Variational Inference for Heteroscedastic and Longitudinal Regression Models

by

Marianne Menictas

Submitted to the School of Mathematical Sciences, Faculty of Science
in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

at the

UNIVERSITY OF TECHNOLOGY SYDNEY

2015

CERTIFICATE OF ORIGINAL AUTHORSHIP

I 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

My deepest thank you goes to my advisor and academic father, Matt. You opened my eyes to a new way of thinking. I could not have asked for a more inspirational, supportive and patient role model to guide me through this journey and will forever be grateful for what you have taught me. I extend my appreciation to my co-supervisor Peter, who spent many hours guiding me with helpful advice throughout my PhD journey. To my honours advisor Debbie, thank you for showing me the importance of rigorous research and for always having your door open to me. Cathy, my PhD buddy and great friend, discussing the intricacies of our work was the highlight of my day so thank you. To Mum and Nick, thank you for your constant unconditional love and support and for always being there for me. Finally, to Dad, my quest would not have taken place if it weren't for you encouraging me to be unreservedly curious and hungry for the pursuit of knowledge. You are the biggest inspiration in my life.

Contents

Abstract

The focus of this thesis is on the development and assessment of mean field variational Bayes (MFVB), which is a fast, deterministic tool for inference in a Bayesian hierarchical model setting. We assess the performance of MFVB via the use of comprehensive comparisons against a Markov chain Monte Carlo (MCMC) benchmark. Each of the models considered are special cases of semiparametric regression. In particular, we focus on the development and assessment of the performance of MFVB for heteroscedastic and longitudinal semiparametric regression models. Generally, the new MFVB methodology performs well in its assessment of accuracy against MCMC for the semiparametric and nonparametric regression models considered in this thesis. It is also much faster and is shown to be applicable to real-time analyses. Several real data illustrations are provided. Altogether, MFVB proves to be a credible inference tool and a good alternative to MCMC, especially when analysis is hindered by time constraints.

Chapter 1

Introduction

"Information is the oil of the 21st century & analytics is the combustion engine"

- Peter Sondergaard, Senior Vice President, Gartner Research.

The rapid growth of information is one of the most challenging issues the world faces today. To appreciate the degree by which an information revolution is under way, consider for instance the Sloan Digital Sky Survey telescope in New Mexico, which in 2000, collected more data in its first few weeks of operation than had been amassed in the entire history of astronomy. An even more salient example is Facebook, having just celebrated its 10th birthday, has more than 10 million photos uploaded every hour and a *like* button clicked almost 3 billion times per day. This recent growth of information has become known as the *big data revolution* and acts as the underlying motivation for this thesis. In some instances, the amount of information has grown so much, that the volume being handled no longer fits into standard computer memory. As a result, engineers need to re-develop tools for analysing it all. This is the drive behind new data processing technologies such as Map Reduce and Hadoop, which allows one to manage larger troves of data than ever before.

In a similar vein, the pace of processing information in recent years has given impetus for the development of faster data analysis techniques to replace conventional methods that, however accurate, take much longer to run. Specifically, this thesis focuses on the development and assessment of mean field variational Bayes (MFVB), which is a fast alternative to Markov chain Monte Carlo (MCMC). These methods are considered in a hierarchical Bayesian model setting, with comprehensive comparisons between the two

being a theme of this work. We commence our discussion by considering a review of the literature surrounding MFVB.

1.1 Literature review

The Bayesian statistical literature is dominated by simulation-based approximations such as MCMC. Since its emergence in Bayesian statistics in the 1990s, and its ensuing construction of the Bayesian inference engines BUGS (Bayesian inference Using Gibbs Sampling) (Spiegelhalter *et al.*, 2003) and Stan (Stan Development Team, 2014), MCMC has made previously intractable inference problems solvable. When challenged with a new and possibly complex model that requires fitting and inference, the Bayesian analyst can choose between at least two approaches: (a) to manually program MCMC using Metropolis-Hastings and/or slice sampling as a way of dealing with arduous full conditional distributions; or (b) to implement the model in the BUGS inference engine by the simple press of a button.

Albeit an extremely accurate inference tool, MCMC's main setback from an applied standpoint is the waiting time required for a program to run, sometimes taking days or even weeks to get results. When considering applications that depend on speedy analysis, such as speech recognition and robot vision, these times are unacceptably slow. A rapid deterministic alternative to MCMC and Monte Carlo methods in general is the variational approximation, which involves approximate inference for parameters in complex statistical models. Let us next discuss variational approximations and their emergence into the computer science and statistical literature.

1.1.1 Variational approximations

Variational approximations originate from the mathematical topic of *variational calculus*, which has been part of the mathematical literature for over two centuries. Variational calculus involves optimising a functional over a specified class of functions, somewhat akin to the way in which a function can depend on a numerical variable. Variational approximations arise when constraints are fixed on the specified class of functions, usually to enforce tractability (e.g. Ormerod & Wand, 2010).

The posterior distribution of the model parameters is of primary interest in Bayesian inference. However, even for relatively simple models, the derivation of the posterior

density can sometimes involve intractable integrals. As mentioned previously, our interest lies in the achievement of faster inference for hierarchical Bayesian models. Specifically, we are interested in providing a deterministic alternative to the stochastic MCMC by combining the notions of Bayesian inference together with the variational approximation, in order to approximate these previously intractable integrals. MFVB arises when the variational approximation being used is subject to a specific factorisation, and is therefore, limited in its approximation accuracy.

Variational approximations have made *eye-catching* appearances in the computer science literature, where they have acted as the underlying methodology behind sophisticated areas of research such as Markov random fields, error-control coding and language processing (Jordan *et al.*, 1999; Jordan, 2004). Comprehensive summaries of the variational approximation may be found in Bishop (2006, Chapter 10) and Ormerod & Wand (2010). What is lacking in the computer science literature, however, is research on the accuracy of the variational approximation. This provides the statistical sciences an opportunity to consolidate the existence of an accuracy assessment where we provide a quantitative assessment of the performance of the variational approximation. Research into the quality of the variational approximation has already manifested in statistical literature for specific models. For example, Wang & Titterington (2003) study the consistency properties of variational Bayesian estimators for mixture models involving known densities. It was shown that with probability 1, as the iterations approach infinity, the algorithm converges locally to the maximum likelihood estimator. In addition, Wang *et al.* (2006) introduce a general algorithm for computing variational Bayesian estimates and look into its convergence properties for a normal mixture model. For a comprehensive listing of other relevant literature on the accuracy of the variational approximation, Jordan (2004) and Titterington (2004) discuss further sources.

The specific focus of this thesis therefore, is to provide an in-depth investigation into a specific type of variational approximation, which we refer to as MFVB. The review of the literature surrounding MFVB follows.

1.1.2 Mean field variational Bayes

Mean field variational Bayes (Wainwright & Jordan, 2008) involves Bayesian estimation in graphical models, specifically catered to the variational posterior distribution restricted

to a factorised form. It is one of the most common types of variational approximation used today (Parisi, 1988). The use of MFVB has become more prevalent in the statistical literature in the past decade. For example, Teschendorff *et al.* (2005) use MFVB as a means of accurately categorising tumour types through gene-expression profiling. McGrory & Titterington (2007) applied MFVB to the Bayesian analysis of mixtures of Gaussian distributions. Further, in 2008, a MFVB-based software package named Infer.NET (Minka *et al.*, 2008) was developed. This allowed one to use MFVB on a wide range of statistical problems, making MFVB more accessible than ever before.

In more recent years, MFVB has appeared in the *Journal of the American Statistical Association* (Braun & McAuliffe, 2010; Faes *et al.*, 2011), exposing MFVB to a wider audience and demonstrating its potential application. A new major area of application based on MFVB is real-time semiparametric regression analysis (Luts *et al.*, 2013). Other recent MFVB examples are given in, but not limited to, Hall *et al.* (2011), Huang *et al.* (2013), Wand (2014), McGrory & Titterington (2007), Wang *et al.* (2006) and Neville *et al.* (2014).

1.2 Contribution

Even though we have seen MFVB appear in many different statistical settings, there is still a wide range of research needed to be carried out. This thesis aims to contribute to the *missing parts* of research and application dedicated to MFVB. In particular, our contribution lies in our ability to expose MFVB, yet again, as a faster, more prominent alternative to more widely used inference methods such as the Laplace approximation and MCMC. In addition, we aim to contribute by quantitatively assessing the accuracy of MFVB in various statistical settings.

1.3 Notational guide

There are two main notations used throughout this thesis, namely vector and matrix notation, and probability notation. These are discussed in the following two sections.

1.3.1 Vector and matrix notation

Lower case bold fonts are used to denote vectors and upper case bold fonts denote matrices. The real line is denoted by $\mathbb{R}$ and m-dimensional space as $\mathbb{R}^m$. The $m \times 1$ column vector with all entries equal to 1 is denoted by $\mathbf{1}_m$. The $m \times m$ identity matrix is denoted by $\boldsymbol{I}_m$. For vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^m$, denoted by $\boldsymbol{a} = [a_1, \ldots, a_m]^\top$ and $\boldsymbol{b} = [b_1, \ldots, b_m]^\top$ we define the following notation.

The element-wise multiplication of two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ is denoted by

$$\boldsymbol{a} \odot \boldsymbol{b} = \begin{bmatrix} a_1 b_1 \\ \vdots \\ a_m b_m \end{bmatrix}.$$

The norm of a vector $\boldsymbol{a}$ is defined to be $\| \boldsymbol{a} \| = \sqrt{\boldsymbol{a}^\top \boldsymbol{a}}$. We use $\boldsymbol{a}_{-i}$ to denote the vector $\boldsymbol{a}$ with the ith element removed. A scalar valued function of a vector is evaluated element-wise. For example,

$$\log(\boldsymbol{a}) = \begin{bmatrix} \log(a_1) \\ \vdots \\ \log(a_m) \end{bmatrix}.$$

We let $\mathrm{diag}(\boldsymbol{a})$ denote the $m \times m$ diagonal matrix formed by assigning the elements of $\boldsymbol{a}$ to the main diagonal. For instance,

$$\mathrm{diag}(\boldsymbol{a}) = \begin{bmatrix} a_1 & 0 & \ldots & 0 \\ 0 & a_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & a_m \end{bmatrix}.$$

For any $m \times m$ matrix $\boldsymbol{A}$, $\mathrm{diagonal}(\boldsymbol{A})$ refers to the $m \times 1$ vector consisting of the diagonal entries of $\boldsymbol{A}$. For square matrices $(\boldsymbol{A}_1, \ldots, \boldsymbol{A}_d)$, we define $\mathrm{blockdiag}(\boldsymbol{A}_1, \ldots, \boldsymbol{A}_d)$ to be the block diagonal matrix, with the ith block equal to $\boldsymbol{A}_i$. For a symmetric $m \times m$ matrix $\boldsymbol{A}$, the determinant and trace are denoted by $|\boldsymbol{A}|$ and $\mathrm{tr}(\boldsymbol{A})$ respectively, and are defined in the usual way.

For the matrices $\boldsymbol{A}(m \times n)$ and $\boldsymbol{B}(p \times q)$, defined as

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \ldots & a_{mn} \end{bmatrix} \quad \text{and} \quad \boldsymbol{B} = \begin{bmatrix} b_{11} & \vdots & b_{1q} \\ \vdots & \ddots & \vdots \\ b_{p1} & \ldots & b_{pq} \end{bmatrix},$$

the following notation is provided.

The transpose of $\boldsymbol{A}$ is denoted by $\boldsymbol{A}^\top$. We let $\boldsymbol{A}^{-1}$ denote the matrix inverse, given that $\boldsymbol{A}$ is an invertible matrix. The Kronecker product of two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ is denoted by

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & \dots & a_{1n}\boldsymbol{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\boldsymbol{B} & \dots & a_{mn}\boldsymbol{B} \end{bmatrix}.$$

We define $\text{vec}(\boldsymbol{A})$ to be the $m^2 \times 1$ vector obtained by stacking the columns of $\boldsymbol{A}$ underneath each other from left to right. If $\boldsymbol{a}$ is a $m^2 \times 1$ vector then $\text{vec}^{-1}(\boldsymbol{a})$ is defined to be the $m \times m$ matrix obtained by listing the entries of $\boldsymbol{a}$ in a column-wise fashion from left to right. Also, vec^{-1} denotes the usual function inverse since the domain of vec is restricted to square matrices.

1.3.2 Probability notation

If x and y represent scalar random variables, the mean and variance of x are represented as $\text{E}(x)$ and $\text{Var}(x)$. The covariance between x and y is represented as $\text{Cov}(x, y)$.

If $\boldsymbol{\theta}$ is a random vector, its density function is denoted by $p(\boldsymbol{\theta})$. If $\boldsymbol{\phi}$ is also a random vector, the conditional density function of $\boldsymbol{\theta}$ given $\boldsymbol{\phi}$ is denoted by $p(\boldsymbol{\theta}|\boldsymbol{\phi})$. In a Bayesian model setting, the full conditional distribution of $\boldsymbol{\theta}$ is given by $p(\boldsymbol{\theta}|\text{rest})$ and represents the conditional distribution of $\boldsymbol{\theta}$ given the rest of the random variables in the model. The mean and covariance of a random vector $\boldsymbol{\theta}$ is given by $\text{E}(\boldsymbol{\theta})$ and $\text{Cov}(\boldsymbol{\theta})$.

When using MFVB, the approximate densities are denoted using the letter q. We call these densities q-densities. The mean and variance of a scalar random variable θ under the MFVB criterion is given by:

$$E_q(\theta) = \mu_{q(\theta)} \quad \text{and} \quad \text{Var}_q(\theta) = \sigma^2_{q(\theta)}.$$

Similarly, the mean and covariance of a random vector $\boldsymbol{\theta}$ under the MFVB criterion is given by:

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

Finally, if $x_1, \dots, x_n$ is a sequence of independent random variables following distribution D, we denote that as $x_i \overset{\text{ind.}}{\sim} D$, for $1 \leq i \leq n$.

1.4 Theorems, definitions and results

1.4.1 Vector differential calculus

Suppose $f(\boldsymbol{x})$ is a function with argument $\boldsymbol{x} \in \mathbb{R}^p$ that returns a scalar. The *derivative vector* of $f(\boldsymbol{x})$ is the $1 \times p$ vector $\mathsf{D}_{\boldsymbol{x}} f$ with ith entry equal to $\partial f(\boldsymbol{x})/\partial x_i$ where x_i is the ith entry of $\boldsymbol{x}$. $\mathsf{D}_{\boldsymbol{x}} f$ is found using the *First Identification Theorem* (Magnus & Neudecker, 1999):

Theorem 1.4.1. *If $\boldsymbol{a}$ is a $1 \times p$ vector where $d f(\boldsymbol{x}) = \boldsymbol{a}\, d\boldsymbol{x}$, then the derivative vector is*

$$\boldsymbol{a} = D_{\boldsymbol{x}} f.$$

Differentiation again with respect to $\boldsymbol{x}$ gives the $p \times p$ *Hessian matrix* $\mathsf{H}_{\boldsymbol{x}} f$ with (i, j)th entry equal to $\partial^2 f(\boldsymbol{x})/\partial x_i \partial x_j$ where x_i and x_j are the ith and jth entry of $\boldsymbol{x}$. $\mathsf{H}_{\boldsymbol{x}} f$ is found using the *Second Identification Theorem* (Magnus & Neudecker, 1999):

Theorem 1.4.2. *If $\boldsymbol{A}$ is a $p \times p$ matrix where $d^2 f(\boldsymbol{x}) = (d\boldsymbol{x})^\top \boldsymbol{A}\, d\boldsymbol{x}$, then the Hessian matrix is*

$$\boldsymbol{A} = H_{\boldsymbol{x}} f.$$

The following results are useful in vector differential calculus.

Result 1.4.1. *If $\boldsymbol{U}$ and $\boldsymbol{V}$ are matrix functions, then*

$$\begin{aligned}
&(a) \quad vec(d\boldsymbol{U}) = d\,vec(\boldsymbol{U}) \\
&(b) \quad d|\boldsymbol{U}| = |\boldsymbol{U}| tr(\boldsymbol{U}^{-1} d\boldsymbol{U}) \\
&(c) \quad d\boldsymbol{U}^{-1} = -\boldsymbol{U}^{-1}(d\boldsymbol{U})\boldsymbol{U}^{-1} \\
&(d) \quad d(tr\boldsymbol{U}) = tr(d\,\boldsymbol{U}) \\
&(e) \quad d(\boldsymbol{U} \otimes \boldsymbol{V}) = d\,\boldsymbol{U} \otimes \boldsymbol{V} + \boldsymbol{U} \otimes d\,\boldsymbol{V} \\
&(f) \quad d\,(\log|\boldsymbol{U}|) = tr\left(\boldsymbol{U}^{-1}\, d\,\boldsymbol{U}\right) \\
&(g) \quad d\,(\boldsymbol{U}\boldsymbol{V}) = (d\boldsymbol{U})\,\boldsymbol{V} + \boldsymbol{U}\,(d\boldsymbol{V})
\end{aligned}$$

1.4.2 Special functions

Definition 1.4.1. *The digamma function is the logarithmic derivative of the gamma function and is defined by*

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

1.4.3 Probability distributions

Here we summarize the main properties of the distributions used throughout this thesis.

1.4.3.1 Uniform distribution

Definition 1.4.2. *A scalar random variable x has a uniform distribution with minimum value a and maximum value b, written $x \sim uniform(a, b)$ if its density function is*

$$p(x) = \begin{cases} \frac{1}{b-a} & , \ a \le x \le b \\ 0 & , \ otherwise. \end{cases}$$

1.4.3.2 Bernoulli distribution

Definition 1.4.3. *A scalar random variable x has a Bernoulli distribution, written $x \sim Bernoulli(p)$, where $0 \le p \le 1$, if its density function is defined to be*

$$p(x) = \begin{cases} p^x(1-p)^{1-x} & , \ x = 0, 1 \\ 0 & , \ otherwise. \end{cases}$$

1.4.3.3 Beta distribution

Definition 1.4.4. *A scalar random variable x has a Beta distribution with shape parameters $\alpha, \beta > 0$, written $x \sim Beta(\alpha, \beta)$, if its density function is defined to be*

$$p(x) = \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1} (1-x)^{\beta-1} & , \ 0 \le x \le 1 \\ 0 & , \ otherwise. \end{cases}$$

The normalising constant $B(\alpha, \beta)$ is known as the Beta function and is expressed as $\Gamma(\alpha+\beta)/\Gamma(\alpha)\Gamma(\beta)$.

1.4.3.4 Normal distribution

The Normal distribution, also known as the Gaussian distribution, is the most widely used distribution for continuous variables.

Definition 1.4.5. *A scalar random variable x has a Normal distribution with mean μ and variance $\sigma^2 > 0$, written $x \sim N(\mu, \sigma^2)$ if its density function is*

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

Definition 1.4.6. *The notation $\phi(\cdot)$ is used to denote the standard normal density function, that is $x \sim N(0,1)$:*

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}.$$

Thus, for general mean μ and standard deviation σ, that is if $x \sim \mathrm{N}(\mu, \sigma^2)$, then

$$p(x) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right).$$

Definition 1.4.7. *An $n \times 1$ vector of random variables $\boldsymbol{x}$ has a Multivariate normal distribution with $n \times 1$ mean vector $\boldsymbol{\mu}$ and $n \times n$ positive-definite covariance matrix $\boldsymbol{\Sigma}$, written $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ if its density function is*

$$p(\boldsymbol{x}) = (2\pi)^{-n/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{ \tfrac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right\}.$$

1.4.3.5 Gamma distribution

Definition 1.4.8. *A scalar random variable x has a Gamma distribution with shape and rate parameters $A, B > 0$, written $x \sim Gamma(A, B)$, if its density function is*

$$p(x) = \begin{cases} \frac{B^A}{\Gamma(A)} x^{A-1} \exp\{-Bx\} & , \ x > 0 \\ 0 & , \ otherwise. \end{cases}$$

Result 1.4.2. *If $x \sim Gamma(A, B)$, then for any $c > 0$,*

$$cx \sim Gamma(A, B/c).$$

1.4.3.6 Inverse-Gamma distribution

Perhaps the main use of the Inverse-Gamma distribution is in Bayesian statistics, where it is used as an analytically tractable prior distribution for the variance parameters.

Definition 1.4.9. *A scalar random variable x has an Inverse Gamma distribution with shape and rate parameters $A, B > 0$, written $x \sim Inverse\text{-}Gamma(A, B)$ if its density function is*

$$p(x) = \begin{cases} \frac{B^A}{\Gamma(A)} x^{-A-1} \exp\left\{-\frac{B}{x}\right\} & , \ x > 0 \\ 0 & , \ otherwise. \end{cases}$$

Result 1.4.3. *If* $x \sim$ *Inverse-Gamma*(A, B)*, then*

$$E\,(1/x) = A/B, \quad E\,\{\log(x)\} = \log(B) - \psi(A),$$

where ψ *is the digamma function as defined in Definition 1.4.1*

Result 1.4.4. *If* $x \sim$ *Gamma*(A, B)*, then*

$$1/x \sim \textit{Inverse-Gamma}(A, B).$$

Result 1.4.5. *If* $x \sim$ *Gamma*$\left(\frac{\nu}{2}, \frac{1}{2}\right)$*, then* x *is identical to the chi-squared distribution with* ν *degrees of freedom* χ^2_ν*.*

1.4.3.7 Half-Cauchy distribution

Often, alternative scale parameter priors are desired. Gelman (2006) argues that Half-Cauchy densities are better for achieving non-informativeness of scale parameters. The Half-Cauchy distribution is a special case of the conditionally-conjugate folded-noncentral t family of prior distributions for the standard deviation parameter and tends to give cleaner inferences.

Definition 1.4.10. *A scalar random variable* x *has a Half-Cauchy distribution with scale parameter* $A > 0$*, written* $x \sim$ *Half-Cauchy*(A)*, if its density function is*

$$p(x) = \begin{cases} \frac{2}{\pi A\left\{1+(x/A)^2\right\}} & , \ x > 0 \\ 0 & , \ \textit{otherwise.} \end{cases}$$

The following result is Result 5 of Wand, Ormerod, Padoan and Frühwirth (2011):

Result 1.4.6. *Suppose that x and a are random variables such that*

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

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

1.4.3.8 Wishart distribution

Definition 1.4.11. *An $n \times n$ matrix $\boldsymbol{X}$ has a Wishart distribution with degrees of freedom $a > 0$ and positive definite rate matrix $\boldsymbol{B}$, written $\boldsymbol{X} \sim$ Wishart$(a, \boldsymbol{B})$, if its density function is*

$$p(\boldsymbol{X}) = \begin{cases} C_{n,a}^{-1} \, |\boldsymbol{B}|^{a/2} \, |\boldsymbol{X}|^{(a-n-1)/2} \exp\{-\tfrac{1}{2}\mathrm{tr}\,(\boldsymbol{B}\boldsymbol{X})\} & , \ \boldsymbol{X} \textit{ symmetric and positive definite} \\ 0 & , \ \textit{otherwise,} \end{cases}$$

where $C_{n,a} \equiv 2^{an/2}\pi^{n(n-1)/4} \prod_{i=1}^{n} \Gamma\left((a+1-i/2)\right)$.

Result 1.4.7. *If $\boldsymbol{X} \sim$ Wishart$(a, \boldsymbol{B})$, then*

$$E(\boldsymbol{X}) = a\boldsymbol{B}^{-1}.$$

1.4.3.9 Inverse-Wishart distribution

The Inverse-Wishart distribution is commonly used as the conjugate prior for the covariance matrix of a multivariate normal distribution.

Definition 1.4.12. *An $n \times n$ matrix $\boldsymbol{X}$ has an Inverse-Wishart distribution with degrees of freedom $a > 0$ and positive definite scale matrix $\boldsymbol{B}$, written $\boldsymbol{X} \sim$ Inverse-Wishart$(a, \boldsymbol{B})$, if its density function is*

$$p(\boldsymbol{X}) = \begin{cases} C_{n,a}^{-1} \, |\boldsymbol{B}|^{a/2} \, |\boldsymbol{X}|^{-(a+n+1)/2} \exp\{-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{B}\boldsymbol{X}^{-1}\right)\} & , \ \boldsymbol{X} \text{ symmetric and positive definite} \\ 0 & , \ \text{otherwise,} \end{cases}$$

where $C_{n,a} \equiv 2^{an/2} \pi^{n(n-1)/4} \prod_{i=1}^{n} \Gamma\left((a + 1 - i/2)\right)$.

Result 1.4.8. *If $\boldsymbol{X} \sim$ Wishart$(a, \boldsymbol{B})$, then*

$$\boldsymbol{X}^{-1} \sim \text{Inverse-Wishart}(a, \boldsymbol{B}).$$

1.4.3.10 Multinomial distribution

Definition 1.4.13. *A vector of random variables $\boldsymbol{x} = (x_1, \ldots, x_k)$ where x_i takes on values $0, \ldots, n$, has a Multinomial distribution with number of independent trials $n > 0$ and probabilities of success $\boldsymbol{p} = (p_1, \ldots, p_k)$, written $x_1, \ldots, x_k \sim$ Multinomial$(n, p_1, \ldots, p_k)$, if its density function is*

$$p(x_1, \ldots, x_k) = \begin{cases} \frac{n!}{x_1! \ldots x_k!} {p_1}^{x_1} \ldots {p_k}^{x_k}, & \text{when} \quad \sum_{i=1}^{k} x_i = n \\ 0, & \text{otherwise} \end{cases}.$$

Result 1.4.9. *Suppose that $x_1, \ldots, x_k \sim$ Multinomial$(n, p_1, \ldots, p_k)$. Then,*

$$E(x_i) = np_i.$$

1.4.3.11 Dirichlet distribution

Definition 1.4.14. *A vector of random variables* $\boldsymbol{x} = (x_1, \ldots, x_k)$ *has a Dirichlet distribution with parameters* $\alpha_1, \ldots, \alpha_k > 0$ *written* $p(x_1, \ldots, x_k) \sim Dirichlet(\alpha_1, \ldots, \alpha_k)$ *if its density function is*

$$p(x_1, \ldots, x_k) = \frac{1}{B(\alpha_1, \ldots, \alpha_k)} \prod_{i=1}^{k} {x_i}^{\alpha_i - 1}$$

for all $x_i, \ldots, x_k > 0$ *satisfying* $\sum_{i=1}^{k} x_i = 1$. *The normalizing constant is known as the multinomial Beta function, which can be expressed in terms of the Gamma function as:*

$$B(\alpha_1, \ldots, \alpha_k) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}.$$

Result 1.4.10. *Suppose that* $x_1, \ldots, x_k \sim Dirichlet(\alpha_1, \ldots, \alpha_k)$. *Then,*

$$E(x_i) = \alpha_i \Big/ \sum_{k=1}^{K} \alpha_k, \quad E\left\{\log(x_i)\right\} = \psi(\alpha_i) - \psi\left(\sum_{k=1}^{K} \alpha_k\right).$$

1.4.3.12 Normal mixture distribution

Definition 1.4.15. *A scalar random variable* x *has a normal mixture distribution with* K *weighted sum components, written* $p(x; \boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \sim Normal\text{-}Mixture(\boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$, *if its density function is*

$$p(x; \boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \sum_{k=1}^{K} \omega_k \left(2\pi\sigma_k^2\right)^{-1/2} \exp\left\{\tfrac{1}{2}(x - \mu_k)^2 / \sigma_k^2\right\},$$

where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_K)$ *are the mixture weights,* $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$ *are the mixture means and* $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_K^2)$ *are the mixture variances associated with each* k*th component density.*

1.4.4 Useful vector and matrix results

Result 1.4.11. *If* $\boldsymbol{a}$ *and* $\boldsymbol{b}$ *are vectors of equal length then* $diag(\boldsymbol{a})\boldsymbol{b} = \boldsymbol{a} \odot \boldsymbol{b}$.

Result 1.4.12. *If* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *are matrices such that the product* $\boldsymbol{AB}$ *can be formed, then*

$$\mathrm{tr}(\boldsymbol{AB}) = \mathrm{tr}(\boldsymbol{BA}).$$

Result 1.4.13. *If* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *are* $m \times n$ *matrices, then* $\mathrm{tr}(\boldsymbol{A}^\top \boldsymbol{B}) = \mathrm{vec}(\boldsymbol{A})^\top \mathrm{vec}(\boldsymbol{B})$.

Result 1.4.14. *If* $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{D}$ *are matrices such that the products* $\boldsymbol{AC}$ *and* $\boldsymbol{BD}$ *can be formed, then*

$$(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}) = (\boldsymbol{AC}) \otimes (\boldsymbol{BD}).$$

Result 1.4.15. *If* $\boldsymbol{A}$ *is an* $m \times m$ *matrix and* $\boldsymbol{B}$ *is an* $n \times n$ *matrix, then*

$$\mathrm{tr}(\boldsymbol{A} \otimes \boldsymbol{B}) = \mathrm{tr}(\boldsymbol{A})\mathrm{tr}(\boldsymbol{B}).$$

Result 1.4.16. *If* $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{D}$ *are matrices such that* $\boldsymbol{ABCD}$ *is a square matrix, then*

$$\mathrm{tr}(\boldsymbol{ABCD}) = \mathrm{vec}(\boldsymbol{D})^\top (\boldsymbol{A} \otimes \boldsymbol{C}^\top) \mathrm{vec}(\boldsymbol{B}^\top).$$

Result 1.4.17. *Let* $\boldsymbol{x}$ *be a random vector. Then*

$$E(\| \boldsymbol{x} \|^2) = \| E(\boldsymbol{x}) \|^2 + \mathrm{tr}\{Cov(\boldsymbol{x})\} \quad \text{and} \quad E(\boldsymbol{x}\boldsymbol{x}^\top) = E(\boldsymbol{x})E(\boldsymbol{x})^\top + Cov(\boldsymbol{x}).$$

Result 1.4.18. *Let $\boldsymbol{x}$ be a random vector and $\boldsymbol{A}$ be a random matrix. Then*

$$E(\boldsymbol{u}^\top \boldsymbol{A}\boldsymbol{u}) = \{E(\boldsymbol{u})\}^\top \boldsymbol{A}\{E(\boldsymbol{u})\} + \mathrm{tr}\{Cov(\boldsymbol{u})\boldsymbol{A}\}.$$

Result 1.4.19. *Let $\boldsymbol{a}$ be constant vector, let $\boldsymbol{b}$ be a random vector, and let $\boldsymbol{C}$ be a constant matrix with the same number of rows as b. Then*

$$E(\| \boldsymbol{a} - \boldsymbol{C}\boldsymbol{b} \|^2) = \| \boldsymbol{a} - \boldsymbol{C}E(\boldsymbol{b}) \|^2 + \mathrm{tr}\left\{\boldsymbol{C}Cov(\boldsymbol{b})\boldsymbol{C}^\top\right\}.$$

Result 1.4.20. *Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be $n \times 1$ random vectors such that $\boldsymbol{x}$ is conditioned on $\boldsymbol{y}$. Then,*

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

1.5 Mean field variational Bayes

Let us consider a general Bayesian model with parameter vector $\boldsymbol{\theta}$ and observed data vector $\boldsymbol{y}$. As discussed earlier, even for simple models, it is often the case that the posterior density function $p(\boldsymbol{\theta}|\boldsymbol{y})$ is intractable. To overcome this problem, we employ MFVB where the objective is to find a tractable distribution $q(\boldsymbol{\theta})$ that closely approximates the true posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$. For example, we wish to replace the joint posterior density function $p(\boldsymbol{\theta}|\boldsymbol{y})$ with an approximating density function $q(\boldsymbol{\theta})$ that assumes the product density form

$$q(\boldsymbol{\theta}) = \prod_{i=1}^{k} q_i(\boldsymbol{\theta}_i) \tag{1.1}$$

for some partition $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k\}$ of $\boldsymbol{\theta}$. Tractability is a major factor on the choice of partition of the approximate density function $q(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$. The so-called q-densities are then chosen to minimise the Kullback-Leibler distance between the two density functions:

$$\int \log\left\{\frac{p(\boldsymbol{\theta}|\boldsymbol{y})}{q(\boldsymbol{\theta})}\right\} q(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

subject to (1.1). Note that the integrals are replaced with sums in the case of discrete variables, however in this thesis we assume $\boldsymbol{\theta}$ to be continuous over the parameter space $\boldsymbol{\Theta}$. An equivalent optimization problem, attained through standard manipulations, is to maximise:

$$\underline{p}(\boldsymbol{y}; q) \equiv \exp \int q_1(\boldsymbol{\theta}_1) \dots q_k(\boldsymbol{\theta}_k) \log \left\{ \frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{q_1(\boldsymbol{\theta}_1) \dots q_k(\boldsymbol{\theta}_k)} \right\} d\boldsymbol{\theta}_1 \dots d\boldsymbol{\theta}_k \tag{1.2}$$

where $\underline{p}(\boldsymbol{y}; q)$ represents the lower bound on the marginal likelihood $p(\boldsymbol{y})$ for all q-densities. Taking the logarithm of (1.2) gives

$$\log \underline{p}(\boldsymbol{y}; q) \quad \propto \quad \int q_1(\boldsymbol{\theta}_1) \log \tilde{p}(\boldsymbol{y}, \boldsymbol{\theta}_1) d\boldsymbol{\theta}_1 - \int q_1(\boldsymbol{\theta}_1) \log q_1(\boldsymbol{\theta}_1) d\boldsymbol{\theta}_1 \tag{1.3}$$

where

$$\begin{aligned} \log \tilde{p}(\boldsymbol{y}, \boldsymbol{\theta}_1) &= \int \log p(\boldsymbol{y}, \boldsymbol{\theta}) q_2(\boldsymbol{\theta}_2) \times \dots \times q_k(\boldsymbol{\theta}_k) d\boldsymbol{\theta}_2 \dots d\boldsymbol{\theta}_k \\ &= E_{-\boldsymbol{\theta}_1} \left\{ \log p(\boldsymbol{y}, \boldsymbol{\theta}) \right\}, \end{aligned}$$

where $E_{-\boldsymbol{\theta}_1}$ is the expectation with respect to all factors except $q_1(\boldsymbol{\theta}_1)$. By recognising that (1.3) is the negative Kullback Leibler divergence between $q_1(\boldsymbol{\theta}_1)$ and $\tilde{p}(\boldsymbol{y}, \boldsymbol{\theta}_1)$, we note that the minimum occurs when $q_1(\boldsymbol{\theta}_1) = \tilde{p}(\boldsymbol{y}, \boldsymbol{\theta}_1)$. Thus we obtain a general expression for the optimal solution $q_1^*(\boldsymbol{\theta}_1)$ given by

$$q_1^*(\boldsymbol{\theta}_1) \propto \exp \left\{ E_{-\boldsymbol{\theta}_1} \log p(\boldsymbol{y}, \boldsymbol{\theta}) \right\}.$$

This can be generalized to an expression for the optimal q-density for each of the i factors:

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

It is easy to show that an alternative expression for $q_i^*(\boldsymbol{\theta}_i)$ is

$$q_i^*(\boldsymbol{\theta}_i) \propto \exp \left\{ E_{-\boldsymbol{\theta}_i} \log p(\boldsymbol{\theta}_i | \text{rest}) \right\}, \quad 1 \leq i \leq k, \tag{1.4}$$

where 'rest' is the set comprising the random variables in the model, except for $\boldsymbol{\theta}_i$. An iterative coordinate ascent scheme induced by (1.4) is then used to find the optimal parameters in these q-densities. If conjugate priors have been used, then the optimal q-densities belong to known density families (e.g., Winn & Bishop, 2005).

Assuming that each iteration results in an increase in $\log \underline{p}(\boldsymbol{y}; q)$, and the search restricted to a compact set, then convergence is guaranteed to a local maximiser of $\log \underline{p}(\boldsymbol{y}; q)$

(Luenberger & Ye, 2008, p. 253). Convergence can be monitored using successive values of $\log \underline{p}(\boldsymbol{y}; q)$.

1.6 Variational message passing

Variational message passing (VMP) is an alternative to MFVB which results in exactly the same answers, but instead works with "messages" that are passed between neighbouring nodes on the factor graph of the model. Factor graphs will be discussed in detail in Chapter 7. VMP is primarily concerned with conjugate exponential family distributions, represented by their natural parameter versions of the distribution. An in-depth explanation as well as a discussion on the advantages and disadvantages of VMP is given in Chapter 7.

1.7 Non-conjugate variational message passing

Non-conjugate variational message passing (NCVMP) is an extension of VMP which obviates the restriction of only being able to work with conjugate exponential family distributions, whilst maintaining the purely algebraic form of the optimal parameters (Knowles & Minka, 2011). The modularity of NCVMP allows modification only to those model parameters that involve previously intractable solutions, whilst keeping the VMP or MFVB solutions of the other parameters in the model. As a result, NCVMP broadens the class of tractable models for approximate inference using VMP and/or MFVB.

Suppose we approximate the joint posterior density function $p(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{y})$ by the product density form

$$q_1(\boldsymbol{\theta}_1) \ldots q_k(\boldsymbol{\theta}_k)\, q_{\boldsymbol{\phi}}(\boldsymbol{\phi}), \tag{1.5}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k\}$. From Section 1.5 we see that the solutions satisfy

$$\begin{aligned} q_i^*(\boldsymbol{\theta}_i) &\propto \exp\left\{E_{q(-\boldsymbol{\theta}_i)} \log p(\boldsymbol{\theta}_i|\boldsymbol{y}, \boldsymbol{\theta}\backslash\boldsymbol{\theta}_i, \boldsymbol{\phi})\right\}, \quad 1 \le i \le k, \\ q_{\boldsymbol{\phi}}^*(\boldsymbol{\phi}) &\propto \exp\left\{E_{q(-\boldsymbol{\phi})} \log p(\boldsymbol{\phi}|\boldsymbol{y}, \boldsymbol{\theta})\right\}, \end{aligned}$$

where $\boldsymbol{\theta}\backslash\boldsymbol{\theta}_i$ represents the model parameters $\boldsymbol{\theta}$ without $\boldsymbol{\theta}_i$. When $E_{q(-\boldsymbol{\phi})} \log p(\boldsymbol{\phi}|\boldsymbol{y}, \boldsymbol{\theta})$ does not offer a tractable solution, NCVMP can be used by replacing (1.5) with

$$q_1(\boldsymbol{\theta}_1) \ldots q_k(\boldsymbol{\theta}_k)\, q_{\boldsymbol{\phi}}(\boldsymbol{\phi}; \boldsymbol{\eta}),$$

where $q_{\boldsymbol{\phi}}(\boldsymbol{\phi};\boldsymbol{\eta})$ is forced to have an exponential family density function with natural parameter vector $\boldsymbol{\eta}$ and natural statistic $\boldsymbol{T}(\boldsymbol{\phi})$. Then, using Theorem 1 of Knowles & Minka (2011) to find an update for $\boldsymbol{\eta}$ and regular MFVB for the other parameters, we get

$$\begin{aligned} q_i^*(\boldsymbol{\theta}_i) &\propto \exp\left\{E_{q(-\boldsymbol{\theta}_i)} \log p(\boldsymbol{\theta}_i|\boldsymbol{y},\boldsymbol{\theta}\backslash\boldsymbol{\theta}_i,\boldsymbol{\phi})\right\}, \quad 1 \le i \le k, \\ \boldsymbol{\eta} &\leftarrow \left[\text{var}\left\{\boldsymbol{T}(\boldsymbol{\phi})\right\}\right]^{-1}\left[\mathsf{D}_{\boldsymbol{\eta}} E_{q(\boldsymbol{\theta},\boldsymbol{\phi})}\left\{\log p(\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\phi})\right\}\right]^{\top}. \end{aligned} \tag{1.6}$$

This thesis only considers the case where $q(\boldsymbol{\phi};\boldsymbol{\eta})$ corresponds to a d-dimensional Multivariate Normal density function. For this special case, Wand (2014) explains that the number of entries in $\text{var}\left\{\boldsymbol{T}(\boldsymbol{\phi})\right\}$ is quartic in d and thus the $\boldsymbol{\eta}$ update in (1.6) becomes numerically challenging. To get around this, Wand (2014) proves that

$$\boldsymbol{\eta} \leftarrow \left[\text{var}\left\{\boldsymbol{T}(\boldsymbol{\phi})\right\}\right]^{-1}\left[\mathsf{D}_{\boldsymbol{\eta}} E_{q(\boldsymbol{\theta},\boldsymbol{\phi})}\left\{\log p(\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\phi})\right\}\right]^{\top}$$

is mathematically equivalent to the following updates

$$\begin{aligned} \boldsymbol{v}_{q(\phi)} &\leftarrow \left(\mathsf{D}_{\boldsymbol{\mu}_{q(\boldsymbol{\phi})}} S\right)^{\top} \\ \boldsymbol{\Sigma}_{q(\phi)} &\leftarrow \left\{-2\text{vec}^{-1}\left((\mathsf{D}_{\text{vec}(\boldsymbol{\Sigma}_{q(\boldsymbol{\phi})})} S)^{\top}\right)\right\}^{-1} \\ \boldsymbol{\mu}_{q(\phi)} &\leftarrow \boldsymbol{\mu}_{q(\boldsymbol{\phi})} + \boldsymbol{\Sigma}_{q(\boldsymbol{\phi})}\boldsymbol{v}_{q(\phi)}, \end{aligned}$$

where

$$S = E_{q(\boldsymbol{\theta},\boldsymbol{\phi})}\left\{\log p(\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\phi})\right\}.$$

vec and vec^{-1} are defined in Section 1.3.1 and D represents the derivative vector which is defined in Section 1.4.1. However, to get around using the vec operator, a result given in the appendix of Opper & Archambeau (2009) shows that an equivalent representation of the second of these updates is

$$\boldsymbol{\Sigma}_{q(\phi)} \leftarrow \left(-\mathsf{H}_{\boldsymbol{\mu}_{q(\boldsymbol{\phi})}} S\right)^{-1}$$

where H represents the Hessian matrix and is defined in Section 1.4.1. The proof of this result is given in Appendix 3.C of this thesis. We work with this alternative form when using NCVMP.

1.8 Graphical models

One of the engaging aspects of graphical models is that a particular graph has the ability to make probabilistic statements for an extensive class of distributions. In particular, we focus

on the use of graph representations of hierarchical Bayesian models, as they are an effective tool in understanding the conditional dependence structure of these models. Graphical models aid in the calculation of the imposed product factorisation on the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ which is involved in the initial stages of MFVB. As a result, an additional partition of the dependence structure in the model, known as an induced factorization (Bishop, 2006), is then possible. Both of these factorizations are subject to the type of imposed factorization and the underlying structure of the model. Graphical models are a very useful tool in aiding our understanding of these factorizations. These are defined and outlined in more detail in the following section.

1.8.1 Directed acyclic graphs

A graph comprises a set of nodes together with a set of edges, where edges are placed between pairs of nodes. In this thesis, nodes are depicted as circles and edges are depicted as line segments. Also, we use the notation $a \perp\!\!\!\perp b|c$ to denote that a is conditionally independent to b given c, in a graphical model setting.

Definition 1.8.1. *A graph is an* ***undirected graph*** *if it comprises a set of nodes connected by undirected edges.*

Definition 1.8.2. *A graph is a* ***directed graph*** *if it comprises a set of nodes connected by directed edges.*

Definition 1.8.3. *In a directed graph, a path between a set of nodes is a* ***cycle*** *if all directed edges in the path meet the nodes head-to-tail throughout the path.*

Figure 1.1 illustrates examples of an undirected graph, directed graph and a cycle.

Definition 1.8.4. *A directed graph which has no cycles is* ***acyclic****. This is referred to as a* ***directed acyclic graph (DAG)****.*

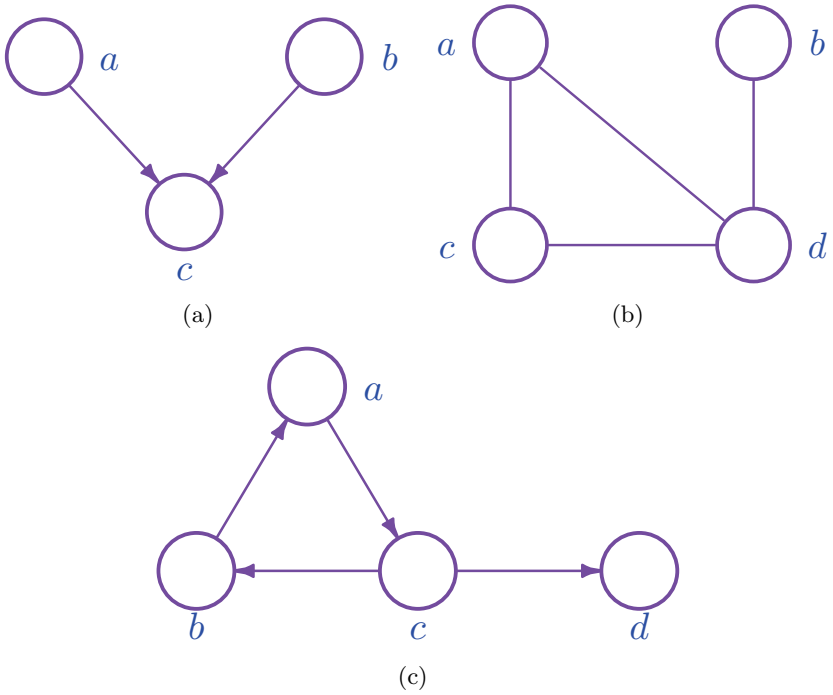


Figure 1.1: *(a) Example of a directed graph. (b) Example of an undirected graph. (c) A directed graph which has cycle* (a, c, b).

Throughout this thesis, we will use *DAG* as shorthand for *directed acyclic graph*. Figure 1.2 displays a simple example of a DAG.

As mentioned previously, graphical representations of hierarchical Bayesian models are an effective tool in understanding the conditional dependence structure of such models. This is achieved when we consider the nodes on a DAG to represent the random variables in the model, and the directed edges to represent the conditional dependence structure of the model. It is often easier to think of a DAG as a *family tree*, where a directed edge represents a *child-parent* relationship within the family.

Definition 1.8.5. *In a DAG, if two nodes are connected by a directed edge then the node connected by the arrow-head is the* ***child*** *and the node connected by the tail is the* ***parent****. If two nodes have a child in common, they are known as* ***co-parents****.*

Definition 1.8.6. *Within a DAG, the* ***Markov blanket*** *of a node is the set containing the node's children, parents and co-parents.*

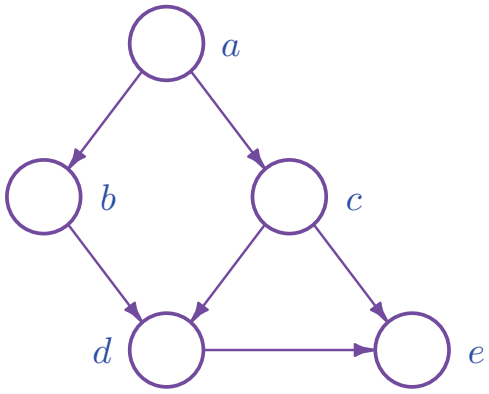
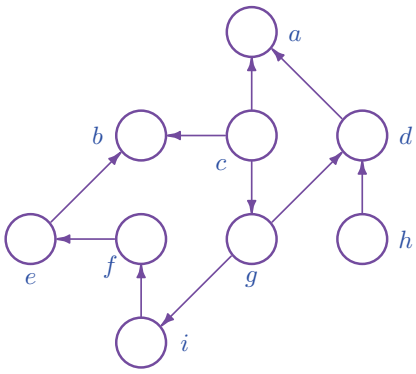


Figure 1.2: *A simple DAG containing nodes a, b, c, d and e.*



node	children	parents	co-parents	Markov blanket
a	$\{\}$	$\{c,d\}$	$\{\}$	$\{c,d\}$
b	$\{\}$	$\{c,e\}$	$\{\}$	$\{c,e\}$
c	$\{a,b,g\}$	$\{\}$	$\{d,e\}$	$\{a,b,d,e,g\}$
d	$\{a\}$	$\{g,h\}$	$\{c\}$	$\{a,c,g,h\}$
e	$\{b\}$	$\{f\}$	$\{c\}$	$\{b,c,f\}$
f	$\{e\}$	$\{i\}$	$\{\}$	$\{e,i\}$
g	$\{d,i\}$	$\{c\}$	$\{h\}$	$\{c,d,h,i\}$
h	$\{d\}$	$\{\}$	$\{g\}$	$\{d,g\}$
i	$\{f\}$	$\{g\}$	$\{\}$	$\{f,g\}$

Figure 1.3: *Left panel: Example of a DAG. Right panel: Table of each node's corresponding children, parents, co-parents and Markov blanket from the given DAG.*

Figure 1.3 illustrates the *family tree* terminology for a given DAG, including its Markov blanket. We now turn our attention to probabilistic DAGs where the nodes represent random variables $\{x_1, \ldots, x_k\}$ with joint density function $p(x_1, \ldots, x_k)$, given by Definition 1.8.7.

Definition 1.8.7. *The joint density function of the random variables $x_1, \ldots, x_k$ represented by nodes in a DAG is given by*

$$p(x_1, \ldots, x_k) = \prod_{i=1}^{k} p(x_i \,|\, \textit{parents of } x_i).$$

DAGs aid in the calculation of the full conditional distributions as is clear from the following definition.

Definition 1.8.8. *The full conditional distribution of a node x_i is the distribution of x_i given all of the other random variables in the DAG, and is denoted by $p(x_i|rest)$.*

The following result shows that probabilistically we can separate each node and its corresponding Markov blanket with the remainder of the DAG.

Result 1.8.1. *Let $p(x_1, \ldots, x_k)$ be a density function defined on a DAG. Then*

$$p(x_i|rest) = p(x_i|Markov\ blanket\ of\ x_i), \quad 1 \leq i \leq k.$$

Result 1.8.1 corresponds to Section 3.2 of Jordan (2004). Making use of Result 1.8.1 in equation (1.4) we now have the optimal q-densities satisfying

$$q_i^*(x_i) \propto \exp\left\{E_{-x_i} \log p(x_i|\text{Markov blanket of } x_i)\right\}, \quad 1 \leq i \leq k. \tag{1.7}$$

This is known as the *locality property* of DAGs and can be extremely important when performing MFVB on large models.

1.8.2 Moral graphs

Moral graphs are undirected graphs that are constructed from DAGs by forcing the *immoral* parents of a common node to get *married*.

Definition 1.8.9. *The **moral graph** of a DAG is constructed by adding an undirected edge between all pairs of parents of a node that are not already connected and turning all directed edges into undirected edges. This procedure is referred to as **moralisation**.*

The following two definitions deal with ancestral sets, which are useful for results concerning probabilistic graphs.

Definition 1.8.10. *An **ancestral set** is a set of nodes in a DAG such that for each node in the set, all of its parents are also in the set.*

Definition 1.8.11. *Let S be a subset of nodes in a DAG. Then the ancestral set containing S that has the fewest number of nodes is the **smallest ancestral set containing S**.*

The above definitions aid in the construction of the Theorem 1.8.1.

Theorem 1.8.1. *If A, B and C are disjoint node subsets in a DAG then $A \perp\!\!\!\perp B|C$ if C separates A from B in the moral graph of the smallest ancestral set containing $A \cup B \cup C$.*

Theorem 1.8.1 corresponds to Corollary 3.23 of Lauritzen (1996).

1.8.3 Graphical models viewpoint of mean field variational Bayes

When considering the conditional independence structure of a DAG, the main benefit of incorporating moral graphs and ancestral sets for MFVB is their extremely simple nature compared to that of d-separation (Bishop, 2006). For example, consider the Bayesian hierarchical Gaussian linear mixed model:

$$\begin{gathered}\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2 \sim \mathrm{N}\left(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u},\sigma_\varepsilon^2\boldsymbol{I}\right), \quad \boldsymbol{u}|\sigma_{u_1}^2,\sigma_{u_2}^2,\sigma_{u_3}^2 \sim \mathrm{N}\left(\boldsymbol{0},\mathrm{blockdiag}\left(\sigma_{u_1}^2,\sigma_{u_2}^2,\sigma_{u_3}^2\right)\right), \\ \boldsymbol{\beta} \sim \mathrm{N}\left(\boldsymbol{0},\sigma_\beta^2\boldsymbol{I}\right), \quad \sigma_{u_\ell}^2 \sim \text{Inverse-Gamma} \sim \left(A_{u_\ell},B_{u_\ell}\right), \quad 1 \le \ell \le 3, \\ \sigma_\varepsilon^2 \sim \text{Inverse-Gamma} \sim \left(A_\varepsilon,B_\varepsilon\right),\end{gathered} \tag{1.8}$$

for $\sigma_\beta^2, A_{u_\ell}, B_{u_\ell}, A_\varepsilon, B_\varepsilon > 0$ and where $\boldsymbol{y}$ is a vector of responses and $\boldsymbol{X}$ and $\boldsymbol{Z}$ are design matrices. Also, $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are vectors of fixed effects and random effects, respectively, and $\sigma_{u_1}^2, \sigma_{u_2}^2, \sigma_{u_3}^2$ and σ_ε^2 are variance parameters. Figure 1.4 shows the DAG and corresponding moral graph for (1.8) where the shaded node corresponds to the observed data vector. Random effects and auxiliary variables are referred to as hidden nodes. The conditional

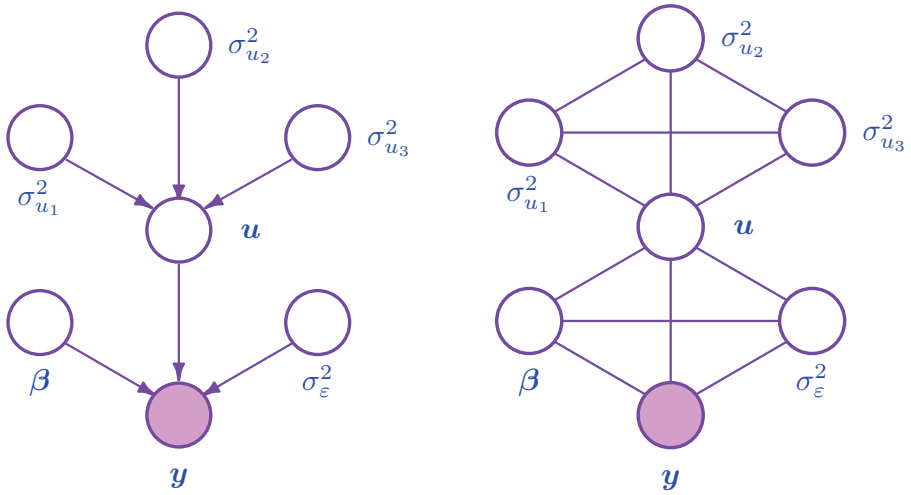


Figure 1.4: *(a) DAG representing model (1.8). (b) Moral graph of DAG in (a).*

dependence structure of (1.8) is illustrated through the directed edges in the DAG in Figure 1.4 (a).

MFVB is focused around imposing a product restriction on the joint posterior density function. For instance, we may impose the restriction

$$q(\boldsymbol{\beta}, \boldsymbol{u}, \sigma^2_{u_1}, \sigma^2_{u_2}, \sigma^2_{u_3}, \sigma^2_{\varepsilon}) = q(\boldsymbol{\beta}, \boldsymbol{u})\, q(\sigma^2_{u_1}, \sigma^2_{u_2}, \sigma^2_{u_3}, \sigma^2_{\varepsilon}). \tag{1.9}$$

The right-hand side of (1.9) corresponds to removing the edges connecting $\boldsymbol{\beta}$ with σ^2_{ε}, $\boldsymbol{u}$ with σ^2_{ε}, $\boldsymbol{u}$ with $\sigma^2_{u_1}$, $\boldsymbol{u}$ with $\sigma^2_{u_2}$, and $\boldsymbol{u}$ with $\sigma^2_{u_3}$ in the moral graph in Figure 1.4 (b). A tractable solution indeed arises for this product restriction, however to illustrate the simplicity of using moralisation for induced factorisations, we assume that a further breakdown is necessary.

Let $A = \left\{\sigma^2_{u_1}, \sigma^2_{u_2}, \sigma^2_{u_3}\right\}$, $B = \sigma^2_{\varepsilon}$ and $C = \{\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{u}\}$. The smallest ancestral set containing $A \cup B \cup C$ is the entire DAG. It is also evident from the moral graph that all paths between $\left\{\sigma^2_{u_1}, \sigma^2_{u_2}, \sigma^2_{u_3}\right\}$ and σ^2_{ε} must pass through at least one of $\{\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{u}\}$. Therefore, adopting Theorem 1.8.1 we see that

$$\left\{\sigma^2_{u_1}, \sigma^2_{u_2}, \sigma^2_{u_3}\right\} \perp\!\!\!\perp \sigma^2_{\varepsilon} \,|\, \{\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{u}\},$$

and hence (1.9) is updated to

$$q(\boldsymbol{\beta}, \boldsymbol{u}, \sigma^2_{u_1}, \sigma^2_{u_2}, \sigma^2_{u_3}, \sigma^2_{\varepsilon}) = q(\boldsymbol{\beta}, \boldsymbol{u})\, q(\sigma^2_{u_1}, \sigma^2_{u_2}, \sigma^2_{u_3})\, q(\sigma^2_{\varepsilon}).$$

Often, the use of induced factorisations is very beneficial, especially when considering increasingly complex models. The use of moral graphs enables detection of these factorisations which results in efficient derivations of MFVB solutions.

1.9 Accuracy measure of mean field variational Bayes

The accuracy of MFVB can range from excellent to relatively poor, depending on the structure of the DAG and type of product restriction imposed on the model i.e. number of edges removed. This section gives detail on the accuracy measure used to assess the performance of MFVB against its MCMC benchmark.

Let θ represent a generic univariate parameter. We are interested in measuring the accuracy of an MFVB approximate density $q^*(\theta)$ with respect to the exact posterior density $p(\theta|\boldsymbol{y})$. There are various ways in which this can be done. Faes *et al.* (2011) recommend working with the L_1 loss, or integrated absolute error (IAE) of $q^*(\theta)$. This quantity is given by

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

The fact that $\text{IAE}\left\{q^*(\theta)\right\}$ lies between 0 and 2 gives motivation to the accuracy measure

$$\text{accuracy}\left\{q^*(\theta)\right\} = 1 - \tfrac{1}{2}\text{IAE}\left\{q^*(\theta)\right\},$$

where the quantity $\frac{1}{2}\text{IAE}\left\{q^*(\theta)\right\}$ is the total variation metric corresponding to the probability measure induced by $p(\theta|\boldsymbol{y})$ and $q^*(\theta)$. Since $0 \leq \text{accuracy}\left\{q^*(\theta)\right\} \leq 1$ and thus can be expressed as a percentage. The posterior $p(\theta|\boldsymbol{y})$ can be approximated quite well by performing MCMC with sufficiently large sample sizes.

1.10 O'Sullivan penalised splines

O'Sullivan splines are an immediate generalisation of smoothing splines (e.g. Green & Silverman, 1994). Prior to the mid-1990s, smoothing splines were of particular interest in the spline-based non-parametric regression literature. This involved the number of basis functions approximately equal to the sample size (e.g. Wahba, 1990; Green & Silverman, 1994). O'Sullivan penalised splines possess the attractive feature of using considerably fewer basis functions, and their natural boundary conditions make them a very reliable choice of basis

in semiparametric and nonparametric regression. Wand & Ormerod (2008) give a detailed description of how O'Sullivan penalised splines can be added to the mixed model-based regression setting including examples with R code. We summarise the approach here.

Consider the simple non-parametric regression setting:

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

where $(x_i, y_i) \in \mathbb{R} \times \mathbb{R}$. Suppose we are interested in an estimate of f over the interval $[a, b]$ comprising the x_is. For $K \le n$, let $B_1, \ldots, B_{K+4}$ be the cubic-B-spline basis functions defined by the knot sequence $a = \kappa_1 = \kappa_2 = \kappa_3 = \kappa_4 < \kappa_5 < \ldots < \kappa_{K+4} < \kappa_{K+5} = \kappa_{K+6} = \kappa_{K+7} = \kappa_{K+8} = b$. We require a function that minimises the penalised residual sum of squares

$$\text{RSS}(f, \lambda) = \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \int_a^b \left\{f''(x)\right\} dx, \tag{1.10}$$

where $\lambda > 0$ is a smoothing parameter. The solution is the O'Sullivan penalised spline $f(\boldsymbol{x}) = \boldsymbol{B}\boldsymbol{\nu}$ and thus (1.10) reduces to

$$\text{RSS}(\boldsymbol{\nu}, \lambda) = (\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\nu})^\top (\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\nu}) + \lambda \boldsymbol{\nu}^\top \boldsymbol{\Omega} \boldsymbol{\nu} \tag{1.11}$$

where $B_{ik} = B_k(x_i)$, and $\Omega_{kk'} = \int_a^b B_k''(x) B_{k'}''(x) dx$. It is easy to see that the solution to (1.11) is

$$\hat{\boldsymbol{\nu}} = (\boldsymbol{B}^\top \boldsymbol{B} + \lambda \boldsymbol{\Omega})^{-1} \boldsymbol{B}^\top \boldsymbol{y} \tag{1.12}$$

and the fitted O'Sullivan penalised spline is thus given by $\hat{f}(\boldsymbol{x}) = \boldsymbol{B}\hat{\boldsymbol{\nu}}$. Computation of the design matrix $\boldsymbol{B}$ is fairly straightforward, however computation of $\boldsymbol{\Omega}$ can be somewhat challenging. Using the theorem in Section 6 of Wand & Ormerod (2008) we get an expression for $\boldsymbol{\Omega}$:

$$\boldsymbol{\Omega} = (\tilde{\boldsymbol{B}}'')^\top \text{diag}(\boldsymbol{\omega}) \tilde{\boldsymbol{B}}'' \tag{1.13}$$

where $\tilde{\boldsymbol{B}}''$ is the $3(K+7) \times (K+4)$ matrix with (i, j)th entry equal to $B_j''(\tilde{x}_i)$ and $\tilde{x}_i$ is the ith entry of $\tilde{\boldsymbol{x}} = (\kappa_1, (\kappa_1 + \kappa_2)/2, \kappa_2, \kappa_2, (\kappa_2 + \kappa_3)/2, \kappa_3, \ldots, \kappa_{K+7}, (\kappa_{K+7} + \kappa_{K+8})/2, \kappa_{K+8})$. Also, $\boldsymbol{\omega}$ is the $3(K+7) \times 1$ vector given by

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

where $(\Delta_{\boldsymbol{\kappa}})_k \equiv \kappa_{K+1} - \kappa_k, 1 \leq k \leq K$. The expression in (1.13) involves Simpson's rule, which is applied over each of the inter-knot differences. For more detail on the construction of (1.13) see Section 2 of Wand & Ormerod (2008). Throughout this thesis we work with mixed model representations of Bayesian hierarchical models and thus show how to incorporate O-Sullivan penalised splines using this representation in the following section.

1.10.1 Mixed model representation

The penalised splines can be written in a linear mixed model form:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon}, \quad \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \sigma_u^2 \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_\varepsilon^2 \boldsymbol{I} \end{bmatrix} \right), \tag{1.14}$$

where $\boldsymbol{X} = [1 \;\; x_i]_{1 \leq i \leq n}$ and $\boldsymbol{Z}$ are design matrices and $\boldsymbol{Z}$ contains B-spline basis functions. A convenient form of expressing $\hat{\boldsymbol{\nu}}$ in (1.12) is by using best linear unbiased prediction (BLUP) of $\boldsymbol{\beta}$ and $\boldsymbol{u}$ (e.g. Ruppert *et al.*, 2003, Section 4.5.3). This leads to

$$\hat{\boldsymbol{\nu}} = \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{u}} \end{bmatrix} = \left(\boldsymbol{C}^\top \boldsymbol{C} + \lambda \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \right)^{-1} \boldsymbol{C}^\top \boldsymbol{y}, \tag{1.15}$$

where $\lambda = \sigma_u^2 / \sigma_\varepsilon^2$ and $\boldsymbol{C} = [\boldsymbol{X} | \boldsymbol{Z}]$. The equivalence of (1.15) and (1.12) can be quantified by using a $(K+4) \times (K+4)$ linear transformation matrix $\boldsymbol{L}$ such that

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

Wand & Ormerod (2008) show that the spectral decomposition of $\boldsymbol{\Omega}$ has the form $\boldsymbol{\Omega} = \boldsymbol{U}\mathrm{diag}(\boldsymbol{d})\boldsymbol{U}^\top$, where $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}$ and $\boldsymbol{d}$ is a $(K+4)\times 1$ vector with 2 zero entries and all others positive. Also, the linear transformation matrix $\boldsymbol{L}$ has the form $\boldsymbol{L} = [\boldsymbol{U}_{\boldsymbol{X}} | \boldsymbol{U}_{\boldsymbol{Z}} \, \mathrm{diag}(\boldsymbol{d}_{\boldsymbol{Z}}^{-1/2})]$, where $\boldsymbol{d}_{\boldsymbol{Z}}$ is the $(K+2)\times 1$ sub-vector of $\boldsymbol{d}$ containing the positive entries and $\boldsymbol{U}_{\boldsymbol{Z}}$ is the $(K+4) \times (K+2)$ sub-matrix of $\boldsymbol{U}$ with columns corresponding to the positive entries in $\boldsymbol{d}$. Therefore, O-Sullivan penalised splines can be used for fitting (1.14) with the following design matrices:

$$\boldsymbol{X} = \boldsymbol{B}\boldsymbol{U}_{\boldsymbol{X}} \quad \text{and} \quad \boldsymbol{Z} = \boldsymbol{B}\boldsymbol{U}_{\boldsymbol{Z}} \mathrm{diag}\left(\boldsymbol{d}_{\boldsymbol{Z}}^{-1/2}\right),$$

whilst keeping in mind that $\boldsymbol{B}\boldsymbol{U}_{\boldsymbol{X}}$ is a basis for straight lines (Speed, 1991), and using $\boldsymbol{X}$ instead will not affect the fit.

1.11 Statistical software

Many semiparametric and nonparametric regression models can be formulated as hierarchical Bayesian models. An upside to this type of formulation is that MCMC software for these type of models can easily be used for fitting and inference. It is possible to tackle different semiparametric regression contexts through the use of BUGS (Bayesian inference Using Gibbs Sampling) inference engine (Spiegelhalter *et al.*, 2003).

In the R computing environment (R Core Team, 2013) BUGS can be accessed through the package BRugs (Ligges *et al.*, 2009), thus enabling complete analysis through a single R script. At the moment, BRugs is only compatible with the Windows operating system communicating through a version of BUGS known as OpenBUGS (Thomas *et al.*, 2006).

MCMC samples can also be obtained using Stan (Stan Development Team, 2014) through the R computing environment, made possible using the rstan package (Stan Development Team, 2014). Stan implements Bayesian inference using Hamiltonian Monte Carlo (HMC) sampling, written in C ++. An alternative to MCMC, which we explore in this thesis, is the variational approximation and in particular MFVB. Some inference engines have come to light recently for performing variational inference using graphical models. These include VIBES (Bishop *et al.*, 2003), short for Variational Inference for BayESian networks and Infer.NET (Minka *et al.*, 2008). Even though a very useful tool in variational approximations, we do not make use of VIBES or Infer.NET as they do not cater to the complexity of most of the models considered in this thesis.

In summary, graphical model-based Bayesian inference engines such as BUGS, Stan and Infer.NET, currently entertain a large variety of semiparametric and nonparametric regression settings (e.g. Marley & Wand, 2010; Luts *et al.*, 2014). BUGS and Stan both make use of Monte Carlo methodology, which is extremely accurate but can be quite slow. Infer.NET on the other hand makes use of variational approximations, such as MFVB, which are slightly less accurate but much faster.

1.11.1 Example in BRugs and RStan

Here we present R code for the simple Binomial model with Beta prior:

$$\boldsymbol{X}|p \sim \text{Binomial}(n, p)$$
$$p \sim \text{Beta}(A, B).$$

After specification of the data, sample size n, probability p, and the two hyperparameters A and B, the model in BRugs takes the form:

```
# Specify model in BRugs:
model
{
   for(i in 1:n)
   {
      x[i] ~ dbern(p)
   }
   p ~ dbeta(A,B)
}
```

Whilst coding the model in RStan is of the form:

```
# Specify model in RStan:
binomBetaModel <- '
   data
   {
      int<lower=1> n;
      int<lower=0,upper=1> x[n];
   }
   parameters
   {
      real<lower=0,upper=1> p;
   }
   model
   {
      for (i in 1:n)
         x[i] ~ bernoulli(p);
      p ~ beta(A,B);binomBetaModel
   }'
```

Throughout this thesis, we make use of both BRugs and RStan for MCMC sampling.

1.12 The flow of the thesis

The primary aim of this thesis is to determine whether MFVB can be used as an alternative to traditional methods for performing inference in semiparametric and nonparametric regression settings. We show that this is possible for various regression settings to allow for fast approximate inference when traditional inference is time consuming or intractable. Each chapter in this thesis is dedicated to the development of an MFVB algorithm catered to a particular model setting. In some instances this requires extension of previously developed MFVB methodology to the model being considered, whereas in others, it involves the original development of an MFVB algorithm. We then use various diagnostic techniques to assess the efficacy of the MFVB algorithm in terms of convergence, timing and accuracy.

Chapter 2 sees the development of an MFVB algorithm for the marginal longitudinal semiparametric regression setting. There are numerous contributions to the topic, however we focus on the Bayesian penalised spline approach of Al Kadiri *et al.* (2010) who use MCMC as an effective way to fitting their class of models and thus we present a comparison of both approaches. The main contribution here lies in the MFVB algorithm being one of the first variational algorithms to involve estimation of an unstructured covariance matrix. In Chapter 3, a modification of MFVB is discussed, known as *non-conjugate variational message passing*, which aids in the incorporation of heteroscedasticity, whilst also involving only closed form algebraic expressions. Chapter 4 looks into the development of MFVB algorithms for three extensions of the heteroscedastic setting given in Chapter 3. This includes an algorithm catered to performing semiparametric regression in real time, bivariate predictor nonparametric regression and in an additive model setting. Accuracy scores and time comparisons are also considered as an effective assessment of the algorithms used. In Chapter 5 we focus on the development of MFVB algorithms for fitting large data sets that exhibit multilevel and longitudinal structures. In particular, we take advantage of a streamlined approach to fitting these type of data. Two types of models are considered: two-level and three-level Gaussian response models. Each are motivated by a corresponding real dataset. Chapter 6 is concerned with MFVB inference for mixture models in measurement error problems. This chapter is heavily motivated by the area of epidemiology, where measurements often do not get measured accurately on all subjects. In Chapter 7, we present a discussion about an alternative approach to MFVB known as

variational message passing, which produces identical results to MFVB but is based on a different algebraic system that allows easier extension to arbitrarily large models. We show this approach for the models considered in Chapter 2 and 6.

Chapter 2

Mean field variational Bayes for marginal longitudinal semiparametric regression

2.1 Introduction

The definitive property of a longitudinal data set is repeated measurements on subjects enabling direct study of change over time. Longitudinal data hence require non-traditional statistical methods, since the set of measurements on one subject tends to be intercorrelated, which must be taken into account when drawing statistical inferences. Over the past decade, a great deal of exposure of longitudinal data in semiparametric regression has emerged. In particular, the special case in which the covariance matrix for each subject's responses is treated as an unspecified parameter to be estimated. This is known as the *marginal longitudinal semiparametric regression* problem. Benefits to marginal longitudinal semiparametric regression analysis include flexible estimation of regression functions and marginal estimation of the covariance matrix of the response vector for each subject.

A proliferation of contributions to the topic emerged in wake of Lin & Carroll (2001) showing that ordinary kernel smoothers are more efficient if *working-independence* is assumed. A summary of research on this problem prior to 2007 is outlined in Ruppert *et al.* (2009, section 3.9). One estimation method for marginal models is MCMC. The primary

The content of this chapter is published as: Menictas, M. and Wand. M.P. (2013). Variational inference for marginal longitudinal semiparametric regression. *Stat*, **2**, 61–71. This research was also presented at the *Young Statisticians Conference*, Melbourne, 2013.

aim of this chapter is to further discussions on the benefit of using a fast deterministic alternative to MCMC known as MFVB, for the marginal longitudinal semiparametric regression problem and some of its semiparametric extensions.

We focus on the Bayesian penalized spline approach of Al Kadiri *et al.* (2010) for function estimation. For fitting, Al Kadiri *et al.* (2010) use MCMC as a fitting and inference tool. However, they admit that such an approach is quite slow in estimation. Their real data example takes almost an hour to run on a contemporary laptop when using BUGS for the MCMC. The variational algorithm developed in this chapter, based on the MFVB paradigm, fits the data in seconds with very similar results. In addition, our algorithm is one of the first variational algorithms involving estimation of an unstructured covariance matrix.

Section 2.2 provides a brief overview of the marginal longitudinal semiparametric regression models considered in Al Kadiri *et al.* (2010). The corresponding MFVB algorithm is developed in Section 2.3. Section 2.4 provides a simulation study for assessing the performance of MFVB against that of MCMC. We provide a real data example in Section 2.5 and the appendices provide algebraic details tied to the methodology in this chapter.

2.2 Marginal longitudinal semiparametric regression models

Nonparametric techniques have recently come into view as an effective way to model longitudinal data. Longitudinal data comprise repeated measurements which are recorded over time on the same subject. We use y_{ij} to denote a measurement for the ith subject at the jth time point. Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in})$ be the vector of responses, $\mathbf{x}_i = (x_{i1}, \ldots, x_{in})$ be the vector of predictors for the ith subject, where $1 \leq i \leq m$ and $1 \leq j \leq n$.

As such, the marginal longitudinal nonparametric regression model is

$$E(y_{ij}) = f(x_{ij}), \quad \text{Cov}\{\boldsymbol{y}_i | f(\mathbf{x}_i)\} = \boldsymbol{\Sigma}, \quad 1 \leq i \leq m, \;\; 1 \leq j \leq n, \tag{2.1}$$

where we assume that the mean function f is a real-valued smooth function, and the covariance matrix $\boldsymbol{\Sigma}$ has dimension $n \times n$. In order to allow for cases where $f(\mathbf{x}_i)$ is random as specified by the model, rather than using $\text{Cov}(\boldsymbol{y}_i)$ to denote the covariance of $\boldsymbol{y}_i$, we use $\text{Cov}\{\boldsymbol{y}_i | f(\mathbf{x}_i)\}$. Figure 2.1 shows, by example, a simulated data set for model

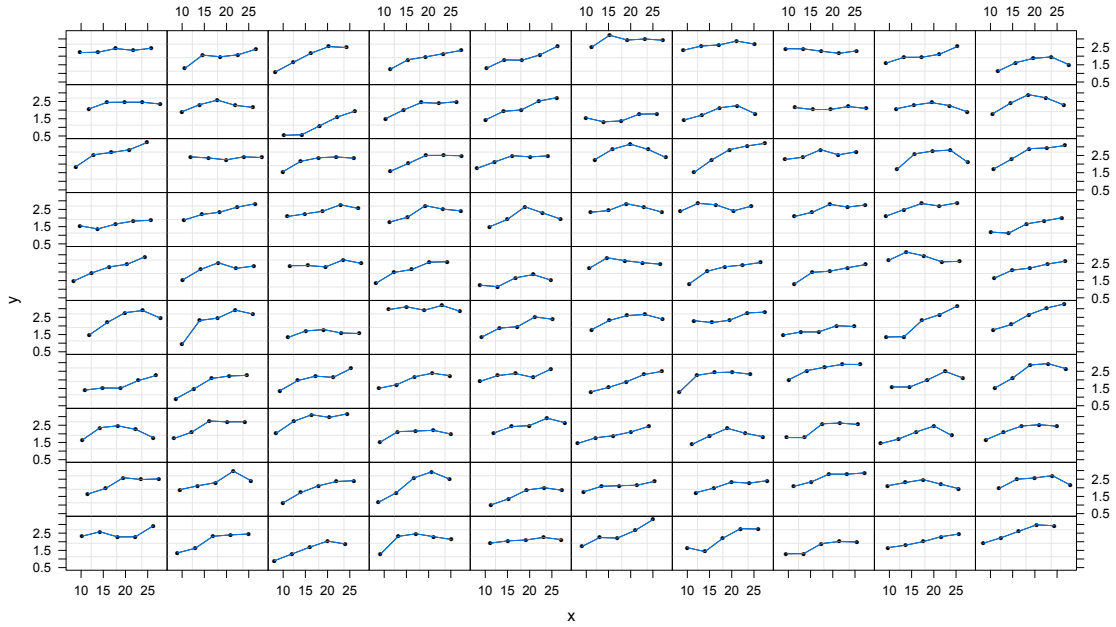


Figure 2.1: *A data set simulated from the marginal longitudinal nonparametric regression model (2.1) with* $m = 100$, $n = 5$ *and* f *and* Σ *are as described in (2.2).*

(2.1), with $m = 100$, $n = 5$,

$$f(x) = 1.5 + \sin\left(0.2(x/2 - 4)\right) \text{ and } \mathbf{\Sigma} = \begin{bmatrix} 0.160 & 0.120 & 0.090 & 0.068 & 0.051 \\ 0.120 & 0.160 & 0.120 & 0.090 & 0.068 \\ 0.090 & 0.120 & 0.160 & 0.120 & 0.090 \\ 0.068 & 0.090 & 0.120 & 0.160 & 0.120 \\ 0.051 & 0.068 & 0.090 & 0.120 & 0.160 \end{bmatrix}. \quad (2.2)$$

We are interested in efficient estimation of f and $\mathbf{\Sigma}$ from data such as that shown in Figure 2.1. The procedure we use to obtain the mean function estimate involves spline models for f of the form

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} u_k z_k(x), \quad (2.3)$$

where z_k, $1 \le k \le K$ represents a set of suitably transformed cubic O'Sullivan splines, as described in Section 1.10. The quantity of spline basis functions K has a small effect on the adequacy of (2.3). However, in practice, $K = 25$ is a sufficient choice for most spline basis functions (Li & Ruppert, 2008). In order to prevent over fitting of the function, we penalize the spline basis function coefficients u_k, $1 \le k \le K$, by treating them as a random sample from a distribution which has a mean of zero and a variance of σ^2. This allows for

the following linear mixed model representation of (2.1) and (2.3):

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

where

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_m \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{1} & \mathbf{x}_1 \\ \vdots & \vdots \\ \mathbf{1} & \mathbf{x}_m \end{bmatrix}, \quad \mathbf{Z} = \begin{bmatrix} z_1(\mathbf{x}_1) & \dots & z_K(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ z_1(\mathbf{x}_m) & \dots & z_K(\mathbf{x}_m) \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_m \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \quad \text{and} \quad \boldsymbol{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_K \end{bmatrix}.$$

The random effects have a mean of zero and a covariance matrix:

$$\text{Cov}\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_m \end{bmatrix} = \begin{bmatrix} \sigma^2\mathbf{I} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\Sigma} & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{\Sigma} \end{bmatrix} = \begin{bmatrix} \sigma^2\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_m \otimes \boldsymbol{\Sigma} \end{bmatrix}.$$

The model for the distribution of $\boldsymbol{u}$ and $\boldsymbol{\varepsilon}$ is

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \sigma^2\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_m \otimes \boldsymbol{\Sigma} \end{bmatrix}\right). \tag{2.5}$$

Sections 2.2.1, 2.2.2 and 2.2.3 provide semiparametric extensions to model (2.1).

2.2.1 Additive models extension

There are various semiparametric regressions extensions of (2.1) that could be considered. Here we consider the additive model extension. This involves the incorporation of several continuous predictor variables for each scalar response variable. We restrict our discussion to the incorporation of two continuous predictor variables x_{1ij} and x_{2ij} corresponding to each y_{ij}. In this case, the marginal longitudinal additive model is

$$\begin{gathered} E(y_{ij}) = \beta_0 + f_1(x_{1ij}) + f_2(x_{2ij}), \\ \text{Cov}\{\boldsymbol{y}_i | f_1(\mathbf{x}_{1i}), f_2(\mathbf{x}_{2i})\} = \boldsymbol{\Sigma}, \quad 1 \le i \le m, \;\; 1 \le j \le n, \end{gathered} \tag{2.6}$$

where we assume that the mean functions f_1 and f_2 are smooth functions and the covariance matrix $\boldsymbol{\Sigma}$ has dimension $n \times n$. The procedure we use to obtain the mean function

estimates involve spline models for f_1 and f_2 of the following form:

$$f_1(x_1) = \beta_{11}\mathbf{x}_1 + \sum_{k=1}^{K_1} u_{1k} z_{1k}(\mathbf{x}_1) \quad \text{and} \quad f_2(\mathbf{x}_2) = \beta_{21}x_2 + \sum_{k=1}^{K_2} u_{2k} z_{2k}(\mathbf{x}_2), \tag{2.7}$$

with coefficients $u_{1k} \overset{\text{ind.}}{\sim} N\left(0, {\sigma_1}^2\right)$, $1 \le k \le K_1$ and $u_{2k} \overset{\text{ind.}}{\sim} N\left(0, {\sigma_2}^2\right)$, $1 \le k \le K_2$. This allows for the following Gaussian linear mixed model representation of (2.6) and (2.7):

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}.$$

The differences between this model and the one considered in Section 2.2 lie in the structure of the design matrices. The design matrices are now

$$\mathbf{X} = \begin{bmatrix} \mathbf{1} & \mathbf{x}_{11} & \mathbf{x}_{21} \\ \vdots & \vdots & \vdots \\ \mathbf{1} & \mathbf{x}_{1m} & \mathbf{x}_{2m} \end{bmatrix}, \quad \text{and}$$

$$\mathbf{Z} = \begin{bmatrix} z_{11}(\mathbf{x}_{11}) & \dots & z_{1K_1}(\mathbf{x}_{11}) & z_{21}(\mathbf{x}_{21}) & \dots & z_{2K_2}(\mathbf{x}_{21}) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ z_{11}(\mathbf{x}_{1m}) & \dots & z_{1K_1}(\mathbf{x}_{1m}) & z_{21}(\mathbf{x}_{2m}) & \dots & z_{2K_2}(\mathbf{x}_{2m}) \end{bmatrix},$$

where $\mathbf{x}_{1i}$ is the $n \times 1$ vector containing the x_{1ij} measurements and $\mathbf{x}_{2i}$ is the $n \times 1$ vector containing the x_{2ij} measurements. The coefficient vectors are

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_{11} \\ \beta_{12} \end{bmatrix}, \quad \text{and} \quad \boldsymbol{u} = \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \end{bmatrix},$$

where $\boldsymbol{u}_1$ is the $K_1 \times 1$ vector containing the u_{1k} and $\boldsymbol{u}_2$ is the $K_2 \times 1$ vector containing the u_{2k}. A convenient assumption for estimating ${\sigma_1}^2$, ${\sigma_2}^2$, and $\boldsymbol{\Sigma}$ is now

$$\begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \\ \boldsymbol{\varepsilon} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} {\sigma_1}^2\mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & {\sigma_2}^2\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_m \otimes \boldsymbol{\Sigma} \end{bmatrix}\right). \tag{2.8}$$

Fitting via MFVB is analogous to that described in Section 2.3. The difference here is that there are two variance parameters ${\sigma_1}^2$ and ${\sigma_2}^2$ (where in extensions to additive models with d smooth functions there will be d such variance parameters) as well as the error covariance matrix $\boldsymbol{\Sigma}$. A simpler type of additive model is

$$\begin{gathered} E(y_{ij}) = \beta_0 + \beta_1 x_{1ij} + f_2(x_{2ij}), \quad \operatorname{Cov}\{\boldsymbol{y}_i | \mathbf{x}_{1i}, f_2(\mathbf{x}_{2i})\} = \boldsymbol{\Sigma}, \\ 1 \le i \le m, \;\; 1 \le j \le n. \end{gathered} \tag{2.9}$$

This model represents a semiparametric regression model in the true sense, since the right-hand side comprises a parametric component $\beta_1 x_{1ij}$ where the effect of the x_{1ij}s are modelled parametrically *and* a nonparametric component $f_2(\mathbf{x}_{2i})$, where the effect of the x_{2ij}s are modelled nonparametrically. This is a simpler model to that of (2.6), since there is only one smooth function component. However, the linear mixed model corresponding to this model is similar to that in Section 2.2, where the random component structure (2.5) applies to (2.9).

2.2.2 Incorporation of interactions

Given the situation where an additive model is not guaranteed to provide a satisfactory fit of the data, one may opt to incorporate an interaction term. Such an extension of (2.6) is

$$\begin{aligned} E(y_{ij}) &= \beta_0 + f_{x_{3ij}}(x_{1ij}) + f_2(x_{2ij}), \\ \text{Cov}\left\{\boldsymbol{y}_i | f_{x_{3ij}}(\mathbf{x}_{1i}),\ f_2(\mathbf{x}_{2i})\right\} &= \boldsymbol{\Sigma}, \quad 1 \le i \le m,\ \ 1 \le j \le n, \end{aligned} \tag{2.10}$$

where the x_{3ij} correspond to measurements on the categorical variable x_3. For example, if $x_{3ij} \in \{0, 1\}$ then we have $E(y_{ij}) = \beta_0 + f_0(x_{1ij}) + f_2(x_{2ij})$ if $x_{3ij} = 0$ and $E(y_{ij}) = \beta_1 + f_1(x_{1ij}) + f_2(x_{2ij})$ if $x_{3ij} = 1$. Interaction models such as (2.10) have two important uses. One of which is the ability to check for additivity in the model. This can be done by testing the null hypothesis of a true additive model i.e., one without any interaction terms, against the alternative hypothesis of the additive model involving interaction terms. If the null hypothesis is accepted, we assume that using the additive model will provide a reasonable fit to the data. Another important use of interaction models is the ability to use alternative models when an additive model fails to fit the data well. When the null hypothesis of an additive model is rejected, appropriate interaction terms are added to the additive model.

The Gaussian linear mixed model for fitting (2.10) has the same structure as that of fitting the additive model (2.6), where the only difference is the form of the design matrices $\mathbf{X}$ and $\mathbf{Z}$. An in-depth explanation including examples may be found in Ruppert *et al.* (2003).

2.2.3 Varying coefficient models

Varying coefficient models are another type of multiple-predictor semiparametric regression model, where the coefficients are allowed to vary as smooth functions of other variables. The model

$$E\left(y_{ij}\right) = f_0\left(s_{ij}\right) + f_1\left(s_{ij}\right) x_{ij}, \quad \operatorname{Cov}\left\{\boldsymbol{y}_i | f_0\left(\mathbf{s}_i\right)\right\} = \boldsymbol{\Sigma}, \\ 1 \leq i \leq m, \;\; 1 \leq j \leq n, \tag{2.11}$$

is the simplest marginal longitudinal varying coefficient model, where s_{ij} are the longitudinal measurements on a continuous predictor variable s and x_{ij} are the measurements on a second predictor x. Model (2.11) shows that s has a modifying effect on the linear relationship between $\mathrm{E}(\boldsymbol{y})$ and x, which is modelled through the varying coefficients $f_0(s)$ and $f_1(s)$. The gaussian linear mixed model for fitting (2.11), has the same structure as that for fitting the additive model (2.6), where the only difference is the form of the design matrices $\mathbf{X}$ and $\mathbf{Z}$. Relevant details are provided in Section 12.4 of Ruppert *et al.* (2003).

2.3 Mean field variational Bayes methodology

Each of the marginal longitudinal semiparametric regression models in section 2.2, and their extensions to d smooth functions, can be treated using the Gaussian linear mixed model

$$\boldsymbol{y}|\boldsymbol{u}, \sim N\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{u},\, \mathbf{I}_m \otimes \boldsymbol{\Sigma}\right), \quad \boldsymbol{u}|\boldsymbol{\sigma}^2 \sim N\left(\mathbf{0}, \underset{1\leq\ell\leq d}{\operatorname{blockdiag}}\left(\sigma_\ell^2 \mathbf{I}_{K_\ell}\right)\right), \tag{2.12}$$

where K_ℓ corresponds to the number of spline basis functions used in the ℓth smooth function estimate. Maximum likelihood and best prediction can be used as inference strategies for constructing estimates of the regression function f. Al Kadiri *et al.* (2010) discuss that, despite the Gaussian linear mixed model (2.4) being a simple model, fitting via standard mixed model software such as `lme()` (Pinheiro *et al.*, 2009) in the R computing language has not yet been successful. This led Al Kadiri *et al.* (2010) to consider the Bayesian inference version and implementation via MCMC. In this section, we discuss and define a deterministic approximation alternative to MCMC known as mean field variational Bayes. MFVB scales well to large applications with simple implementation and requires less computation time than MCMC. However, MFVB cannot generate exact results and seems to be limited in its approximation accuracy. Nevertheless, in certain circumstances

MFVB may be *very* accurate.

We consider working with a hierarchical Bayesian version of (2.12), where $\boldsymbol{\beta}(p \times 1)$, $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_d^2)$ and $\boldsymbol{\Sigma}(n \times n)$ are treated as random. Common prior distributions for these parameters are:

$$\boldsymbol{\beta} \sim N(\mathbf{0}, \mathbf{F}), \quad \sigma_\ell \sim \text{Half-Cauchy}(A_\ell), \quad \boldsymbol{\Sigma} \sim \text{Inverse-Wishart}(A_{\boldsymbol{\Sigma}}, B_{\boldsymbol{\Sigma}}). \tag{2.13}$$

An important concern to fitting and inference is the choice of the hyperparameters $\boldsymbol{F}$, A_ℓ, $A_{\boldsymbol{\Sigma}}$ and $B_{\boldsymbol{\Sigma}}$. When prior beliefs about the model parameters are held, quantities may be used to represent these beliefs about the hyperparameters. However in most instances, prior beliefs are unavailable. In cases such as these one may use vague priors. We assume that the data have been standardised before we impose the following priors for both the fixed effects and variance hyperparameters:

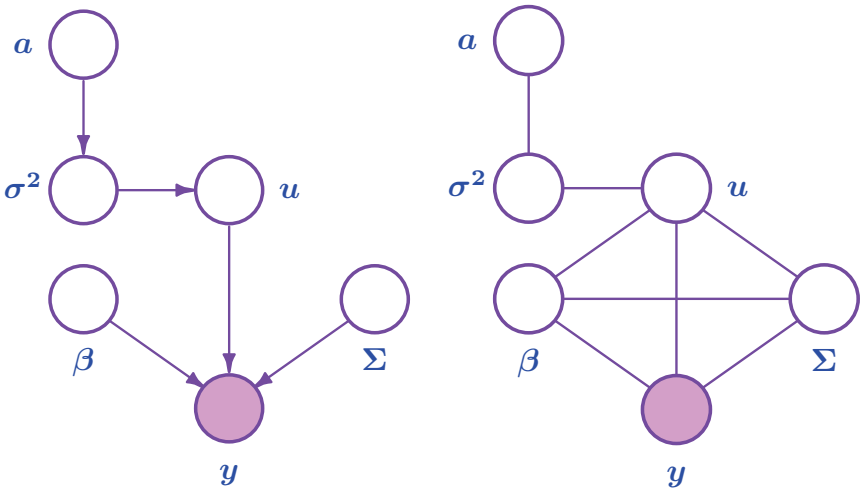
$$\boldsymbol{F} = \boldsymbol{I} \times 10^{10}, \quad A_\ell = 1 \times 10^5, \quad A_{\boldsymbol{\Sigma}} = n, \quad \text{and} \quad B_{\boldsymbol{\Sigma}} = 0.01\mathbf{I}_n. \tag{2.14}$$

Approximate inference via MCMC is aided by the distribution theoretical result given in Result 1.4.6. We achieve the following equivalent form of (2.12) and (2.13):

$$\begin{gathered} \boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\Sigma} \sim N(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u},\ \boldsymbol{I}_m \otimes \boldsymbol{\Sigma}), \quad \boldsymbol{u}|\sigma^2 \sim N\left(\mathbf{0}, \underset{1\le\ell\le d}{\text{blockdiag}}\left(\sigma_\ell^2 \boldsymbol{I}_{K_\ell}\right)\right), \\ \boldsymbol{\beta} \sim N(0, \boldsymbol{F}), \quad {\sigma_\ell}^2|a_\ell \sim \text{Inverse-Gamma}\left(\tfrac{1}{2}, \tfrac{1}{a_\ell}\right), \ a_\ell \sim \text{Inverse-Gamma}\left(\tfrac{1}{2}, \tfrac{1}{{A_\ell}^2}\right), \\ \boldsymbol{\Sigma} \sim \text{Inverse-Wishart}(A_{\boldsymbol{\Sigma}}, B_{\boldsymbol{\Sigma}}), \end{gathered} \tag{2.15}$$

where A_ℓ and $A_{\boldsymbol{\Sigma}}$ are positive constants, $\mathbf{F}$ and $B_{\boldsymbol{\Sigma}}$ are both positive definite matrices. Figure 2.2(a) illustrates the DAG corresponding to (2.15) and is effective in determining its full conditional distributions. The posterior distributions of $\boldsymbol{\beta}$, $\boldsymbol{u}$, σ_ℓ, $\boldsymbol{\Sigma}$, and a_ℓ, are not available in closed form. However the distributions of the full conditionals are shown to be:

$$\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} \Bigg| \text{rest} \sim N\left(\left(\boldsymbol{C}^\top (\boldsymbol{I}_m \otimes \boldsymbol{\Sigma})^{-1} \boldsymbol{C} + \begin{bmatrix} \mathbf{F}^{-1} & \mathbf{0} \\ \mathbf{0} & \underset{1\le\ell\le d}{\text{blockdiag}}\left(\frac{1}{\sigma_\ell^2}\boldsymbol{I}_{K_\ell}\right) \end{bmatrix}\right)^{-1} \mathbf{C}^\top (\boldsymbol{I}_m \otimes \boldsymbol{\Sigma})\, \boldsymbol{y}, \right.$$
$$\left. \left(\boldsymbol{C}^\top (\boldsymbol{I}_m \otimes \boldsymbol{\Sigma})\, \mathbf{C} + \begin{bmatrix} \mathbf{F}^{-1} & \mathbf{0} \\ \mathbf{0} & \underset{1\le\ell\le d}{\text{blockdiag}}\left(\frac{1}{\sigma_\ell^2} I_{K_\ell}\right) \end{bmatrix}\right)^{-1}\right),$$



(a) (b)

Figure 2.2: *(a) Directed acyclic graph representation of the Bayesian penalized spline model (2.15), where* $\boldsymbol{a} = (a_1, \ldots, a_d)$*, and* $\boldsymbol{\sigma}^2 = (\sigma_1{}^2, \ldots, \sigma_d{}^2)$*. (b) Moral graph representation of the DAG in (a).*

$$\begin{aligned}
\sigma_\ell^2|\text{rest} &\sim \text{Inverse-Gamma}\left(\tfrac{1}{2}\left(K_\ell + 1\right), \tfrac{1}{2} \parallel u_\ell \parallel^2 + a_\ell^{-1}\right), \\
a_\ell|\text{rest} &\sim \text{Inverse-Gamma}\left(1, \left(\sigma_\ell^{-2} + A^{-2}\right)\right) \text{ and} \\
\boldsymbol{\Sigma}|\text{rest} &\sim \text{Inverse-Wishart}\left(A_{\boldsymbol{\Sigma}} + m, \sum_{i=1}^{m} \left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{u}\right)\left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{u}\right)^\top + B_{\boldsymbol{\Sigma}}\right).
\end{aligned}$$

An elementary form of MCMC sampling, called Gibbs sampling, can be used to draw successive samples from these full conditional distributions. However, we demonstrate fitting via MFVB as it provides for faster fitting and inference to MCMC.

We next consider implementation of MFVB methodology, which involves restricting the full posterior density function to have the approximate product form

$$p(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{a}, \boldsymbol{\sigma}^2, \boldsymbol{\Sigma}|\boldsymbol{y}) \approx q(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{a}) q(\boldsymbol{\sigma}^2, \boldsymbol{\Sigma}). \tag{2.16}$$

Additional factorizations, referred to as *induced factorizations*, arise from an interaction between the factorization assumed in the variational posterior distribution and the conditional independence properties of the true joint distribution (Bishop, 2006). Such induced factorizations can easily be detected using moralisation (described in Sections 1.8.2 and 1.8.3). We show this using the moral graph of Figure 2.2(b). Let $A = \sigma^2$, $B = \boldsymbol{\Sigma}$, 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 from the moral graph in Figure 2.2(b) that all paths between σ^2 and $\boldsymbol{\Sigma}$ must

pass through at least one of $\{\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{y}\}$. Thus, applying Theorem 1.8.1 we see that

$$\sigma^2 \perp\!\!\!\perp \boldsymbol{\Sigma} | \{\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{y}\}.$$

Similarly, we can say that $\{\boldsymbol{\beta}, \boldsymbol{u}\} \perp\!\!\!\perp \boldsymbol{a} | \left\{\boldsymbol{y}, \sigma^2, \boldsymbol{\Sigma}\right\}$, and hence (2.16) is updated to

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

Then, as shown through the derivations in Appendix 2.A,

$$q^*(\boldsymbol{\beta}, \boldsymbol{u}) \text{ is a Multivariate Normal density function,}$$
$$q^*\left(\sigma_\ell^2\right) \text{ and } q^*(a_\ell) \text{ are each Inverse Gamma density functions,}$$
$$q^*(\boldsymbol{\Sigma}) \text{ is an Inverse-Wishart density function.}$$

Note that $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ denote the mean vector and covariance matrix for $q^*(\boldsymbol{\beta}, \boldsymbol{u})$, and $A_{q(\sigma_\ell^2)}$ and $B_{q(\sigma_\ell^2)}$ denote the shape and rate parameters for $q^*\left(\sigma_\ell^2\right)$. Similar definitions apply for the other density functions. The optimal values of these parameters are determined using Algorithm 1. The lower bound on the marginal log-likelihood is (the

---

Set up initial values:

$\mu_{q(1/\sigma_\ell^2)} > 0,\ \mu_{q(1/a_\ell)} > 0,\ 1 \le \ell \le d,\ M_{q(\boldsymbol{\Sigma}^{-1})}$ positive definite.

Cycle through:

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \left\{ \boldsymbol{C}^\top \left(\boldsymbol{I}_m \otimes \boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1})} \right) \boldsymbol{C} + \begin{pmatrix} \mathbf{F}^{-1} & 0 \\ 0 & \underset{1 \le \ell \le d}{\text{blockdiag}} \left(\mu_{q(1/\sigma_\ell^2)} \boldsymbol{I}_{K_\ell} \right) \end{pmatrix} \right\}^{-1}$$

$$\mu_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \mathbf{C}^\top (\mathbf{I}_m \otimes \mathbf{M}_{q(\boldsymbol{\Sigma}^{-1})}) \boldsymbol{y}$$

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

$$B_{q(a_\ell)} \leftarrow \mu_{q(1/\sigma_\ell^2)} + A^{-2} \ ; \ \mu_{q(1/a_\ell)} \leftarrow 1/B_{q(a_\ell)}$$
$$B_{q(\sigma_\ell^2)} \leftarrow \tfrac{1}{2} \left(\| \boldsymbol{\mu}_{q(\boldsymbol{u})} \|^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u})}) \right) + \mu_{q(1/a_\ell)}$$
$$\mu_{q(1/\sigma_\ell^2)} \leftarrow \tfrac{1}{2}(K_\ell + 1)/B_{q(\sigma_\ell^2)}$$

$$\boldsymbol{B}_{q(\boldsymbol{\Sigma})} \leftarrow \mathbf{B} + (\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})})(\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})})^\top + \boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^\top$$

$$M_{q(\boldsymbol{\Sigma}^{-1})} \leftarrow (a + m) B_{q(\boldsymbol{\Sigma})}^{-1}$$

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

Algorithm 1: *Mean field variational Bayes algorithm for determination of the optimal parameters in* $q^*(\boldsymbol{\beta}, \boldsymbol{u})$, $q^*(\boldsymbol{\sigma}^2)$, $q^*(\boldsymbol{a})$, *and* $q^*(\boldsymbol{\Sigma})$.

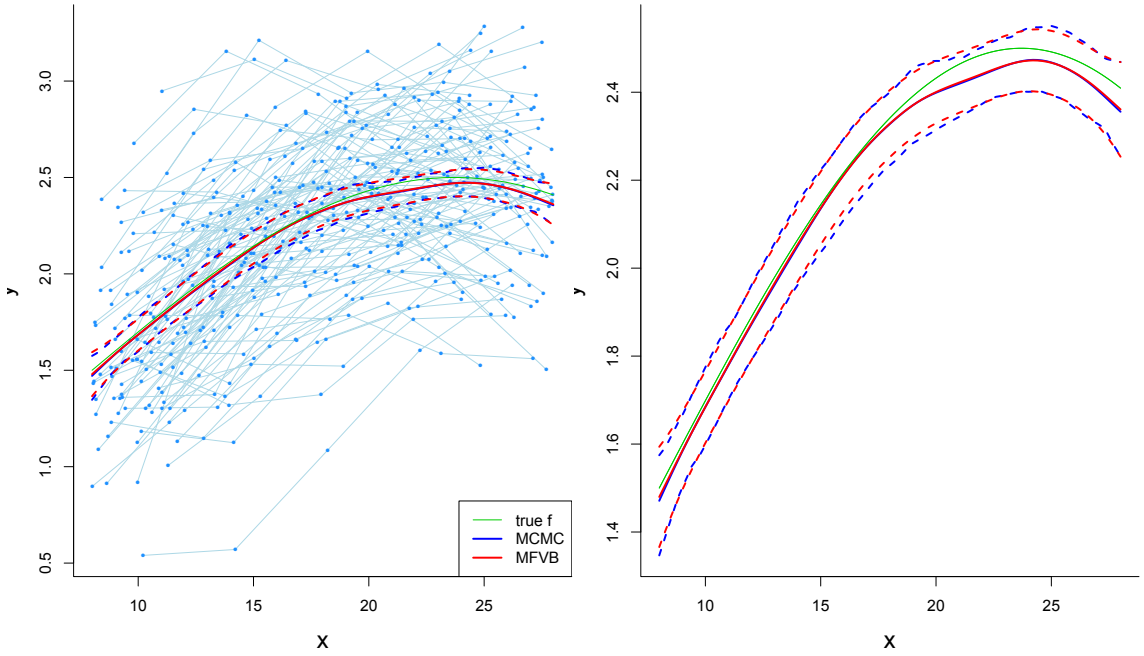


Figure 2.3: *Fitted curve estimates and pointwise 95% credible sets for both MCMC and MFVB approaches for fitting the Bayesian penalized spline model (2.15) to simulated data. Right panel: zoomed view with data points removed.*

derivation of which can be found in Appendix 2.B):

$$\begin{aligned}\log\, \underline{p}\,(\boldsymbol{y};q) \;=\; & \tfrac{1}{2}\left(\textstyle\sum_{\ell=1}^{d} K_\ell + p\right) - \tfrac{1}{2}\, m\, \log(2\pi) - \log(\pi) + \textstyle\sum_{\ell=1}^{d} \log\,\Gamma\left(\tfrac{1}{2}\,(K_\ell+1)\right)\\ & -\tfrac{1}{2}\,\log|\boldsymbol{F}| - \log(A) - \tfrac{1}{2}\mathrm{tr}\left\{\boldsymbol{F}^{-1}\left(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}^{\mathsf{T}} + \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right)\right\}\\ & +\tfrac{1}{2}\,\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},u)}| - \tfrac{1}{2}\textstyle\sum_{\ell=1}^{d}(K_\ell+1)\,\log\left(B_q\left(\sigma_\ell^2\right)\right)\\ & -\tfrac{1}{2}\,(A_{\boldsymbol{\Sigma}}+m)\,\log\,|B_{q(\boldsymbol{\Sigma})}| - \textstyle\sum_{\ell=1}^{d}\log\left(\mu_{q(1/\sigma_\ell^2)} + A^{-2}\right)\\ & +\textstyle\sum_{\ell=1}^{d}\mu_{q(1/\sigma_\ell^2)}\mu_{q(1/a_\ell)} + \tfrac{1}{2}\,A_{\boldsymbol{\Sigma}}\,(\log|B_\ell|) - \log(C_{n,A})\\ & +\log(C_{n,A+m})\,.\end{aligned} \tag{2.17}$$

Figure 2.3 illustrates the fitted curves using both MCMC and MFVB for the Bayesian penalized spline model (2.15). The data were simulated from a version of the marginal longitudinal nonparametric regression model (2.1) with $m = 100$, $n = 5$ and f and $\boldsymbol{\Sigma}$ as described in (2.2). The x_{ij} are equally spaced but with the starting positions x_{i1} generated uniformly from the interval $(8, 8+\frac{20}{n})$ and $\boldsymbol{\varepsilon} \sim N(0,1)$. We used diffuse priors given by (2.14). For the MCMC approach, samples of size 10,000 were generated where the first 5,000 values were discarded. For the MFVB approach the iterations were terminated when the relative change in $\log \underline{p}\,(\boldsymbol{y};q)$ fell below 10^{-7}. For this example the MCMC and MFVB fits and pointwise 95% credible sets are almost identical. This suggests that MFVB

achieves high accuracy for Gaussian response Bayesian penalized spline regression.

2.3.1 Heuristic justification of mean field variational Bayes

Here we are interested in justifying the initial MFVB approximate product restriction given in (2.16) for the model in (2.15). We consider the marginal longitudinal frequentist regression model

$$\boldsymbol{y} \sim \mathrm{N}\left(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{I}_m \otimes \boldsymbol{\Sigma}\right). \tag{2.18}$$

The likelihood-based estimates and confidence intervals attained from (2.18) will be similar to the Bayes estimates and credible sets based on the marginal longitudinal Bayesian regression model

$$\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\Sigma} \sim \mathrm{N}\left(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{I}_m \otimes \boldsymbol{\Sigma}\right),$$

given that the priors on $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are noninformative. In addition, asymptotic independence (or parameter orthogonality) of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ in (2.18) is implied by the block-diagonal form of the Fisher information matrix:

$$\boldsymbol{I}(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \begin{bmatrix} \boldsymbol{X}^{\mathsf{T}}(\boldsymbol{I}_m \otimes \boldsymbol{\Sigma}^{-1})\boldsymbol{X} & \boldsymbol{0} \\ \boldsymbol{0} & \frac{m}{2}(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}) \end{bmatrix}. \tag{2.19}$$

The derivation of (2.19) is given in Appendix 2.C. In the *Bayesian world*, such asymptotic independence corresponds to the approximate product form

$$p(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\boldsymbol{y}) \approx p(\boldsymbol{\beta}|\boldsymbol{y})p(\boldsymbol{\Sigma}|\boldsymbol{y}),$$

where the MFVB approximate product restriction $q(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = q(\boldsymbol{\beta})q(\boldsymbol{\Sigma})$ corresponds to the replacement of approximate with exact equality in this expression. Approaches such as best prediction (Robinson, 1991) need to be considered for frequentist inference of (2.15) for the estimation of the random effects vector $\boldsymbol{u}$. Even so, orthogonality between the mean and variance parameters of the model is still maintained. Hence, we would expect the MFVB approximate product restriction in (2.16) to be reasonable and not incur heavy losses in accuracy.

2.3.2 Mean field variational Bayes approximate density functions for entries of $\boldsymbol{\Sigma}$

In Section 2.4, we present the results from a comprehensive simulation study that compares the MFVB methodology against MCMC. We do this comparison on the mean function hexiles, $f(H_1), \ldots, f(H_5)$ and the entries of the variance covariance matrix, $\boldsymbol{\Sigma}_{11}, \boldsymbol{\Sigma}_{12}, \ldots, \boldsymbol{\Sigma}_{45}$. The mean function hexiles are straightforward to attain, however, the entries of the variance covariance matrix require additional effort. The derivations of the distributions of these entries are given in this section.

2.3.2.1 Diagonal entries

The distributions of the diagonal entries of $q^*(\boldsymbol{\Sigma})$ are obtained using Theorem 2.3.1.

Theorem 2.3.1. *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:*

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

Details of the proof of Theorem 2.3.1 are given in Appendix 2.D. From Algorithm 1 we see that the optimal q-density of $\boldsymbol{\Sigma}$ has the form

$$q^*(\boldsymbol{\Sigma}) \text{ has an Inverse-Wishart } \left(A_{q(\boldsymbol{\Sigma})},\ \boldsymbol{B}_{q(\boldsymbol{\Sigma})}\right) \text{ density function,}$$

where

$$\begin{aligned} A_{q(\boldsymbol{\Sigma})} &= a + m, \text{ and} \\ \boldsymbol{B}_{q(\boldsymbol{\Sigma})} &= \boldsymbol{B} + \left(\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\right)\left(\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\right)^\top + \boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^\top. \end{aligned}$$

Now, using Theorem 2.3.1, the diagonal entries take the form

$$q^*(\boldsymbol{\Sigma}_{jj}) \sim \text{Inverse-Gamma}\Bigg(\tfrac{1}{2}(a+m-n+1),\ \tfrac{1}{2}\left\{\boldsymbol{B} + \left(\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\right)\left(\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\right)^\top + \boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^\top\right\}_{jj}\Bigg),$$

where $j \in 1, \ldots, n$.

2.3.2.2 Off-diagonal entries

We considered two main approaches to finding the distributions of the off-diagonal entries of $q^*(\boldsymbol{\Sigma})$. These included: *(Approach 1)* Numerical characteristic function inversion based on the Fast Fourier Transform; or *(Approach 2)* Monte Carlo sampling of the off-diagonal elements in $q^*(\boldsymbol{\Sigma})$. It is most desirable to attain these densities exactly, however when exploring *Approach 1* we found that there was no straightforward way of obtaining the characteristic function of an Inverse-Wishart random matrix. As a result, we used *Approach 2*, subject to Monte Carlo error, to finding the distributions of these off-diagonal elements.

Whilst initially exploring *Approach 1* however, we came across some exciting results involving the characteristic function of the off-diagonal elements of a Wishart random matrix. This is presented in Theorem 2.3.2.

Theorem 2.3.2. *The characteristic function of the off-diagonal entries of a Wishart*$(a, \boldsymbol{B})$ *random matrix* $\boldsymbol{X}$ *take the form:*

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

where the 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.*

The proof of Theorem 2.3.2 is given in Appendix 2.E. This Theorem enables one to find the distributions of the off-diagonal entries of a Wishart random matrix analytically by making use of the Fast Fourier Transform inversion of the characteristic function.

2.4 Simulation study

A simulation study was carried out for Algorithm 1 in order to asses the accuracy of its Bayesian inference compared to that of MCMC. 1000 data-sets were generated, which corresponds to the simulation setting described in Section 5.1 of Al Kadiri *et al.* (2010). The nonparametric regression model is of the form

$$E(y_{ij}|\boldsymbol{u}) = f(x_{ij}), \quad \text{Cov}(\boldsymbol{y}_i|\boldsymbol{u}) = \boldsymbol{\Sigma}, \quad 1 \le i \le 100, \;\; 1 \le j \le 5, \tag{2.20}$$

with

$$f(x_i) = 1 + \tfrac{1}{2}\Phi\left(\frac{2x-36}{5}\right) \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 0.122 & 0.098 & 0.078 & 0.063 & 0.050 \\ 0.098 & 0.122 & 0.098 & 0.078 & 0.063 \\ 0.078 & 0.098 & 0.122 & 0.098 & 0.078 \\ 0.063 & 0.078 & 0.098 & 0.122 & 0.098 \\ 0.050 & 0.063 & 0.078 & 0.098 & 0.122 \end{bmatrix},$$

where Φ is the standard normal distribution function and (2.20) is an example of the special case of the Bayesian hierarchical model (2.15) with $d = 1$.

There are various approaches by which the accuracy of a variational Bayes approximate density function $q^*(\theta)$ may be assessed with respect to the exact posterior density $p(\theta|\boldsymbol{y})$. Avoiding approximate inference methods would be ideal, however, using MCMC with sufficiently large samples can easily be used to approximate $p(\theta|\boldsymbol{y})$ quite well. A detailed explanation is provided in Section 1.9.

Many parameters are of interest when conducting an assessment of MFVB against MCMC. We have chosen to work with the mean function hexiles, $f(H_1), \ldots, f(H_5)$ and the entries of the variance covariance matrix, $\boldsymbol{\Sigma}_{11}, \boldsymbol{\Sigma}_{12}, \ldots, \boldsymbol{\Sigma}_{45}$. The mean function hexiles are straightforward to attain, however, the entries of the variance covariance matrix require additional effort. The derivations of the distributions of these entries were given in the previous section. The accuracy assessments that follow are based on MCMC samples using BUGS iterations of size 5,000. A thinning factor of 5 was applied to post burn-in samples of size 5,000. This resulted in MCMC samples of size 1,000 for density estimation. The density estimates were obtained using the binned kernel density estimate `bkde()` function in the R package `Kernsmooth` (Wand & Ripley, 2009). The MFVB iterations were terminated when the increase in $\log \underline{p}(\boldsymbol{y}; q)$ fell below 10^{-7}. Figure 2.4 summarizes the accuracy results whilst Figure 2.5 plots the variational Bayes and MCMC approximate posteriors for the first simulated dataset. The mean function at each hexile and each entry of $\boldsymbol{\Sigma}$ shows high accuracy, with most accuracy levels above 85%.

Another crucial type of accuracy assessment is the comparison between the advertised coverage of variational Bayes approximate credible intervals and the true coverage. Table 2.1 shows the percentages of the true parameter coverage for the approximate 95% credible intervals formed from the variational Bayes posterior densities with 0.025 probability mass in each tail. The coverage overall is very good and does not fall below 93%.

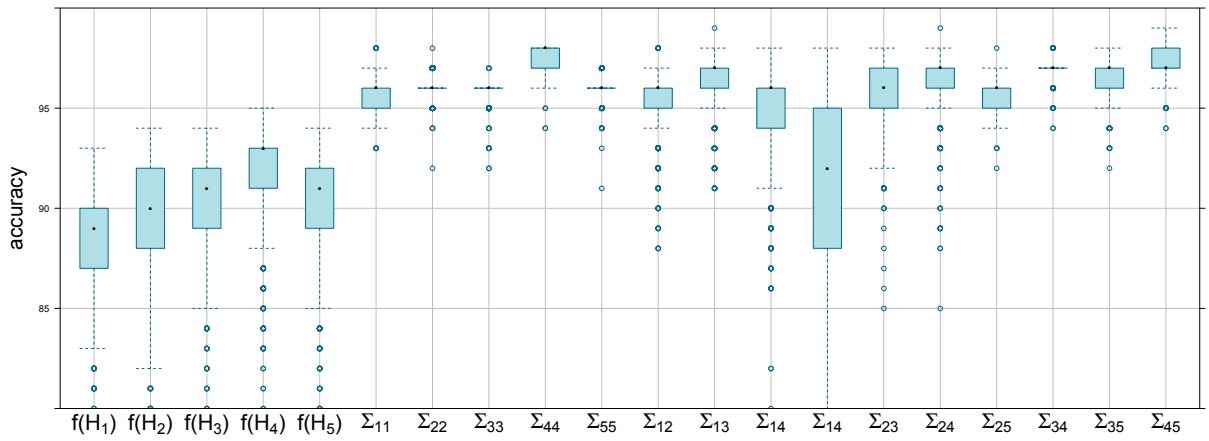


Figure 2.4: *Summary of the simulation for the marginal longitudinal nonparametric regression model (2.1), where $f(H_i)$ is the mean function at hexile i. For the mean function at each hexile and each entry of $\boldsymbol{\Sigma}$, the accuracy values are summarized as a boxplot.*

$\boldsymbol{\Sigma}_{11}$	$\boldsymbol{\Sigma}_{22}$	$\boldsymbol{\Sigma}_{33}$	$\boldsymbol{\Sigma}_{44}$	$\boldsymbol{\Sigma}_{55}$	$\boldsymbol{\Sigma}_{12}$	$\boldsymbol{\Sigma}_{13}$	$\boldsymbol{\Sigma}_{14}$	$\boldsymbol{\Sigma}_{15}$	$\boldsymbol{\Sigma}_{23}$
95%	95%	96%	94%	95%	95%	94%	94%	95%	95%
$\boldsymbol{\Sigma}_{24}$	$\boldsymbol{\Sigma}_{25}$	$\boldsymbol{\Sigma}_{34}$	$\boldsymbol{\Sigma}_{35}$	$\boldsymbol{\Sigma}_{45}$	$f(H_1)$	$f(H_2)$	$f(H_3)$	$f(H_4)$	$f(H_5)$
95%	95%	94%	95%	94%	95%	95%	95%	93%	94%

Table 2.1: *Percentage coverage of true parameter values by approximate 95% credible intervals based on variational Bayes approximate posterior density functions. The percentages are based on 1000 replications.*

2.4.1 Assessment of speed

To assess the gain in time savings by using MFVB, we monitored the time taken to fit each model based on a different replication in the simulation study. The computations for this study were performed on a laptop computer (Mac OS X; 2.8 GHz processor, 16 GBytes random access memory). The computation times for each approach are summarised in Table 2.2.

MCMC	MFVB
282.555 (4.506)	3.064 (0.278)

Table 2.2: *Average (standard deviation) times in seconds for the simulation study comparing MCMC and MFVB fitting of the model in (2.20).*

The average computing time for MFVB is about 3 seconds which is substantially faster

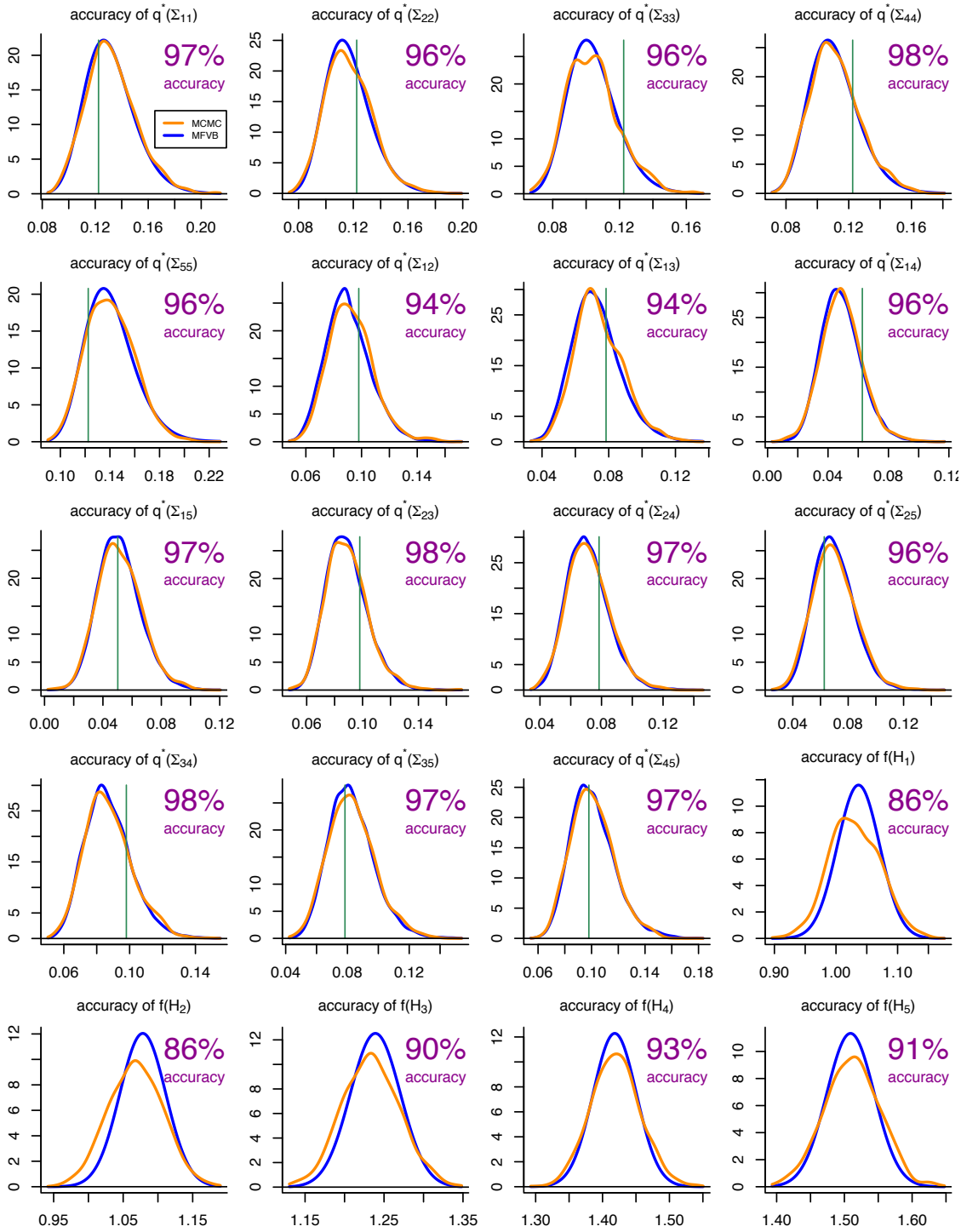


Figure 2.5: *Variational Bayes and MCMC approximate posterior densities for the first simulated data set.*

than the 4.7 minutes taken by MCMC i.e., 94 times faster. It is evident that the speed gains achieved by MFVB needs to be traded off against the accuracy losses which are incurred by restriction (2.16).

Making time comparisons completely fair between MFVB and MCMC is difficult, since convergence for each approximation method is different. The number of iterations used for MFVB and the number of samples used for MCMC was sufficient for convergence in this particular simulation study setting. However, these values will change depending on the type of application.

2.5 Application

The Bayesian additive/interaction model (2.10) was fitted to data from a nutritional epidemiology study (source: Kipnis et al., 2003), which was presented in Section 5.2 of Al Kadiri *et al.* (2010). Convergence for both MCMC and MFVB was assessed in the same way as given in the previous section. Keeping with the notation of (2.10), the variables are:

y = logarithm of intake of protein as measured by the biomarker urinary nitrogen,
x_1 = body mass index,
x_2 = logarithm of intake of protein as measured by a 24-hour recall instrument,
x_3 = gender.

Figure 2.6 shows the MCMC and MFVB estimates of f_{male}, f_{female} and f_2 and the corresponding 95% pointwise credible sets. The MCMC fit took about 1 hour and 5 minutes,

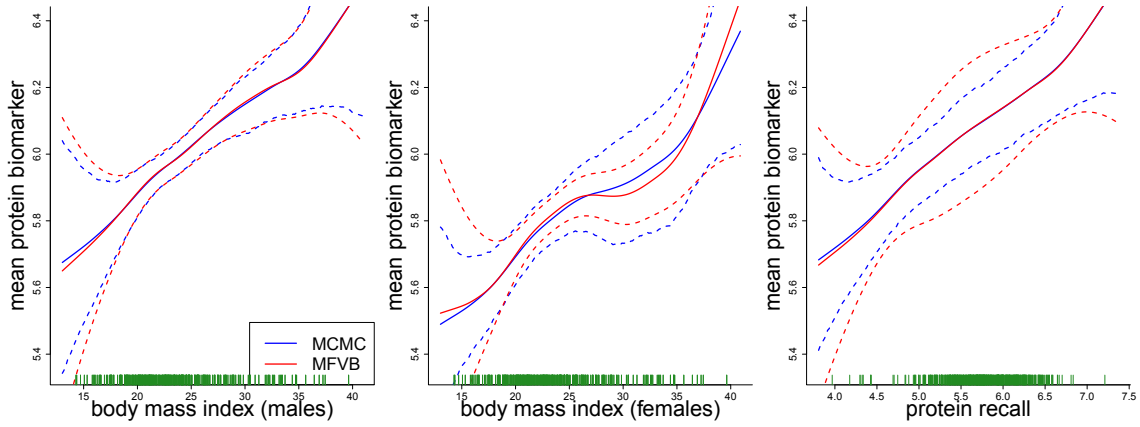


Figure 2.6: *Estimated functions for the additive/interaction model for the nutritional epidemiology data using MCMC and MFVB. The dashed curves correspond to 95% pointwise credible sets.*

whilst the MFVB fit took 20 seconds.

2.6 Discussion

The aim of this chapter was to develop a fast deterministic approach to MCMC to shorten estimation time. We also sought to address the marginal longitudinal regression problem and some of its semiparametric extensions. As a result we have derived a MFVB algorithm for fast approximate inference in semiparametric regression. The central finding here is that, for using the marginal longitudinal semiparametric regression model, MFVB achieves excellent accuracy when compared to MCMC for the main parameters of interest. In addition, the algorithm is one of the first variational algorithms to involve estimation of an unstructured covariance matrix. The real data example presented in Al Kadiri *et al.* (2010) takes almost an hour for MCMC to run on a contemporary laptop using BUGS. The MFVB algorithm developed in this chapter, fits the same data in seconds with very similar results. The simulation study shows evidence of significant speed and accuracy attributes of MFVB against the MCMC benchmark.

2.A Derivation of algorithm 1

Algorithm 1 depends on the following derivations of the optimal density functions. Constants with respect to the function argument are denoted by 'const'. The MFVB calculations rely primarily on the following expressions for the full conditional distributions:

$$\begin{aligned}
\log p\left(\boldsymbol{\beta},\boldsymbol{u}|\text{rest}\right) &= -\tfrac{1}{2}\left(\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}^{\top}\left(\boldsymbol{C}^{\top}\left(\boldsymbol{I}_m\otimes\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{C}+\begin{bmatrix}\mathbf{F}^{-1} & 0\\ 0 & \underset{1\le\ell\le d}{\text{blockdiag}}\left(\sigma_\ell^2\boldsymbol{I}_{K_\ell}\right)\end{bmatrix}\right)\right.\\
&\qquad\left.\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}-2\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}^{\top}\boldsymbol{C}^{\top}\left(\boldsymbol{I}_m\otimes\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{y}\right)+\text{const},\\
\log\ p\left(\sigma_\ell^2|\text{rest}\right) &= \left(\tfrac{1}{2}\left(K_\ell+1\right)-1\right)\log\left(\sigma_\ell^2\right)-\left(\tfrac{1}{2}\|\boldsymbol{u}_\ell\|^2+a_\ell^{-1}\right)/\sigma_\ell^2+\text{const},\\
\log\ p\left(a_\ell|\text{rest}\right) &= -2\log\left(a_\ell\right)-\left(\sigma_\ell^2+A^{-2}\right)/a_\ell+\text{const},\\
\log\ p\left(\boldsymbol{\Sigma}|\text{rest}\right) &= -\frac{A_{\boldsymbol{\Sigma}}+m+n+1}{2}\log|\boldsymbol{\Sigma}|\\
&\quad -\tfrac{1}{2}\text{tr}\left\{\left[\sum_{i=1}^{m}\left(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}-\boldsymbol{Z}_i\boldsymbol{u}\right)\left(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}-\boldsymbol{Z}_i\boldsymbol{u}\right)^{\top}+B_{\boldsymbol{\Sigma}}\right]\boldsymbol{\Sigma}^{-1}\right\}+\text{const}.
\end{aligned}$$

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

We begin with

$$\begin{aligned}
\log\ q^*\left(\boldsymbol{\beta},\boldsymbol{u}\right) &= E_q\left\{\log p\left(\boldsymbol{\beta},\boldsymbol{u}|\text{rest}\right)\right\}+\text{const}\\
&= -\tfrac{1}{2}\left(\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}^{\top}\left(\boldsymbol{C}^{\top}\left(\boldsymbol{I}_m\otimes M_{q(\boldsymbol{\Sigma}^{-1})}\right)\boldsymbol{C}\right.\right.\\
&\qquad\left.+\begin{bmatrix}\mathbf{F}^{-1} & \mathbf{0}\\ \mathbf{0} & \underset{1\le\ell\le d}{\text{blockdiag}}\left(\mu_{q(1/\sigma_\ell^2)}\boldsymbol{I}_{K_\ell}\right)\end{bmatrix}\right)\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}\\
&\qquad\left.-2\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}^{\top}\boldsymbol{C}^{\top}\left(\boldsymbol{I}_m\otimes M_{q(\boldsymbol{\Sigma}^{-1})}\right)\boldsymbol{y}\right)+\text{const}.
\end{aligned}$$

Therefore,

$$q^*\left(\boldsymbol{\beta},\boldsymbol{u}\right)\ \text{is the}\ N\left(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})},\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\right)\ \text{density function},$$

where

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}=\left(\boldsymbol{C}^{\top}\left(I_m\otimes M_{q(\boldsymbol{\Sigma}^{-1})}\right)\boldsymbol{C}+\begin{bmatrix}\mathbf{F}^{-1} & \mathbf{0}\\ \mathbf{0} & \underset{1\le\ell\le d}{\text{blockdiag}}\left(\mu_{q(1/\sigma_\ell^2)}\boldsymbol{I}_{K_\ell}\right)\end{bmatrix}\right)^{-1}$$

and

$$\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} = \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \boldsymbol{C}^{\top} \left(\boldsymbol{I}_m \otimes M_{q\left(\boldsymbol{\Sigma}^{-1}\right)} \right) \boldsymbol{y}.$$

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

Next, we have

$$\begin{aligned} \log\, q^*\left(\sigma_\ell^2\right) &= E_q\left\{\log\, p\left(\sigma_\ell^2 | \text{rest}\right)\right\} + \text{const} \\ &= \left(-\tfrac{1}{2}\left(K_\ell + 1\right) - 1\right) \log\left(\sigma_\ell^2\right) - \left(\tfrac{1}{2} E_q \|\boldsymbol{u}_\ell\|^2 + \mu_{q(1/a_\ell)}\right) / \sigma_\ell^2 + \text{const}. \end{aligned}$$

Therefore,

$$q^*\left(\sigma_\ell^2\right) \text{ is the Inverse-Gamma} \left(\tfrac{1}{2}\left(K_\ell + 1\right), B_{q(\sigma_\ell^2)} \right) \text{ density function,}$$

where

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

In addition, using Result 1.4.3, this gives

$$\mu_{q(1/\sigma_\ell^2)} = \tfrac{1}{2}\left(K_\ell + 1\right) / B_{q(\sigma_\ell^2)}.$$

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

Similarly,

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

Therefore

$$q^*\left(a_\ell\right) \text{ is the Inverse-Gamma}\left(1, B_{q(a_\ell)}\right) \text{ density function,}$$

where

$$B_{q(a_\ell)} = \mu_{q(1/\sigma_\ell^2)} + A^{-2}.$$

Also, making use of Result 1.4.3, we get

$$\mu_{q(1/a_\ell)} = 1 / B_{q(a_\ell)}.$$

Expressions for $B_{q(\boldsymbol{\Sigma})}$ and $M_{q(\boldsymbol{\Sigma}^{-1})}$

Finally,

$$\begin{aligned}
\log\, q^{*}(\boldsymbol{\Sigma}) &= E_q\{\log\, p(\boldsymbol{\Sigma}|\text{rest})\} + \text{const} \\
&= -\tfrac{1}{2}(A_{\boldsymbol{\Sigma}} + m + n + 1)\log|\boldsymbol{\Sigma}| \\
&\quad -\tfrac{1}{2}\text{tr}\left[\sum_{i=1}^{m} E_{q(\boldsymbol{\beta},\boldsymbol{u})}\left\{(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{u})(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{u})^{\top} + B_{\boldsymbol{\Sigma}}\right\}\boldsymbol{\Sigma}^{-1}\right] \\
&\quad +\text{const} \\
&= -\tfrac{1}{2}(A_{\boldsymbol{\Sigma}} + m + n + 1)\log|\boldsymbol{\Sigma}| \\
&\quad -\tfrac{1}{2}\text{tr}\left(\left[\sum_{i=1}^{m}\left\{\left(\boldsymbol{y}_i - \boldsymbol{C}_i\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\right)\left(\boldsymbol{y}_i - \boldsymbol{C}_i\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\right)^{\top} + \boldsymbol{C}_i\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}_i^{T}\right\}\right.\right. \\
&\quad \left.\left. +B_{\boldsymbol{\Sigma}}\right]\boldsymbol{\Sigma}^{-1}\right) + \text{const}.
\end{aligned}$$

Therefore,

$$q^{*}(\boldsymbol{\Sigma}) \text{ is the Inverse-Wishart}\left(A_{\boldsymbol{\Sigma}} + m,\ B_{q(\boldsymbol{\Sigma})}\right) \text{ density function,}$$

where

$$B_{q(\boldsymbol{\Sigma})} = B_{\boldsymbol{\Sigma}} + \sum_{i=1}^{m}\left[\left(\boldsymbol{y}_i - \boldsymbol{C}_i\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\right)\left(\boldsymbol{y}_i - \boldsymbol{C}_i\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\right)^{\top} + \boldsymbol{C}_i\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}_i^{\top}\right].$$

Lastly, making use of Result 1.4.7 gives

$$M_{q(\boldsymbol{\Sigma}^{-1})} = (A_{\boldsymbol{\Sigma}} + m)\, B_{q(\boldsymbol{\Sigma})}^{-1}.$$

2.B Derivation for the marginal log-likelihood lower bound

The expression for the lower bound on the marginal log-likelihood given at (2.17) is

$$\begin{aligned}\log\, \underline{p}\,(\boldsymbol{y};q) \;=\;& \tfrac{1}{2}\left(\sum_{\ell=1}^{d} K_\ell + p\right) - \tfrac{m}{2}\log(2\pi) - \log(\pi) + \sum_{\ell=1}^{d}\log\,\Gamma\left(\tfrac{1}{2}\left(K_\ell+1\right)\right)\\ &-\tfrac{p}{2}\log\left(\sigma^2_{\boldsymbol{\beta}}\right) - \log(A) - \tfrac{1}{2\sigma^2_{\boldsymbol{\beta}}}\left\{\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \,\mathrm{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right)\right\} + \tfrac{1}{2}\,\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}|\\ &-\tfrac{1}{2}\sum_{\ell=1}^{d}\left(K_\ell+1\right)\,\log\left(B_{q\left(\sigma^2_\ell\right)}\right) - \tfrac{1}{2}\left(A_{\boldsymbol{\Sigma}}+m\right)\,\log|B_{q(\boldsymbol{\Sigma})}| + \sum_{\ell=1}^{d}\mu_{q\left(1/\sigma^2_\ell\right)}\mu_{q(1/a_\ell)}\\ &-\sum_{\ell=1}^{d}\log\left(\mu_{q\left(1/\sigma^2_\ell\right)} + A^{-2}\right) + \tfrac{1}{2}\,A_{\boldsymbol{\Sigma}}\left(\log|B_\ell|\right) - \log\left(C_{n,A}\right) + \log\left(C_{n,A+m}\right).\end{aligned}$$

Derivation: The lower bound on the marginal log-likelihood is achieved through the following expression:

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

First we note that

$$\begin{aligned}\log\, p\left(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{\Sigma}\right) \;=\;& -\tfrac{m}{2}\,\log(2\pi) - \tfrac{m}{2}\,log|\boldsymbol{\Sigma}|\\ &-\tfrac{1}{2}\,\mathrm{tr}\{\left(\boldsymbol{I}_m\otimes\boldsymbol{\Sigma}\right)^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{u}\right)\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{u}\right)^{\top}\}.\end{aligned}$$

Therefore,

$$\begin{aligned}E_q\{\log\, p\left(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{\Sigma}\right)\} \;=\;& -\tfrac{m}{2}\,\log(2\pi) - \tfrac{m}{2}\,E_q\left(\log|\boldsymbol{\Sigma}|\right)\\ &-\tfrac{1}{2}\,\mathrm{tr}\left\{\left(\boldsymbol{I}_m\otimes M_{q\left(\boldsymbol{\Sigma}^{-1}\right)}\right)\left[\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},u)}\boldsymbol{C}^{\top}\right.\right.\\ &\left.\left.+\left(\boldsymbol{y}-\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\right)\left(\boldsymbol{y}-\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\right)^{\top}\right]\right\}.\end{aligned}$$

Next,

$$\begin{aligned}&E_q\{\log\, p\left(\boldsymbol{\beta},\boldsymbol{u}\,|\,\boldsymbol{\sigma}^2\right) - \log\, q^*\left(\boldsymbol{\beta},\boldsymbol{u}\right)\}\\ &\quad= -\tfrac{p}{2}\log\left(\sigma^2_{\boldsymbol{\beta}}\right) - \tfrac{1}{2}\sum_{\ell=1}^{d}K_\ell\,E_q\left\{\log\left(\sigma^2_\ell\right)\right\} - \tfrac{1}{2}\sum_{\ell=1}^{d}E_q\left(1/\sigma^2_\ell\right)E_q\left(\|\boldsymbol{u}_\ell\|^2\right) + \tfrac{1}{2}\,\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}|\\ &\quad\quad+\tfrac{1}{2}\left(\sum_{\ell=1}^{d}K_\ell + p\right)\end{aligned}$$

$$
\begin{aligned}
= \; & -\tfrac{1}{2} p \log\left(\sigma_\beta^2\right) - \tfrac{1}{2} \sum_{\ell=1}^{d} K_\ell \; E_q\left(\log\left(\sigma_\ell^2\right)\right) - \tfrac{1}{2\sigma_\beta^2}\left\{\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \operatorname{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right)\right\} \\
& -\tfrac{1}{2} \sum_{\ell=1}^{d} \mu_{q(1/\sigma_\ell^2)}\left\{\|\boldsymbol{\mu}_{q(\boldsymbol{u}_\ell)}\|^2 + \operatorname{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_\ell)}\right)\right\} + \tfrac{1}{2} \; \log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}| + \tfrac{1}{2}\left(\sum_{\ell=1}^{d} K_\ell + p\right).
\end{aligned}
$$

Next,

$$
\begin{aligned}
& \log \; p\left(\boldsymbol{\sigma}^2 \,|\, \boldsymbol{a}\right) - \log \; q^*\left(\boldsymbol{\sigma}^2\right) \\
& \quad = \log\left[\prod_{\ell=1}^{d} \frac{a_\ell^{\frac{1}{2}}}{\Gamma\left(\frac{1}{2}\right)} \left(\sigma_\ell^2\right)^{-\frac{1}{2}-1} \exp\left(-\frac{a_\ell^{\frac{1}{2}}}{\sigma_\ell^2}\right)\right] \\
& \qquad - \log\left[\prod_{\ell=1}^{d} \frac{B_{q\left(\sigma_\ell^2\right)}^{\frac{1}{2}(K_\ell+1)}}{\Gamma\left(\frac{1}{2}\left(K_\ell+1\right)\right)} \left(\sigma_\ell^2\right)^{-\frac{1}{2}(K_\ell+1)-1} \exp\left(\frac{-B_{q\left(\sigma_\ell^2\right)}}{\sigma_\ell^2}\right)\right] \\
& \quad = -\tfrac{1}{2}\sum_{\ell=1}^{d} \log\left(a_\ell\right) - \log \; \Gamma\left(\tfrac{1}{2}\right) - \tfrac{3}{2}\sum_{\ell=1}^{d} \log\left(\sigma_\ell^2\right) - \sum_{\ell=1}^{d} a_\ell^{-1}\left(1/\sigma_\ell^2\right) + \sum_{\ell=1}^{d} B_{q\left(\sigma_\ell^2\right)}/\sigma_\ell^2 \\
& \qquad + \sum_{\ell=1}^{d}\left(\tfrac{1}{2}\left(K_\ell+1\right)-1\right)\log\left(\sigma_\ell^2\right) + \sum_{\ell=1}^{d} \log \; \Gamma\left(\tfrac{1}{2}\left(K_\ell+1\right)\right) - \tfrac{1}{2}\sum_{\ell=1}^{d}\left(K_\ell+1\right)\log\left(B_{q\left(\sigma_\ell^2\right)}\right) \\
& \quad = -\tfrac{1}{2}\sum_{\ell=1}^{d} \log\left(a_\ell\right) + \sum_{\ell=1}^{d} \log \; \Gamma\left(\tfrac{1}{2}\left(K_\ell+1\right)\right) - \tfrac{1}{2}\log\left(\pi\right) + \sum_{\ell=1}^{d}\left(B_{q\left(\sigma_\ell^2\right)} - a_\ell^{-1}\right)\left(1/\sigma_\ell^2\right) \\
& \qquad + \tfrac{1}{2}\sum_{\ell=1}^{d} K_\ell \; \log\left(\sigma_\ell^2\right) - \tfrac{1}{2}\sum_{\ell=1}^{d}\left(K_\ell+1\right)\log\left(B_{q\left(\sigma_\ell^2\right)}\right).
\end{aligned}
$$

This means that

$$
\begin{aligned}
E_q\left\{\log \; p\left(\boldsymbol{\sigma}^2 \,|\, \boldsymbol{a}\right) - \log \; q^*\left(\boldsymbol{\sigma}^2\right)\right\} \quad = \quad & -\tfrac{1}{2}\sum_{\ell=1}^{d} E_q\left\{\log\left(a_\ell\right)\right\} + \sum_{\ell=1}^{d} \log \; \Gamma\left\{\tfrac{1}{2}\left(K_\ell+1\right)\right\} \\
& -\tfrac{1}{2}\log\left(\pi\right) + \sum_{\ell=1}^{d}\left(B_{q\left(\sigma_\ell^2\right)} - \mu_{q(1/a_\ell)}\right)\mu_{q\left(1/\sigma_\ell^2\right)} \\
& + \tfrac{1}{2}\sum_{\ell=1}^{d} K_\ell \; E_q\left\{\log\left(\sigma_\ell^2\right)\right\} - \tfrac{1}{2}\sum_{\ell=1}^{d}\left(K_\ell+1\right)\log\left(B_{q\left(\sigma_\ell^2\right)}\right).
\end{aligned}
$$

In addition,

$$
\begin{aligned}
& \log \; p\left(\boldsymbol{a}\right) - \log \; q^*\left(\boldsymbol{a}\right) \\
& \quad = -\log\left(A\right) - \tfrac{1}{2}\log\left(\pi\right) + \tfrac{1}{2}\sum_{\ell=1}^{d} \log\left(\boldsymbol{a}_\ell\right) - \sum_{\ell=1}^{d} \log\left(B_{q(a_\ell)}\right) + \sum_{\ell=1}^{d}\left(B_{q(a_\ell)} - A^{-2}\right)\left(1/a_\ell\right) \\
& \quad = -\log\left(A\right) - \tfrac{1}{2}\log\left(\pi\right) + \tfrac{1}{2}\sum_{\ell=1}^{d} \log\left(a_\ell\right) - \sum_{\ell=1}^{d} \log\left(\mu_{q\left(1/\sigma_\ell^2\right)} + A^{-2}\right) + \sum_{\ell=1}^{d} \mu_{q\left(1/\sigma_\ell^2\right)}/a_\ell
\end{aligned}
$$

and so

$$\begin{aligned} E_q\left\{\log\ p\left(\boldsymbol{a}\right)-\log\ q^*\left(\boldsymbol{a}\right)\right\} &= -\log\left(A\right)-\tfrac{1}{2}\log\left(\pi\right)+\tfrac{1}{2}\sum_{\ell=1}^{d}E_q\left(\log\left(\boldsymbol{a}_\ell\right)\right) \\ &\quad -\sum_{\ell=1}^{d}\log\left(\mu_{q(1/\sigma_\ell^2)}+A^{-2}\right)+\sum_{\ell=1}^{d}\mu_{q(1/\sigma_\ell^2)}\mu_{q(1/a_\ell)}. \end{aligned}$$

Next,

$$\begin{aligned} \log\ p\left(\boldsymbol{\Sigma}\right)-\log\ q^*\left(\boldsymbol{\Sigma}\right) &= -\log\left(C_{n,A}\right)+\log\left(C_{n,A+m}\right)+\tfrac{m}{2}\log|\boldsymbol{\Sigma}|+\tfrac{1}{2}A\log|B| \\ &\quad -\tfrac{1}{2}(A+m)\log|B_{q(\boldsymbol{\Sigma})}|-\tfrac{1}{2}\mathrm{tr}\left(B\boldsymbol{\Sigma}^{-1}\right)+\tfrac{1}{2}\mathrm{tr}\left(B_{q(\boldsymbol{\Sigma})}\boldsymbol{\Sigma}^{-1}\right). \end{aligned}$$

Hence,

$$\begin{aligned} E_q\left\{\log\ p\left(\boldsymbol{\Sigma}\right)-\log\ q^*\left(\boldsymbol{\Sigma}\right)\right\} &= -\log\left(C_{n,A}\right)+\log\left(C_{n,A+m}\right)+\tfrac{m}{2}E_q\left(\log|\boldsymbol{\Sigma}|\right)+\tfrac{1}{2}A\log|B| \\ &\quad -\tfrac{1}{2}(A+m)\log|B_{q(\boldsymbol{\Sigma})}|+\tfrac{1}{2}\mathrm{tr}\left(\left(B_{q(\boldsymbol{\Sigma})}-B\right)M_{q\left(\boldsymbol{\Sigma}^{-1}\right)}\right). \end{aligned}$$

Next, note that

$$\begin{aligned} \sum_{\ell=1}^{d}\left(B_{q(\sigma_\ell^2)}-\mu_{q(1/a_\ell)}\right)\mu_{q(1/\sigma_\ell^2)} &= \sum_{\ell=1}^{d}\left[\tfrac{1}{2}\left\{\|\boldsymbol{\mu}_{q(\boldsymbol{u}_\ell)}\|^2+\mathrm{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_\ell)}\right)\right\}+\mu_{q(1/a_\ell)}\right. \\ &\quad \left.-\mu_{q(1/a_\ell)}\right]\mu_{q(1/\sigma_\ell^2)} \\ &= \tfrac{1}{2}\sum_{\ell=1}^{d}\mu_{q(1/\sigma_\ell^2)}\left(\|\boldsymbol{\mu}_{q(u_\ell)}\|^2+\mathrm{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_\ell)}\right)\right). \end{aligned}$$

We also note that

$$\begin{aligned} &\tfrac{1}{2}\ \mathrm{tr}\left\{\left(B_{q(\boldsymbol{\Sigma})}-B\right)M_{q\left(\boldsymbol{\Sigma}^{-1}\right)}\right\} \\ &\quad =\tfrac{1}{2}\ \mathrm{tr}\left(\left[\sum_{i=1}^{m}\left\{\left(\boldsymbol{y}_i-\boldsymbol{C}_i\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\right)\left(\boldsymbol{y}_i-\boldsymbol{C}_i\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\right)^\top+\boldsymbol{C}_i\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}_i^\top\right\}\right]\boldsymbol{M}_{q\left(\boldsymbol{\Sigma}^{-1}\right)}\right) \\ &\quad =\tfrac{1}{2}\ \mathrm{tr}\left[\left(\boldsymbol{I}_m\otimes\boldsymbol{M}_{q\left(\boldsymbol{\Sigma}^{-1}\right)}\right)\left\{\left(\boldsymbol{y}-\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\right)\left(\boldsymbol{y}-\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\right)^\top+\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^\top\right\}\right] \end{aligned}$$

The cancellations lead by these equalities leads to the lower bound expression given in (2.17).

2.C Derivation of the Fisher information matrix for the linear marginal longitudinal model

We derive the Fisher information matrix for the frequentist Gaussian marginal longitudinal regression model:

$$\boldsymbol{y} \sim \mathrm{N}\left(\boldsymbol{X}\boldsymbol{\beta},\, \boldsymbol{I}_m \otimes \boldsymbol{\Sigma}\right).$$

The log-likelihood of $(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ is

$$\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = -\tfrac{nm}{2}\log(2\pi) - \tfrac{m}{2}\log|\boldsymbol{\Sigma}| - \tfrac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{I}_m \otimes \boldsymbol{\Sigma}^{-1})(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

To begin, with the help of Result 1.4.1 (g), we take the first differential of $\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ with respect to $\boldsymbol{\beta}$:

$$\begin{aligned} d_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma}) &= \tfrac{1}{2}\left(-\boldsymbol{X}d\boldsymbol{\beta}\right)^\top(\boldsymbol{I}_m \otimes \boldsymbol{\Sigma}^{-1})(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) - \tfrac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{I}_m \otimes \boldsymbol{\Sigma}^{-1})\boldsymbol{X}d\boldsymbol{\beta} \\ &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{I}_m \otimes \boldsymbol{\Sigma}^{-1})\boldsymbol{X}d\boldsymbol{\beta} \end{aligned}$$

and then the second differential of $\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ with respect to $\boldsymbol{\beta}$ is

$$\begin{aligned} d^2_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma}) &= \left(-\boldsymbol{X}d\boldsymbol{\beta}\right)^\top(\boldsymbol{I}_m \otimes \boldsymbol{\Sigma}^{-1})\boldsymbol{X}d\boldsymbol{\beta} \\ &= -\left(d\boldsymbol{\beta}\right)^\top\boldsymbol{X}^\top(\boldsymbol{I}_m \otimes \boldsymbol{\Sigma}^{-1})\boldsymbol{X}\left(d\boldsymbol{\beta}\right), \end{aligned}$$

so, from Theorem 1.4.2, we see that

$$\mathsf{H}_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = -\boldsymbol{X}^\top(\boldsymbol{I}_m \otimes \boldsymbol{\Sigma}^{-1})\boldsymbol{X},$$

and so, the $\boldsymbol{\beta}$ block of $\boldsymbol{I}(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ is

$$-\mathrm{E}\left\{\mathsf{H}_{\boldsymbol{\beta}}\,\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma})\right\} = \boldsymbol{X}^\top(\boldsymbol{I}_m \otimes \boldsymbol{\Sigma}^{-1})\boldsymbol{X}.$$

In order to check that the maximum likelihood estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are indeed uncorrelated, we next deal with

$$d_{\mathrm{vec}(\boldsymbol{\Sigma})}\left\{d_{\boldsymbol{\beta}}\,\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma})\right\} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top d\left\{(\boldsymbol{I}_m \otimes \boldsymbol{\Sigma}^{-1})\right\}d\boldsymbol{\beta}.$$

However, we see that

$$\mathrm{E}\left[d_{\mathrm{vec}(\boldsymbol{\Sigma})}\left\{d_{\boldsymbol{\beta}}\,\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma})\right\}\right] = 0.$$

It then follows that $\boldsymbol{I}(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ is of block-diagonal form. We next move on to find the $\boldsymbol{\Sigma}$ block of the information matrix.

$$d_{\mathrm{vec}(\boldsymbol{\Sigma})}\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = -\tfrac{m}{2} d\left(\log|\boldsymbol{\Sigma}|\right) - \tfrac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top d(\boldsymbol{I}_m \otimes \boldsymbol{\Sigma}^{-1})(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

Then, using Results 1.4.1 (c), (e) and (f), we get

$$d_{\mathrm{vec}(\boldsymbol{\Sigma})}\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = -\tfrac{m}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1} d\boldsymbol{\Sigma}\right) + \tfrac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \left\{\boldsymbol{I}_m \otimes \left(\boldsymbol{\Sigma}^{-1} d\boldsymbol{\Sigma}\, \boldsymbol{\Sigma}^{-1}\right)\right\}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

and so

$$\begin{aligned} d^2_{\mathrm{vec}(\boldsymbol{\Sigma})}\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma}) &= \tfrac{m}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}(d\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}(d\boldsymbol{\Sigma})\right\} \\ &\quad -\tfrac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \left\{\boldsymbol{I}_m \otimes \left(\boldsymbol{\Sigma}^{-1}(d\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}(d\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\right)\right\}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &\quad -\tfrac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \left\{\boldsymbol{I}_m \otimes \left(\boldsymbol{\Sigma}^{-1}(d\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}(d\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\right)\right\}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \tfrac{m}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}(d\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}(d\boldsymbol{\Sigma})\right\} \\ &\quad -(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \left\{\boldsymbol{I}_m \otimes \left(\boldsymbol{\Sigma}^{-1}(d\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}(d\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\right)\right\}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \end{aligned}$$

With the help of Results 1.4.18, 1.4.14 and 1.4.15, we see that

$$\begin{aligned} -\mathrm{E}\left\{d^2_{\mathrm{vec}(\boldsymbol{\Sigma})}\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma})\right\} &= -\tfrac{m}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}(d\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}(d\boldsymbol{\Sigma})\right\} \\ &\quad +\mathrm{tr}\left[\boldsymbol{I}_m \otimes \left\{\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}(d\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}(d\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\right\}\right] \\ &= -\tfrac{m}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}(d\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}(d\boldsymbol{\Sigma})\right\} \\ &\quad +m\,\mathrm{tr}\left\{\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}(d\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}(d\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\right\} \\ &= \tfrac{m}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}(d\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}(d\boldsymbol{\Sigma})\right\}. \end{aligned}$$

Then, using Result 1.4.16, we get

$$-\mathrm{E}\left\{d^2_{\mathrm{vec}(\boldsymbol{\Sigma})}\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma})\right\} = \tfrac{m}{2}\mathrm{vec}(d\boldsymbol{\Sigma})^\top(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathrm{vec}(d\boldsymbol{\Sigma}).$$

Therefore,

$$\mathsf{H}_{\mathrm{vec}(\boldsymbol{\Sigma})}\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = -\tfrac{m}{2}(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}),$$

and the full information matrix is:

$$\boldsymbol{I}(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \begin{bmatrix} \boldsymbol{X}^\top(\boldsymbol{I}_m \otimes \boldsymbol{\Sigma}^{-1})\boldsymbol{X} & \boldsymbol{0} \\ \boldsymbol{0} & \frac{m}{2}(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}) \end{bmatrix}.$$

2.D Proof of Theorem 2.3.1

Theorem 3.4.7 (a) of Mardia *et al.* (1979) states that if

$$\boldsymbol{X} \sim \text{Wishart}\left(a, \boldsymbol{B}\right), \quad a > n, \quad \boldsymbol{B} \text{ positive definite}, \tag{2.21}$$

then the ratio $\boldsymbol{c}^\top \boldsymbol{B}\boldsymbol{c}/\boldsymbol{c}^\top \boldsymbol{X}^{-1}\boldsymbol{c}$ has a ${\chi_{a-n+1}}^2$ distribution for any fixed $n \times 1$ vector $\boldsymbol{c}$. If we let $\boldsymbol{c} = [0 \;\; 1 \;\; 0]^\top$, such that $\boldsymbol{B}_{22} = \boldsymbol{c}^\top \boldsymbol{B}\boldsymbol{c}$ and $\left(\boldsymbol{X}^{-1}\right)_{22} = \boldsymbol{c}^\top \boldsymbol{X}^{-1}\boldsymbol{c}$. Then,

$$\frac{\boldsymbol{B}_{22}}{\left(\boldsymbol{X}^{-1}\right)_{22}} \sim \chi^2_{a-3+1}$$

and making use of Result 1.4.5, this is equivalent to

$$\frac{\boldsymbol{B}_{22}}{\left(\boldsymbol{X}^{-1}\right)_{22}} \sim \text{Gamma}\left(\frac{a-3+1}{2}, \frac{1}{2}\right).$$

Using Result 1.4.4 we see that:

$$\frac{\left(\boldsymbol{X}^{-1}\right)_{22}}{\boldsymbol{B}_{22}} \sim \text{Inverse-Gamma}\left(\frac{a-3+1}{2}, \frac{1}{2}\right).$$

Therefore using Results 1.4.2 and 1.4.8, the $[2,2]$ entry of an Inverse-Wishart random matrix $\boldsymbol{X}^{-1}$ has distribution

$$\left(\boldsymbol{X}^{-1}\right)_{22} \sim \text{Inverse-Gamma}\left(\frac{a-3+1}{2}, \tfrac{1}{2}\boldsymbol{B}_{22}\right).$$

The distribution of the $[1,1]$ and $[3,3]$ entries of $\boldsymbol{X}^{-1}$ have similar forms. We have shown this for a 3×3 random matrix, however this applies for general $n \times n$ random matrices also. Thus, for a general $n \times n$ positive definite matrix $\boldsymbol{X}$ that is distributed

$$\boldsymbol{X} \sim \text{Inverse-Wishart}(a, \boldsymbol{B}),$$

its diagonal entries have the following distribution:

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

2.E Proof of Theorem 2.3.2

The characteristic function of a Wishart random matrix $\boldsymbol{X}$ with distribution

$$\boldsymbol{X} \sim \text{Inverse-Wishart}\,(a, \boldsymbol{B})$$

is

$$E\left[\exp\left\{i\,\text{tr}\,(\boldsymbol{XT})\right\}\right] = |\boldsymbol{I} - 2i\boldsymbol{TB}^{-1}|^{-a/2}.$$

If we let $\boldsymbol{X}$ be of dimension 3×3 and focus on finding the characteristic function of the $[1,2]$ and $[2,1]$ off-diagonal entries of $\boldsymbol{X}$, we set

$$\boldsymbol{T} = \begin{bmatrix} 0 & t & 0 \\ t & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

so that $\text{tr}\,(\boldsymbol{XT}) = 2t\boldsymbol{X}_{12}$. Therefore the characteristic function becomes

$$\begin{aligned} E\left\{\exp\left(2it\boldsymbol{X}_{12}\right)\right\} &= |\boldsymbol{I} - 2i\boldsymbol{TB}^{-1}|^{-a/2} \\ &= \exp\left\{-\tfrac{a}{2}\log|\boldsymbol{I} - 2i\boldsymbol{TB}^{-1}|\right\}, \end{aligned}$$

and since the determinant of a matrix is equal to the product of the eigenvalues of that matrix, we get

$$E\left\{\exp\left(2it\boldsymbol{X}_{12}\right)\right\} = \exp\left\{-\frac{a}{2}\sum_{\ell=1}^{n}\log\left(\lambda_\ell\right)\right\},$$

where the $\lambda_\ell, 1 \le \ell \le n$, are the eigenvalues of the matrix $|\boldsymbol{I} - 2i\boldsymbol{TB}^{-1}|$. This can also be shown for Wishart random matrices of dimension $n \times n$.

Chapter 3

Variational inference for heteroscedastic semiparametric regression

3.1 Introduction

In many situations considering real data, the assumption of constant conditional variance on the regressors is erroneous, albeit used routinely in parametric and nonparametric regression. Ignorance of significant heteroscedasticity will lead to misguided inferential statistics such as prediction intervals. A remedy for this is to employ heteroscedastic nonparametric regression which overcomes these complications. For a set of data, (x_i, y_i), $1 \leq i \leq n$, we replace the homoscedastic nonparametric regression model

$$E(y_i) = f(x_i), \quad \text{Var}(y_i) = \sigma^2$$

with

$$E(y_i) = f(x_i), \quad \text{Var}(y_i) = g(x_i), \tag{3.1}$$

where the variance function g is estimated simultaneously with the mean function f. Several approaches already exist for fitting (3.1), some of which are encompassed in Ruppert *et al.* (2003) and Rigby & Stasinopoulos (2005). However, graphical model representations of mixed model-based penalized splines (Wand, 2009) have shown to be an effective

The content of this chapter is published as: Menictas, M and Wand. M.P. (2015). Variational inference for heteroscedastic semiparametric regression. *Australian and New Zealand Journal of Statistics*, **57**, Number 1, 119–138. This research was also presented at the *Australian Statistical Conference in conjunction with the Institute of Mathematical Statistics Annual Meeting*, Sydney, 2014.

approach, especially when considering the extendability of models arising from (3.1).

In this chapter we investigate a relatively new modification of MFVB known as *non-conjugate variational message passing* (Knowles & Minka, 2011), which aids in the adaptation of heteroscedasticity to the appropriate model whilst also involving only closed form algebraic expressions. Even though we focus on univariate nonparametric regression, the modularity of our approach allows straightforward extensions to more complex models. This is known as the *locality* property of MFVB and is described in section 1.8.1 and in greater detail in section 3 of Wand *et al.* (2011). For example, if a complex graphical model includes a component where one variable is modelled as a heteroscedastic nonparametric regression function of another variable, the variational inference for the parameters in that section of the graph can be developed using the methodology in this chapter.

Lázaro-Gredilla & Titsias (2011) and Nott *et al.* (2012) have also developed variational inference for heteroscedastic regression models, although the latter contribution was restricted to linear mean and log-variance functions. Simultaneous nonparametric mean and variance function estimation aided by an elegant Gaussian process approach was achieved by Lázaro-Gredilla & Titsias (2011). Their strategy involved full-rank nonparametric function estimators with Gauss-Hermite quadrature for variance function estimation. Bugbee *et al.* (2015) use MFVB with an embedded Laplace approximation to account for semi-parametric regression with heteroscedasticity. Our approach is different by using low-rank penalized splines, as well as a non-conjugate MFVB algorithm that involves closed form updates, making it more susceptible to increasingly larger data sets.

In Section 3.2, we provide details on the Bayesian penalized spline model for simultaneous mean and variance function estimation based on univariate data. Section 3.3 provides a description of the variational inference methodology used to formulate Algorithm 2. In Sections 3.4 and 3.5 we present numerical analyses and results which confirm the speed achieved by MFVB.

3.2 Model description

The generic form of the Gaussian heteroscedastic nonparametric regression model is of the form

$$y_i \overset{\text{ind.}}{\sim} \mathrm{N}\left(f(x_i),\, g(x_i)\right), \quad 1 \le i \le n, \tag{3.2}$$

where (x_i, y_i) represents the ith predictor/response pair of a regression data-set and the functions f and g are real valued smooth functions. The quantity g is referred to as the variance function, and $\sqrt{g}$ is the standard deviation function. Fitting of (3.2) using mixed model-based penalized splines (e.g. Ruppert *et al.*, 2003) requires the following structural forms of f and g:

$$\begin{aligned} f(x) &= \beta_0 + \beta_1 x + \sum_{k=1}^{K_u} u_k z_k^u(x), \qquad u_k \overset{\text{ind.}}{\sim} N(0, \sigma_u^2) \\ \text{and} \quad g(x) &= \exp\left(\gamma_0 + \gamma_1 x + \sum_{k=1}^{K_v} v_k z_k^v(x)\right), \quad v_k \overset{\text{ind.}}{\sim} N(0, \sigma_v^2). \end{aligned} \tag{3.3}$$

The z_k^u, $1 \le k \le K_u$ and z_k^v, $1 \le k \le K_v$ are spline bases of sizes K_u and K_v respectively. Here we use suitably transformed cubic O'Sullivan splines as our default for the z_k^u and z_k^u, as explained in Section 4 of Wand & Ormerod (2008). Further, we allow for K_u and K_v to comprise different sizes. The prior specifications on the model parameters are

$$\beta_0, \beta_1 \overset{\text{ind.}}{\sim} N(0, \sigma_\beta^2), \ \gamma_0, \gamma_1 \overset{\text{ind.}}{\sim} N(0, \sigma_\gamma^2), \ \sigma_u \sim \text{Half-Cauchy}(A_u) \text{ and } \sigma_v \sim \text{Half-Cauchy}(A_v),$$

where $\sigma_\beta, \sigma_\gamma, A_u, A_v > 0$ are hyperparameters to be specified by the user. Under the assumption that the data has been transformed to have zero mean and unit variance, the default setting we employ for the hyperparameters throughout this chapter are:

$$\sigma_\beta = \sigma_\gamma = A_u = A_v = 10^5. \tag{3.4}$$

The complete Bayesian hierarchical model corresponding to (3.2) is

$$\begin{aligned} \boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{u}, \boldsymbol{v} &\sim N\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_u\boldsymbol{u}, \ \text{diag}\left\{\exp\left(\boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{Z}_v\boldsymbol{v}\right)\right\}\right), \\ \boldsymbol{u}|\sigma_u^2 \sim N\left(\boldsymbol{0}, \ \sigma_u^2\boldsymbol{I}_{K_u}\right), &\quad \boldsymbol{v}|\sigma_v^2 \sim N\left(\boldsymbol{0}, \ \sigma_v^2\boldsymbol{I}_{K_v}\right), \\ \boldsymbol{\beta} \sim N\left(0, \ \sigma_\beta^2\boldsymbol{I}_2\right), &\quad \boldsymbol{\gamma} \sim N\left(0, \ \sigma_\gamma^2\boldsymbol{I}_2\right), \\ \sigma_u^2|a_u \sim \text{Inverse-Gamma}(\tfrac{1}{2}, 1/a_u), &\quad a_u \sim \text{Inverse-Gamma}(\tfrac{1}{2}, 1/A_u^2), \\ \sigma_v^2|a_v \sim \text{Inverse-Gamma}(\tfrac{1}{2}, 1/a_v), &\quad a_v \sim \text{Inverse-Gamma}(\tfrac{1}{2}, 1/A_v^2), \end{aligned} \tag{3.5}$$

where $\boldsymbol{\beta}$ is the 2×1 vector of fixed effects containing (β_0, β_1) and $\boldsymbol{\gamma}$ defined similarly. Further, $\boldsymbol{u}$ is the $K_u \times 1$ vector containing $(u_1, \ldots, u_{K_u})$, $\boldsymbol{v}$ is a $K_v \times 1$ vector defined similarly, and σ_u^2 and σ_v^2 are variance components corresponding to $\boldsymbol{u}$ and $\boldsymbol{v}$ respectively. The design matrices, $\boldsymbol{X}$, $\boldsymbol{Z}_u$ and $\boldsymbol{Z}_v$, are defined to be

$$\boldsymbol{X} \equiv [1 \ x_i]_{1 \le i \le n}, \quad \boldsymbol{Z}_u \equiv \left[z_k^u(x_i) \right]_{\substack{1 \le k \le K_u \\ 1 \le i \le n}} \quad \text{and} \quad \boldsymbol{Z}_v \equiv \left[z_k^v(x_i) \right]_{\substack{1 \le k \le K_v \\ 1 \le i \le n}}.$$

It is beneficial to merge the mean function coefficients and variance function coefficients into single vectors:

$$\boldsymbol{\nu} \equiv \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\omega} \equiv \begin{bmatrix} \boldsymbol{\gamma} \\ \boldsymbol{v} \end{bmatrix}. \tag{3.6}$$

Combining the corresponding design matrices $\boldsymbol{C}_\nu \equiv [\boldsymbol{X}\ \boldsymbol{Z}_u]$ and $\boldsymbol{C}_\omega \equiv [\boldsymbol{X}\ \boldsymbol{Z}_v]$ then allows us to use the following notation

$$\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_u\boldsymbol{u} = \boldsymbol{C}_\nu\boldsymbol{\nu} \quad \text{and} \quad \boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{Z}_v\boldsymbol{v} = \boldsymbol{C}_\omega\,\boldsymbol{\omega}.$$

The directed acyclic graph corresponding to the model shown in (3.5) is shown in Figure 3.1.

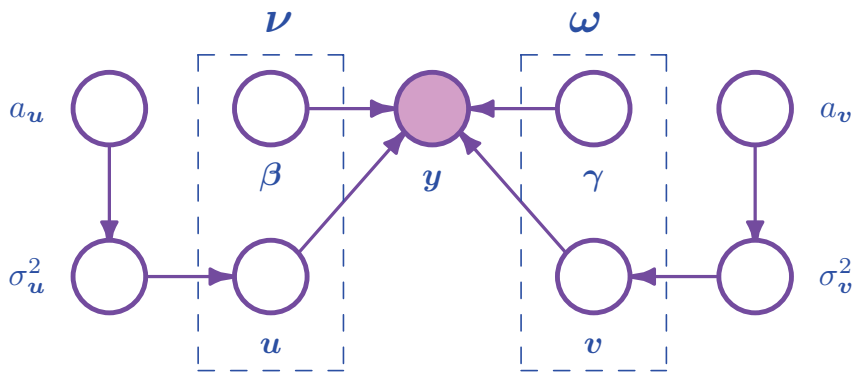


Figure 3.1: *Directed acyclic graph for the model conveyed in (3.5). The shaded node corresponds to the observed data vector. Random effects and auxiliary variables are referred to as hidden nodes. The dashed boxes indicate that* $\boldsymbol{\beta}$ *and* $\boldsymbol{u}$ *are combined into a vector denoted by* $\boldsymbol{\nu}$ *and* $\boldsymbol{\omega}$ *is the combination of* $\boldsymbol{\gamma}$ *and* $\boldsymbol{v}$.

3.3 Variational inference algorithm

Having discussed the elementary design of our model, we next provide the methodology required to achieve variational bayesian inference for (3.5). Ordinary implementation of MFVB methodology involves restricting the full posterior density function to have the approximate product form

$$p(\boldsymbol{\nu}, \boldsymbol{\omega}, \boldsymbol{a}, \boldsymbol{\sigma}^2|\,\boldsymbol{y}) \approx q(\boldsymbol{\nu})q(\boldsymbol{\omega})\,q(\sigma_u^2, \sigma_v^2)\,q(a_u, a_v), \tag{3.7}$$

where $\boldsymbol{a} = (a_u, a_v)$ and $\boldsymbol{\sigma}^2 = (\sigma_u^2, \sigma_v^2)$. Additional factorisations can easily be detected using moralisation. After moralising (see Definition 1.8.9) it becomes clear that all paths be-

tween σ_u^2 and σ_v^2 must visit at least one of the components in the set $\{\boldsymbol{y}, \boldsymbol{\nu}, \boldsymbol{\omega}\}$. Using Theorem 1.8.1, σ_u^2 is separated from σ_v^2 given $\{\boldsymbol{y}, \boldsymbol{\nu}, \boldsymbol{\omega}\}$. Similarly, $a_u \perp\!\!\!\perp a_v \,|\, \{\boldsymbol{y}, \boldsymbol{\nu}, \boldsymbol{\omega}, \sigma_u^2, \sigma_v^2\}$. Therefore, (3.7) reduces to

$$q(\boldsymbol{\nu})q(\boldsymbol{\omega})\, q(\sigma_u^2, \sigma_v^2)\, q(a_u, a_v) = q(\boldsymbol{\nu})q(\boldsymbol{\omega})\, q(\sigma_u^2)\, q(\sigma_v^2)\, q(a_u)\, q(a_v). \tag{3.8}$$

Under this product restriction, the parameters $\boldsymbol{\nu}, \sigma_u^2, \sigma_v^2, a_u$ and a_v all have closed form expressions for their full conditional distributions

$$\begin{aligned}
p(\boldsymbol{\nu}|\text{rest}) \;\sim\; \mathrm{N}\Bigg(& \left\{ \boldsymbol{C}_\nu^\top \mathrm{diag}\left(e^{-\boldsymbol{C}_\omega \boldsymbol{\omega}}\right) \boldsymbol{C}_\nu + \begin{bmatrix} \sigma_{\boldsymbol{\beta}}^{-2} \boldsymbol{I}_2 & 0 \\ 0 & \sigma_u^{-2} \boldsymbol{I}_{K_u} \end{bmatrix} \right\} \boldsymbol{C}_\nu^\top \mathrm{diag}\left(e^{-\boldsymbol{C}_\omega \boldsymbol{\omega}}\right) \boldsymbol{y}, \\
& \left\{ \boldsymbol{C}_\nu^\top \mathrm{diag}\left(e^{-\boldsymbol{C}_\omega \boldsymbol{\omega}}\right) \boldsymbol{C}_\nu + \begin{bmatrix} \sigma_{\boldsymbol{\beta}}^{-2} \boldsymbol{I}_2 & 0 \\ 0 & \sigma_u^{-2} \boldsymbol{I}_{K_u} \end{bmatrix} \right\}^{-1} \Bigg) \\
p(\sigma_u^2|\text{rest}) \;\sim\; & \text{Inverse-Gamma}\left(\tfrac{1}{2}\left(K_u + 1\right),\, \tfrac{1}{2}\|\boldsymbol{u}\|^2 + a_u^{-1}\right) \\
p(\sigma_v^2|\text{rest}) \;\sim\; & \text{Inverse-Gamma}\left(\tfrac{1}{2}\left(K_v + 1\right),\, \tfrac{1}{2}\|\boldsymbol{v}\|^2 + a_v^{-1}\right) \\
p(a_u) \;\sim\; & \text{Inverse-Gamma}\left(1,\, \sigma_u^{-2} + A_u^{-2}\right) \\
p(a_v) \;\sim\; & \text{Inverse-Gamma}\left(1,\, \sigma_v^{-2} + A_v^{-2}\right).
\end{aligned}$$

However, the optimal posterior density function of $\boldsymbol{\omega}$ is

$$q^*\left(\boldsymbol{\omega}\right) \propto \exp\left[E_{q(-\boldsymbol{\omega})}\left\{\log p\left(\boldsymbol{\omega}|\text{rest}\right)\right\}\right], \tag{3.9}$$

which involves intractable multivariate integrals that are not available in closed form expressions. A remedy in this instance is to work with the product restriction

$$p(\boldsymbol{\nu},\, \boldsymbol{\omega},\, \boldsymbol{a},\, \boldsymbol{\sigma}^2|\, \boldsymbol{y}) \approx q(\boldsymbol{\nu})\, q(\boldsymbol{\omega}; \boldsymbol{\mu}_{q(\boldsymbol{\omega})},\, \boldsymbol{\Sigma}_{q(\boldsymbol{\omega})})\, q(\boldsymbol{a})\, q(\boldsymbol{\sigma}^2), \tag{3.10}$$

where

$$q\left(\boldsymbol{\omega}; \boldsymbol{\mu}_{q(\boldsymbol{\omega})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\omega})}\right) \text{ is the } N\left(\boldsymbol{\mu}_{q(\boldsymbol{\omega})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\omega})}\right) \text{ density function.} \tag{3.11}$$

Thus, the marginal log-likelihood is

$$\begin{aligned}
\log \underline{p}(\boldsymbol{y};\, q,\, \boldsymbol{\mu}_{q(\boldsymbol{\omega})},\, \boldsymbol{\Sigma}_{q(\boldsymbol{\omega})}) \;\equiv\; & E_q\big[\log p(\boldsymbol{\nu},\, \boldsymbol{\omega},\, \boldsymbol{a},\, \boldsymbol{\sigma}^2,\, \boldsymbol{y}) \\
& - \log\left\{q(\boldsymbol{\nu})\, q(\boldsymbol{\omega}; \boldsymbol{\mu}_{q(\boldsymbol{\omega})},\, \boldsymbol{\Sigma}_{q(\boldsymbol{\omega})})\, q(\boldsymbol{a})\, q(\boldsymbol{\sigma}^2)\right\}\big]
\end{aligned} \tag{3.12}$$

and choosing $\boldsymbol{\mu}_{q(\boldsymbol{\omega})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\omega})}$ to maximise (3.12) corresponds to minimization of the Kullback-Leibler divergence. Knowles & Minka (2011) termed this approach *non-conjugate variational message passing*, since it offers a way of getting around the non-conjugacies of

ordinary MFVB. This approach is explained in more detail in Section 1.7.

For fixed $(\boldsymbol{\mu}_{q(\boldsymbol{\omega})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\omega})})$, application of ordinary MFVB, with $q^*(\boldsymbol{\omega})$ omitted, gives

$$\begin{aligned} & q^*(\boldsymbol{\nu}) \text{ is a Normal } \left(\boldsymbol{\mu}_{q(\boldsymbol{\nu})},\ \boldsymbol{\Sigma}_{q(\boldsymbol{\nu})}\right) \text{ density function,} \\ & q^*(\sigma_u^2) \text{ is an Inverse-Gamma } \left(\tfrac{1}{2}(K_u+1),\ B_{q(\sigma_u^2)}\right) \text{ density function,} \\ & q^*(\sigma_v^2) \text{ is an Inverse-Gamma } \left(\tfrac{1}{2}(K_v+1),\ B_{q(\sigma_v^2)}\right) \text{ density function,} \\ & q^*(a_u) \text{ is an Inverse-Gamma } \left(1,\ B_{q(a_u)}\right) \text{ density function} \\ \text{and} \quad & q^*(a_v) \text{ is an Inverse-Gamma } \left(1,\ B_{q(a_v)}\right) \text{ density function,} \end{aligned} \tag{3.13}$$

for parameters $\boldsymbol{\mu}_{q(\boldsymbol{\nu})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\nu})}$, the mean and covariance matrix of $q^*(\boldsymbol{\nu})$, $B_{q(\sigma_u^2)}$, the rate parameter of $q^*(\sigma_u^2)$, $B_{q(\sigma_v^2)}$, the rate parameter of $q^*(\sigma_v^2)$, $B_{q(a_u)}$, the rate parameter of $q^*(a_u)$ and $B_{q(a_v)}$, the rate parameter of $q^*(a_v)$.

With assistance from Knowles & Minka (2011) who propose a fixed-point iteration scheme for maximizing (3.12) and Wand (2014) who provides simplified fixed point updates for the Multivariate Normal q-density parameters, such as that shown in (3.11), we can now present Algorithm 2.

Convergence in Algorithm 2 is assessed using the variational lower bound on the approximate marginal log-likelihood, denoted by $\log \underline{p}(\boldsymbol{y}; q, \boldsymbol{\mu}_{q(\boldsymbol{\omega})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\omega})})$ and has the following expression:

$$\begin{aligned} \log \underline{p}(\boldsymbol{y}; q, \boldsymbol{\mu}_{q(\boldsymbol{\omega})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\omega})}) \quad = \quad & \tfrac{1}{2}(K_u + K_v + 4) - \tfrac{n}{2}\log(2\pi) + \log\Gamma\left(\tfrac{1}{2}(K_u+1)\right) - 2\log(\pi) \\ & + \log\Gamma\left(\tfrac{1}{2}(K_v+1)\right) - \log(A_u) - \log(A_v) - \tfrac{1}{2}\mathbf{1}^T(\boldsymbol{C}_\omega \boldsymbol{\mu}_{q(\boldsymbol{\omega})}) \\ & - \tfrac{1}{2}\mathbf{1}^\top\left\{\boldsymbol{\mu}_{q(r_{\boldsymbol{\nu}}^2)} \odot \exp\left(\boldsymbol{\psi}_{q(\boldsymbol{\omega})}\right)\right\} - \log(\sigma_\beta^2) - \log(\sigma_\gamma^2) \\ & + \tfrac{1}{2}\log|\Sigma_{q(\boldsymbol{\nu})}| + \tfrac{1}{2}\log|\Sigma_{q(\boldsymbol{\omega})}| - \tfrac{1}{2\sigma_\beta^2}\left(\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\right) \\ & - \tfrac{1}{2\sigma_\gamma^2}\left\{\|\boldsymbol{\mu}_{q(\boldsymbol{\gamma})}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\gamma})})\right\} - \tfrac{1}{2}(K_u+1)\log(B_{q(\sigma_u^2)}) \\ & - \tfrac{1}{2}(K_v+1)\log(B_{q(\sigma_v^2)}) - \log\left(\mu_{q(1/\sigma_u^2)} + A_u^{-2}\right) \\ & - \log\left(\mu_{q(1/\sigma_v^2)} + A_v^{-2}\right) + \mu_{q(1/\sigma_u^2)}\mu_{q(1/a_u)} + \mu_{q(1/\sigma_v^2)}\mu_{q(1/a_v)}. \end{aligned}$$

The optimal parameters shown in Algorithm 2 are all interdependent and are obtained by cycling through this iterative coordinate ascent procedure. The algorithm also provides fixed point iterative updates for $\boldsymbol{\mu}_{q(\boldsymbol{\omega})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\omega})}$. Details on the derivation of these parameters, as well as the parameters in (3.13) are given in Appendix 3.A. In addition, $\boldsymbol{\mu}_{q(\boldsymbol{u})}$ is defined to be the sub-vector of $\boldsymbol{\mu}_{q(\boldsymbol{\nu})}$ corresponding to $\boldsymbol{u}$, and $\boldsymbol{\Sigma}_{q(\boldsymbol{u})}$ is the sub-matrix of $\boldsymbol{\Sigma}_{q(\boldsymbol{\nu})}$ corresponding to $\boldsymbol{u}$. The parameters $\boldsymbol{\mu}_{q(\boldsymbol{v})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{v})}$ are defined similarly. It is important to note that unlike ordinary MFVB, there is no guarantee that there will be an increase at each iteration for $\underline{p}(\boldsymbol{y}; q, \boldsymbol{\mu}_{q(\boldsymbol{\omega})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\omega})})$. Therefore, the cycle should end

Set up initial values:

$\boldsymbol{\mu}_{q(\boldsymbol{\omega})}$ a $(K_v+2)\times 1$ vector, $\boldsymbol{\Sigma}_{q(\boldsymbol{\omega})}$ a $(K_v+2)\times(K_v+2)$ positive definite matrix, $\boldsymbol{\mu}_{q(1/\sigma_u^2)}, \boldsymbol{\mu}_{q(1/\sigma_v^2)} > 0$, and $\boldsymbol{\mu}_{q(\boldsymbol{r}_\nu^2)}$ an $n\times 1$ vector.

Cycle through:

$$\boldsymbol{\psi}_{q(\boldsymbol{\omega})} \leftarrow \exp\left\{-\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\omega})} + \tfrac{1}{2}\text{diagonal}\left(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\omega})}\boldsymbol{C}^\top\right)\right\}$$

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\nu})} \leftarrow \left(\boldsymbol{C}^\top \text{diag}\left\{\boldsymbol{\psi}_{q(\boldsymbol{\omega})}\right\}\boldsymbol{C} + \begin{bmatrix} \sigma_{\boldsymbol{\beta}}^{-2}\boldsymbol{I}_2 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\mu}_{q(1/\sigma_{\boldsymbol{u}}^2)}\boldsymbol{I}_K \end{bmatrix}\right)^{-1}$$

$$\boldsymbol{\mu}_{q(\boldsymbol{\nu})} \leftarrow \boldsymbol{\Sigma}_{q(\boldsymbol{\nu})}\boldsymbol{C}^\top \text{diag}\left\{\boldsymbol{\psi}_{q(\boldsymbol{\omega})}\right\}\boldsymbol{y}$$

$$\boldsymbol{\mu}_{q(\boldsymbol{r}_\nu^2)} \leftarrow \text{diagonal}\left\{\left(\boldsymbol{y}-\boldsymbol{C}_\nu\boldsymbol{\mu}_{q(\boldsymbol{\nu})}\right)\left(\boldsymbol{y}-\boldsymbol{C}_\nu\boldsymbol{\mu}_{q(\boldsymbol{\nu})}\right)^\top + \boldsymbol{C}_\nu\boldsymbol{\Sigma}_{q(\boldsymbol{\nu})}\boldsymbol{C}_\nu^\top\right\}$$

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\omega})} \leftarrow \left(\boldsymbol{C}^\top \text{diag}\left\{\boldsymbol{\mu}_{q(\boldsymbol{r}_\nu^2)} \odot \boldsymbol{\psi}_{q(\boldsymbol{\omega})}\right\}\boldsymbol{C} + \begin{bmatrix} \sigma_{\boldsymbol{\gamma}}^{-2}\boldsymbol{I}_2 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\mu}_{q(1/\sigma_{\boldsymbol{v}}^2)}\boldsymbol{I}_K \end{bmatrix}\right)^{-1}$$

$$\boldsymbol{\mu}_{q(\boldsymbol{\omega})} \leftarrow \boldsymbol{\mu}_{q(\boldsymbol{\omega})} + \boldsymbol{\Sigma}_{q(\boldsymbol{\omega})}\left\{\boldsymbol{C}^\top\left(\boldsymbol{\mu}_{q(\boldsymbol{r}_\nu^2)} \odot \boldsymbol{\psi}_{q(\boldsymbol{\omega})} - \boldsymbol{1}\right) - \begin{bmatrix} \sigma_{\boldsymbol{\gamma}}^{-2}\boldsymbol{I}_2 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\mu}_{q(1/\sigma_{\boldsymbol{v}}^2)}\boldsymbol{I}_K \end{bmatrix}\boldsymbol{\mu}_{q(\boldsymbol{\omega})}\right\}$$

$$\mu_{q(1/a_{\boldsymbol{u}})} \leftarrow 1/\left(\mu_{q(1/\sigma_{\boldsymbol{u}}^2)} + A_{\boldsymbol{u}}^2\right) \quad ; \quad \mu_{q(1/a_{\boldsymbol{v}})} \leftarrow 1/\left(\mu_{q(1/\sigma_{\boldsymbol{v}}^2)} + A_{\boldsymbol{v}}^2\right)$$

$$\mu_{q(1/\sigma_{\boldsymbol{u}}^2)} \leftarrow (K_{\boldsymbol{u}}+1)\,/\left\{2\mu_{q(1/a_{\boldsymbol{u}})} + \|\boldsymbol{\mu}_{q(\boldsymbol{u})}\|^2 + \text{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{u})}\right)\right\}$$

$$\mu_{q(1/\sigma_{\boldsymbol{v}}^2)} \leftarrow (K_{\boldsymbol{v}}+1)\,/\left\{2\mu_{q(1/a_{\boldsymbol{v}})} + \|\boldsymbol{\mu}_{q(\boldsymbol{v})}\|^2 + \text{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{v})}\right)\right\}$$

until the absolute relative change in $\log \underline{p}(\boldsymbol{y};\, q,\, \boldsymbol{\mu}_{q(\boldsymbol{\omega})},\, \boldsymbol{\Sigma}_{q(\boldsymbol{\omega})})$ is negligible.

Algorithm 2: *MFVB algorithm for the determination of the optimal parameters in* $q^*(\boldsymbol{\omega})$, $q^*(\boldsymbol{\nu})$, $q^*(\sigma_u^2)$, $q^*(\sigma_v^2)$, $q^*(a_u)$ *and* $q^*(a_v)$.

Setting	$f(x)$	$\log g(x)$
A	$\sin(3\pi x^2)$	$0.1 + \cos(4\pi x)$
B	$-1.02x + 0.018x^2 + 5\phi\left(x; 0.38, 0.08\right)$ $+\frac{8}{3}\phi\left(x; 0.75, 0.03\right)$	$-0.5 - \Phi\left(x; 0.2, 0.1\right) + 0.3x^2$
C	$\frac{35}{8}\phi\left(x; 0.01, 0.08\right) + \frac{190}{23}\phi\left(x; 0.45, 0.23\right)$ $+1.8\left\{1 - \frac{100}{14}\phi\left(x; 0.7, 0.14\right)\right\}$	$\frac{3}{2}\phi\left(x; 0, 0.2\right) + 4\phi\left(x; 1, 0.1\right)$
D	$\sin(3\pi x^2) - 1.02x + 0.018x^2$ $+5\phi\left(x; 0.38, 0.08\right)$	$\cos\left(4\pi x\right) - 0.4 + 0.3x^2$ $-\Phi\left(x; 0.2, 10\right)$

Table 3.1: *Details of simulation study settings.*

once the absolute relative change in its logarithm falls below a negligible amount.

Algorithm 2 has been applied to simulated data and accuracy against MCMC has been assessed. The results are summarized in Section 3.4.

3.4 Assessment of performance

In order to assess the performance of Algorithm 2, we carried out an extensive simulation study to address the accuracy and computing time of model (3.2). Data were simulated according to model (3.2) with $n = 500$ and the x_is uniform on $(0, 1)$. The four mean and variance function pairs that were used are listed in Table 3.1, where $\phi(\cdot; \mu, \sigma)$ and $\Phi(\cdot; \mu, \sigma)$ denote the density and distribution functions of the Normal distribution with mean μ and standard deviation σ, respectively. 100 data-sets were generated for each of the function pairs shown in Table 3.1. Each model corresponding to a new replication was fitted using MFVB corresponding to Algorithm 2 and MCMC based on a burnin of 5000, kept sample of 5000 and thinning factor of 5. The MFVB iterations were terminated when the relative change in $\log \underline{p}(\boldsymbol{y}; q)$ fell below 10^{-7}.

In the sections that follow, we provide details on the accuracy, coverage and timing of MFVB against the MCMC benchmark. Let us next discuss each of these assessments in more detail.

3.4.1 Assessment of accuracy

Fast approximate inference for the model parameters is achieved by Algorithm 2, however there is no guarantee that an acceptable level of accuracy will occur. Figure 3.2 provides

an accuracy assessment of Algorithm 2, where the accuracy scores are summarized by boxplots for each parameter of interest. The accuracy score is defined in Section 1.9.

The parameters of interest for which accuracy is monitored are the $f(H_k)$ and $g(H_k)$, $1 \leq k \leq 5$, where the H_k are the sample hexiles of the x_is. It is evident from Figure 3.2 that most of the accuracies for the $f(H_k)$ lie around 90%, while accuracies for the $g(H_k)$ lie around 80%. These favourable accuracy results are in keeping with the heuristics given in Section 2.3.1 of Chapter 2.

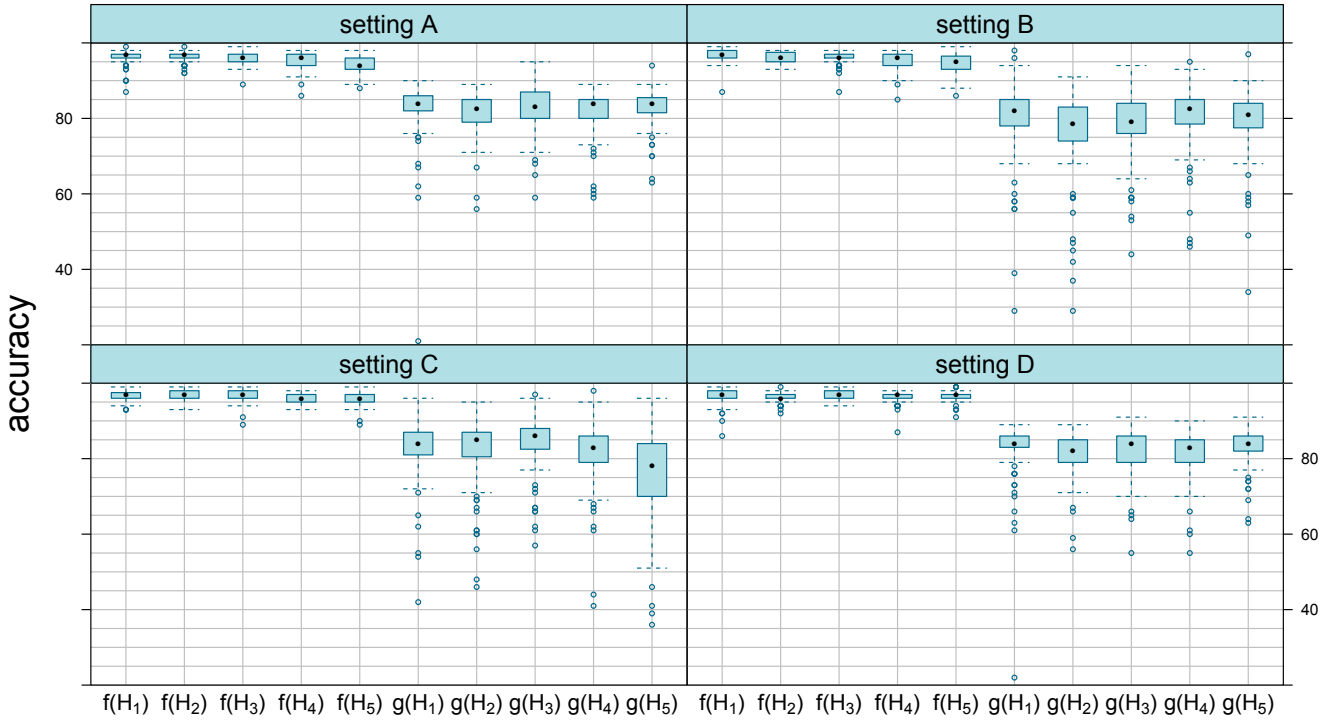


Figure 3.2: *Summary of simulation study where the accuracy values are summarized using boxplots.*

The fitted mean and standard deviation functions for the first replication in each of the four simulation settings is shown in Figure 3.3, with comparisons of MCMC and MFVB. As illustrated, the MCMC and MFVB fits show very good agreement for the mean functions and relatively good agreement for the standard deviation functions.

It is also of interest to see how well the joint distribution of certain parameters are doing in terms of approximation accuracy. Figure 3.4 shows a visual assessment of the approximation accuracy for the joint posterior density functions of the mean and variance hexile pairs according to simulation setting A.

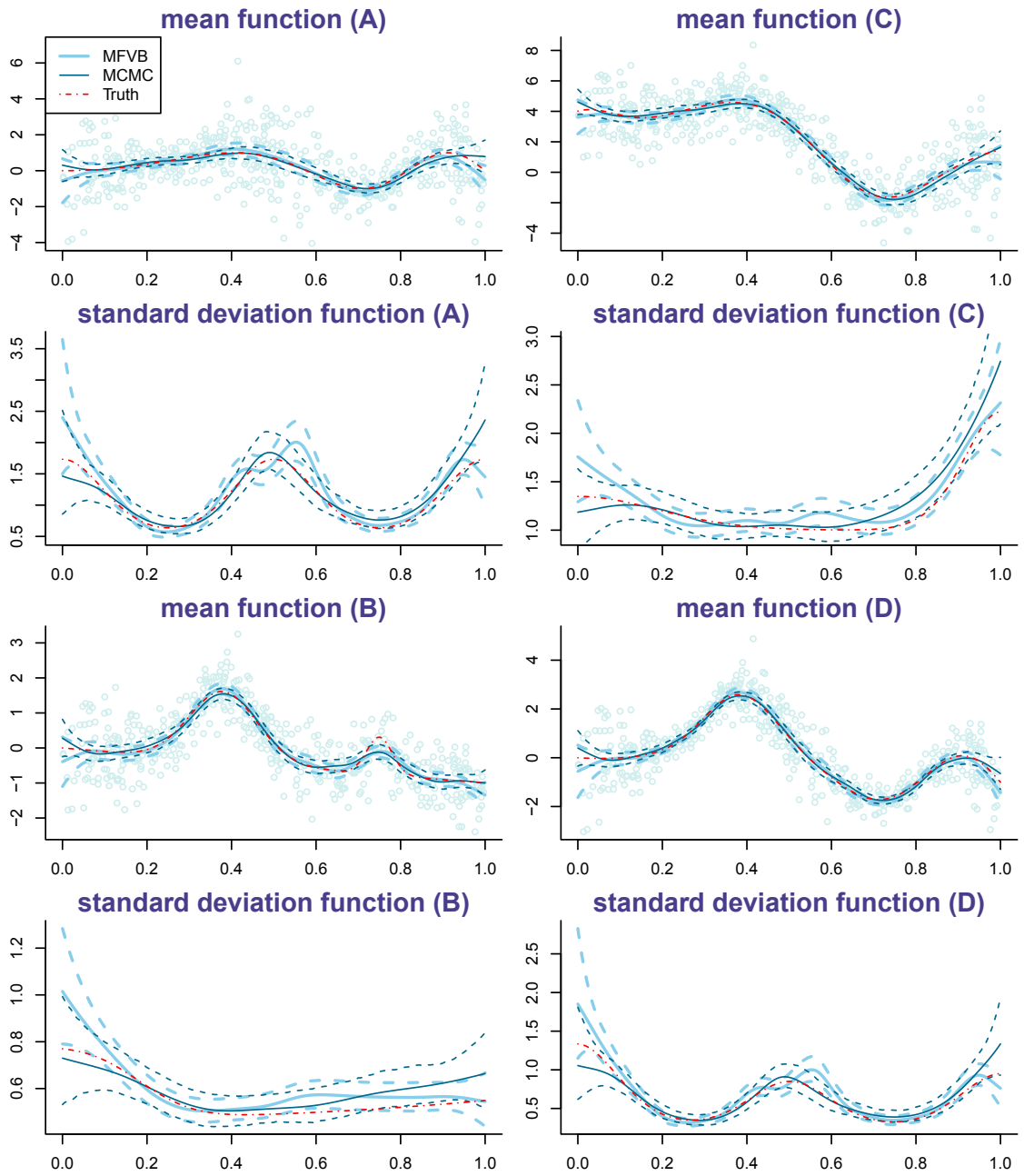


Figure 3.3: *Comparison of MCMC and MFVB fitted functions for the first data-set in each of the four simulation settings.*

In many cases, as shown by Figure 3.2, the accuracies can provide evidence of poor MFVB performance. Figure 3.5 illustrates some of the poorer approximations for this simulation study.

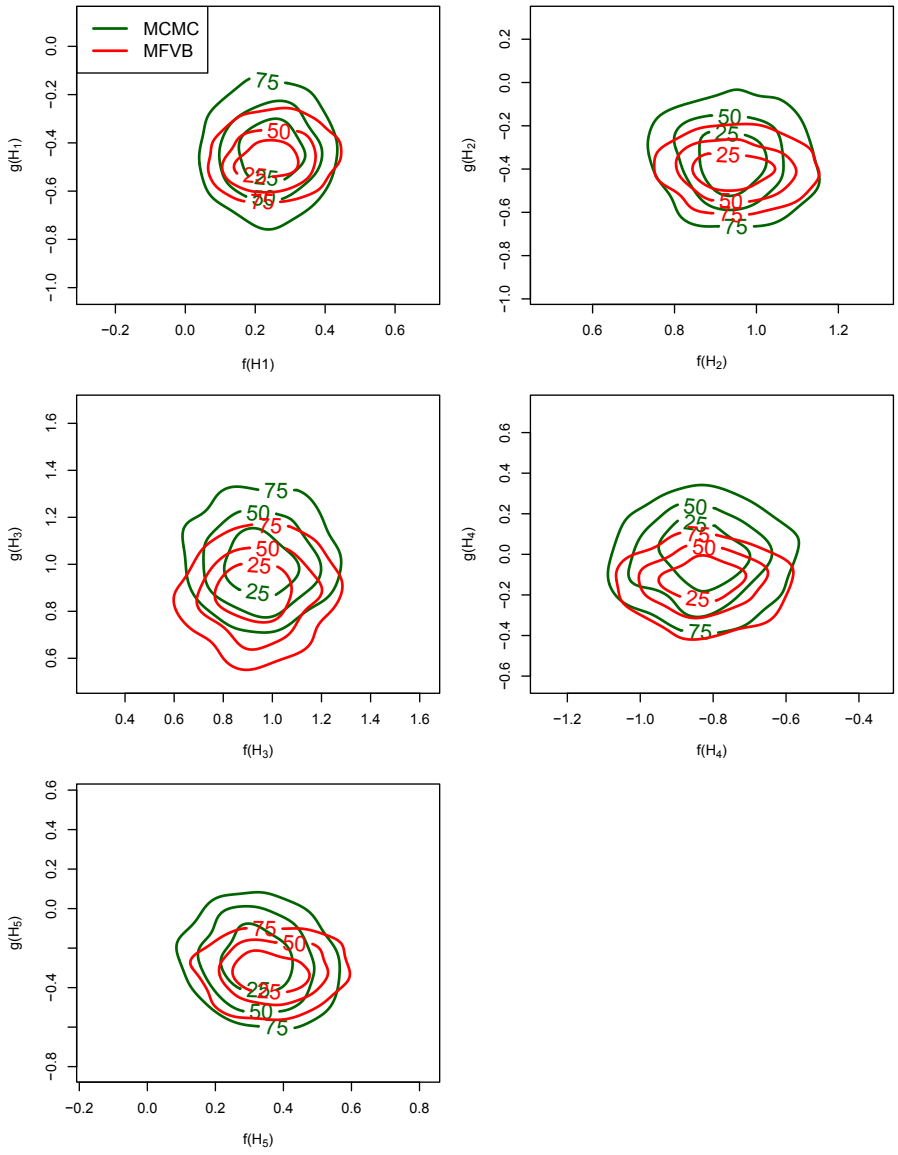


Figure 3.4: *Example of approximation accuracy for the joint posterior distributions of the mean and variance hexiles given for the first dataset in simulation setting A.*

3.4.2 Assessment of coverage

Another assessment of the performance of MFVB, involves the comparison between the true coverage and the coverage gained by the MFVB approximate credible intervals. The percentages of the true parameter coverage based on the approximate 95% credible intervals attained from the MFVB posterior densities are given in Table 3.2. As can be seen, the coverage overall is good and does not fall below 86%. Here we provide further confirmation that the performance of MFVB is excellent for the mean function and relatively

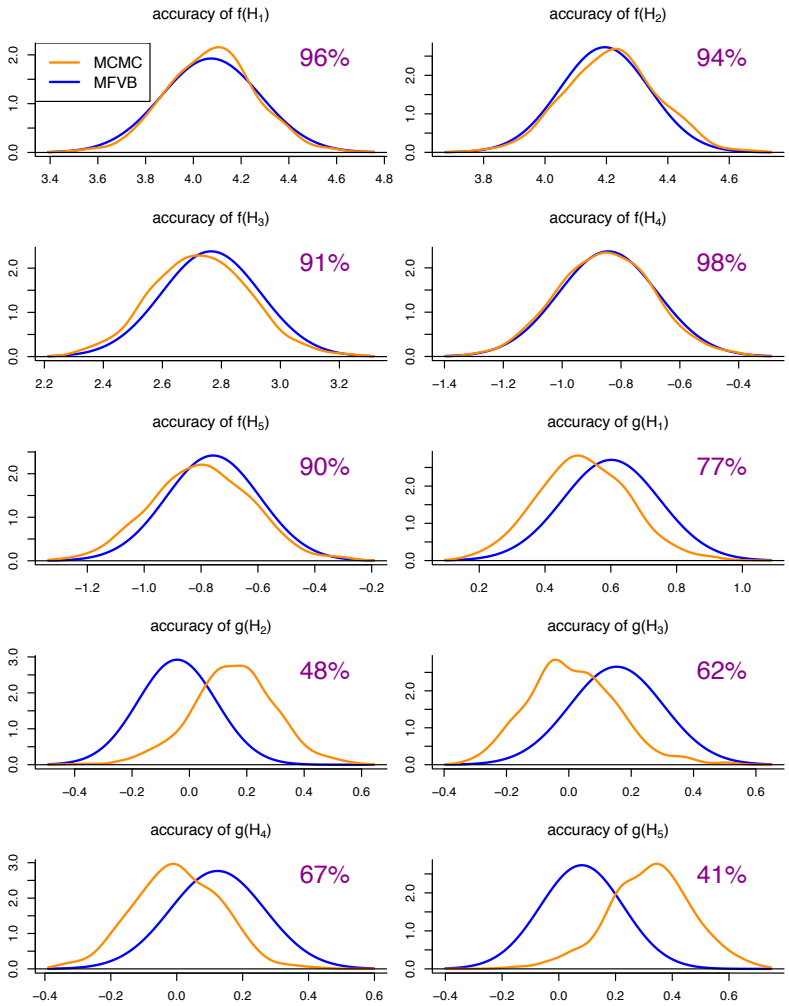


Figure 3.5: *Example of a poor approximation (for the variance function) for a dataset corresponding to simulation setting C.*

good for the variance function.

	$f(H_1)$	$f(H_2)$	$f(H_3)$	$f(H_4)$	$f(H_5)$	$g(H_1)$	$g(H_2)$	$g(H_3)$	$g(H_4)$	$g(H_5)$
sett. A	98	98	94	98	97	89	87	83	87	82
sett. B	98	98	95	96	95	83	83	90	89	83
sett. C	99	98	96	99	99	90	91	92	90	76
sett. D	98	98	95	98	98	89	89	83	86	82

Table 3.2: *Percentage coverage of the true parameter values by approximate* 95% *credible intervals based on variational Bayes approximate posterior density functions. The percentages are based on 100 replications.*

3.4.3 Assessment of speed

In order to demonstrate the time savings to be gained by using MFVB when compared to MCMC, we monitored the time taken to fit each model for each approach. The study was performed on a desktop computer (Intel Core i5-2400 3.10 GHz processor, 8 GBytes of random access memory). The resulting times are summarized in Table 3.3.

As we discussed previously, convergence for both approaches was assessed differently. Specifically, the MFVB iterations were terminated when the relative change in $\log \underline{p}(\boldsymbol{y}; q)$ fell below 10^{-7}, whilst MCMC was based on a kept sample of 1000. In addition, the speed achieved by MFVB is traded off against slight accuracy losses which are invoked by the product restriction given in (3.8).

However, even when allowing for these slight differences, the results clearly demonstrate that MFVB is at least approximately 200 times faster than MCMC across the models. Therefore, we conclude that a model which takes minutes to run using MCMC, will only take seconds to run using our MFVB algorithm.

	MCMC	MFVB
setting A	(1225.49, 1228.45)	(1.69, 1.87)
setting B	(1232.96, 1238.53)	(4.95, 5.89)
setting C	(1185.45, 1188.01)	(3.29, 4.67)
setting D	(1217.77, 1364.56)	(1.26, 1.38)

Table 3.3: *99% Wilcoxon confidence intervals based on run times in seconds for MCMC and MFVB fitting.*

3.5 Application

We have seen the time benefits and good accuracy results for MFVB from the results discussed thus far using simulated data. We next evaluate the performance of MFVB using real data in order to understand its potential application.

Firstly, we consider two nonparametric regression examples without taking heteroscedasticity into account. Secondly, we use our new MFVB methodology by applying Algorithm 2 to the same regression examples. This enables us to clearly understand the implications of ignoring heteroscedasticity, as it might on occasion be tempting to immediately fit the model in a naïve fashion. Our process involved fitting a cubic smoothing spline to two nonparametric regression examples, based on the R function `smooth.spline()` (R Core

Team, 2013) with default settings. Figure 3.6 illustrates the fits. For the interested reader, a complete description of the two data sets that were used can be found in Sections 2.7 and 5.3 of Ruppert *et al.* (2003).

As can be evidenced from the standardized residual plots in Figure 3.6, there exists significant heteroscedasticity. Ignoring the heteroscedastic nature of these data sets will lead to erroneous inferential statements, such as prediction intervals. In order to account

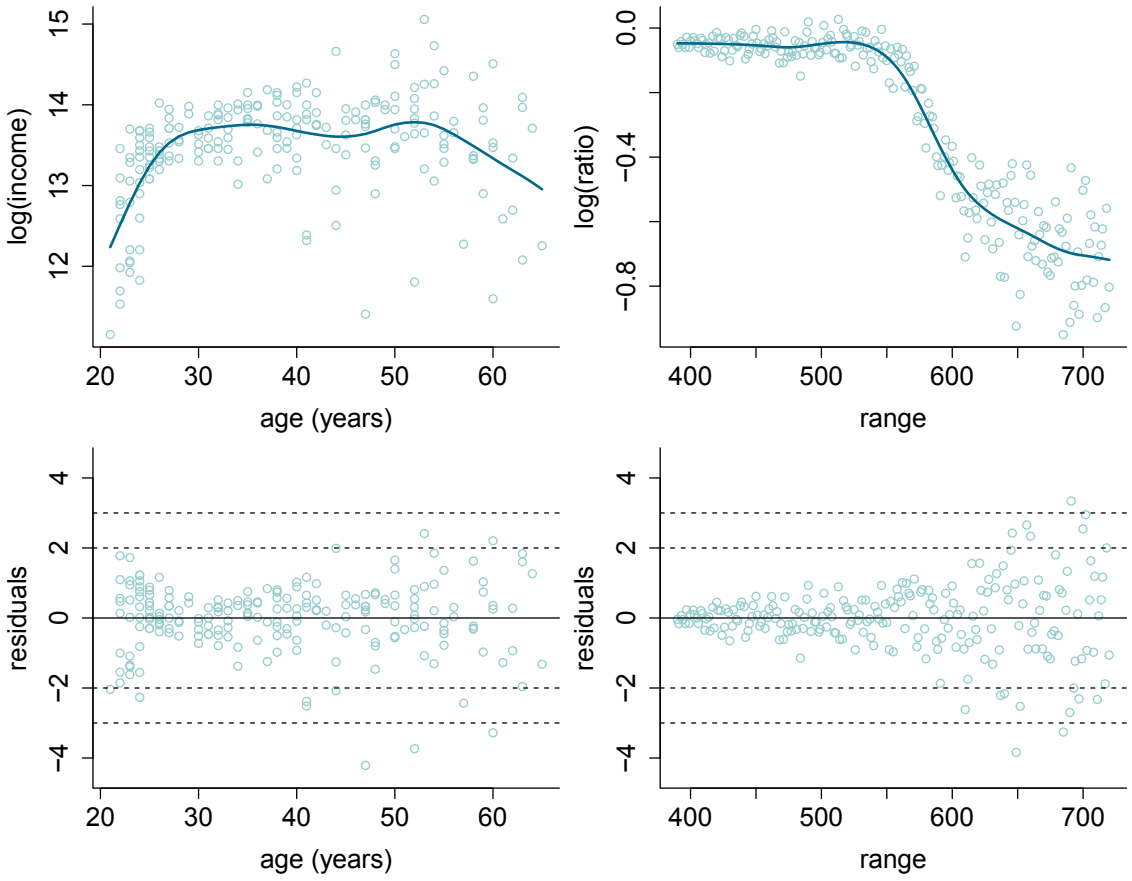


Figure 3.6: *Two example nonparametric regression fits and corresponding standardized residual plots. Horizontal dashed lines at* ±2 *and* ±3 *aid assessment of standard normality of the standardized residuals.*

for the significant heteroscedasticity as suggested by Figure 3.6, we also fitted (3.5) using Algorithm 2 to both these data sets. This is shown in Figure 3.7. The top panels illustrate the fitted mean functions with pointwise 95% credible sets. The middle panels illustrate the fitted standard deviation functions, corresponding to $\sqrt{g}$ in the notation of (3.2), and corresponding 95% credible sets. The model was fitted using MFVB corresponding to Algorithm 2 and MCMC based on BUGS (**?**) with a burnin of 5000, a kept sample of

5000 and a thinning factor of 5. The MFVB iterations were terminated when the relative change in $\log \underline{p}(\boldsymbol{y}; q)$ fell below 10^{-7}. This value is determined by graphically checking if convergence is achieved and can change depending on the type of application. We also conducted a comprehensive check to confirm that the relative change in the approximate marginal log-likelihood does not lead to early stopping. Data was generated from different scenarios, over several hundred runs and Bayes estimates of f and g were recorded at the stopping point and again with iterations continuing 25% beyond the stopping point. The differences in the estimates were negligible.

The agreement between the MFVB and the MCMC fits for the mean function is excellent. For the standard deviation function fit, the agreement between MFVB and MCMC is good, rather than excellent. Finally, the $\sqrt{g}$-standardized residual plots in the bottom panels of Figure 3.7, indicate proper accounting for the heteroscedasticity and agreement with normality.

3.6 Discussion

At the outset of this chapter, we sought to develop a variational Bayesian algorithm for semiparametric regression models that have a heteroscedastic component. We sought to do this because ignoring heteroscedasticity, especially amongst complex statistical models, could lead to problematic inferential results and prediction intervals. As a result of the work carried out in this chapter, we have successfully developed a variational Bayes methodology which incorporates a new modification of MFVB, known as *non-conjugate variational message passing*, involving a closed form algorithm for univariate heteroscedastic semiparametric regression. The methodology also applies to larger more complex models, as provided for by the locality property of mean field variational inference methods.

Comparisons between variational inference and the MCMC benchmark for both simulated and actual data, shows that this new methodology performs very well. In addition, we have been able to achieve significant time savings by using our new methodology when compared to MCMC. For instance, as was shown in Section 3.4.3, we witness at least a 200 fold increase in estimation speed. This is equivalent to saying that a model which takes minutes to run using MCMC, will only take seconds to run using our MFVB methodology.

Having considered the impact of heteroscedasticity and remedies via our new MFVB-type methodology, we next turn our attention to extensions of this work which we discuss

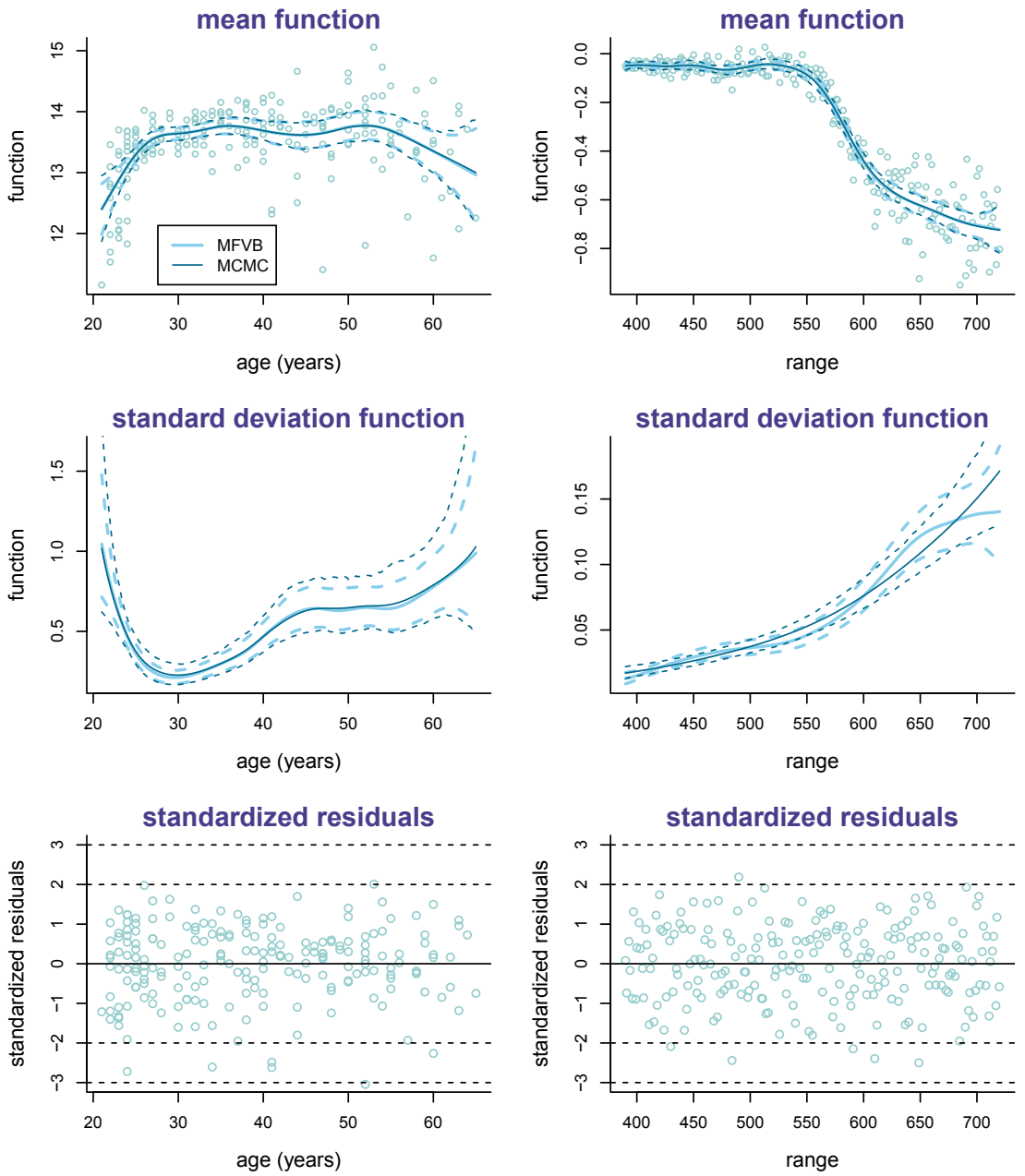


Figure 3.7: *Top panels: fitted mean functions for two example data sets from Wand & Jones (1995) and Ruppert et al. (2003). The solid curves are approximate pointwise posterior means whilst the dashed curves are corresponding pointwise 95% credible sets. The approximate fits are based on MFVB via Algorithm 2 and MCMC via* BUGS*. Middle panels: similar to top panels but for the standard deviation function. Bottom panels: standardized residual plots based on* $\{y - \widehat{f}(x_i)\}/\sqrt{\{\widehat{g}(x_i)\}}$ *where* $\widehat{f}$ *and* $\widehat{g}$ *are the MFVB-approximate Bayes estimates of* f *and* g*.*

in Chapter 4.

3.A Derivation of algorithm 2

Expressions for $\boldsymbol{\mu}_{q(\boldsymbol{\nu})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\nu})}$

First note that

$$\begin{aligned} p(\boldsymbol{\nu}|\text{rest}) &\propto p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\omega})p(\boldsymbol{\beta}) \\ &= \exp\left\{-\tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{C}_\nu\boldsymbol{\nu})^\top \text{diag}\left(e^{-\boldsymbol{C}_\omega\boldsymbol{\omega}}\right)(\boldsymbol{y}-\boldsymbol{C}_\nu\boldsymbol{\nu})\right\} \times \exp\left\{-\tfrac{1}{2\sigma_u^2}\|\boldsymbol{u}\|^2\right\} \\ &\quad \times \exp\left\{-\tfrac{1}{2\sigma_\beta^2}\|\boldsymbol{\beta}\|^2\right\} + \text{const}, \end{aligned}$$

where 'const' denotes terms not depending on the argument of $\boldsymbol{\nu}$. Taking the logarithm of both sides gives

$$\begin{aligned} \log p(\boldsymbol{\nu}|\text{rest}) &\propto -\tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{C}_\nu\boldsymbol{\nu})^\top \text{diag}\left(e^{-\boldsymbol{C}_\omega\boldsymbol{\omega}}\right)(\boldsymbol{y}-\boldsymbol{C}_\nu\boldsymbol{\nu}) - \tfrac{1}{2\sigma_u^2}\|\boldsymbol{u}\|^2 - \tfrac{1}{2\sigma_\beta^2}\|\boldsymbol{\beta}\|^2 \\ &\propto -\tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{C}_\nu\boldsymbol{\nu})^\top \text{diag}\left(e^{-\boldsymbol{C}_\omega\boldsymbol{\omega}}\right)(\boldsymbol{y}-\boldsymbol{C}_\nu\boldsymbol{\nu}) - \tfrac{1}{2}\boldsymbol{\nu}^\top \begin{bmatrix} \sigma_\beta^{-2}\boldsymbol{I}_2 & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_u^{-2}bI_{K_u} \end{bmatrix}\boldsymbol{\nu} \\ &\propto -\tfrac{1}{2}\left\{\boldsymbol{\nu}^\top\boldsymbol{C}_\nu^\top \text{diag}\left(e^{-\boldsymbol{C}_\omega\boldsymbol{\omega}}\right)\boldsymbol{C}_\nu\boldsymbol{\nu} - 2\boldsymbol{C}_\nu^\top\boldsymbol{\nu}^\top \text{diag}\left(e^{-\boldsymbol{C}_\omega\boldsymbol{\omega}}\right)\boldsymbol{y}\right. \\ &\qquad \left. +\boldsymbol{\nu}^\top \begin{bmatrix} \sigma_\beta^{-2}\boldsymbol{I}_2 & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_u^{-2}bI_{K_u} \end{bmatrix}\boldsymbol{\nu}\right\} \\ &= -\tfrac{1}{2}\left[\boldsymbol{\nu}^\top\left\{\boldsymbol{C}_\nu^\top \text{diag}\left(e^{-\boldsymbol{C}_\omega\boldsymbol{\omega}}\right)\boldsymbol{C}_\nu + \begin{bmatrix} \sigma_\beta^{-2}\boldsymbol{I}_2 & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_u^{-2}bI_{K_u} \end{bmatrix}\right\}\boldsymbol{\nu}\right. \\ &\qquad \left. -2\boldsymbol{\nu}^\top\left\{\boldsymbol{C}_\nu^\top \text{diag}\left(e^{-\boldsymbol{C}_\omega\boldsymbol{\omega}}\right)\boldsymbol{y}\right\}\right] + \text{const}. \end{aligned}$$

Now taking expectations with respect to all parameters except $\boldsymbol{\nu}$:

$$\begin{aligned} \log q^*(\boldsymbol{\nu}) &= -\tfrac{1}{2}\left[\boldsymbol{\nu}^\top\left\{\boldsymbol{C}_\nu^\top \text{diag}\left(\boldsymbol{\psi}_{q(\boldsymbol{\omega})}\right)\boldsymbol{C}_\nu + \begin{bmatrix} \sigma_\beta^{-2}\boldsymbol{I}_2 & \boldsymbol{0} \\ \boldsymbol{0} & \mu_{q(1/\sigma_u^2)}\boldsymbol{I}_{K_u} \end{bmatrix}\right\}\boldsymbol{\nu}\right. \\ &\qquad \left. -2\boldsymbol{\nu}^\top\left\{\boldsymbol{C}_\nu^\top \text{diag}\left(\boldsymbol{\psi}_{q(\boldsymbol{\omega})}\right)\right\}\right] + \text{const}, \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{\psi}_{q(\boldsymbol{\omega})} &= E_q\left\{\exp\left(-\boldsymbol{C}_\omega\boldsymbol{\omega}\right)\right\} \\ &= \exp\left\{-\boldsymbol{C}_\omega\boldsymbol{\mu}_{q(\boldsymbol{\omega})} + \tfrac{1}{2}\text{diagonal}\left(\boldsymbol{C}_\omega\boldsymbol{\Sigma}_{q(\boldsymbol{\omega})}\boldsymbol{C}_\omega^\top\right)\right\}. \end{aligned}$$

Completing the square gives

$$\log q^*(\boldsymbol{\nu}) = -\tfrac{1}{2}\left\{\boldsymbol{\nu} - \boldsymbol{C}_\nu^\top \text{diag}\left(\boldsymbol{\psi}_{q(\boldsymbol{\omega})}\right)\boldsymbol{y}\boldsymbol{\Omega}\right\}^\top \boldsymbol{\Omega}\left\{\boldsymbol{\nu} - \boldsymbol{C}_\nu^\top \text{diag}\left(\boldsymbol{\psi}_{q(\boldsymbol{\omega})}\right)\boldsymbol{y}\boldsymbol{\Omega}\right\}$$

where

$$\boldsymbol{\Omega} = \left(\boldsymbol{C}_\nu^\top \text{diag}\left(\boldsymbol{\psi}_{q(\boldsymbol{\omega})}\right)\boldsymbol{C}_\nu + \begin{bmatrix} \sigma_\beta^{-2}\boldsymbol{I}_2 & \boldsymbol{0} \\ \boldsymbol{0} & \mu_{q(1/\sigma_u^2)}\boldsymbol{I}_{K_u} \end{bmatrix}\right)^{-1}.$$

Therefore,

$$q^*(\boldsymbol{\nu}) \text{ is the } N\left(\boldsymbol{\mu}_{q(\boldsymbol{\nu})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\nu})}\right) \text{ density function,}$$

where

$$\boldsymbol{\mu}_{q(\boldsymbol{\nu})} = \boldsymbol{\Sigma}_{q(\boldsymbol{\nu})}\, \boldsymbol{C}_\nu^T \mathrm{diag}(\boldsymbol{\psi}_{q(\boldsymbol{\omega})})\, \boldsymbol{y},$$

and

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\nu})} = \left(\boldsymbol{C}_\nu^T \mathrm{diag}\left\{\boldsymbol{\psi}_{q(\boldsymbol{\omega})}\right\} \boldsymbol{C}_\nu + \begin{bmatrix} \sigma_{\boldsymbol{\beta}}^{-2} \boldsymbol{I}_2 & \boldsymbol{0} \\ \boldsymbol{0} & \mu_{q(1/\sigma_u^2)} \boldsymbol{I}_{K_u} \end{bmatrix} \right)^{-1}.$$

Expressions for $B_{q(\sigma_u^2)}$ and $\mu_{q(1/\sigma_u^2)}$

$$\begin{aligned} p(\sigma_u^2|\text{rest}) &\propto p(\boldsymbol{u}|\sigma_u^2) p(\sigma_u^2|a_u) \\ &= |\sigma_u^2 \boldsymbol{I}_{K_u}|^{-1/2} \exp\left\{-\tfrac{1}{2\sigma_u^2}\|\boldsymbol{u}\|^2\right\} \left(\sigma_u^2\right)^{-\frac{1}{2}-1} \exp\left\{(1/a_u)/\sigma_u^2\right\} + \text{const.} \end{aligned}$$

Taking the logarithm of both sides, we have

$$\begin{aligned} \log p(\sigma_u^2|\text{rest}) &= -\tfrac{1}{2} K_u \log(\sigma_u^2) - \tfrac{1}{2\sigma_u^2}\|\boldsymbol{u}\|^2 - \tfrac{3}{2}\log(\sigma_u^2) - (1/a_u)/\sigma_u^2 + \text{const.} \\ &= -\left\{\tfrac{1}{2}(K_u+1)+1\right\}\log(\sigma_u^2) - \left(\tfrac{1}{2}\|\boldsymbol{u}\|^2 + a_u^{-1}\right)/\sigma_u^2 + \text{const.} \end{aligned}$$

Taking expectations, we get

$$\begin{aligned} \log q^*(\sigma_u^2) &= E_q\left\{\log p(\sigma_u^2|\text{rest})\right\} + \text{const.} \\ &= -\left\{\tfrac{1}{2}(K_u+1)+1\right\}\log(\sigma_u^2) - \left(\tfrac{1}{2}E_q\|\boldsymbol{u}\|^2 + \mu_{q(1/a_u)}\right)/\sigma_u^2 + \text{const} \end{aligned}$$

Noting that $E_q\|\boldsymbol{u}\|^2 = \|\boldsymbol{\mu}_{q(\boldsymbol{u})}\|^2 + \mathrm{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{u})}\right)$, we have

$$\begin{aligned} \log q^*(\sigma_u^2) &= -\left\{\tfrac{1}{2}(K_u+1)+1\right\}\log(\sigma_u^2) - \left(\tfrac{1}{2}\left\{\|\boldsymbol{\mu}_{q(\boldsymbol{u})}\|^2 + \mathrm{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{u})}\right)\right\} + \mu_{q(1/a_u)}\right)/\sigma_u^2 \\ &\quad + \text{const} \end{aligned}$$

Therefore,

$$q^*(\sigma_u^2) \propto (\sigma_u^2)^{-\left\{\frac{1}{2}(K_u+1)+1\right\}} \exp\left(-\tfrac{1}{\sigma_u^2}\left[\tfrac{1}{2}\left\{\|\boldsymbol{\mu}_{q(\boldsymbol{u})}\|^2 + \mathrm{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{u})}\right)\right\} + \mu_{q(1/a_u)}\right]\right)$$

which is of the form of an Inverse-Gamma distribution with parameters

$$A_{q(\sigma_u^2)} = \tfrac{1}{2}(K_u+1), \quad \text{and}$$

$$B_{q(\sigma_u^2)} = \tfrac{1}{2}\left\{\|\boldsymbol{\mu}_{q(\boldsymbol{u})}\|^2 + \mathrm{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{u})}\right)\right\} + \mu_{q(1/a_u)}.$$

In addition, using Result 1.4.3,

$$\mu_{q(1/\sigma_u^2)} = \tfrac{1}{2}(K_u + 1)/B_{q(\sigma_u^2)}.$$

Expressions for $B_{q(\sigma_v^2)}$ and $\mu_{q(1/\sigma_v^2)}$

The derivation of $B_{q(\sigma_v^2)}$ and $\mu_{q(1/\sigma_v^2)}$ is similar to that for $q^*(\sigma_u^2)$. The optimal q density satisfies

$$q^*(\sigma_v^2) \quad \propto \quad (\sigma_v^2)^{-\left\{\frac{1}{2}(K_v+1)+1\right\}} \exp\left(-\tfrac{1}{\sigma_v^2}\left[\tfrac{1}{2}\left\{\|\boldsymbol{\mu}_{q(\boldsymbol{v})}\|^2 + \operatorname{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{v})}\right)\right\} + \mu_{q(1/a_v)}\right]\right)$$

which is of the form of an Inverse-Gamma distribution with parameters

$$A_{q(\sigma_u^2)} = \tfrac{1}{2}(K_v + 1) \quad \text{and}$$

$$B_{q(\sigma_v^2)} = \tfrac{1}{2}\left\{\|\boldsymbol{\mu}_{q(\boldsymbol{v})}\|^2 + \operatorname{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{v})}\right)\right\} + \mu_{q(1/a_v)}.$$

Also,

$$\mu_{q(1/\sigma_v^2)} = \tfrac{1}{2}(K_v + 1)/B_{q(\sigma_v^2)}.$$

Expressions for $B_{q(a_u)}$ and $\mu_{q(1/a_u)}$

$$\begin{aligned} p(a_u|\text{rest}) \quad &\propto \quad p(\sigma_u^2|a_u)p(a_u) \\ &= \quad (1/a_u)^{1/2} \exp\left\{-\left(1/a_u\right)/\sigma_u^2\right\} (a_u)^{-\frac{1}{2}-1} \exp\left\{-\left(1/A_u^2\right)/a_u\right\} + \text{const.} \end{aligned}$$

Taking the logarithm, we get

$$\log p(a_u|\text{rest}) \quad = \quad -2\log(a_u) - (\sigma_u^{-2} + A_u^{-2})/a_u + \text{const.}$$

Taking expectations:

$$\begin{aligned} \log q^*(a_u) \quad &= \quad E_q\left\{\log p(a_u|\text{rest})\right\} + \text{const} \\ &= \quad -2\log(a_u) - E_q\left(\sigma_u^{-2} + A_u^{-2}\right)/a_u + \text{const.} \\ &= \quad -2\log(a_u) - \left(\mu_{q(1/\sigma_u^2)} + A_u^{-2}\right)/a_u + \text{const} \end{aligned}$$

Therefore,

$$q^*(a_u) \propto (a_u)^{-2} \exp\left\{-\left(\mu_{q(1/\sigma_u^2)} + A_u^{-2}\right)/a_u\right\}$$

which is of the form of an Inverse-Gamma distribution with parameters

$$A_{q(a_u)} = 1, \quad \text{and}$$

$$B_{q(a_u)} = \mu_{q(1/\sigma_u^2)} + A_u^{-2}.$$

In addition, using Result 1.4.3,

$$\mu_{q(1/a_u)} = 1/B_{q(a_u)}.$$

Expressions for $B_{q(a_v)}$, and $\mu_{q(1/a_v)}$

The derivation of $B_{q(a_v)}$, and $\mu_{q(1/a_v)}$ is equivalent to that for $q^*(a_u)$. The optimal q density satisfies

$$q^*(a_v) \propto (a_v)^{-2} \exp\left\{-\left(\mu_{q(1/\sigma_v^2)} + A_v^{-2}\right)/a_v\right\}$$

which is of the form of an Inverse-Gamma distribution with parameters

$$A_{q(a_v)} = 1, \quad \text{and}$$

$$B_{q(a_v)} = \mu_{q(1/\sigma_v^2)} + A_v^{-2}.$$

Also,

$$\mu_{q(a_v)} = \mu_{q(1/a_v)} = 1/B_{q(a_v)}.$$

Derivation of the $\boldsymbol{\mu}_{q(\boldsymbol{\omega})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\omega})}$ updates

The updates for $\boldsymbol{\mu}_{q(\boldsymbol{\omega})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\omega})}$ are based on maximisation of the current value of the marginal log-likelihood lower bound $\log \underline{p}(\boldsymbol{y}; q, \boldsymbol{\mu}_{q(\boldsymbol{\omega})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\omega})})$ over these parameters using fixed point iteration. Wand (2014) shows that the updates reduce to

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

where

$$S \equiv E_q\left\{\log p(\boldsymbol{y}|\boldsymbol{\nu}, \boldsymbol{\omega}) + \log p(\boldsymbol{\omega}|\sigma_v^2)\right\},$$

D denotes derivative vector, as defined in Magnus & Neudecker (1999) and vec and vec^{-1} are as defined in Wand (2014). However, an equivalent expression for the first of these

updates is

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\omega})} \quad \leftarrow \quad \left(-\mathsf{H}_{\boldsymbol{\mu}_{q(\boldsymbol{\omega})}} S\right)^{-1} \tag{3.14}$$

where H denotes the Hessian matrix as defined in Magnus & Neudecker (1999). We work with this alternative form here. Proposition 3.A.1 gives justification for expression (3.14). Full details of the proof of Proposition 3.A.1 are given in Appendix 3.C.

Proposition 3.A.1. *Consider the $d \times 1$ random vector $\boldsymbol{x}$ which follows the multivariate normal distribution and has density function of the form $f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \equiv |\boldsymbol{\Sigma}|^{-\frac{1}{2}} exp\left\{\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$, then*

$$\tfrac{1}{2} H_{\boldsymbol{\mu}} f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = vec^{-1}\left(D_{vec(\boldsymbol{\Sigma})} f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\right).$$

An explicit expression for S is

$$\begin{aligned} S \quad \equiv \quad & -\mathbf{1}^T \boldsymbol{C}_\omega \boldsymbol{\mu}_{q(\boldsymbol{\omega})} - \boldsymbol{\mu}_{q(r_{\boldsymbol{\nu}}^2)}^T \exp\left\{-\boldsymbol{C}_\omega \boldsymbol{\mu}_{q(\boldsymbol{\omega})} + \tfrac{1}{2}\text{diagonal}\left(\boldsymbol{C}_\omega \boldsymbol{\Sigma}_{q(\boldsymbol{\omega})} \boldsymbol{C}_\omega^T\right)\right\} \\ & -\tfrac{1}{2}\text{tr}\left(\begin{bmatrix} \sigma_\gamma^{-2}\boldsymbol{I}_2 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\mu}_{q(1/\sigma_v^2)}\boldsymbol{I} \end{bmatrix}\left(\boldsymbol{\mu}_{q(\boldsymbol{\omega})}\boldsymbol{\mu}_{q(\boldsymbol{\omega})}^T + \boldsymbol{\Sigma}_{q(\boldsymbol{\omega})}\right)\right) \\ & -\left(n + \tfrac{K}{2} + 1\right)\log(2\pi) - \tfrac{K}{2}\log\left(\sigma_v^2\right) - \log\left(\sigma_\gamma^2\right). \end{aligned}$$

This leads to

$$\begin{aligned} d_{\boldsymbol{\mu}_{q(\boldsymbol{\omega})}} S \quad = \quad & -\mathbf{1}^T \boldsymbol{C}_\omega d\boldsymbol{\mu}_{q(\boldsymbol{\omega})} \\ & +\boldsymbol{\mu}_{q(r_{\boldsymbol{\nu}}^2)}^T \text{diag}\left(\exp\left\{-\boldsymbol{C}_\omega \boldsymbol{\mu}_{q(\boldsymbol{\omega})} + \tfrac{1}{2}\text{diagonal}\left(\boldsymbol{C}_\omega \boldsymbol{\Sigma}_{q(\boldsymbol{\omega})} \boldsymbol{C}_\omega^T\right)\right\}\right) \boldsymbol{C}_\omega d\boldsymbol{\mu}_{q(\boldsymbol{\omega})} \\ & -\boldsymbol{\mu}_{q(\boldsymbol{\omega})}^T \begin{bmatrix} \sigma_\gamma^{-2}\boldsymbol{I}_2 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\mu}_{q(1/\sigma_v^2)}\boldsymbol{I} \end{bmatrix} d\boldsymbol{\mu}_{q(\boldsymbol{\omega})} \\ = \quad & \left(\left[\boldsymbol{\mu}_{q(r_{\boldsymbol{\nu}}^2)} \odot \exp\left\{-\boldsymbol{C}_\omega \boldsymbol{\mu}_{q(\boldsymbol{\omega})} + \tfrac{1}{2}\text{diagonal}\left(\boldsymbol{C}_\omega \boldsymbol{\Sigma}_{q(\boldsymbol{\omega})} \boldsymbol{C}_\omega^T\right)\right\} - \mathbf{1}\right]^T \boldsymbol{C}_\omega \\ & -\boldsymbol{\mu}_{q(\boldsymbol{\omega})}^T \begin{bmatrix} \sigma_\gamma^{-2}\boldsymbol{I}_2 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\mu}_{q(1/\sigma_v^2)}\boldsymbol{I} \end{bmatrix}\Bigg) d\boldsymbol{\mu}_{q(\boldsymbol{\omega})}. \end{aligned}$$

Then, by Theorem 6, Chapter 5, of Magnus & Neudecker (1999),

$$\begin{aligned}\left(\mathsf{D}_{\boldsymbol{\mu}_{q(\boldsymbol{\omega})}}S\right)^T &= \boldsymbol{C}_\omega^T\left[\boldsymbol{\mu}_{q(r_{\boldsymbol{\nu}}^2)}\odot\exp\left\{-\boldsymbol{C}_\omega\boldsymbol{\mu}_{q(\boldsymbol{\omega})}+\tfrac{1}{2}\text{diagonal}\left(\boldsymbol{C}_\omega\boldsymbol{\Sigma}_{q(\boldsymbol{\omega})}\boldsymbol{C}_\omega^T\right)\right\}-\mathbf{1}\right]\\ &\quad-\begin{bmatrix}\sigma_\gamma^{-2}\boldsymbol{I}_2 & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\mu}_{q(1/\sigma_v^2)}\boldsymbol{I}\end{bmatrix}\boldsymbol{\mu}_{q(\boldsymbol{\omega})}.\end{aligned}$$

Next,

$$\begin{aligned}d^2\boldsymbol{\mu}_{q(\boldsymbol{\omega})}S &= \Bigg(\boldsymbol{\mu}_{q(r_{\boldsymbol{\nu}}^2)}\odot\exp\Big[\left(d\boldsymbol{\mu}_{q(\boldsymbol{\omega})}\right)^T\left\{-\boldsymbol{C}_\omega\boldsymbol{\mu}_{q(\boldsymbol{\omega})}+\tfrac{1}{2}\text{diagonal}\left(\boldsymbol{C}_\omega\boldsymbol{\Sigma}_{q(\boldsymbol{\omega})}\boldsymbol{C}_\omega^T\right)\right\}\Big]^T\boldsymbol{C}_\omega\\ &\qquad-\left(d\boldsymbol{\mu}_{q(\boldsymbol{\omega})}\right)^T\begin{bmatrix}\sigma_\gamma^{-2}\boldsymbol{I}_2 & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\mu}_{q(1/\sigma_v^2)}\boldsymbol{I}\end{bmatrix}\Bigg)d\boldsymbol{\mu}_{q(\boldsymbol{\omega})}\\ &= \Big[\Big\{\boldsymbol{\mu}_{q(r_{\boldsymbol{\nu}}^2)}\odot\Big(\text{diag}\Big[\exp\left\{-\boldsymbol{C}_\omega\boldsymbol{\mu}_{q(\boldsymbol{\omega})}+\tfrac{1}{2}\text{diagonal}\left(\boldsymbol{C}_\omega\boldsymbol{\Sigma}_{q(\boldsymbol{\omega})}\boldsymbol{C}_\omega^T\right)\right\}\Big]\\ &\qquad\times\left(\boldsymbol{C}_\omega d\boldsymbol{\mu}_{q(\boldsymbol{\omega})}\right)\Big)\Big\}^T\boldsymbol{C}_\omega-\left(d\boldsymbol{\mu}_{q(\boldsymbol{\omega})}\right)^T\begin{bmatrix}\sigma_\gamma^{-2}\boldsymbol{I}_2 & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\mu}_{q(1/\sigma_v^2)}\boldsymbol{I}\end{bmatrix}\Big]d\boldsymbol{\mu}_{q(\boldsymbol{\omega})}\end{aligned}$$

Then, using Result 1.4.11,

$$\begin{aligned}d^2\boldsymbol{\mu}_{q(\boldsymbol{\omega})}S &= \Big\{\Big(\text{diag}\left(\boldsymbol{\mu}_{q(r_{\boldsymbol{\nu}}^2)}\right)\text{diag}\Big[\exp\left\{-\boldsymbol{C}_\omega\boldsymbol{\mu}_{q(\boldsymbol{\omega})}+\tfrac{1}{2}\text{diagonal}\left(\boldsymbol{C}_\omega\boldsymbol{\Sigma}_{q(\boldsymbol{\omega})}\boldsymbol{C}_\omega^T\right)\right\}\Big]\\ &\qquad\times\left(-\boldsymbol{C}_\omega d\boldsymbol{\mu}_{q(\boldsymbol{\omega})}\right)\Big)^T\boldsymbol{C}_\omega-\left(d\boldsymbol{\mu}_{q(\boldsymbol{\omega})}\right)^T\begin{bmatrix}\sigma_\gamma^{-2}\boldsymbol{I}_2 & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\mu}_{q(1/\sigma_v^2)}\boldsymbol{I}\end{bmatrix}\Big\}d\boldsymbol{\mu}_{q(\boldsymbol{\omega})}\\ &= \left(d\boldsymbol{\mu}_{q(\boldsymbol{\omega})}\right)^T\Big(-\boldsymbol{C}_\omega^T\text{diag}\big[\boldsymbol{\mu}_{q(r_{\boldsymbol{\nu}}^2)}\odot\exp\big\{-\boldsymbol{C}_\omega\boldsymbol{\mu}_{q(\boldsymbol{\omega})}\\ &\qquad+\tfrac{1}{2}\text{diagonal}\big(\boldsymbol{C}_\omega\boldsymbol{\Sigma}_{q(\boldsymbol{\omega})}\boldsymbol{C}_\omega^T\big)\big\}\big]\Big)\boldsymbol{C}_\omega d\boldsymbol{\mu}_{q(\boldsymbol{\omega})}\\ &\qquad-\left(d\boldsymbol{\mu}_{q(\boldsymbol{\omega})}\right)^T\begin{bmatrix}\sigma_\gamma^{-2}\boldsymbol{I}_2 & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\mu}_{q(1/\sigma_v^2)}\boldsymbol{I}\end{bmatrix}d\boldsymbol{\mu}_{q(\boldsymbol{\omega})}\end{aligned}$$

Therefore, by Theorem 6, Chapter 6, of Magnus & Neudecker (1999),

$$\begin{aligned}-\mathsf{H}_{\boldsymbol{\mu}_{q(\boldsymbol{\omega})}}S &= \boldsymbol{C}_\omega^T\text{diag}\left[\boldsymbol{\mu}_{q(r_{\boldsymbol{\nu}}^2)}\odot\exp\left\{\boldsymbol{C}_\omega\boldsymbol{\mu}_{q(\boldsymbol{\omega})}+\tfrac{1}{2}\text{diagonal}\left(\boldsymbol{C}_\omega\boldsymbol{\Sigma}_{q(\boldsymbol{\omega})}\boldsymbol{C}_\omega^T\right)\right\}\right]\boldsymbol{C}_\omega\\ &\quad+\begin{bmatrix}\sigma_\gamma^{-2}\boldsymbol{I}_2 & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\mu}_{q(1/\sigma_v^2)}\boldsymbol{I}\end{bmatrix}.\end{aligned}$$

Combining these expressions leads to the updates in Algorithm 2.

3.B Derivation for the marginal log-likelihood lower bound

The expression for the lower bound on the marginal log-likelihood given in (3.14) is

$$\begin{aligned}
\log \underline{p}(\boldsymbol{y}; q, \boldsymbol{\mu}_{q(\boldsymbol{\omega})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\omega})}) &= \tfrac{1}{2}(K_u + K_v + 4) - \tfrac{n}{2}\log(2\pi) + \log\Gamma\left(\tfrac{1}{2}(K_u+1)\right) - 2\log(\pi) \\
&\quad + \log\Gamma\left(\tfrac{1}{2}(K_v+1)\right) - \log(A_u) - \log(A_v) - \tfrac{1}{2}\mathbf{1}^T(\boldsymbol{C}_\omega \boldsymbol{\mu}_{q(\boldsymbol{\omega})}) \\
&\quad - \tfrac{1}{2}\mathbf{1}^\top\left\{\boldsymbol{\mu}_{q(r_{\boldsymbol{\nu}}^2)} \odot \exp\left(\boldsymbol{\psi}_{q(\boldsymbol{\omega})}\right)\right\} - \log(\sigma_\beta^2) - \log(\sigma_\gamma^2) \\
&\quad + \tfrac{1}{2}\log|\Sigma_{q(\boldsymbol{\nu})}| + \tfrac{1}{2}\log|\Sigma_{q(\boldsymbol{\omega})}| - \tfrac{1}{2\sigma_\beta^2}\left(\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\right) \\
&\quad - \tfrac{1}{2\sigma_\gamma^2}\left\{\|\boldsymbol{\mu}_{q(\boldsymbol{\gamma})}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\gamma})})\right\} - \tfrac{1}{2}(K_u+1)\log(B_{q(\sigma_u^2)}) \\
&\quad - \tfrac{1}{2}(K_v+1)\log(B_{q(\sigma_v^2)}) - \log\left(\mu_{q(1/\sigma_u^2)} + A_u^{-2}\right) \\
&\quad - \log\left(\mu_{q(1/\sigma_v^2)} + A_v^{-2}\right) + \mu_{q(1/\sigma_u^2)}\mu_{q(1/a_u)} + \mu_{q(1/\sigma_v^2)}\mu_{q(1/a_v)}.
\end{aligned}$$

The derivation of the logarithm lower bound on the marginal likelihood is due to

$$\begin{aligned}
\log \underline{p}(\boldsymbol{y}; q) &= E_q\left\{\log p(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\gamma}, \boldsymbol{v}, \sigma_u^2, \sigma_v^2, a_u, a_v) - \log q^*(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\gamma}, \boldsymbol{v}, \sigma_u^2, \sigma_v^2, a_u, a_v)\right\} \\
&= E_q\left\{\log p(\boldsymbol{y}, \boldsymbol{\nu}, \boldsymbol{\omega}, \sigma_u^2, \sigma_v^2, a_u, a_v) - \log q^*(\boldsymbol{\nu}, \boldsymbol{\omega}, \sigma_u^2, \sigma_v^2, a_u, a_v)\right\}.
\end{aligned}$$

We can further factorise the true joint density since the distribution of $\boldsymbol{y}$ does not depend on $\sigma_u^2, \sigma_v^2, a_u$ or a_v:

$$p(\boldsymbol{y}, \boldsymbol{\nu}, \boldsymbol{\omega}, \sigma_u^2, \sigma_v^2, a_u, a_v) = p(\boldsymbol{y}|\boldsymbol{\nu}, \boldsymbol{\omega})\, p(\boldsymbol{\nu}, \boldsymbol{\omega}, \sigma_u^2, \sigma_v^2, a_u, a_v).$$

Continuing in this fashion, we get

$$\begin{aligned}
p(\boldsymbol{y}, \boldsymbol{\nu}, \boldsymbol{\omega}, \sigma_u^2, \sigma_v^2, a_u, a_v) &= p(\boldsymbol{y}|\boldsymbol{\nu}, \boldsymbol{\omega})\, p(\boldsymbol{\nu}|\sigma_u^2)\, p(\boldsymbol{\omega}, \sigma_u^2, \sigma_v^2, a_u, a_v) \\
&= p(\boldsymbol{y}|\boldsymbol{\nu}, \boldsymbol{\omega})\, p(\boldsymbol{\nu}|\sigma_u^2)\, p(\boldsymbol{\omega}|\sigma_v^2)\, p(\sigma_u^2, \sigma_v^2, a_u, a_v) \\
&= p(\boldsymbol{y}|\boldsymbol{\nu}, \boldsymbol{\omega})\, p(\boldsymbol{\nu}|\sigma_u^2)\, p(\boldsymbol{\omega}|\sigma_v^2)\, p(\sigma_u^2)\, p(\sigma_v^2)\, p(a_u)\, p(a_v)
\end{aligned}$$

since the distribution of $\boldsymbol{\nu}$ does not depend on $\boldsymbol{\omega}, \sigma_v^2, a_u$ or a_v, and the distribution of $\boldsymbol{\omega}$ does not depend on a_u or a_v. Also, $\sigma_u^2, \sigma_v^2, a_u$ and a_v are independent.

The approximate joint density also factorises according to (3.10):

$$q(\boldsymbol{\nu}, \boldsymbol{\omega}, \sigma_u^2, \sigma_v^2, a_u, a_v) = q(\boldsymbol{\nu})\, q(\boldsymbol{\omega})\, q(\sigma_u^2)\, q(\sigma_v^2)\, q(a_u)\, q(a_v).$$

Using the above factorisations, the new expression for the logarithm lower bound is

now:

$$\begin{aligned}\log \underline{p}(\boldsymbol{y};q) &= E_q\left\{\log p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\omega})\right\} + E_q\left\{\log p(\boldsymbol{\nu}|\sigma_u^2) - q^*(\boldsymbol{\nu})\right\} \\ &\quad + E_q\left\{\log p(\boldsymbol{\omega}|\sigma_v^2) - \log q^*(\boldsymbol{\omega})\right\} + E_q\left\{\log p(\sigma_u^2|a_u) - \log q^*(\sigma_u^2)\right\} \\ &\quad + E_q\left\{\log p(\sigma_v^2|a_v) - \log q^*(\sigma_v^2)\right\} + E_q\left\{\log p(a_u) - \log q^*(a_u)\right\} \\ &\quad + E_q\left\{\log p(a_v) - \log q^*(a_v)\right\}.\end{aligned}$$

The derivation of each component is given as follows.

Expression for $E_q\left\{\log p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\omega})\right\}$

Firstly,

$$\begin{aligned}p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\omega}) &= (2\pi)^{-n/2}\left|\operatorname{diag}\left\{\exp(\boldsymbol{C}_\omega\boldsymbol{\omega})\right\}\right|^{-1/2} \\ &\quad \exp\left\{-\tfrac{1}{2}\left(\boldsymbol{y}-\boldsymbol{C}_\nu\boldsymbol{\nu}\right)^\top \operatorname{diag}\left\{\exp(-\boldsymbol{C}_\omega\boldsymbol{\omega})\right\}\right. \\ &\quad \left.\left(\boldsymbol{y}-\boldsymbol{C}_\nu\boldsymbol{\nu}\right)\right\}.\end{aligned}$$

Note that

$$\begin{aligned}\left|\operatorname{diag}\left\{\exp(\boldsymbol{C}_\omega\boldsymbol{\omega})\right\}\right| &= \begin{bmatrix} \exp\left\{(\boldsymbol{C}_\omega\boldsymbol{\omega})_1\right\} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \exp\left\{(\boldsymbol{C}_\omega\boldsymbol{\omega})_n\right\} \end{bmatrix} \\ &= \exp\left\{(\boldsymbol{C}_\omega\boldsymbol{\omega})_1\right\} \times \ldots \times \exp\left\{(\boldsymbol{C}_\omega\boldsymbol{\omega})_n\right\} \\ &= \exp\left\{\boldsymbol{1}^\top\left(\boldsymbol{C}_\omega\boldsymbol{\omega}\right)\right\}\end{aligned}$$

and

$$\begin{aligned}&-\tfrac{1}{2}\left(\boldsymbol{y}-\boldsymbol{C}_\nu\boldsymbol{\nu}\right)^\top \operatorname{diag}\left\{\exp(-\boldsymbol{C}_\omega\boldsymbol{\omega})\right\}\left(\boldsymbol{y}-\boldsymbol{C}_\nu\boldsymbol{\nu}\right) \\ &= -\tfrac{1}{2}\left(\boldsymbol{C}_\nu\boldsymbol{\nu}\right)^\top \operatorname{diag}\left\{\exp\left(-\boldsymbol{C}_\omega\boldsymbol{\omega}\right)\right\}\left(\boldsymbol{C}_\nu\boldsymbol{\nu}\right) \\ &= -\tfrac{1}{2} r_{\boldsymbol{\nu}}^\top \operatorname{diag}\left\{\exp\left(-\boldsymbol{C}_\omega\boldsymbol{\omega}\right)\right\} r_{\boldsymbol{\nu}} \\ &= -\tfrac{1}{2}\boldsymbol{1}^\top\left\{r_{\boldsymbol{\nu}}^2 \odot \exp\left(-\boldsymbol{C}_\omega\boldsymbol{\omega}\right)\right\}\end{aligned}$$

where

$$\begin{aligned}r_{\boldsymbol{\nu}} &= \boldsymbol{C}_\nu\boldsymbol{\nu}, \quad \text{and} \\ r_{\boldsymbol{\nu}}^2 &= \operatorname{diagonal}\left\{\left(\boldsymbol{C}_\nu\boldsymbol{\nu}\right)\left(\boldsymbol{C}_\nu\boldsymbol{\nu}\right)^\top\right\}.\end{aligned}$$

Therefore

$$\log p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\omega}) = -\tfrac{n}{2}\log(2\pi) - \tfrac{1}{2}\boldsymbol{1}^\top\left(\boldsymbol{C}_\omega\boldsymbol{\omega}\right) - \tfrac{1}{2}\boldsymbol{1}^\top\left\{r_{\boldsymbol{\nu}}^2 \odot \exp\left(-\boldsymbol{C}_\omega\boldsymbol{\omega}\right)\right\}.$$

Taking expectations we get

$$E_q\left\{\log p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\omega})\right\} = -\tfrac{n}{2}\log(2\pi) - \tfrac{1}{2}\boldsymbol{1}^\top\left(\boldsymbol{C}_\omega\boldsymbol{\mu}_{q(\boldsymbol{\omega})}\right) - \tfrac{1}{2}\boldsymbol{1}^\top\left\{\boldsymbol{\mu}_{q(r_{\boldsymbol{\nu}}^2)} \odot \exp\left(\boldsymbol{\psi}_{q(\boldsymbol{\omega})}\right)\right\},$$

where

$$\begin{aligned}\boldsymbol{\mu}_{q(r_\nu^2)} &= E_q\left[\text{diagonal}\left\{(\boldsymbol{C}_\nu\boldsymbol{\nu})(\boldsymbol{C}_\nu\boldsymbol{\nu})^\top\right\}\right] \\ &= \text{diagonal}\left[E_q\left\{(\boldsymbol{C}_\nu\boldsymbol{\nu})(\boldsymbol{C}_\nu\boldsymbol{\nu})^\top\right\}\right] \\ &= \text{diagonal}\left\{\left(\boldsymbol{C}_\nu\boldsymbol{\mu}_{q(\boldsymbol{\nu})}\right)\left(\boldsymbol{C}_\nu\boldsymbol{\mu}_{q(\boldsymbol{\nu})}\right)^\top + \boldsymbol{C}_\nu\boldsymbol{\Sigma}_{q(\boldsymbol{\nu})}\boldsymbol{C}_\nu^\top\right\}.\end{aligned}$$

Expression for $E_q\left\{\log p(\boldsymbol{\nu}|\sigma_u^2) - \log q^*(\boldsymbol{\nu})\right\}$

First note that

$$\log p(\boldsymbol{\nu}|\sigma_u^2) = -\log(2\pi) - (\tfrac{k}{2}+1)\log(2\pi) - \log(\sigma_\beta^2) - \tfrac{k}{2}\log(\sigma_u^2) - \tfrac{1}{2\sigma_\beta^2}\|\boldsymbol{\beta}\|^2 - \tfrac{1}{2\sigma_u^2}\|\boldsymbol{u}\|^2.$$

Taking expectations we get

$$\begin{aligned}E_q\left\{\log p(\boldsymbol{\nu}|\sigma_u^2)\right\} &= -\log(2\pi) - (\tfrac{k}{2}+1)\log(2\pi) - \log(\sigma_\beta^2) - \tfrac{k}{2}E_q\left\{\log(\sigma_u^2)\right\} \\ &\quad -\tfrac{1}{2\sigma_\beta^2}E_q\left(\|\boldsymbol{\beta}\|^2\right) - \tfrac{1}{2\sigma_u^2}E_q\left(\|\boldsymbol{u}\|^2\right) \\ &= -\log(2\pi) - (\tfrac{k}{2}+1)\log(2\pi) - \log(\sigma_\beta^2) - \tfrac{k}{2}E_q\left\{\log(\sigma_u^2)\right\} \\ &\quad -\tfrac{1}{2\sigma_\beta^2}\left\{\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \text{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right)\right\} - \tfrac{1}{2}\mu_{q(1/\sigma_u^2)}\left\{\|\boldsymbol{\mu}_{q(\boldsymbol{u})}\|^2 + \text{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{u})}\right)\right\}.\end{aligned}$$

Also,

$$E_q\left\{\log q^*(\boldsymbol{\nu})\right\} = \tfrac{1}{2}(K_u+2)\log(2\pi) + \tfrac{1}{2}(K_u+2) + \tfrac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\nu})}|.$$

Thus, we get

$$\begin{aligned}E_q\left\{\log p(\boldsymbol{\nu}|\sigma_u^2) - \log q^*(\boldsymbol{\nu})\right\} &= -\log(\sigma_\beta^2) - \tfrac{1}{2}K_u\,E_q\left(\log(\sigma_u^2)\right) + \tfrac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\nu})}| \\ &\quad -\tfrac{1}{2\sigma_\beta^2}\left\{\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \text{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right)\right\} + \tfrac{1}{2}(K_u+2) \\ &\quad -\tfrac{1}{2}\mu_{q(1/\sigma_u^2)}\left\{\|\boldsymbol{\mu}_{q(\boldsymbol{u})}\|^2 + \text{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{u})}\right)\right\}.\end{aligned}$$

Expression for $E_q\left\{\log p(\boldsymbol{\omega}|\sigma_v^2) - \log q^*(\boldsymbol{\omega})\right\}$

Similar to the previous derivation we get

$$\begin{aligned}E_q\left\{\log p(\boldsymbol{\omega}|\sigma_v^2) - \log q^*(\boldsymbol{\omega})\right\} &= -\log(\sigma_\gamma^2) - \tfrac{1}{2}K_v\,E_q\left(\log(\sigma_v^2)\right) + \tfrac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\omega})}| \\ &\quad -\tfrac{1}{2\sigma_\gamma^2}\left\{\|\boldsymbol{\mu}_{q(\boldsymbol{\gamma})}\|^2 + \text{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\gamma})}\right)\right\} + \tfrac{1}{2}(K_v+2) \\ &\quad -\tfrac{1}{2}\mu_{q(1/\sigma_v^2)}\left\{\|\boldsymbol{\mu}_{q(\boldsymbol{v})}\|^2 + \text{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{v})}\right)\right\}.\end{aligned}$$

Expression for $E_q\left\{\log p(\sigma_u^2|a_u) - \log q^*(\sigma_u^2)\right\}$

Firstly,

$$\begin{aligned}\log p(\sigma_u^2|a_u) - \log q^*(\sigma_u^2) &= \log\left\{\frac{\left(\frac{1}{a_u}\right)^{\frac{1}{2}}}{\Gamma\left(\frac{1}{2}\right)}\left(\sigma_u^2\right)^{-\frac{1}{2}-1}\exp\left(-\frac{\frac{1}{a_u}}{\sigma_u^2}\right)\right\} \\ &\quad - \log\left\{\frac{B_{q(\sigma_u^2)}^{\frac{1}{2}(K_u+1)}}{\Gamma\left(\frac{1}{2}(K_u+1)\right)}(\sigma_u^2)^{-\frac{1}{2}(K_u+1)-1}\exp\left(-\frac{B_{q(\sigma_u^2)}}{\sigma_u^2}\right)\right\} \\ &= -\tfrac{1}{2}\log(a_u) + \log\Gamma\left\{\tfrac{1}{2}(K_u+1)\right\} - \tfrac{1}{2}\log(\pi) + \tfrac{1}{2}K_u\log(\sigma_u^2) \\ &\quad + \left(1/\sigma_u^2\right)\left(B_{q(\sigma_u^2)} - a_u^{-1}\right) - \tfrac{1}{2}\left(K_u+1\right)\log\left(B_{q(\sigma_u^2)}\right).\end{aligned}$$

Taking expectations we get

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

Expression for $E_q\left\{\log p(\sigma_v^2|a_v) - \log q^*(\sigma_v^2)\right\}$

Similar to the previous derivation we get

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

Expression $E_q\left\{\log p(a_u) - \log q^*(a_u)\right\}$

The difference of the logarithms are

$$\begin{aligned}\log p(a_u) - \log q^*(a_u) &= \log\left\{\frac{\frac{1}{A_u^2}^{\frac{1}{2}}}{\Gamma(\frac{1}{2})}(a_u)^{-\frac{1}{2}-1}\exp\left(-\frac{\frac{1}{A_u^2}}{a_u}\right)\right\} \\ &\quad - \log\left\{\frac{B_{q(a_u)}}{\Gamma(1)}(a_u)^{-2}\exp\left(-\frac{B_{q(a_u)}}{a_u}\right)\right\} \\ &= -\log(A_u) - \log\Gamma(\tfrac{1}{2}) + \tfrac{1}{2}\log(a_u) + (1/a_u)\left(B_{q(a_u)} - A_u^{-2}\right) \\ &\quad - \log(B_{q(a_u)}).\end{aligned}$$

Then, substituting $B_{q(a_u)}$ from Algorithm 2 gives

$$\begin{aligned}\log p(a_u) - \log q^*(a_u) &= -\log(A_u) - \tfrac{1}{2}\log(\pi) + \tfrac{1}{2}\log(a_u) + \mu_{q(1/\sigma_u^2)}/a_u \\ &\quad - \log(\mu_{q(1/\sigma_u^2)} + A_u^{-2}).\end{aligned}$$

Taking expectations, we get

$$\begin{aligned}E_q\left\{\log p(a_u) - \log q^*(a_u)\right\} &= -\log(A_u) - \tfrac{1}{2}\log(\pi) + \tfrac{1}{2}E_q\left\{\log(a_u)\right\} + \mu_{q(1/\sigma_u^2)}\mu_{q(1/a_u)} \\ &\quad - \log(\mu_{q(1/\sigma_u^2)} + A_u^{-2}).\end{aligned}$$

Expression $E_q\{\log p(a_v) - \log q^*(a_v)\}$

Similar to the previous derivation we get

$$\begin{aligned} E_q\{\log p(a_v) - \log q^*(a_v)\} &= -\log(A_v) - \tfrac{1}{2}\log(\pi) + \tfrac{1}{2}E_q\{\log(a_v)\} + \mu_{q(1/\sigma_v^2)}\mu_{q(1/a_v)} \\ &\quad - \log(\mu_{q(1/\sigma_v^2)} + A_v^{-2}). \end{aligned}$$

Adding all expressions together, we get the lower bound expression

$$\begin{aligned} \log \underline{p}(\boldsymbol{y}; q, \boldsymbol{\mu}_{q(\boldsymbol{\omega})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\omega})}) &= \tfrac{1}{2}(K_u + K_v + 4) - \tfrac{n}{2}\log(2\pi) + \log\Gamma\left(\tfrac{1}{2}(K_u+1)\right) - 2\log(\pi) \\ &\quad + \log\Gamma\left(\tfrac{1}{2}(K_v+1)\right) - \log(A_u) - \log(A_v) - \tfrac{1}{2}\mathbf{1}^T(\boldsymbol{C}_\omega \boldsymbol{\mu}_{q(\boldsymbol{\omega})}) \\ &\quad - \tfrac{1}{2}\mathbf{1}^\top\left\{\boldsymbol{\mu}_{q(r_{\boldsymbol{\nu}}^2)} \odot \exp\left(\boldsymbol{\psi}_{q(\boldsymbol{\omega})}\right)\right\} - \log(\sigma_\beta^2) - \log(\sigma_\gamma^2) \\ &\quad + \tfrac{1}{2}\log|\Sigma_{q(\boldsymbol{\nu})}| + \tfrac{1}{2}\log|\Sigma_{q(\boldsymbol{\omega})}| - \tfrac{1}{2\sigma_\beta^2}\left(\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\right) \\ &\quad - \tfrac{1}{2\sigma_\gamma^2}\left\{\|\boldsymbol{\mu}_{q(\boldsymbol{\gamma})}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\gamma})})\right\} - \tfrac{1}{2}(K_u+1)\log(B_{q(\sigma_u^2)}) \\ &\quad - \tfrac{1}{2}(K_v+1)\log(B_{q(\sigma_v^2)}) - \log\left(\mu_{q(1/\sigma_u^2)} + A_u^{-2}\right) \\ &\quad - \log\left(\mu_{q(1/\sigma_v^2)} + A_v^{-2}\right) + \mu_{q(1/\sigma_u^2)}\mu_{q(1/a_u)} + \mu_{q(1/\sigma_v^2)}\mu_{q(1/a_v)}. \end{aligned}$$

Note the following cancellations:

$$-\tfrac{1}{2}\mu_{q(1/\sigma_u^2)}\left\{\|\boldsymbol{\mu}_{q(\boldsymbol{u})}\|^2 + \mathrm{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{u})}\right)\right\} + \mu_{q(1/\sigma_u^2)}\left(B_{q(\sigma_u^2)} - \mu_{q(1/a_u)}\right),$$

where

$$B_{q(\sigma_u^2)} = \tfrac{1}{2}\left\{\|\boldsymbol{\mu}_{q(\boldsymbol{u})}\|^2 + \mathrm{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{u})}\right)\right\} + \mu_{q(1/a_u)}$$

and

$$-\tfrac{1}{2}\mu_{q(1/\sigma_v^2)}\left\{\|\boldsymbol{\mu}_{q(\boldsymbol{v})}\|^2 + \mathrm{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{v})}\right)\right\} + \mu_{q(1/\sigma_v^2)}\left(B_{q(\sigma_v^2)} - \mu_{q(1/a_v)}\right),$$

where

$$B_{q(\sigma_v^2)} = \tfrac{1}{2}\left\{\|\boldsymbol{\mu}_{q(\boldsymbol{v})}\|^2 + \mathrm{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{v})}\right)\right\} + \mu_{q(1/a_v)}.$$

3.C Proof of Proposition 3.A.1

We begin by expanding the first differential with respect to vec $(\boldsymbol{\Sigma})$ in the right-hand side of Proposition 3.A.1:

$$\begin{aligned} d_{\mathrm{vec}(\boldsymbol{\Sigma})} f(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \left(d|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\right)\exp\left\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \\ &\quad +|\boldsymbol{\Sigma}|^{-\frac{1}{2}} d\left[\exp\left\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}\right] \\ &= -\tfrac{1}{2}|\boldsymbol{\Sigma}|^{-\frac{3}{2}} d|\boldsymbol{\Sigma}|\exp\left\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \\ &\quad +|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\left(-\tfrac{1}{2}\mathrm{tr}\left\{(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^{\top} d\boldsymbol{\Sigma}^{-1}\right\}\right) \\ &\quad \times\exp\left\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}. \end{aligned}$$

Using Results 1.4.1 (b) and (c), we get

$$\begin{aligned} d_{\mathrm{vec}(\boldsymbol{\Sigma})} f(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) &= -\tfrac{1}{2}|\boldsymbol{\Sigma}|^{-\frac{3}{2}}|\boldsymbol{\Sigma}|\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}d\boldsymbol{\Sigma}\right)\exp\left\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \\ &\quad -\tfrac{1}{2}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\mathrm{tr}\left\{(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\left(-\boldsymbol{\Sigma}^{-1}(d\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\right)\right\} \\ &\quad \times\exp\left\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}. \end{aligned}$$

Next, using Results 1.4.13 and 1.4.1 (a), we see that

$$\begin{aligned} d_{\mathrm{vec}(\boldsymbol{\Sigma})} f(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) &= -\tfrac{1}{2}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}\mathrm{vec}(\boldsymbol{\Sigma}^{-1})d\mathrm{vec}(\boldsymbol{\Sigma}) \\ &\quad +\tfrac{1}{2}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \\ &\quad \times\mathrm{vec}(\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1})d\mathrm{vec}(\boldsymbol{\Sigma}). \end{aligned}$$

Thus,

$$\begin{aligned} \mathrm{vec}^{-1}\left(\mathsf{D}_{\mathrm{vec}(\boldsymbol{\Sigma})} f(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})\right) &= -\tfrac{1}{2}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}\boldsymbol{\Sigma}^{-1} \\ &\quad +\tfrac{1}{2}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}\boldsymbol{\Sigma}^{-1} \\ &\quad \times(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1} \\ &= -\tfrac{1}{2}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \\ &\quad \times\boldsymbol{\Sigma}^{-1}\left\{\boldsymbol{I}-(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}\right\}. \end{aligned}$$

Now expanding the left-hand side of Proposition 3.A.1 we have

$$\begin{aligned} d_{\boldsymbol{\mu}} f(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \left(d|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\right)\exp\left\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \\ &\quad +|\boldsymbol{\Sigma}|^{-\frac{1}{2}} d\left[\exp\left\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}\right] \\ &= |\boldsymbol{\Sigma}|^{-\frac{1}{2}}\left[-\tfrac{1}{2}\left\{-(d\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})+(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(-d\boldsymbol{\mu})\right\}\right] \\ &\quad \times\exp\left\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \\ &= |\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}d\boldsymbol{\mu}. \end{aligned}$$

Next, we take the second differential with respect to $\boldsymbol{\mu}$:

$$\begin{aligned} d^2_{\boldsymbol{\mu}} f(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) &= d\boldsymbol{\mu}\left[|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}\right]d\boldsymbol{\mu} \\ &= \left[d\boldsymbol{\mu}\left\{|\boldsymbol{\Sigma}|^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}\right\}\exp\left\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}\right. \\ &\quad \left. +|\boldsymbol{\Sigma}|^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}d\boldsymbol{\mu}\left[\exp\left\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}\right]\right]d\boldsymbol{\mu} \\ &= \left[|\boldsymbol{\Sigma}|^{-\frac{1}{2}}(-d\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}\exp\left\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}\right. \\ &\quad +|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(d\boldsymbol{\mu})^{\top} \\ &\quad \left.\times\exp\left\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}\right]d\boldsymbol{\mu} \\ &= (d\boldsymbol{\mu})^{\top}\left[-|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}\boldsymbol{\Sigma}^{-1}\right. \\ &\quad \left.\times\left\{\boldsymbol{I}-(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}\right\}\right]d\boldsymbol{\mu}. \end{aligned}$$

Therefore,

$$\mathsf{H}_{\boldsymbol{\mu}} f(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = -|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}\boldsymbol{\Sigma}^{-1}\left\{\boldsymbol{I}-(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}\right\}.$$

Proposition 3.A.1 follows immediately. This proposition was inspired by a result in the appendix of Opper & Archambeau (2009).

Chapter 4

Heteroscedastic semiparametric regression extensions

4.1 Introduction

The 2010s has engendered an unprecedented explosion of data, spurring a need for faster and increasingly flexible techniques to extrapolate and make sense of such data. The use of semiparametric regression (e.g. Ruppert *et al.*, 2003, 2009) provides a significant collection of tractable statistical techniques, however more inline with datasets comprising moderate sample sizes and batch processing. Of recent, Luts *et al.* (2013) developed algorithms to perform semiparametric regression analysis in real time, where the data is processed as it is collected and promptly made available through modern technologies. This new nonparametric regression methodology is especially geared toward high volume/velocity data. Part of this chapter extends their approach to adjust for possible heteroscedasticity.

Chapter 3 has shown that ignorance of significant heteroscedasticity is likely to lead to inaccurate inferential statistics. A remedy for this is the employment of heteroscedastic nonparametric regression. In this chapter we explore two other extensions to the univariate nonparametric regression model. Here we instead focus on bivariate predictor nonparametric regression and additive models. The modularity of our approach is utilised here since the extension to bivariate and additive models has been relatively straightforward.

The content of this chapter is published as: Menictas, M and Wand. M.P. (2015). Variational inference for heteroscedastic semiparametric regression. *Australian and New Zealand Journal of Statistics*, **57**, Number 1, 119–138. This research was also presented at the *Australian Statistical Conference in conjunction with the Institute of Mathematical Statistics Annual Meeting*, Sydney, 2014.

Much like in Chapter 3 we take a Bayesian approach here as well by implementing a non-conjugate MFVB strategy.

The following section describes the bivariate nonparametric regression extension methodology and illustrates this approach on simulated and real data. Accuracy scores and time comparisons are also given. Section 4.3 describes the additive model extension and also provides some illustrations on real data. Finally in Section 4.4 we extend the work of Luts *et al.* (2013) to accommodate heteroscedasticity.

4.2 Bivariate predictor nonparametric regression

Bivariate nonparametric regression is of particular interest to a number of disciplines, such as geology, oceanography, public health and mining. This diverse area of statistics deals with tractable smoothing of *point clouds* where we would be interested in the mean response as a bivariate function of, for instance, the longitude and latitude of a certain geographical location. Bivariate smoothing and *geostatistics* have many similarities, for instance, geostatistics is concerned with converting geographically referenced observations to increasingly more legible maps (e.g. Cressie, 1993) and hence deals with a bivariate predictor component.

As seen in previous chapters, the use of penalized spline smoothing requires a set of basis functions that enables the possibility of nonlinear structure. Bivariate smoothing on the other hand requires bivariate basis functions. This extension can be treated in at least two specific ways: *(i) tensor product bases* and *(ii) radial basis functions*. More detail of these approaches is given in Ruppert *et al.* (2003). A possible difficulty of *tensor product bases* is their dependence on the position of the coordinate axes, causing the results to change if the geographical locations were measured on axes with different positions. On the other hand, *radial basis functions* are invariant to rotation of the axes. In nongeographical applications rotational invariance is not an obstacle, but for geographical data it is important to know that the answers are not affected by axis position. Hence, for this extension we use *radial basis functions* for bivariate smoothing, in particular the low-rank thin plate spline bases.

4.2.1 Model description

Algorithm 2 is relatively easy to extend to its bivariate extension. The data are of the form

$$(\boldsymbol{x}_i, y_i), \quad \boldsymbol{x}_i \in \mathbb{R}^2, \quad y_i \in \mathbb{R}. \tag{4.1}$$

In classical geostatistics applications the $\boldsymbol{x}_i$s represent geographical locations, such as latitude and longitude. However, in (4.1) the $\boldsymbol{x}_i$s could also represent pairs of non-geographical measurements. The bivariate version of (3.2) is

$$y_i \overset{\text{ind.}}{\sim} N(f(\boldsymbol{x}_i), g(\boldsymbol{x}_i)), \quad 1 \le i \le n, \tag{4.2}$$

where f and g are real-valued smooth functions on $\mathbb{R}^2$. The bivariate analogue of (3.3) is

$$\begin{aligned} f(\boldsymbol{x}) &= \beta_0 + \boldsymbol{\beta}_1^\top \boldsymbol{x} + \sum_{k=1}^{K_u} u_k \, z_k^u(\boldsymbol{x}), \qquad u_k \overset{\text{ind.}}{\sim} N(0, \sigma_u^2) \\ \text{and} \quad g(\boldsymbol{x}) &= \exp\left(\gamma_0 + \boldsymbol{\gamma}_1^\top \boldsymbol{x} + \sum_{k=1}^{K_v} v_k \, z_k^v(\boldsymbol{x})\right), \quad v_k \overset{\text{ind.}}{\sim} N(0, \sigma_v^2), \end{aligned} \tag{4.3}$$

where $\boldsymbol{\beta}_1$ and $\boldsymbol{\gamma}_1$ are each 2×1 vectors of fixed effects. The functions $\{z_k^u : 1 \le k \le K_u\}$ and $\{z_k^v : 1 \le k \le K_v\}$ now represent bivariate spline basis functions. A reasonable default for the z_k^u (Ruppert *et al.*, 2003) is the low-rank thin plate spline basis with kth element:

$$z_k^u(\boldsymbol{x}) = r(\|\boldsymbol{x} - \boldsymbol{\kappa}_k^u\|) \, [\, r(\|\boldsymbol{\kappa}_k^u - \boldsymbol{\kappa}_{k'}^u\|)]^{-1/2}_{1 \le k, k' \le K_u}, \tag{4.4}$$

where $\boldsymbol{\kappa}_1^u, \ldots, \boldsymbol{\kappa}_{K_u}^u$ is a set of bivariate knot locations that efficiently cover the space of the $\boldsymbol{x}_i$s and $r(x) \equiv x^2 \log(x)$. The default z_k^vs have an analogous definition.

Comparing this set-up to the univariate nonparametric heteroscedastic regression model, treated in Chapter 3, the only differences are the basis functions and their coefficients. Hence, Algorithm 2 can be used to fit the bivariate nonparametric heteroscedastic regression model by replacing $\boldsymbol{\nu}$, $\boldsymbol{\omega}$, $\boldsymbol{C}_\nu$ and $\boldsymbol{C}_\omega$ from Section 3.2 with

$$\boldsymbol{\nu} \equiv \begin{bmatrix} \beta_0 \\ \boldsymbol{\beta}_1 \\ \boldsymbol{u} \end{bmatrix}, \; \boldsymbol{\omega} \equiv \begin{bmatrix} \gamma_0 \\ \boldsymbol{\gamma}_1 \\ \boldsymbol{v} \end{bmatrix}, \; \boldsymbol{C}_\nu \equiv \Big[1 \;\; \boldsymbol{x}_i^T \;\; \underset{1 \le k \le K^u}{z_k^u(\boldsymbol{x}_i)}\Big]_{1 \le i \le n} \text{ and } \boldsymbol{C}_\omega \equiv \Big[1 \;\; \boldsymbol{x}_i^T \;\; \underset{1 \le k \le K^v}{z_k^v(\boldsymbol{x}_i)}\Big]_{1 \le i \le n}.$$

We have carried out extensive assessment of the bivariate version of Algorithm 2 to assess accuracy and computing time of model (4.2). Sections 4.2.2 and 4.2.3 give the details of a simulation study and real data example.

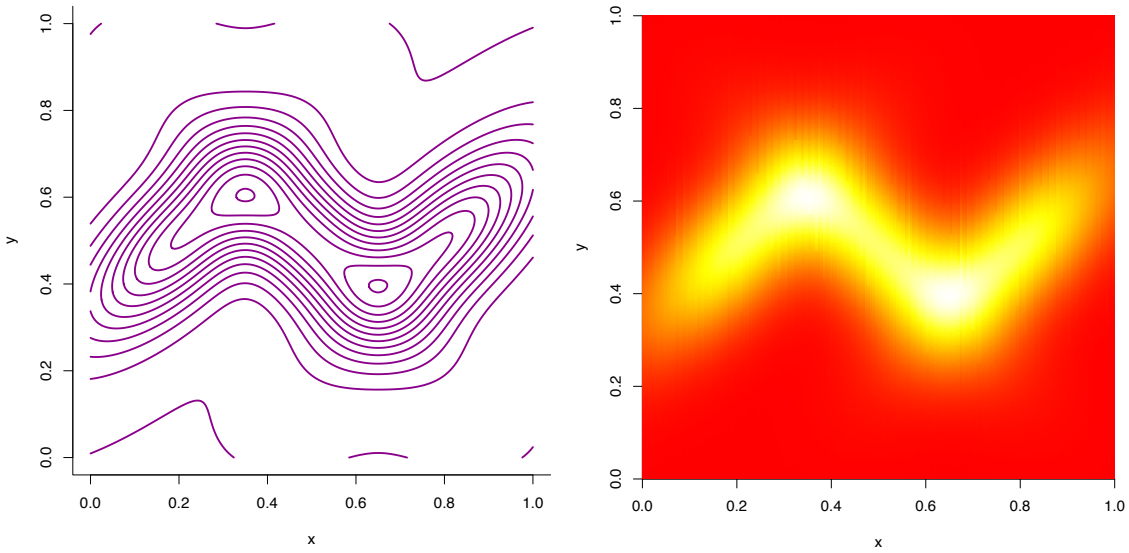


Figure 4.1: *Left panel: Contour plot illustrating the simulation setting given by (4.5). Right panel: Heat plot corresponding to (4.5).*

4.2.2 Simulation study

We consider the following true mean and log variance functions for our simulation setting. This is illustrated in Figure 4.1.

$$\begin{aligned} f(x_1,x_2) &= 0.199 + 12.924\left\{\phi\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}; \begin{bmatrix} 0.2 \\ 0.5 \end{bmatrix}, \begin{bmatrix} 0.360 & 0.252 \\ 0.252 & 0.360 \end{bmatrix}\right)\right. \\ &\quad + \phi\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}; \begin{bmatrix} 0.375 \\ 0.5 \end{bmatrix}, \begin{bmatrix} 0.360 & 0.252 \\ 0.252 & 0.360 \end{bmatrix}\right) \\ &\quad \left. + \phi\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}; \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}, \begin{bmatrix} 0.360 & -0.252 \\ -0.252 & 0.360 \end{bmatrix}\right)\right\}, \\ \log g(x_1,x_2) &= -5.470 + 0.873\cos\left(0.3\pi(x_1+x_2)\right). \end{aligned} \tag{4.5}$$

From this setting we generated 1000 datasets of size $n = 200$. Each model corresponding to a new replication was fitted using MFVB corresponding to the bivariate version of Algorithm 2 and MCMC based on a burnin of 5000, kept sample of 5000 and thinning factor of 5. The MFVB iterations were terminated when the relative change in $\log \underline{p}(\boldsymbol{y}; q, \boldsymbol{\mu}_{q(\boldsymbol{\omega})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\omega})})$ fell below 10^{-7}.

Figure 4.2 illustrates an accuracy assessment where the accuracy scores for each parameter of interest are summarized as a boxplot. The parameters on the horizontal axis of Figure 4.2 are the estimated approximate posterior density functions for f and g, evaluated at the sample quartiles of x_1 and x_2. We use Q_1, Q_2 and Q_3 to denote these sample

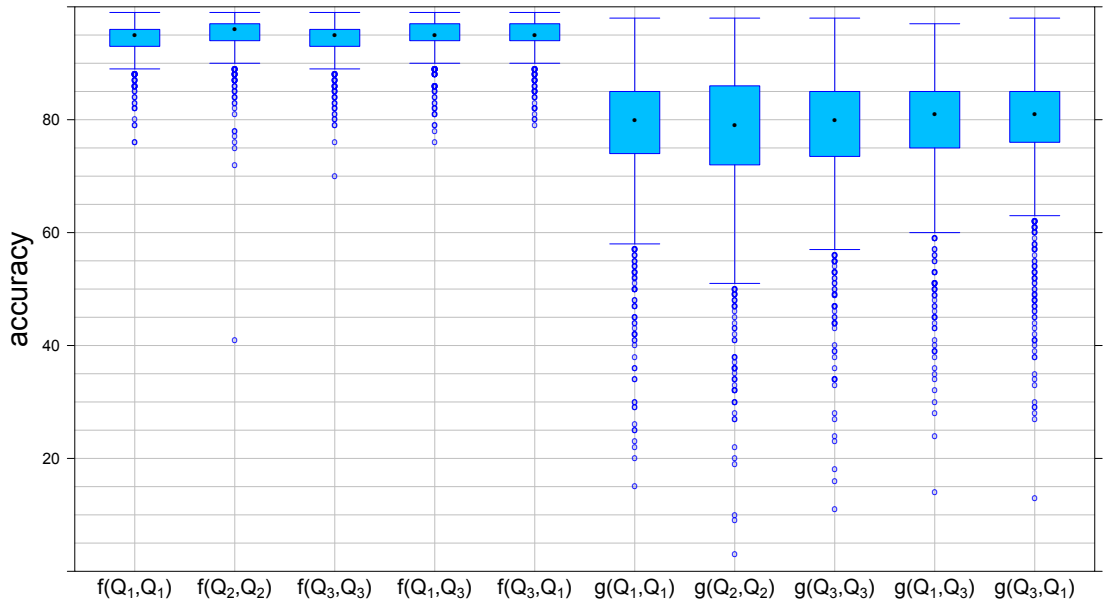


Figure 4.2: *Boxplot representation of accuracy values for MFVB against an MCMC benchmark, where the boxplots correspond to the mean function hexiles and variance function hexiles.*

quartiles. It is pleasing to see that most of the accuracy values for the bivariate mean function lie above 90% and above 70% for the bivariate variance function. Figure 4.3 shows visual assessment of the approximate posterior density functions as given by using MFVB in comparison with the MCMC benchmark for a single replication in the simulation study. MFVB accuracy is seen to be excellent for the mean function hexiles and quite good for the variance function hexiles.

The percentages of the true parameter coverage based on the approximate 95% credible sets attained from the MFVB posterior densities are given in Table 4.1. All of the mean function coverage values are showing excellent coverage of the truth, except for $f(Q_2, Q_2)$ and this may be due to the lack of data surrounding this region in our simulated datasets. Also, the variance function coverage values are doing well in terms of covering the truth and this is consistent with the accuracy scores achieved by MFVB. An assessment of the speed of the competing approaches has also been carried out. The run time of each replication was monitored for both MCMC and MFVB. These times are summarized in Table 4.2. This study was performed on a laptop computer (Intel Core i7 2.8GHz processor, 16 GBytes of random access memory). These results show that on average

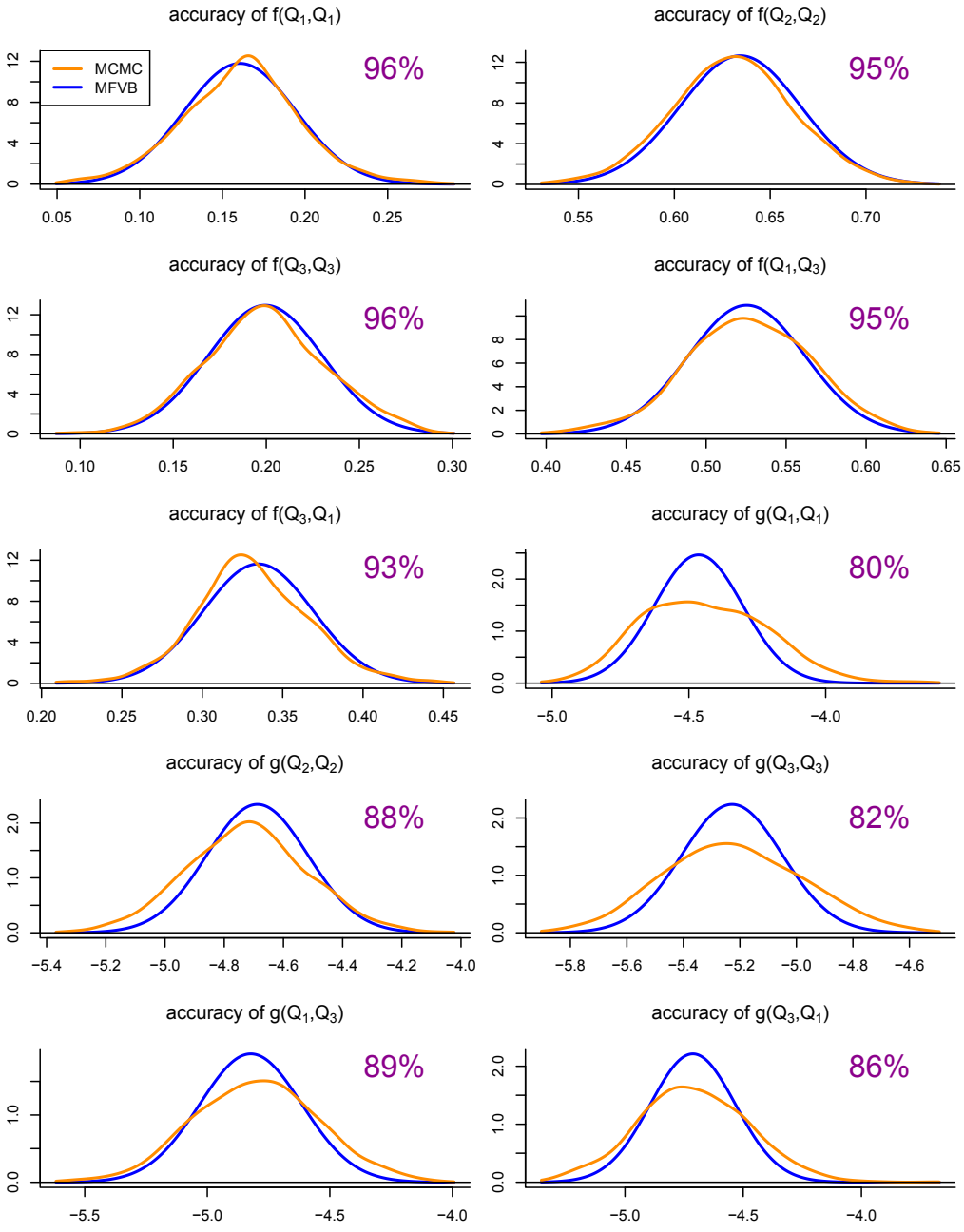


Figure 4.3: *Approximate posterior density functions obtained via MFVB and MCMC for a single replication of the simulation study. The pairs of density functions shown correspond to the mean and variance functions hexiles. The accuracy scores as defined in Section 1.9 are also displayed.*

$f(Q_1, Q_1)$	$f(Q_2, Q_2)$	$f(Q_3, Q_3)$	$f(f(Q_1, Q_3)$	$f(Q_3, Q_1)$
94	68	92	90	91
$g(Q_1, Q_1))$	$g(Q_2, Q_2)$	$g(Q_3, Q_3)$	$g(Q_1, Q_3)$	$g(Q_3, Q_1)$
88	79	86	85	86

Table 4.1: *Percentage coverage of the true parameter values by approximate* 95% *credible intervals based on variational Bayes approximate posterior density functions. The percentages are based on 1000 replications.*

MCMC	MFVB
(533.68, 534.18)	(4.05, 4.52)

Table 4.2: *99% Wilcoxon confidence intervals based on run times in seconds for MCMC and MFVB fitting.*

MFVB is at least 100 times faster than our MCMC benchmark.

4.2.3 Application

We fitted (4.3) to geo-referenced data on sea-floor sediment pollution in the North Sea (source: Pebesma & Duin, 2005). The data are stored in the `pcb` data-frame within the R package `gstat` (Pebesma, 2004). The response variable is a measurement of polychlorinated biphenyl with Ballschmiter-Zell congener number 138 (PCB-138). The motivating study is concerned with spatial and temporal variability of PCB-138. For the purposes of illustration, we ignore the temporal aspect and focus on geographical variability in the mean and variance of the response. In the notation of model (4.2)–(4.3), the variables are $\boldsymbol{x} = (x_1, x_2)$ where

x_1 = x-coordinate in the Universal Mobile Telecommunications System for Zone 31,

x_2 = y-coordinate in the Universal Mobile Telecommunications System for Zone 31, and

y = PCB-138 measured on the sediment fraction smaller than 63 parts per million, in μg/kg dry matter.

The sample size is $n = 216$ and $K_u = K_v = 50$ thin plate spline bases were used for each functional fit. The estimated mean and standard deviation functions are shown in Figure 4.4. Both functions are seen to exhibit pronounced spatial effects. Simple geostatistical

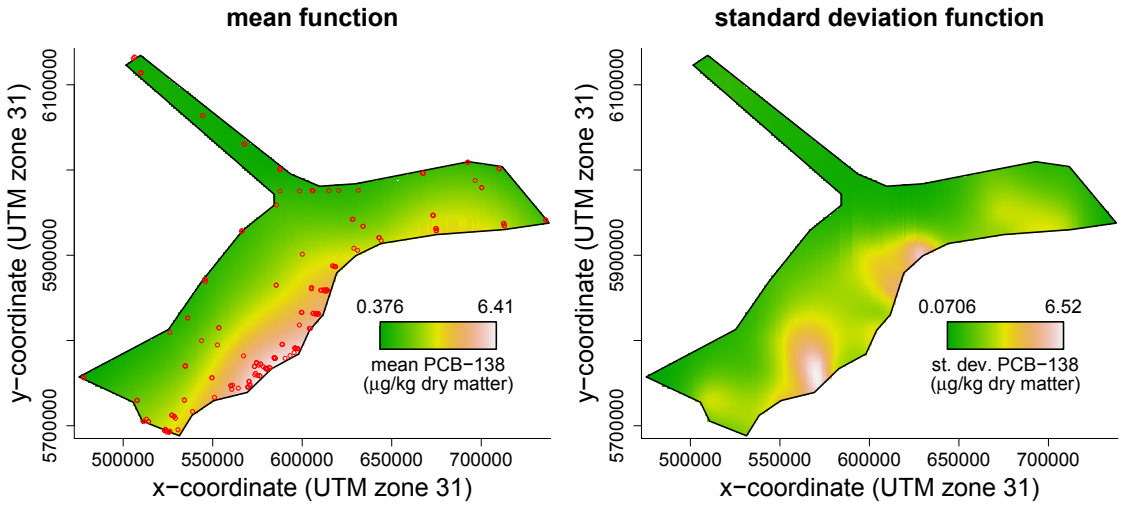


Figure 4.4: *Left panel: Fitted mean function for the polychlorinated biphenyl data described in the text, based on MFVB via Algorithm 2. Right panel: similar to top panel but for the standard deviation function.*

models that ignore the heteroscedasticity described by the right panel of Figure 4.4 would have erroneous prediction intervals.

Figure 4.5 gives a visual assessment of the convergence of the bivariate analogue of Algorithm 2 as it monitors successive values of $\log \underline{p}(\boldsymbol{y}; q, \boldsymbol{\mu}_{q(\boldsymbol{\omega})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\omega})})$. It is evident that after approximately 75 iterations the Algorithm has reached convergence, after which inference was made to achieve Figure 4.4.

Higher dimensional heteroscedastic nonparametric regression can be achieved via Algorithm 2 with little notational change from the bivariate case treated here. The only required modification involves higher-dimensional thin plate spline basis functions instead of those given by (4.4).

4.3 Extension to additive models

The previous section showed how to construct tractable regression models for a bivariate set of continuous predictors modelled as a smooth function. However in many regression settings involving several continuous predictors, incorporation of multiple smooth functions may be more suited. The assumption made here is that of additivity, hence they are referred to as *additive models* (Ezekiel, 1924; Friedman & Stuetzle, 1981). Furthermore, we are dealing with models that need to accommodate heteroscedasticity, thus we

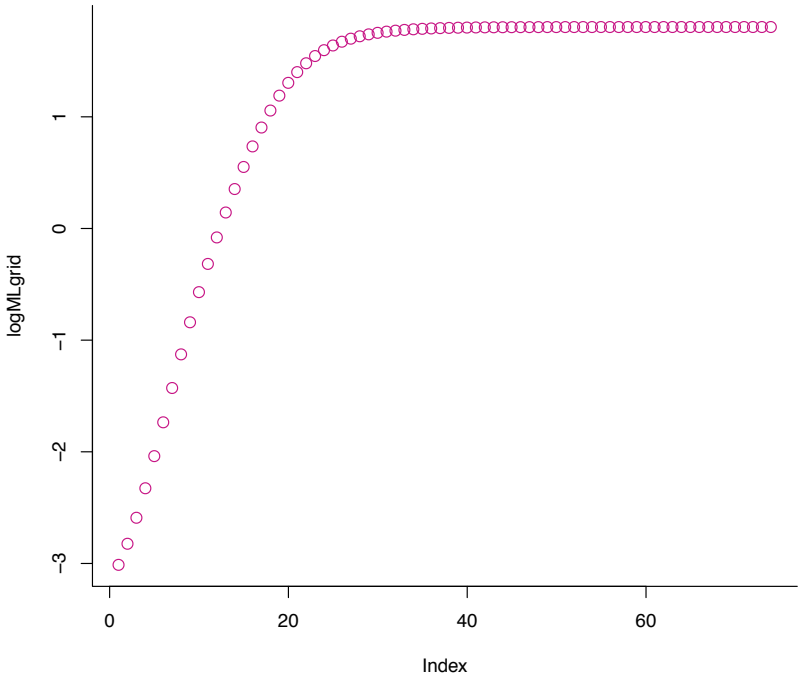


Figure 4.5: *Successive values of* $\log \underline{p}(\boldsymbol{y}; q, \boldsymbol{\mu}_{q(\boldsymbol{\omega})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\omega})})$ *for monitoring the convergence of non-conjugate MFVB given by the bivariate analogue of Algorithm 2.*

incorporate the variance as a multiplicative function. Models of this type have a small literature with Rigby & Stasinopoulos (2005) being a key reference. Their generic form is

$$y_i \sim N\Big(\beta_0 + \sum_{j=1}^{d} f_j(x_{ji}), \exp\big(\gamma_0 + \sum_{j=1}^{d} h_j(x_{ji})\big)\Big), \quad 1 \le i \le n. \tag{4.6}$$

Here f_j and h_j, $1 \le j \le d$, are smooth but otherwise arbitrary functions. Penalised splines are easily extended to handle (4.6). Here we use penalized spline models of the form

$$\begin{aligned} f_j(x) &= \beta_j\, x + \sum_{k=1}^{K_j^u} u_{jk}\, z_{jk}^u(x), \; u_{jk} \overset{\text{ind.}}{\sim} N(0, \sigma_{uj}^2) \\ \text{and} \quad h_j(x) &= \gamma_j\, x + \sum_{k=1}^{K_j^v} v_{jk}\, z_{jk}^v(x), \; v_{jk} \overset{\text{ind.}}{\sim} N(0, \sigma_{vj}^2) \end{aligned} \tag{4.7}$$

where the $\{z_{jk}^u : 1 \le k \le K_j^u\}$, $1 \le j \le d$, are spline basis functions of sizes K_j^u, analogue to those presented in Section 3.2. In addition the $\{z_{jk}^v : 1 \le k \le K_j^v\}$, $1 \le j \le d$, are similarly defined. The priors on the regression coefficients and standard deviation parameters are

$$\begin{aligned} &\beta_j \overset{\text{ind.}}{\sim} N(0, \sigma_{\boldsymbol{\beta}}^2), \quad \gamma_j \overset{\text{ind.}}{\sim} N(0, \sigma_{\boldsymbol{\gamma}}^2), \\ &\sigma_{uj} \overset{\text{ind.}}{\sim} \text{Half-Cauchy}(A_u), \quad \sigma_{vj} \overset{\text{ind.}}{\sim} \text{Half-Cauchy}(A_v), \quad 1 \le j \le d. \end{aligned} \tag{4.8}$$

The complete Bayesian hierarchical model corresponding to (4.6) is:

$$\begin{gathered}
\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{u},\boldsymbol{v} \sim \mathrm{N}\left(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}_u\boldsymbol{u},\ \mathrm{diag}\left\{\exp\left(\boldsymbol{X}\boldsymbol{\gamma}+\boldsymbol{Z}_v\boldsymbol{v}\right)\right\}\right), \\
\boldsymbol{u}|\sigma_u^2 \sim \mathrm{N}\left(\boldsymbol{0},\ \underset{1\le j\le d}{\mathrm{blockdiag}}\left(\sigma^2_{\boldsymbol{u}_j}\boldsymbol{I}_{K_{\boldsymbol{u}_j}}\right)\right), \quad \boldsymbol{v}|\sigma_v^2 \sim \mathrm{N}\left(\boldsymbol{0},\ \underset{1\le j\le d}{\mathrm{blockdiag}}\left(\sigma^2_{\boldsymbol{v}_j}\boldsymbol{I}_{K_{\boldsymbol{v}_j}}\right)\right), \\
\sigma^2_{\boldsymbol{u}_j}|a_{\boldsymbol{u}_j} \sim \text{Inverse-Gamma}(\tfrac{1}{2}, 1/a_{\boldsymbol{u}_j}), \quad a_{\boldsymbol{u}_j} \sim \text{Inverse-Gamma}(\tfrac{1}{2}, 1/A^2_{\boldsymbol{u}_j}), \\
\sigma^2_{\boldsymbol{v}_j}|a_{\boldsymbol{v}_j} \sim \text{Inverse-Gamma}(\tfrac{1}{2}, 1/a_{\boldsymbol{v}_j}), \quad a_{\boldsymbol{v}_j} \sim \text{Inverse-Gamma}(\tfrac{1}{2}, 1/A^2_{\boldsymbol{v}_j}), \\
1 \le j \le d, \quad \boldsymbol{\beta} \sim \mathrm{N}\left(0,\ \sigma^2_\beta \boldsymbol{I}_{d+1}\right), \quad \boldsymbol{\gamma} \sim \mathrm{N}\left(0,\ \sigma^2_\gamma \boldsymbol{I}_{d+1}\right).
\end{gathered} \tag{4.9}$$

The differences of this model and that of Section 3.2 are that the design matrices are now

$$\begin{aligned}
\boldsymbol{X} &\equiv \left[1\ x_{1i}\ x_{2i}\ \cdots\ x_{di}\right]_{1\le i\le n}, \\
\boldsymbol{Z}_u &\equiv \left[\begin{array}{cccc} \left[z_k^{\boldsymbol{u}_1}(x_{1i})\right]_{1\le k\le K_{\boldsymbol{u}_1}} & \left[z_k^{\boldsymbol{u}_2}(x_{2i})\right]_{1\le k\le K_{\boldsymbol{u}_2}} & \cdots & \left[z_k^{\boldsymbol{u}_d}(x_{di})\right]_{1\le k\le K_{\boldsymbol{u}_d}} \end{array}\right]_{1\le i\le n}, \\
\boldsymbol{Z}_v &\equiv \left[\begin{array}{cccc} \left[z_k^{\boldsymbol{v}_1}(x_{1i})\right]_{1\le k\le K_{\boldsymbol{v}_1}} & \left[z_k^{\boldsymbol{v}_2}(x_{2i})\right]_{1\le k\le K_{\boldsymbol{v}_2}} & \cdots & \left[z_k^{\boldsymbol{v}_d}(x_{di})\right]_{1\le k\le K_{\boldsymbol{v}_d}} \end{array}\right]_{1\le i\le n}
\end{aligned} \tag{4.10}$$

The coefficient vectors are

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{bmatrix}, \quad \boldsymbol{\gamma} = \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \vdots \\ \gamma_d \end{bmatrix}, \quad \boldsymbol{u} = \begin{bmatrix} \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_d \end{bmatrix}, \quad \text{and} \quad \boldsymbol{v} = \begin{bmatrix} \boldsymbol{v}_1 \\ \vdots \\ \boldsymbol{v}_d \end{bmatrix}, \tag{4.11}$$

where $\boldsymbol{u}_j$ is the $K_{\boldsymbol{u}_j} \times 1$ vector containing the u_{jk}. It is again convenient to combine the mean function coefficients and variance function coefficients into single vectors as shown in (3.6). The Bayesian model given by (4.9) admits a closed form non-conjugate MFVB algorithm, with the regression coefficients for the full mean and variance functions each being Multivariate Normal. Algorithm 3 gives the details of these closed form updates.

Convergence of Algorithm 3 is assessed using the lower-bound of the marginal log-likelihood:

$$\begin{aligned}
\log \underline{p}(\boldsymbol{y}; q, \boldsymbol{\mu}_{q(\boldsymbol{\omega})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\omega})}) \quad = \quad & \tfrac{1}{2}\left\{\textstyle\sum_{\ell=1}^d (K_{u_\ell} + K_{v_\ell}) + 4\right\} - \tfrac{n}{2}\log(2\pi) + \log\Gamma\left(\tfrac{1}{2}(K_u+1)\right) \\
& -2\log(\pi) + \log\Gamma\left(\tfrac{1}{2}\left(\textstyle\sum_{\ell=1}^d K_{v_\ell} + d - 1\right)\right) - \textstyle\sum_{\ell=1}^d \log(A_{u_\ell}) \\
& - \textstyle\sum_{\ell=1}^d \log(A_{v_\ell}) - \tfrac{1}{2}\mathbf{1}^T(\boldsymbol{C}_\omega \boldsymbol{\mu}_{q(\boldsymbol{\omega})}) - \tfrac{1}{2}\mathbf{1}^\top\left\{\boldsymbol{\mu}_{q(r^2_{\boldsymbol{\nu}})} \odot \exp\left(\boldsymbol{\psi}_{q(\boldsymbol{\omega})}\right)\right\} \\
& -\tfrac{1}{2}(d+1)\log(\sigma^2_\beta) - \tfrac{1}{2}(d+1)\log(\sigma^2_\gamma) + \tfrac{1}{2}\log|\Sigma_{q(\boldsymbol{\nu})}| \\
& +\tfrac{1}{2}\log|\Sigma_{q(\boldsymbol{\omega})}| - \tfrac{1}{2\sigma^2_\beta}\left(\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\right) \\
& -\tfrac{1}{2\sigma^2_\gamma}\left\{\|\boldsymbol{\mu}_{q(\boldsymbol{\gamma})}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\gamma})})\right\} - \tfrac{1}{2}(\textstyle\sum_{\ell=1}^d K_{u_\ell} + 1)\log(B_{q(\sigma^2_{u_\ell})}) \\
& -\tfrac{1}{2}(\textstyle\sum_{\ell=1}^d K_{v_\ell} + 1)\log(B_{q(\sigma^2_{v_\ell})}) - \log\left(\mu_{q(1/\sigma^2_{u_\ell})} + A^{-2}_{u_\ell}\right) \\
& - \log\left(\mu_{q(1/\sigma^2_{v_\ell})} + A^{-2}_{v_\ell}\right) + \mu_{q(1/\sigma^2_{u_\ell})}\mu_{q(1/a_{u_\ell})} + \mu_{q(1/\sigma^2_{v_\ell})}\mu_{q(1/a_{v_\ell})}.
\end{aligned}$$

Set up initial values:

$\boldsymbol{\mu}_{q(\boldsymbol{\omega})}$ a $\left(\sum_j^d K_{v_j} + d + 1\right)\times 1$ vector, $\boldsymbol{\Sigma}_{q(\boldsymbol{\omega})}$ a $\left(\sum_j^d K_{v_j} + d + 1\right)\times\left(\sum_j^d K_{v_j} + d + 1\right)$ positive definite matrix, $\mu_{q\left(1/\sigma_{u_j}^2\right)}, \mu_{q\left(1/\sigma_{v_j}^2\right)} > 0$, and $\boldsymbol{\mu}_{q(\boldsymbol{r}_{\boldsymbol{\nu}}^2)}$ an $n \times 1$ vector.

Cycle through:

$$\boldsymbol{\psi}_{q(\boldsymbol{\omega})} \leftarrow \exp\left\{-\boldsymbol{C}_\omega \boldsymbol{\mu}_{q(\boldsymbol{\omega})} + \tfrac{1}{2}\text{diagonal}\left(\boldsymbol{C}_\omega \boldsymbol{\Sigma}_{q(\boldsymbol{\omega})} \boldsymbol{C}_\omega^\top\right)\right\}$$

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\nu})} \leftarrow \left(\boldsymbol{C}_\nu^\top \text{diag}\left\{\boldsymbol{\psi}_{q(\boldsymbol{\omega})}\right\} \boldsymbol{C}_\nu + \begin{bmatrix} \sigma_{\boldsymbol{\beta}}^{-2}\boldsymbol{I}_3 & \boldsymbol{0} \\ \boldsymbol{0} & \underset{1\le j\le d}{\text{blockdiag}}\left(\mu_{q\left(1/\sigma_{u_j}^2\right)} I_{K_j}^u\right) \end{bmatrix}\right)^{-1}$$

$$\boldsymbol{\mu}_{q(\boldsymbol{\nu})} \leftarrow \boldsymbol{\Sigma}_{q(\boldsymbol{\nu})} \boldsymbol{C}_\nu^\top \text{diag}\left\{\boldsymbol{\psi}_{q(\boldsymbol{\omega})}\right\} \boldsymbol{y}$$

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\omega})} \leftarrow \left(\boldsymbol{C}_\omega^\top \text{diag}\left\{\boldsymbol{\mu}_{q(\boldsymbol{r}_{\boldsymbol{\nu}}^2)} \odot \boldsymbol{\psi}_{q(\boldsymbol{\omega})}\right\} \boldsymbol{C}_\omega + \begin{bmatrix} \sigma_{\boldsymbol{\gamma}}^{-2}\boldsymbol{I}_3 & \boldsymbol{0} \\ \boldsymbol{0} & \underset{1\le j\le d}{\text{blockdiag}}\left(\mu_{q\left(1/\sigma_{v_j}^2\right)} I_{K_j}^v\right) \end{bmatrix}\right)^{-1}$$

$$\boldsymbol{\mu}_{q(\boldsymbol{\omega})} \leftarrow \boldsymbol{\mu}_{q(\boldsymbol{\omega})} + \boldsymbol{\Sigma}_{q(\boldsymbol{\omega})}\left\{\boldsymbol{C}_\omega^\top\left(\boldsymbol{\mu}_{q(\boldsymbol{r}_{\boldsymbol{\nu}}^2)} \odot \boldsymbol{\psi}_{q(\boldsymbol{\omega})} - \boldsymbol{1}\right) - \begin{bmatrix} \sigma_{\boldsymbol{\gamma}}^{-2}\boldsymbol{I}_3 & \boldsymbol{0} \\ \boldsymbol{0} & \underset{1\le j\le d}{\text{blockdiag}}\left(\mu_{q\left(1/\sigma_{v_j}^2\right)} I_{K_j}^v\right) \end{bmatrix}\boldsymbol{\mu}_{q(\boldsymbol{\omega})}\right\}$$

For $j = 1, \ldots, d$:

$$\mu_{q\left(1/a_{\boldsymbol{u}_j}\right)} \leftarrow 1/\left(\mu_{q\left(1/\sigma_{u_j}^2\right)} + A_{u_j}^2\right) \quad ; \quad \mu_{q\left(1/a_{v_j}\right)} \leftarrow 1/\left(\mu_{q\left(1/\sigma_{v_j}^2\right)} + A_{v_j}^2\right)$$

$$\mu_{q\left(1/\sigma_{u_j}^2\right)} \leftarrow \frac{K_{u_j}+1}{2\mu_{q\left(1/a_{u_j}\right)} + \|\boldsymbol{\mu}_{q\left(u_j\right)}\|^2 + \text{tr}\left(\boldsymbol{\Sigma}_{q\left(u_j\right)}\right)}$$

$$\mu_{q\left(1/\sigma_{v_j}^2\right)} \leftarrow \frac{K_{v_j}+1}{2\mu_{q\left(1/a_{v_j}\right)} + \|\boldsymbol{\mu}_{q\left(v_j\right)}\|^2 + \text{tr}\left(\boldsymbol{\Sigma}_{q\left(v_j\right)}\right)}$$

until the absolute relative change in $\log \underline{p}(\boldsymbol{y};\, q,\, \boldsymbol{\mu}_{q(\boldsymbol{\omega})},\, \boldsymbol{\Sigma}_{q(\boldsymbol{\omega})})$ is negligible.

Algorithm 3: *MFVB algorithm for the determination of the optimal parameters in* $q^*(\boldsymbol{\omega})$, $q^*(\boldsymbol{\nu})$, $q^*(\sigma_{uj}^2)$, $q^*(\sigma_{vj}^2)$, $q^*(a_{uj})$ *and* $q^*(a_{vj})$, $1 \le j \le d$.

The derivations of the optimal parameters in Algorithm 3 and the calculations corresponding to the lower-bound of the marginal log-likelihood are similar to that described in the Appendix 3.A and 3.B.

4.3.1 Application

Algorithm 3 has been applied to a data set from the Californian air pollution study described in Breiman & Friedman (1985). The data set consists of 345 observations on the four variables $x_{i1}, x_{i2}, x_{i3}, y_i$, $1 \leq i \leq 345$. The predictors are defined as

x_1 = pressure gradient (mm Hg) from Los Angeles International Airport to Daggett, California,

x_2 = inversion base height (feet)

x_3 = inversion base temperature (degrees Fahrenheit),

and the response variable is

y = ozone concentration (ppm) at Sandburg Air Force Base.

Our model is

$$y_i \overset{\text{ind.}}{\sim} \mathrm{N}\left(\beta_0 + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}), \right. \\ \left. \exp\left\{\gamma_0 + h_1(x_{1i}) + h_2(x_{2i}) + h_3(x_{3i})\right\}\right). \tag{4.12}$$

The data was standardised and the hyperparameters were set to have values $\sigma_{\boldsymbol{\beta}} = \sigma_{\boldsymbol{\gamma}} = A_u = A_v = 10^5$, to achieve non-informativity. After fitting, the resulting values were transformed to their original units. In addition the spline basis function sizes for modelling f_1, f_2 and f_3 where 18 each. The same values were used for modelling h_1, h_2 and h_3. The approximation of the mean function and standard deviation function is shown in Figure 4.6. Inference based on MCMC was carried out using the same convergence criteria as explained in Section 4.2.2.

The estimated mean function f_1 has been vertically aligned to correspond to the response data by plotting, for $\boldsymbol{x}_1$, the response with each of $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$ set at their average. In other words, we have evaluated the estimate of f_2 at $\bar{x}_2$ and estimate of f_3 at $\bar{x}_3$ in order to vertically align the estimated f_1 curve. Similar alignments have been used for f_2, f_3, $\exp(h_1/2)$, $\exp(h_2/2)$, and $\exp(h_3/2)$.

As is seen from Figure 4.6 MFVB and MCMC are showing excellent agreement in

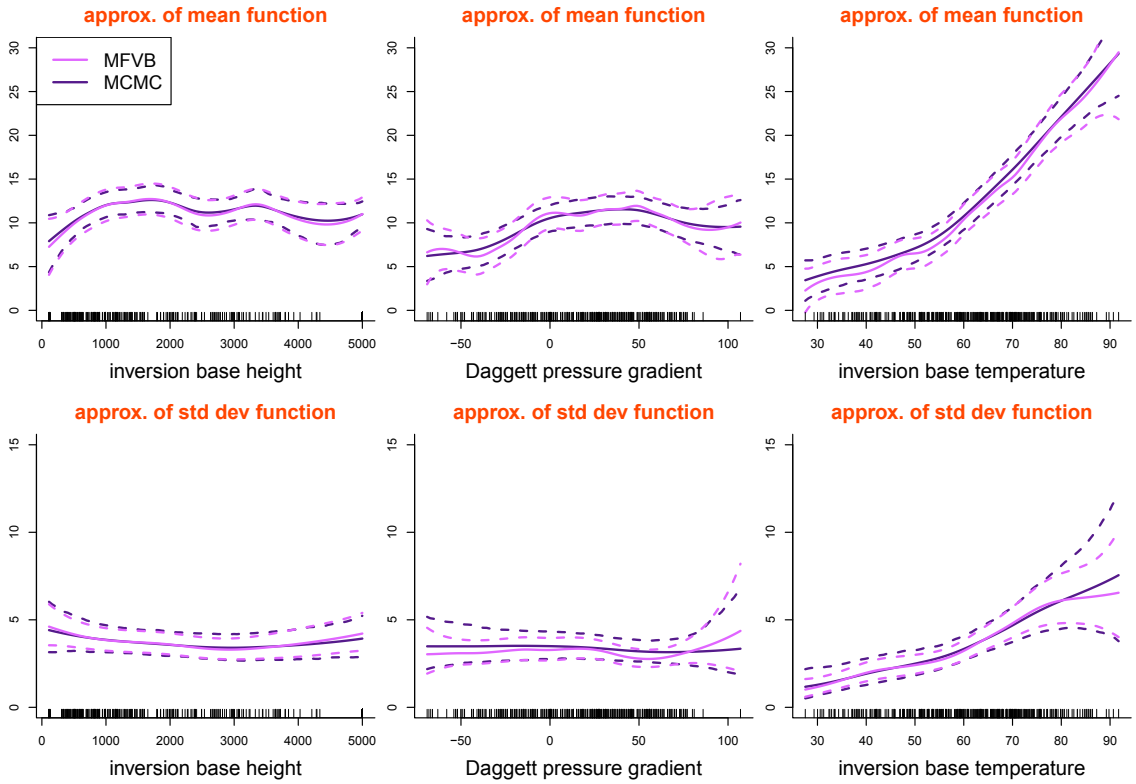


Figure 4.6: *Top panels: approximation to the posterior mean functions* f_1, f_2, *and* f_3 *and their* 95% *credible sets. Bottom panels: approximation to the posterior standard deviation functions* $exp(h_1/2)$, $exp(h_2/2)$, *and* $exp(h_3/2)$ *and their* 95% *credible sets.*

their fits. The MFVB approach was markedly faster for this example. Even though the existence of heteroscedasticity in *inversion base height* and *Dagget pressure gradient* is somewhat low, there is noticeable heteroscedasticity in *inversion base temperature* which has been captured by model (4.12).

4.4 Real-time heteroscedastic nonparametric regression

The constant expansion of technological advancements in the 2010s so far has generated larger volumes of data than ever before. Not only is this information growing bigger but also increasing with higher speed.

Almost all nonparametric regression methodology presented to date make the assumption that the data are processed in batch, that is, at the same time. However, some disadvantages of batch processing include the requirement that analysis must wait until the entire data set has been collected. In the real-time case, the analysis updates once each new data point is collected. This is beneficial, and sometimes essential, for both high volume and/or velocity data. Many procedures used in single predictor nonparametric regression analysis use fully automated batch procedures however do not yet have an online/real-time modification. Recently, Luts *et al.* (2013) developed MFVB-based algorithms for performing semiparametric regression analysis in real time.

Algorithms that deal with the arrival of a stream of data points or just one data point per iteration have recently been developed for variational Bayesian inference by the machine learning community (e.g. Hoffman *et al.*, 2010; Wang *et al.*, 2011). This methodology has become known as *online mean field variational Bayes* or *online variational Bayes* for short. While most procedures to date involve storing a small amount of data in memory, they still need knowledge of the amount of data points from the beginning of the algorithm. In contrast, Luts *et al.* (2013) develop an online algorithm that uses past data only in the form of sufficient statistics and is relevant to fitting mixed models. We extend their approach by developing a variation of Algorithm 2 that allows for real-time heteroscedastic nonparametric regression. Unlike Luts *et al.* (2013) however, our real time MFVB algorithm is hindered by the fact that it requires storage of the entire data set. An extension to our algorithm, that only uses storage of the sufficient statistics would be worthwhile for future investigation.

The dependence on the data in Algorithm 2 is evident only in the response vector $\boldsymbol{y}$

and the design matrices $\boldsymbol{C}_\nu$ and $\boldsymbol{C}_\omega$. These quantities are simple to update when a new observation y_{new} and its corresponding $\{(2+K_u)\times 1\}$ and $\{(2+K_v)\times 1\}$ vectors of design components $\boldsymbol{c}_{\boldsymbol{\nu},\text{new}}$ and $\boldsymbol{c}_{\boldsymbol{\omega},\text{new}}$ arrives. For instance, the new $\boldsymbol{C}_\nu$ matrix is

$$\boldsymbol{C}_\nu \leftarrow \left[\boldsymbol{C}_\nu^\top \;\; \boldsymbol{c}_{\boldsymbol{\nu},\text{new}}\right]^\top.$$

The corresponding updates for y_{new} and $\boldsymbol{C}_\omega$ are given in Algorithm 4. The starting values

1. Use Algorithm 2 to perform batch-based tuning runs, analogous to those described in Algorithm 2' of Luts *et al.* (2014), and determine a warm-up sample size n_{warm} for which convergence is validated.

2. Set $\boldsymbol{\mu}_{q(\boldsymbol{\nu})}$, $\boldsymbol{\Sigma}_{q(\boldsymbol{\nu})}$, $\boldsymbol{\mu}_{q(\boldsymbol{\omega})}$, $\boldsymbol{\Sigma}_{q(\boldsymbol{\omega})}$, $\boldsymbol{\mu}_{q(1/\sigma_u^2)}$, and $\boldsymbol{\mu}_{q(1/\sigma_v^2)}$ to their values obtained in the warm up batch-based tuning run with sample size n_{warm}. Next set y_{warm} to be the response vector on the first n_{warm} observations. Also set $\boldsymbol{C}_{\boldsymbol{\nu},\text{warm}}$ and $\boldsymbol{C}_{\boldsymbol{\omega},\text{warm}}$ to be the design matrices based on the first n_{warm} observations. Lastly assign $n \leftarrow n_{\text{warm}}$.

3. Cycle:

Read in y_{new} (1×1), $\boldsymbol{c}_{\boldsymbol{\nu},\text{new}}$ $\{(2+K_u)\times 1\}$ and $\boldsymbol{c}_{\boldsymbol{\omega},\text{new}}$ $\{(2+K_v)\times 1\}$; $n \leftarrow n+1$

$$\boldsymbol{C}_\nu \leftarrow \left[\boldsymbol{C}_\nu^\top \;\; \boldsymbol{c}_{\boldsymbol{\nu},\text{new}}\right]^\top \quad ; \quad \boldsymbol{C}_\omega \leftarrow \left[\boldsymbol{C}_\omega^\top \;\; \boldsymbol{c}_{\boldsymbol{\omega},\text{new}}\right]^\top \quad ; \quad \boldsymbol{y} \leftarrow [\boldsymbol{y}^\top \;\; y_{\text{new}}]^\top$$

$$\boldsymbol{\mu}_{q(r_{\boldsymbol{\nu}}^2)} \leftarrow \text{diagonal}\left\{\left(\boldsymbol{y}-\boldsymbol{C}_\nu\boldsymbol{\mu}_{q(\boldsymbol{\nu})}\right)\left(\boldsymbol{y}-\boldsymbol{C}_\nu\boldsymbol{\mu}_{q(\boldsymbol{\nu})}\right)^\top + \boldsymbol{C}_\nu\boldsymbol{\Sigma}_{q(\boldsymbol{\nu})}\boldsymbol{C}_\nu^\top\right\}$$

$$\boldsymbol{\psi}_{q(\boldsymbol{\omega})} \leftarrow \exp\left\{-\boldsymbol{C}_\omega\boldsymbol{\mu}_{q(\boldsymbol{\omega})} + \tfrac{1}{2}\text{diagonal}\left(\boldsymbol{C}_\omega\boldsymbol{\Sigma}_{q(\boldsymbol{\omega})}\boldsymbol{C}_\omega^\top\right)\right\}$$

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\nu})} \leftarrow \left(\boldsymbol{C}_\nu^\top\text{diag}\left(\boldsymbol{\psi}_{q(\boldsymbol{\omega})}\right)\boldsymbol{C}_\nu + \begin{bmatrix}\sigma_\beta^{-2}\boldsymbol{I}_2 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{\mu}_{q(1/\sigma_u^2)}\boldsymbol{I}_{K_u}\end{bmatrix}\right)^{-1}$$

$$\boldsymbol{\mu}_{q(\boldsymbol{\nu})} \leftarrow \boldsymbol{\Sigma}_{q(\boldsymbol{\nu})}\boldsymbol{C}_\nu^\top\text{diag}\left(\boldsymbol{\psi}_{q(\boldsymbol{\omega})}\right)\boldsymbol{y}$$

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\omega})} \leftarrow \left(\boldsymbol{C}_\omega^\top\text{diag}\left(\boldsymbol{\mu}_{q(\boldsymbol{r}_{\boldsymbol{\nu}}^2)}\odot\boldsymbol{\psi}_{q(\boldsymbol{\omega})}\right)\boldsymbol{C}_\omega + \begin{bmatrix}\sigma_\gamma^{-2}\boldsymbol{I}_2 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{\mu}_{q(1/\sigma_v^2)}\boldsymbol{I}_{K_v}\end{bmatrix}\right)^{-1}$$

$$\boldsymbol{\mu}_{q(\boldsymbol{\omega})} \leftarrow \boldsymbol{\mu}_{q(\boldsymbol{\omega})} + \boldsymbol{\Sigma}_{q(\boldsymbol{\omega})}\left\{\boldsymbol{C}_\omega^\top\left(\boldsymbol{\mu}_{q(\boldsymbol{r}_{\boldsymbol{\nu}}^2)}\odot\boldsymbol{\psi}_{q(\boldsymbol{\omega})}-\boldsymbol{1}\right)-\begin{bmatrix}\sigma_\gamma^{-2}\boldsymbol{I}_2 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{\mu}_{q(1/\sigma_v^2)}\boldsymbol{I}_{K_v}\end{bmatrix}\boldsymbol{\mu}_{q(\boldsymbol{\omega})}\right\}$$

$$\mu_{q(1/a_{\boldsymbol{u}})} \leftarrow 1/\left(\mu_{q(1/\sigma_{\boldsymbol{u}}^2)} + A_{\boldsymbol{u}}^{-2}\right) \quad ; \quad \mu_{q(1/a_{\boldsymbol{v}})} \leftarrow 1/\left(\mu_{q(1/\sigma_{\boldsymbol{v}}^2)} + A_{\boldsymbol{v}}^{-2}\right)$$

$$\mu_{q(1/\sigma_{\boldsymbol{u}}^2)} \leftarrow (K_u+1)/\{2\mu_{q(1/a_{\boldsymbol{u}})} + \parallel\boldsymbol{\mu}_{q(\boldsymbol{u})}\parallel^2 + \text{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{u})}\right)\}$$

$$\mu_{q(1/\sigma_{\boldsymbol{v}}^2)} \leftarrow (K_v+1)/\{2\mu_{q(1/a_{\boldsymbol{v}})} + \parallel\boldsymbol{\mu}_{q(\boldsymbol{v})}\parallel^2 + \text{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{v})}\right)\}$$

until the analysis is complete or data are no longer available.

Algorithm 4: *MFVB algorithm for real-time determination of the optimal parameters in* $q^*(\boldsymbol{\omega})$, $q^*(\boldsymbol{\nu})$, $q^*(\sigma_u^2)$, $q^*(\sigma_v^2)$, $q^*(a_u)$ *and* $q^*(a_v)$.

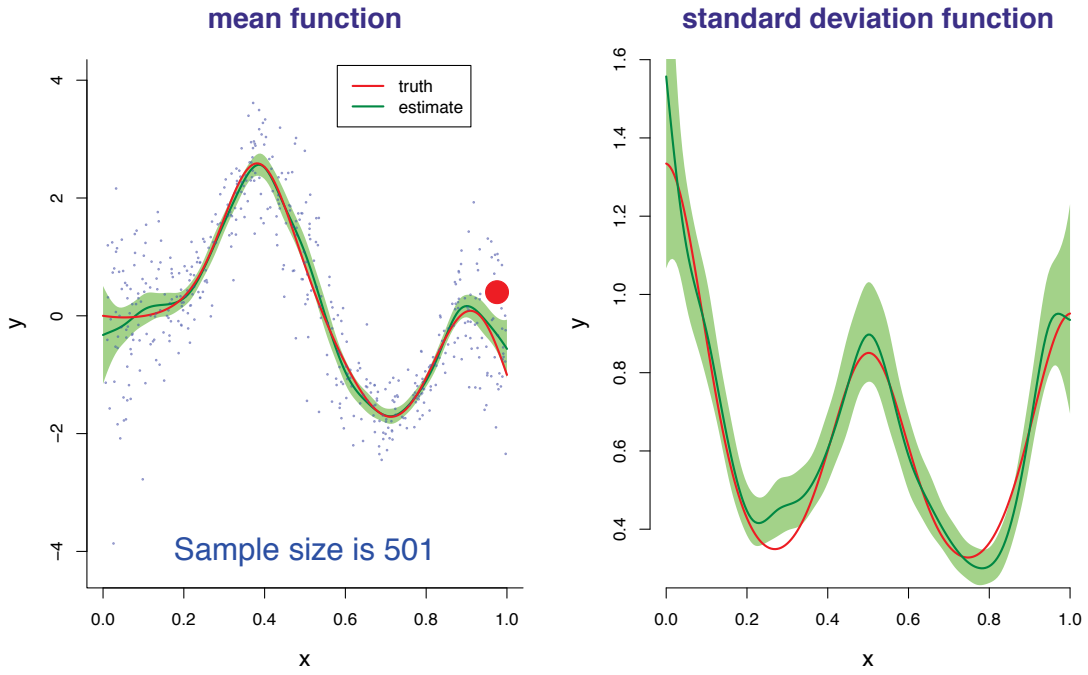
for the real-time procedure are determined by performing a sufficiently large batch fit.

When implementing Algorithm 4, the approximate posterior density functions of the model parameters are updated continually as each new data point arrives.

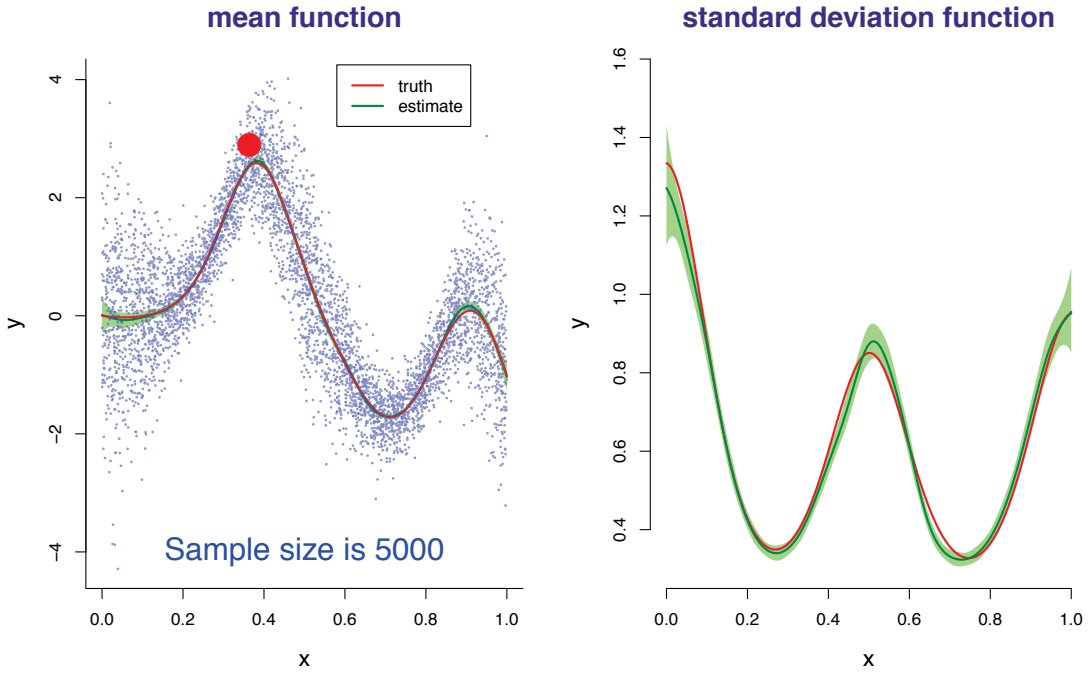
The web-site `realtime-semiparametric-regression.net` features a movie that illustrates Algorithm 4 for data simulated according to setting D from Table 3.1. The warm-up sample size is $n_{\text{warm}} = 500$. The link for the movie is titled `Heteroscedastic nonparametric regression` and portrays the effectiveness of real time processing for mean and standard deviation function fitting. Figure 4.7 illustrates the mean and variance function fits for the first online data point at $n = 501$ and again at the last online data point where $n = 5000$. As each new data point is processed, the mean and variance function fits improve as seen from this comparison.

4.5 Discussion

This chapter extends the work in Chapter 3 as we consider a wider class of models for fitting and inference using non-conjugate MFVB. We have developed closed form algorithms for both fast batch and real-time fitting and inference for a range of semiparametric regression models that have a heteroscedastic component. The methodology developed here applies to increasingly complex models as a result of the locality property of MFVB inference methods in general. In both the simulated and actual data settings, the new methodology is seen to perform very well.



(a) *Resulting mean and standard deviation function fits after first new data point arrives.*



(b) *Resulting mean and standard deviation function fits once all 5000 data points have arrived.*

Figure 4.7: *Real-time approximate non-conjugate MFVB regression fits for the simulated data according to setting D from Table 3.1.*

Chapter 5

Mean field variational Bayes for group-specific curve models

5.1 Introduction

In this chapter we develop MFVB algorithms catered to fitting large data sets that exhibit multilevel and longitudinal structures. In addition to developing the general MFVB algorithms that we have so far seen in the previous chapters of this thesis, here we also look into streamlining these algorithms in terms of storage and number of operations needed. Streamlining of such algorithms is essential when using MFVB for fitting and inference in multilevel and longitudinal models. This is often the case since a naïve implementation would result in cubic dependence on the number of groups within a level of the model being considered.

Lee & Wand (2015) develop streamlined algorithms for treatment of two-level models with Gaussian and binary responses and show that the number of operations required are linear in the number of groups in each level. This allows for much faster and highly accurate Bayesian inference for large longitudinal and multilevel models. We adopt the methodology used in Lee & Wand (2015) to streamline the two-level MFVB algorithm used in Section 5.2.3 and also extend their methodology to cater to the three-level model considered in Section 5.3.2. These algorithms have some inferential inaccuracy, however, as represented by the diagnostic illustrations in this chapter, these inaccuracies are relatively insignificant when compared to the computational time and storage savings afforded by the streamlined MFVB algorithm. The software package `Infer.NET` allows treatment of

such models, however, its non-streamlined implementation makes MFVB very slow when the number of groups is large.

Multilevel and longitudinal models, e.g. Diggle *et al.* (2002), Fitzmaurice *et al.* (2012), Gelman & Hill (2006) and Goldstein (2011), is a prominent area in Statistics, however the union of these areas and that of variational methods has not yet appeared to be eminent. Recent contributions are found in Ormerod & Wand (2010) and Luts *et al.* (2014), which use MFVB for longitudinal/multilevel data but invert the effects covariance matrix parameter updates in a naïve fashion. The main idea of this chapter is to produce MFVB algorithms for multilevel/longitudinal data based on the naïve inversion of the effects covariance matrix and compare this to the algorithms using the streamlined inversion of the effects covariance matrix. We restrict attention to the two-level and three-level Gaussian response models.

The motivation for using a streamlined MFVB algorithm in this chapter comes from a three-level longitudinal data set that represents frequency dependent backscatter coefficients for induced tumors in rodents (Simpson, 2013). The structure of this data led to the consideration of a variant of the general MFVB approach, referred to as streamlined MFVB, which we explore in further detail in this chapter.

Section 5.2 gives description of the two-level Gaussian response model in terms of its construction and corresponding MFVB methodology and Section 5.3 provides similar descriptions for the three-level Gaussian response model. Comprehensive simulation studies and real data examples are considered for both the two-level and three-level Gaussian response models. Appendix 5.A and 5.B provide the derivations for the naïve two-level model Algorithm and its corresponding lower bound on the marginal log-likelihood.

5.2 Two-level Gaussian response model

We consider the development of a model for fitting longitudinal and multilevel data such as that illustrated in Figure 5.1. There are a number of things to consider when developing a model that best represents data of this type. For instance, the repeated measures or longitudinal aspect is crucial since one would expect, upon inspection of Figure 5.1, that utilising *within-subject* information would be helpful. So, at the very least we would want to include a variance component that controls for the deviation within each subject, i.e., a random intercept. In addition, one might also consider a random slope which would

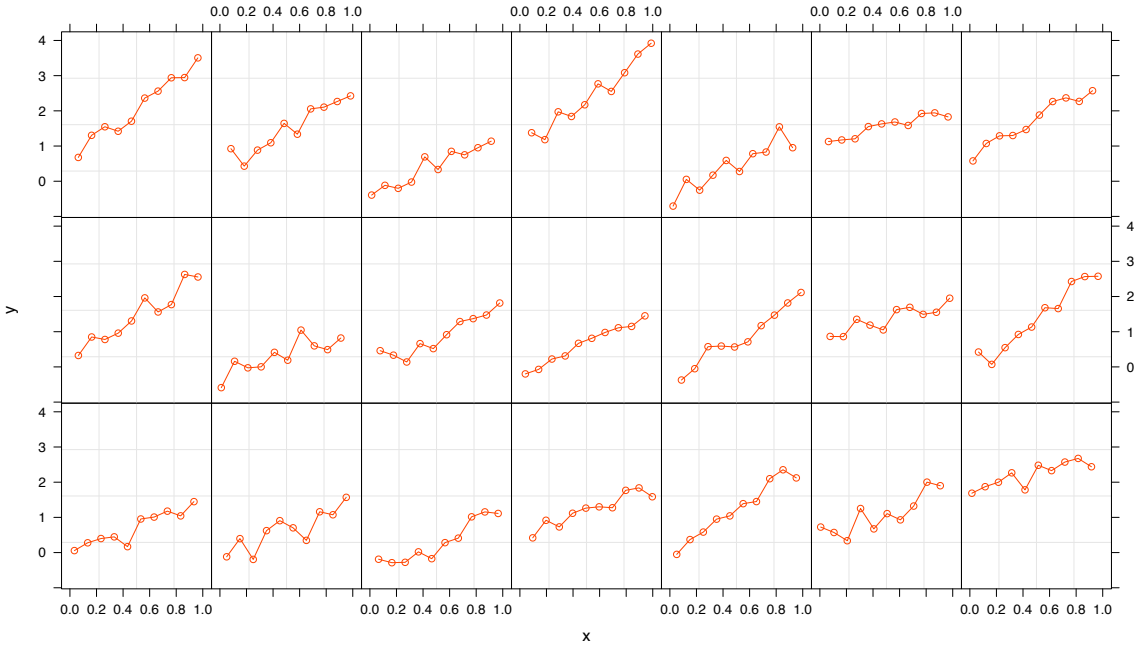


Figure 5.1: *Example of a longitudinal two-level model data-set.*

account for possible variability in the slopes of the growth curves.

The nonlinearity of such data suggests that incorporation of a nonparametric function f would be beneficial. However, f would represent a global function which does not depend on each subject. In order to cater for each individual's nonlinear component, we might consider a model of the following form:

$$y_{ij} = f(x_{ij}) + g_i(x_{ij}) + \varepsilon_{ij}, \qquad 1 \le i \le m, \quad 1 \le j \le n_i, \tag{5.1}$$

where f is a smooth function and the g_i, $1 \le i \le m$, are random functions with mean zero. Both f and g can be modeled as regression splines and then placed into the mixed model framework.

The regression splines associated with f have a fixed component that caters for the linear part of the model, and a random component that allows for departures from linearity. It is important to note, on the other hand, since the g_i are random with zero mean, the linear part of the regression spline is random, rather than fixed. A linear penalised spline

model for (5.1) is

$$\begin{aligned} f(x_{ij}) &= \beta_0 + \beta_1 x_{ij} + \sum_{\ell=1}^{L^{\text{gbl}}} u_\ell^{\text{gbl}} z_\ell^{\text{gbl}}(x_{ij}) \\ \text{and} \quad g_i(x_{ij}) &= \delta_{0i} + \delta_{1i}\, x_{ij} + \sum_{\ell=1}^{L^{\text{grp}}} u_{il}^{\text{grp}} z_\ell^{\text{grp}}(x_{ij}), \quad 1 \le i \le m, \quad 1 \le j \le n_i, \end{aligned} \tag{5.2}$$

where the u_ℓ^{gbl} and $z_\ell^{\text{gbl}}(x_{ij})$ are the spline coefficient and spline basis function for the ℓth knot in the global part of the model, that is, corresponding to the highest level of the two-level model structure. The u_{il}^{grp} and $z_\ell^{\text{grp}}(x_{ij})$ are the spline coefficient and spline basis function for the ℓth knot of the ith subject, that corresponding to the lower level of the two-level model structure. A popular choice for the z_ℓ^{gbl} and z_ℓ^{grp} are suitably transformed cubic O'Sullivan splines, as described in Section 1.10. Equations (5.1) and (5.2) can be represented in the mixed model representation as follows.

5.2.1 Mixed model representation

We define the following linear and non-linear predictors:

$$\boldsymbol{x}_{ij} = [1 \;\; x_{ij}], \qquad \boldsymbol{z}_{ij}^{\text{gbl}} = \left[z_1^{\text{gbl}}(x_{ij}) \;\ldots\; z_{\text{L}^{\text{gbl}}}^{\text{gbl}}(x_{ij})\right] \quad \text{and} \quad \boldsymbol{z}_{ij}^{\text{grp}} = \left[z_1^{\text{grp}}(x_{ij}) \;\ldots\; z_{\text{L}^{\text{grp}}}^{\text{grp}}(x_{ij})\right].$$

In addition, we define

$$\begin{aligned} \boldsymbol{\beta} = [\beta_0 \;\; \beta_1]^\top, \qquad \boldsymbol{\delta}_i = [\delta_{0i} \;\; \delta_{1i}]^\top, \\ \boldsymbol{u}^{\text{gbl}} = \left[u_1^{\text{gbl}} \;\ldots\; u_{\text{L}^{\text{gbl}}}^{\text{gbl}}\right]^\top, \qquad \boldsymbol{u}_i^{\text{grp}} = \left[u_{i1}^{\text{grp}} \;\ldots\; u_{i\text{L}^{\text{grp}}}^{\text{grp}}\right]^\top. \end{aligned} \tag{5.3}$$

We can now write (5.2) as

$$\begin{aligned} f(x_{ij}) &= \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{\text{gbl}}\boldsymbol{u}^{\text{gbl}}, \\ g_i(x_{ij}) &= \boldsymbol{x}_{ij}\boldsymbol{\delta}_i + \boldsymbol{z}_{ij}^{\text{grp}}\boldsymbol{u}_i^{\text{grp}}. \end{aligned}$$

We set $\text{Cov}(\boldsymbol{u}^{\text{gbl}}) = \sigma_{\text{gbl}}^2 \boldsymbol{I}$, $\text{Cov}(\boldsymbol{\delta}_i) = \boldsymbol{\Sigma}$, an unstructured 2×2 covariance matrix which allows for deviation from the typical linear component and lastly we set $\text{Cov}(\boldsymbol{u}_i^{\text{grp}}) = \sigma_{\text{grp}}^2 \boldsymbol{I}$.

We can now incorporate (5.1) into a mixed model framework by letting:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_{11} \\ \vdots \\ \boldsymbol{x}_{1n_1} \\ \vdots \\ \boldsymbol{x}_{m1} \\ \vdots \\ \boldsymbol{x}_{mn_m} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \text{and} \tag{5.4}$$

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{z}_{11}^{\text{gbl}} & \boldsymbol{x}_{11} & \boldsymbol{z}_{11}^{\text{grp}} & \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{0} & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{z}_{1n_1}^{\text{gbl}} & \boldsymbol{x}_{1n_1} & \boldsymbol{z}_{1n_1}^{\text{grp}} & \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{z}_{21}^{\text{gbl}} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{x}_{21} & \boldsymbol{z}_{21}^{\text{grp}} & \dots & \boldsymbol{0} & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{z}_{2n_2}^{\text{gbl}} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{x}_{2n_2} & \boldsymbol{z}_{2n_2}^{\text{grp}} & \dots & \boldsymbol{0} & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{z}_{m1}^{\text{gbl}} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{x}_{m1} & \boldsymbol{z}_{m1}^{\text{grp}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{z}_{mn_m}^{\text{gbl}} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{x}_{mn_m} & \boldsymbol{z}_{mn_m}^{\text{grp}} \end{bmatrix}, \quad \boldsymbol{u} = \begin{bmatrix} \boldsymbol{u}^{\text{gbl}} \\ \boldsymbol{\delta}_1 \\ \boldsymbol{u}_1^{\text{grp}} \\ \boldsymbol{\delta}_2 \\ \boldsymbol{u}_2^{\text{grp}} \\ \vdots \\ \boldsymbol{\delta}_m \\ \boldsymbol{u}_m^{\text{grp}} \end{bmatrix}, \tag{5.5}$$

so we can use the mixed model representation

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

$$\text{where} \quad \text{Cov}\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \sigma_{\text{gbl}}^2 \boldsymbol{I}_{L_{\text{gbl}}} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \underset{1 \le i \le m}{\text{blockdiag}} \left(\boldsymbol{\Sigma}, \sigma_{\text{grp}}^2 \boldsymbol{I}_{L_{\text{grp}}} \right) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \sigma_\varepsilon^2 \boldsymbol{I} \end{bmatrix}. \tag{5.6}$$

This semiparametric regression model can be treated using the following Gaussian linear mixed model:

$$\boldsymbol{y}|\boldsymbol{u} \sim \text{N}\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \sigma_\varepsilon^2 \boldsymbol{I}\right), \quad \boldsymbol{u} \sim \text{N}\left(\boldsymbol{0}, \begin{bmatrix} \sigma_{\text{gbl}}^2 \boldsymbol{I}_{L_{\text{gbl}}} & \boldsymbol{0} \\ \boldsymbol{0} & \underset{1 \le i \le m}{\text{blockdiag}} \left(\boldsymbol{\Sigma}, \sigma_{\text{grp}}^2 \boldsymbol{I}_{L_{\text{grp}}} \right) \end{bmatrix}\right). \tag{5.7}$$

5.2.2 Bayesian inference

Fitting (5.7) using standard mixed model software such as `lme()` (Pinheiro *et al.*, 2009) in the `R` computing language is achievable, but time consuming. Alternatively, we can work with a hierarchical Bayesian version of (5.7) which warrants direct implementation in standard software. We enforce σ_{gbl}, the $\sigma_{\text{grp},i}$, $1 \le i \le m$, and the square-rooted diagonal entries of $\boldsymbol{\Sigma}$ to possess Half-t prior distributions. This specification also allows for corre-

lation parameters within $\boldsymbol{\Sigma}$ to have uniform distributions on $(-1, 1)$, as shown in Huang *et al.* (2013). Such priors are attained through auxiliary variable constructions exhibited in Result 1.4.6.

In its entirety, the Bayesian hierarchical two-level Gaussian response model we deal with here is:

$$\begin{aligned}
\boldsymbol{y}\,|\,\boldsymbol{\beta},\,\boldsymbol{u},\,\sigma_{\varepsilon}^2 &\sim \mathrm{N}\left(\boldsymbol{X}\,\boldsymbol{\beta}+\boldsymbol{Z}\,\boldsymbol{u},\,\sigma_{\varepsilon}^2\boldsymbol{I}\right), \qquad \boldsymbol{\beta}\sim \mathrm{N}\left(\boldsymbol{0},\sigma_{\beta}^2\boldsymbol{I}_2\right),\\
\boldsymbol{u}|\sigma_{\mathrm{gbl}}^2,\,\Sigma,\,\sigma_{\mathrm{grp}}^2 &\sim \mathrm{N}\left(\boldsymbol{0},\begin{bmatrix}\sigma_{\mathrm{gbl}}^2\boldsymbol{I}_{L_{\mathrm{gbl}}} & \boldsymbol{0}\\ \boldsymbol{0} & \underset{1\le i\le m}{\mathrm{blockdiag}}\left(\boldsymbol{\Sigma},\,\sigma_{\mathrm{grp}}^2\boldsymbol{I}_{L_{\mathrm{grp}}}\right)\end{bmatrix}\right),\\
\sigma_{\varepsilon}^2\,|\,a_{\varepsilon} &\sim \text{Inverse-Gamma}\left(\tfrac{1}{2},\,1/a_{\varepsilon}\right), \qquad a_{\varepsilon}\sim \text{Inverse-Gamma}\left(\tfrac{1}{2},\,1/A_{\varepsilon}^2\right),\\
\sigma_{\mathrm{gbl}}^2\,|\,a_{\mathrm{gbl}} &\sim \text{Inverse-Gamma}\left(\tfrac{1}{2},\,1/a_{\mathrm{gbl}}\right), \qquad a_{\mathrm{gbl}}\sim \text{Inverse-Gamma}\left(\tfrac{1}{2},\,1/A_{\mathrm{gbl}}^2\right),\\
\boldsymbol{\Sigma}\,|\,a_{\Sigma,1},\,a_{\Sigma,2} &\sim \text{Inverse-Wishart}\left(\nu+2-1,\,2\,\nu\,\mathrm{diag}\left(1/a_{\Sigma,1},\,1/a_{\Sigma,2}\right)\right),\\
a_{\Sigma,j} &\sim \text{Inverse-Gamma}\left(\tfrac{1}{2},\,1/A_{\Sigma,j}^2\right), \quad 1\le j\le 2,\\
\sigma_{\mathrm{grp}}^2\,|\,a_{\mathrm{grp}} &\sim \text{Inverse-Gamma}\left(\tfrac{1}{2},\,1/a_{\mathrm{grp}}\right), \qquad a_{\mathrm{grp}}\sim \text{Inverse-Gamma}\left(\tfrac{1}{2},\,1/A_{\mathrm{grp}}^2\right).
\end{aligned} \tag{5.8}$$

where a_{ε}, a_{gbl}, $a_{\Sigma,j}$, and a_{grp} represent the auxiliary variables that impose the aforementioned non-informative prior distributions on σ_{ε}^2, σ_{gbl}^2, $\boldsymbol{\Sigma}$, and σ_{grp}^2, respectively. We also note that (5.8) encompasses the setting of hyperparameters ν, σ_{β}, A_{ε}, A_{gbl}, $A_{\Sigma,1}$, $A_{\Sigma,2}$ and A_{grp}, all of which are bound to be positive. The recommended setting for ν is 2, which ensures that Property 4 of Huang *et al.* (2013) holds. Non-informativity is achieved for the remaining hyperparameters when they are set to very large values. Our default setting, assuming the data has been transformed to have mean zero and unit standard deviation, is 10^5.

Figure 5.2 shows a directed acyclic graph representation of model (5.8) where $\boldsymbol{a}_{\Sigma} = [a_{\Sigma,1},\,a_{\Sigma,2}]^{\top}$, $\boldsymbol{\delta} = [\boldsymbol{\delta}_1,\ldots,\boldsymbol{\delta}_m]^{\top}$ and $\boldsymbol{u}_{\mathrm{grp}} = [\boldsymbol{u}_1^{\mathrm{grp}},\ldots,\boldsymbol{u}_m^{\mathrm{grp}}]^{\top}$.

5.2.3 Mean field variational Bayes methodology

The underpinning assumption of the MFVB approach for model (5.8) is to impose an approximate product density restriction to the joint posterior density function of the form

$$\begin{aligned}
p(\boldsymbol{\beta},\boldsymbol{u},a_{\varepsilon},a_{\mathrm{gbl}},\boldsymbol{a}_{\Sigma},a_{\mathrm{grp}},\sigma_{\varepsilon}^2,\sigma_{\mathrm{gbl}}^2,\boldsymbol{\Sigma},\sigma_{\mathrm{grp}}^2|\boldsymbol{y}) \;\approx\; & q(\boldsymbol{\beta},\boldsymbol{u},a_{\varepsilon},a_{\mathrm{gbl}},\boldsymbol{a}_{\Sigma},a_{\mathrm{grp}})\\
& \times q(\sigma_{\varepsilon}^2,\sigma_{\mathrm{gbl}}^2,\boldsymbol{\Sigma},\sigma_{\mathrm{grp}}^2).
\end{aligned} \tag{5.9}$$

Simplicity of the requisite calculations arise as a result of the graphical layout of model (5.8) as conveyed by Figure 5.2. The graphical structure of (5.8) is useful in determining

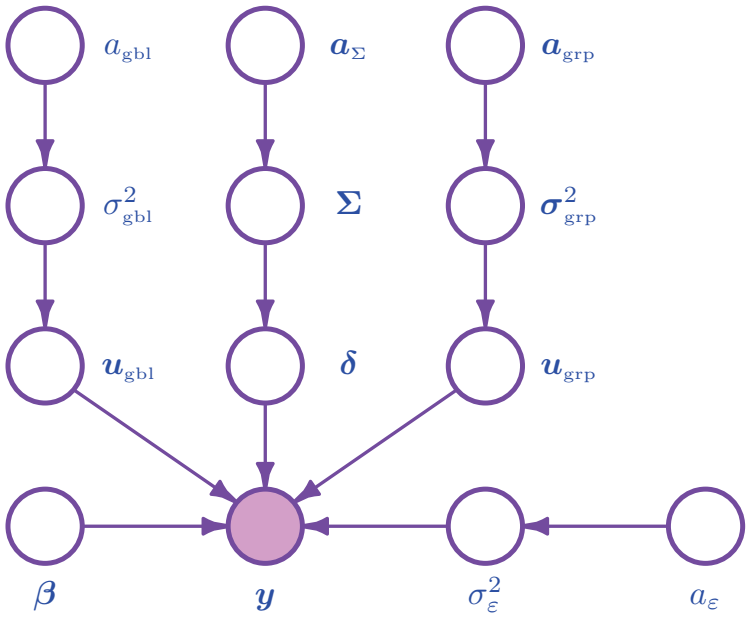


Figure 5.2: *Directed acyclic graph corresponding to the Bayesian hierarchical two-level Gaussian response model in (5.8).*

additional factorizations, i.e., *induced factorizations*. Factorizations such as these, can be observed using a straightforward graphical test, known as *moralisation*, as explained in Section 1.8.2. For example, all paths between σ^2_{ε} and $\{\sigma^2_{\text{gbl}}, \boldsymbol{\Sigma}, \sigma^2_{\text{grp}}\}$ in Figure 5.2 must visit at least one of $\{\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{u}\}$. That is, the set of nodes $\{\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{u}\}$ separates σ^2_{ε} from the set $\{\sigma^2_{\text{gbl}}, \boldsymbol{\Sigma}, \sigma^2_{\text{grp}}\}$. The use of Theorem 1.8.1 gives the result

$$\sigma^2_{\varepsilon} \perp\!\!\!\perp \{\sigma^2_{\text{gbl}}, \boldsymbol{\Sigma}, \sigma^2_{\text{grp}}\} | \{\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{u}\}.$$

Continuing this process for all nodes in Figure (5.2), we achieve the following induced factorization:

$$\begin{aligned} &q(\boldsymbol{\beta}, \boldsymbol{u}, a_{\varepsilon}, a_{\text{gbl}}, \boldsymbol{a}_{\Sigma}, a_{\text{grp}}, \sigma^2_{\varepsilon}, \sigma^2_{\text{gbl}}, \sigma^2_{\text{grp}}, \boldsymbol{\Sigma}) \approx \\ &\qquad q(\boldsymbol{\beta}, \boldsymbol{u})\, q(\sigma^2_{\varepsilon})\, q(\sigma^2_{\text{gbl}})\, q(\boldsymbol{\Sigma})\, q(\sigma^2_{\text{grp}})\, q(a_{\varepsilon})\, q(a_{\text{gbl}}) \left\{ \prod_{j=1}^{2} q(a_{\Sigma,j}) \right\} q(a_{\text{grp}}). \end{aligned} \tag{5.10}$$

The q-densities are chosen to minimize the Kullback-Leibler distance between the left-hand-side and right-hand-side of (5.9). As outlined in Section 1.5 the optimal q-densities, denoted by q^*, can be obtained through an iterative scheme arising from relationships such as

$$q^*(\boldsymbol{\Sigma}) \propto \exp E_q \{\log p(\boldsymbol{\Sigma}|\text{rest})\}, \tag{5.11}$$

where 'rest' denotes the random variables in the model not including $\boldsymbol{\Sigma}$. The full conditional distributions each have standard forms resulting in closed form expressions for the optimal q-densities. As a result, the optimal factors on the right-hand side of (5.10) have the following forms (details of the derivation can be found in Appendix 5.A):

$$
\begin{aligned}
&q^*(\boldsymbol{\beta}, \boldsymbol{u}) \text{ is the } N(\boldsymbol{\mu}_{q(\boldsymbol{\beta},u)}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},u)}) \text{ density function,} \\
&q^*(\sigma^2_\varepsilon) \text{ is the Inverse-Gamma}\left(\tfrac{1}{2}\left(\sum_{i=1}^{m} n_i + 1\right), B_{q(\sigma^2_\varepsilon)}\right) \text{ density function,} \\
&q^*(\sigma^2_{\text{gbl}}) \text{ is the Inverse-Gamma}(\tfrac{1}{2}(L_{\text{gbl}} + 1), B_{q(\sigma^2_{\text{gbl}})}) \text{ density function,} \\
&q^*(\boldsymbol{\Sigma}) \text{ is the Inverse-Wishart}(\nu + m + 1, B_{q(\boldsymbol{\Sigma})}) \text{ density function,} \\
&q^*(\sigma^2_{\text{grp}}) \text{ is the Inverse-Gamma}(\tfrac{1}{2}(m\, L_{\text{grp}} + 1), B_{q(\sigma^2_{\text{grp}})}) \text{ density function,} \\
&q^*(a_\varepsilon) \text{ is the Inverse-Gamma}(1, B_{q(a_\varepsilon)}) \text{ density function,} \\
&q^*(a_{\text{gbl}}) \text{ is the Inverse-Gamma}(1, B_{q(a_{\text{gbl}})}) \text{ density function,} \\
&q^*(a_{\Sigma,j}) \text{ is the Inverse-Gamma}(\tfrac{\nu}{2} + 1, B_{q(a_{\Sigma,j})}) \text{ density function,} \quad 1 \le j \le 2, \\
&q^*(a_{\text{grp}}) \text{ is the Inverse-Gamma}(1, B_{q(a_{\text{grp}})}) \text{ density function,}
\end{aligned}
\tag{5.12}
$$

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 rate parameter of $q^*(\sigma^2_\varepsilon)$, $B_{q(\sigma^2_{\text{gbl}})}$, the rate parameter of $q^*(\sigma^2_{\text{gbl}})$, $B_{q(\boldsymbol{\Sigma})}$, the rate matrix of $q^*(\boldsymbol{\Sigma})$, $B_{q(\sigma^2_{\text{grp}})}$, the rate parameter of $q^*(\sigma^2_{\text{grp}})$, $B_{q(a_\varepsilon)}$, the rate parameter of $q^*(a_\varepsilon)$, $B_{q(a_{\text{gbl}})}$, the rate parameter of $q^*(a_{\text{gbl}})$, $B_{q(a_{\Sigma,j})}$, the rate parameter of $q^*(a_{\Sigma,j})$ and $B_{q(a_{\text{grp}})}$, the rate parameter of $q^*(a_{\text{grp}})$. The parameters in these optimal q-densities are obtained through a naïve implementation which results in Algorithm 5 and uses the notation $\boldsymbol{C} \equiv [\boldsymbol{X} \,|\, \boldsymbol{Z}]$. As seen from (5.5) the $\boldsymbol{Z}$ matrix requires storage of zero's which can become computationally expensive given a large value for m. We aim to alleviate the restriction of storing these zero's in Section 5.2.4.

Convergence of Algorithm 5 is determined using the marginal log-likelihood lower

Set up initial values:

$\mu_{q(1/\sigma_\varepsilon^2)}, \mu_{q\left(1/\sigma_{\text{gbl}}^2\right)}, \mu_{q\left(1/\sigma_{\text{grp}}^2\right)}, \mu_{q(1/a_\varepsilon)}, \mu_{q\left(1/a_{\text{gbl}}\right)}, \mu_{q(1/a_{\text{grp}})} > 0$, $\mu_{q\left(1/a_{\Sigma,j}\right)} > 0$,

$1 \le j \le 2$, $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$, and $\boldsymbol{M}_{q\left(\boldsymbol{\Sigma}^{-1}\right)}$ positive definite.

Cycle through:

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

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \Big(\mu_{q(1/\sigma_\varepsilon^2)} \boldsymbol{C}^\top \boldsymbol{C} + \begin{bmatrix} \sigma_{\boldsymbol{\beta}}^{-2} \boldsymbol{I}_2 & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \mu_{q\left(1/\sigma_{\text{gbl}}^2\right)} \boldsymbol{I}_{L_{\text{gbl}}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \underset{1\le i\le m}{\text{blockdiag}} \left(\boldsymbol{M}_{q\left(\boldsymbol{\Sigma}^{-1}\right)}, \mu_{q\left(1/\sigma_{\text{grp}}^2\right)} \boldsymbol{I}_{L_{\text{grp}}} \right) \end{bmatrix} \Big)^{-1}$$

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

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

$$B_{q(a_\varepsilon)} \leftarrow \mu_{q(1/\sigma_\varepsilon^2)} + A_\varepsilon^{-2}; \quad \mu_{q(1/a_\varepsilon)} \leftarrow 1/B_{q(a_\varepsilon)}$$

$$B_{q\left(\sigma_{\text{gbl}}^2\right)} \leftarrow \tfrac{1}{2} \left\{ \| \, \boldsymbol{\mu}_{q(\boldsymbol{u}^{\text{gbl}})} \, \|^2 + \text{tr} \left(\boldsymbol{\Sigma}_{q(\boldsymbol{u}^{\text{gbl}})} \right) \right\} + \mu_{q\left(1/a_{\text{gbl}}\right)}$$

$$\mu_{q\left(1/\sigma_{\text{gbl}}^2\right)} \leftarrow \tfrac{1}{2} \left(L_{\text{gbl}} + 1 \right) / B_{q\left(\sigma_{\text{gbl}}^2\right)}$$

$$B_{q(a_{\text{gbl}})} \leftarrow \mu_{q\left(1/\sigma_{\text{gbl}}^2\right)} + A_{\text{gbl}}^{-2}; \quad \mu_{q\left(1/a_{\text{gbl}}\right)} \leftarrow 1/B_{q(a_{\text{gbl}})}$$

$$B_{q\left(\sigma_{\text{grp}}^2\right)} \leftarrow \tfrac{1}{2} \left\{ \| \, \boldsymbol{\mu}_{q(\boldsymbol{u}^{\text{grp}})} \, \|^2 + \text{tr} \left(\boldsymbol{\Sigma}_{q(\boldsymbol{u}^{\text{grp}})} \right) \right\} + \mu_{q(1/a_{\text{grp}})}$$

$$\mu_{q\left(1/\sigma_{\text{grp}}^2\right)} \leftarrow \tfrac{1}{2} \left(m \, L_{\text{grp}} + 1 \right) / B_{q\left(\sigma_{\text{grp}}^2\right)}$$

$$B_{q(a_{\text{grp}})} \leftarrow \mu_{q\left(1/\sigma_{\text{grp}}^2\right)} + A_{\text{grp}}^{-2}; \quad \mu_{q(1/a_{\text{grp}})} \leftarrow 1/B_{q(a_{\text{grp}})}$$

$$\boldsymbol{B}_{q(\boldsymbol{\Sigma})} \leftarrow \sum_{i=1}^m \left\{ \boldsymbol{\mu}_{q(\boldsymbol{\delta}_i)} \boldsymbol{\mu}_{q(\boldsymbol{\delta}_i)}^\top + \boldsymbol{\Sigma}_{q(\boldsymbol{\delta}_i)} \right\} + 2\nu \text{diag} \left(\mu_{q\left(1/a_{\Sigma,1}\right)}, \mu_{q\left(1/a_{\Sigma,2}\right)} \right)$$

$$\boldsymbol{M}_{q\left(\boldsymbol{\Sigma}^{-1}\right)} \leftarrow (\nu + m + 1) \, \boldsymbol{B}_{q(\boldsymbol{\Sigma})}^{-1}$$

For $j = 1, 2$:

$$B_{q(a_{\Sigma,j})} \leftarrow \nu \left(\boldsymbol{M}_{q\left(\boldsymbol{\Sigma}^{-1}\right)} \right)_{jj} + 1/A_{\Sigma,j}^2, \quad \mu_{q\left(1/a_{\Sigma,j}\right)} \leftarrow \left(\tfrac{\nu}{2} + 1 \right) / B_{q(a_{\Sigma,j})}$$

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

Algorithm 5: *Naïve MFVB algorithm for the estimation of the optimal parameters in* $q^* (\boldsymbol{\beta}, \boldsymbol{u})$, $q^* \left(\sigma_\varepsilon^2\right)$, $q^* (a_\varepsilon)$, $q^* \left(\sigma_{gbl}^2\right)$, $q^* (a_{gbl})$, $q^* \left(\sigma_{grp}^2\right)$, $q^* (a_{grp})$, $q^* (\boldsymbol{\Sigma})$, *and* $q^* (\boldsymbol{a}_\Sigma)$.

bound (Appendix 5.B shows details on this derivation):

$$
\begin{aligned}
\log \underline{p}(\boldsymbol{y}; q) \;=\; & \tfrac{1}{2}(\nu + p - 1)\log(2\nu) - \tfrac{1}{2}\sum_{i=1}^{m} n_i \log(2\pi) - 4\log(\pi) - \tfrac{p}{2}\log(\sigma_\beta^2) \\
& - \tfrac{1}{\sigma_\beta^2}\left\{ \| \boldsymbol{\mu}_{q(\boldsymbol{\beta})} \|^2 + \mathrm{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right)\right\} + \tfrac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}| \\
& + \tfrac{1}{2}\left\{p + L_{\text{gbl}} + m(p + L_{\text{grp}})\right\} - \log\left(C_{p,\nu+p-1}\right) + \log\left(C_{p,\nu+m+p-1}\right) \\
& - \tfrac{1}{2}\left(\nu + p + m - 1\right)\log|\boldsymbol{B}_{q(\boldsymbol{\Sigma})}| + \log\Gamma\left\{\tfrac{1}{2}\left(\sum_{i=1}^{m} n_i + 1\right)\right\} \\
& - \tfrac{1}{2}\left(\sum_{i=1}^{m} n_i + 1\right)\log\left(B_{q(\sigma_\varepsilon^2)}\right) + \log\Gamma\left\{\tfrac{1}{2}\left(L_{\text{gbl}} + 1\right)\right\} \\
& - \tfrac{1}{2}\left(L_{\text{gbl}} + 1\right)\log\left(B_{q\left(\sigma_{\text{gbl}}^2\right)}\right) + \log\Gamma\left\{\tfrac{1}{2}\left(m \times L_{\text{grp}} + 1\right)\right\} \\
& - \tfrac{1}{2}\left(m \times L_{\text{grp}} + 1\right)\log\left(B_{q\left(\sigma_{\text{grp}}^2\right)}\right) - \log\left(A_\varepsilon\right) - \log\left(B_{q(a_\varepsilon)}\right) \\
& + \mu_{q(1/\sigma_\varepsilon^2)}\mu_{q(1/a_\varepsilon)} - \log\left(A_{\text{gbl}}\right) - \log\left(B_{q\left(a_{\text{gbl}}\right)}\right) + \mu_{q\left(1/\sigma_{\text{gbl}}^2\right)}\mu_{q\left(1/a_{\text{gbl}}\right)} \\
& - \log\left(A_{\text{grp}}\right) - \log\left(B_{q(a_{\text{grp}})}\right) + \mu_{q(1/\sigma_{\text{grp}}^2)}\mu_{q(1/a_{\text{grp}})} - \sum_{j=1}^{p}\log\left(A_{\Sigma,j}\right) \\
& + p\log\Gamma\left\{\tfrac{1}{2}\left(\nu + p\right)\right\} - \tfrac{1}{2}\left(\nu + p - 1\right)\sum_{j=1}^{p}\log\left(B_{q\left(a_{\Sigma,j}\right)}\right) \\
& + \sum_{j=1}^{p}\nu\left(\boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1})}\right)_{jj}\mu_{q(1/a_{\Sigma,j})},
\end{aligned}
\tag{5.13}
$$

where $p = 2$ is the number of columns in the $\boldsymbol{X}$ matrix, which also corresponds to the dimension of $\boldsymbol{\Sigma}$.

5.2.4 Streamlining Mean field variational Bayes for the two-level Gaussian response model

The naïve nature of the implementation given by Algorithm 5, stems from the fact that the update for the covariance matrix $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ requires storing and inverting a matrix which has its row length equal to that of the $\boldsymbol{C}$ matrix and thus can become infeasible for large multilevel/longitudinal datasets. Lee & Wand (2015) develop a streamlined MFVB algorithm for efficient fitting and inference in large longitudinal and multilevel data sets. In particular they develop streamlined MFVB algorithms for the two-level Gaussian response model and the two-level binary response model. We focus on the former algorithm and slightly adjust their Algorithm 2 to cater to (5.8).

In keeping with the notation used in Lee & Wand (2015) we further partition the

following vectors and matrices:

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_m \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_m \end{bmatrix}, \quad \boldsymbol{Z}^{\text{gbl}} = \begin{bmatrix} \boldsymbol{Z}_1^{\text{gbl}} \\ \vdots \\ \boldsymbol{Z}_m^{\text{gbl}} \end{bmatrix}, \quad \text{and} \quad \boldsymbol{Z}^{\text{R}} = \underset{1 \leq i \leq m}{\text{blockdiag}} \left(\boldsymbol{Z}_i^{\text{R}} \right),$$

where

$$\boldsymbol{y}_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{in_i} \end{bmatrix}, \quad \boldsymbol{X}_i = \begin{bmatrix} 1 & x_{i1} \\ \vdots & \vdots \\ 1 & x_{in_i} \end{bmatrix}, \quad \boldsymbol{Z}_i^{\text{gbl}} = \begin{bmatrix} z_{i1}^{\text{gbl}\top} \\ \vdots \\ z_{in_i}^{\text{gbl}\top} \end{bmatrix}, \quad \text{and} \quad \boldsymbol{Z}_i^{\text{R}} = \begin{bmatrix} x_{i1}^{\top} & z_{i1}^{\text{grp}\top} \\ \vdots & \vdots \\ x_{in_i}^{\top} & z_{in_i}^{\text{grp}\top} \end{bmatrix}.$$

In addition, we partition the following global and random effects:

$$\boldsymbol{u} = \begin{bmatrix} \boldsymbol{u}^{\text{gbl}} \\ \boldsymbol{u}^{\text{R}} \end{bmatrix}, \quad \text{and} \quad \boldsymbol{Z} = \left[\boldsymbol{Z}^{\text{gbl}} \;\; \boldsymbol{Z}^{\text{R}} \right],$$

where $\boldsymbol{u}^{\text{R}} = \left[(\boldsymbol{u}_1^{\text{R}})^{\top} \ldots (\boldsymbol{u}_m^{\text{R}})^{\top} \right]^{\top}$, $\boldsymbol{u}_i^{\text{R}} = \left[\boldsymbol{\delta}_i^{\top} \; \left(\boldsymbol{u}_i^{\text{grp}} \right)^{\top} \right]^{\top}$ and $\boldsymbol{u}^{\text{gbl}}$ is defined in (5.3). Finally, Lee & Wand (2015) suggest the useful notation $\boldsymbol{C}^{\text{gbl}} = \left[\boldsymbol{X} \; \boldsymbol{Z}^{\text{gbl}} \right]$ where $\boldsymbol{C}_i^{\text{gbl}} = \left[\boldsymbol{X}_i \; \boldsymbol{Z}_i^{\text{gbl}} \right]$ is the sub-matrix of $\boldsymbol{C}^{\text{gbl}}$ corresponding to the ith subject. With the aid of this new notation, a streamlined version of Algorithm 5 is presented in Algorithm 6.

Convergence of Algorithm 6 is determined using the marginal log-likelihood lower bound:

$$\begin{aligned}
\log \underline{p}(\boldsymbol{y}; q) \;=\; & \tfrac{1}{2} p(\nu + p - 1) \log(2\nu) - \tfrac{1}{2} \sum_{i=1}^{m} n_i \log(2\pi) - 4 \log(\pi) - \tfrac{p}{2} \log(\sigma_\beta^2) \\
& - \tfrac{1}{2\sigma_\beta^2} \left\{ \| \, \boldsymbol{\mu}_{q(\boldsymbol{\beta})} \, \|^2 + \text{tr} \left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} \right) \right\} + \tfrac{1}{2} (p + L_{\text{gbl}} + m(p + L_{\text{grp}})) \\
& - \tfrac{1}{2} \sum_{i=1}^{m} \log \left| \mu_{q(1/\sigma_\varepsilon^2)} \left(\boldsymbol{Z}_i^{\text{R}} \right)^{\top} \boldsymbol{Z}_i^{\text{R}} + \text{blockdiag} \left(M_{q(\boldsymbol{\Sigma}^{-1})}, \mu_{q(1/\sigma_{\text{grp}}^2)} \boldsymbol{I}_{L_{\text{grp}}} \right) \right| \\
& - \tfrac{1}{2} \log |\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}, \boldsymbol{u}^{\text{gbl}})}| - \log \left(C_{p, \nu+p-1} \right) + \log \left(C_{p, \nu+m+p-1} \right) \\
& + \log \Gamma \left\{ \tfrac{1}{2} \left(\sum_{i=1}^{m} n_i + 1 \right) \right\} - \tfrac{1}{2} \left(\sum_{i=1}^{m} n_i + 1 \right) \log \left(B_{q(\sigma_\varepsilon^2)} \right) + \log \Gamma \left(\tfrac{1}{2} \left(L_{\text{gbl}} + 1 \right) \right) \\
& - \tfrac{1}{2} \left(L_{\text{gbl}} + 1 \right) \log \left(B_{q\left(\sigma_{\text{gbl}}^2\right)} \right) + \log \Gamma \left\{ \tfrac{1}{2} \left(m \times L_{\text{grp}} + 1 \right) \right\} + p \log \Gamma \left(\tfrac{1}{2} \left(\nu + p \right) \right) \\
& - \tfrac{1}{2} \left(m L_{\text{grp}} + 1 \right) \log \left(B_{q\left(\sigma_{\text{grp}}^2\right)} \right) - \log \left(A_\varepsilon \right) - \log \left(B_{q(a_\varepsilon)} \right) + \mu_{q(1/\sigma_\varepsilon^2)} \mu_{q(1/a_\varepsilon)} \\
& - \log \left(A_{\text{gbl}} \right) - \log \left(B_{q\left(a_{\text{gbl}}\right)} \right) + \mu_{q\left(1/\sigma_{\text{gbl}}^2\right)} \mu_{q(1/a_{\text{gbl}})} - \log \left(A_{\text{grp}} \right) - \log \left(B_{q\left(a_{\text{grp}}\right)} \right) \\
& + \mu_{q(1/\sigma_{\text{grp}}^2)} \mu_{q(1/\sigma_{\text{grp}}^2)} - \sum_{j=1}^{p} \log \left(A_{\Sigma, j} \right) - \tfrac{1}{2} \left(\nu + p - 1 \right) \sum_{j=1}^{p} \log \left(B_{q\left(a_{\Sigma, j}\right)} \right) \\
& + \sum_{j=1}^{p} \nu \left(\boldsymbol{M}_{q\left(\boldsymbol{\Sigma}^{-1}\right)} \right)_{jj} \mu_{q\left(1/a_{\Sigma, j}\right)} - \tfrac{1}{2} \left(\nu + p + m - 1 \right) \log |\boldsymbol{B}_{q(\boldsymbol{\Sigma})}|.
\end{aligned}$$

Set up initial values:

$\mu_{q(1/\sigma_\varepsilon^2)} > 0$, $\mu_{q(1/a_\varepsilon)} > 0$, $\mu_{q(1/\sigma^2_{\text{gbl}})} > 0$, $\mu_{q(1/\sigma^2_{\text{grp}})} > 0$, $\mu_{q\left(1/a_{\Sigma,j}\right)} > 0$, $1 \le j \le 2$, $\boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1})}$ positive definite.

Cycle through:

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

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

$$\boldsymbol{G}_i \leftarrow \mu_{q(1/\sigma_\varepsilon^2)} \left(\boldsymbol{C}_i^{\text{gbl}}\right)^\top \boldsymbol{Z}_i^{\text{R}}$$

$$\boldsymbol{H}_i \leftarrow \left\{\mu_{q(1/\sigma_\varepsilon^2)} \left(\boldsymbol{Z}_i^{\text{R}}\right)^\top \boldsymbol{Z}_i^{\text{R}} + \text{blockdiag}\left(M_{q(\boldsymbol{\Sigma}^{-1})}, \mu_{q\left(1/\sigma^2_{\text{grp}}\right)} \boldsymbol{I}_{L_{\text{grp}}}\right)\right\}^{-1}$$

$$\boldsymbol{S} \leftarrow \boldsymbol{S} + \boldsymbol{G}_i \boldsymbol{H}_i \boldsymbol{G}_i^\top; \quad \boldsymbol{s} \leftarrow \boldsymbol{s} + \boldsymbol{G}_i \boldsymbol{H}_i \left(\boldsymbol{Z}_i^{\text{R}}\right)^\top \boldsymbol{y}_i$$

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\text{gbl}})} \leftarrow \left\{\mu_{q(1/\sigma_\varepsilon^2)} \left(\boldsymbol{C}^{\text{gbl}}\right)^\top \boldsymbol{C}^{\text{gbl}} + \begin{bmatrix} \sigma_{\boldsymbol{\beta}}^{-2} \boldsymbol{I}_2 & \boldsymbol{0} \\ \boldsymbol{0} & \mu_{q\left(1/\sigma^2_{\text{gbl}}\right)} \boldsymbol{I}_{L_{\text{gbl}}} \end{bmatrix} - S\right\}^{-1}$$

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

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

$$\boldsymbol{\Sigma}_{q\left(\boldsymbol{u}_i^{\text{R}}\right)} \leftarrow \boldsymbol{H}_i + \boldsymbol{H}_i \boldsymbol{G}_i^\top \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\text{gbl}})} \boldsymbol{G}_i \boldsymbol{H}_i$$

$$\boldsymbol{\mu}_{q\left(\boldsymbol{u}_i^{\text{R}}\right)} \leftarrow \boldsymbol{H}_i \left\{\mu_{q(1/\sigma_\varepsilon^2)} \left(\boldsymbol{Z}_i^{\text{R}}\right)^\top \boldsymbol{y}_i - \boldsymbol{G}_i^\top \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\text{gbl}})}\right\}$$

$$B_{q(\sigma_\varepsilon^2)} \leftarrow \mu_{q(1/a_\varepsilon)} + \frac{1}{2}\Bigg[\left\|\boldsymbol{y} - \boldsymbol{C}^{\text{gbl}} \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\text{gbl}})} - \begin{bmatrix} \boldsymbol{Z}_1^{\text{R}} \boldsymbol{\mu}_{q\left(\boldsymbol{u}_1^{\text{R}}\right)} \\ \vdots \\ \boldsymbol{Z}_m^{\text{R}} \boldsymbol{\mu}_{q\left(\boldsymbol{u}_m^{\text{R}}\right)} \end{bmatrix}\right\|^2$$

$$+\text{tr}\left\{\left(\boldsymbol{C}^{\text{gbl}}\right)^\top \boldsymbol{C}^{\text{gbl}} \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\text{gbl}})}\right\} + \sum_{i=1}^m \text{tr}\left\{\left(\boldsymbol{Z}_i^{\text{R}}\right)^\top \boldsymbol{Z}_i^{\text{R}} \boldsymbol{\Sigma}_{q\left(\boldsymbol{u}_i^{\text{R}}\right)}\right\}$$

$$-2\mu_{q(1/\sigma_\varepsilon^2)}^{-1} \sum_{i=1}^m \text{tr}\left(\boldsymbol{G}_i \boldsymbol{H}_i \boldsymbol{G}_i^\top \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\text{gbl}})}\right)\Bigg]$$

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

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

$$\mu_{q\left(1/\sigma^2_{\text{gbl}}\right)} \leftarrow \frac{1}{2}\left(L_{\text{gbl}} + 1\right) / B_{q\left(\sigma^2_{\text{gbl}}\right)}, \quad \mu_{q(1/a_{\text{gbl}})} \leftarrow 1 \Big/ \left\{\mu_{q\left(1/\sigma^2_{\text{gbl}}\right)} + A_{\text{gbl}}^{-2}\right\}$$

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

$$\mu_{q\left(1/\sigma^2_{\text{grp}}\right)} \leftarrow \frac{1}{2}\left(m\, L_{\text{grp}} + 1\right) / B_{q\left(\sigma^2_{\text{grp}}\right)}; \quad \mu_{q(1/a_{\text{grp}})} \leftarrow 1 \Big/ \left\{\mu_{q\left(1/\sigma^2_{\text{grp}}\right)} + A_{\text{grp}}^{-2}\right\},$$

$$\boldsymbol{B}_{q(\boldsymbol{\Sigma})} \leftarrow \sum_{i=1}^{m} \left\{ \boldsymbol{\mu}_{q(\boldsymbol{\delta}_i)} \boldsymbol{\mu}_{q(\boldsymbol{\delta}_i)}^{\top} + \boldsymbol{\Sigma}_{q(\boldsymbol{\delta}_i)} \right\} + 2\nu \operatorname{diag}\left(\mu_{q\left(1/a_{\Sigma,1}\right)}, \mu_{q\left(1/a_{\Sigma,2}\right)} \right)$$

$$\boldsymbol{M}_{q\left(\boldsymbol{\Sigma}^{-1}\right)} \leftarrow (\nu + m + 1)\, \boldsymbol{B}_{q(\boldsymbol{\Sigma})}^{-1}$$

For $j = 1, 2$:

$$B_{q(a_{\Sigma,j})} \leftarrow \nu \left(\boldsymbol{M}_{q\left(\boldsymbol{\Sigma}^{-1}\right)} \right)_{jj} + 1/A_{\Sigma,j}^2, \quad \mu_{q\left(1/a_{\Sigma,j}\right)} \leftarrow \left(\tfrac{\nu}{2} + 1 \right) / B_{q(a_{\Sigma,j})}$$

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

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

$$\boldsymbol{\Lambda}_q(\boldsymbol{\beta}, \boldsymbol{u}^{\text{gbl}}, \boldsymbol{u}_i^{\text{R}}) \equiv E_q \left[\left\{ \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u}^{\text{gbl}} \end{bmatrix} - \boldsymbol{\mu}_{q(\boldsymbol{\beta}, \boldsymbol{u}^{\text{gbl}})} \right\} \left\{ \boldsymbol{u}_i^{\text{R}} - \boldsymbol{\mu}_q(\boldsymbol{u}_i^{\text{R}}) \right\}^{\top} \right] \leftarrow -\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}, \boldsymbol{u}^{\text{gbl}})}\, \boldsymbol{G}_i\, \boldsymbol{H}_i$$

Algorithm 6: *Streamlined MFVB algorithm for the two-level Gaussian response model given in (5.8).*

The difference between this lower bound expression and the one given in (5.2.3) is replacement of the term

$$\log |\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}, \boldsymbol{u})}|$$

$$\text{with} \quad -\sum_{i=1}^{m} \log \left| \mu_{q(1/\sigma_\varepsilon^2)} \left(\boldsymbol{Z}_i^{\text{R}} \right)^{\top} \boldsymbol{Z}_i^{\text{R}} + \operatorname{blockdiag}\left(M_{q\left(\boldsymbol{\Sigma}^{-1}\right)}, \mu_{q(1/\sigma_{\text{grp}}^2)} \boldsymbol{I}_{L_{\text{grp}}} \right) \right| - \log |\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}, \boldsymbol{u}^{\text{gbl}})}|,$$

which is used to streamline the calculation of $\log |\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}, \boldsymbol{u})}|$. Details on this calculation are given in the appendix of Lee & Wand (2015).

In order to produce variability estimates when plotting subject-specific mean estimates, Lee & Wand (2015) propose the following partition of the q-density covariance matrix of the vector of coefficients for the ith subject:

$$\operatorname{Cov}_q \left(\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u}^{\text{gbl}} \\ \boldsymbol{u}_i^{\text{R}} \end{bmatrix} \right) = \begin{bmatrix} \boldsymbol{\Sigma}_{q(\boldsymbol{\beta}, \boldsymbol{u}^{\text{gbl}})} & \boldsymbol{\Lambda}_q(\boldsymbol{\beta}, \boldsymbol{u}^{\text{gbl}}, \boldsymbol{u}_i^{\text{R}}) \\ \boldsymbol{\Lambda}_q(\boldsymbol{\beta}, \boldsymbol{u}^{\text{gbl}}, \boldsymbol{u}_i^{\text{R}})^{\top} & \boldsymbol{\Sigma}_{q(\boldsymbol{u}_i^{\text{R}})} \end{bmatrix}.$$

It is important to keep in mind that the time savings when using a streamlined MFVB approach for the Bayesian hierarchical two-level Gaussian response model given in (5.19) will not be as impressive as the ones achieved by using the two-level model in Lee & Wand (2015). The reason for this lies in the structure of the $\boldsymbol{H}_i$ matrix. For example, in Lee & Wand (2015) there exists no zero's in $\boldsymbol{H}_i$, however in (5.19), since we incorporate the variance of the spline components for the group specific curve of each individual there exists

blocks of zeros concerning the $\mu_{q(1/\sigma^2_{\text{grp}})}\boldsymbol{I}_{L_{\text{grp}}}$ term in Algorithm 6. As a result, this should not be considered a fully streamlined version of Algorithm 5 and further investigation into a more robust version of Algorithm 6 is certainly warranted.

5.2.5 Simulation study

We carried out an extensive simulation study in order to assess the speed of Algorithm 6 against that of Algorithm 5. We generated 25 data-sets according to (5.1) where the $x_{ij} \overset{\text{ind.}}{\sim} \text{Uniform}(0, 1)$, $u^{\text{gbl}}_{\ell} \overset{\text{ind.}}{\sim} \text{N}(0, \sigma^2_{\text{gbl}})$, $\varepsilon_{ij} \overset{\text{ind.}}{\sim} \text{N}(0, \sigma^2_{\varepsilon})$, $\boldsymbol{\delta}_i \overset{\text{ind.}}{\sim} \text{N}(\mathbf{0}, \boldsymbol{\Sigma})$ and the $u^{\text{grp}}_{i\ell} \overset{\text{ind.}}{\sim} \text{N}(0, \sigma^2_{\text{grp}})$. The true parameter values were specified as

$$\beta_0 = 0.2, \quad \beta_1 = 1.8, \quad \sigma^2_{\text{gbl}} = 1.5, \quad \sigma^2_{\varepsilon} = 0.05, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 0.4 & 0.1 \\ 0.1 & 0.4 \end{bmatrix} \text{ and } \sigma^2_{\text{grp}} = 0.29.$$

The number of subjects between simulation studies was varied, where $m \in \{25, 75, 125, 175\}$ and the within subject sample size remained constant at $n_i = 10$, $1 \leq i \leq m$.

Algorithm 5 and 6 were implemented in the R programming language. These computations were performed on a laptop computer (Mac OS X; 2.8 GHz processor, 16 GBytes of random access memory). Table 5.1 summarises the computation times for each approach. As the number of subjects m increase moderately, the average computing time for the naïve approach increases very quickly. The average time increase for the naïve approach from $m = 25$ to $m = 175$ is approximately 3 hours, compared to the streamlined approach which was just under 1 minute. This is represented in the *Ratio* column of Table 5.1 where we can see exactly how much faster the streamlined approach is in comparison to the naïve approach. Keeping in mind that the largest value of m is 175, the possibilities of time saving given larger values of m seem endless.

m	Naïve	Streamlined	Ratio
25	37.16 (17.49)	9.52 (3.75)	3.86
75	761.16 (248.39)	20.44 (6.13)	37.24
125	3178.27 (562.68)	31.45 (5.22)	101.06
175	11181.61 (4587.89)	56.49 (22.36)	197.94

Table 5.1: *Average (standard deviation) run time in seconds for naïve and streamlined MFVB fitting of the two-level Gaussian response model in (5.1).*

5.2.6 Application to data from a growth study

In this section we provide an example of streamlined MFVB fitting for growth data on children in Indianapolis (source: Pratt *et al.*, 1989). These data possess a two-level structure where the repeated measurements on height and age are at the first level, and the persons being measured are at the second level. In this data-set a person is identified as being either *black* or *white* by the use of the indicator variable `black`. We only use the information from this data-set relating to male individuals. This gives $m = 116$, and each n_i takes on a value between 10 and 26 repeated measurements each. The variables and their description are given in Table 5.2.

Variable	Description
`id`	person identifier
`height`	person's height in centimeters
`age`	person's age
`black`	indicator of person being black

Table 5.2: *Description of the Indiana growth data for male subjects.*

We are interested in what the effect of being either *black* or *white* would have on a person's height. Thus an interaction term would be necessary in our model to account for this. The Indiana growth data is divided into two groups:

$$A \equiv \textit{white males} \quad B \equiv \textit{black males},$$

and to quantify the difference between the two groups we extend (5.2) to:

$$\begin{aligned} f_A(\texttt{age}) &= \beta_0^{\mathrm{A}} + \beta_1^{\mathrm{A}}\,\texttt{age} + \sum_{\ell=1}^{L^{\mathrm{gbl}}} u_{\mathrm{A},\ell}^{\mathrm{gbl}}\, z_\ell^{\mathrm{gbl}}(\texttt{age}), \\ f_B(\texttt{age}) &= \beta_0^{\mathrm{A}} + \beta_0^{\mathrm{BvsA}} + (\beta_1^{\mathrm{A}} + \beta_1^{\mathrm{BvsA}})\,\texttt{age} + \sum_{\ell=1}^{L^{\mathrm{gbl}}} u_{\mathrm{B},\ell}^{\mathrm{gbl}}\, z_\ell^{\mathrm{gbl}}(\texttt{age}), \\ \text{and } g_i(\texttt{age}) &= \delta_{0i} + \delta_{1i}\,\texttt{age} + \sum_{\ell=1}^{L^{\mathrm{grp}}} u_{il}^{\mathrm{grp}}\, z_\ell^{\mathrm{grp}}(\texttt{age}). \end{aligned} \tag{5.14}$$

This allows one to estimate the contrast function given in (5.16). We define the parameter vectors to be

$$\boldsymbol{\beta} \equiv \begin{bmatrix} \beta_0^{\mathrm{A}} \\ \beta_1^{\mathrm{A}} \\ \beta_0^{\mathrm{BvsA}} \\ \beta_1^{\mathrm{BvsA}} \end{bmatrix}, \qquad \boldsymbol{u} \equiv \begin{bmatrix} \boldsymbol{u}_{\mathrm{A}}^{\mathrm{gbl}} \\ \boldsymbol{u}_{\mathrm{B}}^{\mathrm{gbl}} \\ \boldsymbol{u}^{\mathrm{R}} \end{bmatrix}$$

where

$$\boldsymbol{u}_{\mathrm{A}}^{\mathrm{gbl}} \equiv \begin{bmatrix} u_{\mathrm{A},1}^{\mathrm{gbl}} \\ \vdots \\ u_{\mathrm{A},L^{\mathrm{gbl}}}^{\mathrm{gbl}} \end{bmatrix}, \qquad \boldsymbol{u}_{\mathrm{B}}^{\mathrm{gbl}} \equiv \begin{bmatrix} u_{\mathrm{B},1}^{\mathrm{gbl}} \\ \vdots \\ u_{\mathrm{B},L^{\mathrm{gbl}}}^{\mathrm{gbl}} \end{bmatrix}.$$

In addition, the design matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ are

$$\boldsymbol{X} \equiv \begin{bmatrix} \mathbf{1} & \mathtt{age}_1 & \boldsymbol{I}_1^{\mathrm{A}} \odot \mathbf{1} & \boldsymbol{I}_1^{\mathrm{A}} \odot \mathtt{age}_1 \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{1} & \mathtt{age}_m & \boldsymbol{I}_m^{\mathrm{A}} \odot \mathbf{1} & \boldsymbol{I}_m^{\mathrm{A}} \odot \mathtt{age}_m \end{bmatrix}, \qquad \boldsymbol{Z} \equiv \left[\boldsymbol{Z}_{\mathrm{A}}^{\mathrm{gbl}} \mid \boldsymbol{Z}_{\mathrm{B}}^{\mathrm{gbl}} \mid \boldsymbol{Z}^{\mathrm{R}} \right],$$

where

$$\boldsymbol{Z}_{\mathrm{A}}^{\mathrm{gbl}} \equiv \begin{bmatrix} \boldsymbol{I}_1^{\mathrm{A}} \odot z_1^{\mathrm{gbl}}(\mathtt{age}_1) & \dots & \boldsymbol{I}_1^{\mathrm{A}} \odot z_{L_{\mathrm{gbl}}}^{\mathrm{gbl}}(\mathtt{age}_1) \\ \vdots & \ddots & \vdots \\ \boldsymbol{I}_m^{\mathrm{A}} \odot z_1^{\mathrm{gbl}}(\mathtt{age}_m) & \dots & \boldsymbol{I}_m^{\mathrm{A}} \odot z_{L_{\mathrm{gbl}}}^{\mathrm{gbl}}(\mathtt{age}_m) \end{bmatrix}$$

and

$$I_{ij}^{\mathrm{A}} \equiv \begin{cases} 1 & \text{if } (\mathtt{age}_{ij}, \mathtt{height}_{ij}) \text{ is of type A,} \\ 0 & \text{if } (\mathtt{age}_{ij}, \mathtt{height}_{ij}) \text{ is of type B.} \end{cases}$$

The $n_i \times 1$ vector $\boldsymbol{I}_i^{\mathrm{A}}$ consists of the I_{ij}^{A} and the design matrix $\boldsymbol{Z}_{\mathrm{B}}^{\mathrm{gbl}}$ is defined in a similar way to $\boldsymbol{Z}_{\mathrm{A}}^{\mathrm{gbl}}$ but with $\boldsymbol{I}_i^{\mathrm{A}}$ replaced by $\mathbf{1} - \boldsymbol{I}_i^{\mathrm{A}}$. The full Bayesian model is then

$$\begin{aligned} \mathtt{height} \,|\, \boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2 &\sim \mathrm{N}\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \sigma_\varepsilon^2 \boldsymbol{I}\right), \qquad \boldsymbol{\beta} \sim \mathrm{N}\left(\mathbf{0}, \sigma_\beta^2 \boldsymbol{I}_2\right), \\ \boldsymbol{u} \,|\, \left(\sigma_{\mathrm{gbl}}^{\mathrm{A}}\right)^2, \left(\sigma_{\mathrm{gbl}}^{\mathrm{B}}\right)^2, \Sigma, \sigma_{\mathrm{grp}}^2 &\sim \mathrm{N}\left(\mathbf{0}, \begin{bmatrix} \left(\sigma_{\mathrm{gbl}}^{\mathrm{A}}\right)^2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \left(\sigma_{\mathrm{gbl}}^{\mathrm{B}}\right)^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \underset{1 \le i \le m}{\mathrm{blockdiag}} \left(\boldsymbol{\Sigma}, \sigma_{\mathrm{grp}}^2 \boldsymbol{I}_{L_{\mathrm{grp}}}\right) \end{bmatrix}\right), \\ \sigma_\varepsilon &\sim \text{Half-Cauchy}\left(A_\varepsilon\right), \qquad \sigma_{\mathrm{gbl}}^{\mathrm{A}} \sim \text{Half-Cauchy}\left(A_{\mathrm{gbl}}^{\mathrm{A}}\right), \\ \sigma_{\mathrm{gbl}}^{\mathrm{B}} &\sim \text{Half-Cauchy}\left(A_{\mathrm{gbl}}^{\mathrm{B}}\right), \qquad \sigma_{\mathrm{grp}} \sim \text{Half-Cauchy}\left(A_{\mathrm{grp}}\right). \\ \boldsymbol{\Sigma} \,|\, a_{\Sigma,1}, a_{\Sigma,2} &\sim \text{Inverse-Wishart}\left(\nu + 2 - 1, 2\nu\, \mathrm{diag}\left(1/a_{\Sigma,1}, 1/a_{\Sigma,2}\right)\right), \\ a_{\Sigma,j} &\sim \text{Inverse-Gamma}\left(\tfrac{1}{2}, 1/A_{\Sigma,j}^2\right), \quad 1 \le j \le 2. \end{aligned} \tag{5.15}$$

We use an equivalent set-up and partition of the mixed model framework as shown in Sections 5.2.1 and 5.2.4. Model (5.15) was fit using the streamlined MFVB algorithm as shown in Algorithm 6, where the variables `height` and `age` were standardised to have zero mean and unit variance and the values of the hyperparameters were all set to 10^5. The `rstan` package which uses Stan code, was used for performing MCMC with a burn-in of 5000, number of further iterations of 5000 and a thinning factor of 5. The MFVB iterations were terminated when the increase in the corresponding $\log \underline{p}(\mathtt{height}; q)$ fell below 10^{-8}.

As mentioned in Chapter 2, making time comparisons fair between MFVB and MCMC is quite difficult, due to convergence for each approximation method being vastly different.

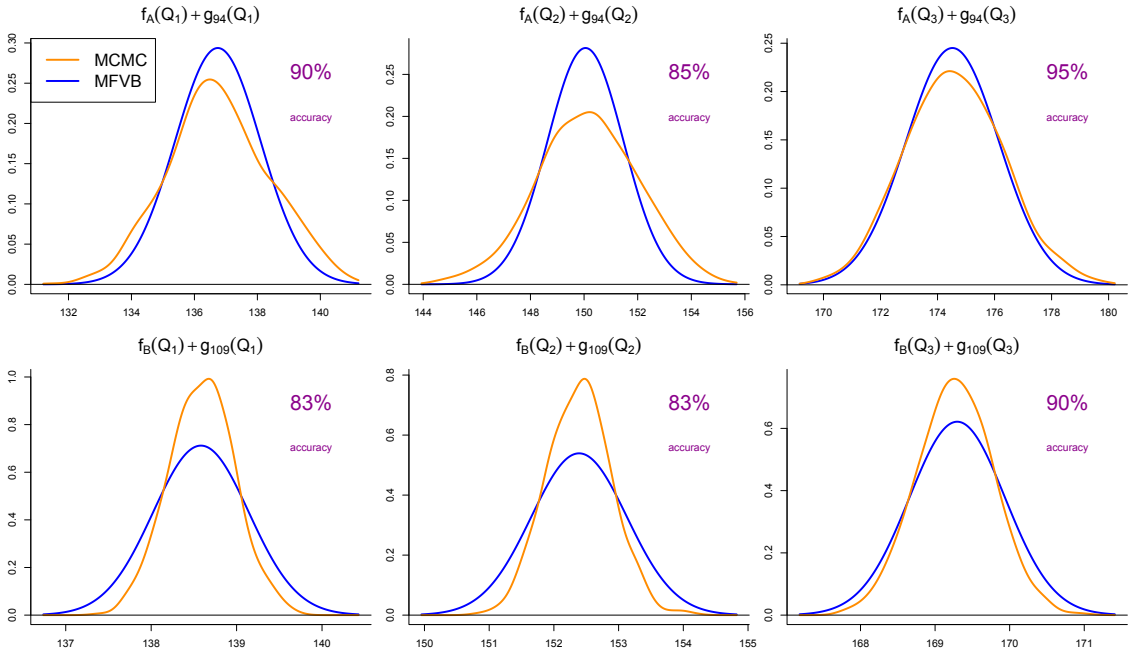


Figure 5.3: *Approximate posterior density functions obtained via MCMC and streamlined MFVB for the growth Indiana data set. MFVB accuracy scores are displayed.*

Here we make use of Stan as opposed to BUGS, which we have been using until now. Stan is known to generally converge faster than BUGS as it is based on Hamiltonian Monte Carlo sampling. Stan is also preferred for certain problems with a complex structure, such as multilevel models.

Figure 5.3 illustrates the accuracy of the streamlined MFVB approach against the MCMC benchmark. The parameters being monitored are the quartiles of the curve estimates of the 94th and 109th individual. These are represented as $f_A(Q_i) + g_{94}(Q_i)$ and $f_B(Q_i) + g_{109}(Q_i)$ for $i = 1, 2, 3$, where the 94th individual is *white* and the 109th individual is *black*. The accuracy of the curve estimates for these parameters are reasonably high considering the complexity of the model.

Figure 5.4 provides illustration of the approximate inference achieved by both streamlined MFVB and MCMC for the two-level Gaussian response model. It appears that the subject-specific curve estimates and pointwise 95% credible sets are almost indistinguishable for both approaches, implying that the streamlined MFVB approach incurs little loss of accuracy. It is immediately obvious that the streamlined MFVB fits are agreeing very well with the longitudinal data for each individual. The MCMC fit took almost 3 hours, whilst the streamlined MFVB fit took only 4 seconds.

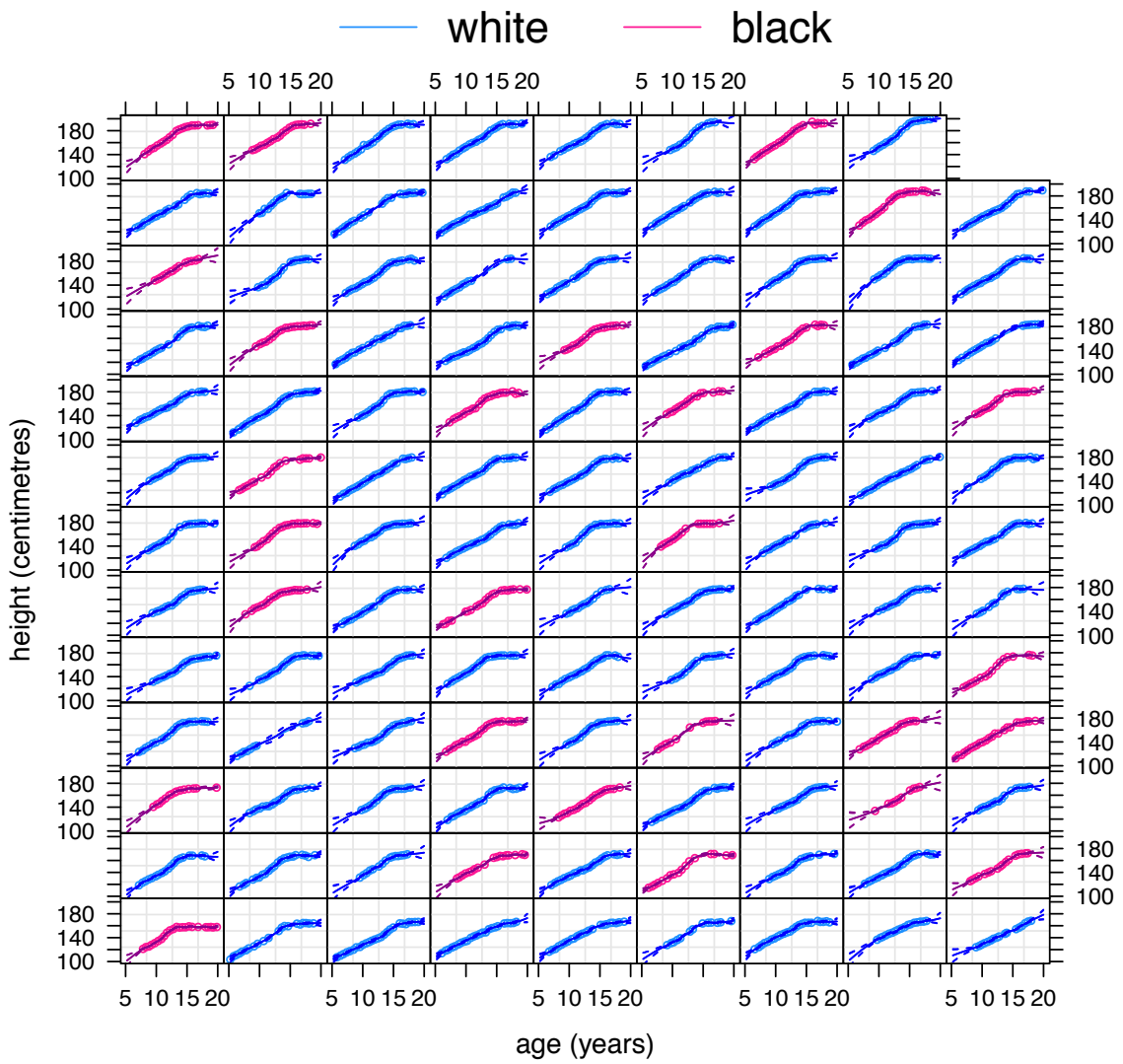


Figure 5.4: *Group-specific estimates and pointwise 95% credible sets obtained via streamlined MFVB and MCMC for the Indiana growth data-set for male subjects.*

Figure 5.5 illustrates the effect of the variable `black` on the height of male individuals. This is given by the *contrast* curve

$$\begin{aligned} c(\mathsf{age}) &\equiv f_B(\mathsf{age}) - f_A(\mathsf{age}) \\ &= \beta_0^{\mathrm{BvsA}} + \beta_1^{\mathrm{BvsA}}\mathsf{age} + \sum_{\ell=1}^{L_{\mathrm{gbl}}} \left(u_{\mathrm{B},\ell}^{\mathrm{gbl}} - u_{\ell,\mathrm{W}}^{\mathrm{gbl}}\right) z_\ell^{\mathrm{gbl}}(\mathsf{age}). \end{aligned} \tag{5.16}$$

A contrast curve was generated for both MCMC and streamlined MFVB. It is evident that at age 5, on average, *black* individuals are approximately 3.5cm taller than *white* individuals. However, as they reach adolescence, *black* individuals are, on average, taller by approximately 6cm. Finally, as males reach their early 20's, the height differences decrease to zero. In addition, the agreement of the curves produced by MCMC and

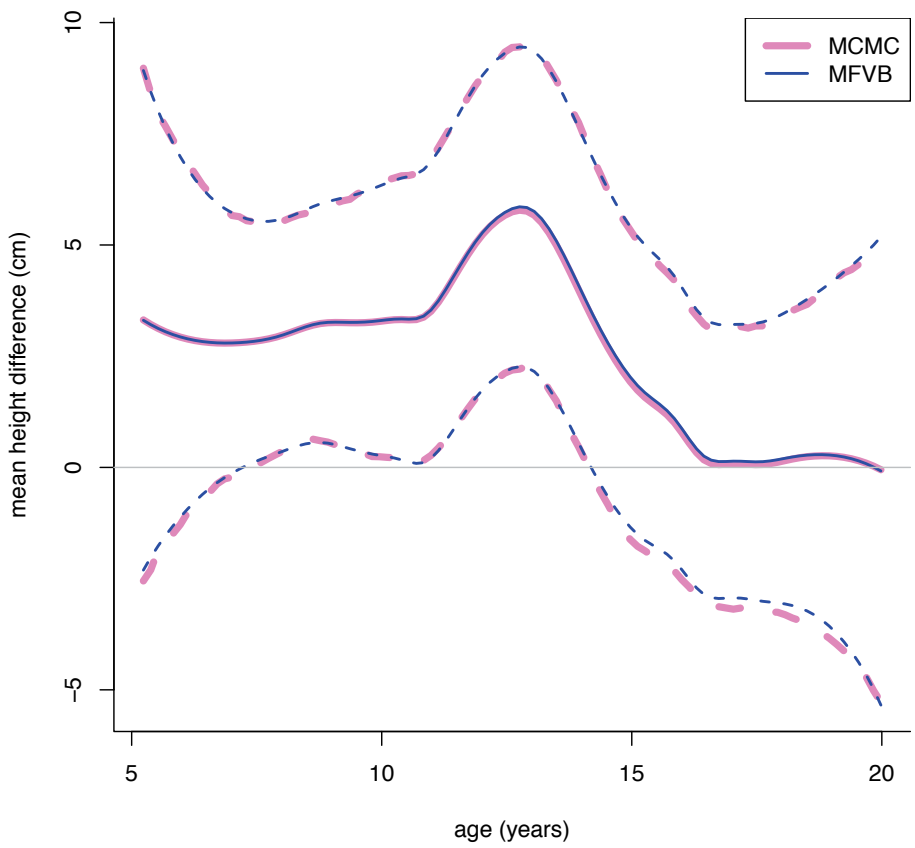


Figure 5.5: *Approximate contrast curves obtained via MCMC and streamlined MFVB for the mean height difference between black and white males.*

streamlined MFVB are quite pleasing, indicating little loss of accuracy.

5.3 Three-level Gaussian response model

In this section, we extend the methodology shown in the previous section to cater to the three-level Gaussian response model for both the naïve and streamlined MFVB situations. We apply this methodology to a data set which exhibits a three-level longitudinal structure. These data represent frequency dependent backscatter coefficients for induced tumors in rodents (source: Simpson, 2013). There are several *slices* or probe locations in which each tumor is scanned. Five different transducers were used on various slices to perform scanning. The variables and their descriptions are given in Table 5.3. Figure 5.6

Variable	Description
`freq`	frequency measurement
`bsc`	backscatter coefficient (converted to decibel scale)
`L9h4`	indicator of transducer L9-4 being used
`L40`	indicator of transducer L40 being used
`MS200`	indicator of transducer MS200 being used
`MS400`	indicator of transducer MS400 being used

Table 5.3: *Description of the rodent tumor dataset.*

illustrates the information from 10 rodents given in this dataset. Each panel represents a different tumor corresponding to a rodent. The x-axis represents the frequency and the y-axis represents the backscatter coefficients for each tumor. Each curve in each panel corresponds to a different slice. Furthermore, within each panel, the colours correspond to different transducers being used according to:

green `L14-5`
red `L9-4`
blue `L40`
orange `MS200`
purple `MS400`.

We use similar notation to that used in Section 5.2, where for example

$$\texttt{bsc}_{ijk} \equiv k\text{th backscatter coefficient for the } j\text{th slice within the } i\text{th tumor.}$$

The generic three-level Gaussian response model extension, specific to the data presented in Figure 5.6 is

$$\begin{gathered}\texttt{bsc}_{ijk} \sim N\left(f(\texttt{freq}_{ijk}) + g_i(\texttt{freq}_{ijk}) + h_{ij}(\texttt{freq}_{ijk}), \sigma_\varepsilon^2\right), \\ 1 \le i \le m, \quad 1 \le j \le n_i, \quad 1 \le k \le o_{ij},\end{gathered} \tag{5.17}$$

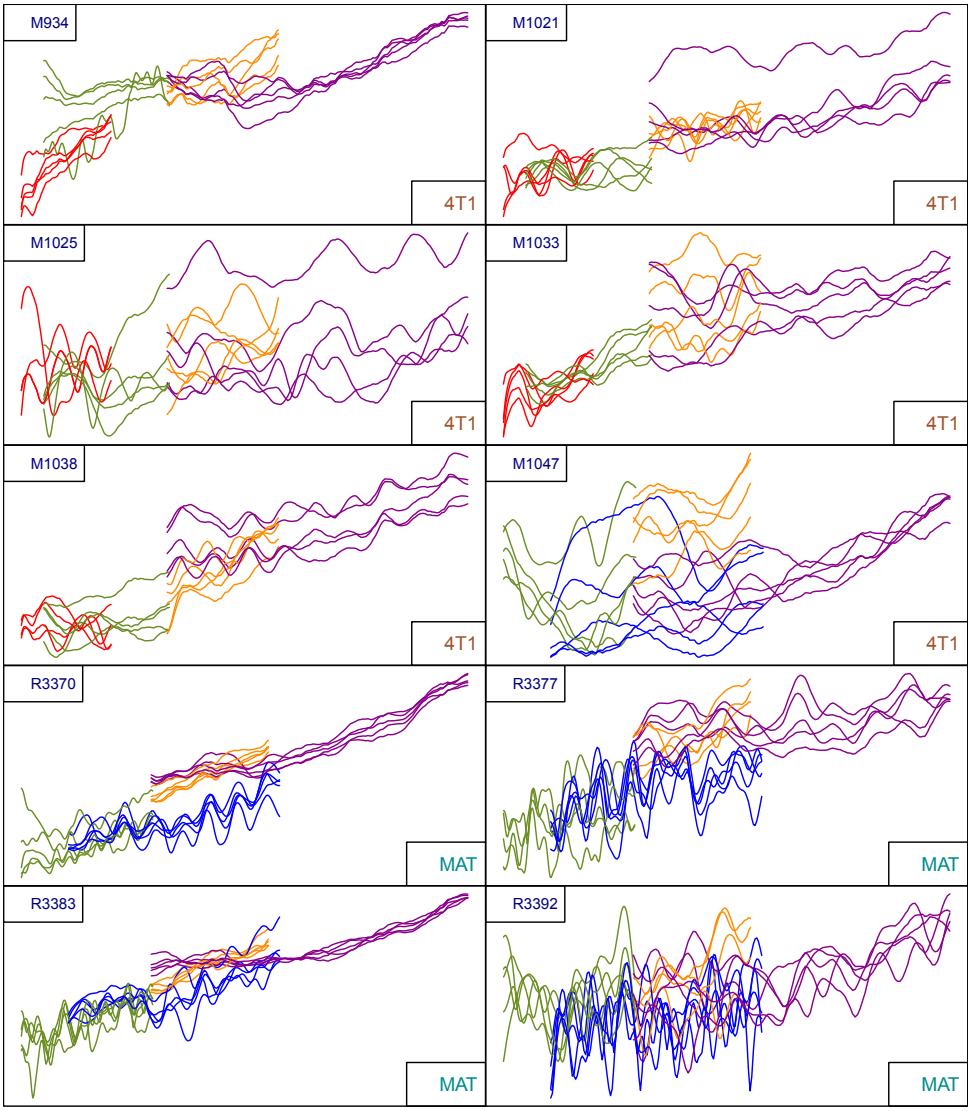


Figure 5.6: *Rodent tumor data set showing frequency dependent backscatter coefficients for two different types of induced tumors in rodents. Each panel is for a different tumor (name shown in top left corner). The type of tumor is shown in the bottom right corner of each panel.*

where we have $m = 10$ tumors, the n_i range from being either 18, 19 or 20 slices (after removing outlier slices) and the $o_{ij} = 128$ is constant throughout and we use o from now on to represent the measurements in each slice. The global mean function has the form

$$f(x) = \beta_0 + \beta_1\, x + \beta_2 \texttt{L9h4} + \beta_3 \texttt{L40} + \beta_4 \texttt{MS200} + \beta_5 \texttt{MS400} + \sum_{\ell=1}^{L_{\text{gbl}}} u_\ell^{\text{gbl}}\, z_\ell^{\text{gbl}}(x),$$

where $u_\ell^{\text{gbl}} \,|\, \sigma^2_{\text{gbl}} \overset{\text{ind.}}{\sim} N\left\{0, \sigma^2_{\text{gbl}}\right\}$. In order to model the deviation from f for the ith tumor, we further extend our model to include the function

$$g_i(x) = \delta_{0i} + \delta_{1i}x + \sum_{\ell=1}^{L_{\text{grpout}}} u_\ell^{\text{grpout}}\, z_\ell^{\text{grpout}}(x), \quad \begin{bmatrix} \delta_{0i} \\ \delta_{1i} \end{bmatrix} \overset{\text{ind.}}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma}_{\text{grpout}}),$$
$$u_\ell^{\text{grpout}} \,|\, \sigma^2_{\text{grpout}} \overset{\text{ind.}}{\sim} N(0, \sigma^2_{\text{grpout}}),$$

where 'grpout' refers to each rodent's information. In addition, to model the deviation from g_i for the ijth slice, we also make use of the function

$$h_{ij}(x) = \gamma_{0ij} + \gamma_{1ij}x + \sum_{\ell=1}^{L_{\text{grpinn}}} u_\ell^{\text{grpinn}}\, z_\ell^{\text{grpinn}}(x), \quad \begin{bmatrix} \gamma_{0ij} \\ \gamma_{1ij} \end{bmatrix} \overset{\text{ind.}}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma}_{\text{grpinn}}),$$
$$u_\ell^{\text{grpinn}} \,|\, \sigma^2_{\text{grpinn}} \overset{\text{ind.}}{\sim} N(0, \sigma^2_{\text{grpinn}}),$$

where 'grpinn' refers to the information involving the measurements for each slice. In order to express our model in a mixed model framework, we first define the following vectors and matrices:

$$\boldsymbol{\beta} \equiv \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix}, \quad \boldsymbol{u}^{\text{gbl}} \equiv \begin{bmatrix} u_1^{\text{gbl}} \\ \vdots \\ u_{\text{L}^{\text{gbl}}}^{\text{gbl}} \end{bmatrix}, \quad \boldsymbol{u}_{\text{full}}^{\text{grpout}} \equiv \begin{bmatrix} \boldsymbol{\delta}_1 \\ \boldsymbol{u}_1^{\text{grpout}} \\ \vdots \\ \boldsymbol{\delta}_m \\ \boldsymbol{u}_m^{\text{grpout}} \end{bmatrix}, \quad \boldsymbol{u}_{\text{full}}^{\text{grpinn}} \equiv \begin{bmatrix} \boldsymbol{\gamma}_{11} \\ \boldsymbol{u}_{11}^{\text{grpinn}} \\ \vdots \\ \boldsymbol{\gamma}_{1n_1} \\ \boldsymbol{u}_{1n_1}^{\text{grpinn}} \\ \vdots \\ \boldsymbol{\gamma}_{m1} \\ \boldsymbol{u}_{m1}^{\text{grpinn}} \\ \vdots \\ \boldsymbol{\gamma}_{mn_m} \\ \boldsymbol{u}_{mn_m}^{\text{grpinn}} \end{bmatrix},$$

$$\boldsymbol{u} \equiv \begin{bmatrix} \boldsymbol{u}^{\text{gbl}} \\ \boldsymbol{u}_{\text{full}}^{\text{grpout}} \\ \boldsymbol{u}_{\text{full}}^{\text{grpinn}} \end{bmatrix},$$

where the $\boldsymbol{\delta}_i$ and $\boldsymbol{u}_i^{\text{grpout}}$ are defined as in section 5.2.1. We define $\boldsymbol{\gamma}_{ij} \equiv \left[\gamma_{0ij} \;\; \gamma_{1ij}\right]^\top$ and $\boldsymbol{u}_{ij}^{\text{grpinn}} \equiv \left[u_1^{\text{grpinn}}, \ldots, u_{\text{L}^{\text{grpinn}}}^{\text{grpinn}}\right]^\top$. The design matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ are

$$\boldsymbol{X} \equiv \begin{bmatrix} \mathbf{1} & \texttt{freq}_1 & \texttt{L9h4}_1 & \texttt{L40}_1 & \texttt{MS200}_1 & \texttt{MS400}_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{1} & \texttt{freq}_m & \texttt{L9h4}_m & \texttt{L40}_m & \texttt{MS200}_m & \texttt{MS400}_m \end{bmatrix},$$

$$\boldsymbol{Z} \equiv \left[\boldsymbol{Z}^{\text{gbl}} \;\middle|\; \underset{1\leq i\leq m}{\text{blockdiag}} \left(\boldsymbol{X}_i \,|\, \boldsymbol{Z}_i^{\text{grpout}} \right) \;\middle|\; \underset{1\leq i\leq m}{\text{blockdiag}} \left(\underset{1\leq j\leq n_i}{\text{blockdiag}} \left(\boldsymbol{X}_{ij} \,|\, \boldsymbol{Z}_{ij}^{\text{grpinn}} \right) \right) \right],$$

where

$$\boldsymbol{Z}^{\text{gbl}} \equiv \begin{bmatrix} z_1^{\text{gbl}}(\texttt{freq}_1) & \dots & z_{\text{L}^{\text{gbl}}}^{\text{gbl}}(\texttt{freq}_1) \\ \vdots & \ddots & \vdots \\ z_1^{\text{gbl}}(\texttt{freq}_m) & \dots & z_{\text{L}^{\text{gbl}}}^{\text{gbl}}(\texttt{freq}_m) \end{bmatrix},$$

$$\boldsymbol{Z}_i^{\text{grpout}} \equiv \begin{bmatrix} z_1^{\text{grpout}}(\texttt{freq}_{i1}) & \dots & z_{L_{\text{grpout}}}^{\text{grpout}}(\texttt{freq}_{i1}) \\ \vdots & \ddots & \vdots \\ z_1^{\text{grpout}}(\texttt{freq}_{in_i}) & \dots & z_{L_{\text{grpout}}}^{\text{grpout}}(\texttt{freq}_{in_i}) \end{bmatrix},$$

$$\boldsymbol{Z}_{ij}^{\text{grpinn}} \equiv \begin{bmatrix} z_1^{\text{grpinn}}(\texttt{freq}_{ij1}) & \dots & z_{\text{L}^{\text{grpinn}}}^{\text{grpinn}}(\texttt{freq}_{ij1}) \\ \vdots & \ddots & \vdots \\ z_1^{\text{grpinn}}(\texttt{freq}_{ijo}) & \dots & z_{\text{L}^{\text{grpinn}}}^{\text{grpinn}}(\texttt{freq}_{ijo}) \end{bmatrix},$$

with $\texttt{freq}_i$ being the $n_i\, o \times 1$ vector containing the $\texttt{freq}_{ijk}$ and $\boldsymbol{X}_i$ is the subset of $\boldsymbol{X}$ corresponding to the ith tumor. We can now represent our model using the mixed model framework:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon}, \tag{5.18}$$

$\boldsymbol{G} \equiv \text{Cov}(\boldsymbol{u}) =$

$$\begin{bmatrix} \sigma_{\text{gbl}}^2 \boldsymbol{I}_{L_{\text{gbl}}} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \underset{1\leq i\leq m}{\text{blockdiag}} \left(\boldsymbol{\Sigma}_{\text{grpout}}\,,\, \sigma_{\text{grpout}}^2 \boldsymbol{I}_{L_{\text{grpout}}} \right) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \underset{1\leq j\leq \sum_{i=1}^m n_i}{\text{blockdiag}} \left(\boldsymbol{\Sigma}_{\text{grpinn}}\,,\, \sigma_{\text{grpinn}}^2 \boldsymbol{I}_{L_{\text{grpinn}}} \right) \end{bmatrix}$$

5.3.1 Bayesian three-level Gaussian response model

Much like in the two-level case, fitting (5.18) using standard mixed model software such as `lme()` in R is possible, but extremely time intensive. We opt to work with a Bayesian version of (5.18) which allows direct implementation in standard Bayesian software. We enforce σ_ε, σ_{gbl}, $\boldsymbol{\Sigma}_{\text{grpout}}$, σ_{grpout}, $\boldsymbol{\Sigma}_{\text{grpinn}}$ and σ_{grpinn} to possess half-t prior distributions on $(-1, 1)$, as suggested by Huang *et al.* (2013). The standard deviation parameter priors are achieved through auxiliary variable constructions as exhibited in Result 1.4.6. The full

Bayesian three-level Gaussian response model for the rodent tumor data is:

$$
\begin{aligned}
&\boldsymbol{y}\,|\,\boldsymbol{\beta},\,\boldsymbol{u},\,\sigma_\varepsilon^2 \sim \mathrm{N}\left(\boldsymbol{X}\,\boldsymbol{\beta}+\boldsymbol{Z}\,\boldsymbol{u},\,\sigma_\varepsilon^2\boldsymbol{I}\right), \quad \boldsymbol{\beta} \sim \mathrm{N}\left(\boldsymbol{0},\sigma_\beta^2\boldsymbol{I}_6\right), \quad \boldsymbol{u}|\boldsymbol{G} \sim \mathrm{N}\left(\boldsymbol{0},\boldsymbol{G}\right),\\
&\sigma_\varepsilon^2\,|\,a_\varepsilon \sim \text{Inverse-Gamma}\left(\tfrac{1}{2},\,1/a_\varepsilon\right), \quad a_\varepsilon \sim \text{Inverse-Gamma}\left(\tfrac{1}{2},\,1/A_\varepsilon^2\right),\\
&\sigma_{\text{gbl}}^2\,|\,a_{\text{gbl}} \sim \text{Inverse-Gamma}\left(\tfrac{1}{2},\,1/a_{\text{gbl}}\right), \quad a_{\text{gbl}} \sim \text{Inverse-Gamma}\left(\tfrac{1}{2},\,A_{\text{gbl}}^{-2}\right),\\
&\boldsymbol{\Sigma}_{\text{grpout}}\,|\,a_{\Sigma_{\text{grpout}},1},\,a_{\Sigma_{\text{grpout}},2} \sim \text{Inverse-Wishart}\left(\nu_{\text{grpout}}+2-1,\right.\\
&\qquad\qquad\left.2\,\nu_{\text{grp}}\,\text{diag}\left(1/a_{\Sigma_{\text{grpout}},1},\,1/a_{\Sigma_{\text{grpout}},2}\right)\right),\\
&a_{\Sigma_{\text{grpout}},j} \sim \text{Inverse-Gamma}\left(\tfrac{1}{2},\,1/A_{\Sigma_{\text{grpout}},j}^2\right), \quad 1 \le j \le 2,\\
&\sigma_{\text{grpout}}^2\,|\,a_{\text{grpout}} \sim \text{Inverse-Gamma}\left(\tfrac{1}{2},\,1/a_{\text{grpout}}\right),\\
&a_{\text{grpout}} \sim \text{Inverse-Gamma}\left(\tfrac{1}{2},\,1/A_{\text{grpout}}^2\right),\\
&\boldsymbol{\Sigma}_{\text{grpinn}}\,|\,a_{\Sigma_{\text{grpinn}},1},\,a_{\Sigma_{\text{grpinn}},2} \sim \text{Inverse-Wishart}\left(\nu_{\text{grpinn}}+2-1,\right.\\
&\qquad\qquad\left.2\,\nu_{\text{grpinn}}\,\text{diag}\left(1/a_{\Sigma_{\text{grpinn}},1},\,1/a_{\Sigma_{\text{grpinn}},2}\right)\right),\\
&a_{\Sigma_{\text{grpinn}},j} \sim \text{Inverse-Gamma}\left(\tfrac{1}{2},\,1/A_{\Sigma_{\text{grpinn}},j}^2\right), \quad 1 \le j \le 2,\\
&\sigma_{\text{grpinn}}^2\,|\,a_{\text{grpinn}} \sim \text{Inverse-Gamma}\left(\tfrac{1}{2},\,1/a_{\text{grpinn}}\right),\\
&a_{\text{grpinn}} \sim \text{Inverse-Gamma}\left(\tfrac{1}{2},\,1/A_{\text{grpinn}}^2\right),
\end{aligned}
\tag{5.19}
$$

where a_ε, a_{gbl}, $\boldsymbol{a}_{\Sigma_{\text{grpout}}}$, a_{grpout}, $\boldsymbol{a}_{\Sigma_{\text{grpinn}}}$ and a_{grpinn} are the auxiliary variables used for the scale parameters in the model. Fitting (5.19) involves the setting of hyperparameters σ_β, A_ε, A_{gbl}, ν_{grpout}, $A_{\Sigma_{\text{grpout}},1}$, $A_{\Sigma_{\text{grpout}},2}$, A_{grpout}, ν_{grpinn}, $A_{\Sigma_{\text{grpinn}},1}$, $A_{\Sigma_{\text{grpinn}},2}$ and A_{grpinn}. To ensure that Property 4 of Huang *et al.* (2013) holds, we set $\nu_{\text{grpout}} = \nu_{\text{grpinn}} = 2$. For the remaining hyperparameters, non-informativity is achieved when their values are large, e.g., 10^5.

5.3.2 Mean field variational Bayes methodology

The essence of the MFVB approach for (5.19) is to approximate the joint posterior density function with an approximate product density form

$$
\begin{aligned}
&q(\boldsymbol{\beta},\,\boldsymbol{u},\,a_\varepsilon,\,a_{\text{gbl}},\,\boldsymbol{a}_{\Sigma_{\text{grpout}}},\,a_{\text{grpout}},\,\boldsymbol{a}_{\Sigma_{\text{grpinn}}},\,a_{\text{grpinn}})\\
&\qquad\times q\left\{\sigma_\varepsilon^2,\,\sigma_{\text{gbl}}^2,\,\boldsymbol{\Sigma}_{\text{grpout}},\,\sigma_{\text{grpout}}^2,\,\boldsymbol{\Sigma}_{\text{grpinn}},\,\sigma_{\text{grpinn}}^2\right\}.
\end{aligned}
$$

Using induced factorisation theory, it is shown that a further factorisation is of the form:

$$
\begin{aligned}
&q(\boldsymbol{\beta},\boldsymbol{u})\,q(\sigma_\varepsilon^2)\,q\left(\sigma_{\text{gbl}}^2\right)\,q(\boldsymbol{\Sigma}_{\text{grpout}})\,q(\sigma_{\text{grpout}}^2)\,q(a_\varepsilon)\,q(a_{\text{gbl}})\\
&\quad\times\left\{\prod_{j=1}^{2}q(a_{\Sigma_{\text{grpout}},j})\right\}q(a_{\text{grpout}})\,\left\{\prod_{j=1}^{2}q(a_{\Sigma_{\text{grpinn}},j})\right\}\,q(a_{\text{grpinn}}).
\end{aligned}
\tag{5.20}
$$

As mentioned in section 5.2.3 the optimal q-densities can be obtained through expressions such as that given in (5.11). The full conditional distributions that we deal with all

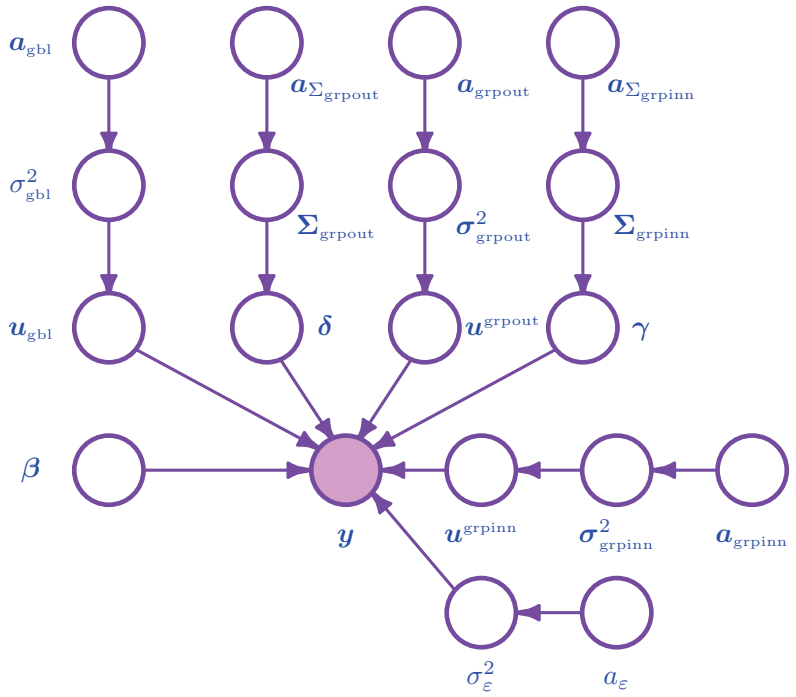


Figure 5.7: *DAG corresponding to the Bayesian hierarchical three-level Gaussian response model in (5.19).*

have closed form expressions, resulting in our optimal q-densities possessing closed form expressions. In light of this, the factors given in (5.20) have the following forms:

$$
\begin{aligned}
q^*(\boldsymbol{\beta}, \boldsymbol{u}) &\sim \mathrm{N}(\boldsymbol{\mu}_{q(\boldsymbol{\beta},u)}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},u)}) \\
q^*(\sigma^2_{\varepsilon}) &\sim \text{Inverse-Gamma}\left(\tfrac{1}{2}\left(\sum_{i=1}^{m} n_i + 1\right), B_{q(\sigma^2_{\varepsilon})}\right) \\
q^*\left(\sigma^2_{\text{gbl}}\right) &\sim \text{Inverse-Gamma}\left(\tfrac{1}{2}(L_{\text{gbl}} + 1), B_{q\left(\sigma^2_{\text{gbl}}\right)}\right) \\
q^*(\boldsymbol{\Sigma}_{\text{grpout}}) &\sim \text{Inverse-Wishart}\left(\nu_{\text{grpout}} + m + 1, B_{q(\boldsymbol{\Sigma}_{\text{grpout}})}\right) \\
q^*(\sigma^2_{\text{grpout}}) &\sim \text{Inverse-Gamma}\left(\tfrac{1}{2}(m\, L_{\text{grpout}} + 1), B_{q(\sigma^2_{\text{grpout}})}\right) \\
q^*(\boldsymbol{\Sigma}_{\text{grpinn}}) &\sim \text{Inverse-Wishart}(\nu_{\text{grpinn}} + \sum_{i=1}^{n} n_i + 1, B_{q(\boldsymbol{\Sigma}_{\text{grpinn}})}), \\
q^*(\sigma^2_{\text{grpinn}}) &\sim \text{Inverse-Gamma}\left\{\tfrac{1}{2}\left(L_{\text{grpinn}} \sum_{i=1}^{m} n_i + 1\right), B_{q(\sigma^2_{\text{grpinn}})}\right\}, \\
q^*(a_{\varepsilon}) \sim \text{Inverse-Gamma}(1, B_{q(a_{\varepsilon})}), &\quad q^*(a_{\text{gbl}}) \sim \text{Inverse-Gamma}(1, B_{q(a_{\text{gbl}})}) \\
q^*(a_{\text{grpout}}) &\sim \text{Inverse-Gamma}(1, B_{q(a_{\text{grpout}})}), \\
q^*(a_{\text{grpinn}}) &\sim \text{Inverse-Gamma}(1, B_{q(a_{\text{grpinn}})}), \\
q^*(a_{\Sigma_{\text{grpout}},j}) &\sim \text{Inverse-Gamma}(\tfrac{\nu_{\text{grpout}}}{2} + 1, B_{q(a_{\Sigma_{\text{grpout}},j})}) \quad 1 \le j \le 2, \\
q^*(a_{\Sigma_{\text{grpinn}},j}) &\sim \text{Inverse-Gamma}(\tfrac{\nu_{\text{grpinn}}}{2} + 1, B_{q(a_{\Sigma_{\text{grpinn}},j})}) \quad 1 \le j \le 2.
\end{aligned}
\tag{5.21}
$$

The derivation of (5.21) is similar to that shown for the two-level model in Appendix 5.A and the parameters in (5.21) are defined analogously to that described in Section 5.2.3. The values of these optimal parameters are obtained through an iterative scheme presented as Algorithm 7, where

$$\boldsymbol{M}_{q(\boldsymbol{G}^{-1})} = \begin{bmatrix} \mu_{q(1/{\sigma_{\text{gbl}}}^2)}\boldsymbol{I}_{L_{\text{gbl}}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \underset{1\le i\le m}{\text{blockdiag}}\left(\boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1}_{\text{grpout}})}, \mu_{q(1/\sigma^2_{\text{grpout}})}\boldsymbol{I}_{L_{\text{grpout}}}\right) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \underset{1\le j\le\sum_{i=1}^m n_i}{\text{blockdiag}}\left(\boldsymbol{M}_{q\left(\boldsymbol{\Sigma}^{-1}_{\text{grpinn}}\right)}, \mu_{q(1/\sigma^2_{\text{grpinn}})}\boldsymbol{I}_{L_{\text{grpinn}}}\right) \end{bmatrix}.$$

We define

$$\boldsymbol{\mu}_{q(\boldsymbol{u}^{\text{gbl}})} \equiv \text{sub-vector of } \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \text{ corresponding to } \boldsymbol{u}^{\text{gbl}}$$

and

$$\boldsymbol{\Sigma}_{q(\boldsymbol{u}^{\text{gbl}})} \equiv \text{sub-matrix of } \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \text{ corresponding to } \boldsymbol{u}^{\text{gbl}}.$$

The vectors $\boldsymbol{\mu}_{q(\boldsymbol{u}^{\text{grpout}})}$, $\boldsymbol{\mu}_{q(\boldsymbol{\delta}_i)}$, $\boldsymbol{\mu}_{q(\boldsymbol{u}^{\text{grpinn}})}$, $\boldsymbol{\mu}_{q(\boldsymbol{\gamma}_{ij})}$ and their corresponding variance covariance matrices are defined analogously. Algorithm 7 makes use of the variational lower bound on the marginal log-likelihood. For model (5.19) and product restriction (5.20) it

Set up initial values:

$\mu_{q(1/\sigma^2_\varepsilon)}$, $\mu_{q\left(1/\sigma^2_{\text{gbl}}\right)}$, $\mu_{q(1/\sigma^2_{\text{grpout}})}$, $\mu_{q(1/\sigma^2_{\text{grpinn}})}$, $\mu_{q(1a_\varepsilon)}$, $\mu_{q(1/a_{\text{gbl}})}$, $\mu_{q(1/a_{\text{grpout}})}$,

$\mu_{q(1/a_{\text{grpinn}})} > 0$, $\mu_{q(1/a_{\Sigma_{\text{grpout}},j})}$, $\mu_{q(1/a_{\Sigma_{\text{grpinn}},j})} > 0$, $\quad 1 \le j \le 2$, $\quad \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$,

$\boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1}_{\text{grpout}})}$, $\boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1}_{\text{grpinn}})}$ positive definite.

Cycle through:

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

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \left(\mu_{q(1/\sigma^2_\varepsilon)} \boldsymbol{C}^\top \boldsymbol{C} + \begin{bmatrix} \sigma^{-2}_{\boldsymbol{\beta}} \boldsymbol{I}_p & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{M}_{q(\boldsymbol{G}^{-1})} \end{bmatrix} \right)^{-1}$$

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

$$\mu_{q(1/\sigma^2_\varepsilon)} \leftarrow \tfrac{1}{2} \left(o \sum_{i=1}^{m} n_i + 1 \right) \Big/ B_{q(\sigma^2_\varepsilon)}; \qquad B_{q(a_\varepsilon)} \leftarrow \mu_{q(1/\sigma^2_\varepsilon)} + A^{-2}_\varepsilon; \qquad \mu_{q(1/a_\varepsilon)} \leftarrow 1/B_{q(a_\varepsilon)}$$

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

$$\mu_{q\left(1/\sigma^2_{\text{gbl}}\right)} \leftarrow \tfrac{1}{2} \left(L_{\text{gbl}} + 1 \right) \Big/ B_{q\left(\sigma^2_{\text{gbl}}\right)}$$

$$B_{q(a_{\text{gbl}})} \leftarrow \mu_{q\left\{1/\sigma^2_{\text{gbl}}\right\}} + \left(A_{\text{gbl}} \right)^{-2}; \qquad \mu_{q(1/a_{\text{gbl}})} \leftarrow 1/B_{q(a_{\text{gbl}})}$$

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

$$\mu_{q\left(1/\sigma^2_{\text{grpout}}\right)} \leftarrow \tfrac{1}{2} \left(m \, L_{\text{grpout}} + 1 \right) / B_{q\left(\sigma^2_{\text{grpout}}\right)}$$

$$B_{q(a_{\text{grpout}})} \leftarrow \mu_{q\left(1/\sigma^2_{\text{grpout}}\right)} + A^{-2}_{\text{grpout}}, \qquad \mu_{q(1/a_{\text{grpout}})} \leftarrow 1/B_{q(a_{\text{grpout}})}$$

$$\boldsymbol{B}_{q(\boldsymbol{\Sigma}_{\text{grpout}})} \leftarrow \sum_{i=1}^{m} \left\{ \boldsymbol{\mu}_{q(\boldsymbol{\delta}_i)} \boldsymbol{\mu}^\top_{q(\boldsymbol{\delta}_i)} + \boldsymbol{\Sigma}_{q(\boldsymbol{\delta}_i)} \right\} + 2\nu_{\text{grpout}} \text{diag} \left(\mu_{q\left(1/a_{\Sigma_{\text{grpout}},1}\right)}, \mu_{q\left(1/a_{\Sigma_{\text{grpout}},2}\right)} \right)$$

$$\boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1}_{\text{grpout}})} \leftarrow \left(\nu_{\text{grpout}} + m + 1 \right) \boldsymbol{B}^{-1}_{q(\boldsymbol{\Sigma}_{\text{grpout}})}$$

For $j = 1, 2$: $\quad B_{q\left(a_{\Sigma_{\text{grpout}},j}\right)} \leftarrow \nu_{\text{grpout}} \left(\boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1}_{\text{grpout}})} \right)_{jj} + 1/A^2_{\Sigma_{\text{grpout}},j}$

$$\mu_{q\left(1/a_{\Sigma_{\text{grpout}},j}\right)} \leftarrow \left(\frac{\nu_{\text{grpout}}}{2} + 1 \right) \Big/ B_{q\left(a_{\Sigma_{\text{grpout}},j}\right)}$$

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

$$\mu_{q\left(1/\sigma^2_{\text{grpinn}}\right)} \leftarrow \tfrac{1}{2} \left(o \, m \, L_{\text{grpinn}} + 1 \right) / B_{q\left(\sigma^2_{\text{grpinn}}\right)}$$

$$B_{q(a_{\text{grpinn}})} \leftarrow \mu_{q\left(1/\sigma^2_{\text{grpinn}}\right)} + A^{-2}_{\text{grpinn}}; \qquad \mu_{q(1/a_{\text{grpinn}})} \leftarrow 1/B_{q(a_{\text{grpinn}})}$$

$$\boldsymbol{B}_{q(\boldsymbol{\Sigma}_{\text{grpinn}})} \leftarrow \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left\{\boldsymbol{\mu}_{q(\boldsymbol{\gamma}_{ij})}\boldsymbol{\mu}_{q(\boldsymbol{\gamma}_{ij})}^{\top} + \boldsymbol{\Sigma}_{q(\boldsymbol{\gamma}_{ij})}\right\} + 2\nu_{\text{grpinn}}\text{diag}\left(\mu_{q\left(1/a_{\Sigma_{\text{grpinn},1}}\right)}, \mu_{q\left(1/a_{\Sigma_{\text{grpinn},2}}\right)}\right)$$

$$\boldsymbol{M}_{q\left(\boldsymbol{\Sigma}_{\text{grpinn}}^{-1}\right)} \leftarrow \left(\nu_{\text{grpinn}} + \sum_{i=1}^{m} n_i + 1\right)\boldsymbol{B}_{q(\boldsymbol{\Sigma}_{\text{grpinn}})}^{-1}$$

$$\text{For } j = 1, 2: \quad B_{q\left(a_{\Sigma_{\text{grpinn},j}}\right)} \leftarrow \nu_{\text{grpinn}}\left(\boldsymbol{M}_{q\left(\boldsymbol{\Sigma}_{\text{grpinn}}^{-1}\right)}\right)_{jj} + 1/A_{\Sigma_{\text{grpinn},j}}^2$$

$$\mu_{q\left(1/a_{\Sigma_{\text{grpinn},j}}\right)} \leftarrow \left(\frac{\nu_{\text{grpinn}}}{2} + 1\right)\Big/ B_{q\left(a_{\Sigma_{\text{grpinn},j}}\right)}$$

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

Algorithm 7: *Naïve MFVB algorithm for the estimation of the optimal parameters in (5.21).*

is shown that $\log\left\{\underline{p}(\boldsymbol{y}; q)\right\}$ has the following expression:

$$\begin{aligned}
&\log \underline{p}(\boldsymbol{y}; q) = \\
&\tfrac{1}{2}(\nu_{\text{grpout}} + q_{\text{grpout}} - 1)\log(2\nu_{\text{grpout}}) + \tfrac{1}{2}(\nu_{\text{grpinn}} + q_{\text{grpinn}} - 1)\log(2\nu_{\text{grpinn}}) \\
&-\tfrac{1}{2}o\sum_{i=1}^{m} n_i \log(2\pi) - 5\log(\pi) - \tfrac{p}{2}\log(\sigma_\beta^2) - \tfrac{1}{\sigma_\beta^2}\left\{\| \boldsymbol{\mu}_{q(\boldsymbol{\beta})} \|^2 + \text{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right)\right\} \\
&+\tfrac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}| + \log\Gamma\left\{\tfrac{1}{2}(L_{\text{gbl}} + 1)\right\} - \log\left(C_{q_{\text{grpout}},\nu_{\text{grpout}}+q_{\text{grpout}}-1}\right) \\
&+\tfrac{1}{2}\left\{p + 2L_{\text{grpout}} + m(q_{\text{grpout}} + L_{\text{grpout}}) + \sum_{i=1}^{m} n_i\,(q_{\text{grpinn}} + L_{\text{grpinn}})\right\} - \log\left(B_{q(a_{\text{gbl}})}\right) \\
&+\log\left(C_{q_{\text{grpout}},\nu_{\text{grpout}}+m+q_{\text{grpout}}-1}\right) - \log\left(C_{q_{\text{grpinn}},\nu_{\text{grpinn}}+q_{\text{grpinn}}-1}\right) \\
&+\log\left(C_{q_{\text{grpinn}},\nu_{\text{grpinn}}+\sum_{i=1}^{m} n_i+q_{\text{grpinn}}-1}\right) + \log\Gamma\left\{\tfrac{1}{2}\left(o\sum_{i=1}^{m} n_i + 1\right)\right\} - \log(A_\varepsilon) \\
&-\tfrac{1}{2}(\nu_{\text{grpout}} + q_{\text{grpout}} + m - 1)\log|\boldsymbol{B}_{q(\boldsymbol{\Sigma}_{\text{grpout}})}| - \log\left(B_{q(a_\varepsilon)}\right) - \log(A_{\text{gbl}}) \\
&-\tfrac{1}{2}\left(\nu_{\text{grpinn}} + q_{\text{grpinn}} + \sum_{i=1}^{m} n_i - 1\right)\log|\boldsymbol{B}_{q(\boldsymbol{\Sigma}_{\text{grpinn}})}| - \tfrac{1}{2}\left(o\sum_{i=1}^{m} n_i + 1\right)\log\left(B_{q(\sigma_\varepsilon^2)}\right) \\
&+\log\Gamma\left\{\tfrac{1}{2}\left(L_{\text{grpinn}}\sum_{i=1}^{m} n_i + 1\right)\right\} - \tfrac{1}{2}(L_{\text{gbl}} + 1)\log\left(B_{q\left(\sigma_{\text{gbl}}^2\right)}\right) - \log(A_{\text{grpinn}}) \\
&-\tfrac{1}{2}(mL_{\text{grpout}} + 1)\log\left(B_{q\left(\sigma_{\text{grpout}}^2\right)}\right) - \tfrac{1}{2}\left(L_{\text{grpinn}}\sum_{i=1}^{m} n_i + 1\right)\log\left(B_{q\left(\sigma_{\text{grpinn}}^2\right)}\right) \\
&+\mu_{q(1/\sigma_\varepsilon^2)}\mu_{q(1/a_\varepsilon)} + \mu_{q(1/\sigma_{\text{gbl}}^2)}\mu_{q(1/a_{\text{gbl}})} + \log\Gamma\left\{\tfrac{1}{2}(mL_{\text{grpout}} + 1)\right\} \\
&-\log\left(B_{q(a_{\text{grpout}})}\right) + \mu_{q(1/\sigma_{\text{grpout}}^2)}\mu_{q(1/a_{\text{grpout}})} - \sum_{j=1}^{q_{\text{grpinn}}}\log\left(A_{\Sigma_{\text{grpinn},j}}\right) \\
&-\log(A_{\text{grpout}}) - \sum_{j=1}^{q_{\text{grpout}}}\log\left(A_{\Sigma_{\text{grpout},j}}\right) + q_{\text{grpout}}\log\Gamma\left\{\tfrac{1}{2}(\nu_{\text{grpout}} + q_{\text{grpout}})\right\} \\
&-\tfrac{1}{2}(\nu_{\text{grpout}} + q_{\text{grpout}} - 1)\sum_{j=1}^{q_{\text{grpout}}}\log\left(B_{q(a_{\Sigma_{\text{grpout},j}})}\right) + q_{\text{grpinn}}\log\Gamma\left\{\tfrac{1}{2}(\nu_{\text{grpinn}} + q_{\text{grpinn}})\right\}
\end{aligned} \tag{5.22}$$

$$
\begin{aligned}
&+\textstyle\sum_{j=1}^{q_{\text{grpout}}} \nu_{\text{grpout}} \left(\boldsymbol{M}_{q(\boldsymbol{\Sigma}_{\text{grpout}}^{-1})}\right)_{jj} \mu_{q(1/a_{\Sigma_S,j})} - \log\left(B_{q(a_{\text{grpinn}})}\right) \\
&-\tfrac{1}{2}\left(\nu_{\text{grpinn}} + q_{\text{grpinn}} - 1\right) \sum_{j=1}^{q_{\text{grpinn}}} \log\left(B_{q(a_{\Sigma_{\text{grpinn}},j})}\right) + \mu_{q(1/\sigma^2_{\text{grpinn}})}\mu_{q(1/a_{\text{grpinn}})} \\
&+\textstyle\sum_{j=1}^{q_{\text{grpinn}}} \nu_{\text{grpinn}} \left(\boldsymbol{M}_{q(\boldsymbol{\Sigma}_{\text{grpinn}}^{-1})}\right)_{jj} \mu_{q(1/a_{\Sigma_{\text{grpinn}},j})}.
\end{aligned}
$$

where $p = 6$ is the number of columns in the $\boldsymbol{X}$ matrix, $q_{\text{grpout}} = 2$ is the dimension of $\boldsymbol{\Sigma}_{\text{grpout}}$ and $q_{\text{grpinn}} = 2$ is the dimension of $\boldsymbol{\Sigma}_{\text{grpinn}}$.

5.3.3 Streamlining mean field variational Bayes for the three-level Gaussian response model

Here we extend the streamlined MFVB algorithm for the two-level Gaussian response model as shown in Section 5.2.4 to cater to our three-level structured rodent tumor data. Once again, in keeping with the notation in Lee & Wand (2015) we partition the following vectors and matrices:

$$
\boldsymbol{Z}^{\text{R}} \equiv \underset{1\le i\le m}{\text{blockdiag}} \left(\boldsymbol{Z}_i^{\text{R}}\right),
$$

where

$$
\boldsymbol{Z}_i^{\text{R}} \equiv \left[\boldsymbol{X}_i \,|\, \boldsymbol{Z}_i^{\text{grpout}} \,|\, \underset{1\le i\le m}{\text{blockdiag}} \left(\boldsymbol{X}_{ij}\, \boldsymbol{Z}_{ij}^{\text{grpinn}}\right)\right].
$$

The global and random effects are partitioned according to:

$$
\boldsymbol{u} \equiv \begin{bmatrix} \boldsymbol{u}^{\text{gbl}} \\ \boldsymbol{u}^{\text{R}} \end{bmatrix}, \qquad \boldsymbol{Z} \equiv \left[\boldsymbol{Z}^{\text{gbl}} \;\; \boldsymbol{Z}^{\text{R}}\right],
$$

where

$$
\boldsymbol{u}^{\text{R}} \equiv \begin{bmatrix} \boldsymbol{u}_1^{\text{R}} \\ \vdots \\ \boldsymbol{u}_m^{\text{R}} \end{bmatrix}, \qquad \boldsymbol{u}_i^{\text{R}} = \begin{bmatrix} \boldsymbol{\delta}_i \\ \boldsymbol{u}_i^{\text{grpout}} \\ \boldsymbol{\gamma}_{i1} \\ \boldsymbol{u}_{i1}^{\text{grpinn}} \\ \vdots \\ \boldsymbol{\gamma}_{in_i} \\ \boldsymbol{u}_{in_i}^{\text{grpinn}} \end{bmatrix}.
$$

With these partitions in mind, a streamlined version of Algorithm 7 is presented in Algorithm 8. Successive values of the marginal log-likelihood lower bound are used to diagnose convergence of Algorithm 8. These values are monitored using the expression in (5.22) where

$$
\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}|
$$

Set up initial values:

$\mu_{q(1/\sigma^2_\varepsilon)}$, $\mu_{q(1/\sigma^2_{\text{gbl}})}$, $\mu_{q(1/\sigma^2_{\text{grpout}})}$, $\mu_{q(1/\sigma^2_{\text{grpinn}})}$, $\mu_{q(1/a_\varepsilon)}$, $\mu_{q(1/a_{\text{gbl}})}$, $\mu_{q(1/a_{\text{grpout}})}$, $\mu_{q(1/a_{\text{grpinn}})} > 0$, $\mu_{q\left(1/a_{\Sigma_{\text{grpout}},j}\right)}$, $\mu_{q\left(1/a_{\Sigma_{\text{grpinn}},j}\right)} > 0$, $1 \le j \le 2$, $\boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1}_{\text{grpout}})}$, $\boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1}_{\text{grpinn}})}$ positive definite.

Cycle through:

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

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

$$\boldsymbol{G}_i \leftarrow \mu_{q(1/\sigma^2_\varepsilon)} \left(\boldsymbol{C}^{\text{gbl}}_i\right)^\top \boldsymbol{Z}^{\text{R}}_i$$

$$\boldsymbol{H}_i \leftarrow \left[\mu_{q(1/\sigma^2_\varepsilon)} \left(\boldsymbol{Z}^{\text{R}}_i\right)^\top \boldsymbol{Z}^{\text{R}}_i + \text{blockdiag}\left\{M_{q(\boldsymbol{\Sigma}^{-1}_{\text{grpout}})}, \mu_{q\left(1/\sigma^2_{\text{grpout}}\right)} \boldsymbol{I}_{L_{\text{grpout}}}, \boldsymbol{I}_{n_i} \otimes \left(M_{q\left(\boldsymbol{\Sigma}^{-1}_{\text{grpinn}}\right)}, \mu_{q\left(1/\sigma^2_{\text{grpinn}}\right)} \boldsymbol{I}_{L_{\text{grpinn}}}\right)\right\}\right]^{-1}$$

$$\boldsymbol{S} \leftarrow \boldsymbol{S} + \boldsymbol{G}_i \boldsymbol{H}_i \boldsymbol{G}_i^\top; \quad \boldsymbol{s} \leftarrow \boldsymbol{s} + \boldsymbol{G}_i \boldsymbol{H}_i \left(\boldsymbol{Z}^{\text{R}}_i\right)^\top \boldsymbol{y}_i$$

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\text{G}})} \leftarrow \left\{\mu_{q(1/\sigma^2_\varepsilon)} \left(\boldsymbol{C}^{\text{gbl}}\right)^\top \boldsymbol{C}^{\text{gbl}} + \begin{bmatrix} \sigma_{\boldsymbol{\beta}}^{-2} \boldsymbol{I}_6 & \boldsymbol{0} \\ \boldsymbol{0} & \mu_{q\left(1/\sigma^2_{\text{gbl}}\right)} \boldsymbol{I}_{L_{\text{gbl}}} \end{bmatrix} - \boldsymbol{S}\right\}^{-1}$$

$$\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\text{G}})} \leftarrow \mu_{q(1/\sigma^2_\varepsilon)} \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\text{G}})} \left\{\left(\boldsymbol{C}^{\text{gbl}}\right)^\top \boldsymbol{y} - s\right\}$$

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

$$\boldsymbol{\Sigma}_{q\left(\boldsymbol{u}^{\text{R}}_i\right)} \leftarrow \boldsymbol{H}_i + \boldsymbol{H}_i \boldsymbol{G}_i^\top \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\text{G}})} \boldsymbol{G}_i \boldsymbol{H}_i$$

$$\boldsymbol{\mu}_{q\left(\boldsymbol{u}^{\text{R}}_i\right)} \leftarrow \boldsymbol{H}_i \left\{\mu_{q(1/\sigma^2_\varepsilon)} \left(\boldsymbol{Z}^{\text{R}}_i\right)^\top \boldsymbol{y}_i - \boldsymbol{G}_i^\top \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\text{G}})}\right\}$$

$$B_{q(\sigma^2_\varepsilon)} \leftarrow \mu_{q(1/a_\varepsilon)} + \frac{1}{2}\left[\left\| \boldsymbol{y} - \boldsymbol{C}^{\text{gbl}} \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\text{G}})} - \begin{bmatrix} \boldsymbol{Z}^{\text{R}}_1 \boldsymbol{\mu}_{q\left(\boldsymbol{u}^{\text{R}}_1\right)} \\ \vdots \\ \boldsymbol{Z}^{\text{R}}_m \boldsymbol{\mu}_{q\left(\boldsymbol{u}^{\text{R}}_m\right)} \end{bmatrix} \right\|^2 + \text{tr}\left\{\left(\boldsymbol{C}^{\text{gbl}}\right)^\top \boldsymbol{C}^{\text{gbl}} \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\text{G}})}\right\} + \sum_{i=1}^{m} \text{tr}\left\{\left(\boldsymbol{Z}^{\text{R}}_i\right)^\top \boldsymbol{Z}^{\text{R}}_i \boldsymbol{\Sigma}_{q\left(\boldsymbol{u}^{\text{R}}_i\right)}\right\} - 2\mu^{-1}_{q(1/\sigma^2_\varepsilon)} \sum_{i=1}^{m} \text{tr}\left(\boldsymbol{G}_i \boldsymbol{H}_i \boldsymbol{G}_i^\top \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\text{G}})}\right)\right]$$

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

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

$$\mu_{q\left(1/\sigma^2_{\text{gbl}}\right)} \leftarrow \tfrac{1}{2}\left(L_{\text{gbl}} + 1\right) / B_{q\left(\sigma^2_{\text{gbl}}\right)}$$

$$B_{q\left(a_{\text{gbl}}\right)} \leftarrow \mu_{q\left(1/\sigma^2_{\text{gbl}}\right)} + \left(A_{\text{gbl}}\right)^{-2}; \quad \mu_{q\left(1/a_{\text{gbl}}\right)} \leftarrow 1/B_{q\left(a_{\text{gbl}}\right)}$$

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

$$\mu_{q\left(1/\sigma^2_{\text{grpout}}\right)} \leftarrow \tfrac{1}{2}\left(m\,L_{\text{grpout}}+1\right)/B_{q\left(\sigma^2_{\text{grpout}}\right)}$$

$$B_{q(a_{\text{grpout}})} \leftarrow \mu_{q\left(1/\sigma^2_{\text{grpout}}\right)} + A^{-2}_{\text{grpout}}, \qquad \mu_{q(1/a_{\text{grpout}})} \leftarrow 1/B_{q(a_{\text{grpout}})}$$

$$\boldsymbol{B}_{q(\boldsymbol{\Sigma}_{\text{grpout}})} \leftarrow \sum_{i=1}^{m}\left\{\boldsymbol{\mu}_{q(\boldsymbol{\delta}_i)}\boldsymbol{\mu}^{\top}_{q(\boldsymbol{\delta}_i)} + \boldsymbol{\Sigma}_{q(\boldsymbol{\delta}_i)}\right\} + 2\nu_{\text{grpout}}\text{diag}\left(\mu_{q\left(1/a_{\Sigma_{\text{grpout}},1}\right)}, \mu_{q\left(1/a_{\Sigma_{\text{grpout}},2}\right)}\right)$$

$$\boldsymbol{M}_{q\left(\boldsymbol{\Sigma}^{-1}_{\text{grpout}}\right)} \leftarrow \left(\nu_{\text{grpout}}+m+1\right)\boldsymbol{B}^{-1}_{q(\boldsymbol{\Sigma}_{\text{grpout}})}$$

For $j = 1, 2$:

$$B_{q\left(a_{\Sigma_{\text{grpout}},j}\right)} \leftarrow \nu_{\text{grpout}}\left(\boldsymbol{M}_{q\left(\boldsymbol{\Sigma}^{-1}_{\text{grpout}}\right)}\right)_{jj} + 1/A^2_{\Sigma_{\text{grpout}},j}$$

$$\mu_{q\left(1/a_{\Sigma_{\text{grpout}},j}\right)} \leftarrow \left(\frac{\nu_{\text{grpout}}}{2}+1\right)\Big/B_{q\left(a_{\Sigma_{\text{grpout}},j}\right)}$$

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

$$\mu_{q\left(1/\sigma^2_{\text{grpinn}}\right)} \leftarrow \tfrac{1}{2}\left(o\,m\,L_{\text{grpinn}}+1\right)/B_{q\left(\sigma^2_{\text{grpinn}}\right)}$$

$$B_{q(a_{\text{grpinn}})} \leftarrow \mu_{q\left(1/\sigma^2_{\text{grpinn}}\right)} + A^{-2}_{\text{grpinn}}, \qquad \mu_{q(1/a_{\text{grpinn}})} \leftarrow 1/B_{q(a_{\text{grpinn}})}$$

$$\boldsymbol{B}_{q(\boldsymbol{\Sigma}_{\text{grpinn}})} \leftarrow \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left\{\boldsymbol{\mu}_{q(\boldsymbol{\gamma}_{ij})}\boldsymbol{\mu}^{\top}_{q(\boldsymbol{\gamma}_{ij})} + \boldsymbol{\Sigma}_{q(\boldsymbol{\gamma}_{ij})}\right\} + 2\nu_{\text{grpinn}}\text{diag}\left(\mu_{q\left(1/a_{\Sigma_{\text{grpinn}},1}\right)}, \mu_{q\left(1/a_{\Sigma_{\text{grpinn}},2}\right)}\right)$$

$$\boldsymbol{M}_{q\left(\boldsymbol{\Sigma}^{-1}_{\text{grpinn}}\right)} \leftarrow \left(\nu_{\text{grpinn}} + \sum_{i=1}^{m} n_i + 1\right)\boldsymbol{B}^{-1}_{q(\boldsymbol{\Sigma}_{\text{grpinn}})}$$

For $j = 1, 2$:

$$B_{q\left(a_{\Sigma_{\text{grpinn}},j}\right)} \leftarrow \nu_{\text{grpinn}}\left(\boldsymbol{M}_{q\left(\boldsymbol{\Sigma}^{-1}_{\text{grpinn}}\right)}\right)_{jj} + 1/A^2_{\Sigma_{\text{grpinn}},j}$$

$$\mu_{q\left(1/a_{\Sigma_{\text{grpinn}},j}\right)} \leftarrow \left(\frac{\nu_{\text{grpinn}}}{2}+1\right)\Big/B_{q\left(a_{\Sigma_{\text{grpinn}},j}\right)}$$

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

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

$$\boldsymbol{\Lambda}_q(\boldsymbol{\beta}, \boldsymbol{u}^{\text{gbl}}, \boldsymbol{u}^{\text{R}}_i) \equiv E_q\left[\left\{\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}^{\text{gbl}}\end{bmatrix} - \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\text{gbl}})}\right\}\left\{\boldsymbol{u}^{\text{R}}_i - \boldsymbol{\mu}_q(\boldsymbol{u}^{\text{R}}_i)\right\}^{\top}\right] \leftarrow -\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\text{gbl}})}\,\boldsymbol{G}_i\,\boldsymbol{H}_i$$

Algorithm 8: *Streamlined MFVB algorithm for the three-level Gaussian response model given in (5.19).*

is replaced by

$$-\sum_{i=1}^{m}\log\Big|\mu_{q(1/\sigma^2_\varepsilon)}\left(\boldsymbol{Z}_i^{\mathrm{R}}\right)^\top\boldsymbol{Z}_i^{\mathrm{R}}+\text{blockdiag}\Big\{M_{q(\boldsymbol{\Sigma}^{-1})},\mu_{q\left(1/\sigma^2_{\text{grpout}}\right)}\boldsymbol{I}_{L_{\text{grpout}}},\\ \boldsymbol{I}_{n_i}\otimes\left(\boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1}_{\text{grpinn}})},\mu_{q(1/\sigma^2_{\text{grpinn}})}\boldsymbol{I}_{L_{\text{grpinn}}}\right)\Big\}\Big|-\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\text{gbl}})}|.$$

Once again, Algorithm 8 should not be considered a fully streamlined version of Algorithm 7 because of the structure of $\boldsymbol{H}_i$. In addition to incorporating the variance structure of the spline components for each subject's curve, there is also the variance structure associated with the slice specific curves. This is represented in $\boldsymbol{H}_i$ as $\boldsymbol{I}_{\sum_{i=1}^m n_i}\otimes\left(\boldsymbol{M}_{q\left(\boldsymbol{\Sigma}^{-1}_{\text{grpinn}}\right)},\mu_{q(1/\sigma^2_{\text{grpinn}})}\boldsymbol{I}_{L_{\text{grpinn}}}\right)$. This block diagonal structure, comprising storage of a large amount of zeros, is hindering the potential time savings that would be possible given further research into this area.

5.3.3.1 Results

The three-level model given in (5.19) was fit using the streamlined MFVB algorithm as shown in Algorithm 8. For fitting this data we set $L_{\text{gbl}} = 17$, $L_{\text{grpout}} = 12$ and $L_{\text{grpinn}} = 9$ and set all hyperparameters to 10^5. The `rstan` package was used for performing MCMC with a burn-in of 1000 and 1000 additional iterations. The MFVB iterations were terminated when the increase in the corresponding $\log \underline{p}(\texttt{bsc}; q)$ fell bellow 10^{-8}. The MCMC fit took approximately 18 hours whilst the streamlined MFVB fit took just under 2 hours to run. Figure 5.8 illustrates the accuracy of the streamlined MFVB approach against the MCMC benchmark. The parameters being monitored are the quartiles of the curve estimates of the 1st rodent for three of its slice curves. These are represented as $f(Q_i) + g_1(Q_i) + h_{11}(Q_i)$, $f(Q_i) + g_1(Q_i) + h_{12}(Q_i)$ and $f(Q_i) + g_1(Q_i) + h_{13}(Q_i)$ for $i = 1, 2, 3$. The accuracy of the curve estimates for these parameters are extremely high, especially considering the complexity of the model.

Figures 5.9 and 5.10 provide illustration of the inference provided by the streamlined MFVB approach and MCMC for the three-level Gaussian response model. Each of these two figures represents a rodent with the interior panels corresponding to the slices or probe locations in which each tumor was scanned. The streamlined MFVB fits are almost indistinguishable from the MCMC fits for each slice.

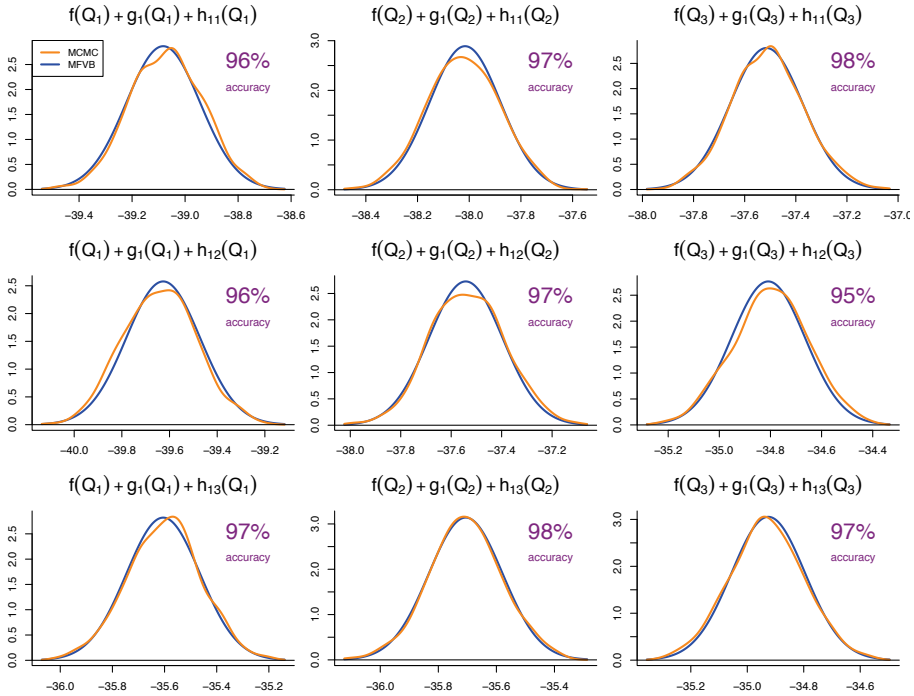


Figure 5.8: *Approximate posterior density functions obtained via MCMC and streamlined MFVB for the quartiles of three specific slice curves from the rodent tumor data-set. MFVB accuracy scores are displayed.*

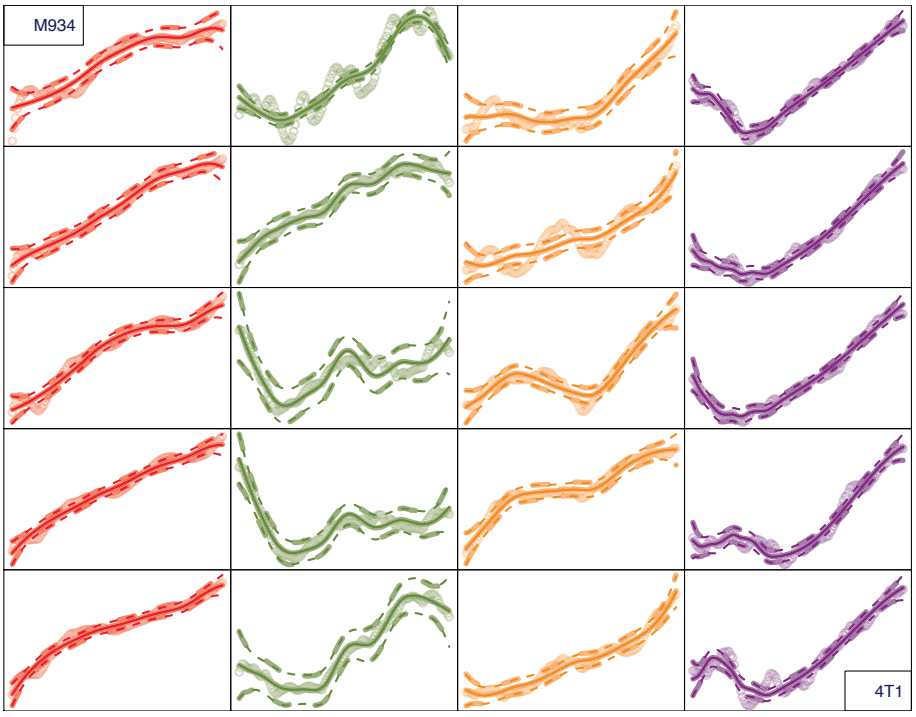


Figure 5.9: *Slice-specific estimates and pointwise 95 % credible sets for a rodent given in Figure 5.6. The different colours correspond to the different transducers. The thicker curve estimates correspond to MCMC and the thinner ones correspond to streamlined MFVB fitting. The data points for each slice are also plotted.*

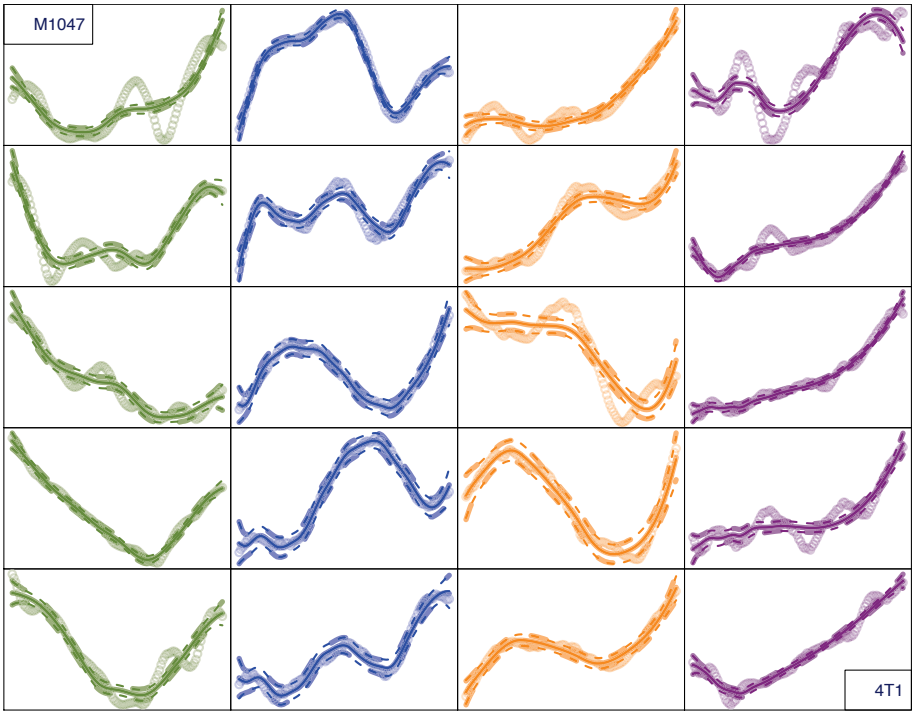


Figure 5.10: *Slice-specific estimates and pointwise 95 % credible sets for a rodent given in Figure 5.6. The different colours correspond to the different transducers. The thicker curve estimates correspond to MCMC and the thinner ones correspond to streamlined MFVB fitting. The data points for each slice are also plotted.*

5.3.4 Simulation study

An extensive simulation study was carried out to assess the speed of Algorithm 8 against that of Algorithm 7. We generated 25 data-sets according to

$$y_{ijk} = f(x_{ijk}) + g_i(x_{ijk}) + h_{ij}(x_{ijk}) + \varepsilon_{ijk}, \qquad 1 \leq i \leq m, \;\; 1 \leq j \leq n_i, \;\; 1 \leq k \leq o_{ij},$$

where the $x_{ijk} \overset{\text{ind.}}{\sim} \text{Uniform}(0, 1)$, $\varepsilon_{ijk} \overset{\text{ind.}}{\sim} \text{N}(0, \sigma^2_\varepsilon)$, $u^{\text{gbl}}_\ell \overset{\text{ind.}}{\sim} \text{N}(0, \sigma^2_{\text{gbl}})$, $\boldsymbol{\delta}_i \overset{\text{ind.}}{\sim} \text{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\text{grpout}})$, $u^{\text{grpout}}_{i\ell} \overset{\text{ind.}}{\sim} \text{N}(0, \sigma^2_{\text{grpout}})$, $\boldsymbol{\gamma}_{ij} \overset{\text{ind.}}{\sim} \text{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\text{grpinn}})$, and $u^{\text{grpinn}}_{ij\ell} \overset{\text{ind.}}{\sim} \text{N}(0, \sigma^2_{\text{grpinn}})$. The true parameter values were specified as

$$\beta_0 = 0.2, \quad \beta_1 = 1.8, \quad \sigma^2_\varepsilon = 0.02, \quad \sigma^2_{\text{gbl}} = 1.5, \quad \boldsymbol{\Sigma}_{\text{grpout}} = \begin{bmatrix} 0.4 & 0.1 \\ 0.1 & 0.4 \end{bmatrix}, \quad \sigma^2_{\text{grpout}} = 0.29,$$

$$\boldsymbol{\Sigma}_{\text{grpinn}} = \begin{bmatrix} 0.4 & 0.15 \\ 0.15 & 0.4 \end{bmatrix} \quad \text{and} \quad \sigma^2_{\text{grpinn}} = 0.17.$$

The number of subjects between simulation studies was varied, where $m \in \{5, 15, 25, 35\}$, the within subject curves remained constant at $n_i = 5$, $1 \leq i \leq m$ and the within curve

sample size remained constant at $o_{ij} = 10$, $1 \leq i \leq m$, $1 \leq j \leq n_i$. The computations were performed on a laptop computer (Mac OS X; 2.8 GHz processor, 16 GBytes of random access memory). The computation times for each approach are summarised in Table 5.4. The time savings here will not be as impressive as the time savings given in Table 5.1,

m	Naïve	Streamlined	Ratio
5	28.49 (3.98)	4.34 (0.15)	6.57
15	617.65 (11.37)	44.79 (1.77)	13.79
25	2638.30 (20.81)	177.42 (0.19)	14.87
35	7051.99 (22.47)	455.89 (2.06)	15.47

Table 5.4: *Average (standard deviation) run time in seconds for naïve and streamlined MFVB fitting of the three-level Gaussian response model.*

due to the storage issues contained in $\boldsymbol{H}_i$. However, we still notice that the streamlined approach for the three-level Gaussian response model is significantly faster than the Naïve approach.

5.4 Discussion

The original aim of this chapter was to develop a fast deterministic approach to MCMC to shorten estimation time for the two-level and three-level Gaussian response model. The rodent tumor dataset that was introduced in Section 5.3 led to the consideration of a possible streamlined approach to MFVB. Meanwhile, research was being conducted by Lee & Wand (2015) that led to the fully streamlined MFVB algorithm for a similar two-level Gaussian response model. This led to the incorporation of a variant of this streamlined algorithm for this chapter. As is evident from the real datasets and simulation studies for both the two-level and three-level Gaussian response models considered here, MFVB in general has achieved faster inference, but a streamlined version of MFVB in particular has achieved even faster inference with little loss in accuracy. The models considered in this chapter are of reasonably high complexity and further research into a fully streamlined algorithm for both models would indeed be beneficial.

5.A Derivation of Algorithm 5

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

Firstly, we define

$$\boldsymbol{G} \equiv \begin{bmatrix} \sigma^2_{\text{gbl}} \boldsymbol{I}_{L_{\text{gbl}}} & \boldsymbol{0} \\ \boldsymbol{0} & \underset{1 \le i \le m}{\text{blockdiag}} \left(\boldsymbol{\Sigma}, \sigma^2_{\text{grp}} \boldsymbol{I}_{L_{\text{grp}}} \right) \end{bmatrix}.$$

Next, we note that

$$\begin{aligned} p(\boldsymbol{\beta}, \boldsymbol{u}|\text{rest}) &\propto p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma^2_\varepsilon)\, p(\boldsymbol{u}|\sigma^2_{\text{gbl}}, \boldsymbol{\Sigma}, \sigma^2_{\text{grp}})\, p(\boldsymbol{\beta}) \\ &\propto \exp\left\{ -\tfrac{1}{2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})^\top \left(\sigma^2_\varepsilon \boldsymbol{I}_n \right)^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u}) \right\} \times \exp\left\{ -\tfrac{1}{2} \boldsymbol{u}^\top \boldsymbol{G}^{-1} \boldsymbol{u} \right\} \\ &\quad \times \exp\left(-\tfrac{1}{2\sigma^2_{\boldsymbol{\beta}}} \parallel \boldsymbol{\beta} \parallel^2 \right). \end{aligned}$$

We next combine the design matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ to form $\boldsymbol{C} = [\boldsymbol{X} \,|\, \boldsymbol{Z}]$. Taking the logarithm of both sides gives

$$\begin{aligned} &\log p(\boldsymbol{\beta}, \boldsymbol{u}|\text{rest}) \\ &\quad \propto -\tfrac{1}{2} \left(\boldsymbol{y} - \boldsymbol{C} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} \right)^\top \left(\sigma^{-2}_\varepsilon \boldsymbol{I}_n \right) \left(\boldsymbol{y} - \boldsymbol{C} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} \right) - \tfrac{1}{2} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix}^\top \begin{bmatrix} \sigma^{-2}_{\boldsymbol{\beta}} \boldsymbol{I}_p & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Omega}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} \\ &\quad = -\tfrac{1}{2} \left\{ \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix}^\top \left(\boldsymbol{C}^\top \left(\sigma^{-2}_\varepsilon \boldsymbol{I}_n \right) \boldsymbol{C} + \begin{bmatrix} \sigma^{-2}_{\boldsymbol{\beta}} \boldsymbol{I}_p & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Omega}^{-1} \end{bmatrix} \right) \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} - 2 \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix}^\top \boldsymbol{C}^\top \left(\sigma^{-2}_\varepsilon \boldsymbol{I}_n \right) \boldsymbol{y} \right\} \\ &\qquad + \text{const} \end{aligned}$$

where 'const' denotes terms not depending on the parameter vector $(\boldsymbol{\beta}, \boldsymbol{u})$. Then, taking expectations with respect to all parameters except $(\boldsymbol{\beta}, \boldsymbol{u})$:

$$\begin{aligned} \log q^* (\boldsymbol{\beta}, \boldsymbol{u}) &= E_q \left\{ \log p(\boldsymbol{\beta}, \boldsymbol{u}|\text{rest}) \right\} + \text{const} \\ &= -\tfrac{1}{2} \left\{ \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix}^\top \left(\boldsymbol{C}^\top \left(\mu_{q(1/\sigma^2_\varepsilon)} \boldsymbol{I}_n \right) \boldsymbol{C} + \begin{bmatrix} \sigma^{-2}_{\boldsymbol{\beta}} \boldsymbol{I}_p & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{M}_{q(\boldsymbol{G}^{-1})} \end{bmatrix} \right) \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} \right. \\ &\qquad \left. -2 \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix}^\top \boldsymbol{C}^\top \left(\mu_{q(1/\sigma^2_\varepsilon)} \boldsymbol{I}_n \right) \boldsymbol{y} \right\} + \text{const} \end{aligned}$$

where

$$\boldsymbol{M}_{q(\boldsymbol{G}^{-1})} \equiv \begin{bmatrix} \mu_{q(1/\sigma^2_{\text{gbl}})}\boldsymbol{I}_{L_{\text{gbl}}} & \boldsymbol{0} \\ \boldsymbol{0} & \underset{1\leq i\leq m}{\text{blockdiag}}\left(\boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1})}\,,\, \mu_{q(1/\sigma^2_{\text{grp}})}\boldsymbol{I}_{L_{\text{grp}}}\right) \end{bmatrix},$$

and

$$\boldsymbol{M}_{q(\boldsymbol{X})} \equiv E_{q(\boldsymbol{X})}.$$

Completing the square from above gives

$$q^*\left(\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}\right) \sim \text{N}\left(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})},\, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\right)$$

where

$$\begin{aligned}
\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} &= \mu_{q(1/\sigma^2_\varepsilon)}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^\top\boldsymbol{y}, \text{ and} \\
\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} &= \left(\mu_{q(1/\sigma^2_\varepsilon)}\boldsymbol{C}^\top\boldsymbol{C} + \begin{bmatrix}\sigma_{\boldsymbol{\beta}}^{-2}\boldsymbol{I}_p & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{M}_{q(\boldsymbol{G}^{-1})}\end{bmatrix}\right)^{-1}
\end{aligned}$$

where $p = 2$ is the column length in $\boldsymbol{X}$.

Expressions for $B_{q(\sigma^2_\varepsilon)}$ and $\mu_{q(1/\sigma^2_\varepsilon)}$

The full conditional distribution for σ^2_ε is given by

$$\begin{aligned}
p(\sigma^2_\varepsilon|\text{rest}) &\propto p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2_\varepsilon)\, p(\sigma^2_\varepsilon|a_\varepsilon) \\
&= |\sigma^2_\varepsilon\boldsymbol{I}_n|^{-1/2}\exp\left\{-\tfrac{1}{2}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{u}\right)^\top\left(\sigma^2_\varepsilon\boldsymbol{I}_n\right)^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{u}\right)\right\} \\
&\quad \left(\sigma^2_\varepsilon\right)^{-3/2}\exp\left\{-\left(1/a_\varepsilon\right)/\sigma^2_\varepsilon\right\} + \text{const}
\end{aligned}$$

Taking the logarithm of both sides gives

$$\begin{aligned}
\log p(\sigma^2_\varepsilon|\text{rest}) &= -\frac{n}{2}\log(\sigma^2_\varepsilon) - \frac{1}{2\sigma^2_\varepsilon}\parallel \boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{u}\parallel^2 - \frac{3}{2}\log(\sigma^2_\varepsilon) - \left(1/a_\varepsilon\right)/\sigma^2_\varepsilon + \text{const} \\
&= \tfrac{1}{2}(n+3)\log(\sigma^2_\varepsilon) - \left(\tfrac{1}{2}\parallel \boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{u}\parallel^2 + 1/a_\varepsilon\right)/\sigma^2_\varepsilon + \text{const}
\end{aligned}$$

Taking expectations:

$$\begin{aligned}
\log q^*(\sigma^2_\varepsilon) &= E_q\left\{\log p(\sigma^2_\varepsilon|\text{rest})\right\} + \text{const} \\
&= \tfrac{1}{2}(n+3)\log(\sigma^2_\varepsilon) - \left(\tfrac{1}{2}E_q\parallel \boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{u}\parallel^2 + \mu_{q(1/a_\varepsilon)}\right)/\sigma^2_\varepsilon + \text{const}
\end{aligned}$$

Using result 1.4.19 it follows that

$$\log q^*(\sigma_\varepsilon^2) = \tfrac{1}{2}(n+3)\log(\sigma_\varepsilon^2) - \left(\tfrac{1}{2}\left\{\| \boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \|^2 + \mathrm{tr}\left(\boldsymbol{C}^\top \boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\right)\right\} + \mu_{q(1/a_\varepsilon)}\right)/\sigma_\varepsilon^2$$
$$+\text{const}$$

This is of the form of an Inverse-Gamma distribution, therefore

$$q^*(\sigma_\varepsilon^2) \sim \text{Inverse-Gamma}\left(\tfrac{1}{2}(n+1),\, B_{q(\sigma_\varepsilon^2)}\right),$$

where

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

and using Result 1.4.3, we get

$$\mu_{q(1/\sigma_\varepsilon^2)} = \tfrac{1}{2}(n+1)/B_{q(\sigma_\varepsilon^2)}.$$

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

The full conditional distribution of a_ε is given by

$$\begin{aligned} p(a_\varepsilon \,|\, \text{rest}) &\propto p(\sigma_\varepsilon^2|a_\varepsilon)\, p(a_\varepsilon) \\ &= (1/a_\varepsilon)^{1/2} \exp\left\{-(1/a_\varepsilon)/\sigma_\varepsilon^2\right\} \times (a_\varepsilon)^{-\frac{1}{2}-1} \exp\left\{-(1/A_\varepsilon^2)/a_\varepsilon\right\} + \text{const} \end{aligned}$$

Next, taking the logarithm gives

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

Taking expectations:

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

This is of the form of an Inverse-Gamma distribution, therefore

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

where

$$B_{q(a_\varepsilon)} = \mu_{q(1/\sigma_\varepsilon^2)} + A_\varepsilon^{-2}$$

and using Result 1.4.3, we get

$$\mu_{q(1/a_\varepsilon)} = 1/B_{q(a_\varepsilon)}.$$

Expressions for $B_{q\left(\sigma^2_{\mathbf{gbl}}\right)}$, and $\mu_{q\left(1/\sigma^2_{\mathbf{gbl}}\right)}$

The full conditional distribution of σ^2_ε is attained through

$$\begin{aligned} p(\sigma^2_{\text{gbl}}\,|\,\text{rest}) &\propto p(\boldsymbol{u}\,|\,\sigma^2_{\text{gbl}},\,\boldsymbol{\Sigma},\,\sigma^2_{\text{grp}})\,p(\sigma^2_{\text{gbl}}\,|\,a_{\text{gbl}}) \\ &= |\boldsymbol{G}|^{-\frac{1}{2}}\exp\left\{-\tfrac{1}{2}\boldsymbol{u}^\top\boldsymbol{G}^{-1}\boldsymbol{u}\right\}\,(\sigma^2_{\text{gbl}})^{-\frac{1}{2}-1}\exp\left\{(1/a_{\text{gbl}})\,/\sigma^2_{\text{gbl}}\right\}+\text{const} \end{aligned}$$

Taking the logarithm of both sides gives

$$\begin{aligned} \log p(\sigma^2_{\text{gbl}}|\text{rest}) &= -\tfrac{L_{\text{gbl}}}{2}\log(\sigma^2_{\text{gbl}}) - \tfrac{3}{2}\log(\sigma^2_{\text{gbl}}) - \tfrac{1}{\sigma^2_{\text{gbl}}}\parallel\boldsymbol{u}^{\text{gbl}}\parallel^2 - (1/a_{\text{gbl}})/\sigma^2_{\text{gbl}} + \text{const} \\ &= -\tfrac{1}{2}(L_{\text{gbl}}+3)\log(\sigma^2_{\text{gbl}}) - \left\{\tfrac{1}{2}\parallel\boldsymbol{u}^{\text{gbl}}\parallel^2 + 1/a_{\text{gbl}}\right\}/\sigma^2_{\text{gbl}} + \text{const} \end{aligned}$$

Taking expectations with respect to all parameters but σ^2_{gbl}, with the help of Result 1.4.17, we get

$$\begin{aligned} \log q^*(\sigma^2_{\text{gbl}}) &= E_q\left\{\log p(\sigma^2_{\text{gbl}}|\text{rest})\right\} \\ &= -\tfrac{1}{2}(L_{\text{gbl}}+3)\log(\sigma^2_{\text{gbl}}) - \left[\tfrac{1}{2}\left\{\parallel\boldsymbol{\mu}_{q(\boldsymbol{u}^{\text{gbl}})}\parallel^2 + \text{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{u}^{\text{gbl}})}\right)\right\} + \mu_{(1/a_{\text{gbl}})}\right]/\sigma^2_{\text{gbl}} \\ &\quad + \text{const} \end{aligned}$$

Thus, we can represent the optimal q-density of σ^2_{gbl} as:

$$q^*(\sigma^2_{\text{gbl}}) \sim \text{Inverse-Gamma}\left(\tfrac{1}{2}(L_{\text{gbl}}+1),\, B_{q\left(\sigma^2_{\text{gbl}}\right)}\right),$$

where

$$B_{q\left(\sigma^2_{\text{gbl}}\right)} = \tfrac{1}{2}\left\{\parallel\boldsymbol{\mu}_{q(\boldsymbol{u}^{\text{gbl}})}\parallel^2 + \text{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{u}^{\text{gbl}})}\right)\right\} + \mu_{(1/a_{\text{gbl}})}$$

and using Result 1.4.3, we get

$$\mu_{q\left(1/\sigma^2_{\text{gbl}}\right)} = \tfrac{1}{2}(L_{\text{gbl}}+1)/B_{q\left(\sigma^2_{\text{gbl}}\right)}.$$

Expressions for $B_{q(a_{\mathbf{gbl}})}$ and $\mu_{q(1/a_{\mathbf{gbl}})}$

The derivations for $B_{q(a_{\text{gbl}})}$ and $\mu_{q(1/a_{\text{gbl}})}$ are similar to that for $B_{q(\sigma^2_\varepsilon)}$ and $\mu_{q(1/\sigma^2_\varepsilon)}$, respectively. Without going into further detail, the optimal q-density for a_{gbl} satisfies

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

where

$$B_{q(a_{\text{gbl}})} = \mu_{q(1/\sigma^2_{\text{gbl}})} + A^{-2}_{\text{gbl}}$$

and

$$\mu_{q(1/a_{\text{gbl}})} = 1/B_{q(a_{\text{gbl}})}.$$

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

The full conditional distribution for $\boldsymbol{\Sigma}$ is given by:

$$\begin{aligned} p(\boldsymbol{\Sigma}|\text{rest}) &\propto p(\boldsymbol{u}\,|\,\sigma^2_{\text{gbl}},\, \boldsymbol{\Sigma},\, \sigma^2_{\text{grp}})\, p(\boldsymbol{\Sigma}\,|\,a_{\Sigma,1},\, a_{\Sigma,2}) \\ &= |\boldsymbol{G}|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}\boldsymbol{u}^\top \boldsymbol{G}^{-1}\boldsymbol{u}\right) |\boldsymbol{\Sigma}|^{-\frac{(\nu+p-1)+p+1}{2}} \\ &\quad \times \exp\left[-\tfrac{1}{2}\text{tr}\left\{2\nu\,\text{diag}\left(1/a_{\Sigma,1}, 1/a_{\Sigma,2}\right)\right\}\boldsymbol{\Sigma}^{-1}\right] + \text{const} \end{aligned}$$

where $p = 2$ is the dimension of $\boldsymbol{\Sigma}$. Taking the logarithm of both sides, we get:

$$\begin{aligned} \log p(\boldsymbol{\Sigma}\,|\,\text{rest}) &= -\tfrac{m}{2}\log|\boldsymbol{\Sigma}| - \tfrac{1}{2}\boldsymbol{\delta}^\top(\boldsymbol{I}_m \otimes \boldsymbol{\Sigma})\boldsymbol{\delta} - \tfrac{\nu+4}{2}\log|\boldsymbol{\Sigma}| \\ &\quad -\text{tr}\left\{2\nu\,\text{diag}\left(1/a_{\Sigma,1}, 1/a_{\Sigma,2}\right)\boldsymbol{\Sigma}^{-1}\right\} + \text{const} \end{aligned}$$

where $\boldsymbol{\delta} = \left[\boldsymbol{\delta}_1^\top, \ldots, \boldsymbol{\delta}_m^\top\right]^\top$. Further simplification results in

$$\begin{aligned} \log p(\boldsymbol{\Sigma}|\text{rest}) &= -\tfrac{1}{2}(\nu+m+4)\log|\boldsymbol{\Sigma}| - \tfrac{1}{2}\text{tr}\left[\left\{\textstyle\sum_{i=1}^m \boldsymbol{\delta}_i\boldsymbol{\delta}_i^\top + 2\nu\,\text{diag}\left(1/a_{\Sigma,1}, 1/a_{\Sigma,2}\right)\right\}\boldsymbol{\Sigma}^{-1}\right] \\ &\quad +\text{const}. \end{aligned}$$

Taking expectations:

$$\begin{aligned} \log q^*(\boldsymbol{\Sigma}) &= E_q\left\{\log p(\boldsymbol{\Sigma}|\text{rest})\right\} \\ &= -\tfrac{1}{2}(\nu+m+4)\log|\boldsymbol{\Sigma}| - \tfrac{1}{2}\left\{\left[\textstyle\sum_{i=1}^m \left\{\boldsymbol{\mu}_{q(\boldsymbol{\delta}_i)}\boldsymbol{\mu}_{q(\boldsymbol{\delta}_i)}^\top + \boldsymbol{\Sigma}_{q(\boldsymbol{\delta}_i)}\right\}\right. \right. \\ &\quad \left.\left. +2\nu\,\text{diag}\left(\mu_{q(1/a_{\Sigma,1})}, \mu_{q(1/a_{\Sigma,2})}\right)\right]\boldsymbol{\Sigma}^{-1}\right\} + \text{const} \end{aligned}$$

Therefore, we can represent the optimal q-density of $\boldsymbol{\Sigma}$ as:

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

where

$$B_{q(\boldsymbol{\Sigma})} = \sum_{i=1}^m \left\{\boldsymbol{\mu}_{q(\boldsymbol{\delta}_i)}\boldsymbol{\mu}_{q(\boldsymbol{\delta}_i)}^\top + \boldsymbol{\Sigma}_{q(\boldsymbol{\delta}_i)}\right\} + 2\nu\,\text{diag}\left(\mu_{q(1/a_{\Sigma,1})}, \mu_{q(1/a_{\Sigma,2})}\right)$$

and using Results 1.4.7 and 1.4.8, we get

$$\boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1})} \equiv E_{q(\boldsymbol{\Sigma}^{-1})} = (\nu+m+1)\boldsymbol{B}_{q(\boldsymbol{\Sigma})}^{-1}.$$

Expressions for the $B_{q(a_{\Sigma,j})}$ and $\mu_{q(1/a_{\Sigma,j})}$

We start with

$$\begin{aligned} p(a_{\Sigma,j}\,|\,\text{rest}) &\propto p(\boldsymbol{\Sigma}\,|\,a_{\Sigma,1},a_{\Sigma,2})\,p(a_{\Sigma,j}) \\ &= |2\nu\,\text{diag}\,(1/a_{\Sigma,1},1/a_{\Sigma,2})\,|^{\frac{\nu+1}{2}}\exp\left\{-\tfrac{1}{2}\text{tr}\left(2\nu\,\text{diag}\,(1/a_{\Sigma,1},1/a_{\Sigma,2})\,\boldsymbol{\Sigma}^{-1}\right)\right\} \\ &\quad\times(a_{\Sigma,j})^{-\frac{1}{2}-1}\exp\left\{-\left(1/A_{\Sigma,j}\right)/a_{\Sigma,j}\right\}+\text{const} \end{aligned}$$

Taking the logarithm of both sides gives

$$\begin{aligned} \log p(a_{\Sigma,j}\,|\,\text{rest}) &= -\tfrac{\nu+1}{2}\log(a_{\Sigma,j})-\tfrac{3}{2}\log(a_{\Sigma,j})-\left(1/A_{\Sigma,j}\right)/a_{\Sigma,j}-\left\{\nu\left(\boldsymbol{\Sigma}^{-1}\right)_{jj}\right\}/a_{\Sigma,j} \\ &\quad+\text{const} \\ &= -\tfrac{\nu+4}{2}\log(a_{\Sigma,j})-\left\{\nu\left(\boldsymbol{\Sigma}^{-1}\right)_{jj}+A_{\Sigma,j}^{-2}\right\}/a_{\Sigma,j} \end{aligned}$$

Taking expectations gives

$$\begin{aligned} \log q^*(a_{\Sigma,j}) &= E_q\left\{\log p(a_{\Sigma,j}\,|\,\text{rest})\right\} \\ &= -\tfrac{\nu+4}{2}\log(a_{\Sigma,j})-\left\{\nu\left(\boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1})}\right)_{jj}+A_{\Sigma,j}^{-2}\right\}/a_{\Sigma,j}+\text{const} \end{aligned}$$

Therefore,

$$q^*(a_{\Sigma,j})\sim\text{Inverse-Gamma}\left(\tfrac{\nu}{2}+1,\,B_{q(a_{\Sigma,j})}\right),\qquad j=1,2,$$

where

$$B_{q(a_{\Sigma,j})}=\nu\left(\boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1})}\right)_{jj}+A_{\Sigma,j}^{-2}$$

and through the use of Result 1.4.3

$$\mu_{q(1/a_{\Sigma,j})}=\left(\tfrac{\nu}{2}+1\right)/B_{q(a_{\Sigma,j})}.$$

Expressions for $B_{q(\sigma^2_{\mathbf{grp}})}$ and $\mu_{q(1/\sigma^2_{\mathbf{grp}})}$

The derivations for $B_{q(\sigma^2_{\text{grp}})}$ and $\mu_{q(1/\sigma^2_{\text{grp}})}$ are similar to that for $B_{q\left(\sigma^2_{\text{gbl}}\right)}$ and $\mu_{q(1/\sigma^2_{\text{gbl}})}$, respectively. Without going through further detail, the optimal q-density for σ^2_{grp} is shown to be of the form:

$$q^*(\sigma^2_{\text{grp}})\sim\text{Inverse-Gamma}\left(\tfrac{1}{2}(m\,L_{\text{grp}}+1),\,B_{q(\sigma^2_{\text{grp}})}\right),$$

where

$$B_{q(\sigma^2_{\text{grp}})}=\tfrac{1}{2}\left\{\|\,\boldsymbol{\mu}_{q(\boldsymbol{u}^{\text{grp}})}\,\|^2+\text{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{u}^{\text{grp}})}\right)\right\}+\mu_{(1/a_{\text{grp}})}$$

and using Result 1.4.3, we get

$$\mu_{q(1/\sigma^2_{\text{grp}})} = \tfrac{1}{2}(m\,L_{\text{grp}} + 1)/B_{q(\sigma^2_{\text{grp}})}.$$

5.B Derivation of the marginal log-likelihood lower bound

The marginal log-likelihood lower bound expression given in (5.13) is:

$$\begin{aligned}
\log \underline{p}(\boldsymbol{y}; q) \;=\; & \tfrac{1}{2}(\nu + p - 1)\log(2\nu) - \tfrac{1}{2}\sum_{i=1}^{m} n_i \log(2\pi) - 4\log(\pi) - \tfrac{p}{2}\log(\sigma^2_\beta) \\
& -\tfrac{1}{\sigma^2_\beta}\left\{\| \boldsymbol{\mu}_{q(\boldsymbol{\beta})} \|^2 + \text{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right)\right\} + \tfrac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}| \\
& +\tfrac{1}{2}\left\{p + L_{\text{gbl}} + m(p + L_{\text{grp}})\right\} - \log\left(C_{p,\nu+p-1}\right) + \log\left(C_{p,\nu+m+p-1}\right) \\
& -\tfrac{1}{2}\left(\nu + p + m - 1\right)\log|\boldsymbol{B}_{q(\boldsymbol{\Sigma})}| + \log\Gamma\left\{\tfrac{1}{2}\left(\sum_{i=1}^{m} n_i + 1\right)\right\} \\
& -\tfrac{1}{2}\left(\sum_{i=1}^{m} n_i + 1\right)\log\left(B_{q(\sigma^2_\varepsilon)}\right) + \log\Gamma\left\{\tfrac{1}{2}\left(L_{\text{gbl}} + 1\right)\right\} \\
& -\tfrac{1}{2}\left(L_{\text{gbl}} + 1\right)\log\left(B_{q\left(\sigma^2_{\text{gbl}}\right)}\right) + \log\Gamma\left\{\tfrac{1}{2}\left(m \times L_{\text{grp}} + 1\right)\right\} \\
& -\tfrac{1}{2}\left(m \times L_{\text{grp}} + 1\right)\log\left(B_{q\left(\sigma^2_{\text{grp}}\right)}\right) - \log\left(A_\varepsilon\right) - \log\left(B_{q(a_\varepsilon)}\right) \\
& +\mu_{q(1/\sigma^2_\varepsilon)}\mu_{q(1/a_\varepsilon)} - \log\left(B_{q(a_{\text{grp}})}\right) + \sum_{j=1}^{p} \nu\left(\boldsymbol{M}_{q\left(\boldsymbol{\Sigma}^{-1}\right)}\right)_{jj} \mu_{q\left(1/a_{\Sigma,j}\right)}. \\
& -\log\left(A_{\text{gbl}}\right) - \log\left(B_{q(a_{\text{gbl}})}\right) + \mu_{q\left(1/\sigma^2_{\text{gbl}}\right)}\mu_{q(1/a_{\text{gbl}})} - \log\left(A_{\text{grp}}\right) \\
& +\mu_{q(1/\sigma^2_{\text{grp}})}\mu_{q(1/\sigma^2_{\text{grp}})} - \sum_{j=1}^{p}\log\left(A_{\Sigma,j}\right) + p\log\Gamma\left\{\tfrac{1}{2}\left(\nu + p\right)\right\} \\
& -\tfrac{1}{2}\left(\nu + p - 1\right)\sum_{j=1}^{p}\log\left(B_{q\left(a_{\Sigma,j}\right)}\right)
\end{aligned}$$

The details of the derivation are as follows:

$$\begin{aligned}
\log \underline{p}\left(\boldsymbol{y}; q\right) &= E_q\left\{\log p(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2, \sigma_{\text{gbl}}^2, \boldsymbol{\Sigma}, \sigma_{\text{grp}}^2, a_\varepsilon, a_{\text{gbl}}, \boldsymbol{a}_\Sigma, a_{\text{grp}})\right. \\
&\qquad \left. - \log q^*\left(\boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2, \sigma_{\text{gbl}}^2, \boldsymbol{\Sigma}, \sigma_{\text{grp}}^2, a_\varepsilon, a_{\text{gbl}}, \boldsymbol{a}_\Sigma, a_{\text{grp}}\right)\right\} \\
&= E_q\left\{\log p\left(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2\right)\right\} \\
&\qquad + E_q\left\{\log p\left(\boldsymbol{\beta}, \boldsymbol{u}|\sigma_{\text{gbl}}^2, \boldsymbol{\Sigma}, \sigma_{\text{grp}}^2\right) - \log q^*\left(\boldsymbol{\beta}, \boldsymbol{u}\right)\right\} \\
&\qquad + E_q\left\{\log p\left(\sigma_\varepsilon^2|a_\varepsilon\right) - \log q^*\left(\sigma_\varepsilon^2\right)\right\} \\
&\qquad + E_q\left\{\log p\left(\sigma_{\text{gbl}}^2|a_{\text{gbl}}\right) - \log q^*\left(\sigma_{\text{gbl}}^2\right)\right\} \\
&\qquad + E_q\left\{\log p\left(\boldsymbol{\Sigma}|a_{\Sigma,1}, a_{\Sigma,2}\right) - \log q^*\left(\boldsymbol{\Sigma}\right)\right\} \\
&\qquad + E_q\left\{\log p\left(\sigma_{\text{grp}}^2|a_{\text{grp}}\right) - \log q^*\left(\sigma_{\text{grp}}^2\right)\right\} \\
&\qquad E_q\left\{\log p\left(a_\varepsilon\right) - \log q^*\left(a_\varepsilon\right)\right\} + E_q\left\{\log p\left(a_{\text{gbl}}\right) - \log q^*\left(a_{\text{gbl}}\right)\right\} \\
&\qquad E_q\left\{\log p\left(\boldsymbol{a}_\Sigma\right) - \log q^*\left(\boldsymbol{a}_\Sigma\right)\right\} + E_q\left\{\log p\left(a_{\text{grp}}\right) - \log q^*\left(a_{\text{grp}}\right)\right\}.
\end{aligned} \tag{5.23}$$

First note that

$$\log p\left(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2\right) = \tfrac{1}{2}\sum_{i=1}^{m} n_i \log(2\pi) - \tfrac{1}{2}\sum_{i=1}^{m} n_i \log(\sigma_\varepsilon^2) - \tfrac{1}{2\sigma_\varepsilon^2} \parallel \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u} \parallel^2$$

Taking expectations we get

$$\begin{aligned}
E_q\left\{\log p\left(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2\right)\right\} &= -\tfrac{1}{2}\sum_{i=1}^{m} n_i \log(2\pi) - \tfrac{1}{2}\sum_{i=1}^{m} n_i E_q\left\{\log(\sigma_\varepsilon^2)\right\} \\
&\qquad -\tfrac{1}{2}\mu_{q(1/\sigma_\varepsilon^2)} E_q\left(\parallel \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u} \parallel^2\right).
\end{aligned} \tag{5.24}$$

Next, using the concatenated form of the design matrices, $\boldsymbol{C} = [\boldsymbol{X} \,|\, \boldsymbol{Z}]$ and Results 1.4.19 and 1.4.12, we see that

$$\begin{aligned}
E_q\left(\parallel \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u} \parallel^2\right) &= E_q\left(\left\| \boldsymbol{y} - \boldsymbol{C}\begin{bmatrix}\boldsymbol{\beta} \\ \boldsymbol{u}\end{bmatrix}\right\|^2\right) \\
&= \left\| \boldsymbol{y} - \boldsymbol{C}\, E_q\left(\begin{bmatrix}\boldsymbol{\beta} \\ \boldsymbol{u}\end{bmatrix}\right)\right\|^2 + \operatorname{tr}\left\{\boldsymbol{C}\operatorname{Cov}\left(\begin{bmatrix}\boldsymbol{\beta} \\ \boldsymbol{u}\end{bmatrix}\right)\boldsymbol{C}^\top\right\} \\
&= \parallel \boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \parallel^2 + \operatorname{tr}\left(\boldsymbol{C}\,\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^\top\right) \\
&= \parallel \boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \parallel^2 + \operatorname{tr}\left(\boldsymbol{C}^\top\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\right).
\end{aligned}$$

Substituting into (5.24), we have

$$\begin{aligned}
E_q\left\{\log p\left(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2\right)\right\} &= -\tfrac{1}{2}\sum_{i=1}^{m} n_i \log(2\pi) - \tfrac{1}{2}\sum_{i=1}^{m} n_i E_q\left\{\log(\sigma_\varepsilon^2)\right\} \\
&\qquad -\tfrac{1}{2}\mu_{q(1/\sigma_\varepsilon^2)}\left\{\parallel \boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \parallel^2 + \operatorname{tr}\left(\boldsymbol{C}^\top\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\right)\right\}.
\end{aligned}$$

Moving on to the next term in (5.23), we see that

$$\begin{aligned}\log p\left(\boldsymbol{\beta}, \boldsymbol{u} | \sigma^2_{\text{gbl}}, \boldsymbol{\Sigma}, \sigma^2_{\text{grp}}\right) = & -\log(2\pi) - \log\left(\sigma^2_\beta\right) - \tfrac{1}{2\sigma^2_\beta} \parallel \boldsymbol{\beta} \parallel^2 \\ & -\tfrac{1}{2}\left\{L_{\text{gbl}} + m(2 + L_{\text{grp}})\right\}\log(2\pi) - \tfrac{L_{\text{gbl}}}{2}\log\left(\sigma^2_{\text{gbl}}\right) \\ & -\tfrac{m}{2}\log|\boldsymbol{\Sigma}| - \tfrac{1}{2}m \times L_{\text{grp}}\log\left(\sigma^2_{\text{grp}}\right) - \tfrac{1}{2\sigma^2_{\text{gbl}}} \parallel \boldsymbol{u}^{\text{gbl}} \parallel^2 \\ & -\tfrac{1}{2}\text{tr}\left(\sum_{i=1}^{m} \boldsymbol{\delta}_i \boldsymbol{\delta}_i^\top \boldsymbol{\Sigma}^{-1}\right) - \tfrac{1}{2\sigma^2_{\text{grp}}} \parallel \boldsymbol{u}^{\text{grp}} \parallel^2 .\end{aligned}$$

Next, note that the entropy of the multivariate normal random vector $\left[\boldsymbol{\beta}^\top, \boldsymbol{u}^\top\right]^\top$ is

$$\begin{aligned}-E_q\left\{\log q^*(\boldsymbol{\beta}, \boldsymbol{u})\right\} &= \tfrac{1}{2}\log\left\{(2\pi e)^{2+L_{\text{gbl}}+m(2+L_{\text{grp}})}|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}|\right\} \\ &= \tfrac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}| + \tfrac{2+L_{\text{gbl}}+m(2+L_{\text{grp}})}{2}\log(2\pi) + \tfrac{2+L_{\text{gbl}}+m(2+L_{\text{grp}})}{2}.\end{aligned}$$

Therefore,

$$\begin{aligned}& E_q\left\{\log p\left(\boldsymbol{\beta}, \boldsymbol{u} | \sigma^2_{\text{gbl}}, \boldsymbol{\Sigma}, \sigma^2_{\text{grp}}\right) - \log q^*(\boldsymbol{\beta}, \boldsymbol{u})\right\} \\ &\quad = -\log(2\pi) - \log\left(\sigma^2_\beta\right) - \tfrac{1}{2\sigma^2_\beta}\left\{\parallel \boldsymbol{\mu}_{q(\boldsymbol{\beta})} \parallel^2 + \text{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right)\right\} \\ &\qquad -\tfrac{1}{2}\left\{L_{\text{gbl}} + m(2 + L_{\text{grp}})\right\}\log(2\pi) - \tfrac{L_{\text{gbl}}}{2}E_q\left\{\log\left(\sigma^2_{\text{gbl}}\right)\right\} \\ &\qquad -\tfrac{m}{2}E_q\left\{\log|\boldsymbol{\Sigma}|\right\} - \tfrac{1}{2}\,m\,L_{\text{grp}}\,E_q\left\{\log\left(\sigma^2_{\text{grp}}\right)\right\} \\ &\qquad -\tfrac{1}{2}\mu_{q\left(1/\sigma^2_{\text{gbl}}\right)}\left\{\parallel \boldsymbol{\mu}_{q(\boldsymbol{u}^{\text{gbl}})} \parallel^2 + \text{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{u}^{\text{gbl}})}\right)\right\} \\ &\qquad -\tfrac{1}{2}\text{tr}\left\{\textstyle\sum_{i=1}^{m}\left(\boldsymbol{\mu}_{q(\boldsymbol{\delta}_i)}\boldsymbol{\mu}^\top_{q(\boldsymbol{\delta}_i)} + \boldsymbol{\Sigma}_{q(\boldsymbol{\delta}_i)}\right)\boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1})}\right\} \\ &\qquad -\tfrac{1}{2}\mu_{q\left(1/\sigma^2_{\text{grp}}\right)}\left\{\parallel \boldsymbol{\mu}_{q(\boldsymbol{u}^{\text{grp}})} \parallel^2 + \text{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{u}^{\text{grp}})}\right)\right\} \\ &\qquad +\tfrac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}| + \tfrac{1}{2}(2 + L_{\text{gbl}} + m(2 + L_{\text{grp}}))\log(2\pi) \\ &\qquad +\tfrac{1}{2}(2 + L_{\text{gbl}} + m(2 + L_{\text{grp}})).\end{aligned}$$

Collecting like terms and removing cancellations, this equates to

$$\begin{aligned}& = -\log\left(\sigma^2_\beta\right) - \tfrac{1}{2\sigma^2_\beta}\left\{\parallel \boldsymbol{\mu}_{q(\boldsymbol{\beta})} \parallel^2 + \text{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right)\right\} - \tfrac{L_{\text{gbl}}}{2}E_q\left\{\log\left(\sigma^2_{\text{gbl}}\right)\right\} \\ &\quad -\tfrac{m}{2}E_q\left\{\log|\boldsymbol{\Sigma}|\right\} - \tfrac{1}{2}\,m\,L_{\text{grp}}\,E_q\left\{\log\left(\sigma^2_{\text{grp}}\right)\right\} \\ &\quad -\tfrac{1}{2}\mu_{q\left(1/\sigma^2_{\text{gbl}}\right)}\left\{\parallel \boldsymbol{\mu}_{q(\boldsymbol{u}^{\text{gbl}})} \parallel^2 + \text{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{u}^{\text{gbl}})}\right)\right\} \\ &\quad -\tfrac{1}{2}\text{tr}\left\{\textstyle\sum_{i=1}^{m}\left(\boldsymbol{\mu}_{q(\boldsymbol{\delta}_i)}\boldsymbol{\mu}^\top_{q(\boldsymbol{\delta}_i)} + \boldsymbol{\Sigma}_{q(\boldsymbol{\delta}_i)}\right)\boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1})}\right\} \\ &\quad -\tfrac{1}{2}\mu_{q\left(1/\sigma^2_{\text{grp}}\right)}\left\{\parallel \boldsymbol{\mu}_{q(\boldsymbol{u}^{\text{grp}})} \parallel^2 + \text{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{u}^{\text{grp}})}\right)\right\} \\ &\quad +\tfrac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}| + \tfrac{1}{2}(2 + L_{\text{gbl}} + m(2 + L_{\text{grp}})).\end{aligned}$$

Next,

$$\begin{aligned}
&\log p(\sigma_\varepsilon^2 \,|\, a_\varepsilon) - \log q^*(\sigma_\varepsilon^2) = \\
&\qquad \log\left\{\frac{(1/a_\varepsilon)^{\frac{1}{2}}}{\Gamma\left(\frac{1}{2}\right)}\left(\sigma_\varepsilon^2\right)^{-\frac{1}{2}-1}\exp\left(-\frac{1/a_\varepsilon}{\sigma_\varepsilon^2}\right)\right\} \\
&\qquad -\log\left\{\frac{B_{q(\sigma_\varepsilon^2)}{}^{\frac{1}{2}\left(\sum_{i=1}^m n_i+1\right)}}{\Gamma\left(\frac{1}{2}\left(\sum_{i=1}^m n_i+1\right)\right)}\left(\sigma_\varepsilon^2\right)^{-\frac{1}{2}\left(\sum_{i=1}^m n_i+1\right)-1}\exp\left(-\frac{B_{q(\sigma_\varepsilon^2)}}{\sigma_\varepsilon^2}\right)\right\} \\
&= -\tfrac{1}{2}\log(a_\varepsilon) - \tfrac{3}{2}\log(\sigma_\varepsilon^2) - (1/a_\varepsilon)/\sigma_\varepsilon^2 - \tfrac{1}{2}\log(\pi) \\
&\quad -\tfrac{1}{2}\left(\sum_{i=1}^m n_i + 1\right)\log\left(B_{q(\sigma_\varepsilon^2)}\right) + \left\{\tfrac{1}{2}\left(\sum_{i=1}^m n_i+1\right)+1\right\}\log(\sigma_\varepsilon^2) \\
&\quad +B_{q(\sigma_\varepsilon^2)}/\sigma_\varepsilon^2 + \log\left\{\Gamma\left(\tfrac{1}{2}\textstyle\sum_{i=1}^m n_i + 1\right)\right\}
\end{aligned}$$

Thus,

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

The derivation for the term

$$E_q\left\{\log p(\sigma_{\text{gbl}}^2 \,|\, a_{\text{gbl}}) - \log q^*(\sigma_{\text{gbl}}^2)\right\}$$

is similar to that for

$$E_q\left\{\log p(\sigma_\varepsilon^2 \,|\, a_\varepsilon) - \log q^*(\sigma_\varepsilon^2)\right\}.$$

So, without going into further detail,

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

We now move on to expanding the fifth term in (5.23):

$$\begin{aligned}
\log p(\boldsymbol{\Sigma} \,|\, a_{\Sigma,1}, a_{\Sigma,2}) - \log q^*(\boldsymbol{\Sigma}) \;=\; & -\log(C_{2,\,\nu+2-1}) + \tfrac{\nu+2-1}{2}\log\left|2\nu\,\text{diag}\left(1/a_{\Sigma,1},\, 1/a_{\Sigma,2}\right)\right| \\
& -\tfrac{\nu+4}{2}\log|\boldsymbol{\Sigma}| - \tfrac{1}{2}\text{tr}\left\{2\nu\,\text{diag}\left(1/a_{\Sigma,1}, 1/a_{\Sigma,2}\right)\boldsymbol{\Sigma}^{-1}\right\} \\
& +\log(C_{2,\,\nu+m+1}) - \tfrac{\nu+m+1}{2}\log|\boldsymbol{B}_{q(\boldsymbol{\Sigma})}| + \tfrac{\nu+4+m}{2}\log|\boldsymbol{\Sigma}| \\
& +\tfrac{1}{2}\text{tr}\left\{\boldsymbol{B}_{q(\boldsymbol{\Sigma})}\,\boldsymbol{\Sigma}^{-1}\right\}
\end{aligned}$$

Therefore,

$$\begin{aligned} E_q\left\{\log p(\boldsymbol{\Sigma}\,|\,a_{\Sigma,1},a_{\Sigma,2})-\log q^*(\boldsymbol{\Sigma})\right\} = \\ &-\log(C_{2,\,\nu+1})+\log(C_{2,\,\nu+m+1}) \\ &+\tfrac{\nu+1}{2}E_q\left\{\log|2\nu\,\mathrm{diag}\,(1/a_{\Sigma,1},\,1/a_{\Sigma,2})\,|\right\} \\ &-\tfrac{\nu+m+1}{2}\log|\boldsymbol{B}_{q(\boldsymbol{\Sigma})}|+\tfrac{m}{2}E_q\left\{\log|\boldsymbol{\Sigma}|\right\} \\ &+\tfrac{1}{2}\mathrm{tr}\left\{\left(\boldsymbol{B}_{q(\boldsymbol{\Sigma})}-2\nu\,\mathrm{diag}\left(\mu_{q(1/a_{\Sigma,1})},\mu_{q(1/a_{\Sigma,2})}\right)\right)\boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1})}\right\}. \end{aligned}$$

The derivation for the term

$$E_q\left\{\log p(\sigma^2_{\mathrm{grp}}\,|\,a_{\mathrm{grp}})-\log q^*(\sigma^2_{\mathrm{grp}})\right\}$$

is similar to that for

$$E_q\left\{\log p(\sigma^2_{\varepsilon}\,|\,a_{\varepsilon})-\log q^*(\sigma^2_{\varepsilon})\right\}.$$

Without going into further detail, it is shown that

$$\begin{aligned} E_q\left\{\log p(\sigma^2_{\mathrm{grp}}\,|\,a_{\mathrm{grp}})-\log q^*(\sigma^2_{\mathrm{grp}})\right\} = &-\tfrac{1}{2}E_q\left\{\log(a_{\mathrm{grp}})\right\}+\tfrac{1}{2}m\,L_{\mathrm{grp}}E_q\left\{\log(\sigma^2_{\mathrm{grp}})\right\} \\ &-\tfrac{1}{2}\log(\pi)+\left(B_{q(\sigma^2_{\mathrm{grp}})}-\mu_{q(1/a_{\mathrm{grp}})}\right)\mu_{q(1/\sigma^2_{\mathrm{grp}})} \\ &+\log\left\{\Gamma\left(\tfrac{1}{2}m\,L_{\mathrm{grp}}+1\right)\right\} \\ &-\tfrac{1}{2}\left(m\,L_{\mathrm{grp}}+1\right)\log\left(B_{q(\sigma^2_{\mathrm{grp}})}\right). \end{aligned}$$

Next, we have

$$\begin{aligned} \log p(a_{\varepsilon})-\log q^*(a_{\varepsilon}) = &\log\left\{\frac{\left(1/A^2_{\varepsilon}\right)^{\frac{1}{2}}}{\Gamma(\frac{1}{2})}\,(a_{\varepsilon})^{-\frac{1}{2}-1}\,\exp\left(-\frac{1/A^2_{\varepsilon}}{a_{\varepsilon}}\right)\right\} \\ &-\log\left\{\frac{B_{q(a_{\varepsilon})}}{\Gamma(1)}\,(a_{\varepsilon})^{-1-1}\,\exp\left(-\frac{B_{q(a_{\varepsilon})}}{a_{\varepsilon}}\right)\right\} \\ = &-\log(A_{\varepsilon})-\tfrac{1}{2}\log(\pi)-\tfrac{3}{2}\log(a_{\varepsilon})-\left(1/A^2_{\varepsilon}\right)/a_{\varepsilon}-\log\left(B_{q(a_{\varepsilon})}\right) \\ &+2\log(a_{\varepsilon})+B_{q(a_{\varepsilon})}/a_{\varepsilon}. \end{aligned}$$

From algorithm 5, we see that

$$B_{q(a_{\varepsilon})} = \mu_{q(1/\sigma^2_{\varepsilon})}+A^{-2}_{\varepsilon},$$

so making use of this and taking expectations we get:

$$\begin{aligned} E_q\left\{\log p(a_{\varepsilon})-\log q^*(a_{\varepsilon})\right\} = &-\log(A_{\varepsilon})-\tfrac{1}{2}\log(\pi)+\tfrac{1}{2}E_q\left\{\log(a_{\varepsilon})\right\}-\log\left(B_{q(a_{\varepsilon})}\right) \\ &+\mu_{q(1/\sigma^2_{\varepsilon})}\mu_{q(1/a_{\varepsilon})} \end{aligned}$$

The derivation of the following term in (5.23) is similar to that just shown above, thus we have

$$E_q\left\{\log p(a_{\text{gbl}}) - \log q^*(a_{\text{gbl}})\right\} = -\log(A_{\text{gbl}}) - \tfrac{1}{2}\log(\pi) + \tfrac{1}{2}E_q\left\{\log(a_{\text{gbl}})\right\} - \log\left(B_{q(a_{\text{gbl}})}\right) + \mu_{q(1/\sigma^2_{\text{gbl}})}\mu_{q(1/a_{\text{gbl}})}.$$

Next,

$$\log p(\boldsymbol{a}_\Sigma) - \log q^*(\boldsymbol{a}_\Sigma) = \log\left\{\prod_{j=1}^{2}\frac{\left(1/A^2_{\Sigma,j}\right)^{\frac{1}{2}}}{\Gamma\left(\frac{1}{2}\right)}\left(a_{\Sigma,j}\right)^{-\frac{1}{2}-1}\exp\left(-\frac{1/A^2_{\Sigma,j}}{a_{\Sigma,j}}\right)\right\} - \log\left\{\prod_{j=1}^{2}\frac{B_{q(a_{\Sigma,j})}{}^{\frac{\nu}{2}+1}}{\Gamma\left(\frac{\nu}{2}+1\right)}\left(a_{\Sigma,j}\right)^{-\left(\frac{\nu}{2}+1\right)-1}\exp\left(-\frac{B_{q(a_{\Sigma,j})}}{a_{\Sigma,j}}\right)\right\}$$

$$= -\sum_{j=1}^{2}\log\left(A_{\Sigma,j}\right) - \log(\pi) + 2\log\left\{\Gamma\left(\tfrac{\nu}{2}+1\right)\right\} + \tfrac{1}{2}(\nu+1)\sum_{j=1}^{2}\log(a_{\Sigma,j}) - \sum_{j=1}^{2}\left(1/A^2_{\Sigma,j}\right)/a_{\Sigma,j} + \sum_{j=1}^{2}B_{q(a_{\Sigma,j})}/a_{\Sigma,j} - \left(\tfrac{\nu}{2}+1\right)\sum_{j=1}^{2}\log\left(B_{q(a_{\Sigma,j})}\right).$$

Using the fact that

$$B_{q(a_{\Sigma,j})} = \nu\left(\boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1})}\right)_{jj} + 1/A^2_{\Sigma,j},$$

and taking expectations, we get

$$E_q\left\{\log p(\boldsymbol{a}_\Sigma) - \log q^*(\boldsymbol{a}_\Sigma)\right\} = -\sum_{j=1}^{2}\log\left(A_{\Sigma,j}\right) - \log(\pi) + 2\log\left\{\Gamma\left(\tfrac{\nu}{2}+1\right)\right\} + \tfrac{1}{2}(\nu+1)\sum_{j=1}^{2}E_q\left\{\log(a_{\Sigma,j})\right\} - \left(\tfrac{\nu}{2}+1\right)\sum_{j=1}^{2}\log\left(B_{q(a_{\Sigma,j})}\right) + \sum_{j=1}^{2}\nu\left(\boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1})}\right)_{jj}\mu_{q(1/a_{\Sigma,j})}.$$

The derivation for the last term in (5.23) is similar to the derivation for the a_ε parameter. Thus, the expression is:

$$E_q\left\{\log p(a_{\text{grp}}) - \log q^*(a_{\text{grp}})\right\} = -\log(A_{\text{grp}}) - \tfrac{1}{2}\log(\pi) + \tfrac{1}{2}E_q\left\{\log(a_{\text{grp}})\right\} - \log\left(B_{q(a_{\text{grp}})}\right) + \mu_{q(1/\sigma^2_{\text{grp}})}\mu_{q(1/a_{\text{grp}})}$$

Bringing all these terms together, and noting the cancellations:

$$-\tfrac{1}{2}\mu_{q(1/\sigma^2_\varepsilon)}\left\{\|\,\boldsymbol{y} - \boldsymbol{C}\,\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\,\|^2 + \text{tr}\left(\boldsymbol{C}^\top\boldsymbol{C}\,\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\right)\right\} = \left(B_{q(\sigma^2_\varepsilon)} - \mu_{q(1/a_\varepsilon)}\right)\mu_{q(1/\sigma^2_\varepsilon)},$$

$$-\tfrac{1}{2}\mu_{q\left(1/\sigma^2_{\text{gbl}}\right)}\left\{\| \boldsymbol{\mu}_{q(\boldsymbol{u}^{\text{gbl}})} \|^2 + \text{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{u}^{\text{gbl}})}\right)\right\} = \left(B_{q(\sigma^2_{\text{gbl}})} - \mu_{q(1/a_{\text{gbl}})}\right)\mu_{q(1/\sigma^2_{\text{gbl}})},$$

$$\begin{aligned}-\tfrac{1}{2}\text{tr}\left\{\textstyle\sum_{i=1}^{m}\left(\boldsymbol{\mu}_{q(\boldsymbol{\delta}_i)}\boldsymbol{\mu}^{\top}_{q(\boldsymbol{\delta}_i)} + \boldsymbol{\Sigma}_{q(\boldsymbol{\delta}_i)}\right)\boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1})}\right\} \\ = \tfrac{1}{2}\text{tr}\left\{\left(\boldsymbol{B}_{q(\boldsymbol{\Sigma})} - 2\nu\,\text{diag}\left(\mu_{q\left(1/a_{\Sigma,1}\right)}, \mu_{q\left(1/a_{\Sigma,2}\right)}\right)\right)\boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1})}\right\},\end{aligned}$$

$$-\tfrac{1}{2}\mu_{q\left(1/\sigma^2_{\text{grp}}\right)}\left\{\| \boldsymbol{\mu}_{q(\boldsymbol{u}^{\text{grp}})} \|^2 + \text{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{u}^{\text{grp}})}\right)\right\} = \left(B_{q(\sigma^2_{\text{grp}})} - \mu_{q(1/a_{\text{grp}})}\right)\mu_{q(1/\sigma^2_{\text{grp}})}$$

and the fact that

$$\tfrac{\nu+1}{2}E_q\left\{\log|2\nu\,\text{diag}\left(1/a_{\Sigma,1}, 1/a_{\Sigma,2}\right)|\right\} + \tfrac{\nu+1}{2}\textstyle\sum_{j=1}^{2}E_q\left\{\log(a_{\Sigma,j})\right\} = \tfrac{\nu+1}{2}\log(2\nu),$$

gives the final lower bound expression in (5.13) where $p = 2$ is the column length in $\boldsymbol{X}$.

Chapter 6

Mean field variational Bayes for mixture models in measurement error problems

6.1 Introduction

In the area of epidemiology it is generally recognised that covariates, such as risk factors, are not being measured accurately on all subjects. This can heavily affect the estimate of the relationship between these covariates and the outcome. These are known as *measurement error* or *errors-in-variables* problems. In this chapter we examine the Bayesian formulation of such a problem, in particular fast mean field variational Bayes approximation and compare this to the MCMC benchmark. The field of epidemiology has lent great motivation to this chapter, as measurement error has long been a problem there. The model we present, inspired by the model used in Richardson *et al.* (2002), can be applied in a variety of settings involving measurement error.

The main concern for problems with measurement error is to make inference on the relationship between a response $\boldsymbol{y}$ and covariates $\boldsymbol{x}$, in situations where the only other information we have on $\boldsymbol{x}$ is through the recording of an error contaminated surrogate $\boldsymbol{o}$, as $\boldsymbol{x}$ may not have been measured accurately on all subjects. An attempt to regress $\boldsymbol{y}$ on $\boldsymbol{o}$ whilst ignoring the problem of measurement error can in most cases lead to extremely inaccurate inferential statistics. An extensive literature has formed for flexible estimation of regression models where precise measurements on the predictor/s are not available. An

extensive reference to the area is Carroll *et al.* (2006). The 2000s in particular saw an influx of research into the area, e.g. Mallick *et al.* (2002), Liang *et al.* (2003), Carroll *et al.* (2004), Ganguli *et al.* (2005) and Carroll *et al.* (2007). Many contributions have discussed the use of MCMC based inference to fitting models impaired by measurement error, however inference based on MCMC can at times be extremely time intensive. As an alternative to MCMC, we focus on MFVB fitting to the model structure given in Richardson *et al.* (2002) where the response follows a normal distribution.

Section 6.2 discusses the different components necessary for the measurement error model with normal mixture components. In Section 6.3 we present the full measurement error model that we deal with. We present the MFVB algorithm in Section 6.4 and Section 6.5 provides a simulation study based on the new MFVB algorithm. The appendices provide details on the calculations required for the MFVB algorithm and corresponding lower bound expression.

6.2 Model structure

We set up the covariate of interest $\boldsymbol{x}$ by splitting it up into two subgroups: $\boldsymbol{x}_{\text{obs}}$ which involves the accurate measurements of the covariate; and $\boldsymbol{x}_{\text{unobs}}$ which involves the part of the covariate that has not been measured accurately. The observations in $\boldsymbol{x}_{\text{obs}}$ are generally recorded using a *gold standard method*, which is an error-free method for measuring the covariate, however this is costly to use on a large scale, hence $\boldsymbol{x}_{\text{unobs}}$ is usually much larger than $\boldsymbol{x}_{\text{obs}}$. The subjects belonging to $\boldsymbol{x}_{\text{obs}}$ are usually referred to as the *validation group*, and the subjects belonging to $\boldsymbol{x}_{\text{unobs}}$ as the *main study*. This is in keeping with the terminology used in Richardson *et al.* (2002).

Throughout this chapter we let i refer to a particular individual, where $1 \leq i \leq n$. Also $\boldsymbol{y}$ denotes the known outcome (e.g. disease status), $\boldsymbol{x}$ the true covariate comprising observed quantities $\boldsymbol{x}_{\text{obs}}$ and unobserved quantities $\boldsymbol{x}_{\text{unobs}}$, $\boldsymbol{o}$ the observed surrogate for $\boldsymbol{x}$ and $\boldsymbol{c}$ denotes a known covariate. We focus on the case where $\boldsymbol{x}$ and $\boldsymbol{o}$ are univariate. The general model is divided into three submodels:

1. a *regression model* expressing the relationship between the covariates $\boldsymbol{c}$ and $\boldsymbol{x}$ and the outcome $\boldsymbol{y}$, denoted by $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\beta})$;

2. a *measurement model*, which expresses the relationship between the surrogate mea-

sures $\boldsymbol{o}$ and the true unknown covariate $\boldsymbol{x}$, denoted by $p(\boldsymbol{o}|\boldsymbol{x}, \boldsymbol{\alpha})$; and

3. the *prior model* for $\boldsymbol{x}$, which specifies the distribution of the unknown covariate $\boldsymbol{x}$ in the general population and is denoted by $p(\boldsymbol{x}\,|\,\boldsymbol{\theta})$,

where $\boldsymbol{\beta}$ is the regression coefficient vector for $\boldsymbol{y}$, $\boldsymbol{\alpha}$ is the regression coefficient vector for $\boldsymbol{o}$ and $\boldsymbol{\theta}$ are the parameters in the prior model for $\boldsymbol{x}$.

6.2.1 Nondifferential measurement error

An important assumption for our model is one of nondifferential measurement error which holds when $\boldsymbol{o}$ has no information about $\boldsymbol{y}$ other than what is given in $\boldsymbol{x}$ and $\boldsymbol{c}$. Thus, if $\boldsymbol{o}$ is conditionally independent of $\boldsymbol{y}$ given $\boldsymbol{x}$ and $\boldsymbol{c}$, then it is a surrogate. This can be expressed as

$$p(\boldsymbol{y}\,|\,\boldsymbol{o},\, \boldsymbol{x},\, \boldsymbol{c},\, \boldsymbol{\beta}) = p(\boldsymbol{y}\,|\,\boldsymbol{x},\, \boldsymbol{c},\, \boldsymbol{\beta}) \qquad \text{or} \qquad \boldsymbol{y} \perp \boldsymbol{o}\,|\,\boldsymbol{x}.$$

This leads to the joint distribution

$$p(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{x}, \boldsymbol{o}, \boldsymbol{y}) = p(\boldsymbol{\theta})\, p(\boldsymbol{\alpha})\, p(\boldsymbol{\beta}) \prod_{i=1}^{n} p(x_i\,|\,\boldsymbol{\theta}) \prod_{i=1}^{n} p(o_i\,|\,x_i,\, \boldsymbol{\alpha}) \prod_{i=1}^{n} p(y_i\,|\,x_i,\, c_i,\, \boldsymbol{\beta}), \quad (6.1)$$

where $\boldsymbol{\theta}$ denotes all the parameters of the normal mixture model and $p(\boldsymbol{\beta})$, $p(\boldsymbol{\alpha})$ and $p(\boldsymbol{\theta})$ are the prior distributions for the parameters in the three submodels.

6.2.2 Finite normal mixture component

A question that has been heavily researched in the last decade is that of how to model $p(\boldsymbol{x}\,|\,\boldsymbol{\theta})$ in such that we allow the data to influence its shape, since regression parameters are known to be susceptible to the particular shape of $p(\boldsymbol{x}\,|\,\boldsymbol{\theta})$. Many contributions have been made to this particular question in the measurement error context, some of which are described in Carroll *et al.* (1993), Roeder *et al.* (1996), Schafer (2001), Aitkin & Rocci (2002) and Müller & Roeder (1997). Richardson *et al.* (2002) however, suggest mixture models with alterable number of components as a natural alternative. In particular, mixtures of normal distributions, as they make model misspecification for $p(\boldsymbol{x}\,|\,\boldsymbol{\theta})$ increasingly more robust. Thus, we also impose a mixture of normal distributions on $p(\boldsymbol{x}\,|\,\boldsymbol{\theta})$ of the form:

$$\begin{gathered} p(x_i\,|\,a^x_{i1}, \cdots, a^x_{iK}) = \prod_{k=1}^{K} \left[(2\pi\sigma_k^2)^{-1/2} \exp\left\{ -\tfrac{1}{2}(x_i - \mu_k)^2/\sigma_k^2 \right\} \right]^{a^x_{ik}}, \\ (a^x_{i1}, \cdots, a^x_{iK})|(\omega_1, \cdots, \omega_K) \overset{\text{ind.}}{\sim} \text{Multinomial}(1;\, \omega_1, \cdots, \omega_K), \quad 1 \le i \le n, \end{gathered} \quad (6.2)$$

where the a_{ik} are auxiliary variables introduced to make the mixture model tractable and K refers to the number of mixture components used. An example of this use of auxiliary variables can be found in Section 2.2.4 of Ormerod & Wand (2010). The prior distributions have the following forms:

$$\begin{gathered}\mu_k \overset{\text{ind.}}{\sim} \mathrm{N}\left(\mu_{\mu_k}, \sigma^2_{\mu_k}\right), \quad \sigma_k \overset{\text{ind.}}{\sim} \text{Half-Cauchy}\left(A_k\right), \quad 1 \le k \le K, \\ (\omega_1, \cdots, \omega_K) \sim \text{Dirichlet}(\alpha, \cdots, \alpha), \quad \alpha > 0,\end{gathered} \tag{6.3}$$

where according to this notation, $\sum_{k=1}^{K} a^x_{ik} = 1$, indicating which mixture component the current observation belongs to, and $\omega_k = P(a^x_{ik} = 1)$.

6.2.3 Joint model

Bringing the distributions in (6.1), (6.2) and (6.3) together and assuming natural conditional independencies, our joint model for the measurement error problem is expressed as

$$\begin{aligned}&\prod_{i=1}^{n}\left\{p(y_i \,|\, x_i, \boldsymbol{c}_i, \boldsymbol{\beta})\right\}\prod_{i=1}^{n}\left\{p(o_i \,|\, x_i, \boldsymbol{\alpha})\right\}\prod_{i=1}^{n}\prod_{k=1}^{K}\left\{p(x_i \,|\, a_{ik}, \mu_k, \sigma^2_k)\right\} \\ &\qquad\times\prod_{i=1}^{n}\left\{p(a^x_{i1}, \cdots, a^x_{ik})\right\}p(\omega_1, \cdots, \omega_K)\prod_{k=1}^{K}\left\{p(\mu_k)\,p(\sigma_k)\right\}p(\boldsymbol{\beta})p(\boldsymbol{\alpha}).\end{aligned}$$

The model components discussed in the previous sections are brought together next to present the full model, considered in a measurement error setting.

6.3 The full model

We consider the model given in (6.4) which exploits the structure described in Section 6.2.

$$\left.\begin{array}{l} y_i \,|\, x_i,\, \boldsymbol{\beta},\, \sigma_\varepsilon^2 \overset{\text{ind.}}{\sim} \mathrm{N}\left\{(\boldsymbol{X}\boldsymbol{\beta})_i,\, \sigma_\varepsilon^2\right\}, \quad \sigma_\varepsilon^2 \,|\, a_\varepsilon \sim \text{Inverse-Gamma}\left(\tfrac{1}{2},\, 1/a_\varepsilon\right), \\ a_\varepsilon \sim \text{Inverse-Gamma}\left(\tfrac{1}{2},\, 1/A_\varepsilon^2\right), \\ \boldsymbol{\beta} \sim \mathrm{N}\left(\mathbf{0},\, \sigma_\beta^2 \boldsymbol{I}_p\right), \end{array}\right\} \textit{regression model}$$

$$\left.\begin{array}{l} o_i \,|\, x_i,\, \boldsymbol{\alpha},\, \sigma_o^2 \overset{\text{ind.}}{\sim} \mathrm{N}\left\{\left(\tilde{\boldsymbol{X}}\,\boldsymbol{\alpha}\right)_i,\, \sigma_o^2\right\}, \quad \boldsymbol{\alpha} \sim \mathrm{N}\left(\mathbf{0},\, \sigma_\alpha^2\, \boldsymbol{I}_q\right), \\ \sigma_o^2 \,|\, a_o \sim \text{Inverse-Gamma}\left(\tfrac{1}{2},\, 1/a_o\right), \quad a_o \sim \text{Inverse-Gamma}\left(\tfrac{1}{2},\, 1/A_o^2\right), \end{array}\right\} \begin{array}{c}\textit{measurement}\\ \textit{error}\\ \textit{model}\end{array} \tag{6.4}$$

$$\left.\begin{array}{l} x_i \,|\, \boldsymbol{a}_i,\, \boldsymbol{\mu}^x,\, (\boldsymbol{\sigma}^x)^2 = \displaystyle\prod_{k=1}^{K}\left[\left\{2\pi\,(\sigma_k^x)^2\right\}^{-1/2} \exp\left\{-\tfrac{1}{2}(x_i-\mu_k^x)^2/\,(\sigma_k^x)^2\right\}\right]^{a_{ik}}, \\ a_{i1},\cdots,a_{iK} \sim \text{Multinomial}\,(1;\, \omega_1,\cdots,\omega_K), \quad \omega_1,\cdots,\omega_K \sim \text{Dirichlet}\,(\alpha,\cdots,\alpha), \\ \mu_k^x \sim \mathrm{N}\left(\mu_\mu,\, \sigma_\mu^2\right), \quad \left(\sigma_k^x\right)^2 \,|\, a_k^x \sim \text{Inverse-Gamma}\left(\tfrac{1}{2},\, 1/a_k^x\right), \\ a_k^x \sim \text{Inverse-Gamma}\left(\tfrac{1}{2},\, 1/\left(A_k^x\right)^2\right), \quad 1 \le k \le K, \end{array}\right\} \begin{array}{c}\textit{prior}\\ \textit{specification}\end{array}$$

where $\boldsymbol{a}$ is an $n \times K$ matrix of auxiliary variables and $\boldsymbol{a}_i$ is the ith column of $\boldsymbol{a}$ with entries $a_{i1}, \ldots, a_{iK}$. In addition, $\boldsymbol{\mu}^x \equiv (\mu_1^x, \ldots, \mu_K^x)$ and $(\boldsymbol{\sigma}^x)^2 \equiv \left\{(\sigma_1^x)^2, \ldots, (\sigma_K^x)^2\right\}$. We also define

$$\boldsymbol{\beta} \equiv \begin{bmatrix} \beta_0 \\ \beta_{\boldsymbol{c}} \\ \beta_{\boldsymbol{x}} \end{bmatrix}, \quad \boldsymbol{\alpha} \equiv \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix}, \quad \boldsymbol{X} \equiv \begin{bmatrix} 1 & c_1 & x_1 \\ \vdots & \vdots & \vdots \\ 1 & c_n & x_n \end{bmatrix}, \quad \tilde{\boldsymbol{X}} \equiv \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}.$$

Here $\boldsymbol{y} = (y_1, \cdots, y_n)$ is an $n \times 1$ vector of response variables, $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are the $p \times 1$ and $q \times 1$ vectors of fixed effects, where $p = 3$ and $q = 2$, $\boldsymbol{X}$ and $\widetilde{\boldsymbol{X}}$ are corresponding design matrices. Also, σ_ε^2 and σ_o^2 are the variance parameters corresponding to the fixed effects vector and a_ε and a_o are the corresponding auxiliary variables as represented in Result 1.4.6. The variables in the *prior specification* of (6.4) are as described in Section 6.2.2.

We define the following additional notation which is useful for the remaining sections of this chapter. Let n_{obs} denote the number of observed x_i measurements and n_{unobs} denote the number of unobserved x_i measurements. We define $\boldsymbol{x}_{\text{obs}}$ to be the $n_{\text{obs}} \times 1$ vector containing the observed x_i measurements and $\boldsymbol{x}_{\text{unobs}}$ to be the $n_{\text{unobs}} \times 1$ vector containing the unobserved x_i measurements. In addition, let $\boldsymbol{y}_{x\text{obs}}$ and $\boldsymbol{y}_{x\text{unobs}}$ be the $n_{\text{obs}} \times 1$ and $n_{\text{unobs}} \times 1$ vectors corresponding to the measurements in $\boldsymbol{x}_{\text{obs}}$ and $\boldsymbol{x}_{\text{unobs}}$. $\boldsymbol{c}_{x\text{obs}}$, and $\boldsymbol{c}_{x\text{unobs}}$

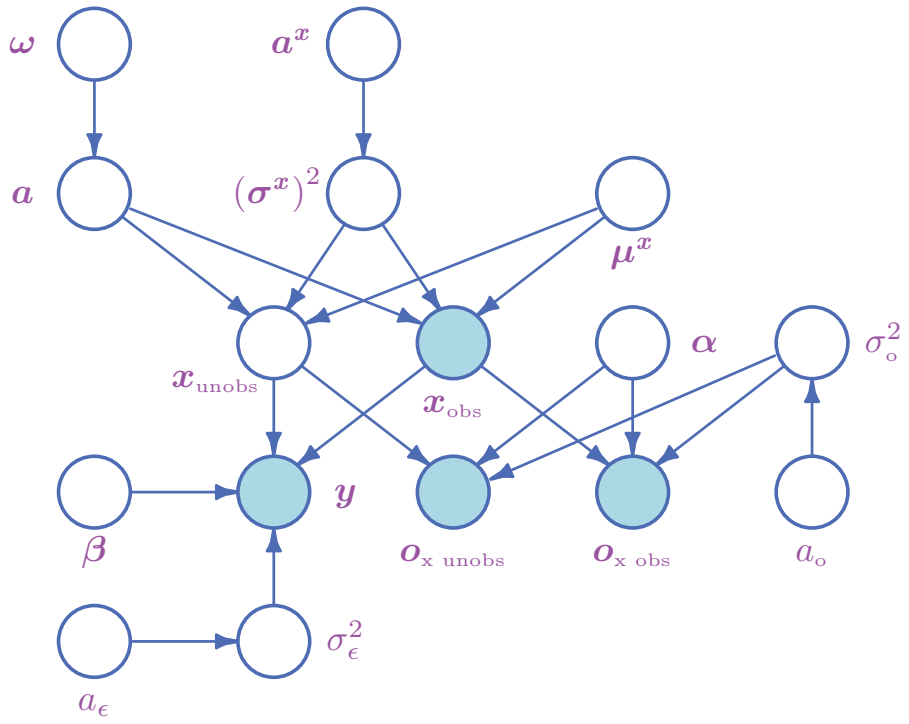


Figure 6.1: *DAG corresponding to the model in (6.4). The shaded nodes correspond to observed data and the open nodes correspond to hidden or latent variables.*

are defined similarly. It is easier to work with a re-ordered set of variables, that is

$$\boldsymbol{y} \equiv \begin{bmatrix} \boldsymbol{y}_{x\text{obs}} \\ \boldsymbol{y}_{x\text{unobs}} \end{bmatrix}, \quad \boldsymbol{c} \equiv \begin{bmatrix} \boldsymbol{c}_{x\text{obs}} \\ \boldsymbol{c}_{x\text{unobs}} \end{bmatrix}, \quad \text{and} \quad \boldsymbol{x} \equiv \begin{bmatrix} \boldsymbol{x}_{\text{obs}} \\ \boldsymbol{x}_{\text{unobs}} \end{bmatrix}.$$

The conditional dependence structure of (6.4) can be visualised in the DAG given in Figure 6.1. This DAG illustrates the relationship between the regression parameters and the unobserved parameters in (6.4). Lastly, we note that the known covariate vector $\boldsymbol{c}$ is omitted from the DAG since it does not require inference.

6.4 Mean field variational Bayes

We provide detail on MFVB fitting catered to the Normal response regression model setting with measurement error as described in model (6.4). The assumption behind MFVB is that we impose a product restriction on the joint posterior distribution of our

model parameters. For instance, we may impose the product factorisation

$$\begin{aligned} p\left\{\boldsymbol{x}_{\text{unobs}},\, \boldsymbol{\beta},\, \boldsymbol{\alpha},\, \boldsymbol{\mu}^{\boldsymbol{x}},\, \boldsymbol{a}^{\boldsymbol{x}},\, \boldsymbol{a},\, \boldsymbol{\omega},\, a_\varepsilon,\, a_o,\, \sigma_\varepsilon^2,\, \sigma_o^2,\, (\boldsymbol{\sigma}^{\boldsymbol{x}})^2 \,|\, \boldsymbol{y},\, \boldsymbol{x}_{\text{obs}},\, \boldsymbol{o}\right\} \\ \approx q\left\{\boldsymbol{x}_{\text{unobs}},\, \boldsymbol{\beta},\, \boldsymbol{\alpha},\, \boldsymbol{\mu}^{\boldsymbol{x}},\, \boldsymbol{a}^{\boldsymbol{x}},\, \boldsymbol{a},\, \boldsymbol{\omega},\, a_\varepsilon,\, a_o,\, \sigma_\varepsilon^2,\, \sigma_o^2,\, (\boldsymbol{\sigma}^{\boldsymbol{x}})^2\right\} \\ = q(\boldsymbol{x}_{\text{unobs}},\, \boldsymbol{\beta},\, \boldsymbol{\alpha},\, \boldsymbol{\mu}^{\boldsymbol{x}},\, \boldsymbol{a}^{\boldsymbol{x}},\, \boldsymbol{a},\, \boldsymbol{\omega},\, a_\varepsilon,\, a_o)\; q\left\{\sigma_\varepsilon^2,\, \sigma_o^2,\, (\boldsymbol{\sigma}^{\boldsymbol{x}})^2\right\}. \end{aligned} \tag{6.5}$$

Using the concept of moralisation (see Definition 1.8.9) we can further reduce the product restriction in (6.5) to have the form

$$\begin{aligned} q(\boldsymbol{x}_{\text{unobs}})\, q(\boldsymbol{\beta})\, q(\boldsymbol{\alpha}) \left\{\prod_{k=1}^{K} q(\mu_k^{\boldsymbol{x}})\, q(a_k^{\boldsymbol{x}})\right\} \left\{\prod_{i=1}^{n} q(a_{i1}, \ldots, a_{iK})\right\} q(\omega_1, \ldots, \omega_K) \\ \times q(a_\varepsilon)\, q(a_o)\, q(\sigma_\varepsilon^2)\, q(\sigma_o^2) \left[\prod_{k=1}^{K} q\left\{(\sigma_k^{\boldsymbol{x}})^2\right\}\right]. \end{aligned} \tag{6.6}$$

In Appendix 6.B, we show that the optimal q-densities for the parameters in (6.6) are:

$q^*(\boldsymbol{x}_{\text{unobs}})$ is the $N\left(\boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs}})},\, \sigma^2_{q(\boldsymbol{x}_{\text{unobs}})} \boldsymbol{I}_{n_{\text{unobs}}}\right)$ density function,

$q^*(\boldsymbol{\beta})$ is the $N\left(\boldsymbol{\mu}_{q(\boldsymbol{\beta})},\, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right)$ density function,

$q^*(\boldsymbol{\alpha})$ is the $N\left(\boldsymbol{\mu}_{q(\boldsymbol{\alpha})},\, \boldsymbol{\Sigma}_{q(\boldsymbol{\alpha})}\right)$ density function,

$q^*(\boldsymbol{\mu}^x) \equiv \prod_{k=1}^{K} q(\mu_k^x)$ is the product of $N\left(\mu_{q(\mu_k^x)},\, \sigma^2_{q(\mu_k^x)}\right)$ density functions,

$q^*(\boldsymbol{a}^x) \equiv \prod_{k=1}^{K} q(a_k^x)$ is the product of Inverse-Gamma$\left(1,\, B_{q(a_k^x)}\right)$ density functions,

$q^*(\boldsymbol{a}) \equiv \prod_{i=1}^{n} q(a_{i1}, \ldots, a_{iK})$ is the product of Multinomial$\left(1,\, \mu_{q(a_{i1})}, \ldots, \mu_{q(a_{iK})}\right)$ density functions,

$q^*(\omega_1, \ldots, \omega_K)$ is the Dirichlet$\left(\alpha_{q(\omega_1)}, \ldots, \alpha_{q(\omega_K)}\right)$ density function,

$q^*(a_\varepsilon)$ is the Inverse-Gamma$\left(1,\, B_{q(a_\varepsilon)}\right)$ density function,

$q^*(a_o)$ is the Inverse-Gamma$\left(1,\, B_{q(a_o)}\right)$ density function,

$q^*(\sigma_\varepsilon^2)$ is the Inverse-Gamma$\left(\frac{1}{2}(n+1),\, B_{q(\sigma_\varepsilon^2)}\right)$ density function,

$q^*(\sigma_o^2)$ is the Inverse-Gamma$\left(\frac{1}{2}(n+1),\, B_{q(\sigma_o^2)}\right)$ density function and

$q^*\left\{(\boldsymbol{\sigma}^x)^2\right\} \equiv q^*\left\{\left(\sigma_k^x\right)^2, \ldots, \left(\sigma_k^x\right)^2\right\}$ is the product of Inverse-Gamma$\left(A_{\left\{(\sigma_k^x)^2\right\}},\, B_{\left\{(\sigma_k^x)^2\right\}}\right)$ density functions.

The optimal parameters in these q-densities possess interdependencies between each other, resulting in an iterative scheme for their solution. This is encompassed in Algorithm 9. The lower bound on the marginal log-likelihood for model (6.4) is given by (the derivation

of which is given in Appendix 6.C):

$$
\begin{aligned}
&\log \underline{p}(\boldsymbol{y}; q) = \\
&\quad -\tfrac{1}{2}(K+5)\log(\pi) + \tfrac{1}{2}(p+q+K+n_{\text{unobs}}) - \tfrac{p}{2}\log\left(\sigma_\beta^2\right) + \tfrac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}| \\
&\quad -\log(A_\varepsilon) - \tfrac{1}{2}\left(2n + n_{\text{unobs}} + \sum_{k=1}^{K}\sum_{i=1}^{n}\mu_{q(a_{ik})}\right)\log(2\pi) + \tfrac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\alpha})}| \\
&\quad -\tfrac{1}{2\sigma_\beta^2}\left\{\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \operatorname{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right)\right\} + \log\Gamma\left\{\tfrac{1}{2}(n+1)\right\} - \tfrac{q}{2}\log\left(\sigma_\alpha^2\right) \\
&\quad -\tfrac{1}{2}(n+1)\log\left(B_{q(\sigma_\varepsilon^2)}\right) - \log\left(B_{q(a_\varepsilon)}\right) + \mu_{q(1/\sigma_\varepsilon^2)}\mu_{q(1/a_\varepsilon)} - \log(A_o) \\
&\quad -\tfrac{1}{2\sigma_\alpha^2}\left\{\|\boldsymbol{\mu}_{q(\boldsymbol{\alpha})}\|^2 + \operatorname{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\alpha})}\right)\right\} - \sum_{k=1}^{K}\log(A_k^x) + \log\Gamma\left\{\tfrac{1}{2}(n+1)\right\} \\
&\quad -\tfrac{K}{2}\log\left(\sigma_\mu^2\right) - \tfrac{1}{2}(n+1)\log\left(B_{q(\sigma_o^2)}\right) - \log\left(B_{q(a_o)}\right) + \mu_{q(1/\sigma_o^2)}\mu_{q(1/a_o)} \\
&\quad +\tfrac{n_{\text{unobs}}}{2}\log\left(\sigma^2_{q(\boldsymbol{x}_{\text{unobs}})}\right) - \tfrac{1}{2\sigma_\mu^2}\sum_{k=1}^{K}\left\{\left(\mu_{q(\mu_k^x)} - \mu_\mu\right)^2 + \sigma^2_{q(\mu_k^x)}\right\} + \tfrac{1}{2}\sum_{k=1}^{K}\log\left(\sigma^2_{q(\mu_k^x)}\right) \\
&\quad +\sum_{k=1}^{K}\log\Gamma\left\{\tfrac{1}{2}\left(\mu_{q(a_{\bullet k})}+1\right)\right\} - \tfrac{1}{2}\sum_{k=1}^{K}\left(\mu_{q(a_{\bullet k})}+1\right)\log\left(B_{q\left\{(\sigma_k^x)^2\right\}}\right) \\
&\quad +\sum_{k=1}^{K}\mu_{q\left\{1/(\sigma_k^x)^2\right\}}\mu_{q\left(1/a_k^x\right)} - \sum_{i=1}^{n}\sum_{k=1}^{K}\log\left(\mu_{q(a_{ik})}\right)\mu_{q(a_{ik})} - \log\Gamma\left(\sum_{k=1}^{K}\alpha_{q(\omega_k)}\right) \\
&\quad +\log\Gamma\left(\sum_{k=1}^{K}\alpha_k\right) - \sum_{k=1}^{K}\log\Gamma(\alpha_k) + \sum_{k=1}^{K}\log\Gamma\left(\alpha_{q(\omega_k)}\right) - \sum_{k=1}^{K}\log\left(B_{q\left(a_k^x\right)}\right).
\end{aligned}
\tag{6.7}
$$

6.5 Simulation study

We conducted a simulation study for the measurement error model given in (6.4), where each ith response measurement takes on a Normal distribution and we make use of a normal mixture model with $K = 2$ mixture components for the prior model on $\boldsymbol{x}$. The true parameter values were specified as

$$
\begin{gathered}
\beta_0 = -0.18, \quad \beta_c = 0.44, \quad \beta_x = 0.79, \quad \sigma_\varepsilon^2 = 0.2, \quad \alpha_o = 0.24, \quad \alpha_1 = 0.83, \quad \sigma_o^2 = 0.33, \\
\omega_1 = 0.25, \quad \omega_2 = 0.75, \quad \mu_1^x = -1.3, \quad \mu_2^x = 1.25, \quad \left(\sigma_1^x\right)^2 = 0.45, \quad \left(\sigma_2^x\right)^2 = 0.45,
\end{gathered}
$$

and the proportion of observed x's, p_{obs}, was varied using $p_{\text{obs}} \in \{0.3, 0.5, 0.7, 0.9\}$. We generated 25 datasets for each of these settings. Algorithm 9 was implemented in the R programming language and MCMC was used through the BUGS inference engine. These computations were performed on a laptop computer (Mac OS X; 2.8 GHz processor, 16 GBytes of random access memory). Figure 6.2 shows MFVB accuracy summaries for different types of p_{obs} values. The parameters that are monitored are σ_ε^2, σ_o^2 and the

Set up initial values:

$\mu_{q(1/\sigma_o^2)}$, $\mu_{q(1/a_o)}$, $\mu_{q(a_{\bullet k})}$, $\mu_{q\left(1/(\sigma_k^x)^2\right)}$, $\mu_{q(\mu_k^x)}$, $\sigma^2_{q(\mu_k^x)}$, $\alpha_{q(\omega_k)}$, $A_{q\left((\sigma_k^x)^2\right)}$, $A_{q\left((\sigma_k^x)^2\right)} > 0$,

$\mu_{q(1/\sigma_\varepsilon^2)}, \mu_{q(1/a_\varepsilon)} > 0$.

Cycle through:

Update the expectations given in Appendix 6.A.

Regression model updates:

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} \leftarrow \left\{\mu_{q(1/\sigma_\varepsilon^2)}\, E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{X}^\top\boldsymbol{X}\right) + \tfrac{1}{\sigma_\beta^2}\boldsymbol{I}_p\right\}^{-1}$$

$$\boldsymbol{\mu}_{q(\boldsymbol{\beta})} \leftarrow \mu_{q(1/\sigma_\varepsilon^2)}\, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\, E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{X}\right)^\top \boldsymbol{y}$$

$$\begin{aligned} B_{q(\sigma_\varepsilon^2)} \leftarrow \tfrac{1}{2}\Big(&\|\boldsymbol{y}_{\boldsymbol{x}_{\text{obs}}} - \boldsymbol{X}_{x\text{obs}}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \text{tr}\left(\boldsymbol{X}_{x\text{obs}}^\top\boldsymbol{X}_{x\text{obs}}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right) + \|\boldsymbol{y}_{\boldsymbol{x}_{\text{unobs}}}\|^2 \\ &-2\boldsymbol{y}_{\boldsymbol{x}_{\text{unobs}}}^\top E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{X}_{x\text{unobs}}\right)\boldsymbol{\mu}_{q(\boldsymbol{\beta})} \\ &+\text{tr}\left[E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{X}_{x\text{unobs}}^\top\boldsymbol{X}_{x\text{unobs}}\right)\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} + \boldsymbol{\mu}_{q(\boldsymbol{\beta})}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}^\top\right)\right]\Big) + \mu_{q(1/a_\varepsilon)} \end{aligned}$$

$$\mu_{q(1/\sigma_\varepsilon^2)} \leftarrow \tfrac{1}{2}(n+1)\big/B_{q(\sigma_\varepsilon^2)}; \qquad B_{q(a_\varepsilon)} \leftarrow \mu_{q(1/\sigma_\varepsilon^2)} + A_\varepsilon^{-2}; \qquad \mu_{q(1/a_\varepsilon)} \leftarrow 1/B_{q(a_\varepsilon)}$$

Measurement model updates:

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\alpha})} \leftarrow \left\{\mu_{q(1/\sigma_o^2)}\, E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\widetilde{\boldsymbol{X}}^\top\widetilde{\boldsymbol{X}}\right) + \tfrac{1}{\sigma_\alpha^2}\boldsymbol{I}_q\right\}^{-1}$$

$$\boldsymbol{\mu}_{q(\boldsymbol{\alpha})} \leftarrow \mu_{q(1/\sigma_o^2)}\, \boldsymbol{\Sigma}_{q(\boldsymbol{\alpha})}\, E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\widetilde{\boldsymbol{X}}\right)^\top \boldsymbol{o}$$

$$\begin{aligned} B_{q(\sigma_o^2)} \leftarrow \tfrac{1}{2}\Big(&\|\boldsymbol{o}_{\boldsymbol{x}_{\text{obs}}} - \tilde{\boldsymbol{X}}_{x\text{obs}}\boldsymbol{\mu}_{q(\boldsymbol{\alpha})}\|^2 + \text{tr}\left(\tilde{\boldsymbol{X}}_{x\text{obs}}^\top\tilde{\boldsymbol{X}}_{x\text{obs}}\boldsymbol{\Sigma}_{q(\boldsymbol{\alpha})}\right) + \|\boldsymbol{o}_{\boldsymbol{x}_{\text{unobs}}}\|^2 \\ &-2\boldsymbol{o}_{\boldsymbol{x}_{\text{unobs}}}^\top E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\tilde{\boldsymbol{X}}_{x\text{unobs}}\right)\boldsymbol{\mu}_{q(\boldsymbol{\alpha})} \\ &+\text{tr}\left[E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\tilde{\boldsymbol{X}}_{x\text{unobs}}^\top\tilde{\boldsymbol{X}}_{x\text{unobs}}\right)\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\alpha})} + \boldsymbol{\mu}_{q(\boldsymbol{\alpha})}\boldsymbol{\mu}_{q(\boldsymbol{\alpha})}^\top\right)\right]\Big) + \mu_{q(1/a_o)} \end{aligned}$$

$$\mu_{q(1/\sigma_o^2)} \leftarrow \tfrac{1}{2}(n+1)\big/B_{q(\sigma_o^2)}; \qquad B_{q(a_o)} \leftarrow \mu_{q(1/\sigma_o^2)} + A_o^{-2}; \qquad \mu_{q(1/a_o)} \leftarrow 1/B_{q(a_o)}$$

Prior specification updates:

$$\begin{aligned} \sigma^2_{q(\boldsymbol{x}_{\text{unobs}})} \leftarrow 1\Big/\Big(&\mu_{q(1/\sigma_\varepsilon^2)}\left[\left\{\left(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\right)_3\right\}^2 + \left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right)_{33}\right] + \textstyle\sum_{k=1}^K \mu_{q(a_{\cdot k})}\mu_{q\left(1/(\sigma_k^x)^2\right)} \\ &+\mu_{q(1/\sigma_o^2)}\left[\left\{\left(\boldsymbol{\mu}_{q(\boldsymbol{\alpha})}\right)_2\right\}^2 + \left(\boldsymbol{\Sigma}_{q(\boldsymbol{\alpha})}\right)_{22}\right]\Big) \end{aligned}$$

$$\boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs}})} \leftarrow \sigma^2_{q(\boldsymbol{x}_{\text{unobs}})} \Big(\mu_{q(1/\sigma^2_\varepsilon)} \Big[\big(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\big)_3 \boldsymbol{y}_{\boldsymbol{x}_{\text{unobs}}}$$
$$- \mathbf{1}_{n_{\text{unobs}}} \Big\{ \big(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\big)_1 + \big(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\big)_3 + \big(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\big)_{13} \Big\}$$
$$- \boldsymbol{c}_{\boldsymbol{x}_{\text{unobs}}} \Big\{ \big(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\big)_2 + \big(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\big)_3 + \big(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\big)_{23} \Big\}$$
$$+ \mathbf{1}_{n_{\text{unobs}}} \textstyle\sum_{k=1}^K \mu_{q(a_{\cdot k})} \mu_{q(\mu^x_k)} \mu_{q\left(1/(\sigma^x_k)^2\right)}$$
$$+ \mu_{q(1/\sigma^2_o)} \Big[\big(\boldsymbol{\mu}_{q(\boldsymbol{\alpha})}\big)_1 \boldsymbol{o}_{\boldsymbol{x}_{\text{unobs}}} - \mathbf{1}_{n_{\text{unobs}}} \Big\{ \big(\boldsymbol{\mu}_{q(\boldsymbol{\alpha})}\big)_1 + \big(\boldsymbol{\mu}_{q(\boldsymbol{\alpha})}\big)_2$$
$$+ \big(\boldsymbol{\Sigma}_{q(\boldsymbol{\alpha})}\big)_{12} \Big\} \Big] \Big] \Big),$$

For $i = 1, \cdots, n$ and $k = 1, \cdots, K$:

$$\nu_{ik} \leftarrow \psi\left(\alpha_{q(\omega_k)}\right) + \tfrac{1}{2}\psi\left(A_{q\left((\sigma^x_k)^2\right)}\right) - \tfrac{1}{2}\log\left(B_{q\left((\sigma^x_k)^2\right)}\right)$$
$$- \tfrac{1}{2}\mu_{q\left(1/(\sigma^x_k)^2\right)} \left(\left[\left\{ E_{q(\boldsymbol{x}_{\text{unobs}})}(\boldsymbol{x}) \right\}_i - \mu_{q(\mu^x_k)} \right]^2 + \left\{ \sigma^2_{q(\boldsymbol{x}_{\text{unobs}})}(\boldsymbol{x}) \right\}_i + \sigma^2_{q(\mu^x_k)} \right)$$

For $i = 1, \cdots, n$ and $k = 1, \cdots, K$: $\quad \mu_{q(a_{ik})} \leftarrow \exp(\nu_{ik}) \big/ \textstyle\sum_{k=1}^K \exp(\nu_{ik})$

For $k = 1, \cdots, K$:

$$\mu_{q(a_{\bullet k})} \leftarrow \sum_{i=1}^n \mu_{q(a_{ik})}; \qquad \sigma^2_{q(\mu^x_k)} \leftarrow 1 \Big/ \left\{ \mu_{q\left(1/(\sigma^x_k)^2\right)} \mu_{q(a_{\bullet k})} + 1/\sigma^2_\mu \right\}$$
$$\mu_{q(\mu^x_k)} \leftarrow \sigma^2_{q(\mu^x_k)} \left\{ \mu_{q\left(1/(\sigma^x_k)^2\right)} \mu_{q(a_{\bullet k})} \sum_{i=1}^n \left\{ E_{q(\boldsymbol{x}_{\text{unobs}})}(\boldsymbol{x}) \right\}_i + \mu_\mu/\sigma^2_\mu \right\}$$
$$\alpha_{q(\omega_k)} \leftarrow \mu_{q(a_{\bullet k})} + \alpha; \qquad A_{q\left((\sigma^x_k)^2\right)} \leftarrow \tfrac{1}{2}\left(\mu_{q(a_{\bullet k})} + 1\right)$$
$$B_{q\left((\sigma^x_k)^2\right)} \leftarrow \tfrac{1}{2}\mu_{q(a_{\bullet k})} \sum_{i=1}^n \left(\left[\left\{ E_{q(\boldsymbol{x}_{\text{unobs}})}(\boldsymbol{x}) \right\}_i - \mu_{q(\mu^x_k)} \right]^2 + \left\{ \sigma^2_{q(\boldsymbol{x}_{\text{unobs}})}(\boldsymbol{x}) \right\}_i + \sigma^2_{q(\mu^x_k)} \right)$$
$$+ \mu_{q(1/a^x_k)}$$
$$\mu_{q\left(1/(\sigma^x_k)^2\right)} \leftarrow A_{q\left((\sigma^x_k)^2\right)} \Big/ B_{q\left((\sigma^x_k)^2\right)}; \qquad B_{q(a^x_k)} \leftarrow \mu_{q\left(1/(\sigma^x_k)^2\right)} + \left(A^x_k\right)^{-2}$$
$$\mu_{q(1/a^x_k)} \leftarrow 1 \Big/ B_{q(a^x_k)}$$

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

Algorithm 9: *MFVB algorithm for obtaining the parameters in the optimal q densities for the normal response measurement error model (6.4).*

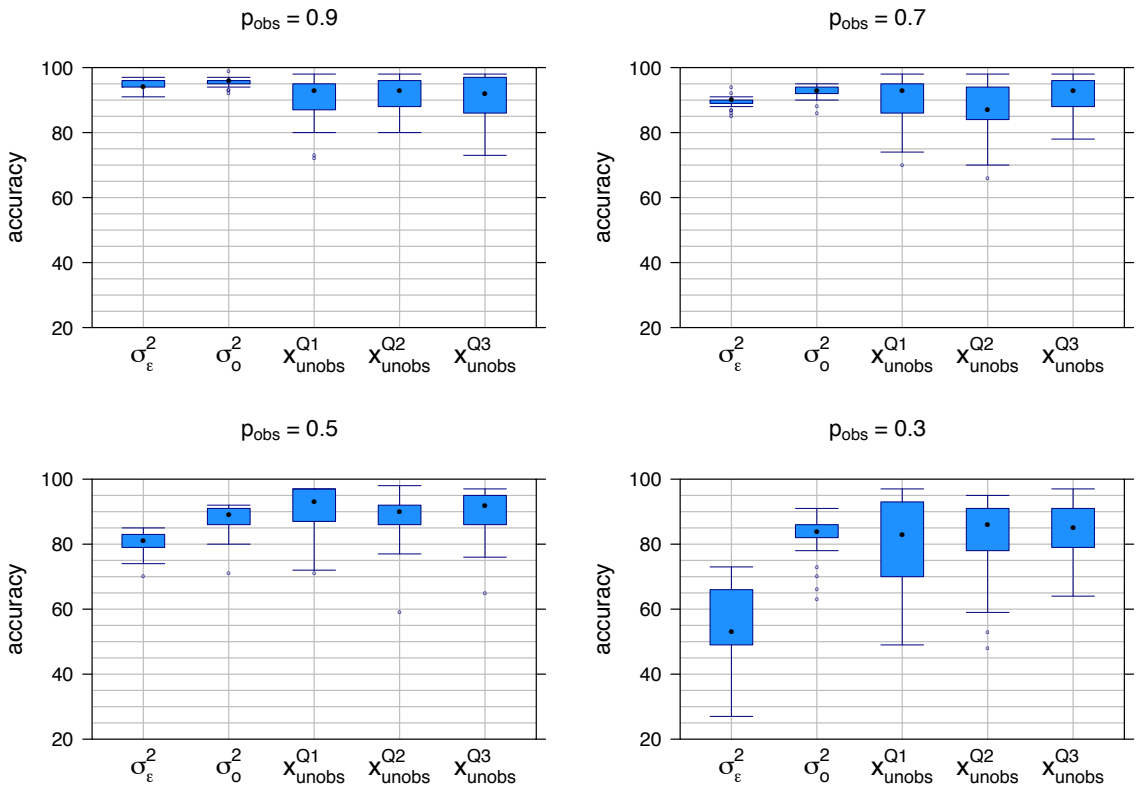


Figure 6.2: *Boxplots of parameter accuracies for MFVB in Algorithm 9 applied to the measurement error model given in (6.4) for the simulation setting described in Section 6.5.*

quartiles of the unobserved x's. These are represented as $x_{\text{unobs}}^{\text{Q}_i}$, $i = 1, 2, 3$. We can see from Figure 6.2 that the accuracies seem to be quite good for σ_o^2 and the quartiles of $\boldsymbol{x}_{\text{unobs}}$ for all values of p_{obs} except for $p_{\text{obs}} = 0.3$. However, accuracies are very poor for σ_ε^2 when $p_{\text{obs}} = 0.3$. Overall, the accuracies are looking good when $p_{\text{obs}} \in \{0.5, 0.7, 0.9\}$.

We also assessed the 95 % credible interval coverage achieved by the MFVB optimal q-densities against the true parameter coverage for the same parameters. From Table 6.1 we see that the coverage probabilities are generally high for all parameters except σ_ε^2 and σ_o^2. As p_{obs} decreases, so do the coverage probabilities for σ_ε^2 and σ_o^2. The run time for

p_{obs}	0.9	0.7	0.5	0.3
σ_ε^2	92	96	76	68
σ_o^2	92	88	84	72
$x_{\text{unobs}}^{\text{Q}_1}$	96	84	96	84
$x_{\text{unobs}}^{\text{Q}_2}$	96	92	100	100
$x_{\text{unobs}}^{\text{Q}_3}$	92	96	100	92

Table 6.1: *95% credible interval coverage probabilities for the MFVB approximation applied to (6.4).*

p_{obs}	0.9	0.7	0.5	0.3
MCMC	51.80 (0.61)	55.69 (0.51)	54.46 (1.70)	59.80 (0.67)
MFVB	1.46 (0.10)	1.46 (0.11)	1.42 (0.03)	1.41 (0.04)
Ratio	35.48	38.17	38.44	42.46

Table 6.2: *Average (standard deviation) run time in seconds for MFVB and MCMC fitting of the Gaussian response measurement error model.*

MCMC and MFVB was also assessed. The resulting times are given in table 6.2. Based on the timing results, we see that MFVB is approximately 40 times faster than MCMC. The time savings afforded by MFVB in this instance is not of a large magnitude compared with other models considered in this thesis, however extension to arbitrarily large and complex models involving measurement error would more clearly represent the time benefits of using MFVB as an alternative to MCMC.

From Figure 6.3 we can see that the means of the MFVB fits are agreeing well to the fits obtained via MCMC. More importantly, both of these fits are agreeing with the true mean function even when $p_{\text{obs}} = 0.3$. Thus, overall this simulation study has represented the effectiveness of using MFVB as an alternative to MCMC in models impaired

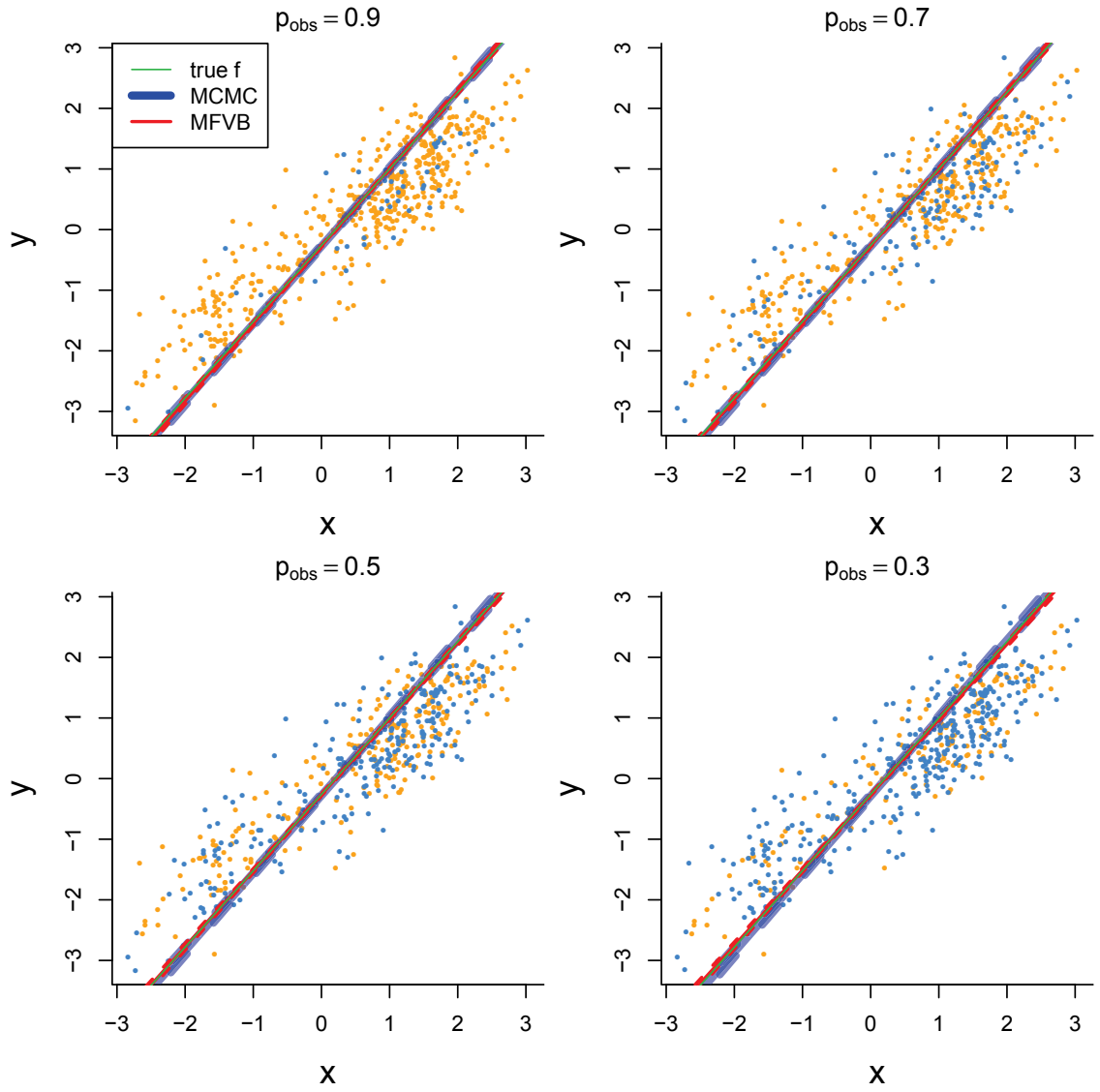


Figure 6.3: *Comparative plots of the MFVB and MCMC mean fitted functions with corresponding 95% credible sets for a single replication of the simulation study. The orange data points represent the observed data and the blue data points represent the unobserved data.*

by measurement error.

6.6 Discussion

We have derived a fast MFVB algorithm to a regression model with classical measurement error. This approach was shown to have good accuracy with evidence of time saving compared to the MCMC approach. Even though the time comparisons weren't of high magnitude for MFVB compared to MCMC, extension of such methodology to arbitrarily large and complex models involving measurement error would exhibit larger time savings.

6.A Expectation updates

This section lists the updates in Algorithm 9 that involve the expectation operator with respect to $\boldsymbol{x}_{\text{unobs}}$:

$$E_{q(\boldsymbol{x}_{\text{unobs}})}(\boldsymbol{x}) \leftarrow \begin{bmatrix} \boldsymbol{x}_{\text{obs}} \\ \boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs}})} \end{bmatrix},$$

$$E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{x}^\top \boldsymbol{x}\right) \leftarrow \parallel \boldsymbol{x}_{\text{obs}} \parallel^2 + \parallel \boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs}})} \parallel^2 + \textstyle\sum_{i=1}^{n_{\text{unobs}}} \sigma^2_{q(\boldsymbol{x}_{\text{unobs}})},$$

$$E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\widetilde{\boldsymbol{X}}\right) \leftarrow \begin{bmatrix} \mathbf{1} & \boldsymbol{x}_{\text{obs}} \\ \mathbf{1} & \boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs}})} \end{bmatrix},$$

$$E_{q(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}})}\left(\widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}}\right) \leftarrow \begin{bmatrix} n & \mathbf{1}^\top \boldsymbol{x}_{\text{obs}} + \mathbf{1}^\top \boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs}})} \\ \mathbf{1}^\top \boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs}})} & E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{x}^\top \boldsymbol{x}\right) \end{bmatrix},$$

$$E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\widetilde{\boldsymbol{X}}_{\boldsymbol{x}_{\text{unobs}}}\right) \leftarrow \begin{bmatrix} \mathbf{1} & \boldsymbol{\mu}_{(\boldsymbol{x}_{\text{unobs}})} \end{bmatrix},$$

$$E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\widetilde{\boldsymbol{X}}^\top_{\boldsymbol{x}_{\text{unobs}}} \widetilde{\boldsymbol{X}}_{\boldsymbol{x}_{\text{unobs}}}\right) \leftarrow \begin{bmatrix} n_{\text{unobs}} & \mathbf{1}^\top \boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs}})} \\ \mathbf{1}^\top \boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs}})} & \parallel \boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs}})} \parallel^2 + \sum_{i=1}^{n_{\text{unobs}}} \sigma^2_{q(\boldsymbol{x}_{\text{unobs}})} \end{bmatrix},$$

$$E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{X}\right) \leftarrow \begin{bmatrix} \mathbf{1} & \boldsymbol{c}_{x\text{obs}} & x_{\text{obs}} \\ \mathbf{1} & \boldsymbol{c}_{x\text{unobs}} & \boldsymbol{\mu}_{q(x_{\text{unobs}})} \end{bmatrix},$$

$$E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{X}^\top \boldsymbol{X}\right) \leftarrow \begin{bmatrix} n & \mathbf{1}^\top \boldsymbol{c} & \mathbf{1}^\top \boldsymbol{x}_{\text{obs}} + \mathbf{1}^\top \boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs}})} \\ \mathbf{1}^\top \boldsymbol{c} & \boldsymbol{c}^\top \boldsymbol{c} & \boldsymbol{c}^\top_{x\text{obs}} \boldsymbol{x}_{\text{obs}} + \boldsymbol{c}^\top_{x\text{unobs}} \boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs}})} \\ \mathbf{1}^\top \boldsymbol{x}_{\text{obs}} + \mathbf{1}^\top \boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs}})} & \boldsymbol{c}^\top_{x\text{obs}} \boldsymbol{x}_{\text{obs}} + \boldsymbol{c}^\top_{x\text{unobs}} \mu_{q(\boldsymbol{x}_{\text{unobs}})} & E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{x}^\top \boldsymbol{x}\right) \end{bmatrix},$$

$$E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{X}_{x\text{unobs}}\right) \leftarrow \begin{bmatrix} \mathbf{1} & \boldsymbol{c}_{x\text{unobs}} & \boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs}})} \end{bmatrix},$$

$$E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{X}^\top_{x\text{unobs}} \boldsymbol{X}_{x\text{unobs}}\right) \leftarrow \begin{bmatrix} n_{\text{unobs}} & \mathbf{1}^\top \boldsymbol{c}_{x\text{unobs}} & \mathbf{1}^\top \boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs}})} \\ \mathbf{1}^\top \boldsymbol{c}_{x\text{unobs}} & \parallel \boldsymbol{c}_{x\text{unobs}} \parallel^2 & \boldsymbol{c}^\top_{x\text{unobs}} \boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs}})} \\ \mathbf{1}^\top \boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs}})} & \boldsymbol{c}^\top_{x\text{unobs}} \boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs}})} & \parallel \boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs}})} \parallel^2 + \sum_{i=1}^{n_{\text{unobs}}} \sigma^2_{q(\boldsymbol{x}_{\text{unobs}})} \end{bmatrix}.$$

Also,

$$\boldsymbol{\sigma}^2_{q(\boldsymbol{x}_{\text{unobs}})}(\boldsymbol{x}) \equiv \begin{bmatrix} \mathbf{0}_{n_{\text{obs}}} \\ \sigma^2_{q(\boldsymbol{x}_{\text{unobs}})} \mathbf{1}_{n_{\text{unobs}}} \end{bmatrix},$$

where $\mathbf{0}_{n_{\text{obs}}}$ is the $n_{\text{obs}} \times 1$ column vector with all entries equal to 0.

6.B Derivation of Algorithm 9

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

First note that

$$\begin{aligned} p(\boldsymbol{\beta} \mid \text{rest}) &\propto p(\boldsymbol{y} \mid \boldsymbol{\beta},\, \boldsymbol{x},\, \sigma^2_\varepsilon)\, p(\boldsymbol{\beta}) \\ &= \exp\left\{ -\tfrac{1}{2\sigma^2_\varepsilon} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 - \tfrac{1}{2\sigma^2_\beta} \|\boldsymbol{\beta}\|^2 \right\} + \text{const}, \end{aligned}$$

where 'const' denotes all terms not depending on $\boldsymbol{\beta}$. Taking the logarithm of both sides gives

$$\begin{aligned} \log p(\boldsymbol{\beta} \mid \text{rest}) &\propto -\tfrac{1}{2\sigma^2_\varepsilon} \parallel \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \parallel^2 - \tfrac{1}{2\sigma^2_\beta} \|\boldsymbol{\beta}\|^2 \\ &\propto -\tfrac{1}{2\sigma^2_\varepsilon} \left(\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{\beta} - 2\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y} \right) - \tfrac{1}{2\sigma^2_\beta} \|\boldsymbol{\beta}\|^2 \\ &= -\tfrac{1}{2} \left\{ \boldsymbol{\beta}^\top \left(\tfrac{1}{\sigma^2_\varepsilon} \boldsymbol{X}^\top \boldsymbol{X} + \tfrac{1}{\sigma^2_\beta} \boldsymbol{I}_3 \right) \boldsymbol{\beta} - 2\boldsymbol{\beta}^\top \left(\tfrac{1}{\sigma^2_\varepsilon} \boldsymbol{X}^\top \boldsymbol{y} \right) \right\} + \text{const}. \end{aligned}$$

Taking expectations with respect to all parameters except $\boldsymbol{\beta}$, we get

$$\begin{aligned} \log q^*(\boldsymbol{\beta}) &= E_q \left\{ \log p(\boldsymbol{\beta} \mid \text{rest}) \right\} + \text{const}. \\ &= -\tfrac{1}{2} \Big\{ \boldsymbol{\beta}^\top \left(\mu_{q(1/\sigma^2_\varepsilon)} E_{q(\boldsymbol{x}_{\text{unobs}})} \left(\boldsymbol{X}^\top \boldsymbol{X} \right) + \tfrac{1}{7} \sigma^2_\beta \boldsymbol{I}_3 \right) \boldsymbol{\beta} \\ &\qquad -2\boldsymbol{\beta}^\top \left(\mu_{q(1/\sigma^2_\varepsilon)} E_{q(\boldsymbol{x}_{\text{unobs}})} \left(\boldsymbol{X} \right)^\top \boldsymbol{y} \right) \Big\} + \text{const}. \end{aligned}$$

Therefore,

$$q^*\left(\boldsymbol{\beta}\right) \text{ is the } N\left(\boldsymbol{\mu}_{q(\boldsymbol{\beta})},\, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right) \text{ density function,}$$

where

$$\begin{aligned} \boldsymbol{\mu}_{q(\boldsymbol{\beta})} &= \boldsymbol{\mu}_{q(1/\sigma^2_\varepsilon)} \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} E_{q(\boldsymbol{x}_{\text{unobs}})} \left(\boldsymbol{X} \right)^\top \boldsymbol{y}, \quad \text{and} \\ \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} &= \left(\boldsymbol{\mu}_{q(1/\sigma^2_\varepsilon)}\, E_{q(\boldsymbol{x}_{\text{unobs}})} \left(\boldsymbol{X}^\top \boldsymbol{X} \right) + \frac{1}{\sigma^2_\beta} \boldsymbol{I}_3 \right)^{-1}. \end{aligned}$$

Expressions for $B_{q(\sigma^2_\varepsilon)}$ and $\mu_{q(1/\sigma^2_\varepsilon)}$

$$
\begin{aligned}
p\left(\sigma_\varepsilon^2|\text{rest}\right) &\propto p\left(\boldsymbol{y} \mid \boldsymbol{\beta},\, \boldsymbol{x},\, \sigma_\varepsilon^2\right)\, p\left(\sigma_\varepsilon^2 \mid a_\varepsilon\right) \\
&= \left|\sigma_\varepsilon^2 \boldsymbol{I}_n\right|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2\sigma_\varepsilon^2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2\right\}\left(\sigma_\varepsilon^2\right)^{-\frac{1}{2}-1} \exp\left\{-(1/a_\varepsilon)/\sigma_\varepsilon^2\right\} + \text{const.}
\end{aligned}
$$

Taking the logarithm of both sides gives

$$
\begin{aligned}
\log p\left(\sigma_\varepsilon^2|\text{rest}\right) &= -\tfrac{n}{2}\log\left(\sigma_\varepsilon^2\right) - \tfrac{1}{2\sigma_\varepsilon^2}\parallel \boldsymbol{y}-\boldsymbol{X}\,\boldsymbol{\beta}\parallel^2 - \tfrac{3}{2}\log\left(\sigma_\varepsilon^2\right) - (1/a_\varepsilon)/\sigma_\varepsilon^2 \\
&\quad +\text{const.} \\
&= -\left\{\tfrac{1}{2}(n+1)+1\right\}\log\left(\sigma_\varepsilon^2\right) - \left(\tfrac{1}{2}\parallel \boldsymbol{y}-\boldsymbol{X}\,\boldsymbol{\beta}\parallel^2 + 1/a_\varepsilon\right)/\sigma_\varepsilon^2 + \text{const.}
\end{aligned}
$$

Taking expectations with respect to all parameters except σ_ε^2, we get

$$
\begin{aligned}
\log q^*(\sigma_\varepsilon^2) &= E_q\left\{\log\, q^*\left(\sigma_\varepsilon^2 \mid \text{rest}\right)\right\} + \text{const.} \\
&= -\left\{\tfrac{1}{2}(n+1)+1\right\}\log\left(\sigma_\varepsilon^2\right) - \left(\tfrac{1}{2}E_q\|\boldsymbol{y}-\boldsymbol{X}\,\boldsymbol{\beta}\|^2 + \boldsymbol{\mu}_{q(1/a_\varepsilon)}\right)/\sigma_\varepsilon^2 \\
&\quad +\text{const.}
\end{aligned}
$$

Now, breaking $E_q\|\boldsymbol{y}-\boldsymbol{X}\,\boldsymbol{\beta}\|^2$ up into the *observed* and *unobserved* parts, we get

$$
E_q\|\boldsymbol{y}-\boldsymbol{X}\,\boldsymbol{\beta}\|^2 = E_q\|\boldsymbol{y}_{x\text{obs}} - \boldsymbol{X}_{x\text{obs}}\,\boldsymbol{\beta}\|^2 + E_q\|\boldsymbol{y}_{x\text{unobs}} - \boldsymbol{X}_{x\text{unobs}}\,\boldsymbol{\beta}\|^2.
$$

Using Result 1.4.19,

$$
\begin{aligned}
E_q\|\boldsymbol{y}-\boldsymbol{X}\,\boldsymbol{\beta}\|^2 &= \|\boldsymbol{y}_{x\text{obs}} - \boldsymbol{X}_{x\text{obs}}\,\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \text{tr}\left(\boldsymbol{X}_{x\text{obs}}^\top \boldsymbol{X}_{x\text{obs}}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right) \\
&\quad +\|\boldsymbol{y}_{x\text{unobs}} - E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{X}_{x\text{unobs}}\right)\,\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \text{tr}\left\{\text{Cov}_q\left(\boldsymbol{X}_{x\text{unobs}}\boldsymbol{\beta}\right)\right\}
\end{aligned}
$$

where, using Result 1.4.20,

$$
\begin{aligned}
\text{Cov}_q\left(\boldsymbol{X}_{x\text{unobs}}\,\boldsymbol{\beta}\right) &= E_q\left\{\text{Cov}_q\left(\boldsymbol{X}_{x\text{unobs}}\,\boldsymbol{\beta}\,|\,\boldsymbol{x}_{\text{unobs}}\right)\right\} + \text{Cov}_q\left\{E_q\left(\boldsymbol{X}_{x\text{unobs}}\,\boldsymbol{\beta}\,|\,\boldsymbol{x}_{\text{unobs}}\right)\right\} \\
&= E_q\left\{\boldsymbol{X}_{x\text{unobs}}\text{Cov}_q\left(\boldsymbol{\beta}\,|\,\boldsymbol{x}_{\text{unobs}}\right)\boldsymbol{X}_{x\text{unobs}}^\top\right\} + \text{Cov}_q\left\{\boldsymbol{X}_{x\text{unobs}}E_q\left(\boldsymbol{\beta}\,|\,\boldsymbol{x}_{\text{unobs}}\right)\right\}
\end{aligned}
$$

Taking into account the independence between $\boldsymbol{x}_{\text{unobs}}$ and $\boldsymbol{\beta}$, as shown in product restriction (6.6), this gives

$$
\begin{aligned}
\text{Cov}_q\left(\boldsymbol{X}_{x\text{unobs}}\,\boldsymbol{\beta}\right) &= E_q\left\{\boldsymbol{X}_{x\text{unobs}}\text{Cov}_q\left(\boldsymbol{\beta}\right)\boldsymbol{X}_{x\text{unobs}}^\top\right\} + \text{Cov}_q\left\{\boldsymbol{X}_{x\text{unobs}}E_q\left(\boldsymbol{\beta}\right)\right\} \\
&= E_q\left(\boldsymbol{X}_{x\text{unobs}}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\boldsymbol{X}_{x\text{unobs}}^\top\right) + \text{Cov}_q\left(\boldsymbol{X}_{x\text{unobs}}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\right) \\
&= E_q\left(\boldsymbol{X}_{x\text{unobs}}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\boldsymbol{X}_{x\text{unobs}}^\top\right) + E_q\left(\boldsymbol{X}_{x\text{unobs}}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}^\top\boldsymbol{X}_{x\text{unobs}}^\top\right) \\
&\quad -\left\{E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{X}_{x\text{unobs}}\right)\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\right\}\left\{E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{X}_{x\text{unobs}}\right)\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\right\}^\top \\
&= E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{X}_{x\text{unobs}}^\top\boldsymbol{X}_{x\text{unobs}}\right)\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} + \boldsymbol{\mu}_{q(\boldsymbol{\beta})}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}^\top\right) \\
&\quad -\left\{E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{X}_{x\text{unobs}}\right)\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\right\}\left\{E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{X}_{x\text{unobs}}\right)\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\right\}^\top.
\end{aligned}
$$

Therefore,

$$\begin{aligned} E_q\|\boldsymbol{y}-\boldsymbol{X}\,\boldsymbol{\beta}\|^2 \;=\; & \|\boldsymbol{y}_{x\text{obs}}-\boldsymbol{X}_{x\text{obs}}\,\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2+\operatorname{tr}\left(\boldsymbol{X}_{x\text{obs}}^{\top}\boldsymbol{X}_{x\text{obs}}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right) \\ & +\|\boldsymbol{y}_{x\text{unobs}}\|^2-2\boldsymbol{y}_{x\text{unobs}}^{\top}E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{X}_{x\text{unobs}}\right)\boldsymbol{\mu}_{q(\boldsymbol{\beta})} \\ & +\operatorname{tr}\left[E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{X}_{x\text{unobs}}^{\top}\boldsymbol{X}_{x\text{unobs}}\right)\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}+\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}^{\top}\right)\right]. \end{aligned}$$

And so,

$$q^*\left(\sigma_\varepsilon^2\right) \text{ follows an Inverse-Gamma}\left(\tfrac{1}{2}\left(n+1\right),\; B_{q(\sigma_\varepsilon^2)}\right) \text{ distribution,}$$

where

$$B_{q(\sigma_\varepsilon^2)}=\tfrac{1}{2}E_q\parallel\boldsymbol{y}-\boldsymbol{X}\,\boldsymbol{\beta}\parallel^2+\boldsymbol{\mu}_{q(1/a_\varepsilon)}.$$

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

We begin with

$$\begin{aligned} p\left(a_\varepsilon\mid\text{rest}\right) &\propto p\left(\sigma_\varepsilon^2\mid a_\varepsilon\right)\,p\left(a_\varepsilon\right) \\ &= \left(1/a_\varepsilon\right)^{1/2}\exp\left\{-\left(1/a_\varepsilon\right)/\sigma_\varepsilon^2\right\}\left(a_\varepsilon\right)^{-\frac{1}{2}-1}\exp\left\{-\left(1/A_\varepsilon^2/a_\varepsilon\right)\right\}+\text{const.} \end{aligned}$$

Taking the logarithm of both sides, gives

$$\begin{aligned} \log p\left(a_\varepsilon\mid\text{rest}\right) &= -\tfrac{1}{2}\log(a_\varepsilon)-\left(1/a_\varepsilon\right)/\sigma_\varepsilon^2-\tfrac{3}{2}\log\left(a_\varepsilon\right)-\left(1/A_\varepsilon^2\right)/a_\varepsilon+\text{const} \\ &= -2\log(a_\varepsilon)-\left(\sigma_\varepsilon^{-2}+A_\varepsilon^{-2}\right)/a_\varepsilon+\text{const.} \end{aligned}$$

Taking expectations, we get

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

Therefore,

$$q^*\left(a_\varepsilon\right) \text{ is the Inverse-Gamma}\left(1,\; B_{q(a_\varepsilon)}\right) \text{ density function,}$$

where

$$B_{q(a_\varepsilon)}=\boldsymbol{\mu}_{q(1/\sigma_\varepsilon^2)}+A_\varepsilon^{-2}.$$

In addition, using Result 1.4.3,

$$\boldsymbol{\mu}_{q(1/a_\varepsilon)}=1/B_{q(a_\varepsilon)}.$$

Expressions for $\boldsymbol{\mu}_{q(\boldsymbol{x}_{\mathbf{unobs}})}$ and $\sigma^2_{q(\boldsymbol{x}_{\mathbf{unobs}})}$

We first note that

$$\begin{aligned} p\left(\boldsymbol{x}_{\text{unobs}} \mid \text{rest}\right) &\propto p\left(\boldsymbol{y} \mid \boldsymbol{\beta},\, \boldsymbol{x},\, \sigma_\varepsilon^2\right)\, p\left(\boldsymbol{x} \mid \boldsymbol{a},\, \boldsymbol{\mu},\, \sigma^2\right)\, p\left(\boldsymbol{o} \mid \boldsymbol{x},\, \boldsymbol{\alpha},\, \sigma_o^2\right) \\ &= \exp\left\{-\tfrac{1}{2\sigma_\varepsilon^2}\|\boldsymbol{y}_{x\text{unobs}} - \boldsymbol{X}_{x\text{unobs}}\boldsymbol{\beta}\|^2\right\} \\ &\quad \times \prod_{i=n_{\text{obs}}+1}^{n_{\text{unobs}}} \prod_{k=1}^{K}\left[\left\{2\pi\left(\sigma_k^x\right)^2\right\}^{-1/2} \exp\left\{-\tfrac{1}{2}\left(x_i - \mu_k^x\right)^2 / \left(\sigma_k^x\right)^2\right\}\right]^{a_{ik}} \\ &\quad \times \exp\left\{-\tfrac{1}{2\sigma_o^2}\|\boldsymbol{o}_{x\text{unobs}} - \tilde{\boldsymbol{X}}_{x\text{unobs}}\boldsymbol{\alpha}\|^2\right\} + \text{const.} \end{aligned}$$

Taking the logarithm of both sides gives

$$\begin{aligned} \log\, p\left(\boldsymbol{x}_{\text{unobs}} \mid \text{rest}\right) &= -\tfrac{1}{2\sigma_\varepsilon^2} \parallel \boldsymbol{y}_{x\text{unobs}} - \beta_0 - \beta_c \boldsymbol{c}_{x\text{unobs}} - \beta_x \boldsymbol{x}_{\text{unobs}} \parallel^2 \\ &\quad -\tfrac{1}{2}\textstyle\sum_{k=1}^{K}\sum_{i=1}^{n_{unobs}} a_{ik}\frac{(\xi-\mu_k)^2}{\sigma_k^2} \\ &\quad -\tfrac{1}{2\sigma_o^2} \parallel \boldsymbol{o}_{x\text{unobs}} - \alpha_0 - \alpha_1 \boldsymbol{x}_{\text{unobs}} \parallel^2 + \text{const} \\ &= -\tfrac{1}{2\sigma_\varepsilon^2}\left(\beta_x^2 \parallel \boldsymbol{x}_{\text{unobs}} \parallel^2 + 2\beta_0\beta_x \mathbf{1}_{n_{\text{unobs}}}^\top \boldsymbol{x}_{\text{unobs}} - 2\beta_x \boldsymbol{y}_{x\text{unobs}}^\top \boldsymbol{x}_{\text{unobs}}\right. \\ &\quad \left. + 2\beta_c\beta_x \boldsymbol{c}_{x\text{unobs}}^\top \boldsymbol{x}_{\text{unobs}}\right) - \tfrac{1}{2}\left(\parallel \boldsymbol{x}_{\text{unobs}} \parallel^2 \textstyle\sum_{k=1}^{K}\frac{a_{.k}}{\sigma_k^2}\right. \\ &\quad \left. -2\left(\mathbf{1}_{n_{\text{unobs}}}^\top \boldsymbol{x}_{\text{unobs}}\right)\textstyle\sum_{k=1}^{K}\frac{a_{.k}\mu_k}{\sigma_k^2}\right) - \tfrac{1}{2\sigma_o^2}\left(\alpha_1^2 \parallel \boldsymbol{x}_{\text{unobs}} \parallel^2\right. \\ &\quad \left. + 2\alpha_0\alpha_1 \mathbf{1}_{n_{\text{unobs}}}^\top \boldsymbol{x}_{\text{unobs}} - 2\alpha_1 \boldsymbol{o}_{x\text{unobs}}^\top \boldsymbol{x}_{\text{unobs}}\right) + \text{const} \\ &= -\tfrac{1}{2}\boldsymbol{x}_{\text{unobs}}^\top\left(\tfrac{\beta_x^2}{\sigma_\varepsilon^2} + \textstyle\sum_{k=1}^{K}\frac{a_{.k}}{\sigma_k^2} + \frac{\alpha_1^2}{\sigma_o^2}\right)\boldsymbol{x}_{\text{unobs}} \\ &\quad + \boldsymbol{x}_{\text{unobs}}^\top\left(\tfrac{\beta_x}{\sigma_\varepsilon^2}\left(\boldsymbol{y}_{x\text{unobs}} - \beta_0 \mathbf{1}_{n_{\text{unobs}}} - \beta_c \boldsymbol{c}_{x\text{unobs}}\right)\right. \\ &\quad \left. + \mathbf{1}_{n_{\text{unobs}}}\textstyle\sum_{k=1}^{K}\frac{a_{.k}\mu_k}{\sigma_k^2} + \frac{\alpha_1}{\sigma_o^2}\left(\boldsymbol{o}_{x\text{unobs}} - \alpha_0 \mathbf{1}_{n_{\text{unobs}}}\right)\right) + \text{const.} \end{aligned}$$

Taking expectations, we get

$$\begin{aligned} \log q^*\left(\boldsymbol{x}_{\text{unobs}}\right) &= E_q\left\{\log\, p\left(\boldsymbol{x}_{\text{unobs}} \mid \text{rest}\right)\right\} + \text{const} \\ &= -\tfrac{1}{2}\left[\boldsymbol{x}_{\text{unobs}}^\top\, E_q\left(\tfrac{\beta_x^2}{\sigma_\varepsilon^2} + \textstyle\sum_{k=1}^{K}\frac{a_{.k}}{\sigma_k^2} + \frac{\alpha_1^2}{\sigma_o^2}\right)\boldsymbol{x}_{\text{unobs}}\right. \\ &\quad -2\boldsymbol{x}_{\text{unobs}}^\top\, E_q\left(\tfrac{\beta_x}{\sigma_\varepsilon^2}\left(\boldsymbol{y}_{x\text{unobs}} - \beta_0 \mathbf{1}_{n_{\text{unobs}}} - \beta_c \boldsymbol{c}_{x\text{unobs}}\right)\right. \\ &\quad \left.\left. + \mathbf{1}_{n_{\text{unobs}}}\textstyle\sum_{k=1}^{K}\frac{a_{.k}\mu_k}{\sigma_k^2} + \frac{\alpha_1}{\sigma_o^2}\left(\boldsymbol{o}_{x\text{unobs}} - \alpha_0 \mathbf{1}_{n_{\text{unobs}}}\right)\right)\right] \\ &\quad + \text{const.} \end{aligned}$$

Therefore,

$$q^*\left(\boldsymbol{x}_{\text{unobs}}\right) \text{ is the } N\left(\boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs}})},\ \sigma^2_{q(\boldsymbol{x}_{\text{unobs}})}\boldsymbol{I}_{n_{\text{unobs}}}\right) \text{ density function,}$$

where

$$\begin{aligned}
\sigma^2_{q(\boldsymbol{x}_{\text{unobs}})} &= 1\Big/E_q\left(\tfrac{\beta_x^2}{\sigma_\varepsilon^2} + \textstyle\sum_{k=1}^K \tfrac{a_{.k}}{\sigma_k^2} + \tfrac{\alpha_1^2}{\sigma_o^2}\right) \\
&= 1\Big/\left[\boldsymbol{\mu}_{q(1/\sigma_\varepsilon^2)}\left\{\boldsymbol{\mu}^2_{q(\beta_x)} + \left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right)_{33}\right\} + \textstyle\sum_{k=1}^K \boldsymbol{\mu}_{q(a_{.k})}\boldsymbol{\mu}_{q\left(1/\sigma_k^2\right)}\right. \\
&\quad \left. + \boldsymbol{\mu}_{q(1/\sigma_o^2)}\left\{\boldsymbol{\mu}^2_{q(\alpha_1)} + \left(\boldsymbol{\Sigma}_{q(\boldsymbol{\alpha})}\right)_{22}\right\}\right], \text{ and} \\
\boldsymbol{\mu}_{q(\boldsymbol{x}_{\text{unobs}})} &= E_q\left(\tfrac{\beta_x}{\sigma_\varepsilon^2}\left(\boldsymbol{y}_{x\text{unobs}} - \beta_0\mathbf{1}_{n_{\text{unobs}}} - \beta_c\boldsymbol{c}_{x\text{unobs}}\right)\right. \\
&\quad \left. + \mathbf{1}_{n_{\text{unobs}}}\textstyle\sum_{k=1}^K \tfrac{a_{.k}\mu_k}{\sigma_k^2} + \tfrac{\alpha_1}{\sigma_o^2}\left(\boldsymbol{o}_{x\text{unobs}} - \alpha_0\mathbf{1}_{n_{\text{unobs}}}\right)\right) \\
&= \sigma^2_{q(\boldsymbol{x}_{\text{unobs}})}\left[\boldsymbol{\mu}_{q(1/\sigma_\varepsilon^2)}\left[\boldsymbol{\mu}_{q(\beta_x)}\boldsymbol{y}_{x\text{unobs}} - \mathbf{1}_{n_{\text{unobs}}}\left(\boldsymbol{\mu}_{q(\beta_0)}\boldsymbol{\mu}_{q(\beta_x)} + \left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right)_{13}\right)\right.\right. \\
&\quad \left. - \boldsymbol{c}_{x\text{unobs}}\left\{\boldsymbol{\mu}_{q(\beta_c)}\boldsymbol{\mu}_{q(\beta_x)} + \left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right)_{23}\right\}\right] + \boldsymbol{\mu}_{q(1/\sigma_o^2)}\left\{\boldsymbol{\mu}_{q(\alpha_1)}\boldsymbol{o}_{x\text{unobs}}\right. \\
&\quad \left.\left. - \mathbf{1}_{n_{\text{unobs}}}\left(\boldsymbol{\mu}_{q(\alpha_0)}\boldsymbol{\mu}_{q(\alpha_1)} + \left(\boldsymbol{\Sigma}_{q(\boldsymbol{\alpha})}\right)_{12}\right)\right\} + \mathbf{1}_{n_{\text{unobs}}}\textstyle\sum_{k=1}^K \boldsymbol{\mu}_{q(a_{.k})}\boldsymbol{\mu}_{q(\mu_k)}\boldsymbol{\mu}_{q\left(1/\sigma_k^2\right)}\right].
\end{aligned}$$

Expressions for $\boldsymbol{\mu}_{q(\boldsymbol{\alpha})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\alpha})}$

This derivation is similar to the derivation for $\boldsymbol{\mu}_{q(\boldsymbol{\beta})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}$, therefore

$$q^*\left(\boldsymbol{\alpha}\right) \text{ is the } N\left(\boldsymbol{\mu}_{q(\boldsymbol{\alpha})},\ \boldsymbol{\Sigma}_{q(\boldsymbol{\alpha})}\right) \text{ density function,}$$

where

$$\begin{aligned}
\boldsymbol{\mu}_{q(\boldsymbol{\alpha})} &= \boldsymbol{\mu}_{q(1/\sigma_o^2)}\boldsymbol{\Sigma}_{q(\boldsymbol{\alpha})}E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\tilde{\boldsymbol{X}}\right)^\top\boldsymbol{o}, \text{ and} \\
\boldsymbol{\Sigma}_{q(\boldsymbol{\alpha})} &= \left(\boldsymbol{\mu}_{q(1/\sigma_o^2)}E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}\right) + \tfrac{1}{\sigma_\alpha^2}\boldsymbol{I}_2\right)^{-1}.
\end{aligned}$$

Expressions for $B_{q(\sigma_o^2)}$, $\mu_{q(1/\sigma_o^2)}$, $B_{q(a_o)}$ and $\mu_{q(1/a_o)}$

These derivations are similar to those given for $B_{q(\sigma_\varepsilon^2)}$, $\mu_{q(1/\sigma_\varepsilon^2)}$, $B_{q(a_\varepsilon)}$ and $\mu_{q(1/a_\varepsilon)}$ and so we get the following updates:

$$q^*\left(\sigma_o^2\right) \text{ is the Inverse-Gamma}\left(\tfrac{1}{2}\left(n+1\right),\ B_{q(\sigma_o^2)}\right) \text{ density function,}$$

where

$$\begin{aligned} B_{q(\sigma_o^2)} &= \tfrac{1}{2} E_q \left\{ \| \, \boldsymbol{o} - \tilde{\boldsymbol{X}} \, \boldsymbol{\alpha} \, \|^2 \right\} + \mu_{q(1/a_o)} \\ &= \tfrac{1}{2} \Big[\| \, \boldsymbol{o}_{x\text{obs}} - \tilde{\boldsymbol{X}}_{x\text{obs}} \, \boldsymbol{\mu}_{q(\boldsymbol{\alpha})} \, \|^2 + \text{tr} \left(\tilde{\boldsymbol{X}}_{x\text{obs}} \, \boldsymbol{\Sigma}_{q(\boldsymbol{\alpha})} \, \tilde{\boldsymbol{X}}_{x\text{obs}}^\top \right) \\ &\quad + \| \, \boldsymbol{o}_{x\text{unobs}} \, \|^2 - 2 \, \boldsymbol{o}_{x\text{unobs}} \, E_{q(\boldsymbol{x}_{\text{unobs}})} \left(\tilde{\boldsymbol{X}}_{x\text{unobs}} \right) \, \mu_{q(\boldsymbol{\alpha})} \\ &\quad + \text{tr} \left\{ E_{q(\boldsymbol{x}_{\text{unobs}})} \left(\tilde{\boldsymbol{X}}_{x\text{unobs}}^\top \tilde{\boldsymbol{X}}_{x\text{unobs}} \right) \left(\boldsymbol{\Sigma}_{q(\boldsymbol{\alpha})} + \boldsymbol{\mu}_{q(\boldsymbol{\alpha})} \boldsymbol{\mu}_{q(\boldsymbol{\alpha})}^\top \right) \right\} \Big] + \boldsymbol{\mu}_{q(1/\sigma_o^2)}. \end{aligned}$$

In addition, using Result 1.4.3, we get

$$\mu_{q(1/\sigma_o^2)} = \tfrac{1}{2} \, (n+1) \, / B_{q(\sigma_o^2)}.$$

Also,

$$q^* \, (a_o) \text{ is the Inverse-Gamma} \left(1, \, B_{q(a_o)} \right) \text{ density function,}$$

where

$$\begin{aligned} B_{q(a_o)} &= \mu_{q(1/\sigma_o^2)} + A_o^{-2}, \text{ and} \\ \mu_{q(1/a_o)} &= 1 / B_{q(a_o)}. \end{aligned}$$

Expressions for $\mu_{q(\mu_k^x)}$ and $\sigma^2_{q(\mu_k^x)}$, $1 \le k \le K$

Keeping in mind that $\boldsymbol{\mu}^x = (\mu_1^x, \ldots, \mu_K^x)$, we begin with

$$\begin{aligned} p \, (\boldsymbol{\mu}^x | \text{rest}) &\propto p \left\{ \boldsymbol{x} | \boldsymbol{a}, \, \boldsymbol{\mu}^x, \, (\boldsymbol{\sigma}^x)^2 \right\} p \, (\boldsymbol{\mu}^x) \\ &= \prod_{k=1}^{K} \prod_{i=1}^{n} \left[\left\{ 2\pi \, (\sigma_k^x)^2 \right\}^{-1/2} \exp \left\{ - \frac{\left(x_i - \mu_k^x \right)^2}{2 \left(\sigma_k^x \right)^2} \right\} \right]^{a_{ik}} \\ &\quad \times \prod_{k=1}^{K} \left[\left(2\pi \sigma_\mu^2 \right)^{-1/2} \exp \left\{ - \frac{\left(\mu_k^x - \mu_\mu \right)^2}{\sigma_\mu^2} \right\} \right]. \end{aligned}$$

Taking the logarithm of both sides gives

$$\begin{aligned} \log p \, (\boldsymbol{\mu}^x \mid \text{rest}) &= -\tfrac{1}{2} \sum_{k=1}^{K} \sum_{i=1}^{n} a_{ik} \left\{ \frac{\left(\mu_k^x \right)^2 - 2 x_i \mu_k^x}{\left(\sigma_k^x \right)^2} \right\} - \frac{1}{2\sigma_\mu^2} \left\{ (\mu_k^x)^2 - 2 \mu_k^x \, \mu_\mu \right\} + \text{const} \\ &= -\tfrac{1}{2} \sum_{k=1}^{K} \left[(\mu_k^x)^2 \left\{ \frac{a_{\bullet k}}{\left(\sigma_k^x \right)^2} + \frac{1}{\sigma_\mu^2} \right\} - 2 \mu_k^x \left\{ \sum_{i=1}^{n} \frac{\sum_{i=1}^{n} a_{ik} x_i}{\left(\sigma_k^x \right)^2} + \frac{\mu_\mu}{\sigma_\mu^2} \right\} \right]. \end{aligned}$$

Taking expectations, we get

$$E_q\left\{\log\left(\boldsymbol{\mu}|\text{rest}\right)\right\} = -\tfrac{1}{2}\sum_{k=1}^{K}\left(\left(\mu_k^x\right)^2\left[\mu_{q\left\{1/\left(\sigma_k^x\right)^2\right\}}\mu_{q(a_{\bullet k})}+\frac{1}{\sigma_\mu^2}\right]\right.$$
$$\left.-2\mu_k^x\left[\mu_{q\left\{1/\left(\sigma_k^x\right)^2\right\}}\sum_{i=1}^{n}\mu_{q(a_{ik})}\left\{E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{x}\right)\right\}_i+\frac{\mu_\mu}{\sigma_\mu^2}\right]\right).$$

Therefore,

$$q^*\left(\boldsymbol{\mu}\right) \text{ is the product of } \mathrm{N}\left(\mu_{q\left(\mu_k^x\right)},\, \sigma^2_{q\left(\mu_k^x\right)}\right) \text{ density functions,}$$

$$\text{where}\quad \mu_{q\left(\mu_k^x\right)} = \sigma^2_{q\left(\mu_k^x s\right)}\left(\mu_{q\left(1/\left(\sigma_k^x\right)^2\right)}\sum_{i=1}^{n}\mu_{q(a_{ik})}\left\{E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{x}\right)\right\}_i+\frac{\mu_\mu}{\sigma_\mu^2}\right),\ \text{and}$$
$$\sigma^2_{q\left(\mu_k^x\right)} = 1\Big/\left(\tfrac{1}{\sigma_\mu^2}+\mu_{q\left(1/\left(\sigma_k^x\right)^2\right)}\mu_{q(a_{\bullet k})}\right),\qquad 1\le k\le K.$$

Expressions for $A_{\left\{\left(\sigma_k^x\right)^2\right\}}$, $B_{\left\{\left(\sigma_k^x\right)^2\right\}}$ and $\mu_{q\left\{1/\left(\sigma_k^x\right)^2\right\}}$, $1\le k\le K$

To begin, we note that

$$p\left\{\left(\boldsymbol{\sigma}^x\right)^2|\text{rest}\right\} \propto p\left(\boldsymbol{x}|\boldsymbol{a},\,\mu^x,\,\boldsymbol{\sigma}^x\right)p\left\{\left(\boldsymbol{\sigma}^x\right)^2|\boldsymbol{a}^x\right\}$$
$$= \prod_{k=1}^{K}\prod_{i=1}^{n}\left[\left\{2\pi\left(\sigma_k^x\right)^2\right\}^{-1/2}\exp\left\{-\frac{\left(x_i-\mu_k^x\right)^2}{2\left(\sigma_k^x\right)^2}\right\}\right]^{a_{ik}}.$$

Taking the logarithm of both sides gives

$$\log p\left\{\left(\boldsymbol{\sigma}^x\right)^2|\text{rest}\right\} = -\tfrac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{n}a_{ik}\left[\log\left\{\left(\sigma_k^x\right)^2\right\}+\frac{\left(x_i-\mu_k^x\right)^2}{\left(\sigma_k^x\right)^2}\right]-\sum_{k=1}^{K}\tfrac{3}{2}\log\left\{\left(\sigma_k^x\right)^2\right\}$$
$$-\sum_{k=1}^{K}\left(1/a_k^x\right)/\left(\sigma_k^x\right)^2+\text{const}$$
$$= -\sum_{k=1}^{K}\left\{\tfrac{1}{2}\left(a_{\bullet k}+1\right)+1\right\}\log\left\{\left(\sigma_k^x\right)^2\right\}$$
$$-\tfrac{1}{2}\sum_{k=1}^{K}\left(a_{\bullet k}\sum_{i=1}^{n}\left(x_i-\mu_k^x\right)^2+1/a_k^x\right)\Big/\left(\sigma_k^x\right)+\text{const.}$$

Taking expectations:

$$E_q\left[\log\left\{\left(\boldsymbol{\sigma}^x\right)^2|\text{rest}\right\}\right] = -\sum_{k=1}^{K}\left\{\tfrac{1}{2}\left(\mu_{q(a_{\bullet k})}+1\right)+1\right\}\log\left\{\left(\sigma_k^x\right)^2\right\}$$
$$-\tfrac{1}{2}\sum_{k=1}^{K}\left(\mu_{q(a_{\bullet k})}\sum_{i=1}^{n}E_q\left\{\left(x_i-\mu_k^x\right)^2\right\}+\mu_{q\left(1/a_k^x\right)}\right)\Big/\left(\sigma_k^x\right)+\text{const}$$

where

$$E_q\left\{(x_i-\mu_k^x)^2\right\}=\left\{\left(\left\{E_{q(\boldsymbol{x}_{\text{unobs}})}(\boldsymbol{x})\right\}_i-\mu_{q(\mu_k^x)}\right)^2+\left\{\sigma^2_{q(\boldsymbol{x}_{\text{unobs}})}(\boldsymbol{x})\right\}_i+\sigma^2_{q(\mu_k^x)}\right\}.$$

Therefore,

$$q^*\left\{(\boldsymbol{\sigma}^x)^2\right\} \text{ is the product of Inverse-Gamma}\left(A_{q\{(\sigma_k^x)\}},\, B_{q\{(\sigma_k^x)\}}\right) \text{ density functions,}$$

$$\begin{aligned}\text{where}\quad A_{q\{(\sigma_k^x)\}} &= \tfrac{1}{2}\mu_{q(a_{\bullet k})}+\tfrac{1}{2},\\ B_{q\{(\sigma_k^x)\}} &= \mu_{q(1/a_k^x)}+\tfrac{1}{2}\sum_{i=1}^{n}\mu_{q(a_{ik})}E_q\left\{(x_i-\mu_k^x)^2\right\}, 1\le k\le K.\end{aligned}$$

In addition, using Result 1.4.3,

$$\mu_{q\left\{1/(\sigma_k^x)^2\right\}}=A_{q\{(\sigma_k^x)\}}\Big/B_{q\{(\sigma_k^x)\}}.$$

Expressions for $B_{(a_k^x)}$ and $\mu_{q(1/a_k^x)}$, $1\le k\le K$

This derivation is similar to the one for $B_{q(\sigma_\varepsilon^2)}$ and $\mu_{q(1/\sigma_\varepsilon^2)}$, and so this gives

$$q^*(a_k^x) \text{ is the product of Inverse-Gamma}\left(1,\, B_{q(a_k^x)}\right) \text{ density functions,}$$

where

$$\begin{aligned}B_{q(a_k^x)} &= \mu_{q\left\{1/(\sigma_k^x)^2\right\}}+\left(A_k^x\right)^{-2}, \text{ and}\\ \mu_{q(1/a_k^x)} &= 1/B_{q(a_k^x)}, 1\le k\le K.\end{aligned}$$

Expressions for $\mu_{q(a_{ik})}$ and $\nu_{ik}, 1\le i\le n, 1\le k\le K$

Firstly, we note that

$$p(\boldsymbol{a}|\text{rest})=\prod_{i=1}^{n}p\left(\boldsymbol{a}_{i1},\ldots,\boldsymbol{a}_{iK}|\text{rest}\right),$$

where each $p\left(\boldsymbol{a}_{i1},\ldots,\boldsymbol{a}_{iK}|\text{rest}\right)$ has a Multinomial$\left(1;\omega_1,\ldots,\omega_K\right)$ distribution. Next, we work with each ith component of this distribution.

$$\begin{aligned}p\left(\boldsymbol{a}_{i1},\ldots,\boldsymbol{a}_{iK}|\text{rest}\right) &\propto p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}p\left(\boldsymbol{a}_{i1},\ldots,\boldsymbol{a}_{iK}|\omega_1,\ldots,\omega_K\right)\\ &= \prod_{k=1}^{K}\left[\left\{2\pi\left(\sigma_k^x\right)^2\right\}^{-1/2}\exp\left\{-\frac{\left(x_i-\mu_k^x\right)^2}{\left(\sigma_k^x\right)^2}\right\}\right]^{a_{ik}}\times\prod_{k=1}^{K}\frac{\omega_k^{a_{ik}}}{a_{ik}!}\end{aligned}$$

Taking the logarithm of both sides gives

$$\log p\left(\boldsymbol{a}_{i1},\ldots,\boldsymbol{a}_{iK}|\text{rest}\right) = \sum_{k=1}^{K} a_{ik}\left(\log\left(\omega_k\right) - \frac{1}{2}\left[\log\left(2\pi\right) + \log\left\{\left(\sigma_k^x\right)^2\right\} + \frac{\left(x_i-\mu_k^x\right)^2}{\left(\sigma_k^x\right)^2}\right]\right) + \text{const.}$$

Taking expectations, we get

$$E_q\left\{\log p\left(\boldsymbol{a}_{i1},\ldots,\boldsymbol{a}_{iK}|\text{rest}\right)\right\} = \sum_{k=1}^{K} a_{ik}\left(E_q\left\{\log\left(\omega_k\right)\right\} - \frac{1}{2}\left[\log\left(2\pi\right) + E_q\left\{\log\left(\sigma_k^x\right)^2\right\} + E_q\left\{\frac{\left(x_i-\mu_k^x\right)^2}{\left(\sigma_k^x\right)^2}\right\}\right]\right) + \text{const.}$$

Using Results 1.4.3 and 1.4.10, this gives

$$\begin{aligned}
&E_q\left\{\log p\left(\boldsymbol{a}_{i1},\ldots,\boldsymbol{a}_{iK}|\text{rest}\right)\right\} \\
&\quad= \sum_{k=1}^{K} a_{ik}\left(\psi\left(\alpha_{q(\omega_k)}\right) - \psi\left(\sum_{k=1}^{K}\alpha_{q(\omega_k)}\right) - \frac{1}{2}\left\{\log(2\pi) + \log\left(B_{q\left\{\left(\sigma_k^x\right)^2\right\}}\right)\right.\right. \\
&\qquad -\psi\left(A_{q\left\{\left(\sigma_k^x\right)^2\right\}}\right) + \mu_{q\left\{1/\left(\sigma_k^x\right)^2\right\}}\left(\left\{E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{x}^\top\boldsymbol{x}\right)\right\}_i + \left\{\sigma^2_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{x}\right)\right\}_i\right. \\
&\qquad \left.\left.\left. -2\left\{E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{x}\right)\right\}_i \boldsymbol{\mu}_{q(\mu_k^x)} + \boldsymbol{\mu}^2_{q(\mu_k^x)} + \sigma^2_{q(\boldsymbol{\mu}_k^x)}\right)\right\}\right) + \text{const} \\
&\quad= \sum_{k=1}^{K} a_{ik}\left(\nu_{ik} + \underline{\text{const}}\right) + \text{const}
\end{aligned}$$

where $\underline{const}$ does not depend on k (e.g. $\underline{const}$ includes $-\frac{1}{2}\log(2\pi)$ and $-\psi\left(\sum_{k=1}^{K}\alpha_{q(\omega_k)}\right)$). Note that the 'proportional to' sign $\propto$ allows us to get rid of 'const' below.

$$\begin{aligned}
q^*(\boldsymbol{a}_{i1},\ldots,\boldsymbol{a}_{iK}) &\propto \prod_{k=1}^{K}\left(e^{\nu_{ik}+\underline{const}}\right)^{a_{ik}} \\
&\propto \prod_{k=1}^{K}\left(\frac{\exp^{\nu_{ik}+\underline{const}}}{\sum_{k'=1}^{K} e^{\nu_{ik}+\underline{const}}}\right)^{a_{ik}},
\end{aligned}$$

since $\mathcal{C} = \frac{1}{\sum_{k'=1}^{K} e^{\nu_{ik}+\underline{const}}}$ does not depend on k, and $\prod_{k=1}^{K}\left(\frac{1}{\mathcal{C}}\right)^{a_{ik}} = \frac{1}{\mathcal{C}}$ given that $\sum_{k=1}^{K} a_{ik} = 1$. Hence the cancellation of $\underline{const}$. Therefore,

$$q^*\left(\boldsymbol{a}_{i1},\ldots,\boldsymbol{a}_{iK}\right) \text{ is the Multinomial}\left(1\,;\,\mu_{q(a_{i1})},\ldots,\mu_{q(a_{iK})}\right)\text{distribution,}$$

where

$$\mu_{q(a_{ik})} = \frac{\exp^{\nu_{ik}}}{\sum_{k'=1}^{K} e^{\nu_{ik}}}, \quad 1 \le k \le K$$

and

$$\begin{aligned}\nu_{ik} \;=\; & \psi\left(\alpha_{q(\omega_k)}\right) + \tfrac{1}{2}\psi\left(A_{q\left\{(\sigma_k^x)^2\right\}}\right) - \tfrac{1}{2}\log\left(B_{q\left\{(\sigma_k^x)^2\right\}}\right) \\ & -\tfrac{1}{2}A_{q\left\{(\sigma_k^x)^2\right\}}\left(\left\{E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{x}^\top\boldsymbol{x}\right)\right\}_i + \left\{\sigma^2_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{x}\right)\right\}_i\right. \\ & \left.-2\left\{E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{x}\right)\right\}_i \boldsymbol{\mu}_{q(\mu_k^x)} + \boldsymbol{\mu}^2_{q(\mu_k^x)} + \sigma^2_{q(\boldsymbol{\mu}_k^x)}\right) \Big/ B_{q\left\{(\sigma_k^x)^2\right\}}.\end{aligned}$$

Expression for $\alpha_{q(\omega_k)}, 1 \le k \le K$

We begin with

$$\begin{aligned}p\left(\omega_1,\ldots,\omega_K|\text{rest}\right) \;&\propto\; p(\omega_1,\ldots,\omega_K) \times \prod_{i=1}^{n} p(\boldsymbol{a}_{i1},\ldots,\boldsymbol{a}_{iK}|\omega_1,\ldots,\omega_K) \\ &=\; \prod_{i=1}^{n}\prod_{k=1}^{K}\frac{\omega_k^{a_{ik}}}{a_{ik}!} \times \frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\Gamma(\alpha_k)}\left(\omega_k\right)^{\alpha_k-1}.\end{aligned}$$

Taking the logarithm of both sides gives

$$\begin{aligned}\log p\left(\omega_1,\ldots,\omega_K|\text{rest}\right) \;&=\; \sum_{i=1}^{n}\sum_{k=1}^{K}\left(a_{ik}\log(\omega_k) - \log(a_{ik}!)\right) + \log\left(\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)\right) \\ &\quad -\sum_{k=1}^{K}\log\left(\Gamma\left(\alpha_k\right)\right) + \sum_{k=1}^{K}\left(\alpha_k-1\right)\log\left(\omega_k\right) + \text{const} \\ &=\; \sum_{i=1}^{n}\sum_{k=1}^{K}a_{ik}\log\left(\omega_k\right) + \sum_{k=1}^{K}\left(\alpha_k-1\right)\log\left(\omega_k\right) + \text{const}.\end{aligned}$$

Taking expectations, we get

$$E_q\left\{\log p\left(\omega_1,\ldots,\omega_K|\text{rest}\right)\right\} \;=\; \sum_{i=1}^{n}\sum_{k=1}^{K}\mu_{q(a_{ik})}\log\left(\omega_k\right) + \sum_{k=1}^{K}\left(\alpha_k-1\right)\log\left(\omega_k\right).$$

Therefore,

$$q^*\left(\omega_1,\ldots,\omega_K\right) \text{ is the Dirichlet}\left(\alpha_{q(\omega_1)},\cdots,\alpha_{q(\omega_k)}\right) \text{ density function,}$$

where

$$\alpha_{q(\omega_k)} = \mu_{q(a_{\cdot k})} + \alpha_k, \quad 1 \le k \le K.$$

6.C Derivation of the marginal log-likelihood lower bound

The expression for the lower bound on the marginal log-likelihood given in (6.7) is

$$
\begin{aligned}
&\log \underline{p}(\boldsymbol{y}; q) = \\
&\quad -\tfrac{1}{2}(K+5)\log(\pi) + \tfrac{1}{2}(p+q+K+n_{\text{unobs}}) - \tfrac{p}{2}\log\left(\sigma_\beta^2\right) + \tfrac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}| - \log(A_\varepsilon) \\
&\quad -\tfrac{1}{2}\left(2n + n_{\text{unobs}} + \sum_{k=1}^{K}\sum_{i=1}^{n}\mu_{q(a_{ik})}\right)\log(2\pi) - \tfrac{1}{2\sigma_\beta^2}\left\{\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \operatorname{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right)\right\} + \log\Gamma\left\{\tfrac{1}{2}(n+1)\right\} \\
&\quad -\tfrac{1}{2}(n+1)\log\left(B_{q(\sigma_\varepsilon^2)}\right) - \log\left(B_{q(a_\varepsilon)}\right) + \mu_{q(1/\sigma_\varepsilon^2)}\mu_{q(1/a_\varepsilon)} - \tfrac{q}{2}\log\left(\sigma_\alpha^2\right) + \tfrac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\alpha})}| \\
&\quad -\log(A_o) - \tfrac{1}{2\sigma_\alpha^2}\left\{\|\boldsymbol{\mu}_{q(\boldsymbol{\alpha})}\|^2 + \operatorname{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\alpha})}\right)\right\} - \sum_{k=1}^{K}\log\left(A_k^x\right) + \log\Gamma\left\{\tfrac{1}{2}(n+1)\right\} \\
&\quad -\tfrac{K}{2}\log\left(\sigma_\mu^2\right) - \tfrac{1}{2}(n+1)\log\left(B_{q(\sigma_o^2)}\right) - \log\left(B_{q(a_o)}\right) + \mu_{q(1/\sigma_o^2)}\mu_{q(1/a_o)} \\
&\quad +\tfrac{n_{\text{unobs}}}{2}\log\left(\sigma^2_{q(\boldsymbol{x}_{\text{unobs}})}\right) - \tfrac{1}{2\sigma_\mu^2}\sum_{k=1}^{K}\left\{\left(\mu_{q(\mu_k^x)} - \mu_\mu\right)^2 + \sigma^2_{q(\mu_k^x)}\right\} + \tfrac{1}{2}\sum_{k=1}^{K}\log\left(\sigma^2_{q(\mu_k^x)}\right) \\
&\quad +\sum_{k=1}^{K}\log\Gamma\left\{\tfrac{1}{2}\left(\mu_{q(a_{\bullet k})}+1\right)\right\} - \tfrac{1}{2}\sum_{k=1}^{K}\left(\mu_{q(a_{\bullet k})}+1\right)\log\left(B_{q\left\{(\sigma_k^x)^2\right\}}\right) - \sum_{k=1}^{K}\log\left(B_{q(a_k^x)}\right) \\
&\quad +\sum_{k=1}^{K}\mu_{q\left\{1/(\sigma_k^x)^2\right\}}\mu_{q(1/a_k^x)} - \sum_{i=1}^{n}\sum_{k=1}^{K}\log\left(\mu_{q(a_{ik})}\right)\mu_{q(a_{ik})} - \log\Gamma\left(\sum_{k=1}^{K}\alpha_{q(\omega_k)}\right) \\
&\quad +\log\Gamma\left(\sum_{k=1}^{K}\alpha_k\right) - \sum_{k=1}^{K}\log\Gamma(\alpha_k) + \sum_{k=1}^{K}\log\Gamma\left(\alpha_{q(\omega_k)}\right).
\end{aligned}
$$

Derivation: The lower bound on the marginal log-likelihood is achieved through the following expression:

$$
\begin{aligned}
\log \underline{p}(\boldsymbol{y}; q) \;=\; & E_q\left[\log p\left\{\boldsymbol{y}, \boldsymbol{\beta}, \sigma_\varepsilon^2, a_\varepsilon, \boldsymbol{o}, \boldsymbol{\alpha}, \sigma_o^2, a_o, \boldsymbol{x}, \boldsymbol{\mu}^x, (\boldsymbol{\sigma}^x)^2, \boldsymbol{a}^x, \boldsymbol{a}, \boldsymbol{\omega}\right\}\right. \\
& \left. -\log q^*\left\{\boldsymbol{\beta}, \sigma_\varepsilon^2, a_\varepsilon, \boldsymbol{\alpha}, \sigma_o^2, a_o, \boldsymbol{x}_{\text{unobs}}, \boldsymbol{\mu}^x, (\boldsymbol{\sigma}^x)^2, \boldsymbol{a}^x, \boldsymbol{a}, \boldsymbol{\omega}\right\}\right] \\
=\; & E_q\left\{\log p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{x}, \sigma_\varepsilon^2)\right\} + E_q\left\{\log p(\boldsymbol{o}|\boldsymbol{\alpha}, \boldsymbol{x}, \sigma_o^2)\right\} \\
& + E_q\left[\log p\left\{\boldsymbol{x}_{\text{obs}}|\boldsymbol{a}, \boldsymbol{\mu}^x, (\boldsymbol{\sigma}^x)^2\right\}\right] + E_q\left\{\log p(\boldsymbol{\beta}) - \log q^*(\boldsymbol{\beta})\right\} \\
& + E_q\left\{\log 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{\alpha}) - \log q^*(\boldsymbol{\alpha})\right\} + E_q\left\{\log p(\sigma_o^2|a_o) - \log q^*(\sigma_o^2)\right\} \\
& + E_q\left\{\log p(a_o) - \log q^*(a_o)\right\} + E_q\left\{\log p(\boldsymbol{\mu}^x) - \log q^*(\boldsymbol{\mu}^x)\right\} \\
& + E_q\left[\log p\left\{\boldsymbol{x}_{\text{unobs}}|\boldsymbol{a}, \boldsymbol{\mu}^x, (\boldsymbol{\sigma}^x)^2\right\} - \log q^*(\boldsymbol{x}_{\text{unobs}})\right] \\
& + E_q\left[\log p\left\{(\boldsymbol{\sigma}^x)^2|\boldsymbol{a}^x\right\} - \log q^*\left\{(\boldsymbol{\sigma}^x)^2\right\}\right] + E_q\left\{\log p(\boldsymbol{a}^x) - \log q^*(\boldsymbol{a}^x)\right\} \\
& + E_q\left\{\log\prod_{i=1}^{n} p\left(a_{i1}, \ldots, a_{iK}\,|\,\omega_1, \ldots, \omega_K\right) - \log\prod_{i=1}^{n} q^*(a_{i1}, \ldots, a_{iK})\right\} \\
& + E_q\left\{\log p(\omega_1, \ldots, \omega_K) - \log q^*(\omega_1, \ldots, \omega_K)\right\}.
\end{aligned}
$$

First we note that

$$\log p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{x}, \sigma_\varepsilon^2) = -\tfrac{n}{2}\log(2\pi) - \tfrac{n}{2}\log(\sigma_\varepsilon^2) - \tfrac{1}{2\sigma_\varepsilon^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2.$$

Therefore,

$$E_q\left\{\log p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{x}, \sigma_\varepsilon^2)\right\} = -\tfrac{n}{2}\log(2\pi) - \tfrac{n}{2}E_q\left\{\log(\sigma_\varepsilon^2)\right\} - \tfrac{1}{2}\mu_{q(1/\sigma_\varepsilon^2)}E_q\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$$

where

$$\begin{aligned} E_q\|\boldsymbol{y} - \boldsymbol{X}\,\boldsymbol{\beta}\|^2 = {} & \|\boldsymbol{y}_{x\text{obs}} - \boldsymbol{X}_{x\text{obs}}\,\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \operatorname{tr}\left(\boldsymbol{X}_{x\text{obs}}^\top \boldsymbol{X}_{x\text{obs}}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right) \\ & + \|\boldsymbol{y}_{x\text{unobs}}\|^2 - 2\boldsymbol{y}_{x\text{unobs}}^\top E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{X}_{x\text{unobs}}\right)\boldsymbol{\mu}_{q(\boldsymbol{\beta})} \\ & + \operatorname{tr}\left[E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{X}_{x\text{unobs}}^\top \boldsymbol{X}_{x\text{unobs}}\right)\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} + \boldsymbol{\mu}_{q(\boldsymbol{\beta})}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}^\top\right)\right]. \end{aligned}$$

Similarly,

$$E_q\left\{\log p(\boldsymbol{o}|\boldsymbol{\alpha}, \boldsymbol{x}, \sigma_o^2)\right\} = -\tfrac{n}{2}\log(2\pi) - \tfrac{n}{2}E_q\left\{\log(\sigma_o^2)\right\} - \tfrac{1}{2}\mu_{q(1/\sigma_o^2)}E_q\|\boldsymbol{o} - \tilde{\boldsymbol{X}}\boldsymbol{\alpha}\|^2$$

where

$$\begin{aligned} E_q\|\boldsymbol{o} - \tilde{\boldsymbol{X}}\,\boldsymbol{\alpha}\|^2 = {} & \|\boldsymbol{o}_{x\text{obs}} - \tilde{\boldsymbol{X}}_{x\text{obs}}\,\boldsymbol{\mu}_{q(\boldsymbol{\alpha})}\|^2 + \operatorname{tr}\left(\tilde{\boldsymbol{X}}_{x\text{obs}}^\top \tilde{\boldsymbol{X}}_{x\text{obs}}\boldsymbol{\Sigma}_{q(\boldsymbol{\alpha})}\right) \\ & + \|\boldsymbol{o}_{x\text{unobs}}\|^2 - 2\boldsymbol{o}_{x\text{unobs}}^\top E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\tilde{\boldsymbol{X}}_{x\text{unobs}}\right)\boldsymbol{\mu}_{q(\boldsymbol{\alpha})} \\ & + \operatorname{tr}\left[E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\tilde{\boldsymbol{X}}_{x\text{unobs}}^\top \tilde{\boldsymbol{X}}_{x\text{unobs}}\right)\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\alpha})} + \boldsymbol{\mu}_{q(\boldsymbol{\alpha})}\boldsymbol{\mu}_{q(\boldsymbol{\alpha})}^\top\right)\right]. \end{aligned}$$

Next,

$$\begin{aligned} \log p\left\{\boldsymbol{x}_{\text{obs}}|\boldsymbol{a}, \boldsymbol{\mu}^x, (\boldsymbol{\sigma}^x)^2\right\} &= -\tfrac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{n_{\text{obs}}} a_{ik}\left[\log\left\{2\pi\left(\sigma_k^x\right)^2\right\} + \frac{\left(x_i - \mu_k^x\right)^2}{\left(\sigma_k^x\right)^2}\right] \\ &= -\tfrac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{n_{\text{obs}}} a_{ik}\log(2\pi) - \tfrac{1}{2}\sum_{k=1}^{K} a_{\bullet k}\log\left\{\left(\sigma_k^x\right)^2\right\} \\ &\quad -\tfrac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{n_{\text{obs}}} a_{ik}\frac{\left(x_i - \mu_k^x\right)}{\left(\sigma_k^x\right)^2}. \end{aligned}$$

Taking expectations of both sides gives

$$\begin{aligned} E_q\left[\log p\left\{\boldsymbol{x}_{\text{obs}}|\boldsymbol{a}, \boldsymbol{\mu}^x, (\boldsymbol{\sigma}^x)^2\right\}\right] = {} & -\tfrac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{n_{\text{obs}}}\mu_{q(a_{ik})}\log(2\pi) - \tfrac{1}{2}\sum_{k=1}^{K}\mu_{q(a_{\bullet k})}E_q\left[\log\left\{\left(\sigma_k^x\right)^2\right\}\right] \\ & -\tfrac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{n_{\text{obs}}}\mu_{q(a_{ik})}\mu_{q\left\{1/\left(\sigma_k^x\right)^2\right\}}E_q\left\{\left(x_i - \mu_k^x\right)\right\} \end{aligned}$$

where

$$E_q\left\{\left(x_i - \mu_k^x\right)^2\right\} = \left\{\left(x_i - \mu_{q(\mu_k^x)}\right)^2 + \sigma_{q(\mu_k^x)}^2\right\}.$$

Moving on to the next expectation update, we have

$$\log p(\boldsymbol{\beta}) = -\tfrac{3}{3}\log(2\pi) - \tfrac{3}{2}\log(\sigma_\beta^2) - \tfrac{1}{2\sigma_\beta^2}\|\boldsymbol{\beta}\|^2$$

Taking the expectation of both sides gives:

$$\begin{aligned} E_q\left\{\log p(\boldsymbol{\beta})\right\} &= -\tfrac{p}{2}\log(2\pi) - \tfrac{p}{2}\log(\sigma_\beta^2) - \tfrac{1}{2\sigma_\beta^2}E_q\left\{\|\boldsymbol{\beta}\|^2\right\} \\ &= -\tfrac{p}{2}\log(2\pi) - \tfrac{p}{2}\log(\sigma_\beta^2) - \tfrac{1}{2\sigma_\beta^2}\left\{\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \operatorname{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right)\right\}. \end{aligned}$$

The entropy of $q^*(\boldsymbol{\beta})$ is

$$\begin{aligned} -E_q\left\{\log q^*(\boldsymbol{\beta})\right\} &= \tfrac{1}{2}\log\left\{(2\pi e)^p\left|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right|\right\} \\ &= \tfrac{1}{2}\log\left|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right| + \tfrac{p}{2}\log(2\pi) + \tfrac{p}{2}. \end{aligned}$$

Therefore

$$E_q\left\{\log p(\boldsymbol{\beta}) - \log q^*(\boldsymbol{\beta})\right\} = -\tfrac{p}{2}\log(\sigma_\beta^2) - \tfrac{1}{2\sigma_\beta^2}\left\{\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \operatorname{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right)\right\} + \tfrac{1}{2}\log\left|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right| + \tfrac{p}{2}.$$

Next,

$$\begin{aligned} \log p(\sigma_\varepsilon^2|a_\varepsilon) &= -\tfrac{1}{2}\log(a_\varepsilon) - \log\Gamma\left(\tfrac{1}{2}\right) - \tfrac{3}{2}\log(\sigma_\varepsilon^2) - (1/a_\varepsilon)/\sigma_\varepsilon^2 \\ &= -\tfrac{1}{2}\log(a_\varepsilon) - \tfrac{1}{2}\log(\pi) - \tfrac{3}{2}\log(\sigma_\varepsilon^2) - (1/a_\varepsilon)/\sigma_\varepsilon^2. \end{aligned}$$

Taking expectations, we get

$$E_q\left\{\log p(\sigma_\varepsilon^2|a_\varepsilon)\right\} = -\tfrac{1}{2}E_q\left\{\log(a_\varepsilon)\right\} - \tfrac{1}{2}\log(\pi) - \tfrac{3}{2}E_q\left\{\log(\sigma_\varepsilon^2)\right\} - \mu_{q(1/a_\varepsilon)}\mu_{q(1/\sigma_\varepsilon^2)}.$$

Also,

$$\begin{aligned} E_q\left\{\log q^*(\sigma_\varepsilon^2)\right\} = \tfrac{1}{2}(n+1)\log B_{q(\sigma_\varepsilon^2)} - \log\Gamma\left\{\tfrac{1}{2}(n+1)\right\} - \left\{\tfrac{1}{2}(n+1)+1\right\}\log(\sigma_\varepsilon^2) \\ -B_{q(\sigma_\varepsilon^2)}/\sigma_\varepsilon^2. \end{aligned}$$

Therefore,

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

Next we have

$$\log p(a_\varepsilon) = -\log(A_\varepsilon) - \tfrac{3}{2}\log(a_\varepsilon) - \tfrac{1}{2}\log(\pi) - \left(1/A_\varepsilon^2\right)/a_\varepsilon.$$

Taking the expectation of both sides gives

$$E_q\left\{\log p(a_\varepsilon)\right\} = -\log(A_\varepsilon) - \tfrac{3}{2}E_q\left\{\log(a_\varepsilon)\right\} - \tfrac{1}{2}\log(\pi) - \left(1/A_\varepsilon^2\right)\mu_{q(1/a_\varepsilon)}.$$

Also,

$$E_q\left\{\log q^*(a_\varepsilon)\right\} = \log\left(B_{q(a_\varepsilon)}\right) - 2E_q\left\{\log(a_\varepsilon)\right\} - B_{q(a_\varepsilon)}\mu_{q(1/a_\varepsilon)}.$$

Therefore,

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

Similarly to $E_q\left\{\log p(\boldsymbol{\beta}) - \log q^*(\boldsymbol{\beta})\right\}$,

$$E_q\left\{\log p(\boldsymbol{\alpha}) - \log q^*(\boldsymbol{\alpha})\right\} = -\tfrac{q}{2}\log(\sigma_\alpha^2) - \tfrac{1}{2\sigma_\alpha^2}\left\{\|\boldsymbol{\mu}_{q(\boldsymbol{\alpha})}\|^2 + \operatorname{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\alpha})}\right)\right\} + \tfrac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\alpha})}| + \tfrac{q}{2}.$$

Similarly to $E_q\left\{\log p(\sigma_\varepsilon^2|a_\varepsilon) - \log q^*(\sigma_\varepsilon^2)\right\}$,

$$\begin{aligned} E_q\left\{\log p(\sigma_o^2|a_o) - \log q^*(\sigma_o^2)\right\} = & -\tfrac{1}{2}E_q\left\{\log(a_o)\right\} + \log\Gamma\left\{\tfrac{1}{2}(n+1)\right\} - \tfrac{1}{2}log(\pi) \\ & + \left(B_{q(\sigma_o^2)} - \mu_{q(1/a_o)}\right)\mu_{q(1/\sigma_o^2)} + \tfrac{n}{2}E_q\left\{\log(\sigma_o^2)\right\} \\ & - \tfrac{1}{2}(n+1)\log\left(B_{q(\sigma_o^2)}\right). \end{aligned}$$

Similarly to $E_q\left\{\log p(a_\varepsilon) - \log q^*(a_\varepsilon)\right\}$,

$$\begin{aligned} E_q\left\{\log p(a_o) - \log q^*(a_o)\right\} = & -\log(A_o) - \tfrac{1}{2}\log(\pi) - \log\left(B_{q(a_o)}\right) \\ & + \left(B_{q(a_o)} - 1/A_o^2\right)\mu_{q(1/a_o)} + \tfrac{1}{2}E_q\left\{\log(a_o)\right\}. \end{aligned}$$

Next,

$$\begin{aligned} \log p\left\{\boldsymbol{x}_{\text{unobs}}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\} = & -\tfrac{1}{2}\sum_{k=1}^{K}\sum_{i=n_{\text{obs}}+1}^{n} a_{ik}\log(2\pi) - \tfrac{1}{2}\sum_{k=1}^{K} a_{\bullet k}\log\left\{\left(\sigma_k^x\right)^2\right\} \\ & - \tfrac{1}{2}\sum_{k=1}^{K}\sum_{i=n_{\text{obs}}+1}^{n} a_{ik}\frac{\left(x_i-\mu_k^x\right)}{\left(\sigma_k^x\right)^2}. \end{aligned}$$

Taking expectations of both sides gives

$$\begin{aligned} E_q\left[\log p\left\{\boldsymbol{x}_{\text{unobs}}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\right] = & -\tfrac{1}{2}\sum_{k=1}^{K}\sum_{i=n_{\text{obs}}+1}^{n}\mu_{q(a_{ik})}\log(2\pi) - \tfrac{1}{2}\sum_{k=1}^{K}\mu_{q(a_{\bullet k})}E_q\left[\log\left\{\left(\sigma_k^x\right)^2\right\}\right] \\ & - \tfrac{1}{2}\sum_{k=1}^{K}\sum_{i=n_{\text{obs}}+1}^{n}\mu_{q(a_{ik})}\mu_{q\left\{1/\left(\sigma_k^x\right)^2\right\}}E_q\left\{\left(x_i-\mu_k^x\right)\right\} \end{aligned}$$

where

$$E_q\left\{\left(x_i-\mu_k^x\right)^2\right\} = \left\{\left(\left\{E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{x}\right)\right\}_i - \mu_{q(\mu_k^x)}\right)^2 + \left\{\sigma^2_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{x}\right)\right\}_i + \sigma^2_{q(\mu_k^x)}\right\}.$$

The entropy of $q^*(\boldsymbol{x}_{\text{unobs}})$ is

$$-E_q\left\{\log q^*(\boldsymbol{x}_{\text{unobs}})\right\} = \tfrac{1}{2}\log|\sigma^2_{q(\boldsymbol{x}_{\text{unobs}})}\boldsymbol{I}_{n_{\text{unobs}}}| + \tfrac{n_{\text{unobs}}}{2}\log(2\pi) + \tfrac{n_{\text{unobs}}}{2}.$$

Therefore,

$$\begin{aligned}
&E_q\left[\log p\left\{\boldsymbol{x}_{\text{unobs}}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\} - \log q^*(\boldsymbol{x}_{\text{unobs}})\right] = \\
&\quad -\tfrac{1}{2}\sum_{k=1}^{K}\sum_{n_{\text{obs}}+1}^{n}\mu_{q(a_{ik})}\log(2\pi) - \tfrac{1}{2}\sum_{k=1}^{K}\mu_{q(a_{\bullet k})}E_q\left[\log\left\{(\sigma^x_k)^2\right\}\right] \\
&\quad -\tfrac{1}{2}\sum_{k=1}^{K}\sum_{i=n_{\text{obs}}+1}^{n}\mu_{q(a_{ik})}\mu_{q\left\{1/(\sigma^x_k)^2\right\}}E_q\left\{(x_i-\mu^x_k)\right\} + \tfrac{n_{\text{unobs}}}{2}\log(\sigma^2_{q(\boldsymbol{x}_{\text{unobs}})}) \\
&\quad +\tfrac{n_{\text{unobs}}}{2}\log(2\pi) + \tfrac{n_{\text{unobs}}}{2}.
\end{aligned}$$

Moving on, we next have

$$\log p(\boldsymbol{\mu}^x) = -\tfrac{K}{2}\log(2\pi) - \tfrac{K}{2}\log(\sigma^2_\mu) - \tfrac{1}{2\sigma^2_\mu}\sum_{k=1}^{K}\|\mu^x_k-\mu_\mu\|^2.$$

Taking expectations, we get

$$\begin{aligned}
E_q\left\{\log p(\boldsymbol{\mu}^x)\right\} &= -\tfrac{K}{2}\log(2\pi) - \tfrac{K}{2}\log(\sigma^2_\mu) - \tfrac{1}{2\sigma^2_\mu}\sum_{k=1}^{K}E_q\left\{\|\mu^x_k-\mu_\mu\|^2\right\} \\
&= -\tfrac{K}{2}\log(2\pi) - \tfrac{K}{2}\log(\sigma^2_\mu) - \tfrac{1}{2\sigma^2_\mu}\sum_{k=1}^{K}\left\{\|\mu_{q(\mu^x_k)}-\mu_\mu\|^2 + \sigma_{q(\mu^x_k)}\right\}.
\end{aligned}$$

The entropy of $q^*(\boldsymbol{\mu}^x)$ is

$$-E_q\left\{\log q^*(\boldsymbol{\mu}^x)\right\} = \tfrac{K}{2} + \tfrac{K}{2}\log(2\pi) + \tfrac{1}{2}\sum_{k=1}^{K}\log\left(\sigma_{q(\mu^x_k)}\right).$$

Therefore,

$$\begin{aligned}
E_q\left\{\log p(\boldsymbol{\mu}^x) - \log q^*(\boldsymbol{\mu}^x)\right\} &= -\tfrac{K}{2}\log(\sigma^2_\mu)\tfrac{1}{2\sigma^2_\mu}\sum_{k=1}^{K}\left\{\|\mu_{q(\mu^x_k)}-\mu_\mu\|^2 + \sigma_{q(\mu^x_k)}\right\} + \tfrac{K}{2} \\
&\quad +\tfrac{1}{2}\sum_{k=1}^{K}\log\left(\sigma_{q(\mu^x_k)}\right).
\end{aligned}$$

Next,

$$\log p\left\{(\boldsymbol{\sigma}^x)^2|\boldsymbol{a}^x\right\} = -\tfrac{1}{2}\sum_{k=1}^{K}\log\left(a^x_k\right) - \tfrac{K}{2}\log(\pi) - \tfrac{3}{2}\sum_{k=1}^{K}\log\left\{(\sigma^x_k)^2\right\} - \sum_{k=1}^{K}\left(1/a^x_k\right)/(\sigma^x_k)^2$$

Taking expectations, we get

$$\begin{aligned}
E_q\left[\log p\left\{(\boldsymbol{\sigma}^x)^2|\boldsymbol{a}^x\right\}\right] &= -\tfrac{1}{2}\sum_{k=1}^{K}E_q\left\{\log\left(a^x_k\right)\right\} - \tfrac{K}{2}\log(\pi) - \tfrac{3}{2}\sum_{k=1}^{K}E_q\left[\log\left\{(\sigma^x_k)^2\right\}\right] \\
&\quad -\sum_{k=1}^{K}\mu_{q(1/a^x_k)}\mu_{q\left\{1/(\sigma^x_k)^2\right\}}.
\end{aligned}$$

Also,

$$\begin{aligned} E_q\left[\log q^*\left\{\left(\boldsymbol{\sigma}^x\right)^2\right\}\right] = & \tfrac{1}{2}\sum_{k=1}^{K}\left(\mu_{q(a_{\bullet k})}+1\right)\log\left(B_{q\left\{\left(\sigma_k^x\right)^2\right\}}\right)-\sum_{k=1}^{K}\log\Gamma\left\{\tfrac{1}{2}\left(\mu_{q(a_{\bullet k})}+1\right)\right\} \\ & -\sum_{k=1}^{K}\left\{\tfrac{1}{2}\left(\mu_{q(a_{\bullet k})}+1\right)+1\right\}E_q\left[\log\left\{\left(\sigma_k^x\right)^2\right\}\right] \\ & -B_{q\left\{\left(\sigma_k^x\right)^2\right\}}\mu_{q\left\{1/\left(\sigma_k^x\right)^2\right\}}. \end{aligned}$$

Therefore,

$$\begin{aligned} & E_q\left[\log p\left\{\left(\boldsymbol{\sigma}^x\right)^2|\boldsymbol{a}^x\right\}-\log q^*\left\{\left(\boldsymbol{\sigma}^x\right)^2\right\}\right] \\ & \quad = -\tfrac{1}{2}\sum_{k=1}^{K}E_q\left\{\log\left(a_k^x\right)\right\}-\tfrac{K}{2}\log(\pi)+\sum_{k=1}^{K}\log\Gamma\left\{\tfrac{1}{2}\left(\mu_{q(a_{\bullet k})}+1\right)\right\} \\ & \qquad +\tfrac{1}{2}\sum_{k=1}^{K}\mu_{q(a_{\bullet k})}E_q\left[\log\left\{\left(\sigma_k^x\right)^2\right\}\right]-\tfrac{1}{2}\sum_{k=1}^{K}\left(\mu_{q(a_{\bullet k})}+1\right)\log\left(B_{q\left\{\left(\sigma_k^x\right)^2\right\}}\right) \\ & \qquad +\sum_{k=1}^{K}\mu_{q\left\{1/\left(\sigma_k^x\right)^2\right\}}\left\{B_{q\left\{\left(\sigma_k^x\right)^2\right\}}-\mu_{q\left(1/a_k^x\right)}\right\}. \end{aligned}$$

Similarly to $E_q\left\{\log p(a_\varepsilon)-\log q^*(a_\varepsilon)\right\}$,

$$\begin{aligned} E_q\left\{\log p(\boldsymbol{a}^x)-\log q^*(\boldsymbol{a}^x)\right\} = & -\sum_{k=1}^{K}\log(A_k^x)-\tfrac{1}{2}\log(\pi)-\sum_{k=1}^{K}\log\left(B_{q(a_k^x)}\right) \\ & +\sum_{k=1}^{K}\left(B_{q(a_k^x)}-1/\left(A_k^x\right)^2\right)\mu_{q\left(1/a_k^x\right)}+\tfrac{1}{2}\sum_{k=1}^{K}E_q\left\{\log\left(a_k^x\right)\right\}. \end{aligned}$$

Moving on, we have

$$\log\prod_{i=1}^{n}p\left(a_{i1},\ldots,a_{iK}\,|\,\omega_1,\ldots,\omega_K\right)=\sum_{i=1}^{n}\sum_{k=1}^{K}\left\{a_{ik}\log(\omega_k)-\log(a_{ik})\right\}.$$

Taking expectations of both sides, we get

$$E_q\left\{\log\prod_{i=1}^{n}p\left(a_{i1},\ldots,a_{iK}\,|\,\omega_1,\ldots,\omega_K\right)\right\}=\sum_{i=1}^{n}\sum_{k=1}^{K}\left[\mu_{q(a_{ik})}E_q\left\{\log(\omega_k)\right\}-E_q\left\{\log(a_{ik})\right\}\right].$$

Also,

$$E_q\left\{\log\prod_{i=1}^{n}q^*\left(a_{i1},\ldots,a_{iK}\right)\right\}=\sum_{i=1}^{n}\sum_{k=1}^{K}\left[\mu_{q(a_{ik})}\log\left(\mu_{q(a_{ik})}\right)-E_q\left\{\log\left(a_{ik}\right)\right\}\right].$$

Therefore,

$$\begin{aligned} & E_q\left\{\log\prod_{i=1}^{n}p\left(a_{i1},\ldots,a_{iK}\,|\,\omega_1,\ldots,\omega_K\right)-\log\prod_{i=1}^{n}q^*\left(a_{i1},\ldots,a_{iK}\right)\right\} \\ & \qquad =\sum_{i=1}^{n}\sum_{k=1}^{K}\mu_{q(a_{ik})}\left[E_q\left\{\log(\omega_k)\right\}-\log\left(\mu_{q(a_{ik})}\right)\right]. \end{aligned}$$

Next,

$$\log p(\omega_1,\ldots,\omega_K) = \log\Gamma\left(\sum_{k=1}^{K}\alpha_k\right) - \sum_{k=1}^{K}\log\Gamma\left(\alpha_k\right) + \sum_{k=1}^{K}\left(\alpha_k - 1\right)\log\left(\omega_k\right).$$

Taking the expectation of both sides gives

$$E_q\left\{\log p(\omega_1,\ldots,\omega_K)\right\} = \log\Gamma\left(\sum_{k=1}^{K}\alpha_k\right) - \sum_{k=1}^{K}\log\Gamma\left(\alpha_k\right) + \sum_{k=1}^{K}\left(\alpha_k - 1\right)E_q\left\{\log\left(\omega_k\right)\right\}.$$

Also,

$$\begin{aligned} E_q\left\{\log q^*(\omega_1,\ldots,\omega_K)\right\} = \log\Gamma\left(\sum_{k=1}^{K}\alpha_{q(\omega_k)}\right) - \sum_{k=1}^{K}\log\Gamma\left(\alpha_{q(\omega_k)}\right) \\ + \sum_{k=1}^{K}\left(\alpha_{q(\omega_k)} - 1\right)E_q\left\{\log\left(\omega_k\right)\right\}. \end{aligned}$$

Therefore,

$$\begin{aligned} E_q\left\{\log p(\omega_1,\ldots,\omega_K) - \log q^*(\omega_1,\ldots,\omega_K)\right\} = \log\Gamma\left(\sum_{k=1}^{K}\alpha_k\right) - \log\Gamma\left(\sum_{k=1}^{K}\alpha_{q(\omega_k)}\right) \\ - \sum_{k=1}^{K}\log\Gamma\left(\alpha_k\right) + \sum_{k=1}^{K}\log\Gamma\left(\alpha_{q(\omega_k)}\right) \\ + \left(\alpha_k - \alpha_{q(\omega_k)}\right)E_q\left\{\log\left(\omega_k\right)\right\}. \end{aligned}$$

Adding all expressions together, we get the lower bound expression given in (6.7). We also note that the following expressions equate to zero.

$$\sum_{k=1}^{K}\left(\alpha_k - \alpha_{q(\omega_k)}\right)E_q\left\{\log\left(\omega_k\right)\right\} + \sum_{i=1}^{n}\sum_{k=1}^{K}\mu_{q(a_{ik})}E_q\left\{\log\left(\omega_k\right)\right\},$$

$$\left\{B_{q(\sigma_\varepsilon^2)} - \mu_{q(1/a_\varepsilon)}\right\}\mu_{q(1/\sigma_\varepsilon^2)} - \tfrac{1}{2}\mu_{q(1/\sigma_\varepsilon^2)}E_q\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2,$$

$$\left\{B_{q(\sigma_o^2)} - \mu_{q(1/a_o)}\right\}\mu_{q(1/\sigma_o^2)} - \tfrac{1}{2}\mu_{q(1/\sigma_o^2)}E_q\|\boldsymbol{o} - \tilde{\boldsymbol{X}}\boldsymbol{\alpha}\|^2,$$

and

$$\begin{aligned} &\sum_{k=1}^{K}\mu_{q\left\{1/(\sigma_k^x)^2\right\}}\left\{B_{q\left\{(\sigma_k^x)^2\right\}} - \mu_{q(1/a_k^x)}\right\} \\ &-\tfrac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\mu_{q(a_{ik})}\mu_{q\left\{1/(\sigma_k^x)^2\right\}}\left\{\left(\left\{E_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{x}\right)\right\}_i - \mu_{q(\mu_k^x)}\right)^2 + \left\{\sigma^2_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{x}\right)\right\}_i + \sigma^2_{q(\mu_k^x)}\right\}. \end{aligned}$$

Chapter 7

Alternative approach based on variational message passing

7.1 Introduction

Throughout this thesis, we have dealt with the situation where approximate inference of a possibly large complex graphical model is required. Until now we have shown the details and benefits of using MFVB for fitting these models. An alternative approach, which produces the same approximation but with a different algebraic system, is *variational message passing (VMP)*. The MFVB inference used by Infer.NET is based on the VMP approach. The advantage of VMP compared with MFVB is that the messages are obtained by using only a handful of algebraic rules. This allows more straightforward extension to arbitrarily large models. In this chapter we only derive the VMP alternative algorithms for two models from earlier chapters. Hence, the extension to arbitrarily large models is not covered here.

The essence of VMP is to pass *messages* on a *factor graph*. These two notions are introduced in Section 7.2. Often, these messages are *exponential family* density functions of model parameters and are generally summarised by their natural parameter vector. In addition, the use of natural parameterisations mean that these rules involve simple algebraic manipulations. As will become apparent in section 7.7.1, VMP makes use of the fact that particular forms of messages occur repeatedly, so algebra corresponding to these forms only need to be carried out once.

Even though VMP can be shown to be more efficient than MFVB in terms of per-

forming the necessary algebra, it can also be notationally demanding to outline the rules involved for general graphical models. This is mainly due to the notational and graphical forms contained in Winn & Bishop (2005), Minka (2005) and Minka & Winn (2009) to differ from each other in terms of the underlying rules associated with VMP. We have found it best to work with the description of VMP as given in Minka (2005) and thus have adopted the same notation here. The primary aim of this chapter is to provide a notationally user friendly guide to the general VMP approach, with examples considering formulation and comparison of the algorithms in Chapters 2 and 6 of this thesis.

Section 7.2 gives description of the factor graph representation of Bayesian hierarchical models. Sections 7.3, 7.4 and 7.5 introduce the natural parameter forms, primitives and function definitions corresponding to the density functions that are used within this chapter. The general VMP algorithm is given in Section 7.6 and the reader is provided a walk through using two examples in Section 7.7. Finally, the derivations of the algorithms presented are given in the Appendices that follow.

7.2 Factor graph representation

Factor graphs play a major role in VMP as they are used to illustrate the connection between random variables in a specified model. The aim is to compute messages which are then passed between factors and corresponding nodes on a factor graph. Probabilistic DAGs have direct factor graph representations, e.g., given a DAG that illustrates the conditional independence structure of, say, the random variables $x_1, \ldots, x_k$, these representations arise from relationships such as

$$p(x_1, \ldots, x_k) = \prod_{i=1}^{k} p(x_i \,|\, \text{parents of } x_i). \tag{7.1}$$

Another way to represent a model by its factor graph is to derive the joint density function of the random variables and the observed data in terms of their corresponding factors. For example, consider the simple linear regression model

$$\begin{gathered} \boldsymbol{y} \,|\, \boldsymbol{\beta}, \sigma^2 \sim \mathrm{N}\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}\right), \quad \boldsymbol{\beta} \sim \mathrm{N}\left(\boldsymbol{0}, \sigma_{\boldsymbol{\beta}}^2 \boldsymbol{I}\right), \\ \sigma^2 \,|\, a \sim \text{Inverse-Gamma}\left(\tfrac{1}{2}, 1/a\right), \quad a \sim \text{Inverse-Gamma}\left(\tfrac{1}{2}, 1/A^2\right). \end{gathered} \tag{7.2}$$

The joint density function of $\boldsymbol{y}$, $\boldsymbol{\beta}$, σ^2 and a is

$$p(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, a) = p(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \sigma^2) p(\sigma^2 \,|\, a) p(\boldsymbol{\beta}) p(a), \tag{7.3}$$

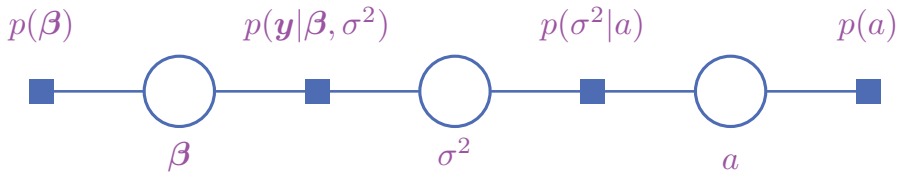


Figure 7.1: *Factor graph corresponding to model (7.2).*

where each of the density functions $p(\boldsymbol{y}\,|\,\boldsymbol{\beta}, \sigma^2)$, $p(\sigma^2\,|\,a)$, $p(\boldsymbol{\beta})$ and $p(a)$ are referred to as *factors*. The corresponding factor graph is given in Figure 7.1. The circular nodes correspond to each of the random variables in (7.2) and the solid squares correspond to each of the factors in the right-hand side of (7.3), which can also be derived from (7.1). To keep notation simple, we suppress the dependence of $\boldsymbol{\beta}$ and σ^2 on the observed data vector $\boldsymbol{y}$ in the factor graph. Minka (2005) argues that factor graph representations of models are useful for VMP approximate inference as will be illustrated in the sections to follow.

7.2.1 Additional notation

When using VMP it is convenient to take advantage of the benefits of using a factor graph representation of a statistical model. Factor neighbours notation play an important role in the understanding of VMP in the context of factor graph representations. The following definitions are key to the introduction of Algorithm 10 in Section 7.6.

Definition 7.2.1. *Let A and B be sets such that $B \subseteq A$. Then*

$$A \backslash B = \textit{set containing elements of } A \textit{ that are not in } B.$$

We define N_{hid} to be the number of hidden nodes and N_{fac} to be the number of factors in a factor graph. For example, in Figure 7.2, $N_{\text{hid}} = 10$ and $N_{\text{fac}} = 9$ since $\theta_1, \ldots, \theta_9$ and ϕ are hidden nodes and $f_1, \ldots, f_9$ are factor nodes.

Definition 7.2.2. *Consider a factor graph with factors f_j, $1 \leq j \leq N_{\text{fac}}$ and non-factor nodes θ_i, $1 \leq i \leq N_{\text{hid}}$. Then for each $j = 1, \ldots, N_{\text{fac}}$,*

$$\textit{neighbours}(j) = \{1 \leq i \leq N_{\text{hid}} : \theta_i \textit{ is a neighbour of } f_j\}.$$

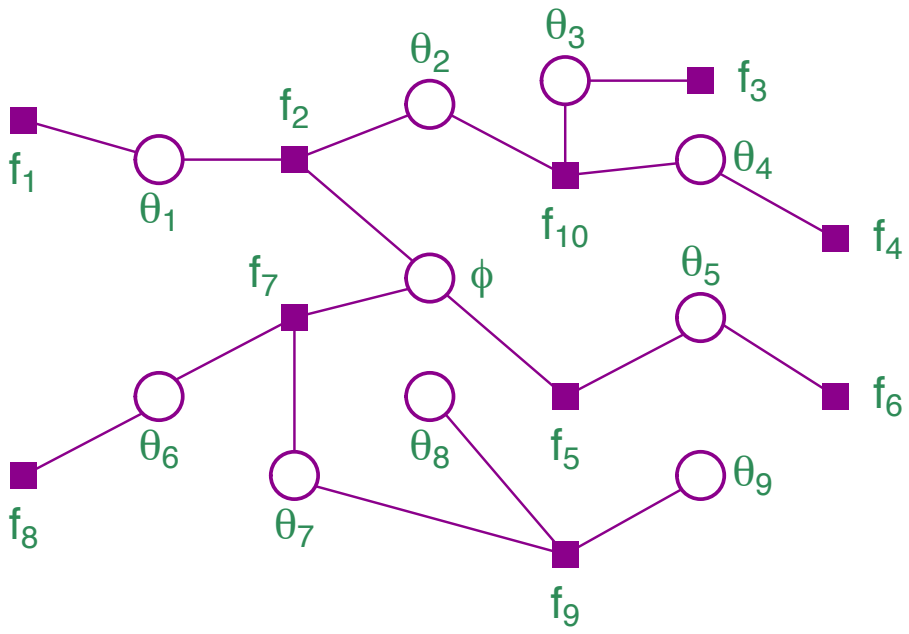


Figure 7.2: *Example of a factor graph.*

It follows that each f_j is a function of its corresponding neighbours on the factor graph.

7.3 Natural parameter forms

Here we provide the forms of the density functions used for the derivations of Algorithms 11 and 12 given later in this chapter.

Univariate normal distribution

The natural parameter form of the univariate normal distribution with mean μ and variance σ^2 is

$$\begin{aligned} p(x;\mu,\sigma^2) &= \exp\left\{\begin{bmatrix} x \\ x^2 \end{bmatrix}^\top \begin{bmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{bmatrix} - \left(\tfrac{\mu}{2\sigma^2} + \tfrac{\log(\sigma^2)}{2}\right) - \tfrac{1}{2}\log(2\pi)\right\} \\ &= \exp\left\{\boldsymbol{T}(\boldsymbol{x})^\top \boldsymbol{\eta} - \left(\tfrac{\mu}{2\sigma^2} + \tfrac{\log(\sigma^2)}{2}\right) - \tfrac{1}{2}\log(2\pi)\right\} \end{aligned}$$

where

$$\boldsymbol{T}(\boldsymbol{x}) \equiv \begin{bmatrix} x \\ x^2 \end{bmatrix}, \quad \boldsymbol{\eta} \equiv \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} \equiv \begin{bmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{bmatrix}$$

are the *natural statistic* and *natural parameter* vectors. Under this representation, we define $\mathrm{N}_{\mathrm{nat}}$ to be such that

$$x \sim \mathrm{N}_{\mathrm{nat}}(\eta_1, \eta_2).$$

Fact 7.3.1. *The ordinary and natural parameters of the univariate normal distribution are mapped between each other through:*

$$\begin{cases} \eta_1 = \mu/\sigma^2 \\ \eta_2 = -1/2\sigma^2 \end{cases} \quad \textit{and} \quad \begin{cases} \mu = -\eta_1/2\eta_2 \\ \sigma^2 = -1/2\eta_2. \end{cases}$$

Multivariate normal distribution

The multivariate Normal density function with $n \times 1$ mean vector $\boldsymbol{\mu}$ and $n \times n$ variance-covariance matrix $\boldsymbol{\Sigma}$ is given in Definition 1.4.7. This density function can also be written in its natural parameter form as

$$p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \exp\left\{ \begin{bmatrix} \boldsymbol{x} \\ \text{vec}\left(\boldsymbol{x}\boldsymbol{x}^\top\right) \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\text{vec}\left(\boldsymbol{\Sigma}^{-1}\right) \end{bmatrix} - \tfrac{1}{2}\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \tfrac{1}{2}\log|2\pi\boldsymbol{\Sigma}| \right\}$$

$$= \exp\left\{ \boldsymbol{T}(\boldsymbol{x})^\top\boldsymbol{\eta} - \tfrac{1}{2}\left(\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \log|\boldsymbol{\Sigma}|\right) - \tfrac{n}{2}\log(2\pi) \right\}$$

where

$$\boldsymbol{T}(\boldsymbol{x}) \equiv \begin{bmatrix} \boldsymbol{x} \\ \text{vec}\left(\boldsymbol{x}\boldsymbol{x}^\top\right) \end{bmatrix}, \quad \boldsymbol{\eta} \equiv \begin{bmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_{2,\text{vec}} \end{bmatrix} \equiv \begin{bmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\text{vec}\left(\boldsymbol{\Sigma}^{-1}\right) \end{bmatrix}$$

are the *natural statistic* and *natural parameters vectors*. We define $\text{N}_{\text{nat,vec}}$ to be such that

$$\boldsymbol{x} \sim \text{N}_{\text{nat,vec}}\left(\boldsymbol{\eta}_1, \boldsymbol{\eta}_{2,\text{vec}}\right).$$

Fact 7.3.2. *The ordinary and natural parameters of the multivariate normal distribution are mapped between each other through the relationship:*

$$\begin{cases} \boldsymbol{\eta}_1 = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ \boldsymbol{\eta}_{2,vec} = -\frac{1}{2}vec\left(\boldsymbol{\Sigma}^{-1}\right) \end{cases} \quad \textit{and} \quad \begin{cases} \boldsymbol{\mu} = -\frac{1}{2}\left\{vec^{-1}\left(\boldsymbol{\eta}_{2,vec}\right)\right\}^{-1}\boldsymbol{\eta}_1 \\ \boldsymbol{\Sigma} = -\frac{1}{2}\left\{vec^{-1}\left(\boldsymbol{\eta}_{2,vec}\right)\right\}^{-1}. \end{cases}$$

Inverse-Gamma distribution

The Inverse-Gamma density function for a scalar variable x is given in Definition 1.4.9 but can also be written in its natural parameter form as

$$p(x; A, B) = \exp\left\{ \begin{bmatrix} \log x \\ 1/x \end{bmatrix}^\top \begin{bmatrix} -A-1 \\ -B \end{bmatrix} + A\log B - \log\Gamma(A) \right\}$$

$$= \exp\left\{ \boldsymbol{T}(x)^\top\boldsymbol{\eta} + A\log B - \log\Gamma(A) \right\}$$

where

$$\boldsymbol{T}(x) \equiv \begin{bmatrix} \log x \\ 1/x \end{bmatrix}, \quad \boldsymbol{\eta} \equiv \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} \equiv \begin{bmatrix} -A-1 \\ -B \end{bmatrix}$$

are the *natural statistic* and *natural parameter* vectors for the Inverse-Gamma distribution. We define $\mathrm{IG}_{\mathrm{nat}}$ to be such that

$$x \sim \mathrm{IG}_{\mathrm{nat}}\left(\eta_1, \eta_2\right).$$

Fact 7.3.3. *The ordinary and natural parameters of the Inverse-Gamma distribution are mapped between each other via:*

$$\begin{cases} \eta_1 = -A-1 \\ \eta_2 = -B \end{cases} \quad \textit{and} \quad \begin{cases} A = -\eta_1 - 1 \\ B = -\eta_2. \end{cases}$$

Inverse-Wishart distribution

The Inverse-Wishart density function with $n \times n$ positive definite scale matrix $\boldsymbol{B}$ and degrees of freedom a is given in Definition 1.4.3.9 but has another representation using its natural parameter form:

$$\begin{aligned} p(\boldsymbol{X}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \exp\left\{ \begin{bmatrix} \log|\boldsymbol{X}| \\ \mathrm{vec}\left(\boldsymbol{X}^{-1}\right) \end{bmatrix}^{\top} \begin{bmatrix} -\frac{1}{2}(a+n+1) \\ -\frac{1}{2}\mathrm{vec}\left(\boldsymbol{B}\right) \end{bmatrix} - \tfrac{A}{2}\log|\boldsymbol{B}| - \log C_{n,a} \right\} \\ &= \exp\left\{ \boldsymbol{T}(\boldsymbol{X})^{\top}\boldsymbol{\eta} - \tfrac{A}{2}\log|\boldsymbol{B}| - \log C_{n,a} \right\} \end{aligned}$$

where

$$\boldsymbol{T}(\boldsymbol{X}) \equiv \begin{bmatrix} \log|\boldsymbol{X}| \\ \mathrm{vec}\left(\boldsymbol{X}^{-1}\right) \end{bmatrix}, \quad \boldsymbol{\eta} \equiv \begin{bmatrix} \eta_1 \\ \boldsymbol{\eta}_2 \end{bmatrix} \equiv \begin{bmatrix} -\frac{1}{2}(a+n+1) \\ -\frac{1}{2}\mathrm{vec}\left(\boldsymbol{B}\right) \end{bmatrix}$$

are the *natural statistic* and *natural parameter* vectors for the Inverse-Wishart distribution. We define $\mathrm{IW}_{\mathrm{nat}}$ to be such that

$$\boldsymbol{X} \sim \mathrm{IW}_{\mathrm{nat}}\left(\eta_1, \boldsymbol{\eta}_2\right).$$

Fact 7.3.4. *The ordinary and natural parameters of the Inverse-Wishart distribution can be mapped between each other via:*

$$\begin{cases} \eta_1 = -\frac{1}{2}(a+n+1) \\ \boldsymbol{\eta}_2 = -\frac{1}{2}vec\left(\boldsymbol{B}\right) \end{cases} \quad \textit{and} \quad \begin{cases} A = -2\eta_1 - n - 1 \\ \boldsymbol{B} = -2vec^{-1}(\boldsymbol{\eta}_2). \end{cases}$$

Multinomial distribution

The Multinomial density function with 1 independent trial and probability of success $\boldsymbol{p} = (p_1, \ldots, p_k)$ is given in Definition 1.4.13 but has another representation using its natural parameter form:

$$\begin{aligned} p(x_1, \ldots, x_k; 1, p_1, \ldots, p_k) &= \exp\left\{ \begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix}^\top \begin{bmatrix} \log(p_1) \\ \vdots \\ \log(p_k) \end{bmatrix} \right\} \\ &= \exp\left\{ \boldsymbol{T}(\boldsymbol{x})^\top \boldsymbol{\eta} \right\} \end{aligned}$$

where

$$\boldsymbol{T}(\boldsymbol{X}) \equiv \begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix}, \quad \boldsymbol{\eta} \equiv \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_k \end{bmatrix} \equiv \begin{bmatrix} \log(p_1) \\ \vdots \\ \log(p_k) \end{bmatrix}$$

are the *natural statistic* and *natural parameter* vectors. We define $\mathrm{M}_{\mathrm{nat}}$ to be such that

$$x_1, \ldots, x_k \sim \mathrm{M}_{\mathrm{nat}}(1, \eta_1, \ldots, \eta_k).$$

Fact 7.3.5. *The ordinary and natural parameters of the Multinomial distribution can be mapped between each other via:*

$$\begin{cases} \eta_1 = \log(p_1) \\ \quad \vdots \\ \eta_k = \log(p_k) \end{cases} \qquad \textit{and} \qquad \begin{cases} p_1 = e^{\eta_1} \\ \quad \vdots \\ p_k = e^{\eta_k}. \end{cases}$$

Dirichlet distribution

The Dirichlet density function with parameters $\alpha_1, \ldots \alpha_k$ is given in Definition 1.4.14. Its natural parameter form is expressed as

$$\begin{aligned} p(x_1, \ldots, x_k | \alpha_1, \ldots, \alpha_k) &= \exp\left\{ \begin{bmatrix} \log(x_1) \\ \vdots \\ \log(x_k) \end{bmatrix}^\top \begin{bmatrix} \alpha_1 - 1 \\ \vdots \\ \alpha_k - 1 \end{bmatrix} \right. \\ &\qquad \left. + \sum_{k=1}^{K} \log\{\Gamma(\alpha_k)\} - \log\left\{ \Gamma\left(\sum_{k=1}^{K} \alpha_k \right) \right\} \right\} \\ &= \exp\left\{ \boldsymbol{T}(\boldsymbol{x})^\top \boldsymbol{\eta} + \sum_{k=1}^{K} \log\{\Gamma(\alpha_k)\} - \log\left\{ \Gamma\left(\sum_{k=1}^{K} \alpha_k \right) \right\} \right\} \end{aligned}$$

where

$$\boldsymbol{T}(\boldsymbol{x}) \equiv \begin{bmatrix} \log(x_1) \\ \vdots \\ \log(x_k) \end{bmatrix}, \quad \boldsymbol{\eta} \equiv \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_k \end{bmatrix} \equiv \begin{bmatrix} \alpha_1 - 1 \\ \vdots \\ \alpha_k - 1 \end{bmatrix}$$

are the *natural statistic* and *natural parameter* vectors. We define $\mathrm{D}_{\mathrm{nat}}$ to be such that

$$x_1, \ldots, x_k \sim \mathrm{D}_{\mathrm{nat}}(\eta_1, \ldots, \eta_k).$$

Fact 7.3.6. *The ordinary and natural parameters of the Dirichlet distribution can be mapped between each other via:*

$$\begin{cases} \eta_1 = \alpha_1 - 1 \\ \quad \vdots \\ \eta_k = \alpha_k - 1 \end{cases} \quad \textit{and} \quad \begin{cases} p_1 = \eta_1 + 1 \\ \quad \vdots \\ p_k = \eta_k + 1. \end{cases}$$

7.4 Primitive integrals and results

The following primitives and results follow immediately from the results involving moments of random variables which take on distributions from Section 7.3. These moments aid in the derivation of the algorithms considered in this chapter as they are presented in their natural parameter forms.

We define

$p_{\mathrm{N}_{\mathrm{nat}}}\left(x; \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix}\right)$ to be the $\mathrm{N}_{\mathrm{nat}}(\eta_1, \eta_2)$ density function in x,

$p_{\mathrm{N}_{\mathrm{nat,vec}}}\left(\boldsymbol{x}; \begin{bmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_{2,\mathrm{vec}} \end{bmatrix}\right)$ to be the $\mathrm{N}_{\mathrm{nat,vec}}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_{2,\mathrm{vec}})$ density function in $\boldsymbol{x}$,

$p_{\mathrm{IG}_{\mathrm{nat}}}\left(x; \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix}\right)$ to be the $\mathrm{IG}_{\mathrm{nat}}(\eta_1, \eta_2)$ density function in x and

$p_{\mathrm{IW}_{\mathrm{nat}}}\left(\boldsymbol{X}; \begin{bmatrix} \eta_1 \\ \boldsymbol{\eta}_2 \end{bmatrix}\right)$ to be the $\mathrm{IW}_{\mathrm{nat}}(\eta_1, \boldsymbol{\eta}_2)$ density function in $\boldsymbol{X}$.

We now present various primitives and results concerning moments of the density functions given in Section 7.3.

Primitive 7.4.1.

$$\int_{-\infty}^{\infty} x^2 \, p_{N_{nat}}\left(x; \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix}\right) dx = \frac{\eta_1^2 - 2\eta_2}{4\eta_2^2}$$

Primitive 7.4.2.

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-y)^2\ p_{N_{nat}}\left(x,y;\begin{bmatrix}\eta_{x,1}\\ \eta_{x,2}\end{bmatrix},\begin{bmatrix}\eta_{y,1}\\ \eta_{y,2}\end{bmatrix}\right)dxdy$$
$$=\frac{\eta_{x,2}\left(\eta_{x,1}^2-2\eta_{x,2}\right)}{4\eta_{x,2}^3}+\frac{\eta_{y,2}\left(\eta_{y,1}^2-2\eta_{y,2}\right)}{4\eta_{y,2}^3}-\frac{\eta_{x,1}\eta_{y,1}}{2\eta_{x,2}\eta_{y,2}}$$

where $p_{N_{nat}}\left(x,y;\begin{bmatrix}\eta_{x,1}\\ \eta_{x,2}\end{bmatrix},\begin{bmatrix}\eta_{y,1}\\ \eta_{y,2}\end{bmatrix}\right)$ *is the joint density function of* x *and* y*.*

Primitive 7.4.3. *If* $\boldsymbol{x}$ *is a* $d\times 1$ *vector*

$$\int_{\mathbb{R}^d}\boldsymbol{x}^\top\boldsymbol{x}\ p_{N_{nat,vec}}\left(\boldsymbol{x};\begin{bmatrix}\boldsymbol{\eta}_1\\ \boldsymbol{\eta}_{2,vec}\end{bmatrix}\right)d\boldsymbol{x}$$
$$=\tfrac{1}{4}tr\left(\left\{vec^{-1}\left(\boldsymbol{\eta}_{2,vec}\right)\right\}^{-1}\left[\boldsymbol{\eta}_1\boldsymbol{\eta}_1^\top\left\{vec^{-1}\left(\boldsymbol{\eta}_{2,vec}\right)\right\}^{-1}-2\boldsymbol{I}\right]\right)$$

Primitive 7.4.4. *For any* $n\times 1$ *vector* $\boldsymbol{a}$ *and* $n\times p$ *matrix* B*,*

$$\int_{\mathbb{R}^p}\left(\boldsymbol{a}-\boldsymbol{B}\boldsymbol{x}\right)\left(\boldsymbol{a}-\boldsymbol{B}\boldsymbol{x}\right)^\top\ p_{N_{nat,vec}}\left(\boldsymbol{x};\begin{bmatrix}\boldsymbol{\eta}_1\\ \boldsymbol{\eta}_{2,vec}\end{bmatrix}\right)d\boldsymbol{x}$$
$$=\boldsymbol{a}\boldsymbol{a}^\top+\left\{vec^{-1}\left(\boldsymbol{\eta}_{2,vec}\right)\right\}^{-1}\left[\boldsymbol{a}\boldsymbol{\eta}_1^\top\boldsymbol{B}^\top+\tfrac{1}{4}\boldsymbol{B}\boldsymbol{\eta}_1\boldsymbol{\eta}_1^\top\boldsymbol{B}^\top\left\{vec^{-1}\left(\boldsymbol{\eta}_{2,vec}\right)\right\}^{-1}-\tfrac{1}{2}\boldsymbol{B}\boldsymbol{B}^\top\right]$$

Primitive 7.4.5. *If* $[x_1,\ldots,x_d]^\top\sim N_{nat,vec}\left(\boldsymbol{\eta}_1,\boldsymbol{\eta}_{2,vec}\right)$*, then*

$$\int_{\mathbb{R}^d}x_ix_j\ p_{N_{nat,vec}}\left(\boldsymbol{x};\begin{bmatrix}\boldsymbol{\eta}_1\\ \boldsymbol{\eta}_{2,vec}\end{bmatrix}\right)d\boldsymbol{x}$$
$$=-\tfrac{1}{8}\left[\left\{vec^{-1}\left(\boldsymbol{\eta}_{2,vec}\right)\right\}^{-1}\boldsymbol{\eta}_1\right]_i\left[\left\{vec^{-1}\left(\boldsymbol{\eta}_{2,vec}\right)\right\}^{-1}\boldsymbol{\eta}_1\right]_j+\left[\left\{vec^{-1}\left(\boldsymbol{\eta}_{2,vec}\right)\right\}^{-1}\right]_{ij}$$

Primitive 7.4.6.

$$\int_0^\infty\frac{1}{x}\ p_{IG_{nat}}\left(x;\begin{bmatrix}\eta_1\\ \eta_2\end{bmatrix}\right)dx=\frac{\eta_1+1}{\eta_2}$$

Primitive 7.4.7.

$$\int_0^\infty\log(x)\ p_{IG_{nat}}\left(x;\begin{bmatrix}\eta_1\\ \eta_2\end{bmatrix}\right)dx=\log(-\eta_2)-\psi(-\eta_1-1)$$

Primitive 7.4.8.

$$\int_S \boldsymbol{X}^{-1}\, p_{IW_{nat}}\left(\boldsymbol{X}; \begin{bmatrix}\eta_1 \\ \boldsymbol{\eta}_2\end{bmatrix}\right) d\boldsymbol{X} = \left(\eta_1 + \frac{n+1}{2}\right)\left\{vec^{-1}\left(\boldsymbol{\eta}_2\right)\right\}^{-1}$$

where S is the set of all symmetric positive definite $n \times n$ matrices.

Result 7.4.1. *Suppose that $x_1, \ldots, x_K \sim M_{nat}(1; \eta_1, \ldots, \eta_K)$. Then*

$$E(x_k) = e^{\eta_k}, \quad 1 \le k \le K.$$

Result 7.4.2. *Suppose that $x_1, \ldots, x_K \sim D_{nat}(\eta_1, \ldots, \eta_K)$. Then*

$$E\left\{\log\left(x_k\right)\right\} = \psi\left(\eta_k + 1\right) - \psi\left\{\sum_{k=1}^{K}\left(\eta_k + 1\right)\right\}, \quad 1 \le k \le K.$$

7.5 Function definitions

The following functions are useful for describing various VMP updates used throughout this chapter.

$$\mathcal{F}\left(\begin{bmatrix}a_1 \\ a_2\end{bmatrix}\right) \equiv \tfrac{1}{4}\mathrm{tr}\left(\left\{\mathrm{vec}^{-1}(a_2)\right\}^{-1}\left[a_1 a_1^{\top}\left\{\mathrm{vec}^{-1}(a_2)\right\}^{-1} - 2\boldsymbol{I}\right]\right),$$

where a_1 and a_2 are scalars.

$$\mathcal{G}\left(\begin{bmatrix}a_1 \\ \boldsymbol{a}_2\end{bmatrix}; \boldsymbol{b}, \boldsymbol{C}\right) \equiv \left(a_1 + \frac{n+1}{2}\right)\left\{\mathrm{vec}^{-1}(\boldsymbol{a}_2)\right\}^{-1},$$

where a_1 is a scalar and $\boldsymbol{a}_2$ is an $n^2 \times 1$ vector.

$$\mathcal{H}\left(\begin{bmatrix}\boldsymbol{a}_1 \\ \boldsymbol{a}_2\end{bmatrix}; \boldsymbol{b}, \boldsymbol{C}\right) \equiv \boldsymbol{b}\boldsymbol{b}^{\top} + \left\{\mathrm{vec}^{-1}\left(\boldsymbol{a}_2\right)\right\}^{-1}\left[\boldsymbol{b}\boldsymbol{a}_1^{\top}\boldsymbol{C}^{\top} + \tfrac{1}{4}\boldsymbol{C}\boldsymbol{a}_1\boldsymbol{a}_1^{\top}\boldsymbol{C}^{\top}\left\{\mathrm{vec}^{-1}\left(\boldsymbol{a}_2\right)\right\}^{-1} - \tfrac{1}{2}\boldsymbol{C}\boldsymbol{C}^{\top}\right],$$

where $\boldsymbol{a}_1$ is a $q \times 1$ vector, $\boldsymbol{a}_2$ a $q^2 \times 1$ vector, $\boldsymbol{b}$ a $p \times 1$ vector and $\boldsymbol{C}$ a $p \times q$ matrix.

$$\mathcal{I}\left(\begin{bmatrix}a_1 \\ a_2\end{bmatrix}, \begin{bmatrix}b_1 \\ b_2\end{bmatrix}\right) \equiv \frac{a_2(a_1^2 - 2a_2)}{4a_2^3} + \frac{b_2(b_1^2 - 2b_2)}{4b_2^3} - \frac{a_1 b_1}{2a_2 b_2},$$

where a_1, a_2, b_1 and b_2 are scalars.

$$\mathcal{J}\left(\begin{bmatrix}\boldsymbol{a}_1 \\ \boldsymbol{a}_2\end{bmatrix}; b, c\right) \equiv -\tfrac{1}{8}\left[\left\{\mathrm{vec}^{-1}(\boldsymbol{a}_2)\right\}^{-1}\boldsymbol{a}_1\right]_b\left[\left\{\mathrm{vec}^{-1}(\boldsymbol{a}_2)\right\}^{-1}\boldsymbol{a}_1\right]_c + \left[\left\{\mathrm{vec}^{-1}(\boldsymbol{a}_2)\right\}^{-1}\right]_{bc},$$

where $\boldsymbol{a}_1$ is a $q \times 1$ vector, $\boldsymbol{a}_2$ is a $q^2 \times 1$ vector and b and c both scalars.

$$\mathcal{K}\left(\begin{bmatrix} a_1 \\ a_2 \end{bmatrix}\right) \equiv \frac{a_1^2 - 2a_2}{4a_2^2},$$

where a_1 and a_2 are scalars.

7.6 The variational message passing algorithm

Our goal is to obtain the mean field approximation to the joint posterior density function:

$$p(\theta_1, \ldots, \theta_{N_{\text{hid}}} | \boldsymbol{x}) \approx q(\theta_1) \ldots q(\theta_{N_{\text{hid}}}) \tag{7.4}$$

where $\theta_1, \ldots, \theta_{N_{\text{hid}}}$ are the parameters or hidden nodes and $\boldsymbol{x}$ contains the data or evidence nodes in the corresponding statistical model. VMP aims to obtain the optimal q-densities through an iterative scheme concerned with updating messages of the form:

$$m_{f_j \to \theta_i} \quad \text{and} \quad m_{\theta_i \to f_j} \quad \text{for each} \quad i \in \text{neighbours}\,(j)\,, \quad 1 \le j \le N_{\text{fac}}.$$

The general iterative VMP scheme, labelled Algorithm 10, can be used to solve the optimal q densities, by updating such messages.

Convergence of Algorithm 10 is assessed by monitoring successive values of the lower bound on the marginal log-likelihood $\log \underline{p}(\boldsymbol{x}; q)$ and stopping once the successive differences are very small. The lower bound on the marginal log-likelihood is given by

$$\log \underline{p}(\boldsymbol{x}; q) = E_q \left[\sum_{j=1}^{N_{\text{fac}}} \log f_j \left(\theta_{neighbours(j)} \right) - \sum_{i=1}^{N_{\text{hid}}} \log q(\theta_i) \right]. \tag{7.5}$$

Within Algorithm 10, if neighbours$(j)\backslash\{i\} = \varnothing$, the empty set, the expression

$$\prod_{i' \in \text{ neighbours}(j)\backslash\{i\}} m_{\theta_{i'} \to f_j}(\theta_{i'}) \; m_{f_j \to \theta_{i'}}(\theta_{i'})$$

is taken to be 1, since there would exist no i'. Integration is replaced by summation if $\theta_{\text{neighbours}(j)\backslash\{i\}}$ in Step 2 (a) contains discrete random variables. It is easier to work with messages appearing under the same product sign to be from the same exponential family. This is referred to as conjugate VMP. However, VMP is a general scheme and not necessarily restricted to conjugate situations. When conjugacy exists within the message updates, the functional forms of these messages are exposed and it becomes clear that the parameters of these messages are inter-related, much like what happens with MFVB.

Step 1: Initialise

$m_{f_j \to \theta_i}(\theta_i)$ for all $i \in \text{neighbours}(j)$, $1 \le j \le N_{\text{fac}}$, to be arbitrary density functions.

Step 2: Cycle

(a) For $i \in \text{neighbours}(j)$, $1 \le j \le N_{\text{fac}}$,

$$m_{\theta_i \to f_j}(\theta_i) \leftarrow \prod_{\{j' \neq j \,:\, i \in \text{neighbours}(j')\}} m_{f_{j'} \to \theta_i}(\theta_i),$$

(b) For $i \in \text{neighbours}(j)$, $1 \le j \le N_{\text{fac}}$,

$$m_{f_j \to \theta_i}(\theta_i) \leftarrow \exp\left[\int \left(\frac{1}{Z} \prod_{i' \in \text{neighbours}(j)\setminus\{i\}} m_{\theta_{i'} \to f_j}(\theta_{i'}) \; m_{f_j \to \theta_{i'}}(\theta_{i'}) \right) \times \log f_j\left(\theta_{\text{neighbours}(j)}\right) d\theta_{\text{neighbours}(j)\setminus\{i\}} \right],$$

where $Z = \int \prod_{i' \in \text{neighbours}(j)\setminus\{i\}} m_{\theta_{i'} \to f_j}(\theta_{i'}) \; m_{f_j \to \theta_{i'}}(\theta_{i'}) \, d\theta_{\text{neighbours}(j)\setminus\{i\}}$.

until all messages converge.

Step 3: For $1 \le i \le N_{\text{hid}}$,

$$q^*(\theta_i) \leftarrow \prod_{\{j \,:\, i \in \text{neighbours}(j)\}} m^*_{f_j \to \theta_i}(\theta_i) \Big/ \int \prod_{\{j \,:\, i \in \text{neighbours}(j)\}} m^*_{f_j \to \theta_i}(\theta_i) \, d\theta_i.$$

Algorithm 10: *The general variational message passing algorithm.*

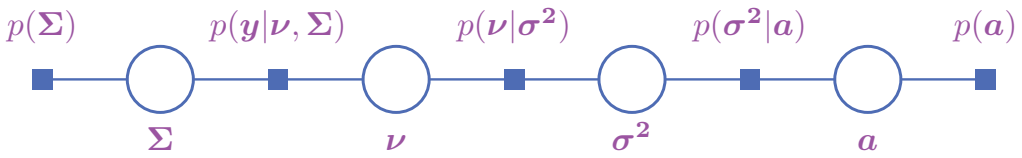


Figure 7.3: *Factor graph corresponding to the marginal longitudinal semiparametric regression model in (2.15).*

These updates begin to exhibit particular patterns and *remembering* these patterns allows one to streamline the calculations of VMP. This is especially useful for larger complex statistical models. Examples of these patterns are given in Appendix 7.4 for the two models considered in the following sections. In addition, when computing the VMP updates in Algorithm 10, it is appropriate to work with the natural parameter forms of exponential family density functions.

7.7 Examples

We consider two regression examples used throughout this thesis, namely the marginal longitudinal semiparametric regression model (2.15) given in Chapter 2 and the mixture model with measurement error (6.4) given in Chapter 6. We provide detail on the VMP methodology specifically for these two models.

7.7.1 Marginal longitudinal semiparametric regression model

The factor graph corresponding to model (2.15) is given in Figure 7.3, where the circular nodes correspond to each of the random variables, vectors or matrices and the solid squares represent each of the factors in the model. To make notation simple we have combined the mean function coefficients into a single vector, i.e.,

$$\boldsymbol{\nu} \equiv \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix}.$$

We aim to obtain a mean field approximation to the joint posterior density function:

$$p(\boldsymbol{\nu}, \boldsymbol{a}, \boldsymbol{\sigma}^2, \boldsymbol{\Sigma} \,|\, \boldsymbol{y}) \approx q(\boldsymbol{\nu}) q(\boldsymbol{a}) q(\boldsymbol{\sigma}^2) q(\boldsymbol{\Sigma}).$$

Following the rules of Algorithm 10 and referring to the factor graph in Figure 7.3 for the stochastic node $\boldsymbol{\Sigma}$, the two factor to node messages to initialise are $\boldsymbol{m}_{p(\boldsymbol{\Sigma}) \to \boldsymbol{\Sigma}}(\boldsymbol{\Sigma})$ and $\boldsymbol{m}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma}) \to \boldsymbol{\Sigma}}(\boldsymbol{\Sigma})$. We can set these to be any density functions, however to make the

calculations simple and impose conjugacy we choose them to be the Inverse-Wishart$(n, \boldsymbol{I}_n)$ density function. This can be written as:

$$\boldsymbol{m}_{p(\boldsymbol{\Sigma}) \to \boldsymbol{\Sigma}}(\boldsymbol{\Sigma}) \quad \propto \exp\left\{\begin{bmatrix} \log|\boldsymbol{\Sigma}| \\ \text{vec}\left(\boldsymbol{\Sigma}^{-1}\right) \end{bmatrix}^{\top} \boldsymbol{\eta}_{p(\boldsymbol{\Sigma}) \to \boldsymbol{\Sigma}}\right\}$$

$$\boldsymbol{m}_{p(\boldsymbol{y}|\boldsymbol{\nu}, \boldsymbol{\Sigma}) \to \boldsymbol{\Sigma}}(\boldsymbol{\Sigma}) \propto \exp\left\{\begin{bmatrix} \log|\boldsymbol{\Sigma}| \\ \text{vec}\left(\boldsymbol{\Sigma}^{-1}\right) \end{bmatrix}^{\top} \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu}, \boldsymbol{\Sigma}) \to \boldsymbol{\Sigma}}\right\}$$

where the natural parameter vectors $\boldsymbol{\eta}_{p(\boldsymbol{\Sigma}) \to \boldsymbol{\Sigma}}$ and $\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu}, \boldsymbol{\Sigma}) \to \boldsymbol{\Sigma}}$ are initialised to be

$$\boldsymbol{\eta}_{p(\boldsymbol{\Sigma}) \to \boldsymbol{\Sigma}} \propto \begin{bmatrix} -n-1 \\ -\frac{1}{2}\text{vec}\left(\boldsymbol{I}_n\right) \end{bmatrix}, \quad \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu}, \boldsymbol{\Sigma}) \to \boldsymbol{\Sigma}} \propto \begin{bmatrix} -n-1 \\ -\frac{1}{2}\text{vec}\left(\boldsymbol{I}_n\right) \end{bmatrix}.$$

The natural parameter forms of the density functions used for this model were given in Section 7.3. Step 2 (a) of Algorithm 10 is concerned with updating the stochastic node to factor messages. If we begin with the message $\boldsymbol{m}_{\boldsymbol{\Sigma} \to p(\boldsymbol{\Sigma})}(\boldsymbol{\Sigma})$, we notice that the only other neighbour of $\boldsymbol{\Sigma}$ apart from $p(\boldsymbol{\Sigma})$ is $p(\boldsymbol{y}|\boldsymbol{\nu}, \boldsymbol{\Sigma})$. This means that $\boldsymbol{m}_{\boldsymbol{\Sigma} \to p(\boldsymbol{\Sigma})}(\boldsymbol{\Sigma})$ has the update

$$\begin{aligned} \boldsymbol{m}_{\boldsymbol{\Sigma} \to p(\boldsymbol{\Sigma})}(\boldsymbol{\Sigma}) &\propto \boldsymbol{m}_{p(\boldsymbol{y}|\boldsymbol{\nu}, \boldsymbol{\Sigma}) \to \boldsymbol{\Sigma}}(\boldsymbol{\Sigma}) \\ &= \exp\left\{\begin{bmatrix} \log|\boldsymbol{\Sigma}| \\ \text{vec}\left(\boldsymbol{\Sigma}^{-1}\right) \end{bmatrix}^{\top} \boldsymbol{\eta}_{\boldsymbol{\Sigma} \to p(\boldsymbol{\Sigma})}\right\} \end{aligned}$$

where

$$\boldsymbol{\eta}_{\boldsymbol{\Sigma} \to p(\boldsymbol{\Sigma})} \propto \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu}, \boldsymbol{\Sigma}) \to \boldsymbol{\Sigma}}.$$

Similarly, $\boldsymbol{m}_{\boldsymbol{\Sigma} \to p(\boldsymbol{y}|\boldsymbol{\nu}, \boldsymbol{\Sigma})}(\boldsymbol{\Sigma}) \propto \boldsymbol{m}_{p(\boldsymbol{\Sigma}) \to \boldsymbol{\Sigma}}(\boldsymbol{\Sigma})$, which corresponds to the natural parameter update:

$$\boldsymbol{\eta}_{\boldsymbol{\Sigma} \to p(\boldsymbol{y}|\boldsymbol{\nu}, \boldsymbol{\sigma}^2)} \propto \boldsymbol{\eta}_{p(\boldsymbol{\Sigma}) \to \boldsymbol{\Sigma}}.$$

Step 2 (b) of Algorithm 10 is concerned with updating the factor to stochastic node messages and focusing on stochastic node $\boldsymbol{\Sigma}$ which involves updating $\boldsymbol{m}_{p(\boldsymbol{\Sigma}) \to \boldsymbol{\Sigma}}(\boldsymbol{\Sigma})$ and $\boldsymbol{m}_{p(\boldsymbol{y}|\boldsymbol{\nu}, \boldsymbol{\Sigma}) \to \boldsymbol{\Sigma}}(\boldsymbol{\Sigma})$. The first factor to stochastic node message to update is $\boldsymbol{m}_{p(\boldsymbol{\Sigma}) \to \boldsymbol{\Sigma}}(\boldsymbol{\Sigma})$. Since the only neighbour of factor $p(\boldsymbol{\Sigma})$ is the stochastic node $\boldsymbol{\Sigma}$ this message has the fol-

lowing form

$$\begin{aligned}\boldsymbol{m}_{p(\boldsymbol{\Sigma})\to\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}) &\leftarrow \exp\{\log p(\boldsymbol{\Sigma})\}\\ &\propto \exp\left\{\begin{bmatrix}\log|\boldsymbol{\Sigma}|\\ \operatorname{vec}\left(\boldsymbol{\Sigma}^{-1}\right)\end{bmatrix}^{\top}\begin{bmatrix}-\frac{1}{2}\left(A_{\Sigma}+n+1\right)\\ -\frac{1}{2}\operatorname{vec}\left(\boldsymbol{B}_{\Sigma}\right)\end{bmatrix}\right\}.\end{aligned}$$

This means that

$$\boldsymbol{\eta}_{p(\boldsymbol{\Sigma})\to\boldsymbol{\Sigma}}\leftarrow\begin{bmatrix}-\frac{1}{2}\left(A_{\Sigma}+n+1\right)\\ -\frac{1}{2}\operatorname{vec}\left(\boldsymbol{B}_{\Sigma}\right)\end{bmatrix}$$

and remains fixed at this value as it doesn't depend on any other messages. The next message has the form:

$$\begin{aligned}\boldsymbol{m}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}) \leftarrow \exp\Bigg\{\int_{\mathbb{R}^v}&\boldsymbol{m}_{\boldsymbol{\nu}\to p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})}(\boldsymbol{\nu})\times\boldsymbol{m}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{\nu}}(\boldsymbol{\nu})\\ &\times\log p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\,d(\boldsymbol{\nu})\Big\},\end{aligned}$$

where $v = p + \sum_{\ell=1}^{d} K_{\ell}$, the number of columns in $\boldsymbol{C} = [\boldsymbol{X}\,|\,\boldsymbol{Z}]$. Further breakdown of $\boldsymbol{m}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{\Sigma}}(\boldsymbol{\Sigma})$ gives

$$\begin{aligned}&\boldsymbol{m}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{\Sigma}}(\boldsymbol{\Sigma})\\ &\quad\propto\exp\Bigg[\int_{\mathbb{R}^v}\left\{\begin{bmatrix}\boldsymbol{\nu}\\ \operatorname{vec}\left(\boldsymbol{\nu}\boldsymbol{\nu}^{\top}\right)\end{bmatrix}^{\top}\left(\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to\boldsymbol{\nu}}+\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{\nu}}\right)\right\}\\ &\qquad\times\left\{\begin{bmatrix}\log|\boldsymbol{\Sigma}|\\ \operatorname{vec}(\boldsymbol{\Sigma}^{-1})\end{bmatrix}^{\top}\begin{bmatrix}-\frac{m}{2}\\ -\frac{1}{2}\sum_{i=1}^{m}\operatorname{vec}\left\{\left(\boldsymbol{y}_i-\boldsymbol{C}_i\boldsymbol{\nu}_i\right)\left(\boldsymbol{y}_i-\boldsymbol{C}_i\boldsymbol{\nu}_i\right)^{\top}\right\}\end{bmatrix}\right\}d\boldsymbol{\nu}\Bigg]\\ &\quad\propto\exp\left\{\begin{bmatrix}\log|\boldsymbol{\Sigma}|\\ \operatorname{vec}(\boldsymbol{\Sigma}^{-1})\end{bmatrix}^{\top}\begin{bmatrix}-\frac{m}{2}\\ -\frac{1}{2}\sum_{i=1}^{m}\operatorname{vec}\left\{\int_{\mathbb{R}^v}\left(\boldsymbol{y}_i-\boldsymbol{C}_i\boldsymbol{\nu}_i\right)\left(\boldsymbol{y}_i-\boldsymbol{C}_i\boldsymbol{\nu}_i\right)^{\top}\right.\\ \times p_{\mathrm{N}_{\mathrm{nat,vec}}}\left(\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to\boldsymbol{\nu}}\right.\\ \left.\left.+\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{\nu}}\right)d\boldsymbol{\nu}\right\}\end{bmatrix}\right\}\end{aligned}$$

where $p_{\mathrm{N}_{\mathrm{nat,vec}}}\left(\boldsymbol{x};\left[\boldsymbol{\eta}_1^{\top}\ \boldsymbol{\eta}_{2,\mathrm{vec}}^{\top}\right]^{\top}\right)$ is the $\mathrm{N}_{\mathrm{nat,vec}}(\boldsymbol{\eta}_1,\boldsymbol{\eta}_{2,\mathrm{vec}})$ density function in $\boldsymbol{x}$. Further detail on this notation is given previously in Sections 7.3 and 7.4. Using Primitive 7.4.4

we get

$$\begin{aligned}
&\boldsymbol{m}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}) \\
&\qquad\propto \exp\left\{\begin{bmatrix}\log|\boldsymbol{\Sigma}| \\ \mathrm{vec}(\boldsymbol{\Sigma}^{-1})\end{bmatrix}^{\top}\begin{bmatrix} -\frac{m}{2} \\ -\frac{1}{2}\sum_{i=1}^{m}\mathrm{vec}\left\{\mathcal{H}\left(\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to\boldsymbol{\nu}} \right.\right. \\ \left.\left. +\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{\nu}};\, \boldsymbol{y}_i,\, \boldsymbol{C}_i\right)\right\}\end{bmatrix}\right\},
\end{aligned}$$

where the function $\mathcal{H}$ is defined in Section 7.5. This corresponds to the natural parameter update having the form

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{\Sigma}} \leftarrow \begin{bmatrix} -\frac{m}{2} \\ -\frac{1}{2}\sum_{i=1}^{m}\mathrm{vec}\left\{\mathcal{H}\left(\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to\boldsymbol{\nu}}+\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{\nu}};\, \boldsymbol{y}_i,\, \boldsymbol{C}_i\right)\right\}\end{bmatrix}.$$

Step 2 (b) is now finished for the updates corresponding to the stochastic node $\boldsymbol{\Sigma}$. Once convergence is achieved for the messages in Steps 2 (a) and (b) corresponding to all stochastic nodes in factor graph (7.3), we move on to Step 3 of Algorithm 10. For $\boldsymbol{\Sigma}$ this involves the update:

$$\boldsymbol{\eta}^{*}_{q(\boldsymbol{\Sigma})} \leftarrow \boldsymbol{\eta}^{*}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to(\boldsymbol{\nu})}+\boldsymbol{\eta}^{*}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to(\boldsymbol{\nu})}.$$

This means that

$$q^{*}(\boldsymbol{\Sigma}) \text{ is the } \mathrm{IW}_{\mathrm{nat}}\left(\boldsymbol{\eta}^{*}_{q(\boldsymbol{\Sigma}),1},\, \boldsymbol{\eta}^{*}_{q(\boldsymbol{\Sigma}),2}\right) \text{ density function in } \boldsymbol{\Sigma}$$

and can also be transformed back to the original parameters, such as those used in Chapter 2, by writing this as the

$$\text{Inverse-Wishart}\left\{-\left(2\boldsymbol{\eta}^{*}_{q(\boldsymbol{\Sigma}),2}+n+1\right),\, -2\mathrm{vec}\left(\boldsymbol{\eta}^{*}_{q(\boldsymbol{\Sigma}),1}\right)\right\} \text{ density function in } \boldsymbol{\Sigma}.$$

Algorithm 11 gives the updates for all stochastic nodes in Figure 7.3 and the derivation is given in Appendix 7.A. The lower bound on the marginal log-likelihood for this example is derived by using the expression given in (7.5) and gives the same result as in expression (2.17) in Chapter 2. Therefore, once the natural parameters are transformed back to the original parameters, convergence is assessed in the same manner as before.

Initialise:

$$\boldsymbol{\eta}_{p(a_\ell)\to a_\ell} \leftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix}; \quad \boldsymbol{\eta}_{p(\sigma_\ell^2|a_\ell)\to a_\ell} \leftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix}; \quad \boldsymbol{\eta}_{p(\sigma_\ell^2|a_\ell)\to \sigma_\ell^2} \leftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix};$$
$$\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to \sigma_\ell^2} \leftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix}; \quad \boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to \boldsymbol{\nu}} \leftarrow \begin{bmatrix} \mathbf{0}_v \\ -\frac{1}{2}\mathrm{vec}\left(\boldsymbol{I}_v\right) \end{bmatrix};$$
$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to \boldsymbol{\nu}} \leftarrow \begin{bmatrix} \mathbf{0}_v \\ -\frac{1}{2}\mathrm{vec}\left(\boldsymbol{I}_v\right) \end{bmatrix}; \quad \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to \boldsymbol{\Sigma}} \leftarrow \begin{bmatrix} -n-1 \\ -\frac{1}{2}\mathrm{vec}\left(\boldsymbol{I}_n\right) \end{bmatrix};$$
$$\boldsymbol{\eta}_{p(\boldsymbol{\Sigma})\to \boldsymbol{\Sigma}} \leftarrow \begin{bmatrix} -n-1 \\ -\frac{1}{2}\mathrm{vec}\left(\boldsymbol{I}_n\right) \end{bmatrix}$$

Cycle:

(a)
$$\boldsymbol{\eta}_{a_\ell\to p(a_\ell)} \leftarrow \boldsymbol{\eta}_{p(\sigma_\ell^2|a_\ell)\to a_\ell} \;\;;\;\; \boldsymbol{\eta}_{a_\ell\to p(\sigma_\ell^2|a_\ell)} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{a})\to \boldsymbol{a}}$$
$$\boldsymbol{\eta}_{\sigma_\ell^2\to p(\sigma_\ell^2|a_\ell)} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to \sigma_\ell^2} \;\;;\;\; \boldsymbol{\eta}_{\sigma_\ell^2\to p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)} \leftarrow \boldsymbol{\eta}_{p(\sigma_\ell^2|a_\ell)\to \sigma_\ell^2}$$
$$\boldsymbol{\eta}_{\boldsymbol{\nu}\to p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to \boldsymbol{\nu}} \;\;;\;\; \boldsymbol{\eta}_{\boldsymbol{\nu}\to p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\sigma}^2)} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to \boldsymbol{\nu}}$$
$$\boldsymbol{\eta}_{\boldsymbol{\Sigma}\to p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\sigma}^2)} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{\Sigma})\to \boldsymbol{\Sigma}} \;\;;\;\; \boldsymbol{\eta}_{\boldsymbol{\Sigma}\to p(\boldsymbol{\Sigma})} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to \boldsymbol{\Sigma}}$$

(b)
$$\boldsymbol{\eta}_{p(\sigma_\ell^2|a_\ell)\to a_\ell} \leftarrow \begin{bmatrix} -\frac{1}{2} \\ -\dfrac{\left(\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to \sigma_\ell^2} + \boldsymbol{\eta}_{p(\sigma_\ell^2|a_\ell)\to \sigma_\ell^2}\right)_1 + 1}{\left(\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to \sigma_\ell^2} + \boldsymbol{\eta}_{p(\sigma_\ell^2|a_\ell)\to \sigma_\ell^2}\right)_2} \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\sigma_\ell^2|a_\ell)\to \sigma_\ell^2} \leftarrow \begin{bmatrix} -\frac{3}{2} \\ -\dfrac{\left(\boldsymbol{\eta}_{p(a_\ell)\to a_\ell} + \boldsymbol{\eta}_{p(\sigma_\ell^2|a_\ell)\to a_\ell}\right)_1 + 1}{\left(\boldsymbol{\eta}_{p(a_\ell)\to a_\ell} + \boldsymbol{\eta}_{p(\sigma_\ell^2|a_\ell)\to a_\ell}\right)_2} \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to \sigma_\ell^2} \leftarrow \begin{bmatrix} -\frac{1}{2}K_\ell \\ -\frac{1}{2}\mathcal{F}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to \boldsymbol{u}_\ell} + \boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to \boldsymbol{u}_\ell}\right) \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to \boldsymbol{\nu}}$$
$$\leftarrow \begin{bmatrix} \mathbf{0} \\ -\frac{1}{2}\mathrm{vec}\left(\begin{bmatrix} \boldsymbol{F}^{-1} & \mathbf{0} \\ \mathbf{0} & \underset{1\le\ell\le d}{\mathrm{blockdiag}}\left\{\left(\dfrac{\left(\boldsymbol{\eta}_{p(\sigma_\ell^2|a_\ell)\to \sigma_\ell^2} + \boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to \sigma_\ell^2}\right)_1 + 1}{\left(\boldsymbol{\eta}_{p(\sigma_\ell^2|a_\ell)\to \sigma_\ell^2} + \boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to \sigma_\ell^2}\right)_2}\right)\boldsymbol{I}_{K_\ell}\right\} \end{bmatrix}\right) \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to \boldsymbol{\nu}} \leftarrow \begin{bmatrix} \boldsymbol{C}^\top\left\{\boldsymbol{I}_m \otimes \mathcal{G}\left(\boldsymbol{\eta}_{p(\boldsymbol{\Sigma})\to \boldsymbol{\Sigma}} + \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to \boldsymbol{\Sigma}} \;;\; n\right)\right\}\boldsymbol{y} \\ \frac{1}{2}\mathrm{vec}\left[\boldsymbol{C}^\top\left\{\boldsymbol{I}_m \otimes \mathcal{G}\left(\boldsymbol{\eta}_{p(\boldsymbol{\Sigma})\to \boldsymbol{\Sigma}} + \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to \boldsymbol{\Sigma}} \;;\; n\right)\right\}\boldsymbol{C}\right] \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to \boldsymbol{\Sigma}} \leftarrow \begin{bmatrix} -\frac{m}{2} \\ -\frac{1}{2}\sum_{i=1}^m \mathrm{vec}\left\{\mathcal{H}\left(\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to \boldsymbol{\nu}} + \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to \boldsymbol{\nu}} \;;\; \boldsymbol{y}_i \;,\; \boldsymbol{C}_i\right)\right\} \end{bmatrix}$$

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

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{\Sigma}} \leftarrow \begin{bmatrix} -\frac{m}{2} \\ -\frac{1}{2}\sum_{i=1}^{m}\operatorname{vec}\left\{\mathcal{H}\left(\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to\boldsymbol{\nu}}+\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{\nu}}\,;\,\boldsymbol{y}_i\,,\,\boldsymbol{C}_i\right)\right\} \end{bmatrix}$$

$$\boldsymbol{\eta}^*_{q(\boldsymbol{\nu})} \leftarrow \boldsymbol{\eta}^*_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to(\boldsymbol{\nu})}+\boldsymbol{\eta}^*_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to(\boldsymbol{\nu})}$$
$$\boldsymbol{\eta}^*_{q(\boldsymbol{\Sigma})} \leftarrow \boldsymbol{\eta}^*_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{\Sigma}}+\boldsymbol{\eta}^*_{p(\boldsymbol{\Sigma})\to\boldsymbol{\Sigma}}$$
For $1\le\ell\le d$:
$$\boldsymbol{\eta}^*_{q(a_\ell)} \leftarrow \boldsymbol{\eta}^*_{p(a_\ell)\to a_\ell}+\boldsymbol{\eta}^*_{p(\sigma^2_\ell|a_\ell)\to a_\ell}$$
$$\boldsymbol{\eta}^*_{q(\sigma^2_\ell)} \leftarrow \boldsymbol{\eta}^*_{p(\sigma^2_\ell|a_\ell)\to\sigma^2_\ell}+\boldsymbol{\eta}^*_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to\sigma^2_\ell}$$

Algorithm 11: *VMP algorithm for determining the natural parameter vectors* $\boldsymbol{\eta}^*_{q(a_\ell)}$, $\boldsymbol{\eta}^*_{q(\sigma^2_\ell)}$, $\boldsymbol{\eta}^*_{q(\boldsymbol{\nu})}$ *and* $\boldsymbol{\eta}^*_{q(\boldsymbol{\Sigma})}$ *of the optimal density functions* $q^*(\boldsymbol{\nu})$, $q^*(\boldsymbol{\sigma}^2)$, $q^*(\boldsymbol{a})$ *and* $q^*(\boldsymbol{\Sigma})$ *for approximate inference in Model (2.15).*

7.7.2 Mixture model with measurement error

The factor graph corresponding to model (6.4) is given in Figure 7.4. We impose the same mean field approximations on the joint posterior density functions as given in Chapter 6. Algorithm 12 gives the VMP updates corresponding to this model. Details of the derivations for the updates given in Algorithm 12 are given in Appendix 7.B. In order to impose conjugacy, we initialise the messages corresponding to stochastic nodes $\boldsymbol{a}^x$, $(\boldsymbol{\sigma}^x)^2$, σ^2_ε, a_ε, σ^2_o and a_o to be the Inverse-Gamma $(1,1)$ density function, corresponding to the natural parameter vector being set to $\begin{bmatrix}-2 & -1\end{bmatrix}^\top$. The messages corresponding to the stochastic nodes $\boldsymbol{x}_{\text{unobs}}$ and $\boldsymbol{\mu}^x$ are initialised to be the $\mathrm{N}(0,1)$ density function, which corresponds to the natural parameter vector begin set to $\begin{bmatrix}0 & -\frac{1}{2}\end{bmatrix}^\top$. The messages relating to the columns in node $\boldsymbol{a}$ are initialised to be the Multinomial $\left(a_{i1},\ldots,a_{iK};1,\frac{1}{K},\ldots,\frac{1}{K}\right)$ density function, where K is the number of mixture components used. This corresponds to the natural parameters being set to $\log(1/K),\ldots,\log(1/K)$. The messages relating to $\boldsymbol{\omega}$ are initialised to be the Dirichlet $(\omega_1,\ldots,\omega_K;2,\ldots,2)$ density function and this corresponds to the natural parameters begin set to $1,\ldots,1$. Lastly, the messages corresponding to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are initialised to be the $\mathrm{N}(\boldsymbol{0}_d,\boldsymbol{I}_d)$ density function, where d is the dimension of the corresponding stochastic vector node, and this corresponds to the natural parameter vector being set to $\begin{bmatrix}\boldsymbol{0}_d & -\frac{1}{2}\operatorname{vec}(\boldsymbol{I}_d)\end{bmatrix}^\top$.

Initialise:

$$\boldsymbol{\eta}_{p(a_k^x)\to a_k^x} \leftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix}; \quad \boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to a_k^x} \leftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix}; \quad \boldsymbol{\eta}_{p(\mu_k^x)\to \mu_k^x} \leftarrow \begin{bmatrix} 0 \\ -\frac{1}{2} \end{bmatrix}$$

$$\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to (\sigma_k^x)^2} \leftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix}; \ \boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to (\sigma_k^x)^2} \leftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix}; \ \eta_{p(\omega_k)\to\omega_k} \leftarrow 1$$

$$\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to a_{ik}} \leftarrow \log(1/K); \quad \boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to \mu_k^x} \leftarrow \begin{bmatrix} 0 \\ -\frac{1}{2} \end{bmatrix}$$

$$\boldsymbol{\eta}_{p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to x_{\text{unobs,i}}} \leftarrow \begin{bmatrix} 0 \\ -\frac{1}{2} \end{bmatrix}; \quad \eta_{p(a_{ik}|\omega_k)\to a_{ik}} \leftarrow \log(1/K)$$

$$\eta_{p(\boldsymbol{a}_k|\omega_k)\to\omega_k} \leftarrow 1; \quad \boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\sigma_o^2} \leftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix}; \quad \boldsymbol{\eta}_{p(\boldsymbol{\alpha})\to\boldsymbol{\alpha}} \leftarrow \begin{bmatrix} \mathbf{0}_2 \\ -\frac{1}{2}\text{vec}\left(\boldsymbol{I}_2\right) \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to a_o} \leftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix}; \quad \boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to x_{\text{unobs,i}}} \leftarrow \begin{bmatrix} 0 \\ -\frac{1}{2} \end{bmatrix}; \quad \boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to\sigma_o^2} \leftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\boldsymbol{\alpha}} \leftarrow \begin{bmatrix} \mathbf{0}_2 \\ -\frac{1}{2}\text{vec}\left(\boldsymbol{I}_2\right) \end{bmatrix}; \quad \boldsymbol{\eta}_{p(a_o)\to a_o} \leftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{\beta})\to\boldsymbol{\beta}} \leftarrow \begin{bmatrix} \mathbf{0}_3 \\ -\frac{1}{2}\text{vec}\left(\boldsymbol{I}_3\right) \end{bmatrix}; \quad \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to x_{\text{unobs,i}}} \leftarrow \begin{bmatrix} 0 \\ -\frac{1}{2} \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\sigma_\varepsilon^2} \leftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix}; \quad \boldsymbol{\eta}_{p(a_\varepsilon)\to a_\varepsilon} \leftarrow \begin{bmatrix} -3/2 \\ -1/A_\varepsilon^2 \end{bmatrix}; \quad \boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to a_\varepsilon} \leftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\boldsymbol{\beta}} \leftarrow \begin{bmatrix} \mathbf{0}_3 \\ -\frac{1}{2}\text{vec}\left(\boldsymbol{I}_3\right) \end{bmatrix}; \quad \boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to\sigma_\varepsilon^2} \leftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix}$$

Cycle:

(a) $\boldsymbol{\eta}_{a_k^x\to p(a_k^x)} \leftarrow \boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to a_k^x}; \quad \boldsymbol{\eta}_{a_k^x\to p\left\{(\sigma_k^x)^2|a_k^x\right\}} \leftarrow \boldsymbol{\eta}_{p(a_k^x)\to a_k^x}$

$\boldsymbol{\eta}_{(\sigma_k^x)^2\to p\left\{(\sigma_k^x)^2|a_k^x\right\}} \leftarrow \boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to(\sigma_k^x)^2}$

$\boldsymbol{\eta}_{(\sigma_k^x)^2\to p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma_k^x)^2\right\}} \leftarrow \boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to(\sigma_k^x)^2}$

$\boldsymbol{\eta}_{a_{ik}\to p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma_k^x)^2\right\}} \leftarrow \eta_{p(a_{ik}|\omega_k)\to a_{ik}}; \quad \eta_{a_{ik}\to p(a_{ik}|\omega_k)} \leftarrow \eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to a_{ik}}$

$\eta_{\omega_k\to p(a_{ik}|\omega_k)} \leftarrow \eta_{p(\omega_k)\to\omega_k}; \quad \eta_{\omega_k\to p(\omega_k)} \leftarrow \eta_{p(\boldsymbol{a}_k|\omega_k)\to\omega_k}$

$\boldsymbol{\eta}_{\mu_k^x\to p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma_k^x)^2\right\}} \leftarrow \boldsymbol{\eta}_{p(\mu_k^x)\to\mu_k^x}; \quad \boldsymbol{\eta}_{\mu_k^x\to p(\mu_k^x)} \leftarrow \boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to\mu_k^x}$

$\boldsymbol{\eta}_{\boldsymbol{\alpha}\to p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{\alpha})\to\boldsymbol{\alpha}}; \quad \boldsymbol{\eta}_{\boldsymbol{\alpha}\to p(\boldsymbol{\alpha})} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\boldsymbol{\alpha}}$

$\boldsymbol{\eta}_{\sigma_o^2\to p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)} \leftarrow \boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to\sigma_o^2}; \quad \boldsymbol{\eta}_{\sigma_o^2\to p(\sigma_o^2|a_o)} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\sigma_o^2}$

$\boldsymbol{\eta}_{a_o\to p(\sigma_o^2|a_o)} \leftarrow \boldsymbol{\eta}_{p(a_o)\to a_o}; \quad \boldsymbol{\eta}_{a_o\to p(a_o)} \leftarrow \boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to a_o}$

$\boldsymbol{\eta}_{x_{\text{unobs,i}}\to p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to x_{\text{unobs,i}}} + \boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to x_{\text{unobs,i}}}$

$$\boldsymbol{\eta}_{x_{\text{unobs,i}} \to p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)} \leftarrow \boldsymbol{\eta}_{p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\} \to x_{\text{unobs,i}}} + \boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2) \to x_{\text{unobs,i}}}$$

$$\boldsymbol{\eta}_{x_{\text{unobs,i}} \to p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)} \leftarrow \boldsymbol{\eta}_{p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\} \to x_{\text{unobs,i}}} + \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2) \to x_{\text{unobs,i}}}$$

$$\boldsymbol{\eta}_{\boldsymbol{\beta} \to p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{\beta}) \to \boldsymbol{\beta}}; \quad \boldsymbol{\eta}_{\boldsymbol{\beta} \to p(\boldsymbol{\beta})} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2) \to \boldsymbol{\beta}}$$

$$\boldsymbol{\eta}_{\sigma_\varepsilon^2 \to p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)} \leftarrow \boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon) \to \sigma_\varepsilon^2}; \quad \boldsymbol{\eta}_{\sigma_\varepsilon^2 \to p(\sigma_\varepsilon^2|a_\varepsilon)} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2) \to \sigma_\varepsilon^2}$$

$$\boldsymbol{\eta}_{a_\varepsilon \to p(\sigma_\varepsilon^2|a_\varepsilon)} \leftarrow \boldsymbol{\eta}_{p(a_\varepsilon) \to a_\varepsilon}; \quad \boldsymbol{\eta}_{a_\varepsilon \to p(a_\varepsilon)} \leftarrow \boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon) \to a_\varepsilon}$$

(b) $$\boldsymbol{\eta}_{p(a_k^x) \to a_k^x} \leftarrow \begin{bmatrix} -3/2 \\ -1/\left(A_k^x\right)^2 \end{bmatrix} \quad ; \quad \boldsymbol{\eta}_{p(\mu_k^x) \to \mu_k^x} \leftarrow \begin{bmatrix} \mu_\mu\ \sigma_\mu^2 \\ -1/2\sigma_\mu^2 \end{bmatrix}$$

$$\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\} \to a_k^x} \leftarrow \begin{bmatrix} -\frac{1}{2} \\ -\dfrac{\left(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\} \to (\sigma_k^x)^2} + \boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\} \to (\sigma_k^x)^2}\right)_1^{+1}}{\left(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\} \to (\sigma_k^x)^2} + \boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\} \to (\sigma_k^x)^2}\right)_2} \end{bmatrix}$$

$$\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\} \to (\sigma_k^x)^2} \leftarrow \begin{bmatrix} -3/2 \\ -\dfrac{\left(\boldsymbol{\eta}_{p(a_k^x) \to a_k^x} + \boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\} \to a_k^x}\right)_1^{+1}}{\left(\boldsymbol{\eta}_{p(a_k^x) \to a_k^x} + \boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\} \to a_k^x}\right)_2} \end{bmatrix}$$

$$\boldsymbol{\eta}_{p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\} \to x_{\text{unobs,i}}}$$

$$\leftarrow \begin{bmatrix} -\frac{1}{2}\exp\left(\eta_{p(a_{ik}|\omega_k) \to a_{ik}} + \eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\} \to a_{ik}}\right) \\ \quad \times \dfrac{\left(\boldsymbol{\eta}_{p(\mu_k^x) \to \mu_k^x} + \boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\} \to \mu_k^x}\right)_1}{\left(\boldsymbol{\eta}_{p(\mu_k^x) \to \mu_k^x} + \boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\} \to \mu_k^x}\right)_2} \\ \times \dfrac{\left(\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\} \to (\sigma_k^x)^2} + \boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\} \to (\sigma_k^x)^2}\right)_1^{+1}}{\left(\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\} \to (\sigma_k^x)^2} + \boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\} \to (\sigma_k^x)^2}\right)_2} \\ \\ -\frac{1}{2}\exp\left(\eta_{p(a_{ik}|\omega_k) \to a_{ik}} + \eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\} \to a_{ik}}\right) \\ \times \dfrac{\left(\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\} \to (\sigma_k^x)^2} + \boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\} \to (\sigma_k^x)^2}\right)_1^{+1}}{\left(\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\} \to (\sigma_k^x)^2} + \boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\} \to (\sigma_k^x)^2}\right)_2} \end{bmatrix}$$

$$\eta_{p(a_{ik}|\omega_k) \to a_{ik}} \leftarrow \left(\psi\left\{\left(\eta_{p(\omega_k) \to \omega_k} + \eta_{p(a_{ik}|\omega_k) \to \omega_k}\right)_k + 1\right\} - \psi\left[\sum_{k=1}^{K}\left\{\left(\eta_{p(\omega_k) \to \omega_k} + \eta_{p(a_{ik}|\omega_k) \to \omega_k}\right)_k + 1\right\}\right]\right)$$

$$\eta_{p(a_{ik}|\omega_k) \to \omega_k} \leftarrow \exp\left(\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\} \to a_{ik}} + \eta_{p(a_{ik}|\omega_k) \to a_{ik}}\right)$$

$$\eta_{p(\omega_k) \to \omega_k} \leftarrow \alpha_k - 1 \quad ; \quad \boldsymbol{\eta}_{p(\boldsymbol{\beta}) \to \boldsymbol{\beta}} \leftarrow \begin{bmatrix} \mathbf{0} \\ -\frac{1}{2}\text{vec}\left(\frac{1}{\sigma_\beta^2}\boldsymbol{I}_3\right) \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to x_{\text{unobs,i}}} \leftarrow \begin{bmatrix} \dfrac{\left(\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to\sigma_o^2}{}^{+1}\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\sigma_o^2}\right)_1{}^{+1}}{\left(\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to\sigma_o^2}{}^{+1}\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\sigma_o^2}\right)_2} \\ \times \left\{ -\frac{1}{2}\boldsymbol{o}_{x_{\text{unobs},i}} \dfrac{\left(\boldsymbol{\eta}_{p(\alpha_1)\to\alpha_1}{}^{+1}\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\alpha_1}\right)_1}{\left(\boldsymbol{\eta}_{p(\alpha_1)\to\alpha_1}{}^{+1}\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\alpha_1}\right)_2} \right. \\ \left. +\mathcal{J}\left(\boldsymbol{\eta}_{p(\boldsymbol{\alpha})\to\boldsymbol{\alpha}} + \boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\boldsymbol{\alpha}}; 1, 2\right)\right\} \\ -\dfrac{\left(\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to\sigma_o^2}{}^{+1}\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\sigma_o^2}\right)_1{}^{+1}}{2\left(\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to\sigma_o^2}{}^{+1}\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\sigma_o^2}\right)_2} \\ \times\mathcal{K}\left(\boldsymbol{\eta}_{p(\alpha_1)\to\alpha_1} + \boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\alpha_1}\right) \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\boldsymbol{\alpha}}$$

$$\leftarrow \begin{bmatrix} \dfrac{\left(\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to\sigma_o^2}{}^{+1}\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\sigma_o^2}\right)_1{}^{+1}}{\left(\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to\sigma_o^2}{}^{+1}\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\sigma_o^2}\right)_2} \times E^{nat}_{q(\boldsymbol{x}_{\text{unobs}})}\left(\widetilde{\boldsymbol{X}}\right)^\top \boldsymbol{o} \\ -\dfrac{\left(\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to\sigma_o^2}{}^{+1}\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\sigma_o^2}\right)_1{}^{+1}}{2\left(\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to\sigma_o^2}{}^{+1}\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\sigma_o^2}\right)_2} \times \text{vec}\left\{E^{nat}_{q(\boldsymbol{x}_{\text{unobs}})}\left(\widetilde{\boldsymbol{X}}^\top\widetilde{\boldsymbol{X}}\right)\right\} \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{\alpha})\to\boldsymbol{\alpha}} \leftarrow \begin{bmatrix} \mathbf{0} \\ -\frac{1}{2}\text{vec}\left(\frac{1}{\sigma_\alpha^2}\boldsymbol{I}_2\right) \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\sigma_o^2} \leftarrow \begin{bmatrix} -\frac{n}{2} \\ -\frac{1}{2}E^{nat}_{q(\boldsymbol{x}_{\text{unobs}},\boldsymbol{\alpha})}\left(\|\boldsymbol{o}-\widetilde{\boldsymbol{X}}\boldsymbol{\alpha}\|^2\right) \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to\sigma_o^2} \leftarrow \begin{bmatrix} -\frac{3}{2} \\ -\dfrac{\left(\boldsymbol{\eta}_{p(a_o)\to a_o}{}^{+1}\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to a_o}\right)_1{}^{+1}}{\left(\boldsymbol{\eta}_{p(a_o)\to a_o}{}^{+1}\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to a_o}\right)_2} \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to a_o} \leftarrow \begin{bmatrix} -\frac{1}{2} \\ -\dfrac{\left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\sigma_o^2}{}^{+1}\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to\sigma_o^2}\right)_1{}^{+1}}{\left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\sigma_o^2}{}^{+1}\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to\sigma_o^2}\right)_2} \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\sigma_\varepsilon^2} \leftarrow \begin{bmatrix} -\frac{n}{2} \\ -\frac{1}{2}E^{nat}_{q(\boldsymbol{x}_{\text{unobs}},\boldsymbol{\beta})}\left(\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2\right) \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to a_\varepsilon} \leftarrow \begin{bmatrix} -\frac{1}{2} \\ -\dfrac{\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\sigma_\varepsilon^2}{}^{+1}\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to\sigma_\varepsilon^2}\right)_1{}^{+1}}{\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\sigma_\varepsilon^2}{}^{+1}\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to\sigma_\varepsilon^2}\right)_2} \end{bmatrix}$$

$$\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\rightarrow(\sigma_k^x)^2}$$

$$\leftarrow\begin{bmatrix} -\frac{1}{2}\exp\left(\eta_{p(a_{ik}|\omega_k)\rightarrow a_{ik}}+\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\rightarrow a_{ik}}\right) \\ -\frac{1}{2}\exp\left(\eta_{p(a_{ik}|\omega_k)\rightarrow a_{ik}}+\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\rightarrow a_{ik}}\right) \\ \times\mathcal{I}\left\{\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\rightarrow x_{\text{unobs,i}}}+\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\rightarrow x_{\text{unobs,i}}}\right.\right. \\ \left.+\boldsymbol{\eta}_{p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\rightarrow x_{\text{unobs,i}}}\right),\left(\boldsymbol{\eta}_{p(\mu_k^x)\rightarrow\mu_k^x}\right. \\ \left.\left.+\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\rightarrow\mu_k^x}\right)\right\}\end{bmatrix}$$

$$\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\rightarrow a_{ik}}$$

$$\begin{aligned}\leftarrow-\frac{1}{2}\Bigg[&\log\left\{-\left(\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\rightarrow(\sigma_k^x)^2}+\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\rightarrow(\sigma_k^x)^2}\right)_2\right\} \\ &-\psi\left\{-\left(\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\rightarrow(\sigma_k^x)^2}+\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\rightarrow(\sigma_k^x)^2}\right)_1-1\right\} \\ &-\left\{\frac{\left(\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\rightarrow(\sigma_k^x)^2}+\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\rightarrow(\sigma_k^x)^2}\right)_1^{+1}}{\left(\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\rightarrow(\sigma_k^x)^2}+\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\rightarrow(\sigma_k^x)^2}\right)_2}\right\} \\ &\times\mathcal{I}\left\{\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\rightarrow x_{\text{unobs,i}}}+\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\rightarrow x_{\text{unobs,i}}}\right.\right. \\ &\left.+\boldsymbol{\eta}_{p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\rightarrow x_{\text{unobs,i}}}\right), \\ &\left.\left(\boldsymbol{\eta}_{p(\mu_k^x)\rightarrow\mu_k^x}+\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\rightarrow\mu_k^x}\right)\right\}\Bigg]\end{aligned}$$

$$\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\rightarrow\mu_k^x}\leftarrow\begin{bmatrix} -\frac{1}{2}\,\exp\left(\eta_{p(a_{ik}|\omega_k)\rightarrow a_{ik}}+\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\rightarrow a_{ik}}\right) \\ \times\left\{E^{\,nat}_{q(\boldsymbol{x}_{\text{unobs}})}(\boldsymbol{x})\right\}_i \\ \times\dfrac{\left(\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\rightarrow(\sigma_k^x)^2}+\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\rightarrow(\sigma_k^x)^2}\right)_1^{+1}}{\left(\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\rightarrow(\sigma_k^x)^2}+\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\rightarrow(\sigma_k^x)^2}\right)_2} \\ -\frac{1}{2}\,\exp\left(\eta_{p(a_{ik}|\omega_k)\rightarrow a_{ik}}+\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\rightarrow a_{ik}}\right) \\ \times\dfrac{\left(\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\rightarrow(\sigma_k^x)^2}+\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\rightarrow(\sigma_k^x)^2}\right)_1^{+1}}{\left(\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\rightarrow(\sigma_k^x)^2}+\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\rightarrow(\sigma_k^x)^2}\right)_2}\end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\rightarrow\boldsymbol{\beta}}\leftarrow\begin{bmatrix}\dfrac{\left(\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\rightarrow a_\varepsilon}+\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\rightarrow\sigma_\varepsilon^2}\right)_1^{+1}}{\left(\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\rightarrow a_\varepsilon}+\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\rightarrow\sigma_\varepsilon^2}\right)_2} \\ \times E^{\,nat}_{q(\boldsymbol{x}_{\text{unobs}})}(\boldsymbol{X})^\top\boldsymbol{y} \\ -\dfrac{\left(\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\rightarrow\sigma_\varepsilon^2}+\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\rightarrow\sigma_\varepsilon^2}\right)_1^{+1}}{2\left(\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\rightarrow\sigma_\varepsilon^2}+\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\rightarrow\sigma_\varepsilon^2}\right)_2} \\ \times\text{vec}\left\{E^{\,nat}_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)\right\}\end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to\sigma_\varepsilon^2} \leftarrow \begin{bmatrix} -\frac{3}{2} \\ -\frac{\left(\boldsymbol{\eta}_{p(a_\varepsilon)\to a_\varepsilon} + \boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to a_\varepsilon}\right)_1 + 1}{\left(\boldsymbol{\eta}_{p(a_\varepsilon)\to a_\varepsilon} + \boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to a_\varepsilon}\right)_2} \end{bmatrix}; \quad \boldsymbol{\eta}_{p(a_\varepsilon)\to a_\varepsilon} \leftarrow \begin{bmatrix} -\frac{3}{2} \\ -\frac{1}{A_\varepsilon^2} \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to x_{\text{unobs,i}}} \leftarrow \begin{bmatrix} \frac{\left(\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to\sigma_\varepsilon^2} + \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\sigma_\varepsilon^2}\right)_1 + 1}{\left(\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to\sigma_\varepsilon^2} + \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\sigma_\varepsilon^2}\right)_2} \\ \times \left\{ -\frac{1}{2}\boldsymbol{y}_{x_{\text{unobs},i}} \frac{\left(\boldsymbol{\eta}_{p(\beta_x)\to\beta_x} + \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\beta_x}\right)_1}{\left(\boldsymbol{\eta}_{p(\beta_x)\to\beta_x} + \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\beta_x}\right)_2} \right. \\ -\mathcal{J}\left(\boldsymbol{\eta}_{p(\boldsymbol{\beta})\to\boldsymbol{\beta}} + \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\boldsymbol{\beta}}; 1, 3\right) \\ \left. -\boldsymbol{c}_{x_{\text{unobs},i}}\, \mathcal{J}\left(\boldsymbol{\eta}_{p(\boldsymbol{\beta})\to\boldsymbol{\beta}} + \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\boldsymbol{\beta}}; 2, 3\right) \right\} \\ \\ \frac{1}{4}\frac{\left(\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to\sigma_\varepsilon^2} + \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\sigma_\varepsilon^2}\right)_1 + 1}{\left(\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to\sigma_\varepsilon^2} + \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\sigma_\varepsilon^2}\right)_2} \\ \times \frac{\left(\boldsymbol{\eta}_{p(\beta_x)\to\beta_x} + \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\beta_x}\right)_1}{\left(\boldsymbol{\eta}_{p(\beta_x)\to\beta_x} + \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\beta_x}\right)_2} \end{bmatrix}$$

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

$$\boldsymbol{\eta}^*_{q(\boldsymbol{\alpha})} \leftarrow \boldsymbol{\eta}^*_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\boldsymbol{\alpha}} + \boldsymbol{\eta}^*_{p(\boldsymbol{\alpha})\to\boldsymbol{\alpha}}; \quad \boldsymbol{\eta}^*_{q(\sigma_o^2)} \leftarrow \boldsymbol{\eta}^*_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\sigma_o^2} + \boldsymbol{\eta}^*_{p(\sigma_o^2|a_o)\to\sigma_o^2}$$
$$\boldsymbol{\eta}^*_{q(a_o)} \leftarrow \boldsymbol{\eta}^*_{p(\sigma_o^2|a_o)\to a_o} + \boldsymbol{\eta}^*_{p(a_o)\to a_o}$$

For $1 \le k \le K$:
$$\boldsymbol{\eta}^*_{q(a_k^x)} \leftarrow \boldsymbol{\eta}^*_{p(a_k^x)\to a_k^x} + \boldsymbol{\eta}^*_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to a_k^x}$$
$$\boldsymbol{\eta}^*_{q\left\{(\sigma_k^x)^2\right\}} \leftarrow \boldsymbol{\eta}^*_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to(\sigma_k^x)^2} + \boldsymbol{\eta}^*_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to(\sigma_k^x)^2}$$
$$\boldsymbol{\eta}^*_{q(\mu_k^x)} \leftarrow \boldsymbol{\eta}^*_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to\mu_k^x} + \boldsymbol{\eta}^*_{p(\mu_k^x)\to\mu_k^x}$$

For $1 \le i \le n$ and $1 \le k \le K$:
$$\boldsymbol{\eta}^*_{q(a_{ik})} \leftarrow \eta^*_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to a_{ik}} + \eta^*_{p(a_{ik}|\omega_k)\to a_{ik}}$$
$$\boldsymbol{\eta}^*_{q(\omega_k)} \leftarrow \eta^*_{p(a_{ik}|\omega_k)\to\omega_k} + \eta^*_{p(\omega_k)\to\omega_k}$$

$$\boldsymbol{\eta}^*_{q(\boldsymbol{\beta})} \leftarrow \boldsymbol{\eta}^*_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\boldsymbol{\beta}} + \boldsymbol{\eta}^*_{p(\boldsymbol{\beta})\to\boldsymbol{\beta}}; \quad \boldsymbol{\eta}^*_{q(\sigma_\varepsilon^2)} \leftarrow \boldsymbol{\eta}^*_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\sigma_\varepsilon^2} + \boldsymbol{\eta}^*_{p(\sigma_\varepsilon^2|a_\varepsilon)\to\sigma_\varepsilon^2}$$
$$\boldsymbol{\eta}^*_{q(a_\varepsilon)} \leftarrow \boldsymbol{\eta}^*_{p(\sigma_\varepsilon^2|a_\varepsilon)\to a_\varepsilon} + \boldsymbol{\eta}^*_{p(a_\varepsilon)\to a_\varepsilon}$$

For $1 \le i \le n_{\text{unobs}}$:

$$\boldsymbol{\eta}^*_{q(x_{\text{unobs,i}})} \leftarrow \boldsymbol{\eta}^*_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to x_{\text{unobs,i}}} + \boldsymbol{\eta}^*_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to x_{\text{unobs,i}}} + \boldsymbol{\eta}^*_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to x_{\text{unobs,i}}}$$

Algorithm 12: *VMP algorithm for determining the natural parameter vectors of the optimal density functions for approximate inference in Model (6.4).*

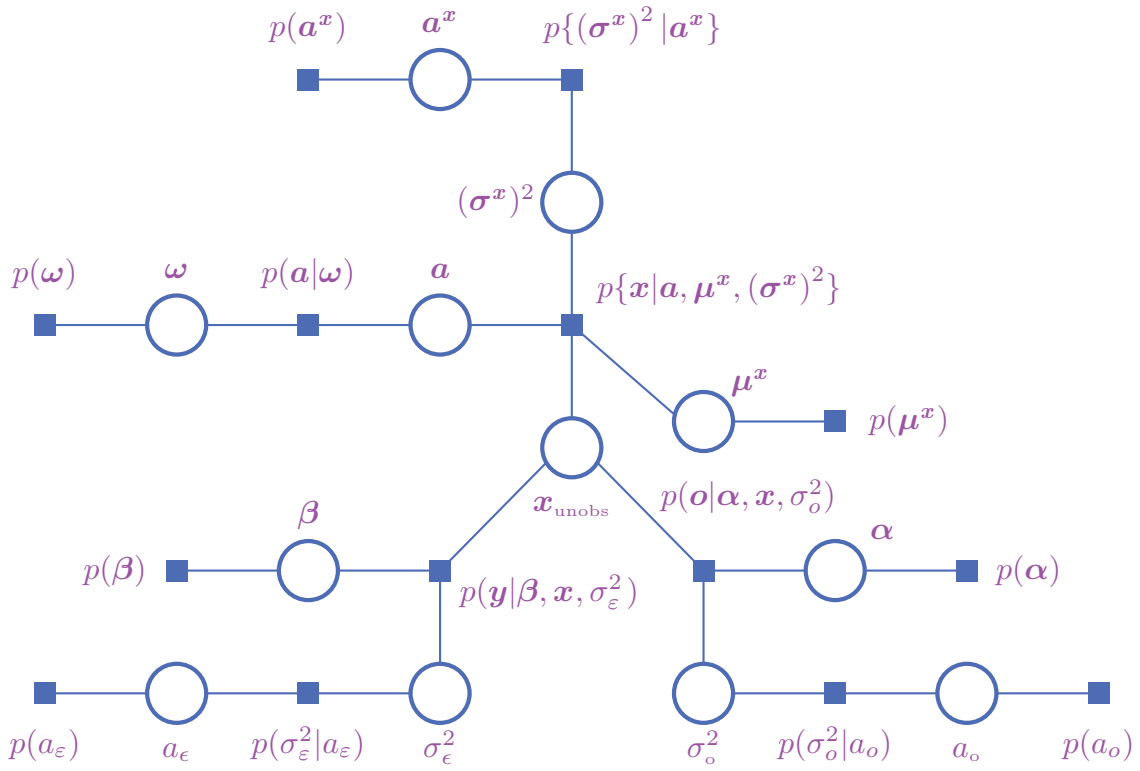


Figure 7.4: *Factor graph corresponding to the mixture model with measurement error given in (6.4).*

7.8 Discussion

The general VMP algorithm, presented as Algorithm 10, is a useful guide, in terms of notation, for representing the three main phases in a VMP scheme. The two examples in this chapter provided further details on the use of natural parameter forms and primitives for certain density functions when using VMP. The extension of VMP to arbitrarily large models may not be immediately obvious from the examples discussed here, however the ease by which certain primitives and results can be used for larger and more complex models makes VMP an attractive alternative to MFVB.

For the interested reader, the efficiency of the VMP approach can be seen from the examples given in Wand (2015). Wand (2015) defines the notion of a *fragment*, which is a sub-graph of a factor graph that consists only of a single factor and its corresponding stochastic neighbor nodes. These *fragments* represent the underlying structure of a wide range of large semiparametric regression models and gives VMP the advantage of being more amenable to modulatization and extension to larger compex models.

7.A Derivation of Algorithm 11

In this section we describe in detail, the derivations for Algorithm 11. These derivations have made use of the natural parameter forms, primitives and functions defined in Sections 7.3, 7.4 and 7.5.

7.A.1 Step 1: Initialise factor to stochastic node messages

The derivation of the updates corresponding to the $\boldsymbol{\Sigma}$ node in Figure 7.3 were given in section 7.7.1. Here we give the derivations of the updates corresponding to the rest of the nodes in the factor graph. The Inverse-Gamma messages are initialised to be the Inverse-Gamma$(1,1)$ density function and the multivariate normal messages are initialised to be the $\mathrm{N}(\mathbf{0}_v, \boldsymbol{I}_v)$ density function, where $v = p + \sum_{\ell=1}^{d} K_\ell$. These messages can be written as

$$
\begin{aligned}
\boldsymbol{m}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{\nu}}(\boldsymbol{\beta},\boldsymbol{u}) &\leftarrow \exp\left\{\begin{bmatrix}\boldsymbol{\nu}\\ \mathrm{vec}\left(\boldsymbol{\nu}\boldsymbol{\nu}^\top\right)\end{bmatrix}\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{\nu}}\right\}\\
\boldsymbol{m}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to\boldsymbol{\nu}}(\boldsymbol{\beta},\boldsymbol{u}) &\leftarrow \exp\left\{\begin{bmatrix}\boldsymbol{\nu}\\ \mathrm{vec}\left(\boldsymbol{\nu}\boldsymbol{\nu}^\top\right)\end{bmatrix}\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to\boldsymbol{\nu}}\right\}\\
\boldsymbol{m}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to\boldsymbol{\sigma}^2}\left(\boldsymbol{\sigma}^2\right) &\leftarrow \exp\left\{\sum_{\ell=1}^{d}\begin{bmatrix}\log\left(\sigma_\ell^2\right)\\ 1/\sigma_\ell^2\end{bmatrix}^\top\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to\sigma_\ell^2}\right\}\\
\boldsymbol{m}_{p(\boldsymbol{\sigma}^2|\boldsymbol{a})\to\boldsymbol{\sigma}^2}\left(\boldsymbol{\sigma}^2\right) &\leftarrow \exp\left\{\sum_{\ell=1}^{d}\begin{bmatrix}\log\left(\sigma_\ell^2\right)\\ 1/\sigma_\ell^2\end{bmatrix}^\top\boldsymbol{\eta}_{p(\sigma_\ell^2|a_\ell)\to\sigma_\ell^2}\right\}\\
\boldsymbol{m}_{p(\boldsymbol{\sigma}^2|\boldsymbol{a})\to\boldsymbol{a}}(\boldsymbol{a}) &\leftarrow \exp\left\{\sum_{\ell=1}^{d}\begin{bmatrix}\log\left(a_\ell\right)\\ 1/a_\ell\end{bmatrix}^\top\boldsymbol{\eta}_{p(\sigma_\ell^2|a_\ell)\to a_\ell}\right\}\\
\boldsymbol{m}_{p(\boldsymbol{a})\to\boldsymbol{a}}(\boldsymbol{a}) &\leftarrow \exp\left\{\sum_{\ell=1}^{d}\begin{bmatrix}\log\left(a_\ell\right)\\ 1/a_\ell\end{bmatrix}^\top\boldsymbol{\eta}_{p(a_\ell)\to a_\ell}\right\}
\end{aligned}
$$

and these settings correspond to the natural parameter vectors having the following initial values:

$$
\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{\nu}} \leftarrow \begin{bmatrix}\mathbf{0}_v\\ -\frac{1}{2}\mathrm{vec}\left(\boldsymbol{I}_v\right)\end{bmatrix},\ \boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to\boldsymbol{\nu}} \leftarrow \begin{bmatrix}\mathbf{0}_v\\ -\frac{1}{2}\mathrm{vec}\left(\boldsymbol{I}_v\right)\end{bmatrix},
$$

$$\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2) \to \sigma_\ell^2} \leftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix}, \; \boldsymbol{\eta}_{p(\sigma_\ell^2|a_\ell) \to \sigma_\ell^2} \leftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix},$$

$$\boldsymbol{\eta}_{p(\sigma_\ell^2|a_\ell) \to a_\ell} \leftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix}, \; \boldsymbol{\eta}_{p(a_\ell) \to a_\ell} \leftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix}.$$

7.A.2 Step 2 (a): Update stochastic node to factor messages

The next stochastic node to factor message to update is $\boldsymbol{m}_{\boldsymbol{\nu} \to p(\boldsymbol{y}|\boldsymbol{\nu}, \boldsymbol{\Sigma})}(\boldsymbol{\nu})$, and since the only other neighbour of $\boldsymbol{\nu}$ apart from $p(\boldsymbol{y}|\boldsymbol{\nu}, \boldsymbol{\Sigma})$ is $p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)$, this message takes the form

$$\boldsymbol{m}_{\boldsymbol{\nu} \to p(\boldsymbol{y}|\boldsymbol{\nu}, \boldsymbol{\Sigma})}(\boldsymbol{\nu}) \leftarrow \boldsymbol{m}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2) \to \boldsymbol{\nu}}(\boldsymbol{\nu})$$

and can be written as

$$\boldsymbol{m}_{\boldsymbol{\nu} \to p(\boldsymbol{y}|\boldsymbol{\nu}, \boldsymbol{\Sigma})}(\boldsymbol{\nu}) \leftarrow \exp\left\{ \begin{bmatrix} \boldsymbol{\nu} \\ \operatorname{vec}(\boldsymbol{\nu}\boldsymbol{\nu}^\top) \end{bmatrix} \boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2) \to \boldsymbol{\nu}} \right\}$$

where

$$\boldsymbol{\eta}_{\boldsymbol{\nu} \to p(\boldsymbol{y}|\boldsymbol{\nu}, \boldsymbol{\sigma}^2)} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2) \to \boldsymbol{\nu}}.$$

Similarly,

$$\begin{aligned}
\boldsymbol{m}_{\boldsymbol{\nu} \to p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)}(\boldsymbol{\nu}) &\leftarrow \boldsymbol{m}_{p(\boldsymbol{y}|\boldsymbol{\nu}, \boldsymbol{\Sigma}) \to \boldsymbol{\nu}}(\boldsymbol{\nu}) \\
\boldsymbol{m}_{\boldsymbol{\sigma}^2 \to p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)}(\boldsymbol{\sigma}^2) &\leftarrow \boldsymbol{m}_{p(\boldsymbol{\sigma}^2|\boldsymbol{a}) \to \boldsymbol{\sigma}^2}(\boldsymbol{\sigma}^2) \\
\boldsymbol{m}_{\boldsymbol{\sigma}^2 \to p(\boldsymbol{\sigma}^2|\boldsymbol{a})}(\boldsymbol{\sigma}^2) &\leftarrow \boldsymbol{m}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2) \to \boldsymbol{\sigma}^2}(\boldsymbol{\sigma}^2) \\
\boldsymbol{m}_{\boldsymbol{a} \to p(\boldsymbol{\sigma}^2|\boldsymbol{a})}(\boldsymbol{a}) &\leftarrow \boldsymbol{m}_{p(\boldsymbol{a}) \to \boldsymbol{a}}(\boldsymbol{a}) \\
\boldsymbol{m}_{\boldsymbol{a} \to p(\boldsymbol{a})}(\boldsymbol{a}) &\leftarrow \boldsymbol{m}_{p(\boldsymbol{\sigma}^2|\boldsymbol{a}) \to \boldsymbol{a}}(\boldsymbol{a})
\end{aligned}$$

which is equivalent to the following natural parameter updates:

$$\begin{aligned}
\boldsymbol{\eta}_{\boldsymbol{\nu} \to p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)} &\leftarrow \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu}, \boldsymbol{\Sigma}) \to \boldsymbol{\nu}} \\
\boldsymbol{\eta}_{\sigma_\ell^2 \to p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)} &\leftarrow \boldsymbol{\eta}_{p(\sigma_\ell^2|a_\ell) \to \sigma_\ell^2} \\
\boldsymbol{\eta}_{\sigma_\ell^2 \to p(\sigma_\ell^2|a_\ell)} &\leftarrow \boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2) \to \sigma_\ell^2} \\
\boldsymbol{\eta}_{a_\ell \to p(\sigma_\ell^2|a_\ell)} &\leftarrow \boldsymbol{\eta}_{p(a_\ell) \to a_\ell} \\
\boldsymbol{\eta}_{a_\ell \to p(a_\ell)} &\leftarrow \boldsymbol{\eta}_{p(\sigma_\ell^2|a_\ell) \to a_\ell}
\end{aligned}$$

7.A.3 Step 2 (b): Update factor to stochastic node messages

The next factor to stochastic node message to update is $\boldsymbol{m}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{\nu}}(\boldsymbol{\nu})$ and takes the form

$$\boldsymbol{m}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{\nu}}(\boldsymbol{\nu}) \leftarrow \exp\Big\{\int_{\mathbb{R}^{n\times n}} \frac{1}{Z}\boldsymbol{m}_{\boldsymbol{\Sigma}\to p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})}(\boldsymbol{\Sigma}) \times \boldsymbol{m}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}) \times \log p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\, d\boldsymbol{\Sigma}\Big\}$$

where Z is the normalising constant.

$$\begin{aligned}
&\boldsymbol{m}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{\nu}}(\boldsymbol{\nu}) \\
&\propto \exp\Bigg[\int_{\mathbb{R}^{n\times n}} \left\{\begin{bmatrix}\log|\boldsymbol{\Sigma}| \\ \operatorname{vec}\left(\boldsymbol{\Sigma}^{-1}\right)\end{bmatrix}^{\top}\left(\boldsymbol{\eta}_{p(\boldsymbol{\Sigma})\to\boldsymbol{\Sigma}} + \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{\Sigma}}\right)\right\} \\
&\qquad \times \left\{\begin{bmatrix}\boldsymbol{\nu} \\ \operatorname{vec}\left(\boldsymbol{\nu}\boldsymbol{\nu}^{\top}\right)\end{bmatrix}^{\top}\begin{bmatrix}\boldsymbol{C}^{\top}\left(\boldsymbol{I}_m \otimes \boldsymbol{\Sigma}^{-1}\right)\boldsymbol{y} \\ \frac{1}{2}\operatorname{vec}\left(\boldsymbol{C}^{\top}\left(\boldsymbol{I}_m \otimes \boldsymbol{\Sigma}^{-1}\right)\boldsymbol{C}\right)\end{bmatrix}\right\} d\boldsymbol{\Sigma}\Bigg] \\
&= \exp\left\{\begin{bmatrix}\boldsymbol{\nu} \\ \operatorname{vec}\left(\boldsymbol{\nu}\boldsymbol{\nu}^{\top}\right)\end{bmatrix}^{\top}\begin{bmatrix}\boldsymbol{C}^{\top}\left\{\boldsymbol{I}_m \otimes \int_{\mathbb{R}^{n\times n}} \boldsymbol{\Sigma}^{-1} p_{\mathrm{IW}_{\mathrm{nat}}}\left(\boldsymbol{\Sigma};\, \boldsymbol{\eta}_{p(\boldsymbol{\Sigma})\to\boldsymbol{\Sigma}} + \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{\Sigma}}\right) d\boldsymbol{\Sigma}\right\}\boldsymbol{y} \\ \frac{1}{2}\operatorname{vec}\left[\boldsymbol{C}^{\top}\left\{\boldsymbol{I}_m \otimes \int_{\mathbb{R}^{n\times n}} \boldsymbol{\Sigma}^{-1} p_{\mathrm{IW}_{\mathrm{nat}}}\left(\boldsymbol{\Sigma};\, \boldsymbol{\eta}_{p(\boldsymbol{\Sigma})\to\boldsymbol{\Sigma}} + \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{\Sigma}}\right) d\boldsymbol{\Sigma}\right\}\boldsymbol{C}\right]\end{bmatrix}\right\} \\
&= \exp\left\{\begin{bmatrix}\boldsymbol{\nu} \\ \operatorname{vec}\left(\boldsymbol{\nu}\boldsymbol{\nu}^{\top}\right)\end{bmatrix}^{\top}\begin{bmatrix}\boldsymbol{C}^{\top}\left\{\boldsymbol{I}_m \otimes \mathcal{G}\left(\boldsymbol{\eta}_{p(\boldsymbol{\Sigma})\to\boldsymbol{\Sigma}} + \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{\Sigma}}\right)\right\}\boldsymbol{y} \\ \frac{1}{2}\operatorname{vec}\left[\boldsymbol{C}^{\top}\left\{\boldsymbol{I}_m \otimes \mathcal{G}\left(\boldsymbol{\eta}_{p(\boldsymbol{\Sigma})\to\boldsymbol{\Sigma}} + \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{\Sigma}}\right)\right\}\boldsymbol{C}\right]\end{bmatrix}\right\}.
\end{aligned}$$

This means that the natural parameter vector has the following update:

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{\nu}} \leftarrow \begin{bmatrix}\boldsymbol{C}^{\top}\left\{\boldsymbol{I}_m \otimes \mathcal{G}\left(\boldsymbol{\eta}_{p(\boldsymbol{\Sigma})\to\boldsymbol{\Sigma}} + \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{\Sigma}}\right)\right\}\boldsymbol{y} \\ \frac{1}{2}\operatorname{vec}\left[\boldsymbol{C}^{\top}\left\{\boldsymbol{I}_m \otimes \mathcal{G}\left(\boldsymbol{\eta}_{p(\boldsymbol{\Sigma})\to\boldsymbol{\Sigma}} + \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{\Sigma}}\right)\right\}\boldsymbol{C}\right]\end{bmatrix}.$$

The next message begins with

$$\boldsymbol{m}_{p\left(\boldsymbol{\nu}|\boldsymbol{\sigma}^2\right)\to\boldsymbol{\nu}}(\boldsymbol{\nu}) \leftarrow \exp\Big\{\int_{\mathbb{R}^d_{\geq 0}} \frac{1}{Z}\boldsymbol{m}_{\boldsymbol{\sigma}^2\to p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)}\left(\boldsymbol{\sigma}^2\right) \times \boldsymbol{m}_{p\left(\boldsymbol{\nu}|\boldsymbol{\sigma}^2\right)\to\boldsymbol{\sigma}^2}\left(\boldsymbol{\sigma}^2\right) \times \log p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\, d\boldsymbol{\sigma}^2\Big\}$$

Collecting like terms in $\boldsymbol{m}_{\boldsymbol{\sigma}^2 \to p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)}(\boldsymbol{\sigma}^2)$ and $\boldsymbol{m}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2) \to \boldsymbol{\sigma}^2}(\boldsymbol{\sigma}^2)$ gives:

$$
\begin{aligned}
&\boldsymbol{m}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2) \to \boldsymbol{\nu}}(\boldsymbol{\nu}) \\
&\quad \propto \exp\left[\int_{\mathbb{R}^d_{\geq 0}} \left\{\sum_{\ell=1}^{d} \begin{bmatrix} \log\left(\sigma_\ell^2\right) \\ 1/\sigma_\ell^2 \end{bmatrix}^\top \left(\boldsymbol{\eta}_{p(\sigma_\ell^2|a_\ell) \to \sigma_\ell^2} + \boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2) \to \sigma_\ell^2}\right)\right\}\right. \\
&\qquad \left.\times \left\{\begin{bmatrix} \boldsymbol{\nu} \\ \operatorname{vec}\left(\boldsymbol{\nu}\boldsymbol{\nu}^\top\right) \end{bmatrix}^\top \begin{bmatrix} \mathbf{0} \\ -\frac{1}{2}\operatorname{vec}\left(\begin{bmatrix} \boldsymbol{F}^{-1} & \mathbf{0} \\ \mathbf{0} & \underset{1\leq\ell\leq d}{\operatorname{blockdiag}}\left(\frac{1}{\sigma_\ell^2}\boldsymbol{I}_{K_\ell}\right) \end{bmatrix}\right) \end{bmatrix}\right\} d\boldsymbol{\Sigma}\right] \\
&= \exp\left\{\begin{bmatrix} \boldsymbol{\nu} \\ \operatorname{vec}\left(\boldsymbol{\nu}\boldsymbol{\nu}^\top\right) \end{bmatrix}^\top \right. \\
&\qquad \left.\times \begin{bmatrix} \mathbf{0} \\ -\frac{1}{2}\operatorname{vec}\left(\begin{bmatrix} \boldsymbol{F}^{-1} & \mathbf{0} \\ \mathbf{0} & \underset{1\leq\ell\leq d}{\operatorname{blockdiag}}\left\{\left(\int_0^\infty \frac{1}{\sigma_\ell^2} p_{\mathrm{IG}_{\mathrm{nat}}}\left(\boldsymbol{\eta}^\oplus\right) d\sigma_\ell^2\right)\boldsymbol{I}_{K_\ell}\right\} \end{bmatrix}\right) \end{bmatrix}\right\}
\end{aligned}
$$

where $\boldsymbol{\eta}^\oplus \equiv \boldsymbol{\eta}_{p(\sigma_\ell^2|a_\ell) \to \sigma_\ell^2} + \boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2) \to \sigma_\ell^2} \equiv \begin{bmatrix} \eta_1^\oplus & \eta_2^\oplus \end{bmatrix}^\top$. Using Primitive 7.4.6, we get

$$
= \exp\left\{\begin{bmatrix} \boldsymbol{\nu} \\ \operatorname{vec}\left(\boldsymbol{\nu}\boldsymbol{\nu}^\top\right) \end{bmatrix}^\top \begin{bmatrix} \mathbf{0} \\ -\frac{1}{2}\operatorname{vec}\left(\begin{bmatrix} \boldsymbol{F}^{-1} & \mathbf{0} \\ \mathbf{0} & \underset{1\leq\ell\leq d}{\operatorname{blockdiag}}\left\{\left(\frac{\eta_1^\oplus + 1}{\eta_2^\oplus}\right)\boldsymbol{I}_{K_\ell}\right\} \end{bmatrix}\right) \end{bmatrix}\right\}.
$$

This leads to the natural parameter update

$$
\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2) \to \boldsymbol{\nu}} \leftarrow \begin{bmatrix} \mathbf{0} \\ -\frac{1}{2}\operatorname{vec}\left(\begin{bmatrix} \boldsymbol{F}^{-1} & \mathbf{0} \\ \mathbf{0} & \underset{1\leq\ell\leq d}{\operatorname{blockdiag}}\left\{\left(\frac{\eta_1^\oplus + 1}{\eta_2^\oplus}\right)\boldsymbol{I}_{K_\ell}\right\} \end{bmatrix}\right) \end{bmatrix}.
$$

Next we look at the factor to stochastic node message going from $p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)$ to $\boldsymbol{\sigma}^2$:

$$
\begin{aligned}
\boldsymbol{m}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2) \to \boldsymbol{\sigma}^2}(\boldsymbol{\sigma}^2) &\leftarrow \exp\left\{\int_{\mathbb{R}^v} \frac{1}{Z}\boldsymbol{m}_{\boldsymbol{\nu} \to p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)}(\boldsymbol{\nu}) \times \boldsymbol{m}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2) \to \boldsymbol{\nu}}(\boldsymbol{\nu})\right. \\
&\qquad \left.\times \log p(\boldsymbol{\nu}\,|\,\boldsymbol{\sigma}^2)\, d\boldsymbol{\nu}\right\} \\
&\propto \exp\left[\int_{\mathbb{R}^v} \left\{\begin{bmatrix} \boldsymbol{\nu} \\ \operatorname{vec}\left(\boldsymbol{\nu}\boldsymbol{\nu}^\top\right) \end{bmatrix}^\top \left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma}) \to \boldsymbol{\nu}} + \boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2) \to \boldsymbol{\nu}}\right)\right\}\right. \\
&\qquad \left.\times \left\{\sum_{\ell=1}^{d} \begin{bmatrix} \log\left(\sigma_\ell^2\right) \\ 1/\sigma_\ell^2 \end{bmatrix}^\top \begin{bmatrix} -\frac{1}{2}K_\ell \\ -\frac{1}{2}\|\boldsymbol{u}_\ell\|^2 \end{bmatrix}\right\} d\boldsymbol{\nu}\right].
\end{aligned}
$$

Since $\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{\nu}}$ is a vector that can be partitioned by each of its components, i.e.,

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{\nu}} \equiv \begin{bmatrix} \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{\beta}} \\ \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{u}_1} \\ \vdots \\ \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{u}_d} \end{bmatrix},$$

and similarly for $\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to\boldsymbol{\nu}}$, then we can write $\boldsymbol{m}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to\boldsymbol{\sigma}^2}(\boldsymbol{\sigma}^2)$ as

$$\exp\left\{\sum_{\ell=1}^{d}\begin{bmatrix}\log\left(\sigma_\ell^2\right)\\ 1/\sigma_\ell^2\end{bmatrix}^\top \times \begin{bmatrix} -\frac{1}{2}K_\ell \\ -\frac{1}{2}\int_{\mathbb{R}^{\boldsymbol{u}_\ell}} \|\boldsymbol{u}_\ell\|^2 p_{\mathrm{N}_{\mathrm{nat,vec}}}\left(\boldsymbol{u}_\ell\,;\,\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{u}_\ell}+\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to\boldsymbol{u}_\ell}\right)d\,\boldsymbol{u}_\ell \end{bmatrix}\right\}.$$

Using Primitive 7.4.3, this becomes

$$\exp\left\{\sum_{\ell=1}^{d}\begin{bmatrix}\log\left(\sigma_\ell^2\right)\\ 1/\sigma_\ell^2\end{bmatrix}^\top \begin{bmatrix} -\frac{1}{2}K_\ell \\ -\frac{1}{2}\mathcal{F}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{u}_\ell}+\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to\boldsymbol{u}_\ell}\right) \end{bmatrix}\right\}$$

and thus the natural parameter update is

$$\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to\sigma_\ell^2} \leftarrow \begin{bmatrix} -\frac{1}{2}K_\ell \\ -\frac{1}{2}\mathcal{F}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{u}_\ell}+\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to\boldsymbol{u}_\ell}\right) \end{bmatrix}.$$

The next message to update is

$$\begin{aligned}\boldsymbol{m}_{p(\boldsymbol{\sigma}^2|\boldsymbol{a})\to\boldsymbol{\sigma}^2}(\boldsymbol{\sigma}^2) &\leftarrow \exp\left\{\int_{\mathbb{R}^d_{\geq 0}}\frac{1}{Z}\boldsymbol{m}_{\boldsymbol{a}\to p(\boldsymbol{\sigma}^2|\boldsymbol{a})}(\boldsymbol{a})\times\boldsymbol{m}_{p(\boldsymbol{\sigma}^2|\boldsymbol{a})\to\boldsymbol{a}}(\boldsymbol{a})\right.\\ &\qquad\left.\times\log p(\boldsymbol{\sigma}^2|\boldsymbol{a})\,d\,\boldsymbol{a}\right\}\\ &\propto \exp\left[\int_{\mathbb{R}^d_{\geq 0}}\left\{\sum_{\ell=1}^{d}\begin{bmatrix}\log(a_\ell)\\ 1/a_\ell\end{bmatrix}^\top\left(\boldsymbol{\eta}_{p(a_\ell)\to a_\ell}+\boldsymbol{\eta}_{p(\sigma_\ell^2|a_\ell)\to a_\ell}\right)\right\}\right.\\ &\qquad\left.\times\left\{\sum_{\ell=1}^{d}\begin{bmatrix}\log\left(\sigma_\ell^2\right)\\ 1/\sigma_\ell^2\end{bmatrix}^\top\begin{bmatrix}-3/2\\ -1/a_\ell\end{bmatrix}\right\}d\,\boldsymbol{a}\right]\end{aligned}$$

$$= \exp\left\{\sum_{\ell=1}^{d}\begin{bmatrix}\log\left(\sigma_\ell^2\right)\\ 1/\sigma_\ell^2\end{bmatrix}^\top\begin{bmatrix}-3/2\\ -\int_0^\infty\frac{1}{a_\ell}\,p_{\mathrm{IG}_{\mathrm{nat}}}\left(a_\ell\,;\,\boldsymbol{\eta}_{p(a_\ell)\to a_\ell}+\boldsymbol{\eta}_{p(\sigma_\ell^2|a_\ell)\to a_\ell}\right)d\,a_\ell\end{bmatrix}\right\}.$$

The use of Primitive 7.4.6 gives

$$\boldsymbol{m}_{p(\boldsymbol{\sigma}^2|\boldsymbol{a})\to\boldsymbol{\sigma}^2}(\boldsymbol{\sigma}^2) \propto \exp\left\{\sum_{\ell=1}^{d}\begin{bmatrix}\log\left(\sigma_\ell^2\right)\\ 1/\sigma_\ell^2\end{bmatrix}^\top\begin{bmatrix}-3/2\\ -\frac{\left(\boldsymbol{\eta}_{p(a_\ell)\to a_\ell}+\boldsymbol{\eta}_{p(\sigma_\ell^2|a_\ell)\to a_\ell}\right)_1+1}{\left(\boldsymbol{\eta}_{p(a_\ell)\to a_\ell}+\boldsymbol{\eta}_{p(\sigma_\ell^2|a_\ell)\to a_\ell}\right)_2}\end{bmatrix}\right\}.$$

Therefore, the natural parameter update is given as

$$\boldsymbol{\eta}_{p(\sigma_\ell^2|a_\ell)\to\sigma_\ell^2} \leftarrow \begin{bmatrix} -3/2 \\ -\dfrac{\left(\boldsymbol{\eta}_{p(a_\ell)\to a_\ell} + \boldsymbol{\eta}_{p(\sigma_\ell^2|a_\ell)\to a_\ell}\right)_1 + 1}{\left(\boldsymbol{\eta}_{p(a_\ell)\to a_\ell} + \boldsymbol{\eta}_{p(\sigma_\ell^2|a_\ell)\to a_\ell}\right)_2} \end{bmatrix}.$$

The second last factor to stochastic node message to update is

$$\begin{aligned}\boldsymbol{m}_{p(\boldsymbol{\sigma}^2|\boldsymbol{a})\to\boldsymbol{a}}(\boldsymbol{a}) &\leftarrow \exp\left\{\int_{\mathbb{R}^d_{\geq 0}} \frac{1}{Z}\boldsymbol{m}_{\boldsymbol{\sigma}^2\to p(\boldsymbol{\sigma}^2|\boldsymbol{a})}(\boldsymbol{\sigma}^2)\times \boldsymbol{m}_{p(\boldsymbol{\sigma}^2|\boldsymbol{a})\to\boldsymbol{\sigma}^2}(\boldsymbol{\sigma}^2)\right. \\ &\qquad \left.\times \log p(\boldsymbol{\sigma}^2\,|\,\boldsymbol{a})\, d\boldsymbol{\sigma}^2\right\} \\ &\propto \exp\left[\int_{\mathbb{R}^d_{\geq 0}} \left\{\sum_{\ell=1}^d \begin{bmatrix} \log\left(\sigma_\ell^2\right) \\ 1/\sigma_\ell^2 \end{bmatrix}^\top \left(\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to\sigma_\ell^2} + \boldsymbol{\eta}_{p(\sigma_\ell^2|a_\ell)\to\sigma_\ell^2}\right)\right\}\right. \\ &\qquad \left.\times \left\{\sum_{\ell=1}^d \begin{bmatrix} \log\left(a_\ell\right) \\ 1/a_\ell \end{bmatrix}^\top \begin{bmatrix} -\frac{1}{2} \\ -1/\sigma_\ell^2 \end{bmatrix}\right\} d\,\sigma_\ell^2\right] \\ &= \exp\left\{\sum_{\ell=1}^d \begin{bmatrix} \log\left(a_\ell\right) \\ 1/a_\ell \end{bmatrix}^\top \begin{bmatrix} -\frac{1}{2} \\ -\displaystyle\int_0^\infty \frac{1}{\sigma_\ell^2}\, p_{\mathrm{IG}_{\mathrm{nat}}}\left(\sigma_\ell^2\,;\, \boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to\sigma_\ell^2} + \boldsymbol{\eta}_{p(\sigma_\ell^2|a_\ell)\to\sigma_\ell^2}\right) d\,\sigma_\ell^2 \end{bmatrix}\right\}\end{aligned}$$

Using Primitive 7.4.6, this becomes

$$\exp\left\{\sum_{\ell=1}^d \begin{bmatrix} \log\left(a_\ell\right) \\ 1/a_\ell \end{bmatrix}^\top \begin{bmatrix} -\frac{1}{2} \\ -\dfrac{\left(\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to\sigma_\ell^2} + \boldsymbol{\eta}_{p(\sigma_\ell^2|a_\ell)\to\sigma_\ell^2}\right)_1 + 1}{\left(\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to\sigma_\ell^2} + \boldsymbol{\eta}_{p(\sigma_\ell^2|a_\ell)\to\sigma_\ell^2}\right)_2} \end{bmatrix}\right\},$$

and so the natural parameter update is

$$\boldsymbol{\eta}_{p(\sigma_\ell^2|a_\ell)\to a_\ell} \leftarrow \begin{bmatrix} -\frac{1}{2} \\ -\dfrac{\left(\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to\sigma_\ell^2} + \boldsymbol{\eta}_{p(\sigma_\ell^2|a_\ell)\to\sigma_\ell^2}\right)_1 + 1}{\left(\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to\sigma_\ell^2} + \boldsymbol{\eta}_{p(\sigma_\ell^2|a_\ell)\to\sigma_\ell^2}\right)_2} \end{bmatrix}.$$

The last factor to stochastic node message to update is $\boldsymbol{m}_{p(\boldsymbol{a})\to\boldsymbol{a}}(\boldsymbol{a})$ and has the form

$$\begin{aligned}\boldsymbol{m}_{p(\boldsymbol{a})\to\boldsymbol{a}}(\boldsymbol{a}) &\leftarrow \exp\left\{\tfrac{1}{Z}\log p(\boldsymbol{a})\right\} \\ &\propto \exp\left\{\sum_{\ell=1}^d \begin{bmatrix} \log(a_\ell) \\ 1/a_\ell \end{bmatrix}^\top \begin{bmatrix} -3/2 \\ -1/A_\ell^2 \end{bmatrix}\right\}\end{aligned}$$

which means that the natural parameter update corresponding to this message is

$$\boldsymbol{\eta}_{p(a_\ell)\to a_\ell} \leftarrow \begin{bmatrix} -3/2 \\ -1/A_\ell^2 \end{bmatrix}$$

and remains fixed at this constant value since A_ℓ is a hyperparameter to be specified by the user.

7.A.4 Step 3 : Optimal q-densities

Once convergence is achieved from Steps 2(a) and (b) we get the optimal q-densities of each stochastic node through:

$$\boldsymbol{\eta}^*_{q(\boldsymbol{\nu})} \leftarrow \boldsymbol{\eta}^*_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to(\boldsymbol{\nu})} + \boldsymbol{\eta}^*_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to(\boldsymbol{\nu})}$$

$$\boldsymbol{\eta}^*_{q(\boldsymbol{\Sigma})} \leftarrow \boldsymbol{\eta}^*_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\Sigma})\to\boldsymbol{\Sigma}} + \boldsymbol{\eta}^*_{p(\boldsymbol{\Sigma})\to\boldsymbol{\Sigma}}$$

and for $1 \le \ell \le d$:

$$\boldsymbol{\eta}^*_{q(a_\ell)} \leftarrow \boldsymbol{\eta}^*_{p(a_\ell)\to a_\ell} + \boldsymbol{\eta}^*_{p(\sigma^2_\ell|a_\ell)\to a_\ell}$$

$$\boldsymbol{\eta}^*_{q(\sigma^2_\ell)} \leftarrow \boldsymbol{\eta}^*_{p(\sigma^2_\ell|a_\ell)\to\sigma^2_\ell} + \boldsymbol{\eta}^*_{p(\boldsymbol{\nu}|\boldsymbol{\sigma}^2)\to\sigma^2_\ell}.$$

7.B Derivation of Algorithm 12

This section gives the derivations for the updates given in Algorithm 12. Here, we also make use of the natural parameter forms, primitives, results and functions defined in Sections 7.3, 7.4 and 7.5.

7.B.1 Step 1: Initialise factor to stochastic node messages

We initialise the factor to stochastic node messages to be

$$\boldsymbol{m}_{\,p(\boldsymbol{a}^x)\to\boldsymbol{a}^x}(\boldsymbol{a}^x) \leftarrow \exp\left\{\sum_{k=1}^{K}\begin{bmatrix}\log\left(a_k^x\right)\\ 1/a_k^x\end{bmatrix}^{\top}\boldsymbol{\eta}_{p(a_k^x)\to a_k^x}\right\}$$

$$\boldsymbol{m}_{\,p\left\{(\boldsymbol{\sigma}^x)^2|\boldsymbol{a}^x\right\}\to\boldsymbol{a}^x}(\boldsymbol{a}^x) \leftarrow \exp\left\{\sum_{k=1}^{K}\begin{bmatrix}\log\left(a_k^x\right)\\ 1/a_k^x\end{bmatrix}^{\top}\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to a_k^x}\right\}$$

$$\boldsymbol{m}_{\,p\left\{(\boldsymbol{\sigma}^x)^2|\boldsymbol{a}^x\right\}\to(\boldsymbol{\sigma}^x)^2}\left\{(\boldsymbol{\sigma}^x)^2\right\} \leftarrow \exp\left\{\sum_{k=1}^{K}\begin{bmatrix}\log\left\{\left(\sigma_k^x\right)^2\right\}\\ 1/\left(\sigma_k^x\right)^2\end{bmatrix}^{\top}\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to(\sigma_k^x)^2}\right\}$$

$$\boldsymbol{m}_{\,p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to(\boldsymbol{\sigma}^x)^2}\left\{(\boldsymbol{\sigma}^x)^2\right\} \leftarrow \exp\left\{\sum_{k=1}^{K}\begin{bmatrix}\log\left\{\left(\sigma_k^x\right)^2\right\}\\ 1/\left(\sigma_k^x\right)^2\end{bmatrix}^{\top}\times\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to(\sigma_k^x)^2}\right\}$$

$$\boldsymbol{m}_{\,p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to\boldsymbol{a}}(\boldsymbol{a}) \leftarrow \exp\left\{\sum_{i=1}^{n}\begin{bmatrix}a_{i1}\\ \vdots\\ a_{iK}\end{bmatrix}^{\top}\begin{bmatrix}\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to a_{i1}}\\ \vdots\\ \eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to a_{iK}}\end{bmatrix}\right\}$$

$$\boldsymbol{m}_{\,p(\boldsymbol{a}\,|\,\boldsymbol{\omega})\to\boldsymbol{a}}(\boldsymbol{a}) \leftarrow \exp\left\{\sum_{i=1}^{n}\begin{bmatrix}a_{i1}\\ \vdots\\ a_{iK}\end{bmatrix}^{\top}\begin{bmatrix}\eta_{p(a_{i1}|\omega_1)\to a_{i1}}\\ \vdots\\ \eta_{p(a_{iK}|\omega_K)\to a_{iK}}\end{bmatrix}\right\}$$

$$\boldsymbol{m}_{\,p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to\boldsymbol{x}_{\text{unobs}}}(\boldsymbol{x}_{\text{unobs}}) \leftarrow \exp\left\{\sum_{i=1}^{n_{\text{unobs}}}\begin{bmatrix}x_{\text{unobs,i}}\\ x_{\text{unobs,i}}^2\end{bmatrix}^{\top}\times\boldsymbol{\eta}_{p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to x_{\text{unobs,i}}}\right\}$$

$$\boldsymbol{m}_{\,p(\boldsymbol{o}\,|\,\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\boldsymbol{x}_{\text{unobs}}}(\boldsymbol{x}_{\text{unobs}}) \leftarrow \exp\left\{\sum_{i=1}^{n_{\text{unobs}}}\begin{bmatrix}x_{\text{unobs,i}}\\ x_{\text{unobs,i}}^2\end{bmatrix}^{\top}\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to x_{\text{unobs,i}}}\right\}$$

$$\boldsymbol{m}_{\,p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to\boldsymbol{\mu}^x}\left(\boldsymbol{\mu}^x\right)\leftarrow\exp\left\{\sum_{k=1}^{K}\begin{bmatrix}\mu_k\\ \mu_k^2\end{bmatrix}^{\top}\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to\mu_k^x}\right\}$$

$$\boldsymbol{m}_{\,p(\boldsymbol{\mu}^x)\to\boldsymbol{\mu}^x}\left(\boldsymbol{\mu}^x\right)\leftarrow\exp\left\{\sum_{k=1}^{K}\begin{bmatrix}\mu_k\\ \mu_k^2\end{bmatrix}^{\top}\boldsymbol{\eta}_{p(\mu_k^x)\to\mu_k^x}\right\}$$

$$\boldsymbol{m}_{\,p(\boldsymbol{a}\,|\,\boldsymbol{\omega})\to\boldsymbol{\omega}}\left(\boldsymbol{\omega}\right)\leftarrow\exp\left\{\sum_{k=1}^{K}\begin{bmatrix}\log\left(\omega_1\right)\\ \vdots\\ \log\left(\omega_K\right)\end{bmatrix}^{\top}\begin{bmatrix}\eta_{p(\boldsymbol{a}_1|\omega_1)\to\omega_1}\\ \vdots\\ \eta_{p(\boldsymbol{a}_K|\omega_K)\to\omega_K}\end{bmatrix}\right\}$$

$$\boldsymbol{m}_{\,p(\boldsymbol{\omega})\to\boldsymbol{\omega}}\left(\boldsymbol{\omega}\right)\leftarrow\exp\left\{\sum_{k=1}^{K}\begin{bmatrix}\log\left(\omega_1\right)\\ \vdots\\ \log\left(\omega_K\right)\end{bmatrix}^{\top}\begin{bmatrix}\eta_{p(\omega_1)\to\omega_1}\\ \vdots\\ \eta_{p(\omega_K)\to\omega_K}\end{bmatrix}\right\}$$

$$\boldsymbol{m}_{\,p(\boldsymbol{o}\,|\,\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\boldsymbol{\alpha}}\left(\boldsymbol{\alpha}\right)\leftarrow\exp\left\{\begin{bmatrix}\boldsymbol{\alpha}\\ \operatorname{vec}\left(\boldsymbol{\alpha}\boldsymbol{\alpha}^{\top}\right)\end{bmatrix}^{\top}\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\boldsymbol{\alpha}}\right\}$$

$$\boldsymbol{m}_{\,p(\boldsymbol{\alpha})\to\boldsymbol{\alpha}}\left(\boldsymbol{\alpha}\right)\leftarrow\exp\left\{\begin{bmatrix}\boldsymbol{\alpha}\\ \operatorname{vec}\left(\boldsymbol{\alpha}\boldsymbol{\alpha}^{\top}\right)\end{bmatrix}^{\top}\boldsymbol{\eta}_{p(\boldsymbol{\alpha})\to\boldsymbol{\alpha}}\right\}$$

$$\boldsymbol{m}_{\,p(\boldsymbol{o}\,|\,\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\sigma_o^2}\left(\sigma_o^2\right)\leftarrow\exp\left\{\begin{bmatrix}\log\left(\sigma_o^2\right)\\ 1/\sigma_o^2\end{bmatrix}^{\top}\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\sigma_o^2}\right\}$$

$$\boldsymbol{m}_{\,p(\sigma_o^2|a_o)\to\sigma_o^2}\left(\sigma_o^2\right)\leftarrow\exp\left\{\begin{bmatrix}\log\left(\sigma_o^2\right)\\ 1/\sigma_o^2\end{bmatrix}^{\top}\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to\sigma_o^2}\right\}$$

$$\boldsymbol{m}_{\,p(\sigma_o^2|a_o)\to a_o}\left(a_o\right)\leftarrow\exp\left\{\begin{bmatrix}\log\left(a_o\right)\\ 1/a_o\end{bmatrix}^{\top}\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to a_o}\right\}$$

$$\boldsymbol{m}_{\,p(a_o)\to a_o}\left(a_o\right)\leftarrow\exp\left\{\begin{bmatrix}\log\left(a_o\right)\\ 1/a_o\end{bmatrix}^{\top}\boldsymbol{\eta}_{p(a_o)\to a_o}\right\}$$

$$\boldsymbol{m}_{\,p(\boldsymbol{\beta})\to\boldsymbol{\beta}}\left(\boldsymbol{\beta}\right)\leftarrow\exp\left\{\begin{bmatrix}\boldsymbol{\beta}\\ \operatorname{vec}\left(\boldsymbol{\beta}\boldsymbol{\beta}^{\top}\right)\end{bmatrix}^{\top}\boldsymbol{\eta}_{p(\boldsymbol{\beta})\to\boldsymbol{\beta}}\right\}$$

$$\boldsymbol{m}_{\,p(\boldsymbol{y}\,|\,\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\boldsymbol{x}_{\text{unobs}}}\left(\boldsymbol{x}_{\text{unobs}}\right)\leftarrow\exp\left\{\sum_{i=1}^{n_{\text{unobs}}}\begin{bmatrix}x_{\text{unobs,i}}\\ x_{\text{unobs,i}}^2\end{bmatrix}^{\top}\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to x_{\text{unobs,i}}}\right\}$$

$$\boldsymbol{m}_{\,p(\boldsymbol{y}\,|\,\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\,\to\,\boldsymbol{\beta}}\left(\boldsymbol{\beta}\right) \leftarrow \exp\left\{\begin{bmatrix}\boldsymbol{\beta}\\ \mathrm{vec}\left(\boldsymbol{\beta}\boldsymbol{\beta}^\top\right)\end{bmatrix}^\top \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\,\to\,\boldsymbol{\beta}}\right\}$$

$$\boldsymbol{m}_{\,p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\,\to\,\sigma_\varepsilon^2}\left(\sigma_\varepsilon^2\right) \leftarrow \exp\left\{\begin{bmatrix}\log\left(\sigma_\varepsilon^2\right)\\ 1/\sigma_\varepsilon^2\end{bmatrix}^\top \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\,\to\,\sigma_\varepsilon^2}\right\}$$

$$\boldsymbol{m}_{\,p(\sigma_\varepsilon^2|a_\varepsilon)\,\to\,\sigma_\varepsilon^2}\left(\sigma_\varepsilon^2\right) \leftarrow \exp\left\{\begin{bmatrix}\log\left(\sigma_\varepsilon^2\right)\\ 1/\sigma_\varepsilon^2\end{bmatrix}^\top \boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\,\to\,\sigma_\varepsilon^2}\right\}$$

$$\boldsymbol{m}_{\,p(\sigma_\varepsilon^2|a_\varepsilon)\,\to\,a_\varepsilon}\left(a_\varepsilon\right) \leftarrow \exp\left\{\begin{bmatrix}\log\left(a_\varepsilon\right)\\ 1/a_\varepsilon\end{bmatrix}^\top \boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\,\to\,a_\varepsilon}\right\}$$

$$\boldsymbol{m}_{\,p(a_\varepsilon)\,\to\,a_\varepsilon}\left(a_\varepsilon\right) \leftarrow \exp\left\{\begin{bmatrix}\log\left(a_\varepsilon\right)\\ 1/a_\varepsilon\end{bmatrix}^\top \boldsymbol{\eta}_{p(a_\varepsilon)\,\to\,a_\varepsilon}\right\}.$$

The initial values for the corresponding natural parameter vectors are set as:

$$\boldsymbol{\eta}_{p(a_k^x)\,\to\,a_k^x} \leftarrow \begin{bmatrix}-2\\ -1\end{bmatrix};\quad \boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\,\to\,a_k^x} \leftarrow \begin{bmatrix}-2\\ -1\end{bmatrix};\quad \boldsymbol{\eta}_{p(\sigma_o^2|a_o)\,\to\,\sigma_o^2} \leftarrow \begin{bmatrix}-2\\ -1\end{bmatrix}$$

$$\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\,\to\,(\sigma_k^x)^2} \leftarrow \begin{bmatrix}-2\\ -1\end{bmatrix};\quad \boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\,\to\,(\sigma_k^x)^2} \leftarrow \begin{bmatrix}-2\\ -1\end{bmatrix};$$

$$\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\,\to\,a_{ik}} \leftarrow \log\left(\tfrac{1}{K}\right);\quad \eta_{p(a_{ik}|\omega_k)\,\to\,a_{ik}} \leftarrow \log\left(\tfrac{1}{K}\right);$$

$$\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\,\to\,\mu_k^x} \leftarrow \begin{bmatrix}0\\ -\frac{1}{2}\end{bmatrix};\quad \boldsymbol{\eta}_{p(\mu_k^x)\,\to\,\mu_k^x} \leftarrow \begin{bmatrix}0\\ -\frac{1}{2}\end{bmatrix};\quad \eta_{p(\boldsymbol{a}_k|\omega_k)\,\to\,\omega_k} \leftarrow 1;$$

$$\eta_{p(\omega_k)\,\to\,\omega_k} \leftarrow \alpha_k - 1;\quad \boldsymbol{\eta}_{p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\,\to\,x_{\mathrm{unobs,i}}} \leftarrow \begin{bmatrix}0\\ -\frac{1}{2}\end{bmatrix};$$

$$\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\to\,x_{\mathrm{unobs,i}}} \leftarrow \begin{bmatrix}0\\ -\frac{1}{2}\end{bmatrix};\quad \boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\to\,\boldsymbol{\alpha}} \leftarrow \begin{bmatrix}\mathbf{0}_2\\ -\frac{1}{2}\mathrm{vec}\left(\boldsymbol{I}_2\right)\end{bmatrix};$$

$$\boldsymbol{\eta}_{p(\boldsymbol{\alpha})\,\to\,\boldsymbol{\alpha}} \leftarrow \begin{bmatrix}\mathbf{0}_2\\ -\frac{1}{2}\mathrm{vec}\left(\boldsymbol{I}_2\right)\end{bmatrix};\quad \boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\to\,\sigma_o^2} \leftarrow \begin{bmatrix}-2\\ -1\end{bmatrix};$$

$$\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\,\to\,a_o} \leftarrow \begin{bmatrix}-2\\ -1\end{bmatrix};\quad \boldsymbol{\eta}_{p(a_o)\,\to\,a_o} \leftarrow \begin{bmatrix}-2\\ -1\end{bmatrix};\quad \boldsymbol{\eta}_{p(\boldsymbol{\beta})\,\to\,\boldsymbol{\beta}} \leftarrow \begin{bmatrix}\mathbf{0}_3\\ -\frac{1}{2}\mathrm{vec}\left(\boldsymbol{I}_3\right)\end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to x_{\text{unobs,i}}} \leftarrow \begin{bmatrix} 0 \\ -\frac{1}{2} \end{bmatrix}; \quad \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\sigma_\varepsilon^2} \leftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to\sigma_\varepsilon^2} \leftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix}; \quad \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\boldsymbol{\beta}} \leftarrow \begin{bmatrix} \mathbf{0}_3 \\ -\frac{1}{2}\text{vec}\left(\boldsymbol{I}_3\right) \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to a_\varepsilon} \leftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix}; \quad \boldsymbol{\eta}_{p(a_\varepsilon)\to a_\varepsilon} \leftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix}$$

7.B.2 Step 2 (a): Update stochastic node to factor messages

The stochastic node to factor messages have the following updates:

$$\boldsymbol{m}_{\boldsymbol{a}^x\to p(\boldsymbol{a}^x)}(\boldsymbol{a}^x) \leftarrow \boldsymbol{m}_{p\left\{(\boldsymbol{\sigma}^x)^2|\boldsymbol{a}^x\right\}\to\boldsymbol{a}^x}(\boldsymbol{a}^x)$$

$$\boldsymbol{m}_{(\boldsymbol{\sigma}^x)^2\to p\left\{(\boldsymbol{\sigma}^x)^2|\boldsymbol{a}^x\right\}} \leftarrow \boldsymbol{m}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to(\boldsymbol{\sigma}^x)^2}$$

$$\boldsymbol{m}_{(\boldsymbol{\sigma}^x)^2\to p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}} \leftarrow \boldsymbol{m}_{p\left\{(\boldsymbol{\sigma}^x)^2|\boldsymbol{a}^x\right\}\to(\boldsymbol{\sigma}^x)^2}$$

$$\boldsymbol{m}_{\boldsymbol{a}\to p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,\left(\boldsymbol{\sigma}^2\right)^x\right\}} \leftarrow \boldsymbol{m}_{p(\boldsymbol{a})\to\boldsymbol{a}}; \quad \boldsymbol{m}_{\boldsymbol{a}\to p(\boldsymbol{a})} \leftarrow \boldsymbol{m}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to\boldsymbol{a}}$$

$$\boldsymbol{m}_{\boldsymbol{\omega}\to p(\boldsymbol{a}|\boldsymbol{\omega})} \leftarrow \boldsymbol{m}_{p(\boldsymbol{\omega})\to\boldsymbol{\omega}}; \quad \boldsymbol{m}_{\boldsymbol{\omega}\to p(\boldsymbol{\omega})} \leftarrow \boldsymbol{m}_{p(\boldsymbol{a}|\boldsymbol{\omega})\to\boldsymbol{\omega}}$$

$$\boldsymbol{m}_{\boldsymbol{\mu}^x\to p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}} \leftarrow \boldsymbol{m}_{p(\boldsymbol{\mu}^x)\to\boldsymbol{\mu}^x}$$

$$\boldsymbol{m}_{\boldsymbol{\mu}^x\to p(\boldsymbol{\mu}^x)} \leftarrow \boldsymbol{m}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to\boldsymbol{\mu}^x}$$

$$\boldsymbol{m}_{\boldsymbol{\alpha}\to p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)} \leftarrow \boldsymbol{m}_{p(\boldsymbol{\alpha})\to\boldsymbol{\alpha}}; \quad \boldsymbol{m}_{\boldsymbol{\alpha}\to p(\boldsymbol{\alpha})} \leftarrow \boldsymbol{m}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\boldsymbol{\alpha}}$$

$$\boldsymbol{m}_{\sigma_o^2\to p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)} \leftarrow \boldsymbol{m}_{p(\sigma_o^2|a_o)\to\sigma_o^2}; \quad \boldsymbol{m}_{\sigma_o^2\to p(\sigma_o^2|a_o)} \leftarrow \boldsymbol{m}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\sigma_o^2}$$

$$\boldsymbol{m}_{a_o\to p(\sigma_o^2|a_o)} \leftarrow \boldsymbol{m}_{p(a_o)\to a_o}; \quad \boldsymbol{m}_{a_o\to p(a_o)} \leftarrow \boldsymbol{m}_{p(\sigma_o^2|a_o)\to a_o}$$

$$\boldsymbol{m}_{\boldsymbol{x}_{\text{unobs}}\to p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}} \leftarrow \boldsymbol{m}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\boldsymbol{x}_{\text{unobs}}} + \boldsymbol{m}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\boldsymbol{x}_{\text{unobs}}}$$

$$\boldsymbol{m}_{\boldsymbol{x}_{\text{unobs}}\to p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)} \leftarrow \boldsymbol{m}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to\boldsymbol{x}_{\text{unobs}}} + \boldsymbol{m}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\boldsymbol{x}_{\text{unobs}}}$$

$$\boldsymbol{m}_{\boldsymbol{x}_{\text{unobs}}\to p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)} \leftarrow \boldsymbol{m}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to\boldsymbol{x}_{\text{unobs}}} + \boldsymbol{m}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\boldsymbol{x}_{\text{unobs}}}$$

$$\boldsymbol{m}_{\boldsymbol{\beta}\to p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)} \leftarrow \boldsymbol{m}_{p(\boldsymbol{\beta})\to\boldsymbol{\beta}}; \quad \boldsymbol{m}_{\boldsymbol{\beta}\to p(\boldsymbol{\beta})} \leftarrow \boldsymbol{m}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\boldsymbol{\beta}}$$

$$\boldsymbol{m}_{\sigma_\varepsilon^2\to p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)} \leftarrow \boldsymbol{m}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to\sigma_\varepsilon^2}; \quad \boldsymbol{m}_{\sigma_\varepsilon^2\to p(\sigma_\varepsilon^2|a_\varepsilon)} \leftarrow \boldsymbol{m}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\sigma_\varepsilon^2}$$

$$\boldsymbol{m}_{a_\varepsilon\to p(\sigma_\varepsilon^2|a_\varepsilon)} \leftarrow \boldsymbol{m}_{p(a_\varepsilon)\to a_\varepsilon}; \quad \boldsymbol{m}_{a_\varepsilon\to p(a_\varepsilon)} \leftarrow \boldsymbol{m}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to a_\varepsilon}$$

and this is equivalent to the following natural parameter updates:

$$\boldsymbol{\eta}_{a_k^x\to p(a_k^x)} \leftarrow \boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to a_k^x}; \quad \boldsymbol{\eta}_{a_k^x\to p\left\{(\sigma_k^x)^2|a_k^x\right\}} \leftarrow \boldsymbol{\eta}_{p(a_k^x)\to a_k^x}$$

$$\boldsymbol{\eta}_{(\sigma_k^x)^2\to p\left\{(\sigma_k^x)^2|a_k^x\right\}} \leftarrow \boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to(\sigma_k^x)^2}$$

$$\boldsymbol{\eta}_{(\sigma_k^x)^2\to p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma_k^x)^2\right\}} \leftarrow \boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to(\sigma_k^x)^2}$$

$$\boldsymbol{\eta}_{a_{ik}\to p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma_k^x)^2\right\}} \leftarrow \eta_{p(a_{ik}|\omega_k)\to a_{ik}}; \quad \eta_{a_{ik}\to p(a_{ik}|\omega_k)} \leftarrow \eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to a_{ik}}$$

$$\eta_{\omega_k\to p(a_{ik}|\omega_k)} \leftarrow \eta_{p(\omega_k)\to\omega_k}; \quad \eta_{\omega_k\to p(\omega_k)} \leftarrow \eta_{p(\boldsymbol{a}_k|\omega_k)\to\omega_k}$$

$$\boldsymbol{\eta}_{\mu_k^x \to p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma_k^x)^2\right\}} \leftarrow \boldsymbol{\eta}_{p(\mu_k^x)\to\mu_k^x}; \quad \boldsymbol{\eta}_{\mu_k^x \to p(\mu_k^x)} \leftarrow \boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to\mu_k^x}$$
$$\boldsymbol{\eta}_{\boldsymbol{\alpha}\to p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{\alpha})\to\boldsymbol{\alpha}}; \quad \boldsymbol{\eta}_{\boldsymbol{\alpha}\to p(\boldsymbol{\alpha})} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\boldsymbol{\alpha}}$$
$$\boldsymbol{\eta}_{\sigma_o^2\to p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)} \leftarrow \boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to\sigma_o^2}; \quad \boldsymbol{\eta}_{\sigma_o^2\to p(\sigma_o^2|a_o)} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\sigma_o^2}$$
$$\boldsymbol{\eta}_{a_o\to p(\sigma_o^2|a_o)} \leftarrow \boldsymbol{\eta}_{p(a_o)\to a_o}; \quad \boldsymbol{\eta}_{a_o\to p(a_o)} \leftarrow \boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to a_o}$$
$$\boldsymbol{\eta}_{x_{\text{unobs,i}}\to p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to x_{\text{unobs,i}}} + \boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to x_{\text{unobs,i}}}$$
$$\boldsymbol{\eta}_{x_{\text{unobs,i}}\to p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)} \leftarrow \boldsymbol{\eta}_{p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to x_{\text{unobs,i}}} + \boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to x_{\text{unobs,i}}}$$
$$\boldsymbol{\eta}_{x_{\text{unobs,i}}\to p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)} \leftarrow \boldsymbol{\eta}_{p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to x_{\text{unobs,i}}} + \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to x_{\text{unobs,i}}}$$
$$\boldsymbol{\eta}_{\boldsymbol{\beta}\to p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{\beta})\to\boldsymbol{\beta}}; \quad \boldsymbol{\eta}_{\boldsymbol{\beta}\to p(\boldsymbol{\beta})} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\boldsymbol{\beta}}$$
$$\boldsymbol{\eta}_{\sigma_\varepsilon^2\to p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)} \leftarrow \boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to\sigma_\varepsilon^2}; \quad \boldsymbol{\eta}_{\sigma_\varepsilon^2\to p(\sigma_\varepsilon^2|a_\varepsilon)} \leftarrow \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\sigma_\varepsilon^2}$$
$$\boldsymbol{\eta}_{a_\varepsilon\to p(\sigma_\varepsilon^2|a_\varepsilon)} \leftarrow \boldsymbol{\eta}_{p(a_\varepsilon)\to a_\varepsilon}; \quad \boldsymbol{\eta}_{a_\varepsilon\to p(a_\varepsilon)} \leftarrow \boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to a_\varepsilon}$$

7.B.3 Step 2 (b): Update factor to stochastic node messages

We begin with the update for the factor to stochastic node message $\boldsymbol{m}_{\,p(\boldsymbol{a}^x)\to\boldsymbol{a}^x}(\boldsymbol{a}^x)$. Since the only neighbour of factor $p(\boldsymbol{a}^x)$ is the stochastic node $\boldsymbol{a}^x$ this message takes on the form

$$\begin{aligned}\boldsymbol{m}_{\,p(\boldsymbol{a}^x)\to\boldsymbol{a}^x}(\boldsymbol{a}^x) &\leftarrow \exp\left\{\tfrac{1}{Z}\log p(\boldsymbol{a}^x)\right\}\\ &\propto \exp\left\{\sum_{k=1}^{K}\begin{bmatrix}\log\left(a_k^x\right)\\ 1/a_k^x\end{bmatrix}^{\top}\begin{bmatrix}-3/2\\ -1/\left(A_k^x\right)^2\end{bmatrix}\right\}.\end{aligned}$$

This means that the update for the natural parameter vector is

$$\boldsymbol{\eta}_{p(a_k^x)\to a_k^x} \leftarrow \begin{bmatrix}-3/2\\ -1/\left(A_k^x\right)^2\end{bmatrix}$$

and remains constant at this value since A_k^x is a hyperparameter to be specified by the user. Next we consider the message going from factor $p\left\{(\boldsymbol{\sigma}^x)^2|\boldsymbol{a}^x\right\}$ to stochastic node $\boldsymbol{a}^x$:

$$\begin{aligned}\boldsymbol{m}_{\,p\left\{(\boldsymbol{\sigma}^x)^2|\boldsymbol{a}^x\right\}\to\boldsymbol{a}^x}(\boldsymbol{a}^x) \leftarrow \exp\Bigg[\int_{\mathbb{R}^K_{\geq 0}}&\frac{1}{Z}\boldsymbol{m}_{\,(\boldsymbol{\sigma}^x)^2\to p\left\{(\boldsymbol{\sigma}^x)^2|\boldsymbol{a}^x\right\}}\left\{(\boldsymbol{\sigma}^x)^2\right\}\\ &\times\boldsymbol{m}_{\,p\left\{(\boldsymbol{\sigma}^x)^2|\boldsymbol{a}^x\right\}\to(\boldsymbol{\sigma}^x)^2}\left\{(\boldsymbol{\sigma}^x)^2\right\}\\ &\times\log p\left\{(\boldsymbol{\sigma}^x)^2|\boldsymbol{a}^x\right\}d\left(\boldsymbol{\sigma}^2\right)^2\Bigg]\end{aligned}$$

$$\propto \exp\left[\int_{\mathbb{R}^K_{\geq 0}} \left\{\sum_{k=1}^K \begin{bmatrix} \log\left\{\left(\sigma_k^x\right)^2\right\} \\ 1/\left(\sigma_k^x\right)^2 \end{bmatrix}^\top \left(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to(\sigma_k^x)^2} + \boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to(\sigma_k^x)^2}\right)\right\}\right.$$
$$\left.\times\left\{\sum_{k=1}^K \begin{bmatrix} \log\left\{a_k^x\right\} \\ 1/a_k^x \end{bmatrix}^\top \begin{bmatrix} -\frac{1}{2} \\ -\frac{1}{\left(\sigma_k^x\right)^2} \end{bmatrix}\right\} d\left(\boldsymbol{\sigma}^x\right)^2\right]$$

$$\propto \exp\left\{\sum_{k=1}^K \begin{bmatrix} \log\left\{a_k^x\right\} \\ 1/a_k^x \end{bmatrix}^\top \begin{bmatrix} -\frac{1}{2} \\ -\int_0^\infty \frac{1}{\left(\sigma_k^x\right)^2}\, p_{\mathrm{IG}_{\mathrm{nat}}}\left((\sigma_k^x)^2\ ;\ \boldsymbol{\eta}^{\boxplus}\right) d\left(\sigma_k^x\right)^2 \end{bmatrix}\right\}$$

where $\boldsymbol{\eta}^{\boxplus} \equiv \boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to(\sigma_k^x)^2} + \boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to(\sigma_k^x)^2} \equiv \left[\eta_1^{\boxplus}\ \eta_2^{\boxplus}\right]^\top$. Using Primitive 7.4.6, we get

$$\boldsymbol{m}_{\,p\left\{(\boldsymbol{\sigma}^x)^2|\boldsymbol{a}^x\right\}\to\boldsymbol{a}^x}\left(\boldsymbol{a}^x\right) \leftarrow \exp\left\{\sum_{k=1}^K \begin{bmatrix} \log\left\{a_k^x\right\} \\ 1/a_k^x \end{bmatrix}^\top \begin{bmatrix} -\frac{1}{2} \\ -\frac{\eta_1^{\boxplus}+1}{\eta_2^{\boxplus}} \end{bmatrix}\right\}$$

and thus, the natural parameter update is

$$\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to a_k^x} \leftarrow \begin{bmatrix} -\frac{1}{2} \\ -\frac{\eta_1^{\boxplus}+1}{\eta_2^{\boxplus}} \end{bmatrix}.$$

The next message is expressed as

$$\boldsymbol{m}_{\,p\left\{(\boldsymbol{\sigma}^x)^2|\boldsymbol{a}^x\right\}\to(\boldsymbol{\sigma}^x)^2}\left\{(\boldsymbol{\sigma}^x)^2\right\} \leftarrow \exp\left[\int_{\mathbb{R}^K_{\geq 0}} \frac{1}{Z}\boldsymbol{m}_{\,\boldsymbol{a}^x\to p\left\{(\boldsymbol{\sigma}^x)^2|\boldsymbol{a}^x\right\}}\left(\boldsymbol{a}^x\right)\right.$$
$$\times\boldsymbol{m}_{\,p\left\{(\boldsymbol{\sigma}^x)^2|\boldsymbol{a}^x\right\}\to\boldsymbol{a}^x}\left(\boldsymbol{a}^x\right)$$
$$\left.\times\log p\left\{(\boldsymbol{\sigma}^x)^2|\boldsymbol{a}^x\right\} d\,\boldsymbol{a}^x\right]$$

$$\propto \exp\left[\int_{\mathbb{R}^K_{\geq 0}} \left\{\sum_{k=1}^K \begin{bmatrix} \log\left\{a_k^x\right\} \\ 1/a_k^x \end{bmatrix}^\top \left(\boldsymbol{\eta}_{p(a_k^x)\to a_k^x} + \boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to a_k^x}\right)\right\}\right.$$
$$\left.\times\left\{\sum_{k=1}^K \begin{bmatrix} \log\left\{\left(\sigma_k^x\right)^2\right\} \\ 1/\left(\sigma_k^x\right)^2 \end{bmatrix}^\top \begin{bmatrix} -3/2 \\ -\frac{1}{a_k^x} \end{bmatrix}\right\} d\,\boldsymbol{a}^x\right]$$

$$\propto \exp\left\{\sum_{k=1}^K \begin{bmatrix} \log\left\{\left(\sigma_k^x\right)^2\right\} \\ 1/\left(\sigma_k^x\right)^2 \end{bmatrix}^\top \begin{bmatrix} -3/2 \\ -\int_0^\infty \frac{1}{a_k^x}\, p_{\mathrm{IG}_{\mathrm{nat}}}\left(a_k^x\ ;\ \boldsymbol{\eta}_{p(a_k^x)\to a_k^x}\right. \\ \left. +\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to a_k^x}\right) d\,a_k^x \end{bmatrix}\right\}.$$

Using Primitive 7.4.6, this gives

$$\begin{aligned}\boldsymbol{m}_{\,p\left\{(\boldsymbol{\sigma}^x)^2|\boldsymbol{a}^x\right\}\to(\boldsymbol{\sigma}^x)^2}\left\{(\boldsymbol{\sigma}^x)^2\right\}&\\ \leftarrow\exp\left\{\sum_{k=1}^{K}\begin{bmatrix}\log\left\{\left(\sigma_k^x\right)^2\right\}\\ 1/\left(\sigma_k^x\right)^2\end{bmatrix}^{\top}\begin{bmatrix}-3/2\\ -\dfrac{\left(\boldsymbol{\eta}_{p(a_k^x)\to a_k^x}+\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to a_k^x}\right)_1+1}{\left(\boldsymbol{\eta}_{p(a_k^x)\to a_k^x}+\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to a_k^x}\right)_2}\end{bmatrix}\right\}&\end{aligned}$$

and so, the natural parameter update is

$$\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to(\sigma_k^x)^2}\leftarrow\begin{bmatrix}-3/2\\ -\dfrac{\left(\boldsymbol{\eta}_{p(a_k^x)\to a_k^x}+\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to a_k^x}\right)_1+1}{\left(\boldsymbol{\eta}_{p(a_k^x)\to a_k^x}+\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to a_k^x}\right)_2}\end{bmatrix}.$$

Next,

$$\begin{aligned}\boldsymbol{m}_{\,p(\boldsymbol{\mu}^x)\to\boldsymbol{\mu}^x}(\boldsymbol{\mu}^x)&\leftarrow\exp\left\{\tfrac{1}{Z}\log p(\boldsymbol{\mu}^x)\right\}\\ &\propto\exp\left\{\sum_{k=1}^{K}\begin{bmatrix}\mu_k^x\\ \left(\mu_k^x\right)^2\end{bmatrix}^{\top}\begin{bmatrix}\mu_\mu\ \sigma_\mu^2\\ -1/2\sigma_\mu^2\end{bmatrix}\right\}.\end{aligned}$$

So, the constant value for the natural parameter vector is

$$\boldsymbol{\eta}_{p(\mu_k^x)\to\mu_k^x}\leftarrow\begin{bmatrix}\mu_\mu\ \sigma_\mu^2\\ -1/2\sigma_\mu^2\end{bmatrix}.$$

Moving along, we next consider the message going from factor $p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}$ to the stochastic node $\boldsymbol{x}_{\text{unobs}}$:

$$\begin{aligned}\boldsymbol{m}_{\,p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to\boldsymbol{x}_{\text{unobs}}}(\boldsymbol{x}_{\text{unobs}})&\\ \leftarrow\exp\Bigg[\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{a_{ik}=0}^{1}\frac{1}{Z}\int_{\mathbb{R}^K}\int_{\mathbb{R}^K}&\boldsymbol{m}_{\,\boldsymbol{a}\to p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,\left(\boldsymbol{\sigma}^2\right)^x\right\}}(\boldsymbol{a})\\ &\times\boldsymbol{m}_{\,p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to\boldsymbol{a}}(\boldsymbol{a})\times\boldsymbol{m}_{\,\boldsymbol{\mu}^x\to p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}}(\boldsymbol{\mu}^x)\\ &\times\boldsymbol{m}_{\,p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to\boldsymbol{\mu}^x}(\boldsymbol{\mu}^x)\times\boldsymbol{m}_{\,(\boldsymbol{\sigma}^x)^2\to p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}}\left\{(\boldsymbol{\sigma}^x)^2\right\}\\ &\times\boldsymbol{m}_{\,p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to(\boldsymbol{\sigma}^x)^2}\left\{(\boldsymbol{\sigma}^x)^2\right\}\times\log p\left\{\boldsymbol{a}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\\ &d\,\boldsymbol{a}\,d\,\boldsymbol{\mu}^x\,d\,(\boldsymbol{\sigma}^x)^2\Bigg]\end{aligned}$$

$$
\begin{aligned}
&\propto \exp\left[\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{a_{ik}=0}^{1}\int_{\mathbb{R}^K}\int_{\mathbb{R}^K}\left\{a_{ik}\left(\eta_{p(a_{ik}|\omega_k)\to a_{ik}}+\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to a_{ik}}\right)\right\}\right.\\
&\quad\times\left\{\begin{bmatrix}\mu_k^x\\ \left(\mu_k^x\right)^2\end{bmatrix}^{\top}\left(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to\mu_k^x}+\boldsymbol{\eta}_{p(\mu_k^x)\to\mu_k^x}\right)\right\}\\
&\quad\times\left\{\begin{bmatrix}\log\left\{\left(\sigma_k^x\right)^2\right\}\\ 1/\left(\sigma_k^x\right)^2\end{bmatrix}^{\top}\left(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to(\sigma_k^x)^2}+\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to(\sigma_k^x)^2}\right)\right\}\\
&\quad\left.\times\begin{bmatrix}x_{\text{unobs,i}}\\ x_{\text{unobs,i}}^2\end{bmatrix}^{\top}\begin{bmatrix}a_{ik}\mu_k^x/\left(\sigma_k^x\right)^2\\ -a_{ik}/2\left(\sigma_k^x\right)^2\end{bmatrix}d\,\boldsymbol{a}\,d\,\boldsymbol{\mu}^x\,d\,(\boldsymbol{\sigma}^x)^2\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \exp\left\{\sum_{i=1}^{n_{\text{unobs}}}\sum_{k=1}^{K}\begin{bmatrix}x_{\text{unobs,i}}\\ x_{\text{unobs,i}}^2\end{bmatrix}^{\top}\right.\\
&\quad\times\left[\begin{array}{l}
\sum_{a_{ik}=0}^{1}a_{ik}\times a_{ik}\times\left(\eta_{p(a_{ik}|\omega_k)\to a_{ik}}+\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to a_{ik}}\right)\\
\quad\times\int_{\mathbb{R}}\mu_k^x\,p_{\mathrm{N}_{\text{nat}}}\left(\mu_k^x\,;\,\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to\mu_k^x}\right.\\
\quad\left.+\boldsymbol{\eta}_{p(\mu_k^x)\to\mu_k^x}\right)d\,\mu_k^x\times\int_{\mathbb{R}_{\geq 0}}\frac{1}{\left(\sigma_k^x\right)^2}\,p_{\mathrm{IG}_{\text{nat}}}\left((\sigma_k^x)^2\,;\right.\\
\quad\left.\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to(\sigma_k^x)^2}+\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to(\sigma_k^x)^2}\right)\\
\quad d\left(\sigma_k^x\right)^2\\
\\
-\frac{1}{2}\sum_{a_{ik}=0}^{1}a_{ik}\times a_{ik}\times\left(\eta_{p(a_{ik}|\omega_k)\to a_{ik}}\right.\\
\quad\left.+\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to a_{ik}}\right)\times\int_{\mathbb{R}_{\geq 0}}\frac{1}{\left(\sigma_k^x\right)^2}\,p_{\mathrm{IG}_{\text{nat}}}\left((\sigma_k^x)^2\,;\right.\\
\quad\left.\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to(\sigma_k^x)^2}+\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to(\sigma_k^x)^2}\right)\\
\quad d\left(\sigma_k^x\right)^2
\end{array}\right]\Bigg\}.
\end{aligned}
$$

This gives

$$\boldsymbol{m}_{\,p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to\boldsymbol{x}_{\text{unobs}}}(\boldsymbol{x}_{\text{unobs}})\leftarrow$$

$$= \exp\left\{\sum_{i=1}^{n_{\text{unobs}}}\sum_{k=1}^{K}\begin{bmatrix} x_{\text{unobs,i}} \\ x_{\text{unobs,i}}^2\end{bmatrix}^{\top} \times\begin{bmatrix} -\frac{1}{2}\exp\left(\eta_{p(a_{ik}|\omega_k)\to a_{ik}}+\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to a_{ik}}\right) \\ \quad\times\dfrac{\left(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to\mu_k^x}+\boldsymbol{\eta}_{p(\mu_k^x)\to\mu_k^x}\right)_1}{\left(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to\mu_k^x}+\boldsymbol{\eta}_{p(\mu_k^x)\to\mu_k^x}\right)_2} \\ \quad\times\dfrac{\left(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to(\sigma_k^x)^2}+\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to(\sigma_k^x)^2}\right)_1^{+1}}{\left(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to(\sigma_k^x)^2}+\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to(\sigma_k^x)^2}\right)_2} \\ -\frac{1}{2}\exp\left(\eta_{p(a_{ik}|\omega_k)\to a_{ik}}+\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to a_{ik}}\right) \\ \quad\times\dfrac{\left(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to(\sigma_k^x)^2}+\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to(\sigma_k^x)^2}\right)_1^{+1}}{\left(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to(\sigma_k^x)^2}+\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to(\sigma_k^x)^2}\right)_2} \end{bmatrix}\right\}$$

and so the update for the corresponding natural parameter is

$$\boldsymbol{\eta}_{p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to x_{\text{unobs,i}}}\leftarrow$$

$$\begin{bmatrix} -\frac{1}{2}\exp\left(\eta_{p(a_{ik}|\omega_k)\to a_{ik}}+\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to a_{ik}}\right) \\ \quad\times\dfrac{\left(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to\mu_k^x}+\boldsymbol{\eta}_{p(\mu_k^x)\to\mu_k^x}\right)_1}{\left(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to\mu_k^x}+\boldsymbol{\eta}_{p(\mu_k^x)\to\mu_k^x}\right)_2} \\ \quad\times\dfrac{\left(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to(\sigma_k^x)^2}+\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to(\sigma_k^x)^2}\right)_1^{+1}}{\left(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to(\sigma_k^x)^2}+\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to(\sigma_k^x)^2}\right)_2} \\ -\frac{1}{2}\exp\left(\eta_{p(a_{ik}|\omega_k)\to a_{ik}}+\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to a_{ik}}\right) \\ \quad\times\dfrac{\left(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to(\sigma_k^x)^2}+\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to(\sigma_k^x)^2}\right)_1^{+1}}{\left(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to(\sigma_k^x)^2}+\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to(\sigma_k^x)^2}\right)_2} \end{bmatrix}.$$

The next message has the form:

$$
\begin{aligned}
&\boldsymbol{m}_{\,p(\boldsymbol{o}\,|\,\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\to\,\boldsymbol{x}_{\text{unobs}}}\left(\boldsymbol{x}_{\text{unobs}}\right) \\
&\quad\leftarrow \exp\Bigg\{\int_{\mathbb{R}^3}\int_{\mathbb{R}_{\geq 0}} \frac{1}{Z}\boldsymbol{m}_{\,\boldsymbol{\alpha}\,\to\,p(\boldsymbol{o}\,|\,\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)}\left(\boldsymbol{\alpha}\right)\times \boldsymbol{m}_{\,p(\boldsymbol{o}\,|\,\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\to\,\boldsymbol{\alpha}}\left(\boldsymbol{\alpha}\right) \\
&\qquad\times \boldsymbol{m}_{\,\sigma_o^2\,\to\,p(\boldsymbol{o}\,|\,\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)}\left(\sigma_o^2\right)\times \boldsymbol{m}_{\,p(\boldsymbol{o}\,|\,\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\to\,\sigma_o^2}\left(\sigma_o^2\right) \\
&\qquad\times \log p(\boldsymbol{o}\,|\,\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,d\,\boldsymbol{\alpha}\,d\,\sigma_o^2\Bigg\} \\
&\quad\propto \exp\Bigg[\int_{\mathbb{R}^3}\int_{\mathbb{R}_{\geq 0}}\left\{\begin{bmatrix}\boldsymbol{\alpha}\\ \operatorname{vec}\left(\boldsymbol{\alpha}\boldsymbol{\alpha}^\top\right)\end{bmatrix}^\top\left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\to\,\boldsymbol{\alpha}}+\boldsymbol{\eta}_{p(\boldsymbol{\alpha})\,\to\,\boldsymbol{\alpha}}\right)\right\} \\
&\qquad\times\left\{\begin{bmatrix}\log\left(\sigma_o^2\right)\\ 1/\sigma_o^2\end{bmatrix}^\top\left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\to\,\sigma_o^2}+\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\,\to\,\sigma_o^2}\right)\right\} \\
&\qquad\times\sum_{i=1}^{n_{\text{unobs}}}\begin{bmatrix}x_{\text{unobs,i}}\\ x_{\text{unobs,i}}^2\end{bmatrix}^\top\begin{bmatrix}\frac{\alpha_1}{\sigma_o^2}\left(\boldsymbol{o}_{x_{\text{unobs},i}}+\alpha_0\right)\\ -\frac{\alpha_1^2}{2\sigma_o^2}\end{bmatrix}d\,\boldsymbol{\alpha}\,d\,\sigma_o^2\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
=\exp\Bigg\{&\sum_{i=1}^{n_{\text{unobs}}}\begin{bmatrix}x_{\text{unobs,i}}\\ x_{\text{unobs,i}}^2\end{bmatrix}^\top \\
&\times\begin{bmatrix}
\int_{\mathbb{R}_{\geq 0}}\frac{1}{\sigma_o^2}p_{\text{IG}_{\text{nat}}}\left(\sigma_o^2\,;\,\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\to\,\sigma_o^2}+\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\,\to\,\sigma_o^2}\right)d\,\sigma_o^2 \\
\quad\times\Big\{\boldsymbol{o}_{x_{\text{unobs},i}}\int_{\mathbb{R}}\alpha_1 p_{\text{N}_{\text{nat}}}\left(\alpha_1\,;\,\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\to\,\alpha_1}\right. \\
\quad\left.+\boldsymbol{\eta}_{p(\alpha_1)\,\to\,\alpha_1}\right)d\alpha_1+\int_{\mathbb{R}}\int_{\mathbb{R}}\alpha_0\alpha_1 p_{\text{N}_{\text{nat}}}\left(\alpha_0,\alpha_1\,;\right. \\
\quad\left.\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\to\,\alpha_0}\boldsymbol{\eta}_{p(\alpha_0)\,\to\,\alpha_0},\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\to\,\alpha_1}\right. \\
\quad\left.+\boldsymbol{\eta}_{p(\alpha_1)\,\to\,\alpha_1}\right)d\,\alpha_0\,d\,\alpha_1\Big\} \\
-\frac{1}{2}\int_{\mathbb{R}_{\geq 0}}\frac{1}{\sigma_o^2}p_{\text{IG}_{\text{nat}}}\left(\sigma_o^2\,;\,\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\to\,\sigma_o^2}+\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\,\to\,\sigma_o^2}\right)d\,\sigma_o^2 \\
\quad\times\int_{\mathbb{R}}\alpha_1^2 p_{\text{N}_{\text{nat}}}\left(\alpha_1\,;\,\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\to\,\alpha_1}+\boldsymbol{\eta}_{p(\alpha_1)\,\to\,\alpha_1}\right)d\,\alpha_1
\end{bmatrix}\Bigg\}.
\end{aligned}
$$

Using Primitives 7.4.6, 7.4, and 7.4.1, this gives

$$\boldsymbol{m}_{\,p(\boldsymbol{o}\,|\,\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\rightarrow\,\boldsymbol{x}_{\text{unobs}}}\left(\boldsymbol{x}_{\text{unobs}}\right)$$

$$\propto \exp\left\{ \sum_{i=1}^{n_{\text{unobs}}} \begin{bmatrix} x_{\text{unobs,i}} \\ x_{\text{unobs,i}}^2 \end{bmatrix}^{\top} \begin{bmatrix} \dfrac{\left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\rightarrow\,\sigma_o^2} + \boldsymbol{\eta}_{p(\sigma_o^2|a_o)\,\rightarrow\,\sigma_o^2}\right)_1 + 1}{\left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\rightarrow\,\sigma_o^2} + \boldsymbol{\eta}_{p(\sigma_o^2|a_o)\,\rightarrow\,\sigma_o^2}\right)_2} \\ \quad \times \left\{ -\frac{1}{2}\boldsymbol{o}_{x_{\text{unobs},i}} \dfrac{\left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\rightarrow\,\alpha_1} + \boldsymbol{\eta}_{p(\alpha_1)\,\rightarrow\,\alpha_1}\right)_1}{\left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\rightarrow\,\alpha_1} + \boldsymbol{\eta}_{p(\alpha_1)\,\rightarrow\,\alpha_1}\right)_2} \right. \\ \quad \left. + \mathcal{J}\left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\rightarrow\,\boldsymbol{\alpha}} + \boldsymbol{\eta}_{p(\boldsymbol{\alpha})\,\rightarrow\,\boldsymbol{\alpha}}\,;\,1,2\right)\right\} \\ -\dfrac{\left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\rightarrow\,\sigma_o^2} + \boldsymbol{\eta}_{p(\sigma_o^2|a_o)\,\rightarrow\,\sigma_o^2}\right)_1 + 1}{2\left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\rightarrow\,\sigma_o^2} + \boldsymbol{\eta}_{p(\sigma_o^2|a_o)\,\rightarrow\,\sigma_o^2}\right)_2} \\ \quad \times \mathcal{K}\left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\rightarrow\,\alpha_1} + \boldsymbol{\eta}_{p(\alpha_1)\,\rightarrow\,\alpha_1}\right) \end{bmatrix} \right\}.$$

Therefore, the corresponding natural parameter update is

$$\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\rightarrow\,x_{\text{unobs,i}}} \leftarrow \begin{bmatrix} \dfrac{\left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\rightarrow\,\sigma_o^2} + \boldsymbol{\eta}_{p(\sigma_o^2|a_o)\,\rightarrow\,\sigma_o^2}\right)_1 + 1}{\left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\rightarrow\,\sigma_o^2} + \boldsymbol{\eta}_{p(\sigma_o^2|a_o)\,\rightarrow\,\sigma_o^2}\right)_2} \\ \quad \times \left\{ -\frac{1}{2}\boldsymbol{o}_{x_{\text{unobs},i}} \dfrac{\left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\rightarrow\,\alpha_1} + \boldsymbol{\eta}_{p(\alpha_1)\,\rightarrow\,\alpha_1}\right)_1}{\left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\rightarrow\,\alpha_1} + \boldsymbol{\eta}_{p(\alpha_1)\,\rightarrow\,\alpha_1}\right)_2} \right. \\ \quad \left. + \mathcal{J}\left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\rightarrow\,\boldsymbol{\alpha}} + \boldsymbol{\eta}_{p(\boldsymbol{\alpha})\,\rightarrow\,\boldsymbol{\alpha}}\,;\,1,2\right)\right\} \\ -\dfrac{\left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\rightarrow\,\sigma_o^2} + \boldsymbol{\eta}_{p(\sigma_o^2|a_o)\,\rightarrow\,\sigma_o^2}\right)_1 + 1}{2\left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\rightarrow\,\sigma_o^2} + \boldsymbol{\eta}_{p(\sigma_o^2|a_o)\,\rightarrow\,\sigma_o^2}\right)_2} \\ \quad \times \mathcal{K}\left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\rightarrow\,\alpha_1} + \boldsymbol{\eta}_{p(\alpha_1)\,\rightarrow\,\alpha_1}\right) \end{bmatrix}.$$

Moving on, we next look at

$$\begin{aligned} \boldsymbol{m}_{\,p(\boldsymbol{a}\,|\,\boldsymbol{\omega})\,\rightarrow\,\boldsymbol{a}}\left(\boldsymbol{a}\right) &\leftarrow \exp\left\{ \sum_{k=1}^{K} \sum_{\omega_k=0}^{1} \frac{1}{Z} \boldsymbol{m}_{\,\boldsymbol{\omega}\,\rightarrow\,p(\boldsymbol{a}\,|\,\boldsymbol{\omega})}\left(\boldsymbol{\omega}\right) \times \boldsymbol{m}_{\,p(\boldsymbol{a}\,|\,\boldsymbol{\omega})\,\rightarrow\,\boldsymbol{\omega}}\left(\boldsymbol{\omega}\right) \right. \\ &\qquad \left. \times \log p(\boldsymbol{a}\,|\,\boldsymbol{\omega})\, d\,\boldsymbol{\omega} \right\} \\ &\propto \exp\left\{ \sum_{k=1}^{K} \sum_{\omega_k=0}^{1} \sum_{i=1}^{n} \log\left(\omega_k\right) \left(\eta_{p(a_{ik}|\omega_k)\,\rightarrow\,\omega_k} + \eta_{p(\omega_k)\,\rightarrow\,\omega_k}\right) \right. \\ &\qquad \left. \times a_{ik} \log\left(\omega_k\right)\, d\,\omega_k \right\}. \end{aligned}$$

Using Result 7.4.2 gives

$$\begin{aligned} \boldsymbol{m}_{\,p(\boldsymbol{a}\,|\,\boldsymbol{\omega})\,\rightarrow\,\boldsymbol{a}}\left(\boldsymbol{a}\right) \propto \exp\left\{ \sum_{k=1}^{K} \sum_{i=1}^{n} a_{ik} \left(\psi\left\{ \left(\eta_{p(\omega_k)\,\rightarrow\,\omega_k} + \eta_{p(a_{ik}|\omega_k)\,\rightarrow\,\omega_k}\right)_k + 1 \right\} \right. \right. \\ \left. \left. -\psi\left[\sum_{k=1}^{K} \left\{ \left(\eta_{p(\omega_k)\,\rightarrow\,\omega_k} + \eta_{p(a_{ik}|\omega_k)\,\rightarrow\,\omega_k}\right)_l + 1 \right\} \right] \right) \right\}. \end{aligned}$$

Thus, the corresponding natural parameter is

$$\eta_{p(a_{ik}|\omega_k)\to a_{ik}} \leftarrow \Big(\psi\left\{\left(\eta_{p(\omega_k)\to\omega_k}+\eta_{p(a_{ik}|\omega_k)\to\omega_k}\right)_k+1\right\} - \psi\left[\sum_{k=1}^{K}\left\{\left(\eta_{p(\omega_k)\to\omega_k}+\eta_{p(a_{ik}|\omega_k)\to\omega_k}\right)_l+1\right\}\right]\Big).$$

Next, we consider the message

$$\begin{aligned}\boldsymbol{m}_{\,p(\boldsymbol{a}\,|\,\boldsymbol{\omega})\to\boldsymbol{\omega}}(\boldsymbol{\omega}) &\leftarrow \exp\left\{\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{a_{ik}=0}^{1}\frac{1}{Z}\boldsymbol{m}_{\,\boldsymbol{a}\to p(\boldsymbol{a}\,|\,\boldsymbol{\omega})}(\boldsymbol{a})\times\boldsymbol{m}_{\,p(\boldsymbol{a}\,|\,\boldsymbol{\omega})\to\boldsymbol{a}}(\boldsymbol{a})\right.\\ &\qquad\left.\times\log p\left(\boldsymbol{a}\,|\boldsymbol{\omega}\right)\right\}\\ &\propto \exp\left\{\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{a_{ik}=0}^{1}a_{ik}\left(\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to a_{ik}}+\eta_{p(a_{ik}|\omega_k)\to a_{ik}}\right)\right.\\ &\qquad\left.\times\log\left(\omega_k\right)a_{ik}\right\}.\end{aligned}$$

The use of Result 7.4.1 gives

$$\boldsymbol{m}_{\,p(\boldsymbol{a}\,|\,\boldsymbol{\omega})\to\boldsymbol{\omega}}(\boldsymbol{\omega}) \propto \exp\left[\sum_{i=1}^{n}\sum_{k=1}^{K}\log\left(\omega_k\right)\left\{\exp\left(\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to a_{ik}}+\eta_{p(a_{ik}|\omega_k)\to a_{ik}}\right)\right\}\right].$$

This means that the update for the natural parameter vector is

$$\eta_{p(a_{ik}|\omega_k)\to\omega_k} \leftarrow \exp\left(\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to a_{ik}}+\eta_{p(a_{ik}|\omega_k)\to a_{ik}}\right).$$

Next we derive the update for the factor to node message $\boldsymbol{m}_{\,p(\boldsymbol{\omega})\to\boldsymbol{\omega}}(\boldsymbol{\omega})$:

$$\begin{aligned}\boldsymbol{m}_{\,p(\boldsymbol{\omega})\to\boldsymbol{\omega}}(\boldsymbol{\omega}) &\leftarrow \exp\left\{\tfrac{1}{Z}\log p\left(\boldsymbol{\omega}\right)\right\}\\ &\propto \exp\left\{\sum_{k=1}^{K}\log\left(\omega_k\right)\left(\alpha_k-1\right)\right\}.\end{aligned}$$

Therefore the update for the corresponding natural parameter vector is

$$\eta_{p(\omega_k)\to\omega_k} \leftarrow \alpha_k - 1.$$

Similarly, we next consider another constant update, whose corresponding message takes the form

$$\begin{aligned}\boldsymbol{m}_{\,p(\boldsymbol{\beta})\to\boldsymbol{\beta}}(\boldsymbol{\beta}) &\leftarrow \exp\left\{\tfrac{1}{Z}\log p\left(\boldsymbol{\beta}\right)\right\}\\ &\propto \exp\left\{\begin{bmatrix}\boldsymbol{\beta}\\ \operatorname{vec}\left(\boldsymbol{\beta}\boldsymbol{\beta}^{\top}\right)\end{bmatrix}^{\top}\begin{bmatrix}\mathbf{0}\\ -\frac{1}{2}\operatorname{vec}\left(\frac{1}{\sigma_\beta^2}\boldsymbol{I}_3\right)\end{bmatrix}\right\},\end{aligned}$$

and so the natural parameter vector takes on the constant value

$$\boldsymbol{\eta}_{p(\boldsymbol{\beta}) \to \boldsymbol{\beta}} \leftarrow \begin{bmatrix} \mathbf{0} \\ -\frac{1}{2}\mathrm{vec}\left(\frac{1}{\sigma_\beta^2}\boldsymbol{I}_3\right) \end{bmatrix}.$$

Following on, we next consider

$$\begin{aligned}
\boldsymbol{m}_{\,p(\boldsymbol{o}\,|\,\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2) \to \boldsymbol{\alpha}}(\boldsymbol{\alpha}) \leftarrow \exp\Bigg\{ & \int_{\mathbb{R}^{n_{\text{unobs}}}} \int_{\mathbb{R}_{\geq 0}} \frac{1}{Z} \boldsymbol{m}_{\,\boldsymbol{x}_{\text{unobs}} \to p(\boldsymbol{o}\,|\,\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)}(\boldsymbol{x}_{\text{unobs}}) \\
& \times \boldsymbol{m}_{\,p(\boldsymbol{o}\,|\,\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2) \to \boldsymbol{x}_{\text{unobs}}}(\boldsymbol{x}_{\text{unobs}}) \\
& \times \boldsymbol{m}_{\,\sigma_o^2 \to p(\boldsymbol{o}\,|\,\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)}(\sigma_o^2) \\
& \times \boldsymbol{m}_{\,p(\boldsymbol{o}\,|\,\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2) \to \sigma_o^2}(\sigma_o^2) \\
& \times \log p(\boldsymbol{o}\,|\,\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\, d\,\boldsymbol{x}_{\text{unobs}}\, d\,\sigma_o^2 \Bigg\}
\end{aligned}$$

$$\begin{aligned}
\propto \exp\Bigg[& \int_{\mathbb{R}^{n_{\text{unobs}}}} \int_{\mathbb{R}_{\geq 0}} \Bigg\{ \sum_{i=1}^{n_{\text{unobs}}} \begin{bmatrix} x_{\text{unobs,i}} \\ x_{\text{unobs,i}}^2 \end{bmatrix}^\top \Big(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\} \to x_{\text{unobs,i}}} \\
& + \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2) \to x_{\text{unobs,i}}} + \boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2) \to x_{\text{unobs,i}}} \Big) \Bigg\} \\
& \times \left\{ \begin{bmatrix} \log\left(\sigma_o^2\right) \\ 1/\sigma_o^2 \end{bmatrix}^\top \left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2) \to \sigma_o^2} + \boldsymbol{\eta}_{p(\sigma_o^2|a_o) \to \sigma_o^2} \right) \right\} \\
& \times \begin{bmatrix} \boldsymbol{\alpha} \\ \mathrm{vec}\left(\boldsymbol{\alpha}\boldsymbol{\alpha}^\top\right) \end{bmatrix}^\top \begin{bmatrix} \frac{1}{\sigma_o^2}\tilde{\boldsymbol{X}}^\top \boldsymbol{o} \\ -\frac{1}{2\sigma_o^2}\mathrm{vec}\left(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}\right) \end{bmatrix} d\,\boldsymbol{x}_{\text{unobs}}\, d\,\sigma_o^2 \Bigg]
\end{aligned}$$

$$\begin{aligned}
= \exp\Bigg\{ & \begin{bmatrix} \boldsymbol{\alpha} \\ \mathrm{vec}\left(\boldsymbol{\alpha}\boldsymbol{\alpha}^\top\right) \end{bmatrix}^\top \\
& \times \begin{bmatrix} \displaystyle\int_{\mathbb{R}_{\geq 0}} \frac{1}{\sigma_o^2}\, p_{\text{IG}_{\text{nat}}}\left(\sigma_o^2\,;\, \boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2) \to \sigma_o^2} + \boldsymbol{\eta}_{p(\sigma_o^2|a_o) \to \sigma_o^2}\right) \\ \quad \times E^{\text{nat}}_{q(\boldsymbol{x}_{\text{unobs}})}\left(\tilde{\boldsymbol{X}}\right)^\top \boldsymbol{o} \\ -\frac{1}{2}\displaystyle\int_{\mathbb{R}_{\geq 0}} \frac{1}{\sigma_o^2}\, p_{\text{IG}_{\text{nat}}}\left(\sigma_o^2\,;\, \boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2) \to \sigma_o^2} + \boldsymbol{\eta}_{p(\sigma_o^2|a_o) \to \sigma_o^2}\right) \\ \quad \times \mathrm{vec}\left\{ E^{\text{nat}}_{q(\boldsymbol{x}_{\text{unobs}})}\left(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}\right) \right\} \end{bmatrix} \Bigg\}
\end{aligned}$$

$$= \exp\left\{ \begin{bmatrix} \boldsymbol{\alpha} \\ \mathrm{vec}\left(\boldsymbol{\alpha}\boldsymbol{\alpha}^\top\right) \end{bmatrix}^\top \begin{bmatrix} \frac{\left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\sigma_o^2} + \boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to\sigma_o^2}\right)_1 + 1}{\left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\sigma_o^2} + \boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to\sigma_o^2}\right)_2} \\ \quad \times E^{\mathrm{nat}}_{q(\boldsymbol{x}_{\mathrm{unobs}})}\left(\tilde{\boldsymbol{X}}\right)^\top \boldsymbol{o} \\ -\frac{\left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\sigma_o^2} + \boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to\sigma_o^2}\right)_1 + 1}{2\left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\sigma_o^2} + \boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to\sigma_o^2}\right)_2} \\ \quad \times \mathrm{vec}\left\{ E^{\mathrm{nat}}_{q(\boldsymbol{x}_{\mathrm{unobs}})}\left(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}\right)\right\} \end{bmatrix} \right\}$$

where

$$E^{\mathrm{nat}}_{q(\boldsymbol{x}_{\mathrm{unobs}})}\left(\tilde{\boldsymbol{X}}\right) = \begin{bmatrix} \mathbf{1}_{n_{\mathrm{obs}}} & \boldsymbol{x}_{\mathrm{obs}} \\ \mathbf{1}_{n_{\mathrm{unobs}}} & -\frac{1}{2}\left\{\mathrm{vec}^{-1}\left(\eta^{\circledast}_{2,\mathrm{vec}}\right)\right\}^{-1}\eta^{\circledast}_1 \end{bmatrix},$$

$E^{\mathrm{nat}}_{q(\boldsymbol{x}_{\mathrm{unobs}})}\left(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}\right) =$

$$\left[\begin{array}{c|c} n & \begin{array}{c} \mathbf{1}^\top_{n_{\mathrm{obs}}}\boldsymbol{x}_{\mathrm{obs}} \\ -\frac{1}{2}\mathbf{1}^\top_{n_{\mathrm{unobs}}}\left\{\mathrm{vec}^{-1}\left(\eta^{\circledast}_{2,\mathrm{vec}}\right)\right\}^{-1}\eta^{\circledast}_1 \end{array} \\ \hline \begin{array}{c} \mathbf{1}^\top_{n_{\mathrm{obs}}}\boldsymbol{x}_{\mathrm{obs}} \\ -\frac{1}{2}\mathbf{1}^\top_{n_{\mathrm{unobs}}}\left\{\mathrm{vec}^{-1}\left(\eta^{\circledast}_{2,\mathrm{vec}}\right)\right\}^{-1}\eta^{\circledast}_1 \end{array} & \begin{array}{c} \left\|\boldsymbol{x}_{\mathrm{obs}}\right\|^2 + \\ \left\| -\frac{1}{2}\left\{\mathrm{vec}^{-1}\left(\eta^{\circledast}_{2,\mathrm{vec}}\right)\right\}^{-1}\eta^{\circledast}_1\right\|^2 \\ -\frac{1}{2}\left\{\mathrm{vec}^{-1}\left(\eta^{\circledast}_{2,\mathrm{vec}}\right)\right\}^{-1} \end{array} \end{array}\right]$$

and

$$\boldsymbol{\eta}^{\circledast} \equiv \left[\eta^{\circledast}_1 \ \eta^{\circledast}_{2,\mathrm{vec}}\right]^\top \equiv \boldsymbol{\eta}_{p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to\boldsymbol{x}_{\mathrm{unobs}}} + \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\boldsymbol{x}_{\mathrm{unobs}}} + \boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\boldsymbol{x}_{\mathrm{unobs}}}.$$

Therefore the corresponding natural parameter update is

$$\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\boldsymbol{\alpha}} \leftarrow \begin{bmatrix} \frac{\left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\sigma_o^2} + \boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to\sigma_o^2}\right)_1 + 1}{\left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\sigma_o^2} + \boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to\sigma_o^2}\right)_2} \\ \quad \times E^{\mathrm{nat}}_{q(\boldsymbol{x}_{\mathrm{unobs}})}\left(\tilde{\boldsymbol{X}}\right)^\top \boldsymbol{o} \\ -\frac{\left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\sigma_o^2} + \boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to\sigma_o^2}\right)_1 + 1}{2\left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\sigma_o^2} + \boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to\sigma_o^2}\right)_2} \\ \quad \times \mathrm{vec}\left\{ E^{\mathrm{nat}}_{q(\boldsymbol{x}_{\mathrm{unobs}})}\left(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}\right)\right\} \end{bmatrix}.$$

Next we have

$$\begin{aligned} \boldsymbol{m}_{\,p(\boldsymbol{\alpha})\to\boldsymbol{\alpha}}(\boldsymbol{\alpha}) &\leftarrow \exp\left\{\tfrac{1}{Z}\log p(\boldsymbol{\alpha})\right\} \\ &\propto \exp\left\{ \begin{bmatrix} \boldsymbol{\alpha} \\ \mathrm{vec}\left(\boldsymbol{\alpha}\boldsymbol{\alpha}^\top\right) \end{bmatrix}^\top \begin{bmatrix} \mathbf{0} \\ -\frac{1}{2}\mathrm{vec}\left(\frac{1}{\sigma_\alpha^2}\boldsymbol{I}_2\right) \end{bmatrix} \right\} \end{aligned}$$

Therefore the natural parameter update stays constant at:

$$\boldsymbol{\eta}_{p(\boldsymbol{\alpha})\to\boldsymbol{\alpha}} \leftarrow \begin{bmatrix} \mathbf{0} \\ -\frac{1}{2}\mathrm{vec}\left(\frac{1}{\sigma_\alpha^2}\boldsymbol{I}_2\right) \end{bmatrix}.$$

The next message we look at is from $p(\boldsymbol{o}\,|\,\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)$ to σ_o^2:

$$\begin{aligned}
&\boldsymbol{m}_{\,p(\boldsymbol{o}\,|\,\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\sigma_o^2}\left(\sigma_o^2\right) \\
&\quad\leftarrow \exp\Bigg\{\int_{\mathbb{R}^2}\int_{\mathbb{R}^{n_{\text{unobs}}}} \frac{1}{Z}\boldsymbol{m}_{\,\boldsymbol{\alpha}\to p(\boldsymbol{o}\,|\,\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)}(\boldsymbol{\alpha})\times \boldsymbol{m}_{\,p(\boldsymbol{o}\,|\,\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\boldsymbol{\alpha}}(\boldsymbol{\alpha}) \\
&\qquad\quad \times \boldsymbol{m}_{\,\boldsymbol{x}_{\text{unobs}}\to p(\boldsymbol{o}\,|\,\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)}(\boldsymbol{x}_{\text{unobs}})\times \boldsymbol{m}_{\,p(\boldsymbol{o}\,|\,\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\boldsymbol{x}_{\text{unobs}}}(\boldsymbol{x}_{\text{unobs}}) \\
&\qquad\quad \times \log p(\boldsymbol{o}\,|\,\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,d\boldsymbol{\alpha},\,d\boldsymbol{x}_{\text{unobs}}\Bigg\} \\
&\quad\propto \exp\Bigg[\int_{\mathbb{R}^2}\int_{\mathbb{R}^{n_{\text{unobs}}}} \left\{\begin{bmatrix}\boldsymbol{\alpha}\\ \mathrm{vec}\left(\boldsymbol{\alpha}\boldsymbol{\alpha}^\top\right)\end{bmatrix}^\top \left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\boldsymbol{\alpha}} + \boldsymbol{\eta}_{p(\boldsymbol{\alpha})\to\boldsymbol{\alpha}}\right)\right\} \\
&\qquad\quad \times\Bigg\{\sum_{i=1}^{n_{\text{unobs}}}\begin{bmatrix}x_{\text{unobs,i}}\\ x_{\text{unobs,i}}^2\end{bmatrix}^\top \Big(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to x_{\text{unobs,i}}} + \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to x_{\text{unobs,i}}} \\
&\qquad\qquad + \boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to x_{\text{unobs,i}}}\Big)\Bigg\}\times\begin{bmatrix}\log\left(\sigma_o^2\right)\\ 1/\sigma_o^2\end{bmatrix}^\top\begin{bmatrix}-n/2\\ -\frac{1}{2}\|\boldsymbol{o}-\tilde{\boldsymbol{X}}\boldsymbol{\alpha}\|^2\end{bmatrix} d\boldsymbol{\alpha}\,d\boldsymbol{x}_{\text{unobs}}\Bigg] \\
&\quad= \exp\left\{\begin{bmatrix}\log\left(\sigma_o^2\right)\\ 1/\sigma_o^2\end{bmatrix}^\top\begin{bmatrix}-n/2\\ -\frac{1}{2}E^{\text{nat}}_{q(\boldsymbol{x}_{\text{unobs}})}\left(\|\boldsymbol{o}-\tilde{\boldsymbol{X}}\boldsymbol{\alpha}\|^2\right)\end{bmatrix}\right\}
\end{aligned}$$

where

$$\begin{aligned}
&E^{\text{nat}}_{q(\boldsymbol{x}_{\text{unobs}})}\left(\|\boldsymbol{o}-\tilde{\boldsymbol{X}}\boldsymbol{\alpha}\|^2\right) = \\
&\left\|\boldsymbol{o}_{x\text{obs}} + \tfrac{1}{2}\tilde{\boldsymbol{X}}_{x\text{obs}}\left\{\mathrm{vec}^{-1}\left(\eta_{2,\text{vec}}^{\bigstar}\right)\right\}^{-1}\eta_1^{\bigstar}\right\|^2 - \tfrac{1}{2}\mathrm{tr}\left[\tilde{\boldsymbol{X}}_{x\text{obs}}^\top\tilde{\boldsymbol{X}}_{x\text{obs}}\left\{\mathrm{vec}^{-1}\left(\eta_{2,\text{vec}}^{\bigstar}\right)\right\}^{-1}\right] \\
&+\|\boldsymbol{o}_{x\text{unobs}}\|^2 + \boldsymbol{o}_{x\text{unobs}}^\top E^{\text{nat}}_{q(\boldsymbol{x}_{\text{unobs}})}\left(\tilde{\boldsymbol{X}}_{x\text{unobs}}\right)\left\{\mathrm{vec}^{-1}\left(\eta_{2,\text{vec}}^{\bigstar}\right)\right\}^{-1}\eta_1^{\bigstar} \\
&+\tfrac{1}{4}\mathrm{tr}\Big\{E^{\text{nat}}_{q(\boldsymbol{x}_{\text{unobs}})}\left(\tilde{\boldsymbol{X}}_{x\text{unobs}}^\top\tilde{\boldsymbol{X}}_{x\text{unobs}}\right)\left\{\mathrm{vec}^{-1}\left(\eta_{2,\text{vec}}^{\bigstar}\right)\right\}^{-1}\Big[\eta_1^{\bigstar}\left(\eta_1^{\bigstar}\right)^\top\left\{\mathrm{vec}^{-1}\left(\eta_{2,\text{vec}}^{\bigstar}\right)\right\}^{-1} \\
&-2\boldsymbol{I}]\Big\},
\end{aligned}$$

$$E^{\text{nat}}_{q(\boldsymbol{x}_{\text{unobs}})}\left(\tilde{\boldsymbol{X}}_{x\text{unobs}}\right) = \left[\mathbf{1}_{n_{\text{unobs}}} \quad -\tfrac{1}{2}\left\{\mathrm{vec}^{-1}\left(\eta_{2,\text{vec}}^{\circledast}\right)\right\}^{-1}\eta_1^{\circledast}\right],$$

$$E^{\text{nat}}_{q(\boldsymbol{x}_{\text{unobs}})}\left(\tilde{\boldsymbol{X}}_{x\text{unobs}}^\top\tilde{\boldsymbol{X}}_{x\text{unobs}}\right) =$$

$$\left[\begin{array}{c|c} n_{\text{unobs}} & -\frac{1}{2}\mathbf{1}_{n_{\text{unobs}}}^\top\left\{\mathrm{vec}^{-1}\left(\eta_{2,\text{vec}}^{\circledast}\right)\right\}^{-1}\eta_1^{\circledast} \\ \hline -\frac{1}{2}\mathbf{1}_{n_{\text{unobs}}}^\top\left\{\mathrm{vec}^{-1}\left(\eta_{2,\text{vec}}^{\circledast}\right)\right\}^{-1}\eta_1^{\circledast} & \begin{array}{c}\|-\frac{1}{2}\left\{\mathrm{vec}^{-1}\left(\eta_{2,\text{vec}}^{\circledast}\right)\right\}^{-1}\eta_1^{\circledast}\|^2 \\ -\frac{1}{2}\left\{\mathrm{vec}^{-1}\left(\eta_{2,\text{vec}}^{\circledast}\right)\right\}^{-1}\end{array}\end{array}\right],$$

and

$$\boldsymbol{\eta}^{\bigstar} \equiv \left[\eta_1^{\bigstar}\ \eta_{2,\mathrm{vec}}^{\bigstar}\right]^{\top} \equiv \boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\to\,\boldsymbol{\alpha}} + \boldsymbol{\eta}_{p(\boldsymbol{\alpha})\,\to\,\boldsymbol{\alpha}}.$$

Therefore, the update for the corresponding natural parameter is

$$\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\to\,\sigma_o^2}\left(\sigma_o^2\right) \leftarrow \begin{bmatrix} -n/2 \\ -\frac{1}{2}E^{\mathrm{nat}}_{q(\boldsymbol{x}_{\mathrm{unobs}})}\left(\|\boldsymbol{o}-\tilde{\boldsymbol{X}}\boldsymbol{\alpha}\|^2\right) \end{bmatrix}.$$

Following on, the next message is

$$\begin{aligned}
\boldsymbol{m}_{\,p(\sigma_o^2|a_o)\,\to\,\sigma_o^2}\left(\sigma_o^2\right) &\leftarrow \exp\Bigg\{\int_{\mathbb{R}_{\geq 0}} \frac{1}{Z}\boldsymbol{m}_{\,a_o\,\to\,p(\sigma_o^2|a_o)}\left(a_o\right) \times \boldsymbol{m}_{\,p(\sigma_o^2|a_o)\,\to\,a_o}\left(a_o\right) \\
&\qquad \times \log p(\sigma_o^2|a_o)\, d\,a_o\Bigg\} \\
&\propto \exp\Bigg\{\int_{\mathbb{R}_{\geq 0}} \begin{bmatrix} \log\left(a_o\right) \\ 1/a_o \end{bmatrix}^{\top} \left(\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\,\to\,a_o} + \boldsymbol{\eta}_{p(a_o)\,\to\,a_o}\right) \\
&\qquad \times \begin{bmatrix} \log\left(\sigma_o^2\right) \\ 1/\sigma_o^2 \end{bmatrix}^{\top} \begin{bmatrix} -3/2 \\ -1/a_o \end{bmatrix} d\,a_o\Bigg\} \\
&= \exp\left\{\begin{bmatrix} \log\left(\sigma_o^2\right) \\ 1/\sigma_o^2 \end{bmatrix}^{\top} \begin{bmatrix} -3/2 \\ -\int_{\mathbb{R}_{\geq 0}} \frac{1}{a_o}\, p_{\mathrm{IG}_{\mathrm{nat}}}\left(\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\,\to\,a_o} + \boldsymbol{\eta}_{p(a_o)\,\to\,a_o}\right) d\,a_o \end{bmatrix}\right\}.
\end{aligned}$$

With use of Primitive 7.4.6, this gives

$$\boldsymbol{m}_{\,p(\sigma_o^2|a_o)\,\to\,\sigma_o^2}\left(\sigma_o^2\right) \propto \exp\left\{\begin{bmatrix} \log\left(\sigma_o^2\right) \\ 1/\sigma_o^2 \end{bmatrix}^{\top} \begin{bmatrix} -3/2 \\ -\frac{\left(\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\,\to\,a_o} + \boldsymbol{\eta}_{p(a_o)\,\to\,a_o}\right)_1 + 1}{\left(\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\,\to\,a_o} + \boldsymbol{\eta}_{p(a_o)\,\to\,a_o}\right)_2} \end{bmatrix}\right\},$$

therefore the corresponding natural parameter update is

$$\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\,\to\,\sigma_o^2} \leftarrow \begin{bmatrix} -3/2 \\ -\frac{\left(\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\,\to\,a_o} + \boldsymbol{\eta}_{p(a_o)\,\to\,a_o}\right)_1 + 1}{\left(\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\,\to\,a_o} + \boldsymbol{\eta}_{p(a_o)\,\to\,a_o}\right)_2} \end{bmatrix}.$$

Next, we look at

$$\begin{aligned}\boldsymbol{m}_{\,p(\sigma_o^2|a_o)\to a_o}(a_o) &\leftarrow \exp\Big\{\int_{\mathbb{R}_{\geq 0}} \frac{1}{Z}\boldsymbol{m}_{\,\sigma_o^2\to p(\sigma_o^2|a_o)}\left(\sigma_o^2\right)\times \boldsymbol{m}_{\,p(\sigma_o^2|a_o)\to\sigma_o^2}\left(\sigma_o^2\right)\\ &\qquad\qquad \times \log p(\sigma_o^2|a_o)\, d\,\sigma_o^2\Big\}\\ &\propto \exp\left\{\int_{\mathbb{R}_{\geq 0}}\begin{bmatrix}\log\left(\sigma_o^2\right)\\ 1/\sigma_o^2\end{bmatrix}^{\top}\left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\sigma_o^2}+\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to\sigma_o^2}\right)\right.\\ &\qquad\qquad \left.\times\begin{bmatrix}\log\left(a_o\right)\\ 1/a_o\end{bmatrix}^{\top}\begin{bmatrix}-\frac{1}{2}\\ -1/\sigma_o^2\end{bmatrix} d\,\sigma_o^2\right\}\end{aligned}$$

$$= \exp\left\{\begin{bmatrix}\log\left(a_o\right)\\ 1/a_o\end{bmatrix}^{\top}\begin{bmatrix}-\frac{1}{2}\\ -\int_{\mathbb{R}_{\geq 0}}\frac{1}{\sigma_o^2}\, p_{\mathrm{IG}_{\mathrm{nat}}}\left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\sigma_o^2}+\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to\sigma_o^2}\right) d\,\sigma_o^2\end{bmatrix}\right\}.$$

Once again, making use of Primitive 7.4.6, we get

$$\boldsymbol{m}_{\,p(\sigma_o^2|a_o)\to a_o}(a_o)\propto\exp\left\{\begin{bmatrix}\log\left(a_o\right)\\ 1/a_o\end{bmatrix}^{\top}\begin{bmatrix}-3/2\\ -\frac{\left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\sigma_o^2}+\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to\sigma_o^2}\right)_1+1}{\left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\sigma_o^2}+\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to\sigma_o^2}\right)_2}\end{bmatrix}\right\},$$

and so, the corresponding natural parameter update is

$$\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to a_o}\leftarrow\begin{bmatrix}-3/2\\ -\frac{\left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\sigma_o^2}+\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to\sigma_o^2}\right)_1+1}{\left(\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\sigma_o^2}+\boldsymbol{\eta}_{p(\sigma_o^2|a_o)\to\sigma_o^2}\right)_2}\end{bmatrix}.$$

Next, we have

$$\begin{aligned}\boldsymbol{m}_{\,p(a_o)\to a_o}(a_o) &\leftarrow \exp\left\{\tfrac{1}{Z}\log\left(a_o\right)\right\}\\ &\propto \exp\left\{\begin{bmatrix}\log\left(a_o\right)\\ 1/a_o\end{bmatrix}^{\top}\begin{bmatrix}-3/2\\ -1/A_o^2\end{bmatrix}\right\},\end{aligned}$$

thus the corresponding natural parameter vector has the constant value

$$\boldsymbol{\eta}_{p(a_o)\to a_o}\leftarrow\begin{bmatrix}-3/2\\ -1/A_o^2\end{bmatrix}.$$

Next, we have

$$
\begin{aligned}
&\boldsymbol{m}_{\,p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\,\to\,(\boldsymbol{\sigma}^x)^2}\left\{(\boldsymbol{\sigma}^x)^2\right\} \\
&\quad\leftarrow \exp\Bigg[\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{a_{ik}=0}^{1}\int_{\mathbb{R}^{n_{\text{unobs}}}}\int_{\mathbb{R}^K}\frac{1}{Z}\boldsymbol{m}_{\,\boldsymbol{a}\,\to\,p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^2)^x\right\}}(\boldsymbol{a}) \\
&\qquad\times\boldsymbol{m}_{\,p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\,\to\,\boldsymbol{a}}(\boldsymbol{a}) \\
&\qquad\times\boldsymbol{m}_{\,\boldsymbol{x}_{\text{unobs}}\,\to\,p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}}(\boldsymbol{x}_{\text{unobs}})\times\boldsymbol{m}_{\,p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\,\to\,\boldsymbol{x}_{\text{unobs}}}(\boldsymbol{x}_{\text{unobs}}) \\
&\qquad\times\boldsymbol{m}_{\,\boldsymbol{\mu}^x\,\to\,p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}}(\boldsymbol{\mu}^x)\times\boldsymbol{m}_{\,p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\,\to\,\boldsymbol{\mu}^x}(\boldsymbol{\mu}^x) \\
&\qquad\times\log p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}d\,\boldsymbol{a}\,d\,\boldsymbol{x}_{\text{unobs}}\,d\,\boldsymbol{\mu}^x\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
&\propto \exp\Bigg[\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{a_{ik}=0}^{1}\int_{\mathbb{R}^{n_{\text{unobs}}}}\int_{\mathbb{R}^K}\left\{a_{ik}\left(\eta_{p(a_{ik}|\omega_k)\,\to\,a_{ik}}+\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\,\to\,a_{ik}}\right)\right\} \\
&\qquad\times\left\{\sum_{i=1}^{n_{\text{unobs}}}\begin{bmatrix}x_{\text{unobs,i}}\\ x_{\text{unobs,i}}^2\end{bmatrix}^{\top}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\,\to\,x_{\text{unobs,i}}}+\boldsymbol{\eta}_{p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\,\to\,x_{\text{unobs,i}}}\right.\right. \\
&\qquad\left.\left.+\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\to\,x_{\text{unobs,i}}}\right)\right\}\times\left\{\sum_{k=1}^{K}\begin{bmatrix}\mu_k^x\\ (\mu_k^x)^2\end{bmatrix}^{\top}\left(\boldsymbol{\eta}_{p(\mu_k^x)\,\to\,\mu_k^x}\right.\right. \\
&\qquad\left.\left.+\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\,\to\,\mu_k^x}\right)\right\}\left\{\sum_{i=1}^{n}\sum_{k=1}^{K}\begin{bmatrix}\log\left\{\left(\sigma_k^x\right)^2\right\}\\ 1/\left(\sigma_k^x\right)^2\end{bmatrix}^{\top}\begin{bmatrix}-\frac{1}{2}a_{ik}\\ -\frac{1}{2}a_{ik}\left(x_i-\mu_k^x\right)^2\end{bmatrix}\right\} \\
&\qquad d\,\boldsymbol{a}\,d\,\boldsymbol{x}_{\text{unobs}}\,d\,\boldsymbol{\mu}^x\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
&\propto \exp\Bigg\{\sum_{i=1}^{n}\sum_{k=1}^{K}\begin{bmatrix}\log\left\{\left(\sigma_k^x\right)^2\right\}\\ 1/\left(\sigma_k^x\right)^2\end{bmatrix}^{\top} \\
&\qquad\begin{bmatrix}-\frac{1}{2}\sum_{a_{ik}=0}^{1}a_{ik}\times a_{ik}\times\left(\eta_{p(a_{ik}|\omega_k)\,\to\,a_{ik}}+\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\,\to\,a_{ik}}\right)\\ \\ -\frac{1}{2}\times a_{ik}\times\left(\eta_{p(a_{ik}|\omega_k)\,\to\,a_{ik}}+\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\,\to\,a_{ik}}\right)\\ \times\int_{\mathbb{R}}\int_{\mathbb{R}}(x_i-\mu_k)^2\,p_{\text{N}_{\text{nat}}}\left(x_i,\mu_k\,;\,\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\,\to\,x_{\text{unobs,i}}}\right.\\ \left.+\boldsymbol{\eta}_{p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\,\to\,x_{\text{unobs,i}}}+\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\to\,x_{\text{unobs,i}}}\,,\right.\\ \left.\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\,\to\,\mu_k^x}+\boldsymbol{\eta}_{p(\mu_k^x)\,\to\,\mu_k^x}\right)d\,x_{\text{unobs,i}}\,d\,\mu_k^x\end{bmatrix}\Bigg\}.
\end{aligned}
$$

This implies that the $(a_{i1},\ldots,a_{iK})$ follow the

$$
\begin{aligned}
\text{M}_{\text{nat}}\Bigg\{1;&\left(\eta_{p(\boldsymbol{a}_i|\boldsymbol{\omega})\,\to\,\boldsymbol{a}_i}+\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\,\to\,\boldsymbol{a}_i}\right)_1,\ldots,\\
&\left(\eta_{p(\boldsymbol{a}_i|\boldsymbol{\omega})\,\to\,\boldsymbol{a}_i}+\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\,\to\,\boldsymbol{a}_i}\right)_K\Bigg)
\end{aligned}
$$

distribution and making use of Result 7.4.1 and Primitive 7.4.2, we get

$$\boldsymbol{m}_{\,p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to(\boldsymbol{\sigma}^x)^2}\left\{\left(\boldsymbol{\sigma}^x\right)^2\right\}\propto$$
$$\exp\left\{\sum_{i=1}^{n}\sum_{k=1}^{K}\begin{bmatrix}\log\left\{\left(\sigma_k^x\right)^2\right\}\\ 1/\left(\sigma_k^x\right)^2\end{bmatrix}^{\top}\right.$$
$$\left.\times\begin{bmatrix}-\frac{1}{2}\exp\left(\eta_{p(a_{ik}|\omega_k)\to a_{ik}}+\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to a_{ik}}\right)\\ -\frac{1}{2}\exp\left(\eta_{p(a_{ik}|\omega_k)\to a_{ik}}+\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to a_{ik}}\right)\\ \times\mathcal{I}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to x_{\text{unobs,i}}}+\boldsymbol{\eta}_{p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to x_{\text{unobs,i}}}\right.\\ \left.+\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to x_{\text{unobs,i}}}\,,\,\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to\mu_k^x}\right.\\ \left.+\boldsymbol{\eta}_{p(\mu_k^x)\to\mu_k^x}\right)\end{bmatrix}\right\}.$$

Therefore, the update for the corresponding natural parameter vector is

$$\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to(\sigma_k^x)^2}\leftarrow$$
$$\begin{bmatrix}-\frac{1}{2}\exp\left(\eta_{p(a_{ik}|\omega_k)\to a_{ik}}+\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to a_{ik}}\right)\\ -\frac{1}{2}\exp\left(\eta_{p(a_{ik}|\omega_k)\to a_{ik}}+\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to a_{ik}}\right)\\ \times\mathcal{I}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to x_{\text{unobs,i}}}+\boldsymbol{\eta}_{p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to x_{\text{unobs,i}}}\right.\\ \left.+\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to x_{\text{unobs,i}}}\,,\,\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to\mu_k^x}\right.\\ \left.+\boldsymbol{\eta}_{p(\mu_k^x)\to\mu_k^x}\right)\end{bmatrix}.$$

Next we look at

$$\begin{aligned}&\boldsymbol{m}_{\,p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to\boldsymbol{a}}(\boldsymbol{a})\\ &\quad\leftarrow\exp\left[\int_{\mathbb{R}^K}\int_{\mathbb{R}^K}\int_{\mathbb{R}^{n_{\text{unobs}}}}\frac{1}{Z}\boldsymbol{m}_{\,(\boldsymbol{\sigma}^x)^2\to p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}}\left\{\left(\boldsymbol{\sigma}^x\right)^2\right\}\right.\\ &\qquad\times\boldsymbol{m}_{\,p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to(\boldsymbol{\sigma}^x)^2}\left\{\left(\boldsymbol{\sigma}^x\right)^2\right\}\times\boldsymbol{m}_{\,\boldsymbol{\mu}^x\to p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}}(\boldsymbol{\mu}^x)\\ &\qquad\times\boldsymbol{m}_{\,p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to\boldsymbol{\mu}^x}(\boldsymbol{\mu}^x)\times\boldsymbol{m}_{\,\boldsymbol{x}_{\text{unobs}}\to p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}}(\boldsymbol{x}_{\text{unobs}})\\ &\qquad\times\boldsymbol{m}_{\,p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to\boldsymbol{x}_{\text{unobs}}}(\boldsymbol{x}_{\text{unobs}})\times\log p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\\ &\qquad d\,(\boldsymbol{\sigma}^x)^2\,d\,\boldsymbol{\mu}^x\,d\,\boldsymbol{x}_{\text{unobs}}\Big]\end{aligned}$$

$$
\begin{aligned}
&\propto \exp\Bigg[\int_{\mathbb{R}^K}\int_{\mathbb{R}^K}\int_{\mathbb{R}^{n_{\text{unobs}}}}\Bigg\{\sum_{k=1}^{K}\begin{bmatrix}\log\left\{\left(\sigma_k^x\right)^2\right\}\\ 1/\left(\sigma_k^x\right)^2\end{bmatrix}^{\top}\Big(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\rightarrow(\sigma_k^x)^2}\\
&\quad+\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\rightarrow(\sigma_k^x)^2}\Big)\Bigg\}\times\Bigg\{\sum_{k=1}^{K}\begin{bmatrix}\mu_k^x\\ \left(\mu_k^x\right)^2\end{bmatrix}^{\top}\Big(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\rightarrow\mu_k^x}\\
&\quad+\boldsymbol{\eta}_{p(\mu_k^x)\rightarrow\mu_k^x}\Big)\Bigg\}\times\Bigg\{\sum_{i=1}^{n_{\text{unobs}}}\sum_{k=1}^{K}\begin{bmatrix}x_{\text{unobs,i}}\\ x_{\text{unobs,i}}^2\end{bmatrix}^{\top}\Big(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\rightarrow x_{\text{unobs,i}}}\\
&\quad+\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\rightarrow x_{\text{unobs,i}}}+\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\rightarrow x_{\text{unobs,i}}}\Big)\Bigg\}\\
&\quad\times\sum_{i=1}^{n}\sum_{k=1}^{K}a_{ik}\left(\log\left[\left\{2\pi\left(\sigma_k^x\right)^2\right\}^{-1/2}\exp\left\{-\tfrac{1}{2}\left(x_i-\mu_k^x\right)^2/\left(\sigma_k^x\right)^2\right\}\right]\right)\\
&\quad d\left(\boldsymbol{\sigma}^x\right)^2\,d\boldsymbol{\mu}^x\,d\boldsymbol{x}_{\text{unobs}}\Big]
\end{aligned}
$$

$$
\begin{aligned}
&\propto \exp\Bigg\{\sum_{i=1}^{n}\sum_{k=1}^{K}a_{ik}\Bigg(-\tfrac{1}{2}\Bigg[\int_{\mathbb{R}_{\geq 0}}\log\left\{(\sigma_k^x)^2\right\}\,p_{\text{IG}_{\text{nat}}}\Big((\sigma_k^x)^2\;;\;\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\rightarrow(\sigma_k^x)^2}\\
&\quad+\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\rightarrow(\sigma_k^x)^2}\Big)\,d\left(\sigma_k^x\right)^2-\int_{\mathbb{R}_{\geq 0}}\frac{1}{\left(\sigma_k^x\right)^2}p_{\text{IG}_{\text{nat}}}\Big((\sigma_k^x)^2\;;\;\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\rightarrow(\sigma_k^x)^2}\\
&\quad+\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\rightarrow(\sigma_k^x)^2}\Big)\,d\left(\sigma_k^x\right)^2\times\int_{\mathbb{R}}\int_{\mathbb{R}}\left(x_i-\mu_k^x\right)^2\,p_{\text{N}_{\text{nat}}}\left(x_i,\mu_k^x\;;\right.\\
&\quad\boldsymbol{\eta}_{p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\rightarrow x_{\text{unobs,i}}}+\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\rightarrow x_{\text{unobs,i}}}+\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\rightarrow x_{\text{unobs,i}}},\\
&\quad\left.\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\rightarrow\mu_k^x}+\boldsymbol{\eta}_{p(\mu_k^x)\rightarrow\mu_k^x}\Big)\,d\,x_{\text{unobs,i}}\,d\,\mu_k^x\Big]\Big)\Bigg\}.
\end{aligned}
$$

Making use of Primitives 7.4.2, 7.4.6 and 7.4.7, this gives

$$
\begin{aligned}
&\boldsymbol{m}_{\,p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\rightarrow\boldsymbol{a}}(\boldsymbol{a})\propto\\
&\quad\exp\Bigg\{\sum_{i=1}^{n}\sum_{k=1}^{K}a_{ik}\Bigg(-\tfrac{1}{2}\Bigg[\log\Bigg\{-\Big(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\rightarrow(\sigma_k^x)^2}+\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\rightarrow(\sigma_k^x)^2}\Big)_2\Bigg\}\\
&\qquad-\psi\Bigg\{-\Big(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\rightarrow(\sigma_k^x)^2}+\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\rightarrow(\sigma_k^x)^2}\Big)_1-1\Bigg\}\\
&\qquad-\Bigg\{\frac{\Big(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\rightarrow(\sigma_k^x)^2}+\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\rightarrow(\sigma_k^x)^2}\Big)_1^{+1}}{\Big(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\rightarrow(\sigma_k^x)^2}+\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\rightarrow(\sigma_k^x)^2}\Big)_2}\Bigg\}\\
&\qquad\times\mathcal{I}\Big(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\rightarrow x_{\text{unobs,i}}}\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\rightarrow x_{\text{unobs,i}}}+\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\rightarrow x_{\text{unobs,i}}},\\
&\qquad\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\rightarrow\mu_k^x}+\boldsymbol{\eta}_{p(\mu_k^x)\rightarrow\mu_k^x}\Big)\Big]\Big)\Bigg\}.
\end{aligned}
$$

Recognising that $\boldsymbol{a}$ follows a series of n Multinomial distributions, the update for the

corresponding natural parameter vector is

$$
\begin{aligned}
&\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\rightarrow a_{ik}} \leftarrow \\
&\quad -\tfrac{1}{2}\Bigg[\log\Bigg\{-\left(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\rightarrow(\sigma_k^x)^2}+\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\rightarrow(\sigma_k^x)^2}\right)_2\Bigg\} \\
&\quad -\psi\Bigg\{-\left(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\rightarrow(\sigma_k^x)^2}+\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\rightarrow(\sigma_k^x)^2}\right)_1-1\Bigg\} \\
&\quad -\left\{\frac{\left(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\rightarrow(\sigma_k^x)^2}+\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\rightarrow(\sigma_k^x)^2}\right)_1^{+1}}{\left(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\rightarrow(\sigma_k^x)^2}+\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\rightarrow(\sigma_k^x)^2}\right)_2}\right\} \\
&\quad \times\mathcal{I}\Big(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\rightarrow x_{\text{unobs,i}}}\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\rightarrow x_{\text{unobs,i}}} \\
&\quad +\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\rightarrow x_{\text{unobs,i}}},\ \boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\rightarrow\mu_k^x}+\boldsymbol{\eta}_{p(\mu_k^x)\rightarrow\mu_k^x}\Big)\Bigg].
\end{aligned}
$$

Next,

$$
\begin{aligned}
&\boldsymbol{m}_{\,p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\rightarrow\boldsymbol{\mu}^x}\left(\boldsymbol{\mu}^x\right) \\
&\quad\leftarrow\exp\Bigg\{\int_{\mathbb{R}_{\geq 0}}\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{\boldsymbol{a}=0}^{1}\int_{\mathbb{R}^{n_{\text{unobs}}}}\frac{1}{Z}\boldsymbol{m}_{\,(\boldsymbol{\sigma}^x)^2\rightarrow p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}}\left\{\left(\boldsymbol{\sigma}^x\right)^2\right\} \\
&\qquad\times\boldsymbol{m}_{\,p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\rightarrow(\boldsymbol{\sigma}^x)^2}\left\{\left(\boldsymbol{\sigma}^x\right)^2\right\}\times\boldsymbol{m}_{\,\boldsymbol{a}\rightarrow p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,\left(\boldsymbol{\sigma}^2\right)^x\right\}}\left(\boldsymbol{a}\right) \\
&\qquad\times\boldsymbol{m}_{\,p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\rightarrow\boldsymbol{a}}\left(\boldsymbol{a}\right)\times\boldsymbol{m}_{\,\boldsymbol{x}_{\text{unobs}}\rightarrow p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}}\left(\boldsymbol{x}_{\text{unobs}}\right) \\
&\qquad\times\boldsymbol{m}_{\,p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\rightarrow\boldsymbol{x}_{\text{unobs}}}\left(\boldsymbol{x}_{\text{unobs}}\right)\times\log p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,\left(\boldsymbol{\sigma}^x\right)^2\right\} \\
&\qquad d\left(\boldsymbol{\sigma}^x\right)^2\,d\,\boldsymbol{a}\,d\,\boldsymbol{x}_{\text{unobs}}\Bigg\}
\end{aligned}
$$

$$
\begin{aligned}
&\propto\Bigg[\int_{\mathbb{R}_{\geq 0}}\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{a_{ik}=0}^{1}\int_{\mathbb{R}^{n_{\text{unobs}}}}\Bigg\{\sum_{k=1}^{K}\begin{bmatrix}\log\left\{\left(\sigma_k^x\right)^2\right\}\\ 1/\left(\sigma_k^x\right)^2\end{bmatrix}^{\top}\Big(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\rightarrow(\sigma_k^x)^2} \\
&\quad+\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\rightarrow(\sigma_k^x)^2}\Big)\Bigg\}\times\Bigg\{\sum_{i=1}^{n}\sum_{k=1}^{K}a_{ik}\left(\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\rightarrow a_{ik}}+\eta_{p(a_{ik}|\omega_k)\rightarrow a_{ik}}\right)\Bigg\} \\
&\quad\times\Bigg\{\sum_{i=1}^{n_{\text{unobs}}}\begin{bmatrix}x_{\text{unobs,i}}\\ x_{\text{unobs,i}}^2\end{bmatrix}^{\top}\Big(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\rightarrow x_{\text{unobs,i}}}+\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\rightarrow x_{\text{unobs,i}}} \\
&\quad+\boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\rightarrow x_{\text{unobs,i}}}\Big)\Bigg\}\times\Bigg\{\sum_{i=1}^{n}\sum_{k=1}^{K}\begin{bmatrix}\mu_k^x\\ \left(\mu_k^x\right)^2\end{bmatrix}^{\top}\begin{bmatrix}a_{ik}\left(x_i/\left(\sigma_k^x\right)^2\right)\\ -\frac{1}{2}a_{ik}/\left(\sigma_k^x\right)^2\end{bmatrix}\Bigg\} \\
&\quad d\left(\boldsymbol{\sigma}^x\right)^2\,d\,\boldsymbol{a}\,d\,\boldsymbol{x}_{\text{unobs}}\Bigg]
\end{aligned}
$$

$$
\propto \exp\left\{ \sum_{i=1}^{n}\sum_{k=1}^{K} \begin{bmatrix} \mu_k^x \\ \left(\mu_k^x\right)^2 \end{bmatrix}^\top \right.
$$
$$
\left. \times \begin{bmatrix}
\sum\limits_{a_{ik}=0}^{1} a_{ik} \times a_{ik} \times \left(\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\} \to a_{ik}} + \eta_{p(a_{ik}|\omega_k)\to a_{ik}} \right) \\
\times \int_{\mathbb{R}} x_i \, p_{\mathrm{N}_{\mathrm{nat}}}\left(x_i \,;\, \boldsymbol{\eta}_{p\left\{\boldsymbol{x}\,|\,\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to x_{\mathrm{unobs,i}}} \right. \\
\left. + \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to x_{\mathrm{unobs,i}}} + \boldsymbol{\eta}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to x_{\mathrm{unobs,i}}} \right) d\, x_{\mathrm{unobs,i}} \\
\times \int_{\mathbb{R}_{\geq 0}} \frac{1}{\left(\sigma_k^x\right)^2} \, p_{\mathrm{IG}_{\mathrm{nat}}}\left(\left(\sigma_k^x\right)^2 \,;\, \boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to(\sigma_k^x)^2} \right. \\
\left. + \boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to(\sigma_k^x)^2} \right) d\left(\sigma_k^x\right)^2 \\
\\
-\frac{1}{2}\sum\limits_{a_{ik}=0}^{1} a_{ik} \times a_{ik} \times \left(\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\} \to a_{ik}} + \eta_{p(a_{ik}|\omega_k)\to a_{ik}} \right) \\
\times \int_{\mathbb{R}_{\geq 0}} \frac{1}{\left(\sigma_k^x\right)^2} \, p_{\mathrm{IG}_{\mathrm{nat}}}\left(\left(\sigma_k^x\right)^2 \,;\, \boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to(\sigma_k^x)^2} \right. \\
\left. + \boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to(\sigma_k^x)^2} \right) d\left(\sigma_k^x\right)^2
\end{bmatrix} \right\}.
$$

Using Primitive 7.4.6 and Result 7.4.1, we get

$$
\propto \exp\left\{ \sum_{i=1}^{n}\sum_{k=1}^{K} \begin{bmatrix} \mu_k^x \\ \left(\mu_k^x\right)^2 \end{bmatrix}^\top \right.
$$
$$
\left. \begin{bmatrix}
-\frac{1}{2}\exp\left(\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\} \to a_{ik}} + \eta_{p(a_{ik}|\omega_k)\to a_{ik}} \right) \\
\times \left\{ E^{\,nat}_{q(\boldsymbol{x}_{\mathrm{unobs}})}\left(\boldsymbol{x}\right) \right\}_i \\
\times \dfrac{\left(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to(\sigma_k^x)^2} + \boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to(\sigma_k^x)^2} \right)_1^{+1}}{\left(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to(\sigma_k^x)^2} + \boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to(\sigma_k^x)^2} \right)_2} \\
\\
-\frac{1}{2}\exp\left(\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\} \to a_{ik}} + \eta_{p(a_{ik}|\omega_k)\to a_{ik}} \right) \\
\times \dfrac{\left(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to(\sigma_k^x)^2} + \boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to(\sigma_k^x)^2} \right)_1^{+1}}{\left(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to(\sigma_k^x)^2} + \boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to(\sigma_k^x)^2} \right)_2}
\end{bmatrix} \right\}
$$

where

$$
E^{\,nat}_{q(\boldsymbol{x}_{\mathrm{unobs}})}\left(\boldsymbol{x}\right) \equiv \begin{bmatrix} \boldsymbol{x}_{\mathrm{obs}} \\ -\frac{1}{2}\left\{ \mathrm{vec}^{-1}\left(\eta^{\circledast}_{2,\mathrm{vec}} \right) \right\}^{-1} \eta^{\circledast}_{1} \end{bmatrix}
$$

therefore the corresponding natural parameter vector is

$$\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to\mu_k^x} \leftarrow \begin{bmatrix} -\frac{1}{2}\exp\left(\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to a_{ik}}+\eta_{p(a_{ik}|\omega_k)\to a_{ik}}\right) \\ \times\left\{E^{\,nat}_{q(\boldsymbol{x}_{\text{unobs}})}(\boldsymbol{x})\right\}_i \\ \times\dfrac{\left(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to(\sigma_k^x)^2}+\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to(\sigma_k^x)^2}\right)_1^{+1}}{\left(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to(\sigma_k^x)^2}+\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to(\sigma_k^x)^2}\right)_2} \\ -\frac{1}{2}\exp\left(\eta_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to a_{ik}}+\eta_{p(a_{ik}|\omega_k)\to a_{ik}}\right) \\ \times\dfrac{\left(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to(\sigma_k^x)^2}+\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to(\sigma_k^x)^2}\right)_1^{+1}}{\left(\boldsymbol{\eta}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\boldsymbol{\sigma}^x)^2\right\}\to(\sigma_k^x)^2}+\boldsymbol{\eta}_{p\left\{(\sigma_k^x)^2|a_k^x\right\}\to(\sigma_k^x)^2}\right)_2} \end{bmatrix}.$$

Continuing on, we next consider

$$\begin{aligned} &\boldsymbol{m}_{\,p(\boldsymbol{y}\,|\,\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\boldsymbol{x}_{\text{unobs}}}(\boldsymbol{x}_{\text{unobs}}) \\ &\quad\leftarrow \exp\left\{\int_{\mathbb{R}^3}\int_{\mathbb{R}_{\geq 0}}\frac{1}{Z}\boldsymbol{m}_{\,\boldsymbol{\beta}\to p(\boldsymbol{y}\,|\,\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)}(\boldsymbol{\beta})\times\boldsymbol{m}_{\,p(\boldsymbol{y}\,|\,\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\boldsymbol{\beta}}(\boldsymbol{\beta})\right. \\ &\qquad\times\boldsymbol{m}_{\,\sigma_\varepsilon^2\to p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)}\left(\sigma_\varepsilon^2\right)\times\boldsymbol{m}_{\,p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\sigma_\varepsilon^2}\left(\sigma_\varepsilon^2\right) \\ &\qquad\left.\times\log p(\boldsymbol{y}\,|\,\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\,d\boldsymbol{\beta}\,d\sigma_\varepsilon^2\right\} \\ &\quad\propto \exp\left[\int_{\mathbb{R}^3}\int_{\mathbb{R}_{\geq 0}}\left\{\begin{bmatrix}\boldsymbol{\beta}\\ \text{vec}\left(\boldsymbol{\beta}\boldsymbol{\beta}^\top\right)\end{bmatrix}^\top\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\boldsymbol{\beta}}+\boldsymbol{\eta}_{p(\boldsymbol{\beta})\to\boldsymbol{\beta}}\right)\right\}\right. \\ &\qquad\times\left\{\begin{bmatrix}\log\left(\sigma_\varepsilon^2\right)\\ 1/\sigma_\varepsilon^2\end{bmatrix}^\top\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\sigma_\varepsilon^2}+\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to\sigma_\varepsilon^2}\right)\right\} \\ &\qquad\left.\times\sum_{i=1}^{n_{\text{unobs}}}\begin{bmatrix}x_{\text{unobs,i}}\\ x_{\text{unobs,i}}^2\end{bmatrix}^\top\begin{bmatrix}\frac{1}{\sigma_\varepsilon^2}\left(\beta_x\boldsymbol{y}_{x\text{unobs},i}-\beta_0\beta_x-\beta_c\beta_x\boldsymbol{c}_{x\text{unobs},i}\right)\\ -\frac{\beta_x}{2\sigma_\varepsilon^2}\end{bmatrix}d\boldsymbol{\beta}\,d\sigma_\varepsilon^2\right] \end{aligned}$$

$$= \exp\left\{ \sum_{i=1}^{n_{\text{unobs}}} \begin{bmatrix} x_{\text{unobs,i}} \\ x_{\text{unobs,i}}^2 \end{bmatrix}^\top \times \begin{bmatrix} \int_{\mathbb{R}_{\geq 0}} \frac{1}{\sigma_\varepsilon^2}\, p_{\text{IG}_{\text{nat}}}\left(\sigma_\varepsilon^2 \,;\, \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\sigma_\varepsilon^2} + \boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to\sigma_\varepsilon^2}\right) d\sigma_\varepsilon^2 \\ \times \left\{ \boldsymbol{y}_{x\text{unobs},i} \int_{\mathbb{R}} \beta_x p_{\text{N}_{\text{nat}}}\left(\beta_x \,;\, \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\beta_x} + \boldsymbol{\eta}_{p(\beta_x)\to\beta_x}\right) d\beta_x \right. \\ - \int_{\mathbb{R}^2} \beta_0\beta_x p_{\text{N}_{\text{nat}}}\left(\beta_0,\beta_x \,;\, \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\beta_0} + \boldsymbol{\eta}_{p(\beta_0)\to\beta_0},\right. \\ \left. \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\beta_x} + \boldsymbol{\eta}_{p(\beta_x)\to\beta_x}\right) d\beta_0\, d\beta_x \\ - \boldsymbol{c}_{x\text{unobs},i} \int_{\mathbb{R}^2} \beta_c\beta_x p_{\text{N}_{\text{nat}}}\left(\beta_c,\beta_x \,;\, \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\beta_c} + \boldsymbol{\eta}_{p(\beta_c)\to\beta_c},\right. \\ \left. \left. \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\beta_x} + \boldsymbol{\eta}_{p(\beta_x)\to\beta_x}\right) d\beta_c\, d\beta_x \right\} \\ \\ -\frac{1}{2} \int_{\mathbb{R}_{\geq 0}} \frac{1}{\sigma_\varepsilon^2}\, p_{\text{IG}_{\text{nat}}}\left(\sigma_\varepsilon^2 \,;\, \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\sigma_\varepsilon^2} + \boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to\sigma_\varepsilon^2}\right) d\sigma_\varepsilon^2 \\ \times \int_{\mathbb{R}} \beta_x p_{\text{IG}_{\text{nat}}}\left(\beta_x \,;\, \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\beta_x} + \boldsymbol{\eta}_{p(\beta_x)\to\beta_x}\right) d\beta_x \end{bmatrix} \right\}.$$

With the use of Primitives 7.4.5 and 7.4.6, this gives

$$\boldsymbol{m}_{\,p(\boldsymbol{y}\,|\,\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\boldsymbol{x}_{\text{unobs}}}(\boldsymbol{x}_{\text{unobs}})$$

$$\propto \exp\left\{ \sum_{i=1}^{n_{\text{unobs}}} \begin{bmatrix} x_{\text{unobs,i}} \\ x_{\text{unobs,i}}^2 \end{bmatrix}^\top \begin{bmatrix} \dfrac{\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\sigma_\varepsilon^2} + \boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to\sigma_\varepsilon^2}\right)_1 + 1}{\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\sigma_\varepsilon^2} + \boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to\sigma_\varepsilon^2}\right)_2} \\ \times \left\{ -\frac{1}{2}\boldsymbol{y}_{x\text{unobs},i} \dfrac{\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\beta_x} + \boldsymbol{\eta}_{p(\beta_x)\to\beta_x}\right)_1}{\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\beta_x} + \boldsymbol{\eta}_{p(\beta_x)\to\beta_x}\right)_2} \right. \\ -\mathcal{J}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\boldsymbol{\beta}} + \boldsymbol{\eta}_{p(\boldsymbol{\beta})\to\boldsymbol{\beta}} \,;\, 1,\, 3\right) \\ \left. -\boldsymbol{c}_{x\text{unobs},i}\mathcal{J}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\boldsymbol{\beta}} + \boldsymbol{\eta}_{p(\boldsymbol{\beta})\to\boldsymbol{\beta}} \,;\, 2,\, 3\right)\right\} \\ \\ \dfrac{1}{4} \dfrac{\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\sigma_\varepsilon^2} + \boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to\sigma_\varepsilon^2}\right)_1 + 1}{\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\sigma_\varepsilon^2} + \boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to\sigma_\varepsilon^2}\right)_2} \dfrac{\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\beta_x} + \boldsymbol{\eta}_{p(\beta_x)\to\beta_x}\right)_1}{\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\beta_x} + \boldsymbol{\eta}_{p(\beta_x)\to\beta_x}\right)_2} \end{bmatrix} \right\}$$

and so the corresponding natural parameter update is

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to x_{\text{unobs,i}}} \leftarrow$$

$$\left[\begin{array}{c}
\frac{\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\sigma_\varepsilon^2}+\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to\sigma_\varepsilon^2}\right)_1+1}{\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\sigma_\varepsilon^2}+\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to\sigma_\varepsilon^2}\right)_2} \\
\times\left\{\begin{array}{l}
-\frac{1}{2}\boldsymbol{y}_{x_{\text{unobs},i}}\frac{\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\beta_x}+\boldsymbol{\eta}_{p(\beta_x)\to\beta_x}\right)_1}{\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\beta_x}+\boldsymbol{\eta}_{p(\beta_x)\to\beta_x}\right)_2} \\
-\mathcal{J}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\boldsymbol{\beta}}+\boldsymbol{\eta}_{p(\boldsymbol{\beta})\to\boldsymbol{\beta}}\,;\,1,\,3\right) \\
-\boldsymbol{c}_{x_{\text{unobs},i}}\mathcal{J}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\boldsymbol{\beta}}+\boldsymbol{\eta}_{p(\boldsymbol{\beta})\to\boldsymbol{\beta}}\,;\,2,\,3\right)
\end{array}\right\} \\
\frac{1}{4}\frac{\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\sigma_\varepsilon^2}+\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to\sigma_\varepsilon^2}\right)_1+1}{\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\sigma_\varepsilon^2}+\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to\sigma_\varepsilon^2}\right)_2} \\
\times\frac{\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\beta_x}+\boldsymbol{\eta}_{p(\beta_x)\to\beta_x}\right)_1}{\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\beta_x}+\boldsymbol{\eta}_{p(\beta_x)\to\beta_x}\right)_2}
\end{array}\right].$$

Next we consider the message from factor $p(\boldsymbol{y}\,|\,\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)$ to stochastic node $\boldsymbol{\beta}$. The derivation for this message is similar to that of $\boldsymbol{m}_{\,p(\boldsymbol{o}\,|\,\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\to\boldsymbol{\alpha}}(\boldsymbol{\alpha})$ and has the form:

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\boldsymbol{\beta}} \leftarrow \left[\begin{array}{c}
\frac{\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\sigma_\varepsilon^2}+\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to\sigma_\varepsilon^2}\right)_1+1}{\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\sigma_\varepsilon^2}+\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to\sigma_\varepsilon^2}\right)_2} \\
\quad\times E^{\text{nat}}_{q(\boldsymbol{x}_{\text{unobs}})}(\boldsymbol{X})^\top\boldsymbol{y} \\
-\frac{\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\sigma_\varepsilon^2}+\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to\sigma_\varepsilon^2}\right)_1+1}{2\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\to\sigma_\varepsilon^2}+\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\to\sigma_\varepsilon^2}\right)_2} \\
\quad\times\text{vec}\left\{E^{\text{nat}}_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)\right\}
\end{array}\right],$$

where

$$E^{\text{nat}}_{q(\boldsymbol{x}_{\text{unobs}})}(\boldsymbol{X}) = \left[\begin{array}{ccc}
\mathbf{1}_{n_{\text{obs}}} & \boldsymbol{c}_{x_{\text{obs}}} & \boldsymbol{x}_{\text{obs}} \\
\mathbf{1}_{n_{\text{unobs}}} & \boldsymbol{c}_{x_{\text{unobs}}} & -\frac{1}{2}\left\{\text{vec}^{-1}\left(\eta^{\circledast}_{2,\text{vec}}\right)\right\}^{-1}\eta^{\circledast}_1
\end{array}\right]$$

and

$$E^{\text{nat}}_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{X}^\top\boldsymbol{X}\right) =$$

$$\begin{bmatrix} n & \mathbf{1}_n^\top \boldsymbol{c} & \begin{array}{c}\mathbf{1}_{n_{\text{obs}}}^\top \boldsymbol{x}_{\text{obs}} \\ -\frac{1}{2}\mathbf{1}_{n_{\text{unobs}}}^\top \left\{\text{vec}^{-1}\left(\eta_{2,\text{vec}}^{\circledast}\right)\right\}^{-1} \\ \times \eta_1^{\circledast}\end{array} \\ \hline \mathbf{1}_n^\top \boldsymbol{c} & \|\boldsymbol{c}\|^2 & \begin{array}{c}\boldsymbol{c}_{x\text{obs}}^\top \boldsymbol{x}_{\text{obs}} \\ -\frac{1}{2}\boldsymbol{c}_{x\text{unobs}}^\top \left\{\text{vec}^{-1}\left(\eta_{2,\text{vec}}^{\circledast}\right)\right\}^{-1} \\ \times \eta_1^{\circledast}\end{array} \\ \hline \begin{array}{c}\mathbf{1}_{n_{\text{obs}}}^\top \boldsymbol{x}_{\text{obs}} \\ -\frac{1}{2}\mathbf{1}_{n_{\text{unobs}}}^\top \left\{\text{vec}^{-1}\left(\eta_{2,\text{vec}}^{\circledast}\right)\right\}^{-1} \\ \times \eta_1^{\circledast}\end{array} & \begin{array}{c}\boldsymbol{c}_{x\text{obs}}^\top \boldsymbol{x}_{\text{obs}} \\ -\frac{1}{2}\boldsymbol{c}_{x\text{unobs}}^\top \left\{\text{vec}^{-1}\left(\eta_{2,\text{vec}}^{\circledast}\right)\right\}^{-1} \\ \times \eta_1^{\circledast}\end{array} & \begin{array}{c}\|\boldsymbol{x}_{\text{obs}}\|^2 + \\ \| - \frac{1}{2}\left\{\text{vec}^{-1}\left(\eta_{2,\text{vec}}^{\circledast}\right)\right\}^{-1}\eta_1^{\circledast}\|^2 \\ -\frac{1}{2}\left\{\text{vec}^{-1}\left(\eta_{2,\text{vec}}^{\circledast}\right)\right\}^{-1}\end{array} \end{bmatrix}.$$

The last four messages we consider are $\boldsymbol{m}_{\,p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\,\to\,\sigma_\varepsilon^2}\left(\sigma_\varepsilon^2\right)$, $\boldsymbol{m}_{\,p(\sigma_\varepsilon^2|a_\varepsilon)\,\to\,\sigma_\varepsilon^2}\left(\sigma_\varepsilon^2\right)$, $\boldsymbol{m}_{\,p(\sigma_\varepsilon^2|a_\varepsilon)\,\to\,a_\varepsilon}\left(a_\varepsilon\right)$ and $\boldsymbol{m}_{\,p(a_\varepsilon)\,\to\,a_\varepsilon}\left(a_\varepsilon\right)$. The derivation of these messages is similar to that shown for $\boldsymbol{m}_{\,p(\boldsymbol{o}|\,\boldsymbol{\alpha},\boldsymbol{x},\sigma_o^2)\,\to\,\sigma_o^2}\left(\sigma_o^2\right)$, $\boldsymbol{m}_{\,p(\sigma_o^2|a_o)\,\to\,\sigma_o^2}\left(\sigma_o^2\right)$, $\boldsymbol{m}_{\,p(\sigma_o^2|a_o)\,\to\,a_o}\left(a_o\right)$ and $\boldsymbol{m}_{\,p(a_o)\,\to\,a_o}\left(a_o\right)$, respectively. Consequently, we only give their corresponding natural parameter updates:

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\,\to\,\sigma_\varepsilon^2}\left(\sigma_\varepsilon^2\right) \leftarrow \begin{bmatrix} -n/2 \\ -\frac{1}{2}E^{\text{nat}}_{q(\boldsymbol{x}_{\text{unobs}})}\left(\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2\right) \end{bmatrix},$$

$$\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\,\to\,\sigma_\varepsilon^2} \leftarrow \begin{bmatrix} -3/2 \\ -\dfrac{\left(\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\,\to\,a_\varepsilon} + \boldsymbol{\eta}_{p(a_\varepsilon)\,\to\,a_\varepsilon}\right)_1 + 1}{\left(\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\,\to\,a_\varepsilon} + \boldsymbol{\eta}_{p(a_\varepsilon)\,\to\,a_\varepsilon}\right)_2} \end{bmatrix},$$

$$\boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\,\to\,a_\varepsilon} \leftarrow \begin{bmatrix} -3/2 \\ -\dfrac{\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\,\to\,\sigma_\varepsilon^2} + \boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\,\to\,\sigma_\varepsilon^2}\right)_1 + 1}{\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma_\varepsilon^2)\,\to\,\sigma_\varepsilon^2} + \boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\,\to\,\sigma_\varepsilon^2}\right)_2} \end{bmatrix},$$

$$\boldsymbol{\eta}_{p(a_\varepsilon)\,\to\,a_\varepsilon} \leftarrow \begin{bmatrix} -3/2 \\ -1/A_\varepsilon^2 \end{bmatrix},$$

where

$$\begin{aligned} &E^{\text{nat}}_{q(\boldsymbol{x}_{\text{unobs}})}\left(\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2\right) = \\ &\left\|\boldsymbol{y}_{x\text{obs}} + \tfrac{1}{2}\boldsymbol{X}_{x\text{obs}}\left\{\text{vec}^{-1}\left(\eta_{2,\text{vec}}^{\blacklozenge}\right)\right\}^{-1}\eta_1^{\blacklozenge}\right\|^2 - \tfrac{1}{2}\text{tr}\left[\boldsymbol{X}_{x\text{obs}}^\top\boldsymbol{X}_{x\text{obs}}\left\{\text{vec}^{-1}\left(\eta_{2,\text{vec}}^{\blacklozenge}\right)\right\}^{-1}\right] \\ &+\|\boldsymbol{y}_{x\text{unobs}}\|^2 + \boldsymbol{y}_{x\text{unobs}}^\top E^{\text{nat}}_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{X}_{x\text{unobs}}\right)\left\{\text{vec}^{-1}\left(\eta_{2,\text{vec}}^{\blacklozenge}\right)\right\}^{-1}\eta_1^{\blacklozenge} \\ &+\tfrac{1}{4}\text{tr}\left\{E^{\text{nat}}_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{X}_{x\text{unobs}}^\top\boldsymbol{X}_{x\text{unobs}}\right)\left\{\text{vec}^{-1}\left(\eta_{2,\text{vec}}^{\blacklozenge}\right)\right\}^{-1}\left[\eta_1^{\blacklozenge}\left(\eta_1^{\blacklozenge}\right)^\top\left\{\text{vec}^{-1}\left(\eta_{2,\text{vec}}^{\blacklozenge}\right)\right\}^{-1}\right.\right. \\ &\left.\left.-2\boldsymbol{I}\right]\right\}, \end{aligned}$$

$$E^{\text{nat}}_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{X}_{x\text{unobs}}\right) = \left[\mathbf{1}_{n_{\text{unobs}}} \quad \boldsymbol{c}_{x\text{unobs}} \quad -\tfrac{1}{2}\left\{\text{vec}^{-1}\left(\eta^{\circledast}_{2,\text{vec}}\right)\right\}^{-1}\eta^{\circledast}_{1}\right],$$

$$E^{\text{nat}}_{q(\boldsymbol{x}_{\text{unobs}})}\left(\boldsymbol{X}^{\top}_{x\text{unobs}}\boldsymbol{X}_{x\text{unobs}}\right) =$$

$$\left[\begin{array}{c|c|c} n_{\text{unobs}} & \mathbf{1}^{\top}_{n_{\text{unobs}}}\boldsymbol{c}_{x\text{unobs}} & -\tfrac{1}{2}\mathbf{1}^{\top}_{n_{\text{unobs}}}\left\{\text{vec}^{-1}\left(\eta^{\circledast}_{2,\text{vec}}\right)\right\}^{-1} \times\eta^{\circledast}_{1} \\ \hline \mathbf{1}^{\top}_{n_{\text{unobs}}}\boldsymbol{c}_{x\text{unobs}} & \|\boldsymbol{c}_{x\text{unobs}}\|^2 & -\tfrac{1}{2}\boldsymbol{c}^{\top}_{x\text{unobs}}\left\{\text{vec}^{-1}\left(\eta^{\circledast}_{2,\text{vec}}\right)\right\}^{-1} \times\eta^{\circledast}_{1} \\ \hline -\tfrac{1}{2}\mathbf{1}^{\top}_{n_{\text{unobs}}}\left\{\text{vec}^{-1}\left(\eta^{\circledast}_{2,\text{vec}}\right)\right\}^{-1} \times\eta^{\circledast}_{1} & -\tfrac{1}{2}\boldsymbol{c}^{\top}_{x\text{unobs}}\left\{\text{vec}^{-1}\left(\eta^{\circledast}_{2,\text{vec}}\right)\right\}^{-1} \times\eta^{\circledast}_{1} & \|-\tfrac{1}{2}\left\{\text{vec}^{-1}\left(\eta^{\circledast}_{2,\text{vec}}\right)\right\}^{-1}\eta^{\circledast}_{1}\|^2 \\ & & -\tfrac{1}{2}\left\{\text{vec}^{-1}\left(\eta^{\circledast}_{2,\text{vec}}\right)\right\}^{-1} \end{array}\right],$$

and $\boldsymbol{\eta}^{\blacklozenge} \equiv \left[\eta^{\blacklozenge}_{1} \ \eta^{\blacklozenge}_{2,\text{vec}}\right]^{\top} \equiv \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma^2_{\varepsilon})\to\boldsymbol{\beta}} + \boldsymbol{\eta}_{p(\boldsymbol{\beta})\to\boldsymbol{\beta}}$.

7.B.4 Step 3 : Optimal q-densities

Once convergence is achieved from Steps 2(a) and (b) we get the optimal q-densities of each stochastic node through the following updates:

$$\boldsymbol{\eta}^{*}_{q(\boldsymbol{\alpha})} \leftarrow \boldsymbol{\eta}^{*}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma^2_o)\to\boldsymbol{\alpha}} + \boldsymbol{\eta}^{*}_{p(\boldsymbol{\alpha})\to\boldsymbol{\alpha}}$$
$$\boldsymbol{\eta}^{*}_{q(\sigma^2_o)} \leftarrow \boldsymbol{\eta}^{*}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma^2_o)\to\sigma^2_o} + \boldsymbol{\eta}^{*}_{p(\sigma^2_o|a_o)\to\sigma^2_o}$$
$$\boldsymbol{\eta}^{*}_{q(a_o)} \leftarrow \boldsymbol{\eta}^{*}_{p(\sigma^2_o|a_o)\to a_o} + \boldsymbol{\eta}^{*}_{p(a_o)\to a_o}.$$

For $1 \le k \le K$:

$$\boldsymbol{\eta}^{*}_{q(a^x_k)} \leftarrow \boldsymbol{\eta}^{*}_{p(a^x_k)\to a^x_k} + \boldsymbol{\eta}^{*}_{p\left\{(\sigma^x_k)^2|a^x_k\right\}\to a^x_k}$$
$$\boldsymbol{\eta}^{*}_{q\left\{(\sigma^x_k)^2\right\}} \leftarrow \boldsymbol{\eta}^{*}_{p\left\{(\sigma^x_k)^2|a^x_k\right\}\to(\sigma^x_k)^2} + \boldsymbol{\eta}^{*}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to(\sigma^x_k)^2}$$
$$\boldsymbol{\eta}^{*}_{q(\mu^x_k)} \leftarrow \boldsymbol{\eta}^{*}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to\mu^x_k} + \boldsymbol{\eta}^{*}_{p(\mu^x_k)\to\mu^x_k}.$$

For $1 \le i \le n$ and $1 \le k \le K$:

$$\boldsymbol{\eta}^{*}_{q(a_{ik})} \leftarrow \eta^{*}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to a_{ik}} + \eta^{*}_{p(a_{ik}|\omega_k)\to a_{ik}}$$
$$\boldsymbol{\eta}^{*}_{q(\omega_k)} \leftarrow \eta^{*}_{p(a_{ik}|\omega_k)\to\omega_k} + \eta^{*}_{p(\omega_k)\to\omega_k}$$

$$\boldsymbol{\eta}^{*}_{q(\boldsymbol{\beta})} \leftarrow \boldsymbol{\eta}^{*}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma^2_{\varepsilon})\to\boldsymbol{\beta}} + \boldsymbol{\eta}^{*}_{p(\boldsymbol{\beta})\to\boldsymbol{\beta}}$$
$$\boldsymbol{\eta}^{*}_{q(\sigma^2_{\varepsilon})} \leftarrow \boldsymbol{\eta}^{*}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma^2_{\varepsilon})\to\sigma^2_{\varepsilon}} + \boldsymbol{\eta}^{*}_{p(\sigma^2_{\varepsilon}|a_{\varepsilon})\to\sigma^2_{\varepsilon}}$$
$$\boldsymbol{\eta}^{*}_{q(a_{\varepsilon})} \leftarrow \boldsymbol{\eta}^{*}_{p(\sigma^2_{\varepsilon}|a_{\varepsilon})\to a_{\varepsilon}} + \boldsymbol{\eta}^{*}_{p(a_{\varepsilon})\to a_{\varepsilon}}.$$

For $1 \le i \le n_{\text{unobs}}$:

$$\boldsymbol{\eta}^{*}_{q(x_{\text{unobs,i}})} \leftarrow \boldsymbol{\eta}^{*}_{p\left\{\boldsymbol{x}|\boldsymbol{a},\boldsymbol{\mu}^x,(\sigma^x)^2\right\}\to x_{\text{unobs,i}}} + \boldsymbol{\eta}^{*}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x},\sigma^2_{\varepsilon})\to x_{\text{unobs,i}}} + \boldsymbol{\eta}^{*}_{p(\boldsymbol{o}|\boldsymbol{\alpha},\boldsymbol{x},\sigma^2_o)\to x_{\text{unobs,i}}}.$$

Chapter 8

Conclusion

This thesis has aimed to increase the body of algorithms and results on MFVB inference in heteroscedastic and longitudinal regression. It has helped emphasise the importance of MFVB in both semiparametric and nonparametric regression settings and has demonstrated that MFVB provides a fast deterministic alternative to the stochastic MCMC with often little degradation in accuracy. Overall, we have developed extensions of previously used MFVB methodology for some model settings, whereas in others, we have developed original MFVB algorithms which had not previously been considered. Even though the simulations and real datasets used in this thesis have not exceeded over 10 megabytes in size, the idea behind the construction of these fast algorithms is to apply such methodology to larger and even more complex datasets that require further storage than what is available on a standard computer.

In Chapter 2, we extended the work of Al Kadiri *et al.* (2010) to obtain an MFVB algorithm catered to the marginal longitudinal semiparametric regression setting. This resulted in faster inference for the nutritional epidemiology study and it was clear that MFVB estimation of the model parameters in a simulation study was very high.

Chapter 3 considered the setting involving heteroscedastic semiparametric regression. This led to a modification of ordinary MFVB, called non-conjugate variational message passing which aids the incorporation of a heteroscedastic component without the need to stray from closed form algebraic expressions. Overall, the accuracy of the mean functions is very high and the accuracy of the variance functions, whilst not as accurate as the mean, are still producing considerably favorable accuracy scores.

In Chapter 4, we looked into the development of MFVB algorithms for three extensions

of the heteroscedastic setting considered in Chapter 3. We explore a real-time, bivariate predictor and additive model analogue of Chapter 3. Through the use of real-data examples, simulation studies and a real-time movie, we saw the time benefits of using such an algorithm.

Chapter 5 looks into the development of an MFVB algorithm for subject-specific curve models making use of longitudinal and multilevel structures. This chapter takes advantage of the structure of the model and looks into streamlining a naïve implementation of an MFVB algorithm, resulting in even faster inference.

In Chapter 6, we derived a fast MFVB algorithm to fitting a linear regression model with classical measurement error. The time savings here were evident, however extension of such methodology to arbitrarily large models involving measurement error would show time savings of higher magnitudes.

Chapter 7 discussed VMP, an alternative to MFVB, which produces the same approximation but with a different algebraic system. There, we derived the VMP updates for two models used throughout the thesis and provided the details of the methodology behind VMP by presenting a general VMP scheme.

It is generally not known ahead of time whether the MFVB approximation will produce an acceptable level of accuracy. The accuracy of MFVB varies depending on the type of application. As exhibited in this thesis, MFVB estimation of mean curves is very good but MFVB credible sets are generally underestimated. However, even with the shortcomings of poor estimation of credible sets, the speed and computational gains afforded by MFVB make it a good choice for complex models combined with large datasets.

Each chapter has shed light on the overall effectiveness of MFVB inference as a fast alternative to MCMC when time and/or storage is of concern. We have seen that the accuracy of MFVB varies depending on many details involved in the type of application. As indicated by the variety of models used, MFVB is of no doubt a dynamic area of statistical research where the potential to further contributions is profound.

Bibliography

Aitkin, M. & Rocci, R. (2002). A general maximum likelihood analysis of measurement error in generalized linear models. *Statistics and Computing*, **12**, no. 2, 163–174.

Al Kadiri, M., Carroll, R. & Wand, M. (2010). Marginal longitudinal semiparametric regression via penalized splines. *Statistics & probability letters*, **80**, no. 15, 1242–1252.

Bishop, C. M. (2006). *Pattern recognition and machine learning*, volume 1. springer New York.

Bishop, C. M., Spiegelhalter, D. J. & Winn, J. (2003). *VIBES: A variational inference engine for Bayesian networks*. In advances in Neural Information Processing Systems, eds S. Becker, S. Thrun and K. Obermayer. Cambridge, MA: MIT Press.

Braun, M. & McAuliffe, J. (2010). Variational inference for large-scale models of discrete choice. *Journal of the American Statistical Association*, **105**, no. 489, 324–335.

Breiman, L. & Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, **80**, no. 391, 580–598.

Bugbee, B. D., Breidt, F. J. & van der Woerd, M. J. (2015). Laplace variational approximation for semiparametric regression in the presence of heteroskedastic errors. *Journal of Computational and Graphical Statistics*, in press.

Carroll, R., Gail, M. & Lubin, J. (1993). Case-control studies with errors in covariates. *Journal of the American Statistical Association*, **88**, no. 421, 185–199.

Carroll, R. J., Delaigle, A. & Hall, P. (2007). Non-parametric regression estimation from data contaminated by a mixture of berkson and classical errors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**, no. 5, 859–878.

Carroll, R. J., Ruppert, D., Crainiceanu, C. M., Tosteson, T. D. & Karagas, M. R. (2004). Nonlinear and nonparametric regression and instrumental variables. *Journal of the American Statistical Association*, **99**, no. 467.

Carroll, R. J., Ruppert, D., Stefanski, L. A. & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. CRC press.

Cressie, N. (1993). Statistics for spatial data: Wiley series in probability and statistics.

Diggle, P., Heagerty, P., Liang, K.-Y. & Zeger, S. (2002). *Analysis of longitudinal data*. Oxford University Press.

Ezekiel, M. (1924). A method of handling curvilinear correlation for any number of variables. *Journal of the American Statistical Association*, **19**, no. 148, 431–453.

Faes, C., Ormerod, J. & Wand, M. (2011). Variational bayesian inference for parametric and nonparametric regression with missing data. *Journal of the American Statistical Association*, **106**, no. 495.

Fitzmaurice, G. M., Laird, N. M. & Ware, J. H. (2012). *Applied longitudinal analysis*, volume 998. John Wiley & Sons.

Friedman, J. H. & Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American statistical Association*, **76**, no. 376, 817–823.

Ganguli, B., Staudenmayer, J. & Wand, M. (2005). Additive models with predictors subject to measurement error. *Australian & New Zealand Journal of Statistics*, **47**, no. 2, 193–202.

Gelman, A. & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Goldstein, H. (2011). *Multilevel statistical models*, volume 922. John Wiley & Sons.

Green, P. & Silverman, B. (1994). Nonparametric regression and generalized linear models, vol. 58 of. *Monographs on Statistics and Applied Probability*.

Hall, P., Pham, T., Wand, M. P., Wang, S. S. *et al.* (2011). Asymptotic normality and valid inference for gaussian variational approximation. *The Annals of Statistics*, **39**, no. 5, 2502–2532.

Hoffman, M., Bach, F. R. & Blei, D. M. (2010). Online learning for latent dirichlet allocation. In *advances in neural information processing systems*. pages 856–864.

Huang, A., Wand, M. *et al.* (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis*, **8**, no. 2, 439–452.

Jordan, M. I. (2004). Graphical models. *Statistical Science*, 140–155.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, **37**, no. 2, 183–233.

Knowles, D. A. & Minka, T. (2011). Non-conjugate variational message passing for multinomial and binary regression. In *Advances in Neural Information Processing Systems*. pages 1701–1709.

Lauritzen, S. L. (1996). *Graphical models*. Oxford University Press.

Lázaro-Gredilla, M. & Titsias, M. (2011). Variational heteroscedastic gaussian process regression. *Proceedings of the 28th International Conference on Machine Learning*.

Lee, C. Y. & Wand, M. (2015). Streamlined mean field variational bayes for longitudinal and multilevel data analysis. *Unpublished manuscript*.

Li, Y. & Ruppert, D. (2008). On the asymptotics of penalized splines. *Biometrika*, **95**, no. 2, 415–436.

Liang, H., Wu, H. & Carroll, R. J. (2003). The relationship between virologic and immunologic responses in aids clinical research using mixed-effects varying-coefficient models with measurement error. *Biostatistics*, **4**, no. 2, 297–312.

Ligges, U., Thomas, A., Spiegelhalter, D., Best, N. G., Lunn, D., Rice, K. & Strutz, S. (2009). *BRugs 0.5: OpenBUGS and Its R/S-PLUS Interface BRugs*. `http://www.stats.ox.ac.uk/pub/RWin/src/contrib/`.

Lin, X. & Carroll, R. J. (2001). Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association*, **96**, no. 455.

Luenberger, D. G. & Ye, Y. (2008). *Linear and nonlinear programming*, volume 116. Springer.

Luts, J., Broderick, T. & Wand, M. P. (2013). Real-time semiparametric regression. *Journal of Computational and Graphical Statistics*, , no. just-accepted.

Luts, J., Wang, S., Ormerod, J. & Wand, M. (2014). Semiparametric regression analysis via infer. net.

Magnus, J. & Neudecker, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics (revised ed.)*. New York: Wiley.

Mallick, B., Hoffman, F. O. & Carroll, R. J. (2002). Semiparametric regression modeling with mixtures of berkson and classical error, with application to fallout from the nevada test site. *Biometrics*, **58**, no. 1, 13–20.

Mardia, K. V., Kent, J. T. & Bibby, J. M. (1979). *Multivariate analysis*. Academic press.

Marley, J. & Wand, M. (2010). Non-standard semiparametric regression via brugs. *Journal of Statistical Software*, **37**, no. 5.

McGrory, C. A. & Titterington, D. (2007). Variational approximations in bayesian model selection for finite mixture distributions. *Computational Statistics & Data Analysis*, **51**, no. 11, 5352–5367.

Minka, T. (2005). Divergence measures and message passing. Technical report, Technical report, Microsoft Research.

Minka, T. & Winn, J. (2009). Gates. In D. Koller, D. Schuurmans, Y. Bengio & L. Bottou, editors, *Advances in Neural Information Processing Systems 21*. Curran Associates, Inc., pages 1073–1080.

Minka, T., Winn, J., Guiver, G. & Kannan, A. (2008). *Infer.Net*. Microsoft Research Cambridge, Cambridge, UK.

Müller, P. & Roeder, K. (1997). A bayesian semiparametric model for case-control studies with errors in variables. *Biometrika*, **84**, no. 3, 523–537.

Neville, S. E., Ormerod, J. T. & Wand, M. P. (2014). Mean field variational bayes for continuous sparse signal shrinkage: Pitfalls and remedies. *Electronic Journal of Statistics*, **8**, no. 1, 1113–1151.

Nott, D. J., Tran, M.-N. & Leng, C. (2012). Variational approximation for heteroscedastic linear models and matching pursuit algorithms. *Statistics and Computing*, **22**, no. 2, 497–512.

Opper, M. & Archambeau, C. (2009). The variational gaussian approximation revisited. *Neural computation*, **21**, no. 3, 786–792.

Ormerod, J. & Wand, M. (2010). Explaining variational approximations. *The American Statistician*, **64**, no. 2, 140–153.

Parisi, G. (1988). *Statistical field theory*. Addison-Wesley.

Pebesma, E. J. (2004). Multivariate geostatistics in R: the `gstat` package. *Computers and Geosciences*, **30**, 683–691.

Pebesma, E. J. & Duin, R. N. M. (2005). Spatial patterns of temporal change in north sea sediment quality on different spatial scales. in p. renard, h. demougeot-renard and r. froidevaux (eds). *Geostatistics for Environmental Application: Proceedings of the Fifth European Conference on Geostatistics for Environmental Applications (pp. 367-378)*.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & The R Core Team (2009). *nlme 3.1: linear and nonlinear mixed effects models. R package.* `www.R-project.org`.

Pratt, J. H., Jones, J. J., Miller, J. Z., Wagner, M. A. & Fineberg, N. S. (1989). Racial differences in aldosterone excretion and plasma aldosterone concentrations in children. *New England Journal of Medicine*, **321**, no. 17, 1152–1157. doi:10.1056/NEJM198910263211703. PMID: 2677724.

R Core Team (2013). *R: a language and environment for statistical computing*. R foundation for Statistical Computing, Vienna, Austria, ISBN: 3-900051-07-0. `www.R-project.org`.

Richardson, S., Leblond, L., Jaussent, I. & 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**, no. 3, 549–566.

Rigby, R. & Stasinopoulos, D. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**, no. 3, 507–554.

Robinson, G. K. (1991). That blup is a good thing: The estimation of random effects. *Statistical science*, 15–32.

Roeder, K., Carroll, R. J. & Lindsay, B. G. (1996). A semiparametric mixture approach to case-control studies with errors in covariables. *Journal of the American Statistical Association*, **91**, no. 434, 722–732.

Ruppert, D., Wand, M. & Carroll, R. J. (2009). Semiparametric regression during 2003–2007. *Electronic Journal of Statistics*, **3**, 1193.

Ruppert, D., Wand, M. P. & Carroll, R. J. (2003). *Semiparametric regression*. New York: Cambridge University Press.

Schafer, D. W. (2001). Semiparametric maximum likelihood for measurement error model regression. *Biometrics*, **57**, no. 1, 53–61.

Simpson, D. (2013). Rodent tumor dataset monitoring frequency dependent backscatter coefficients. Department of Statistics, University of Illinois at Urbana-Champaign.

Speed, T. (1991). [that blup is a good thing: The estimation of random effects]: Comment. *Statistical Science*, 42–44.

Spiegelhalter, D. J., Thomas, A., Best, N. G., Gilks, W. R. & Lunn, D. (2003). *BUGS: Bayesian inference using Gibbs sampling*. Medical Research Council Biostatistics Unit, Cambridge, England. `http://www.mrc-bsu.cam.ac.uk/bugs`.

Stan Development Team (2014). *Stan: A C++ Library for Probability and Sampling*. Version 2.2. `http://mc-stan.org`.

Teschendorff, A. E., Wang, Y., Barbosa-Morais, N. L., Brenton, J. D. & Caldas, C. (2005). A variational bayesian mixture modelling framework for cluster analysis of gene-expression data. *Bioinformatics*, **21**, no. 13, 3025–3033.

Thomas, A., O'Hara, B., Ligges, U. & Strutz, S. (2006). *Making BUGS Open*. R News. `http://CRAN.R-project.org/doc/Rnews/`.

Titterington, D. (2004). Bayesian methods for neural networks and related models. *Statistical Science*, 128–139.

Wahba, G. (1990). *Spline models for observational data*, volume 59. Siam.

Wainwright, M. J. & Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, **1**, no. 1-2, 1–305.

Wand, M. (2009). Semiparametric regression and graphical models. *Australian & New Zealand Journal of Statistics*, **51**, no. 1, 9–41.

Wand, M. (2015). Fast approximate inference for arbitrarily large semiparametric regression models via message passing.

Wand, M. & Ormerod, J. (2008). On semiparametric regression with o'sullivan penalized splines. *Australian & New Zealand Journal of Statistics*, **50**, no. 2, 179–198.

Wand, M. & Ripley, B. (2009). Kernsmooth: Functions for kernel smoothing for wand and jones (1995)kernel smoothing. r package.

Wand, M. P. (2014). Fully simplified multivariate normal updates in non-conjugate variational message passing. *Journal of Machine Learning Research*, **15**, 1351–1369.

Wand, M. P., Ormerod, J. T., Padoan, S. A., Fuhrwirth, R. *et al.* (2011). Mean field variational bayes for elaborate distributions. *Bayesian Analysis*, **6**, no. 4, 847–900.

Wang, B. & Titterington, D. (2003). Local convergence of variational bayes estimators for mixing coefficients. *Preprint*.

Wang, B., Titterington, D. *et al.* (2006). Convergence properties of a general algorithm for calculating variational bayesian estimates for a normal mixture model. *Bayesian Analysis*, **1**, no. 3, 625–650.

Wang, C., Paisley, J. W. & Blei, D. M. (2011). Online variational inference for the hierarchical dirichlet process. In *International Conference on Artificial Intelligence and Statistics*. pages 752–760.

Winn, J. M. & Bishop, C. M. (2005). Variational message passing. In *Journal of Machine Learning Research*. pages 661–694.