) as an exponential family form gives

$$p(\boldsymbol{\upsilon}) = \exp\left\{\boldsymbol{T}(\boldsymbol{\upsilon})^{\mathsf{T}}\boldsymbol{\eta} - A(\boldsymbol{\eta}) - \tfrac{d}{2}\log(2\pi)\right\},$$

where

$$\begin{aligned} \boldsymbol{T}(\boldsymbol{\upsilon}) &\equiv \begin{bmatrix}\boldsymbol{\upsilon}\\ \operatorname{vech}(\boldsymbol{\upsilon}\boldsymbol{\upsilon}^{\mathsf{T}})\end{bmatrix}, \quad \boldsymbol{\eta} \equiv \begin{bmatrix}\boldsymbol{\Sigma}^{-1}\boldsymbol{u}\\ -\frac{1}{2}\boldsymbol{D}_d^T\operatorname{vec}(\boldsymbol{\Sigma}^{-1})\end{bmatrix}\\ \text{and} \quad A(\boldsymbol{\eta}) &= -\tfrac{1}{4}\boldsymbol{\eta}_1^{\mathsf{T}}\left\{\operatorname{vec}^{-1}(\boldsymbol{D}_2^{+\mathsf{T}}\boldsymbol{\eta}_2)\right\}^{-1} - \tfrac{1}{2}\log\left|-2\operatorname{vec}^{-1}(\boldsymbol{D}_d^{+\mathsf{T}}\boldsymbol{\eta}_2)\right| \end{aligned}$$

are the natural statistic, natural parameter and log-partition function respectively. The optimal q-density function $q^*(\boldsymbol{\beta}, \boldsymbol{u})$ can instead be found using

$$\boldsymbol{\eta} \leftarrow [\operatorname{var}\{\boldsymbol{T}(\boldsymbol{\upsilon})\}]^{-1}\left\{\mathsf{D}_{\boldsymbol{\eta}}E_{q(\boldsymbol{\Omega},\boldsymbol{\upsilon})}\left\{\log p(\boldsymbol{y}, \boldsymbol{\Omega}, \boldsymbol{\upsilon})\right\}\right\}, \tag{2.25}$$

where $\boldsymbol{\Omega} \in \{\boldsymbol{\Sigma}^{\mathrm{R}}, \boldsymbol{a}^{\mathrm{R}}, \boldsymbol{\sigma}_u^2, \boldsymbol{a}_u\}$. This approach, first introduced in Knowles and Minka

(2011), has been labelled as the *non-conjugate variational message passing* since it offers a way of circumventing non-conjugacies in ordinary MFVB. The exponential family distribution parameters are chosen via fixed-point iteration, which we describe in Theorem 2.8.1 following a straight adaption from Rhode and Wand (2015). Wand (2014) provides fully simplified fixed point update expressions for the multivariate normal q-density parameters, in terms of derivatives with respect to the corresponding mean and covariance matrix:

$$\begin{aligned} \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} &\leftarrow \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \left(\mathsf{D}_{\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}} \mathsf{E}_{q(\boldsymbol{\Omega},\boldsymbol{\beta},\boldsymbol{u})} \left\{\log p(\boldsymbol{y},\boldsymbol{\Omega},\boldsymbol{\beta},\boldsymbol{u})\right\}\right)^{\mathsf{T}}, \\ \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} &\leftarrow -\left(\mathsf{H}_{\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}} \mathsf{E}_{q(\boldsymbol{\Omega},\boldsymbol{\beta},\boldsymbol{u})} \left\{\log p(\boldsymbol{y},\boldsymbol{\Omega},\boldsymbol{\beta},\boldsymbol{u})\right\}\right)^{-1}, \end{aligned} \tag{2.26}$$

where H denotes the Hessian matrix as defined in Definition 1.9.2. The term $\mathsf{D}_{\boldsymbol{x}} f$ is the derivative factor of f with respect to $\boldsymbol{x}$.

Theorem 2.8.1. The aim of fixed-point iterations is to find points, denoted by $\boldsymbol{x}^*$, that satisfy the following stationary point condition

$$\mathsf{D}_{\boldsymbol{x}} f(\boldsymbol{x})^{\mathsf{T}} = \mathbf{0}, \tag{2.27}$$

assuming that $\mathsf{D}_{\boldsymbol{x}} f(\boldsymbol{x})$ exists. Such points are considered as the maxima or minima of f. Equation (2.27) can be rewritten in the form $\boldsymbol{x} = \boldsymbol{g}(\boldsymbol{x})$ for some function $\boldsymbol{g} : D \subseteq \mathbb{R}^d \leftarrow \mathbb{R}^d$. Given $\boldsymbol{g}$, fixed-point iterations simply involve repeated evaluation of $\boldsymbol{g}$ as given in Algorithm 5.

Initialise: $\boldsymbol{x} \leftarrow \boldsymbol{x}_{\text{init}}$ for some $\boldsymbol{x}_{\text{init}} \in D$.

Cycle:

$\boldsymbol{x} \leftarrow \boldsymbol{g}(\boldsymbol{x})$

until convergence.

Algorithm 5: The fixed-point iteration algorithm in generic form.

Similar to Subsection 2.5.1, the q-densities are subject to the following product form:

$$q(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{a}^{\mathrm{R}},\boldsymbol{a}_u,\boldsymbol{\Sigma}^{\mathrm{R}},\boldsymbol{\sigma}_u^2) = q(\boldsymbol{\beta},\boldsymbol{u})\, q(\boldsymbol{\Sigma}^{\mathrm{R}}) \left\{\prod_{r=1}^{q^{\mathrm{R}}} q(a_r^{\mathrm{R}})\right\} \left\{\prod_{\ell=1}^{L} q(\sigma_{u\ell}^2)\right\} \left\{\prod_{\ell=1}^{L} q(a_{u\ell})\right\}. \tag{2.28}$$

As shown through the derivations in Appendix 2.D, the optimal q-densities admit the

following forms:

$$
\begin{aligned}
&q^*(\boldsymbol{\beta}, \boldsymbol{u}\,; \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}) \text{ is the } N(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})},\, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}) \text{ density function,}\\
&q^*(\sigma^2_{u\ell}) \text{ is the Inverse-Gamma } \left(\tfrac{1}{2}\,(q^{\mathrm{G}}_\ell + 1),\, B_{q(\sigma^2_{u\ell})}\right) \text{ density function,}\\
&q^*(a_{u\ell}) \text{ is the Inverse-Gamma}(1,\, B_{q(a_{u\ell})}) \text{ density function,}\\
&q^*(\boldsymbol{\Sigma}^{\mathrm{R}}) \text{ is the Inverse-Wishart } \left(\nu + m + q^{\mathrm{R}} + 1,\, \boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\mathrm{R}})}\right) \text{ density function,}\\
&q^*(a^{\mathrm{R}}_r) \text{ is the Inverse-Gamma } \left(\tfrac{1}{2}(\nu + q^{\mathrm{R}}),\, B_{q(a^{\mathrm{R}}_r)}\right) \text{ density function,}
\end{aligned}
$$

for parameters $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$, the mean vector and covariance matrix of $q^*(\boldsymbol{\beta}, \boldsymbol{u})$, $B_{q(\sigma^2_{u\ell})}$, the scale parameter of $q^*(\sigma^2_{u\ell})$, $B_{q(a_{u\ell})}$, the scale parameter of $q^*(a_{u\ell})$, $B_{q(a^{\mathrm{R}}_r)}$, the scale parameter of $q^*(a^{\mathrm{R}}_r)$ and $\boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\mathrm{R}})}$, the scale parameter of $q^*(\boldsymbol{\Sigma}^{\mathrm{R}})$.

The interdependencies between the parameters in these optimal q-density functions, combined with the updates for $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ given by (2.26), give rise to an iterative scheme for their solution and is encompassed in Algorithm 6. Algorithm 6 also uses the variational lower bound on the marginal log-likelihood. For model (2.22) under product restriction (2.28) it has the explicit expression as follows:

$$
\begin{aligned}
\log \underline{p}(\boldsymbol{y}; q) \;=\;& \tfrac{1}{2}\, q^{\mathrm{R}}(\nu + q^{\mathrm{R}} - 1)\log(2\nu) - \left(\tfrac{1}{2}q^{\mathrm{R}} + L\right)\log(\pi)\\
&+ \boldsymbol{y}^{\mathsf{T}}\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} - \mathbf{1}^{\mathsf{T}}\exp\{\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \tfrac{1}{2}\mathrm{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})\boldsymbol{C}^{\mathsf{T}}\}\\
&- \mathbf{1}^{\mathsf{T}}\log(\boldsymbol{y}!) + \left(\boldsymbol{y} - \tfrac{1}{2}\mathbf{1}\right)^{\mathsf{T}}\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})} + \tfrac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}|\\
&- \tfrac{1}{2}p\log(\sigma^2_\beta) - \tfrac{1}{2}\sigma^{-2}_\beta\left(\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\right)\\
&- \log(\mathcal{C}_{q^{\mathrm{R}},\nu+q^{\mathrm{R}}-1}) + \log(\mathcal{C}_{q^{\mathrm{R}},\nu+m+q^{\mathrm{R}}-1})\\
&- \tfrac{1}{2}\,(\nu + m + q^{\mathrm{R}} - 1)\log|B_{q(\boldsymbol{\Sigma}^{\mathrm{R}})}|\\
&+ \textstyle\sum_{\ell=1}^{L}\log\Gamma\left\{\tfrac{1}{2}\,(q^{\mathrm{G}}_\ell + 1)\right\} - \tfrac{1}{2}\,\sum_{\ell=1}^{L}(q^{\mathrm{G}}_\ell + 1)\,\log(B_{q(\sigma^2_{u\ell})})\\
&+ q^{\mathrm{R}}\log\Gamma\left\{\tfrac{1}{2}(\nu + q^{\mathrm{R}})\right\} + \textstyle\sum_{r=1}^{q^{\mathrm{R}}}\nu(\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})})_{rr}\,\mu_{q(1/a^{\mathrm{R}}_r)}\\
&- \tfrac{1}{2}\,(\nu + q^{\mathrm{R}})\,\textstyle\sum_{r=1}^{q^{\mathrm{R}}}\log(B_{q(a^{\mathrm{R}}_r)}) - \sum_{\ell=1}^{L}\log(A_{u\ell})\\
&- \textstyle\sum_{\ell=1}^{L}\log\left(B_{q(a_{u\ell})}\right) + \sum_{\ell=1}^{L}\mu_{q(1/a_{u\ell})}\mu_{q(1/\sigma^2_{u\ell})}\\
&+ \tfrac{1}{2}\left(\textstyle\sum_{\ell=1}^{L}q^{\mathrm{G}}_\ell + p + m\right) - \sum_{r=1}^{q^{\mathrm{R}}}\log(A_{\mathrm{R}r}),
\end{aligned}
$$

where $\boldsymbol{\mu}_{q(\boldsymbol{\beta})}$ is defined to be the subvector of $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ corresponding to $\boldsymbol{\beta}$ and similarly, $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}$ is the sub-matrix of $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ corresponding to $\boldsymbol{\beta}$. The terms $\mathcal{C}_{q^{\mathrm{R}},\nu+q^{\mathrm{R}}-1}$ and $\mathcal{C}_{q^{\mathrm{R}},\nu+m+q^{\mathrm{R}}-1}$ are defined in Table 1.1. Details on the derivation for the optimal q-densities and lower bound expression are deferred to Appendix 2.D. It is worth mentioning that, unlike ordinary MFVB, there is no guarantee that there will be an increase at each iteration for this variational lower bound. Therefore, we continue the cycle of updates once the absolute relative change in its logarithm falls below a negligible amount.

Initialise: $\nu = 2$, $\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})}$, a $q^{\mathrm{R}} \times q^{\mathrm{R}}$ positive definite matrix, $\mu_{q(1/a_r^{\mathrm{R}})} > 0$, $1 \leq r \leq q^{\mathrm{R}}$, $\mu_{q(1/\sigma^2_{u\ell})} > 0$, $1 \leq \ell \leq L$, $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$, a $(p + q^{\mathrm{R}}m + \sum_{l=1}^{L} q_\ell^{\mathrm{G}}) \times 1$ vector and $\boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})}$, a $(\sum_{i=1}^{m} n_i) \times 1$ vector of positive entries.

Cycle through updates:

$$\textbf{Define: } \boldsymbol{M}_{q(\boldsymbol{\Omega}^{-1})} \equiv \begin{bmatrix} \sigma_\beta^{-2}\boldsymbol{I}_p & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \underset{1\leq\ell\leq L}{\text{blockdiag}}\left(\mu_{q(1/\sigma^2_{u\ell})}\boldsymbol{I}_{q_\ell^{\mathrm{G}}}\right) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{I}_m \otimes \boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})} \end{bmatrix}$$

Update multivariate normal $q^*(\boldsymbol{\beta},\boldsymbol{u})$ and $\boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ parameters:

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \left\{\boldsymbol{C}^{\mathsf{T}}\text{diag}(\boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})})\boldsymbol{C} + \boldsymbol{M}_{q(\boldsymbol{\Omega}^{-1})}\right\}^{-1}$$

$$\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\left\{\boldsymbol{C}^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})}) - \boldsymbol{M}_{q(\boldsymbol{\Omega}^{-1})}\,\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\right\}$$

$$\boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \exp\left\{\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \tfrac{1}{2}\text{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^{\mathsf{T}})\right\}$$

2. Update inverse-Gamma $q^*(a_{u\ell})$ and inverse-Gamma $q^*(\sigma^2_{u\ell})$ parameters:

For $\ell = 1, \ldots, L$:

$$\mu_{q(1/a_{u\ell})} \leftarrow 1/(\mu_{q(1/\sigma^2_{u\ell})} + A_{u\ell}^{-2})$$

$$\mu_{q(\sigma^2_{u\ell})} \leftarrow \frac{q_\ell^{\mathrm{G}} + 1}{||\boldsymbol{\mu}_{q(\boldsymbol{u}_\ell^{\mathrm{G}})}||^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_\ell^{\mathrm{G}})}) + 2\,\mu_{q(1/a_{u\ell})}}$$

3. Update inverse-Gamma $q^*(a_r^{\mathbf{R}})$ and inverse-Wishart $q^*(\boldsymbol{\Sigma}^{\mathbf{R}})$parameters:

For $r = 1, \ldots, q^{\mathrm{R}}$:

$$B_{q(a_r^{\mathrm{R}})} \leftarrow \nu(\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})})_{rr} + A_{\mathrm{R}r}^{-2} \quad ; \quad \mu_{q(a_r^{\mathrm{R}})} \leftarrow \tfrac{1}{2}(\nu + q^{\mathrm{R}})/B_{q(a_r^{\mathrm{R}})}$$

$$\boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\mathrm{R}})} \leftarrow \sum_{i=1}^{m}(\boldsymbol{\mu}_{q(\boldsymbol{u}_i^{\mathrm{R}})}\boldsymbol{\mu}_{q(\boldsymbol{u}_i^{\mathrm{R}})}^{\mathsf{T}} + \boldsymbol{\Sigma}_{q(\boldsymbol{u}_i^{\mathrm{R}})}) + 2\,\nu\,\text{diag}(\mu_{q(1/a_1^{\mathrm{R}})}, \ldots, \mu_{q(1/a_{q^{\mathrm{R}}}^{\mathrm{R}})})$$

$$\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})} \leftarrow (\nu + m + q^{\mathrm{R}} - 1)\;\boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\mathrm{R}})}^{-1}$$

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

Algorithm 6: Mean field variational Bayes algorithm for obtaining model parameters in the optimal q-density functions for the two-level Bayesian semiparametric mixed model with Poisson response.

2.9 Displaying approximate posterior means of regression function fits

The majority of this chapter describes the process for obtaining approximating posterior distributions for various mixed model parameters. In this section we translate these into meaningful graphical displays consisting of fitted curves and their corresponding variability bands.

Essentially, we plot for each predictor, the profile of the response outcome with each of the other predictors set at their mean value. In practice, we set up a grid over the domain of each predictor. Take the Gaussian response model with $L = 1$ as an example, at the global level, this would involve plotting the overall mean of the penalised spline regression function

$$\hat{h}(\bar{x}; s) = \hat{\beta}_0 + \hat{\beta}_x \bar{x} + \hat{\beta}_s \, s + \sum_{k=1}^{q^{\mathrm{G}}} u_k^{\mathrm{G}} z_k(s).$$

Here we simply ignore the random effects terms that represent group-specific departures from that overall mean $\hat{h}(\bar{x}; s)$.

Penalised regression splines lend themselves to fairly straightforward pointwise credible bands as we now describe. Suppose that we want a pointwise credible band for function $\hat{h}(\bar{x}; s)$ over a grid of M s-values, i.e. $\boldsymbol{g} = (g_{s1}, \ldots, g_{sM})$. Define $\hat{h}_g$ to be the MFVB approximate posterior mean of the penalised spline function over $\boldsymbol{g}$ and $\mathrm{Cov}_q(\cdot)$ be the q-density covariance matrix corresponding to $(\boldsymbol{\beta}, \boldsymbol{u}^{\mathrm{G}})$:

$$\hat{h}_g \equiv \begin{bmatrix} \hat{h}_g(\bar{x}, g_{s1}) \\ \vdots \\ \hat{h}_g(\bar{x}, g_{sM}) \end{bmatrix} = \boldsymbol{C}_g \boldsymbol{\mu}_{q(\boldsymbol{\beta}, \boldsymbol{u}^{\mathrm{G}})} \quad \text{and} \quad \mathrm{Cov}_q \left(\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u}^{\mathrm{G}} \end{bmatrix} \right) = \begin{bmatrix} \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} & \boldsymbol{\Lambda}_{q(\boldsymbol{\beta}, \boldsymbol{u}^{\mathrm{G}})} \\ \boldsymbol{\Lambda}_{q(\boldsymbol{\beta}, \boldsymbol{u}^{\mathrm{G}})}^{\mathsf{T}} & \boldsymbol{\Sigma}_{q(\boldsymbol{u}^{\mathrm{G}})} \end{bmatrix},$$

where $\boldsymbol{\mu}_{q(\boldsymbol{\beta}, \boldsymbol{u}^{\mathrm{G}})}$ is defined to be the subvector of $\boldsymbol{\mu}_{q(\boldsymbol{\beta}, \boldsymbol{u})}$ corresponding to $\boldsymbol{\beta}$ and $\boldsymbol{u}^{\mathrm{G}}$, and the remaining terms are defined analogously. Variability bands in $\hat{h}_g$ can then be obtained from

$$\mathrm{Var}_q(\hat{h}_g) = \boldsymbol{C}_g \mathrm{Cov}_q \left(\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u}^{\mathrm{G}} \end{bmatrix} \right) \boldsymbol{C}_g^{\mathsf{T}},$$

where

$$\boldsymbol{C}_g \equiv \begin{bmatrix} 1 & \bar{x} & g_{s1} & z_1(g_{s1}) & \cdots & z_{q^{\mathrm{G}}}(g_{s1}) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \bar{x} & g_{sM} & z_1(g_{sM}) & \cdots & z_{q^{\mathrm{G}}}(g_{sM}) \end{bmatrix}.$$

The MFVB approximation of $\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}^{\mathrm{G}}\end{bmatrix}$, i.e. $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})}$ as presented previously, has a multivariate normal distribution. Thus, our problem reduced to finding the 0.025 and 0.975 quantiles of a multivariate normal at each point on our grid. It follows that an approximate $100(1-\alpha)\%$ credible interval for $\hat{h}_g$ is

$$\hat{h}_g \pm z(1-\alpha/2)\sqrt{\mathrm{Var}_q(\hat{h}_g)},$$

and the display of this fit is achieved by plotting $\hat{h}_g$ against $\boldsymbol{g}$. The variability estimates in the group-specific mean function estimates can be obtained in a similar vein.

2.10 Numerical evaluation

We conducted a series of comprehensive simulation studies to assess the performance of Algorithms 3, 4 and 6 in terms of inferential accuracy, credible interval coverage and computational speed. We generated 30 datasets according to the following simulation settings:

Gaussian response

$$\begin{aligned}
&y_{ij}|\beta_0,\beta_x,u^{\mathrm{R}}_{0i},u^{\mathrm{R}}_{1i},\sigma^2_\varepsilon \overset{\text{ind.}}{\sim} N\left(\beta_0+u^{\mathrm{R}}_{0i}+(\beta_x+u^{\mathrm{R}}_{1i})\,x_{ij}+f(s_{ij}),\,\sigma^2_\varepsilon\right),\\
\text{where}\quad &f(s)=1-2.6\,\phi(s;0.15,0.1)-(2.3s-0.07s^2)+0.5\{1-\Phi(s;0.8,0.07)\},\\
&x_{ij}\overset{\text{ind.}}{\sim}\text{Uniform}(0,1),\quad s_{ij}\overset{\text{ind.}}{\sim}\text{Uniform}(0,1),\quad [u^{\mathrm{R}}_{0i}\quad u^{\mathrm{R}}_{1i}]^{\mathsf{T}}\overset{\text{ind.}}{\sim}N(\mathbf{0},\boldsymbol{\Sigma}^{\mathrm{R}}).\\
\text{True values:}\quad &\beta_0=0.58,\quad \beta_x=1.89,\quad \sigma^2_\varepsilon=0.1\quad\text{and}\quad \boldsymbol{\Sigma}^{\mathrm{R}}=\begin{bmatrix}2.58&0.22\\0.22&1.73\end{bmatrix}.
\end{aligned}$$

Student-t response

$$\begin{aligned}
&y_{ij}|\beta_0,\beta_x,u^{\mathrm{R}}_{0i},u^{\mathrm{R}}_{1i},\sigma^2_\varepsilon,\nu \overset{\text{ind.}}{\sim} t\left(\beta_0+u^{\mathrm{R}}_{0i}+(\beta_x+u^{\mathrm{R}}_{1i})\,x_{ij}+f(s_{ij}),\,\sigma^2_\varepsilon,\,\nu\right),\\
\text{where}\quad &f(s)=0.1+\cos(4\pi s),\quad x_{ij}\overset{\text{ind.}}{\sim}\text{Uniform}(0,1),\\
&s_{ij}\overset{\text{ind.}}{\sim}\text{Uniform}(0,1),\quad [u^{\mathrm{R}}_{0i}\quad u^{\mathrm{R}}_{1i}]^{\mathsf{T}}\overset{\text{ind.}}{\sim}N\left(\mathbf{0},\boldsymbol{\Sigma}^{\mathrm{R}}\right).\\
\text{True values:}\quad &\nu=2,\quad \beta_0=-1.58,\quad \beta_x=0.30,\quad \sigma^2_\varepsilon=0.1\quad\text{and}\quad \boldsymbol{\Sigma}^{\mathrm{R}}=\begin{bmatrix}1.91&0.34\\0.34&3.27\end{bmatrix}.
\end{aligned}$$

Bernoulli response

$$\begin{aligned}
&y_{ij}|\beta_0,\beta_x,u^{\mathrm{R}}_{0i},u^{\mathrm{R}}_{1i}\overset{\text{ind.}}{\sim}\text{Bernoulli}\left(\text{logit}^{-1}\left\{\beta_0+u^{\mathrm{R}}_{0i}+(\beta_x+u^{\mathrm{R}}_{1i})\,x_{ij}+f(s_{ij})\right\}\right),\\
\text{where}\quad &f(s)=1-2.6\,\phi(s;0.15,0.1)-(2.3s-0.07s^2)+0.5\{1-\Phi(s;0.8,0.07)\},
\end{aligned}$$

$$x_{ij} \overset{\text{ind.}}{\sim} \text{Uniform}(0,1), \quad s_{ij} \overset{\text{ind.}}{\sim} \text{Uniform}(0,1), \quad [u_{0i}^{\text{R}} \quad u_{1i}^{\text{R}}]^{\mathsf{T}} \overset{\text{ind.}}{\sim} N(\mathbf{0}, \mathbf{\Sigma}^{\text{R}}).$$

True values: $\beta_0 = 0.58, \quad \beta_x = 1.89$ and $\mathbf{\Sigma}^{\text{R}} = \begin{bmatrix} 2.58 & 0.22 \\ 0.22 & 1.73 \end{bmatrix}$.

Poisson response

$$y_{ij}|\beta_0, \beta_x, u_{0i}^{\text{R}}, u_{1i}^{\text{R}} \overset{\text{ind.}}{\sim} \text{Poisson}\left(\exp\left\{\beta_0 + u_{0i}^{\text{R}} + (\beta_x + u_{1i}^{\text{R}})\, x_{ij} + f(s_{ij})\right\}\right),$$

where $f(s) = 0.1 + \cos(4\pi s)$; $\quad x_{ij} \overset{\text{ind.}}{\sim} \text{Uniform}(0,1)$,

$s_{ij} \overset{\text{ind.}}{\sim} \text{Uniform}(0,1), \quad [u_{0i}^{\text{R}} \quad u_{1i}^{\text{R}}]^{\mathsf{T}} \overset{\text{ind.}}{\sim} N(\mathbf{0}, \mathbf{\Sigma}^{\text{R}})$.

True values: $\beta_0 = -1.58, \quad \beta_x = 0.30, \quad \sigma_\varepsilon^2 = 0.1$ and $\mathbf{\Sigma}^{\text{R}} = \begin{bmatrix} 1.91 & 0.34 \\ 0.34 & 3.27 \end{bmatrix}$.

The above settings are for $1 \leq i \leq m$ and $1 \leq j \leq n_i$. The number of groups m is 50 and the within-group sample size n_i ranged between 40 and 50. All hyperparameters are set to 10 000 and lastly we use $q^{\text{G}} = 25$ interior knots when using the O'Sullivan splines with spacing as described in Wand and Ormerod (2008).

2.10.1 Fitting a Bayesian hierarchical model in Stan

We describe in detail the implementation of model (2.11) in Section 2.5 via Stan. Note that the necessary changes for other response models and more complex models are relatively minor. These steps include writing the model in Stan, using Stan to set up the data and starting values, calling Stan and graphing the results.

```
'data
{
   int <lower=1> n;          int <lower=1> m;
   int <lower=1> numKnots;   int <lower=1> idnum[n];
   int <lower=1> ncX;        matrix[n,ncX] X;
   vector[n] y;              matrix[n,numKnots] ZG;
   real x[n];                real <lower=0> sigmaBeta;
   real <lower=1> Aeps;      real <lower=1> Au;
   real <lower=1> Ar;        vector[2] zeroVec;
}
parameters
{
   vector[ncX] beta;         matrix[2,m] uR;
   vector[numKnots] uG;      cov_matrix[2] SigmaR;
   real <lower=0> sigmauG;   real <lower=0> sigmaEps;
   vector <lower=1e-10> [2] aR;
}
transformed parameters
{
```

```
   vector[n] fmean;          real mu[n];
   fmean <- X*beta + ZG*uG;
   for (i in 1:n)
      mu[i] <- (fmean[i] + uR[1,idnum[i]] + uR[2,idnum[i]]*x[i]);
}
model
{
   matrix[2,2] rateForWish;

   y ~ normal(mu,sigmaEps);

   for (j in 1:m)
      col(uR,j) ~ multi_normal(zeroVec,SigmaR);

   rateForWish[1,2] <- 0 ; rateForWish[2,1] <- 0 ;
   for (r in 1:2)
   {
      aR[r] ~ inv_gamma(0.5,pow(Ar,-2));
      rateForWish[r,r] <- 4/aR[r];
   }
   SigmaR ~ inv_wishart(3,rateForWish);

   uG ~ normal(0,sigmauG);
   beta ~ normal(0,sigmaBeta);
   sigmaEps ~ cauchy(0,Aeps);
   sigmauG ~ cauchy(0,Au);
}'
```

The first paragraph of the above code specifies the data: the number of observations, N; the number of groups, m; the response vector, $\boldsymbol{y}$; the number of spline knots, q^{G}; the group identification number, `idnum`; the respective fixed and random effects design matrices, $\boldsymbol{X}$, $\boldsymbol{Z}^{\mathrm{R}}$ and $\boldsymbol{Z}^{\mathrm{G}}$; and the hyperparameters, σ^2_β, A_ε, A_{R} and A_u. Data are labelled as integer or real and can be vectors if dimensions are specified. Data can also be constrained, for example all hyperparameters must be positive. The code next introduces the unknowns to be estimated in the model fit. These are the fixed effects vector, $\boldsymbol{\beta}$; the random effects vectors, $\boldsymbol{u}^{\mathrm{R}}$ and $\boldsymbol{u}^{\mathrm{G}}$; the covariance parameters and matrix, σ^2_{u}, σ^2_ε, $\boldsymbol{\Sigma}^{\mathrm{R}}$; and the auxiliary variables, $\boldsymbol{a}^{\mathrm{R}}$. In addition, we parametrise $\boldsymbol{\mu} = \boldsymbol{X\beta} + \boldsymbol{Zu}$ to be a transformation parameter in order to ensure the sampler runs more efficiently. Finally, we write the model in vector notation, which is clearer and much faster in Stan (by making use of more efficient autodifferentiation). The first line specifics the outcome regression model, followed by the prior distributions for the fixed and random effects parameters.

2.10.2 Assessment of accuracy

For each simulation setting, we fitted each replicated dataset using both the MFVB and MCMC. The MFVB fits were obtained using the respective Algorithms 3, 4 and 6 with the iterations terminated when the relative increase in $\log \underline{p}(\boldsymbol{y}; q)$ fell below 10^{-7}. The MCMC samples were obtained using Stan (Stan Development Team, 2015) with R (R Development Core Team, 2015) interfacing via the `Rstan` package (Stan Development Team, 2015). In each case, MCMC samples of size 10 000 were generated. The first 5000 values of each sample were discarded as burn-in and the remaining 5000 values were thinned by a factor of 5.

The accuracy of MFVB approximation for a generic parameter θ is assessed using the accuracy score defined and justified in Faes *et al.* (2011),

$$\text{accuracy}(q^*(\theta)) = 100 \left(1 - \tfrac{1}{2} \int_{-\infty}^{\infty} |q^*(\theta) - p_{\text{MCMC}}(\theta|\boldsymbol{y})| d\theta\right) \%,$$

This score is derived from the L_1 error, or otherwise known as the *integrated absolute error*. Kernel density estimates, based on the MCMC samples using the binned kernel density estimate `bkde`() function in the R package `KernSmooth` (Wand and Ripley (2006)), were used as a proxy for the exact posterior densities. The default bandwidth is the `dpik`() function in `KernSmooth`, corresponding to the direct plug-in rule. The quality of these proxies dependent upon the MCMC sample sizes and whether or not the posteriors are well-behaved. Nonetheless, it is likely that they are subject to errors inherent in density estimation and bandwidth selection.

Figure 2.7 is an example of summarised MCMC outputs for fitting model (2.11) to the simulated data. Rudimentary diagnostic plots of the MCMC samples given in the first few columns have no discernible pattern, and thus the sampler appears to have mixed well. The kernel density estimated marginal posterior densities of all model parameters, Bayes estimates and 95% credible sets, based on the MCMC samples, are also shown.

Figure 2.8 displays the approximate posterior density functions for the covariance parameters obtained via MFVB and MCMC for a typical realisation from each of the simulation studies. Figure 2.10 displays the side-by-side boxplots of the accuracy scores for the model parameters β_x, $\boldsymbol{\Sigma}^{\text{R}}_{11}$, $\boldsymbol{\Sigma}^{\text{R}}_{12}$, $\boldsymbol{\Sigma}^{\text{R}}_{22}$, σ^2_ε and $f(Q_k)$, $1 \leq k \leq 4$, where Q_k are the kth sample quintiles of the s_{ij}s. The boxplots show that the majority of the accuracy scores exceed 90% and they rarely drop below 85%. The accuracy drops considerably for the covariance parameters corresponding to the Bernoulli response model. This might be due to the limitations of the Jaakkola and Jordan's approximation as described in Subsection 2.7.1. This deficiency can potentially be remedied through a more elaborate variational approximation - for example, one allows posterior dependence between $(\boldsymbol{\beta}, \boldsymbol{u})$ and $\boldsymbol{\Sigma}^{\text{R}}$. However, such a elaboration would bring computational costs, which needs to be

traded off against the importance of making inference about the covariance parameters.

In addition, Figure 2.8 indicates that the MFVB and MCMC penalised splines fits are virtually identical, and when overlaid, are indistinguishable from one another, across all response models.

2.10.3 Assessment of coverage

Another important type of accuracy assessment is the comparison between the advertised coverage of MFVB credible intervals and the actual coverage. Table 2.2 shows the percentages of true parameter coverage for the approximate 95% credible intervals from the MFVB posterior densities with 0.025 probability mass in each tail, based on 1000 MFVB runs. Across all response models the coverage is generally very good and does not fall below 86% for the majority of the model parameters. For the Bernoulli response model, there is some degradation in coverage accuracy for the covariance parameters.

2.10.4 Assessment of speed

To quantify the efficiency gain offered by MFVB, we monitored the time taken for each model to be fitted and the results are summarised in Table 2.1. All computations were performed on a Mac OS X laptop with a 2.6 GHz Intel Core i5 processor and 8 GBytes of random access memory.

As with most speed comparisons, one should be aware of the potential limitations which are: the MFVB and MCMC answers were computed using different programming languages; and both methods have arbitrarily chosen stopping criteria. Despite these caveats, Table 2.1 gives an impression of the relative computational time involved if an "off-the-shelf" MCMC implementation is employed. The results indicate that MFVB is hundred to thousand times faster than MCMC across all models.

Distribution	MCMC	MFVB	Ratio
Gaussian	2768.55 (577.60)	0.52 (0.02)	5324
Student-t	2651.33 (297.09)	13.31 (0.44)	199
Bernoulli	3366.91 (596.45)	29.07 (2.16)	116
Poisson	1061.38 (237.32)	7.60 (1.30)	140

Table 2.1: Average (standard error) elapsed of the computing times in seconds for MCMC and MFVB fitting of the two-level Bayesian semiparametric mixed models to the simulated data.

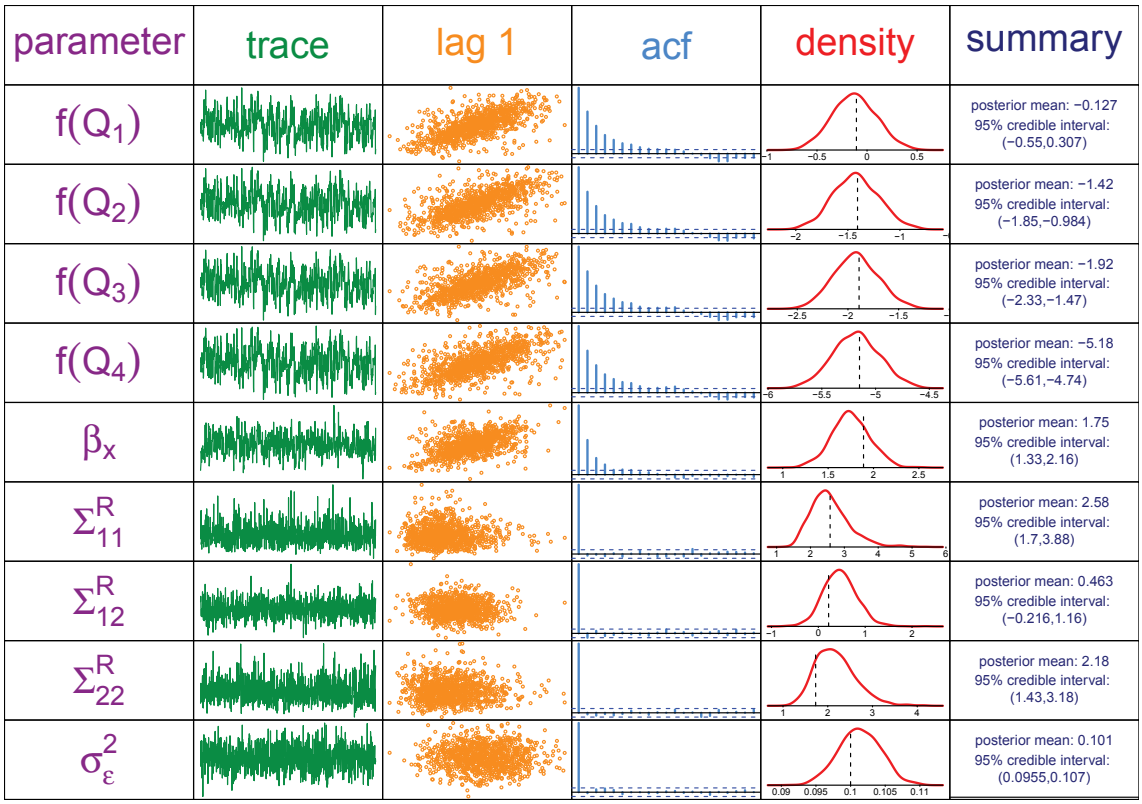

Figure 2.7: Summary of MCMC samples for fitting the two-level Bayesian semiparametric mixed model with Gaussian response to the simulated data. The columns are: parameter name, trace plot of the MCMC sample, plot of sample against its lag-one sample, sample autocorrelation function, kernel density estimate of posterior density function and numerical summaries of posterior density function.

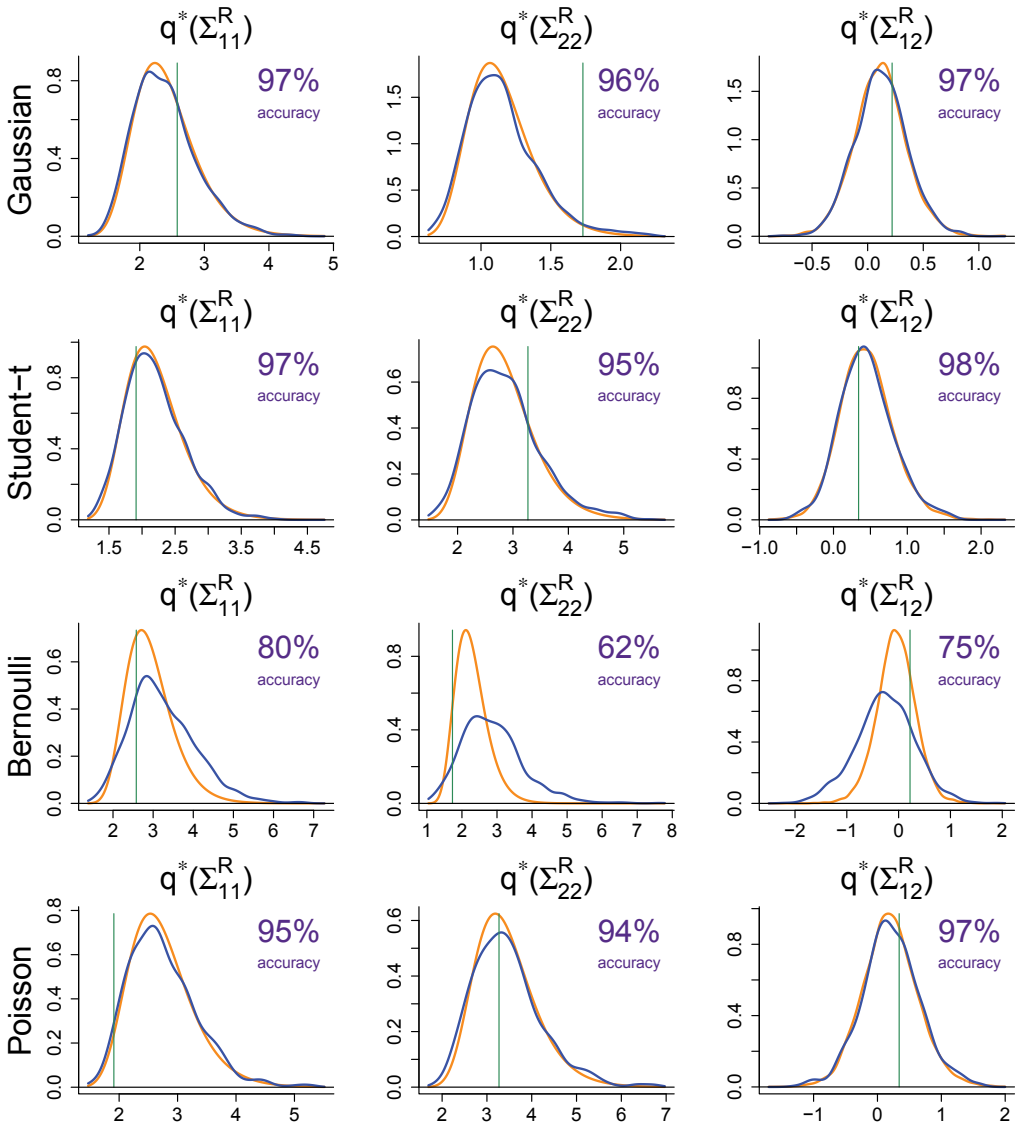

Figure 2.8: Approximate posterior density functions for the covariance parameters obtained via MFVB approximation (orange curves) and MCMC (blue curves) for a single replication from each of the simulation studies described in the text. The green vertical lines represent the true parameter values. The accuracy score on the top right of each plot represents the accuracy of MFVB approximation compared against a MCMC benchmark.

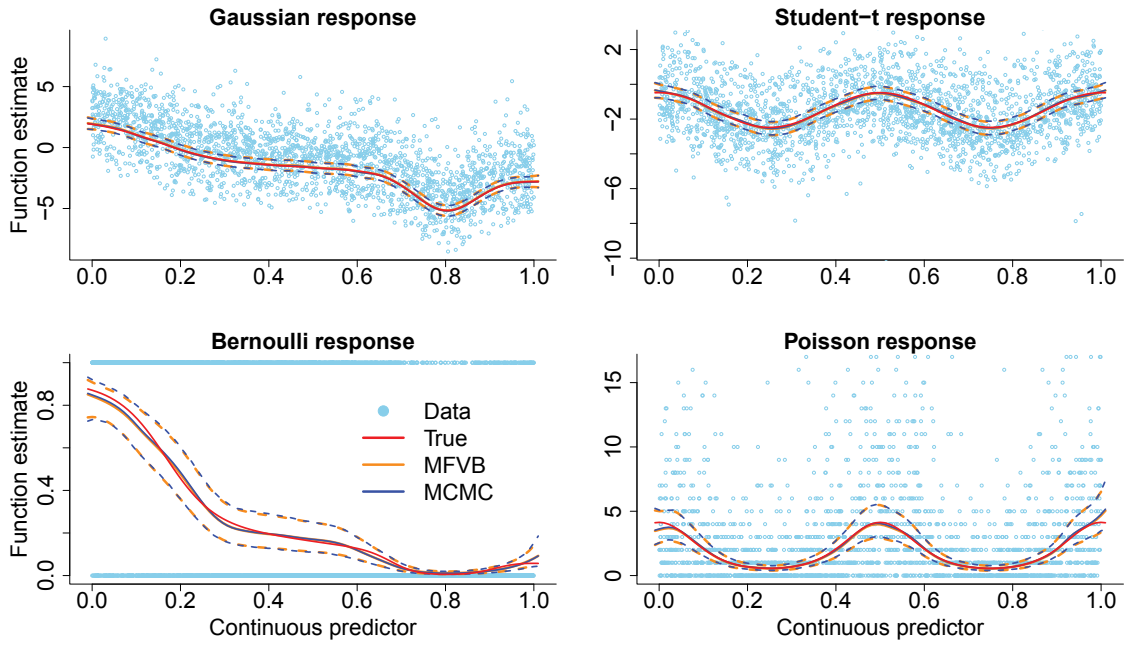

Figure 2.9: Approximate posterior means of the penalised regression functions and corresponding pointwise 95% credible sets obtained via MFVB approximation (orange curves) and MCMC (blue curves) for a single replication of the simulation study described in the text. The red curves represent the true mean functions and the sky blue circles represent the simulated data.

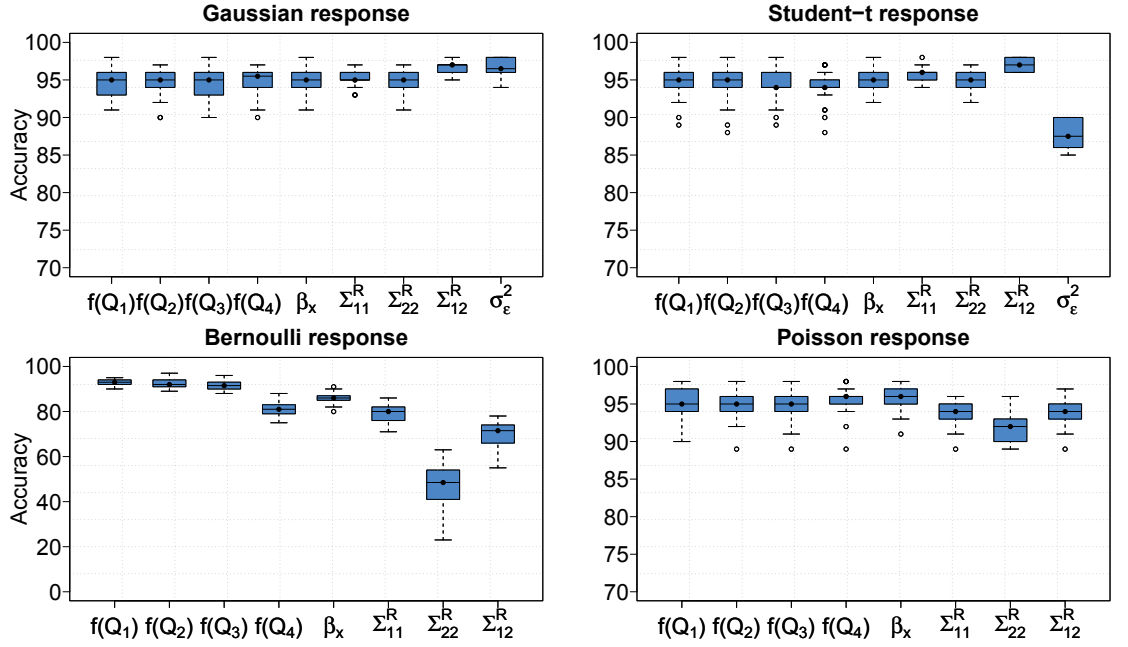

Figure 2.10: Side-by-side boxplots of accuracy scores for MFVB approximation compared against MCMC over 30 runs for the two-level Bayesian semiparametric mixed models.

Parameter	Gaussian	Student-t	Bernoulli	Poisson
$f(Q_1)$	94	98	94	93
$f(Q_2)$	97	97	94	95
$f(Q_3)$	97	93	92	93
$f(Q_4)$	95	96	95	97
β_x	97	93	92	95
$\mathbf{\Sigma}_{11}^{\mathrm{R}}$	95	91	97	94
$\mathbf{\Sigma}_{12}^{\mathrm{R}}$	93	91	67	90
$\mathbf{\Sigma}_{22}^{\mathrm{R}}$	97	99	68	96
σ_{ε}^2	93	86	–	–

Table 2.2: Percentage coverage of true parameter values by approximate 95% credible sets based on the MFVB approximate posterior density functions. The percentages are based on 1000 replications.

2.11 Concluding remarks

Mean field variational Bayes provides a fast and deterministic alternative to MCMC when speed is at a premium. In this chapter we significantly enriched the class of longitudinal and multilevel models which can be handled via MFVB paradigm. The numerical studies in Section 2.10 show that MFVB for various response distributions entails some loss in accuracy for the convenient product restrictions used in our illustrations. Further, particular variational approximations, depending on the problem at hand, can be of hundreds to thousands times faster than the standard MCMC methods. Implementation of these algorithms is often elegant, adding to their practical appeal.

2.A Optimal q-densities derivation for Gaussian semiparametric mixed models

We now derive Algorithms 3, 4 and 6 concerning MFVB fitting of the Bayesian semiparametric mixed models. Throughout these appendices, we use the notation "const" to denote additive constants with respect to the function argument. The MFVB calculations rely primarily on the results for the full conditional density functions.

Recall that

$$N \equiv \sum_{i=1}^{m} n_i, \quad \boldsymbol{v} \equiv \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix}, \quad \boldsymbol{C} \equiv \begin{bmatrix} \boldsymbol{X} & \boldsymbol{Z} \end{bmatrix}, \quad \boldsymbol{C}_{12} \equiv \operatorname*{diag}_{1 \le i \le N}(b_i),$$
$$\text{and} \quad \boldsymbol{\Omega} \equiv \begin{bmatrix} \sigma_\beta^2 \boldsymbol{I}_p & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \operatorname*{blockdiag}_{1 \le \ell \le L}(\sigma_{u\ell}^2 \boldsymbol{I}_{q_\ell^{\mathrm{G}}}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_m \otimes \boldsymbol{\Sigma}^{\mathrm{R}} \end{bmatrix}.$$

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

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

where

$$\begin{aligned} \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} &= \mu_{q(1/\sigma_\varepsilon^2)} \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \boldsymbol{C}^{\mathsf{T}} \boldsymbol{y} \quad \text{and} \\ \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} &= \left(\mu_{q(1/\sigma_\varepsilon^2)} \boldsymbol{C}^{\mathsf{T}} \boldsymbol{C} + \begin{bmatrix} \sigma_\beta^{-2} \boldsymbol{I}_p & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \operatorname*{blockdiag}_{1 \le \ell \le L}(\mu_{q(1/\sigma_{u\ell}^2)} \boldsymbol{I}_{q_\ell^{\mathrm{G}}}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_m \otimes \boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})} \end{bmatrix} \right)^{-1}. \end{aligned}$$

Derivation:
First we note that

$$\begin{aligned} p(\boldsymbol{\beta}, \boldsymbol{u} | \text{rest}) &\propto p(\boldsymbol{\beta}, \boldsymbol{u} | \boldsymbol{y}, \sigma_\varepsilon^2, \boldsymbol{\Sigma}^{\mathrm{R}}, \boldsymbol{\sigma}_u^2) \\ &\propto p(\boldsymbol{y} | \boldsymbol{\beta}, \boldsymbol{u}^{\mathrm{R}}, \sigma_\varepsilon^2, \boldsymbol{\sigma}_u^2)\ p(\boldsymbol{u}^{\mathrm{R}} | \boldsymbol{\Sigma}^{\mathrm{R}})\ p(\boldsymbol{u}^{\mathrm{G}} | \boldsymbol{\sigma}_u^2)\ p(\boldsymbol{\beta}). \end{aligned}$$

Taking logarithms on both sides gives

$$\begin{aligned} \log p(\boldsymbol{\beta}, \boldsymbol{u} | \text{rest}) &= -\tfrac{1}{2\sigma_\varepsilon^2} (\boldsymbol{y} - \boldsymbol{C}\boldsymbol{v})^{\mathsf{T}} (\boldsymbol{y} - \boldsymbol{C}\boldsymbol{v}) - \tfrac{1}{2} \boldsymbol{v}^{\mathsf{T}} \boldsymbol{\Omega}^{-1} \boldsymbol{v} + \text{const} \\ &= -\tfrac{1}{2} \left\{ \boldsymbol{v}^{\mathsf{T}} (\sigma_\varepsilon^{-2} \boldsymbol{C}^{\mathsf{T}} \boldsymbol{C} + \boldsymbol{\Omega}^{-1}) \boldsymbol{v} - 2\, \sigma_\varepsilon^{-2}\, \boldsymbol{v}^{\mathsf{T}} \boldsymbol{C}^{\mathsf{T}} \boldsymbol{y} \right\} + \text{const}, \end{aligned}$$

which then leads to the full conditional

$$\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix} \Bigg| \text{ rest} \sim N\left((\sigma_\varepsilon^{-2}\boldsymbol{C}^\mathsf{T}\boldsymbol{C}+\boldsymbol{\Omega}^{-1})^{-1}\sigma_\varepsilon^{-2}\,\boldsymbol{C}^\mathsf{T}\boldsymbol{y},\,(\sigma_\varepsilon^{-2}\,\boldsymbol{C}^\mathsf{T}\boldsymbol{C}+\boldsymbol{\Omega}^{-1})^{-1}\right).$$

Taking expectations with respect to all parameters except $(\boldsymbol{\beta},\boldsymbol{u})$ gives

$$\begin{aligned}\log q^*(\boldsymbol{\beta},\boldsymbol{u}) &= E_q\left\{\log p(\boldsymbol{\beta},\boldsymbol{u}|\text{rest})\right\}+\text{const}\\ &= -\tfrac{1}{2}\Big\{\boldsymbol{v}^\mathsf{T}\left(\mu_{q(1/\sigma_\varepsilon^2)}\boldsymbol{C}^\mathsf{T}\boldsymbol{C}+\boldsymbol{M}_{q(\boldsymbol{\Omega}^{-1})}\right)\boldsymbol{v}-2\mu_{q(1/\sigma_\varepsilon^2)}\boldsymbol{v}^\mathsf{T}\boldsymbol{C}^\mathsf{T}\boldsymbol{y}\Big\}+\text{const},\end{aligned}$$

where

$$\boldsymbol{M}_{q(\boldsymbol{\Omega}^{-1})}\begin{bmatrix}\sigma_\beta^{-2}\boldsymbol{I}_p & \boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \underset{1\le\ell\le L}{\text{blockdiag}}(\mu_{q(1/\sigma_{u\ell}^2)}\boldsymbol{I}_{q_\ell^{\text{G}}}) & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_m\otimes\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\text{R}})^{-1})}\end{bmatrix}.$$

The stated results follow from standard "completion of square" manipulations.

Expressions for $q^*(\sigma_\varepsilon^2), B_{q(\sigma_\varepsilon^2)}$ and $\mu_{q(1/\sigma_\varepsilon^2)}$

$$q^*(\sigma_\varepsilon^2)\sim\text{Inverse-Gamma}\left(\tfrac{1}{2}\left(\textstyle\sum_{i=1}^m n_i+1\right),\,B_{q(\sigma_\varepsilon^2)}\right),$$

where

$$\begin{aligned}B_{q(\sigma_\varepsilon^2)} &= \tfrac{1}{2}\left\{\|\boldsymbol{y}-\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\|^2+\text{tr}(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^\mathsf{T})\right\}+\mu_{q(1/a_\varepsilon)}\\ \text{and}\quad \mu_{q(\sigma_\varepsilon^2)} &= \tfrac{1}{2}\left(\textstyle\sum_{i=1}^m n_i+1\right)/B_{q(\sigma_\varepsilon^2)}.\end{aligned}$$

Derivation:
First we note that

$$p(\sigma_\varepsilon^2|\text{rest})\propto p(\sigma_\varepsilon^2|\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{u},a_\varepsilon)\propto p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2)\;p(\sigma_\varepsilon^2|a_\varepsilon).$$

Taking logarithms on both sides gives

$$\log p(\sigma_\varepsilon^2|\text{rest})=\left\{-\tfrac{1}{2}\left(\textstyle\sum_{i=1}^m n_i+1\right)-1\right\}\log(\sigma_\varepsilon^2)-\left(\tfrac{1}{2}\|\boldsymbol{y}-\boldsymbol{C}\boldsymbol{v}\|^2+1/a_\varepsilon\right)/\sigma_\varepsilon^2+\text{const},$$

which then leads to the full conditional

$$\sigma_\varepsilon^2|\text{rest} \sim \text{Inverse-Gamma}\left(\tfrac{1}{2}(\textstyle\sum_{i=1}^m n_i + 1),\ \tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{C}\boldsymbol{v}\|^2 + 1/a_\varepsilon\right).$$

Taking expectations with respect to all parameters except σ_ε^2 gives

$$\begin{aligned}\log q^*(\sigma_\varepsilon^2) &= E_q\left\{\log p(\sigma_\varepsilon^2|\text{rest})\right\} + \text{const} \\ &= \left\{-\tfrac{1}{2}(\textstyle\sum_{i=1}^m n_i + 1) - 1\right\}\log(\sigma_\varepsilon^2) - \left\{\tfrac{1}{2}\left(\|\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\|^2 + \text{tr}(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^\mathsf{T})\right)\right. \\ &\quad \left. +\ \mu_{q(1/a_\varepsilon)}\right\}/\sigma_\varepsilon^2 + \text{const}.\end{aligned}$$

The form of $q^*(\sigma_\varepsilon^2)$ and $B_{q(\sigma_\varepsilon^2)}$ follows from Result 1.11.6

$$E_q(\|\boldsymbol{\mu}\|^2) = \|E_q(\boldsymbol{\mu})\|^2 + \text{tr}(\text{Cov}_q(\boldsymbol{\mu})).$$

The expression for $\mu_{q(1/\sigma_\varepsilon^2)}$ follows from elementary manipulations involving inverse-Gamma density functions.

Expressions for $q^*(\sigma_{u\ell}^2), B_{q(\sigma_{u\ell}^2)}$ and $\mu_{q(1/\sigma_{u\ell}^2)}$

For $\ell = 1, \ldots, L$,

$$q^*(\sigma_{u\ell}^2) \sim \text{Inverse-Gamma}(\tfrac{1}{2}\,(q_\ell^\text{G} + 1),\ B_{q(\sigma_{u\ell}^2)}),$$

where

$$B_{q(\sigma_{u\ell}^2)} = \tfrac{1}{2}\left\{\|\boldsymbol{\mu}_{q(\boldsymbol{u}_\ell^\text{G})}\|^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_\ell^\text{G})})\right\} + \mu_{q(1/a_{u\ell})} \quad \text{and} \quad \mu_{q(\sigma_{u\ell}^2)} = \tfrac{1}{2}\,(q_\ell^\text{G} + 1)/B_{q(\sigma_{u\ell}^2)}.$$

Derivation:
First we note that

$$p(\sigma_{u\ell}^2|\text{rest}) \propto p(\sigma_{u\ell}^2|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{u}, a_{u\ell}) \propto p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma_{u\ell}^2)\ p(\sigma_{u\ell}^2|a_{u\ell}).$$

Taking logarithms on both sides gives

$$\log p(\sigma_{u\ell}^2|\text{rest}) = \left\{-\tfrac{1}{2}(q_\ell^\text{G} + 1) - 1\right\}\log(\sigma_{u\ell}^2) - \left(\tfrac{1}{2}\|\boldsymbol{u}_\ell^\text{G}\|^2 + 1/a_{u\ell}\right)/\sigma_{u\ell}^2 + \text{const},$$

which then leads to the full conditional

$$\sigma^2_{u\ell}|\text{rest} \sim \text{Inverse-Gamma}\left(\tfrac{1}{2}(q^{\text{G}}_\ell + 1),\ \tfrac{1}{2}\|\boldsymbol{u}^{\text{G}}_\ell\|^2 + 1/a_{u\ell}\right).$$

Taking expectations with respect to all parameters except $\sigma^2_{u\ell}$ and using Result 1.11.6 gives

$$\begin{aligned}\log q^*(\sigma^2_{u\ell}) &= E_q\left\{\log p(\sigma^2_{u\ell}|\text{rest})\right\} + \text{const}\\ &= \left\{-\tfrac{1}{2}(q^{\text{G}}_\ell + 1) - 1\right\}\log(\sigma^2_{u\ell}) - \left\{\tfrac{1}{2}\left(\|\boldsymbol{\mu}_{q(\boldsymbol{u}^{\text{G}}_\ell)}\|^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}^{\text{G}}_\ell)})\right)\right.\\ &\quad \left. + \mu_{q(1/\sigma^2_{u\ell})}\right\}\Big/\sigma^2_{u\ell} + \text{const}.\end{aligned}$$

Expressions for $q^*(a_\varepsilon), B_{q(a_\varepsilon)}$ and $\mu_{q(1/a_\varepsilon)}$

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

where

$$B_{q(a_\varepsilon)} = \mu_{q(1/\sigma^2_\varepsilon)} + A_\varepsilon^{-2} \quad \text{and} \quad \mu_{q(1/a_\varepsilon)} = 1/\left\{\mu_{q(1/\sigma^2_\varepsilon)} + A_\varepsilon^{-2}\right\}.$$

Derivation:
First we note that

$$p(a_\varepsilon|\text{rest}) \propto p(a_\varepsilon|\sigma^2_\varepsilon) = p(\sigma^2_\varepsilon|a_\varepsilon)\ p(a_\varepsilon).$$

Taking logarithms on both sides gives

$$\log p(a_\varepsilon|\text{rest}) = -2\log(a_\varepsilon) - \left(\sigma_\varepsilon^{-2} + A_\varepsilon^{-2}\right)/a_\varepsilon + \text{const},$$

which then leads to the full conditional

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

Taking expectations with respect to all parameters except a_ε gives

$$\begin{aligned}\log q^*(a_\varepsilon) &= -2\log(a_\varepsilon) - E_q(\sigma_\varepsilon^{-2} + A_\varepsilon^{-2})/a_\varepsilon + \text{const}\\ &= (-1-1)\log(a_\varepsilon) - \left(\mu_{q(1/\sigma^2_\varepsilon)} + 1/A_\varepsilon\right)/a_\varepsilon + \text{const}.\end{aligned}$$

The expressions for $B_{q(a_\varepsilon)}$ and $\mu_{q(1/a_\varepsilon)}$ follow immediately.

Expressions for $q^*(a_{u\ell}), B_{q(a_{u\ell})}$ and $\mu_{q(1/a_{u\ell})}$

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

where

$$B_{q(a_{u\ell})} = \mu_{q(1/\sigma^2_{u\ell})} + A_u^{-2} \quad \text{and} \quad \mu_{q(1/a_{u\ell})} = 1/\{\mu_{q(1/\sigma^2_{u\ell})} + A_u^{-2}\}.$$

Derivation:

The derivation is analogous to that of $q^*(a_\varepsilon)$ and related quantities.

Expressions for $q^*(\boldsymbol{\Sigma}^{\mathrm{R}}), \boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\mathrm{R}})}$ and $\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})}$

$$q^*(\boldsymbol{\Sigma}^{\mathrm{R}}) \sim \text{Inverse-Wishart}\left(\nu + m + q^{\mathrm{R}} - 1,\, \boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\mathrm{R}})}\right),$$

where

$$\begin{aligned}
\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})} &= (\nu + m + q^{\mathrm{R}} - 1)\, \boldsymbol{B}^{-1}_{q(\boldsymbol{\Sigma}^{\mathrm{R}})} \quad \text{and} \\
\boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\mathrm{R}})} &= \sum_{i=1}^{m} (\boldsymbol{\mu}_{q(\boldsymbol{u}^{\mathrm{R}}_i)} \boldsymbol{\mu}^{\mathsf{T}}_{q(\boldsymbol{u}^{\mathrm{R}}_i)} + \boldsymbol{\Sigma}_{q(\boldsymbol{u}^{\mathrm{R}}_i)}) + 2\,\nu\, \mathrm{diag}(\mu_{q(1/a^{\mathrm{R}}_1)}, \ldots, \mu_{q(1/a^{\mathrm{R}}_{q^{\mathrm{R}}})}).
\end{aligned}$$

Derivation:

First we note that

$$p(\boldsymbol{\Sigma}^{\mathrm{R}}|\text{rest}) \propto p(\boldsymbol{\Sigma}^{\mathrm{R}}|\boldsymbol{u}^{\mathrm{R}}, \boldsymbol{a}^{\mathrm{R}}) = p(\boldsymbol{u}^{\mathrm{R}}|\boldsymbol{\Sigma}^{\mathrm{R}})\; p(\boldsymbol{\Sigma}^{\mathrm{R}}|\boldsymbol{a}^{\mathrm{R}}).$$

Taking logarithms on both sides gives

$$\begin{aligned}
\log p(\boldsymbol{\Sigma}^{\mathrm{R}}|\text{rest}) = \; & - \tfrac{1}{2}\left(A_{\boldsymbol{\Sigma}^{\mathrm{R}}} + m + q^{\mathrm{R}} + 1\right) \log(|\boldsymbol{\Sigma}^{\mathrm{R}}|) \\
& - \tfrac{1}{2}\,\mathrm{tr}\left[\left\{\sum_{i=1}^{m}\left(\boldsymbol{u}^{\mathrm{R}}_i (\boldsymbol{u}^{\mathrm{R}}_i)^{\mathsf{T}}\right) + 2\,\nu\,\mathrm{diag}(1/a^{\mathrm{R}}_1, \ldots, 1/a^{\mathrm{R}}_{q^{\mathrm{R}}})\right\} (\boldsymbol{\Sigma}^{\mathrm{R}})^{-1}\right] + \text{const},
\end{aligned}$$

which then leads to the full conditional

$$\boldsymbol{\Sigma}^{\mathrm{R}}|\text{rest} \sim \text{Inverse-Wishart}\left(\nu + m + q^{\mathrm{R}} - 1,\, \sum_{i=1}^{m}\left(\boldsymbol{u}_i^{\mathrm{R}}(\boldsymbol{u}_i^{\mathrm{R}})^{\mathsf{T}}\right) + 2\,\nu\,\mathrm{diag}(1/a_1^{\mathrm{R}}, \ldots, 1/a_{q^{\mathrm{R}}}^{\mathrm{R}})\right).$$

Taking expectations with respect to all parameters except $\boldsymbol{\Sigma}^{\mathrm{R}}$ gives

$$\begin{aligned}
\log q^*(\boldsymbol{\Sigma}^{\mathrm{R}}) = & - \tfrac{1}{2}\left(A_{\boldsymbol{\Sigma}^{\mathrm{R}}} + m + q^{\mathrm{R}} + 1\right)\log(|\boldsymbol{\Sigma}^{\mathrm{R}}|) \\
& - \tfrac{1}{2}\,\mathrm{tr}\left[E_q\left\{\sum_{i=1}^{m}\left(\boldsymbol{u}_i^{\mathrm{R}}(\boldsymbol{u}_i^{\mathrm{R}})^{\mathsf{T}}\right) + 2\,\nu\,\mathrm{diag}(1/a_1^{\mathrm{R}}, \ldots, 1/a_{q^{\mathrm{R}}}^{\mathrm{R}})\right\}(\boldsymbol{\Sigma}^{\mathrm{R}})^{-1}\right] + \text{const} \\
= & - \tfrac{1}{2}\left(A_{\boldsymbol{\Sigma}^{\mathrm{R}}} + m + q^{\mathrm{R}} + 1\right)\log(|\boldsymbol{\Sigma}^{\mathrm{R}}|) \\
& - \tfrac{1}{2}\,\mathrm{tr}\Bigg[\Bigg\{\sum_{i=1}^{m}\left(\boldsymbol{\mu}_{q(\boldsymbol{u}_i^{\mathrm{R}})}\boldsymbol{\mu}_{q(\boldsymbol{u}_{ii}^{\mathrm{R}})}^{\mathsf{T}} + \boldsymbol{\Sigma}_{q(\boldsymbol{u}_i^{\mathrm{R}})}\right) \\
& \qquad + 2\,\nu\,\mathrm{diag}(\mu_{q(1/a_1^{\mathrm{R}})}, \ldots, \mu_{q(1/a_{q^{\mathrm{R}}}^{\mathrm{R}})})\Bigg\}(\boldsymbol{\Sigma}^{\mathrm{R}})^{-1}\Bigg] + \text{const}.
\end{aligned}$$

The expressions for $\boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\mathrm{R}})}$ and $\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})}$ follow immediately.

Expressions for $q^*(a_r^{\mathbf{R}}), B_{q(a_r^{\mathbf{R}})}$ and $\mu_{q(a_r^{\mathbf{R}})}$

For $r = 1, \ldots, q^{\mathrm{R}}$,

$$q^*(a_r^{\mathrm{R}}) \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2}(\nu + q^{\mathrm{R}}), B_{q(a_r^{\mathrm{R}})}),$$

where

$$B_{q(a_r^{\mathrm{R}})} = \nu(M_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})})_{rr} + A_{\mathrm{R}r}^{-2} \quad \text{and} \quad \mu_{q(a_r^{\mathrm{R}})} = \tfrac{1}{2}(\nu + q^{\mathrm{R}})/B_{q(a_r^{\mathrm{R}})}.$$

Derivation:
First we note that

$$p(a_r^{\mathrm{R}}|\text{rest}) \propto p(a_r^{\mathrm{R}}|\boldsymbol{\Sigma}^{\mathrm{R}}) = p(\boldsymbol{\Sigma}^{\mathrm{R}}|a_r^{\mathrm{R}})\; p(a_r^{\mathrm{R}}).$$

Taking logarithms on both sides gives

$$\begin{aligned}
\log p(a_r^{\mathrm{R}}|\text{rest}) \;=\; & -\{\tfrac{1}{2}(\nu + q^{\mathrm{R}} - 1)\}\;(\log(2\nu) + \log(a_r^{\mathrm{R}})) - \{\nu(\boldsymbol{\Sigma}^{\mathrm{R}})^{-1}\}_{rr}/a_r^{\mathrm{R}} \\
& - (a_r^{\mathrm{R}})^{-1}A_{\mathrm{R}r}^{-2},
\end{aligned}$$

which then leads to the full conditional

$$a_r^{\mathrm{R}}|\text{rest} \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2}(\nu+q^{\mathrm{R}}),\, \nu(\boldsymbol{\Sigma}^{\mathrm{R}})_{rr}^{-1} + A_{\mathrm{R}r}^{-2}).$$

Taking expectations with respect to all parameters except a_r^{R} gives

$$\begin{aligned}\log q^*(a_r^{\mathrm{R}}) &= -\{\tfrac{1}{2}(\nu+q^{\mathrm{R}}-1)\}\,(\log(2\nu)+\log(a_r^{\mathrm{R}})) - \nu(\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})})_{rr}/a_r^{\mathrm{R}} \\ &\quad - (a_r^{\mathrm{R}})^{-1}A_{\mathrm{R}r}^{-2}.\end{aligned}$$

The expressions for $B_{q(a_r^{\mathrm{R}})}$ and $\mu_{q(a_r^{\mathrm{R}})}$ follow immediately.

2.A.1 Derivation of the marginal log-likelihood lower bound

The expression for the marginal log-likelihood lower bound is

$$\begin{aligned}\log \underline{p}(\boldsymbol{y};q) &= \tfrac{1}{2}\,q^{\mathrm{R}}(\nu+q^{\mathrm{R}}-1)\log(2\nu) - \tfrac{1}{2}\log(2\pi)\textstyle\sum_{i=1}^m n_i - \left(\tfrac{1}{2}q^{\mathrm{R}}+L+1\right)\log(\pi) \\ &\quad - \tfrac{1}{2}p\log(\sigma_\beta^2) - \tfrac{1}{2}\sigma_\beta^{-2}\left(\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\right) + \tfrac{1}{2}\left(\textstyle\sum_{\ell=1}^L q_\ell^{\mathrm{G}} + p + m\right) \\ &\quad + \tfrac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}| - \log(\mathcal{C}_{q^{\mathrm{R}},\,\nu+q^{\mathrm{R}}-1}) + \log(\mathcal{C}_{q^{\mathrm{R}},\,\nu+m+q^{\mathrm{R}}-1}) \\ &\quad - \tfrac{1}{2}\,(\nu+m+q^{\mathrm{R}}-1)\log|\boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\mathrm{R}})}| \\ &\quad + \textstyle\sum_{\ell=1}^L \log\Gamma\left\{\tfrac{1}{2}\,(q_\ell^{\mathrm{G}}+1)\right\} - \tfrac{1}{2}\,\sum_{\ell=1}^L (q_\ell^{\mathrm{G}}+1)\,\log(B_{q(\sigma_{u\ell}^2)}) \\ &\quad + \log\Gamma\left\{\tfrac{1}{2}(\textstyle\sum_{i=1}^m n_i+1)\right\} - \tfrac{1}{2}\,(\textstyle\sum_{i=1}^m n_i+1)\log\left(B_{q(\sigma_\varepsilon^2)}\right) \\ &\quad - \textstyle\sum_{r=1}^{q^{\mathrm{R}}}\log(A_{\mathrm{R}r}) + q^{\mathrm{R}}\log\Gamma\left\{\tfrac{1}{2}(\nu+q^{\mathrm{R}})\right\} \\ &\quad - \textstyle\sum_{\ell=1}^L\log(A_{u\ell}) - \sum_{\ell=1}^L\log\left(B_{q(a_{u\ell})}\right) + \sum_{\ell=1}^L \mu_{q(1/a_{u\ell})}\mu_{q(1/\sigma_{u\ell}^2)} \\ &\quad + \textstyle\sum_{r=1}^{q^{\mathrm{R}}}\nu(\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})})_{rr}\,\mu_{q(1/a_r^{\mathrm{R}})} - \tfrac{1}{2}\,(\nu+q^{\mathrm{R}})\,\sum_{r=1}^{q^{\mathrm{R}}}\log(B_{q(a_r^{\mathrm{R}})}) \\ &\quad - \log(A_\varepsilon) - \log\left(B_{q(a_\varepsilon)}\right) + \mu_{q(1/a_\varepsilon)}\mu_{q(1/\sigma_\varepsilon^2)}.\end{aligned}$$

Derivation:
The logarithm of the lower bound on the marginal likelihood is given by

$$\begin{aligned}\log \underline{p}(\boldsymbol{y};q) &= E_q\left\{\log p(\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{R}},\boldsymbol{u}^{\mathrm{G}},\boldsymbol{\Sigma}^{\mathrm{R}},\boldsymbol{\sigma}_u^2,\sigma_\varepsilon^2,\boldsymbol{a}^{\mathrm{R}},\boldsymbol{a}_u,a_\varepsilon)\right. \\ &\qquad \left. - \log q^*(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{R}},\boldsymbol{u}^{\mathrm{G}},\boldsymbol{\Sigma}^{\mathrm{R}},\boldsymbol{\sigma}_u^2,\sigma_\varepsilon^2,\boldsymbol{a}^{\mathrm{R}},\boldsymbol{a}_u,a_\varepsilon)\,\right\}.\end{aligned}$$

According to the product restriction (2.12), our expression for the lower bound of the marginal log-likelihood becomes:

$$\log \underline{p}(\boldsymbol{y};q) = E_q\left\{\log p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{R}},\boldsymbol{u}^{\mathrm{G}},\sigma_\varepsilon^2)\right\}$$

$$
\begin{aligned}
&+ E_q\left\{\log p(\boldsymbol{\beta}, \boldsymbol{u}^{\mathrm{R}}, \boldsymbol{u}^{\mathrm{G}}|\boldsymbol{\Sigma}^{\mathrm{R}}, \boldsymbol{\sigma}_u^2) - \log q^*(\boldsymbol{\beta}, \boldsymbol{u}^{\mathrm{R}}, \boldsymbol{u}^{\mathrm{G}})\right\} \\
&+ E_q\left\{\log p(\boldsymbol{\Sigma}^{\mathrm{R}}|\boldsymbol{a}^{\mathrm{R}}) - \log q^*(\boldsymbol{\Sigma}^{\mathrm{R}})\right\} + E_q\left\{p(\boldsymbol{\sigma}_u^2|\boldsymbol{a}_u) - \log q^*(\boldsymbol{\sigma}_u^2)\right\} \\
&+ E_q\left\{\log p(\boldsymbol{a}^{\mathrm{R}}) - \log q^*(\boldsymbol{a}^{\mathrm{R}})\right\} + E_q\left\{\log p(\boldsymbol{a}_u) - \log q^*(\boldsymbol{a}_u)\right\} \\
&+ E_q\left\{p(\sigma_\varepsilon^2|a_\varepsilon) - \log q^*(\sigma_\varepsilon^2)\right\} + E_q\left\{\log p(a_\varepsilon) - \log q^*(a_\varepsilon)\right\}.
\end{aligned}
$$

First note that

$$
\log p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}^{\mathrm{R}}, \boldsymbol{u}^{\mathrm{G}}, \sigma_\varepsilon^2) = \tfrac{N}{2}\log(2\pi) - \tfrac{N}{2}\log(\sigma_\varepsilon^2) - \tfrac{1}{2\sigma_\varepsilon^2}\|\boldsymbol{y} - \boldsymbol{C}\boldsymbol{v}\|^2.
$$

Taking expectations gives

$$
\begin{aligned}
E_q\{\log p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}^{\mathrm{R}}, \boldsymbol{u}^{\mathrm{G}}, \sigma_\varepsilon^2)\} \;=\; & \tfrac{N}{2}\log(2\pi) - \tfrac{N}{2}E_q\{\log(\sigma_\varepsilon^2)\} \\
& - \tfrac{1}{2}\,\mu_{q(1/\sigma_\varepsilon^2)}\left\{\|\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\|^2 + \mathrm{tr}(\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})\right\}.
\end{aligned}
$$

Using the well-established result concerning entropy of a multivariate random normal vector results in

$$
\begin{aligned}
&E_q\{\log p(\boldsymbol{\beta}, \boldsymbol{u}^{\mathrm{R}}, \boldsymbol{u}^{\mathrm{G}}|\boldsymbol{\Sigma}^{\mathrm{R}}, \boldsymbol{\sigma}_u^2) - \log q^*(\boldsymbol{\beta}, \boldsymbol{u}^{\mathrm{R}}, \boldsymbol{u}^{\mathrm{G}})\} = \\
&\quad -\tfrac{p}{2}\log(2\pi) - \tfrac{p}{2}\log(\sigma_\beta^2) - \tfrac{1}{2\sigma_\beta^2}(\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})) \\
&\quad -\tfrac{m}{2}\log(2\pi) - \tfrac{m}{2}\,E_q\{\log(|\boldsymbol{\Sigma}^{\mathrm{R}}|)\} - \tfrac{1}{2}\,(\boldsymbol{I}_m \otimes \boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})})(\|\boldsymbol{\mu}_{q(\boldsymbol{u}^{\mathrm{R}})}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}^{\mathrm{R}})})) \\
&\quad -\tfrac{1}{2}\sum_{\ell=1}^{L}\left\{q_\ell^{\mathrm{G}}\log(2\pi) - \,q_\ell^{\mathrm{G}}E_q\{\log(|\sigma_{u\ell}^2|)\} - \tfrac{1}{2}\,\mu_{q(1/\sigma_{u\ell}^2)}(\|\boldsymbol{\mu}_{q(\boldsymbol{u}_\ell^{\mathrm{G}})}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_\ell^{\mathrm{G}})}))\right\} \\
&\quad -\tfrac{1}{2}\log(|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}|) - \tfrac{1}{2}\,(m + p + \textstyle\sum_{\ell=1}^{L} q_\ell^{\mathrm{G}})\log(2\pi) - \tfrac{1}{2}\,(m + p + \textstyle\sum_{\ell=1}^{L} q_\ell^{\mathrm{G}}).
\end{aligned}
$$

The next term is

$$
\begin{aligned}
&\log p(\boldsymbol{\Sigma}^{\mathrm{R}}|\boldsymbol{a}^{\mathrm{R}}) - \log q^*(\boldsymbol{\Sigma}^{\mathrm{R}}) = \\
&\qquad \log(\mathcal{C}_{q^{\mathrm{R}},v+m-1}) + \log(\mathcal{C}_{q^{\mathrm{R}},v+m+q^{\mathrm{R}}-1}) + \tfrac{1}{2}(v + m - 1)\log(|\boldsymbol{B}_{\boldsymbol{\Sigma}^{\mathrm{R}}}|) \\
&\qquad - \tfrac{1}{2}\,(v + 2m - 1)\log(|\boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\mathrm{R}})}|) + \tfrac{1}{2}\,(v + m + q^{\mathrm{R}})\log(|\boldsymbol{\Sigma}^{\mathrm{R}}|) \\
&\qquad + \tfrac{1}{2}\,(v + 2m + q^{\mathrm{R}})\,\log(|\boldsymbol{\Sigma}^{\mathrm{R}}|) - \,\tfrac{1}{2}\,\mathrm{tr}\left(\boldsymbol{B}_{\boldsymbol{\Sigma}^{\mathrm{R}}}\,(\boldsymbol{\Sigma}^{\mathrm{R}})^{-1}\right) + \mathrm{tr}\left(\boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\mathrm{R}})}\,(\boldsymbol{\Sigma}^{\mathrm{R}})^{-1}\right).
\end{aligned}
$$

Taking expectations gives

$$
\begin{aligned}
&E_q\left\{\log p(\boldsymbol{\Sigma}^{\mathrm{R}}|\boldsymbol{a}^{\mathrm{R}}) - \log q^*(\boldsymbol{\Sigma}^{\mathrm{R}})\right\} = \\
&\qquad \log(\mathcal{C}_{q^{\mathrm{R}},v+m-1}) + \log(\mathcal{C}_{q^{\mathrm{R}},v+2m-1}) + \tfrac{1}{2}(v + m - 1)\log(|\boldsymbol{B}_{\boldsymbol{\Sigma}^{\mathrm{R}}}|) \\
&\qquad - \tfrac{1}{2}\,(v + 2m - 1)\log(|\boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\mathrm{R}})}|) + \tfrac{m}{2}\,E_q\left\{\log(|\boldsymbol{\Sigma}^{\mathrm{R}}|)\right\}
\end{aligned}
$$

$$+\tfrac{1}{2}\,\mathrm{tr}\left\{\left(\boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\mathrm{R}})}-\boldsymbol{B}_{\boldsymbol{\Sigma}^{\mathrm{R}}}\right)\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})}\right\},$$

where

$$\begin{aligned}\tfrac{1}{2}\,\mathrm{tr}\left\{\left(\boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\mathrm{R}})}-\boldsymbol{B}_{\boldsymbol{\Sigma}^{\mathrm{R}}}\right)\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})}\right\} &= \tfrac{1}{2}\,\mathrm{tr}\left\{\sum_{i=1}^{m}\left(\boldsymbol{\mu}_{q(\boldsymbol{u}_i^{\mathrm{R}})}\boldsymbol{\mu}_{q(\boldsymbol{u}_i^{\mathrm{R}})}^{\mathsf{T}}+\boldsymbol{\Sigma}_{q(\boldsymbol{u}_i^{\mathrm{R}})}\right)\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})}\right\}\\ &= \tfrac{1}{2}\left(\boldsymbol{I}_m\otimes\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})}\right)\left(\boldsymbol{\mu}_{q(\boldsymbol{u}^{\mathrm{R}})}^{\mathsf{T}}\boldsymbol{\mu}_{q(\boldsymbol{u}^{\mathrm{R}})}+\boldsymbol{\Sigma}_{q(\boldsymbol{u}^{\mathrm{R}})}\right).\end{aligned}$$

The next term is

$$\begin{aligned}&\log p(\sigma_{u\ell}^2|a_{u\ell})-\log q^*(\sigma_{u\ell}^2)=\\ &\quad\log\left\{\frac{a_{u\ell}^{-0.5}}{\Gamma(0.5)}\,(\sigma_{u\ell}^2)^{-1.5}\exp\left(-\frac{1}{a_{u\ell}\,\sigma_{u\ell}^2}\right)\right\}\\ &\quad-\log\left\{\frac{(B_{q(\sigma_{u\ell}^2)})^{0.5\,(q_\ell^{\mathrm{G}}+1)}}{\Gamma\left\{\frac{1}{2}\,(q_\ell^{\mathrm{G}}+1)\right\}}-(\sigma_{u\ell}^2)^{\left\{-\frac{1}{2}\,(q_\ell^{\mathrm{G}}+1)-1\right\}}\exp\left(-\frac{B_{q(\sigma_{u\ell}^2)}}{\sigma_{u\ell}^2}\right)\right\}.\end{aligned}$$

Taking expectations gives

$$\begin{aligned}&E_q\{\log p(\sigma_{u\ell}^2|a_{u\ell})-\log q^*(\sigma_{u\ell}^2)\}=\\ &\quad-\tfrac{1}{2}\,E_q\left\{\log(a_{u\ell})\right\}-\tfrac{1}{2}\log(\pi)+\log\Gamma\left\{\tfrac{1}{2}\,(q_\ell^{\mathrm{G}}+1)\right\}\\ &\quad+\tfrac{1}{2}\,q_\ell^{\mathrm{G}}\,E_q\left\{\log(\sigma_{u\ell}^2)\right\}+\left(B_{q(\sigma_{u\ell}^2)}-\mu_{q(1/a_{u\ell})}\right)\mu_{q(1/\sigma_{u\ell}^2)}\\ &\quad-\tfrac{1}{2}\,(q_\ell^{\mathrm{G}}+1)\,\log(B_{q(\sigma_{u\ell}^2)}),\end{aligned}$$

where

$$(B_{q(\sigma_{u\ell}^2)}-\mu_{q(1/a_{u\ell})})\mu_{q(1/\sigma_{u\ell}^2)}=\tfrac{1}{2}\,\mu_{q(1/\sigma_{u\ell}^2)}(\|\boldsymbol{u}_\ell^{\mathrm{G}}\|^2+\mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_\ell^{\mathrm{G}})})).$$

Similarly, for

$$\begin{aligned}E_q\{\log p(\sigma_\varepsilon^2|a_\varepsilon)-\log q^*(\sigma_\varepsilon^2)\} &= -\tfrac{1}{2}\,E_q\left\{\log(a_\varepsilon)\right\}-\tfrac{1}{2}\log(\pi)+\log\Gamma\left\{\tfrac{1}{2}\,(N+1)\right\}\\ &\quad+\tfrac{N}{2}\,E_q\left\{\log(\sigma_\varepsilon^2)\right\}+(B_{q(\sigma_\varepsilon^2)}-\mu_{q(1/a_\varepsilon)})\mu_{q(1/\sigma_\varepsilon^2)}\\ &\quad-\tfrac{1}{2}\,(N+1)\log\left\{B_{q(\sigma_\varepsilon^2)}\right\}.\end{aligned}$$

The next term is

$$\log p(a_r^{\mathrm{R}})-\log q^*(a_r^{\mathrm{R}}) = \log\left\{\prod_{r=1}^{q^{\mathrm{R}}}\frac{A_{\mathrm{R}r}^{-1}}{\Gamma(0.5)}\,(a_r^{\mathrm{R}})^{-1.5}\exp\left(\frac{-1}{a_r^{\mathrm{R}}A_{\mathrm{R}r}}\right)\right\}$$

$$- \log \left\{ \prod_{r=1}^{q^{\mathrm{R}}} \frac{(B_{q(a_r^{\mathrm{R}})})^{0.5\,(\nu+2)}}{\Gamma\left\{\frac{1}{2}\,(\nu+2)\right\}} (a_r^{\mathrm{R}})^{\left\{-\frac{1}{2}(\nu+1)-1\right\}} \exp\left(-\frac{B_{q(a_r^{\mathrm{R}})}}{a_r^{\mathrm{R}}}\right) \right\}.$$

Taking expectations gives

$$\begin{aligned} & E_q\{\log p(a_r^{\mathrm{R}}) - \log q^*(a_r^{\mathrm{R}})\} = \\ & \quad \tfrac{1}{2} \sum_{r=1}^{q^{\mathrm{R}}} \log A_{\mathrm{R}r} - \log(\pi) + 2\log\Gamma\left\{\tfrac{1}{2}\,(\nu+2)\right\} + \tfrac{1}{2}\,(\nu+q^{\mathrm{R}}-1) \sum_{r=1}^{q^{\mathrm{R}}} E_q\left\{\log(a_r^{\mathrm{R}})\right\} \\ & \quad + \sum_{r=1}^{q^{\mathrm{R}}} \left\{(B_{q(a_r^{\mathrm{R}})}-1)/A_{\mathrm{R}r}\right\} \mu_{q(1/a_r^{\mathrm{R}})} - \tfrac{1}{2}\,(\nu+q^{\mathrm{R}}) \sum_{r=1}^{q^{\mathrm{R}}} \log\left(B_{q(a_r^{\mathrm{R}})}\right), \end{aligned}$$

where

$$\sum_{r=1}^{q^{\mathrm{R}}} \left\{(B_{q(a_r^{\mathrm{R}})}-1)/A_{\mathrm{R}r}\right\} \mu_{q(1/a_r^{\mathrm{R}})} = \sum_{r=1}^{q^{\mathrm{R}}} \nu(\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})})_{rr}\, \mu_{q(1/a_r^{\mathrm{R}})}.$$

The next term is

$$\begin{aligned} \log p(a_{u\ell}) - \log q^*(a_{u\ell}) \;\; = \;\; & \log\left\{\frac{(A_{u\ell})^{-1}}{\Gamma(0.5)}\,(a_{u\ell})^{-1.5} \exp\left(-\frac{1}{a_{u\ell}A_{u_\ell}^2}\right)\right\} \\ & - \log\left\{\frac{B_{q(a_{u\ell})}}{\Gamma(1)}(a_{u\ell})^{-2} \exp\left(-\frac{B_{q(a_{u\ell})}}{a_{u\ell}}\right)\right\}. \end{aligned}$$

Taking expectations gives

$$\begin{aligned} E_q\{\log p(a_{u\ell}) - \log q^*(a_{u\ell})\} \;\; = \;\; & -\tfrac{1}{2}\log(A_{u_\ell}^2) - \log\left(B_{q(a_{u\ell})}\right) - \tfrac{1}{2}\log(\pi) + \tfrac{1}{2}\,E_q\{\log(a_{u\ell})\} \\ & + \left(B_{q(a_{u\ell})} - 1/A_{u_\ell}^2\right) \mu_{q(1/a_{u\ell})}, \end{aligned}$$

where

$$\left(B_{q(a_{u\ell})} - 1/A_{u_\ell}^2\right) \mu_{q(1/a_{u\ell})} = \mu_{q(1/a_{u\ell})}\mu_{q(1/\sigma_{u\ell}^2)}.$$

The stated results then follows from cancellations lead by these quantities, as well as more obvious ones.

2.B Optimal q-densities derivation for Student-t semiparametric mixed models

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

$$q^*(\boldsymbol{\beta}, \boldsymbol{u}) \sim \text{Normal}(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})},\, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}),$$

where

$$\begin{aligned}
\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} &= \mu_{q(1/\sigma_\varepsilon^2)} \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \boldsymbol{C}^{\mathsf{T}} \boldsymbol{C}_{12} \boldsymbol{y} \quad \text{and} \\
\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} &= \left(\mu_{q(1/\sigma_\varepsilon^2)} \boldsymbol{C}^{\mathsf{T}} \boldsymbol{C}_{12} \boldsymbol{C} + \begin{bmatrix} \sigma_\beta^{-2} \boldsymbol{I}_p & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \underset{1\le\ell\le L}{\text{blockdiag}}(\mu_{q(1/\sigma_{u\ell}^2)} \boldsymbol{I}_{q_\ell^{\text{G}}}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_m \otimes \boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\text{R}})^{-1})} \end{bmatrix}^{-1} \right).
\end{aligned}$$

Derivation:
First we note that

$$\begin{aligned}
p(\boldsymbol{\beta}, \boldsymbol{u}|\text{rest}) &\propto p(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{y}, \sigma_\varepsilon^2, \boldsymbol{\Sigma}^{\text{R}}, \boldsymbol{\sigma}_u^2) \\
&\propto p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}^{\text{R}}, \boldsymbol{u}^{\text{G}}, \sigma_\varepsilon^2)\; p(\boldsymbol{u}^{\text{G}}|\boldsymbol{\sigma}_u^2)\; p(\boldsymbol{u}^{\text{R}}|\boldsymbol{\Sigma}^{\text{R}})\; p(\boldsymbol{\beta}).
\end{aligned}$$

Taking logarithms on both sides gives

$$\log p(\boldsymbol{\beta}, \boldsymbol{u}|\text{rest}) = -\tfrac{1}{2} \left\{ \boldsymbol{v}^{\mathsf{T}} (\sigma_\varepsilon^{-2} \boldsymbol{C}^{\mathsf{T}} \boldsymbol{C}_{12}\, \boldsymbol{C} + \boldsymbol{\Omega}^{-1}) \boldsymbol{v} - 2\sigma_\varepsilon^{-2} \boldsymbol{v}^{\mathsf{T}} \boldsymbol{C}^{\mathsf{T}} \boldsymbol{C}_{12}\, \boldsymbol{y} \right\} + \text{const},$$

which the leads to the full conditional

$$\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} \Bigg| \text{ rest} \sim N\left((\sigma_\varepsilon^{-2} \boldsymbol{C}^{\mathsf{T}} \boldsymbol{C}_{12}\, \boldsymbol{C} + \boldsymbol{\Omega}^{-1})^{-1} \sigma_\varepsilon^{-2} \boldsymbol{C}^{\mathsf{T}} \boldsymbol{C}_{12}\, \boldsymbol{y},\; (\sigma_\varepsilon^{-2} \boldsymbol{C}^{\mathsf{T}} \boldsymbol{C}_{12}\, \boldsymbol{C} + \boldsymbol{\Omega}^{-1})^{-1} \right).$$

Taking expectations with respect to all parameters except $(\boldsymbol{\beta}, \boldsymbol{u})$ gives

$$\begin{aligned}
\log q^*(\boldsymbol{\beta}, \boldsymbol{u}) &= E_q \left\{ \log p(\boldsymbol{\beta}, \boldsymbol{u}|\text{rest}) \right\} + \text{const} \\
&= -\tfrac{1}{2} \left\{ \boldsymbol{v}^{\mathsf{T}} \left(\mu_{q(1/\sigma_\varepsilon^2)} \boldsymbol{C}^{\mathsf{T}} \boldsymbol{C}_{12}\, \boldsymbol{C} + \boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\text{R}})^{-1})} \right) \boldsymbol{v} - 2\sigma_\varepsilon^{-2} \boldsymbol{v}^{\mathsf{T}} \boldsymbol{C}^{\mathsf{T}} \boldsymbol{C}_{12}\, \boldsymbol{y} \right\} + \text{const}.
\end{aligned}$$

The stated results follows from standard "completion of square" manipulations.

The remaining q^* densities involve straightforward adaptation of the derivations given in the Gaussian response case.

2.B.1 Derivation of the marginal log-likelihood lower bound

The expression for the marginal log-likelihood lower bound is

$$
\begin{aligned}
\underline{p}(\boldsymbol{y};q) \;=\;& \tfrac{1}{2}\,q^{\mathrm{R}}(\nu+q^{\mathrm{R}}-1)\log(2\nu) - \tfrac{1}{2}\log(2\pi)\textstyle\sum_{i=1}^{m} n_i - \left(\tfrac{1}{2}q^{\mathrm{R}}+L+1\right)\log(\pi)\\
&- \tfrac{1}{2}p\log(\sigma^2_\beta) - \tfrac{1}{2}\sigma^{-2}_\beta\left(\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\right) + \tfrac{1}{2}\left(\textstyle\sum_{\ell=1}^{L} q^{\mathrm{G}}_\ell + p + m\right)\\
&+ \tfrac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}| - \log(\mathcal{C}_{q^{\mathrm{R}},\,\nu+q^{\mathrm{R}}-1}) + \log(\mathcal{C}_{q^{\mathrm{R}},\,\nu+m+q^{\mathrm{R}}-1})\\
&- \tfrac{1}{2}\,(\nu+m+q^{\mathrm{R}}-1)\log|\boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\mathrm{R}})}|\\
&+ \textstyle\sum_{\ell=1}^{L}\log\Gamma\left\{\tfrac{1}{2}\,(q^{\mathrm{G}}_\ell+1)\right\} - \tfrac{1}{2}\,\sum_{\ell=1}^{L}(q^{\mathrm{G}}_\ell+1)\,\log(B_{q(\sigma^2_{u\ell})})\\
&+ \log\Gamma\left\{\tfrac{1}{2}(\textstyle\sum_{i=1}^{m} n_i+1)\right\} - \tfrac{1}{2}\,(\textstyle\sum_{i=1}^{m} n_i+1)\log\left(B_{q(\sigma^2_\varepsilon)}\right)\\
&- \textstyle\sum_{r=1}^{q^{\mathrm{R}}}\log(A_{\mathrm{R}r}) + q^{\mathrm{R}}\log\Gamma\left\{\tfrac{1}{2}(\nu+q^{\mathrm{R}})\right\}\\
&- \textstyle\sum_{\ell=1}^{L}\log(A_{u\ell}) - \sum_{\ell=1}^{L}\log\left(B_{q(a_{u\ell})}\right) + \sum_{\ell=1}^{L}\mu_{q(1/a_{u\ell})}\mu_{q(1/\sigma^2_{u\ell})}\\
&+ \textstyle\sum_{r=1}^{q^{\mathrm{R}}}\nu(\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})})_{rr}\,\mu_{q(1/a^{\mathrm{R}}_r)} - \tfrac{1}{2}\,(\nu+q^{\mathrm{R}})\,\sum_{r=1}^{q^{\mathrm{R}}}\log(B_{q(a^{\mathrm{R}}_r)})\\
&- \log(A_\varepsilon) - \log\left(B_{q(a_\varepsilon)}\right) + \mu_{q(1/a_\varepsilon)}\mu_{q(1/\sigma^2_\varepsilon)}\\
&- \tfrac{1}{2}\mu_{q(\nu+1)}\textstyle\sum_{i=1}^{N}\log B_{q(b_i)} + \log\Gamma\left\{\tfrac{1}{2}(\mu_{q(\nu)}+1)\right\} + \sum_{i=1}^{N} B_{q(b_i)}\mu_{q(1/b_i)}\\
&- \log(\nu_{\max}-\nu_{\min}) - \tfrac{1}{2}\mu_{q(\nu)}\textstyle\sum_{i=1}^{N}\mu_{q(\log(b_i))} + \log\mathcal{F}(0,\sum_{i=1}^{m} n_i, C_1, \nu_{\max}, \nu_{\min}),
\end{aligned}
$$

Derivation:
The logarithm of the lower bound on the marginal likelihood is given by

$$
\begin{aligned}
\log\underline{p}(\boldsymbol{y};q) \;=\;& E_q\left\{\log p(\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{R}},\boldsymbol{u}^{\mathrm{G}},\boldsymbol{\Sigma}^{\mathrm{R}},\boldsymbol{\sigma}^2_u,\sigma^2_\varepsilon,\boldsymbol{a}^{\mathrm{R}},\boldsymbol{a}_u,a_\varepsilon,\boldsymbol{b},v)\right.\\
&\left. - \log q^*(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{R}},\boldsymbol{u}^{\mathrm{G}},\boldsymbol{\Sigma}^{\mathrm{R}},\boldsymbol{\sigma}^2_u,\sigma^2_\varepsilon,\boldsymbol{a}^{\mathrm{R}},\boldsymbol{a}_u,a_\varepsilon,\boldsymbol{b},v)\;\right\}.
\end{aligned}
$$

According to the product restriction (2.17), our expression for the lower bound of the marginal log-likelihood becomes:

$$
\begin{aligned}
\log\underline{p}(\boldsymbol{y};q) \;=\;& E_q\left\{\log p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{R}},\boldsymbol{u}^{\mathrm{G}},\sigma^2_\varepsilon,\boldsymbol{b})\right\}\\
&+ E_q\left\{\log p(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{R}},\boldsymbol{u}^{\mathrm{G}}|\boldsymbol{\Sigma}^{\mathrm{R}},\boldsymbol{\sigma}^2_u) - \log q^*(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{R}},\boldsymbol{u}^{\mathrm{G}})\right\}\\
&+ E_q\left\{\log p(\boldsymbol{\Sigma}^{\mathrm{R}}|\boldsymbol{a}^{\mathrm{R}}) - \log q^*(\boldsymbol{\Sigma}^{\mathrm{R}})\right\} + E_q\left\{p(\boldsymbol{\sigma}^2_u|\boldsymbol{a}_u) - \log q^*(\boldsymbol{\sigma}^2_u)\right\}\\
&+ E_q\left\{\log p(\boldsymbol{a}^{\mathrm{R}}) - \log q^*(\boldsymbol{a}^{\mathrm{R}})\right\} + E_q\left\{\log p(\boldsymbol{a}_u) - \log q^*(\boldsymbol{a}_u)\right\}\\
&+ E_q\left\{p(\sigma^2_\varepsilon|a_\varepsilon) - \log q^*(\sigma^2_\varepsilon)\right\} + E_q\left\{\log p(a_\varepsilon) - \log q^*(a_\varepsilon)\right\}\\
&+ E_q\left\{\log p(\boldsymbol{b}|v) - \log q^*(\boldsymbol{b})\right\} + E_q\left\{\log p(v) - q^*(v)\right\}.
\end{aligned}
$$

We are only interested in the first and last terms (highlighted in darker colour) since they are the only differences between the Gaussian and Student-t response cases.

First note that

$$\begin{aligned}\log p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{R}},\boldsymbol{u}^{\mathrm{G}},\sigma_\varepsilon^2,\boldsymbol{b}) &= \tfrac{N}{2}\log(2\pi)-\tfrac{N}{2}\log(\sigma_\varepsilon^2)-\tfrac{1}{2}\textstyle\sum_{i=1}^{N}\log(1/b_i)\\ &\quad -\tfrac{1}{2}\left\{(\boldsymbol{y}-\boldsymbol{C}\boldsymbol{v})^{\mathsf{T}}(\sigma_\varepsilon^2\boldsymbol{C}_{12})^{-1}(\boldsymbol{y}-\boldsymbol{C}\boldsymbol{v})\right\}.\end{aligned}$$

Taking expectations gives

$$\begin{aligned}E_q\{\log p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{R}},\boldsymbol{u}^{\mathrm{G}},\sigma_\varepsilon^2,\boldsymbol{b})\} &= \tfrac{N}{2}\log(2\pi)-\tfrac{N}{2}E_q\{\log(\sigma_\varepsilon^2)\}-\tfrac{1}{2}\textstyle\sum_{i=1}^{N}\log(1/\mu_{q(b_i)})\\ &\quad -\tfrac{1}{2}\,\mu_{q(1/\sigma_\varepsilon^2)}\left\{(\boldsymbol{y}-\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})})^{\mathsf{T}}\boldsymbol{C}_{12}(\boldsymbol{y}-\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})})\right.\\ &\quad \left.+\,\mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C}_{12}\boldsymbol{C})\right\}.\end{aligned}$$

The next term is

$$\begin{aligned}\log p(\boldsymbol{b}|v)-\log q^*(\boldsymbol{b}) &= N\left\{\tfrac{v}{2}\log(v/2)-\log\Gamma(v/2)\right\}-\tfrac{v}{2}\sum_{i=1}^{N}\{\log(b_i)+(1/b_i)\}\\ &\quad -\tfrac{1}{2}(\mu_{q(v)}+1)\sum_{i=1}^{N}\log B_{q(b_i)}+N\log\Gamma\{\tfrac{1}{2}\mu_{q(v)}+1\}\\ &\quad +\tfrac{1}{2}\,(\mu_{q(v)}+3)\sum_{i=1}^{N}\log(b_i)-\sum_{i=1}^{N}B_{q(b_i)}/b_i.\end{aligned}$$

Taking expectations gives

$$\begin{aligned}E_q\left\{\log p(\boldsymbol{b}|v)-\log q^*(\boldsymbol{b})\right\} &= E_q\left\{N\left(\tfrac{v}{2}\log(v/2)-\log\Gamma(v/2)\right)\right\}\\ &\quad -\tfrac{1}{2}(\mu_{q(v)}+1)\sum_{i=1}^{N}\log B_{q(b_i)}+N\log\Gamma\{\tfrac{1}{2}(\mu_{q(v)}+1)\}\\ &\quad -\tfrac{1}{2}\sum_{i=1}^{N}\mu_{q(\log(b_i))}+\sum_{i=1}^{N}\left(B_{q(b_i)}-\tfrac{1}{2}\mu_{q(v)}\right)\mu_{q(1/b_i)}.\end{aligned}$$

The next term is

$$\begin{aligned}\log p(v)-\log q^*(v) &= -\log(v_{\max}-v_{\min})-N\left(\tfrac{v}{2}\log(v/2)-\log\Gamma(v/2)\right)\\ &\quad +\tfrac{v}{2}\sum_{i=1}^{N}\{\log(b_i)+(1/b_i)\}+\log\mathcal{F}(0,N,\boldsymbol{C}_1,v_{\max},v_{\min}).\end{aligned}$$

Taking expectations gives

$$\begin{aligned} E_q\left\{\log p(v) - \log q^*(v)\right\} &= -\log\left(v_{\max} - v_{\min}\right) - E_q\left(N\left(\tfrac{v}{2}\log(v/2) - \log\Gamma(v/2)\right)\right) \\ &\quad + \tfrac{1}{2}\mu_{q(v)}\sum_{i=1}^{N}\{\mu_{q(\log(b_i))} + \mu_{q(1/b_i)}\} + \log\mathcal{F}(0, N, \boldsymbol{C}_1, v_{\max}, v_{\min}). \end{aligned}$$

Expressions for $q^*(\sigma_\varepsilon^2), B_{q(\sigma_\varepsilon^2)}$ and $\mu_{q(1/\sigma_\varepsilon^2)}$

$$q^*(\sigma_\varepsilon^2) \sim \text{Inverse-Gamma}(\tfrac{1}{2}\left(\textstyle\sum_{i=1}^m n_i + 1\right),\, B_{q(\sigma_\varepsilon^2)}),$$

where

$$\begin{aligned} \mu_{q(\sigma_\varepsilon^2)} &= \tfrac{1}{2}\left(\textstyle\sum_{i=1}^m n_i + 1\right)/B_{q(\sigma_\varepsilon^2)} \quad \text{and} \\ B_{q(\sigma_\varepsilon^2)} &= \tfrac{1}{2}\left\{(\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})})^{\mathsf{T}}\boldsymbol{C}_{12}(\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}) + \text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C}_{12}\boldsymbol{C})\right\} + \mu_{q(1/a_\varepsilon)}. \end{aligned}$$

Derivation:
First we note that

$$p(\sigma_\varepsilon^2|\text{rest}) \propto p(\sigma_\varepsilon^2|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{u}, a_\varepsilon) \propto p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2)\; p(\sigma_\varepsilon^2|a_\varepsilon).$$

Taking logarithms on both sides gives

$$\begin{aligned} \log p(\sigma_\varepsilon^2|\text{rest}) &= \left\{-\tfrac{1}{2}(\textstyle\sum_{i=1}^m n_i + 1) - 1\right\} \\ &\quad - \log(\sigma_\varepsilon^2)\left\{\tfrac{1}{2}(\boldsymbol{y} - \boldsymbol{C}\boldsymbol{v})^{\mathsf{T}}\boldsymbol{C}_{12}(\boldsymbol{y} - \boldsymbol{C}\boldsymbol{v}) + 1/a_\varepsilon\right\}\Big/\sigma_\varepsilon^2 + \text{const}, \end{aligned}$$

which then leads to the full conditional

$$\sigma_\varepsilon^2|\text{rest} \;\sim\; \text{Inverse-Gamma}\left(\tfrac{1}{2}(\textstyle\sum_{i=1}^m n_i + 1),\, \tfrac{1}{2}(\boldsymbol{y} - \boldsymbol{C}\boldsymbol{v})^{\mathsf{T}}\boldsymbol{C}_{12}(\boldsymbol{y} - \boldsymbol{C}\boldsymbol{v}) + a_\varepsilon^{-1}\right).$$

Taking expectations with respect to all parameters except σ_ε^2 gives

$$\begin{aligned} \log q^*(\sigma_\varepsilon^2) &= E_q\left\{\log p(\sigma_\varepsilon^2|\text{rest})\right\} + \text{const} \\ &= \left\{-\tfrac{1}{2}(\textstyle\sum_{i=1}^m n_i + 1) - 1\right\}\log(\sigma_\varepsilon^2) - \left\{\tfrac{1}{2}(\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})})^{\mathsf{T}}\boldsymbol{C}_{12}(\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})})\right. \\ &\quad \left. + \tfrac{1}{2}\text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C}_{12}\boldsymbol{C}) + \mu_{q(1/a_\varepsilon)}\right\}\Big/\sigma_\varepsilon^2 + \text{const}. \end{aligned}$$

The expression for $\mu_{q(1/\sigma_\varepsilon^2)}$ follows from elementary manipulations involving inverse-Gamma density functions.

Expressions for $q^*(b_i), B_{q(b_i)}$ and $\mu_{q(b_i)}$

$$q^*(\boldsymbol{b}) = \prod_{i=1}^{N} q^*(b_i); \quad q^*(b_i) \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2}(\mu_{q(\nu)}+1),\, B_{q(b_i)})$$

where

$$B_{q(b_i)} = \tfrac{1}{2}\left\{\mu_{q(\nu)} + (\boldsymbol{y}-\boldsymbol{C}\boldsymbol{v})_i^2 \,/\sigma_\varepsilon^2\right\} \quad \text{and} \quad \mu_{q(b_i)} = \tfrac{1}{2}(\mu_{q(\nu)}+1)/B_{q(b_i)}.$$

Derivation:
First we note that

$$p(\boldsymbol{b}|\text{rest}) \propto p(\boldsymbol{b}|\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2,\nu) \propto p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{b},a_\varepsilon)\; p(\boldsymbol{b}|\nu).$$

Taking logarithms on both sides gives

$$\log p(\boldsymbol{b}|\text{rest}) = \left\{-\tfrac{1}{2}(\nu+1)-1\right\}\sum_{i=1}^{N}\left[\log(b_i) - \tfrac{1}{2b_i}\left\{\nu + (\boldsymbol{y}-\boldsymbol{C}\boldsymbol{v})_i^2\,/\sigma_\varepsilon^2\right\}\right] + \text{const},$$

which then leads to the full conditional

$$b_i|\text{rest} \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(\tfrac{1}{2}(\nu+1),\, \tfrac{1}{2}\left\{\nu + (\boldsymbol{y}-\boldsymbol{C}\boldsymbol{v})_i^2\,/\sigma_\varepsilon^2\right\}\right).$$

Taking expectations with respect to all parameters except $\boldsymbol{b}$ gives

$$\begin{aligned}\log q^*(\boldsymbol{b}) \;=\; & \left\{-\tfrac{1}{2}(\mu_{q(\nu)}+1)-1\right\}\sum_{i=1}^{N}\Big[\log(b_i) - \tfrac{1}{2b_i}\Big\{\mu_{q(\nu)} + \Big((\boldsymbol{y}-\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})})_i^2 \\ & + (\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^{\mathsf{T}})_{ii}\Big/\mu_{q(\sigma_\varepsilon^2)}\Big)\Big\}\Big] + \text{const}.\end{aligned}$$

The expressions for $B_{q(b_i)}$ and $\mu_{q(b_i)}$ follow immediately.

Expressions for $p(\nu|\text{rest})$ and $q^*(\nu)$

The optimal density $q^*(\nu)$ is not of a standard from:

$$\begin{aligned} p(\nu|\text{rest}) \;&\propto\; p(\nu|\boldsymbol{b}) = p(\boldsymbol{b}|\nu)\; p(\nu) \\ &=\; \frac{1}{(\nu_{\max}-\nu_{\min})}\prod_{i=1}^{N}\left\{\frac{(0.5\,\nu)^{0.5\nu}}{\Gamma(0.5\,\nu)}(b_i)^{-\left(\frac{1}{2}\nu+1\right)}\exp\left(-\frac{\nu}{2b_i}\right)\right\}\end{aligned}$$

$$\begin{aligned}\log p(\nu|\text{rest}) &= N\left\{\tfrac{\nu}{2}\log(\nu/2) - \log\Gamma(\nu/2)\right\} - \tfrac{\nu}{2}\sum_{i=1}^{N}\left(\log(b_i) + b_i^{-1}\right) + \text{const} \\ \log q^*(\nu) &= N\left\{\tfrac{\nu}{2}\log(\nu/2) - \log\Gamma(\nu/2)\right\} - \tfrac{\nu}{2}\sum_{i=1}^{N}\left(\log\left(\mu_{q(b_i)}\right) + \mu_{q(1/b_i)}\right) + \text{const.}\end{aligned}$$

The remaining q^* densities involve straightforward adaptation of the derivations given in the Gaussian response case.

2.C Optimal q-densities derivation for Bernoulli semiparametric mixed models

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

$$q^*(\boldsymbol{\beta}, \boldsymbol{u}) \sim \text{Normal}(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u};\,\boldsymbol{\xi})},\ \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\,\boldsymbol{\xi})}),$$

where

$$\begin{aligned}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} &= \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\,\boldsymbol{\xi})}\boldsymbol{C}^{\mathsf{T}}\left(\boldsymbol{y} - \tfrac{1}{2}\mathbf{1}\right) \quad \text{and} \\ \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\,\boldsymbol{\xi})} &= \left(2\boldsymbol{C}^{\mathsf{T}}\text{diag}\{\lambda(\boldsymbol{\xi})\}\boldsymbol{C} + \begin{bmatrix}\sigma_\beta^{-2}\boldsymbol{I}_p & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \underset{1\le\ell\le L}{\text{blockdiag}}(\mu_{q(1/\sigma^2_{u\ell})}\boldsymbol{I}_{q_\ell^{\text{G}}}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{I}_m \otimes \boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\text{R}})^{-1})}\end{bmatrix}\right)^{-1}.\end{aligned}$$

Derivation:
First note that

$$\begin{aligned}p(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{u};\, \boldsymbol{\xi}) &= \exp\Big\{\boldsymbol{y}^{\mathsf{T}}\boldsymbol{C}\boldsymbol{v} - \left(\boldsymbol{v}^{\mathsf{T}}\boldsymbol{C}^{\mathsf{T}}\,\text{diag}\{\lambda(\boldsymbol{\xi})\}\,\boldsymbol{C}\boldsymbol{v} - \tfrac{1}{2}\mathbf{1}^{\mathsf{T}}\boldsymbol{C}\boldsymbol{v} + \tfrac{1}{2}\mathbf{1}^{\mathsf{T}}\,\psi(\boldsymbol{\xi})\right) \\ &\quad - \tfrac{1}{2}\boldsymbol{v}^{\mathsf{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{v}\Big\} + \text{const.}\end{aligned}$$

Taking logarithms on both sides gives

$$\log p(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{u};\, \boldsymbol{\xi}) = -\tfrac{1}{2}\left\{\boldsymbol{v}^{\mathsf{T}}\left(2\boldsymbol{C}^{\mathsf{T}}\text{diag}\{\lambda(\boldsymbol{\xi})\}\boldsymbol{C} + \boldsymbol{\Omega}^{-1}\right)\boldsymbol{v} - 2\boldsymbol{C}^{\mathsf{T}}\left(\boldsymbol{y} - \tfrac{1}{2}\mathbf{1}\right)\boldsymbol{v}^{\mathsf{T}}\right\} + \text{const},$$

which then leads to the full conditional

$$\left.\begin{bmatrix}\boldsymbol{\beta} \\ \boldsymbol{u}\end{bmatrix}\right| \text{rest} \sim N\left((2\boldsymbol{C}^{\mathsf{T}}\text{diag}\{\lambda(\boldsymbol{\xi})\}\boldsymbol{C} + \boldsymbol{\Omega}^{-1})^{-1}\boldsymbol{C}^{\mathsf{T}}\left(\boldsymbol{y} - \tfrac{1}{2}\mathbf{1}\right),\right.$$

$$(2\boldsymbol{C}^{\mathsf{T}}\mathrm{diag}\{\lambda(\boldsymbol{\xi})\}\boldsymbol{C}+\boldsymbol{\Omega}^{-1})^{-1}\Big).$$

Taking expectations with respect to all parameters except $(\boldsymbol{\beta},\boldsymbol{u})$ gives

$$\begin{aligned}\log q^*(\boldsymbol{\beta},\boldsymbol{u};\,\boldsymbol{\xi}) &= E_q\left\{\log p\left(\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{u};\,\boldsymbol{\xi}\right)|\mathrm{rest}\right\}+\mathrm{const}\\ &= -\tfrac{1}{2}\Big\{\boldsymbol{v}^{\mathsf{T}}\Big(2\boldsymbol{C}^{\mathsf{T}}\mathrm{diag}\{\lambda(\boldsymbol{\xi})\}\boldsymbol{C}+\boldsymbol{M}_{q(\boldsymbol{\Omega}^{-1})}\Big)\boldsymbol{v}-2\boldsymbol{C}^{\mathsf{T}}\left(\boldsymbol{y}-\tfrac{1}{2}\mathbf{1}\right)\boldsymbol{v}^{\mathsf{T}}\Big\}+\mathrm{const},\end{aligned}$$

where the optimal update for the additional variational vector $\boldsymbol{\xi}$ is

$$\boldsymbol{\xi}\leftarrow\sqrt{\mathrm{diagonal}\left\{\boldsymbol{C}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\,\boldsymbol{\xi})}+\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u};\,\boldsymbol{\xi})}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u};\,\boldsymbol{\xi})}^{\mathsf{T}}\right)\boldsymbol{C}^{\mathsf{T}}\right\}}.$$

This follows the Expectation-Maximisation arguments presented in Jaakkola and Jordan (2000). The remaining q^* densities involve straightforward adaptation of the derivations given in the Gaussian response case.

2.C.1 Derivation of the marginal log-likelihood lower bound

The expression for the marginal log-likelihood lower bound is

$$\begin{aligned}\log\underline{p}(\boldsymbol{y};q) &= \tfrac{1}{2}\,q^{\mathrm{R}}(\nu+q^{\mathrm{R}}-1)\log(2\nu)-\left(\tfrac{1}{2}q^{\mathrm{R}}+L\right)\log(\pi)-\lambda(\boldsymbol{\xi})^{\mathsf{T}}(\boldsymbol{\xi}^2)\\ &\quad+\mathbf{1}^{\mathsf{T}}\psi(\boldsymbol{\xi})+\left(\boldsymbol{y}-\tfrac{1}{2}\mathbf{1}\right)^{\mathsf{T}}\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}+\tfrac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}|\\ &\quad-\tfrac{1}{2}p\log(\sigma_\beta^2)-\tfrac{1}{2}\sigma_\beta^{-2}\left(\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2+\mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\right)+\tfrac{1}{2}\left(\textstyle\sum_{\ell=1}^{L}q_\ell^{\mathrm{G}}+p+m\right)\\ &\quad-\log(\mathcal{C}_{q^{\mathrm{R}},\nu+q^{\mathrm{R}}-1})+\log(\mathcal{C}_{q^{\mathrm{R}},\nu+m+q^{\mathrm{R}}-1})-\tfrac{1}{2}\,(\nu+m+q^{\mathrm{R}}-1)\log|B_{q(\boldsymbol{\Sigma}^{\mathrm{R}})}|\\ &\quad+\textstyle\sum_{\ell=1}^{L}\log\Gamma\left\{\frac{1}{2}\,(q_\ell^{\mathrm{G}}+1)\right\}-\frac{1}{2}\,\sum_{\ell=1}^{L}(q_\ell^{\mathrm{G}}+1)\,\log(B_{q(\sigma_{u\ell}^2)})\\ &\quad-\textstyle\sum_{r=1}^{q^{\mathrm{R}}}\log(A_{\mathrm{R}r})+q^{\mathrm{R}}\log\Gamma\left\{\frac{1}{2}(\nu+q^{\mathrm{R}})\right\}\\ &\quad+\textstyle\sum_{r=1}^{q^{\mathrm{R}}}\nu(\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})})_{rr}\,\mu_{q(1/a_r^{\mathrm{R}})}\\ &\quad-\tfrac{1}{2}\,(\nu+q^{\mathrm{R}})\,\textstyle\sum_{r=1}^{q^{\mathrm{R}}}\log(B_{q(a_r^{\mathrm{R}})})-\sum_{\ell=1}^{L}\log(A_{u\ell})\\ &\quad-\textstyle\sum_{\ell=1}^{L}\log\left(B_{q(a_{u\ell})}\right)+\sum_{\ell=1}^{L}\mu_{q(1/a_{u\ell})}\mu_{q(1/\sigma_{u\ell}^2)},\end{aligned}$$

Derivation:
The logarithm of the lower bound on the marginal likelihood is given by

$$\begin{aligned}\log\underline{p}(\boldsymbol{y};q) &= E_q\left\{\log p(\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{R}},\boldsymbol{u}^{\mathrm{G}},\boldsymbol{\Sigma}^{\mathrm{R}},\boldsymbol{\sigma}_u^2,\sigma_\varepsilon^2,\boldsymbol{a}^{\mathrm{R}},\boldsymbol{a}_u)\right.\\ &\qquad\left.-\log q^*(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{R}},\boldsymbol{u}^{\mathrm{G}},\boldsymbol{\Sigma}^{\mathrm{R}},\boldsymbol{\sigma}_u^2,\sigma_\varepsilon^2,\boldsymbol{a}^{\mathrm{R}},\boldsymbol{a}_u)\right\}.\end{aligned}$$

According to the product restriction (2.21), our expression for the lower bound of the

marginal log-likelihood becomes:

$$\begin{aligned}\log \underline{p}(\boldsymbol{y};q) &= E_q\left\{\log p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{R}},\boldsymbol{u}^{\mathrm{G}})\right\}\\ &\quad + E_q\left\{\log p(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{R}},\boldsymbol{u}^{\mathrm{G}}|\boldsymbol{\Sigma}^{\mathrm{R}},\boldsymbol{\sigma}_u^2) - \log q^*(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{R}},\boldsymbol{u}^{\mathrm{G}}\right\}\\ &\quad + E_q\left\{\log p(\boldsymbol{\Sigma}^{\mathrm{R}}|\boldsymbol{a}^{\mathrm{R}}) - \log q^*(\boldsymbol{\Sigma}^{\mathrm{R}})\right\} E_q\left\{p(\boldsymbol{\sigma}_u^2|\boldsymbol{a}_u) - \log q^*(\boldsymbol{\sigma}_u^2)\right\}\\ &\quad + E_q\left\{\log p(\boldsymbol{a}^{\mathrm{R}}) - \log q^*(\boldsymbol{a}^{\mathrm{R}})\right\} + E_q\left\{\log p(\boldsymbol{a}_u) - \log q^*(\boldsymbol{a}_u)\right\}.\end{aligned}$$

We are only interested in the first term (highlighted in darker colour) since it is the only difference between the Gaussian and Bernoulli response cases.

Replacing q^* by $q^*(\cdot;\boldsymbol{\xi})$ we then have the following

$$\begin{aligned}\log p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{R}},\boldsymbol{u}^{\mathrm{G}}) &= \boldsymbol{y}^{\mathsf{T}}\boldsymbol{C}\boldsymbol{v} - \mathbf{1}^{\mathsf{T}}\log(\mathbf{1}+\exp(\boldsymbol{C}\boldsymbol{v}))\\ &\geq \boldsymbol{y}^{\mathsf{T}}\boldsymbol{C}\boldsymbol{v} + \boldsymbol{v}^{\mathsf{T}}\boldsymbol{C}^{\mathsf{T}}\mathrm{diag}\{\lambda(\boldsymbol{\xi})\}\boldsymbol{C}\boldsymbol{v} - \tfrac{1}{2}\,\mathbf{1}^{\mathsf{T}}\boldsymbol{C}\boldsymbol{v} + \mathbf{1}^{\mathsf{T}}\psi(\boldsymbol{\xi})\\ \log \underline{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{R}},\boldsymbol{u}^{\mathrm{G}};\boldsymbol{\xi}) &= (\boldsymbol{y}-\tfrac{1}{2}\mathbf{1})^{\mathsf{T}}\boldsymbol{C}\boldsymbol{v} + \boldsymbol{v}^{\mathsf{T}}\boldsymbol{C}^{\mathsf{T}}\mathrm{diag}\{\lambda(\boldsymbol{\xi})\}\boldsymbol{C}\boldsymbol{v} + \mathbf{1}^{\mathsf{T}}\psi(\boldsymbol{\xi}).\end{aligned}$$

Taking expectations,

$$\begin{aligned}E_{\underline{q}}\left\{\log \underline{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{R}},\boldsymbol{u}^{\mathrm{G}};\boldsymbol{\xi})\right\} &= (\boldsymbol{y}-\tfrac{1}{2}\mathbf{1})^{\mathsf{T}}\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})} + \mathrm{tr}\Big(\boldsymbol{C}^{\mathsf{T}}\mathrm{diag}\{\lambda(\boldsymbol{\xi})\}\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}\\ &\quad + \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}^{\mathsf{T}}\Big) + \mathbf{1}^{\mathsf{T}}\psi(\boldsymbol{\xi}).\end{aligned}$$

2.D Optimal q-densities derivation for Poisson semiparametric mixed models

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

$$\begin{aligned}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} &\leftarrow \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\left(\mathsf{D}_{\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}}\boldsymbol{S}\right)^{\mathsf{T}},\\ \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} &\leftarrow \left\{-2\mathrm{vec}^{-1}\left(\left(\mathsf{D}_{\mathrm{vec}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})}\boldsymbol{S}\right)^{\mathsf{T}}\right)\right\}^{-1},\end{aligned}$$

where

$$S \equiv E_q\left\{\log p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{R}},\boldsymbol{u}^{\mathrm{G}}) + \log p(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{R}},\boldsymbol{u}^{\mathrm{G}}|\boldsymbol{\Sigma}^{\mathrm{R}},\boldsymbol{\sigma}_u^2)\right\}$$

and the notation D denotes derivative vector.

Derivation:
First note that

$$\begin{aligned}
\log p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{R}},\boldsymbol{u}^{\mathrm{G}}) &= \boldsymbol{y}^{\mathsf{T}}(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u}) - \boldsymbol{1}^{\mathsf{T}}\exp(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u}) - \boldsymbol{1}^{\mathsf{T}}\log(\boldsymbol{y}!).\\
\log p(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{R}},\boldsymbol{u}^{\mathrm{G}}|\boldsymbol{\Sigma}^{\mathrm{R}},\boldsymbol{\sigma}_u^2) &= \log\Big[(2\pi)^{-p/2}(\sigma_\beta^2\,\boldsymbol{I}_p)^{-p/2}\exp(\tfrac{1}{2}\sigma_\beta^{-2}\,\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\beta})\\
&\quad\times(2\pi)^{-m/2}\,(\boldsymbol{\Sigma}^{\mathrm{R}})^{-m/2}\exp\Big\{-\tfrac{1}{2}(\boldsymbol{u}^{\mathrm{R}})^{\mathsf{T}}(\boldsymbol{\Sigma}^{\mathrm{R}})^{-1}\boldsymbol{u}^{\mathrm{R}}\Big\}\\
&\quad\times\prod_{\ell=1}^{L}(2\pi)^{-q_\ell^{\mathrm{G}}/2}\,(\sigma_{u\ell}^2\,\boldsymbol{I}_{q_\ell^{\mathrm{G}}})^{-q_\ell^{\mathrm{G}}/2}\exp\Big\{\tfrac{1}{2}\sigma_{u\ell}^{-2}\,(\boldsymbol{u}_\ell^{\mathrm{G}})^{\mathsf{T}}\boldsymbol{u}_\ell^{\mathrm{G}}\Big\}\Big].
\end{aligned}$$

An explicit expression for S is

$$\begin{aligned}
S &= \boldsymbol{y}^{\mathsf{T}}\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} - \boldsymbol{1}^{\mathsf{T}}\exp\Big\{\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \tfrac{1}{2}\,\mathrm{diagonal}\Big(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^{\mathsf{T}}\Big)\Big\}\\
&\quad -\tfrac{1}{2}\,\mathrm{tr}\Big(\boldsymbol{M}_{q(\boldsymbol{\Omega}^{-1})}\,\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}^{\mathsf{T}} + \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\Big) - \tfrac{1}{2}\,\Big(\textstyle\sum_{i=1}^{m} n_i + \sum_{\ell=1}^{L} q_\ell^{\mathrm{G}} + 1\Big)\,\log(2\pi)\\
&\quad -\tfrac{1}{2}\,\textstyle\sum_{i=1}^{m} n_i\,E_q\{\log(|\boldsymbol{\Sigma}^{\mathrm{R}}|)\} - \tfrac{1}{2}\,\sum_{\ell=1}^{L} q_\ell^{\mathrm{G}}\,E_q\{\log(\sigma_{u\ell}^2)\}\\
&\quad -E_q\{\log(\sigma_\beta^2)\} - \boldsymbol{1}^{\mathsf{T}}\log(\boldsymbol{y}!).
\end{aligned}$$

Differentiate S with respect to $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ gives

$$\begin{aligned}
d_{\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}}S &= \boldsymbol{y}^{\mathsf{T}}\boldsymbol{C}d_{\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}} - \boldsymbol{1}^{\mathsf{T}}\mathrm{diag}\Big\{\exp(\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}) + \tfrac{1}{2}\,\mathrm{diagonal}\,(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^{\mathsf{T}})\Big\}\,\boldsymbol{C}d_{\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}}\\
&\quad -\tfrac{1}{2}\,\Big(2\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}^{\mathsf{T}}\boldsymbol{M}_{q(\boldsymbol{\Omega}^{-1})}\,d_{\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}}\Big).
\end{aligned}$$

Thus, by Theorem 6, Chapter 5 of Magnus and Neudecker (1995)

$$\begin{aligned}
\Big(\mathsf{D}_{\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}}S\Big)^{\mathsf{T}} &= \boldsymbol{C}^{\mathsf{T}}\Big[\boldsymbol{y} - \exp\Big\{\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \tfrac{1}{2}\,\mathrm{diagonal}\Big(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^{\mathsf{T}}\Big)\Big\}\Big]\\
&\quad -\boldsymbol{M}_{q(\boldsymbol{\Omega}^{-1})}\,\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}.
\end{aligned}$$

Differentiate S with respect to $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ gives

$$\begin{aligned}
d_{\mathrm{vec}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})}S &= -\boldsymbol{1}^{\mathsf{T}}\mathrm{diag}\Big[\exp\Big\{\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \tfrac{1}{2}\,Q(\boldsymbol{C})\mathrm{vec}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})\Big\}\Big]\,\tfrac{1}{2}\,Q(\boldsymbol{C})d_{\mathrm{vec}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})}\\
&\quad -\tfrac{1}{2}\,\mathrm{vec}\Big(\boldsymbol{C}^{\mathsf{T}}\mathrm{diag}\Big[\exp\Big\{\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \tfrac{1}{2}\,\mathrm{diagonal}\Big(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^{\mathsf{T}}\Big)\Big\}\,\boldsymbol{C}\\
&\quad +\ \boldsymbol{M}_{q(\boldsymbol{\Omega}^{-1})}\Big]\Big)^{\mathsf{T}}\,d_{\mathrm{vec}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})},
\end{aligned}$$

where $Q(\boldsymbol{C}) = (\boldsymbol{C}\otimes\boldsymbol{1}^{\mathsf{T}})\odot(\boldsymbol{1}^{\mathsf{T}}\otimes\boldsymbol{C})$ is defined in Theorem 2 of Section 4 of Wand (2014).

Therefore,

$$\begin{aligned}\mathrm{vec}^{-1}\left\{(\mathsf{D}_{\mathrm{vec}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})}S)^{\mathsf{T}}\right\} &= \tfrac{1}{2}\left(\boldsymbol{C}^{\mathsf{T}}\mathrm{diag}\left[\exp\left\{\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}+\tfrac{1}{2}\,\mathrm{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^{\mathsf{T}})\right\}\right]\boldsymbol{C}\right.\\ &\quad\left.+\,\boldsymbol{M}_{q(\boldsymbol{\Omega}^{-1})}\right).\end{aligned}$$

The remaining q^* densities involve straightforward adaptation of the derivations given in the Gaussian response case.

2.D.1 Derivation of the marginal log-likelihood lower bound

The expression for the marginal log-likelihood lower bound is

$$\begin{aligned}\log\underline{p}(\boldsymbol{y};q) &= \tfrac{1}{2}\,q^{\mathrm{R}}(\nu+q^{\mathrm{R}}-1)\log(2\nu)-\left(\tfrac{1}{2}q^{\mathrm{R}}+L\right)\log(\pi)\\ &\quad+\boldsymbol{y}^{\mathsf{T}}\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}-\mathbf{1}^{\mathsf{T}}\exp\{\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}+\tfrac{1}{2}\mathrm{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})\boldsymbol{C}^{\mathsf{T}}\}\\ &\quad-\mathbf{1}^{\mathsf{T}}\log(\boldsymbol{y}!)+\left(\boldsymbol{y}-\tfrac{1}{2}\mathbf{1}\right)^{\mathsf{T}}\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}+\tfrac{1}{2}\log\|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}\|\\ &\quad-\tfrac{1}{2}p\log(\sigma_\beta^2)-\tfrac{1}{2}\sigma_\beta^{-2}\left(\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2+\mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\right)\\ &\quad-\log(\mathcal{C}_{q^{\mathrm{R}},\nu+q^{\mathrm{R}}-1})+\log(\mathcal{C}_{q^{\mathrm{R}},\nu+m+q^{\mathrm{R}}-1})\\ &\quad-\tfrac{1}{2}\,(\nu+m+q^{\mathrm{R}}-1)\log|B_{q(\boldsymbol{\Sigma}^{\mathrm{R}})}|\\ &\quad+\textstyle\sum_{\ell=1}^{L}\log\Gamma\left\{\frac{1}{2}\,(q_\ell^{\mathrm{G}}+1)\right\}-\frac{1}{2}\,\sum_{\ell=1}^{L}(q_\ell^{\mathrm{G}}+1)\,\log(B_{q(\sigma_{u\ell}^2)})\\ &\quad+q^{\mathrm{R}}\log\Gamma\left\{\tfrac{1}{2}(\nu+q^{\mathrm{R}})\right\}+\textstyle\sum_{r=1}^{q^{\mathrm{R}}}\nu(\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})})_{rr}\,\mu_{q(1/a_r^{\mathrm{R}})}\\ &\quad-\tfrac{1}{2}\,(\nu+q^{\mathrm{R}})\,\textstyle\sum_{r=1}^{q^{\mathrm{R}}}\log(B_{q(a_r^{\mathrm{R}})})-\,\sum_{\ell=1}^{L}\log(A_{u\ell})\\ &\quad-\textstyle\sum_{\ell=1}^{L}\log\left(B_{q(a_{u\ell})}\right)+\sum_{\ell=1}^{L}\mu_{q(1/a_{u\ell})}\mu_{q(1/\sigma_{u\ell}^2)}\\ &\quad+\tfrac{1}{2}\left(\textstyle\sum_{\ell=1}^{L}q_\ell^{\mathrm{G}}+p+m\right)-\sum_{r=1}^{q^{\mathrm{R}}}\log(A_{\mathrm{R}r}),\end{aligned}$$

Derivation:
The derivation is very similar to that of the Bernoulli case.

Chapter 3

Streamlining Mean Field Variational Bayes Algorithms And Three-Level Model Extensions

Technology is so much fun but we can drown in our technology. The fog of information can drive out knowledge.

Daniel J. Boorstin

3.1 Introduction

Previously we introduced the concept of MFVB algorithms that aim to approximate the intractable posterior by a factorised distribution which can be represented by a directed acyclic graph, and optimisation of the factorised approximating posterior can be decomposed into local computations that involve only neighbouring nodes. Recent examples of these include Algorithm 3 of Ormerod and Wand (2010) and Algorithms 3 and 5 of Luts *et al.* (2014) which use naïve matrix inversion for the effects covariance matrix updates to support MFVB fitting of longitudinal and multilevel data. Following the naïve implementation by these authors, we presented the naïve version of MFVB algorithms which are not fully computationally optimised because of their cubic dependence on the number of groups.

The main content of this chapter is published as: Lee, C. Y. Y. and Wand, M. P. (2015). Streamlined mean field variational Bayes for longitudinal and multilevel data analysis. *Biometrical Journal article*. DOI:10.1002/bimj.201500007. This research has been accepted for presentation at four conferences.

In this chapter, we develop MFVB algorithms for fitting and inference in arbitrarily large longitudinal and multilevel models that are streamlined in terms of number of operations and storage. The number of operations is linear in the number of groups at each level, which constitutes a major improvement over the naïve implementations. The essence of our approach is to take advantage of the sparseness of the matrix requiring inversion and streamline its inversion. This involves omitting correlations of the estimated effects between groups, which are rarely of interest. It is likely that our resultant streamlined algorithms provide the fastest ever means of approximate Bayesian analyses of very large longitudinal and multilevel datasets.

Recently, Tan and Nott (2013) introduce a different strategy for improving the efficiency of variational methods for longitudinal and multilevel models. The product restriction of Tan and Nott (2013) is such that the random effects for each group appear in a separate factor and their *partially noncentered parameterisation* strategy is aimed at improved accuracy in the face of such a restriction. We do not impose their restriction in our variational approximation for the random group effects. Fitting and inference proceeds efficiently via a streamlined approach that takes advantage of the inherent blocked structure of the effects covariance matrix. Our algorithms correspond to the algorithms of Tan and Nott (2013) in which their tuning parameter, that captures correlations between fixed and random effects, does not have to be estimated and, instead, is specified optimally. Our streamlined algorithms are simpler due to minimal product restrictions in the MFVB approximation and apply to a richer class of models.

The next section presents the computational obstacles involved in the naïve MFVB algorithms. Section 3.4 gives details of our novel streamlined approach to sparse covariance estimation vis matrix permutation and decomposition, and is the centrepiece of this chapter. In Section 3.5 we provide numerical evidence of the efficiency gain for using the streamlined approach over the naïve approach. Illustration of real data examples is presented in Section 3.6. Section 3.7 extends the streamlined variational algorithms to cater for higher-level Bayesian semiparametric mixed models and concluding remarks are given in Section 3.8.

3.2 Computational challenges in naïve MFVB algorithms

Recall from Chapter 2 that the general compact form of a two-level Bayesian generalised semiparametric mixed model is

$$\begin{aligned} \boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u} &\sim \text{Dist}\left(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u},\boldsymbol{R}\right), \quad \boldsymbol{\beta}\sim N(0,\sigma_\beta^2\boldsymbol{I}_p), \\ \boldsymbol{u}|\sigma_{u1}^2,\ldots,\sigma_{uL}^2,\boldsymbol{\Sigma}^{\text{R}} &\sim N\left(\boldsymbol{0},\begin{bmatrix}\underset{1\le\ell\le L}{\text{blockdiag}}(\sigma_{u\ell}^2\boldsymbol{I}_{q_\ell^{\text{G}}}) & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}_m\otimes\boldsymbol{\Sigma}^{\text{R}}\end{bmatrix}\right), \end{aligned} \tag{3.1}$$

$$\begin{aligned}
\boldsymbol{\Sigma}^{\mathrm{R}}|a_1^{\mathrm{R}},\ldots,a_{q^{\mathrm{R}}}^{\mathrm{R}} &\sim \text{Inverse-Wishart}\left(\nu+q^{\mathrm{R}}-1,\ 2\,\nu\,\text{diag}(1/a_1^{\mathrm{R}},\ldots,1/a_{q^{\mathrm{R}}}^{\mathrm{R}})\right),\\
a_r^{\mathrm{R}} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(\tfrac{1}{2},1/A_{\mathrm{R}r}^2\right),\quad r=1,\ldots,q^{\mathrm{R}},\\
\sigma_{u\ell}^2|a_{u\ell} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2},\,1/a_{u\ell}),\quad a_{u\ell}\overset{\text{ind.}}{\sim}\text{Inverse-Gamma}(\tfrac{1}{2},\,1/A_{u\ell}^2).
\end{aligned}$$

The notation "Dist" denotes the type of response distribution including Gaussian, Student-t, Bernoulli or Poisson and $\boldsymbol{R}$ represents the additional parameters corresponding to the errors, e.g. $\boldsymbol{R}=\sigma_\varepsilon^2\boldsymbol{I}$ for the Gaussian response mixed model. We illustrated in the previous chapter approximate Bayesian fitting of model (3.1) via iterative coordinate descent MFVB algorithms. While perhaps not intuitively obvious, these algorithms are challenged by the increasing volume and complexity of longitudinal and multilevel data as we now elaborate.

A computational problem arising consistently across Algorithms 3, 4 and 6 lies in the update expression for the effects covariance matrix. Take the Gaussian response model as an example,

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\leftarrow\left(\mu_{q(1/\sigma_\varepsilon^2)}\,\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C}+\begin{bmatrix}\sigma_\beta^{-2}\boldsymbol{I}_p & \boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \underset{1\le\ell\le L}{\text{blockdiag}}(\mu_{q(1/\sigma_{u\ell}^2)}\boldsymbol{I}_{q_\ell^{\mathrm{G}}}) & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_m\otimes\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})}\end{bmatrix}\right)^{-1}. \quad (3.2)$$

This update expression requires storage and inversion of a large sparse matrix of dimension $(p+q^{\mathrm{R}}\,m+\sum_{\ell=1}^L q_\ell^{\mathrm{G}})\times(p+q^{\mathrm{R}}\,m+\sum_{\ell=1}^L q_\ell^{\mathrm{G}})$, where p is the number of predictors, q_ℓ^{G} ($1\le\ell\le L$) is the size of the ℓth penalised spline, typically in the range $15-40$ regardless sample size, and m is the number of groups that can be arbitrarily large. For the real data examples we consider in this chapter, m does not pose serious problems. However, other studies involve a much larger m, the *Six Cities Study of Air Pollution and Health* described in Fitzmaurice *et al.* (2012) with $m=13\,379$ is a practical example. Thus, without doubt the number of groups m dominates the dimension of $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ in many practical situations. Henceforth, if $p=10$, $q^{\mathrm{R}}=2$, $m=10\,000$, $\ell=1$ and $q_1^{\mathrm{G}}=25$, then the dimension of the matrix requiring storage and inversion would exceed $20\,000\times20\,000$. In addition, it is well known from numerical linear algebra that *naïve* computation of $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ is $O(m^3)$, that is cubic dependence on the number of groups. Hence, direct implementation of Algorithms 3, 4 and 6 can be very costly, or even infeasible, in large longitudinal and multilevel studies. Our aim in the next few sections is to *streamline* these MFVB algorithms by removing the computational obstacles involving update expressions such as (3.2) for most practical values of m and, thus, maximising the benefits of MFVB methods for approximate Bayesian inference.

3.3 Predictor structure and matrix notation

Here we introduce new matrix notation for our predictor structure that will benefit our presentation of the streamlined approach later on. Our predictor structure corresponds to the set up notation of Section 2 of Zhao *et al.* (2006) with the spatial correlation structure omitted. We first partition $\boldsymbol{\beta}$, $\boldsymbol{X}$, $\boldsymbol{u}$ and $\boldsymbol{Z}$ into random group effects (superscript R) and general spline components (superscript G) as follows:

$$\boldsymbol{\beta} \equiv \begin{bmatrix} \boldsymbol{\beta}^{\mathrm{R}} \\ \boldsymbol{\beta}^{\mathrm{G}} \end{bmatrix}, \quad \boldsymbol{X} \equiv [\boldsymbol{X}^{\mathrm{R}} \ \boldsymbol{X}^{\mathrm{G}}], \quad \boldsymbol{u} \equiv \begin{bmatrix} \boldsymbol{u}^{\mathrm{R}} \\ \boldsymbol{u}^{\mathrm{G}} \end{bmatrix} \quad \text{and} \quad \boldsymbol{Z} \equiv [\boldsymbol{Z}^{\mathrm{R}} \ \boldsymbol{Z}^{\mathrm{G}}],$$

which then leads to

$$\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} = \boldsymbol{X}^{\mathrm{R}}\boldsymbol{\beta}^{\mathrm{R}} + \boldsymbol{X}^{\mathrm{G}}\boldsymbol{\beta}^{\mathrm{G}} + \boldsymbol{Z}^{\mathrm{R}}\boldsymbol{u}^{\mathrm{R}} + \boldsymbol{Z}^{\mathrm{G}}\boldsymbol{u}^{\mathrm{G}}. \tag{3.3}$$

The random group effects design matrices are of the form

$$\boldsymbol{X}^{\mathrm{R}} \equiv \begin{bmatrix} \boldsymbol{X}^{\mathrm{R}}_1 \\ \vdots \\ \boldsymbol{X}^{\mathrm{R}}_m \end{bmatrix} \quad \text{and} \quad \boldsymbol{Z}^{\mathrm{R}} \equiv \underset{1 \le i \le m}{\mathrm{blockdiag}}(\boldsymbol{X}^{\mathrm{R}}_i),$$

where the $\boldsymbol{X}^{\mathrm{R}}_i$, $1 \le i \le m$, are $n_i \times q^{\mathrm{R}}$ design matrices. The random group effects vector is such that

$$\boldsymbol{u}^{\mathrm{R}} = \begin{bmatrix} \boldsymbol{u}^{\mathrm{R}}_1 \\ \vdots \\ \boldsymbol{u}^{\mathrm{R}}_m \end{bmatrix} \quad \text{and} \quad \mathrm{Cov}(\boldsymbol{u}^{\mathrm{R}}) = \boldsymbol{I}_m \otimes \boldsymbol{\Sigma}^{\mathrm{R}},$$

where $\boldsymbol{\Sigma}^{\mathrm{R}}$ is an unstructured $q^{\mathrm{R}} \times q^{\mathrm{R}}$ covariance matrix. Thus, $\boldsymbol{X}^{\mathrm{R}}\boldsymbol{\beta}^{\mathrm{R}} + \boldsymbol{Z}^{\mathrm{R}}\boldsymbol{u}^{\mathrm{R}}$ corresponds to random intercepts and slopes for repeated measures data on m groups with sample sizes $n_1, \ldots, n_m$.

The matrices $\boldsymbol{X}^{\mathrm{G}}$ and $\boldsymbol{Z}^{\mathrm{G}}$ are general design matrices corresponding to the fixed effects vector $\boldsymbol{\beta}^{\mathrm{G}}$ and spline coefficients vector $\boldsymbol{u}^{\mathrm{G}}$ respectively. Typically, $\boldsymbol{X}^{\mathrm{G}}$ contains polynomial functions of a continuous predictor that enter the model as a penalised spline. The $\boldsymbol{Z}^{\mathrm{G}}$ matrix would then contain spline basis functions of the same predictor. Mixed model based penalised spline fitting of (3.3) (e.g. Ruppert *et al.*, 2003) involves modelling a smooth, but otherwise unspecified, function f according to:

$$f(x) = \beta_x\, x + \sum_{k=1}^{q^{\mathrm{G}}} u^{\mathrm{G}}_k\, z_k(x), \quad u^{\mathrm{G}}_k \overset{\text{ind.}}{\sim} N(0, \sigma^2_u),$$

where $\{z_k : 1 \leq k \leq q^{\mathrm{G}}\}$ are spline bases of size q^{G} and σ_u^2 is the penalised parameter for the spline coefficients $u_1^{\mathrm{G}}, \ldots, u_{q^{\mathrm{G}}}^{\mathrm{G}}$. A common choice for the z_k is the suitably linearly transformed cubic O'Sullivan spline, as described in Section 4 of Wand and Ormerod (2008). The $\boldsymbol{Z}^{\mathrm{G}}\boldsymbol{u}^{\mathrm{G}}$ term can be further decomposed according to

$$\boldsymbol{Z}^{\mathrm{G}}\boldsymbol{u}^{\mathrm{G}} \equiv \sum_{\ell=1}^{L} \boldsymbol{Z}_\ell^{\mathrm{G}}\boldsymbol{u}_\ell^{\mathrm{G}} \quad \text{with} \quad \mathrm{Cov}(\boldsymbol{u}^{\mathrm{G}}) = \underset{1\leq\ell\leq L}{\mathrm{blockdiag}}(\sigma_{u\ell}^2 \boldsymbol{I}_{q_\ell^{\mathrm{G}}}). \tag{3.4}$$

Here the $\boldsymbol{u}_\ell^{\mathrm{G}}$ are $q_\ell^{\mathrm{G}} \times 1$ random spline coefficient vectors for $1 \leq \ell \leq L$, so the blocks of $\mathrm{Cov}(\boldsymbol{u}^{\mathrm{G}})$ correspond to the decomposition of $\boldsymbol{u}^{\mathrm{G}}$. The expression in (3.4) is in keeping with spline penalisation in multi-predictor semiparametric regression models (e.g. Ruppert *et al.*, 2003). Further, it is useful to define $\boldsymbol{C}^{\mathrm{G}} \equiv [\boldsymbol{X}\ \boldsymbol{Z}^{\mathrm{G}}]$ and partition $\boldsymbol{y}$, $\boldsymbol{X}$, $\boldsymbol{Z}$ and $\boldsymbol{C}^{\mathrm{G}}$ row-wise corresponding to the groups in $\boldsymbol{X}^{\mathrm{R}}$:

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_m \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_m \end{bmatrix}, \quad \boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z}_1 \\ \vdots \\ \boldsymbol{Z}_m \end{bmatrix} \quad \text{and} \quad \boldsymbol{C}^{\mathrm{G}} = \begin{bmatrix} \boldsymbol{C}_1^{\mathrm{G}} \\ \vdots \\ \boldsymbol{C}_m^{\mathrm{G}} \end{bmatrix}.$$

Here $\boldsymbol{y}_i \equiv [y_{i1} \ldots y_{in_i}]^{\mathsf{T}}$ denotes the $n_i \times 1$ vector of responses for the ith group. The matrices $\boldsymbol{X}_i$, $\boldsymbol{Z}_i$ and $\boldsymbol{C}_i^{\mathrm{G}}$ are defined in the same manner.

3.4 Streamlining mean field variational Bayes algorithms

The direct approach to obtaining the optimal q-density moments for all key parameters in Algorithms 3, 4 and 6 involves computation of $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$. As mentioned in Section 3.2, the matrix $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ increases in dimension as the number of groups m increases, and the time complexity is as high as $O(m^3)$, making variational algorithms become computationally infeasible in large longitudinal and multilevel studies. We get around this by exploiting the fact that most of the matrix requiring inversion is block-diagonal. Hence we propose a streamlined approach based around block decomposition of a matrix, using the Gaussian response model (2.11) as a demonstrating example. First we re-express (3.2) in a simpler

matrix form as given below:

$$\begin{aligned}\mu_{q(1/\sigma_\varepsilon^2)}\,\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C} + \begin{bmatrix} \sigma_\beta^{-2}\boldsymbol{I}_p & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \underset{1\le\ell\le L}{\text{blockdiag}}(\mu_{q(1/\sigma_{u\ell}^2)}\boldsymbol{I}_{q_\ell^{\text{G}}}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_m \otimes M_{q((\boldsymbol{\Sigma}^{\text{R}})^{-1})} \end{bmatrix} \\ = \begin{bmatrix} \boldsymbol{F} & \boldsymbol{G}_1 & \boldsymbol{G}_2 & \cdots & \boldsymbol{G}_m \\ \boldsymbol{G}_1^T & \boldsymbol{H}_1^{-1} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{G}_2^T & \boldsymbol{0} & \boldsymbol{H}_2^{-1} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{G}_m^T & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{H}_m^{-1} \end{bmatrix},\end{aligned} \tag{3.5}$$

where the submatrices $\boldsymbol{F}$, $\boldsymbol{G}_i$ and $\boldsymbol{H}_i$ are defined as follows:

$$\begin{aligned} \boldsymbol{F} &\equiv \mu_{q(1/\sigma_\varepsilon^2)}\,(\boldsymbol{C}^{\text{G}})^{\mathsf{T}}\boldsymbol{C}^{\text{G}} + \begin{bmatrix} \sigma_\beta^{-2}\boldsymbol{I}_p & \boldsymbol{0} \\ \boldsymbol{0} & \underset{1\le\ell\le L}{\text{blockdiag}}\left(\mu_{q(1/\sigma_{u\ell}^2)}\boldsymbol{I}_{q_\ell^{\text{G}}}\right) \end{bmatrix}, \\ \boldsymbol{G}_i &\equiv \mu_{q(1/\sigma_\varepsilon^2)}\,(\boldsymbol{C}_i^{\text{G}})^{\mathsf{T}}\boldsymbol{X}_i^{\text{R}}, \\ \text{and}\quad \boldsymbol{H}_i &\equiv \left\{\mu_{q(1/\sigma_\varepsilon^2)}\,(\boldsymbol{X}_i^{\text{R}})^{\mathsf{T}}\boldsymbol{X}_i^{\text{R}} + M_{q((\boldsymbol{\Sigma}^{\text{R}})^{-1})}\right\}^{-1}. \end{aligned}$$

The block-diagonal structure in the bottom right submatrix of (3.5), the contribution from the random group effects, is crucial as it enables us to reduce the inversion to $O(m)$ operations.

Result 3.4.1. Let $\boldsymbol{A}$ be a $m \times m$ matrix, $\boldsymbol{B}$ a $m \times n$ matrix, $\boldsymbol{C}$ an $n \times m$ matrix and $\boldsymbol{D}$ an $n \times n$ matrix. Suppose that the partitioned matrix $\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}$ is nonsingular, its inverse can then be obtained by solving a system of simultaneous linear equations in matrix form

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix} \begin{bmatrix} \widetilde{\boldsymbol{A}} & \widetilde{\boldsymbol{B}} \\ \widetilde{\boldsymbol{C}} & \widetilde{\boldsymbol{D}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix},$$

where the submatrices are defined as follows:

$$\begin{aligned} \widetilde{\boldsymbol{A}} &= (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1} \\ \widetilde{\boldsymbol{B}} &= \widetilde{\boldsymbol{A}}\boldsymbol{B}\boldsymbol{D}^{-1}, \quad \widetilde{\boldsymbol{C}} = -\boldsymbol{D}^{-1}\boldsymbol{C}\widetilde{\boldsymbol{A}} \\ \widetilde{\boldsymbol{D}} &= \boldsymbol{D}^{-1} + \boldsymbol{D}^{-1}\boldsymbol{C}\widetilde{\boldsymbol{A}}\boldsymbol{B}\boldsymbol{D}^{-1}, \end{aligned}$$

assuming the inverse matrices $\boldsymbol{D}^{-1}$ and $(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})$ above exist.

The standard Result 3.4.1 (Harville, 2008) on inversion of a block-partitioned matrix leads to the inverse of (3.5) equalling

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \equiv \begin{bmatrix} \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})} & \boldsymbol{\Lambda}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}},\boldsymbol{u}_1^{\mathrm{R}},\ldots,\boldsymbol{u}_m^{\mathrm{R}})} \\ \boldsymbol{\Lambda}^{\mathsf{T}}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}},\boldsymbol{u}_1^{\mathrm{R}},\ldots,\boldsymbol{u}_m^{\mathrm{R}})} & \boldsymbol{\Sigma}_{q(\boldsymbol{u}^{\mathrm{R}})} \end{bmatrix}, \tag{3.6}$$

where

$$\begin{aligned} \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})} &\equiv \left(\boldsymbol{F} - \sum_{i=1}^{m} \boldsymbol{G}_i\,\boldsymbol{H}_i\,\boldsymbol{G}_i^{\mathsf{T}}\right)^{-1} \\ \text{and} \quad \boldsymbol{\Lambda}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}},\boldsymbol{u}_1^{\mathrm{R}},\ldots,\boldsymbol{u}_m^{\mathrm{R}})} &\equiv \left[-\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})}\,\boldsymbol{G}_1\,\boldsymbol{H}_1 \quad \cdots - \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})}\,\boldsymbol{G}_m\,\boldsymbol{H}_m\right]. \end{aligned}$$

The dimension of the matrix $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})}$ is $(p + \sum_{\ell=1}^{L} q_\ell^{\mathrm{G}}) \times (p + \sum_{\ell=1}^{L} q_\ell^{\mathrm{G}})$ and is relatively straightforward to compute. The summation term renders this whole process as $O(m)$. This matrix is all we require to obtain variability bands for the penalised regression splines.

To find the variability estimates to accompany group-specific mean estimates, we need $\boldsymbol{\Sigma}_{q(\boldsymbol{u}^{\mathrm{R}})}$ which is not a block-diagonal matrix. However, since the covariance between the fitted values of two different groups is rarely of interest, it suffices to compute and store the diagonal blocks

$$\boldsymbol{\Sigma}_{q(\boldsymbol{u}_i^{\mathrm{R}})} \equiv \boldsymbol{H}_i + \boldsymbol{H}_i\,\boldsymbol{G}_i^{\mathsf{T}}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})}\,\boldsymbol{G}_i\,\boldsymbol{H}_i, \quad 1 \le i \le m. \tag{3.7}$$

Since $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})}$, $\boldsymbol{G}_i$ and $\boldsymbol{H}_i$ have dimensions much smaller than m, the complexity of the matrix calculations required in these submatrices does not increase as m increases. Therefore, the calculations with the highest order of complexity are the calculations of the m relevant submatrices of $\boldsymbol{\Lambda}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}},\boldsymbol{u}_i^{\mathrm{R}})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{u}_i^{\mathrm{R}})}$. This renders the whole process as $O(m)$, representing an improvement of order m^2 over the naïve approach to matrix inversion. As the number of groups increases, the improvements due to streamlining become enormous. It is worth mentioning that automatic sparse methods could be used, without requiring detailed algebra, but the downside of this is a possibly sub-optimal implementation compared to knowing the block-diagonal structure.

3.4.1 Streamlining update expressions involving $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$

We see in Algorithm 3 that the update expressions for the mean vector of the coefficient estimates and the scale parameter of the error variance involve the matrix $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$. Henceforth we now work towards a streamlined alternative to these updates.

Recall from Algorithm 3 that the naïve update for $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ is

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

By replacing $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ with (3.6) and partitioning $\boldsymbol{C}$ into $\left[(\boldsymbol{C}^\mathrm{G})^\mathsf{T} \, (\boldsymbol{Z}^\mathrm{R})^\mathsf{T}\right]^\mathsf{T}$, we can partition the above update into two expressions corresponding to $(\boldsymbol{\beta}, \boldsymbol{u}^\mathrm{G})$ and $\boldsymbol{u}^\mathrm{R}$ respectively:

$$\begin{bmatrix} \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^\mathrm{G})} \\ \boldsymbol{\mu}_{q(\boldsymbol{u}^\mathrm{R})} \end{bmatrix} \leftarrow \mu_{q(1/\sigma_\varepsilon^2)} \begin{bmatrix} \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^\mathrm{G})} & \boldsymbol{\Lambda}_{q(\boldsymbol{\beta},\boldsymbol{u}^\mathrm{G},\boldsymbol{u}_1^\mathrm{R},\ldots,\boldsymbol{u}_m^\mathrm{R})} \\ \boldsymbol{\Lambda}^\mathsf{T}_{q(\boldsymbol{\beta},\boldsymbol{u}^\mathrm{G},\boldsymbol{u}_1^\mathrm{R},\ldots,\boldsymbol{u}_m^\mathrm{R})} & \boldsymbol{\Sigma}_{q(\boldsymbol{u}^\mathrm{R})} \end{bmatrix} \begin{bmatrix} (\boldsymbol{C}^\mathrm{G})^\mathsf{T}\boldsymbol{y} \\ (\boldsymbol{Z}^\mathrm{R})^\mathsf{T}\boldsymbol{y} \end{bmatrix}.$$

Specifically, after some simple algebraic manipulations, we obtain the streamlined update for $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^\mathrm{G})}$ and $\boldsymbol{\mu}_{q(\boldsymbol{u}^\mathrm{R})}$ respectively:

$$\begin{aligned} \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^\mathrm{G})} &\leftarrow \mu_{q(1/\sigma_\varepsilon^2)} \left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^\mathrm{G})} (\boldsymbol{C}^\mathrm{G})^\mathsf{T}\boldsymbol{y} + \Lambda_{q(\boldsymbol{\beta},\boldsymbol{u}^\mathrm{G},\boldsymbol{u}_1^\mathrm{R}\cdots\boldsymbol{u}_m^\mathrm{R})} (\boldsymbol{Z}^\mathrm{R})^\mathsf{T}\boldsymbol{y}\right) \\ &= \mu_{q(1/\sigma_\varepsilon^2)} \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^\mathrm{G})} \left((\boldsymbol{C}^\mathrm{G})^\mathsf{T}\boldsymbol{y} - \sum_{i=1}^m \boldsymbol{G}_i \boldsymbol{H}_i (\boldsymbol{X}_i^\mathrm{R})^\mathsf{T}\boldsymbol{y}_i\right) \\ \text{and} \quad \boldsymbol{\mu}_{q(\boldsymbol{u}^\mathrm{R})} &\leftarrow \mu_{q(1/\sigma_\varepsilon^2)} \left(\boldsymbol{\Lambda}^\mathsf{T}_{q(\boldsymbol{\beta},\boldsymbol{u}^\mathrm{G},\boldsymbol{u}_1^\mathrm{R},\ldots,\boldsymbol{u}_m^\mathrm{R})} (\boldsymbol{C}^\mathrm{G})^\mathsf{T}\boldsymbol{y} + \boldsymbol{\Sigma}_{q(\boldsymbol{u}^\mathrm{R})} (\boldsymbol{Z}^\mathrm{R})^\mathsf{T}\boldsymbol{y}\right). \end{aligned} \tag{3.8}$$

Using (3.6) and (3.7), we can further break down $\boldsymbol{\mu}_{q(\boldsymbol{u}^\mathrm{R})}$ into components corresponding to the ith group only:

$$\begin{aligned} \boldsymbol{\mu}_{q(\boldsymbol{u}_i^\mathrm{R})} &\leftarrow \mu_{q(1/\sigma_\varepsilon^2)} \left\{ -\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^\mathrm{G})} \boldsymbol{G}_i \boldsymbol{H}_i (\boldsymbol{X}_i^\mathrm{R})^\mathsf{T}\boldsymbol{y}_i \right. \\ &\quad \left. + \left(\boldsymbol{H}_i (\boldsymbol{X}_i^\mathrm{R})^\mathsf{T}\boldsymbol{y}_i + \boldsymbol{H}_i \boldsymbol{G}_i^\mathsf{T} \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^\mathrm{G})} \sum_{i=1}^m \boldsymbol{G}_i \boldsymbol{H}_i (\boldsymbol{X}_i^\mathrm{R})^\mathsf{T}\boldsymbol{y}_i \right) \right\} \\ &= \boldsymbol{H}_i \left\{ \mu_{q(1/\sigma_\varepsilon^2)} (\boldsymbol{X}_i^\mathrm{R})^\mathsf{T}\boldsymbol{y}_i - \boldsymbol{G}_i^\mathsf{T} \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^\mathrm{G})} \right\}. \end{aligned} \tag{3.9}$$

Similarly, recall from Algorithm 3 that the naïve update for $B_{q(\sigma_\varepsilon^2)}$ is

$$B_{q(\sigma_\varepsilon^2)} \leftarrow \mu_{q(1/a_\varepsilon)} + \tfrac{1}{2}\, \mu_{q(1/\sigma_\varepsilon^2)} \left\{ (\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})})^\mathsf{T} (\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}) + \mathrm{tr}(\boldsymbol{C}^\mathsf{T}\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}) \right\}.$$

By replacing $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ with (3.8) and partitioning $\boldsymbol{C}$ into $\left[(\boldsymbol{C}^\mathrm{G})^\mathsf{T} \, (\boldsymbol{Z}^\mathrm{R})^\mathsf{T}\right]^\mathsf{T}$, we can rewrite

$\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ and $\text{tr}(\boldsymbol{C}^\mathsf{T}\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^\text{G})})$ in the following equivalent streamlined form:

$$\begin{aligned}
\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} &= \boldsymbol{C}^\text{G}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^\text{G})} + \begin{bmatrix} \boldsymbol{X}^\text{R}_1\boldsymbol{\mu}_{q(\boldsymbol{u}^\text{R}_1)} \\ \vdots \\ \boldsymbol{X}^\text{R}_m\boldsymbol{\mu}_{q(\boldsymbol{u}^\text{R}_m)} \end{bmatrix} \quad \text{and} \qquad (3.10) \\
\text{tr}\left(\boldsymbol{C}^\mathsf{T}\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^\text{G})}\right) &= \text{tr}\left\{(\boldsymbol{C}^\text{G})^\mathsf{T}\boldsymbol{C}^\text{G}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^\text{G})}\right\} - 2\mu^{-1}_{q(1/\sigma^2_\varepsilon)}\sum_{i=1}^m \text{tr}\left(\boldsymbol{G}_i\boldsymbol{H}_i\boldsymbol{G}_i^\mathsf{T}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^\text{G})}\right) \\
&\quad + \sum_{i=1}^m \text{tr}\left\{(\boldsymbol{X}^\text{R}_i)^\mathsf{T}\boldsymbol{X}^\text{R}_i\boldsymbol{\Sigma}_{q(\boldsymbol{u}^\text{R}_i)}\right\},
\end{aligned}$$

and the streamlined update for $B_{q(\sigma^2_\varepsilon)}$ then follows.

3.4.2 Streamlining lower bound expression involving $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$

The marginal log-likelihood lower bound $\log \underline{p}(\boldsymbol{y}; q)$ for Algorithm 3 is obtained via straightforward, albeit long-winded, algebra. The only change in the lower bound expression for the streamlined approach is to compute $\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}|$ more efficiently, which we now justify. It depends on the following result concerning determinant of block-diagonal matrices:

$$\begin{vmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{vmatrix} = |\boldsymbol{D}|\,|\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}|, \qquad (3.11)$$

where $\boldsymbol{D}$ is a square matrix and is invertible (e.g. Theorem 13.3.8 of Harville, 2008). From (3.11) we then get

$$\begin{aligned}
\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}| &= -\log\begin{vmatrix} \boldsymbol{F} & \boldsymbol{G}_1 & \boldsymbol{G}_2 & \cdots & \boldsymbol{G}_m \\ \boldsymbol{G}_1^\mathsf{T} & \boldsymbol{H}_1^{-1} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{G}_2^\mathsf{T} & \boldsymbol{0} & \boldsymbol{H}_2^{-1} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{G}_m^\mathsf{T} & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{H}_m^{-1} \end{vmatrix} \\
&= -\sum_{i=1}^m \log|\boldsymbol{H}_i^{-1}| - \log|\boldsymbol{\Sigma}^{-1}_{q(\boldsymbol{\beta},\boldsymbol{u}^\text{G})}|.
\end{aligned}$$

We illustrate step-by-step transformations of update expressions in Algorithm 3 concerning the $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ term for the Gaussian response model. It is, however, relatively straightforward to extrapolate the steps discussed and apply to other response models that were covered in Chapter 2. Hence we omit the details and directly present the streamlined algorithms for much faster and memory-efficient MFVB approximate fitting

of longitudinal and multilevel models. These are presented as Algorithms 7, 8 and 9.

Algorithms 7, 8 and 9 are an alternative to (restricted) maximum likelihood and best prediction-based fitting of large longitudinal and multilevel models, the latter being employed by popular software such as `PROC MIXED` in SAS (SAS Institute Inc., 2013) and `lmer()` in the R package `lme4` (Pinheiro *et al.*, 2014). Leaving aside the fact that the latter are based on frequentist inference paradigms, whilst Algorithms 7, 8 and 9 are intrinsically Bayesian, it is worth pointing out that these algorithms are purely matrix algebraic which has advantages such as stability, speed, parallelisability and online fitting. On the other hand, (restricted) maximum likelihood estimation of covariance matrices involves multi-dimensional, nonlinear optimisation (e.g. Wolfinger *et al.*, 1994). Such optimisation procedures become computationally burdensome for very large longitudinal and multilevel datasets, as exemplified by the computing times given in Table 3.5 as part of the speed comparisons given in Section 3.5. Earlier covariance estimation procedures, such as the minimum norm quadratic unbiased estimator method of Rao (1972) and the moment-based approach described in Section 3 of Goldstein (1986), are faster but have drawbacks such as positive definiteness not being guaranteed.

3.5 Computational speed: naïve versus streamlined

We conducted a simulation study to assess the computational speed of the naïve Algorithm 3 against that of the streamlined Algorithm 7. We generated 100 datasets according to the following simulation setting, focusing on the Gaussian response model.

$$y_{ij} \,|\, \beta_0, \beta_x, u^{\mathrm{R}}_{0i}, u^{\mathrm{R}}_{1i}, \sigma^2_\varepsilon \overset{\text{ind.}}{\sim} N\left(\beta_0 + u^{\mathrm{R}}_{0i} + (\beta_x + u^{\mathrm{R}}_{1i})\, x_{ij} + f(s_{ij}),\, \sigma^2_\varepsilon\right),$$

where

$$f(s) = 1 - \frac{13}{5\sqrt{2\pi}}\, e^{-(s-0.15)^2/0.2} - (2.3\, s - 0.07\, s^2) + 0.5\,\{1 - \Phi(s; 0.8, 0.07)\}$$

and $s_{ij} \overset{\text{ind.}}{\sim} \text{Uniform}(0, 1)$, $x_{ij} \overset{\text{ind.}}{\sim} \text{Uniform}(0, 1)$ and $\begin{bmatrix} u^{\mathrm{R}}_{0i} & u^{\mathrm{R}}_{1i} \end{bmatrix}^{\mathsf{T}} \Big| \boldsymbol{\Sigma}^{\mathrm{R}} \overset{\text{ind.}}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma}^{\mathrm{R}})$, for $1 \leq i \leq m$, $1 \leq j \leq n_i$. The true parameter values are:

$$\beta_0 = 0.58, \quad \beta_x = 1.89, \quad \sigma^2_\varepsilon = 0.04 \quad \text{and} \quad \boldsymbol{\Sigma}^{\mathrm{R}} = \begin{bmatrix} 2.58 & 0.22 \\ 0.22 & 1.73 \end{bmatrix}.$$

The number of groups is varied, $m \in \{100, 500, 2500, 12\,500\}$, with the within-group sample sizes n_i ranged between 10 and 20.

Both the naïve Algorithm 3 and streamlined Algorithm 7 were implemented in `Fortran`

Initialise: $\nu = 2$, $\mu_{q(1/\sigma_\varepsilon^2)} > 0$, $\mu_{q(1/a_\varepsilon)} > 0$, $\mu_{q(1/\sigma_{u\ell}^2)} > 0, 1 \leq \ell \leq L$, $\mu_{q(1/a_r^{\mathrm{R}})} > 0$, $1 \leq r \leq q^{\mathrm{R}}$, $\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})}$, a $q^{\mathrm{R}} \times q^{\mathrm{R}}$ positive definite matrix, $\mu_{q(1/b_i)} > 0, 1 \leq i \leq N$ and $\mu_{q(\nu)} > 0$.

Cycle through updates:

$$\textbf{Define: } \boldsymbol{M}_{q(\boldsymbol{\Omega})} \equiv \begin{bmatrix} \sigma_{\boldsymbol{\beta}}^{-2}\boldsymbol{I}_p & \boldsymbol{0} \\ \boldsymbol{0} & \underset{1\leq\ell\leq L}{\mathrm{blockdiag}}(\mu_{q(1/\sigma_{u\ell}^2)}\boldsymbol{I}_{q_\ell^{\mathrm{G}}}) \end{bmatrix}$$

$N \leftarrow \sum_{i=1}^m n_i$; Gaussian: $\boldsymbol{\Psi} = \boldsymbol{I}$; Student-t: $\boldsymbol{\Psi} \leftarrow \underset{1\leq i\leq N}{\mathrm{diag}}\left(\mu_{q(1/b_i)}\right)$

$\boldsymbol{S} \leftarrow \boldsymbol{0}$; $\boldsymbol{s} \leftarrow \boldsymbol{0}$

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

$$\boldsymbol{G}_i \leftarrow \mu_{q(1/\sigma_\varepsilon^2)}(\boldsymbol{C}_i^{\mathrm{G}})^{\mathsf{T}}\boldsymbol{\Psi}_i\boldsymbol{X}_i^{\mathrm{R}}$$
$$\boldsymbol{H}_i \leftarrow \left\{\mu_{q(1/\sigma_\varepsilon^2)}(\boldsymbol{X}_i^{\mathrm{R}})^{\mathsf{T}}\boldsymbol{\Psi}_i\boldsymbol{X}_i^{\mathrm{R}} + \boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})}\right\}^{-1}$$
$$\boldsymbol{S} \leftarrow \boldsymbol{S} + \boldsymbol{G}_i\boldsymbol{H}_i\boldsymbol{G}_i^{\mathsf{T}} \ ; \ \boldsymbol{s} \leftarrow \boldsymbol{s} + \boldsymbol{G}_i\boldsymbol{H}_i(\boldsymbol{X}_i^{\mathrm{R}})^{\mathsf{T}}\boldsymbol{y}_i$$

1. Update multivariate normal $q^*(\boldsymbol{\beta}, \boldsymbol{u}^{\mathrm{G}})$ parameters:

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})} \leftarrow \left\{\mu_{q(1/\sigma_\varepsilon^2)}(\boldsymbol{C}^{\mathrm{G}})^{\mathsf{T}}\boldsymbol{\Psi}\boldsymbol{C}^{\mathrm{G}} + \boldsymbol{M}_{q(\boldsymbol{\Omega})} - \boldsymbol{S}\right\}^{-1}$$

$$\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})} \leftarrow \mu_{q(1/\sigma_\varepsilon^2)}\,\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})}\left\{(\boldsymbol{C}^{\mathrm{G}})^{\mathsf{T}}\boldsymbol{y} - \boldsymbol{s}\right\}$$

Update multivariate normal $q^*(\boldsymbol{u}_i^{\mathrm{R}})$ parameters:

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

$$\boldsymbol{\Sigma}_{q(\boldsymbol{u}_i^{\mathrm{R}})} \leftarrow \boldsymbol{H}_i + \boldsymbol{H}_i\,\boldsymbol{G}_i^{\mathsf{T}}\,\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})}\,\boldsymbol{G}_i\,\boldsymbol{H}_i$$
$$\boldsymbol{\mu}_{q(\boldsymbol{u}_i^{\mathrm{R}})} \leftarrow \boldsymbol{H}_i\left\{\mu_{q(1/\sigma_\varepsilon^2)}(\boldsymbol{X}_i^{\mathrm{R}})^{\mathsf{T}}\boldsymbol{y}_i - \boldsymbol{G}_i^{\mathsf{T}}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})}\right\}$$

2. Update inverse-Gamma $q^*(\sigma_\varepsilon^2)$ and inverse-Gamma $q^*(a_\varepsilon)$ parameters:

$$\boldsymbol{R} \leftarrow \boldsymbol{y} - \boldsymbol{C}^{\mathrm{G}}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})} - \begin{bmatrix} \boldsymbol{X}_1^{\mathrm{R}}\boldsymbol{\mu}_{q(\boldsymbol{u}_1^{\mathrm{R}})} \\ \vdots \\ \boldsymbol{X}_m^{\mathrm{R}}\boldsymbol{\mu}_{q(\boldsymbol{u}_m^{\mathrm{R}})} \end{bmatrix}$$

$$B_{q(\sigma_\varepsilon^2)} \leftarrow \mu_{q(1/a_\varepsilon)} + \tfrac{1}{2}\left[\boldsymbol{R}^{\mathsf{T}}\boldsymbol{\Psi}\boldsymbol{R} + \mathrm{tr}\{(\boldsymbol{C}^{\mathrm{G}})^{\mathsf{T}}\boldsymbol{\Psi}\boldsymbol{C}^{\mathrm{G}}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})}\}\right.$$
$$+\textstyle\sum_{i=1}^m \mathrm{tr}\{(\boldsymbol{X}_i^{\mathrm{R}})^{\mathsf{T}}\boldsymbol{\Psi}_i\boldsymbol{X}_i^{\mathrm{R}}\boldsymbol{\Sigma}_{q(\boldsymbol{u}_i^{\mathrm{R}})}\}$$
$$\left.-2(\mu_{q(1/\sigma_\varepsilon^2)})^{-1}\textstyle\sum_{i=1}^m \mathrm{tr}(\boldsymbol{G}_i\boldsymbol{H}_i\boldsymbol{G}_i^{\mathsf{T}}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})})\right]$$

$$\mu_{q(1/\sigma_\varepsilon^2)} \leftarrow \tfrac{1}{2}\left(\textstyle\sum_{i=1}^m n_i + 1\right)/B_{q(\sigma_\varepsilon^2)} \ ; \ \mu_{q(1/a_\varepsilon)} \leftarrow 1/(\mu_{q(1/\sigma_\varepsilon^2)} + A_\varepsilon^{-2})$$

3. Update inverse-Gamma $q^*(b_i)$ and $q^*(\nu)$ parameters for the Student-t response model:

$\boldsymbol{\zeta} \leftarrow \text{diagonal}\{\boldsymbol{C}^{\text{G}}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\text{G}})}(\boldsymbol{C}^{\text{G}})^{\mathsf{T}}\}$

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

$\boldsymbol{\zeta}_i \leftarrow \boldsymbol{\zeta}_i - 2\,\text{diagonal}\{\boldsymbol{X}_i^{\text{R}}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\text{G}})}\boldsymbol{G}_i\boldsymbol{H}_i)^{\mathsf{T}}(\boldsymbol{C}_i^{\text{G}})^{\mathsf{T}}\}$

$\boldsymbol{\zeta}_i \leftarrow \boldsymbol{\zeta}_i + \text{diagonal}\{\boldsymbol{X}_i^{\text{R}}\boldsymbol{\Sigma}_{q(\boldsymbol{u}_i^{\text{R}})}(\boldsymbol{X}_i^{\text{R}})^{\mathsf{T}}\}$

For $1 \le i \le N$:

$B_{q(b_i)} \leftarrow \frac{1}{2}\left[\mu_{q(\nu)} + \mu_{q(1/\sigma_\varepsilon^2)}\left\{(\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})})_i^2 + \boldsymbol{\zeta}_i\right\}\right]$

$\mu_{q(1/b_i)} \leftarrow \frac{1}{2}(\mu_{q(\nu)} + 1)/B_{q(b_i)}$

$\mu_{q(\log b_i)} \leftarrow \log(B_{q(b_i)}) - \text{digamma}\left\{\frac{1}{2}(\mu_{q(\nu)} + 1)\right\}$

$C_1 \leftarrow \sum_{i=1}^N \left(\mu_{q(\log(b_i))} + \mu_{q(1/b_i)}\right)$

$\mu_{q(\nu)} \leftarrow \exp\left\{\log \mathcal{F}(1, N, C_1, \nu_{\min}, \nu_{\max}) - \log \mathcal{F}(0, N, C_1, \nu_{\min}, \nu_{\max})\right\}$

4. Update inverse-Gamma $q^*(a_{u\ell})$ and inverse-Gamma $q^*(\sigma_{u\ell}^2)$ parameters:

For $\ell = 1, \ldots, L$:

$\mu_{q(1/a_{u\ell})} \leftarrow 1/(\mu_{q(1/\sigma_{u\ell}^2)} + A_{u\ell}^{-2})$

$$\mu_{q(\sigma_{u\ell}^2)} \leftarrow \frac{q_\ell^{\text{G}} + 1}{||\boldsymbol{\mu}_{q(\boldsymbol{u}_\ell^{\text{G}})}||^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_\ell^{\text{G}})}) + 2\,\mu_{q(1/a_{u\ell})}}$$

5. Update inverse-Gamma $q^*(a_r^{\text{R}})$ and inverse-Wishart $q^*(\boldsymbol{\Sigma}^{\text{R}})$ parameters:

For $r = 1, \ldots, q^{\text{R}}$:

$$B_{q(a_r^{\text{R}})} \leftarrow \nu\{\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\text{R}})^{-1})}\}_{rr} + A_{\text{R}r}^{-2} \quad ; \quad \mu_{q(a_r^{\text{R}})} \leftarrow \tfrac{1}{2}(\nu + q^{\text{R}})/B_{q(a_r^{\text{R}})}$$

$$\boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\text{R}})} \leftarrow \sum_{i=1}^m (\boldsymbol{\mu}_{q(\boldsymbol{u}_i^{\text{R}})}\boldsymbol{\mu}_{q(\boldsymbol{u}_i^{\text{R}})}^{\mathsf{T}} + \boldsymbol{\Sigma}_{q(\boldsymbol{u}_i^{\text{R}})}) + 2\,\nu\,\text{diag}(\mu_{q(1/a_1^{\text{R}})}, \ldots, \mu_{q(1/a_{q^{\text{R}}}^{\text{R}})})$$

$$\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\text{R}})^{-1})} \leftarrow (\nu + m + q^{\text{R}} - 1)\ \boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\text{R}})}^{-1}$$

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

Update q-density covariance matrix for the ith group:

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

$$\boldsymbol{\Lambda}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\text{G}},\boldsymbol{u}_i^{\text{R}})} \equiv E_q\left[\left(\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}^{\text{G}}\end{bmatrix} - \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\text{G}})}\right)(\boldsymbol{u}_i^{\text{R}} - \boldsymbol{\mu}_{q(\boldsymbol{u}_i^{\text{R}})})^{\mathsf{T}}\right] \leftarrow -\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\text{G}})}\boldsymbol{G}_i\,\boldsymbol{H}_i$$

Algorithm 7: Streamlined mean field variational Bayes algorithm for obtaining model parameters in the optimal q-density functions for the two-level Bayesian semiparametric mixed models with Gaussian and Student-t responses.

Initialise: $\nu = 2$, $\mu_{q(1/\sigma^2_{u\ell})} > 0$, $1 \leq \ell \leq L$, $\mu_{q(1/a^{\mathrm{R}}_r)} > 0$, $1 \leq r \leq q^{\mathrm{R}}$, $\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})}$, a $q^{\mathrm{R}} \times q^{\mathrm{R}}$ positive definite matrix and $\boldsymbol{\xi}$, a $(\sum_{i=1}^{m} n_i) \times 1$ vector of positive entries.

Cycle through updates:

Define: $\boldsymbol{M}_{q(\boldsymbol{\Omega})} \equiv \begin{bmatrix} \sigma_{\boldsymbol{\beta}}^{-2}\boldsymbol{I}_p & \boldsymbol{0} \\ \boldsymbol{0} & \underset{1\leq\ell\leq L}{\mathrm{blockdiag}}(\mu_{q(1/\sigma^2_{u\ell})}\boldsymbol{I}_{q^{\mathrm{G}}_\ell}) \end{bmatrix}$

$\boldsymbol{S} \leftarrow \boldsymbol{0}$; $\boldsymbol{s} \leftarrow \boldsymbol{0}$

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

$\boldsymbol{G}_i \leftarrow 2(\boldsymbol{C}^{\mathrm{G}}_i)^{\mathsf{T}}\mathrm{diag}\{\lambda(\boldsymbol{\xi}_i)\}\boldsymbol{X}^{\mathrm{R}}_i$

$\boldsymbol{H}_i \leftarrow \left\{2(\boldsymbol{X}^{\mathrm{R}}_i)^{\mathsf{T}}\mathrm{diag}\{\lambda(\boldsymbol{\xi}_i)\}\boldsymbol{X}^{\mathrm{R}}_i + \boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})}\right\}^{-1}$

$\boldsymbol{S} \leftarrow \boldsymbol{S} + \boldsymbol{G}_i\boldsymbol{H}_i\boldsymbol{G}_i^{\mathsf{T}}$; $\boldsymbol{s} \leftarrow \boldsymbol{s} + \boldsymbol{G}_i\boldsymbol{H}_i(\boldsymbol{X}^{\mathrm{R}}_i)^{\mathsf{T}}(\boldsymbol{y}_i - \frac{1}{2}\boldsymbol{1})$

1. Update Multivariate Normal $q^*(\boldsymbol{\beta}, \boldsymbol{u}^{\mathrm{G}})$ parameters:

$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})} \leftarrow \left\{2(\boldsymbol{C}^{\mathrm{G}})^{\mathsf{T}}\mathrm{diag}\{\lambda(\boldsymbol{\xi})\}\boldsymbol{C}^{\mathrm{G}} + \boldsymbol{M}_{q(\boldsymbol{\Omega})} - \boldsymbol{S}\right\}^{-1}$

$\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})} \leftarrow 2\,\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})}\left\{(\boldsymbol{C}^{\mathrm{G}})^{\mathsf{T}}(\boldsymbol{y} - \frac{1}{2}\boldsymbol{1}) - \boldsymbol{s}\right\}$

Update Multivariate Normal $q^*(\boldsymbol{u}^{\mathrm{R}}_i)$ parameters:

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

$\boldsymbol{\Sigma}_{q(\boldsymbol{u}^{\mathrm{R}}_i)} \leftarrow \boldsymbol{H}_i + \boldsymbol{H}_i\,\boldsymbol{G}_i^{\mathsf{T}}\,\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})}\,\boldsymbol{G}_i\,\boldsymbol{H}_i$

$\boldsymbol{\mu}_{q(\boldsymbol{u}^{\mathrm{R}}_i)} \leftarrow \boldsymbol{H}_i\left\{2(\boldsymbol{X}^{\mathrm{R}}_i)^T(\boldsymbol{y}_i - \frac{1}{2}\boldsymbol{1}) - \boldsymbol{G}_i^T\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})}\right\}$

$\boldsymbol{\xi}^2 \leftarrow \mathrm{diagonal}\{\boldsymbol{C}^{\mathrm{G}}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})} + \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})}\boldsymbol{\mu}^{\mathsf{T}}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})})(\boldsymbol{C}^{\mathrm{G}})^{\mathsf{T}}\}$

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

$\boldsymbol{\xi}^2_i \leftarrow \boldsymbol{\xi}^2_i + 2\,\mathrm{diagonal}\{\boldsymbol{C}^{\mathrm{G}}(-\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})}\boldsymbol{G}_i\boldsymbol{H}_i + \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})}\boldsymbol{\mu}^{\mathsf{T}}_{q(\boldsymbol{u}^{\mathrm{R}}_i)})(\boldsymbol{X}^{\mathrm{R}}_i)^{\mathsf{T}}\}$

$\boldsymbol{\xi}^2_i \leftarrow \boldsymbol{\xi}^2_i + \mathrm{diagonal}\{(\boldsymbol{X}^{\mathrm{R}}_i(\boldsymbol{\Sigma}_{q(\boldsymbol{u}^{\mathrm{R}}_i)} + \boldsymbol{\mu}_{q(\boldsymbol{u}^{\mathrm{R}}_i)}\boldsymbol{\mu}^{\mathsf{T}}_{q(\boldsymbol{u}^{\mathrm{R}}_i)})(\boldsymbol{X}^{\mathrm{R}}_i)^{\mathsf{T}}\}$

2. Update inverse-Gamma $q^*(a_{u\ell})$ and inverse-Gamma $q^*(\sigma^2_{u\ell})$ parameters:

For $\ell = 1, \ldots, L$:

$\mu_{q(1/a_{u\ell})} \leftarrow 1/(\mu_{q(1/\sigma^2_{u\ell})} + A^{-2}_{u\ell})$

$\mu_{q(\sigma^2_{u\ell})} \leftarrow \dfrac{q^{\mathrm{G}}_\ell + 1}{||\boldsymbol{\mu}_{q(\boldsymbol{u}^{\mathrm{G}}_\ell)}||^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}^{\mathrm{G}}_\ell)}) + 2\,\mu_{q(1/a_{u\ell})}}$

3. Update inverse-Gamma $q^*(a_r^{\mathrm{R}})$ and inverse-Wishart $q^*(\boldsymbol{\Sigma}^{\mathrm{R}})$ parameters:

For $\ell = 1, \ldots, q^{\mathrm{R}}$:

$$B_{q(a^{\mathrm{R}})} \leftarrow \nu(\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})})_{rr} + A_{\mathrm{R}r}^{-2} \;\; ; \;\; \mu_{q(a_r^{\mathrm{R}})} \leftarrow \tfrac{1}{2}(\nu + q^{\mathrm{R}})/B_{q(a_r^{\mathrm{R}})}$$

$$\boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\mathrm{R}})} \leftarrow \sum_{i=1}^{m}(\boldsymbol{\mu}_{q(\boldsymbol{u}_i^{\mathrm{R}})}\boldsymbol{\mu}_{q(\boldsymbol{u}_i^{\mathrm{R}})}^{\mathsf{T}} + \boldsymbol{\Sigma}_{q(\boldsymbol{u}_i^{\mathrm{R}})}) + 2\,\nu\,\mathrm{diag}(\mu_{q(1/a_1^{\mathrm{R}})}, \ldots, \mu_{q(1/a_{q^{\mathrm{R}}}^{\mathrm{R}})})$$

$$\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})} \leftarrow (\nu + \textstyle\sum_{i=1}^{m} n_i + q^{\mathrm{R}} - 1)\; \boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\mathrm{R}})}^{-1}$$

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

Update q-density covariance matrix for the ith group:

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

$$\boldsymbol{\Lambda}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}},\boldsymbol{u}_i^{\mathrm{R}})} \equiv E_q\left[\left(\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}^{\mathrm{G}}\end{bmatrix} - \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})}\right)(\boldsymbol{u}_i^{\mathrm{R}} - \boldsymbol{\mu}_{q(\boldsymbol{u}_i^{\mathrm{R}})})^{\mathsf{T}}\right] \leftarrow -\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})}\boldsymbol{G}_i\,\boldsymbol{H}_i$$

Algorithm 8: Streamlined mean field variational Bayes algorithm for obtaining model parameters in the optimal q-density functions for the two-level Bayesian semiparametric mixed model with Bernoulli response.

77. We also compared Algorithm 7 with contemporary software for fitting the frequentist version of (3.12). This was achieved using the function `gamm()` in the R package `mgcv` (Wood, 2014). Note that `gamm()` uses the function `lme()` in the R package `nlme` (Pinheiro *et al.*, 2014) as a fitting engine. All computations were performed on a Mac OS X laptop with a 2.6 GHz Intel Core i5 processor and 8 GBytes of random access memory.

Table 3.5 summarises the average (standard error) computing times over 100 runs and shows the practical benefits of the streamlined MFVB approach. As m increases the average computing time for the naïve approach increases rapidly from 0.2 seconds to almost 25 minutes, compared with about a quarter to half of a second for the streamlined approach. For $m = 12\,500$, the naïve approach failed due to required storage of $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\,\boldsymbol{u})}$ exceeding memory restrictions for typical 2015 personal computing environments. However, the streamlined approach just took over 1.8 seconds on average to compute. The impressive speed gains in the streamlined approach are clearly reflected in the ratios of naïve over streamlined and the ratios of `gamm()` over streamlined with respect to the average computing times. In situations where a dataset has a large number of groups, the streamlined approach is more than 3500 times faster than the naïve approach and 1700 times faster than the `gamm`() approach.

Initialise: $\nu = 2$, $\mu_{q(1/\sigma^2_{u\ell})} > 0$, $1 \leq \ell \leq L$, $\mu_{q(1/a^{\mathrm{R}}_r)} > 0$, $1 \leq r \leq q^{\mathrm{R}}$, $\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})}$, a $q^{\mathrm{R}} \times q^{\mathrm{R}}$ positive definite matrix, $\boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})}$, a $(\sum_{i=1}^{m} n_i) \times 1$ vector of positive entries and $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})}$, a $(p + \sum_{l=1}^{L} q^{\mathrm{G}}_{\ell}) \times 1$ vector.

Cycle through updates:

Define: $\boldsymbol{M}_{q(\boldsymbol{\Omega})} \equiv \begin{bmatrix} \sigma_{\boldsymbol{\beta}}^{-2}\boldsymbol{I}_p & \boldsymbol{0} \\ \boldsymbol{0} & \underset{1\leq\ell\leq L}{\mathrm{blockdiag}}(\mu_{q(1/\sigma^2_{u\ell})}\boldsymbol{I}_{q^{\mathrm{G}}_{\ell}}) \end{bmatrix}$

$\boldsymbol{S} \leftarrow \boldsymbol{0}$; $\boldsymbol{s} \leftarrow \boldsymbol{0}$

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

$\boldsymbol{G}_i \leftarrow (\boldsymbol{C}^{\mathrm{G}}_i)^{\mathsf{T}}\mathrm{diag}\{(\boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})})_i\}\boldsymbol{X}^{\mathrm{R}}_i$

$\boldsymbol{H}_i \leftarrow \left\{(\boldsymbol{X}^{\mathrm{R}}_i)^{\mathsf{T}}\mathrm{diag}\{(\boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})})_i\}\boldsymbol{X}^{\mathrm{R}}_i + \boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})}\right\}^{-1}$

$\boldsymbol{S} \leftarrow \boldsymbol{S} + \boldsymbol{G}_i\boldsymbol{H}_i\boldsymbol{G}_i^{\mathsf{T}}$; $\boldsymbol{s} \leftarrow \boldsymbol{s} + \boldsymbol{G}_i\boldsymbol{H}_i(\boldsymbol{X}^{\mathrm{R}}_i)^{\mathsf{T}}\{\boldsymbol{y}_i - (\boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})})_i\}$

1. Update Multivariate Normal $q^*(\boldsymbol{\beta}, \boldsymbol{u}^{\mathrm{G}})$ parameters:

$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})} \leftarrow \left\{(\boldsymbol{C}^{\mathrm{G}})^{\mathsf{T}}\boldsymbol{\Psi}\boldsymbol{C}^{\mathrm{G}} + \boldsymbol{M}_{q(\boldsymbol{\Omega})} - \boldsymbol{S}\right\}^{-1}$

$\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})} \leftarrow \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})} + \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})}\left\{(\boldsymbol{C}^{\mathrm{G}})^{\mathsf{T}}\boldsymbol{y} - \boldsymbol{s} - \mathrm{diagonal}(\boldsymbol{M}_{q(\boldsymbol{\Omega})})\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})}\right\}$

$\boldsymbol{\mu}^{\mathrm{old}}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})} \leftarrow \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})}$

Update Multivariate Normal $q^*(\boldsymbol{u}^{\mathrm{R}}_i)$ parameters:

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

$\boldsymbol{\Sigma}_{q(\boldsymbol{u}^{\mathrm{R}}_i)} \leftarrow \boldsymbol{H}_i + \boldsymbol{H}_i\,\boldsymbol{G}_i^{\mathsf{T}}\,\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})}\,\boldsymbol{G}_i\,\boldsymbol{H}_i$

$\boldsymbol{\mu}_{q(\boldsymbol{u}^{\mathrm{R}}_i)} \leftarrow \boldsymbol{\mu}_{q(\boldsymbol{u}^{\mathrm{R}}_i)} + \boldsymbol{H}_i\left[2(\boldsymbol{X}^{\mathrm{R}}_i)^{\mathsf{T}}\{\boldsymbol{y}_i - (\boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})})_i\} - \boldsymbol{G}_i^{\mathsf{T}}(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})} - \boldsymbol{\mu}^{\mathrm{old}}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})})\right]$

$\boldsymbol{\zeta} \leftarrow \mathrm{diagonal}\{\boldsymbol{C}^{\mathrm{G}}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})}(\boldsymbol{C}^{\mathrm{G}})^{\mathsf{T}}\}$

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

$\boldsymbol{\zeta}_i \leftarrow \boldsymbol{\zeta}_i - 2\,\mathrm{diagonal}\{\boldsymbol{X}^{\mathrm{R}}_i(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})}\boldsymbol{G}_i\boldsymbol{H}_i)^{\mathsf{T}}(\boldsymbol{C}^{\mathrm{G}}_i)^{\mathsf{T}}\}$

$\boldsymbol{\zeta}_i \leftarrow \boldsymbol{\zeta}_i + \mathrm{diagonal}\{\boldsymbol{X}^{\mathrm{R}}_i\boldsymbol{\Sigma}_{q(\boldsymbol{u}^{\mathrm{R}}_i)}(\boldsymbol{X}^{\mathrm{R}}_i)^{\mathsf{T}}\}$

$\boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \exp\left(\boldsymbol{C}^{\mathrm{G}}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})} + \begin{bmatrix} \boldsymbol{X}^{\mathrm{R}}_1\boldsymbol{\mu}_{q(\boldsymbol{u}^{\mathrm{R}}_1)} \\ \vdots \\ \boldsymbol{X}^{\mathrm{R}}_m\boldsymbol{\mu}_{q(\boldsymbol{u}^{\mathrm{R}}_m)} \end{bmatrix} + \tfrac{1}{2}\,\boldsymbol{\zeta}\right)$

4. Update inverse-Gamma $q^*(a_{u\ell})$ and inverse-Gamma $q^*(\sigma^2_{u\ell})$ parameters:

For $\ell = 1, \ldots, L$:

$$\mu_{q(1/a_{u\ell})} \leftarrow 1/(\mu_{q(1/\sigma^2_{u\ell})} + A^{-2}_{u\ell})$$

$$\mu_{q(\sigma^2_{u\ell})} \leftarrow \frac{q^{\mathrm{G}}_{\ell} + 1}{||\boldsymbol{\mu}_{q(\boldsymbol{u}^{\mathrm{G}}_{\ell})}||^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}^{\mathrm{G}}_{\ell})}) + 2\,\mu_{q(1/a_{u\ell})}}$$

5. Update inverse-Gamma $q^*(a^{\mathbf{R}}_r)$ and inverse-Wishart $q^*(\boldsymbol{\Sigma}^{\mathbf{R}})$parameters:

For $r = 1, \ldots, q^{\mathrm{R}}$:

$$B_{q(a^{\mathrm{R}}_r)} \leftarrow \nu(\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})})_{rr} + A^{-2}_{\mathrm{R}r} \quad ; \quad \mu_{q(a^{\mathrm{R}}_r)} \leftarrow \tfrac{1}{2}(\nu + q^{\mathrm{R}})/B_{q(a^{\mathrm{R}}_r)}$$

$$\boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\mathrm{R}})} \leftarrow \sum_{i=1}^{m}(\boldsymbol{\mu}_{q(\boldsymbol{u}^{\mathrm{R}}_i)}\boldsymbol{\mu}^{\mathsf{T}}_{q(\boldsymbol{u}^{\mathrm{R}}_i)} + \boldsymbol{\Sigma}_{q(\boldsymbol{u}^{\mathrm{R}}_i)}) + 2\,\nu\,\mathrm{diag}(\mu_{q(1/a^{\mathrm{R}}_1)}, \ldots, \mu_{q(1/a^{\mathrm{R}}_{q^{\mathrm{R}}})})$$

$$\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})} \leftarrow (\nu + m + q^{\mathrm{R}} - 1)\ \boldsymbol{B}^{-1}_{q(\boldsymbol{\Sigma}^{\mathrm{R}})}$$

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

Update q-density covariance matrix for the ith group:

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

$$\boldsymbol{\Lambda}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}},\boldsymbol{u}^{\mathrm{R}}_i)} \equiv E_q\left[\left(\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}^{\mathrm{G}}\end{bmatrix} - \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})}\right)(\boldsymbol{u}^{\mathrm{R}}_i - \boldsymbol{\mu}_{q(\boldsymbol{u}^{\mathrm{R}}_i)})^{\mathsf{T}}\right] \leftarrow -\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})}\boldsymbol{G}_i\,\boldsymbol{H}_i$$

Algorithm 9: Streamlined mean field variational Bayes algorithm for obtaining model parameters in the optimal q-density functions for the two-level Bayesian semiparametric mixed model with Poisson response.

m	Naïve	Streamlined	`gamm`	Naïve/Stream.	`gamm`/Stream.
100	0.198 (0.023)	0.027 (0.003)	0.17 (0.010)	7.4	5.8
500	14.482 (1.704)	0.106 (0.021)	7.075 (0.165)	136.8	76.2
2500	1488.757 (108.311)	0.419 (0.056)	739.687 (17.793)	3550.4	1765.4
12 500	Failed	1.766 (0.096)	Failed	N/A	N/A

Table 3.1: Average (standard error) elapsed of the computing times in seconds for the simulation described in the text, using the naïve Algorithm 3, streamlined Algorithm 7 and `gam`() function in R `mgcv` package. The ratio of naïve over streamlined and the ratio of `gamm()` over streamlined are also presented.

3.6 Real data applications

We now provide illustration of our streamlined MFVB methodology through a series of re-analysis of real data examples. Throughout this section, we standardise the response and continuous variables to have zero means and unit standard deviations. We use a normal prior for $\boldsymbol{\beta}$ and a half-Cauchy prior for σ^2_{R}, σ^2_{ε} and σ^2_{u}, with the values of the hyperparameters all set to 10 000.

3.6.1 Application to smoking data

The first example is based on the perinatal health dataset from the United States National Center for Health Statistics, which is discussed and analysed by Abrevaya (2006). A 10% random sub-sample is provided by Rabe-Hesketh and Skrondal (2008) and our analysis is of this subset. The data have a two-level structure with 8604 births as units at Level 1 and 3978 mothers as groups at Level 2. There are an average of 2.2 births per mother. The following variables are given in Table 3.2.

In this example, study variables can either vary at the birth level (therefore also at the mother level) or at the mother level only. For instance, `smoke` is a Level-1 variable as maternal smoking status can change from one pregnancy to the next, whereas `black` is a Level-2 variable as mother's ethnicity remains unchanged between pregnancies.

Motivated by a report from the United States Surgeon General: "Infants born to women who smoke during pregnancy have a lower average birthweight and are more likely to be small for gestational age than infants born to women who do not smoke ..." Abrevaya (2006) examined the effect of smoking on birthweight for women in the United States between 1990 and 1998 using a matched panel data approach. Here we re-analyse their data using a different approach via streamlined MFVB. The main study factor is maternal smoking status and the response of interest is infant's birthweight. The following Bayesian Gaussian semiparametric mixed model with a random intercept for each mother was fitted

Variable	Description
`momid`	mother identifier
`birwt`	birthweight in grams
`gestat`	infant's gestational age in weeks
`mage`	mother's age at the birth of the infant in years
`smoke`	indicator for mother smoking during pregnancy
`male`	indicator for infant being male
`married`	indicator for mother being married
`hsgrad`	indicator for mother having some college education, but not degree
`somecoll`	indicator for mother having graduated from college
`black`	indicator for mother being black
`kessner2`	indicator for Kessner index equalling 2
`kessner3`	indicator for Kessner index equalling 3
`novisit`	indicator for no prenatal care visit
`pretri2`	indicator for first prenatal care visit having occurred in second trimester
`pretri3`	indicator for first prenatal care visit having occurred in third trimester

Table 3.2: Description of the United States National Center for Health Statistics perinatal health data as presented in Abrevaya (2006).

via Algorithm 7:

$$\begin{aligned}
\texttt{birwt}_{ij}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2 \overset{\text{ind.}}{\sim} N\big(\, & \beta_0 + u_i^{\text{R}} + \beta_1\,\texttt{smoke}_{ij} + \beta_2\,\texttt{mage}_{ij} + \beta_3\,\texttt{male}_{ij} + \beta_4\,\texttt{married}_{ij} \\
& + \beta_5\,\texttt{hsgrad}_{ij} + \beta_6\,\texttt{somecoll}_{ij} + \beta_7\,\texttt{collgrad}_{ij} + \beta_8\,\texttt{black}_{ij} \\
& + \beta_9\,\texttt{kessner2}_{ij} + \beta_{10}\,\texttt{kessner3}_{ij} + \beta_{11}\,\texttt{novisit}_{ij} \\
& + \beta_{12}\,\texttt{pretri2}_{ij} + \beta_{13}\,\texttt{pretri3}_{ij} + f\,(\texttt{gestat}_{ij}),\, \sigma_\varepsilon^2\,\big)\,, \\
\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma_\beta^2 \boldsymbol{I}_p), \quad u_i^{\text{R}}|\sigma_{\text{R}}^2 \overset{\text{ind.}}{\sim} N(0,\sigma_{\text{R}}^2), \quad \sigma_{\text{R}}^2 \sim \text{Half-Cauchy}\,(0,\, A_{\text{R}}), \\
\sigma_\varepsilon^2 \sim \text{Half-Cauchy}\,(0,\, A_\varepsilon), \quad u_k^{\text{G}}|\sigma_u^2 \overset{\text{ind.}}{\sim} N(0,\sigma_u^2), \quad \sigma_u^2 \sim \text{Half-Cauchy}\,(0, A_u) \qquad (3.12)
\end{aligned}$$

where

$$f(\texttt{gestat}) \;=\; \beta_{14}\,\texttt{gestat} + \sum_{k=1}^{q^{\text{G}}} u_k^{\text{G}}\, z_k(\texttt{gestat})$$

is a penalised spline function for gestational age. Here $z_1(\cdot), \ldots, z_{q^{\text{G}}}(\cdot)$ is a set of O'Sullivan spline basis functions and σ_u^2 represents the amount of penalisation of the spline coefficients $u_1^{\text{G}}, \ldots\, u_{q^{\text{G}}}^{\text{G}}$ as described in Section 1.3.

Figures 3.1 and 3.2 allow a visual assessment of the MFVB-based approximate posterior density functions against a MCMC benchmark for (3.12). MCMC was fitted using Stan with a burn-in of size 5000, a post burn-in of size 5000 with a thinning factor of 5. The

accuracy of approximate posterior density functions of $\boldsymbol{\beta}$ and $f(Q_k)$ is excellent, ranging from 95% to 98%, while it is about 75% to 82% for the variance parameters σ_u^2 and σ_ε^2 respectively. Inspection of Figure 3.1 shows that the MFVB fits and pointwise 95% credible sets are very close to those obtained using MCMC. The MCMC fits took 4.1 days, while the MFVB fits took 1.4 seconds. This represents more than a two hundred thousand-fold improvement in computing time.

3.6.2 Application to student assessment data

The second example is based on a dataset from the 2000 Program for International Student Assessment conducted by the Organisation for Economic Cooperation and Development. The survey assessed education attainment of 15 years old students in 43 counties, with an emphasis on reading proficiency. Our analysis is of the United States sample of the full dataset, provided by Rabe-Hesketh and Skrondal (2008). The data have a two-level structure with 2069 observations as units at Level 1 and 148 schools as groups at Level 2. The following variables are given in Table 3.3.

Variable	Description
`idschool`	school identifier
`passread`	indicator for being proficient in reading
`isei`	International Socioeconomic Index (SES)
`college`	indicator for highest education level by either parent being college
`oneforeign`	indicator for one parent being foreign born
`twoforeign`	indicator for both parents being foreign born
`language`	indicator for test language (English) being spoken at home

Table 3.3: Description of the 2000 Program for International Student Assessment conducted by the Organisation for Economic Cooperation and Development.

We treat reading proficiency as the response variable, an indicator taking the value 1 to indicate proficient and 0 otherwise. The effect of socio-economic status has considerable interest in education, we therefore model its effect the most flexible way via a penalised spline. The following Bayesian semiparametric random intercept logistic regression model for student i in school j was fitted via Algorithm 8:

$$\begin{aligned}\texttt{passread}_{ij}|\boldsymbol{\beta},\boldsymbol{u} &\overset{\text{ind.}}{\sim} \text{Bernoulli}\left(\text{logit}^{-1}\left\{\beta_0 + u_i^{\text{R}} + \beta_1\,\texttt{eduction}_{ij} + \beta_2\,\texttt{oneforeign}_{ij}\right.\right.\\ &\qquad\left.\left. + \beta_3\,\texttt{bothforeign}_{ij} + \beta_4\,\texttt{language}_{ij} + f(\texttt{isei}_{ij})\right\}\right),\\ \boldsymbol{\beta} &\sim N(\mathbf{0},\sigma_\beta^2\boldsymbol{I}_p),\quad u_i^{\text{R}}|\sigma_{\text{R}}^2 \overset{\text{ind.}}{\sim} N(0,\sigma_{\text{R}}^2),\quad \sigma_{\text{R}}^2 \sim \text{Half-Cauchy}\,(0,\,A_{\text{R}}),\\ u_k^{\text{G}}|\sigma_u^2 &\overset{\text{ind.}}{\sim} N(0,\sigma_u^2)\quad\text{and}\quad \sigma_u^2 \sim \text{Half-Cauchy}\,(0,A_u).\end{aligned} \tag{3.13}$$

The function f is a penalised spline for the International Socio-economic Index, defined

analogously in (3.12). Figures 3.1 and 3.3 allow a visual assessment of the MFVB-based approximate posterior density functions against a MCMC benchmark for (3.13). MCMC was fitted using Stan with a burn-in of size 5000, a post burn-in of size 5000 with a thinning factor of 5. The accuracy of the approximate posterior density functions of $\boldsymbol{\beta}$ and $f(Q_k)$ is good, ranging from 87% to 94%. In contrast, the MFVB approximation is poor for the variance parameter σ^2_{R}, as we previously observed in the Bernoulli simulation study results. Nonetheless, inspection of Figure 3.1 shows that the MFVB fits and pointwise 95% credible sets are close to those obtained using MCMC. The MCMC fits took 57 minutes, while the MFVB fits took 1.2 minutes.

3.6.3 Application to German health care data

The final example is based on the German Socio-Economic Panel dataset, an ongoing annual household survey that was started in 1984. The data are presented in Winkelmann (2004), and a subset of the original data is provided by Rabe-Hesketh and Skrondal (2008) and our analysis is of this subset. The data have a two-level structure with 32 837 observations as units at Level 1 and 9197 patients as groups at Level 2. Each patient has one to five observations. The data have a few variables relating to the usage of health service, and of particular interest is the number of visits to a doctor during the previous three months. The following variables are given in Table 3.4.

Variable	Description
`id`	patient identifier
`numvisit`	number of visits to a doctor during the 3 months before the interview
`male`	indicator for being a male
`married`	indicator for being married
`sport`	indicator for being actively in sports
`goodhealth`	indicator for being in good health
`badhealth`	indicator for being in bad health
`welfare`	indicator for one receiving welfare payments
`fulltime`	indicator for being employed full-time
`unemploy`	indicator for being unemployed
`winter`	indicator for being interviewed in winter quarter
`spring`	indicator for being interviewed in spring quarter
`fall`	indicator for being interviewed in fall quarter

Table 3.4: Description of the German health care data as presented in Winkelmann (2004).

We treat the number of doctor visits as the response variable, and of particular interest is the age effects which we modelled flexibly via a penalised spline. The following Bayesian semiparametric random intercept Poisson regression model for patient i in observation j was fitted via Algorithm 9:

$$\begin{aligned}
\texttt{numvisit}_{ij}|\boldsymbol{\beta},\boldsymbol{u} &\overset{\text{ind.}}{\sim} \text{Poisson}\left(\exp\left\{\beta_0 + u_i^{\text{R}} + \beta_1\,\texttt{male}_{ij} + \beta_2\,\texttt{married}_{ij} + \beta_3\,\texttt{sport}_{ij}\right.\right.\\
&\quad + \beta_4\,\texttt{goodhealth}_{ij} + \beta_5\,\texttt{badhealth}_{ij} + \beta_6\,\texttt{welfare}_{ij}\\
&\quad + \beta_7\,\texttt{fulltime}_{ij} + \beta_8\,\texttt{unemploy}_{ij} + \beta_9\,\texttt{winter}_{ij}\\
&\quad \left.\left. + \beta_{10}\,\texttt{spring}_{ij} + \beta_{11}\,\texttt{fall}_{ij} + f(\texttt{isei}_{ij})\right\}\right),\\
\boldsymbol{\beta} \sim N(\mathbf{0},\sigma_\beta^2\boldsymbol{I}_p),&\quad u_i^{\text{R}}|\sigma_{\text{R}}^2 \overset{\text{ind.}}{\sim} N(0,\sigma_{\text{R}}^2),\quad \sigma_{\text{R}}^2 \sim \text{Half-Cauchy}\,(0,\,A_{\text{R}}),\\
u_k^{\text{G}}|\sigma_u^2 \overset{\text{ind.}}{\sim} N(0,\sigma_u^2)&\quad \text{and}\quad \sigma_u^2 \sim \text{Half-Cauchy}\,(0,A_u).
\end{aligned} \tag{3.14}$$

The function f is a penalised spline for age, defined analogously in (3.12). Figures 3.1 and 3.4 allow a visual assessment of the MFVB-based approximate posterior density functions against a MCMC benchmark for (3.13). MCMC was fitted using Stan with a burn-in of size 5000, a post burn-in of size 5000 with a thinning factor of 5. The accuracy of the approximate posterior density functions for all parameters is excellent, ranging from 80% to 97%. Inspection of Figure 3.4 shows that both the MFVB and MCMC spline fits are virtually identical. The MCMC fits took 1.6 hours while the MFVB fits took 48 seconds.

3.7 Extension to three-level semiparametric mixed models

A great advantage of MFVB algorithms is *modularity*. By this we means that concepts like main effects, interaction effects, higher-order random effects and spline regression can be viewed as modules that can be put together into an almost endless variety of statistical models. All that is required is relatively straightforward modifications on the structures of the general design matrices and effects covariance matrix in order to accommodate larger and more complicated mixed models. Consider a general three-level mixed models with the Gaussian response, and with simplicity in mind, we confine the model to the case of a single predictor, say x,

$$\begin{aligned}
y_{ijk} &= \beta_0 + \beta_x\,x_{ijk} + \{(u_{0ij}^{\text{RL2}} + u_{0i}^{\text{RL3}}) + (u_{1ij}^{\text{RL2}} + u_{1i}^{\text{RL3}})\,x_{ijk}\} + \varepsilon_{ijk},\\
\begin{bmatrix} u_{0ij}^{\text{RL2}} & u_{1ij}^{\text{RL2}}\end{bmatrix}^{\mathsf{T}} &\overset{\text{ind.}}{\sim} N\left(\mathbf{0},\boldsymbol{\Sigma}^{\text{RL2}} = \begin{bmatrix}(\sigma_{u_0}^{\text{RL2}})^2 & \sigma_{u_0,u_1}^{\text{RL2}}\\ \sigma_{u_0,u_1}^{\text{RL2}} & (\sigma_{u_1}^{\text{RL2}})^2\end{bmatrix}\right),\\
\begin{bmatrix} u_{0i}^{\text{RL3}} & u_{1i}^{\text{RL3}}\end{bmatrix}^{\mathsf{T}} &\overset{\text{ind.}}{\sim} N\left(\mathbf{0},\boldsymbol{\Sigma}^{\text{RL3}} = \begin{bmatrix}(\sigma_{u_0}^{\text{RL3}})^2 & \sigma_{u_0,u_1}^{\text{RL3}}\\ \sigma_{u_0,u_1}^{\text{RL3}} & (\sigma_{u_1}^{\text{RL3}})^2\end{bmatrix}\right),\\
\text{and}\quad \varepsilon_{ijk} &\overset{\text{ind.}}{\sim} N\left(0,\sigma_\varepsilon^2\right),\quad 1\le i\le m,\,1\le j\le n_i,\,1\le k\le o_{ij},
\end{aligned} \tag{3.15}$$

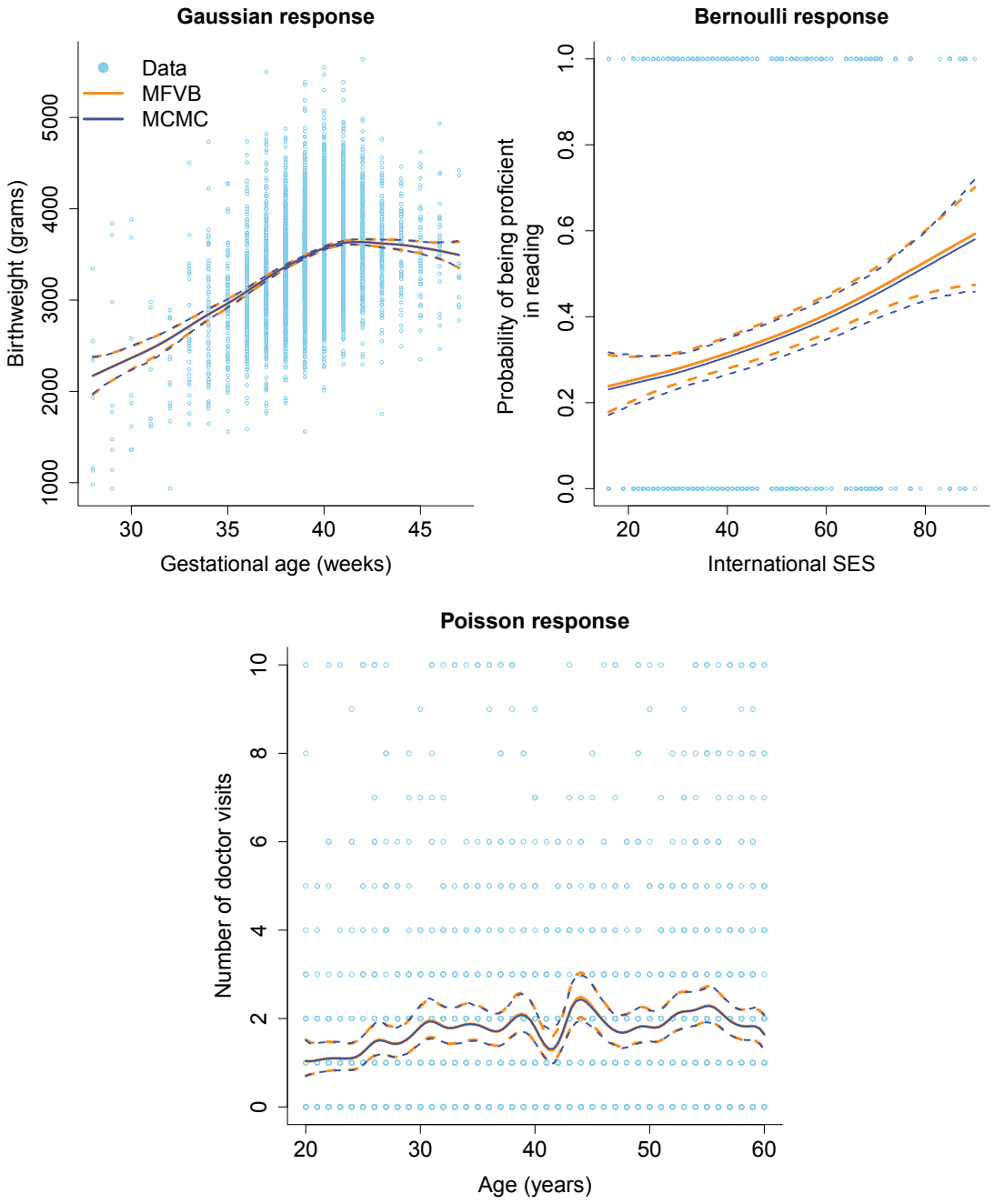

Figure 3.1: Approximate posterior means of the penalised regression functions and corresponding pointwise 95% credible sets obtained via MFVB approximation (orange curves) and MCMC (blue curves) for the two-level Bayesian semiparametric mixed models to the real datasets. The sky blue circles represent the real data.

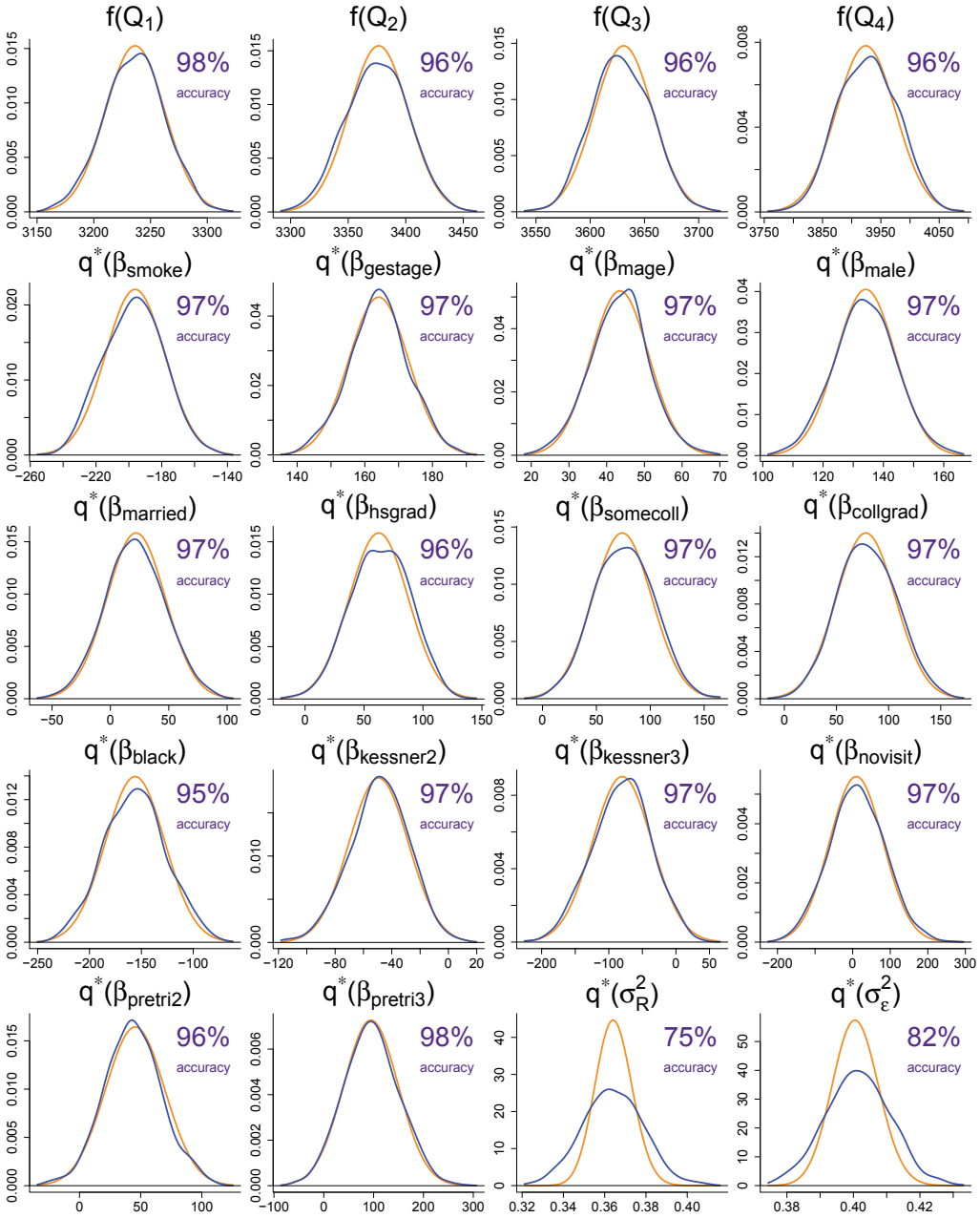

Figure 3.2: Approximate posterior density functions for the model parameters obtained via MFVB approximation (orange curves) and MCMC (blue curves) for the Gaussian response model to the smoking data. The accuracy score on the top right of each plot represents the accuracy of MFVB approximation compared against the MCMC benchmark.

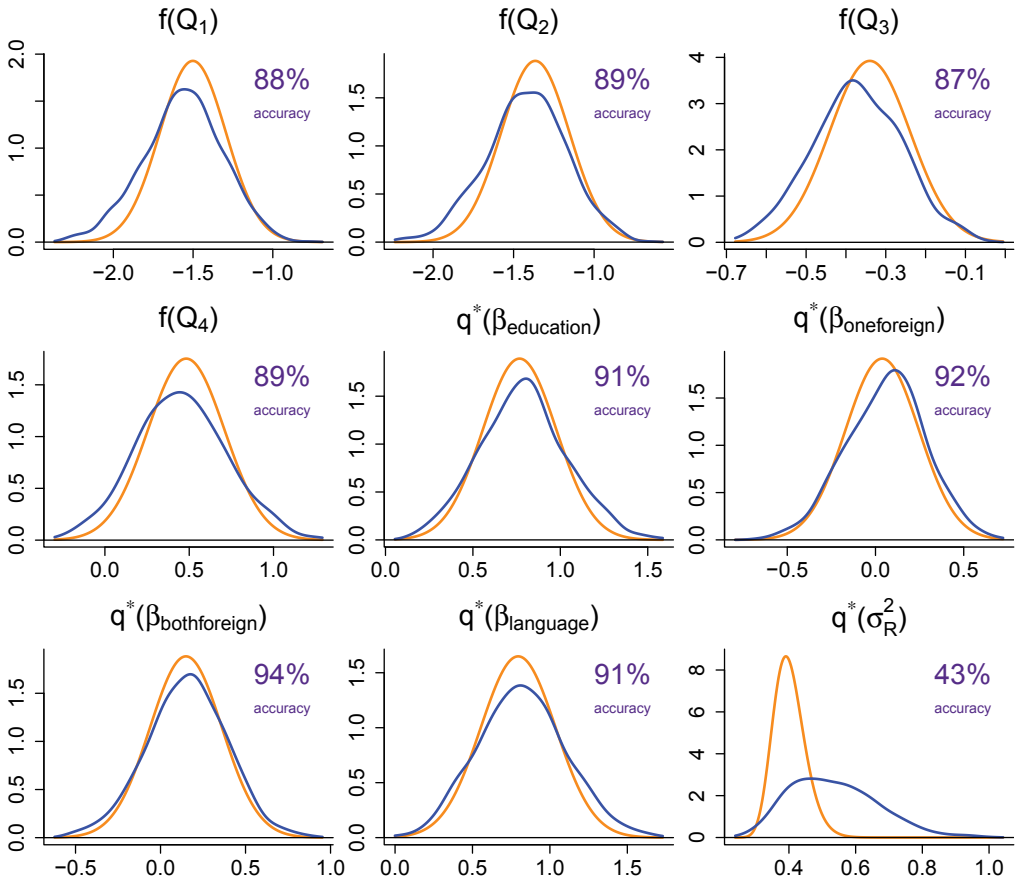

Figure 3.3: Approximate posterior density functions for the model parameters obtained via MFVB approximation (orange curves) and MCMC (blue curves) for the Bernoulli response model to the Program of International Student Assessment data. The accuracy score on the top right of each plot represents the accuracy of MFVB approximation compared against the MCMC benchmark.

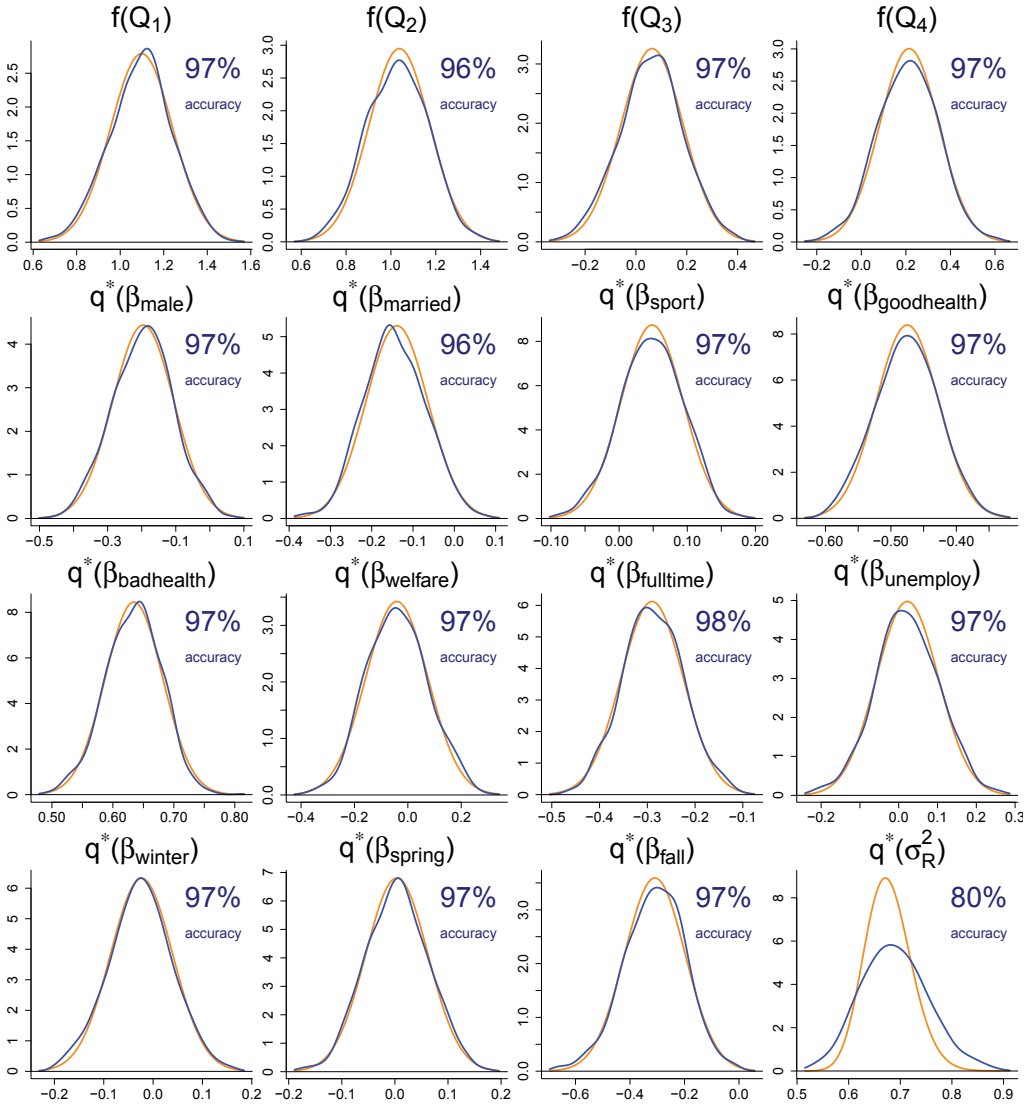

Figure 3.4: Approximate posterior density functions for the model parameters obtained via MFVB approximation (orange curves) and MCMC (blue curves) for the Poisson response model to the German health care data. The accuracy score on the top right of each plot represents the accuracy of MFVB approximation compared against the MCMC benchmark.

where i, j, k denote the *nested* indexing for the group level (Level 3, superscript RL3), subgroup level (Level 2, superscript RL2) and unit level (Level 1) respectively. The superscript notation is deliberately chosen to be in accordance with Section 3.3. The term y_{ijk} and x_{ijk} denote the response and predictor variables for the kth individual nested within the jth subgroup and the ith group, β_0 and β_x are the so-called fixed effects, u_{0ij}^{RL2} and u_{0ij}^{RL2} are the so-called random intercepts and random slopes at the subgroup level, and similarly, u_{0i}^{RL3} and u_{1i}^{RL3} are the random intercepts and random slopes at the group level. Here we use the random effects to model the dependence of ys within groups and subgroups. It is reasonable to assume that the level parameters are independent of each other and each of them is independent of the error terms ε_{ijk}. To avoid confusion, we summarise the dimension variables as follows:

$$
\begin{aligned}
p &= \text{number of fixed effects,} \\
q^{\mathrm{RL3}} &= \text{number of random effects at the group level,} \\
q^{\mathrm{RL2}} &= \text{number of random effects at the subgroup level,} \\
q^{\mathrm{G}} &= \text{number of spline basis functions,} \\
m &= \text{total number of groups} \\
n_i &= \text{number of subgroups in the } i\text{th group,} \\
o_{ij} &= \text{number of observations in the } j\text{th subgroup within the } i\text{th group,} \\
N &= \text{total number of subgroups, i.e. } \textstyle\sum_{i=1}^{m} n_i, \\
O &= \text{total number of observations, i.e. } \textstyle\sum_{i=1}^{m} \sum_{j=1}^{n_i} o_{ij}.
\end{aligned}
$$

Much like in the two-level case, model (3.7) can be easily extended to other response distributions and including semiparametric extensions and hence we omit the details here.

3.7.1 Mixed model representation

In keeping with the notation used in Section 3.3, we express model (3.7) in matrix notation as

$$
\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} = \boldsymbol{X}^{\mathrm{R}}\boldsymbol{\beta}^{\mathrm{R}} + \boldsymbol{X}^{\mathrm{G}}\boldsymbol{\beta}^{\mathrm{G}} + \boldsymbol{Z}^{\mathrm{R}}\boldsymbol{u}^{\mathrm{R}} + \boldsymbol{Z}^{\mathrm{G}}\boldsymbol{u}^{\mathrm{G}} \tag{3.16}
$$

by partitioning the following vectors and matrices:

$$
\boldsymbol{X}^{\mathrm{R}} \equiv \begin{bmatrix} \boldsymbol{X}_1^{\mathrm{R}} \\ \vdots \\ \boldsymbol{X}_m^{\mathrm{R}} \end{bmatrix}, \quad \boldsymbol{X}^{\mathrm{G}} \equiv \begin{bmatrix} \boldsymbol{X}_1^{\mathrm{G}} \\ \vdots \\ \boldsymbol{X}_m^{\mathrm{G}} \end{bmatrix}, \quad \boldsymbol{Z}^{\mathrm{R}} \equiv \underset{1\le i\le m}{\mathrm{blockdiag}} \left(\boldsymbol{X}_i^{\mathrm{R}} \,\middle|\, \underset{1\le j\le n_i}{\mathrm{blockdiag}} \left(\boldsymbol{X}_{ij}^{\mathrm{R}} \right) \right).
$$

$$\boldsymbol{u}^{\mathrm{R}} \equiv \begin{bmatrix} \boldsymbol{u}_1^{\mathrm{R}} \\ \vdots \\ \boldsymbol{u}_m^{\mathrm{R}} \end{bmatrix}, \quad \boldsymbol{u}^{\mathrm{G}} \equiv \begin{bmatrix} \boldsymbol{u}_1^{\mathrm{G}} \\ \vdots \\ \boldsymbol{u}_L^{\mathrm{G}} \end{bmatrix}, \quad \boldsymbol{Z}^{\mathrm{G}} \equiv \begin{bmatrix} z_{\ell 1}(\boldsymbol{X}_1^{\mathrm{G}}) & \cdots & z_{\ell q_\ell^{\mathrm{G}}}(\boldsymbol{X}_1^{\mathrm{G}}) \\ \vdots & \ddots & \vdots \\ z_{\ell 1}(\boldsymbol{X}_m^{\mathrm{G}}) & \cdots & z_{\ell q_\ell^{\mathrm{G}}}(\boldsymbol{X}_m^{\mathrm{G}}) \end{bmatrix},$$

where

$$\boldsymbol{X}_i^{\mathrm{R}} \equiv \begin{bmatrix} \boldsymbol{X}_{i1}^{\mathrm{R}} \\ \vdots \\ \boldsymbol{X}_{in_m}^{\mathrm{R}} \end{bmatrix} \quad \text{given} \quad \boldsymbol{X}_{ij}^{\mathrm{R}} \equiv \begin{bmatrix} 1 & x_{ij1} \\ \vdots & \vdots \\ 1 & x_{ijo_{mn_m}} \end{bmatrix}, \quad \boldsymbol{X}_i^{\mathrm{G}} \equiv \begin{bmatrix} \boldsymbol{X}_{i1}^{\mathrm{C}} \\ \vdots \\ \boldsymbol{X}_{in_m}^{\mathrm{C}} \end{bmatrix}$$

$$\boldsymbol{u}_i^{\mathrm{R}} \equiv \begin{bmatrix} \boldsymbol{u}_i^{\mathrm{RL3}} \\ \boldsymbol{u}_{i1}^{\mathrm{RL2}} \\ \vdots \\ \boldsymbol{u}_{in_i}^{\mathrm{RL2}} \end{bmatrix} \quad \text{given} \quad \boldsymbol{u}_i^{\mathrm{RL3}} \equiv \begin{bmatrix} u_{0i}^{\mathrm{RL3}} \\ u_{1i}^{\mathrm{RL3}} \end{bmatrix} \quad \text{and} \quad \boldsymbol{u}_{ij}^{\mathrm{RL2}} \equiv \begin{bmatrix} u_{0ij}^{\mathrm{RL2}} \\ u_{1ij}^{\mathrm{RL2}} \end{bmatrix}, \quad \boldsymbol{u}_\ell^{\mathrm{G}} \equiv \begin{bmatrix} u_\ell^{\mathrm{G}} \\ \vdots \\ u_{q_\ell^{\mathrm{G}}}^{\mathrm{G}} \end{bmatrix}.$$

The response variable $\boldsymbol{y}$ is an $(\sum_{i=1}^m \sum_{j=1}^{n_i} o_{ij}) \times 1$ vector and $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects. The matrices $\boldsymbol{X}^{\mathrm{R}}$ and $\boldsymbol{Z}^{\mathrm{R}}$ are the $(\sum_{i=1}^m \sum_{j=1}^{n_i} o_{ij}) \times q^{\mathrm{RL3}}$ and $(\sum_{i=1}^m \sum_{j=1}^{n_i} o_{ij}) \times \{m\,(q^{\mathrm{RL2}} + q^{\mathrm{RL3}} \sum_{i=1}^m n_i)\}$ general design matrices corresponding to the fixed effects vector $\boldsymbol{\beta}^{\mathrm{R}}$ and random group effects $\boldsymbol{u}^{\mathrm{R}}$ respectively. The matrices $\boldsymbol{X}^{\mathrm{G}}$ and $\boldsymbol{Z}^{\mathrm{G}}$ are the $(\sum_{i=1}^m \sum_{j=1}^{n_i} o_{ij}) \times 1$ and $(\sum_{i=1}^m \sum_{j=1}^{n_i} o_{ij}) \times \sum_{\ell=1}^L q_\ell^{\mathrm{G}}$ general design matrices corresponding to the fixed effects vector $\boldsymbol{\beta}^{\mathrm{G}}$ and spline coefficients $\boldsymbol{u}^{\mathrm{G}}$ respectively. The matrices $\boldsymbol{\Sigma}^{\mathrm{RL2}}$ and $\boldsymbol{\Sigma}^{\mathrm{RL3}}$ are the $q^{\mathrm{RL2}} \times q^{\mathrm{RL2}}$ and $q^{\mathrm{RL3}} \times q^{\mathrm{RL3}}$ *unstructured* (no pattern is specified) covariance matrices. The random effects covariance matrix of $(\boldsymbol{u}^{\mathrm{R}}, \boldsymbol{u}^{\mathrm{G}})$ is

$$\mathrm{Cov}\left(\begin{bmatrix} \boldsymbol{u}^{\mathrm{R}} \\ \boldsymbol{u}^{\mathrm{G}} \end{bmatrix}\right) = \begin{bmatrix} \underset{1 \le i \le m}{\mathrm{blockdiag}} \begin{bmatrix} \boldsymbol{\Sigma}^{\mathrm{RL3}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{n_i} \otimes \boldsymbol{\Sigma}^{\mathrm{RL2}} \end{bmatrix} & \boldsymbol{0} \\ \boldsymbol{0} & \underset{1 \le \ell \le L}{\mathrm{blockdiag}} \left(\sigma_{u\ell}^2 \boldsymbol{I}_{q_\ell^{\mathrm{G}}}\right) \end{bmatrix}.$$

Here the $\boldsymbol{u}_\ell^{\mathrm{G}}$s are $q_\ell^{\mathrm{G}} \times 1$ vectors for $1 \le \ell \le L$, so the blocks of $\mathrm{Cov}(\boldsymbol{u}^{\mathrm{G}})$ corresponding to the decomposition of $\boldsymbol{u}^{\mathrm{G}}$.

We are now ready to present the full three-level Bayesian semiparametric mixed models. Figure 3.5 is the directed acyclic graph corresponding to model (3.17). We use the same prior specification for the model parameters as described in Section 2.4. Consider the response vector $\boldsymbol{y}$ with either the Gaussian, Student-t, Bernoulli or Poisson distribution as follows:

$$\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2 \sim \mathrm{Dist}\,(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \boldsymbol{R}), \quad \boldsymbol{\beta} \sim N(0, \sigma_\beta^2 \boldsymbol{I}_p), \tag{3.17}$$

$$
\begin{aligned}
\boldsymbol{u}|\sigma^2_{u\ell}, \boldsymbol{\Sigma}^{\text{RL2}}, \boldsymbol{\Sigma}^{\text{RL3}} &\sim N\left(\mathbf{0}, \begin{bmatrix} \underset{1\le\ell\le L}{\text{blockdiag}}\left(\sigma^2_{u\ell}\boldsymbol{I}_{q^{\text{G}}_\ell}\right) & \mathbf{0} \\ \mathbf{0} & \underset{1\le i\le m}{\text{blockdiag}}\begin{bmatrix} \boldsymbol{\Sigma}^{\text{RL3}} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{n_i}\otimes\boldsymbol{\Sigma}^{\text{RL2}} \end{bmatrix} \end{bmatrix}\right), \\
\boldsymbol{\Sigma}^{\text{RL2}}|a^{\text{RL2}}_1,\ldots,a^{\text{RL2}}_{q^{\text{RL2}}} &\sim \text{Inverse-Wishart}\left(\nu+q^{\text{RL2}}-1,\ 2\,\nu\,\text{diag}(1/a^{\text{RL2}}_1,\ldots,1/a^{\text{RL2}}_{q^{\text{RL2}}})\right), \\
\boldsymbol{\Sigma}^{\text{RL3}}|a^{\text{RL3}}_1,\ldots,a^{\text{RL3}}_{q^{\text{RL3}}} &\sim \text{Inverse-Wishart}\left(\nu+q^{\text{RL3}}-1,\ 2\,\nu\,\text{diag}(1/a^{\text{RL3}}_1,\ldots,1/a^{\text{RL3}}_{q^{\text{RL3}}})\right), \\
a^{\text{RL2}}_{r_{\text{L2}}} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(\tfrac{1}{2}, 1/A^2_{\text{RL2}}\right), \quad r_{\text{L2}}=1,\ldots,q^{\text{RL2}}, \\
a^{\text{RL3}}_{r_{\text{L3}}} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(\tfrac{1}{2}, 1/A^2_{\text{RL3}}\right), \quad r_{\text{L3}}=1,\ldots,q^{\text{RL3}}, \\
\sigma^2_{u\ell}|a_{u\ell} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2},\ 1/a_{u\ell}), \quad a_{u\ell} \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2},\ 1/A^2_{u_\ell}),
\end{aligned}
$$

where $a^{\text{RL2}}_{r_{\text{L2}}}\,(1\le r_{\text{L2}}\le q^{\text{RL2}})$, $a^{\text{RL3}}_{r_{\text{L3}}}\,(1\le r_{\text{L3}}\le q^{\text{RL3}})$ and $a_{u\ell}\,(1\le\ell\le L)$ are the auxiliary variables used for the scale parameters in the model. The notation "Dist" denotes the type of response distribution and $\boldsymbol{R}$ represents the additional parameters corresponding to the errors, e.g. $\boldsymbol{R}=\sigma^2_\varepsilon\boldsymbol{I}$ for the Gaussian response mixed model. Fitting (3.17) involves setting of the hyperparameters σ^2_β, A_{RL2}, A_{RL3} and $A_{u\ell}$ by the analyst. To ensure the property of Huang and Wand (2013) holds, we set $\nu=2$.

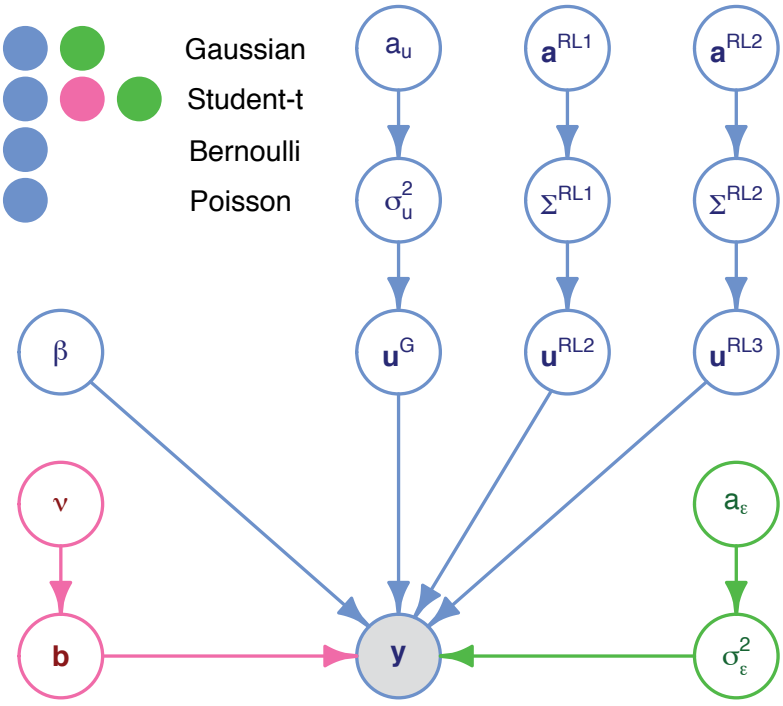

Figure 3.5: Directed acyclic graphs for model (3.17). The shaded node corresponds to the observed data vector. The colour keys at the top of the figure indicate the components of the graph corresponding to each response outcome.

3.7.2 Naïve mean field variational Bayes algorithms

From an extendability standpoint, a general design framework for mixed models is a particularly attractive approach that allows MFVB algorithms to cater for larger and more complicated models with straightforward algebraic adjustments. The structural changes in the random effects matrices $\boldsymbol{Z}$ and $\boldsymbol{G}$ underpin the complexity of a mixed model being transitioned from a two-level to a three-level hierarchical structure. For example, Figure 3.6 shows that the $\boldsymbol{Z}$ matrix starts off as having a simple block-diagonal structure and, as the level of hierarchy increases, it grows into a "nested" block-diagonal structure (each large block is itself block-diagonal, with one or more small blocks on the diagonal). Further, even with the additional random group effects parameters, the covariance matrix shown in (3.17) remains to be a block-diagonal, but certainly larger, matrix with different entries for the variance components corresponding to the random basis coefficients for the overall mean and random group effects. Taking these into consideration, it then follows that the MFVB algorithms for model (3.17) simply involves modifications of Algorithms 7, 8 and 9 to incorporate the aforementioned structural changes in the $\boldsymbol{Z}$ and $\boldsymbol{G}$ matrices, as well as updates for the additional model parameters as follows:

$q^*(\boldsymbol{\Sigma}^{\text{RL2}})$ is the Inverse-Wishart $\left(\nu + \sum_{i=1}^m n_i + q^{\text{RL2}} + 1,\, \boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\text{RL2}})}\right)$ density function,
$q^*(a_{r_{\text{L2}}}^{\text{RL2}})$ is the Inverse-Gamma $\left(\frac{1}{2}(\nu + q^{\text{RL2}}),\, B_{q(a_{r_{\text{L2}}}^{\text{RL2}})}\right)$ density function,
$q^*(\boldsymbol{\Sigma}^{\text{RL3}})$ is the Inverse-Wishart $\left(\nu + m + q^{\text{RL3}} + 1,\, \boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\text{RL3}})}\right)$ density function and
$q^*(a_{r_{\text{L3}}}^{\text{RL3}})$ is the Inverse-Gamma $\left(\frac{1}{2}(\nu + q^{\text{RL3}}),\, B_{q(a_{r_{\text{L3}}}^{\text{RL3}})}\right)$ density function.

These naïve mean field variational Bayes algorithms are summarised in Algorithm 10.

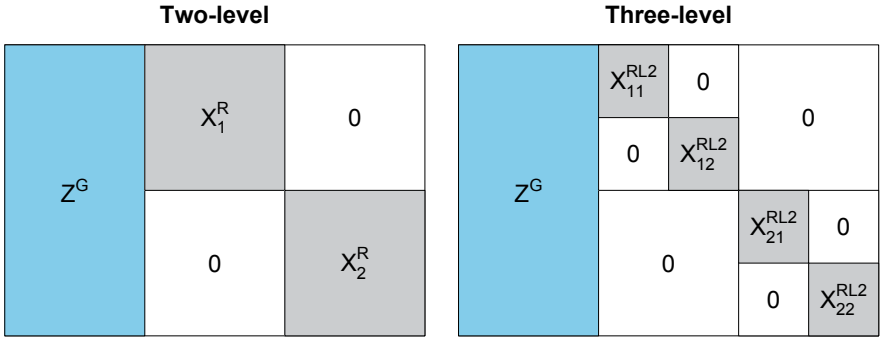

Figure 3.6: Different $\boldsymbol{Z}$ structures for the two-level and three-level models with $m = 2$. The $\boldsymbol{Z}$ matrix starts off as having a simple block-diagonal structure and, as the level of hierarchy increases, it grows into a "nested" block-diagonal structure. The definitions of matrices are described in the text.

3.7.3 Streamlined mean field variational Bayes algorithms

As pointed out in Section 3.4, the naïve implementation of MFVB algorithms for arbitrarily large grouped data is extremely inefficient in terms of speed and storage. The time complexity can be as high as $O(m^3)$, making variational inference impractical for large and complex mixed models. Through exploiting the inherent block-diagonal structure of the effects covariance matrix, we develop fast and memory-efficient MFVB algorithms that streamline inversion and update for $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$. Here we extend the streamlined approach to the three-level model settings. The essential trick for this extension is to permute the $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ matrix into an approximate block-diagonal form for decomposition, as shown in Section 3.4. Then every previously described step can be analogously applied to the three-level mixed model by simply modifying the submatrices $\boldsymbol{F}$, $\boldsymbol{G}_i$ and $\boldsymbol{H}_i$ that are defined in Section 3.4. Take the Gaussian response model as an example,

$$\begin{aligned}
\boldsymbol{F} &\equiv \mu_{q(1/\sigma_\varepsilon^2)}\,(\boldsymbol{C}^{\mathrm{G}})^{\mathsf{T}}\boldsymbol{C}^{\mathrm{G}} + \begin{bmatrix} \sigma_\beta^{-2}\boldsymbol{I}_p & \boldsymbol{0} \\ \boldsymbol{0} & \underset{1\le\ell\le L}{\mathrm{blockdiag}}\left(\mu_{q(1/\sigma_{u\ell}^2)}\boldsymbol{I}_{q_\ell^{\mathrm{G}}}\right) \end{bmatrix}, \\
\boldsymbol{G}_i &\equiv \mu_{q(1/\sigma_\varepsilon^2)}\,(\boldsymbol{C}_i^{\mathrm{G}})^{\mathsf{T}}\boldsymbol{Z}_i^{\mathrm{R}}, \\
\text{and}\quad \boldsymbol{H}_i &\equiv \left\{\mu_{q(1/\sigma_\varepsilon^2)}\,(\boldsymbol{Z}_i^{\mathrm{R}})^{\mathsf{T}}\boldsymbol{Z}_i^{\mathrm{R}} + \begin{bmatrix} \boldsymbol{\Sigma}^{\mathrm{RL3}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{n_i}\otimes\boldsymbol{\Sigma}^{\mathrm{RL2}} \end{bmatrix}\right\}^{-1}.
\end{aligned}$$

Despite the significant amount of time we have already saved, unlike the two-level case, the resultant block diagonal structure in $\boldsymbol{H}_i$ can still be considered large when n_i is large and comprises unnecessary zeros. This hinders the optimal time savings that would be possible and further research into this area is undergoing.

3.8 Concluding remarks

Mean field variational Bayes with streamlining, as exemplified by Algorithms 7, 8 and 9, is a useful addition to the longitudinal and multilevel data analysis arsenal. It allows rapid and accurate approximate Bayesian inference for very large datasets with computational times that increase only linearly in the number of groups. Extension of streamlined MFVB to higher level longitudinal and multilevel models is also of interest, although this requires a considerable amount of analytic computation. This chapter has helped opened up a new branch of methodology for fitting and inference for grouped data in the high volume/velocity era.

Initialise: $\nu = 2$, $\mu_{q(1/\sigma_\varepsilon^2)} > 0$, $\mu_{q(1/a_\varepsilon)} > 0$, $\boldsymbol{M}_{q\left((\boldsymbol{\Sigma}_u^{\mathrm{RL2}})^{-1}\right)}$, a $q^{\mathrm{RL2}} \times q^{\mathrm{RL2}}$ positive definite matrix, $\boldsymbol{M}_{q\left((\boldsymbol{\Sigma}_u^{\mathrm{RL3}})^{-1}\right)}$, a $q^{\mathrm{RL3}} \times q^{\mathrm{RL3}}$ positive definite matrix, $\mu_{q(1/a_{r_{\mathrm{L2}}}^{\mathrm{RL2}})} > 0$, $1 \leq r_{\mathrm{L2}} \leq q^{\mathrm{RL2}}$, $\mu_{q(1/a_{r_{\mathrm{L3}}}^{\mathrm{RL3}})} > 0$, $1 \leq r_{\mathrm{L3}} \leq q^{\mathrm{RL3}}$, $\mu_{q(1/\sigma_{u\ell}^2)} > 0$, $1 \leq \ell \leq L$, $\boldsymbol{\xi}$, a $(\sum_{i=1}^m \sum_{j=1}^{n_i} o_{ij}) \times 1$ vector of positive entries and $\mu_{q(1/b_i)} > 0$, $1 \leq i \leq O$, $\mu_{q(\nu)} > 0$, $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$, a $(p + m\, q^{\mathrm{RL3}} + \sum_{i=1}^m n_i\, q^{\mathrm{RL2}} + \sum_{l=1}^L q_\ell^{\mathrm{G}}) \times 1$ vector and $\boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})}$, a $(\sum_{i=1}^m \sum_{j=1}^{n_i} o_{ij}) \times 1$ vector of positive entries.

Cycle through updates:

Define: $\boldsymbol{M}_{q(\boldsymbol{\Omega})} = \begin{bmatrix} \sigma_\beta^{-2}\boldsymbol{I}_p & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \underset{1\leq\ell\leq L}{\mathrm{blockdiag}}\left(\mu_{q(1/\sigma_{u\ell}^2)}\boldsymbol{I}_{q_\ell^{\mathrm{G}}}\right) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \underset{1\leq i\leq m}{\mathrm{blockdiag}}\begin{bmatrix} \boldsymbol{M}_{q\left((\boldsymbol{\Sigma}_u^{\mathrm{RL2}})^{-1}\right)} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{n_i} \otimes \boldsymbol{M}_{q\left((\boldsymbol{\Sigma}_u^{\mathrm{RL3}})^{-1}\right)} \end{bmatrix} \end{bmatrix}$

If fitting the Gaussian model:

- Update multivariate normal $q^*(\boldsymbol{\beta},\boldsymbol{u})$ parameters:
  $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \left(\mu_{q(1/\sigma_\varepsilon^2)}\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C} + \boldsymbol{M}_{q(\boldsymbol{\Omega})}\right)^{-1}$
  $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \mu_{q(1/\sigma_\varepsilon^2)}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^{\mathsf{T}}\boldsymbol{y}$
- Update inverse-Gamma $q^*(\sigma_\varepsilon^2)$ and inverse-Gamma $q^*(a_\varepsilon)$ parameters:
  $B_{q(\sigma_\varepsilon^2)} \leftarrow \mu_{q(1/a_\varepsilon)} + \frac{1}{2}\left\{||\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}||^2 + \mathrm{tr}(\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})\right\}$
  $\mu_{q(1/\sigma_\varepsilon^2)} \leftarrow \frac{1}{2}(\sum_{i=1}^m \sum_{j=1}^{n_i} o_{ij} + 1)/B_{q(\sigma_\varepsilon^2)} \quad ; \quad \mu_{q(1/a_\varepsilon)} \leftarrow 1/\left(\mu_{q(1/\sigma_\varepsilon^2)} + A_\varepsilon^{-2}\right)$

If fitting the Student-*t* model:

$O \leftarrow \sum_{i=1}^m \sum_{j=1}^{n_i} o_{ij} \quad ; \quad \boldsymbol{C}_{12} \leftarrow \underset{1\leq i\leq O}{\mathrm{diag}}\left(\mu_{q(1/b_i)}\right)$

- Update multivariate normal $q^*(\boldsymbol{\beta},\boldsymbol{u})$ parameters:
  $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \left(\mu_{q(1/\sigma_\varepsilon^2)}\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C}_{12}\boldsymbol{C} + \boldsymbol{M}_{q(\boldsymbol{\Omega})}\right)^{-1}$
  $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \mu_{q(1/\sigma_\varepsilon^2)}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C}_{12}\boldsymbol{y}$
- Update inverse-Gamma $q^*(\sigma_\varepsilon^2)$ and inverse-Gamma $q^*(a_\varepsilon)$ parameters:
  $B_{q(\sigma_\varepsilon^2)} \leftarrow \mu_{q(1/a_\varepsilon)} + \frac{1}{2}\,\mu_{q(1/\sigma_\varepsilon^2)}\left\{(\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})})^{\mathsf{T}}\boldsymbol{C}_{12}(\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}) + \mathrm{tr}(\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C}_{12}\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})\right\}$
  $\mu_{q(1/\sigma_\varepsilon^2)} \leftarrow \frac{1}{2}(\sum_{i=1}^m \sum_{j=1}^{n_i} o_{ij} + 1)/B_{q(\sigma_\varepsilon^2)} \quad ; \quad \mu_{q(1/a_\varepsilon)} \leftarrow 1/\left(\mu_{q(1/\sigma_\varepsilon^2)} + A_\varepsilon^{-2}\right)$
- Update inverse-Gamma $q^*(b_i)$ parameters:
  For $1 \leq i \leq O$:
  $B_{q(b_i)} \leftarrow \frac{1}{2}\left[\mu_{q(\nu)} + \mu_{q(1/\sigma_\varepsilon^2)}\left\{(\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})})_i^2 + (\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^{\mathsf{T}})_{ii}\right\}\right]$
  $\mu_{q(1/b_i)} \leftarrow \frac{1}{2}(\mu_{q(\nu)} + 1)/B_{q(b_i)}$
  $\mu_{q(\log b_i)} \leftarrow \log(B_{q(b_i)}) - \mathrm{digamma}\left\{\frac{1}{2}(\mu_{q(\nu)} + 1)\right\}$

- Update $q^*(\nu)$ parameters:
  $C_1 \leftarrow \sum_{i=1}^{O} \left(\mu_{q(\log(b_i))} + \mu_{q(1/b_i)}\right)$
  $\mu_{q(\nu)} \leftarrow \exp\Big\{\log\mathcal{F}(1, \sum_{i=1}^{m}\sum_{j=1}^{n_i} o_{ij}, C_1, \nu_{\min}, \nu_{\max})$
  $\qquad -\log\mathcal{F}(0, \sum_{i=1}^{m}\sum_{j=1}^{n_i} o_{ij}, C_1, \nu_{\min}, \nu_{\max})\Big\}$

If fitting the Bernoulli model:

- Update Multivariate Normal $q^*(\boldsymbol{\beta}, \boldsymbol{u})$ and $\boldsymbol{\xi}$ parameters:
  $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \left[2\,\boldsymbol{C}^{\mathsf{T}}\text{diag}\{A(\boldsymbol{\xi})\}\boldsymbol{C} + \boldsymbol{M}_{q(\boldsymbol{\Omega})}\right]^{-1}$
  $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^{\mathsf{T}}(\boldsymbol{y} - \frac{1}{2}\mathbf{1})$
  $\boldsymbol{\xi} \leftarrow \sqrt{\text{diagonal}\left\{\boldsymbol{C}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}^{\mathsf{T}}\right)\boldsymbol{C}^{\mathsf{T}}\right\}}$

If fitting the Poisson model:

- Update $q^*(\boldsymbol{\beta}, \boldsymbol{u})$ and $\boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ parameters:
  $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \left\{\boldsymbol{C}^{\mathsf{T}}\text{diag}(\boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})})\boldsymbol{C} + \boldsymbol{M}_{q(\boldsymbol{\Omega})}\right\}^{-1}$
  $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\left\{\boldsymbol{C}^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})}) - \boldsymbol{M}_{q(\boldsymbol{\Omega})}\,\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\right\}$
  $\boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \exp\left\{\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \frac{1}{2}\text{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^{\mathsf{T}})\right\}$

For any response model:

- Update inverse-Gamma $q^*(a_{u\ell})$ and inverse-Gamma $q^*(\sigma^2_{u\ell})$ parameters:
  For $\ell = 1, \ldots, L$:
  $\mu_{q(1/a_{u\ell})} \leftarrow 1/(\mu_{q(1/\sigma^2_{u\ell})} + A_{u\ell}^{-2})$
  $\mu_{q(\sigma^2_{u\ell})} \leftarrow \dfrac{q_\ell^{\mathrm{G}} + 1}{||\boldsymbol{\mu}_{q(\boldsymbol{u}_\ell^{\mathrm{G}})}||^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_\ell^{\mathrm{G}})}) + 2\,\mu_{q(1/a_{u\ell})}}$
- Update inverse-Gamma $q^*(a^{\mathrm{RL2}}_{r_{\mathrm{L2}}})$ and inverse-Wishart $q^*(\boldsymbol{\Sigma}^{\mathrm{RL2}})$ parameters:
  For $r_{\mathrm{L2}} = 1, \ldots, q^{\mathrm{RL2}}$:
  $B_{q(a^{\mathrm{RL2}}_{r_{\mathrm{L2}}})} \leftarrow \nu(\boldsymbol{M}_{q\left((\boldsymbol{\Sigma}_u^{\mathrm{RL2}})^{-1}\right)})_{r_{\mathrm{L2}}r_{\mathrm{L2}}} + A_{\mathrm{RL2}}^{-2}$
  $\mu_{q(a^{\mathrm{RL2}}_{r_{\mathrm{L2}}})} \leftarrow \frac{1}{2}(\nu + q^{\mathrm{RL2}})/B_{q(a^{\mathrm{RL2}}_{r_{\mathrm{L2}}})}$
  $\boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\mathrm{RL2}})} \leftarrow \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\boldsymbol{\mu}_{q(\boldsymbol{u}_{ij}^{\mathrm{RL2}})}\boldsymbol{\mu}^{\mathsf{T}}_{q(\boldsymbol{u}_{ij}^{\mathrm{RL2}})} + \boldsymbol{\Sigma}_{q(\boldsymbol{u}_{ij}^{\mathrm{RL2}})})$
  $\qquad + 2\,\nu\,\text{diag}(\mu_{q(1/a_1^{\mathrm{RL2}})}, \ldots, \mu_{q(1/a^{\mathrm{RL2}}_{q^{\mathrm{RL2}}})})$
  $\boldsymbol{M}_{q\left((\boldsymbol{\Sigma}_u^{\mathrm{RL2}})^{-1}\right)} \leftarrow (\nu + \sum_{i=1}^{m} n_i + q^{\mathrm{RL2}} - 1)\ \boldsymbol{B}^{-1}_{q(\boldsymbol{\Sigma}^{\mathrm{RL2}})}$

- Update inverse-Gamma $q^*(a_{r_{\mathrm{L3}}}^{\mathrm{RL3}})$ and inverse-Wishart $q^*(\boldsymbol{\Sigma}^{\mathrm{RL3}})$ parameters:
  For $r_{\mathrm{L3}} = 1, \ldots, q^{\mathrm{RL3}}$:

$$B_{q(a_{r_{\mathrm{L3}}}^{\mathrm{RL3}})} \leftarrow \nu(\boldsymbol{M}_{q\left((\boldsymbol{\Sigma}_u^{\mathrm{RL3}})^{-1}\right)})_{r_{\mathrm{L3}} r_{\mathrm{L3}}} + A_{\mathrm{RL3}}^{-2}$$

$$\mu_{q(a_{r_{\mathrm{L3}}}^{\mathrm{RL3}})} \leftarrow \tfrac{1}{2}(\nu + q^{\mathrm{RL3}})/B_{q(a_{r_{\mathrm{L3}}}^{\mathrm{RL3}})}$$

$$\boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\mathrm{RL3}})} \leftarrow \sum_{i=1}^{m}(\boldsymbol{\mu}_{q(\boldsymbol{u}_i^{\mathrm{RL3}})}\boldsymbol{\mu}_{q(\boldsymbol{u}_i^{\mathrm{RL3}})}^{\mathsf{T}} + \boldsymbol{\Sigma}_{q(\boldsymbol{u}_i^{\mathrm{RL3}})}) + 2\,\nu\,\mathrm{diag}(\mu_{q(1/a_1^{\mathrm{RL3}})}, \ldots, \mu_{q(1/a_{q^{\mathrm{RL3}}}^{\mathrm{RL3}})})$$

$$\boldsymbol{M}_{q\left((\boldsymbol{\Sigma}_u^{\mathrm{RL3}})^{-1}\right)} \leftarrow (\nu + m + q^{\mathrm{RL3}} - 1)\,\boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\mathrm{RL3}})}^{-1}$$

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

Algorithm 10: Mean field variational Bayes algorithm for obtaining model parameters in the optimal q-density functions for the three-level Bayesian semiparametric mixed models with Gaussian, Student-t, Bernoulli and Poisson responses.

Chapter 4

Measurement Error, Missing Data and Real-Time Extensions

You can use all the quantitative data you can get, but you still have to distrust it and use your own intelligence and judgment.

Alvin Toffler

4.1 Introduction

The beauty of the Bayesian paradigm combined with MFVB algorithms is its tremendous flexibility. Our MFVB algorithms are "modular" in the sense that the methods of handling, for example, predictor effects, spline or wavelet regression, measurement error and/or missing data models and real-time models can be combined relatively easily by incorporating additional nodes in the DAGs. In this chapter, we explore three interesting and challenging extensions of Bayesian semiparametric mixed models with an emphasis on the Gaussian response as previously described in Section 2.5. These include: data subject to measurement error problems, data subject to missing data problems and data that are processed in real-time as they arrive.

All of the examples given so far in this thesis have been for datasets with complete information on all variables and with the assumption that the variables of interest have been measured accurately. Unfortunately, many real-world applications suffer from measurement error and/or missing data problems. For example, in the *New South Wales 45 and*

The real-time content of this chapter was presented at the 16th J.B. Douglas Awards organised by the Statistical Society of Australia, Sydney, Australia, 2015. (Awarded Equal First Prize for oral presentation).

Up Study (http://www.45andUp.org.au) in which the data come from questionnaires filled out by individuals aged 45 years and over, it is common for questions involving smoking and alcohol consumption to be skipped or under-reported. In addition, a potentially important variable, such as weekly protein intake, is not measured by the questionnaires but a surrogate, such as the number of times eating red meats per week, could be formed from a composite of the questionnaire responses. Such a situation is often called the *classical measurement error problem*, in which the truth is measured with additive errors, usually with a constant variance. Methods on how to best account for such impurities in the data have spawned major areas of statistical research. Books devoted to handling measurement error and/or missing data problems include Little and Rubin (2014), Carroll *et al.* (2006) and Daniels and Hogan (2008).

When the data are susceptible to measurement error and/or missingness problems, a Bayesian approach allows a relatively straightforward incorporation of standard measurement error and/or missing data mechanisms (e.g. Faes *et al.*, 2011; Richardson *et al.*, 2002), resulting in a larger Bayesian hierarchical model. Inference via MCMC is simple in principle, but can be costly in computational time. The first two-third of this chapter is concerned with the fast variational analysis of Bayesian semiparametric mixed models with error-contaminated data. Efficiency is achieved by using MFVB posterior approximations. This is a deterministic approach that yields approximate inference, rather than asymptotically exact inference produced by the standard MCMC methods.

The remaining one-third of this chapter is to extend our previously developed streamlined MFVB algorithms to handle real-time variational inference, in which the data are processed as they are collected and made immediately available via modern technologies. Previous variational inference assumes that the data are processed in *batch*, that is, all at the same time. Here we treat data as arriving online and process them in real time, and thus have the advantage of not storing potentially very large datasets.

4.2 Gaussian semiparametric mixed models with measurement error problems

We distinguish between two types of predictors: $\boldsymbol{s}$ represents a predictor that, for all practical purposes, is measured without error (for the present example, $\boldsymbol{s}$ is continuous with non-linear predictor effects), and $\boldsymbol{x}$ represents a predictor that is not measured accurately on all subjects. The distinguishing feature of a measurement error model from model (2.11) is that we can observe an imperfect surrogate of $\boldsymbol{x}$, denoted by $\boldsymbol{w}$, and the classical measurement error model states that $\boldsymbol{w} = \boldsymbol{x} + \boldsymbol{v}$, where $\boldsymbol{w}$ is an unbiased measure of $\boldsymbol{x}$ and $\boldsymbol{v}$ are i.i.d. $N(\boldsymbol{0}, \sigma_w^2 \boldsymbol{I})$ random variables. Since $\boldsymbol{x}$ is not fully observed, it is not

possible to regress y on $(\boldsymbol{x}, \boldsymbol{s})$. The goal of measurement error modelling is therefore to obtain accurate estimates of model parameters indirectly by fitting a model of $\boldsymbol{y}$ in terms of $(\boldsymbol{w}, \boldsymbol{s})$.

Early work on Bayesian analysis of measurement error problems has been developed by Clayton (1992), followed by Richardson *et al.* (2002). Their work are based on the notion of *structural and functional specifications*. We split the predictor $\boldsymbol{x}$ into two subgroups: $\boldsymbol{x}_{\text{obs}}$ that involves the accurate measurements of the predictor and $\boldsymbol{x}_{\text{unobs}}$ that involves the inaccurate measurements of the predictor. The observations $\boldsymbol{x}_{\text{obs}}$ are generally obtained from a *validation sample*, i.e. a subsample of the data where, by design, the predictor $\boldsymbol{x}$ is accurately recorded by the use of the so-called "gold standard" method. Quite often, this is costly to implement on a large scale, hence $\boldsymbol{x}_{\text{unobs}}$ is usually larger than $\boldsymbol{x}_{\text{obs}}$. To be specific, let us distinguish data in the validation sample and the main sample by the superscript index "obs" or "unobs" respectively. Hence, $\boldsymbol{w}_{x_{\text{obs}}}$ partners the validation sample $\boldsymbol{x}_{\text{obs}}$ and $\boldsymbol{w}_{x_{\text{unobs}}}$ contains the surrogates for the $\boldsymbol{x}_{\text{obs}}$.

Throughout this section, we let $\boldsymbol{y}$ denote the known outcome, $\boldsymbol{x}$ denote the true predictor which is unobserved (except in the validation sample), $\boldsymbol{w}$ denote the observed surrogate for $\boldsymbol{x}$ and $\boldsymbol{s}$ denote the known continuous, non-linear predictor. For the present work, we assume that $\boldsymbol{x}$ and $\boldsymbol{w}$ are univariate. As detailed in Richardson *et al.* (2002), the structural specifications of a measurement error model entail the formulation of three submodels:

1. a regression model which expresses the relationship between $\boldsymbol{y}$, $\boldsymbol{x}$ and $\boldsymbol{s}$, denoted by $p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{x}, \boldsymbol{s}, \boldsymbol{u}, \sigma_\varepsilon^2)$;

2. a classical measurement model which express the relationship between $\boldsymbol{w}$ and $\boldsymbol{x}$, denoted by $p(\boldsymbol{w}|\boldsymbol{x}, \sigma_w^2)$; and

3. a prior model which specifies the prior distribution of $\boldsymbol{x}$, denoted by $p(\boldsymbol{x}|\mu_x, \sigma_x^2)$,

The definition of $\boldsymbol{u}$ is given in Section 2.3 and appropriate priors are placed on $\sigma_\varepsilon^2, \sigma_w^2, \mu_x$ and σ_x^2. Following the above, functional forms for the distributions involved in the submodels are required to be chosen. For the present regression and measurement error models, we use linear regression coupled with normal error models corresponding to additive errors, a choice that has been frequently used in epidemiology. To keep this section manageable, we also choose a normal distribution to represent the prior model $p(\boldsymbol{x}|\mu_x, \sigma_x^2)$. Although not the most ideal, this can be relatively straightforward to extend to a mixture model, which avoids possible model misspecification in the prior model.

Putting all of these components together, we now consider the following model which exploits the structure described above:

$$\begin{aligned} \boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{u}, \sigma_\varepsilon^2 &\sim N(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \sigma_\varepsilon^2 \boldsymbol{I}), \quad \boldsymbol{\beta} \sim N(\boldsymbol{0}, \sigma_\beta^2 \boldsymbol{I}) \\ \boldsymbol{x}_{\text{obs}}|\mu_x, \sigma_x^2 &\sim N(\mu_x \boldsymbol{1}_{n_{\text{obs}}}, \sigma_x^2 \boldsymbol{I}), \quad \boldsymbol{x}_{\text{unobs}} \,|\, \mu_x, \sigma_x^2 \sim N(\mu_x \boldsymbol{1}_{n_{\text{unobs}}}, \sigma_x^2 \boldsymbol{I}), \end{aligned} \tag{4.1}$$

$$
\begin{aligned}
\boldsymbol{w}_{x_{\text{obs}}}|\boldsymbol{x}_{\text{obs}},\sigma_w^2 &\sim N(\boldsymbol{x}_{\text{obs}},\sigma_w^2\boldsymbol{I}), \quad \boldsymbol{w}_{x_{\text{unobs}}}|\boldsymbol{x}_{\text{unobs}},\sigma_w^2 \sim N(\boldsymbol{x}_{\text{unobs}},\sigma_w^2\boldsymbol{I})\\
\boldsymbol{u}|\sigma_u^2,\boldsymbol{\Sigma}^{\text{R}} &\sim N\left(\boldsymbol{0},\begin{bmatrix}\sigma_u^2\boldsymbol{I}_{q^{\text{G}}} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}_m\otimes\boldsymbol{\Sigma}^{\text{R}}\end{bmatrix}\right), \quad \mu_x|\sigma_x^2 \sim N(0,\sigma_x^2),\\
\sigma_x^2|a_x &\sim \text{Inverse-Gamma}(\tfrac{1}{2},1/a_x), \quad a_x \sim \text{Inverse-Gamma}(\tfrac{1}{2},1/A_x^2),\\
\boldsymbol{\Sigma}^{\text{R}}|a_1^{\text{R}},\cdots,a_{q^{\text{R}}}^{\text{R}} &\sim \text{Inverse-Wishart}\left(\nu+q^{\text{R}}-1, 2\,\nu\,\text{diag}(1/a_1^{\text{R}},\ldots,1/a_{q^{\text{R}}}^{\text{R}})\right),\\
a_r^{\text{R}} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2},1/A_r^2), \quad 1\le r\le q^{\text{R}},\\
\sigma_u^2|a_u &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2},1/a_u), \quad a_u \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2},1/A_u^2),\\
\sigma_w^2|a_w &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2},1/a_w), \quad a_w \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2},1/A_w^2),\\
\sigma_\varepsilon^2|a_\varepsilon &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2},1/a_\varepsilon), \quad a_\varepsilon \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2},1/A_\varepsilon^2).
\end{aligned}
$$

Whilst model (4.1) extends model (2.11) by including the measurement error mechanism, its follows the same way to be represented as a mixed model as shown in Section 2.3. The prior specification of model (4.1) can be found in Section 2.4. The conditional independence properties of model (4.1) can be visualised in the DAG as depicted in Figure 4.2.

4.2.1 Notation

The following notation is useful in the upcoming subsections. Let n_i^{obs} denote the number of observed x_{ij}s in the ith group and n_i^{unobs} denote the number of unobserved x_{ij}s in the ith group. Let $\boldsymbol{x}_{\text{obs},i}$ be the $n_i^{\text{obs}}\times 1$ vector containing the observed x_{ij}s and $\boldsymbol{x}_{\text{unobs},i}$ be the $n_i^{\text{unobs}}\times 1$ containing the unobserved x_{ij}s. We reorder the data so that the observed data is first within each group. Hence, the full vector of predictor is

$$\boldsymbol{x} \equiv \begin{bmatrix}\boldsymbol{x}_{\text{obs},1} & \boldsymbol{x}_{\text{unobs},1} & \cdots & \boldsymbol{x}_{\text{obs},m} & \boldsymbol{x}_{\text{unobs},m}\end{bmatrix}^{\mathsf{T}}.$$

In addition, let $x_{\text{obs},ij}$ be the value of the predictor corresponding to the jth entry of $\boldsymbol{x}_{\text{obs},i}$ and $x_{\text{unobs},ij'}$ be the value of the predictor corresponding to the j'th entry of $\boldsymbol{x}_{\text{unobs},i}$ for $1\le i\le m$, $1\le j\le n_i^{\text{obs}}$ and $1\le j'\le n_i^{\text{unobs}}$.

Define the following vectors and matrices

$$
\begin{aligned}
\boldsymbol{x}_i &= \begin{bmatrix}\boldsymbol{x}_{\text{obs},i}\\ \boldsymbol{x}_{\text{unobs},i}\end{bmatrix}, \quad \boldsymbol{x}_{\text{obs},i} = \begin{bmatrix}x_{\text{obs},i1}\\ \vdots\\ x_{\text{obs},in_i^{\text{obs}}}\end{bmatrix}, \quad \boldsymbol{x}_{\text{unobs},i} = \begin{bmatrix}x_{\text{unobs},i1}\\ \vdots\\ x_{\text{unobs},in_i^{\text{unobs}}}\end{bmatrix},\\
\boldsymbol{Z}_{x_{\text{obs}}}^{\text{R}} &= \underset{1\le i\le m}{\text{blockdiag}}\left(\begin{bmatrix}\boldsymbol{1} & \boldsymbol{x}_{\text{obs},i}\end{bmatrix}\right), \quad \boldsymbol{Z}_{x_{\text{unobs}}}^{\text{R}} = \underset{1\le i\le m}{\text{blockdiag}}\left(\begin{bmatrix}\boldsymbol{1} & \boldsymbol{x}_{\text{unobs},i}\end{bmatrix}\right),
\end{aligned}
$$

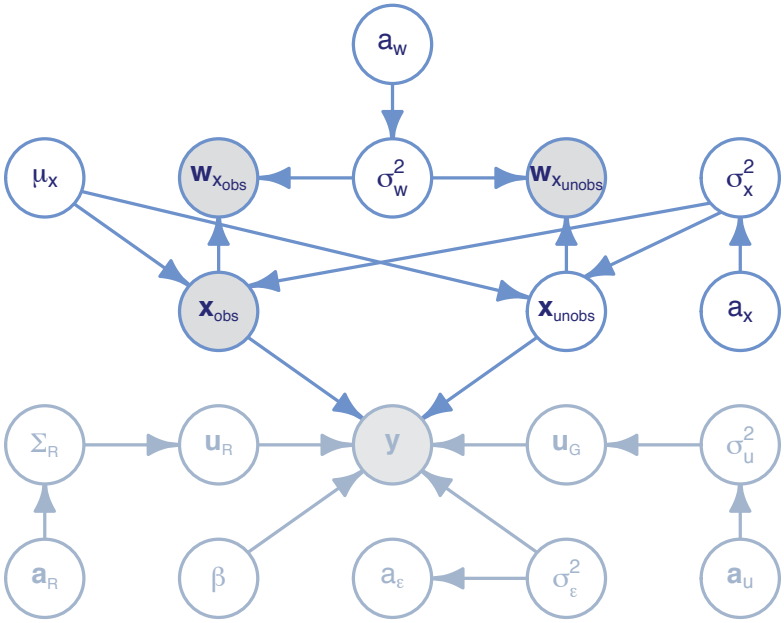

Figure 4.1: Directed acyclic graph corresponds to model (4.1). The shaded node corresponds to the observed data vector and the open nodes correspond to the random or auxiliary variables. The fragments shown in grey are presented in the Gaussian model DAG as shown in Figure 2.3.

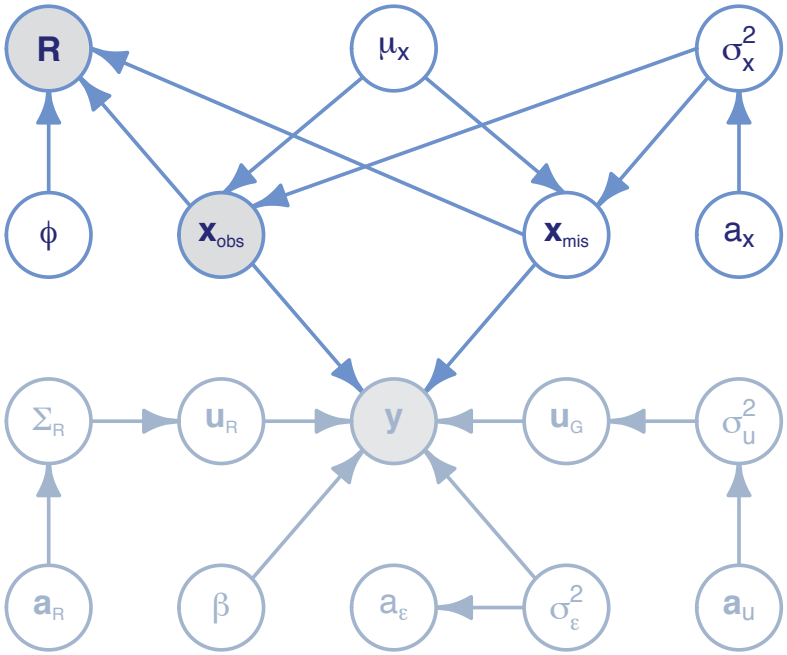

Figure 4.2: Directed acyclic graph corresponds to model (4.7). The shaded node corresponds to the observed data vector and the open nodes correspond to the random or auxiliary variables. The fragments shown in grey are presented in the Gaussian model DAG as shown in Figure 2.3.

$$\boldsymbol{Z}^{\mathrm{G}}_{x_{\mathrm{obs}}} = \begin{bmatrix} z_1(\boldsymbol{s}_{x_{\mathrm{obs},1}}) & \cdots & z_{q^{\mathrm{G}}}(\boldsymbol{s}_{x_{\mathrm{obs},1}}) \\ \vdots & \vdots & \vdots \\ z_1(\boldsymbol{s}_{x_{\mathrm{obs},m}}) & \cdots & z_{q^{\mathrm{G}}}(\boldsymbol{s}_{x_{\mathrm{obs},m}}) \end{bmatrix}, \quad \boldsymbol{Z}^{\mathrm{G}}_{x_{\mathrm{unobs}}} = \begin{bmatrix} z_1(\boldsymbol{s}_{x_{\mathrm{unobs},1}}) & \cdots & z_{q^{\mathrm{G}}}(\boldsymbol{s}_{x_{\mathrm{unobs},1}}) \\ \vdots & \vdots & \vdots \\ z_1(\boldsymbol{s}_{x_{\mathrm{unobs},m}}) & \cdots & z_{q^{\mathrm{G}}}(\boldsymbol{s}_{x_{\mathrm{unobs},m}}) \end{bmatrix}.$$

For conciseness of forthcoming expressions we adopt the following notations:

$$\begin{aligned} \boldsymbol{C} &= \begin{bmatrix} \boldsymbol{C}_{x_{\mathrm{obs}}} \\ \boldsymbol{C}_{x_{\mathrm{unobs}}} \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \boldsymbol{x}_{\mathrm{obs}} & \boldsymbol{s}_{x_{\mathrm{obs}}} & \boldsymbol{Z}^{\mathrm{G}}_{x_{\mathrm{obs}}} & \boldsymbol{Z}^{\mathrm{R}}_{x_{\mathrm{obs}}} \\ \mathbf{1} & \boldsymbol{x}_{\mathrm{unobs}} & \boldsymbol{s}_{x_{\mathrm{unobs}}} & \boldsymbol{Z}^{\mathrm{G}}_{x_{\mathrm{unobs}}} & \boldsymbol{Z}^{\mathrm{R}}_{x_{\mathrm{unobs}}} \end{bmatrix} \\ E_q(\boldsymbol{C}) &= \begin{bmatrix} \boldsymbol{C}_{x_{\mathrm{obs}}} \\ E_q(\boldsymbol{C}_{x_{\mathrm{unobs}}}) \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \boldsymbol{x}_{\mathrm{obs}} & \boldsymbol{s}_{x_{\mathrm{obs}}} & \boldsymbol{Z}^{\mathrm{G}}_{x_{\mathrm{obs}}} & \boldsymbol{Z}^{\mathrm{R}}_{x_{\mathrm{obs}}} \\ \mathbf{1} & E_q(\boldsymbol{x}_{\mathrm{unobs}}) & \boldsymbol{s}_{x_{\mathrm{unobs}}} & \boldsymbol{Z}^{\mathrm{G}}_{x_{\mathrm{unobs}}} & E_q(\boldsymbol{Z}^{\mathrm{R}}_{x_{\mathrm{unobs}}}) \end{bmatrix}, \end{aligned}$$

where

$$\begin{aligned} E_q(\boldsymbol{x}_{\mathrm{unobs}}) &= \begin{bmatrix} \boldsymbol{\mu}_{q(\boldsymbol{x}_{\mathrm{unobs},1})} & \cdots & \boldsymbol{\mu}_{q(\boldsymbol{x}_{\mathrm{unobs},m})} \end{bmatrix}^{\mathsf{T}} \\ \text{and} \quad E_q(\boldsymbol{Z}^{\mathrm{R}}_{x_{\mathrm{unobs}}}) &= \underset{1\le i\le m}{\mathrm{blockdiag}} \left(\begin{bmatrix} \mathbf{1} & \boldsymbol{\mu}_{q(\boldsymbol{x}_{\mathrm{unobs},i})} \end{bmatrix} \right), \end{aligned}$$

and the explicit expression for $\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C}$ is

$$\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C} = \begin{bmatrix} n & \mathbf{1}^{\mathsf{T}}\boldsymbol{x} & \mathbf{1}^{\mathsf{T}}\boldsymbol{s} \\ \boldsymbol{x}^{\mathsf{T}}\mathbf{1} & \boldsymbol{x}^{\mathsf{T}}_{\mathrm{obs}}\boldsymbol{x}_{\mathrm{obs}} + \boldsymbol{x}^{\mathsf{T}}_{\mathrm{unobs}}\boldsymbol{x}_{\mathrm{unobs}} & \boldsymbol{x}^{\mathsf{T}}_{\mathrm{obs}}\boldsymbol{s}_{x_{\mathrm{obs}}} + \boldsymbol{x}^{\mathsf{T}}_{\mathrm{unobs}}\boldsymbol{s}_{x_{\mathrm{unobs}}} \\ \boldsymbol{s}^{\mathsf{T}}\mathbf{1} & \boldsymbol{s}^{\mathsf{T}}_{x_{\mathrm{obs}}}\boldsymbol{x}_{\mathrm{obs}} + \boldsymbol{s}^{\mathsf{T}}_{x_{\mathrm{unobs}}}\boldsymbol{x}_{\mathrm{unobs}} & \boldsymbol{s}^{\mathsf{T}}_{x_{\mathrm{obs}}}\boldsymbol{s}_{x_{\mathrm{obs}}} + \boldsymbol{s}^{\mathsf{T}}_{x_{\mathrm{unobs}}}\boldsymbol{s}_{x_{\mathrm{unobs}}} \\ (\boldsymbol{Z}^{\mathrm{G}})^{\mathsf{T}}\mathbf{1} & (\boldsymbol{Z}^{\mathrm{G}}_{x_{\mathrm{obs}}})^{\mathsf{T}}\boldsymbol{x}_{\mathrm{obs}} + (\boldsymbol{Z}^{\mathrm{G}}_{x_{\mathrm{unobs}}})^{\mathsf{T}}\boldsymbol{x}_{\mathrm{unobs}} & (\boldsymbol{Z}^{\mathrm{G}}_{x_{\mathrm{obs}}})^{\mathsf{T}}\boldsymbol{s}_{x_{\mathrm{obs}}} + (\boldsymbol{Z}^{\mathrm{G}}_{x_{\mathrm{unobs}}})^{\mathsf{T}}\boldsymbol{s}_{x_{\mathrm{unobs}}} \\ (\boldsymbol{Z}^{\mathrm{R}})^{\mathsf{T}}\mathbf{1} & (\boldsymbol{Z}^{\mathrm{R}}_{x_{\mathrm{obs}}})^{\mathsf{T}}\boldsymbol{x}_{\mathrm{obs}} + (\boldsymbol{Z}^{\mathrm{R}}_{x_{\mathrm{unobs}}})^{\mathsf{T}}\boldsymbol{x}_{\mathrm{unobs}} & (\boldsymbol{Z}^{\mathrm{R}}_{x_{\mathrm{obs}}})^{\mathsf{T}}\boldsymbol{s}_{x_{\mathrm{obs}}} + (\boldsymbol{Z}^{\mathrm{R}}_{x_{\mathrm{unobs}}})^{\mathsf{T}}\boldsymbol{s}_{x_{\mathrm{unobs}}} \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{1}^{\mathsf{T}}\boldsymbol{Z}^{\mathrm{G}} & \mathbf{1}^{\mathsf{T}}\boldsymbol{Z}^{\mathrm{R}} \\ \boldsymbol{x}^{\mathsf{T}}_{\mathrm{obs}}\boldsymbol{Z}^{\mathrm{G}}_{x_{\mathrm{obs}}} + \boldsymbol{x}^{\mathsf{T}}_{\mathrm{unobs}}\boldsymbol{Z}^{\mathrm{G}}_{x_{\mathrm{unobs}}} & \boldsymbol{x}^{\mathsf{T}}_{\mathrm{obs}}\boldsymbol{Z}^{\mathrm{R}}_{x_{\mathrm{obs}}} + \boldsymbol{x}^{\mathsf{T}}_{\mathrm{unobs}}\boldsymbol{Z}^{\mathrm{R}}_{x_{\mathrm{unobs}}} \\ \boldsymbol{s}^{\mathsf{T}}_{x_{\mathrm{obs}}}\boldsymbol{Z}^{\mathrm{G}}_{x_{\mathrm{obs}}} + \boldsymbol{s}^{\mathsf{T}}_{x_{\mathrm{unobs}}}\boldsymbol{Z}^{\mathrm{G}}_{x_{\mathrm{unobs}}} & \boldsymbol{s}^{\mathsf{T}}_{x_{\mathrm{obs}}}\boldsymbol{Z}^{\mathrm{R}}_{x_{\mathrm{obs}}} + \boldsymbol{s}^{\mathsf{T}}_{x_{\mathrm{unobs}}}\boldsymbol{Z}^{\mathrm{R}}_{x_{\mathrm{unobs}}} \\ (\boldsymbol{Z}^{\mathrm{G}}_{x_{\mathrm{obs}}})^{\mathsf{T}}\boldsymbol{Z}^{\mathrm{G}}_{x_{\mathrm{obs}}} + (\boldsymbol{Z}^{\mathrm{G}}_{x_{\mathrm{unobs}}})^{\mathsf{T}}\boldsymbol{Z}^{\mathrm{G}}_{x_{\mathrm{unobs}}} & (\boldsymbol{Z}^{\mathrm{G}}_{x_{\mathrm{obs}}})^{\mathsf{T}}\boldsymbol{Z}^{\mathrm{R}}_{x_{\mathrm{obs}}} + (\boldsymbol{Z}^{\mathrm{G}}_{x_{\mathrm{unobs}}})^{\mathsf{T}}\boldsymbol{Z}^{\mathrm{R}}_{x_{\mathrm{unobs}}} \\ (\boldsymbol{Z}^{\mathrm{R}}_{x_{\mathrm{obs}}})^{\mathsf{T}}\boldsymbol{Z}^{\mathrm{G}}_{x_{\mathrm{obs}}} + (\boldsymbol{Z}^{\mathrm{R}}_{x_{\mathrm{unobs}}})^{\mathsf{T}}\boldsymbol{Z}^{\mathrm{G}}_{x_{\mathrm{unobs}}} & (\boldsymbol{Z}^{\mathrm{R}}_{x_{\mathrm{obs}}})^{\mathsf{T}}\boldsymbol{Z}^{\mathrm{R}}_{x_{\mathrm{obs}}} + (\boldsymbol{Z}^{\mathrm{R}}_{x_{\mathrm{unobs}}})^{\mathsf{T}}\boldsymbol{Z}^{\mathrm{R}}_{x_{\mathrm{unobs}}} \end{bmatrix}. \quad (4.2)$$

Let $\boldsymbol{\mu}_{q(\boldsymbol{X}_{\mathrm{unobs},i})} = \begin{bmatrix} \mathbf{1} & \boldsymbol{\mu}_{q(\boldsymbol{x}_{\mathrm{unobs},1})} \end{bmatrix}$, then the expressions corresponding to $E_q(\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C})$ with respect to $\boldsymbol{x}_{\mathrm{unobs}}$ are as follows:

First we note that:

$$\begin{aligned} E_q((\boldsymbol{Z}^{\mathrm{R}}_{x_{\mathrm{unobs}}})^{\mathsf{T}}\mathbf{1}) &= E_q^{\mathsf{T}}(\mathbf{1}^{\mathsf{T}}\boldsymbol{Z}^{\mathrm{R}}_{x_{\mathrm{unobs}}}) \\ E_q((\boldsymbol{x}_{\mathrm{unobs}})^{\mathsf{T}}\boldsymbol{s}_{x_{\mathrm{unobs}}}) &= E_q^{\mathsf{T}}((\boldsymbol{s}_{x_{\mathrm{unobs}}})^{\mathsf{T}}\boldsymbol{x}_{\mathrm{unobs}}) \end{aligned}$$

$$
\begin{aligned}
E_q((\boldsymbol{Z}^{\mathrm{R}}_{x_{\mathrm{unobs}}})^{\mathsf{T}}\boldsymbol{s}_{x_{\mathrm{unobs}}}) &= E_q^{\mathsf{T}}(\boldsymbol{s}^{\mathsf{T}}_{x_{\mathrm{unobs}}}\boldsymbol{Z}^{\mathrm{R}}_{x_{\mathrm{unobs}}}) \\
E_q((\boldsymbol{Z}^{\mathrm{G}}_{x_{\mathrm{unobs}}})^{\mathsf{T}}\boldsymbol{x}_{\mathrm{unobs}}) &= E_q^{\mathsf{T}}(\boldsymbol{x}^{\mathsf{T}}_{\mathrm{unobs}}\boldsymbol{Z}^{\mathrm{G}}_{x_{\mathrm{unobs}}}).
\end{aligned}
$$

The remaining expectation update expressions are derived via straightforward algebra manipulations:

$$
\begin{aligned}
E_q((\boldsymbol{x}_{\mathrm{unobs}})^{\mathsf{T}}\boldsymbol{x}_{\mathrm{unobs}}) &\leftarrow \|\boldsymbol{\mu}_{q(\boldsymbol{x}_{\mathrm{unobs}})}\|^2 + \sum_{i=1}^{m} n_i^{\mathrm{unobs}}\sigma^2_{q(\boldsymbol{x}_{\mathrm{unobs},i})} \qquad (4.3)\\
E_q((\boldsymbol{s}_{x_{\mathrm{unobs}}})^{\mathsf{T}}\boldsymbol{x}_{\mathrm{unobs}}) &\leftarrow \sum_{i=1}^{m} \boldsymbol{s}^{\mathsf{T}}_{x_{\mathrm{unobs},i}}\boldsymbol{\mu}_{q(\boldsymbol{x}_{\mathrm{unobs},i})}\\
E_q((\boldsymbol{Z}^{\mathrm{R}}_{x_{\mathrm{unobs}}})^{\mathsf{T}}\boldsymbol{Z}^{\mathrm{R}}_{x_{\mathrm{unobs}}}) &\leftarrow \operatorname*{blockdiag}_{1\le i\le m}\left(\boldsymbol{\mu}^{\mathsf{T}}_{q(\boldsymbol{X}_{\mathrm{unobs},i})}\boldsymbol{\mu}_{q(\boldsymbol{X}_{\mathrm{unobs},i})}\right) + \sum_{i=1}^{m} n_i^{\mathrm{unobs}}\sigma^2_{q(\boldsymbol{x}_{\mathrm{unobs},i})}\\
E_q(\mathbf{1}^{\mathsf{T}}\boldsymbol{Z}^{\mathrm{R}}_{x_{\mathrm{unobs}}}) &\leftarrow \begin{bmatrix}\mathbf{1}^{\mathsf{T}}\boldsymbol{\mu}_{q(\boldsymbol{X}_{\mathrm{unobs},1})} & \cdots & \mathbf{1}^{\mathsf{T}}\boldsymbol{\mu}_{q(\boldsymbol{X}_{\mathrm{unobs},m})}\end{bmatrix}\\
E_q(\boldsymbol{x}^{\mathsf{T}}_{\mathrm{unobs}}\boldsymbol{Z}^{\mathrm{R}}_{x_{\mathrm{unobs}}}) &\leftarrow \begin{bmatrix}\|\boldsymbol{\mu}_{q(\boldsymbol{x}_{\mathrm{unobs},1})}\|^2 & \cdots & \|\boldsymbol{\mu}_{q(\boldsymbol{x}_{\mathrm{unobs},m})}\|^2\end{bmatrix} + \sum_{i=1}^{m} n_i^{\mathrm{unobs}}\sigma^2_{q(\boldsymbol{x}_{\mathrm{unobs},i})}\\
E_q(\boldsymbol{s}^{\mathsf{T}}_{x_{\mathrm{unobs}}}\boldsymbol{Z}^{\mathrm{R}}_{x_{\mathrm{unobs}}}) &\leftarrow \begin{bmatrix}\boldsymbol{s}^{\mathsf{T}}_{x_{\mathrm{unobs},1}}\boldsymbol{\mu}_{q(\boldsymbol{X}_{\mathrm{unobs},1})} & \cdots & \boldsymbol{s}^{\mathsf{T}}_{x_{\mathrm{unobs},m}}\boldsymbol{\mu}_{q(\boldsymbol{X}_{\mathrm{unobs},m})}\end{bmatrix}\\
E_q((\boldsymbol{x}_{\mathrm{unobs}})^{\mathsf{T}}\boldsymbol{Z}^{\mathrm{G}}_{x_{\mathrm{unobs}}}) &\leftarrow \begin{bmatrix}\sum_{i=1}^{m}\boldsymbol{\mu}^{\mathsf{T}}_{q(\boldsymbol{x}_{\mathrm{unobs},i})}\, z_1(\boldsymbol{s}_{x_{\mathrm{unobs},i}}) & \cdots & \sum_{i=1}^{m}\boldsymbol{\mu}^{\mathsf{T}}_{q(\boldsymbol{x}_{\mathrm{unobs},i})}\, z_{q^{\mathrm{G}}}(\boldsymbol{s}_{x_{\mathrm{unobs},i}})\end{bmatrix}\\
E_q((\boldsymbol{Z}^{\mathrm{R}}_{x_{\mathrm{unobs}}})^{\mathsf{T}}\boldsymbol{Z}^{\mathrm{G}}_{x_{\mathrm{unobs}}}) &\leftarrow \begin{bmatrix}\boldsymbol{\mu}^{\mathsf{T}}_{q(\boldsymbol{X}_{\mathrm{unobs},1})}z_1(\boldsymbol{s}_{x_{\mathrm{unobs},1}}) & \cdots & \boldsymbol{\mu}^{\mathsf{T}}_{q(\boldsymbol{X}_{\mathrm{unobs},1})}z_{q^{\mathrm{G}}}(\boldsymbol{s}_{x_{\mathrm{unobs},1}})\\ \vdots & \vdots & \vdots\\ \boldsymbol{\mu}^{\mathsf{T}}_{q(\boldsymbol{X}_{\mathrm{unobs},m})}z_1(\boldsymbol{s}_{x_{\mathrm{unobs},m}}) & \cdots & \boldsymbol{\mu}^{\mathsf{T}}_{q(\boldsymbol{X}_{\mathrm{unobs},m})}z_{q^{\mathrm{G}}}(\boldsymbol{s}_{x_{\mathrm{unobs},m}})\end{bmatrix}\\
E_q((\boldsymbol{Z}^{\mathrm{G}}_{x_{\mathrm{unobs}}})^{\mathsf{T}}\boldsymbol{Z}^{\mathrm{R}}_{x_{\mathrm{unobs}}}) &\leftarrow \begin{bmatrix}z_1^{\mathsf{T}}(\boldsymbol{s}_{x_{\mathrm{unobs},1}})\boldsymbol{\mu}_{q(\boldsymbol{X}_{\mathrm{unobs},1})} & \cdots & z_1^{\mathsf{T}}(\boldsymbol{s}_{x_{\mathrm{unobs},m}})\boldsymbol{\mu}_{q(\boldsymbol{X}_{\mathrm{unobs},1})}\\ \vdots & \vdots & \vdots\\ z^{\mathsf{T}}_{q^{\mathrm{G}}}(\boldsymbol{s}_{x_{\mathrm{unobs},1}})\boldsymbol{\mu}_{q(\boldsymbol{X}_{\mathrm{unobs},m})} & \cdots & z^{\mathsf{T}}_{q^{\mathrm{G}}}(\boldsymbol{s}_{x_{\mathrm{unobs},m}})\boldsymbol{\mu}_{q(\boldsymbol{X}_{\mathrm{unobs},m})}\end{bmatrix}.
\end{aligned}
$$

4.2.2 Approximate Bayesian inference via mean field variational Bayes

We now provide details on MFVB fitting and inference catered to the Gaussian response model with classical measurement errors as described in model (4.1). The essence of our MFVB approach lies in approximating the full conditional joint posterior density function of the form

$$
\begin{aligned}
p(\boldsymbol{x}_{\mathrm{unobs}}, \boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{a}^{\mathrm{R}}, \boldsymbol{a}_u, a_\varepsilon, \mu_x, \boldsymbol{\Sigma}^{\mathrm{R}}, \boldsymbol{\sigma}_u^2, \sigma_\varepsilon^2, \sigma_x^2, \sigma_w^2 | \boldsymbol{y}, \boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{w}) &\approx \\
q(\boldsymbol{x}_{\mathrm{unobs}}, \boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{a}^{\mathrm{R}}, \boldsymbol{a}_u, a_\varepsilon, \mu_x, \boldsymbol{\Sigma}^{\mathrm{R}}, \boldsymbol{\sigma}_u^2, \sigma_\varepsilon^2, \sigma_x^2, \sigma_w^2)&
\end{aligned}
$$

subject to the q-density product restriction

$$q(\boldsymbol{x}_{\text{unobs}}, \boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{a}^{\text{R}}, \boldsymbol{a}_u, a_\varepsilon, \mu_x, \boldsymbol{\Sigma}^{\text{R}}, \boldsymbol{\sigma}_u^2, \sigma_\varepsilon^2, \sigma_x^2, \sigma_w^2) = q(\boldsymbol{x}_{\text{unobs}}, \boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{a}^{\text{R}}, \boldsymbol{a}_u, a_\varepsilon, \mu_x) \times q(\boldsymbol{\Sigma}^{\text{R}}, \boldsymbol{\sigma}_u^2, \sigma_\varepsilon^2, \sigma_x^2, \sigma_w^2). \tag{4.4}$$

Restriction (4.4) is the minimal factorisation in the MFVB approximation, deriving based on the heuristic arguments provided by Menictas and Wand (2013). One can also use the concepts of induced factorisation and moralisation introduced in Subsection 2.5.1 to establish additional product restriction

$$q(\boldsymbol{x}_{\text{unobs}})\, q(\boldsymbol{\beta}, \boldsymbol{u}) \left\{\prod_{r=1}^{q^{\text{R}}} q(a_r^{\text{R}})\right\} \left\{\prod_{\ell=1}^{L} q(a_{u\ell})\right\} q(a_\varepsilon)\, q(\mu_x)\, q(\boldsymbol{\Sigma}^{\text{R}}) \left\{\prod_{\ell=1}^{L} q(\sigma_{u\ell}^2)\right\} \times q(\sigma_\varepsilon^2)\, q(\sigma_x^2)\, q(\sigma_w^2). \tag{4.5}$$

As laid out in Subsection 2.6.1, the locality property of MFVB means that calculations concerning a particular parameter can be confined to "nearby" parameters through graph-theoretic representations of Bayesian hierarchical models. Therefore, some of the q-densities for this measurement error model are remained the same as in model (2.11). Nonetheless, we show that the optimal q-densities admit the following forms under the production restriction (4.5) in Appendix 4.A:

$$\begin{aligned}
&q^*(\boldsymbol{x}_{\text{unobs}}) \text{ is the } N\left(\boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs}})}, \sigma^2_{q(\boldsymbol{x}_{\text{unobs},i})} \boldsymbol{I}_{n_{\text{unobs}}}\right) \text{ density function,} \\
&q^*(\boldsymbol{\beta}, \boldsymbol{u}) \text{ is the } N(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})},\ \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}) \text{ density function,} \\
&q^*(\mu_x) \text{ is the } N(\mu_{q(\mu_x)}, \sigma^2_{q(\mu_x)}) \text{ density function} \\
&q^*(\sigma_x^2) \text{ is the Inverse-Gamma } \left(\tfrac{1}{2}\left(\textstyle\sum_{i=1}^m n_i + 1\right),\ B_{q(\sigma_x^2)}\right) \text{ density function,} \\
&q^*(\sigma_w^2) \text{ is the Inverse-Gamma } \left(\tfrac{1}{2}\left(\textstyle\sum_{i=1}^m n_i + 1\right),\ B_{q(\sigma_w^2)}\right) \text{ density function,} \\
&q^*(\sigma_\varepsilon^2) \text{ is the Inverse-Gamma } \left(\tfrac{1}{2}\left(\textstyle\sum_{i=1}^m n_i + 1\right),\ B_{q(\sigma_\varepsilon^2)}\right) \text{ density function,} \\
&q^*(a_\varepsilon) \text{ is the Inverse-Gamma}(1,\ B_{q(a_\varepsilon)}) \text{ density function,} \\
&q^*(\sigma_{u\ell}^2) \text{ is the Inverse-Gamma } \left(\tfrac{1}{2}\left(q_\ell^{\text{G}} + 1\right),\ B_{q(\sigma_{u\ell}^2)}\right) \text{ density function,} \\
&q^*(a_{u\ell}) \text{ is the Inverse-Gamma}(1,\ B_{q(a_{u\ell})}) \text{ density function,} \\
&q^*(\boldsymbol{\Sigma}^{\text{R}}) \text{ is the Inverse-Wishart } \left(\nu + m + q^{\text{R}} + 1,\ \boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\text{R}})}\right) \text{ density function,} \\
&q^*(a_r^{\text{R}}) \text{ is the Inverse-Gamma } \left(\tfrac{1}{2}(\nu + q^{\text{R}}),\ B_{q(a_r^{\text{R}})}\right) \text{ density function,}
\end{aligned} \tag{4.6}$$

where the parameters are updated according to Algorithm 11, with $\ell = 1$. The variational lower bound on the marginal log-likelihood at the bottom of the main loop is derived in

Appendix 4.A.1.

4.3 Gaussian semiparametric mixed models with missing data problems

In principle, there are three missingness mechanisms: *missing completely at random*, where the missing-data mechanism is independent of the data; *missing at random*, where the missing-data mechanism depends completely on the observed $\boldsymbol{y}$ but not on the missing $\boldsymbol{x}$; and *missing not at random*, where the missing-data mechanism depends on the unobserved $\boldsymbol{x}$. Faes *et al.* (2011) treat the full array of missingness scenarios but is limited to the simplest parametric and nonparametric models. In this section we extend their work to the Gaussian semiparametric mixed models, focussing on the most challenging missing not at random scenario, where the core tenets can be elucidated without excessive notation and algebra. The other two scenarios enjoy many of the assets of the missing not at random case and therefore are omitted in this section.

We begin with the same Bayesian framework in model (2.11) and incorporate a binary random variable R_{ij} such that

$$R_{ij} = \begin{cases} 1, & \text{if } x_{ij} \text{ is observed,} \\ 0, & \text{if } x_{ij} \text{ is missing,} \end{cases}$$

for $1 \leq i \leq m$ and $1 \leq j \leq n_i$. Bayesian inference for the missing data models differs according to the dependence of distribution of R_{ij} on the observed data as defined in Section 4.2. Putting all of these components together, we now consider the following model which exploits the structure described above:

$$\begin{aligned}
\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{u}, \sigma_\varepsilon^2 &\sim N(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \sigma_\varepsilon^2 \boldsymbol{I}), \quad \boldsymbol{\beta} \sim N(\boldsymbol{0}, \sigma_\beta^2 \boldsymbol{I}) \\
\boldsymbol{x}_{\text{obs}}|\mu_x, \sigma_x^2 &\sim N(\mu_x \boldsymbol{1}_{n_{\text{obs}}}, \sigma_x^2 \boldsymbol{I}), \quad \boldsymbol{x}_{\text{mis}}|\mu_x, \sigma_x^2 \sim N(\mu_x \boldsymbol{1}_{n_{\text{mis}}}, \sigma_x^2 \boldsymbol{I}), \\
\boldsymbol{R}|\boldsymbol{\phi}, \boldsymbol{x} &\sim \text{Bernoulli}\left(\{1 + \exp(-\boldsymbol{X}\boldsymbol{\phi})\}^{-1}\right), \quad \boldsymbol{\phi} \sim N(\boldsymbol{0}, \sigma_\phi^2 \boldsymbol{I}) \\
\boldsymbol{u}|\sigma_u^2, \boldsymbol{\Sigma}^{\text{R}} &\sim N\left(\boldsymbol{0}, \begin{bmatrix} \sigma_u^2 \boldsymbol{I}_{q^{\text{G}}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_m \otimes \boldsymbol{\Sigma}^{\text{R}} \end{bmatrix}\right), \quad \mu_x|\sigma_x^2 \sim N(0, \sigma_x^2), \\
\sigma_x^2|a_x &\sim \text{Inverse-Gamma}(\tfrac{1}{2}, 1/a_x), \quad a_x \sim \text{Inverse-Gamma}(\tfrac{1}{2}, 1/A_x^2), \\
\boldsymbol{\Sigma}^{\text{R}}|a_1^{\text{R}}, \cdots, a_{q^{\text{R}}}^{\text{R}} &\sim \text{Inverse-Wishart}\left(\nu + q^{\text{R}} - 1, 2\,\nu\,\text{diag}(1/a_1^{\text{R}}, \ldots, 1/a_{q^{\text{R}}}^{\text{R}})\right), \qquad (4.7) \\
a_r^{\text{R}} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2}, 1/A_r^2), \quad 1 \leq r \leq q^{\text{R}}, \\
\sigma_u^2|a_u &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2}, 1/a_u), \quad a_u \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2}, 1/A_u^2), \\
\sigma_w^2|a_w &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2}, 1/a_w), \quad a_w \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2}, 1/A_w^2),
\end{aligned}$$

Initialise: $\nu = 2$, $\mu_{q(1/\sigma_\varepsilon^2)} > 0$, $\mu_{q(1/a_\varepsilon)} > 0$, $\mu_{q(1/\sigma_x^2)} > 0$, $\mu_{q(1/a_x)} > 0$, $\mu_{q(1/\sigma_w^2)} > 0$, $\mu_{q(1/a_w)} > 0$, $\mu_{q(\mu_x)} > 0$, $\mu_{q(1/a_r^{\mathrm{R}})} > 0, 1 \leq r \leq q^{\mathrm{R}}$, $\mu_{q(1/\sigma_u^2)} > 0$, $\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})}$, a $q^{\mathrm{R}} \times q^{\mathrm{R}}$ positive definite matrix, $\sigma^2_{q(\boldsymbol{x}_{\mathrm{unobs},i})} > 0$, $1 \leq i \leq m$, $\boldsymbol{\mu}_{q(\boldsymbol{x}_{\mathrm{unobs},i})}$, a $n_i^{\mathrm{unobs}} \times 1$ vector of positive entries, $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$, a $(p + q^{\mathrm{R}}m + q^{\mathrm{G}}) \times 1$ vector and $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$, a $(p + mq^{\mathrm{R}} + q^{\mathrm{G}}) \times (p + mq^{\mathrm{R}} + q^{\mathrm{G}})$ positive definite matrix.

Cycle through updates:

Define: $\boldsymbol{M}_{q((\boldsymbol{\Omega})^{-1})} \equiv \begin{bmatrix} \sigma_\beta^{-2}\boldsymbol{I}_p & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \mu_{q(1/\sigma_u^2)}\boldsymbol{I}_{q^{\mathrm{G}}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_m \otimes \boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})} \end{bmatrix}$

Update Univariate Normal $q^*(\boldsymbol{x}_{\mathbf{unobs},i})$ parameters:

For $i = 1, \cdots, m$:

If $(n_i^{\mathrm{unobs}} > 0)$ then

$$\sigma^2_{q(\boldsymbol{x}_{\mathrm{unobs},i})} \leftarrow 1 \Big/ \Big[\mu_{q(1/\sigma_\varepsilon^2)} \Big\{ \boldsymbol{\mu}^2_{q(\beta_x)} + \boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_x)} + 2\left(\boldsymbol{\mu}_{q(\beta_x)}\,\boldsymbol{\mu}_{q(u_{1i}^{\mathrm{R}})} + \boldsymbol{\Lambda}_{q(\beta_x, u_{1i}^{\mathrm{R}})}\right) \Big\} + \boldsymbol{\mu}^2_{q(u_{1i}^{\mathrm{R}})} + \boldsymbol{\Sigma}_{q(u_{1i}^{\mathrm{R}})} + \mu_{q(1/\sigma_x^2)} + \mu_{q(1/\sigma_w^2)} \Big]$$

$$\begin{aligned} \boldsymbol{\mu}_{q(\boldsymbol{x}_{\mathrm{unobs},i})} \leftarrow \sigma^2_{q(\boldsymbol{x}_{\mathrm{unobs},i})} \boldsymbol{I}_{n_i^{\mathrm{unobs}}} \Big[& \mu_{q(1/\sigma_\varepsilon^2)} \Big\{ \left(\boldsymbol{\mu}_{q(\beta_x)} + \boldsymbol{\mu}_{q(u_{1i}^{\mathrm{R}})}\right) \boldsymbol{y}_{x_{\mathrm{unobs},i}} \\ & - \left(\boldsymbol{\mu}_{q(\beta_0)}\,\boldsymbol{\mu}_{q(\beta_x)} + \boldsymbol{\Lambda}_{q(\beta_0,\beta_x)} + \boldsymbol{\mu}_{q(\beta_0)}\,\boldsymbol{\mu}_{q(u_{1i}^{\mathrm{R}})} + \boldsymbol{\Lambda}_{q(\beta_0, u_{1i}^{\mathrm{R}})}\right) \boldsymbol{1}_{n_i^{\mathrm{unobs}}} \\ & - \left(\boldsymbol{\mu}_{q(\beta_x)}\,\boldsymbol{\mu}_{q(\beta_s)} + \boldsymbol{\Lambda}_{q(\beta_x,\beta_s)} + \boldsymbol{\mu}_{q(\beta_s)}\,\boldsymbol{\mu}_{q(u_{1i}^{\mathrm{R}})} + \boldsymbol{\Lambda}_{q(\beta_s, u_{1i}^{\mathrm{R}})}\right) \boldsymbol{s}_{x_{\mathrm{unobs},i}} \\ & - \textstyle\sum_{k=1}^{q^{\mathrm{G}}} \left(\boldsymbol{\mu}_{q(\beta_x)}\,\boldsymbol{\mu}_{q(u_k^{\mathrm{G}})} + \boldsymbol{\Lambda}_{q(\beta_x, u_k^{\mathrm{G}})}\right) z_{q^{\mathrm{G}}}\left(\boldsymbol{s}_{x_{\mathrm{unobs},i}}\right) \\ & - \textstyle\sum_{k=1}^{q^{\mathrm{G}}} \left(\boldsymbol{\mu}_{q(u_{1i}^{\mathrm{R}})}\,\boldsymbol{\mu}_{q(u_k^{\mathrm{G}})} + \boldsymbol{\Lambda}_{q(u_{1i}^{\mathrm{R}}, u_k^{\mathrm{G}})}\right) z_{q^{\mathrm{G}}}\left(\boldsymbol{s}_{x_{\mathrm{unobs},i}}\right) \\ & - \left(\boldsymbol{\mu}_{q(\beta_x)}\,\boldsymbol{\mu}_{q(u_{0i}^{\mathrm{R}})} + \boldsymbol{\Lambda}_{q(\beta_x, u_{0i}^{\mathrm{R}})} + \boldsymbol{\mu}_{q(u_{0i}^{\mathrm{R}})}\,\boldsymbol{\mu}_{q(u_{1i}^{\mathrm{R}})} + \boldsymbol{\Lambda}_{q(u_{0i}^{\mathrm{R}}, u_{1i}^{\mathrm{R}})}\right) \boldsymbol{1}_{n_i^{\mathrm{unobs}}} \Big\} \\ & + \mu_{q(1/\sigma_x^2)}\,\mu_{q(\mu_x)}\,\boldsymbol{1}_{n_i^{\mathrm{unobs}}} + \mu_{q(1/\sigma_w^2)}\,\boldsymbol{w}_{x_{\mathrm{unobs},i}} \Big] \end{aligned}$$

Update $E_{q(\boldsymbol{x}_{\mathbf{unobs}})}(\boldsymbol{C})$ and $E_{q(\boldsymbol{x}_{\mathbf{unobs}})}(\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C})$ according to Subsection 4.2.1.

Update multivariate normal $q^*(\boldsymbol{\beta}, \boldsymbol{u})$ parameters:

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \{\mu_{q(1/\sigma_\varepsilon^2)} E_{q(\boldsymbol{x}_{\mathrm{unobs}})}(\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C}) + \boldsymbol{M}_{q((\boldsymbol{\Omega})^{-1})}\}^{-1}$$

$$\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\{\mu_{q(1/\sigma_\varepsilon^2)} E_{q(\boldsymbol{x}_{\mathrm{unobs}})}(\boldsymbol{C})^{\mathsf{T}}\boldsymbol{y}\}$$

Update univariate normal $q^*(\mu_x)$ parameters:

$$\sigma^2_{q(\mu_x)} \leftarrow 1/\left(\textstyle\sum_{i=1}^m n_i\,\mu_{q(1/\sigma_x^2)} + 1/\sigma^2_{\mu_x}\right)$$

$$\mu_{q(\mu_x)} \leftarrow \sigma^2_{q(\mu_x)}\,\mu_{q(1/\sigma_x^2)}(\boldsymbol{1}^{\mathsf{T}}\boldsymbol{x}_{\mathrm{obs}} + \boldsymbol{1}^{\mathsf{T}}\boldsymbol{\mu}_{q(\boldsymbol{x}_{\mathrm{unobs}})})$$

Update inverse-Gamma $q^*(\sigma_\varepsilon^2)$ parameters:

$$B_{q(\sigma_\varepsilon^2)} \leftarrow \mu_{q(1/a_\varepsilon)} + \tfrac{1}{2}\Big[\|\boldsymbol{y}_{x_{\text{obs}}} - \boldsymbol{C}_{x_{\text{obs}}}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\|^2 + \text{tr}(\boldsymbol{C}_{x_{\text{obs}}}^\mathsf{T}\boldsymbol{C}_{x_{\text{obs}}}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})$$
$$+ \|\boldsymbol{y}_{x_{\text{unobs}}}\|^2 - 2\,\boldsymbol{y}_{x_{\text{unobs}}}^\mathsf{T}\,E_{q(\boldsymbol{x}_{\text{unobs}})}(\boldsymbol{C}_{x_{\text{unobs}}})\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$$
$$+ \text{tr}\left\{E_{q(\boldsymbol{x}_{\text{unobs}})}(\boldsymbol{C}_{x_{\text{unobs}}}^\mathsf{T}\boldsymbol{C}_{x_{\text{unobs}}})(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}^\mathsf{T} + \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})\right\}\Big]$$

$$\mu_{q(1/\sigma_\varepsilon^2)} \leftarrow \tfrac{1}{2}(\textstyle\sum_{i=1}^m n_i + 1)/B_{q(\sigma_\varepsilon^2)} \quad ; \quad \mu_{q(1/a_\varepsilon)} \leftarrow 1/\left(\mu_{q(1/\sigma_\varepsilon^2)} + A_\varepsilon^{-2}\right)$$

Update inverse-Gamma $q^*(\sigma_x^2)$ parameters:

$$B_{q(\sigma_x^2)} \leftarrow \mu_{q(1/a_x)} + \tfrac{1}{2}\Big(\|\boldsymbol{x}_{\text{obs}} - \mu_{q(\mu_x)}\mathbf{1}\|^2 + \|\boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs}})} - \mu_{q(\mu_x)}\mathbf{1}\|^2$$
$$+ \textstyle\sum_{i=1}^m n_i\,\sigma_{q(\mu_x)}^2 + \sum_{i=1}^m n_i^{\text{unobs}}\sigma_{q(\boldsymbol{x}_{\text{unobs},i})}^2\Big)$$

$$\mu_{q(1/\sigma_x^2)} \leftarrow \tfrac{1}{2}(\textstyle\sum_{i=1}^m n_i + 1)/B_{q(\sigma_x^2)} \quad ; \quad \mu_{q(1/a_x)} \leftarrow 1/\left(\mu_{q(1/\sigma_x^2)} + A_x^{-2}\right)$$

Update inverse-Gamma $q^*(\sigma_w^2)$ parameters:

$$B_{q(\sigma_w^2)} \leftarrow \mu_{q(1/a_w)} + \tfrac{1}{2}\Big(\|\boldsymbol{w}_{x_{\text{obs}}} - \boldsymbol{x}_{\text{obs}}\|^2 + \|\boldsymbol{w}_{x_{\text{unobs}}} - \boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs}})}\|^2$$
$$+ \textstyle\sum_{i=1}^m n_i^{\text{unobs}}\sigma_{q(\boldsymbol{x}_{\text{unobs},i})}^2\Big)$$

$$\mu_{q(1/\sigma_w^2)} \leftarrow \tfrac{1}{2}(\textstyle\sum_{i=1}^m n_i + 1)/B_{q(\sigma_w^2)} \quad ; \quad \mu_{q(1/a_w)} \leftarrow 1/\left(\mu_{q(1/\sigma_w^2)} + A_w^{-2}\right)$$

Update inverse-Gamma $q^*(a_r^{\text{R}})$ and inverse-Gamma $q^*(\boldsymbol{\Sigma}^{\text{R}})$ parameters:

For $r = 1, \ldots, q^{\text{R}}$:

$$B_{q(a_r^{\text{R}})} \leftarrow \nu(\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\text{R}})^{-1})})_{rr} + A_{\text{R}r}^{-2} \;\; ; \;\; \mu_{q(1/a_r^{\text{R}})} \leftarrow \tfrac{1}{2}(\nu + q^{\text{R}})/B_{q(a_r^{\text{R}})}$$

$$\boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\text{R}})} \leftarrow \textstyle\sum_{i=1}^m(\boldsymbol{\mu}_{q(\boldsymbol{u}_i^{\text{R}})}\boldsymbol{\mu}_{q(\boldsymbol{u}_i^{\text{R}})}^\mathsf{T} + \boldsymbol{\Sigma}_{q(\boldsymbol{u}_i^{\text{R}})}) + 2\,\nu\,\text{diag}(\mu_{q(1/a_1^{\text{R}})}, \ldots, \mu_{q(1/a_{q^{\text{R}}}^{\text{R}})})$$

$$\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\text{R}})^{-1})} \leftarrow (\nu + m + q^{\text{R}} - 1)\,\boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\text{R}})}^{-1}$$

Update inverse-Gamma $q^*(a_u)$ and inverse-Gamma $q^*(\sigma_u^2)$ parameters:

$$\mu_{q(1/a_u)} \leftarrow 1/(\mu_{q(1/\sigma_u^2)} + A_u^{-2})$$

$$\mu_{q(1/\sigma_u^2)} \leftarrow \frac{q^{\text{G}} + 1}{2\,\mu_{q(1/a_u)} + \|\boldsymbol{\mu}_{q(\boldsymbol{u}^{\text{G}})}\|^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}^{\text{G}})})}$$

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

Algorithm 11: Mean field variational Bayes algorithm for obtaining model parameters in the optimal q-density functions for the two-level Bayesian semiparametric mixed model with Gaussian response subject to classical measurement errors.

$$\sigma_\varepsilon^2|a_\varepsilon \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2}, 1/a_\varepsilon), \quad a_\varepsilon \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2}, 1/A_\varepsilon^2).$$

Again, whilst model (4.7) extends model (2.11) by including the missingness mechanism, its follows the same way to be represented as a mixed model as described in Section 2.3. The prior specification of model (4.1) can be found in Section 2.4. The DAG in Figure 4.2 shows the interplay between the regression parameters and missing data mechanism parameters.

In Section 2.7, we establish that Bayesian models with logistic regression components benefit from the introduction of the variational parameter vector $\boldsymbol{\xi}$ via Jaakkola and Jordan's approximation (Jaakkola and Jordan, 1997). As will become clear in the Appendix 4.B that the MFVB approximation become completely algebraic (without the need of numerical quadrature or Monte Carlo methods) if the variational parameter is introduced into the model.

4.3.1 Notation

The following notation is useful in the upcoming sections. Let n_i^{obs} denote the number of observed x_{ij}s in the ith group and n_i^{unobs} denote the number of missing x_{ij}s in the ith group. Let $\boldsymbol{x}_{\text{obs},i}$ be the $n_i^{\text{obs}} \times 1$ vector containing the observed x_{ij}s and $\boldsymbol{x}_{\text{mis},i}$ be the $n_i^{\text{mis}} \times 1$ containing the missing x_{ij}s. We reorder the data so that the observed data is first for each group. Hence, the full vector of predictor is

$$\boldsymbol{x} \equiv \begin{bmatrix} \boldsymbol{x}_{\text{obs},1} & \boldsymbol{x}_{\text{mis},1} & \cdots & \boldsymbol{x}_{\text{obs},m} & \boldsymbol{x}_{\text{mis},m} \end{bmatrix}^\mathsf{T}.$$

In addition, let $x_{\text{obs},ij}$ be the value of the predictor corresponding to the jth entry of $\boldsymbol{x}_{\text{obs},i}$ and $x_{\text{mis},ij'}$ be the value of the predictor corresponding to the j'th entry of $\boldsymbol{x}_{\text{mis},i}$ for $1 \leq i \leq m$, $1 \leq j \leq n_i^{\text{obs}}$ and $1 \leq j' \leq n_i^{\text{mis}}$.

Define the following vectors and matrices

$$\begin{aligned}
\boldsymbol{x}_i &= \begin{bmatrix} \boldsymbol{x}_{\text{obs},i} \\ \boldsymbol{x}_{\text{unobs},i} \end{bmatrix}, \quad \boldsymbol{x}_{\text{obs},i} = \begin{bmatrix} x_{\text{obs},i1} \\ \vdots \\ x_{\text{obs},in_i^{\text{obs}}} \end{bmatrix}, \quad \boldsymbol{x}_{\text{mis},i} = \begin{bmatrix} x_{\text{mis},i1} \\ \vdots \\ x_{\text{mis},in_i^{\text{mis}}} \end{bmatrix}, \\
\boldsymbol{Z}_{x_{\text{obs}}}^{\text{R}} &= \underset{1\leq i\leq m}{\text{blockdiag}} \left(\begin{bmatrix} \boldsymbol{1} & \boldsymbol{x}_{\text{obs},i} \end{bmatrix} \right), \quad \boldsymbol{Z}_{x_{\text{mis}}}^{\text{R}} = \underset{1\leq i\leq m}{\text{blockdiag}} \left(\begin{bmatrix} \boldsymbol{1} & \boldsymbol{x}_{\text{mis},i} \end{bmatrix} \right), \\
\boldsymbol{Z}_{x_{\text{obs}}}^{\text{G}} &= \begin{bmatrix} z_1(\boldsymbol{s}_{x_{\text{obs},1}}) & \cdots & z_{q^{\text{G}}}(\boldsymbol{s}_{x_{\text{obs},1}}) \\ \vdots & \vdots & \vdots \\ z_1(\boldsymbol{s}_{x_{\text{obs},m}}) & \cdots & z_{q^{\text{G}}}(\boldsymbol{s}_{x_{\text{obs},m}}) \end{bmatrix}, \quad \boldsymbol{Z}_{x_{\text{mis}}}^{\text{G}} = \begin{bmatrix} z_1(\boldsymbol{s}_{x_{\text{mis},1}}) & \cdots & z_{q^{\text{G}}}(\boldsymbol{s}_{x_{\text{mis},1}}) \\ \vdots & \vdots & \vdots \\ z_1(\boldsymbol{s}_{x_{\text{mis},m}}) & \cdots & z_{q^{\text{G}}}(\boldsymbol{s}_{x_{\text{mis},m}}) \end{bmatrix}.
\end{aligned}$$

For conciseness of forthcoming expressions we adopt the following notations:

$$\begin{aligned} \boldsymbol{C} &= \begin{bmatrix} \boldsymbol{C}_{x_{\text{obs}}} \\ \boldsymbol{C}_{x_{\text{mis}}} \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \boldsymbol{x}_{\text{obs}} & \boldsymbol{s}_{x_{\text{obs}}} & \boldsymbol{Z}^{\text{G}}_{x_{\text{obs}}} & \boldsymbol{Z}^{\text{R}}_{x_{\text{obs}}} \\ \mathbf{1} & \boldsymbol{x}_{\text{mis}} & \boldsymbol{s}_{x_{\text{mis}}} & \boldsymbol{Z}^{\text{G}}_{x_{\text{mis}}} & \boldsymbol{Z}^{\text{R}}_{x_{\text{mis}}} \end{bmatrix} \\ E_q(\boldsymbol{C}) &= \begin{bmatrix} \boldsymbol{C}_{x_{\text{obs}}} \\ E_q(\boldsymbol{C}_{x_{\text{mis}}}) \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \boldsymbol{x}_{\text{obs}} & \boldsymbol{s}_{x_{\text{obs}}} & \boldsymbol{Z}^{\text{G}}_{x_{\text{obs}}} & \boldsymbol{Z}^{\text{R}}_{x_{\text{obs}}} \\ \mathbf{1} & E_q(\boldsymbol{x}_{\text{mis}}) & \boldsymbol{s}_{x_{\text{mis}}} & \boldsymbol{Z}^{\text{G}}_{x_{\text{mis}}} & E_q(\boldsymbol{Z}^{\text{R}}_{x_{\text{mis}}}) \end{bmatrix}, \end{aligned}$$

where

$$\begin{aligned} E_q(\boldsymbol{x}_{\text{mis}}) &= \begin{bmatrix} \boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{mis},1})} & \cdots & \boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{mis},m})} \end{bmatrix}^{\mathsf{T}} \\ \text{and} \quad E_q(\boldsymbol{Z}^{\text{R}}_{x_{\text{mis}}}) &= \underset{1 \le i \le m}{\text{blockdiag}} \left(\begin{bmatrix} \mathbf{1} & \boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{mis},i})} \end{bmatrix} \right). \end{aligned}$$

The update expressions for $E_q(\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C})$ are analogous to those in (4.3) except the subscript "unobs" is now replaced by "mis" and, in addition,

$$E_q[\boldsymbol{X}^{\mathsf{T}}\text{diag}\{\lambda(\boldsymbol{\xi})\}\boldsymbol{X}] = \begin{bmatrix} \mathbf{1}^{\mathsf{T}}_{n_{\text{obs}}}\text{diag}\{\lambda(\boldsymbol{\xi}_{x_{\text{obs}}})\}\mathbf{1}_{n_{\text{obs}}} + \mathbf{1}^{\mathsf{T}}_{n_{\text{mis}}}\text{diag}\{\lambda(\boldsymbol{\xi}_{x_{\text{mis}}})\}\mathbf{1}_{n_{\text{mis}}} & \\ \boldsymbol{x}^{\mathsf{T}}_{\text{obs}}\text{diag}\{\lambda(\boldsymbol{\xi}_{x_{\text{obs}}})\}\boldsymbol{x}_{\text{obs}} + \boldsymbol{\mu}^{\mathsf{T}}_{q(\boldsymbol{x}_{\text{mis}})}\text{diag}\{\lambda(\boldsymbol{\xi}_{x_{\text{mis}}})\}\mathbf{1}_{n_{\text{mis}}} & \\ \mathbf{1}^{\mathsf{T}}_{n_{\text{obs}}}\text{diag}\{\lambda(\boldsymbol{\xi}_{x_{\text{obs}}})\}\mathbf{1}_{n_{\text{obs}}} + \mathbf{1}^{\mathsf{T}}_{n_{\text{mis}}}\text{diag}\{\lambda(\boldsymbol{\xi}_{x_{\text{mis}}})\}\boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{mis}})} & \\ \boldsymbol{x}^{\mathsf{T}}_{\text{obs}}\text{diag}\{\lambda(\boldsymbol{\xi}_{x_{\text{obs}}})\}\boldsymbol{x}_{\text{obs}} + E_q[\boldsymbol{x}^{\mathsf{T}}_{\text{mis}}\text{diag}\{\lambda(\boldsymbol{\xi}_{x_{\text{mis}}})\}\boldsymbol{x}_{\text{mis}}] & \end{bmatrix}, \quad (4.8)$$

where

$$\begin{aligned} E_q[\boldsymbol{x}^{\mathsf{T}}_{\text{mis}}\text{diag}\{\lambda(\boldsymbol{\xi}_{x_{\text{mis}}})\}\boldsymbol{x}_{\text{mis}}] &= \boldsymbol{\mu}^{\mathsf{T}}_{q(\boldsymbol{x}_{\text{mis}})}\text{diag}\{\lambda\boldsymbol{\xi}_{x_{\text{mis}}}\}\boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{mis}})} \\ &\quad + \sum_{i=1}^{m}\sum_{j=1}^{n_i}\lambda(\xi_{x_{\text{mis}},ij})\sigma^2_{q(\boldsymbol{x}_{\text{mis},i})}. \end{aligned}$$

4.3.2 Approximate Bayesian inference via mean field variational Bayes

We now provide details on MFVB fitting and inference catered to the Gaussian response model with missing data as described in model (4.7). The essence of our MFVB approach lies in approximating the full conditional joint posterior density function of the form

$$\begin{aligned} p(\boldsymbol{x}_{\text{mis}}, \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{u}, \boldsymbol{a}^{\text{R}}, \boldsymbol{a}_u, a_\varepsilon, \mu_x, \boldsymbol{\Sigma}^{\text{R}}, \boldsymbol{\sigma}^2_u, \sigma^2_\varepsilon, \sigma^2_x, \sigma^2_\phi | \boldsymbol{y}, \boldsymbol{x}_{\text{obs}}, \boldsymbol{R}) &\approx \\ q(\boldsymbol{x}_{\text{mis}}, \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{u}, \boldsymbol{a}^{\text{R}}, \boldsymbol{a}_u, a_\varepsilon, \mu_x, \boldsymbol{\Sigma}^{\text{R}}, \boldsymbol{\sigma}^2_u, \sigma^2_\varepsilon, \sigma^2_x, \sigma^2_\phi) & \end{aligned}$$

subject to the q-density product restriction

$$q(\boldsymbol{x}_{\text{mis}}, \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{u}, \boldsymbol{a}^{\text{R}}, \boldsymbol{a}_u, a_\varepsilon, \mu_x, \boldsymbol{\Sigma}^{\text{R}}, \boldsymbol{\sigma}_u^2, \sigma_\varepsilon^2, \sigma_x^2, \sigma_\phi^2) = q(\boldsymbol{x}_{\text{mis}}, \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{u}, \boldsymbol{a}^{\text{R}}, \boldsymbol{a}_u, a_\varepsilon, \mu_x) \times q(\boldsymbol{\Sigma}^{\text{R}}, \boldsymbol{\sigma}_u^2, \sigma_\varepsilon^2, \sigma_x^2, \sigma_\phi^2). \tag{4.9}$$

Restriction (4.9) is the minimal factorisation in the MFVB approximation, deriving based on the heuristic arguments provided by Menictas and Wand (2013). One can also use the concepts of induced factorisation and moralisation introduced in Subsection 2.5.1 to establish additional product restriction

$$q(\boldsymbol{x}_{\text{mis}})\, q(\boldsymbol{\beta}, \boldsymbol{u})\, q(\boldsymbol{\phi}) \left\{ \prod_{r=1}^{q^{\text{R}}} q(a_r^{\text{R}}) \right\} \left\{ \prod_{\ell=1}^{L} q(a_{u\ell}) \right\} q(a_\varepsilon)\, q(\mu_x)\, q(\boldsymbol{\Sigma}^{\text{R}}) \left\{ \prod_{\ell=1}^{L} q(\sigma_{u\ell}^2) \right\} \times q(\sigma_\varepsilon^2)\, q(\sigma_x^2)\, q(\sigma_\phi^2). \tag{4.10}$$

In Appendix 4.B, we show that the optimal q-densities admit the following forms under the production restriction (4.10):

$$\begin{aligned}
&\boldsymbol{\xi} \leftarrow \sqrt{\text{diagonal}\{E_{q(\boldsymbol{x}_{\text{mis}})}(\boldsymbol{X})(\boldsymbol{\Sigma}_{q(\boldsymbol{\phi})} + \boldsymbol{\mu}_{q(\boldsymbol{\phi})}\boldsymbol{\mu}_{q(\boldsymbol{\phi})}^{\mathsf{T}})E_{q(\boldsymbol{x}_{\text{mis}})}(\boldsymbol{X})^{\mathsf{T}}\}}, \\
&q^*(\boldsymbol{x}_{\text{mis}}) \text{ is the } N(\boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{mis}})}, \sigma^2_{q(\boldsymbol{x}_{\text{mis}})}\boldsymbol{I}_{n_{\text{miss}}}) \text{ density function,} \\
&q^*(\boldsymbol{\beta}, \boldsymbol{u}) \text{ is the } N(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})},\ \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}) \text{ density function,} \\
&q^*(\boldsymbol{\phi}) \text{ is the } N(\boldsymbol{\mu}_{q(\boldsymbol{\phi})},\ \boldsymbol{\Sigma}_{q(\boldsymbol{\phi})}) \text{ density function,} \\
&q^*(\mu_x) \text{ is the } N(\mu_{q(\mu_x)}, \sigma^2_{q(\mu_x)}) \text{ density function} \\
&q^*(\sigma_x^2) \text{ is the Inverse-Gamma } \left(\tfrac{1}{2}\left(\textstyle\sum_{i=1}^m n_i + 1\right),\ B_{q(\sigma_x^2)}\right) \text{ density function,} \\
&q^*(\sigma_\phi^2) \text{ is the Inverse-Gamma } \left(\tfrac{1}{2}\left(\textstyle\sum_{i=1}^m n_i + 1\right),\ B_{q(\sigma_\phi^2)}\right) \text{ density function,} \\
&q^*(\sigma_\varepsilon^2) \text{ is the Inverse-Gamma } \left(\tfrac{1}{2}\left(\textstyle\sum_{i=1}^m n_i + 1\right),\ B_{q(\sigma_\varepsilon^2)}\right) \text{ density function,} \\
&q^*(a_\varepsilon) \text{ is the Inverse-Gamma}(1,\ B_{q(a_\varepsilon)}) \text{ density function,} \\
&q^*(\sigma_{u\ell}^2) \text{ is the Inverse-Gamma } \left(\tfrac{1}{2}\left(q_\ell^{\text{G}} + 1\right),\ B_{q(\sigma_{u\ell}^2)}\right) \text{ density function,} \\
&q^*(a_{u\ell}) \text{ is the Inverse-Gamma}(1,\ B_{q(a_{u\ell})}) \text{ density function,} \\
&q^*(\boldsymbol{\Sigma}^{\text{R}}) \text{ is the Inverse-Wishart } \left(\nu + m + q^{\text{R}} + 1,\ \boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\text{R}})}\right) \text{ density function,} \\
&q^*(a_r^{\text{R}}) \text{ is the Inverse-Gamma } \left(\tfrac{1}{2}(\nu + q^{\text{R}}),\ B_{q(a_r^{\text{R}})}\right) \text{ density function,}
\end{aligned} \tag{4.11}$$

where the parameters are updated according to Algorithm 12, with $\ell = 1$.

Initialise: $\nu = 2$, $\mu_{q(1/\sigma_\varepsilon^2)} > 0$, $\mu_{q(1/a_\varepsilon)} > 0$, $\mu_{q(1/\sigma_x^2)} > 0$, $\mu_{q(1/a_x)} > 0$, $\mu_{q(1/\sigma_w^2)} > 0$, $\mu_{q(1/a_w)} > 0$, $\mu_{q(1/a_r^{\mathrm{R}})} > 0, 1 \leq r \leq q^{\mathrm{R}}$, $\mu_{q(1/\sigma_u^2)} > 0$, $\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})}$, a $q^{\mathrm{R}} \times q^{\mathrm{R}}$ positive definite matrix, $\sigma_{q(\boldsymbol{x}_{\mathrm{mis}})} > 0$, $1 \leq i \leq m$, $\boldsymbol{\mu}_{q(\boldsymbol{x}_{\mathrm{mis},i})}$, a $n_i^{\mathrm{mis}} \times 1$ vector of positive entries, $\boldsymbol{\mu}_{q(\boldsymbol{\phi})}$, a 2×1 vector of positive entries, $\boldsymbol{\Sigma}_{q(\boldsymbol{\phi})}$, a 2×2 positive definite matrix, $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$, a $(p + q^{\mathrm{R}}m + q^{\mathrm{G}}) \times 1$ vector and $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$, a $(p + mq^{\mathrm{R}} + q^{\mathrm{G}}) \times (p + mq^{\mathrm{R}} + q^{\mathrm{G}})$ positive definite matrix.

Cycle through updates:

Define: $\boldsymbol{M}_{q((\boldsymbol{\Omega})^{-1})} \equiv \begin{bmatrix} \sigma_\beta^{-2}\boldsymbol{I}_p & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \mu_{q(1/\sigma_u^2)}\boldsymbol{I}_{q^{\mathrm{G}}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_m \otimes \boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})} \end{bmatrix}$

Update Univariate Normal $q^*(\boldsymbol{x}_{\mathbf{mis},i})$ parameters:

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

If $(n_i^{\mathrm{mis}} > 0)$ then

$$\begin{aligned}
\sigma_{q(\boldsymbol{x}_{\mathrm{mis},i})} \leftarrow 1 \Big/ \Big[& \mu_{q(1/\sigma_\varepsilon^2)} \Big\{ \boldsymbol{\mu}_{q(\beta_x)}^2 + \boldsymbol{\Sigma}_{q(\beta_x)} + 2\left(\boldsymbol{\mu}_{q(\beta_x)}\boldsymbol{\mu}_{q(u_{1i}^{\mathrm{R}})} + \boldsymbol{\Lambda}_{q(\beta_x, u_{1i}^{\mathrm{R}})}\right) \\
& + \boldsymbol{\mu}_{q(u_{1i}^{\mathrm{R}})}^2 + \boldsymbol{\Sigma}_{q(u_{1i}^{\mathrm{R}})} \Big\} + \mu_{q(1/\sigma_x^2)} \\
& + 2(\boldsymbol{\mu}_{q(\phi_1)}^2 + \boldsymbol{\Sigma}_{q(\boldsymbol{\phi}_1)}) \textstyle\sum_{i=1}^m \lambda(\boldsymbol{\xi}_{x_{\mathrm{mis,i}}}) \Big] \\
\boldsymbol{\mu}_{q(\boldsymbol{x}_{\mathrm{mis},i})} \leftarrow \sigma_{q(\boldsymbol{x}_{\mathrm{mis},i})}\boldsymbol{I}_{n_i^{\mathrm{mis}}} \Big[& \mu_{q(1/\sigma_\varepsilon^2)} \Big\{ \left(\boldsymbol{\mu}_{q(\beta_x)} + \boldsymbol{\mu}_{q(u_{1i}^{\mathrm{R}})}\right) \boldsymbol{y}_{x_{\mathrm{mis},i}} \\
& - \left(\boldsymbol{\mu}_{q(\beta_0)}\boldsymbol{\mu}_{q(\beta_x)} + \boldsymbol{\Lambda}_{q(\beta_0,\beta_x)} + \boldsymbol{\mu}_{q(\beta_0)}\boldsymbol{\mu}_{q(u_{1i}^{\mathrm{R}})} + \boldsymbol{\Lambda}_{q(\beta_0, u_{1i}^{\mathrm{R}})}\right) \mathbf{1}_{n_i^{\mathrm{mis}}} \\
& - \left(\boldsymbol{\mu}_{q(\beta_x)}\boldsymbol{\mu}_{q(\beta_s)} + \boldsymbol{\Lambda}_{q(\beta_x,\beta_s)} + \boldsymbol{\mu}_{q(\beta_s)}\boldsymbol{\mu}_{q(u_{1i}^{\mathrm{R}})} + \boldsymbol{\Lambda}_{q(\beta_s, u_{1i}^{\mathrm{R}})}\right) \boldsymbol{s}_{x_{\mathrm{mis},i}} \\
& - \textstyle\sum_{k=1}^{q^{\mathrm{G}}} \left(\boldsymbol{\mu}_{q(\beta_x)}\boldsymbol{\mu}_{q(u_k^{\mathrm{G}})} + \boldsymbol{\Lambda}_{q(\beta_x, u_k^{\mathrm{G}})}\right) z_{q^{\mathrm{G}}}(\boldsymbol{s}_{x_{\mathrm{mis},i}}) \\
& - \textstyle\sum_{k=1}^{q^{\mathrm{G}}} \left(\boldsymbol{\mu}_{q(u_{1i}^{\mathrm{R}})}\boldsymbol{\mu}_{q(u_k^{\mathrm{G}})} + \boldsymbol{\Lambda}_{q(u_{1i}^{\mathrm{R}}, u_k^{\mathrm{G}})}\right) z_{q^{\mathrm{G}}}(\boldsymbol{s}_{x_{\mathrm{mis},i}}) \\
& - \left(\boldsymbol{\mu}_{q(\beta_x)}\boldsymbol{\mu}_{q(u_{0i}^{\mathrm{R}})} + \boldsymbol{\Lambda}_{q(\beta_x, u_{0i}^{\mathrm{R}})} + \boldsymbol{\mu}_{q(u_{0i}^{\mathrm{R}})}\boldsymbol{\mu}_{q(u_{1i}^{\mathrm{R}})} + \boldsymbol{\Lambda}_{q(u_{0i}^{\mathrm{R}}, u_{1i}^{\mathrm{R}})}\right)\mathbf{1}_{n_i^{\mathrm{mis}}} \Big\} \\
& + \mu_{q(1/\sigma_x^2)}\mu_{q(\mu_x)}\mathbf{1}_{n_i^{\mathrm{mis}}} + 2(\boldsymbol{\mu}_{q(\phi_1)}^2 + \boldsymbol{\Sigma}_{q(\boldsymbol{\phi}_1)})\lambda(\boldsymbol{\xi}_{\boldsymbol{x}_{\mathrm{mis},i}}) \Big]
\end{aligned}$$

Update $E_{q(\boldsymbol{x}_{\mathrm{mis}})}(\boldsymbol{C})$ and $E_{q(\boldsymbol{x}_{\mathrm{mis}})}(\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C})$ according to Subsection 4.3.1.

Update multivariate normal $q^*(\boldsymbol{\beta}, \boldsymbol{u})$ parameters:

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \{\mu_{q(1/\sigma_\varepsilon^2)}E_{q(\boldsymbol{x}_{\mathrm{mis}})}(\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C}) + \boldsymbol{M}_{q((\boldsymbol{\Omega})^{-1})}\}^{-1}$$

$$\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\{\mu_{q(1/\sigma_\varepsilon^2)}E_{q(\boldsymbol{x}_{\mathrm{mis}})}(\boldsymbol{C})^{\mathsf{T}}\boldsymbol{y}\}$$

Update univariate normal $q^*(\mu_x)$ parameters:

$$\sigma_{q(\mu_x)}^2 \leftarrow 1/(\textstyle\sum_{i=1}^m n_i\,\mu_{q(1/\sigma_x^2)} + 1/\sigma_{\mu_x}^2)$$

$$\mu_{q(\mu_x)} \leftarrow \sigma_{q(\mu_x)}^2\,\mu_{q(1/\sigma_x^2)}(\mathbf{1}^{\mathsf{T}}\boldsymbol{x}_{\mathrm{obs}} + \mathbf{1}^{\mathsf{T}}\boldsymbol{\mu}_{q(\boldsymbol{x}_{\mathrm{mis}})})$$

Update inverse-Gamma $q^*(\sigma_\varepsilon^2)$ parameters:

$$B_{q(\sigma_\varepsilon^2)} \leftarrow \mu_{q(1/a_\varepsilon)} + \tfrac{1}{2}\Big[\|\boldsymbol{y}_{x_{\text{obs}}} - \boldsymbol{C}_{x_{\text{obs}}}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\|^2 + \text{tr}(\boldsymbol{C}_{x_{\text{obs}}}^\mathsf{T}\boldsymbol{C}_{x_{\text{obs}}}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})$$
$$+ \|\boldsymbol{y}_{x_{\text{mis}}}\|^2 - 2\,\boldsymbol{y}_{x_{\text{mis}}}^\mathsf{T} E_{q(\boldsymbol{x}_{\text{mis}})}(\boldsymbol{C}_{x_{\text{mis}}})\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$$
$$+ \text{tr}\left\{E_{q(\boldsymbol{x}_{\text{mis}})}(\boldsymbol{C}_{x_{\text{mis}}}^\mathsf{T}\boldsymbol{C}_{x_{\text{mis}}})(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}^\mathsf{T} + \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})\right\}\Big]$$

$$\mu_{q(1/\sigma_\varepsilon^2)} \leftarrow \tfrac{1}{2}(\textstyle\sum_{i=1}^m n_i + 1)/B_{q(\sigma_\varepsilon^2)} \quad ; \quad \mu_{q(1/a_\varepsilon)} \leftarrow 1/\left(\mu_{q(1/\sigma_\varepsilon^2)} + A_\varepsilon^{-2}\right)$$

Update inverse-Gamma $q^*(\sigma_x^2)$ parameters:

$$B_{q(\sigma_x^2)} \leftarrow \mu_{q(1/a_x)} + \tfrac{1}{2}\Big(\|\boldsymbol{x}_{\text{obs}} - \mu_{q(\mu_x)}\boldsymbol{1}\|^2 + \|\boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{mis}})} - \mu_{q(\mu_x)}\boldsymbol{1}\|^2$$
$$+ \textstyle\sum_{i=1}^m n_i\,\sigma_{q(\mu_x)}^2 + \sum_{i=1}^m n_i^{\text{mis}}\sigma_{q(\boldsymbol{x}_{\text{mis},i})}^2\Big)$$

$$\mu_{q(1/\sigma_x^2)} \leftarrow \tfrac{1}{2}(\textstyle\sum_{i=1}^m n_i + 1)/B_{q(\sigma_x^2)} \quad ; \quad \mu_{q(1/a_x)} \leftarrow 1/\left(\mu_{q(1/\sigma_x^2)} + A_x^{-2}\right)$$

Update multivariate normal $q^*(\boldsymbol{\phi})$ parameters:

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\phi})} \leftarrow \left[2E_{q(\boldsymbol{x}_{\text{mis}})}(\boldsymbol{X}^\mathsf{T}\text{diag}\{\lambda(\boldsymbol{\xi}_{\boldsymbol{x}_{\text{mis},i}})\}\boldsymbol{X}) + \tfrac{1}{\sigma_\phi^2}\boldsymbol{I}\right]^{-1}$$

$$\boldsymbol{\mu}_{q(\boldsymbol{\phi})} \leftarrow \boldsymbol{\Sigma}_{q(\boldsymbol{\phi})}E_{q(\boldsymbol{x}_{\text{mis}})}(\boldsymbol{X})^\mathsf{T}\left(\boldsymbol{R} - \tfrac{1}{2}\boldsymbol{1}\right)$$

$$\boldsymbol{\xi} \leftarrow \sqrt{\text{diagonal}\{E_{q(\boldsymbol{x}_{\text{mis}})}(\boldsymbol{X})(\boldsymbol{\Sigma}_{q(\boldsymbol{\phi})} + \boldsymbol{\mu}_{q(\boldsymbol{\phi})}\boldsymbol{\mu}_{q(\boldsymbol{\phi})}^\mathsf{T})E_{q(\boldsymbol{x}_{\text{mis}})}(\boldsymbol{X})^\mathsf{T}\}}$$

Update inverse-Gamma $q^*(a_r^{\text{R}})$ and inverse-Gamma $q^*(\boldsymbol{\Sigma}^{\text{R}})$ parameters:

For $r = 1, \ldots, q^{\text{R}}$:

$$B_{q(a_r^{\text{R}})} \leftarrow \nu(\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\text{R}})^{-1})})_{rr} + A_{\text{R}r}^{-2} \quad ; \quad \mu_{q(1/a_r^{\text{R}})} \leftarrow \tfrac{1}{2}(\nu + q^{\text{R}})/B_{q(a_r^{\text{R}})}$$

$$\boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\text{R}})} \leftarrow \textstyle\sum_{i=1}^m(\boldsymbol{\mu}_{q(\boldsymbol{u}_i^{\text{R}})}\boldsymbol{\mu}_{q(\boldsymbol{u}_i^{\text{R}})}^\mathsf{T} + \boldsymbol{\Sigma}_{q(\boldsymbol{u}_i^{\text{R}})}) + 2\,\nu\,\text{diag}(\mu_{q(1/a_1^{\text{R}})}, \ldots, \mu_{q(1/a_{q^{\text{R}}}^{\text{R}})})$$

$$\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\text{R}})^{-1})} \leftarrow (\nu + m + q^{\text{R}} - 1)\,\boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\text{R}})}^{-1}$$

Update inverse-Gamma $q^*(a_u)$ and inverse-Gamma $q^*(\sigma_u^2)$ parameters:

$$\mu_{q(1/a_u)} \leftarrow 1/(\mu_{q(1/\sigma_u^2)} + A_u^{-2})$$

$$\mu_{q(1/\sigma_u^2)} \leftarrow \frac{q^{\text{G}} + 1}{2\,\mu_{q(1/a_u)} + \|\boldsymbol{\mu}_{q(\boldsymbol{u}^{\text{G}})}\|^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}^{\text{G}})})}$$

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

Algorithm 12: Mean field variational Bayes algorithm for obtaining model parameters in the optimal q-density functions for the two-level Bayesian semiparametric mixed model with Gaussian response subject to missing data problems.

4.4 Numerical evaluation

We conducted a series of comprehensive simulation studies to assess the performance of Algorithms 11 and 12 in terms of inferential accuracy, credible interval coverage and computational speed. We generated 30 datasets according to the following simulation settings. All hyperparameters are set to 10 000 and lastly we use $q^{\mathrm{G}} = 25$ interior knots when using the O'Sullivan splines with spacing as described in Wand and Ormerod (2008).

Classical measurement error

For the measurement error model, we consider the following simulation setting:

$$\begin{aligned}
& y_{ij}|\beta_0, \beta_x, u^{\mathrm{R}}_{0i}, u^{\mathrm{R}}_{1i}, \sigma^2_\varepsilon \overset{\text{ind.}}{\sim} N\left(\beta_0 + u^{\mathrm{R}}_{0i} + (\beta_x + u^{\mathrm{R}}_{1i})\, x_{ij} + f(s_{ij}),\, \sigma^2_\varepsilon\right), \\
\text{where} \quad & f(s) = -\sin(2\pi s), \quad x_{ij} \overset{\text{ind.}}{\sim} N(\mu_x, \sigma^2_x); \quad s_{ij} \overset{\text{ind.}}{\sim} \text{Uniform}(0,1), \\
& w_{ij} \overset{\text{ind.}}{\sim} N(x_{ij}, \sigma^2_w), \quad [u^{\mathrm{R}}_{0i} \quad u^{\mathrm{R}}_{1i}]^{\mathsf{T}} \overset{\text{ind.}}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma}^{\mathrm{R}}). \\
\text{True values:} \quad & \beta_0 = 0.58, \quad \beta_x = 1.89, \quad \sigma^2_\varepsilon = 0.04, \quad \boldsymbol{\Sigma}^{\mathrm{R}} = \begin{bmatrix} 2.58 & 0.22 \\ 0.22 & 1.73 \end{bmatrix}, \quad \mu_x = 0.1,
\end{aligned}$$

where $1 \leq i \leq m$ and $1 \leq j \leq n_i$. The proportion of observed x_{ij}s is varied with $p_{\text{obs}} \in \{0.15, 0.25\}$. The number of groups m is 50 and the within-group sample size n_i ranged between 40 and 50.

A commonly used scale-free measure of measurement error is the so-called *reliability ratio* (RR) (e.g. Pham *et al.*, 2013), defined by

$$RR = \frac{\sigma^2_x}{\sigma^2_x + \sigma^2_w}.$$

In this simulation setting we use the values $RR \in \{0.9, 0.8, 0.7\}$ where $RR = 0.9$ corresponds to a small amount of measurement error and $RR = 0.7$ corresponds to a high amount of measurement error. In our examples We set $\sigma^2_x = 0.1$ and subsequently use the reliability ratio to determine σ^2_w.

Missing not at random

For the missing data model, we consider the following simulation setting:

$$\begin{aligned}
& y_{ij}|\beta_0, \beta_x, u^{\mathrm{R}}_{0i}, u^{\mathrm{R}}_{1i}, \sigma^2_\varepsilon \overset{\text{ind.}}{\sim} N\left(\beta_0 + u^{\mathrm{R}}_{0i} + (\beta_x + u^{\mathrm{R}}_{1i})\, x_{ij} + f(s_{ij}),\, \sigma^2_\varepsilon\right), \\
\text{where} \quad & f(s) = -\sin(2\pi s), \quad x_{ij} \overset{\text{ind.}}{\sim} N(\mu_x, \sigma^2_x), \\
& s_{ij} \overset{\text{ind.}}{\sim} \text{Uniform}(0,1); \quad [u^{\mathrm{R}}_{0i} \quad u^{\mathrm{R}}_{1i}]^{\mathsf{T}} \overset{\text{ind.}}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma}^{\mathrm{R}}).
\end{aligned}$$

$$\text{True values:} \quad \beta_0 = 0.58, \quad \beta_x = 1.89, \quad \sigma_\varepsilon^2 = 0.04, \quad \boldsymbol{\Sigma}^{\mathrm{R}} = \begin{bmatrix} 2.58 & 0.22 \\ 0.22 & 1.73 \end{bmatrix},$$
$$\mu_x = 0.1 \quad \text{and} \quad \sigma_\varepsilon^2 \in \{0.05, 0.2, 0.8\},$$

where $1 \leq i \leq m$ and $1 \leq j \leq n_i$. The number of groups m is 50 and the within-group sample size n_i ranged between 40 and 50. In this simulation setting the missingness is controlled by the two pairs of logistic coefficients:

$$\begin{aligned} (\phi_0, \phi_1) &= (2.95, -2.95) \quad \text{and} \\ (\phi_0, \phi_1) &= (0.85, -1.05), \end{aligned}$$

where the probability of missingness increases as a function of the predictor.

4.4.1 Assessment of accuracy

We now focus on an empirical assessment of the accuracy of MFVB algorithms presented in Sections 4.2 and 4.3. For each simulation setting, we fitted each replicated dataset using both the MFVB and MCMC. The MFVB fits were obtained using the respective Algorithms 11 and 12 with the iterations terminated when the relative increase in $\log \underline{p}(\boldsymbol{y}; q)$ fell below 10^{-7}. The MCMC samples were obtained using Stan (Stan Development Team, 2015) with R (R Development Core Team, 2015) interfacing via the `Rstan` package (Stan Development Team, 2015). In each case, MCMC samples of size 10 000 were generated. The first 5000 values of each sample were discarded as burn-in and the remaining 5000 values were thinned by a factor of 5.

As in Chapter 2 we use a measure of accuracy based on the L_1 or integrated absolute error. Recall that the accuracy score, expressed as a percentage, for a particular generic parameter θ is

$$\text{accuracy}(q^*(\theta)) = 100 \left(1 - \tfrac{1}{2} \int_{-\infty}^{\infty} |q^*(\theta) - p_{\mathrm{MCMC}}(\theta|\boldsymbol{y})| d\theta\right) \%.$$

Computation of accuracy$(q^*(\theta))$ is challenging, since it depends on the posterior $p(\theta|\boldsymbol{y})$ that we are trying to avoid by using MFVB methods. Instead, an approximation is made by replacing $p(\theta|\boldsymbol{y})$ with a kernel density estimate $p(\theta|\boldsymbol{y})$ based on a large number of MCMC samples via the R package `KernSmooth` (Wand and Ripley, 2006).

Figure 4.3 displays the side-by-side boxplots of the accuracy scores for the key model parameters and $f(Q_k)$, $1 \leq k \leq 4$, where Q_k are the kth sample quintiles of the s_{ij}s. For all values of RR and p_{obs}, all parameters except σ_ε^2 and σ_w^2 exhibit high accuracy, with almost all accuracy levels above 95%. The accuracy is poor for σ_ε^2 and σ_w^2, with

performance degrading for smaller values of RR and p_{obs}.

Figure 4.4 displays the side-by-side boxplots of the accuracy scores for the key parameters and $f(Q_k)$, $1 \leq k \leq 4$, where Q_k are the kth sample quintiles of the s_{ij}s. The parameters corresponding to the regression part of the model $(\beta_x, \sigma^2_\varepsilon)$ and the sample quintiles of the s_{ij}s show high accuracy across all simulation settings. However, MFVB approximations are generally poor for the missingness mechanism parameters ϕ_0 and ϕ_1 and the accuracy deteriorates as the amount of missing data gets larger or when the data becomes more noisy.

In addition, Figure 4.5 indicates that the MFVB and MCMC penalised splines fits are virtually identical, and when overlaid, are indistinguishable from one another.

4.4.2 Assessment of coverage

Here we also assess the credible interval coverage against the actual coverage from the MFVB q-densities. Table 4.1 shows the percentages of true parameter coverage for the approximate 95% credible intervals formed from the MFVB densities with 0.025 probability mass in each tail, based on 1000 MFVB runs.

For the measurement error models the coverage is generally very good and does not fall below 87% for most parameters and unobservable x_{ij}s. The coverage for the variance parameters σ^2_x and σ^2_w is generally quite poor, but the coverage is better when $p_{\text{obs}} = 0.25$. For the missing data models the coverage is also very good overall except for the missing data mechanism parameters.

4.4.3 Assessment of speed

While running the simulation studies described in Section 4.4 we kept track of the time taken for each model to be fitted and the results are summarised in Table 4.2. We used the Mac OS X laptop with a 2.6 GHz Intel Core i5 processor and 8 GBytes of random access memory. The results indicate that the MFVB approach appears to be between hundreds and thousands times faster than MCMC. However considerations such as computing environment and convergence criterion need to be taken into account to allow a fairer speed comparison.

4.5 Real-time mean field variational Bayes algorithms

With the ongoing advancements in the fields of electronics, computer and engineering sciences, data are not only growing rapidly in volume, but speed as well. Inevitably, this necessitates the design of statistical tools to gain insights into these high volume and velocity data, and support decision making is of key interest.

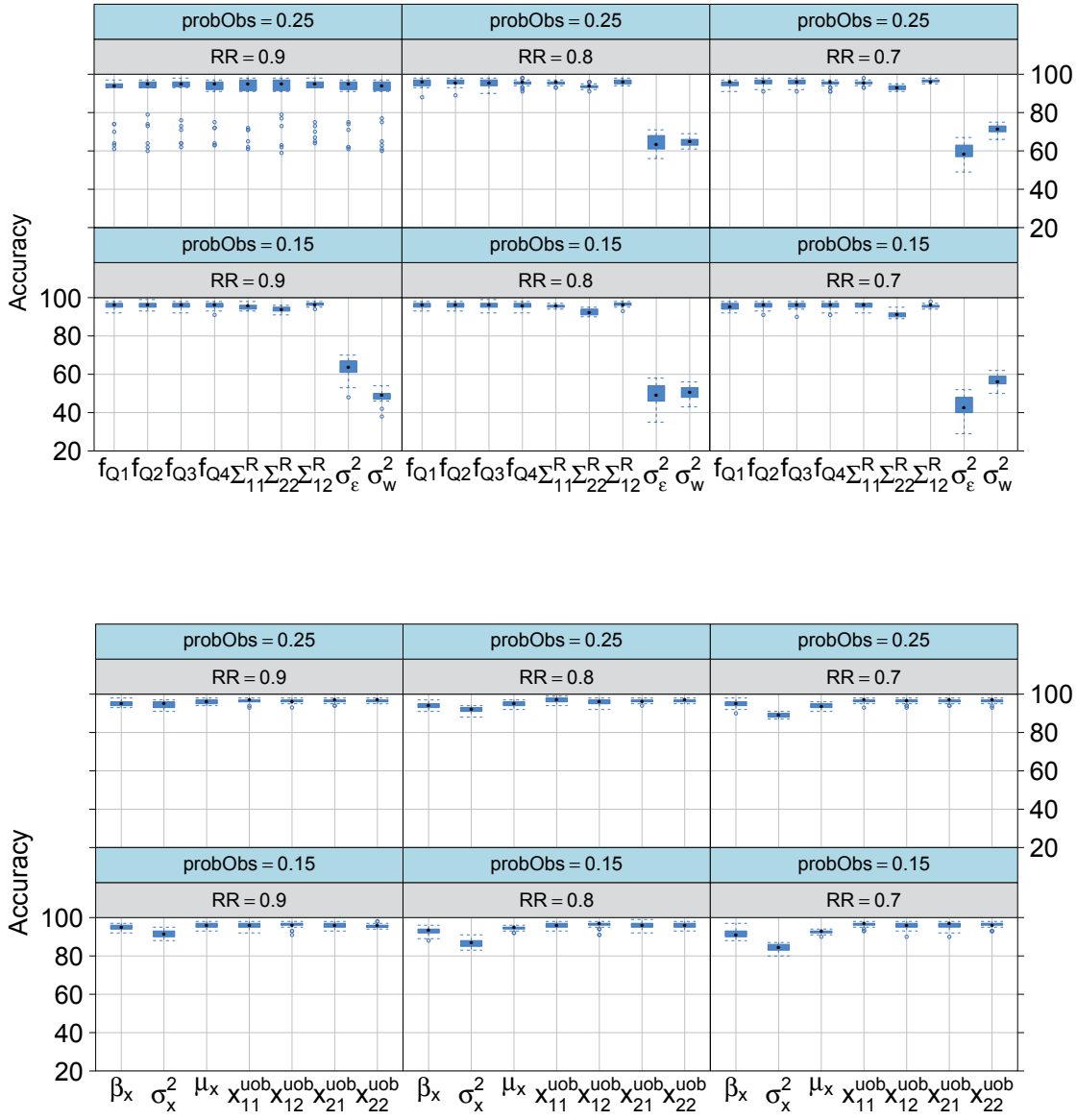

Figure 4.3: Side-by-side boxplots of accuracy scores for the MFVB approximation compared against MCMC over 30 runs for the two-level Bayesian semiparametric mixed model with Gaussian response subject to classical measurement errors.

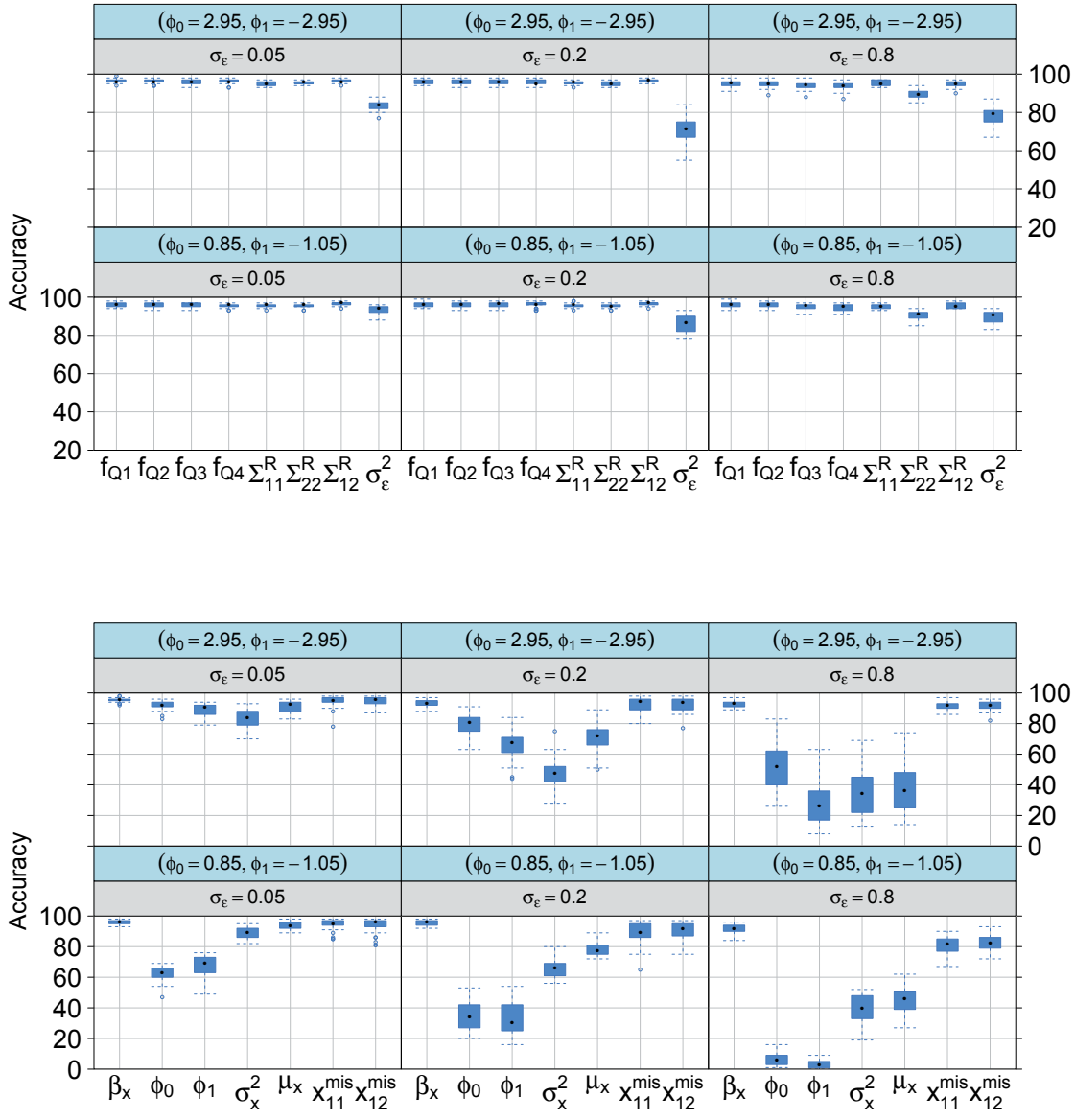

Figure 4.4: Side-by-side boxplots of accuracy scores for the MFVB approximation compared against MCMC over 30 runs for the two-level Bayesian semiparametric mixed model with Gaussian response subject to missingness.

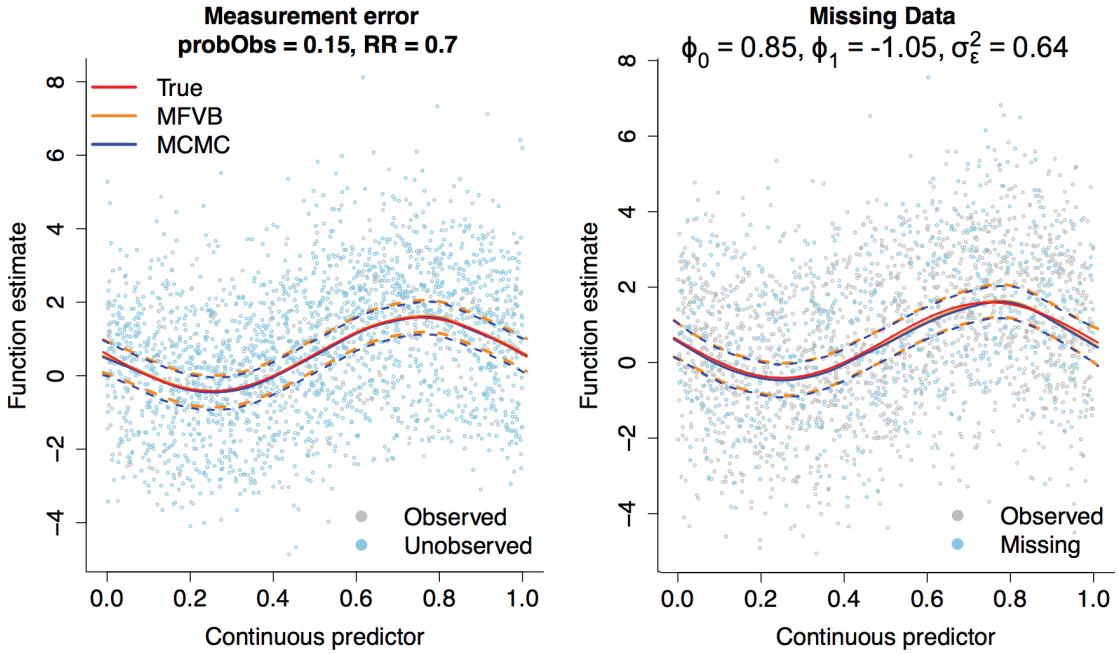

Figure 4.5: Approximate posterior means of the penalised regression functions and corresponding pointwise 95% credible sets obtained via MFVB approximation (orange curves) and MCMC (blue curves) for a single replication of the simulation study described in the text. The red curves represent the true mean functions. The grey circles represent the observed data and the sky blue circles represent the unobserved or missing data.

	ME $p_{\text{obs}} = 0.25$			ME $p_{\text{obs}} = 0.15$				MNAR low miss.			MNAR high miss.		
RR	0.9	0.8	0.7	0.9	0.8	0.7	σ_ε	0.05	0.2	0.8	0.05	0.2	0.8
$f(Q_1)$	97	97	97	97	97	97	$f(Q_1)$	96	96	94	96	96	95
$f(Q_2)$	97	97	97	97	97	97	$f(Q_2)$	96	95	95	96	95	95
$f(Q_3)$	97	97	97	97	96	97	$f(Q_3)$	96	97	97	96	97	97
$f(Q_4)$	96	96	96	95	97	96	$f(Q_4)$	96	96	96	96	96	97
β_x	97	97	97	97	97	98	β_x	98	98	97	98	98	98
$\mathbf{\Sigma}_{11}$	94	94	94	94	94	95	ϕ_0	92	95	67	71	39	0
$\mathbf{\Sigma}_{12}$	95	96	95	96	96	96	ϕ_x	92	84	4	71	32	0
$\mathbf{\Sigma}_{22}$	96	96	96	96	96	96	$\mathbf{\Sigma}_{11}$	95	95	96	95	95	96
σ^2_ε	81	89	65	68	47	35	$\mathbf{\Sigma}_{12}$	93	93	93	94	94	93
σ^2_w	89	82	78	57	54	58	$\mathbf{\Sigma}_{22}$	97	97	95	97	94	96
σ^2_x	88	87	86	86	87	83	σ^2_ε	91	82	88	94	94	95
μ_x	95	94	94	93	93	91	σ^2_x	91	73	46	94	90	65
$x_{\text{unobs},11}$	91	91	94	90	92	94	μ_x	91	88	95	93	94	74
$x_{\text{unobs},21}$	94	95	96	94	96	96	$x_{\text{miss},11}$	97	97	98	96	93	92
–	–	–	–	–	–	–	$x_{\text{miss},12}$	92	91	91	96	91	89

Table 4.1: Percentage coverage of true parameter values by approximate 95% credible sets based on MFVB approximate posterior density functions. A value of $RR = 0.9$ corresponds to a small amount of measurement error and $RR = 0.7$ corresponds to a substantial corruption of the predictor. Low missingness for the MNAR model corresponds to $(\phi_0, \phi_1) = (2.95, -2.95)$ and high missingness to $(\phi_0, \phi_1) = (0.85, -1.05)$. The percentages are based on 1000 replications.

	Measurement error			Missing data		
	MCMC	MFVB	Ratio	MCMC	MFVB	Ratio
Setting 1	5424.84 (936.34)	2.13 (0.13)	2547	46277.54 (7569.84)	4.53 (0.21)	10216
Setting 2	5447.26 (864.55)	2.11 (0.28)	2582	11270.56 (491.04)	3.43 (0.26)	3286
Setting 3	5525.01 (844.71)	2.61 (0.19)	2117	2877.14 (150.42)	2.43 (0.37)	1184
Setting 4	5442.12 (673.22)	2.46 (0.48)	2212	42906.04 (2928.12)	9.36 (0.76)	4584
Setting 5	5611.41 (959.97)	3.05 (0.25)	1840	11110.27 (752.86)	8.70 (0.94)	1277
Setting 6	5548.48 (1050.81)	3.31 (0.32)	1676	2887.62 (157.70)	7.23 (0.65)	399

Table 4.2: Average (standard error) elapsed of the computing times in seconds for the MFVB and MCMC fitting of the two-level Bayesian semiparametric mixed models with Gaussian response subject to classical measurement error or missing data problem.

Up until now all proposed semiparametric regression mixed models assume that the data are processed in *batch*, that is all at the same time. The downsides to such batch processing are that the statistical analysis must wait until the entire dataset has been collected and the necessity of storing a potentially very large dataset in memory. In the *real-time* or *online* scenario, the analysis updates as each new data point (or a subset of new data points) arrive. The real-time updates use only the new data points and summary statistics from previous iterations rather than the full set of available data, thus has advantages of saving a significant amount of memory and avoiding data confidentiality.

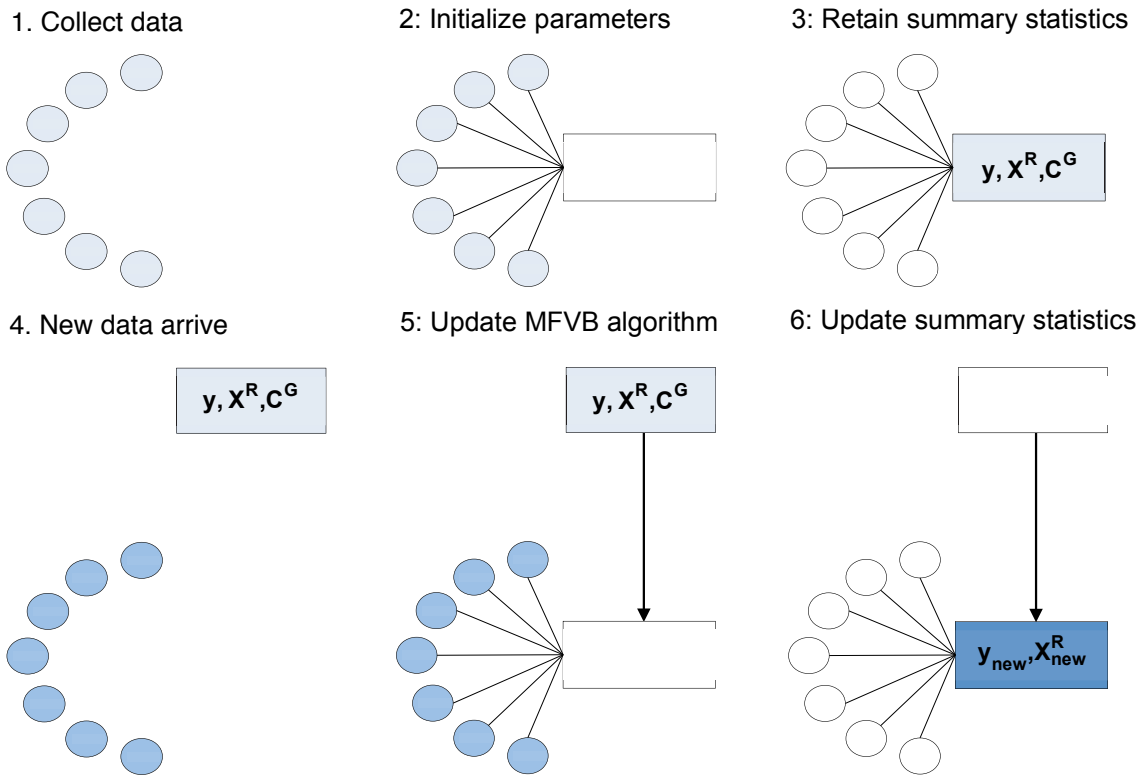

Figure 4.6: Diagrammatic description of the real-time/online variational analysis. The first phase is batch phase, where we start by running a small subset of the initial data in the batch algorithm to initialise model parameters and obtain starting values for the data sufficient statistics (these are all we keep). The second phase is the real-time/online phase where the MFVB algorithm updates as each new data point arrives. Real-time updates use only the new data and summary statistics from previous iterations rather than the full set of available data.

Figure 4.6 is a diagrammatic description of the real-time/online variational analysis.

In the machine learning literature, single-pass algorithms, one iteration per data point or per some small fixed number of data points, have recently been developed for variational inference. A recent work by Hoffman *et al.* (2010) derive a MFVB algorithm for latent Dirichlet allocation with application to topic modelling. Wang *et al.* (2011) extend their work to the hierarchical Dirichlet process. Luts *et al.* (2014) pursue a more classical definition of an "online algorithm" in that each iteration of the MFVB algorithm uses new data and past data only in the form of sufficient statistics. In this section, we utilise the approach of Luts *et al.* (2014) and develop a real-time or online version of Algorithms 3 and 4 for the longitudinal and multilevel data settings with the Gaussian or Bernoulli response. We use Algorithm 3 as the demonstrating example.

The crucial piece that allows extending Algorithm 3 towards real-time or online processing is how it uses the data: it only depends on the data through the response vector

$\boldsymbol{y}$, the design matrices $\boldsymbol{X}^{\mathrm{R}}$ and $\boldsymbol{C}^{\mathrm{G}}$ and the quantities $(\boldsymbol{C}^{\mathrm{G}})^{\mathsf{T}}\boldsymbol{y}$ and $(\boldsymbol{C}^{\mathrm{G}})^{\mathsf{T}}\boldsymbol{C}^{\mathrm{G}}$. These quantities have simple updates when a new response vector of a new group $\boldsymbol{y}_{\mathrm{new}}$ $(n_{\mathrm{new}} \times 1)$ and its corresponding $n_{\mathrm{new}} \times q^{\mathrm{R}}$ matrix of $\boldsymbol{X}^{\mathrm{R}}_{\mathrm{new}}$ and $n_{\mathrm{new}} \times (p + \sum_{\ell=1}^{L} q^{\mathrm{G}}_{\ell})$ matrix of $\boldsymbol{C}^{\mathrm{G}}_{\mathrm{new}}$ arrive. For example, the new $(\boldsymbol{C}^{\mathrm{G}})^{\mathsf{T}}\boldsymbol{y}$ matrix is simply

$$(\boldsymbol{C}^{\mathrm{G}})^{\mathsf{T}}\boldsymbol{y} \leftarrow (\boldsymbol{C}^{\mathrm{G}})^{\mathsf{T}}\boldsymbol{y} + (\boldsymbol{C}^{\mathrm{G}}_{\mathrm{new}})^{\mathsf{T}}\boldsymbol{y}_{\mathrm{new}}.$$

Continuing this way leads to Algorithm 13, the real-time or online modification of Algorithm 3. It should be noted that the starting values for the real-time procedure are determined by performing a sufficiently large batch fit. Section 2.1.1 of Luts *et al.* (2014) explains that good initialisation by the warm-up step can be important for convergence of the real-time approach. For more details regarding the batch-based tuning and convergence diagnosis process, interested readers are encouraged to refer to their paper.

The real-time Algorithm 13 differs from Algorithm 3 in that the data are processed as they arrive and the approximate posterior densities of model parameters are continually updated throughout the data collection process. Thus dynamic graphical displays as depicted in Figure 4.7 could be entertained. Figure 4.7 provides rudimentary illustration of online regression inference for the Gaussian semiparametric mixed model with data simulated according to the setting described in Section 2.10. During the warm-up phase, the number of groups is $m_{\mathrm{warm}} = 20$ with the total number of batch observations of 550.

The real-time or online MFVB Algorithms 13 and 14 are founded upon the same assumption as their batch counterpart. Both batch and real-time algorithms fit the Bayesian semiparametric mixed models, but the latter has the option to perform the fitting in real time for sequentially arriving data. A potential limitation with the proposed method is that the spline basis functions have to be set a priori without having the full dataset available. For example, we need to pre-specify a set of knot positions, and for many practical situations, it is plausible to specify the range of possible values beforehand, such as age. If this assumption is not reasonable, some adjustment needs to be made. In addition, we assume "fixed targets" for the model parameters, in the sense that the parameters remain more or less the same as new data arrive. Extensions to scenarios where the model parameters drift over time are beyond the scope of this thesis and would require further research. Luts (2015) explores similar extensions for the simple Gaussian linear model in the case of distributed dataset.

4.6 Concluding remarks

We derived efficient MFVB algorithms for inference and fitting in Gaussian semiparametric mixed models with classical measurement error and/or missing data problems. The

Batch-based tuning and convergence diagnosis

1. Set m_{warm} to be the warm-up group size and m_{valid} to be the size of the validation period. Read in the first $\sum_{i=1}^{m_{\text{warm}}} n_i + \sum_{i=1}^{m_{\text{valid}}} n_i$ response and predictor values.

2. Create $\boldsymbol{y}_{\text{warm}} \equiv [\boldsymbol{y}_1 \cdots \boldsymbol{y}_{m_{\text{warm}}}]$, $\boldsymbol{X}^{\text{R}}_{\text{warm}} \equiv [\boldsymbol{X}^{\text{R}}_1 \cdots \boldsymbol{X}^{\text{R}}_{m_{\text{warm}}}]$ and $\boldsymbol{C}^{\text{G}}_{\text{warm}} \equiv [\boldsymbol{C}^{\text{G}}_1 \cdots \boldsymbol{C}^{\text{G}}_{m_{\text{warm}}}]$ consisting of the first m_{warm} group of response and predictor values.

3. Feed $\boldsymbol{y}_{\text{warm}}$, $\boldsymbol{X}^{\text{R}}_{\text{warm}}$ and $\boldsymbol{C}^{\text{G}}_{\text{warm}}$ into the batch Algorithm 3 to obtain starting values for $\mu_{q(\sigma^2_\varepsilon)}$, $\mu_{q(a_\varepsilon)}$, $\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\text{R}})^{-1})}$, $\mu_{q(a^{\text{R}}_r)}$ and $\mu_{q(\sigma^2_{u\ell})}$.

4. Set $m \leftarrow m_{\text{warm}}$, $\boldsymbol{n} \equiv [n_1 \cdots n_{m_{\text{warm}}}]$, $(\boldsymbol{C}^{\text{G}})^{\mathsf{T}}\boldsymbol{y} \leftarrow (\boldsymbol{C}^{\text{G}}_{\text{warm}})^{\mathsf{T}}\boldsymbol{y}_{\text{warm}}$ and $(\boldsymbol{C}^{\text{G}})^{\mathsf{T}}\boldsymbol{C}^{\text{G}} \leftarrow (\boldsymbol{C}^{\text{G}}_{\text{warm}})^{\mathsf{T}}\boldsymbol{C}^{\text{G}}_{\text{warm}}$.

5. Run the following online MFVB Algorithm until $m = m_{\text{warm}} + m_{\text{valid}}$.

   Initialise: $\nu = 2$, $\mu_{q(1/\sigma^2_\varepsilon)} > 0$, $\mu_{q(1/a_\varepsilon)} > 0$, $\mu_{q(1/\sigma^2_{u\ell})} > 0$, $\mu_{q(1/a_u)} > 0$, $\mu_{q(1/a^{\text{R}}_r)} > 0$, $1 \le r \le q^{\text{R}}$, $\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\text{R}})^{-1})}$ positive definite.

   Set: $(\boldsymbol{C}^{\text{G}})^{\mathsf{T}}\boldsymbol{y} \leftarrow 0$ and $(\boldsymbol{C}^{\text{G}})^{\mathsf{T}}\boldsymbol{C}^{\text{G}} \leftarrow 0$.

   Cycle:

   Read in $\boldsymbol{y}_{\text{new}}$ $(n_{\text{new}} \times 1)$, $\boldsymbol{X}^{\text{R}}_{\text{new}}$ $(n_{\text{new}} \times q^{\text{R}})$ and $\boldsymbol{C}^{\text{G}}_{\text{new}} \equiv [\boldsymbol{X}_{\text{new}} \quad \boldsymbol{Z}^{\text{G}}_{\text{new}}]$ $\{(n_{\text{new}} \times (p + q^{\text{G}})\}$

   $m \leftarrow m + 1 \quad ; \quad \boldsymbol{n} \leftarrow [\boldsymbol{n} \quad n_{\text{new}}] \quad ; \quad \boldsymbol{y} \leftarrow \left[\boldsymbol{y}^{\mathsf{T}} \; \boldsymbol{y}^{\mathsf{T}}_{\text{new}}\right]^{\mathsf{T}}$

   $\boldsymbol{X}^{\text{R}} \leftarrow \left[(\boldsymbol{X}^{\text{R}})^{\mathsf{T}} \; (\boldsymbol{X}^{\text{R}}_{\text{new}})^{\mathsf{T}}\right]^{\mathsf{T}} \quad ; \quad \boldsymbol{C}^{\text{G}} \leftarrow \left[(\boldsymbol{C}^{\text{G}})^{\mathsf{T}} \; (\boldsymbol{C}^{\text{G}}_{\text{new}})^{\mathsf{T}}\right]^{\mathsf{T}}$

   $(\boldsymbol{C}^{\text{G}})^{\mathsf{T}}\boldsymbol{y} \leftarrow (\boldsymbol{C}^{\text{G}})^{\mathsf{T}}\boldsymbol{y} + (\boldsymbol{C}^{\text{G}}_{\text{new}})^{\mathsf{T}}\boldsymbol{y}_{\text{new}}$

   $(\boldsymbol{C}^{\text{G}})^{\mathsf{T}}\boldsymbol{C}^{\text{G}} \leftarrow (\boldsymbol{C}^{\text{G}})^{\mathsf{T}}\boldsymbol{C}^{\text{G}} + (\boldsymbol{C}^{\text{G}}_{\text{new}})^{\mathsf{T}}\boldsymbol{C}^{\text{G}}_{\text{new}}$

   Repeat update expressions for the batch MFVB Algorithm 3

6. Use convergence diagnostic graphics to assess whether the online parameters are converging to the batch parameters.

   (a) If not converging then return to Step 1 and increase m_{warm}.

   (b) If converging then continue running the online MFVB Algorithm until data no longer available or analysis terminated.

Algorithm 13: Online mean field variational Bayes algorithm for obtaining model parameters in the optimal q-density functions for the two-level Bayesian semiparametric mixed model with Gaussian response.

Batch-based tuning and convergence diagnosis

1. Set m_{warm} to be the warm-up group size and m_{valid} to be the size of the validation period. Read in the first $\sum_{i=1}^{m_{\text{warm}}} n_i + \sum_{i=1}^{m_{\text{valid}}} n_i$ response and predictor values.

2. Create $\boldsymbol{y}_{\text{warm}} \equiv [\boldsymbol{y}_1 \cdots \boldsymbol{y}_{m_{\text{warm}}}]$, $\boldsymbol{\xi}_{\text{warm}} \equiv [\boldsymbol{\xi}_1 \cdots \boldsymbol{\xi}_{m_{\text{warm}}}]$, $\boldsymbol{X}^{\text{R}}_{\text{warm}} \equiv [\boldsymbol{X}^{\text{R}}_1 \cdots \boldsymbol{X}^{\text{R}}_{m_{\text{warm}}}]$ and $\boldsymbol{C}^{\text{G}}_{\text{warm}} \equiv [\boldsymbol{C}^{\text{G}}_1 \cdots \boldsymbol{C}^{\text{G}}_{m_{\text{warm}}}]$ consisting of the first m_{warm} group of response and predictor values.

3. Feed $\boldsymbol{y}_{\text{warm}}$, $\boldsymbol{X}^{\text{R}}_{\text{warm}}$ and $\boldsymbol{C}^{\text{G}}_{\text{warm}}$ into the batch Algorithm 4 to obtain starting values for $\boldsymbol{G}$, $\boldsymbol{H}$, $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\text{G}})}$, $\boldsymbol{\mu}_{q(\boldsymbol{u}^{\text{R}}_i)}$, $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\text{G}})}$, $\boldsymbol{\Sigma}_{q(\boldsymbol{u}^{\text{R}}_i)}$, $\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\text{R}})^{-1})}$, $\mu_{q(a^{\text{R}}_r)}$ and $\mu_{q(\sigma^2_{u\ell})}$.

4. Set $m \leftarrow m_{\text{warm}}$, $\boldsymbol{n} \equiv [n_1 \cdots n_{m_{\text{warm}}}]$ and $(\boldsymbol{C}^{\text{G}})^{\mathsf{T}}\boldsymbol{y} \leftarrow (\boldsymbol{C}^{\text{G}}_{\text{warm}})^{\mathsf{T}}\boldsymbol{y}_{\text{warm}}$.

5. Run the following online MFVB Algorithm until $m = m_{\text{warm}} + m_{\text{valid}}$.

   Initialise: $\nu = 2$, $\mu_{q(1/\sigma^2_{u\ell})} > 0$, $\mu_{q(1/a_u)} > 0$, $\mu_{q(1/a^{\text{R}}_r)} > 0$, $1 \leq r \leq q^{\text{R}}$, $\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\text{R}})^{-1})}$ positive definite, $\boldsymbol{\xi}$, a $(\sum_{i=1}^{m} n_i) \times 1$ vector of positive entries.

   Set: $(\boldsymbol{C}^{\text{G}})^{\mathsf{T}}\boldsymbol{y} \leftarrow 0$.

   Cycle:

   Read in $\boldsymbol{y}_{\text{new}}$ $(n_{\text{new}} \times 1)$, $\boldsymbol{X}^{\text{R}}_{\text{new}}$ $(n_{\text{new}} \times q^{\text{R}})$ and $\boldsymbol{C}^{\text{G}}_{\text{new}} \equiv [\boldsymbol{X}_{\text{new}} \quad \boldsymbol{Z}^{\text{G}}_{\text{new}}]$ $\{(n_{\text{new}} \times (p + q^{\text{G}})\}$)

   $\boldsymbol{\xi}^2_{\text{new}} \leftarrow \text{diagonal}\{\boldsymbol{C}^{\text{G}}_{\text{new}}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\text{G}})} + \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\text{G}})}\boldsymbol{\mu}^T_{q(\boldsymbol{\beta},\boldsymbol{u}^{\text{G}})})(\boldsymbol{C}^{\text{G}}_{\text{new}})^T\}$

   $\boldsymbol{\xi}^2_{\text{new}} \leftarrow \boldsymbol{\xi}^2_{\text{new}} + 2\,\text{diagonal}\{\boldsymbol{C}^{\text{G}}_{\text{new}}(-\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\text{G}})}\boldsymbol{G}_m\boldsymbol{H}_m + \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\text{G}})}\boldsymbol{\mu}^T_{q(\boldsymbol{u}^{\text{R}}_m)})(\boldsymbol{X}^{\text{R}}_{\text{new}})^T\}$

   $\boldsymbol{\xi}^2_{\text{new}} \leftarrow \boldsymbol{\xi}^2_{\text{new}} + \text{diagonal}\{(\boldsymbol{X}^{\text{R}}_{\text{new}}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}^{\text{R}}_m)} + \boldsymbol{\mu}_{q(\boldsymbol{u}^{\text{R}}_m)}\boldsymbol{\mu}^T_{q(\boldsymbol{u}^{\text{R}}_m)})(\boldsymbol{X}^{\text{R}}_{\text{new}})^T\}$

   $m \leftarrow m + 1 \quad ; \quad \boldsymbol{n} \leftarrow [\boldsymbol{n} \quad n_{\text{new}}]$

   $\boldsymbol{y} \leftarrow \left[\boldsymbol{y}^{\mathsf{T}} \; \boldsymbol{y}^{\mathsf{T}}_{\text{new}}\right]^{\mathsf{T}} \quad ; \quad \boldsymbol{\xi} \leftarrow \left[\boldsymbol{\xi}^{\mathsf{T}} \; \boldsymbol{\xi}^{\mathsf{T}}_{\text{new}}\right]^{\mathsf{T}}$

   $\boldsymbol{X}^{\text{R}} \leftarrow \left[(\boldsymbol{X}^{\text{R}})^{\mathsf{T}} \; (\boldsymbol{X}^{\text{R}}_{\text{new}})^{\mathsf{T}}\right]^{\mathsf{T}} \quad ; \quad \boldsymbol{C}^{\text{G}} \leftarrow \left[(\boldsymbol{C}^{\text{G}})^{\mathsf{T}} \; (\boldsymbol{C}^{\text{G}}_{\text{new}})^{\mathsf{T}}\right]^{\mathsf{T}}$

   $(\boldsymbol{C}^{\text{G}})^{\mathsf{T}}\boldsymbol{y} \leftarrow (\boldsymbol{C}^{\text{G}})^{\mathsf{T}}\boldsymbol{y} + (\boldsymbol{C}^{\text{G}}_{\text{new}})^{\mathsf{T}}\boldsymbol{y}_{\text{new}}$

   Repeat update expressions for the batch MFVB Algorithm 4

6. Use convergence diagnostic graphics to assess whether the online parameters are converging to the batch parameters.

   (a) If not converging then return to Step 1 and increase m_{warm}.

   (b) If converging then continue running the online MFVB Algorithm until data no longer available or analysis terminated.

Algorithm 14: Online mean field variational Bayes algorithm for obtaining model parameters in the optimal q-density functions for the two-level Bayesian semiparametric mixed model with Bernoulli response.

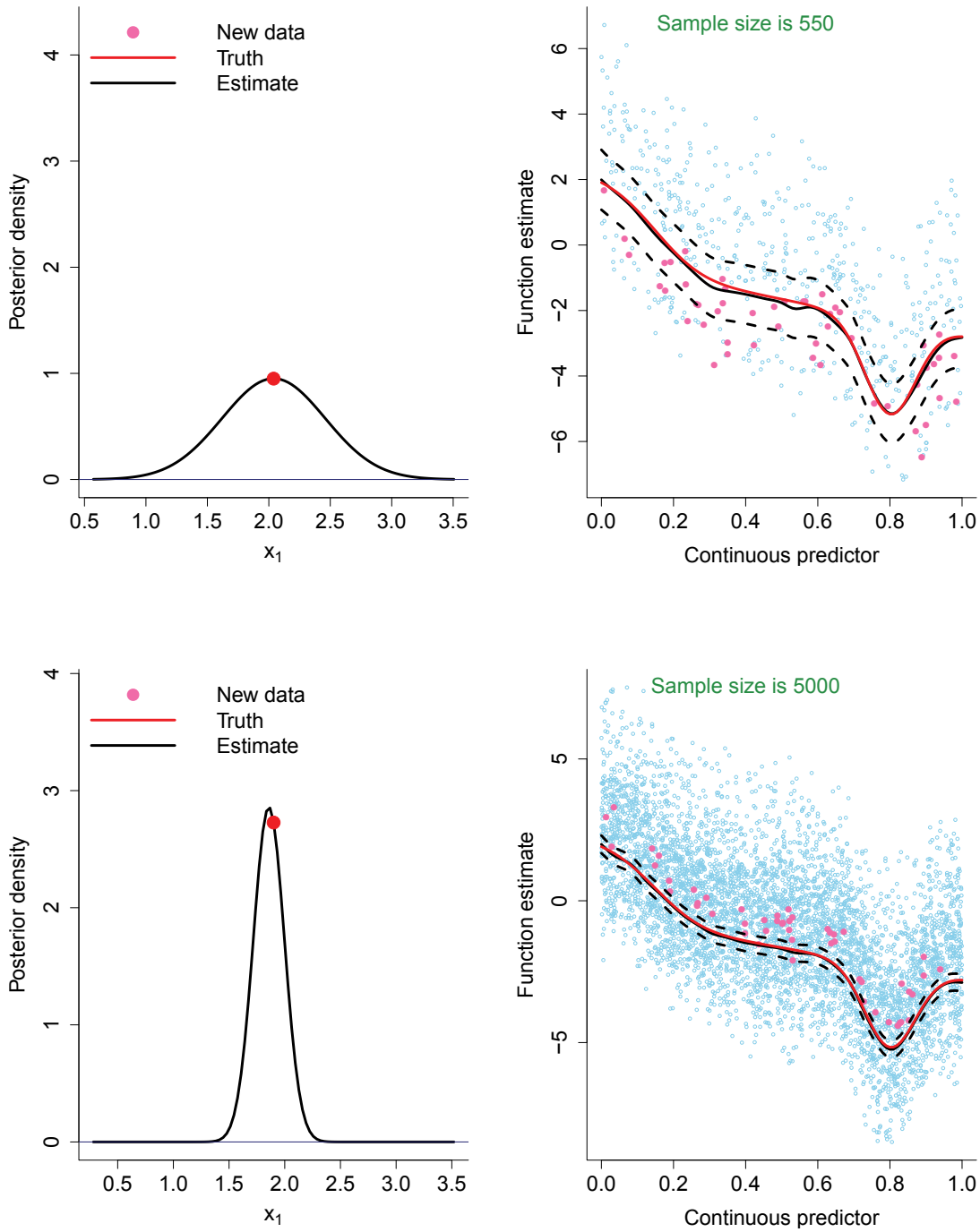

Figure 4.7: Real-time MFVB fitting of the two-level Bayesian semiparametric mixed models with Gaussian response for two different sample sizes. The sky blue circles represent the batch data and the pink circles represent the new data.

MFVB algorithms were shown to have moderately good to excellent accuracy for the main parameters of interest. Poor accuracy is realised for the missing data mechanism parameters, but can potentially be improved via a less stringent, but more computationally demanding mean-field restriction. In addition, the speed comparisons illustrate that MFVB achieves great improvement in efficiency over the MCMC approach for static data. Further, we showed that the advantage of MFVB approaches to approximate inference is their adaptability to real-time or online manner by combining summary statistics instead of the actual data. This aspect is particularly important in the era of high volume and velocity data.

4.A Optimal q-densities derivation for measurement error problems

We now derive Algorithms 11 and 12 concerning MFVB fitting of the Bayesian semiparametric mixed models with data that are subject to either measurement error or missing data problems. Throughout these appendices, we use the notation "const" to denote additive constants with respect to the function argument.

We recall that

$$\begin{aligned} N &\equiv \textstyle\sum_{i=1}^{m} n_i, \quad \boldsymbol{\upsilon} \equiv \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix}, \quad \boldsymbol{C} \equiv \begin{bmatrix} \boldsymbol{X} & \boldsymbol{Z} \end{bmatrix} \\ \text{and} \quad \boldsymbol{\Omega} &\equiv \begin{bmatrix} \sigma_\beta^2 \boldsymbol{I}_p & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_u^2 \boldsymbol{I}_{q^{\mathrm{G}}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_m \otimes \boldsymbol{\Sigma}^{\mathrm{R}} \end{bmatrix}. \end{aligned}$$

We first write out the full model expression

$$y_{ij} = \beta_0 + \beta_x\, x_{ij} + \beta_s\, s_{ij} + \sum_{k=1}^{q^{\mathrm{G}}} u_k^{\mathrm{G}}\, z_k(s_{ij}) + u_{0i}^{\mathrm{R}} + u_{1i}^{\mathrm{R}}\, x_{ij} + \varepsilon_{ij},$$

which can be broken into the observed and unobserved parts:

$$\begin{bmatrix} \boldsymbol{y}_{x_{\text{obs}}} \\ \boldsymbol{y}_{x_{\text{unobs}}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{1}_{n_{\text{obs}}} & \boldsymbol{x}_{\text{obs}} & \boldsymbol{s}_{x_{\text{obs}}} \\ \boldsymbol{1}_{n_{\text{unobs}}} & \boldsymbol{x}_{\text{unobs}} & \boldsymbol{s}_{x_{\text{unobs}}} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_x \\ \beta_s \end{bmatrix} + \begin{bmatrix} \boldsymbol{Z}^{\mathrm{G}}_{x_{\text{obs}}} \\ \boldsymbol{Z}^{\mathrm{G}}_{x_{\text{unobs}}} \end{bmatrix} \begin{bmatrix} u_1^{\mathrm{G}} \\ \vdots \\ u_m^{\mathrm{G}} \end{bmatrix} + \begin{bmatrix} \boldsymbol{Z}^{\mathrm{R}}_{x_{\text{obs}}} \\ \boldsymbol{Z}^{\mathrm{R}}_{x_{\text{unobs}}} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_1^{\mathrm{R}} \\ \vdots \\ \boldsymbol{u}_m^{\mathrm{R}} \end{bmatrix}.$$

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

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

where

$$\begin{aligned} \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} &= \mu_{q(1/\sigma_\varepsilon^2)} \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} E_q(\boldsymbol{C}^{\mathsf{T}}) \boldsymbol{y} \quad \text{and} \\ \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} &= \left(\mu_{q(1/\sigma_\varepsilon^2)}\, E_q(\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C}) + \begin{bmatrix} \sigma_\beta^{-2} \boldsymbol{I}_p & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \mu_{q(1/\sigma_u^2)} \boldsymbol{I}_{q^{\mathrm{G}}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_m \otimes \boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\mathrm{R}})^{-1})} \end{bmatrix} \right)^{-1}. \end{aligned}$$

Derivation:

First we note that

$$\begin{aligned} p(\boldsymbol{\beta}, \boldsymbol{u}|\text{rest}) &\propto p(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{y}, \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{unobs}}, \sigma_\varepsilon^2, \boldsymbol{\Sigma}^{\text{R}}, \boldsymbol{\sigma}_u^2) \\ &\propto p(\boldsymbol{y}|\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{unobs}}, \boldsymbol{\beta}, \boldsymbol{u}^{\text{R}}, \boldsymbol{u}^{\text{G}}, \sigma_\varepsilon^2)\, p(\boldsymbol{\beta})\, p(\boldsymbol{u}^{\text{R}}|\boldsymbol{\Sigma}^{\text{R}})\, p(\boldsymbol{u}^{\text{G}}|\boldsymbol{\sigma}_u^2). \end{aligned}$$

Taking logarithms on both sides gives

$$\begin{aligned} \log p(\boldsymbol{\beta}, \boldsymbol{u}|\text{rest}) &= -\tfrac{1}{2\sigma_\varepsilon^2}(\boldsymbol{y} - \boldsymbol{C}\boldsymbol{v})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{C}\boldsymbol{v}) - \tfrac{1}{2}\boldsymbol{v}^{\mathsf{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{v} + \text{const} \\ &= -\tfrac{1}{2}\left\{\boldsymbol{v}^{\mathsf{T}}(\sigma_\varepsilon^{-2}\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C} + \boldsymbol{\Omega}^{-1})\boldsymbol{v} - 2\sigma_\varepsilon^{-2}\boldsymbol{v}^{\mathsf{T}}\boldsymbol{C}^{\mathsf{T}}\boldsymbol{y}\right\} + \text{const}, \end{aligned}$$

which then leads to the full conditional

$$\begin{bmatrix}\boldsymbol{\beta} \\ \boldsymbol{u}\end{bmatrix} \Bigg| \text{rest} \;\sim\; N\left((\sigma_\varepsilon^{-2}\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C} + \boldsymbol{\Omega}^{-1})^{-1}\sigma_\varepsilon^{-2}\boldsymbol{C}^{\mathsf{T}}\boldsymbol{y},\, (\sigma_\varepsilon^{-2}\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C} + \boldsymbol{\Omega}^{-1})^{-1}\right).$$

Taking expectations with respect to all parameters except $(\boldsymbol{\beta}, \boldsymbol{u})$ gives

$$\begin{aligned} \log q^*(\boldsymbol{\beta}, \boldsymbol{u}) &= E_q\left\{\log p(\boldsymbol{\beta}, \boldsymbol{u}|\text{rest})\right\} + \text{const} \\ &= -\tfrac{1}{2}\left\{\boldsymbol{v}^{\mathsf{T}}\left(\mu_{q(1/\sigma_\varepsilon^2)}E_q(\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C}) + \boldsymbol{M}_{q(\boldsymbol{\Omega}^{-1})}\right)\boldsymbol{v} - 2\sigma_\varepsilon^{-2}\boldsymbol{v}^{\mathsf{T}}E_q(\boldsymbol{C}^{\mathsf{T}})\boldsymbol{y}\right\} + \text{const}, \end{aligned}$$

where $E_q(\boldsymbol{C}^{\mathsf{T}})$ and $E_q(\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C})$ are given in Section 4.2.1 and

$$\boldsymbol{M}_{q(\boldsymbol{\Omega}^{-1})} \begin{bmatrix} \sigma_\beta^{-2}\boldsymbol{I}_p & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \mu_{q(1/\sigma_u^2)}\boldsymbol{I}_{q^{\text{G}}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_m \otimes \boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\text{R}})^{-1})} \end{bmatrix}.$$

The stated results follow via standard "completion of square" manipulations.

Expression for $q^*(\boldsymbol{x}_{\text{unobs},i})$, $\boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs},i})}$ and $\sigma^2_{q(\boldsymbol{x}_{\text{unobs},i})}$

For $i = 1, \ldots, m$

$$q^*(\boldsymbol{x}_{\text{unobs},i}) \sim N(\boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs},i})}, \sigma^2_{q(\boldsymbol{x}_{\text{unobs},i})}\boldsymbol{I}),$$

where

$$\sigma^2_{q(\boldsymbol{x}_{\text{unobs},i})} = E_q\left(\frac{(\beta_x + u_{1i}^{\text{R}})^2}{\sigma_\varepsilon^2} + \frac{1}{\sigma_x^2} + \frac{1}{\sigma_w^2}\right)^{-1}$$

$$\begin{aligned}\boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs},i})} &= \sigma^2_{q(\boldsymbol{x}_{\text{unobs},i})}\boldsymbol{I}_{n_i^{\text{unobs}}}\, E_q\left\{\tfrac{1}{\sigma_\varepsilon^2}(\beta_x + u_{1i}^{\text{R}})\left(\boldsymbol{y}_{x_{\text{unobs},i}} - \beta_0\,\mathbf{1}_{n_i^{\text{unobs}}}\right.\right.\\ &\quad \left.\left. -\beta_s\,\boldsymbol{s}_{x_{\text{unobs},i}} - \textstyle\sum_{k=1}^{q^{\text{G}}} u_k^{\text{G}}\, z_k(\boldsymbol{s}_{x_{\text{unobs},i}}) - u_{0i}^{\text{R}}\,\mathbf{1}_{n_i^{\text{unobs}}}\right)\right\}.\end{aligned}$$

Derivation:
First we note that

$$\begin{aligned}p(\boldsymbol{x}_{\text{unobs}}|\text{rest}) &\propto p(\boldsymbol{x}_{\text{unobs}}|\boldsymbol{x}_{\text{obs}},\boldsymbol{y},\boldsymbol{u}^{\text{R}},\boldsymbol{u}^{\text{G}},\boldsymbol{\beta},\sigma_\varepsilon^2,\boldsymbol{w}_{x_{\text{unobs}}},\sigma_w^2,\mu_x,\sigma_x^2)\\ &\propto p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x}_{\text{obs}},\boldsymbol{x}_{\text{unobs}},\boldsymbol{u}^{\text{R}},\boldsymbol{u}^{\text{G}},\sigma_\varepsilon^2)\,p(\boldsymbol{x}_{\text{unobs}}|\mu_x,\sigma_x^2)\,p(\boldsymbol{w}_{x_{\text{unobs}}}|\boldsymbol{x}_{\text{unobs}},\sigma_w^2).\end{aligned}$$

Taking logarithms on both sides where x is unobserved gives

$$\begin{aligned}\log p(\boldsymbol{x}_{\text{unobs}}|\text{rest}) &= \sum_{i=1}^{m}\left\{-\tfrac{1}{2\sigma_\varepsilon^2}\left\|\boldsymbol{y}_{x_{\text{unobs},i}} - \left(\beta_0\mathbf{1}_{n_i^{\text{unobs}}} + \beta_x\boldsymbol{x}_{\text{unobs},i} + \beta_s\boldsymbol{s}_{x_{\text{unobs},i}}\right.\right.\right.\\ &\quad \left.\left.\left. + \sum_{k=1}^{q^{\text{G}}} u_k^{\text{G}} z_k(\boldsymbol{s}_{x_{\text{unobs},i}}) + u_{0i}^{\text{R}}\mathbf{1}_{n_i^{\text{unobs}}} + u_{1i}^{\text{R}}\boldsymbol{x}_{\text{unobs},i}\right)\right\|^2\right.\\ &\quad \left. - \tfrac{1}{2\sigma_x^2}\left\|\boldsymbol{x}_{\text{unobs},i} - \mu_x\mathbf{1}_{n_i^{\text{unobs}}}\right\|^2 - \tfrac{1}{2\sigma_w^2}\left\|\boldsymbol{w}_{x_{\text{unobs},i}} - \boldsymbol{x}_{\text{unobs},i}\right\|^2\right\} + \text{const}\\ &= \sum_{i=1}^{m}\left[-\tfrac{1}{2\sigma_\varepsilon^2}\left\{(\beta_x + u_{1i}^{\text{R}})^2\|\boldsymbol{x}_{\text{unobs},i}\|^2 - 2\,(\beta_x + u_{1i}^{\text{R}})\,\boldsymbol{x}_{\text{unobs},i}^{\mathsf{T}}\right.\right.\\ &\quad \left.\left.\times\left(\boldsymbol{y}_{x_{\text{unobs},i}} - \beta_0\,\mathbf{1}_{n_i^{\text{unobs}}} - \beta_s\,\boldsymbol{s}_{x_{\text{unobs},i}} - \textstyle\sum_{k=1}^{q^{\text{G}}} u_k^{\text{G}}\, z_k(\boldsymbol{s}_{x_{\text{unobs},i}}) - u_{0i}^{\text{R}}\,\mathbf{1}_{n_i^{\text{unobs}}}\right)\right\}\right.\\ &\quad -\tfrac{1}{2\sigma_x^2}\left(\|\boldsymbol{x}_{\text{unobs},i}\|^2 - 2\,\boldsymbol{x}_{\text{unobs},i}^{\mathsf{T}}\,\mu_x\,\mathbf{1}_{n_i^{\text{unobs}}}\right)\\ &\quad \left.-\tfrac{1}{2\sigma_w^2}\left(\|\boldsymbol{x}_{\text{unobs},i}\|^2 - 2\,\boldsymbol{x}_{\text{unobs},i}^{\mathsf{T}}\boldsymbol{w}_{x_{\text{unobs},i}}\right)\right] + \text{const.}\end{aligned}$$

Taking expectations with respect to all parameters except $\boldsymbol{x}_{\text{unobs},i}$ gives

$$\begin{aligned}E_q\{\log p(\boldsymbol{x}_{\text{unobs},i}|\text{rest})\} &= -\frac{1}{2}\left[\boldsymbol{x}_{\text{unobs},i}^{\mathsf{T}}\,E_q\left(\frac{(\beta_x + u_{1i}^{\text{R}})^2}{\sigma_\varepsilon^2} + \frac{1}{\sigma_x^2} + \frac{1}{\sigma_w^2}\right)\boldsymbol{x}_{\text{unobs},i}\right.\\ &\quad - 2\,\boldsymbol{x}_{\text{unobs},i}^{\mathsf{T}}\left\{E_q\left(\tfrac{1}{\sigma_\varepsilon^2}(\beta_x + u_{1i}^{\text{R}})\left(\boldsymbol{y}_{x_{\text{unobs},i}} - \beta_0\,\mathbf{1}_{n_i^{\text{unobs}}}\right.\right.\right.\\ &\quad \left.\left.\left.-\beta_s\,\boldsymbol{s}_{x_{\text{unobs},i}} - \textstyle\sum_{k=1}^{q^{\text{G}}} u_k^{\text{G}}\, z_k(\boldsymbol{s}_{x_{\text{unobs},i}}) - u_{0i}^{\text{R}}\right)\right)\\ &\quad \left.+\frac{\mu_x\,\mathbf{1}_{n_i^{\text{unobs}}}}{\sigma_x^2} + \frac{\boldsymbol{w}_{x_{\text{unobs},i}}}{\sigma_w^2}\right\}\right] + \text{const},\end{aligned}$$

where

$$E_q\left(\frac{(\beta_x + u_{1i}^{\text{R}})^2}{\sigma_\varepsilon^2} + \frac{1}{\sigma_x^2} + \frac{1}{\sigma_w^2}\right)$$

$$
\begin{aligned}
&= \Big[\mu_{q(1/\sigma_\varepsilon^2)}\Big\{\boldsymbol{\mu}^2_{q(\beta_x)} + (\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})_{22} + 2\left(\boldsymbol{\mu}_{q(\beta_x)}\,\boldsymbol{\mu}_{q(\boldsymbol{u}^{\mathrm{R}}_{1i})} + \boldsymbol{\Lambda}_{q(\beta_x,\,u^{\mathrm{R}}_{1i})}\right)\Big\} \\
&\quad + \boldsymbol{\mu}^2_{q(u^{\mathrm{R}}_{1i})} + \boldsymbol{\Sigma}_{q(u^{\mathrm{R}}_{1i})} + \mu_{q(1/\sigma_x^2)} + \mu_{q(1/\sigma_w^2)}\Big] \\
\boldsymbol{\mu}_{q(\boldsymbol{x}_{\mathrm{unobs},i})} &= \sigma^2_{q(\boldsymbol{x}_{\mathrm{unobs},i})}\boldsymbol{I}_{n_i^{\mathrm{unobs}}}\Big[\mu_{q(1/\sigma_\varepsilon^2)}\Big\{\left(\boldsymbol{\mu}_{q(\beta_x)} + \boldsymbol{\mu}_{q(\boldsymbol{u}^{\mathrm{R}}_{1i})}\right)\boldsymbol{y}_{x_{\mathrm{unobs},i}} \\
&\quad - \left(\boldsymbol{\mu}_{q(\beta_x)}\,\boldsymbol{\mu}_{q(\beta_0)} + \boldsymbol{\Lambda}_{q(\beta_0,\beta_x)} + \boldsymbol{\mu}_{q(\boldsymbol{u}^{\mathrm{R}}_{1i})}\,\boldsymbol{\mu}_{q(\beta_0)} + \boldsymbol{\Lambda}_{q(u^{\mathrm{R}}_{1i},\,\beta_0)}\right)\mathbf{1}_{n_i^{\mathrm{unobs}}} \\
&\quad - \left(\boldsymbol{\mu}_{q(\beta_x)}\,\boldsymbol{\mu}_{q(\beta_s)} + \boldsymbol{\Lambda}_{q(\beta_x,\beta_s)} + \boldsymbol{\mu}_{q(u^{\mathrm{R}}_{1i})}\,\boldsymbol{\mu}_{q(\beta_s)} + \boldsymbol{\Lambda}_{q(u^{\mathrm{R}}_{1i},\,\beta_s)}\right)\boldsymbol{s}_{x_{\mathrm{unobs},i}} \\
&\quad - \textstyle\sum_{k=1}^{q^{\mathrm{G}}}\left(\boldsymbol{\mu}_{q(\beta_x)}\,\boldsymbol{\mu}_{q(u^{\mathrm{G}}_k)} + \boldsymbol{\Lambda}_{q(\beta_x,\,u^{\mathrm{G}}_k)}\right)z_k(\boldsymbol{s}_{x_{\mathrm{unobs},i}}) \\
&\quad - \textstyle\sum_{k=1}^{q^{\mathrm{G}}}\left(\boldsymbol{\mu}_{q(u^{\mathrm{R}}_{1i})}\,\boldsymbol{\mu}_{q(u^{\mathrm{G}}_k)} + \boldsymbol{\Lambda}_{q(u^{\mathrm{R}}_{1i},\,u^{\mathrm{G}}_k)}\right)z_k(\boldsymbol{s}_{x_{\mathrm{unobs},i}}) \\
&\quad - \left(\boldsymbol{\mu}_{q(\beta_x)}\boldsymbol{\mu}_{q(u^{\mathrm{R}}_{0i})} + \boldsymbol{\Lambda}_{q(\beta_x,\,\boldsymbol{u}^{\mathrm{R}}_{0i})} + \mu_{q(u^{\mathrm{R}}_{1i})}\,\boldsymbol{\mu}_{q(u^{\mathrm{R}}_{0i})} + \boldsymbol{\Lambda}_{q(\boldsymbol{u}^{\mathrm{R}}_{0i},\boldsymbol{u}^{\mathrm{R}}_{1i})}\mathbf{1}_{n_i^{\mathrm{unobs}}}\Big\}\right. \\
&\quad \left. + \mu_{q(1/\sigma_x^2)}\,\mu_{q(\mu_x)}\,\mathbf{1}_{n_i^{\mathrm{unobs}}} + \mu_{q(1/\sigma_w^2)}\,\boldsymbol{w}_{x_{\mathrm{unobs},i}}\right].
\end{aligned}
$$

Expressions for $q^*(\sigma_\varepsilon^2), B_{q(\sigma_\varepsilon^2)}$ and $\mu_{q(1/\sigma_\varepsilon^2)}$

$$
q^*(\sigma_\varepsilon^2) \sim \text{Inverse-Gamma}\left(\tfrac{1}{2}\left(\textstyle\sum_{i=1}^m n_i + 1\right),\, B_{q(\sigma_\varepsilon^2)}\right),
$$

where

$$
B_{q(\sigma_\varepsilon^2)} = \tfrac{1}{2}\,E_q(\|\boldsymbol{y} - \boldsymbol{C}\boldsymbol{v}\|^2) + \mu_{q(1/\sigma_\varepsilon^2)} \quad \text{and} \quad \mu_{q(\sigma_\varepsilon^2)} = \tfrac{1}{2}\,\left(\textstyle\sum_{i=1}^m n_i + 1\right)/B_{q(\sigma_\varepsilon^2)}
$$

Derivation:
First we note that

$$
\begin{aligned}
p(\sigma_\varepsilon^2|\text{rest}) &\propto p(\sigma_\varepsilon^2|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{x}_{\mathrm{unobs}} a_\varepsilon) \\
&\propto p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{x}_{\mathrm{unobs}}, \sigma_\varepsilon^2)\; p(\sigma_\varepsilon^2|a_\varepsilon).
\end{aligned}
$$

Taking logarithms on both sides gives

$$
\log p(\sigma_\varepsilon^2|\text{rest}) = \left\{-\tfrac{1}{2}\left(\textstyle\sum_{i=1}^m n_i + 1\right) - 1\right\}\log(\sigma_\varepsilon^2) - \left(\tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{C}\boldsymbol{v}\|^2 + a_\varepsilon^{-1}\right)/\sigma_\varepsilon^2 + \text{const}
$$

which then leads to the full conditional

$$
\sigma_\varepsilon^2|\text{rest} \sim \text{Inverse-Gamma}\left(\tfrac{1}{2}(\textstyle\sum_{i=1}^m n_i + 1),\, \tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{C}\boldsymbol{v}\|^2 + a_\varepsilon^{-1}\right).
$$

Taking expectations with respect to all parameters except σ_ε^2 gives

$$\begin{aligned}\log q^*(\sigma_\varepsilon^2) &= E_q\left\{\log p(\sigma_\varepsilon^2|\text{rest})\right\} + \text{const} \\ &= \left\{-\tfrac{1}{2}(\textstyle\sum_{i=1}^m n_i + 1) - 1\right\}\log(\sigma_\varepsilon^2) - \left[\tfrac{1}{2}\left\{E_q(\|\boldsymbol{y} - \boldsymbol{C}\boldsymbol{v}\|^2)\right\} + \mu_{q(1/a_\varepsilon)}\right]/\sigma_\varepsilon^2 + \text{const.}\end{aligned}$$

The form of $q^*(\sigma_\varepsilon^2)$ and $B_{q(\sigma_\varepsilon^2)}$ follows from Result 1.11.6 where

$$\begin{aligned}E_q(\|\boldsymbol{y} - \boldsymbol{C}\boldsymbol{v}\|^2) &= \|\boldsymbol{y}_{x_{\text{obs}}} - \boldsymbol{C}\boldsymbol{x}_{\text{obs}}\boldsymbol{\mu}_{q(\boldsymbol{v})}\| + \text{tr}(\boldsymbol{C}_{x_{\text{obs}}}^\mathsf{T}\boldsymbol{C}_{x_{\text{obs}}}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}) \\ &\quad + \|\boldsymbol{y}_{x_{\text{unobs}}}\|^2 - 2\boldsymbol{y}^\mathsf{T}\boldsymbol{x}_{\text{unobs}}E_q(\boldsymbol{C}_{x_{\text{unobs}}})\boldsymbol{\mu}_{q(\boldsymbol{v})} \\ &\quad + \text{tr}\{E_q(\boldsymbol{C}_{x_{\text{unobs}}}^\mathsf{T}\boldsymbol{C}_{x_{\text{unobs}}})(\boldsymbol{\mu}_{q(\boldsymbol{v})}\boldsymbol{\mu}_{q(\boldsymbol{v})}^\mathsf{T} + \boldsymbol{\Sigma}_{q(\boldsymbol{v})})\}.\end{aligned}$$

Expressions for $q^*(a_\varepsilon), B_{q(a_\varepsilon)}$ and $\mu_{q(1/a_\varepsilon)}$

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

where

$$B_{q(a_\varepsilon)} = \mu_{q(1/\sigma_\varepsilon^2)} + A_\varepsilon^{-2} \quad \text{and} \quad \mu_{q(1/a_\varepsilon)} = 1/\left\{\mu_{q(1/\sigma_\varepsilon^2)} + A_\varepsilon^{-2}\right\}.$$

Derivation:
First we note that

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

Taking logarithms on both sides gives

$$\log p(a_\varepsilon|\text{rest}) = -2\log(a_\varepsilon) - \left(\sigma_\varepsilon^{-2} + A_\varepsilon^{-2}\right)/a_\varepsilon + \text{const},$$

which then leads to the full conditional

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

Taking expectations with respect to all parameters except a_ε gives

$$\begin{aligned}\log q^*(a_\varepsilon) &= -2\log(a_\varepsilon) - E_q(\sigma_\varepsilon^{-2} + A_\varepsilon^{-2})/a_\varepsilon + \text{const} \\ &= (-1-1)\log(a_\varepsilon) - \left(\mu_{q(1/\sigma_\varepsilon^2)} + A_\varepsilon^{-2}\right)/a_\varepsilon + \text{const.}\end{aligned}$$

The expressions for $B_{q(a_\varepsilon)}$ and $\mu_{q(1/a_\varepsilon)}$ follow immediately.

Expressions for $q^*(\sigma_w^2), B_{q(\sigma_w^2)}$ and $\mu_{q(1/\sigma_w^2)}$

$$q^*(\sigma_w^2) \sim \text{Inverse-Gamma}\left(\tfrac{1}{2}\left(\textstyle\sum_{i=1}^m n_i + 1\right),\, B_{q(\sigma_w^2)}\right),$$

where

$$B_{q(\sigma_w^2)} = \tfrac{1}{2}\, E_q(\|\boldsymbol{w} - \boldsymbol{x}\|^2) + \mu_{q(1/\sigma_w^2)} \quad \text{and} \quad \mu_{q(\sigma_w^2)} = \tfrac{1}{2}\, \left(\textstyle\sum_{i=1}^m n_i + 1\right) / B_{q(\sigma_w^2)}$$

and

$$\begin{aligned} E_q(\|\boldsymbol{w} - \boldsymbol{x}\|^2) &= \|\boldsymbol{w}_{x_{\text{obs}}} - \boldsymbol{x}_{\text{obs}}\|^2 + \|\boldsymbol{w}_{x_{\text{unobs}}}\|^2 - 2\boldsymbol{w}_{x_{\text{unobs}}}^{\mathsf{T}} \boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs}})} \\ &\quad + \|\boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs}})}\|^2 + \text{tr}(\text{Cov}_q(\boldsymbol{x}_{\text{unobs}})) \\ &= \|\boldsymbol{w}_{x_{\text{unobs}}} - \boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs}})}\|^2 + \sum_{i=1}^m n_i^{\text{unobs}} \sigma^2_{q(\boldsymbol{x}_{\text{unobs},i})}. \end{aligned}$$

Derivation:

The derivation is analogous to that of $q^*(\sigma_\varepsilon^2)$ and related quantities.

Expressions for $q^*(a_w), B_{q(a_w)}$ and $\mu_{q(1/a_w)}$

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

where

$$B_{q(a_w)} = \mu_{q(1/\sigma_w^2)} + A_w^{-2} \quad \text{and} \quad \mu_{q(1/a_w)} = 1/\left\{\mu_{q(1/\sigma_w^2)} + A_w^{-2}\right\}.$$

Derivation:

The derivation is analogous to that of $q^*(a_\varepsilon)$ and related quantities.

Expressions for $q^*(\mu_x), B_{q(\mu_x)}$ and $\mu_{q(1/\mu_x)}$

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

where

$$\mu_{q(\mu_x)} = \sigma^2_{q(\mu_x)}\mu_{q(1/\sigma^2_x)}(\mathbf{1}^\mathsf{T}\boldsymbol{x}_{\text{obs}} + \mathbf{1}^\mathsf{T}\boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs}})}) \quad \text{and} \quad \sigma^2_{q(\mu_x)} = 1/\{N\mu_{q(1/\sigma^2_x)} + 1/\sigma^2_{\mu_x}\}.$$

Derivation:
First we note that

$$p(\mu_x|\text{rest}) \quad \propto \quad p(\mu_x|\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{unobs}}, \sigma^2_x) \propto p(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{unobs}}|\mu_x, \sigma^2_x)\, p(\mu_x).$$

Taking logarithms on both sides gives

$$\begin{aligned} \log p(\mu_x|\text{rest}) &= -\tfrac{1}{2\sigma^2_x}\,\|\boldsymbol{x} - \mu_x\mathbf{1}\|^2 - \tfrac{1}{2\sigma^2_{\mu_x}}\mu^2_x \\ &= -\tfrac{1}{2}\left\{\left(\tfrac{N}{\sigma^2_x} + \tfrac{1}{\sigma^2_{\mu_x}}\right)\mu^2_x - 2\tfrac{1}{\sigma^2_x}\mu_x\mathbf{1}^\mathsf{T}\boldsymbol{x}\right\}. \end{aligned}$$

Taking expectations with respect to all parameters except μ_x gives

$$\log p(\mu_x|\text{rest}) \quad = \quad -\tfrac{1}{2}\left\{\left(N\mu_{q(1/\sigma^2_x)} + 1/\sigma^2_{\mu_x}\right)\mu^2_x - 2\mu_{q(1/\sigma^2_x)}\mu_x E_q(\mathbf{1}^\mathsf{T}\boldsymbol{x})\right\}.$$

The form of $q^*(\sigma^2_\varepsilon)$ and $B_{q(\sigma^2_\varepsilon)}$ follows where

$$E_q(\mathbf{1}^\mathsf{T}\boldsymbol{x}) = \mathbf{1}^\mathsf{T}\boldsymbol{x}_{\text{obs}} + \mathbf{1}^\mathsf{T}\boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs}})}.$$

Expressions for $q^*(\sigma^2_x), B_{q(\sigma^2_x)}$ and $\mu_{q(1/\sigma^2_x)}$

$$q^*(\sigma^2_x) \sim \text{Inverse-Gamma}\left(\tfrac{1}{2}\left(\textstyle\sum_{i=1}^m n_i + 1\right),\, B_{q(\sigma^2_x)}\right),$$

where

$$B_{q(\sigma^2_x)} = \tfrac{1}{2}\,E_q(\|\boldsymbol{x} - \mu_x\mathbf{1}\|^2) + \mu_{q(1/\sigma^2_x)} \quad \text{and} \quad \mu_{q(\sigma^2_x)} = \tfrac{1}{2}\,\left(\textstyle\sum_{i=1}^m n_i + 1\right)/B_{q(\sigma^2_\varepsilon)}$$

and

$$\begin{aligned} E_q(\|\boldsymbol{x} - \mu_x\mathbf{1}\|^2) &= \|\boldsymbol{x}_{\text{obs}} - \mu_{q(\mu_x)}\mathbf{1}\|^2 + \|\boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs}})} - \mu_{q(\mu_x)}\mathbf{1}\|^2 \\ &\quad + N\sigma^2_{q(\mu_x)} + \sum_{i=1}^m n_i^{\text{unobs}}\sigma^2_{q(\boldsymbol{x}_{\text{unobs},i})}. \end{aligned}$$

Derivation:

The derivation is analogous to that of $q^*(\sigma_\varepsilon^2)$ and related quantities.

The remaining q^* densities involve straightforward adaptation of the derivations given in Appendix 2.A.

4.A.1 Derivation of the marginal log-likelihood lower bound

The expression for the marginal log-likelihood lower bound is

$$\begin{aligned}
\log \underline{p}(\boldsymbol{y}, \boldsymbol{x}_{\text{obs}}, \boldsymbol{w}; q) \;=\; & -\tfrac{1}{2}\sigma_\beta^{-2}\left(\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\right) + \tfrac{1}{2}\left(\textstyle\sum_{\ell=1}^{L} q_\ell^{\text{G}} + p + m\right) \\
& + \tfrac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}| - \log(\mathcal{C}_{q^{\text{R}},\,\nu+q^{\text{R}}-1}) + \log(\mathcal{C}_{q^{\text{R}},\,\nu+m+q^{\text{R}}-1}) \\
& - \tfrac{1}{2}\left(\nu + m + q^{\text{R}} - 1\right)\log|\boldsymbol{B}_{q(\boldsymbol{\Sigma}^{\text{R}})}| - \tfrac{1}{2}\textstyle\sum_{i=1}^{m} n_i^{\text{unobs}} \log(\sigma^2_{q(\boldsymbol{x}_{\text{unobs},i})}) \\
& - \tfrac{1}{2}\textstyle\sum_{\ell=1}^{L}\left(q_\ell^{\text{G}} + 1\right)\log(B_{q(\sigma^2_{u\ell})}) - \tfrac{1}{2}\left(\textstyle\sum_{i=1}^{m} n_i + 1\right)\log\left(B_{q(\sigma_\varepsilon^2)}\right) \\
& - \tfrac{1}{2}\left(\textstyle\sum_{i=1}^{m} n_i + 1\right)\log\left(B_{q(\sigma_x^2)}\right) - \tfrac{1}{2}\left(\textstyle\sum_{i=1}^{m} n_i + 1\right)\log\left(B_{q(\sigma_w^2)}\right) \\
& - \textstyle\sum_{\ell=1}^{L}\log\left(B_{q(a_{u\ell})}\right) + \sum_{\ell=1}^{L}\mu_{q(1/a_{u\ell})}\mu_{q(1/\sigma^2_{u\ell})} \\
& + \textstyle\sum_{r=1}^{q^{\text{R}}}\nu(\boldsymbol{M}_{q((\boldsymbol{\Sigma}^{\text{R}})^{-1})})_{rr}\,\mu_{q(1/a_r^{\text{R}})} - \tfrac{1}{2}\left(\nu + q^{\text{R}}\right)\textstyle\sum_{r=1}^{q^{\text{R}}}\log(B_{q(a_r^{\text{R}})}) \\
& - \log\left(B_{q(a_\varepsilon)}\right) + \mu_{q(1/a_\varepsilon)}\mu_{q(1/\sigma_\varepsilon^2)} - \log\left(B_{q(a_x)}\right) + \mu_{q(1/a_x)}\mu_{q(1/\sigma_x^2)} \\
& - \log\left(B_{q(a_w)}\right) + \mu_{q(1/a_w)}\mu_{q(1/\sigma_w^2)} \\
& - \tfrac{1}{2}\left\{\log(\sigma^2_{q(\mu_x)}) - \log(\sigma^2_{\mu_x})\right\} - \tfrac{1}{2}\,\sigma_{\mu_x}^{-2}\|\mu_{q(\mu_x)}\|^2 \\
& + \tfrac{1}{2}\,\sigma^2_{q(\mu_x)}\left(1/\sigma^2_{q(\mu_x)} - 1/\sigma^2_{\mu_x}\right) + \text{const},
\end{aligned}$$

where $\ell = 1$.

Derivation:

The logarithm of the lower bound on the marginal likelihood is given by

$$\begin{aligned}
&\log \underline{p}(\boldsymbol{y}, \boldsymbol{x}_{\text{obs}}, \boldsymbol{w}; q) = \\
&\quad E_q\left\{\log p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{w}, \boldsymbol{\beta}, \boldsymbol{u}^{\text{R}}, \boldsymbol{u}^{\text{G}}, \boldsymbol{\Sigma}^{\text{R}}, \boldsymbol{\sigma}_u^2, \boldsymbol{a}^{\text{R}}, \boldsymbol{a}_u, \mu_x, \sigma_x^2, a_x, \sigma_w^2, a_w, \sigma_\varepsilon^2, a_\varepsilon)\right. \\
&\qquad \left. - \log q^*(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{\beta}, \boldsymbol{u}^{\text{R}}, \boldsymbol{u}^{\text{G}}, \boldsymbol{\Sigma}^{\text{R}}, \boldsymbol{\sigma}_u^2, \boldsymbol{a}^{\text{R}}, \boldsymbol{a}_u, \mu_x, \sigma_x^2, a_x, \sigma_w^2, a_w, \sigma_\varepsilon^2, a_\varepsilon)\right\}.
\end{aligned}$$

According to the product restriction (4.5), our expression for the lower bound of the log-likelihood becomes:

$$\begin{aligned}
\log \underline{p}(\boldsymbol{y}, \boldsymbol{x}_{\text{obs}}, \boldsymbol{w}; q) \;=\; & E_q\left\{\log p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}^{\text{R}}, \boldsymbol{u}^{\text{G}}, \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{unobs}}, \sigma_\varepsilon^2)\right\} \\
& + E_q\left\{\log p(\boldsymbol{\beta}, \boldsymbol{u}^{\text{R}}, \boldsymbol{u}^{\text{G}}|\boldsymbol{\Sigma}^{\text{R}}, \boldsymbol{\sigma}_u^2) - \log q^*(\boldsymbol{\beta}, \boldsymbol{u}^{\text{R}}, \boldsymbol{u}^{\text{G}})\right\} \\
& + E_q\left\{\log p(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{unobs}}|\mu_x, \sigma_x^2) - \log q^*(\boldsymbol{x}_{\text{unobs}})\right\}
\end{aligned}$$

$$
\begin{aligned}
&+ E_q\left\{\log p(\boldsymbol{\Sigma}^{\mathrm{R}}|\boldsymbol{a}^{\mathrm{R}}) - \log q^*(\boldsymbol{\Sigma}^{\mathrm{R}})\right\} E_q\left\{p(\boldsymbol{\sigma}_u^2|\boldsymbol{a}_u) - \log q^*(\boldsymbol{\sigma}_u^2)\right\} \\
&+ E_q\left\{p(\sigma_x^2|a_x) - \log q^*(\sigma_x^2)\right\} + E_q\left\{p(\sigma_w^2|a_w) - \log q^*(\sigma_w^2)\right\} \\
&+ E_q\left\{\log p(\boldsymbol{a}^{\mathrm{R}}) - \log q^*(\boldsymbol{a}^{\mathrm{R}})\right\} + E_q\left\{\log p(\boldsymbol{a}_u) - \log q^*(\boldsymbol{a}_u)\right\} \\
&+ E_q\left\{\log p(a_x) - \log q^*(a_x)\right\} + E_q\left\{\log p(a_w) - \log q^*(a_w)\right\} \\
&+ E_q\left\{p(\sigma_\varepsilon^2|a_\varepsilon) - \log q^*(\sigma_\varepsilon^2)\right\} + E_q\left\{\log p(a_\varepsilon) - \log q^*(a_\varepsilon)\right\} \\
&+ E_q\left\{\log p(\boldsymbol{w}_{x_{\mathrm{obs}}}, \boldsymbol{w}_{x_{\mathrm{unobs}}}|\boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{x}_{\mathrm{unobs}}, \sigma_w^2)\right\} \\
&+ E_q\left\{\log p(\mu_x) - \log q^*(\mu_x)\right\}.
\end{aligned}
$$

We are only interested in the ones highlighted in darker colour as they represent the differences between models (2.11) and (4.1).
First note that

$$
\log p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}^{\mathrm{R}}, \boldsymbol{u}^{\mathrm{G}}, \boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{x}_{\mathrm{unobs}}, \sigma_\varepsilon^2) \;=\; -\tfrac{N}{2}\log(2\pi) - \tfrac{N}{2}\log(\sigma_\varepsilon^2) - \tfrac{1}{2\sigma_\varepsilon^2}\,\|\boldsymbol{y} - \boldsymbol{C}\boldsymbol{v}\|^2.
$$

Taking expectations gives

$$
\begin{aligned}
&E_q\{\log p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}^{\mathrm{R}}, \boldsymbol{u}^{\mathrm{G}}, \boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{x}_{\mathrm{unobs}}, \sigma_\varepsilon^2)\} = \\
&\qquad -\tfrac{N}{2}\log(2\pi) - \tfrac{N}{2}\,E_q\{\log(\sigma_\varepsilon^2)\} - \tfrac{1}{2\sigma_\varepsilon^2}\,E_q(\|\boldsymbol{y} - \boldsymbol{C}\boldsymbol{v}\|^2),
\end{aligned}
$$

where

$$
\begin{aligned}
E_q(\|\boldsymbol{y} - \boldsymbol{C}\boldsymbol{v}\|^2) \;=\; &\|\boldsymbol{y}_{x_{\mathrm{obs}}} - \boldsymbol{C}_{x_{\mathrm{obs}}}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\|^2 + \mathrm{tr}(\boldsymbol{C}_{x_{\mathrm{obs}}}^{\mathsf{T}}\boldsymbol{C}_{x_{\mathrm{obs}}}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}) \\
&+ \|\boldsymbol{y}_{x_{\mathrm{unobs}}}\|^2 - 2\,\boldsymbol{y}_{x_{\mathrm{unobs}}}^{\mathsf{T}}\,E_q(\boldsymbol{C}_{x_{\mathrm{unobs}}})\,\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \\
&+ \mathrm{tr}\left\{E_q(\boldsymbol{C}_{x_{\mathrm{unobs}}}^{\mathsf{T}}\boldsymbol{C}_{x_{\mathrm{unobs}}})(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}^{\mathsf{T}} + \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})\right\}.
\end{aligned}
$$

The next term is

$$
\begin{aligned}
&\log p(\boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{x}_{\mathrm{unobs}}|\mu_x, \sigma_x^2) - \log q^*(\boldsymbol{x}_{\mathrm{unobs}}) = \\
&\qquad -\tfrac{1}{2}\left\{\sum_{i=1}^{m}\left(n_i\log(2\pi) - n_i\log(\sigma_x^2) - \sigma_x^{-2}\,\|\boldsymbol{x}_i - \mu_x\mathbf{1}\|^2\right)\right\} \\
&\qquad -\tfrac{1}{2}\left\{\sum_{i=1}^{m}\left(n_i^{\mathrm{unobs}}\log(2\pi) - n_i^{\mathrm{unobs}}\log(\sigma_{q(\boldsymbol{x}_{\mathrm{unobs},i})}^2)\right.\right. \\
&\qquad \left.\left. -\;(1/\sigma_{q(\boldsymbol{x}_{\mathrm{unobs},i})}^2)\,\|\boldsymbol{x}_{\mathrm{unobs},i} - \boldsymbol{\mu}_{q(\boldsymbol{x}_{\mathrm{unobs},i})}\|^2\right)\right\}.
\end{aligned}
$$

Taking expectations gives

$$
\begin{aligned}
&E_q\left\{\log p(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{unobs}}|\mu_x, \sigma_x^2) - \log q^*(\boldsymbol{x}_{\text{unobs}})\right\} = \\
&\quad \sum_{i=1}^{m}\Big\{ -\tfrac{1}{2}\, n_i^{\text{obs}} \log(2\pi) - \tfrac{1}{2}\, n_i^{\text{obs}}\, E_q(\log(\sigma_x^2)) - \tfrac{1}{2}\, n_i^{\text{unobs}} \log(\sigma^2_{q(\boldsymbol{x}_{\text{unobs},i})}) \\
&\quad - \tfrac{1}{2}\, \mu_{q(1/\sigma_x^2)}\Big(\|\boldsymbol{x}_{\text{obs},i} - \mu_{q(\mu_x)} \mathbf{1}_{n_i^{\text{obs}}}\|^2 + \|\boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs},i})} - \mu_{q(\mu_x)}\, \mathbf{1}_{n_i^{\text{unobs}}}\|^2 + n_i\, \sigma^2_{q(\mu_x)}\Big) \\
&\quad + \tfrac{1}{2}\, \Big(1/\sigma^2_{q(\boldsymbol{x}_{\text{unobs},i})} + \mu_{q(1/\sigma_x^2)}\Big)\, n_i^{\text{unobs}} \sigma^2_{q(\boldsymbol{x}_{\text{unobs},i})}\Big\}.
\end{aligned}
$$

The next term is

$$
\begin{aligned}
\log p(\sigma_x^2|a_x) - \log q^*(\sigma_x^2) \;=\; & -\tfrac{1}{2}\log(a_x) - \log\Gamma(1/2) - \tfrac{3}{2}\log(\sigma_x^2) - 1/(a_x\sigma_x^2) \\
& - \tfrac{1}{2}(N+1)\log(B_{q(\sigma_x^2)}) + \log\Gamma\{\tfrac{1}{2}(N+1)\} \\
& + (\tfrac{N}{2} + \tfrac{3}{2})\log(\sigma_x^2) - B_{q(\sigma_x^2)}/\sigma_x^2.
\end{aligned}
$$

Taking expectations gives

$$
\begin{aligned}
E_q\left\{\log p(\sigma_x^2|a_x) - \log q^*(\sigma_x^2)\right\} \;=\; & -\tfrac{1}{2}E_q\{\log(a_x)\} - \log\Gamma(1/2) + \tfrac{N}{2}E_q\{\log(\sigma_x^2)\} \\
& + \mu_{q(1/\sigma_x^2)}\left(B_{q(\sigma_x^2)} - \mu_{q(1/a_x)}\right) \\
& - \tfrac{1}{2}(N+1)\log B_{q(\sigma_x^2)} + \log\Gamma\{\tfrac{1}{2}(N+1)\}.
\end{aligned}
$$

Similar arguments lead to

$$
\begin{aligned}
E_q\left\{\log p(\sigma_w^2|a_w) - \log q^*(\sigma_w^2)\right\} \;=\; & -\tfrac{1}{2}E_q\{\log(a_w)\} - \log\Gamma(1/2) + \tfrac{N}{2}E_q\{\log(\sigma_w^2)\} \\
& + \mu_{q(1/\sigma_w^2)}\left(B_{q(\sigma_w^2)} - \mu_{q(1/a_w)}\right) \\
& - \tfrac{1}{2}(N+1)\log B_{q(\sigma_w^2)} + \log\Gamma\{\tfrac{1}{2}(N+1)\}
\end{aligned}
$$

and

$$
\begin{aligned}
E_q\left\{\log p(\sigma_\varepsilon^2|a_\varepsilon) - \log q^*(\sigma_\varepsilon^2)\right\} \;=\; & -\tfrac{1}{2}E_q\{\log(a_\varepsilon)\} - \log\Gamma(1/2) + \tfrac{N}{2}E_q\{\log(\sigma_\varepsilon^2)\} \\
& + \mu_{q(1/\sigma_\varepsilon^2)}\left(B_{q(\sigma_\varepsilon^2)} - \mu_{q(1/a_\varepsilon)}\right) \\
& - \tfrac{1}{2}(N+1)\log B_{q(\sigma_\varepsilon^2)} + \log\Gamma\{\tfrac{1}{2}(N+1)\}.
\end{aligned}
$$

The next term is

$$
E_q\left\{\log p(\boldsymbol{w}|\boldsymbol{x}, \sigma_w^2)\right\} = -\tfrac{N}{2}\log(2\pi) - \tfrac{N}{2}E_q\{\log(\sigma_w^2)\} - \tfrac{1}{2}\,\mu_{q(1/\sigma_w^2)}E_q(\|\boldsymbol{w} - \boldsymbol{x}\|^2),
$$

where

$$\begin{aligned} E_q(\|\boldsymbol{w}-\boldsymbol{x}\|^2) &= \|\boldsymbol{w}_{x_{\text{obs}}}-\boldsymbol{x}_{\text{obs}}\|^2+\|\boldsymbol{w}_{x_{\text{unobs}}}\|^2-2\boldsymbol{w}_{x_{\text{unobs}}}^{\mathsf{T}}\boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs}})} \\ &\quad +\|\boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs}})}\|^2+\text{tr}(\text{Cov}_q(\boldsymbol{x}_{\text{unobs}})) \\ &= \|\boldsymbol{w}_{x_{\text{unobs}}}-\boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs}})}\|^2+\sum_{i=1}^{m} n_i^{\text{unobs}}\sigma^2_{q(\boldsymbol{x}_{\text{unobs},i})}. \end{aligned}$$

The next term is

$$\begin{aligned} \log p(\mu_x)-\log q^*(\mu_x) &= -\tfrac{1}{2}\log(2\pi)-\tfrac{1}{2}\log(\sigma^2_{\mu_x})-\tfrac{1}{2\sigma^2_{\mu_x}}\|\mu_x\|^2 \\ &\quad +\tfrac{1}{2}\log(2\pi)+\tfrac{1}{2}\log(\sigma^2_{q(\mu_x)})+\tfrac{1}{2\sigma^2_{q(\mu_x)}}(\|\mu_x-\mu_{q(\mu_x)}\|^2). \end{aligned}$$

Taking expectations gives

$$\begin{aligned} E_1\left\{\log p(\mu_x)-\log q^*(\mu_x)\right\} &= -\tfrac{1}{2}\left(\log(\sigma^2_{q(\mu_x)})-\log(\sigma^2_{\mu_x})\right)-\tfrac{1}{2\sigma^2_{\mu_x}}\|\mu_{q(\mu_x)}\|^2 \\ &\quad +\tfrac{1}{2}\sigma^2_{q(\mu_x)}\left(\tfrac{1}{\sigma^2_{q(\mu_x)}}-\tfrac{1}{\sigma^2_{\mu_x}}\right). \end{aligned}$$

The remaining expectation expressions follow similar algebraic calculations and are therefore omitted.

4.B Optimal q-densities derivation for missing data problems

We first write out the full model expression

$$y_{ij} = \beta_0+\beta_x\,x_{ij}+\beta_s\,s_{ij}+\sum_{k=1}^{q^{\text{G}}}u_k^{\text{G}}\,z_k(s_{ij})+u_{0i}^{\text{R}}+u_{1i}^{\text{R}}\,x_{ij}+\varepsilon_{ij},$$

which can be broken into the observed and missing parts:

$$\begin{bmatrix}\boldsymbol{y}_{x_{\text{obs}}}\\ \boldsymbol{y}_{x_{\text{mis}}}\end{bmatrix} = \begin{bmatrix}\mathbf{1}_{n_{\text{obs}}} & \boldsymbol{x}_{\text{obs}} & \boldsymbol{s}_{x_{\text{obs}}}\\ \mathbf{1}_{n_{\text{mis}}} & \boldsymbol{x}_{\text{mis}} & \boldsymbol{s}_{x_{\text{mis}}}\end{bmatrix}\begin{bmatrix}\beta_0\\ \beta_x\\ \beta_s\end{bmatrix}+\begin{bmatrix}\boldsymbol{Z}^{\text{G}}_{x_{\text{obs}}}\\ \boldsymbol{Z}^{\text{G}}_{x_{\text{mis}}}\end{bmatrix}\begin{bmatrix}u_1^{\text{G}}\\ \vdots\\ u_m^{\text{G}}\end{bmatrix}+\begin{bmatrix}\boldsymbol{Z}^{\text{R}}_{x_{\text{obs}}}\\ \boldsymbol{Z}^{\text{R}}_{x_{\text{mis}}}\end{bmatrix}\begin{bmatrix}\boldsymbol{u}_1^{\text{R}}\\ \vdots\\ \boldsymbol{u}_m^{\text{R}}\end{bmatrix}.$$

Expression for $q^*(\boldsymbol{x}_{\mathbf{mis},i})$, $\boldsymbol{\mu}_{q(\boldsymbol{x}_{\mathbf{mis},i})}$ and $\sigma^2_{q(\boldsymbol{x}_{\mathbf{mis},i})}$

For $i = 1, \ldots, m$

$$q^*(\boldsymbol{x}_{\mathrm{mis},i}) \sim N(\boldsymbol{\mu}_{q(\boldsymbol{x}_{\mathrm{mis},i})}, \sigma^2_{q(\boldsymbol{x}_{\mathrm{mis},i})}\boldsymbol{I}),$$

where

$$\begin{aligned}
\sigma^2_{q(\boldsymbol{x}_{\mathrm{mis},i})} &= E_q\left\{\mathbf{1}^{\mathsf{T}}_{n_i^{\mathrm{mis}}}\left(\tfrac{1}{\sigma^2_\varepsilon}(\beta_x + u^{\mathrm{R}}_{1i})^2\mathbf{1}_{n_i^{\mathrm{mis}}} + \tfrac{1}{\sigma^2_x}\mathbf{1}_{n_i^{\mathrm{mis}}} - 2\,\phi_1^2\,\lambda(\boldsymbol{\xi}_{x_{\mathrm{mis,i}}})\right)\right\}^{-1} \\
\text{and}\quad \boldsymbol{\mu}_{q(\boldsymbol{x}_{\mathrm{mis},i})} &= \sigma^2_{q(\boldsymbol{x}_{\mathrm{mis},i})}\boldsymbol{I}_{x_{\mathrm{mis}}}E_q\Big\{\tfrac{1}{\sigma^2_\varepsilon}(\beta_x + u^{\mathrm{R}}_{1i})^2\Big(\boldsymbol{y}_{x_{\mathrm{mis},i}} - \beta_0\mathbf{1}_{n_i^{\mathrm{mis}}} \\
&\quad -\beta_s\boldsymbol{s}_{x_{\mathrm{mis},i}} - \textstyle\sum_{k=1}^{q^{\mathrm{G}}} u^{\mathrm{G}}_k z_k(\boldsymbol{s}_{x_{\mathrm{mis},i}}) - u^{\mathrm{R}}_{0i}\mathbf{1}_{n_i^{\mathrm{mis}}}\Big) \\
&\quad + \Big(\tfrac{\mu_x}{\sigma^2_x} + 2\,\phi_0\,\phi_1\,\mathrm{diag}\{\lambda(\boldsymbol{\xi}_{x_{\mathrm{mis,i}}})\} - \tfrac{1}{2}\phi_1\Big)\mathbf{1}_{n_i^{\mathrm{mis}}}\Big\}
\end{aligned}$$

Derivation:
First we note that

$$\begin{aligned}
p(\boldsymbol{x}_{\mathrm{mis}}|\mathrm{rest}) &\propto p(\boldsymbol{x}_{\mathrm{mis}}|\boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{y}, \boldsymbol{u}^{\mathrm{R}}, \boldsymbol{u}^{\mathrm{G}}, \boldsymbol{\beta}, \sigma^2_\varepsilon, \mu_x, \sigma^2_x, \boldsymbol{R}, \boldsymbol{\phi}) \\
&\propto p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{x}_{\mathrm{mis}}, \boldsymbol{u}^{\mathrm{R}}, \boldsymbol{u}^{\mathrm{G}}, \sigma^2_\varepsilon)\, p(\boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{x}_{\mathrm{mis}}|\mu_x, \sigma^2_x, \boldsymbol{R}, \boldsymbol{\phi})\, p(\boldsymbol{R}|\boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{x}_{\mathrm{mis}}, \boldsymbol{\phi}).
\end{aligned}$$

Taking logarithms on both sides where x is missing gives

$$\begin{aligned}
&\log p(\boldsymbol{x}_{\mathrm{mis}}|\mathrm{rest}) \\
&= \sum_{i=1}^{m}\Big[-\tfrac{1}{2\sigma^2_\varepsilon}\Big\|\boldsymbol{y}_{x_{\mathrm{mis},i}} - \Big(\beta_0\mathbf{1}_{n_i^{\mathrm{mis}}} + \beta_x\boldsymbol{x}_{\mathrm{mis},i} + \beta_s\boldsymbol{s}_{x_{\mathrm{mis},i}} \\
&\quad - \sum_{k=1}^{q^{\mathrm{G}}} u^{\mathrm{G}}_k z_k(\boldsymbol{s}_{x_{\mathrm{mis},i}}) + u^{\mathrm{R}}_{0i}\mathbf{1}_{n_i^{\mathrm{mis}}} + u^{\mathrm{R}}_{1i}\boldsymbol{x}_{\mathrm{mis},i}\Big)\Big\|^2 \\
&\quad - \tfrac{1}{2\sigma^2_x}\|\boldsymbol{x}_{\mathrm{mis},i} - \mu_x\mathbf{1}_{n_i^{\mathrm{mis}}}\|^2 + \{1 + \exp(\phi_0 + \phi_1\boldsymbol{x}_{\mathrm{mis},i})\}^{-1}\Big] \\
&\geq -\tfrac{1}{2}\Big(\mathbf{1}^{\mathsf{T}}_{n_i^{\mathrm{mis}}}\Big[\Big\{\tfrac{1}{\sigma^2_\varepsilon}(\beta_x + u^{\mathrm{R}}_{1i})^2\mathbf{1}_{n_i^{\mathrm{mis}}} + \tfrac{1}{\sigma^2_x}\mathbf{1}_{n_i^{\mathrm{mis}}} - 2\,\phi_1^2\,\lambda\{\boldsymbol{\xi}_{x_{\mathrm{mis,i}}}\}\Big\} \odot (\boldsymbol{x}_{\mathrm{mis},i})^2\Big] \\
&\quad - 2\,\boldsymbol{x}^{\mathsf{T}}_{\mathrm{mis},i}\Big\{\tfrac{1}{\sigma^2_\varepsilon}(\beta_x + u^{\mathrm{R}}_{1i})\Big(\boldsymbol{y}_{x_{\mathrm{mis},i}} - \beta_0\mathbf{1}_{n_i^{\mathrm{mis}}} - \beta_s\mathbf{1}_{n_i^{\mathrm{mis}}} - \sum_{k=1}^{q^{\mathrm{G}}} u^{\mathrm{G}}_k z_k(\boldsymbol{s}_{x_{\mathrm{mis},i}}) - u^{\mathrm{R}}_{0i}\mathbf{1}_{n_i^{\mathrm{mis}}}\Big) \\
&\quad + \tfrac{1}{\sigma^2_x}(\mu_x\mathbf{1}_{n_i^{\mathrm{mis}}}) + (2\,\phi_0\,\phi_1\,\mathrm{diag}\{\lambda(\boldsymbol{\xi}_{x_{\mathrm{mis,i}}})\} - \tfrac{1}{2}\,\phi_1)\mathbf{1}_{n_i^{\mathrm{mis}}}\Big\}\Big) + \mathrm{const},
\end{aligned}$$

where the two expectation terms are as follows:

$$E_q\left\{\mathbf{1}^{\mathsf{T}}_{n_i^{\mathrm{mis}}}\left(\tfrac{1}{\sigma^2_\varepsilon}(\beta_x + u^{\mathrm{R}}_{1i})^2\mathbf{1}_{n_i^{\mathrm{mis}}} + \tfrac{1}{\sigma^2_x}\mathbf{1}_{n_i^{\mathrm{mis}}} - 2\,\phi_1^2\,\lambda(\boldsymbol{\xi}_{x_{\mathrm{mis,i}}})\right)\right\}$$

$$
\begin{aligned}
&= \mu_{q(1/\sigma_\varepsilon^2)} \Big\{ \boldsymbol{\mu}^2_{q(\beta_x)} + \boldsymbol{\Sigma}_{q(\beta_x)} + 2\left(\boldsymbol{\mu}_{q(\beta_x)}\,\boldsymbol{\mu}_{q(u^{\mathrm{R}}_{1i})} + \boldsymbol{\Lambda}_{q(\beta_x,\,u^{\mathrm{R}}_{1i})}\right) \\
&\quad + \boldsymbol{\mu}^2_{q(u^{\mathrm{R}}_{1i})} + \boldsymbol{\Sigma}_{q(u^{\mathrm{R}}_{1i})} \Big\} + \mu_{q(1/\sigma_x^2)} \\
&\quad + 2(\boldsymbol{\mu}^2_{q(\phi_1)} + \boldsymbol{\Sigma}_{q(\boldsymbol{\phi}_1)}) \textstyle\sum_{i=1}^m \lambda(\boldsymbol{\xi}_{x_{\mathrm{mis,i}}}), \\
E_q\Big\{ &\tfrac{1}{\sigma_\varepsilon^2}(\beta_x + u^{\mathrm{R}}_{1i})^2 \Big(\boldsymbol{y}_{x_{\mathrm{mis},i}} - \beta_0 \mathbf{1}_{n_i^{\mathrm{mis}}} - \beta_s \boldsymbol{s}_{x_{\mathrm{mis},i}} - \textstyle\sum_{k=1}^{q^{\mathrm{G}}} u^{\mathrm{G}}_k z_k(\boldsymbol{s}_{x_{\mathrm{mis},i}}) - u^{\mathrm{R}}_{0i} \mathbf{1}_{n_i^{\mathrm{mis}}} \Big) \\
&+ \Big(\tfrac{\mu_x}{\sigma_x^2} + 2\,\phi_0\,\phi_1\,\mathrm{diag}\{\lambda(\boldsymbol{\xi}_{x_{\mathrm{mis,i}}})\} - \tfrac{1}{2}\phi_1 \Big) \mathbf{1}_{n_i^{\mathrm{mis}}} \Big\} \\
&= \sigma_{q(\boldsymbol{x}_{\mathrm{mis},i})} \boldsymbol{I}_{n_i^{\mathrm{mis}}} \Big[\mu_{q(1/\sigma_\varepsilon^2)} \Big\{ \left(\boldsymbol{\mu}_{q(\beta_x)} + \boldsymbol{\mu}_{q(u^{\mathrm{R}}_{1i})}\right) \boldsymbol{y}_{x_{\mathrm{mis},i}} \\
&\quad - \left(\boldsymbol{\mu}_{q(\beta_0)}\,\boldsymbol{\mu}_{q(\beta_x)} + \boldsymbol{\Lambda}_{q(\beta_0,\beta_x)} + \boldsymbol{\mu}_{q(\beta_0)}\,\boldsymbol{\mu}_{q(u^{\mathrm{R}}_{1i})} + \boldsymbol{\Lambda}_{q(\beta_0,\,u^{\mathrm{R}}_{1i})}\right) \mathbf{1}_{n_i^{\mathrm{mis}}} \\
&\quad - \left(\boldsymbol{\mu}_{q(\beta_x)}\,\boldsymbol{\mu}_{q(\beta_s)} + \boldsymbol{\Lambda}_{q(\beta_x,\,\beta_s)} + \boldsymbol{\mu}_{q(\beta_s)}\,\boldsymbol{\mu}_{q(u^{\mathrm{R}}_{1i})} + \boldsymbol{\Lambda}_{q(\beta_s,\,u^{\mathrm{R}}_{1i})}\right) \boldsymbol{s}_{x_{\mathrm{mis},i}} \\
&\quad - \textstyle\sum_{k=1}^{q^{\mathrm{G}}} \left(\boldsymbol{\mu}_{q(\beta_x)}\,\boldsymbol{\mu}_{q(u^{\mathrm{G}}_k)} + \boldsymbol{\Lambda}_{q(\beta_x,\,u^{\mathrm{G}}_k)}\right) z_k(\boldsymbol{s}_{x_{\mathrm{mis},i}}) \\
&\quad - \textstyle\sum_{k=1}^{q^{\mathrm{G}}} \left(\boldsymbol{\mu}_{q(\beta_x)}\,\boldsymbol{\mu}_{q(u^{\mathrm{G}}_k)} + \boldsymbol{\Lambda}_{q(\beta_x,\,u^{\mathrm{G}}_k)}\right) z_k(\boldsymbol{s}_{x_{\mathrm{mis},i}}) \\
&\quad - \textstyle\sum_{k=1}^{q^{\mathrm{G}}} \left(\boldsymbol{\mu}_{q(u^{\mathrm{R}}_{1i})}\,\boldsymbol{\mu}_{q(u^{\mathrm{G}}_k)} + \boldsymbol{\Lambda}_{q(u^{\mathrm{R}}_{1i},\,u^{\mathrm{G}}_k)}\right) z_k(\boldsymbol{s}_{x_{\mathrm{mis},i}}) \\
&\quad - \left(\boldsymbol{\mu}_{q(\beta_x)}\,\boldsymbol{\mu}_{q(u^{\mathrm{R}}_{0i})} + \boldsymbol{\Lambda}_{q(\beta_x,\,u^{\mathrm{R}}_{0i})} + \boldsymbol{\mu}_{q(u^{\mathrm{R}}_{0i})}\,\boldsymbol{\mu}_{q(u^{\mathrm{R}}_{1i})} + \boldsymbol{\Lambda}_{q(u^{\mathrm{R}}_{0i},\,u^{\mathrm{R}}_{1i})}\right) \mathbf{1}_{n_i^{\mathrm{mis}}} \Big\} \\
&\quad + \mu_{q(1/\sigma_x^2)}\,\mu_{q(\mu_x)} \mathbf{1}_{n_i^{\mathrm{mis}}} + 2(\boldsymbol{\mu}^2_{q(\phi_1)} + \boldsymbol{\Sigma}_{q(\boldsymbol{\phi}_1)}) \lambda(\boldsymbol{\xi}_{\boldsymbol{x}_{\mathrm{mis},i}}) \Big] .
\end{aligned}
$$

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

$$q^*(\boldsymbol{\phi}) \sim N(\boldsymbol{\mu}_{q(\boldsymbol{\phi};\boldsymbol{\xi})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\phi};\boldsymbol{\xi})})$$

where

$$\boldsymbol{\mu}_{q(\boldsymbol{\phi};\boldsymbol{\xi})} = \boldsymbol{\Sigma}_{q(\boldsymbol{\phi};\boldsymbol{\xi})} \left\{ (\boldsymbol{R} - \tfrac{1}{2}) E_q(\boldsymbol{X}^{\mathsf{T}}) \right\} \quad \text{and} \quad \boldsymbol{\Sigma}_{q(\boldsymbol{\phi};\boldsymbol{\xi})} = \left(\tfrac{1}{\sigma_\phi^2} \boldsymbol{I} - 2\, E_q[\boldsymbol{X}^{\mathsf{T}} \mathrm{diag}\{\lambda(\boldsymbol{\xi})\} \boldsymbol{X}^{\mathsf{T}}] \right)^{-1}.$$

Derivation:
First we note that

$$
\begin{aligned}
p(\boldsymbol{\phi}|\mathrm{rest}) &\propto p(\boldsymbol{\phi}|\boldsymbol{R}, \boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{x}_{\mathrm{mis}}) = p(\boldsymbol{R}|\boldsymbol{\phi}, \boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{x}_{\mathrm{mis}})\, p(\boldsymbol{\phi}) \\
&\propto \exp[\boldsymbol{R}^{\mathsf{T}}(\boldsymbol{X}\phi) - \mathbf{1}^{\mathsf{T}} \log(1 + \exp(\boldsymbol{X}\phi))] \exp(-\tfrac{1}{2\sigma_\phi^2} \|\boldsymbol{\phi}\|^2)
\end{aligned}
$$

Replacing

$$-\mathbf{1}^{\mathsf{T}} \log(1 + \exp(\boldsymbol{X}\boldsymbol{\phi})) \geq (\boldsymbol{X}\boldsymbol{\phi})^{\mathsf{T}} \mathrm{diag}\{\lambda(\boldsymbol{\xi})\}(\boldsymbol{X}\boldsymbol{\phi}) - \tfrac{1}{2}\mathbf{1}^{\mathsf{T}}(\boldsymbol{X}\boldsymbol{\phi}) + \mathbf{1}^{\mathsf{T}} \psi(\boldsymbol{\xi})$$

and taking logarithms on both sides gives

$$\log p(\boldsymbol{\phi}|\text{rest}) \geq -\tfrac{1}{2}\left\{\boldsymbol{\phi}^{\mathsf{T}}\left(\tfrac{1}{\sigma_{\phi}^{2}}\boldsymbol{I} - 2\left[\boldsymbol{X}^{\mathsf{T}}\text{diag}\{\lambda(\boldsymbol{\xi})\}\boldsymbol{X}\right]\right)\boldsymbol{\phi} - 2\,\boldsymbol{\phi}^{\mathsf{T}}(\boldsymbol{R}-\tfrac{1}{2})\,\boldsymbol{X}^{\mathsf{T}}\right\}.$$

Taking expectations with respect to all parameters except $\boldsymbol{x}_{\text{unobs},i}$ leads to

$$\log q^{*}(\boldsymbol{\phi}) \geq -\tfrac{1}{2}\left\{\boldsymbol{\phi}^{\mathsf{T}}\left(\tfrac{1}{\sigma_{\phi}^{2}}\boldsymbol{I} - 2\,E_{q}[\boldsymbol{X}^{\mathsf{T}}\text{diag}\{\lambda(\boldsymbol{\xi})\}\boldsymbol{X}]\right)\boldsymbol{\phi} - 2\,\boldsymbol{\phi}^{\mathsf{T}}(\boldsymbol{R}-\tfrac{1}{2})\,E_{q}(\boldsymbol{X})^{\mathsf{T}}\right\},$$

where the expressions for $E_q[\boldsymbol{X}^{\mathsf{T}}\text{diag}\{\lambda(\boldsymbol{\xi})\}\boldsymbol{X}]$ and $E_q(\boldsymbol{X})$ are given in Subsection 4.3.1.

Chapter 5

Extension to Group-Specific Curve Models With Contrasting

There is a magic in graphs. The profile of a curve reveals in a flash a whole situation - the life history of an epidemic, a panic, or an era of prosperity. The curve informs the mind, awakens the imagination, convinces.

Henry D. Hubbard

5.1 Introduction

Parametric regression methods for longitudinal and multilevel data have been well developed in the last two decades. Such methods can be broadly classified into estimating equations based methods (Diggle *et al.*, 2002; Zeger and Liang, 1986) and mixed models (Breslow and Clayton, 1993; Gelman and Hill, 2007; Goldstein, 2011; Laird and Ware, 1982). A parametric regression model assumes that the relationship between the mean of a response outcome and predictor variables to be of a known functional form. Although such a parametric assumption offers simplicity, it is inappropriate for situations when the relationship between the mean response and predictors is unknown. As an illustration, Figure 5.1 shows the overall and hospital-specific trends in caesarean section rates for low-risk nulliparous women in the largest state of Australia, New South Wales (NSW),

The full content of this chapter is published as: Lee, C. Y. Y. and Wand, M. P. (2015). Variational methods for fitting complex Bayesian mixed models to health data. *Statistics in Medicine*. DOI: 10.1002/sim.6737. A manuscript related to this chapter is currently under revision in *Birth*: Lee, C. Y. Y., Homer, C., Bisits, A. and Ryan, L. (2015). Increasing variation in hospital caesarean section rates among low-risk nulliparous women in Australia, from 1994 to 2010.

between 1994 and 2010. The trends are clearly non-linear and show substantial within- and between-hospital variability over time. In order to capture these features simultaneously, a more complicated and flexible model is required. For example, a more flexible model is a generalisation to additive models (Hastie and Tibshirani, 1990), where the overall mean and deviation of each group (in this case hospital) from that overall mean are modelled nonparametrically with arbitrary smooth functions of time. We refer to such model as a *group-specific curve model*.

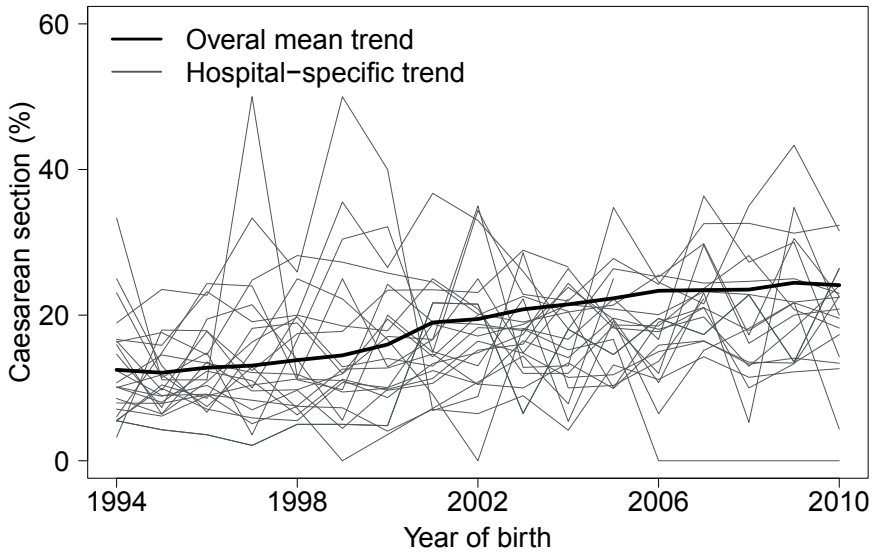

Figure 5.1: Trends in caesarean section rates for low-risk nulliparous women in the largest state of Australia, New South Wales, between 1994 and 2010. The dotted line is the overall mean trend and the solid lines are the selected hospital-specific trends.

Early work on group-specific curve models used different smoothing techniques (e.g. kernels and smoothing splines) to estimate both overall mean curve and group-specific curves, with random effects modelled either by a parametric function or a Gaussian process (Verbyla *et al.*, 1999; Zeger and Diggle, 1994; Zhang *et al.*, 1998). Brumback and Rice (1998) embedded smoothing splines within the mixed model framework. However, they ran into computational problems in the estimation since they assumed fixed intercepts and slopes for the group-specific curves, thus leading to the number of fixed effects being at least twice as large as the number of curves. Rice and Wu (2001) alleviated computational problems by modelling group-specific curves as spline basis functions with random coefficients. However, their approach requires a careful selection of the number and location of knots. Durbán *et al.* (2005) relaxed the importance of this selection by using low rank smoothing splines with a penalty approach, which expresses the group-specific curves as a

linear combination of truncated polynomial spline bases with random coefficients. Their proposed covariance matrix for the random basis coefficients was modelled parametrically with independence assumptions. Chen and Wang (2011) and Ryu *et al.* (2011) extended this work by allowing a more general covariance matrix structure for the random basis coefficients, although the former was confined to functional data analysis.

When the dimension of the spline basis functions and the number of groups become large, many of the above-mentioned approaches become too slow, or even computationally infeasible. These problems motivate the development of fast and scalable methods for fitting group-specific curve models to large longitudinal and multilevel datasets. In Bayesian statistics, exact inference for nonparametric or semiparametric regression models that use penalised spline basis functions is typically intractable, requiring approximate inference methods for use in practice. Markov chain Monte Carlo or MCMC is the most commonly used approximate inference method in this setting, but can be computationally intensive and often suffers from poor convergence in large and complex models. A faster, deterministic alternative to MCMC is *variational approximations* (Bishop, 2006; Ormerod and Wand, 2010). The basic idea behind variational approximations is to recast the problem of computing posterior probabilities as an optimisation problem by introducing a class of more manageable approximating distributions, and then optimizing some criterion to find the distribution within that class that best matches the posterior. Recently, there has been a growing interest in variational methods for longitudinal and multilevel models. Ormerod and Wand (2010) and Luts *et al.* (2014) developed variational algorithms for fitting and inference for grouped data. However, their algorithms are infeasible for datasets with a large number of groups due to naïve inversion of the random effects covariance matrix. Tan and Nott (2013) introduced a partially noncentered nonparametrisation strategy allowing the random effects for each group to be independent. Such a restriction was shown to improve efficiency of the variational algorithms. Lee and Wand (2015a) did not impose such a restriction, but rather, took advantage of the inherent blocked structure and sparseness of the random effects covariance matrix and developed algorithms that streamline its inversion and estimation. Their streamlined algorithms for the longitudinal and multilevel models result in impressive speed improvements (up to the order of thousands) and currently represent the state-of-the-art in this area. Stewart (2014) applied their streamlined algorithms to research studies in social sciences.

In this chapter, we present a fully variational approach to fitting a series of Bayesian logistic mixed models in order to characterise trends in caesarean section rates in New South Wales, Australia. We begin with the standard random intercept and slope model and then generalize the model by replacing linear models for the overall mean and hospital-specific trends with arbitrary smooth functions that are nonparametrically estimated from the data. Section 5.2 reviews the random intercept and slope model and presents various

extensions of the standard model including nonparametric functions and factor-by-curve interactions. A basic framework of approximate inference for a simple Bayesian logistic mixed model is outlined, serving as a building block for the more complicated models. Details on variational fitting and inference for models are also given. In Section 5.3 we present numerical evidence of the efficacy of our developed variational algorithms in terms of inferential accuracy and computational speed. The final proposed model is illustrated in Section 5.4 with the analysis of caesarean section data. The chapter concludes with a brief discussion in Section 5.5.

5.2 Mean field variational Bayes approximations to caesarean section data

In this section, we present an epidemiology study that motivates our methodological development. Caesarean section rates are increasing worldwide and among countries of Organisation for Economic Co-operation and Development the average rate increased from 20% in 2000 to 27% in 2011 (Organisation for Economic Co-Operation and Development, 2011). It is recognised that caesarean section rates vary considerably across regions and hospitals in several countries. For example, in the United States, a four-fold variation was found between low- and high-use areas (Baicker *et al.*, 2006). In the United Kingdom, rates of emergency cesarean section among National Health Service trusts ranged from 15% to 32% (Bragg *et al.*, 2010). In the largest state of Australia, NSW, there were 1 500 964 deliveries from 1994 to 2010, with caesarean section rate increased from 17.4% to 30.6% over this 17-year period. While the statewide trend in caesarean section rate is well reported, little is known about the trend in caesarean section rate for each hospital, especially for small hospitals. Previous work by the author addressed this gap by examining hospital-specific trends in caesarean section rates for low-risk nulliparous women in NSW between 1994 and 2010 (Lee *et al.*, 2015). Here we extend the model applied previously, using the MFVB approximation, to compare hospital-specific trends between subpopulations of women.

Data were obtained from the NSW Perinatal Data Collection. The Perinatal Data Collection is a legislated population-based surveillance system covering all live births, and stillbirths of at least 20 weeks gestation or at least 400 grams birthweight. Information is recorded by the attending midwife or medical practitioner providing maternity care, and includes maternal demographic, medical and obstetric information of the mother, and details of labor, birth and condition of the infant. Data for this analysis were de-identified. Permission for use of data was approved by the NSW Ministry of Health. The study population consists of low-risk women giving birth for the first time in a NSW public

or private hospital between 1994 and 2010. Low-risk women were defined as those women giving birth for the first time (nulliparous women), aged 20-34 years who did not smoke in pregnancy and did not have any pre-existing or gestational medical conditions, such as diabetes and hypertension, gave birth to singleton cephalic (head down position) live infants at term (37 weeks gestation).

The primary outcome is the annual rate of caesarean section for each hospital in NSW over a 17-year period. Public and private hospitals with $\geq$ 50 births per annum for more than half of the 17-year study period were included. The data have a two-level hierarchical structure with women nested within hospitals. To define notation, we use a double subscript convention with i denoting hospital ($i = 1, \ldots, m$) and j denoting woman ($j = 1, \ldots n_i$). For example, y_{ij} denote the binary indicator of caesarean section for woman j in hospital i and x_{ij} the corresponding year of birth. The total number of low-risk nulliparous women in the study population is $\sum_{i=1}^{m} n_i$. To examine the underlying trends in caesarean section rates overall and for each hospital, we consider a series of Bayesian logistic mixed models with increasing complexity. For each model, the year of birth is modelled by adding together two **arbitrary unspecified functions** as covariates, one representing the overall mean $f(x)$ and the other being hospital-specific departures from that overall mean $g_i(x)$. To keep notation simple, we consider only a single predictor in the model. Extension to models with more than one predictor is straightforward. More precisely, we assume

$$y_{ij} \overset{\text{ind.}}{\sim} \text{Bernoulli}(\text{logit}^{-1}\{f(x_{ij}) + g_i(x_{ij})\}), \quad 1 \leq j \leq n_i,\ 1 \leq i \leq m, \tag{5.1}$$

where the functional forms of f and g_i will be reviewed in the later subsections. The function g_i controls for hospital-to-hospital variation, both in magnitude and across the 17-year trend. The sum of f and g_i provides the trend in caesarean section rate for each hospital, which can have an unique intercept, slope and even shape depending upon whether they are of a parametric or nonparametric form. The term $f(x_{ij}) + g_i(x_{ij})$ represents the log odds of caesarean section for woman j in hospital i and will be the focus in the following subsections.

The Bayesian hierarchical model corresponding to (5.1) is

$$\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u} \sim \text{Bernoulli}(\text{logit}^{-1}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u})), \quad \boldsymbol{u}|\boldsymbol{G} \sim N(\boldsymbol{0}, \boldsymbol{G}). \tag{5.2}$$

The notation $\boldsymbol{v} \sim \text{Bernoulli}(\boldsymbol{p})$ used in (5.2) is shorthand for the entries of the random vector $\boldsymbol{v}$ having independent Bernoulli distributions with parameters corresponding to the entries of $\boldsymbol{p}$. The matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the respective $\sum_{i=1}^{m} n_i \times P$ fixed effects design matrix and $\sum_{i=1}^{m} n_i \times d$ random effects design matrix, $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are the so called $P \times 1$

fixed effects vector and $d \times 1$ random effects vector. The random effects vectors are given a multivariate normal prior with zero mean and covariance matrix $\boldsymbol{G}$. Throughout we take the prior distribution of the fixed effects vector $\boldsymbol{\beta}$ to be of the form $\boldsymbol{\beta} \sim N(0, \sigma_\beta^2 \boldsymbol{I})$, where σ_β^2 is to be chosen by the analyst. In logistic regression, situations where a shift in the predictor x corresponds to the probability of response y changing from 0.01 to 0.99 are rare. Hence, a prior distribution that assigns low probabilities to changes of 10 on the logistic scale would be appropriate. Further, we use proper but "diffuse" conditionally conjugate priors for the random effects, this corresponds to a Half-Cauchy distribution for a single variance component (Result 5 of Wand *et al.* (2012))

$$\sigma^2 \,|\, a \sim \text{Inverse-Gamma}(\tfrac{1}{2}, 1/a), \quad a \sim \text{Inverse-Gamma}(\tfrac{1}{2}, 1/A^2).$$

The multivariate extension of the Half-Cauchy distribution is a scaled inverse-Wishart distribution for an unstructured $q^{\mathrm{R}} \times q^{\mathrm{R}}$ random effects covariance matrix (Huang and Wand, 2013)

$$\begin{aligned} \boldsymbol{\Sigma}_{\mathrm{R}} \,|\, a_{\mathrm{R},1}, \ldots, a_{\mathrm{R},q^{\mathrm{R}}} &\sim \text{Inverse-Wishart}\left(\nu + q^{\mathrm{R}} - 1,\ 2\,\nu\,\text{diag}(1/a_{\mathrm{R},1} \ldots, 1/a_{\mathrm{R},q^{\mathrm{R}}})\right), \\ a_{\mathrm{R},r} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2}, 1/A_{\mathrm{R}}^2), \quad 1 \le r \le q^{\mathrm{R}}, \end{aligned} \tag{5.3}$$

with hyperparameters $\nu, A, A_{\mathrm{R}} > 0$. The value $\nu = 2$ corresponds to the correlation parameters having uniform distributions over $(-1,\, 1)$ and the standard deviation parameters having Half-t distributions with 2 degrees of freedom.

It will soon become apparent that the model framework introduced in (5.2) encompasses a wide range of logistic mixed models with different types of group structures by simply changing $\boldsymbol{G}$.

5.2.1 Random intercept and slope model

A natural starting point for these caesarean section data is to assume that the overall mean and deviation of the ith hospital from that overall mean are simply straight lines. This leads to the standard random intercept and slope model (Laird and Ware, 1982), with $f(x)$ and $g_i(x)$ modelling through **linear functions**:

$$\begin{aligned} f(x) &= \beta_0 + \beta_1 x_{ij} \ ; \ g_i(x) = U_{0i} + U_{1i} x_{ij} \\ \text{and} \quad \boldsymbol{G} &= \boldsymbol{I}_m \otimes \boldsymbol{\Sigma}_{\mathrm{R}} = \underset{1 \le i \le m}{\text{blockdiag}} \begin{bmatrix} \sigma_{u_0}^2 & \sigma_{u_0,u_1} \\ \sigma_{u_0,u_1} & \sigma_{u_1}^2 \end{bmatrix}, \end{aligned} \tag{5.4}$$

where β_0 and β_1 are the respective overall intercept and slope, and U_{0i} and U_{1i} are the hospital-specific deviations from that overall intercept and slope, being treated as a random

sample from a bivariate normal distribution with an unstructured 2×2 covariance matrix $\boldsymbol{\Sigma}_{\mathrm{R}}$. The model accounts for possible variability in the intercepts $\sigma^2_{u_0}$ and slopes $\sigma^2_{u_1}$ of hospitals and allows for an intercept-slope correlation ρ_{u_0,u_1}. In common practice, we generalise the fixed and random components of the mixed models to arbitrary general design matrices (Section 2 of Zhao *et al.* (2006)). This allows one to take advantage of the ever-expanding methods and software for inference in these models. Henceforth, we rewrite model (5.4) in matrix notation as

$$\begin{aligned} \boldsymbol{X}_i &\equiv \begin{bmatrix} 1 & x_{i1} \\ \vdots & \vdots \\ 1 & x_{in_i} \end{bmatrix}, \quad \boldsymbol{X} \equiv \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_m \end{bmatrix}, \quad \boldsymbol{\beta} \equiv \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \qquad (5.5) \\ \boldsymbol{Z} &\equiv \begin{bmatrix} \boldsymbol{X}_1 & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \boldsymbol{X}_m \end{bmatrix} \quad \text{and} \quad \boldsymbol{u} \equiv \begin{bmatrix} u_{01} \\ u_{11} \\ \vdots \\ u_{0m} \\ u_{1m} \end{bmatrix}. \end{aligned}$$

Noting that the random effects design matrix $\boldsymbol{Z}$ has a block-diagonal form, where each block corresponds to the ith hospital-specific deviations from the overall mean and slope.

We now combine the ideas introduced in Chapter 1 Subsection 1.5.3 to develop a scalable iterative algorithm for approximate Bayesian inference in model (5.4). We first briefly introduce the following useful notation: for a scalar random variable θ, let $\mu_{q(\theta)} \equiv E_q(\theta)$ and $\sigma_{q(\theta)} \equiv \mathrm{Var}_q(\theta)$ be the mean and variance with respect to the approximating q-distribution. For a random vector parameter $\boldsymbol{\theta}$, we use the analogously defined $\boldsymbol{\mu}_{q(\boldsymbol{\theta})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}$. In addition, we define $\boldsymbol{M}_{q(\boldsymbol{\Theta})}$ to be the mean with respect to the approximating q-distribution for a random matrix $\boldsymbol{\Theta}$. The journey towards a practical MFVB algorithm commences with a q-density product restriction

$$p(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{G}, a_{\mathrm{R},1}, a_{\mathrm{R},2} | \boldsymbol{y}) \approx q(\boldsymbol{\beta}, \boldsymbol{u})\ q(\boldsymbol{G})\ q(a_{\mathrm{R},1})\ q(a_{\mathrm{R},2}),$$

where $a_{\mathrm{R},1}$ and $a_{\mathrm{R},2}$ are auxiliary parameters defined in (5.3). Such a restriction arises from an interaction between the initial factorisation assumed in the approximating posterior and the underlying conditional independence properties of the true joint posterior (Bishop, 2006).

To provide an example of how optimal q-densities are constructed, we derive the optimal density $q^*(\boldsymbol{\beta}, \boldsymbol{u})$. According to Algorithm 2, the optimal q-density for $(\boldsymbol{\beta}, \boldsymbol{u})$ satisfies

$$q^*(\boldsymbol{\beta}, \boldsymbol{u}) \propto E_q\{\log p(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{y}, \boldsymbol{G}, a_{\mathrm{R},1}, a_{\mathrm{R},2})\} \qquad (5.6)$$

$$\begin{aligned} &\propto E_q\{\log p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u}) \times \log p(\boldsymbol{u}|\boldsymbol{G}) \times \log p(\boldsymbol{\beta})\} \\ &\propto E_q\big[\boldsymbol{y}^\mathsf{T}(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u}) - \mathbf{1}^\mathsf{T}\log\{\mathbf{1}+\exp(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u})\} - \tfrac{1}{2}\boldsymbol{u}^\mathsf{T}\boldsymbol{G}^{-1}\boldsymbol{u} - \tfrac{1}{2\sigma_\beta^2}\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\beta}\big]. \end{aligned}$$

We see that the non-quadratic convex term $-\log(1+e^x)$ in the likelihood poses a multivariate intractability problem with regard to approximate inference for $(\boldsymbol{\beta},\boldsymbol{u})$. To get around this, we transform this convex term to a simple quadratic function $f(x;\xi)$, a trick first introduced in Jaakkola and Jordan (1997). Different values of ξ correspond to different parabolas, all of which are smaller than $-\log(1+e^x)$. Thus we can simplify the convex term to be the maxima of a family of parabolas (Jaakkola and Jordan, 1997)

$$-\log(1+e^x) = \max_{\xi\in\mathbb{R}}\left\{\lambda(\xi)\,x^2 - \tfrac{1}{2}x + \psi(\xi)\right\} \quad \text{for all} \quad x\in\mathbb{R}, \tag{5.7}$$

where $\lambda(\xi) \equiv -\tanh(\xi/2)/(4\xi)$ and $\psi(\xi) \equiv \xi/2 - \log(1+e^\xi) + \xi\tanh(\xi/2)/4$. Let $\boldsymbol{C} \equiv [\boldsymbol{X}\ \boldsymbol{Z}]$ and $\boldsymbol{v} \equiv [\boldsymbol{\beta}^\mathsf{T}\ \boldsymbol{u}^\mathsf{T}]^\mathsf{T}$ and substitute (5.7) into (5.6) gives the following lower bound on $q^*(\boldsymbol{\beta},\boldsymbol{u})$:

$$\begin{aligned} q^*(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi}) &\geq \boldsymbol{y}^\mathsf{T}\boldsymbol{C}\boldsymbol{v} - \left[\boldsymbol{v}^\mathsf{T}\boldsymbol{C}^\mathsf{T}\mathrm{diag}\{\lambda(\boldsymbol{\xi})\}\boldsymbol{C}\boldsymbol{v} - \frac{1}{2}\mathbf{1}^\mathsf{T}\boldsymbol{C}\boldsymbol{v}\right] - \frac{1}{2}\boldsymbol{v}^\mathsf{T}\begin{bmatrix}\sigma_\beta^{-2}\boldsymbol{I} & \mathbf{0}\\ \mathbf{0} & E_q(\boldsymbol{G}^{-1})\end{bmatrix}\boldsymbol{v} \\ &= -\frac{1}{2}\left\{\boldsymbol{v}^\mathsf{T}\left(2\boldsymbol{C}^\mathsf{T}\mathrm{diag}\{\lambda(\boldsymbol{\xi})\}\boldsymbol{C} + \begin{bmatrix}\sigma_\beta^{-2}\boldsymbol{I} & \mathbf{0}\\ \mathbf{0} & E_q(\boldsymbol{G}^{-1})\end{bmatrix}\right)\boldsymbol{v} - \boldsymbol{C}^\mathsf{T}\left(\boldsymbol{y}-\frac{1}{2}\mathbf{1}\right)\right\} \\ &= \log\underline{q}^*(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi}), \end{aligned} \tag{5.8}$$

where $E_{q(\boldsymbol{G}^{-1})} \equiv \boldsymbol{I}_m \otimes \boldsymbol{M}_{q(\boldsymbol{\Sigma}_\mathrm{R}^{-1})}$ and $\boldsymbol{\xi}$ is being introduced as an $\sum_{i=1}^m n_i \times 1$ vector of variational parameters. Scalar functions applied to vectors are evaluated element-wise. Noting that since the right hand side of (5.8) is a quadratic form and, by completing the square in the usual way to identify the mean and covariance, we approximate the posterior of $(\boldsymbol{\beta},\boldsymbol{u})$ by a multivariate normal distribution

$$\begin{aligned} q^*(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi}) &\sim N(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}), \\ \text{where}\quad \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})} &\equiv \left(2\,\boldsymbol{C}^\mathsf{T}\mathrm{diag}\{\lambda(\boldsymbol{\xi})\}\boldsymbol{C} + \begin{bmatrix}\sigma_\beta^{-1}\boldsymbol{I} & \mathbf{0}\\ \mathbf{0} & E_q(\boldsymbol{G}^{-1})\end{bmatrix}\right)^{-1} \\ \text{and}\quad \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})} &\equiv \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}\boldsymbol{C}^\mathsf{T}(\boldsymbol{y}-\tfrac{1}{2}\mathbf{1}). \end{aligned}$$

The remaining q-densities involve similar adaptation of the above derivation and hence we

omit the details and state their functional forms directly:

$$\boldsymbol{\xi} \leftarrow \sqrt{\text{diagonal}\{\boldsymbol{C}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})} + \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}\boldsymbol{\mu}^{\mathsf{T}}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})})\boldsymbol{C}^{\mathsf{T}}\}},$$
$$q^*(\boldsymbol{\Sigma}_{\text{R}}) \text{ is the Inverse-Wishart } \left(\nu + m + 1, \boldsymbol{B}_{q(\boldsymbol{\Sigma}_{\text{R}})}\right) \text{ density function,}$$
$$q^*(a_{\text{R},r}) \text{ is the Inverse-Gamma } \left(1, B_{q(a_{\text{R},r})}\right) \text{ density function, } 1 \leq r \leq 2,$$

where the parameter $\boldsymbol{B}_{q(\boldsymbol{\Sigma}_{\text{R}})}$ is the scale matrix of the inverse-Wishart q-density and $B_{q(a_{\text{R},r})}$ is the scale parameter of the inverse-Gamma q-density. Thus all optimal densities belong to parametric families with the parameters explicitly determined by the distributions of the remaining model parameters and observed data. Taken together, these solutions lead to Algorithm 15 for approximate Bayesian inference in model (5.4).

Initialise: $\boldsymbol{\xi}$ ($\sum_{i=1}^{m} n_i \times 1$; all entries positive), $\boldsymbol{M}_{q(\boldsymbol{\Sigma}_{\text{R}}^{-1})}$ positive definite and $\mu_{q(a_{\text{R},r})} > 0$, $r = 1, 2$

Cycle through updates:

Define: $E_{q(\boldsymbol{G}^{-1})} \equiv \boldsymbol{I}_m \otimes \boldsymbol{M}_{q(\boldsymbol{\Sigma}_{\text{R}}^{-1})}$

Update mean and covariance matrix of multivariate normal $q^*(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\xi})$:

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})} \leftarrow \left(2\, \boldsymbol{C}^{\mathsf{T}}\text{diag}\left\{\lambda(\boldsymbol{\xi})\right\}\boldsymbol{C} + \begin{bmatrix} \sigma_{\beta}^{-2}\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & E_{q(\boldsymbol{G}^{-1})} \end{bmatrix}\right)^{-1}$$

$$\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})} \leftarrow \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}\boldsymbol{C}^{\mathsf{T}}(\boldsymbol{y} - \tfrac{1}{2}\boldsymbol{1})$$

$$\boldsymbol{\xi} \leftarrow \sqrt{\text{diagonal}\{\boldsymbol{C}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})} + \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}\boldsymbol{\mu}^{\mathsf{T}}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})})\boldsymbol{C}^{\mathsf{T}}\}}$$

Update mean and scale matrix of inverse-Wishart $q^*(\boldsymbol{\Sigma}_{\text{R}})$:

$$\boldsymbol{B}_{q(\boldsymbol{\Sigma}_{\text{R}})} \leftarrow \sum_{i=1}^{m}\left(\mu_{q(\boldsymbol{u}_i)}\mu_{q(\boldsymbol{u}_i)} + \boldsymbol{\Sigma}_{q(\boldsymbol{u}_i)}\right) + 2\,\nu\,\text{diag}(\mu_{q(1/a_{\text{R},1})},\, \mu_{q(a_{\text{R},2})})$$

$$\boldsymbol{M}_{q(\boldsymbol{\Sigma}_{\text{R}}^{-1})} \leftarrow \tfrac{1}{2}(\nu + m + 1)\boldsymbol{B}^{-1}_{q(\boldsymbol{\Sigma}_{\text{R}})}$$

Update mean and scale parameter of inverse-Gamma $q^*(a_{\boldsymbol{R},r})$:

For $r = 1, 2$:

$$B_{q(a_{\text{R},r})} \leftarrow \nu(\boldsymbol{M}_{q(\boldsymbol{\Sigma}_{\text{R}}^{-1})})_{rr} + 1/A_{\text{R}}^2 \quad ; \quad \mu_{q(a_{\text{R},r})} \leftarrow \tfrac{1}{2}(\nu + 1)/B_{q(a_{\text{R},r})}$$

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

Algorithm 15: Iterative scheme for obtaining the optimal q-density functions in the Bayesian logistic mixed model (5.4).

Algorithm 15 is a basic framework for approximate inference for a simple Bayesian logistic mixed model that serves as a building block for the more complicated models. In what follows, we will gradually increase the complexity of model (5.4) and demonstrate a

great advantage of the MFVB algorithms, *modularity*. By this we mean that concepts like main effects, interaction effects, higher-order random effects and spline regression can be viewed as modules that can be put together into an almost endless variety of statistical models. All that required is relatively straightforward modifications on the structures of the general design matrices and random effects covariance matrix in order to accommodate larger and more complicated mixed models.

5.2.2 Group-specific curve model

Close inspection of Figure 5.1 shows that the straight line assumption imposed by the random intercept and slope model is unreasonable. We seek to relax this linearity assumption by replacing the linear mean and hospital-specific functions with **arbitrary smooth functions**, hereafter group-specific curve model, taking form (5.1), but with

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K_{\text{gbl}}} u_{\text{gbl},k} z_{\text{gbl},k}(x) \;\; ; \;\; g_i(x) = U_{0i} + U_{1i} x + \sum_{k=1}^{K_{\text{grp}}} u_{\text{grp},ik} z_{\text{grp},ik}(x)$$

$$\text{and} \quad \boldsymbol{G} = \begin{bmatrix} \overbrace{\sigma^2_{\text{gbl}} \boldsymbol{I}_{K_{\text{gbl}}}}^{\text{Overall mean}} & \overbrace{\boldsymbol{0}}^{\text{Hospital-specific}} \\ \boldsymbol{0} & \underset{1 \le i \le m}{\text{blockdiag}} \left(\boldsymbol{\Sigma}_{\text{R}} | \sigma^2_{\text{grp}} \boldsymbol{I}_{K_{\text{grp}}} \right) \end{bmatrix}. \tag{5.9}$$

This model is a direct extension of the random intercept and slope model since in (5.9) the overall mean curve and each hospital-specific curve have two components: a linear part (analogous to model (5.4)) and a non-linear part, which allows more flexibility.

There are numerous approaches to modelling and estimating f and g_i nonparametrically. The one which is most conducive to inference via variational methods is penalised regression splines with mixed model representation. An advantage of this approach is that one can think in terms of constructing a series of basis functions that can be used as covariates, such as $z_{\text{gbl},k}$ and $z_{\text{grp},ik}$ being the respective spline bases of size K_{gbl} and K_{grp}. Our preference of $z_{\text{gbl},k}$ and $z_{\text{grp},ik}$ is suitably linearly transformed cubic O'Sullivan penalised splines (Section 4 of Wand and Ormerod (2008)), since this leads to approximate smoothing splines with good boundary and extrapolation properties. In practice, $K_{\text{gbl}} = 25$ is a sufficient choice for most spline basis functions (Li and Ruppert, 2008) and typically K_{grp} is smaller than K_{gbl} since fewer basis functions are needed to handle group-specific deviations. The coefficients $u_{\text{gbl},k}$ and $u_{\text{grp},ik}$ can be considered as a measure of the basis amplitude since they regulate the roughness of the time curves. In order to avoid overfitting the data, we impose a penalty on the basis coefficients by treating them as a random sample from a normal distribution with mean 0 and variance σ^2, i.e. $u_{\text{gbl},k} \overset{\text{ind.}}{\sim} N(0, \sigma^2_{\text{gbl}})$ and $u_{\text{grp},ik} \overset{\text{ind.}}{\sim} N(0, \sigma^2_{\text{grp}})$ respectively. The variances σ^2_{gbl} and σ^2_{grp} are

commonly known as the *smoothing parameters*, thus selection of smoothing parameters in model (5.9) simply reduces to variance component estimation in a mixed effect model.

From an extendability standpoint, a general design framework for mixed effect models is a particularly attractive approach that allows smoothing-type models such as group-specific curve models to be fitted as generalized linear mixed models. Moving from parametric regression to nonparametric regression using mixed model based penalised splines involves replacing $\boldsymbol{Z}$

$$\text{from} \quad \begin{array}{c} \begin{matrix} \text{Hospital 1} & & \text{Hospital } m \end{matrix} \\ \begin{bmatrix} \boldsymbol{X}_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \boldsymbol{X}_m \end{bmatrix} \end{array} \quad \text{to} \quad \begin{array}{c} \begin{matrix} \text{Overall mean} & \text{Hospital 1} & \cdots & \text{Hospital } m \end{matrix} \\ \begin{bmatrix} & \boldsymbol{X}_1 | \boldsymbol{Z}_{\text{grp},1} & \cdots & \mathbf{0} \\ \boldsymbol{Z}_{\text{gbl}} & \vdots & \ddots & \vdots \\ & \mathbf{0} & \cdots & \boldsymbol{X}_m | \boldsymbol{Z}_{\text{grp},m} \end{bmatrix} \end{array}, \tag{5.10}$$

where $\boldsymbol{Z}_{\text{gbl}}$ is an $\sum_{i=1}^{m} n_i \times K_{\text{gbl}}$ random spline basis matrix for the overall mean curve and $\boldsymbol{Z}_{\text{grp},i}$ is an $n_i \times K_{\text{grp}}$ random spline basis matrix for the ith hospital-specific deviation. Further, even with the additional $K_{\text{gbl}} + mK_{\text{grp}}$ random effects parameters, the covariance matrix $\boldsymbol{G}$ shown in (5.9) remains to be a block-diagonal, but certainly larger, matrix with different entries for the variance components corresponding to the random basis coefficients for the overall mean and random effects for each hospital. Taking these into consideration, it then follows that the MFVB algorithm for model (5.9) simply involves modifications of Algorithm 16 to incorporate the above-mentioned structural changes in the $\boldsymbol{Z}$ and $\boldsymbol{G}$ matrices, as well as updates for the additional model parameters as follows:

$$\begin{aligned}
&q^*(a_{\text{gbl}}) \text{ is the Inverse-Gamma } \left(1, B_{q(a_{\text{gbl}})}\right) \text{ density function;} \\
&q^*(\sigma^2_{\text{gbl}}) \text{ is the Inverse-Gamma } \left(\tfrac{1}{2}(K_{\text{gbl}}+1), B_{q(\sigma^2_{\text{gbl}})}\right) \text{ density function;} \\
&q^*(a_{\text{grp}}) \text{ is the Inverse-Gamma } \left(1, B_{q(a_{\text{grp}})}\right) \text{ density function; and} \\
&q^*(\sigma^2_{\text{grp}}) \text{ is the Inverse-Gamma } \left(\tfrac{1}{2}(m\,K_{\text{grp}}+1), B_{q(\sigma^2_{\text{grp}})}\right) \text{ density function,}
\end{aligned}$$

where $B_{q(a_{\text{gbl}})}, B_{q(a_{\text{grp}})}, B_{q(\sigma^2_{\text{gbl}})}$ and $B_{q(\sigma^2_{\text{grp}})}$ are the scale parameters with respect to the corresponding q-density functions.

5.2.3 Factor-by-curve interactions

The second objective of this caesarean section analysis is to examine differences in trends in caesarean section rates between low-risk nulliparous women aged less than 25 years and those aged greater than or equal to 25 years. We are therefore interested in fitting a separate time curve for each maternal age group (see Figure 5.2). This leads to a group-specific curve model with a *factor-by-curve interaction*, where one can view the effect of

maternal age on the probability of caesarean section varies smoothly over time. From here onwards we refer low-risk nulliparous women aged less than 25 years as Group A (younger group) and those greater or equal to 25 years as Group B (older group).

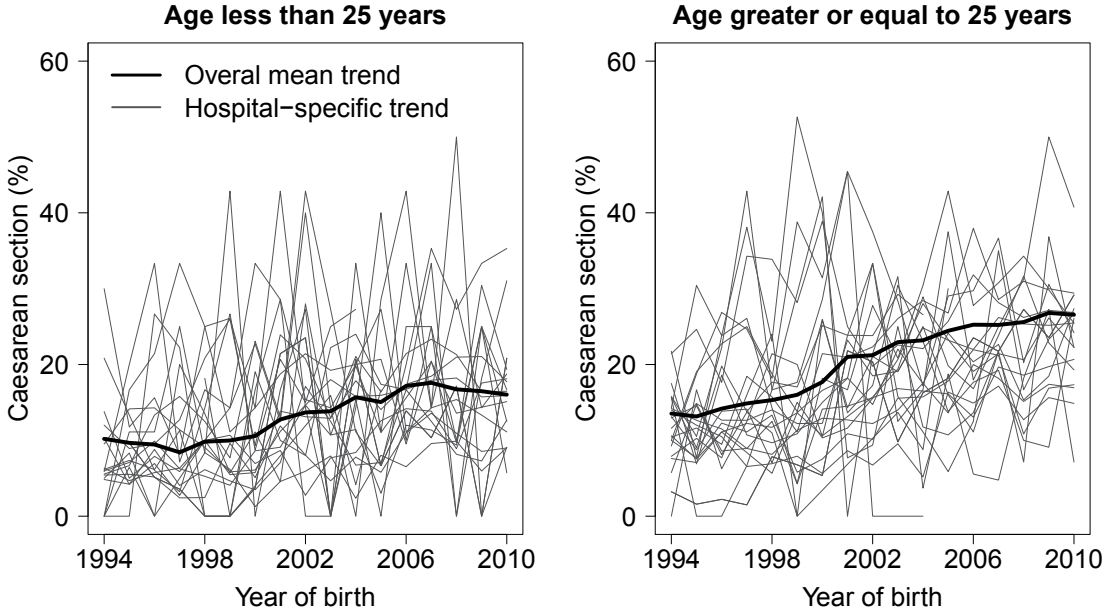

Figure 5.2: Trends in caesarean section rates for low-risk nulliparous women aged less than 25 years and those aged greater or equal to 25 years in New South Wales, Australia, from 1994 to 2010. The dotted line is the overall mean trend and the solid lines are the selected hospital-specific trends.

Additive models that include factor-by-curve interactions are known to be useful but can be computationally very demanding. Coull *et al.* (2000) used a backfitting algorithm, in which, for each backfit iteration corresponding to a curve interaction term, one splits the data into subsets and fits a smooth function to each subset. However, this backfitting approach can be computationally impractical when the number of interaction terms or the number of data subsets become large. Maringwa *et al.* (2008) used penalised regression splines in a mixed model framework to compare population-averaged profiles, but only focused on the random intercept models without any group-specific curves. The penalised spline approach is preferable since the former data subsetting approach must be nested within a backfitting algorithm. We follow on Maringwa *et al.* (2008) and incorporate a factor-by-curve interaction into the group-specific curve model (5.9). Let $I_{ij}^{\mathrm{A}} = 1$ if (x_{ij}, y_{ij}) belongs to a low-risk nulliparous women of aged less than 25 years and zero otherwise, we propose a model of the following form

$$(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u})_{ij} \quad = \quad I_{ij}^{\mathrm{A}} \{f^{\mathrm{A}}(x_{ij}) + g_i^{\mathrm{A}}(x_{ij})\} + (1 - I_{ij}^{\mathrm{A}}) \{f^{\mathrm{B}}(x_{ij}) + g_i^{\mathrm{B}}(x_{ij})\},$$

where

$$
\begin{aligned}
f^{\mathrm{A}}(x) &= \beta_0^{\mathrm{A}} + \beta_1^{\mathrm{A}}\, x + \sum_{k=1}^{K_{\mathrm{gbl}}} u_{\mathrm{gbl},k}^{\mathrm{A}}\; z_{\mathrm{gbl},k}(x) \qquad (5.11)\\
g_i^{\mathrm{A}}(x) &= U_{0i}^{\mathrm{A}} + U_{1i}^{\mathrm{A}}\, x \sum_{k=1}^{K_{\mathrm{grp}}} u_{\mathrm{grp},ik}^{\mathrm{A}}\; z_{\mathrm{grp},k}(x),\\
f^{\mathrm{B}}(x) &= \beta_0^{\mathrm{A}} + \beta_0^{\mathrm{BvsA}} + (\beta_1^{\mathrm{A}} + \beta_1^{\mathrm{BvsA}})\, x + \sum_{k=1}^{K_{\mathrm{gbl}}} u_{\mathrm{gbl},k}^{\mathrm{B}}\; z_{\mathrm{gbl},k}(x)\\
g_i^{\mathrm{B}}(x) &= U_{0i}^{\mathrm{B}} + U_{1i}^{\mathrm{B}}\, x \sum_{k=1}^{K_{\mathrm{grp}}} u_{\mathrm{grp},ik}^{\mathrm{B}}\; z_{\mathrm{grp},k}(x)
\end{aligned}
$$

$$
\text{and} \quad \boldsymbol{G} = \begin{bmatrix} \overbrace{\begin{matrix} (\sigma_{\mathrm{gbl}}^{\mathrm{A}})^2 \boldsymbol{I}_{K_{\mathrm{gbl}}} & \mathbf{0} \\ \mathbf{0} & (\sigma_{\mathrm{gbl}}^{\mathrm{B}})^2 \boldsymbol{I}_{K_{\mathrm{gbl}}} \end{matrix}}^{\text{Overall mean}} & \overbrace{\begin{matrix} \mathbf{0} \\ \mathbf{0} \end{matrix}}^{\text{Hospital-specific}} \\ \mathbf{0} \quad \mathbf{0} & \underset{1\le i\le m}{\mathrm{blockdiag}} \left(\boldsymbol{\Sigma}_{\mathrm{R}} | \sigma_{\mathrm{grp}}^2 \boldsymbol{I}_{K_{\mathrm{grp}}} | \sigma_{\mathrm{grp}}^2 \boldsymbol{I}_{K_{\mathrm{grp}}} \right) \end{bmatrix}.
$$

We consider low-risk women aged less than 25 years as our reference group and define the corresponding f^{A} and g_i^{A} functions the same way as in model (5.9). For low-risk women aged greater or equal to 25 years, the f^{B} function is defined slightly differently by introducing the parameters β_0^{BvsA} and β_1^{BvsA} that represent the differences in overall intercepts and slopes between low-risk nulliparous women of younger and older age. This gives an *asymmetrical structure* in the fixed effects formulation. While asymmetrical formulation of fixed effects is common in the presence of a reference group, we do not recommend such a structure imposed on the random spline coefficients since it would induce restriction on the smoothness of a particular curve. We impose a different variance parameter for each overall mean curve, allowing the level of smoothing to differ between curves. A simpler model would be to assume a common variance parameter for all overall mean curves, i.e. $\sigma_{\mathrm{gbl}}^{\mathrm{A}} = \sigma_{\mathrm{gbl}}^{\mathrm{B}} = \sigma_{\mathrm{gbl}}$, meaning that all curves have equivalent smoothness, but with different forms of shape. In addition, we account for the potential correlation between the random intercepts and slopes of each of the women age groups within the same hospital using an unstructured 4×4 covariance matrix, i.e. $[U_{0i}^{\mathrm{A}}\, U_{1i}^{\mathrm{A}}\, U_{0i}^{\mathrm{B}}\, U_{1i}^{\mathrm{A}}]^{\mathsf{T}} \overset{\text{ind.}}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma}_{\mathrm{R}})$.

The differences in trends in caesarean section rates for low-risk nulliparous women aged greater or equal to 25 years versus those aged less than 25 years can be quantified via estimation of a so-called *contrast function*. At the population level, this is simply obtained

by subtracting the two overall mean curves:

$$c^{\text{BvsA}}(x) \equiv f^{\text{B}}(x) - f^{\text{A}}(x) = \beta_0^{\text{BvsA}} + \beta_1^{\text{BvsA}}\, x + \sum_{k=1}^{K_{\text{gbl}}} (u^{\text{B}}_{\text{gbl},k} - u^{\text{A}}_{\text{gbl},k}) z_{\text{gbl},k}(x). \quad (5.12)$$

This contrast function can be interpreted as the log odds of caesarean section (averaged across hospitals) for low-risk nulliparous women aged greater or equal to 25 years, as a function of time, compared with those aged less than 25 years.

The structural changes in the random effects matrices $\boldsymbol{Z}$ and $\boldsymbol{G}$ underpin the complexity of a mixed effect model being transitioned from the simplest random intercept and slope model to a more complicated model including a factor-by-curve interaction. For example, Figure 5.3 shows that the $\boldsymbol{Z}$ matrix starts off as having a simple block-diagonal structure and, as the model increases in complexity, it grows into a "nested" block-diagonal structure (Each large block is itself block-diagonal, with one or more small blocks on the diagonal). As the number of random spline basis functions increases in the model, the dimensions of the $\boldsymbol{Z}$ and $\boldsymbol{G}$ matrices increase accordingly. While these changes do not directly affect the algebraic aspect of Algorithm 16 because of the advantage of modularity, one must be aware that the update of the covariance matrix $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}$ now requires storage and inversion of a large and sparse matrix. Specifically, the number of columns in $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}$ for each considered model is

1.	Random intercepts and slopes	$P + m\, q^{\text{R}}$
2.	Group-specific curves	$P + K_{\text{gbl}} + m\,(q^{\text{R}} + K_{\text{grp}})$
3.	Factor-by-curve interactions	$P + 2K_{\text{gbl}} + 2\,m\,(q^{\text{R}} + K_{\text{grp}})$

Without doubt the number of groups m dominates the dimension of $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}$. For example, if $P = 10, m = 1000, K_{\text{gbl}} = 25$, $q^{\text{R}} = 2$ and $K_{\text{grp}} = 10$, then the dimension of the matrix requiring storage and inversion for the most complicated model would be $24\,060 \times 24\,060$. This aspect renders Algorithm 16 infeasible for very large and complex longitudinal and multilevel datasets. As pointed out by by Lee and Wand (2015a), the naïve implementation of MFVB algorithms for arbitrarily large grouped data is extremely inefficient in terms of speed and storage. The time complexity can be as high as $O(m^3)$, making variational inference impractical for large and complex mixed models. Through exploiting the inherent block-diagonal structure of the random effects covariance matrix, the authors developed fast and memory-efficient MFVB algorithms that streamline inversion and update for $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}$. Here we briefly reiterate their streamlined approach in more general terms. Recall that the general update expression for $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}$ in Algorithm

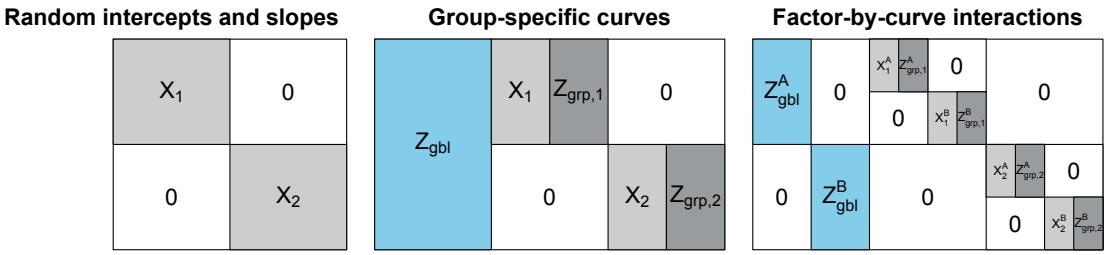

Figure 5.3: Various structures of the $\boldsymbol{Z}$ matrix across logistic mixed models (5.4), (5.9) and (5.11), with the number of groups $m = 2$. The $\boldsymbol{Z}$ matrix starts off as having a simple block-diagonal structure and, as the model increases in complexity, it grows into a "nested" block-diagonal structure. The definitions of matrices are described in Subsection 5.2.2.

15 involves inversion of the following matrix

$$2\,\boldsymbol{C}^{\mathsf{T}}\mathrm{diag}\left\{\lambda(\boldsymbol{\xi})\right\}\boldsymbol{C} + \begin{bmatrix} \sigma_\beta^{-2}\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & E_{q(\boldsymbol{G}^{-1})} \end{bmatrix} = \left[\begin{array}{c|cccc} \boldsymbol{U} & \boldsymbol{V}_1 & \boldsymbol{V}_2 & \cdots & \boldsymbol{V}_m \\ \hline \boldsymbol{V}_1^{\mathsf{T}} & \boldsymbol{W}_1^{-1} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{V}_2^{\mathsf{T}} & \boldsymbol{0} & \boldsymbol{W}_2^{-1} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{V}_m^{\mathsf{T}} & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{W}_m^{-1} \end{array}\right] \tag{5.13}$$

The central idea of the streamlined approach requires permuting the above matrix into an approximate block-diagonal form for decomposition, as shown in (5.13). The matrices $\boldsymbol{U}, \boldsymbol{V}_i$ and $\boldsymbol{W}_i^{-1}$ are usually of general forms with small dimension (Appendix of (Lee and Wand, 2015a)); each can be easily derived via straightforward matrix manipulation. With this transformation we can efficiently invert the block-partitioned form of (5.13) by solving a system of simultaneous linear equations in matrix form (Smith and Wand, 2008)

$$\overbrace{\begin{bmatrix} \boldsymbol{U} & \boldsymbol{V} \\ \boldsymbol{V}^{\mathsf{T}} & \boldsymbol{W} \end{bmatrix}}^{\boldsymbol{\Sigma}^{-1}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}} \overbrace{\begin{bmatrix} \boldsymbol{X} & \boldsymbol{Y} \\ \boldsymbol{Y}^{\mathsf{T}} & \boldsymbol{Z} \end{bmatrix}}^{\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix}.$$

Standard calculations lead to the following forms of submatrices:

$$\begin{aligned} \boldsymbol{X} &\equiv \left(\boldsymbol{U} - \boldsymbol{V}\boldsymbol{W}^{-1}\boldsymbol{V}^{\mathsf{T}}\right)^{-1} \\ \boldsymbol{Y} &\equiv -\boldsymbol{X}\boldsymbol{V}\boldsymbol{W}^{-1} \\ \text{and} \quad \boldsymbol{Z} &\equiv \boldsymbol{W}^{-1} + \boldsymbol{W}^{-1}\boldsymbol{V}^{\mathsf{T}}\boldsymbol{X}\boldsymbol{V}\boldsymbol{W}^{-1}, \end{aligned}$$

It is worth noting that the matrix $\boldsymbol{Z}$ is not a block-diagonal matrix. However, since the covariance between the fitted values of two different groups is rarely of interest, it suffices

to compute and store the diagonal blocks. Appendix 5.A provides details of the derivations for the streamlined MFVB algorithm for model (5.11).

5.3 Numerical evaluation

In this section, we conducted a comprehensive simulation study to evaluate the performance of Algorithm 15 in terms of Bayesian inferential accuracy and computational speed. We simulated 30 independent datasets from model (5.9) and used 25 knots for the overall mean function and 10 knots for the group-specific deviation functions:

$$\begin{aligned} y_{ij} &\overset{\text{ind.}}{\sim} \text{Bernoulli}\left(\text{logit}^{-1}\left\{f(x_{ij}) + g(x_{ij})\right\}\right), \\ \text{where} \quad f(x) &= 10\ \phi(x; 1.5, 0.3) + 6\ \phi(x; 4, 0.6), \\ g_i(x) &= \alpha \sin(2\pi x^{\beta}) \ \ ; \ \ x_{ij} \sim \text{Uniform}(0, 1/n) \\ \text{and} \quad \alpha &\sim N(0, 0.5) \ \ ; \ \ \beta \in \{1,\ 2,\ 3\}. \end{aligned} \tag{5.14}$$

Each dataset consists of $m = 50$ groups and the number of observations per group is balanced with $n_i = n = 50$. The term $\phi(x; a, b)$ is an univariate normal density with mean a and standard deviation b.

We fitted each replicated dataset using both MFVB approximation via a streamlined version of Algorithm 15 and MCMC sampling. The MFVB algorithm was implemented in the R computing environment (R Development Core Team, 2015) and its stopping criterion is when the relative change in the lower bound $\underline{p}(\boldsymbol{y}; q)$ falls below 10^{-8}. The MCMC samples were obtained using the R package `RStan` (Stan Development Team, 2015). In each dataset, MCMC samples of size 10 000 were generated, with the first 5000 values of each sample discarding as burn-in and the remaining 5000 values thinning by a factor of 5. Trace plots and the autocorrelation functions for all model parameters were used to assess MCMC convergence.

5.3.1 Assessment of accuracy

For a generic parameter θ we assessed the accuracy of the MFVB approximation by comparing the optimal q-density function $q^*(\theta)$ with a highly accurate MCMC-based posterior approximation $p_{\text{MCMC}}(\theta|\boldsymbol{y})$. While there are numerous means of measuring accuracy, we recommend working with the a measure that is based on the L_1 distance, also known as the *integrated absolute error* (Faes *et al.*, 2011) of $q^*(\theta)$, given by

$$\text{accuracy}\{q^*(\theta)\} \equiv 100 \left(1 - \tfrac{1}{2} \int_{-\infty}^{\infty} |q^*(\theta) - p_{\text{MCMC}}(\theta|\boldsymbol{y})| d\theta\right) \%.$$

This accuracy measure has the attraction of being invariant to monotone transformations on the parameter θ. Exact computation of $p_{\text{MCMC}}(\theta|\boldsymbol{y})$ is numerically challenging and hence we used binned kernel density estimation with direct plug-in bandwidth selection, as facilitated in the R package `KernSmooth` (Wand and Ripley, 2006).

Figures 5.4 and 5.5 display the approximate posterior density functions and side-by-side boxplots of accuracy scores for the model parameters $f(Q_k)$, $1 \leq k \leq 3$ and $g_i(Q_k)$, $k = 2$, where Q_k is the kth sample quintile of the xs. The boxplots show that the majority of the accuracy scores exceed 80%, with some over 90%. As indicated by Figure 5.4, the MFVB credible intervals have good coverage properties. However, interestingly the MFVB credible intervals tend to be narrower than that of the MCMC samples based on kernel density estimation. The average elapsed time for the MCMC fits is 7.1 hours (standard error 3.0 hours), whilst for the MFVB fits is 1.9 minutes (standard error 4.7 seconds). This corresponds to a speed-up in the order of several hundreds.

5.3.2 Assessment of speed

We now turn our attention to quantification of the speed gains afforded by the streamlined MFVB algorithm. The simulation described in Section 5.3 was re-run using MFVB with a combination of m and n values and MCMC omitted. All computations were performed on a Mac OS X laptop with a 2.6 GHz Intel Core i7 processor and 4 GBytes of random access memory. Table 5.1 summarises the average (standard error) computing times over 30 runs and highlights the practical and scalability benefits of the streamlined MFVB approach.

Number of groups (m)	100	500	1000
Within-group size (n)			
10	65.1 (0.86)	351.6 (5.49)	709.4 (4.42)
50	127.1 (2.57)	826.7 (9.63)	1859.2 (6.15)
100	106.5 (1.51)	929.3 (10.77)	2870.3 (13.62)

Table 5.1: Average (standard error) elapsed of the computing times in seconds for the streamlined MFVB Algorithm15 in the simulation setting described in (5.14).

It is well-established that MCMC can be very slow in situations where complex models are applied to large datasets. For the setting in Table 5.1, we expect the MCMC fitting takes days to weeks to run and therefore a similar timing comparison between MFVB and MCMC is not practical.

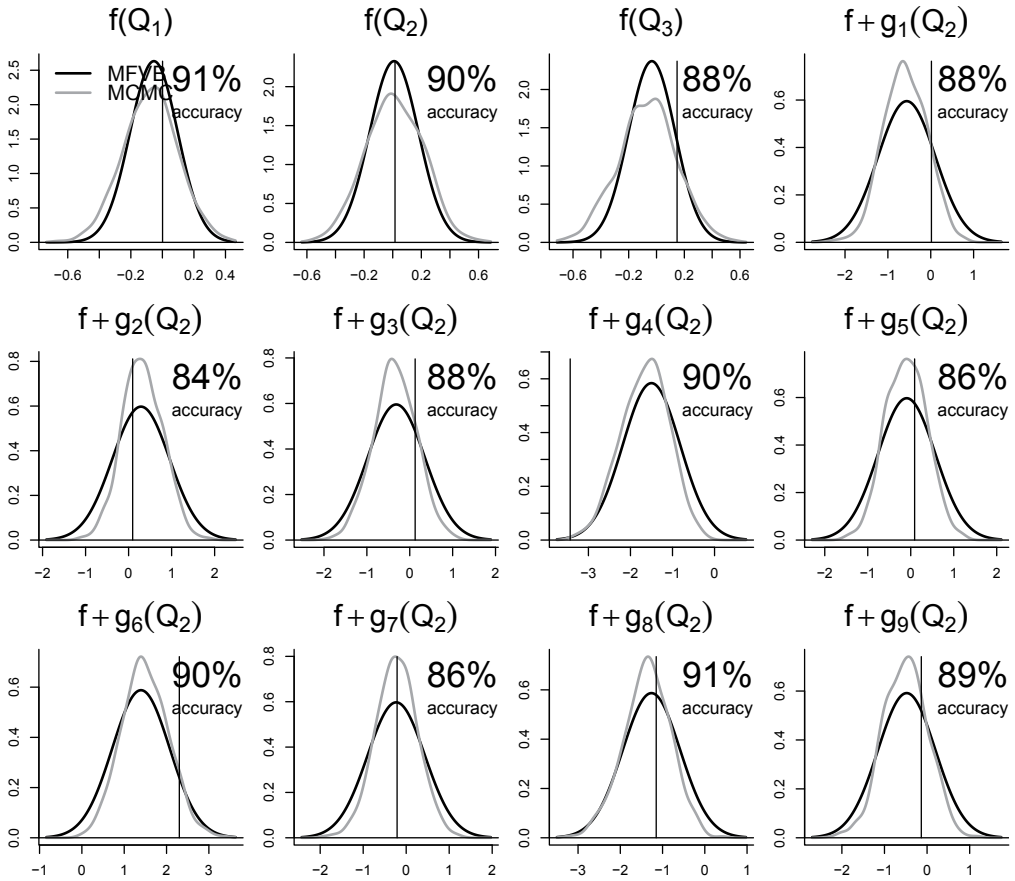

Figure 5.4: The approximate posterior density functions obtained from MFVB and MCMC for a single replication of the simulation study described in the text. Each pair of density function corresponds to a model parameter $f(Q_k)$, $1 \leq k \leq 3$ and $g_i(Q_k)$, $k = 2$, where Q_k is the kth sample quintile of the xs. The vertical lines represent the true parameter values. The accuracy scores on the top right of on each plot show the accuracy of MFVB approximation compared against a MCMC benchmark.

5.4 Application to caesarean section data

As described in Section 5.2, we applied our MFVB algorithm to the caesarean section data, with the aim of characterising trends in caesarean section rates for low-risk nulliparous women aged less than 25 years and those aged greater or equal to 25 years in NSW hospitals between 1994 and 2010. A group-specific curve model with a factor-by-curve interaction was fitted to data, allowing for non-linear estimation of time courses as seen in Figure 5.2.

From 1994 to 2010 there were 295 340 low-risk nulliparous women giving birth for the first time in 99 NSW public or private hospitals. Of which 73 795 women (24.5%) aged less than 25 years and 221 545 women (75.0%) aged greater or equal to 25 years. The annual rate of caesarean section among low-risk nulliparous women has increased radically from

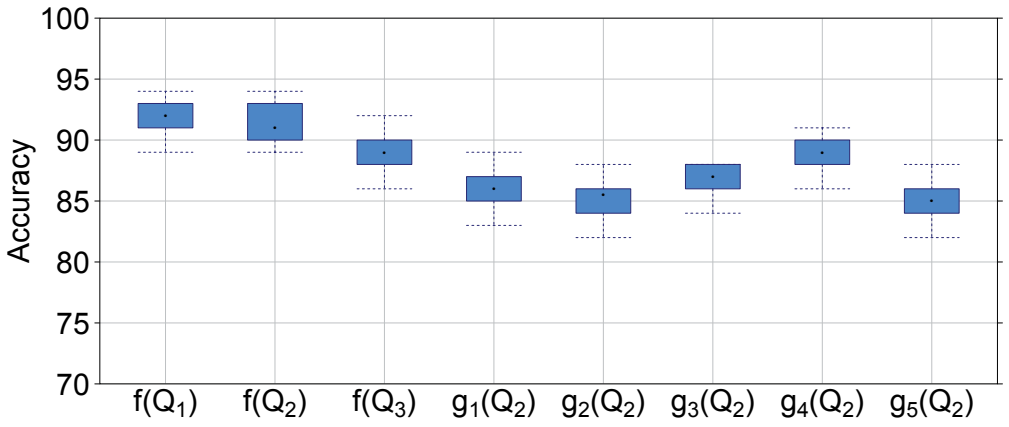

Figure 5.5: Side-by-side boxplots of accuracy scores for the MFVB approximation compared against MCMC over 30 runs. Each boxplot corresponds to a model parameter $f(Q_k)$, $1 \leq k \leq 3$ and $g_i(Q_k)$, $k = 2$, where Q_k is the kth sample quintile of the xs.

12.5% to 24.1% in NSW over this 17-year period; however, rates for women of older age were consistently higher than that of those of younger age (younger group: 10.2% to 16.0%; older group: 13.5% to 26.6%). Figure 5.6 shows the fitted probability function of caesarean section for each hospital between 1994 and 2010. Compared with low-risk women aged less than 25 years, those aged greater or equal to 25 years are more likely to have a caesarean section, rather than a vaginal delivery. The fitted probabilities are seen to vary markedly among hospitals and over time, with some hospitals exhibiting a consistently high probability and others exhibiting an upside down U shape. It is evident that the hospital-specific probability functions of caesarean section cannot be adequately modelled by linear functions.

Figure 5.7 shows the estimated overall and selected hospital-specific contrast curves defined by (5.12), corresponding to the odds of caesarean section for low-risk nulliparous women aged greater or equal to 25 years, as a function of time, compared with those aged less than 25 years. The pointwise 95% credible sets, shown in the first panel by the grey shaded region, suggest that there are significant differences in trends in caesarean section rates between the two maternal age groups over the year of birth. Low-risk nulliparous women of older age are on average 1.3 times more likely to have a caesarean section compared with those of younger age in the year 1994. This odds ratio increases as the year of birth increases.

To examine hospital differences in odds of caesarean section for low-risk nulliparous women of older age compared with those of younger age, we calculated and plotted the hospital-specific odds ratios for the latest year of birth. In 2010, the statewide odds ratio is 1.74, meaning that, on average, a low-risk nulliparous women of older age giving birth by caesarean section is about 1.74 times more likely than that of a women of younger age.

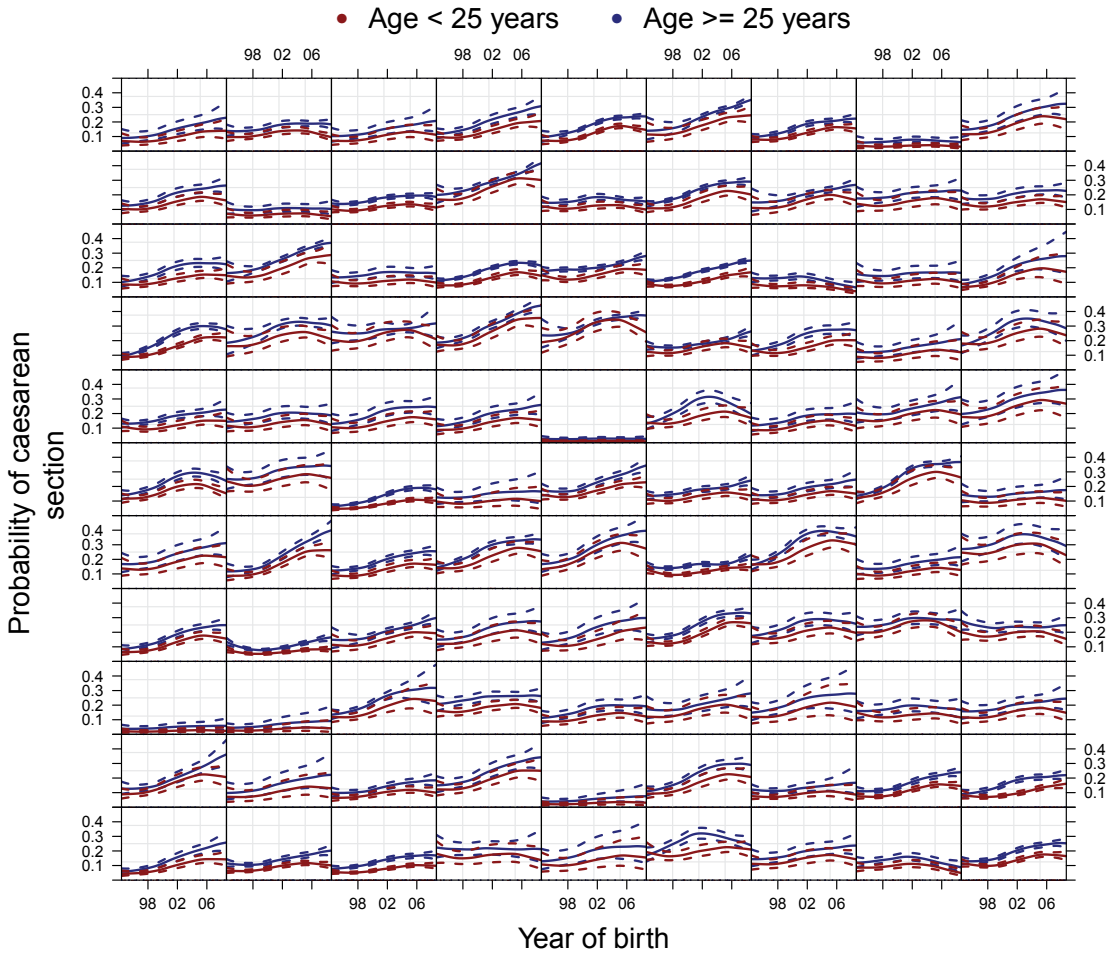

Figure 5.6: The MFVB fitted hospital-specific probability functions of caesarean section for low-risk nulliparous women of aged less than 25 years and those of aged greater or equal to 25 years, as a function of time for each hospital. The dashed curves represent pointwise 95% credible sets. Each panel corresponds to a different hospital.

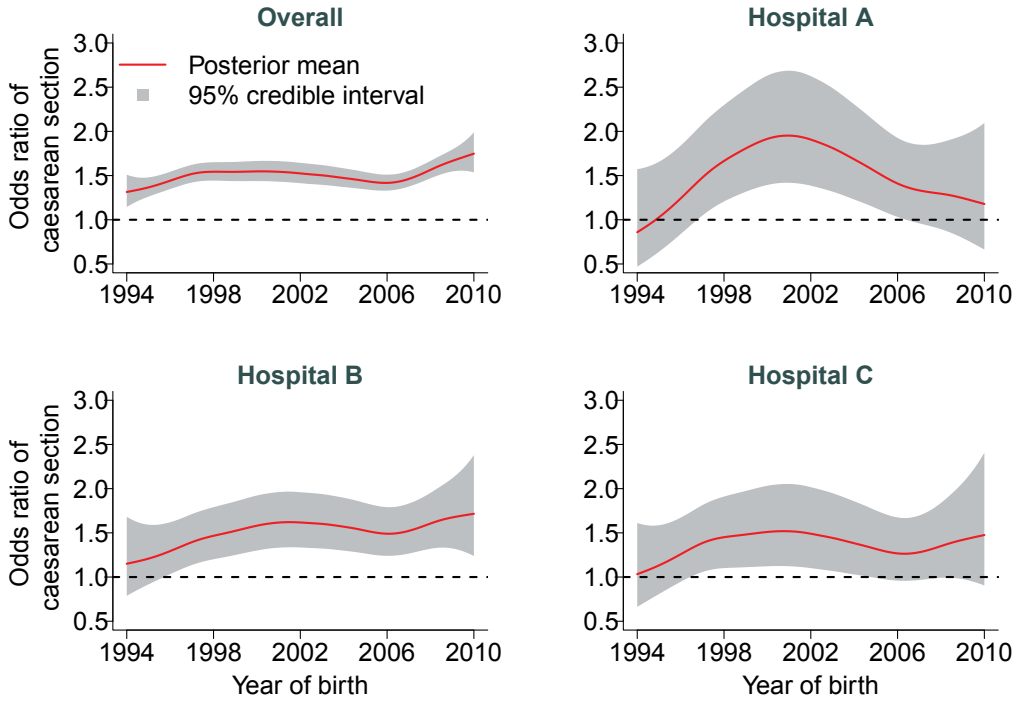

Figure 5.7: The MFVB estimated overall and selected hospital-specific contrast curves defined by (5.12), corresponding to the odds of caesarean section for low-risk nulliparous women aged greater or equal to 25 years, as a function of time, compared with those aged less than 25 years. The shaded regions correspond to pointwise 95% credible sets.

Year	Odds Ratio	95% Credible Interval
1994	1.31	(1.14, 1.50)
1998	1.42	(1.24, 1.62)
2002	1.52	(1.34, 1.74)
2006	1.63	(1.44, 1.86)
2010	1.74	(1.53, 1.98)

Table 5.2: The MFVB estimated odds ratios of caesarean section (averaged across hospitals) for low-risk nulliparous women aged greater than or equal to 25 years compared with those aged less than 25 years for selected years of birth.

There is a wide variation in odds ratios among hospitals, ranging from 1.21 to 2.24 (Figure 5.8).

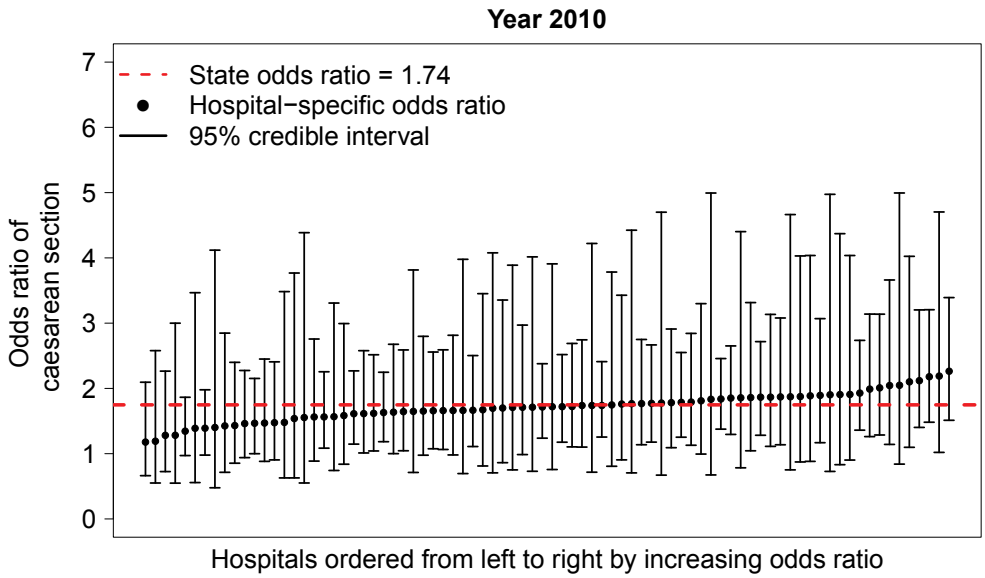

Figure 5.8: The MFVB estimated hospital-specific odds ratios of caesarean section for low-risk nulliparous women of older age compared with those of younger age in the year 2010.

5.5 Concluding remarks

We have described MFVB methods for fast approximate inference pertaining to three Bayesian logistic mixed models. Their utility to describe the overall mean and hospital-specific trends in caesarean section rates in NSW between 1994 and 2010 has been demonstrated. Using MCMC methods as a benchmark, we have evaluated the inferential accuracy and computation speed of our streamlined MFVB algorithms.

Compared to MCMC methods which directly generate Markov chain samples from the target posterior distribution, MFVB methods seek an approximating distribution in a factorised form that has minimum Kullback-Leibler distance to the target posterior. Although both methods are an approximation to the true posterior, MCMC methods are theoretically guaranteed to converge asymptotically to the true posterior (but come at great expense), while MFVB methods are known to underestimate the true variance due to the factorisation assumption (Wainwright and Jordan, 2008). Our simulation study shows moderately good to excellent accuracy of the MFVB approximation for all posterior densities of trend parameters. There is some underestimation in the posterior standard deviations, which may also be attributable to the Jaakkola and Jordan's approach described in (5.7). Despite minor degradation in accuracy, we believe that the MFVB approximation are a strong complement to the standard MCMC methods, as a way to efficiently explore possible models for revealing the complex non-linear temporal evolution of disease rates and identifying regions with unusual patterns at any point in time. Perhaps more practically, the MFVB parameter estimates can be used as starting points for MCMC to speed up convergence. In addition, our simulation results highlight the key advantage of the MFVB methods – computational efficiency. It serves as a great tool for iterative model building/construction. Even with concerns over accuracy in mind, MFVB methods provide an easy way to experiment with a range of models of different specifications, predictors and random effects covariance structures. As illustrated, the computational time for the MFVB algorithms is significantly shorter than that required for MCMC, with answers delivered in minutes rather than hours, although considerations such as computing environment and convergence criterion need to be taken into account to allow a fairer speed comparison. One of the more computationally expensive steps in the MFVB algorithms is the matrix inversion associated with the update of the covariance matrix of $q^*(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\xi})$, a cost that grows cubically in the number of groups. However, through matrix permutation and block decomposition we derived a streamlined approach that reduces the number of operations to be linear in the number of groups and is memory efficient.

The MFVB methods we consider here are fast and versatile and can be easily extended to more complicated scenarios. For example, the methods allow arbitrary priors for the hyperparameters (Neville *et al.*, 2014) and similar types of models with Gaussian response (Lee and Wand, 2015a; Luts, 2015; Tan and Nott, 2013). Stewart (2014) provides great examples of more elaborate models within the context of social sciences. The ability to handle binary outcomes is particularly useful in health science studies, where binary outcomes are common. For variational inference in logistic regression, we followed on Jaakkola and Jordan (1997) justifying based on constructing a lower bound for the marginal likelihood using convex duality Jaakkola and Jordan (1997) and derived closed-form expressions that approximate the posterior distributions for the parameters,

thus numerical techniques are not required. An extension to our work would be to develop approximations with less stringent conditional independence assumptions to see if the accuracy of MFVB approximations improves. This would be especially useful when we have prior knowledge about the conditional independence structure of the variables. Nevertheless, the q-density factorisation we investigated in the chapter remains appealing for its simplicity and ease of use.

In this chapter, we presented a flexible estimation of population- and hospital-level trends in caesarean section rates using a penalised spline basis function approach. Apart from being conceptually appealing, the approach allows the resulting models to be couched within a Bayesian mixed model framework for approximate inference. The presented group-specific curve models incorporate both temporal and hospital variabilities into a cohesive modelling framework, enabling one to visually describe the dynamic evolution of outcome rates across hospitals and over time. Our proposed contrast functions are useful in identifying particular regions of the temporal profiles that show significant differences in odds of caesarean section between low-risk nulliparous women of older and younger age. These regions are potential priority targets for public health interventions that aim at reducing temporal and hospital variations in caesarean section rates.

5.A Derivation of the mean field variational Bayes algorithm for model (5.11)

The full Bayesian representation for model (5.11) is

$$
\begin{aligned}
\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u} &\sim \text{Bernoulli}\left(\text{logit}^{-1}\left(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u}\right)\right), \quad \boldsymbol{\beta}\sim N(\boldsymbol{0},\sigma_\beta^2\boldsymbol{I}),\\
\boldsymbol{u}|\sigma^{\text{A}}_{\text{gbl}},\sigma^{\text{B}}_{\text{gbl}},\boldsymbol{\Sigma}_{\text{R}},\sigma_{\text{grp}} &\sim N\left(\boldsymbol{0},\begin{bmatrix}(\sigma^{\text{A}}_{\text{gbl}})^2\boldsymbol{I}_{K_{\text{gbl}}} & \boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & (\sigma^{\text{B}}_{\text{gbl}})^2\boldsymbol{I}_{K_{\text{gbl}}} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0} & \underset{1\le i\le m}{\text{blockdiag}}\left(\begin{bmatrix}\boldsymbol{\Sigma}_{\text{R}} & \boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \sigma^2_{\text{grp}}\boldsymbol{I}_{K_{\text{grp}}} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0} & \sigma^2_{\text{grp}}\boldsymbol{I}_{K_{\text{grp}}}\end{bmatrix}\right)\end{bmatrix}\right),\\
\boldsymbol{\Sigma}_{\text{R}}|a_{\text{R},1},a_{\text{R},2},a_{\text{R},3},a_{\text{R},4} &\sim \text{Inverse-Wishart}\ \left(\nu+3,2\,\nu\,\text{diag}(1/a_{\text{R},1},\ldots,1/a_{\text{R},4})\right),\\
a_{\text{R},1},a_{\text{R},2},a_{\text{R},3},a_{\text{R},4} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(\tfrac{1}{2},A_{\text{R}}^{-2}\right),\\
(\sigma^{\text{A}}_{\text{gbl}})^2|a^{\text{A}}_{\text{gbl}} &\sim \text{Inverse-Gamma}\left(\tfrac{1}{2},1/a^{\text{A}}_{\text{gbl}}\right), \quad a^{\text{A}}_{\text{gbl}}\sim \text{Inverse-Gamma}\left(\tfrac{1}{2},A^{-2}_{\text{gbl}}\right),\\
(\sigma^{\text{B}}_{\text{gbl}})^2|a^{\text{B}}_{\text{gbl}} &\sim \text{Inverse-Gamma}\left(\tfrac{1}{2},1/a^{\text{B}}_{\text{gbl}}\right), \quad a^{\text{B}}_{\text{gbl}}\sim \text{Inverse-Gamma,}\left(\tfrac{1}{2},A^{-2}_{\text{gbl}}\right),\\
\sigma^2_{\text{grp}}|a_{\text{grp}} &\sim \text{Inverse-Gamma}\left(\tfrac{1}{2},1/a_{\text{grp}}\right), \quad a_{\text{grp}}\sim \text{Inverse-Gamma,}\left(\tfrac{1}{2},A^{-2}_{\text{grp}}\right),
\end{aligned}
$$

Figure 5.9 shows the directed acyclic graph, as a graph theoretic representation of the Bayesian group-specific curve model conveyed in (5.15). The shaded node corresponds to the observed data vector and the non-shaded nodes correspond to the set of random variables. The advantages to such a graphical representation are that it provides a visualisation of the hierarchical structure of its corresponding Bayesian model, and the graph theoretic results can be used to determine probabilistic relationships between nodes.

Calculations similar to those in Subsection 5.2.1 are used to approximate the full joint posterior density function p by a product of approximating q-density functions:

$$
\begin{aligned}
&p\left(\boldsymbol{\beta},\boldsymbol{u},a^{\text{A}}_{\text{gbl}},a^{\text{B}}_{\text{gbl}},a^{\text{A}}_{\text{grp}},a^{\text{B}}_{\text{grp}},\boldsymbol{a}_{\text{R}},\sigma^{\text{A}}_{\text{gbl}},\sigma^{\text{B}}_{\text{gbl}},\sigma_{\text{grp}},\boldsymbol{\Sigma}_{\text{R}}\right)\\
&\qquad\approx q(\boldsymbol{\beta},\boldsymbol{u})\,q(a^{\text{A}}_{\text{gbl}})\,q(a^{\text{B}}_{\text{gbl}})\,q(a_{\text{grp}})\ q(\boldsymbol{a}_{\text{R}})\ q(\sigma^{\text{A}}_{\text{gbl}})\,q(\sigma^{\text{B}}_{\text{gbl}})\,q(\sigma_{\text{grp}})\,q(\boldsymbol{\Sigma}_{\text{R}}),
\end{aligned}
$$

with the optimal q-density functions admitting the following forms:

$$
\begin{aligned}
&\boldsymbol{\xi}\leftarrow\sqrt{\text{diagonal}[\boldsymbol{C}\{\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}+\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}\boldsymbol{\mu}^{\mathsf{T}}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}\}\boldsymbol{C}^{\mathsf{T}}]},\\
&q^*(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi}) \text{ is the } N(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})},\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}) \text{ density function,}\\
&q^*(a^{\text{A}}_{\text{gbl}}) \text{ is the Inverse-Gamma } \left(1,B_{q(a^{\text{A}}_{\text{gbl}})}\right) \text{ density function,}\\
&q^*(a^{\text{B}}_{\text{gbl}}) \text{ is the Inverse-Gamma } \left(1,B_{q(a^{\text{B}}_{\text{gbl}})}\right) \text{ density function,}\\
&q^*(a_{\text{R},r}) \text{ is the Inverse-Gamma } \left(1,B_{q(a_{\text{R},r})}\right) \text{ density function, } 1\le r\le 4,
\end{aligned}
$$

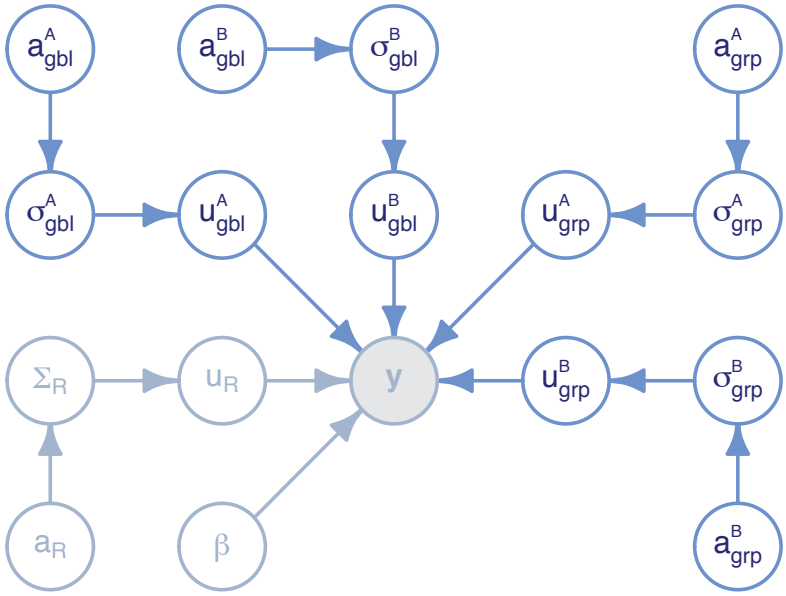

Figure 5.9: Directed Acyclic Graph for the Bernoulli group-specific curve model with contrasting. The shaded node corresponds to the observed data vector and the open nodes correspond to the random or auxiliary variables. The fragments shown in grey are present in the Bernoulli model DAG graph as shown in Figure 2.4.

$q^*((\sigma^{\mathrm{A}}_{\mathrm{gbl}})^2)$ is the Inverse-Gamma $\left(\frac{1}{2}(K_{\mathrm{gbl}}+1), B_{q((\sigma^{\mathrm{A}}_{\mathrm{gbl}})^2)}\right)$ density function,

$q^*((\sigma^{\mathrm{B}}_{\mathrm{gbl}})^2)$ is the Inverse-Gamma $\left(\frac{1}{2}(K_{\mathrm{gbl}}+1), B_{q((\sigma^{\mathrm{B}}_{\mathrm{gbl}})^2)}\right)$ density function,

$q^*(\sigma^2_{\mathrm{grp}})$ is the Inverse-Gamma $\left(\frac{1}{2}(m\,K_{\mathrm{grp}}+1), B_{q(\sigma^2_{\mathrm{grp}})}\right)$ density function and

$q^*(\boldsymbol{\Sigma}_{\mathrm{R}})$ is the Inverse-Wishart $\left(\nu+m+1, \boldsymbol{B}_{q(\boldsymbol{\Sigma}_{\mathrm{R}})}\right)$ density function.

The parameters $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}$ are the respective mean vector and covariance matrix of $q^*(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})$. The parameter $B_{q(\cdot)}$ is the scale parameter of the inverse-Gamma q-density. Appendix B of Wand and Ormerod (2011) provides a step-by step derivation for the optimal q-densities of similar type.

5.A.1 Mixed model representation

Using the equivalence between penalised splines and mixed models, we now aim to express model (5.11) within the mixed model framework. Define the following linear and non-linear predictor vectors and random effects vectors corresponding to the ith group to be:

$$\boldsymbol{X}^*_i \equiv \begin{bmatrix}\mathbf{1} & \boldsymbol{x}_i\end{bmatrix} \;;\; \boldsymbol{Z}_{\mathrm{gbl},i} \equiv \begin{bmatrix} z_{\mathrm{gbl},1}(\boldsymbol{x}_i) & \cdots & z_{\mathrm{gbl},K_{\mathrm{gbl}}}(\boldsymbol{x}_i)\end{bmatrix},$$

$$
\begin{aligned}
\boldsymbol{Z}_{\text{grp},i} &\equiv \begin{bmatrix} z_{\text{grp},1}(\boldsymbol{x}_i) & \cdots & z_{\text{grp},K_{\text{grp}}}(\boldsymbol{x}_i) \end{bmatrix}, \\
\boldsymbol{u}^{\text{A}}_{\text{gbl},i} &\equiv \begin{bmatrix} u^{\text{A}}_{\text{gbl},1} & \cdots & u^{\text{A}}_{\text{gbl},K_{\text{gbl}}} \end{bmatrix}^{\mathsf{T}} \;;\; \boldsymbol{u}^{\text{B}}_{\text{gbl},i} \equiv \begin{bmatrix} u^{\text{B}}_{\text{gbl},1} & \cdots & u^{\text{B}}_{\text{gbl},K_{\text{gbl}}} \end{bmatrix}^{\mathsf{T}}, \\
\boldsymbol{u}^{\text{A}}_{\text{grp},i} &\equiv \begin{bmatrix} u^{\text{A}}_{\text{grp},1} & \cdots u^{\text{A}}_{\text{grp},K_{\text{grp}}} \end{bmatrix}^{\mathsf{T}} \;;\; \boldsymbol{u}^{\text{B}}_{\text{grp},i} \equiv \begin{bmatrix} u^{\text{B}}_{\text{grp},1} & \cdots u^{\text{B}}_{\text{grp},K_{\text{grp}}} \end{bmatrix}^{\mathsf{T}} \\
\text{and} \qquad & \boldsymbol{u}_{\text{R},i} \equiv \begin{bmatrix} U^{\text{A}}_{0i} & U^{\text{A}}_{1i} & U^{\text{B}}_{0i} & U^{\text{B}}_{1i} \end{bmatrix}^{\mathsf{T}},
\end{aligned}
$$

where $\boldsymbol{x}_i$ is an $n_i \times 1$ vector containing the x_{ij}.

In Section 2 of Zhao *et al.* (2006) notation, the compact matrix form of model (5.11) is

$$
\boldsymbol{\beta} \equiv \begin{bmatrix} \boldsymbol{\beta}^{\text{R}} \\ \boldsymbol{\beta}^{\text{G}} \end{bmatrix}, \quad \boldsymbol{X} \equiv [\boldsymbol{X}^{\text{R}} \; \boldsymbol{X}^{\text{G}}], \quad \boldsymbol{u} \equiv \begin{bmatrix} \boldsymbol{u}^{\text{R}} \\ \boldsymbol{u}^{\text{G}} \end{bmatrix} \quad \text{and} \quad \boldsymbol{Z} \equiv [\boldsymbol{Z}^{\text{R}} \; \boldsymbol{Z}^{\text{G}}],
$$

which leads to

$$
\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} = \boldsymbol{X}^{\text{R}}\boldsymbol{\beta}^{\text{R}} + \boldsymbol{X}^{\text{G}}\boldsymbol{\beta}^{\text{G}} + \boldsymbol{Z}^{\text{R}}\boldsymbol{u}^{\text{R}} + \boldsymbol{Z}^{\text{G}}\boldsymbol{u}^{\text{G}}.
$$

Define $\boldsymbol{I}^{\text{A}}_i = \boldsymbol{1}$ if $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ is of type A and zero otherwise. The matrices $\boldsymbol{X}^{\text{R}}$ and $\boldsymbol{Z}^{\text{R}}$ are random design matrices corresponding to the fixed effects vector $\boldsymbol{\beta}^{\text{R}}$ and random group effects vector $\boldsymbol{u}^{\text{R}}$ (superscript R) respectively. They are defined as:

$$
\begin{aligned}
\boldsymbol{X}^{\text{R}} &\equiv \emptyset, \quad \boldsymbol{Z}^{\text{R}} \equiv \underset{1 \le i \le m}{\text{blockdiag}} \left(\boldsymbol{X}^*_i \mid \boldsymbol{I}^{\text{A}} \odot \boldsymbol{Z}_{\text{grp},i} \mid (\boldsymbol{1} - \boldsymbol{I}^{\text{A}}) \odot \boldsymbol{Z}_{\text{grp},i} \right), \\
\boldsymbol{\beta}^{\text{R}} &\equiv \emptyset, \quad \boldsymbol{u}^{\text{R}} \equiv \begin{bmatrix} (\boldsymbol{u}^{\mathsf{T}}_{\text{R},1} & (\boldsymbol{u}^{\text{A}}_{\text{grp},1})^{\mathsf{T}} & (\boldsymbol{u}^{\text{B}}_{\text{grp},1})^{\mathsf{T}} & \cdots & \boldsymbol{u}^{\mathsf{T}}_{\text{R},m} & (\boldsymbol{u}^{\text{A}}_{\text{grp},m})^{\mathsf{T}} & (\boldsymbol{u}^{\text{B}}_{\text{grp},m})^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}}.
\end{aligned}
$$

The matrices $\boldsymbol{X}^{\text{G}}$ and $\boldsymbol{Z}^{\text{G}}$ are general design matrices corresponding to the fixed effects vector $\boldsymbol{\beta}^{\text{G}}$ and random spline coefficients vector $\boldsymbol{u}^{\text{G}}$ (superscript G) respectively. Typically, $\boldsymbol{X}^{\text{G}}$ contains polynomial functions of continuous predictors that are modeled as penalised splines and $\boldsymbol{Z}^{\text{G}}$ would then contain random spline basis functions of the same predictors. They are defined as:

$$
\begin{aligned}
\boldsymbol{\beta}^{\text{G}} &\equiv \begin{bmatrix} \beta^{\text{A}}_0 \\ \beta^{\text{A}}_1 \\ \beta^{\text{BvsA}}_0 \\ \beta^{\text{BvsA}}_1 \end{bmatrix}, \quad \boldsymbol{X}^{\text{G}} \equiv \begin{bmatrix} \boldsymbol{X}^*_i & (\boldsymbol{1} - \boldsymbol{I}^{\text{A}}_i) \odot \boldsymbol{X}^*_i \\ \vdots & \vdots \\ \boldsymbol{X}^*_m & (\boldsymbol{1} - \boldsymbol{I}^{\text{A}}_m) \odot \boldsymbol{X}^*_m \end{bmatrix}, \\
\boldsymbol{Z}^{\text{G}} &\equiv \begin{bmatrix} \boldsymbol{I}^{\text{A}}_i \odot \boldsymbol{Z}_{\text{gbl},i} & (\boldsymbol{1} - \boldsymbol{I}^{\text{A}}_i) \odot \boldsymbol{Z}_{\text{gbl},i} \\ \vdots & \vdots \\ \boldsymbol{I}^{\text{A}}_m \odot \boldsymbol{Z}_{\text{gbl},m} & (\boldsymbol{1} - \boldsymbol{I}^{\text{A}}_m) \odot \boldsymbol{Z}_{\text{gbl},m} \end{bmatrix}, \\
\text{and} \quad \boldsymbol{u}^{\text{G}} &\equiv \begin{bmatrix} (\boldsymbol{u}^{\text{A}}_{\text{gbl},1})^{\mathsf{T}} & \cdots & (\boldsymbol{u}^{\text{A}}_{\text{gbl},m})^{\mathsf{T}} & (\boldsymbol{u}^{\text{B}}_{\text{gbl},1})^{\mathsf{T}} & \cdots & (\boldsymbol{u}^{\text{B}}_{\text{gbl},m})^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}},
\end{aligned}
$$

where $\odot$ denotes the element-wise product of matrices.

Building on Lee and Wand (2015a), we derive a streamlined iterative scheme for obtaining optimal q-density moments for all model parameters of the group-specific curve model with a factor-by-curve interaction. Our presentation of the streamlined MFVB algorithm benefits from the following notation, where $\boldsymbol{y}$, $\boldsymbol{X}^{\text{G}}$, $\boldsymbol{Z}^{\text{R}}$ and $\boldsymbol{Z}^{\text{G}}$ are partitioned row-wise corresponding to the ith group:

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_m \end{bmatrix}, \quad \boldsymbol{X}^{\text{G}} = \begin{bmatrix} \boldsymbol{X}^{\text{G}}_1 \\ \vdots \\ \boldsymbol{X}^{\text{G}}_m \end{bmatrix}, \quad \boldsymbol{Z}^{\text{R}} = \begin{bmatrix} \boldsymbol{Z}^{\text{R}}_1 \\ \vdots \\ \boldsymbol{Z}^{\text{R}}_m \end{bmatrix} \quad \text{and} \quad \boldsymbol{Z}^{\text{G}} = \begin{bmatrix} \boldsymbol{Z}^{\text{G}}_1 \\ \vdots \\ \boldsymbol{Z}^{\text{G}}_m \end{bmatrix}.$$

Here $\boldsymbol{y}_i \equiv [y_{i1} \cdots y_{in_i}]^{\mathsf{T}}$ denotes the $n_i \times 1$ vector of responses for the ith group. The matrices $\boldsymbol{X}^{\text{G}}_i$, $\boldsymbol{Z}^{\text{R}}_i$ and $\boldsymbol{Z}^{\text{G}}_i$ are defined in the same fashion. In addition, it is useful to define

$$\boldsymbol{C}^{\text{G}} \equiv [\boldsymbol{X} \ \boldsymbol{Z}^{\text{G}}] \quad \text{and} \quad \boldsymbol{C}^{\text{R}} \equiv [\boldsymbol{X}^{\text{R}} \ \boldsymbol{Z}^{\text{G}}].$$

5.A.2 Streamlined mean field variational Bayes algorithm

We are now ready to present the streamlined algorithm for fast MFVB approximate fitting and inference for model (5.11).

5.A.3 Fitting the group-specific curve model (5.11) in `Rstan`

Stan is a probabilistic programming language, written in C++, for implementing full Bayesian statistical inference through Hamiltonian Monte Carlo No U-turn sampling, a form of MCMC sampling. We now provide the `Rstan` code for fitting of model (5.11).

Specify the data: the number of observation, $\sum_{i=1}^{m} n_i$; the number of hospitals, m; the hospital identification number, `idnum`; the response vector, $\boldsymbol{y}$; the respective fixed and random effects design matrices, $\boldsymbol{X}$ and $\boldsymbol{Z}$; and the hyperparameters, σ^2_β, A_{R}, A_{gbl} and A_{grp}. Data are labelled as integer or real and can be vectors if dimensions are specified. Data can also be constrained; for example all hyperparameters must be positive.

```
grpSpecCurveModel <-
'data
{
   int<lower=1> numObs;            int<lower=1> m;
   int<lower=1> idnum[numObs];     int<lower=1> ncXR;
   int<lower=1> ncZ;               int<lower=1> ncZgrp;
   real<lower=0> sigmaBeta;        real<lower=0> AR;
   real<lower=0> AuGbl;            real<lower=0> AuGrp;
   matrix[numObs,ncXR] Xbase;      matrix[numObs,ncXR] XTypA;
```

Initialise: $\nu = 4$, $\mu_{q(1/(\sigma^{\mathrm{A}}_{\mathrm{gbl}})^2)} > 0$, $\mu_{q(1/(\sigma^{\mathrm{B}}_{\mathrm{gbl}})^2)} > 0$, $\mu_{q(1/a^{\mathrm{A}}_{\mathrm{gbl}})} > 0$, $\mu_{q(1/a^{\mathrm{B}}_{\mathrm{gbl}})} > 0$, $\mu_{q(1/\sigma^2_{\mathrm{grp}})} > 0$, $\mu_{q(1/a_{\mathrm{grp}})} > 0$, $\mu_{q(1/a_{\mathrm{R},r})} > 0$, $\boldsymbol{M}_{q(\boldsymbol{\Sigma}_{\mathrm{R}}^{-1})}$, a $q^{\mathrm{R}} \times q^{\mathrm{R}}$ positive definite matrix and $\boldsymbol{\xi}$, a $(\sum_{i=1}^{m} n_i) \times 1$ vector of positive entries.

Cycle through updates:

$$\textbf{Define: } \boldsymbol{M}_{q(\boldsymbol{\Omega})} \equiv \begin{bmatrix} \sigma_\beta^{-2}\boldsymbol{I}_p & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \mu_{q(1/(\sigma^{\mathrm{A}}_{\mathrm{gbl}})^2)}\boldsymbol{I}_{K_{\mathrm{gbl}}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \mu_{q(1/(\sigma^{\mathrm{B}}_{\mathrm{gbl}})^2)}\boldsymbol{I}_{K_{\mathrm{gbl}}} \end{bmatrix}$$

$\boldsymbol{S} \leftarrow \boldsymbol{0}$; $\boldsymbol{s} \leftarrow \boldsymbol{0}$

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

$$\boldsymbol{G}_i \leftarrow 2(\boldsymbol{C}_i^{\mathrm{G}})^{\mathsf{T}}\mathrm{diag}\{\lambda(\boldsymbol{\xi}_i)\}\boldsymbol{C}_i^{\mathrm{R}}$$

$$\boldsymbol{H}_i \leftarrow \left\{2(\boldsymbol{C}_i^{\mathrm{R}})^{\mathsf{T}}\mathrm{diag}\{\lambda(\boldsymbol{\xi}_i)\}\boldsymbol{C}_i^{\mathrm{R}} + \begin{bmatrix} \boldsymbol{M}_{q(\boldsymbol{\Sigma}_{\mathrm{R}}^{-1})} & \boldsymbol{0} \\ \boldsymbol{0} & \mu_{q(1/\sigma^2_{\mathrm{grp}})}\boldsymbol{I}_{2K_{\mathrm{grp}}} \end{bmatrix}\right\}^{-1}$$

$$\boldsymbol{S} \leftarrow \boldsymbol{S} + \boldsymbol{G}_i\boldsymbol{H}_i\boldsymbol{G}_i^{\mathsf{T}} \quad ; \quad \boldsymbol{s} \leftarrow \boldsymbol{s} + \boldsymbol{G}_i\boldsymbol{H}_i(\boldsymbol{C}_i^{\mathrm{R}})^{\mathsf{T}}(\boldsymbol{y}_i - \tfrac{1}{2}\boldsymbol{1})$$

Update multivariate Normal $q^*(\boldsymbol{\beta}, \boldsymbol{u}^{\mathbf{A}}_{\mathbf{gbl}}, \boldsymbol{u}^{\mathbf{B}}_{\mathbf{gbl}})$ parameters:

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{A}}_{\mathrm{gbl}},\boldsymbol{u}^{\mathrm{B}}_{\mathrm{gbl}})} \leftarrow \left\{2(\boldsymbol{C}^{\mathrm{G}})^{\mathsf{T}}\mathrm{diag}\{\lambda(\boldsymbol{\xi}_i)\}\boldsymbol{C}^{\mathrm{G}} + \boldsymbol{M}_{q(\boldsymbol{\Omega})} - \boldsymbol{S}\right\}^{-1}$$

$$\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{A}}_{\mathrm{gbl}},\boldsymbol{u}^{\mathrm{B}}_{\mathrm{gbl}})} \leftarrow \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{A}}_{\mathrm{gbl}},\boldsymbol{u}^{\mathrm{B}}_{\mathrm{gbl}})}\left\{(\boldsymbol{C}^{\mathrm{G}})^{\mathsf{T}}(\boldsymbol{y} - \tfrac{1}{2}\boldsymbol{1}) - \boldsymbol{s}\right\}$$

Update multivariate Normal $q^*(\boldsymbol{u}_{\mathbf{R},i}, \boldsymbol{u}^{\mathbf{A}}_{\mathbf{grp},i}, \boldsymbol{u}^{\mathbf{B}}_{\mathbf{grp},i})$ parameters:

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

$$\boldsymbol{\Sigma}_{q(\boldsymbol{u}_{\mathrm{R},i},\boldsymbol{u}^{\mathrm{A}}_{\mathrm{grp},i},\boldsymbol{u}^{\mathrm{B}}_{\mathrm{grp},i})} \leftarrow \boldsymbol{H}_i + \boldsymbol{H}_i\,\boldsymbol{G}_i^{\mathsf{T}}\,\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{A}}_{\mathrm{gbl}},\boldsymbol{u}^{\mathrm{B}}_{\mathrm{gbl}})}\,\boldsymbol{G}_i\,\boldsymbol{H}_i$$

$$\boldsymbol{\mu}_{q(\boldsymbol{u}_{\mathrm{R},i},\boldsymbol{u}^{\mathrm{A}}_{\mathrm{grp},i},\boldsymbol{u}^{\mathrm{B}}_{\mathrm{grp},i})} \leftarrow \boldsymbol{H}_i\left\{(\boldsymbol{X}_i^{\mathrm{R}})^{\mathsf{T}}\boldsymbol{y}_i - \boldsymbol{G}_i^{\mathsf{T}}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{A}}_{\mathrm{gbl}},\boldsymbol{u}^{\mathrm{B}}_{\mathrm{gbl}})}\right\}$$

Update $\boldsymbol{\xi}$ parameters:

$$\boldsymbol{\xi}^2 \leftarrow \mathrm{diagonal}\{\boldsymbol{C}^{\mathrm{G}}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{A}}_{\mathrm{gbl}},\boldsymbol{u}^{\mathrm{B}}_{\mathrm{gbl}})} + \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{A}}_{\mathrm{gbl}},\boldsymbol{u}^{\mathrm{B}}_{\mathrm{gbl}})}\boldsymbol{\mu}^{\mathsf{T}}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{A}}_{\mathrm{gbl}},\boldsymbol{u}^{\mathrm{B}}_{\mathrm{gbl}})})(\boldsymbol{C}^{\mathrm{G}})^{\mathsf{T}}\}$$

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

$$\boldsymbol{\xi}_i^2 \leftarrow \boldsymbol{\xi}_i^2 + 2\,\mathrm{diagonal}\{\boldsymbol{C}^{\mathrm{G}}(-\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{A}}_{\mathrm{gbl}},\boldsymbol{u}^{\mathrm{B}}_{\mathrm{gbl}})}\boldsymbol{G}_i\boldsymbol{H}_i + \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{A}}_{\mathrm{gbl}},\boldsymbol{u}^{\mathrm{B}}_{\mathrm{gbl}})}\boldsymbol{\mu}^{\mathsf{T}}_{q(\boldsymbol{u}_{\mathrm{R},i},\boldsymbol{u}^{\mathrm{A}}_{\mathrm{grp},i},\boldsymbol{u}^{\mathrm{B}}_{\mathrm{grp},i})})(\boldsymbol{C}_i^{\mathrm{R}})^{\mathsf{T}}\}$$

$$\boldsymbol{\xi}_i^2 \leftarrow \boldsymbol{\xi}_i^2 + \mathrm{diagonal}\{(\boldsymbol{C}_i^{\mathrm{R}}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_{\mathrm{R},i},\boldsymbol{u}^{\mathrm{A}}_{\mathrm{grp},i},\boldsymbol{u}^{\mathrm{B}}_{\mathrm{grp},i})} + \boldsymbol{\mu}_{q(\boldsymbol{u}_{\mathrm{R},i},\boldsymbol{u}^{\mathrm{A}}_{\mathrm{grp},i},\boldsymbol{u}^{\mathrm{B}}_{\mathrm{grp},i})}\boldsymbol{\mu}^{\mathsf{T}}_{q(\boldsymbol{u}_{\mathrm{R},i},\boldsymbol{u}^{\mathrm{A}}_{\mathrm{grp},i},\boldsymbol{u}^{\mathrm{B}}_{\mathrm{grp},i})})(\boldsymbol{C}_i^{\mathrm{R}})^{\mathsf{T}}\}$$

Update inverse-Gamma $q^*(\sigma^{\mathbf{A}}_{\mathbf{gbl}})$ and inverse-Gamma $q^*(a^{\mathbf{A}}_{\mathbf{gbl}})$ parameters:

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

$$\mu_{q(1/(\sigma^{\mathrm{A}}_{\mathrm{gbl}})^2)} \leftarrow \tfrac{1}{2}(K_{\mathrm{gbl}}+1)/B_{q((\sigma^{\mathrm{A}}_{\mathrm{gbl}})^2)} \quad ; \quad \mu_{q(1/a^{\mathrm{A}}_{\mathrm{gbl}})} \leftarrow 1/(\mu_{q(1/(\sigma^{\mathrm{A}}_{\mathrm{gbl}})^2)} + A^{-2}_{\mathrm{gbl}})$$

Update inverse-Gamma $q^*(\sigma^{\mathbf{B}}_{\mathbf{gbl}})$ and inverse-Gamma $q^*(a^{\mathbf{B}}_{\mathbf{gbl}})$ parameters:

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

$$\mu_{q(1/(\sigma^{\mathrm{B}}_{\mathrm{gbl}})^2)} \leftarrow \tfrac{1}{2}(K_{\mathrm{gbl}}+1)/B_{q((\sigma^{\mathrm{B}}_{\mathrm{gbl}})^2)} \quad ; \; 1/(\mu_{q(1/(\sigma^{\mathrm{B}}_{\mathrm{gbl}})^2)} + A^{-2}_{\mathrm{gbl}})$$

Update inverse-Gamma $q^*(\sigma_{\mathbf{grp}})$ and inverse-Gamma $q^*(a_{\mathbf{grp}})$ parameters:

$$B_{q(\sigma^2_{\mathrm{grp}})} \leftarrow \mu_{q(1/a_{\mathrm{grp}})} + \tfrac{1}{2}\sum_{i=1}^{m}\left\{\|\boldsymbol{\mu}_{q(\boldsymbol{u}^{\mathrm{A}}_{\mathrm{grp},i},\boldsymbol{u}^{\mathrm{B}}_{\mathrm{grp},i})}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}^{\mathrm{A}}_{\mathrm{grp},i},\boldsymbol{u}^{\mathrm{B}}_{\mathrm{grp},i})})\right\}$$

$$\mu_{q(1/\sigma^2_{\mathrm{grp}})} \leftarrow \tfrac{1}{2}(m\ K_{\mathrm{grp}}+1)/B_{q(\sigma^2_{\mathrm{grp}})}$$

$$B_{q(a_{\mathrm{grp}})} \leftarrow \mu_{q(1/\sigma^2_{\mathrm{grp}})} + A^{-2}_{\mathrm{grp}} \;\; ; \;\; \mu_{q(1/a_{\mathrm{grp}})} \leftarrow 1/B_{q(a_{\mathrm{grp}})}$$

Update inverse-Gamma $q^*(a_{\mathbf{R},r})$ and inverse-Gamma $q^*(\boldsymbol{\Sigma}_{\mathbf{R}})$ parameters:

For $r = 1, \ldots, 4$:

$$B_{q(a_{\mathrm{R},r})} \leftarrow \nu(M_{q(\boldsymbol{\Sigma}^{-1}_{\mathrm{R}})})_{rr} + A^{-2}_{\mathrm{R}r} \;\; ; \;\; \mu_{q(1/a_{\mathrm{R},r})} \leftarrow \tfrac{1}{2}(\nu+3)/B_{q(a_{\mathrm{R},r})}$$

$$\boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1}_{\mathrm{R}})} \leftarrow (\nu+m+1)\,\boldsymbol{B}^{-1}_{q(\boldsymbol{\Sigma}_{\mathrm{R}})}$$

$$\boldsymbol{B}_{q(\boldsymbol{\Sigma}_{\mathrm{R}})} \leftarrow \sum_{i=1}^{m}(\boldsymbol{\mu}_{q(\boldsymbol{u}_{\mathrm{R},i})}\boldsymbol{\mu}^{\mathsf{T}}_{q(\boldsymbol{u}_{\mathrm{R},i})} + \boldsymbol{\Sigma}_{q(\boldsymbol{u}_{\mathrm{R},i})}) + 2\,\nu\,\mathrm{diag}\left(\mu_{q(1/a_{\mathrm{R},1})}, \ldots, \mu_{q(1/a_{\mathrm{R},4})}\right)$$

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

Update the q-density covariance matrix for the ith group:

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

$$\boldsymbol{\Lambda}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{A}}_{\mathrm{gbl}},\boldsymbol{u}^{\mathrm{B}}_{\mathrm{gbl}},\boldsymbol{u}_{\mathrm{R},i},\boldsymbol{u}_{\mathrm{grp},i})} \equiv E_q\left[\left\{\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}^{\mathrm{A}}_{\mathrm{gbl}}\\ \boldsymbol{u}^{\mathrm{B}}_{\mathrm{gbl}}\end{bmatrix} - \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{A}}_{\mathrm{gbl}},\boldsymbol{u}^{\mathrm{B}}_{\mathrm{gbl}})}\right\}\left\{\begin{bmatrix}\boldsymbol{u}_{\mathrm{R},i}\\ \boldsymbol{u}_{\mathrm{grp},i}\end{bmatrix} - \boldsymbol{\mu}_{q(\boldsymbol{u}_{\mathrm{R},i},\boldsymbol{u}_{\mathrm{grp},i})}\right\}^{\mathsf{T}}\right]$$
$$\leftarrow -\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{A}}_{\mathrm{gbl}},\boldsymbol{u}^{\mathrm{B}}_{\mathrm{gbl}})}\boldsymbol{G}_i\,\boldsymbol{H}_i$$

Algorithm 16: Mean field variational Bayes algorithm for the Bernoulli group-specific curve model with contrasting.

```
   matrix[numObs,ncXR] XTypB;       int<lower=0,upper=1> y[numObs];
   real x[numObs];                  matrix[numObs,ncZ] ZTypA;
   matrix[numObs,ncZ] ZTypB;        real ZgrpTypA[numObs,ncZgrp];
   real ZgrpTypB[numObs,ncZgrp];    vector[2*ncXR] zeroVec;
}
```

Specify the model parameters: the unknown to be estimated in the model fit. These are the fixed effects vector, $\boldsymbol{\beta}$; the random effects vectors, $\boldsymbol{u}_{\text{gbl}}$, $\boldsymbol{u}_{\text{grp}}$ and $\boldsymbol{u}_{\text{R}}$; the covariance vectors and matrix, $\boldsymbol{\sigma}^2_{\text{gbl}}$, $\boldsymbol{\sigma}^2_{\text{grp}}$ and $\boldsymbol{\Sigma}_{\text{R}}$. In addition, we parametrize $\mu \equiv \boldsymbol{X\beta} + \boldsymbol{Zu}$ to be a transformation parameter in order to ensure the sampler runs more efficiently.

```
parameters
{
   vector[ncXR] beta;               vector[ncXR] betaTypB;
   vector[ncZ] uGblTypA;            vector[ncZ] uGblTypB;
   vector[2*ncXR] uR[m];            real uGrpTypA[m,ncZgrp];
   real uGrpTypB[m,ncZgrp];         cov_matrix[2*ncXR] SigmaR;
   vector[2*ncXR] aR;               real<lower=0> sigmauGblTypA;
   real<lower=0> sigmauGblTypB;     real<lower=0> sigmauGrp;
}
transformed parameters
{
   vector[numObs] fmean;            vector[numObs] fullMean;
   fmean <- (Xbase*beta + XTypB*betaTypB
             + ZTypA*uGblTypA + ZTypB*uGblTypB);
   for (iAll in 1:numObs)
      fullMean[iAll] <- (fmean[iAll]
             + uR[idnum[iAll],3]*XTypA[iAll,1]
             + uR[idnum[iAll],4]*XTypA[iAll,2]
             + uR[idnum[iAll],1]*XTypB[iAll,1]
             + uR[idnum[iAll],2]*XTypB[iAll,2]
             + dot_product(uGrpTypA[idnum[iAll]],ZgrpTypA[iAll])
             + dot_product(uGrpTypB[idnum[iAll]],ZgrpTypB[iAll]));
}
```

Specify the model statement: The `Bernoulli` specifies that the response vector $\boldsymbol{y}$ has a bernoulli distribution with mean $\text{logit}^{-1}(\mu)$, where the mean is specified to be the sum of variables for the overall mean, maternal age deviation from that overall mean and hospital deviations.

```
model
{
   matrix[2*ncXR,2*ncXR] rateForWish;

   y ~ bernoulli_logit(fullMean);

   for (i in 1:m)
      uR[i] ~ multi_normal(zeroVec,SigmaR);

   uGblTypA ~ normal(0,sigmauGblTypA);
   uGblTypB ~ normal(0,sigmauGblTypB);
```

```
   for (i in 1:m)
   {
      for (k in 1:ncZgrp)
      {
         uGrpTypA[i,k] ~ normal(0,sigmauGrp);
         uGrpTypB[i,k] ~ normal(0,sigmauGrp);
      }
   }

   rateForWish <- rep_matrix(0,4,4);
   for (r in 1:(2*ncXR))
   {
      aR[r] ~ inv_gamma(0.5,pow(AR,-2));
      rateForWish[r,r] <- 4/aR[r];
   }
   SigmaR ~ inv_wishart(5,rateForWish);

   beta ~ normal(0,sigmaBeta);
   betaTypB ~ normal(0,sigmaBeta);
   sigmauGblTypA ~ cauchy(0,AuGbl);
   sigmauGblTypB ~ cauchy(0,AuGbl);
   sigmauGrp ~ cauchy(0,AuGrp);
}'
```

Chapter 6

Alternative Approach Based on Variational Message Passing

Far better an approximate answer to the right question, which is often vague, than the exact answer to the wrong question, which can always be made precise.

John Tukey

6.1 Introduction

In the previous chapters, we presented the full description and implementation of the state-of-the-art MFVB algorithms in longitudinal and multilevel data settings and showed how MFVB approximations can be flexibly scaled to other classes of statistical models including, for example, higher-level random effects, missing data and measurement error models, and group-specific curve models with contrasting. In this chapter, we introduce an alternative approach, which produces the same posterior approximation as MFVB but with a different algebraic system, known as the *variational message passing* (VMP) (e.g. Winn and Bishop, 2005).

We revisit MFVB for Bayesian semiparametric mixed regression but instead work within the VMP framework. Although the MFVB and VMP approaches each lead to different iterative variational algorithms, they converge to the identical approximating posterior density functions in a wide range of models under the same mean-field restrictions. In essence, the approach is guided by the notion of *message passing*, a general principle that is well known in software engineering for efficient computing within distributed systems, but not in mainstream statistics. As explained in Minka (2005), the

parameter update expressions in the MFVB algorithm can be alternatively expressed as *messages* passed between nodes on a suitable factor graph. Factor graphs, earlier introduced in Frey (1998), is a graphical concept tailored to the VMP approach. Often, the VMP messages are model parameters from the exponential family density functions and are generally summarised in natural canonical forms. As pointed out by Wand (2015), the natural parametrisation of VMP messages leads to greater efficiency gains from using VMP over MFVB in terms of modularisation, since the extension to arbitrarily large models is much more straightforward via the introduction of factor graph fragments. The current version of Infer.NET supports VMP fitting of Bayesian hierarchical models, but is only applicable to certain classes of standard models. For the semiparametric mixed regression scenarios, self-implementation of VMP algorithms is our only option.

Contemporary literature on VMP are mainly contributed by Winn and Bishop (2005), Minka (2005) and Minka *et al.* (2009). However, each article offers a slightly different notation system and description of the VMP scheme. We use the same notation convention as described in Minka (2005). The thrust of this chapter is to provide a unified and notationally friendly framework for the general VMP approach regarding the Gaussian response examples considered in Sections 2.2.3 and 3.7.

Sections 6.2 and 6.3 provide relevant background materials relevant to VMP. We provide a general notationally friendly recipe of the VMP approach in Section 6.5. In Section 6.6 we use a simple Bayesian mixed model to convey the main ideas of VMP and then describe the ease of extension to larger statistical models in Section 6.7. Concluding remarks are given in Section 6.8.

6.2 Natural canonical forms for exponential family densities

The natural canonical form of any univariate exponential family distribution in a scalar random variable x can be written as

$$p(x) \equiv \exp\{\boldsymbol{T}(x)^{\mathsf{T}}\boldsymbol{\eta} - A(\boldsymbol{\eta})\}\, h(x), \tag{6.1}$$

where $\boldsymbol{T}(x)$ is the *natural statistic*, $\boldsymbol{\eta}$ is the *natural parameter*, $A(\boldsymbol{\eta})$ is the *log-partition function* and $h(x)$ is the *base measure*. In general, the natural statistic is not unique and is commonly expressed $\boldsymbol{T}(x)$ as the simplest possible algebraic form given $p(x)$. For example, the Inverse-Gamma(A,B) density function with shape parameter $A > 0$ and scale parameter $B > 0$ has the general form

$$p(x; A, B) \;=\; B^A\, \Gamma(A)^{-1}\, x^{-A-1} \exp(-B/x), \quad x > 0, \tag{6.2}$$

$$= \exp\left\{ \begin{bmatrix} \log(x) \\ 1/x \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} -A-1 \\ -B \end{bmatrix} + A\log(B) - \log\Gamma(A) \right\}.$$

Straightforward algebra shows that (6.2) can be expressed in the form of (6.1) with

$$\boldsymbol{T}(x) = \begin{bmatrix} \log(x) \\ 1/x \end{bmatrix}, \quad \boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} -A-1 \\ -B \end{bmatrix} \quad \text{and} \quad h(x) = 1.$$

The log-partition function is

$$A(\boldsymbol{\eta}) = \log\Gamma(-\eta_1 - 1) - (-\eta_1 - 1)\log(-\eta_2).$$

Table 6.1 lists the natural statistics, natural parameter vectors and natural canonical forms of common exponential family densities.

Probability distribution	Natural statistic	Natural parameter vector	Natural canonical form
Bernoulli(p)	x	$\log\{p(1-p)\}$	$\text{Bernoulli}_{\text{nat}}(\eta)$
$N(\mu, \sigma^2)$	$\begin{bmatrix} x \\ x^2 \end{bmatrix}$	$\begin{bmatrix} \mu/\sigma^2 \\ -1/(2\sigma^2) \end{bmatrix}$	$N_{\text{nat}}(\eta_1, \eta_2)$
Inverse-Gamma(A, B)	$\begin{bmatrix} \log(x) \\ 1/x \end{bmatrix}$	$\begin{bmatrix} -A-1 \\ B \end{bmatrix}$	$\text{Inverse-Gamma}_{\text{nat}}(\eta_1, \eta_2)$
$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	$\begin{bmatrix} \boldsymbol{x} \\ \text{vec}(\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}) \end{bmatrix}$	$\begin{bmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\text{vec}(\boldsymbol{\Sigma}^{-1}) \end{bmatrix}$	$N_{\text{nat}}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_{2,\text{vec}})$
Inverse-Wishart($A, \boldsymbol{B}$)	$\begin{bmatrix} \log(\boldsymbol{X}) \\ \text{vec}(\boldsymbol{X}^{-1}) \end{bmatrix}$	$\begin{bmatrix} \frac{1}{2}(A+d+1) \\ -\frac{1}{2}\text{vec}(\boldsymbol{B}) \end{bmatrix}$	$\text{Inverse-Wishart}_{\text{nat}}(\eta_1, \boldsymbol{\eta}_2)$

Table 6.1: Expressions for natural statistics, natural parameter vectors and natural canonical forms of common exponential family densities.

6.3 Primitive integrals and function definitions

The following primitives and function definitions follow immediately from the results involving moments of random variables which take on distributions in Table 6.1. These moments aid in the derivation of the VMP algorithms described in the later sections.

We define

$$p_{\mathrm{N}_{\mathrm{nat,vec}}}\left(\boldsymbol{x}; \begin{bmatrix}\boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_{2,\mathrm{vec}}\end{bmatrix}\right) \text{ to be the } N_{\mathrm{nat,vec}}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_{2,\mathrm{vec}}) \text{ density function in } x,$$

$$p_{\mathrm{IG}_{\mathrm{nat}}}\left(x, \begin{bmatrix}\eta_1 \\ \eta_2\end{bmatrix}\right) \text{ to be the Inverse-Gamma}_{\mathrm{nat}}(\eta_1, \eta_2) \text{ density function in } x,$$

$$p_{\mathrm{IW}_{\mathrm{nat}}}\left(\boldsymbol{X}, \begin{bmatrix}\eta_1 \\ \boldsymbol{\eta}_2\end{bmatrix}\right) \text{ to be the Inverse-Wishart}_{\mathrm{nat}}(\eta_1, \boldsymbol{\eta}_2) \text{ density function in } \boldsymbol{X}.$$

We now present the corresponding algorithmic primitives and function definitions used throughout this chapter:

Primitive 6.3.1. If $\boldsymbol{x}$ is a $d \times 1$ random vector,

$$\begin{aligned}\mathcal{F}\left(\begin{bmatrix}\boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_{2,\mathrm{vec}}\end{bmatrix}\right) &\equiv \int_{\mathbb{R}^d} \boldsymbol{x}^\mathsf{T}\boldsymbol{x}\, p_{\mathrm{N}_{\mathrm{nat,vec}}}\left(\boldsymbol{x}; \begin{bmatrix}\boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_{2,\mathrm{vec}}\end{bmatrix}\right) d\boldsymbol{x} \\ &= \tfrac{1}{4}\mathrm{tr}\left(\left\{\mathrm{vec}^{-1}(\boldsymbol{\eta}_{2,\mathrm{vec}})\right\}^{-1}\left[\boldsymbol{\eta}_1\boldsymbol{\eta}_1^\mathsf{T}\left\{\mathrm{vec}^{-1}(\boldsymbol{\eta}_{2,\mathrm{vec}})\right\}^{-1} - \boldsymbol{I}\right]\right).\end{aligned}$$

Primitive 6.3.2. If $\boldsymbol{x}$ is a $d \times 1$ random vector, $\boldsymbol{b}$ a $p \times 1$ vector and $\boldsymbol{C}$ a $p \times q$ matrix,

$$\begin{aligned}\mathcal{H}\left(\begin{bmatrix}\boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_{2,\mathrm{vec}}\end{bmatrix}; \boldsymbol{b}, \boldsymbol{C}\right) &\equiv \int_{\mathbb{R}^d} (\boldsymbol{b}-\boldsymbol{C}\boldsymbol{x})(\boldsymbol{b}-\boldsymbol{C}\boldsymbol{x})^\mathsf{T} p_{\mathrm{N}_{\mathrm{nat,vec}}}\left(\boldsymbol{x}; \begin{bmatrix}\boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_{2,\mathrm{vec}}\end{bmatrix}\right) d\boldsymbol{x} \\ &= \boldsymbol{b}\boldsymbol{b}^\mathsf{T} + \{\mathrm{vec}^{-1}(\boldsymbol{\eta}_{2,\mathrm{vec}})\}^{-1}\left[\boldsymbol{b}\boldsymbol{\eta}_1^\mathsf{T}\boldsymbol{C}^\mathsf{T} + \tfrac{1}{4}\boldsymbol{C}\boldsymbol{\eta}_1\boldsymbol{\eta}_1^\mathsf{T}\boldsymbol{C}^\mathsf{T}\{\mathrm{vec}^{-1}(\boldsymbol{\eta}_{2,\mathrm{vec}})\}^{-1} - \tfrac{1}{2}\boldsymbol{C}\boldsymbol{C}^\mathsf{T}\right].\end{aligned}$$

Primitive 6.3.3. If $\boldsymbol{X}$ is a $d \times d$ random vector,

$$\begin{aligned}\mathcal{G}\left(\begin{bmatrix}\eta_1 \\ \boldsymbol{\eta}_2\end{bmatrix}\right) &\equiv \int_S \boldsymbol{X}^{-1} p_{\mathrm{IW}_{\mathrm{nat}}}\left(\boldsymbol{X}; \begin{bmatrix}\eta_1 \\ \boldsymbol{\eta}_2\end{bmatrix}\right) d\boldsymbol{x} \\ &= \left\{\eta_1 + \tfrac{1}{2}(d+1)\right\}\{\mathrm{vec}^{-1}(\boldsymbol{\eta}_2)\}^{-1},\end{aligned}$$

where S is the set of all symmetric, positive definite $d \times d$ matrices.

Primitive 6.3.4. If x is a scalar random variable,

$$\mathcal{I}\left(\begin{bmatrix}\eta_1\\ \eta_2\end{bmatrix}\right) \equiv \int_0^\infty \tfrac{1}{x}\, p_{\mathrm{IG_{nat}}}\left(x, \begin{bmatrix}\eta_1\\ \eta_2\end{bmatrix}\right) dx = \frac{\eta_1 + 1}{\eta_2}.$$

6.4 Factor graphs

A *factor graph* is a graphical representation of the mathematical relation between arguments and local factors of a real-valued function. Consider, for example, the function $g(x_1, x_2, x_3, x_4, x_5)$ of five variables defined on $\mathbb{R}^5$ as follows:

$$g(x_1, x_2, x_3, x_4, x_5) \equiv f_A(x_1)\, f_B(x_2, x_3)\, f_C(x_3, x_4, x_5)\, f_D(x_3, x_4)\, f_E(x_5). \tag{6.3}$$

Figure 6.1 is a factor graph corresponding to g. The circular nodes correspond to the arguments of g and the square nodes correspond to the factors in (6.3). Edges are drawn between each factor node and arguments of that factor. In general, factor graphs of functions are non-unique since, for example, f_A and f_B could be aggregated into a single factor resulting in a different factor graph. For the remainder of this chapter, we refer a circular node as a *stochastic node* (i.e. random variables, random vectors or random matrices) and a square node as a *factor*. We use the word *node* to describe either a stochastic node or a factor.

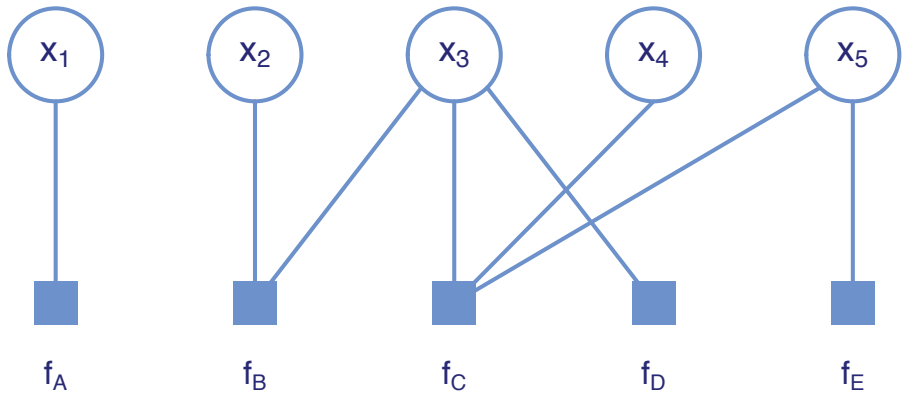

Figure 6.1: A factor graph corresponding to the function $g(x_1, x_2, x_3, x_4, x_5)$ defined by (6.3).

6.5 Variational message passing

Consider a generic Bayesian model with observed data vector y and parameter vector $\boldsymbol{\theta}$ that is continuous over the parameter space $\boldsymbol{\Theta}$. A MFVB approximation to the posterior density function $p(\boldsymbol{\theta}|\boldsymbol{y})$ is

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \approx q^*(\boldsymbol{\theta}),$$

where $q^*(\boldsymbol{\theta})$ is the minimiser of the Kullback-Leibler divergence of $p(\boldsymbol{\theta}|\boldsymbol{y})$ from $q(\boldsymbol{\theta})$ subject to the product density restriction for a subset S of $\{1, \ldots, M\}$

$$q(\boldsymbol{\theta}) = \prod_{i=1}^{M} q_i(\boldsymbol{\theta}_S), \quad \boldsymbol{\theta}_S \equiv \{\boldsymbol{\theta}_j \,:\, j \in S\}. \tag{6.4}$$

The restriction (6.4) means that the joint density function of $\boldsymbol{\theta}$ and $\boldsymbol{y}$ can be expressed as

$$p(\boldsymbol{\theta}, \boldsymbol{y}) = \prod_{j=1}^{N} f_j(\boldsymbol{\theta}_{S_j}) \text{ for subsets } S_j \text{ of } \{1, \ldots, M\} \text{ and factors } f_j, 1 \leq j \leq N. \tag{6.5}$$

Variational message passing is a message-passing algorithmic implementation of the mean-field approximation. The VMP iterative updates are expressed in terms of updating messages passed between nodes on the factor graph. For example, if $p(\boldsymbol{\theta}, \boldsymbol{y})$ is a directed acyclic graphical model with nodes $\boldsymbol{\theta}_1, \ldots \boldsymbol{\theta}_M$ and $\boldsymbol{y}$ then

$$p(\boldsymbol{\theta}, \boldsymbol{y}) = \left\{ \prod_{j=1}^{M} p(\boldsymbol{\theta}_i | \text{parents of } \boldsymbol{\theta}_i) \right\} p(\boldsymbol{y} | \text{parents of } \boldsymbol{y}) \tag{6.6}$$

is an example of (6.5) with the factors f_j corresponding to the density function of $\boldsymbol{\theta}_j$ conditional on its parent nodes and the factor f_{M+1} corresponding to the likelihood. Each factor is a function of the subset of (6.4) representing the parent-child relationships in the directed acyclic graph infrastructure. Figure 6.2 is an example of the factor graph that represents the full factorisation of $p(\boldsymbol{\theta}|\boldsymbol{y})$ with respect to (6.6) with $M = 9$ and $N = 10$. It becomes clear that in factor graphs the conditional independence properties of statistical models are represented by the squared factor nodes in replacement of the directed arrows in directed acyclic graphs, henceforth we introduce a new notation

$$\text{neighbours}(j) = \{1 \leq i \leq M \,:\, \boldsymbol{\theta}_i \text{ is a neighbour of } f_j\} \tag{6.7}$$

and (6.5) is re-expressible as $p(\boldsymbol{\theta}, \boldsymbol{y}) = \prod_{j=1}^{N} f_j(\boldsymbol{\theta}_{\text{neighbours}_{(j)}})$.

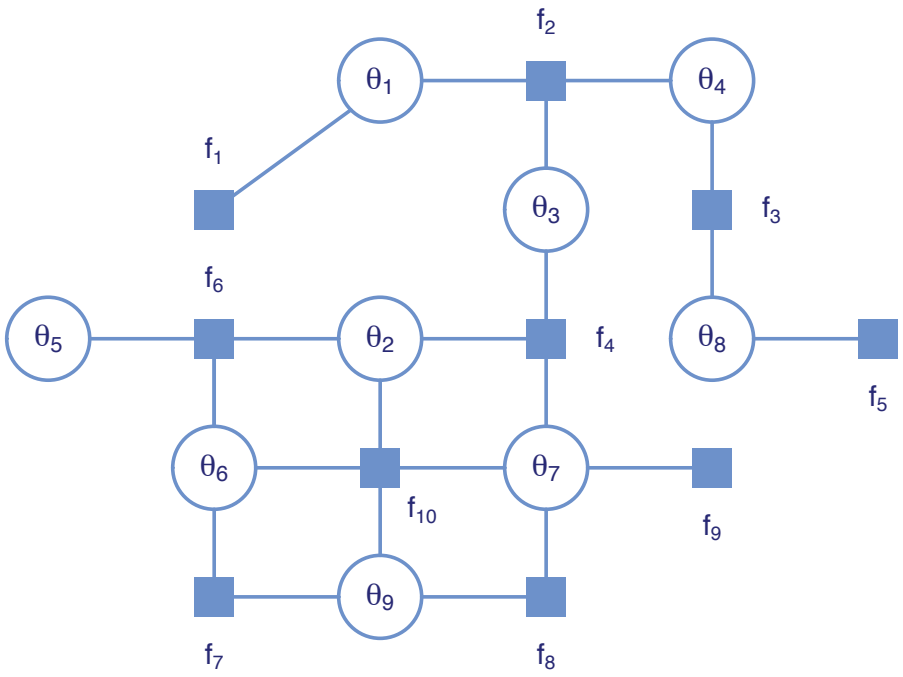

Figure 6.2: A factor graph corresponding to a Bayesian model with stochastic nodes $\boldsymbol{\theta}_1, \ldots \boldsymbol{\theta}_9$ and factors $f_1, \ldots f_{10}$.

Wand (2015) provides a unified framework for obtaining the optimal q-densities through an iterative scheme concerned with updating two types of messages in a factor graph. Here we reiterate his scheme. For each $1 \leq i \leq M$ and $1 \leq j \leq N$, the VMP updates for the first type of messages, *stochastic node to factor messages*, are

$$m_{\boldsymbol{\theta}_i \to f_j}(\boldsymbol{\theta}_i) \longleftarrow \prod_{j' \neq j} \prod_{i \in \text{neighbours}(j')} m_{f_{j'} \to \boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) \tag{6.8}$$

and the second type of messages, *factor to stochastic node messages*, are

$$m_{f_j \to \boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) \longleftarrow \exp\left[E_{f_j \to \boldsymbol{\theta}_i}\left\{\log f_j(\boldsymbol{\theta}_{\text{neighbours}(j)})\right\}\right], \tag{6.9}$$

where the subscript m designates the nodes involved in the message passing and the direction in which the message is passed, and the notation $E_{f_j \to \boldsymbol{\theta}_i}$ denotes expectation with respect to the density function

$$\frac{1}{Z}\left(\prod_{i' \in \text{neighbours}(j) \setminus \{i\}} m_{f_j \to \boldsymbol{\theta}_{i'}}(\boldsymbol{\theta}_{i'}) m_{f_{i'} \to \boldsymbol{\theta}_j}(\boldsymbol{\theta}_{i'})\right) \tag{6.10}$$

given Z is the corresponding normalising constant. The notation $A \setminus B$ denotes the set

containing elements of A that are not in B if A and B are sets such that $B \subseteq A$. If neighbours$(j)\backslash\{i\} = \emptyset$, then the right hand side of 6.9 simplifies to $f_j(\boldsymbol{\theta}_{\text{neighbours}(j)})$. The convergence of these message updates is assessed by monitoring successive values of the lower bound of the marginal log-likelihood

$$\log \underline{p}(q; \boldsymbol{y}) = \sum_{i=1}^{M} E_{q(\boldsymbol{\theta}_i)}\{\log q(\boldsymbol{\theta}_i)\} + \sum_{j=1}^{N} E_q\{\log(f_j)\}. \tag{6.11}$$

The first term within the summation is known as the *entropy* or *differential entropy*. Wand (2015) provides a list of entropy expressions for the common exponential density functions. Upon convergence of the message updates (when the successive differences of (6.11) are very small), the optimal q-densities are obtained via

$$q^*(\boldsymbol{\theta}_i) \propto \prod_{i \in \text{neighbours}(j)} m_{f_j \to \boldsymbol{\theta}_i}(\boldsymbol{\theta}_i). \tag{6.12}$$

6.6 Illustrative example of simple mixed effects regression

Consider the random intercepts and slopes model (2.2),

$$\begin{aligned}
\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2 &\sim N(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \sigma_\varepsilon^2 \boldsymbol{I}), \\
\boldsymbol{u}|\boldsymbol{\Sigma}^{\text{R}} &\sim N\left(\boldsymbol{0}, \boldsymbol{I}_m \otimes \boldsymbol{\Sigma}^{\text{R}}\right), \quad \boldsymbol{\beta} \sim N(0, \sigma_\beta^2 \boldsymbol{I}_p), \\
\boldsymbol{\Sigma}^{\text{R}}|a_1^{\text{R}}, \ldots, a_{q^{\text{R}}}^{\text{R}} &\sim \text{Inverse-Wishart}\left(\nu + q^{\text{R}} - 1,\ 2\,\nu\,\text{diag}(1/a_1^{\text{R}}, \ldots, 1/a_{q^{\text{R}}}^{\text{R}})\right), \\
a_r^{\text{R}} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(\tfrac{1}{2}, 1/A_{\text{R}r}^2\right), \quad r = 1, \ldots, q^{\text{R}} \\
\sigma_\varepsilon^2|a_\varepsilon &\sim \text{Inverse-Gamma}(\tfrac{1}{2},\ 1/a_\varepsilon), \quad a_\varepsilon \sim \text{Inverse-Gamma}(\tfrac{1}{2},\ 1/A_\varepsilon^2),
\end{aligned} \tag{6.13}$$

where the parameters are previously defined in Subsection 2.2.3. The joint posterior density function of model parameters and auxiliary variables is analytically intractable and numerically challenging, requiring variational approximations for fitting in practice. Mean field variational Bayes is a general class of variational methods for approximation of posterior density functions in the directed acyclic graph model infrastructure. Variational message passing is a message-passing version of the mean-field method, which leads to the identical posterior density function approximations since they are each founded upon the same Kullback-Leibler optimisation problem. Here we aim to obtain a mean-field approximation to the joint posterior density function:

$$p(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\Sigma}^{\text{R}}, \sigma_\varepsilon^2, \boldsymbol{a}^{\text{R}}, a_\varepsilon) \approx q(\boldsymbol{\beta}, \boldsymbol{u})\ q(\boldsymbol{\Sigma}^{\text{R}})\ q(\sigma_\varepsilon^2)\ q(\boldsymbol{a}^{\text{R}})\ q(a_\varepsilon).$$

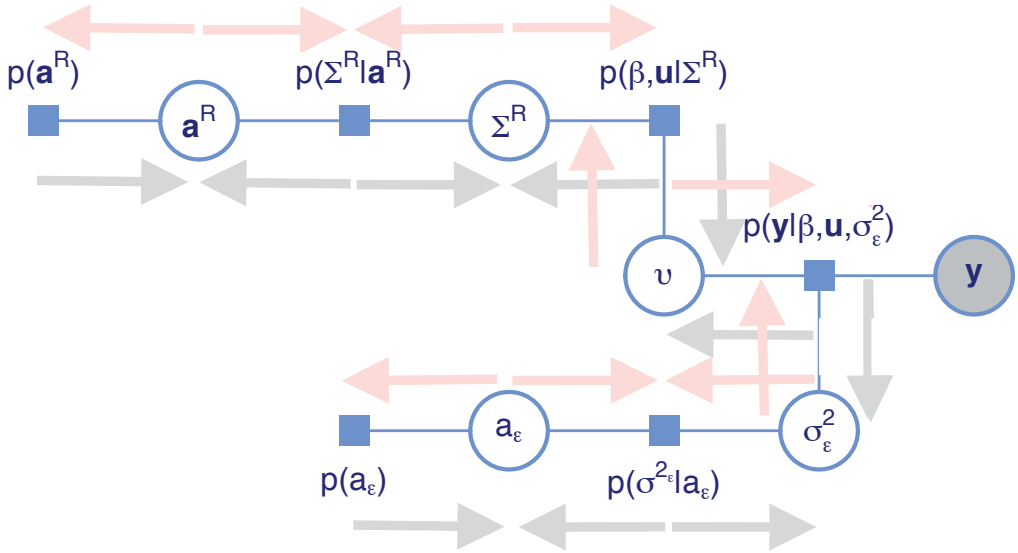

Figure 6.3: Each of the messages between neighbouring nodes on the factor graph for the random intercepts and slopes model illustrative example. The pink arrows depict the direction from the stochastic node to factor, while the grey arrows depict the direction from the factor to stochastic node.

VMP iterations for fitting model (6.13) involve updating 10 messages passed from factor to stochastic node (pink arrows) and 10 messages passed from stochastic node to factor (grey arrows) on the factor graph in Figure 6.3. Each message is a function of the stochastic node that is either received or sent by the message. Let $\boldsymbol{\upsilon} \equiv [\boldsymbol{\beta}^\mathsf{T} \, \boldsymbol{u}^\mathsf{T}]^\mathsf{T}$ and $\boldsymbol{C} \equiv [\boldsymbol{X} \, \boldsymbol{Z}]$, we first initialise each factor to stochastic node message to be an arbitrary density function:

$$
\begin{aligned}
m_{p(a_r^{\mathrm{R}}) \to a_r^{\mathrm{R}}}(a_r^{\mathrm{R}}) &\propto \exp\left\{ \sum_{r=1}^{q^{\mathrm{R}}} \begin{bmatrix} \log(a_r^{\mathrm{R}}) \\ 1/a_r^{\mathrm{R}} \end{bmatrix}^\mathsf{T} \boldsymbol{\eta}_{p(a_r^{\mathrm{R}}) \to a_r^{\mathrm{R}}} \right\}, \\
m_{p(\boldsymbol{\Sigma}^{\mathrm{R}}|a_r^{\mathrm{R}}) \to a_r^{\mathrm{R}}}(a_r^{\mathrm{R}}) &\propto \exp\left\{ \sum_{r=1}^{q^{\mathrm{R}}} \begin{bmatrix} \log(a_r^{\mathrm{R}}) \\ 1/a_r^{\mathrm{R}} \end{bmatrix}^\mathsf{T} \boldsymbol{\eta}_{p(\boldsymbol{\Sigma}^{\mathrm{R}}|a_r^{\mathrm{R}}) \to a_r^{\mathrm{R}}} \right\}, \\
m_{p(\boldsymbol{\Sigma}^{\mathrm{R}}|\boldsymbol{a}^{\mathrm{R}}) \to \boldsymbol{a}^{\mathrm{R}}}(\boldsymbol{\Sigma}^{\mathrm{R}}) &\propto \exp\left\{ \begin{bmatrix} \log|\boldsymbol{\Sigma}^{\mathrm{R}}| \\ \mathrm{vec}\{(\boldsymbol{\Sigma}^{\mathrm{R}})^{-1}\} \end{bmatrix}^\mathsf{T} \boldsymbol{\eta}_{p(\boldsymbol{\Sigma}^{\mathrm{R}}|a^{\mathrm{R}}) \to \boldsymbol{\Sigma}^{\mathrm{R}}} \right\}, \\
m_{p(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{R}}) \to \boldsymbol{\Sigma}^{\mathrm{R}}}(\boldsymbol{\Sigma}^{\mathrm{R}}) &\propto \exp\left\{ \begin{bmatrix} \log|\boldsymbol{\Sigma}^{\mathrm{R}}| \\ \mathrm{vec}\{(\boldsymbol{\Sigma}^{\mathrm{R}})^{-1}\} \end{bmatrix}^\mathsf{T} \boldsymbol{\eta}_{p(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{R}}) \to \boldsymbol{\Sigma}^{\mathrm{R}}} \right\}, \\
m_{p(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{R}}) \to (\boldsymbol{\beta}, \boldsymbol{u})}(\boldsymbol{\beta}, \boldsymbol{u}) &\propto \exp\left\{ \begin{bmatrix} \boldsymbol{\upsilon} \\ \mathrm{vec}\left(\boldsymbol{\upsilon}\boldsymbol{\upsilon}^\mathsf{T}\right) \end{bmatrix}^\mathsf{T} \boldsymbol{\eta}_{p(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{R}}) \to (\boldsymbol{\beta}, \boldsymbol{u})} \right\},
\end{aligned}
$$

$$
\begin{aligned}
m_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2)\to(\boldsymbol{\beta},\boldsymbol{u})}(\boldsymbol{\beta},\boldsymbol{u}) &\propto \exp\left\{\begin{bmatrix}\boldsymbol{\upsilon}\\ \mathrm{vec}\left(\boldsymbol{\upsilon}\boldsymbol{\upsilon}^\mathsf{T}\right)\end{bmatrix}^\mathsf{T}\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2)\to(\boldsymbol{\beta},\boldsymbol{u})}\right\},\\
m_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2)\to\sigma_\varepsilon^2}(\sigma_\varepsilon^2) &\propto \exp\left\{\begin{bmatrix}\log(\sigma_\varepsilon^2)\\ 1/\sigma_\varepsilon^2\end{bmatrix}^\mathsf{T}\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2)\to\sigma_\varepsilon^2}\right\},\\
m_{p(\sigma_\varepsilon^2|a_\varepsilon)\to a_\varepsilon}(\sigma_\varepsilon^2) &\propto \exp\left\{\begin{bmatrix}\log(\sigma_\varepsilon^2)\\ 1/\sigma_\varepsilon^2\end{bmatrix}^\mathsf{T}\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to\sigma_\varepsilon^2}\right\},\\
m_{p(\sigma_\varepsilon^2|a_\varepsilon)\to a_\varepsilon}(a_\varepsilon) &\propto \exp\left\{\begin{bmatrix}\log(a_\varepsilon)\\ 1/a_\varepsilon\end{bmatrix}^\mathsf{T}\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to a_\varepsilon}\right\},\\
m_{p(a_\varepsilon)\to a_\varepsilon}(a_\varepsilon) &\propto \exp\left\{\begin{bmatrix}\log(a_\varepsilon)\\ 1/a_\varepsilon\end{bmatrix}^\mathsf{T}\boldsymbol{\eta}_{p(a_\varepsilon)\to a_\varepsilon}\right\},
\end{aligned}
$$

where the subscripted $\boldsymbol{\eta}$ vectors are natural parameters for a particular exponential family density function and are initialised as, for example,

$$
\boldsymbol{\eta}_{p(a_r^\mathrm{R})\to a_r^\mathrm{R}} \leftarrow \begin{bmatrix}-\frac{3}{2}\\ -\frac{1}{A_{\mathrm{R}r}^2}\end{bmatrix} \quad \text{and} \quad \boldsymbol{\eta}_{p(\boldsymbol{\Sigma}^\mathrm{R}|\boldsymbol{a}^\mathrm{R})\to\boldsymbol{a}^\mathrm{R}} \leftarrow \begin{bmatrix}-2\\ -1\end{bmatrix}.
$$

Continuing this process leads to the first two and last four messages to be proportional to the inverse-Gamma distributions, the third and fourth messages are proportional to the inverse-Wishart distributions, and the fifth and sixth messages are proportional to the $(p+mq^\mathrm{R})$-dimensional multivariate normal distributions.

The stochastic node to factor messages have the same functional forms as their reverse messages. The natural statistic vectors of messages remain unchanged and hence the VMP iterations boil down to computing and updating the corresponding natural parametric vectors. For example,

$$
m_{a_r^\mathrm{R}\to p(a_r^\mathrm{R})}(a_r^\mathrm{R}) \propto \exp\left\{\sum_{r=1}^{q^\mathrm{R}}\begin{bmatrix}\log(a_r^\mathrm{R})\\ 1/a_r^\mathrm{R}\end{bmatrix}^\mathsf{T}\boldsymbol{\eta}_{a_r^\mathrm{R}\to p(a_r^\mathrm{R})}\right\}.
$$

According to (6.8), the next step involves the updates of stochastic node to factor natural parameter vectors which have simple forms as follows:

$$
\begin{aligned}
&\boldsymbol{\eta}_{a_r^\mathrm{R}\to p(a_r^\mathrm{R})} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{\Sigma}^\mathrm{R}|a_r^\mathrm{R})\to a_r^\mathrm{R}}, \quad \boldsymbol{\eta}_{a_r^\mathrm{R}\to p(\boldsymbol{\Sigma}^\mathrm{R}|a_r^\mathrm{R})} \leftarrow \boldsymbol{\eta}_{p(a_r^\mathrm{R})\to a_r^\mathrm{R}}\\
&\boldsymbol{\eta}_{\boldsymbol{\Sigma}^\mathrm{R}\to p(\boldsymbol{\Sigma}^\mathrm{R}|\boldsymbol{a}^\mathrm{R})} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^\mathrm{R})\to\boldsymbol{\Sigma}^\mathrm{R}}, \quad \boldsymbol{\eta}_{\boldsymbol{\Sigma}^\mathrm{R}\to p(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^\mathrm{R})} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{\Sigma}^\mathrm{R}|a^\mathrm{R})\to\boldsymbol{\Sigma}^\mathrm{R}}\\
&\boldsymbol{\eta}_{(\boldsymbol{\beta},\boldsymbol{u})\to p(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^\mathrm{R})} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2)\to(\boldsymbol{\beta},\boldsymbol{u})}, \quad \boldsymbol{\eta}_{(\boldsymbol{\beta},\boldsymbol{u})\to p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2)} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^\mathrm{R})\to(\boldsymbol{\beta},\boldsymbol{u})}
\end{aligned}
$$

$$\boldsymbol{\eta}_{\sigma_\varepsilon^2 \to p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2)} \leftarrow \boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon) \to \sigma_\varepsilon^2}, \quad \boldsymbol{\eta}_{\sigma_\varepsilon^2 \to p(\sigma_\varepsilon^2|a_\varepsilon)} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2) \to \sigma_\varepsilon^2},$$
$$\boldsymbol{\eta}_{a_\varepsilon \to p(a_\varepsilon)} \leftarrow \boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon) \to a_\varepsilon}, \quad \boldsymbol{\eta}_{a_\varepsilon \to p(\sigma_\varepsilon^2|a_\varepsilon)} \leftarrow \boldsymbol{\eta}_{p(a_\varepsilon) \to a_\varepsilon}.$$

For model (6.13) each stochastic node to factor message is solely dependent upon one corresponding factor to stochastic node message due to the linear association between nodes as depicted in Figure 6.3. Based on (6.8) and (6.9),

$$m_{p(a_r^{\mathrm{R}}) \to a_r^{\mathrm{R}}}(a_r^{\mathrm{R}}) \propto p(a_r^{\mathrm{R}}) \quad \text{and} \quad m_{p(a_\varepsilon) \to a_\varepsilon} \propto p(a_\varepsilon),$$

so the natural parameter updates for these two messages are simply

$$\boldsymbol{\eta}_{p(a_r^{\mathrm{R}}) \to a_r^{\mathrm{R}}} \leftarrow \begin{bmatrix} -\frac{3}{2} \\ -\frac{1}{A_{\mathrm{R}r}^2} \end{bmatrix} \quad \boldsymbol{\eta}_{p(a_\varepsilon) \to a_\varepsilon} \leftarrow \begin{bmatrix} -\frac{3}{2} \\ -\frac{1}{A_\varepsilon^2} \end{bmatrix}$$

and remain constant throughout the iterations. The following step according to (6.9) is concerned with updating the factor to stochastic node messages. We focus on the stochastic node $(\boldsymbol{\beta}, \boldsymbol{u})$ that involves updating

$$m_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2) \to (\boldsymbol{\beta},\boldsymbol{u})}(\boldsymbol{\beta},\boldsymbol{u}) \quad \text{and} \quad m_{p(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{R}}) \to (\boldsymbol{\beta},\boldsymbol{u})}(\boldsymbol{\beta},\boldsymbol{u}).$$

We begin with the first message

$$\begin{aligned} & m_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2) \to (\boldsymbol{\beta},\boldsymbol{u})} \\ & \propto \exp\left\{ \int_0^\infty m_{\sigma_\varepsilon^2 \to p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2)}\, m_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2) \to \sigma_\varepsilon^2} \log p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2) \right\} d\sigma_\varepsilon^2 \\ & \propto \exp\left\{ \begin{bmatrix} \boldsymbol{v} \\ \mathrm{vec}\left(\boldsymbol{v}\boldsymbol{v}^\mathsf{T}\right) \end{bmatrix}^\mathsf{T} \begin{bmatrix} \boldsymbol{C}^\mathsf{T} \left(\int_0^\infty \frac{1}{\sigma_\varepsilon^2}\, p_{\mathrm{IG_{nat}}}(\sigma_\varepsilon^2, \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2) \leftrightarrow \sigma_\varepsilon^2})\, d\sigma_\varepsilon^2 \right) \boldsymbol{y} \\ -\frac{1}{2}\mathrm{vec}\left\{ \boldsymbol{C}^\mathsf{T} \left(\int_0^\infty \frac{1}{\sigma_\varepsilon^2}\, p_{\mathrm{IG_{nat}}}(\sigma_\varepsilon^2, \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2) \leftrightarrow \sigma_\varepsilon^2})\, d\sigma_\varepsilon^2 \right) \boldsymbol{C} \right\} \end{bmatrix} \right\} \end{aligned} \tag{6.14}$$

and then the second message

$$\begin{aligned} & m_{p(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{R}}) \to (\boldsymbol{\beta},\boldsymbol{u})} \\ & \propto \exp\left\{ \int_{\mathcal{R}_+^d} m_{\boldsymbol{\Sigma}^{\mathrm{R}} \to p(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{R}})}\, m_{p(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{R}}) \to \boldsymbol{\Sigma}^{\mathrm{R}}} \log p(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{R}}) \right\} d\boldsymbol{\Sigma}^{\mathrm{R}} \end{aligned} \tag{6.15}$$

$$\propto \exp\left\{ \begin{bmatrix} \boldsymbol{v} \\ \text{vec}\left(\boldsymbol{v}\boldsymbol{v}^{\mathsf{T}}\right) \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} \mathbf{0} \\ -\frac{1}{2}\,\text{vec}\left\{ \text{blockdiag}\left(\sigma_\beta^{-2}\boldsymbol{I}_p, \boldsymbol{I}_m \otimes \int_{\mathcal{R}_+^d} \frac{1}{\boldsymbol{\Sigma}^{\text{R}}} \right.\right. \right. \\ \left.\left.\left. \times\, p_{\text{IW}_{\text{nat}}}(\boldsymbol{\Sigma}^{\text{R}}; \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2) \leftrightarrow (\boldsymbol{\beta},\boldsymbol{u})})\, d\boldsymbol{\Sigma}^{\text{R}} \right) \right\} \end{bmatrix} \right\}.$$

The expectation in (6.14) and (6.15) reduces to a linear combination of expected natural statistics. Using the algorithmic primitives defined in Section 6.3 then leads to

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2) \to (\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \begin{bmatrix} \boldsymbol{C}^{\mathsf{T}}\mathcal{I}(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2) \leftrightarrow \sigma_\varepsilon^2})\,\boldsymbol{y} \\ -\frac{1}{2}\text{vec}\left(\boldsymbol{C}^{\mathsf{T}}\mathcal{I}(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2) \leftrightarrow \sigma_\varepsilon^2})\,\boldsymbol{C}\right) \end{bmatrix}$$

and

$$\boldsymbol{\eta}_{p(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\text{R}}) \to (\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \begin{bmatrix} \mathbf{0} \\ -\frac{1}{2}\,\text{vec}\left\{ \text{blockdiag}\left(\sigma_\beta^{-2}\boldsymbol{I}_p, \boldsymbol{I}_m \otimes \mathcal{G}(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2) \leftrightarrow (\boldsymbol{\beta},\boldsymbol{u})}) \right) \right\} \end{bmatrix},$$

where the functions $\mathcal{G}$ and $\mathcal{I}$ are defined in Section 6.3. We repeat the same procedure for the remaining messages in the model. Once convergence of the messages has been attained, the q-density natural parameters can be obtained as:

$$\begin{aligned}
&\boldsymbol{\eta}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\text{R}}) \to (\boldsymbol{\beta},\boldsymbol{u})} + \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2) \to (\boldsymbol{\beta},\boldsymbol{u})}, \\
&\boldsymbol{\eta}_{q(\sigma_\varepsilon^2)} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2) \to \sigma_\varepsilon^2} + \boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon) \to \sigma_\varepsilon^2}, \\
&\boldsymbol{\eta}_{q(a_\varepsilon)} \leftarrow \boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon) \to a_\varepsilon} + \boldsymbol{\eta}_{p(a_\varepsilon) \to a_\varepsilon}, \\
&\boldsymbol{\eta}_{q(\boldsymbol{\Sigma}^{\text{R}})} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{\Sigma}^{\text{R}}|a^{\text{R}}) \to \boldsymbol{\Sigma}^{\text{R}}} + \boldsymbol{\eta}_{p(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\text{R}}) \to \boldsymbol{\Sigma}^{\text{R}}}, \\
&\text{and} \qquad \boldsymbol{\eta}_{q(a_r^{\text{R}})} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{\Sigma}^{\text{R}}|a_r^{\text{R}}) \to a_r^{\text{R}}} + \boldsymbol{\eta}_{p(a_r^{\text{R}}) \to a_r^{\text{R}}},\ 1 \le r \le q^{\text{R}}.
\end{aligned}$$

Hence the resultant optimal q-densities in natural canonical form are:

$q^*(\boldsymbol{\beta},\boldsymbol{u})$ is the $N_{\text{nat}}\left((\boldsymbol{\eta}_{q(\boldsymbol{\beta},\boldsymbol{u})})_1, (\boldsymbol{\eta}_{q(\boldsymbol{\beta},\boldsymbol{u})})_{2,\text{vec}}\right)$ density function,

$q^*(\sigma_\varepsilon^2)$ is the $\text{Inverse-Gamma}_{\text{nat}}((\boldsymbol{\eta}_{q(\sigma_\varepsilon^2)})_1, (\boldsymbol{\eta}_{q(\sigma_\varepsilon^2)})_2))$ density function,

$q^*(a_\varepsilon)$ is the $\text{Inverse-Gamma}_{\text{nat}}((\boldsymbol{\eta}_{q(a_\varepsilon)})_1, (\boldsymbol{\eta}_{q(a_\varepsilon)})_2))$ density function,

$q^*(\boldsymbol{\Sigma}^{\text{R}})$ is the $\text{Inverse-Wishart}_{\text{nat}}((\boldsymbol{\eta}_{q(\boldsymbol{\Sigma}^{\text{R}})})_1, (\boldsymbol{\eta}_{q(\boldsymbol{\Sigma}^{\text{R}})})_2))$ density function,

$q^*(a_r^{\text{R}})$ is the $\text{Inverse-Gamma}_{\text{nat}}((\boldsymbol{\eta}_{p(a_r^{\text{R}}) \to a_r^{\text{R}}})_1, (\boldsymbol{\eta}_{p(a_r^{\text{R}}) \to a_r^{\text{R}}})_2))$ density function.

Algorithm 17 gives the VMP iterative updates for obtaining optimal parameters for model

(6.13).

6.7 Arbitrarily large models viewpoint

VMP benefits from concepts such as natural parameter vectors and factor graphs, which makes it more amenable to modularisation than MFVB. The VMP algorithms can be easily extended to arbitrarily large models via the notion of *factor graph fragments* that were first introduced in Wand (2015):

Definition 6.7.1. A factor graph fragment is a sub-graph of a factor graph consisting of a single factor and each of its neighbouring stochastic nodes.

Wand (2015) described the factor graph fragments as building blocks for a wide range of VMP-based Bayesian semiparametric mixed models involving, for example, multiple predictors and interactions, penalised splines or wavelets, higher-level random effects and non-Gaussian responses. Interested readers would benefit from reading Section 4 of his paper on derivation of the five fundamental fragments. With the help of these fragment definitions, the updates of natural parameters for the factor to stochastic node messages only need to be derived and coded once and then be integrated into different compartmentalised functions. This potentially saves a huge amount of time in terms of derivation and programming.

Consider an example where we extend the simple mixed model (6.13) to a three-level semiparametric mixed model with Gaussian response

$$
\begin{aligned}
\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2 &\sim N\left(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u}\right), \quad \boldsymbol{\beta}\sim N(0,\sigma_\beta^2\boldsymbol{I}_p), \qquad (6.16)\\
\boldsymbol{u}|\sigma_{u\ell}^2,\boldsymbol{\Sigma}^{\text{RL2}},\boldsymbol{\Sigma}^{\text{RL3}} &\sim N\left(\boldsymbol{0},\begin{bmatrix}\underset{1\le\ell\le L}{\text{blockdiag}}\left(\sigma_{u\ell}^2\boldsymbol{I}_{q_\ell^{\text{G}}}\right) & \boldsymbol{0}\\ \boldsymbol{0} & \underset{1\le i\le m}{\text{blockdiag}}\begin{bmatrix}\boldsymbol{\Sigma}^{\text{RL3}} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}_{n_i}\otimes\boldsymbol{\Sigma}^{\text{RL2}}\end{bmatrix}\end{bmatrix}\right),\\
\boldsymbol{\Sigma}^{\text{RL2}}|a_1^{\text{RL2}},\ldots,a_{q^{\text{RL2}}}^{\text{RL2}} &\sim \text{Inverse-Wishart}\left(\nu+q^{\text{RL2}}-1,\ 2\,\nu\,\text{diag}(1/a_1^{\text{RL2}},\ldots,1/a_{q^{\text{RL2}}}^{\text{RL2}})\right),\\
\boldsymbol{\Sigma}^{\text{RL3}}|a_1^{\text{RL3}},\ldots,a_{q^{\text{RL3}}}^{\text{RL3}} &\sim \text{Inverse-Wishart}\left(\nu+q^{\text{RL3}}-1,\ 2\,\nu\,\text{diag}(1/a_1^{\text{RL3}},\ldots,1/a_{q^{\text{RL3}}}^{\text{RL3}})\right),\\
a_{r_{\text{L2}}}^{\text{RL2}} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(\tfrac{1}{2},1/A_{\text{RL2}}^2\right), \quad r_{\text{L2}}=1,\ldots,q^{\text{RL2}},\\
a_{r_{\text{L3}}}^{\text{RL3}} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(\tfrac{1}{2},1/A_{\text{RL3}}^2\right), \quad r_{\text{L3}}=1,\ldots,q^{\text{RL3}},\\
\sigma_{u\ell}^2|a_{u\ell} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2},\,1/a_{u\ell}), \quad a_{u\ell}\overset{\text{ind.}}{\sim}\text{Inverse-Gamma}(\tfrac{1}{2},\,1/A_{u_\ell}^2),
\end{aligned}
$$

where the model parameters are previously defined in Subsection 3.7.1. The joint posterior density function of all random variables in model (6.16) admits the factorisation

$$
p(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{\Sigma}^{\text{RL2}},\boldsymbol{\Sigma}^{\text{RL3}},\boldsymbol{\sigma}_u^2,\sigma_\varepsilon^2,\boldsymbol{a}^{\text{RL2}},\boldsymbol{a}^{\text{RL2}},\boldsymbol{a}_u,a_\varepsilon) = \qquad (6.17)
$$

Initialise all factor to stochastic node messages to be arbitrary density functions:

$$\boldsymbol{\eta}_{p(a_r^{\mathrm{R}}) \to a_r^{\mathrm{R}}} \leftarrow \begin{bmatrix} -\frac{3}{2} \\ -\frac{1}{A_{\mathrm{R}r}^2} \end{bmatrix} \quad ; \quad \boldsymbol{\eta}_{p(\boldsymbol{\Sigma}^{\mathrm{R}}|a_r^{\mathrm{R}}) \to a_r^{\mathrm{R}}} \leftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix} \quad ; \quad \boldsymbol{\eta}_{p(\boldsymbol{\Sigma}^{\mathrm{R}}|a^{\mathrm{R}}) \to \boldsymbol{\Sigma}^{\mathrm{R}}} \leftarrow \begin{bmatrix} -3 \\ -\frac{1}{2}\mathrm{vec}(\boldsymbol{I}_2) \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{R}}) \to \boldsymbol{\Sigma}^{\mathrm{R}}} \leftarrow \begin{bmatrix} -3 \\ -\frac{1}{2}\mathrm{vec}(\boldsymbol{I}_2) \end{bmatrix} \quad ; \quad \boldsymbol{\eta}_{p(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{R}}) \to (\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \begin{bmatrix} -\mathbf{0}_{p+mq^{\mathrm{R}}} \\ -\frac{1}{2}\mathrm{vec}(\boldsymbol{I}_{p+mq^{\mathrm{R}}}) \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2) \to (\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \begin{bmatrix} -\mathbf{0}_{p+mq^{\mathrm{R}}} \\ -\frac{1}{2}\mathrm{vec}(\boldsymbol{I}_{p+mq^{\mathrm{R}}}) \end{bmatrix} \quad ; \quad \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2) \to \sigma_\varepsilon^2} \leftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon) \to \sigma_\varepsilon^2} \leftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix} \quad ; \quad \boldsymbol{\eta}_{p(a_\varepsilon) \to a_\varepsilon} \leftarrow \begin{bmatrix} -\frac{3}{2} \\ -\frac{1}{A_\varepsilon^2} \end{bmatrix}$$

Cycle through updates:

(a) For stochastic node to factor messages:

For $r = 1, \ldots, q^{\mathrm{R}}$

$$\boldsymbol{\eta}_{a_r^{\mathrm{R}} \to p(a_r^{\mathrm{R}})} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{\Sigma}^{\mathrm{R}}|a_r^{\mathrm{R}}) \to a_r^{\mathrm{R}}} \quad ; \quad \boldsymbol{\eta}_{a_r^{\mathrm{R}} \to p(\boldsymbol{\Sigma}^{\mathrm{R}}|a_r^{\mathrm{R}})} \leftarrow \boldsymbol{\eta}_{p(a_r^{\mathrm{R}}) \to a_r^{\mathrm{R}}}$$

$$\boldsymbol{\eta}_{\boldsymbol{\Sigma}^{\mathrm{R}} \to p(\boldsymbol{\Sigma}^{\mathrm{R}}|\boldsymbol{a}^{\mathrm{R}})} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{R}}) \to \boldsymbol{\Sigma}^{\mathrm{R}}} \quad ; \quad \boldsymbol{\eta}_{\boldsymbol{\Sigma}^{\mathrm{R}} \to p(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{R}})} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{\Sigma}^{\mathrm{R}}|a^{\mathrm{R}}) \to \boldsymbol{\Sigma}^{\mathrm{R}}}$$

$$\boldsymbol{\eta}_{(\boldsymbol{\beta},\boldsymbol{u}) \to p(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{R}})} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2) \to (\boldsymbol{\beta},\boldsymbol{u})} \quad ; \quad \boldsymbol{\eta}_{(\boldsymbol{\beta},\boldsymbol{u}) \to p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2)} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{R}}) \to (\boldsymbol{\beta},\boldsymbol{u})}$$

$$\boldsymbol{\eta}_{\sigma_\varepsilon^2 \to p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2)} \leftarrow \boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon) \to \sigma_\varepsilon^2} \quad ; \quad \boldsymbol{\eta}_{\sigma_\varepsilon^2 \to p(\sigma_\varepsilon^2|a_\varepsilon)} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2) \to \sigma_\varepsilon^2}$$

$$\boldsymbol{\eta}_{a_\varepsilon \to p(a_\varepsilon)} \leftarrow \boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon) \to a_\varepsilon} \quad ; \quad \boldsymbol{\eta}_{a_\varepsilon \to p(\sigma_\varepsilon^2|a_\varepsilon)} \leftarrow \boldsymbol{\eta}_{p(a_\varepsilon) \to a_\varepsilon}$$

(b) For factor to stochastic node messages:

For $r = 1, \ldots, q^{\mathrm{R}}$

$$\boldsymbol{\eta}_{p(a_r^{\mathrm{R}}) \to a_r^{\mathrm{R}}} \leftarrow \begin{bmatrix} -\frac{3}{2} \\ -\frac{1}{A_{\mathrm{R}r}^2} \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{\Sigma}^{\mathrm{R}}|a_r^{\mathrm{R}}) \to a_r^{\mathrm{R}}} \leftarrow \begin{bmatrix} -\frac{1}{2}(\nu + q^{\mathrm{R}} - 1) \\ -\nu\, \mathcal{G}\left(\boldsymbol{\eta}_{p(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{R}}) \to \boldsymbol{\Sigma}^{\mathrm{R}}} + \boldsymbol{\eta}_{p(\boldsymbol{\Sigma}^{\mathrm{R}}|a^{\mathrm{R}}) \to \boldsymbol{\Sigma}^{\mathrm{R}}}\right) \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{\Sigma}^{\mathrm{R}}|a^{\mathrm{R}}) \to \boldsymbol{\Sigma}^{\mathrm{R}}} \leftarrow \begin{bmatrix} -\frac{1}{2}(A_{\boldsymbol{\Sigma}^{\mathrm{R}}} + q^{\mathrm{R}} - 1) \\ -\frac{1}{2}\sum_{r=1}^{2} \mathrm{vec}\left\{2\nu \mathrm{diag}\left(\mathcal{I}(\boldsymbol{\eta}_{p(\boldsymbol{\Sigma}^{\mathrm{R}}|a_r^{\mathrm{R}}) \to a_r^{\mathrm{R}}} + \boldsymbol{\eta}_{p(a_r^{\mathrm{R}}) \to a_r^{\mathrm{R}}})\right)\right\} \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{R}}) \to \boldsymbol{\Sigma}^{\mathrm{R}}} \leftarrow \begin{bmatrix} -\frac{1}{2}\sum_{i=1}^{m} n_i \\ \mathcal{F}\left(\boldsymbol{\eta}_{(\boldsymbol{\beta},\boldsymbol{u}) \to p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2)} + \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2) \to (\boldsymbol{\beta},\boldsymbol{u})}\right) \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2) \to (\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \begin{bmatrix} \boldsymbol{C}^{\mathsf{T}}\mathcal{I}\left(\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon) \to a_\varepsilon} + \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2) \to \sigma_\varepsilon^2}\right)\boldsymbol{y} \\ \frac{1}{2}\mathrm{vec}\left\{\boldsymbol{C}^{\mathsf{T}}\mathcal{I}\left(\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon) \to a_\varepsilon} + \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2) \to \sigma_\varepsilon^2}\right)\boldsymbol{C}\right\} \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{R}})\to(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \begin{bmatrix} \begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix} \\ -\frac{1}{2}\mathrm{vec}\left(\begin{bmatrix}\sigma_\beta^{-2}\boldsymbol{I}_p & \mathbf{0}\\ \mathbf{0} & \boldsymbol{I}_m\otimes\mathcal{G}(\boldsymbol{\eta}_{p(\boldsymbol{\Sigma}^{\mathrm{R}}|a^{\mathrm{R}})\to\boldsymbol{\Sigma}^{\mathrm{R}}}+\boldsymbol{\eta}_{p(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{R}})\to\boldsymbol{\Sigma}^{\mathrm{R}}})\end{bmatrix}\right)\end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2)\to\sigma_\varepsilon^2} \leftarrow$$

$$\begin{bmatrix} -\frac{1}{2}\sum_{i=1}^m n_i \\ -\frac{1}{2}\sum_{i=1}^m\sum_{j=1}^{n_i}\mathrm{vec}\left(\mathcal{H}(\boldsymbol{\eta}_{p(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{R}})\to(\boldsymbol{\beta},\boldsymbol{u})}+\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2)\to(\boldsymbol{\beta},\boldsymbol{u})};\boldsymbol{y}_{ij};\boldsymbol{C}_{ij})\right)\end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to\sigma_\varepsilon^2} \leftarrow \begin{bmatrix} -\frac{3}{2} \\ -\mathcal{I}\left(\boldsymbol{\eta}_{p(a_\varepsilon)\to a_\varepsilon}+\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to a_\varepsilon}\right)\end{bmatrix}$$

$$\boldsymbol{\eta}_{p(a_\varepsilon)\to a_\varepsilon} \leftarrow \begin{bmatrix} -\frac{3}{2} \\ -\frac{1}{A_\varepsilon^2}\end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to a_\varepsilon} \leftarrow \begin{bmatrix} -\frac{1}{2} \\ -\mathcal{I}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2)\to\sigma_\varepsilon^2}+\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to\sigma_\varepsilon^2}\right)\end{bmatrix}$$

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

Obtain q-density natural parameters via summation of messages:

$$\boldsymbol{\eta}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{R}})\to(\boldsymbol{\beta},\boldsymbol{u})}+\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2)\to(\boldsymbol{\beta},\boldsymbol{u})}$$

$$\boldsymbol{\eta}_{q(\sigma_\varepsilon^2)} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2)\to\sigma_\varepsilon^2}+\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to\sigma_\varepsilon^2}$$

$$\boldsymbol{\eta}_{q(a_\varepsilon)} \leftarrow \boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to a_\varepsilon}+\boldsymbol{\eta}_{p(a_\varepsilon)\to a_\varepsilon}$$

$$\boldsymbol{\eta}_{q(\boldsymbol{\Sigma}^{\mathrm{R}})} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{\Sigma}^{\mathrm{R}}|a^{\mathrm{R}})\to\boldsymbol{\Sigma}^{\mathrm{R}}}+\boldsymbol{\eta}_{p(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{R}})\to\boldsymbol{\Sigma}^{\mathrm{R}}}$$

For $r=1,\ldots,q^{\mathrm{R}}$

$$\boldsymbol{\eta}_{q(a_r^{\mathrm{R}})} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{\Sigma}^{\mathrm{R}}|a_r^{\mathrm{R}})\to a_r^{\mathrm{R}}}+\boldsymbol{\eta}_{p(a_r^{\mathrm{R}})\to a_r^{\mathrm{R}}}$$

Algorithm 17: Variational message passing iterative scheme for obtaining the optimal parameters for model (6.13).

$$p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2_\varepsilon)\ p(\boldsymbol{\beta},\boldsymbol{u}|(\boldsymbol{\Sigma}^{\text{RL2}},\boldsymbol{\Sigma}^{\text{RL3}},\boldsymbol{\sigma}^2_u)\ p(\boldsymbol{\Sigma}^{\text{RL2}}|\boldsymbol{a}^{\text{RL2}})\ p(\boldsymbol{\Sigma}^{\text{RL3}}|\boldsymbol{a}^{\text{RL3}})$$
$$\times p(\boldsymbol{\sigma}^2_u|\boldsymbol{a}_u)\ \ p(\sigma^2_\varepsilon|a_\varepsilon)\ p(\boldsymbol{a}^{\text{RL2}})\ p(\boldsymbol{a}^{\text{RL3}})\ p(\boldsymbol{a}_u)\ p(a_\varepsilon).$$

Treating (6.17) as a set of parameter functions corresponding to each factor gives the factor graph in Figure 6.4. This factor graph has 10 graph fragments, with five of them have the same form as the fragments of Figure 6.3 and are coloured in light blue. The darker blue fragments are the additional ones.

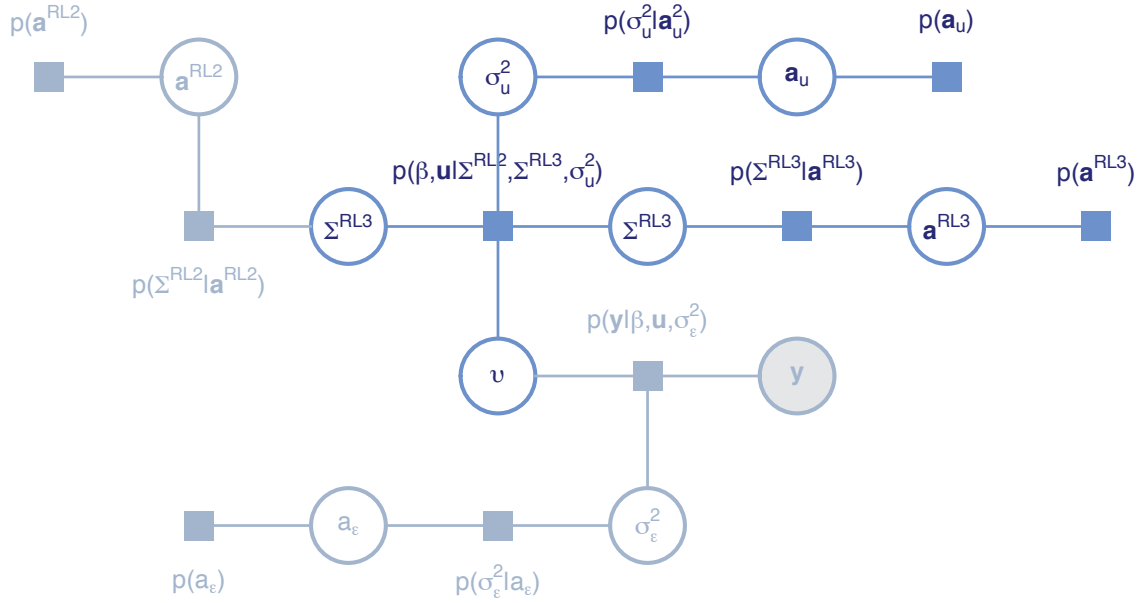

Figure 6.4: Diagrammatic depiction of the extension from simple mixed effects regression to three-level penalised spline mixed regression. The fragments shown in darker blue are the ones that are additional to the fragments in the simple mixed model. The fragments shown in grey are presented in Figure 6.3.

The stochastic node to factor messages in Figure 6.4 have trivial updates analogous to those given in Section 6.6. The factor to stochastic node messages are however more complicated, but the majority of the darker blue fragments have identical or very similar updates as the ones given in factor graph Figure 6.3. For example, the messages passed from $p(\boldsymbol{\sigma}^2_u|\boldsymbol{a}_u)$ to $\boldsymbol{\sigma}^2_u$ and $\boldsymbol{a}_u$ have the same form as those passed from $p(\sigma^2_\varepsilon)$ to σ^2_ε and a_ε for model (6.13). The natural parameter updates $\boldsymbol{\eta}_{p(\boldsymbol{\sigma}^2_u|\boldsymbol{a}_u)\to\boldsymbol{\sigma}^2_u}$ and $\boldsymbol{\eta}_{p(\boldsymbol{\sigma}^2_u|\boldsymbol{a}_u)\to\boldsymbol{a}_u}$ take the same forms as those for $\boldsymbol{\eta}_{p(\sigma^2_\varepsilon|a_\varepsilon)\to\sigma^2_\varepsilon}$ and $\boldsymbol{\eta}_{p(\sigma^2_\varepsilon|a_\varepsilon)\to a_\varepsilon}$ given in Section 6.6. Similarly, the message passed from $p(\boldsymbol{a}_u)$ to $\boldsymbol{a}_u$ has the same form as that passed from $p(a_\varepsilon)$ to a_ε. The natural parameter updates $\boldsymbol{\eta}_{p(\boldsymbol{a}_u)\to\boldsymbol{a}_u}$ takes the same form as that for $\boldsymbol{\eta}_{p(a_\varepsilon)\to a_\varepsilon}$. Similar arguments can be applied to the messages passed from $p(\boldsymbol{\Sigma}^{\text{RL3}}|\boldsymbol{a}^{\text{RL3}})$ to $\boldsymbol{\Sigma}^{\text{RL3}}$ and $\boldsymbol{a}^{\text{RL3}}$, and messages from $p(\boldsymbol{a}^{\text{RL3}})$ to $\boldsymbol{a}^{\text{RL3}}$. It remains to take care of the darker blue fragments that are at the centre of the factor graph. The message passed from

$p(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{RL2}}, \boldsymbol{\Sigma}^{\mathrm{RL3}}, \boldsymbol{\sigma}_u^2)$ to $\boldsymbol{\sigma}_u^2$ is

$$m_{p(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{RL2}}, \boldsymbol{\Sigma}^{\mathrm{RL3}}, \boldsymbol{\sigma}_u^2) \to \sigma_{u\ell}^2}(\sigma_{u\ell}^2) = \exp\left\{ \sum_{\ell=1}^{L} \begin{bmatrix} \log(\sigma_{u\ell}^2) \\ 1/\sigma_{u\ell}^2 \end{bmatrix}^{\mathsf{T}} \boldsymbol{\eta}_{p(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{RL2}}, \boldsymbol{\Sigma}^{\mathrm{RL3}}, \sigma_{u\ell}^2) \to \sigma_{u\ell}^2} \right\},$$

for $\ell = 1, \ldots, L$, where

$$\boldsymbol{\eta}_{p(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{RL2}}, \boldsymbol{\Sigma}^{\mathrm{RL3}}, \sigma_{u\ell}^2) \to \sigma_{u\ell}^2} \leftarrow \begin{bmatrix} -\frac{1}{2} q_\ell^{\mathrm{G}} \\ \mathcal{F}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2) \to (\boldsymbol{\beta}, \boldsymbol{u})} + \boldsymbol{\eta}_{p(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{RL2}}, \boldsymbol{\Sigma}^{\mathrm{RL3}}, \boldsymbol{\sigma}_u^2) \to (\boldsymbol{\beta}, \boldsymbol{u})} \right) \end{bmatrix}.$$

The message passed from $p(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{RL2}}, \boldsymbol{\Sigma}^{\mathrm{RL3}}, \boldsymbol{\sigma}_u^2)$ to $(\boldsymbol{\beta}, \boldsymbol{u})$ is

$$m_{p(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{RL2}}, \boldsymbol{\Sigma}^{\mathrm{RL3}}, \boldsymbol{\sigma}_u^2) \to (\boldsymbol{\beta}, \boldsymbol{u})}(\boldsymbol{\beta}, \boldsymbol{u}) = \exp\left\{ \begin{bmatrix} \boldsymbol{v} \\ \mathrm{vec}\left(\boldsymbol{v}\boldsymbol{v}^{\mathsf{T}}\right) \end{bmatrix}^{\mathsf{T}} \boldsymbol{\eta}_{p(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{RL2}}, \boldsymbol{\Sigma}^{\mathrm{RL3}}, \sigma_{u\ell}^2) \to (\boldsymbol{\beta}, \boldsymbol{u})} \right\},$$

where

$$\boldsymbol{\eta}_{p(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{RL2}}, \boldsymbol{\Sigma}^{\mathrm{RL3}}, \sigma_{u\ell}^2) \to (\boldsymbol{\beta}, \boldsymbol{u})} \leftarrow \begin{bmatrix} \mathbf{0} \\ -\frac{1}{2}\mathrm{vec}\left\{ \mathrm{blockdiag}\left(\sigma_\beta^{-2}\boldsymbol{I},\, \boldsymbol{\Omega}_{\boldsymbol{\Sigma}},\, \boldsymbol{\Omega}_{\boldsymbol{u}} \right) \right\} \end{bmatrix},$$

$$\boldsymbol{\Omega}_{\boldsymbol{\Sigma}} = \mathrm{blockdiag}_{1 \le i \le m} \begin{bmatrix} \boldsymbol{I}_{ni} \otimes \mathcal{F}(\boldsymbol{\eta}_{p(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{RL2}}, \boldsymbol{\Sigma}^{\mathrm{RL3}}, \boldsymbol{\sigma}_u^2) \leftrightarrow \boldsymbol{\Sigma}^{\mathrm{RL2}})} & \mathbf{0} \\ \mathbf{0} & \mathcal{F}(\boldsymbol{\eta}_{p(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{RL2}}, \boldsymbol{\Sigma}^{\mathrm{RL3}}, \boldsymbol{\sigma}_u^2) \leftrightarrow \boldsymbol{\Sigma}^{\mathrm{RL3}})} \end{bmatrix},$$

$$\boldsymbol{\Omega}_{\boldsymbol{u}} = \mathrm{blockdiag}_{1 \le i \le m} \left(\mathcal{I}\left(\boldsymbol{\eta}_{p(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{RL2}}, \boldsymbol{\Sigma}^{\mathrm{RL3}}, \boldsymbol{\sigma}_u^2) \leftrightarrow \sigma_{u\ell}^2} \right) \boldsymbol{I}_{q_\ell^{\mathrm{G}}} \right),$$

where the functions $\mathcal{F}$ and $\mathcal{I}$ are defined in Section 6.3. The message passed from $p(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{RL2}}, \boldsymbol{\Sigma}^{\mathrm{RL3}}, \boldsymbol{\sigma}_u^2)$ to $\boldsymbol{\Sigma}^{\mathrm{RL2}}$ or $\boldsymbol{\Sigma}^{\mathrm{RL2}}$ takes the same form as that passed from $p(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{R}})$ to $\boldsymbol{\Sigma}^{\mathrm{R}}$ in Section 6.6.

6.8 Concluding remarks

We demonstrated that the notion of message passing can be used to streamline algebra calculations and programming for fast approximate inference in large Bayesian Gaussian semiparametric mixed models. The resultant VMP algorithms have ready-to-implement closed-form expressions and are easier to extend to a broad class of arbitrarily large

Bayesian mixed models, via factor graph fragments, than the MFVB approach.

Chapter 7

Conclusions and Future Directions

The goal is to transform data into information, information into insight.

Carly Fiorina

We have exhaustively exemplified the use of variational methods in the setting of Bayesian longitudinal and multilevel data analysis. In this thesis we introduced sophisticated variational inference techniques for fitting and inference in a wide variety of Bayesian semiparametric mixed models. Four common response distributions, including Gaussian, Student-t, Bernoulli and Poisson, were explored. MFVB is a principled approach to approximate inference, where it defines a family of approximating distributions whose members are simpler in some sense than the actual posterior and then seeks the optimal member of that family to minimising the Kullback-Leibler divergence (the dissimilarity criterion) between the approximating distribution and the exact posterior. Through a series of numerical studies and real data examples, we demonstrated that MFVB approximations have attractive computation and accuracy trade-offs. To further improve speed and memory efficiency of the naïve/direct MFVB algorithms we used matrix permutation and block decomposition to streamline the estimation of the large sparse effects covariance matrix. Rather than deciding whether this new methodology dominates the traditional ones in general, we discussed the strengths and weaknesses of the variational and standard MCMC methods. While MCMC sampling provides theoretical guarantees of accuracy, variational methods, in general, suffer some accuracy loss, albeit minimal, by the nature of the restriction on which they are based. In several instances, variational approximations can be shown to be overconfident, i.e. they underestimate posterior variances (e.g. Bishop, 2006; Wang and Titterington, 2006). This raises the question of whether this underestimation

affects model selection when variance components are a focus? As previously shown in You *et al.* (2014) such criticism does not hold for general variational methods in an asymptotic sense. The authors have proven the variational estimators for linear regression models achieve desirable frequentist properties such as consistency and can obtain asymptotically valid standard errors under mild regularity conditions. In addition, Ormerod *et al.* (2014) extended You *et al.* (2014)'s approach by inducing sparsity on the linear regression coefficients via a spike and slab prior and showed that the resultant estimates are consistent with valid standard errors, and that the true model can be found at an exponential rate in n. Extending these theoretical results to non-Gaussian responses and other data scenarios was not considered in the above papers but would be worthwhile to investigate in the future. Nonetheless, variational algorithms provide impressively fast and analytical approximations to otherwise unattainable posteriors. For example, we found that variational methods can be exploited to yield closed-form expressions that approximate the posteriors for the logistic regression effect parameters. It is good to keep in mind that both MCMC and variational methods are computational paradigms, each offers a range of different algorithmic approaches which trade off between speed, accuracy and ease of implementation.

The improvements in speed and memory efficiency offered by MFVB are not a simple novelty. Putting accuracy considerations aside, the algebraic infrastructure we laid out in this thesis has far-reaching implications beyond the examples considered here, for the analysis of big datasets via Bayesian hierarchical models as both continue to grow in size and complexity. We showed the versatility and utility of MFVB algorithms in terms of speed for moderately large data, and easy extendability to complicated, but more realistic scenarios, including semiparametric regression, higher-level mixed models, measurement error and missing data models, models with group-specific curves, and real-time/online processing for high velocity data. The contributions made in this thesis open up a broader class of models considering variational methods as a strong complement to the more traditional MCMC methods for exploratory data analysis and iterative model fitting. Furthermore, in order to get the best of both worlds, the MFVB approximation can always be used to initialise a sampling based approach which tends to asymptotically recover the actual posterior. Thus the fast variational methods can be used to quickly explore and respecify models which then lead to improved convergence of the asymptotically exact MCMC sampling algorithms.

Variational inference opens the door to several promising research directions. The MFVB algorithms derived in this thesis provide an efficient and convenient iterative scheme for updating the variational parameters, but in most examples we used conjugate priors, often chosen for reasons of tractability. If $q_i(\boldsymbol{\theta}_i)$ does not belong to a recognisable density family, some optimisation techniques are required to estimate the marginal likeli-

hood. Future work can move beyond closed-form updates and fully factorised approximate posteriors, as well as allow arbitrary priors for the hyperparameters (e.g. Neville *et al.*, 2014). For examples, Teh *et al.* (2006) introduce *collapsed variational inference*, which marginalises out some of the hidden variables, training simple closed-form updates for low-dimensional posteriors. Hoffman *et al.* (2013) develop *stochastic variational inference* that facilitates stochastic optimisation to optimise the variational objective. The authors use structured variational distributions to relax the mean-field restriction for better approximating complex posteriors such as those arising in time-series models. Tan and Nott (2014) extend the stochastic variational inference approach of Hoffman *et al.* (2013) by combining non-conjugate variational messaging passing with algorithms from stochastic optimisation which work with mini-batches of data, with applications to non-conjugate generalised linear mixed models. Related work on non-conjugate models by Wang and Blei (2013) use the Laplace approximation for updating non-conjugate variational factors and, as described in the previous chapters, Knowles and Minka (2011) introduce the non-conjugate variational message passing framework for variational approximations in the exponential family. Finally, the popular divide and recombine strategy introduced in Nott *et al.* (2013) involves partitioning a large dataset into smaller subsets and then combining the approximating distributions that have been learned in parallel from each separate subset using the *hybrid variational Bayes algorithm*, a combination of the mean-field and fixed-form variational methods.

We live in an increasingly data-rich world, with ever greater desires and responsibilities to extract signals from noisy data. Even though the simulated and real data examples considered in this thesis are not gigabytes in size, coupled with modern data subsetting techniques, the variational inference framework presented in this thesis offers the opportunity to develop many fast algorithms that can be applied to even larger datasets which require storage greater than what is possible on a standard computer.

References

Abrevaya, J. (2006). Estimating the effect of smoking on birth outcomes using a matched panel data approach. *Journal of Applied Econometrics*, **21**:489–519.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, **88**:669–679.

Baicker, K., Buckles, K. S., and Chandra, A. (2006). Geographic variation in the appropriate use of cesarean delivery. *Health Affairs*, **25**:w355–w367.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

Bishop, C. M. (2008). A new framework for machine learning. Em *Computational Intelligence: Research Frontiers*, pgs. 1–24. Springer.

Bishop, C. M., Spiegelhalter, D., and Winn, J. (2002). VIBES: a variational inference engine for Bayesian networks. Em *Advances in neural information processing systems*, pgs. 777–784.

Bragg, F., Cromwell, D. A., Edozien, L. C., Gurol-Urganci, I., Mahmood, T. A., Templeton, A., and van der Meulen, J. H. (2010). Variation in rates of caesarean section among english NHS trusts after accounting for maternal and clinical risk: cross sectional study. *British Medical Journal*, **341**:c5065.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**:9–25.

Brumback, B. A. and Rice, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association*, **93**:961–976.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. CRC press.

Challis, E. and Barber, D. (2013). Gaussian Kullback-Leibler approximate inference. *The*

Journal of Machine Learning Research, **14**:2239–2286.

Chappell, M., Groves, A. R., Whitcher, B., and Woolrich, M. W. (2009). Variational Bayesian inference for a nonlinear forward model. *Signal Processing, IEEE Transactions on*, **57**:223–236.

Chen, H. and Wang, Y. (2011). A penalized spline approach to functional mixed effects model analysis. *Biometrics*, **67**:861–870.

Clayton, D. G. (1992). Models for the analysis of cohort and case-control studies with inaccurately measured exposures. *Statistical models for longitudinal studies of health*, pgs. 301–331.

Coull, B. A., Catalano, P. J., and Godleski, J. J. (2000). Semiparametric analyses of cross-over data with repeated measures. *Journal of Agricultural, Biological, and Environmental Statistics*, **5**:417–429.

Daniels, M. J. and Hogan, J. W. (2008). *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis*. CRC Press.

Dechter, R. and Pearl, J. (1988). *Network-based heuristics for constraint-satisfaction problems*. New York: Springer.

Diggle, P. J., Heagerty, P., Liang, K.-Y., and Zeger, S. (2002). *Analysis of longitudinal data*. Oxford: Oxford University Press.

Durbán, M., Harezlak, J., Wand, M. P., and Carroll, R. J. (2005). Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine*, **24**:1153–1167.

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**:959–971.

Fitzmaurice, G. M., Davidian, M., Verbeke, G., and Molenberghs, G. (2008). *Longitudinal data analysis*. CRC Press.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2012). *Applied longitudinal analysis*, volume 998. New Jersey: John Wiley & Sons.

Fong, Y., Rue, H., and Wakefield, J. (2009). Bayesian inference for generalized linear mixed models. *Biostatistics*, pg. kxp053.

Frey, B. J. (1998). Graphical models for machine learning and digital communication. *Massachusetts: MIT press*.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, **85(410)**:398–409.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **1**:515–533.

Gelman, A. and Hill, J. (2007). *Data analysis using regression and multilevel or hierarchical models*. New York: Cambridge University Press.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **(6)**:721–741.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, **73**:43–56.

Goldstein, H. (2011). *Multilevel statistical models*, volume 922. New Jersey: John Wiley & Sons.

Harville, D. A. (2008). *Matrix algebra from a statistician's perspective*, volume 1. New York: Springer.

Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, **27**:83–85.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*, volume 43. CRC Press.

Hoffman, M., Bach, F. R., and Blei, D. M. (2010). Online learning for latent Dirichlet allocation. Em *advances in neural information processing systems*, pgs. 856–864.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research*, **14**:1303–1347.

Huang, A. and Wand, M. P. (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis*, **8**:439–452.

Jaakkola, T. S. and Jordan, M. I. (1997). A variational approach to Bayesian logistic regression models and their extensions. Em *Sixth International Workshop on Artificial Intelligence and Statistics*.

Jaakkola, T. S. and Jordan, M. I. (2000). Bayesian parameter estimation via variational

methods. *Statistics and Computing*, **10**:25–37.

Jordan, M. I. (2004). Graphical models. *Statistical Science*, **19**:140–155.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variation methods for graphical models. *Machine Learning*, **37**:183–233.

Knowles, D. A. and Minka, T. (2011). Non-conjugate variational message passing for multinomial and binary regression. Em *Advances in Neural Information Processing Systems*, pgs. 1701–1709.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**:963–974.

Lee, C. Y. Y., Homer, C., Bisits, A., and Ryan, L. (2015). Increasing variation in hospital caesarean section rates among low-risk nulliparous women in australia, from 1994 to 2010. *Birth in revision*.

Lee, C. Y. Y. and Wand, M. P. (2015a). Streamlined mean field variational Bayes for longitudinal and multilevel data analysis. *Biometrical Journal in press*.

Lee, C. Y. Y. and Wand, M. P. (2015b). Variational methods for fitting complex Bayesian mixed effects models to health data. *Statistics in Medicine. DOI 10.1002/sim.6737*.

Li, Y. and Ruppert, D. (2008). On the asymptotics of penalized splines. *Biometrika*, **95**:415–436.

Ligges, U., Thomas, A., Spiegelhalter, D., Best, N., Lunn, D., Rice, K., and Sturtz, S. (2009). *`BRugs` 0.5: `OpenBUGS` and its `R/S-PLUS` interface `BRugs`*.

Little, R. J. and Rubin, D. B. (2014). *Statistical analysis with missing data*. New Jersey: John Wiley & Sons.

Luts, J. (2015). Real-time semiparametric regression for distributed data sets. *Knowledge and Data Engineering, IEEE Transactions on*, **27**:545–557.

Luts, J., Broderick, T., and Wand, M. P. (2014). Real-time semiparametric regression. *Journal of Computational and Graphical Statistics*, **23**:589–615.

Magnus, J. R. and Neudecker, H. (1995). *Matrix differential calculus with applications in statistics and econometrics*. New Jersey: John Wiley & Sons.

Maringwa, J. T., Geys, H., Shkedy, Z., Faes, C., Molenberghs, G., Aerts, M., Ammel, K. V., Teisman, A., and Bijnens, L. (2008). Application of semiparametric mixed models and simultaneous confidence bands in a cardiovascular safety experiment

with longitudinal data. *Journal of Biopharmaceutical Statistics*, **18**:1043–1062.

Marley, J. and Wand, M. P. (2010). Non-standard semiparametric regression via `brugs`. *Journal of Statistical Software*, **37**:1–30.

McGrory, C. A. and Titterington, D. (2007). Variational approximations in Bayesian model selection for finite mixture distributions. *Computational Statistics and Data Analysis*, **51**:5352–5367.

Menictas, M. (2015). *Variational inference for hesteroscedastic and longitudinal regression models*. PhD thesis, University of Technology Sydney, School of Mathematical and Physical Sciences.

Menictas, M. and Wand, M. P. (2013). Variational inference for marginal longitudinal semiparametric regression. *Stat*, **2**:61–71.

Minka, T. (2005). Divergence measures and message passing. *Technical report, Microsoft Research*.

Minka, T., Winn, J., Guiver, J., and Kannan, A. (2009). *`Infer.NET 2.3`*. Microsoft Research Cambridge, Cambridge, UK.

Neville, S. E., Ormerod, J. T., and Wand, M. P. (2014). Mean field variational Bayes for continuous sparse signal shrinkage: pitfalls and remedies. *Electronic Journal of Statistics*, **8**:1113–1151.

Nott, D. J., Tran, M.-N., Kuk, A. Y., and Kohn, R. (2013). Efficient variational inference for generalized linear mixed models with large datasets. *Unpublished*.

Opper, M. and Archambeau, C. (2009). The variational Gaussian approximation revisited. *Neural Computation*, **21**:786–792.

Organisation for Economic Co-Operation and Development (2011). *Caesarean sections. In: Health at a glance 2011: OECD Indicators.* OECD Publishing.

Ormerod, J. T. and Wand, M. P. (2010). Explaining variational approximation. *The American Statistician*, **64**: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**:2–17.

Ormerod, J. T., You, C., and Muller, S. (2014). A variational Bayes approach to variable selection. (Unpublished).

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statistical Science*, **1**:502–527.

Pham, T. H., Ormerod, J. T., and Wand, M. P. (2013). Mean field variational Bayesian inference for nonparametric regression with measurement error. *Computational Statistics and Data Analysis*, **68**:375–387.

Pinheiro, J., Bates, D., DebRoy, S., and Sarkar, D. (2014). Mixed `GAM` computation vehicle with `GCV/AIC/REML` smoothness estimation. *R package version 1.8.*

R Development Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabe-Hesketh, S. and Skrondal, A. (2008). *Multilevel and longitudinal modeling using Stata*. Texas: STATA press.

Rao, C. R. (1972). Estimation of variance and covariance components in linear models. *Journal of the American Statistical Association*, **67**:112–115.

Raudenbush, S. W., Yang, M.-L., and Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate laplace approximation. *Journal of computational and Graphical Statistics*, **9**:141–157.

Rhode, D. and Wand, M. P. (2015). Semiparametric mean field variational Bayes: General principles and numerical issues. *Unpublished*.

Rice, J. A. and Wu, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, **57**:253–259.

Richardson, S., Leblond, L., Jaussent, I., and Green, P. J. (2002). Mixture models in measurement error problems, with reference to epidemiological studies. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **165**:549–566.

Robert, C. P. and Casella, G. (2004). *Monte Carlo statistical methods*. New York: Springer Science & Business Media.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*, volume 12. New York: Cambridge university press.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2009). Semiparametric regression during 2003–2007. *Electronic Journal of Statistics*, **3**:1193.

Ryu, D., Li, E., and Mallick, B. K. (2011). Bayesian nonparametric regression analysis of data with random effects covariates from longitudinal measurements. *Biometrics*,

67:454–466.

SAS Institute Inc. (2013). *SAS/STAT 13.1 Users Guide*. North Carolina: SAS Institute Inc.

Smith, A. D. and Wand, M. P. (2008). Streamlined variance calculations for semiparametric mixed models. *Statistics in Medicine*, **27**:435–448.

Stan Development Team (2015). *Stan: A C++ Library for probability and sampling. Version 2.6.*

Stewart, B. (2014). Latent factor regressions for the social sciences. *Unpublished*.

Tan, L. S. and Nott, D. J. (2013). Variational inference for generalized linear mixed models using partially noncentered parametrizations. *Statistical Science*, **28**:168–188.

Tan, L. S. and Nott, D. J. (2014). A stochastic variational framework for fitting and diagnosing generalized linear mixed models. *Bayesian Analysis*, **9**:963–1004.

Teh, Y. W., Newman, D., and Welling, M. (2006). A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. Em *Advances in neural information processing systems*, pgs. 1353–1360.

Teschendorff, A. E., Wang, Y., Barbosa-Morais, N. L., Brenton, J. D., and Caldas, C. (2005). A variational Bayesian mixture modelling framework for cluster analysis of gene-expression data. *Bioinformatics*, **21**:3025–3033.

Titterington, D. (2004). Bayesian methods for neural networks and related models. *Statistical Science*, **23**:128–139.

Verbyla, A. P., Cullis, B. R., Kenward, M. G., and Welham, S. J. (1999). The analysis of designed experiments and longitudinal data by using smoothing splines. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **48**:269–311.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, **1**:1–305.

Wand, M. P. (2014). Fully simplified multivariate normal updates in non-conjugate variational message passing. *The Journal of Machine Learning Research*, **15**:1351–1369.

Wand, M. P. (2015). Fast approximate inference for arbitrarily large semiparametric regression models via message passing. *Unpublished*.

Wand, M. P. and Ormerod, J. T. (2008). On semiparametric regression with O'Sullivan penalized splines. *Australian and New Zealand Journal of Statistics*, **50**:179–198.

Wand, M. P. and Ormerod, J. T. (2011). Penalized wavelets: embedding wavelets into semiparameric regression. *Electronic Journal of Statistics*, **5**:1654–1717.

Wand, M. P., Ormerod, J. T., Padoan, S. A., and Fruhwirth, R. (2012). Mean field variational Bayes for elaborate distributions. *Bayesian Analysis*, **7**:847–900.

Wand, M. P. and Ripley, B. (2006). `KernSmooth`: Functions for kernel smoothing for wand and jones (1995). *R package version*, **2**:19–22.

Wang, B. and Titterington, D. (2006). Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. *Bayesian Analysis*, **1**:625–650.

Wang, C. and Blei, D. M. (2013). Variational inference in nonconjugate models. *The Journal of Machine Learning Research*, **14**:1005–1031.

Wang, C., Paisley, J. W., and Blei, D. M. (2011). Online variational inference for the hierarchical Dirichlet process. Em *International conference on artificial intelligence and statistics*, pgs. 752–760.

Winkelmann, R. (2004). Health care reform and the number of doctor visitsan econometric analysis. *Journal of Applied Econometrics*, **19**:455–472.

Winn, J. M. and Bishop, C. M. (2005). Variational message passing. *Journal of Machine Learning Research*, **6**:661–694.

Wolfinger, R., Tobias, R., and Sall, J. (1994). Computing gaussian likelihoods and their derivatives for general linear mixed models. *SIAM Journal on Scientific Computing*, **14**:1294–1310.

Wood, S. N. (2014). `nlme`: Linear and nonlinear mixed effects models. *R package version 3.1.*

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

Zeger, S. L. and Diggle, P. J. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics*, pgs. 689–699.

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**:79–86.

Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous

outcomes. *Biometrics*, **42**:121–130.

Zhang, D., Lin, X., Raz, J., and Sowers, M. (1998). Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association*, **93**:710–719.

Zhao, Y., Staudenmayer, J., Coull, B. A., and Wand, M. P. (2006). General design Bayesian generalized linear mixed models. *Statistical Science*, **21**:35–51.