UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
Corso di Dottorato di Ricerca in Scienze Statistiche
Ciclo XXXI

On Variational Approximations for Frequentist and Bayesian Inference

Course Coordinator: Prof. Nicola Sartori

Supervisor: Prof. Nicola Sartori

Co-supervisors: Prof. Alessandra Salvan and Prof. Matt P. Wand

Dottorando: Luca Maestrini

Abstract

Variational approximations are approximate inference techniques for complex statistical models providing fast, deterministic alternatives to conventional methods that, however accurate, take much longer to run. We extend recent work concerning variational approximations developing and assessing some variational tools for likelihood based and Bayesian inference. In particular, the first part of this thesis employs a Gaussian variational approximation strategy to handle frequentist generalized linear mixed models with general design random effects matrices such as those including spline basis functions. This method involves approximation to the distributions of random effects vectors, conditional on the responses, via a Gaussian density. The second thread is concerned with a particular class of variational approximations, known as mean field variational Bayes, which is based upon a nonparametric product density restriction on the approximating density. Algorithms for inference and fitting for models with elaborate responses and structures are developed adopting the variational message passing perspective. The modularity of variational message passing is such that extensions to models with more involved likelihood structures and scalability to big datasets are relatively simple. We also derive algorithms for models containing higher level random effects and non-normal responses, which are streamlined in support of computational efficiency. Numerical studies and illustrations are provided, including comparisons with a Markov chain Monte Carlo benchmark.

Sommario

Le approssimazioni variazionali sono tecniche di inferenza approssimata per modelli statistici complessi che si propongono come alternative, più rapide e di tipo deterministico, a metodi tradizionali che, sebbene accurati, necessitano di maggiori tempi per l'adattamento. Vengono qui sviluppati e valutati alcuni strumenti variazionali per l'inferenza basata sulla verosimiglianza e per l'inferenza bayesiana, estendendo dei risultati recenti in letteratura sulle approssimazioni variazionali. In particolare, la prima parte della tesi impiega una strategia basata su un'approssimazione variazionale gaussiana per la funzione di verosimiglianza di modelli lineari generalizzati misti con matrici di disegno degli effetti casuali generiche, includenti, per esempio, funzioni di basi *spline*. Questo metodo consiste nell'approssimare la distribuzione del vettore degli effetti casuali, condizionatamente alle risposte, con una densità gaussiana. Il secondo filone concerne invece una particolare classe di approssimazioni variazionali nota come *mean field variational Bayes*, che impone un prodotto di densità come restrizione non parametrica sulla densità approssimante. Vengono sviluppati algoritmi per l'inferenza e l'adattamento di modelli con risposte elaborate, adottando la prospettiva del *variational message passing*. La modularità del *variational message passing* è tale da consentire estensioni a modelli con strutture di verosimiglianza più complesse e scalabilità a insiemi di dati di grandi dimensioni con relativa semplicità. Vengono inoltre derivati in forma esplicita degli algoritmi per modelli contenenti effetti casuali su più livelli e risposte non normali, introducendo semplificazioni atte a incrementare l'efficienza computazionale. Sono inclusi studi numerici e illustrazioni, considerando come riferimento per un confronto il metodo *Markov chain Monte Carlo*.

To my family

Acknowledgements

Words are simply not enough to express my deepest thank you to my supervisor and co-supervisor in Padova, Nicola and Alessandra, who patiently transmitted all their passion for rigorous scientific research to me.

I extend my sincere gratitude to my co-supervisor in Sydney, Matt, who spent many hours guiding me with helpful advice throughout a great part of my PhD journey and from whom I learnt dedication and determination to achieve gratifying results.

Thank you to the people who dedicated some of their time for discussing my research during my PhD studies, including Ruggero Bellio, Carlo Gaetan, Francis Hui, and John Ormerod, as well as to the reviewers for their insightful comments.

I am always and forever grateful to my family, to Mum and Dad, Simone, my lovely grandparents, aunts, uncles and cousins for raising me or giving me all the support I needed.

Finally, I would like to thank my friends, especially my PhD colleagues in Padova and Sydney who have shared with me one of my life most beautiful experiences.

November, 30 2018

Notational conventions

Lower-case Roman and Greek letters in boldface denote vectors whose entries are subscripts. For example, $\boldsymbol{x}$ denotes a $n \times 1$ vector containing $x_1, \ldots, x_n$. All vectors are column vectors. Upper-case Roman and Greek letters in boldface denote matrices. For example, $\boldsymbol{X}$ denotes a $m \times n$ matrix containing n vectors of dimension $m \times 1$, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$.

$\mathbb{R}$	The set of real numbers.
$\mathbb{R}^n$	The set of real vectors of dimension n.
$\mathbb{R}^{n \times m}$	The set of real matrices with n rows and m columns.
T	Transpose symbol for vectors and matrices.
$(\cdot, \ldots, \cdot)$	Lists of scalars, vectors and matrices within round brackets are concatenated vertically (vertical concatenation), e.g. $\boldsymbol{a} = (a_1, \ldots, a_n)$ is a column vector.
$[\cdot, \ldots, \cdot]$	Lists of scalars, vectors and matrices within square brackets are concatenated horizontally (horizontal concatenation), e.g. $\boldsymbol{a} = [a_1, \ldots, a_n]$ is a row vector. In addition, $(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n) = \left[\boldsymbol{a}_1^T, \ldots, \boldsymbol{a}_n^T\right]^T$.
$\begin{bmatrix} a_{ij} \\ 1 \le i \le n \end{bmatrix}_{1 \le j \le m}$	Matrix with n rows and m columns or vector with n rows if $m = 1$.
$(\boldsymbol{a})_i$ or a_i	The element in ith position of a vector $\boldsymbol{a}$.
$(\boldsymbol{A})_{ij}$ or A_{ij}	The element in (i, j)th position of a matrix $\boldsymbol{A}$.
$\operatorname{diag}(\boldsymbol{a})$	For a vector $\boldsymbol{a} \in \mathbb{R}^n$, diagonal matrix of size $n \times n$ such that $\operatorname{diag}(\boldsymbol{a}) = \begin{bmatrix} a_1 & 0 & \cdots & 0 \\ 0 & a_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_n \end{bmatrix}$.
$\operatorname{dg}(\boldsymbol{A})$	For a square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, a vector of length n whose entries correspond to the diagonal elements of $\boldsymbol{A}$ such that $\operatorname{dg}\left(\begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{bmatrix}\right) = (A_{11}, A_{22}, \ldots, A_{nn})$.
$\boldsymbol{0}$ or $\boldsymbol{O}$	An appropriately-sized vector or matrix of zeroes.

$\mathbf{1}$	An appropriately-sized vector or matrix of ones.
$\boldsymbol{e}_i$	An appropriately-sized vector of zeros, except the ith value which is 1.
$\boldsymbol{I}$	An appropriately-sized identity matrix.
$\boldsymbol{a} \odot \boldsymbol{b}$	The element-wise multiplication of vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$, that is, $(a_1 b_1, \ldots, a_n b_n)$.
$\boldsymbol{a}/\boldsymbol{b}$	The element-wise division of vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$, that is, $(a_1/b_1, \ldots, a_n/b_n)$.
$f(\boldsymbol{x})$	Univariate function $f : \mathbb{R} \to \mathbb{R}$ of a vector $\boldsymbol{x} \in \mathbb{R}^n$ such that $f(\boldsymbol{x}) = (f(x_1), \ldots, f(x_n))$.
$\|\boldsymbol{a}\|$	Vector 2−norm of a vector $\boldsymbol{a} \in \mathbb{R}^n$ such that $\|\boldsymbol{a}\| = (\sum_{i=1}^n a_i^2)^{1/2}$.
$\operatorname{tr}(\boldsymbol{A})$	The trace of the matrix $\boldsymbol{A}$.
$\lvert\boldsymbol{A}\rvert$	The determinant of the matrix $\boldsymbol{A}$.
$\boldsymbol{A}^{-1}$	The inverse matrix of the matrix $\boldsymbol{A}$.
$\boldsymbol{A} \otimes \boldsymbol{B}$	The Kronecker product of matrices $\boldsymbol{A} \in \mathbb{R}^{n \times m}$ and $\boldsymbol{B} \in \mathbb{R}^{p \times q}$, that is, the $np \times mq$ matrix defined by $[a_{ij}\boldsymbol{B}]_{1 \le i \le n, 1 \le j \le m}$.
$\boldsymbol{A} \odot \boldsymbol{B}$	The element-wise multiplication of matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n \times m}$, that is, the $n \times m$ matrix defined by $[a_{ij} b_{ij}]_{1 \le i \le n, 1 \le j \le m}$.
$\operatorname{vec}(\boldsymbol{A})$	For a square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times m}$, the $nm \times 1$ vector obtained by stacking the columns of $\boldsymbol{A}$ underneath each other in order from left to right (*vectorization*) such that $\operatorname{vec}(\boldsymbol{A}) = (A_{11}, \ldots, A_{n1}, A_{12}, \ldots, A_{n2}, \ldots, A_{1m}, \ldots, A_{nm})$.
$\operatorname{vec}^{-1}_{n \times m}(\boldsymbol{a})$	For a vector $\boldsymbol{a} \in \mathbb{R}^{nm}$, the $n \times m$ matrix $\boldsymbol{A}$ formed from listing the entries of $\boldsymbol{a}$ in a column-wise fashion in order from left to right, such that $\operatorname{vec}(\boldsymbol{A}) = \boldsymbol{a}$; if the subscript is not specified, $\boldsymbol{a} \in \mathbb{R}^{n^2}$ and $\boldsymbol{A}$ is of size $n \times n$.
$\operatorname{vech}(\boldsymbol{A})$	For a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, the $n(n+1)/2 \times 1$ vector obtained by vectorizing only the lower triangular part of $\boldsymbol{A}$ (*halph-vectorization*) such that $\operatorname{vech}(\boldsymbol{A}) = \big(A_{11}, \ldots, A_{n1}, A_{22}, \ldots, A_{n2}, \ldots, A_{(n-1)(n-1)}, A_{n(n-1)}, A_{nn}\big)$.
$\operatorname{vech}^{-1}(\boldsymbol{a})$	For a vector $\boldsymbol{a} \in \mathbb{R}^{n(n+1)/2}$, the $n \times n$ symmetric matrix $\boldsymbol{A}$ whose lower triangular part is formed from listing the entries of $\boldsymbol{a}$ in a column-wise fashion in order from left to right such that $\operatorname{vech}(\boldsymbol{A}) = \boldsymbol{a}$.
$\boldsymbol{D}_n$	For a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, the *duplication matrix of order* n and size $n^2 \times n(n+1)/2$ containing all zeroes and ones such that $\boldsymbol{D}_n \operatorname{vech}(\boldsymbol{A}) = \operatorname{vec}(A)$.
$\boldsymbol{D}_n^+$	For a duplication matrix of order n, $\boldsymbol{D}_n$, and a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, the *Moore–Penrose inverse* $\boldsymbol{D}_n^+ \equiv \big(\boldsymbol{D}_n^T \boldsymbol{D}_n\big)^{-1} \boldsymbol{D}_n^T$ such that $\boldsymbol{D}_n^+ \operatorname{vech}(\boldsymbol{A}) = \operatorname{vec}(A)$.

$\underset{1\leq i\leq d}{\text{blockdiag}}(\boldsymbol{A}_i)$ For matrices $\boldsymbol{A}_i \in \mathbb{R}^{n_i \times m_i}$, $1 \leq i \leq d$, a matrix of size $\sum_{i=1}^{d} n_i \times \sum_{i=1}^{d} m_i$ such that

$$\text{blockdiag}_{1\leq i\leq d}(\boldsymbol{A}_i) \equiv \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{O} & \cdots & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{A}_2 & \cdots & \boldsymbol{O} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{O} & \boldsymbol{O} & \cdots & \boldsymbol{A}_d \end{bmatrix}.$$

$\underset{1\leq i\leq d}{\text{stack}}(\boldsymbol{A}_i)$ For matrices $\boldsymbol{A}_i \in \mathbb{R}^{n_i \times m}$, $1 \leq i \leq d$, a matrix of size $\sum_{i=1}^{d} n_i \times m$ such that

$$\text{stack}_{1\leq i\leq d}(\boldsymbol{A}_i) \equiv \begin{bmatrix} \boldsymbol{A}_1 \\ \vdots \\ \boldsymbol{A}_d \end{bmatrix}.$$

$I(x)$ Indicator variable, which takes the value 1 if x is true and 0 otherwise.

$\phi(x)$ Density function of a standard normal distributed random variable x.

$\Phi(x)$ Cumulative distribution function of a standard normal distributed random variable x.

$o(\cdot)$ For f and g real valued functions, both defined on some unbounded set of real positive numbers and $g(x)$ strictly positive for all large enough values of x, $f(x) = o(g(x))$ as $x \to \infty$ if for every positive constant ε there exists a constant N such that $|f(x)| \leq \varepsilon g(x)$ for all $x \geq N$.

$O(\cdot)$ For f and g real valued functions, both defined on some unbounded set of real positive numbers and $g(x)$ strictly positive for all large enough values of x, $f(x) = O(g(x))$ as $x \to \infty$ if and only if there exist a positive real number M and a real number x_0 such that $|f(x)| \leq Mg(x)$ for all $x \geq x_0$.

Contents

List of Figures

List of Tables

List of Algorithms

Introduction

> "Far better an approximate answer to the *right* question, which is often vague, than an *exact* answer to the wrong question, which can always be made precise."
>
> John W. Tukey[1], 1915 – 2000

Overview

The rapid growth of information has revolutionized the concept of data analysis and the instruments for handling data. In some cases, the amount of data to store has grown so much that standard computer memory is unable to manage such a large volume. This gives a new impulse to engineering research on new data processing technologies. In a similar manner, the celerity of processing information in recent years has motivated researchers and practitioners to develop and employ faster data analysis techniques in lieu of conventional methods whose running time does not match practical requirements.

This *big data revolution* encourages the start of new research threads in statistics and, in the same spirit, is the drive behind this thesis. Specifically, the thesis focuses on the development and assessment of variational approximation methods for statistical inference and fitting problems from both likelihood based and Bayesian perspectives.

Variational approximations are a class of techniques for making approximate deterministic inference for complex statistical models. The name derives from the mathematical discipline of variational calculus, since the approximation is obtained from the optimization of a functional over a class of density functions on which that functional depends. Now part of conventional computer science methodology concerning elaborate problems such as natural language processing, speech recognition, document retrieval, genetic linkage analysis, computational biology, computational neuroscience, computer vision and robotics, variational approximations are finding a growing presence in statistics. However, much of the mainstream literature on variational approximations uses

[1] See Tukey (1962).

terminology, notation, and examples from computer science rather than statistics. An introduction to the topic from a mere statistical perspective is contained in Ormerod and Wand (2010), that, in turn, refer to Jordan *et al.* (1999), Jordan (2004), Titterington (2004), and Bishop (2006, Chapter 10) for summaries of variational approximations. Teschendorff *et al.* (2005), McGrory and Titterington (2007) and McGrory *et al.* (2009) are cited as works about variational approximation methodology for particular applications, while Hall *et al.* (2002) and Wang and Titterington (2006) as references about the statistical properties of estimators obtained via variational approximation. Additionally, a description of variational approximations as a machine learning method for approximating probability densities is included in Wainwright and Jordan (2008). A more recent reference, Blei *et al.* (2017), provides an exhaustive overview on variational inference specifically addressed to statisticians. Particular emphasis is placed on the popular stochastic variational inference methodology (Hoffman *et al.*, 2013), which scales variational inference to large datasets using stochastic optimization (Robbins and Monro, 1951).

Blei *et al.* (2017) also assert that the development of variational techniques, initially for Bayesian inference, followed two parallel, yet separate, tracks. Peterson and Anderson (1987) describe presumably the first variational procedure for a particular model, a neural network. Their article, in connection with some intuitions from statistical mechanics (Parisi, 1988), brought to the flowering of variational inference procedures for a wide class of models. In parallel, a variational algorithm for a similar neural network model was proposed in Hinton and Van Camp (1993).

Though the theory around variational inference is not growing at the same speed as the methodology, there are several threads of research that prove theoretical guarantees of variational approximations. References are given in Section 5.2 of Blei *et al.* (2017). Additional results on asymptotic properties are included in Wang and Blei (2018) and Zhang and Gao (2018).

Variational approximations find application in both likelihood based and Bayesian inference problems. However, their use in the literature is far more widespread for Bayesian inference, where intractable calculus abounds and where they provide fast, deterministic alternatives to Monte Carlo methods. The idea behind this methodology is to first propose a family of densities in the integral of interest and then to find the member of that family which is closest to the target density. In the standard version of variational approximations the approximating density produces a lower bound and closeness is measured by Kullback–Leibler divergence (Kullback and Leibler, 1951). Nevertheless, not all variational approximations fit within the Kullback–Leibler divergence framework.

Another variety are what might be called *tangent transform* variational approximations since they are based on tangent-type representations of concave and convex functions (e.g. Jordan *et al.*, 1999; Ormerod and Wand, 2010, Section 3). This thesis is concerned with the first type of variational approximations to which we refer as *density transform* approach. For sake of completeness, Wainwright and Jordan (2008) emphasized that any inferential procedure that uses optimization and alternative divergence measures to approximate a density can be named "variational inference". This includes methods such as *expectation propagation* (Minka, 2001), *belief propagation* (Yedidia *et al.*, 2001) or even the Laplace approximation.

The essence of the density transform variational approach is to provide an approximation which is derived from a lower bound that is more tractable than the likelihood function. Tractability is enhanced by restricting the approximating density to a more manageable class of densities. The most common restrictions are:

(a) the approximating density is a member of a parametric family of density functions;

(b) the approximating density factorizes as a product of densities according to a partition of the set of parameters.

Note that (b) represents a type of nonparametric restriction since the product form is the only assumption being made. Restriction (b) produces the so-called *mean field* approximation, whose roots are in statistical physics (e.g. Parisi, 1988). The term *variational Bayes* has become commonplace for approximate Bayesian inference under product density restrictions. More recently, Wand (2017) reviseted mean field variational Bayes, with a focus on semiparametric regression, working with an approach known as *variational message passing* (Winn and Bishop, 2005). This approach has the advantage that the algorithms it generates are amenable to modularization and extension to arbitrarily large models via the notion of *factor graph* (Frey *et al.*, 1998) fragments.

Main contributions of the thesis

This thesis is developed around Kullback–Leibler-based variational methodologies that employ the restrictions (a) and (b) mentioned above, the former in frequentist settings, the latter for Bayesian inference. Specifically, the parametric restriction (a) is applied as a strategy for fitting generalized linear mixed models with general design matrices, following the framework established in Ormerod and Wand (2012). The approach, named Gaussian variational approximations, consists in approximating the distribution of random effects vectors, conditional on the responses, with a Gaussian

density chosen to minimize a variational lower bound on the model likelihood function. Our contribution lies in exposing Gaussian variational approximations as a fast and effective alternative to more widely used inference methods such as, for instance, penalized quasi likelihood or generalized cross-validation. Recently, Hui *et al.* (2018) have investigated the use of Gaussian variational approximations for generalized additive models. However, their approach puts emphasis on the penalty parameter estimation which in our formulation may arise as an outcome of the optimization problem and their optimization strategy is structured in a different way. Furthermore they get rid of a residual integration step still present in the variational lower bound for Bernoulli models by creating a second lower bound of the initial lower bound, while we prefer a fast and more accurate route with numerical integration.

The second part of the thesis is concerned with the nonparametric restriction (b) and the so called mean field variational Bayes methodology under the variational message passing perspective for factor graph fragments. We extend the work contained in McLean and Wand (2018) to accommodate additional elaborate likelihood fragments. The modularity of variational message passing permits relatively simple extensions to more complicated scenarios in such a way that, for instance, semiparametric regression models can be handled using factor graph fragments. Extension to arbitrarily large models is guaranteed by algorithm updates that are based on natural parameters and sufficient statistics of exponential family densities. Our practical implementations of variational algorithms employ factorised approximating posteriors and priors that belong to the conjugate-exponential families, making the required integrals tractable.

Furthermore, we present explicit algorithms to implement streamlined variational inference for two-level non-normal response models. The algorithms take advantage of the sparse matrix results presented in Nolan *et al.* (2018) and are part of a framework which is prone to extension to high-level random effects models. Scalability for large datasets also benefits of the streamlined variational message passing approach.

The present thesis is organized in the following way. Chapter 1 defines the settings for the variational approximations methodology under consideration. A broader overview on variational inference techniques and some references to theoretical results are also provided. Details on semiparametric regression via O'Sullivan penalized splines (O'Sullivan, 1986) and the generalized linear model formulation through general design matrices are also provided.

Chapter 2 is dedicated to *Gaussian variational approximations*. In detail, Poisson and binomial response models with nonparametric and additive structures are considered.

Inference and confidence interval construction are easily derived from the estimated approximating Gaussian density. Results are then compared to those obtained via classical fast estimation methods and good overall performances in terms of estimation time and inferential properties are observed. Issues involving the lower bound optimization are also investigated. Furthermore, the settings for fitting geoadditive models as an example of application to generalized linear mixed models with spatial structures are described.

Each subsection of Chapter 3 develops variational algorithms catered to a particular likelihood fragment. Three likelihood fragments, including Pareto random sample, support vector regression and skew t regression are explored. For the skew t case, we also investigate how various auxiliary random variable representations of the likelihood impact the variational approximating results. The response likelihoods are re-expressed in terms of auxiliary variables and more common distributions to avoid numerically intractable steps. As a drawback, we show that such a reparametrization may introduce strong posterior dependence among variables which is hard to capture with simple forms of approximating densities.

In Chapter 4 we develop fast variational algorithms for fitting and inference in mixed models, where all algorithmic updates are available in closed form. The centrepieces for Chapter 4 are streamlined variational algorithms for fast and memory efficient fitting and inference in large two-level Bayesian random effects models with Poisson and Binomial responses.

The probability distributions used in the thesis are displayed in Appendix A. The rest of the Appendix reflects the structure of main chapters and essentiallyv complements derivations and computational steps.

All the numerical analyses appearing in this thesis are supported by a comparison with Markov chain Monte Carlo, which provides our benchmark for assessing variational approximation performances.

Chapter 1

Variational inference

1.1 Density transform variational approximations

The class of variational approximations contains a wide range of techniques for deterministic approximate inference. In the present thesis, the variational approximation methods under consideration refer to the most common variant, known as density transform approach, according to the classification of Ormerod and Wand (2010). This method involves approximation of conditional distributions of interest or posterior densities by other densities that minimize the divergence to the densities of interest and for which inference is more tractable. We describe these concepts more in detail with an example in the Bayesian setting. The following illustration easily extends to variational inference for frequentist models.

Consider a generic Bayesian model with parameter vector $\boldsymbol{\theta}$, parameter space Θ and observed data vector $\boldsymbol{y}$, where, for simplicity, we assume that both the random vectors are continuous. Bayesian inference is concerned with the *posterior density function*

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

where $p(\boldsymbol{y}, \boldsymbol{\theta})$ is the *joint density* of $\boldsymbol{y}$ and $\boldsymbol{\theta}$, and $p(\boldsymbol{y})$ is known as the *marginal likelihood* or *model evidence* in the computer science literature on variational approximations. Given an arbitrary density function q defined over Θ, the logarithm of the

marginal likelihood satisfies

$$\begin{aligned}
\log p(\boldsymbol{y}) &= \log p(\boldsymbol{y}) \int q(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\
&= \int q(\boldsymbol{\theta}) \log \left\{ \frac{p(\boldsymbol{y}, \boldsymbol{\theta}) / q(\boldsymbol{\theta})}{p(\boldsymbol{\theta} \mid \boldsymbol{y}) / q(\boldsymbol{\theta})} \right\} d\boldsymbol{\theta} \\
&= \int q(\boldsymbol{\theta}) \log \left\{ \frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right\} d\boldsymbol{\theta} + \int q(\boldsymbol{\theta}) \log \left\{ \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta} \mid \boldsymbol{y})} \right\} d\boldsymbol{\theta} \qquad (1.1) \\
&\geq \int q(\boldsymbol{\theta}) \log \left\{ \frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right\} d\boldsymbol{\theta}. \qquad (1.2)
\end{aligned}$$

Note that the second integral on the right hand side of (1.1) is the Kullback–Leibler divergence between q and $p(\boldsymbol{\theta} \mid \boldsymbol{y})$,

$$\mathrm{KL}\left\{ q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta} \mid \boldsymbol{y}) \right\} = \int q(\boldsymbol{\theta}) \log \left\{ \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta} \mid \boldsymbol{y})} \right\} d\boldsymbol{\theta}, \qquad (1.3)$$

which is non-negative for all densities q, with equality arising if and only if $q(\boldsymbol{\theta}) = p(\boldsymbol{\theta} \mid \boldsymbol{y})$. This gives the inequality (1.2) and a *lower bound* $\underline{p}(\boldsymbol{y}; q)$ on the marginal likelihood

$$p(\boldsymbol{y}) \geq \underline{p}(\boldsymbol{y}; q) = \exp \int q(\boldsymbol{\theta}) \log \left\{ \frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right\} d\boldsymbol{\theta}. \qquad (1.4)$$

Maximization of $\underline{p}(\boldsymbol{y}; q)$ is equivalent to minimization of the Kullback–Leibler divergence between $q(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ and may provide an alternative approach to the direct optimization of the marginal likelihood. The new maximization problem consists in finding the optimal approximating density in terms of Kullback–Leibler divergence that approximates the posterior density function.

The lower bound (1.4) can be also obtained via the Jensen's inequality, if renouncing to quantify the gap between $p(\boldsymbol{y})$ and $\underline{p}(\boldsymbol{y}; q)$ (e.g. Jordan *et al.*, 1999).

The whole idea behind this approach is to propose an approximation of the posterior density $p(\boldsymbol{\theta}|\boldsymbol{y})$ by a $q(\boldsymbol{\theta})$ for which $\underline{p}(\boldsymbol{y}; q)$ is more tractable than $p(\boldsymbol{y})$. Tractability is achieved by restricting $q(\boldsymbol{\theta})$ to a more manageable class of densities. As described in the previous chapter, the two common restrictions are:

(a) $q(\boldsymbol{\theta})$ is a member of a parametric family of density functions;

(b) $q(\boldsymbol{\theta})$ factorizes as $\prod_{i=1}^{M} q_i(\boldsymbol{\theta}_i)$, for some partition $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$ of $\boldsymbol{\theta}$.

Depending on the model at hand, both restrictions can have minor or major impacts on the approximation performances in terms of accuracy.

An approximation based on the parametric restriction (a) is named Gaussian variational approximation (GVA) whenever the approximating density $q(\boldsymbol{\theta})$ is assumed to

be within the family of Gaussian densities. Section 1.2 provides a brief description of GVA.

Also known as mean field approximation in Bayesian statistics, the product density form in case (b) is the only assumption being made and represents a type of nonparametric restriction. Restricting q to a subclass of product densities gives rise to explicit solutions to the optimization problem for each product component in terms of the others and, in turn, leads to an iterative scheme to obtain the simultaneous solutions which is known as mean field variational Bayes (MFVB). Variational message passing (VMP) is a prescription for obtaining mean field variational Bayes (MFVB) approximations to posterior density functions that allows for modularization and extension to arbitrarily large models. Both the MFVB and VMP perspectives are treated in Section 1.3.

Section 1.4 contains a concise review of early references on the mean field methodology and a glance at the class of message passing algorithms.

Section 1.5 provides some references to the theory of variational approximations that present connections with the methods considered in this thesis.

Finally, Section 1.6 is dedicated to an accessory but recurrent topic in the present thesis, that is, semiparametric regression. In particular, the mixed model representation of semiparametric regression models facilitates variational approximate inference under both restrictions above.

1.2 Gaussian variational approximations

Nontrivial frequentist examples where an explicit solution arises by applying the product density methodology are not known. On the other hand, restricting the approximating density to be part of a parametric family of density functions may lead to effective variational approximation strategies.

Frequentist models that stand to benefit from variational approximations are those for which specification of the likelihood involves conditioning on a vector of latent variables $\boldsymbol{u}$. Given a vector of observed data $\boldsymbol{y}$, the log-likelihood of the model parameter vector $\boldsymbol{\theta}$ takes the form

$$\ell(\boldsymbol{\theta}) = \log p(\boldsymbol{y}; \boldsymbol{\theta}) = \log \int p(\boldsymbol{y} \mid \boldsymbol{u}; \boldsymbol{\theta})\, p(\boldsymbol{u}; \boldsymbol{\theta})\, d\boldsymbol{u},$$

and

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \ell(\boldsymbol{\theta})$$

is the *maximum likelihood estimate* of $\boldsymbol{\theta}$.

For some statistical models, the log-likelihood $\ell(\boldsymbol{\theta})$ may not be available in a closed form because of analytically intractable integration. In such a context, variational approximations can provide a more tractable approximation in replacement of the original optimization problem, depending on the forms of $p(\boldsymbol{y} \mid \boldsymbol{u}; \boldsymbol{\theta})$ and $p(\boldsymbol{u}; \boldsymbol{\theta})$.

In this context, the auxiliary function q is a density of the latent variable $\boldsymbol{u}$. Suppose to restrict q to a parametric family of densities $\{q(\boldsymbol{u}; \boldsymbol{\xi}) : \boldsymbol{\xi} \in \Xi\}$, where $\boldsymbol{\xi}$ is a vector of *variational parameters*. Then a log-likelihood lower bound

$$\underline{\ell}(\boldsymbol{\theta}, \boldsymbol{\xi}; q) = \int q(\boldsymbol{u}; \boldsymbol{\xi}) \log \left\{ \frac{p(\boldsymbol{y}, \boldsymbol{u}; \boldsymbol{\theta})}{q(\boldsymbol{u}; \boldsymbol{\xi})} \right\} d\boldsymbol{u} \tag{1.5}$$

can be derived, as shown for the Bayesian counterpart (1.2). The new maximization problem over the model parameters $\boldsymbol{\theta}$ and the variational parameters $\boldsymbol{\xi}$

$$\left(\hat{\underline{\boldsymbol{\theta}}}, \hat{\boldsymbol{\xi}}\right) = \underset{\boldsymbol{\theta}, \boldsymbol{\xi}}{\operatorname{argmax}} \, \underline{\ell}(\boldsymbol{\theta}, \boldsymbol{\xi}; q)$$

is the set-up for variational approximate inference and GVA, when the $q(\boldsymbol{u}; \boldsymbol{\xi})$ is chosen to be a family of normal densities. The vector $\hat{\underline{\boldsymbol{\theta}}}$ is a variational approximation to the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$. Estimated standard errors can be obtained by plugging in $\hat{\underline{\boldsymbol{\theta}}}$ for $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\xi}}$ for $\boldsymbol{\xi}$ in the variational approximate Fisher information matrix arising from replacement of $\ell(\boldsymbol{\theta})$ by $\underline{\ell}(\boldsymbol{\theta}, \boldsymbol{\xi}; q)$ and the corresponding Hessian matrix (e.g. Ormerod and Wand, 2012).

1.3 Mean field variational approximations

Variational message passing is an approach to variational Bayes approximate inference that allows modularization through the notions of *factor graphs* and *message passing*. According to a factor graph message passing approach (e.g. Minka and Winn, 2008, Appendix A), calculations only need to be performed once for a particular fragment and can be integrated with other fragments to construct inference engines for arbitrarily large models.

A *mean field variational approximation* $q^*(\boldsymbol{\theta})$ to $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ is the minimizer of the Kullback–Leibler divergence (1.3) subject to a product density restriction, or mean field restriction,

$$q(\boldsymbol{\theta}) = \prod_{i=1}^{M} q(\boldsymbol{\theta}_i), \tag{1.6}$$

where $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$ is some partition of $\boldsymbol{\theta}$. It can be shown (e.g. Ormerod and Wand, 2010, Section 2.2) that the optimal q-density functions satisfy

$$q^*(\boldsymbol{\theta}_i) \propto \exp\left\{E_{q(\boldsymbol{\theta}\backslash\boldsymbol{\theta}_i)} \log p(\boldsymbol{y}, \boldsymbol{\theta})\right\}, \quad 1 \leq i \leq M, \tag{1.7}$$

or, alternatively,

$$q^*(\boldsymbol{\theta}_i) \propto \exp\left\{E_{q(\boldsymbol{\theta}\backslash\boldsymbol{\theta}_i)} \log p(\boldsymbol{\theta}_i \,|\, \boldsymbol{y}, \boldsymbol{\theta}\backslash\boldsymbol{\theta}_i)\right\}, \quad 1 \leq i \leq M, \tag{1.8}$$

where $\boldsymbol{\theta}\backslash\boldsymbol{\theta}_i$ denotes the entries of $\boldsymbol{\theta}$ with $\boldsymbol{\theta}_i$ omitted and the distribution $p(\boldsymbol{\theta}_i \,|\, \boldsymbol{y}, \boldsymbol{\theta}\backslash\boldsymbol{\theta}_i)$ is known as *full conditional density function*. The previous expression gives rise to the MFVB iterative scheme for obtaining the optimal density functions $q^*(\boldsymbol{\theta}_i)$. A listing of such an algorithm is provided, for instance, in Ormerod and Wand (2010). Alternatively, optimization may be performed via a message passing algorithmic approach for mean field approximation (VMP), which is limited to conjugate exponential family models (Winn and Bishop, 2005).

When expectation steps appearing in algorithm updates require evaluation of definite integrals that do not admit analytic solutions or easily manageable analytic solutions, univariate quadrature schemes may be considered. The trapezoidal rule is a simple and effective quadrature approach, whose accuracy arbitrarily improves by increasing the number of trapezoidal elements. If the integral is over an infinite or semi-infinite region, rather than a compact interval, it is important to define the effective support of the integrand function to guarantee accurate computation. Attention is necessary to avoid overflow and underflow issues concerning the application of the trapezoidal rule.

1.3.1 Coordinate ascent mean field variational Bayes

Expression (1.7) gives rise to a conceptually simple coordinate ascent algorithm. First, initialize $q(\boldsymbol{\theta}_1), \ldots, q(\boldsymbol{\theta}_M)$. Then cycle

$$q_i(\boldsymbol{\theta}_i) \longleftarrow \frac{\exp\left\{E_{q(\boldsymbol{\theta}\backslash\boldsymbol{\theta}_i)} \log p(\boldsymbol{y}, \boldsymbol{\theta})\right\}}{\int \exp\left\{E_{q(\boldsymbol{\theta}\backslash\boldsymbol{\theta}_i)} \log p(\boldsymbol{y}, \boldsymbol{\theta})\right\} d\boldsymbol{\theta}_i},$$

for each $1 \leq i \leq M$, until the increase in the lower bound on the marginal log-likelihood

$$\log \underline{p}(\boldsymbol{y}; q) = E_{q(\boldsymbol{\theta})}\{\log p(\boldsymbol{y}, \boldsymbol{\theta}) - \log q(\boldsymbol{\theta})\} \tag{1.9}$$

is negligible. The optimal $q_i^*(\boldsymbol{\theta}_i)$ densities are obtained at convergence. The $\longleftarrow$ symbol means that the function of $\boldsymbol{\theta}_i$ on the left-hand side is updated according to the expression

on the right-hand side; multiplicative factors not depending on $\boldsymbol{\theta}_i$ can be ignored. The expressions containing such a symbol are technically named "updates".

The term $E_{q(\boldsymbol{\theta})}\{\log q(\boldsymbol{\theta})\}$ in (1.9) is the so called *entropy* of the density function $q(\boldsymbol{\theta})$. Boyd and Vandenberghe (2004) use convexity properties to show that convergence to at least a local optimum is guaranteed. In fact, the MFVB scheme is basically interpretable as a generalisation of the *expectation-maximisation* (EM) approach (Chappell *et al.*, 2009).

In presence of conjugacy with priors, one may take advantage of the optimal form (1.8) and associate the $q^*(\boldsymbol{\theta}_i)$s to recognizable density families. Then the optimization procedure reduces to updating parameters in the $q^*(\boldsymbol{\theta}_i)$ family.

Directed acyclic graph (DAG) representations of Bayesian models are very useful when deriving the optimal q-densities formulated as (1.8). In detail, DAGs yield substantial benefits to MFVB schemes for large models taking advantage of an important probabilistic concept, namely the *Markov blanket theory*. For each node on a probabilistic DAG, the conditional distribution of the node given the rest of the nodes is the same as the conditional distribution of the node given its Markov blanket (Dechter and Pearl, 1988), whose definition is provided in the exemplification below. In the Bayesian models considered here this implies

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

It follows that (1.8) simplifies to

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

This is known as the *locality property* of DAGs. For large DAGs, this property produces considerable algebraic benefits. Such a result means that determination of the required full conditionals involves only a series of local calculations on the DAG. In particular, it shows that the $q^*(\boldsymbol{\theta}_i)$s require only local calculations on the models DAG. We explain this concept with an elementary graphical example.

Consider the following hierarchical Bayesian model similar to the exemplification provided in Wand *et al.* (2011):

$$\begin{aligned}
&\boldsymbol{y} \,|\, \theta_1, \theta_2, \theta_3 \sim p(\boldsymbol{y} \,|\, \theta_1, \theta_2, \theta_3), \\
&\theta_1 \,|\, \theta_4 \sim p(\theta_1 \,|\, \theta_4), \quad \theta_2 \,|\, \theta_4, \theta_5 \sim p(\theta_2 \,|\, \theta_4, \theta_5), \quad \theta_3 \,|\, \theta_6 \sim p(\theta_3 \,|\, \theta_6) \quad \text{indep.}, \\
&\theta_4 \sim p(\theta_4), \quad \theta_5 \sim p(\theta_5), \quad \theta_6 \sim p(\theta_6) \quad \text{indep.},
\end{aligned} \tag{1.11}$$

where $\boldsymbol{y}$ is the observed data vector and $\theta_1, \ldots, \theta_6$ are model parameters. Figure 1.1

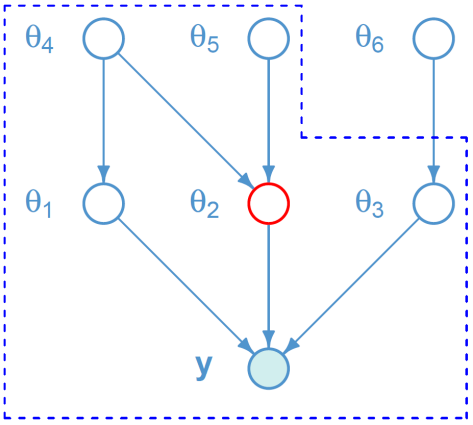

FIGURE 1.1: Directed acyclic graph corresponding to model (1.11). The six parameters (hidden nodes) are associated with circles. The shaded node $\boldsymbol{y}$ corresponds to the obsered data (evidence node). The dashed line indicates the Markov blanket for θ_2.

shows the directed acyclic graph representation of model (1.11). The θ_is, $1 \leq i \leq 6$, are represented by open circles named *hidden nodes* and the data vector $\boldsymbol{y}$, the *evidence node*, is portrayed by a shaded circle. The arrows indicate conditional dependence relationships among the model random variables. A DAG can be interpreted as a *family tree* in which each directed edge conveys a parent-child relationship between the associated nodes. For instance, the node θ_2 is pointed by the arrow-heads departing from nodes θ_4 and θ_5, therefore θ_2 is a *child* of the *parent* nodes θ_4 and θ_5. The nodes θ_1 and θ_3 are *co-parents* of node θ_2 since they all have a common child, $\boldsymbol{y}$. The Markov blanket of a node, say θ_2, is the set including the parents, co-parents and children nodes (θ_1, θ_3, θ_4, θ_5 and $\boldsymbol{y}$) of that particular node, as displayed by the dashed line in Figure 1.1. Briefly, the Markov blanket of a node separates that node from the remainder of the graph and has a probabilistic interpretation known as *locality property*. Because of (1.10), $q^*(\theta_2)$ depends on particular q-density moments of θ_1, θ_3, θ_4 and θ_5 but not on their distribution. Therefore changing, for example, the distributional assumptions on $p(\theta_1 \mid \theta_4)$ will not affect the form of $q^*(\theta_2)$.

The locality property of MFVB ensures that attention can be restricted to the simplest versions of the models of interest and in particular to the response distribution, knowing that the forms of the optimal densities preserve the same structure also in larger models. This property and the factor graph fragment perspective employed in variational message passing have motivated our focus on the simplest forms of likelihood fragments described in Chapter 3.

1.3.2 Variational message passing on factor graphs

Variational message passing arrives at variational Bayes approximation via message passing on an appropriate factor graph. A *factor graph* is a graphical representation of the argument groupings of a real-valued function. Consider, for example, the regression model

$$\begin{aligned} \boldsymbol{y} \,|\, \boldsymbol{\beta}, \sigma^2 &\sim N\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}\right), \quad \boldsymbol{\beta} \sim N\left(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right), \\ \sigma^2 \,|\, a &\sim \text{Inverse-}\chi^2\left(1, 1/a\right), \quad a \sim \text{Inverse-}\chi^2\left(1, 1/A^2\right), \end{aligned} \tag{1.12}$$

where $\boldsymbol{y}$ is an $n \times 1$ vector of response data and $\boldsymbol{X}$ is an $n \times d$ design matrix. The $d \times 1$ vector $\boldsymbol{\mu}_{\boldsymbol{\beta}}$, the $d \times d$ covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ and $A > 0$ are user-specified hyperparameters. A factor graph representation of this model based on the product density restriction

$$q\left(\beta, \sigma^2, a\right) = q\left(\boldsymbol{\beta}\right) q\left(\sigma^2\right) q\left(a\right) \tag{1.13}$$

is that of Figure 1.2.

FIGURE 1.2: Factor graph for the regression model in (1.12) and restriction (1.13).

Each factor graph has a corresponding graphical representation based on nodes connected by edges. The word *node* is used for both a stochastic node $\boldsymbol{\theta}_i$, $1 \leq i \leq M$, and a factor f_j, $1 \leq j \leq N$. In detail, the shaded squares correspond to *factors*, which are single product components of the real-valued function. The unshaded circles are called *stochastic nodes* and refer to parameters expressing product dependencies in the approximating density. An *edge* connects a factor to the stochastic nodes included in that factor. Two nodes are neighbors of each other if they are joined by an edge. In Figure 1.2, stochastic nodes $\boldsymbol{\beta}$ and σ^2 are *neighbors* of the factor $p\left(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \sigma^2\right)$, for instance. We denote by neighbors (j) the θ_i indices connected to the jth factor.

Rather than using result (1.8), the VMP procedure is founded upon the notion of *messages* passed between any two neighboring nodes, which are a particular function of the stochastic node that either sends or receives the message. Among the several variants of VMP in the literature, we follow the approach of Minka (2005), which is described in Section 2.5 of Wand (2017). The approach is here briefly summarized.

Let N be the number of factors. For each $1 \leq i \leq M$ and $1 \leq j \leq N$, the VMP *stochastic node to factor* message updates are

$$m_{\boldsymbol{\theta}_i \to f_j}(\boldsymbol{\theta}_i) \longleftarrow \prod_{j' \neq j : i \in \text{neighbors}(j')} m_{f_{j'} \to \boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) \tag{1.14}$$

and the *factor to stochastic node* message updates have form

$$m_{f_j \to \boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) \longleftarrow \exp\left[E_{f_j \to \boldsymbol{\theta}_i}\left\{\log f_j\left(\boldsymbol{\theta}_{\text{neighbors}(j)}\right)\right\}\right], \tag{1.15}$$

with $E_{f_j \to \boldsymbol{\theta}_i}$ denoting expectation with respect to the density function

$$\frac{\prod_{i' \in \text{neighbors}(j) \setminus \{i\}} m_{f_j \to \boldsymbol{\theta}_{i'}}(\boldsymbol{\theta}_{i'})\, m_{\boldsymbol{\theta}_{i'} \to f_j}(\boldsymbol{\theta}_{i'})}{\prod_{i' \in \text{neighbors}(j) \setminus \{i\}} \int m_{f_j \to \boldsymbol{\theta}_{i'}}(\boldsymbol{\theta}_{i'})\, m_{\boldsymbol{\theta}_{i'} \to f_j}(\boldsymbol{\theta}_{i'})\, d\boldsymbol{\theta}_{i'}}. \tag{1.16}$$

If $\text{neighbors}(j) \setminus \{i\} = \emptyset$, then the expectation in (1.15) can be dropped and the right-hand side of (1.15) is proportional to $f_j\left(\boldsymbol{\theta}_{\text{neighbors}(j)}\right)$. In general, the optimal q-densities are obtained from

$$q^*(\boldsymbol{\theta}_i) \propto \prod_{j : i \in \text{neighbors}(j)} m^*_{f_j \to \boldsymbol{\theta}_i}(\boldsymbol{\theta}_i), \tag{1.17}$$

where $m^*_{f_j \to \boldsymbol{\theta}_i}(\boldsymbol{\theta}_i)$ are the optimal messages at convergence. Formally, convergence of the message updates may be assessed by monitoring at each iteration the lower bound of the marginal log-likelihood

$$\log \underline{p}(q; \boldsymbol{y}) = \sum_{i=1}^{M} E_{q(\boldsymbol{\theta}_i)}\left\{\log q(\boldsymbol{\theta}_i)\right\} + \sum_{j=1}^{N} E_{q(\boldsymbol{\theta}_i)}\left\{\log(f_j)\right\},$$

where the first component is known as the *entropy* or *differential entropy*. In practice, derivation of the lower bound expression may be cumbersome when the approximating densities $q(\boldsymbol{\theta}_i)$s belong to non-standard exponential families. Alternatively, convergence may be determined tracking the parameters of the approximating densities. In the present thesis, we choose the latter strategy.

We also define the notation

$$\boldsymbol{\eta}_{f \leftrightarrow \boldsymbol{\theta}} = \boldsymbol{\eta}_{f \to \boldsymbol{\theta}} + \boldsymbol{\eta}_{\boldsymbol{\theta} \to f}$$

for any natural parameter $\boldsymbol{\eta}$, factor f and stochastic node $\boldsymbol{\theta}$, in support of Chapters 3 and 4. Natural parameters arise from the theory of exponential families described in Section A.2.1 of Appendix A.

1.4 Variational approximations and message passing

One way to orient oneself inside the vast literature on variational approximations is to understand the physical interpretation of variational methods. Starting from the original concepts of mean field approximations derived in statistical physics, we contextualize the variational message passing approach inside a broader class of message passing algorithms.

1.4.1 The origins of mean field approximations

Peterson and Anderson (1987) describe what is arguably the first variational procedure for a particular model, a neural network. The mean field approximation has its roots in physics, where it is known as *mean field theory*. Later, it was then extended to inference in graphical models (e.g. Jordan *et al.*, 1999; Attias, 1999). Subsequently, it gradually penetrated into conventional statistical literature (e.g. Teschendorff *et al.*, 2005; McGrory and Titterington, 2007).

Parisi (1988) is usually cited in the statistics literature as a key reference to the origins of mean field theory in physics. The book treats statistical field theory and is built upon the principles of statistical mechanics, nevertheless it includes interesting analogies with statistics that we highlight here.

Statistical mechanics is involved with deriving thermodynamic properties of macroscopic bodies starting from a description of the motion of microscopic components such as atoms, electrons, etc. Classical mechanics would approach this problem by defining a Hamiltonian system that describes the evolution of the dynamic physical system and initial conditions. Since this problem formulation concerning the physical system is particularly complex, probabilistic methods are introduced. The problem can be treated in two steps:

(1) finding the probability distribution of the microscopic components in thermal equilibrium, for example, after a sufficiently long time;

(2) compute the macroscopic properties of the system given the microscopic probability distribution.

For the rest of this subsection, we preserve some of the classic physics notation and terminology to convey the main concepts and maintain a direct reference to original sources such as Parisi (1988) and Yedidia *et al.* (2005).

Consider a system of N particles, each having a state x_i. The overall state of the system $\boldsymbol{x} = \{x_1, \ldots, x_N\}$ has a corresponding *energy* $E(\boldsymbol{x})$. In thermal equilibrium, the probability of a state will be given by the *Bolthanzmann Law*

$$p(\boldsymbol{x}) = \frac{1}{Z(T)} e^{-E(\boldsymbol{x})/T},$$

where T is the temperature and $Z(T)$ is the normalization constant (*partition function*)

$$Z(T) = \sum_{\boldsymbol{x} \in S} e^{-E(\boldsymbol{x})/T},$$

with S the space of all possible states $\boldsymbol{x}$ of the system.

The *Helmholts free energy* F_H of a system is

$$F_H = -\log Z.$$

Physicists have devoted considerable effort to developing techniques which give good approximations to F_H. The variational approach is based on a trial probability distribution $b(\boldsymbol{x})$ and a corresponding *variational free energy* (*Gibbs free energy*)

$$F(b) = U(b) - H(b),$$

where $U(b)$ is the *variational average energy*

$$U(b) = \sum_{\boldsymbol{x} \in S} b(\boldsymbol{x}) E(\boldsymbol{x})$$

and $H(b)$ is the *variational entropy*

$$H(b) = -\sum_{\boldsymbol{x} \in S} b(\boldsymbol{x}) \log b(\boldsymbol{x}).$$

It follows that

$$F(b) = F_H + \mathrm{KL}(b \| p),$$

where

$$\mathrm{KL}(b \| p) = \sum_{\boldsymbol{x} \in S} b(\boldsymbol{x}) \log \frac{b(\boldsymbol{x})}{p(\boldsymbol{x})}$$

is the Kullback-Leibler divergence between $b(\boldsymbol{x})$ and $p(\boldsymbol{x})$, which drives the choice of the optimal trial probability distribution $b(\boldsymbol{x})$. Since the Kullback–Leibler divergence is always non-negative, $F(b) \geq F_H$, with equality when $b(\boldsymbol{x}) = p(\boldsymbol{x})$. Minimizing the

variational free energy $F(b)$ with respect to the trial probability function $b(\boldsymbol{x})$ can be intractable for large N. A possible solution is to impose a mean field restriction to $b(\boldsymbol{x})$ such that

$$b(\boldsymbol{x}) = \prod_{i=1}^{N} b_i(x_i),$$

which is analogous to the assumption on the approximating density at the base of statistical mean field algorithms.

1.4.2 Message passing algorithms

The literature on variational approximation methods is not limited to the approximation by Kullback–Leibler divergence. Alternative divergences may be hard to optimize but give better approximations (Minka, 2005). Nevertheless, one general way to understand this class of algorithms is to view their cost functions as the aforementioned free-energy functions from statistical physics (Yedidia *et al.*, 2005; Heskes, 2003). From this viewpoint, each algorithm arises as a different way to approximate the entropy of a distribution, in a similar way as in Subsection 1.4.1.

Message passing is a method for fitting variational approximations including several variants, each minimizing a different cost function with different message equations. Minka (2005) presents a unifying view of message passing algorithms as a class of methods that differ only by the divergence they minimize. From such a perspective, the ensemble of message passing techniques may include the following approaches:

- Loopy belief propagation (Frey and MacKay, 1998);
- Expectation propagation (Minka, 2001);
- Fractional belief propagation (Wiegerinck and Heskes, 2003);
- Power expectation propagation (Minka, 2004).
- Variational message passing (Winn and Bishop, 2005);
- Tree-reweighted message-passing (Wainwright *et al.*, 2005).

Minka (2005) describes the behavior of different message-passing algorithms through the illustration of divergence measure properties, with a particular focus on the α-divergence, a generalization of the Kullback–Leibler divergence indexed by $\alpha \in \mathbb{R} \setminus \{0; 1\}$. Given a density p and an approximating density q, the α-divergence measure between

p and q is

$$D_\alpha\left(p \,\|\, q\right) = \frac{\int \left\{\alpha p\left(x\right) + \left(1-\alpha\right) q\left(x\right) - p\left(x\right)^\alpha q\left(x\right)^{1-\alpha}\right\} dx}{\alpha\left(1-\alpha\right)}.$$

This divergence measure corresponds to $KL\left(q \,\|\, p\right)$ for $\alpha \to 0$ and $KL\left(p \,\|\, q\right)$ when $\alpha \to 1$.

Wainwright and Jordan (2008) point out that also other deterministic algorithms, such as the *sum-product algorithm* (e.g. Kschischang *et al.*, 2001) and *semi-denite relaxations* based on *Lasserre sequences* (e.g. Lasserre, 2001), can be couched within the variational methodology framework. Li and Turner (2016) propose a generalized variational inference approach deriving a lower bound via the *Rényi divergence*.

1.5 Theory

Research on the accuracy of the variational approximation is lacking in the computer science literature. This provides statistical sciences with an opportunity to develop interdisciplinary contributions with quantitative performance assessments. Research into the quality of the variational approximation for specific models is present in the statistical literature, but offers extensions towards several directions.

Wang and Titterington (2003) study the consistency properties of variational Bayesian estimators for mixture models of known densities. It was shown that, with probability 1, the proposed algorithm converges locally to the maximum likelihood estimator when iterations approach infinity. Wang and Titterington (2006) describe a general algorithm for computing variational Bayesian estimates and study its convergence properties for a normal mixture model.

Hall *et al.* (2011a) and Hall *et al.* (2011b) use a Gaussian variational approximation to estimate the parameters of a simple Poisson mixed-effects model with a single predictor and a random intercept. They prove consistency of these estimates, provide rates of convergence and show asymptotic normality with asymptotically valid standard errors. Ormerod and Wand (2012) extend with heuristic arguments these results for the consistency of Gaussian variational approximations for more general generalized linear mixed models.

Wang and Blei (2018) describe frequentist consistency and asymptotic normality of variational Bayes methods. Specifically, they connect variational Bayes methods to point estimates based on variational approximations. Zhang and Gao (2018) study

convergence rates of variational posterior distributions for nonparametric and high-dimensional inference, with a focus on the variational Bayes methodology.

Jordan (2004), Titterington (2004) and Blei *et al.* (2017) indicate further sources for a comprehensive listing of other relevant literature on the accuracy of the variational approximation.

1.6 Semiparametric regression

Variational methods may be proposed as a fast and effective tool for handling semiparametric regression models under both the frequentist and Bayesian perspectives. Furthermore, the notion of message passing can be used to streamline the algebra and computer coding for approximate inference in large Bayesian semiparametric regression models, taking advantage of sparse structures of design and covariance matrices. This motivates a brief overview of semiparametric models.

Parametric models such as linear, linear mixed, generalized linear or generalized linear mixed models use a particular functional form of predictors to explain the mean response. Such parametric assumption may offer a simple and interpretable representation of the relationship between response and predictors but it might not be suitable for circumstances in which the mean response is scarcely interpretable as a known function of predictors.

Semiparametric regression extends classical parametric regression analysis to allow the treatment of nonlinear predictor components. This extension can be achieved through penalized basis functions such as, for instance, B-splines or Daubechies wavelets and random effects modelling analogous to the classical longitudinal and multilevel analysis. Consequently, the general framework of mixed models offers a tailored infrastructure also for fitting and inference of semiparametric regression models. Furthermore, in the Bayesian settings, semiparametric regression finds a corresponding directed acyclic graphical model representation which supports the derivation of MCMC and scalable MFVB algorithms.

The variational methodological studies and applications of this thesis concerning semiparametric regression make use of the mixed model representation and O'Sullivan splines, which are an immediate generalization of smoothing splines based on penalized B-splines basis functions (e.g. O'Sullivan, 1986; Green and Silverman, 1994). The classical smoothing splines involve a number of basis functions which approximately equals the sample size. A spline of order k is a continuous piecewise polynomial with

continuous derivatives up to order $k-1$. O'Sullivan penalised splines possess the attractive feature of requiring remarkably fewer basis functions. Furthermore, their natural boundary conditions (e.g. Green and Silverman, 1994, p. 12), computational numerical stability and smoothness make them of particular interest in the spline-based semiparametric literature as well as a reliable choice of basis for standard statistical software implementations.

Wand and Ormerod (2008) provide a detailed description of O'Sullivan penalised splines and their mixed model representation, including examples with `R` code. We briefly summarize the approach here with an example.

1.6.1 Semiparametric regression via O'Sullivan penalized splines

Consider the simple nonparameteric regression setting

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

where $(x_i, y_i) \in \mathbb{R} \times \mathbb{R}$ and ε_is are random variables with $E(\varepsilon_i) = 0$ and variance σ_ε^2. Suppose we are interested in estimating f over the interval $[a, b]$ containing the $x_i s$ via a set of cubic B-spline basis functions $\boldsymbol{B}_x = [B_1(x), \ldots, B_{K+4}(x)]$, for $K \leq n$. The corresponding knot sequence is defined by $a = \kappa_1 = \kappa_2 = \kappa_3 = \kappa_4 < \kappa_5 < \cdots < \kappa_{K+4} < \kappa_{K+5} = \kappa_{K+6} = \kappa_{K+7} = \kappa_{K+8} = b$, where the actual values of the additional knots beyond the boundary are arbitrary and it is customary to make them all the same and equal to a and b, respectively (e.g. Hastie *et al.*, 2009, Chapter 5). We require a function that minimizes the *penalized residual sum of squares* (PRSS)

$$PRSS(f, \lambda) = \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \int_a^b f''(x)^2 \, dx, \tag{1.18}$$

The expression $\lambda \int_a^b f''(x)^2 \, dx$ is the so-called *penalty term* because it penalizes fits that are too rough, thus yielding a smoother result. The amount of smoothing is controlled by $\lambda > 0$, where λ is usually referred to as a *smoothing parameter*. The case $\lambda = 0$ corresponds to the unconstrained problem. The solution to (1.18) is the O'Sullivan penalized spline $f(\boldsymbol{x}) = \boldsymbol{B}\boldsymbol{\nu}$ and thus (1.18) can be rewritten as

$$PRSS(\boldsymbol{\nu}, \lambda) = (\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\nu})^T (\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\nu}) + \lambda \boldsymbol{\nu}^T \boldsymbol{\Omega} \boldsymbol{\nu}, \tag{1.19}$$

where $\boldsymbol{B}$ is the *design matrix* with $B_{ik} = B_k(x_i)$ and $\boldsymbol{\Omega}$ is the *penalty matrix* with $\Omega_{kk'} = \int_a^b B_k''(x) B_{k'}''(x)\,dx$. Straightforward algebraic manipulation leads to the following O'Sullivan penalized spline with a solution to (1.19) such that

$$\hat{f}(\boldsymbol{x}) = \boldsymbol{B}\hat{\boldsymbol{\nu}} \quad \text{and} \quad \hat{\boldsymbol{\nu}} = \left(\boldsymbol{B}^T\boldsymbol{B} + \lambda\boldsymbol{\Omega}\right)^{-1}\boldsymbol{B}^T\boldsymbol{y}. \tag{1.20}$$

In the special case in which the interior knots coincide with the x_is, assumed distinct, $\hat{f}(\boldsymbol{x})$ corresponds to the cubic smoothing spline arising as the minimizer of (1.18) (e.g. Schoenberg, 1964). These results also generalize to order m smoothing splines. Computation of the design matrix $\boldsymbol{B}$ is straightforward and is readily available in the R environment. However, computation of the penalty matrix $\boldsymbol{\Omega}$ can be challenging. In Section 6 of Wand and Ormerod (2008), an exact matrix expression for $\boldsymbol{\Omega}$ is derived by applying the Simpson's rule over each of the inter-knot differences, given as

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

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

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

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

$$\boldsymbol{\omega} = \left\{\frac{1}{6}(\Delta\boldsymbol{\kappa})_1, \frac{2}{3}(\Delta\boldsymbol{\kappa})_1, \frac{1}{6}(\Delta\boldsymbol{\kappa})_1, \frac{1}{6}(\Delta\boldsymbol{\kappa})_2, \frac{2}{3}(\Delta\boldsymbol{\kappa})_2, \frac{1}{6}(\Delta\boldsymbol{\kappa})_2, \ldots, \right.$$
$$\left. \frac{1}{6}(\Delta\boldsymbol{\kappa})_{K+7}, \frac{2}{3}(\Delta\boldsymbol{\kappa})_{K+7}, \frac{1}{6}(\Delta\boldsymbol{\kappa})_{K+7}\right\},$$

where $(\Delta\boldsymbol{\kappa})_k = \kappa_{k+1} - \kappa_k, 1 \le k \le K+7$. A common default choice for the number of knots is

$$K = \min(n_U/4, 35), \tag{1.21}$$

where n_U is the number of unique $x_i s$ and the distribution of knots can either be quantile-based or equally spaced (e.g. Ruppert *et al.*, 2003). In the next section we show how the O'Sullivan penalized splines can be expressed within the mixed model and Bayesian hierarchical model framework.

The positive penalization constant λ in (1.18) remains unspecified. An appropriate value for the constant λ trades the loss term given by the residual sum of squares against the penalty term. As pointed out in Luts and Ormerod (2014) this restricts

the space of solutions, reduces the overfitting effect and allows the extension to new data. Penalized likelihood is a fast but potentially unstable approach for estimation and inference with penalty parameter selection. Popular techniques for tuning the penalty parameter include cross-validation and random sampling methods, but these approaches increase the overall computational overhead. Embedding a model into a mixed effect framework is a convenient and more stable way for choosing the penalty parameter which also allows for automatic selection. However, it may involve intractable integration.

1.6.2 Mixed model representation

Semiparametric regression can be couched within the mixed model infrastructure exploiting the inferential equivalence between penalized likelihood models and mixed model representation (e.g. Ruppert *et al.*, 2003). Consider the regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon}, \quad \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \sigma_u^2 \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_\varepsilon^2 \boldsymbol{I} \end{bmatrix} \right), \tag{1.22}$$

for the general design matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$. The value $\hat{\boldsymbol{\nu}}$ appearing in (1.20) can be conveniently expressed using the least squares estimator to (1.22), which is equivalent to the *best linear unbiased predictor* (BLUP) of $\boldsymbol{\beta}$ and $\boldsymbol{u}$ given as

$$\hat{\boldsymbol{\nu}} = \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{u}} \end{bmatrix} = \left(\boldsymbol{C}^T\boldsymbol{C} + \lambda \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \right)^{-1} \boldsymbol{C}^T\boldsymbol{y}, \tag{1.23}$$

where $\lambda = \sigma_u^2/\sigma_\varepsilon^2$ is the smoothing parameter and $\boldsymbol{C} = [\boldsymbol{X}\ \boldsymbol{Z}]$ (e.g. Ruppert *et al.*, 2003, Section 4.5.3). The equivalence of (1.20) and (1.23) can be achieved if there exists a $(K+4) \times (K+4)$ linear transformation matrix $\boldsymbol{L}$ such that

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

Wand and Ormerod (2008) affirm that a common tool to obtain $\boldsymbol{L}$ and $\boldsymbol{\Omega}$ is spectral decomposition and provide the resultant forms

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

where $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$, $\boldsymbol{d}$ is a $(K+4) \times 1$ vector with exactly two zero entries and all others positive, $\boldsymbol{d}_Z$ is a $(K+2) \times 1$ subvector of $\boldsymbol{d}$ containing the positive entries and $\boldsymbol{U}_Z$ is the $(K+4) \times (K+2)$ submatrix of $\boldsymbol{U}$ with columns corresponding to the positive entries

of $\boldsymbol{d}$. It follows that O'Sullivan penalized splines can be used for fitting (1.22) defining the design matrices

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

Chapter 2

Variational inference for general design generalized linear mixed models

2.1 Introduction

In this chapter we introduce a new framework for estimation and inference for general design generalized linear mixed models (GLMMs) based on Gaussian variational approximations. This involves approximating the distributions of random effects vectors, given the responses, by the Gaussian distribution minimizing the Kullback–Leibler divergence. Standard errors for fixed effects and covariance parameter estimates are an outgrowth of the optimization algorithm. For the random effects, approximate best predictions and prediction variances also arise from the GVA procedure.

Ormerod and Wand (2012) devise an effective variational approximation strategy for fitting GLMMs appropriate for grouped data, affirming that a future challenge is to treat more general GLMMs. We extend their work providing a more general framework to handle GLMMs containing, for instance, spline basis functions in the random effects design matrix or spatial correlation structures.

A recent and parallel work by Hui *et al.* (2018) employs GVA for inference on generalized additive models. Fully tractable variational likelihoods for some common response types are proposed, offering a framework for inference on parametric components and a closed-form approach for smoothing parameter selection. The simulation studies in Hui *et al.* (2018) show the variational approximation approach performs similarly to and sometimes better than software for fitting generalized additive models that are currently in use.

To our knowledge, our work is the first to employ GVA to treat general design generalized linear mixed models. Without loss of generality, we consider two common Poisson and Bernoulli response types and taking advantage of the mixed model representation of semiparametric regression models, we make use of tractable variational likelihood lower bounds. Differently from Hui *et al.* (2018), residual unidimensional integration is present in our lower bounds for the Bernoulli case. However, residual integration can be easily performed via Gauss–Hermite quadrature, which we prefer to a fully tractable, but less accurate, lower bound.

Section 2.2 provides an overview on general design GLMMs, a framework that allows for inclusion of random intercepts and slopes, spline basis functions and spatial correlation structures, for example. Section 2.3 introduces GVA for the treatment of GLMMs, while Section 2.4 is dedicated to the optimization of the variational lower bound. Section 2.5 explains how to obtain approximate standard errors and best prediction of random effects from the optimization procedure. Illustrations and simulation studies are included in Sections 2.6 and 2.7.

2.2 General design GLMMs

Generalized linear models (GLMs), introduced by Nelder and Wedderburn (1972) and systematically formalized by McCullagh and Nelder (1989), permit to analyze the relationship between a response variable and covariates via a linear functional form, given a link function. However, GLM may not be flexible enough when analyzing more complex situations. Generalized additive models (GAMs) extend the GLM framework replacing linear components by a sum of smooth unknown functions of predictor variables (Hastie and Tibshirani, 1990). The mixed model versions of GLMs and GAMs are valid alternatives when one believes that the relationship between the dependent variable and covariates is better explained if the linear predictor includes random effects in addition to the fixed effects. In this sense, the class of generalized linear mixed models is extremely rich allowing, for example, varying coefficient models, cross random effects, nested random effects, spline-type smoothing, additive and semiparametric components or spatial components. The key is to devise a random effects design matrix with a general structure.

Zhao *et al.* (2006) suggest a factorization of fixed and random effects structures for describing the fitting of general design generalized linear mixed models. In particular, it is important to separate out random effects structures for handling grouping, longitudinal data, smoothing regression or spatial models, for example.

In detail, Zhao *et al.* (2006) propose the following breakdown:

$$\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} = \boldsymbol{X}^R\boldsymbol{\beta}^R + \boldsymbol{Z}^R\boldsymbol{u}^R + \boldsymbol{X}^G\boldsymbol{\beta}^G + \boldsymbol{Z}^G\boldsymbol{u}^G + \boldsymbol{Z}^C\boldsymbol{u}^C. \tag{2.1}$$

The decomposition is not necessarily unique for a particular model. However, the matrices

$$\boldsymbol{X}^R = \begin{bmatrix} \boldsymbol{X}_1^R \\ \vdots \\ \boldsymbol{X}_m^R \end{bmatrix}, \quad \boldsymbol{Z}^R = \underset{1\leq i\leq m}{\text{blockdiag}}\left(\boldsymbol{X}_i^R\right)$$

and

$$\text{Cov}\left(\boldsymbol{u}^R\right) = \underset{1\leq i\leq m}{\text{blockdiag}}\left(\boldsymbol{\Sigma}_i^R\right) = \boldsymbol{I}_m \otimes \boldsymbol{\Sigma}^R$$

can be intended to model random intercepts and slopes, as in the case of repeated measures data on m groups with sample sizes $n_1, \ldots, n_m$, with $\boldsymbol{X}_i^R$ an $n_i \times q^R$ design matrix corresponding to the ith group and $\boldsymbol{\Sigma}^R$ a $q^R \times q^R$ covariance matrix.

The matrices $\boldsymbol{X}^G$ and $\boldsymbol{Z}^G$ are general design matrices of different form than those arising in random effects models. For example, $\boldsymbol{X}^G$ may contain indicator variables or polynomial basis functions of a continuous predictor, while $\boldsymbol{Z}^G$ may include spline basis functions. The $\boldsymbol{Z}^G\boldsymbol{u}^G$ term may be further decomposed as

$$\boldsymbol{Z}^G\boldsymbol{u}^G = \sum_{\ell=1}^{L} \boldsymbol{Z}_\ell^G\boldsymbol{u}_\ell^G$$

to allow, for instance, additive model formulations. In a spline penalization context, a possible way to model the corresponding covariance form is

$$\text{Cov}\left(\boldsymbol{u}^G\right) = \underset{1\leq \ell\leq L}{\text{blockdiag}}\left(\sigma_{u\ell}^2\boldsymbol{I}\right).$$

The $\boldsymbol{Z}^C\boldsymbol{u}^C$ component may represent other types of random effects, such as those with spatial correlation structure.

2.2.1 Overview of software implementations

GLMMs are very popular models that have implementations in standard software, including SAS, Stata, and `R` (R Core Team, 2018). More in detail, several `R` packages support GLMM analysis. Many of them, e.g. `MASS` (Ripley, 2012), `gamm4` (Wood and Scheipl, 2017) and `mgcv` (Wood, 2018), use Laplace approximation, but a few, e.g. `glmmBUGS` (Brown and Zhou, 2018), `MCMCglmm` (Hadfield, 2018), `R2BayesX` (Umlauf *et al.*,

2017) and `spikeSlabGAM` (Scheipl and Gruen, 2017), use MCMC methods. Exact maximum likelihood-based inference via quadrature, is supported by `glmmML` (Broström, 2018) and `lme4` (Bates *et al.*, 2018). For the generalized additive mixed model (GAMM) extension, only approximate inference is so far available and it is supported by `gamm4`, `mgcv`, `R2BayesX`, `spikeSlabGAM` and `gammSlice` (Pham and Wand, 2018). Laplace approximation is used by the first two of the these packages. The third and fourth perform GAMM fitting via MCMC. The latter is the only `R` package that provides MCMC based inference for GAMM analyses using the same penalized spline approach that `mgcv` employs.

2.3 GVA for GLMMs

As a starting point to the introduction of Gaussian variational approximations for inference and fitting, consider GLMMs for canonical one-parameter exponential families with Gaussian random effects, taking the general form

$$\boldsymbol{y} \,|\, \boldsymbol{u} \sim \exp\left\{\boldsymbol{y}^T\left(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u}\right)-\mathbf{1}^T b\left(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u}\right)+\mathbf{1}^T c\left(\boldsymbol{y}\right)\right\}, \quad \boldsymbol{u} \sim N\left(\mathbf{0},\boldsymbol{G}\right), \tag{2.2}$$

where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are general design matrices and $\boldsymbol{G}$ is the random effect covariance matrix for the random vector $\boldsymbol{u}$ of length K. The component $\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u}$ can be designed according to the breakdown (2.1). The functions b and c characterize the members of the family. For the cases considered here, we have that $b(x) = e^x$ for Poisson responses or $b(x) = \log(1+e^x)$ in the logistic case. The treatment of other response types goes beyond the illustrative purpose of this thesis. However, the framework proposed here can be easily extended to treat other cases, ranging from normal responses to more elaborate ones. We also use p to denote densities of random vectors and sometimes suppress dependence on parameters to shorten the notation. For example, the log-likelihood function can be written in terms of the joint density of $\boldsymbol{y}$ as $\ell\left(\boldsymbol{\beta},\boldsymbol{G}\right) = \log p\left(\boldsymbol{y};\boldsymbol{\beta},\boldsymbol{G}\right) = \log p\left(\boldsymbol{y}\right)$.

The log-likelihood function coresponding to model (2.2) is

$$\begin{aligned}
\ell\left(\boldsymbol{\beta},\boldsymbol{G}\right) =& \log p\left(\boldsymbol{y};\boldsymbol{\beta},\boldsymbol{G}\right) \\
=& \log \int p\left(\boldsymbol{y} \,|\, \boldsymbol{u};\boldsymbol{\beta}\right) p\left(\boldsymbol{u};\boldsymbol{G}\right) d\boldsymbol{u} \\
=& \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} + \mathbf{1}^T c\left(\boldsymbol{y}\right) - \frac{1}{2}\log\left|\boldsymbol{G}\right| - \frac{K}{2}\log\left(2\pi\right) \\
&+ \log \int \exp\left\{\boldsymbol{y}^T\boldsymbol{Z}\boldsymbol{u} - \mathbf{1}^T b\left(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u}\right) - \frac{1}{2}\boldsymbol{u}^T\boldsymbol{G}^{-1}\boldsymbol{u}\right\} d\boldsymbol{u}.
\end{aligned}$$

Maximization of $\ell(\boldsymbol{\beta}, \boldsymbol{G})$ is hindered by the presence of K dimensional integration, which cannot be solved analytically.

As an alternative, we may resort to variational approximations and make inference on a lower bound in the form of (1.5), which is more tractable than the original function of interest. This produces the following log-likelihood lower bound on $\ell(\boldsymbol{\beta}, \boldsymbol{G})$:

$$\underline{\ell}(\boldsymbol{\beta}, \boldsymbol{G}, \boldsymbol{\xi}; q) = \int q(\boldsymbol{u}; \boldsymbol{\xi}) \log \left\{ \frac{p(\boldsymbol{y}, \boldsymbol{u}; \boldsymbol{\beta}, \boldsymbol{G})}{q(\boldsymbol{u}; \boldsymbol{\xi})} \right\} d\boldsymbol{u}, \tag{2.3}$$

where $q(\boldsymbol{u}; \boldsymbol{\xi})$ is an arbitrarily density function in $\mathbb{R}^K$, depending on the variational parameters $\boldsymbol{\xi}$. The lower-bound expression (2.3) can be derived using the ideas of Kullback–Leibler divergence and decomposing the log-likelihood function as

$$\begin{aligned}
\ell(\boldsymbol{\beta}, \boldsymbol{G}) &= \int q(\boldsymbol{u}; \boldsymbol{\xi}) \log \left\{ \frac{p(\boldsymbol{y}, \boldsymbol{u}; \boldsymbol{\beta}, \boldsymbol{G}) / q(\boldsymbol{u}; \boldsymbol{\xi})}{p(\boldsymbol{u} \mid \boldsymbol{y}) / q(\boldsymbol{u}; \boldsymbol{\xi})} \right\} d\boldsymbol{u} \\
&= \int q(\boldsymbol{u}; \boldsymbol{\xi}) \log \left\{ \frac{p(\boldsymbol{y}, \boldsymbol{u}; \boldsymbol{\beta}, \boldsymbol{G})}{q(\boldsymbol{u}; \boldsymbol{\xi})} \right\} d\boldsymbol{u} + \int q(\boldsymbol{u}; \boldsymbol{\xi}) \log \left\{ \frac{q(\boldsymbol{u}; \boldsymbol{\xi})}{p(\boldsymbol{u} \mid \boldsymbol{y})} \right\} d\boldsymbol{u}.
\end{aligned}$$

Since the last term is the Kullback–Leibler discrepancy between $q(\boldsymbol{u}; \boldsymbol{\xi})$ and $p(\boldsymbol{u} \mid \boldsymbol{y})$, which is always non-negative, we get the lower bound (2.3). Note that $q(\boldsymbol{u}; \boldsymbol{\xi})$ may be interpreted as an approximation to the distribution of random effects given the responses, $p(\boldsymbol{u} \mid \boldsymbol{y})$, and setting $q(\boldsymbol{u}; \boldsymbol{\xi}) = p(\boldsymbol{u} \mid \boldsymbol{y})$ the lower bound corresponds to the log-likelihood function.

Proposition 2.1 provides an expression for the lower bound $\underline{\ell}(\boldsymbol{\beta}, \boldsymbol{G}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ when $q(\boldsymbol{u}; \boldsymbol{\xi})$ is chosen to be a normal distribution.

Proposition 2.1. *Setting $q(\boldsymbol{u}; \boldsymbol{\xi})$ to the $N(\boldsymbol{\mu}, \boldsymbol{\Lambda})$, the likelihood lower bound* (2.3) *takes the form*

$$\begin{aligned}
\underline{\ell}(\boldsymbol{\beta}, \boldsymbol{G}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = & \frac{K}{2} + \boldsymbol{y}^T (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\mu}) - \mathbf{1}^T B\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\mu}, dg\left(\boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{Z}^T\right)\right) + \mathbf{1}^T c(\boldsymbol{y}) \\
& - \frac{1}{2}\left\{\boldsymbol{\mu}^T \boldsymbol{G}^{-1}\boldsymbol{\mu} + tr\left(\boldsymbol{G}^{-1}\boldsymbol{\Lambda}\right)\right\} + \frac{1}{2}\log\left|\boldsymbol{G}^{-1}\boldsymbol{\Lambda}\right|,
\end{aligned} \tag{2.4}$$

where $B(\mu, \sigma^2) = \int_{-\infty}^{\infty} b(\mu + \sigma x)\, \phi(x)\, dx$. For vector arguments, function B is applied in element-wise fashion such that, for instance, $B\left(\begin{bmatrix} a_1 \\ a_2 \end{bmatrix}, \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}\right) = \begin{bmatrix} B(a_1, b_1) \\ B(a_2, b_2) \end{bmatrix}$.

A proof of Proposition 2.1 is given in Section B.1.1 of Appendix B. The lower bound $\underline{\ell}(\boldsymbol{\beta}, \boldsymbol{G}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ is the so called *Gaussian variational approximation* to $\ell(\boldsymbol{\beta}, \boldsymbol{G})$, since $q(\boldsymbol{u}; \boldsymbol{\xi})$ is assumed to be a normal density function, with variational parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$. The advantage of using $\underline{\ell}(\boldsymbol{\beta}, \boldsymbol{G}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ in replacement of $\ell(\boldsymbol{\beta}, \boldsymbol{G})$ is that the former

no longer involves K-dimensional integration but unidimensional integrals appearing in function $B(\mu, \sigma^2)$ at most. Such integrals are even expressible in closed form in the case of Poisson responses with canonical link function. In this case, $B(\mu, \sigma^2) = \exp(\mu + \frac{1}{2}\sigma^2)$. Ormerod and Wand (2012) suggest to treat this integration via adaptive Gauss–Hermite quadrature and provide details in their supplementary material. We follow this scheme for the integration steps in our illustrations involving Bernoulli responses.

Since $\ell(\boldsymbol{\beta}, \boldsymbol{G}) \geq \underline{\ell}(\boldsymbol{\beta}, \boldsymbol{G}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ for all $(\boldsymbol{\mu}, \boldsymbol{\Lambda})$, maximizing over the variational parameters implies pushing the lower bound towards the log-likelihood function. Therefore, we define a new optimization problem in lieu of log-likelihood maximization, which forms the base for variational inference. Let

$$\left(\widehat{\underline{\boldsymbol{\beta}}}, \widehat{\underline{\boldsymbol{G}}}, \widehat{\underline{\boldsymbol{\mu}}}, \widehat{\underline{\boldsymbol{\Lambda}}}\right) = \underset{\boldsymbol{\beta}, \boldsymbol{G}, \boldsymbol{\mu}, \boldsymbol{\Lambda}}{\operatorname{argmax}} \underline{\ell}(\boldsymbol{\beta}, \boldsymbol{G}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$$

then $\widehat{\underline{\boldsymbol{\beta}}}$ and $\widehat{\underline{\boldsymbol{G}}}$ are the Gaussian variational approximate maximum likelihood estimators for $\boldsymbol{\beta}$ and $\boldsymbol{G}$, respectively. The optimization problem requires the application of efficient computational procedures that are able to handle a large set of parameters, especially those entering covariance matrices.

One idea is to conduct the optimization procedure through the Newton-Raphson algorithm, which guarantees fast convergence, given that parameters are initialized "nearby" the optima.

2.4 Lower bound optimization

Despite several routines in the `R` computing environment support efficient optimization, an additional Newton-Raphson optimization step may provide improved estimates with fast convergence, if set close to the optima. An efficient Newton-Raphson scheme adapted to the Gaussian variational lower bound is described in Ormerod and Wand (2012, Section A.5).

Our optimization strategy includes the following three steps:

1. parameters initialization;
2. first optimization step via the `R` function `optim()` and "BFGS" (Broyden, Fletcher, Goldfarb and Shanno) quasi-Newton method;
3. second optimization step via Newton-Raphson, having as input the values from the previous optimization.

There are at least two issues that must be addressed in the lower bound optimization. The first is setting starting values that respect constraints, such as covariance matrices being symmetric and positive definite, and give rise to a converging optimization. The second involves the calculation of derivatives and Hessians of the lower bound (2.4), in order to apply a Newton-Raphson scheme. Section B.1.2 of Appendix B lists first order derivative and Hessian expressions for the lower bound (2.4). However, it should be pointed out that these expressions need to be adapted to single model specifications to guarantee a better optimization procedure. For instance, if the covariance matrix $\boldsymbol{G}$ is specified as $\boldsymbol{G} = \sigma^2\boldsymbol{I}$, than it makes sense to compute derivatives with respect to σ^2, rather than the whole matrix $\boldsymbol{G}$. Alternatively to tedious and somehow tricky derivatives calculations, the `R` package `TMB` Kristensen (2018) may be employed. This tool allows for efficient calculation of first and second order derivatives, taking as input the function to be optimized as a simple C++ template.

An important issue that arises in the optimization procedure over the variational parameters is to maintain the symmetric and positive-definite characteristics of the covariance matrix of the approximating Gaussian density and covariance matrices associated with random effects components. A straightforward solution when working with simple covariance matrices parameterized as, for instance, $\sigma^2\boldsymbol{I}$ is to work with the logarithm of the variance parameter, σ^2. For nontrivial cases, the solution we adopt is the one proposed in Pinheiro and Bates (2000, Subsection 2.2.7), which consists of parameterizing the covariance matrix through its *matrix logarithm*. Noting that any symmetric positive-definite matrix $\boldsymbol{A}$ can be expressed as the *matrix exponential* of another symmetric matrix $\boldsymbol{B}$, it is possible to rewrite $\boldsymbol{A}$ as

$$\boldsymbol{A} = e^{\boldsymbol{B}} = \boldsymbol{I} + \boldsymbol{B} + \frac{\boldsymbol{B}^2}{2!} + \frac{\boldsymbol{B}^3}{3!} + \cdots .$$

Given a real symmetric and positive-definite matrix $\boldsymbol{A}$ of dimension $q \times q$, one way to obtain its matrix logarithm $\boldsymbol{B}$ is to calculate the eigendecomposition

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{L}\boldsymbol{Q}^T,$$

where $\boldsymbol{L}$ is a $q \times q$ diagonal matrix whose main diagonal entries are the eigenvalues of $\boldsymbol{A}$ and $\boldsymbol{Q}$ is a $q \times q$ and orthogonal matrix whose columns are the eigenvectors of $\boldsymbol{A}$. Note that if $\boldsymbol{A}$ is positive-definite, then all the diagonal elements of $\boldsymbol{L}$ must be positive and we are able to define the matrix logarithm of $\boldsymbol{L}$, $\log \boldsymbol{L}$, as the diagonal matrix whose main diagonal elements are the logarithms of the corresponding elements of $\boldsymbol{L}$. Therefore we

get

$$\boldsymbol{B} = \log \boldsymbol{A} = \boldsymbol{Q} \log \boldsymbol{L}\boldsymbol{Q}^T.$$

Unconstrained optimization can be then performed via the matrix logarithm of the matrix of interest.

The covariance matrix of the approximating Gaussian density, $\boldsymbol{\Lambda}$, may have large dimensions, especially in applications to models with grouped data or spline basis functions in the design matrix. Some solutions have already been proposed in the literature. For instance, Ormerod and Wand (2012, Theorem 1) prove that the optimal $\boldsymbol{\Lambda}$ has a simplified block-diagonal form for grouped data GLMMs with same group size. Tan and Nott (2018) impose sparsity in the precision matrix to reflect appropriate conditional independence structures in the model, showing applications to GLMMs and state space models.

In our simulation studies in Section 2.6 we perform GVA with simplified $\boldsymbol{\Lambda}$ matrices reflecting the random effects covariance matrix structure, although without providing in this thesis formal results in support of our choices.

2.5 Approximate standard errors and best prediction of random effects

Ormerod and Wand (2012) suggest a strategy to obtain approximate standard errors and approximate best prediction of random effects that easily extends to a framework involving general design generalized linear mixed models.

Imagine to treat the lower bound as a log-likelihood function and $\left(\boldsymbol{\mu}^T, \text{vech}\left(\boldsymbol{\Lambda}\right)^T\right)$ as a vector of nuisance parameters, setting $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^T, \text{vech}\left(\boldsymbol{G}\right)^T\right)$. Then, via standard theory of inference we get the asymptotic covariance matrix of $\underline{\hat{\boldsymbol{\theta}}} = \left(\underline{\hat{\boldsymbol{\beta}}}^T, \text{vech}\left(\underline{\hat{\boldsymbol{G}}}\right)^T\right)$,

$$\widehat{\boldsymbol{\Sigma}}^{\text{asym}}_{\underline{\hat{\boldsymbol{\theta}}}} = \boldsymbol{\theta} \text{ sub-block of } \underline{\mathbf{I}}\left(\boldsymbol{\beta}, \boldsymbol{G}, \boldsymbol{\mu}, \boldsymbol{\Lambda}\right)^{-1}, \tag{2.5}$$

where $\underline{\mathbf{I}}\left(\boldsymbol{\beta}, \boldsymbol{G}, \boldsymbol{\mu}, \boldsymbol{\Lambda}\right) = E\left\{-\text{H}\underline{\ell}\left(\boldsymbol{\beta}, \boldsymbol{G}, \boldsymbol{\mu}, \boldsymbol{\Lambda}\right)\right\}$ is the *variational approximate Fisher information matrix* and H is the Hessian matrix operator with respect to $(\boldsymbol{\theta}, \boldsymbol{\mu}, \text{vech}\left(\boldsymbol{\Lambda}\right))$. Approximate standard errors can be obtained at convergence of the optimization procedure by extracting the square roots of the diagonal entries of $\widehat{\boldsymbol{\Sigma}}^{\text{asym}}_{\underline{\hat{\boldsymbol{\theta}}}}$.

The *best predictor* of $\boldsymbol{u}$

$$\text{BP}\left(\boldsymbol{u}\right) = E\left(\boldsymbol{u} \,|\, \boldsymbol{y}\right) = \int \boldsymbol{u} p\left(\boldsymbol{u} \,|\, \boldsymbol{y}\right) d\boldsymbol{u}$$

is often of interest, but in our case hindered by intractable integration. We may then resort to the fact that the optimal approximating density $q(\boldsymbol{u}; \boldsymbol{\mu}, \boldsymbol{\Lambda})$ is chosen to minimize the KullbackLeibler divergence with $p(\boldsymbol{u} \mid \boldsymbol{y})$ and propose the following approximation to $\mathrm{BP}(\boldsymbol{u})$:

$$\underline{\mathrm{BP}}(\boldsymbol{u}) = \int \boldsymbol{u} q\left(\boldsymbol{u}; \underline{\widehat{\boldsymbol{\mu}}}, \underline{\widehat{\boldsymbol{\Lambda}}}\right) d\boldsymbol{u} = \underline{\widehat{\boldsymbol{\mu}}}.$$

From best prediction theory (e.g. McCulloch *et al.*, 2008, Chapter 13) we have that

$$\mathrm{Cov}\left\{\mathrm{BP}(\boldsymbol{u}) - \boldsymbol{u}\right\} = E_{\boldsymbol{y}}\left\{\mathrm{Cov}(\boldsymbol{u} \mid \boldsymbol{y})\right\},$$

with $E_{\boldsymbol{y}}$ indicating expectation with respect to $\boldsymbol{y}$. Resorting again to the fact that $q(\boldsymbol{u}; \boldsymbol{\mu}, \boldsymbol{\Lambda})$ approximates $p(\boldsymbol{u} \mid \boldsymbol{y})$, we get the estimated asymptotic covariance matrix

$$\widehat{\boldsymbol{\Sigma}}^{\mathrm{asym}}_{\{\underline{\mathrm{BP}}(\boldsymbol{u})-\boldsymbol{u}\}} = \underline{\widehat{\boldsymbol{\Lambda}}}. \tag{2.6}$$

Proposition 2.2 associates the estimate for the variability of $\underline{\mathrm{BP}}$ to the variational approximate Fisher information matrix.

Proposition 2.2. *Let* $\mathrm{H}_{\boldsymbol{\mu}\boldsymbol{\mu}}\underline{\ell}$ *be the* $\boldsymbol{\mu}$ *block of* $\mathrm{H}\underline{\ell}(\boldsymbol{\beta}, \boldsymbol{G}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$. *Then*

$$\underline{\widehat{\boldsymbol{\Lambda}}} = \left(-\mathrm{H}_{\boldsymbol{\mu}\boldsymbol{\mu}}\underline{\ell}\right)^{-1}$$

According to Proposition 2.2, $\widehat{\boldsymbol{\Sigma}}^{\mathrm{asym}}_{\{\underline{\mathrm{BP}}(\boldsymbol{u})-\boldsymbol{u}\}}$ can be obtained as the $\boldsymbol{\mu}$ sub-block of the variational approximate Fisher information matrix $\underline{\mathbf{I}}(\boldsymbol{\beta}, \boldsymbol{G}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$, in a similar way as for fixed effects. Section B.1.3 of Appendix B provides a proof of Proposition 2.2.

2.5.1 Asymptotic properties

Hall *et al.* (2011a) and Hall *et al.* (2011b) examine a simple Poisson mixed-effects model with a single predictor and a random intercept. They use a GVA approach and estimate parameters with a variational expectation-maximization procedure. They prove consistency of the corresponding estimates at a certain parametric rate as well as asymptotic normality with asymptotically valid standard errors.

Starting from arguments similar to those used by Opper and Archambeau (2009) to show that the Laplace approximation and the GVA are closely related, Ormerod and Wand (2012) give heuristic arguments for the consistency of GVA for simple generalized linear mixed models.

Hui *et al.* (2018) consider a framework for semiparametric regression based on GVA. They demonstrate the consistency of the variational approximation estimates and asymptotic normality for the parametric component.

All these results and the somehow satisfactory performances of GVA highlighted by our next simulation studies, encourage to dedicate further research endeavor to formalize more general results.

2.6 Illustrative examples using simulated data

The general design GLMMs framework we operate in covers a vast range of situations and a complete description is impractical. Therefore we illustrate the application of GVA for some prominent models through simulated datasets, similarly to the description of the `R` package `gammSlice` provided in Pham and Wand (2018). The cases we consider are Poisson nonparametric regression, semiparametric logistic regression and a logistic additive mixed model.

We also provide a rough performance assessment via the comparison of GVA and MCMC results obtained with `rstan` (Stan Development Team, 2018). Such a comparison of frequentist and Bayesian inferential summaries may rise some philosophical criticisms. However, since Bayesian inference is here performed with diffuse priors, there is at least an informal sense that motivates this juxtaposition of the results. Fixed effects parameters, say $\boldsymbol{\beta}$, have priors $N\left(\mathbf{0}, \sigma_{\boldsymbol{\beta}}^2 \boldsymbol{I}\right)$, with $\sigma_{\boldsymbol{\beta}}^2 = 10^{10}$, while priors on standard deviations, say σ, are half Cauchy distributions with scale parameter $A_\sigma = 10^5$. Subsection B.1.4 of Appendix B displays the `rstan` code for fitting Poisson nonparametric regression via MCMC. The code can be easily adapted to fit the other GLMMs with little modification.

The lower bound optimization may benefit of parsimonious parametrization of $\boldsymbol{\Lambda}$. For example, Ormerod and Wand (2012) consider models for grouped data such as the Poisson random intercept model

$$\begin{aligned} y_{ij} &\overset{\text{ind.}}{\sim} \text{Poisson}\left(\exp\left(\beta_0 + \beta_1 x_{ij} + u_i\right)\right), \\ u_i &\overset{\text{ind.}}{\sim} N\left(0, \sigma^2\right), \quad 1 \leq i \leq m, \quad 1 \leq j \leq n. \end{aligned}$$

According to Ormerod and Wand (2012, Theorem 1), the optimal $\boldsymbol{\Lambda}$ takes form

$$\boldsymbol{\Lambda} = \underset{1 \leq i \leq m}{\text{blockdiag}}\left(\boldsymbol{\Lambda}_i\right),$$

where each $\mathbf{\Lambda}_i$ is a $n \times n$ positive definite matrix. Without providing any formal result, we can make similar assumptions on the form of matrix $\mathbf{\Lambda}$ to support efficient optimization in cases such as additive models, where the covariance matrix of random effects has a block-diagonal structure.

2.6.1 Poisson nonparametric regression

We simulate data from

$$y_i \overset{\text{ind.}}{\sim} \text{Poisson}\left(\exp\left\{f\left(x_i\right)\right\}\right), \quad 1 \leq i \leq n,$$

with $n = 500$, $x_i \overset{\text{ind.}}{\sim} \text{Uniform}\left(0, 1\right)$ and

$$f\left(x\right) = \cos\left(4\pi x\right) + 2x - 1.$$

The smooth function f is modeled using penalized splines and we estimate the Poisson nonparametric regression model

$$\begin{aligned} &y_i \overset{\text{ind.}}{\sim} \text{Poisson}\left(\exp\left\{\beta_0 + \beta_1 x_i + \sum_{k=1}^{K} u_k z_k\left(x_i\right)\right\}\right), \quad 1 \leq i \leq n, \\ &u_k \overset{\text{ind.}}{\sim} N\left(0, \sigma^2\right). \end{aligned} \tag{2.7}$$

The $\{z_k\left(\cdot\right) : 1 \leq k \leq K\}$ are the O'Sullivan spline functions described in Section 1.6.1. The choice of K is of relatively minor concern for penalised splines but of impact on the number of parameters to optimize, since the size of matrix $\mathbf{\Lambda}$ directly depends on it. In this case we choose $K = 50$.

Referring to (2.2), an equivalent GLMM formulation involves the design matrices

$$\begin{aligned} &\boldsymbol{X} = \left[\begin{array}{cc} 1 & x_i \end{array}\right]_{1 \leq i \leq n}, \quad \boldsymbol{Z} = \left[z_k\left(x_i\right)_{1 \leq k \leq K}\right]_{1 \leq i \leq n} \\ &\boldsymbol{\beta} = \left[\begin{array}{cc} \beta_0 & \beta_1 \end{array}\right]^T, \quad \boldsymbol{u} = \left[u_k{}_{1 \leq k \leq K}\right]^T, \\ &\boldsymbol{G} = \sigma^2 \boldsymbol{I}. \end{aligned}$$

We model $\mathbf{\Lambda}$ as a symmetric positive definite full matrix of size $K \times K$.

Credible intervals are obtained making use of the expressions (2.5) and (2.6) by assuming asymptotic normality. Results are displayed in Figure 2.1 and show great similarity between MCMC and GVA.

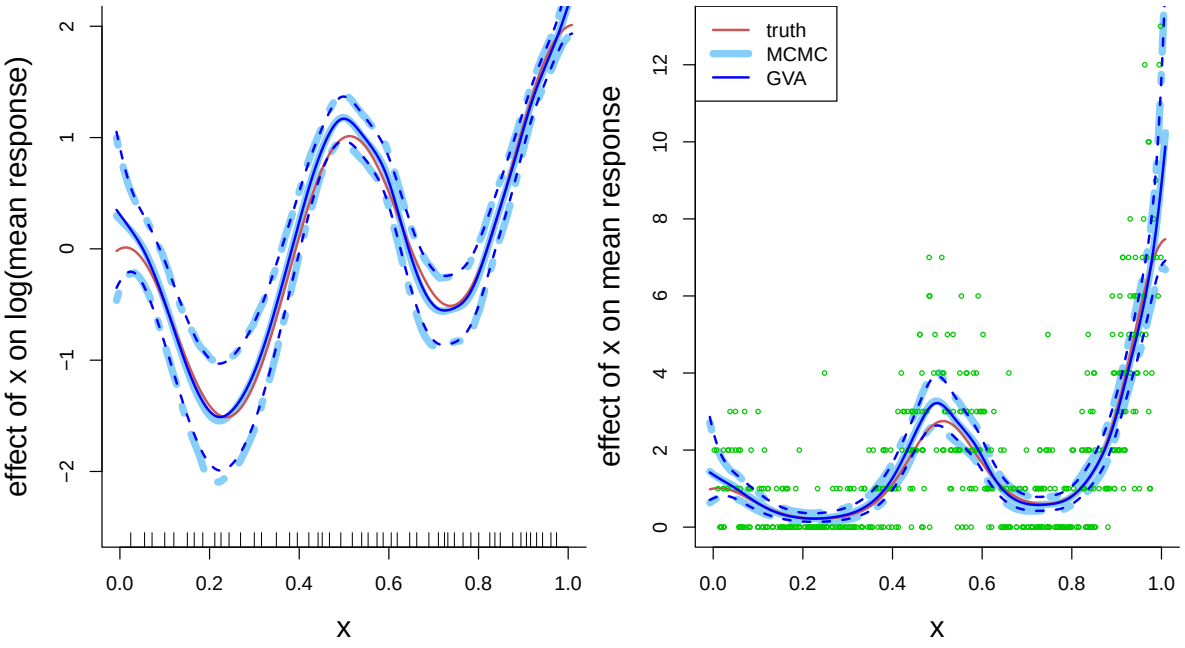

FIGURE 2.1: Left panel: mean estimate of f (solid line) and pointwise 95% credible intervals (dashed lines) obtained via MCMC and GVA for the Poisson nonparametric regression model (2.7). The interior knots are drawn on the x axis. The true f from which the data were generated is shown as a red solid line. Right panel: As for the left panel, but for $\exp(f)$ instead of f. The data are shown as circles.

2.6.2 Semiparametric logistic regression

The nomenclature *semiparametric* is introduced when an explanatory variable enters the model via a nonparametric function and a second one, at least, provides a parametric explanatory component. We simulate data according to the following semiparametric logistic regression model:

$$y_i \overset{\text{ind.}}{\sim} \text{Bernoulli}\left(\text{logit}^{-1}\left\{\beta x_{1i} + f\left(x_{2i}\right)\right\}\right), \quad 1 \leq i \leq n,$$

where $n = 500$, $\text{logit}^{-1}(x) = e^x/(1+e^x)$. We use $\beta = 0.5$, $x_{1i} \overset{\text{ind.}}{\sim} \text{Bernoulli}\left(\frac{1}{2}\right)$, $x_{2i} \overset{\text{ind.}}{\sim} \text{Uniform}(0,1)$ and

$$f(x) = \sin(2\pi x).$$

Then, we estimate the model

$$\begin{aligned} y_i &\overset{\text{ind.}}{\sim} \text{Bernoulli}\left(\text{logit}^{-1}\left\{\beta_0 + \beta_{x_1} x_{1i} + \beta_{x_2} x_{2i} + \sum_{k=1}^{K} u_k z_k\left(x_{2i}\right)\right\}\right), \quad 1 \leq i \leq n, \\ u_k &\overset{\text{ind.}}{\sim} N\left(0, \sigma^2\right), \end{aligned} \tag{2.8}$$

where $\{z_k(\cdot) : 1 \leq k \leq K\}$ is a set of O'Sullivan spline functions. We set $K = 50$.

Referring to (2.2), an equivalent GLMM formulation is

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{1i} & x_{2i} \end{bmatrix}_{1\leq i\leq n}, \quad \boldsymbol{Z} = \begin{bmatrix} z_k(x_{2i}) \\ {\scriptstyle 1\leq k\leq K} \end{bmatrix}_{1\leq i\leq n}$$
$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 & \beta_{x_1} & \beta_{x_2} \end{bmatrix}^T, \quad \boldsymbol{u} = \begin{bmatrix} u_k \\ {\scriptstyle 1\leq k\leq K} \end{bmatrix}^T,$$
$$\boldsymbol{G} = \sigma^2 \boldsymbol{I}.$$

We employ a symmetric positive definite full matrix $\boldsymbol{\Lambda}$ of size $K \times K$. The effect of the covariates entering the model through nonparametric components, $\boldsymbol{x}_2$, on the logit function of the Bernoulli probability is shown in Figure 2.2, together with mean estimate and credible intervals obtained via MCMC and GVA.

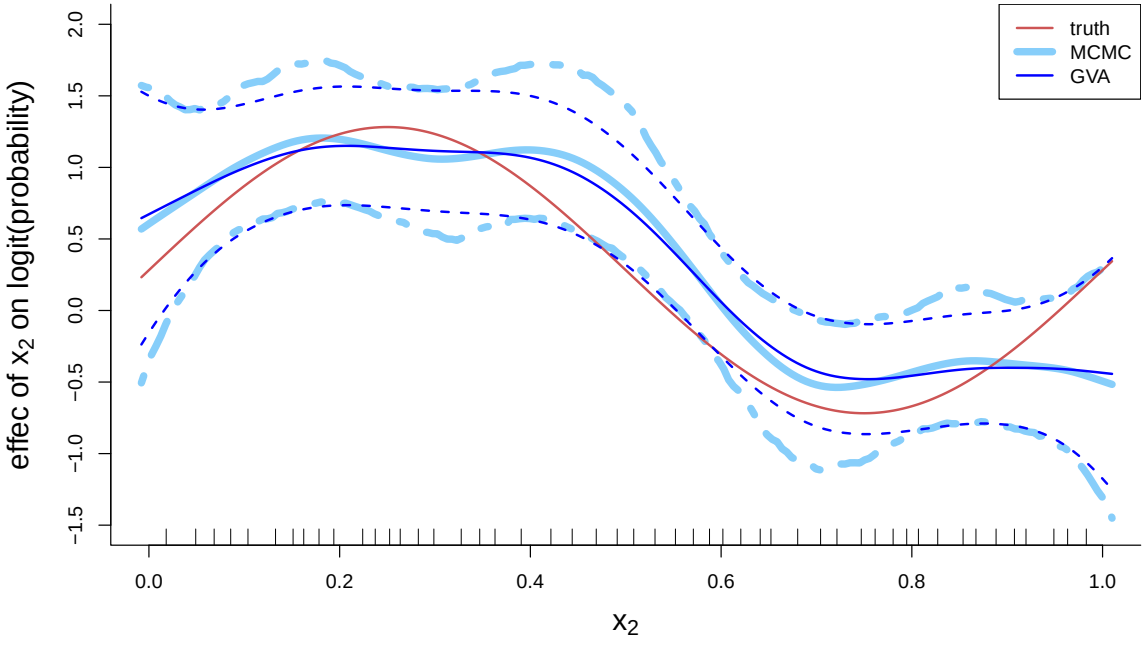

FIGURE 2.2: Mean estimate of f (solid line) and pointwise 95% credible intervals (dashed lines) with respect to $\boldsymbol{x}_2$, keeping $\boldsymbol{x}_1$ fixed to its mean, obtained via MCMC and GVA for the semiparametric logistic regression model (2.8). The interior knots are drawn on the x axis. The true f from which the data were generated is shown as a red solid line line.

2.6.3 Logistic additive model

We simulate a dataset from the model

$$y_i \overset{\text{ind.}}{\sim} \text{Bernoulli}\left(\text{logit}^{-1}\left\{f_1\left(x_{1i}\right) + f_2\left(x_{2i}\right)\right\}\right), \quad 1 \leq i \leq n, \tag{2.9}$$

with $n = 500$, $x_{1i} \overset{\text{ind.}}{\sim} \text{Uniform}\left(0,1\right)$, $x_{2i} \overset{\text{ind.}}{\sim} \text{Uniform}\left(0,1\right)$ and

$$f_1\left(x\right) = \cos\left(4\pi x\right) + 2x, \quad f_2\left(x\right) = \sin\left(2\pi x^2\right).$$

Then, we estimate the logistic additive model

$$\begin{aligned} y_i \overset{\text{ind.}}{\sim} \text{Bernoulli}\Bigg(\text{logit}^{-1}\Bigg\{&\beta_0 + \beta_{x_1} x_{1i} + \sum_{k_1=1}^{K_1} u_{1k_1} z_{1k_1}\left(x_{1i}\right) \\ &+\beta_{x_2} x_{2i} + \sum_{k_2=1}^{K_2} u_{2k_2} z_{2k_2}\left(x_{2i}\right)\Bigg\}\Bigg), \quad 1 \leq i \leq n, \\ u_{1k_1} \sim N\left(0, \sigma_1^2\right), \quad 1 \leq k_1 \leq K_1, \quad u_{1k_2} &\sim N\left(0, \sigma_2^2\right), \quad 1 \leq k_2 \leq K_2, \end{aligned} \tag{2.10}$$

where

$$\left\{z_{1k_1}\left(\cdot\right) : 1 \leq k_1 \leq K_1\right\} \quad \text{and} \quad \left\{z_{2k_2}\left(\cdot\right) : 1 \leq k_2 \leq K_2\right\}$$

are O'Sullivan spline functions over the range of the x_{1i}s and x_{2i}s, respectively. We set $K_1 = K_2 = 30$.

Referring to (2.2), an equivalent GLMM formulation involves matrices and vectors

$$\begin{aligned} \boldsymbol{X} &= \left[\begin{array}{ccc} 1 & x_{1i} & x_{2i} \end{array}\right]_{1 \leq i \leq n}, \quad \boldsymbol{Z} = \left[\begin{array}{cc} z_{1k_1}\left(x_{1i}\right) & z_{2k_2}\left(x_{2i}\right) \\ {\scriptstyle 1 \leq k_1 \leq K_1} & {\scriptstyle 1 \leq k_2 \leq K_2} \end{array}\right]_{1 \leq i \leq n}, \\ \boldsymbol{\beta} &= \left[\begin{array}{ccc} \beta_0 & \beta_{x_1} & \beta_{x_2} \end{array}\right]^T, \quad \boldsymbol{u} = \left[\begin{array}{cc} u_{1k} & u_{2k} \\ {\scriptstyle 1 \leq k \leq K_1} & {\scriptstyle 1 \leq k \leq K_2} \end{array}\right]^T, \\ \boldsymbol{G} &= \left[\begin{array}{cc} \sigma_1^2 \boldsymbol{I}_{K_1} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_2^2 \boldsymbol{I}_{K_2} \end{array}\right]. \end{aligned}$$

Reflecting the structure of $\boldsymbol{G}$, we choose $\boldsymbol{\Lambda}$ to be

$$\boldsymbol{\Lambda} = \left[\begin{array}{cc} \boldsymbol{\Lambda}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Lambda}_2 \end{array}\right],$$

where $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\Lambda}_2$ are $K_1 \times K_1$ and $K_2 \times K_2$ symmetric positive definite matrices, respectively. At the end of the optimization procedure, we are able to produce the plots displayed in Figure 2.3.

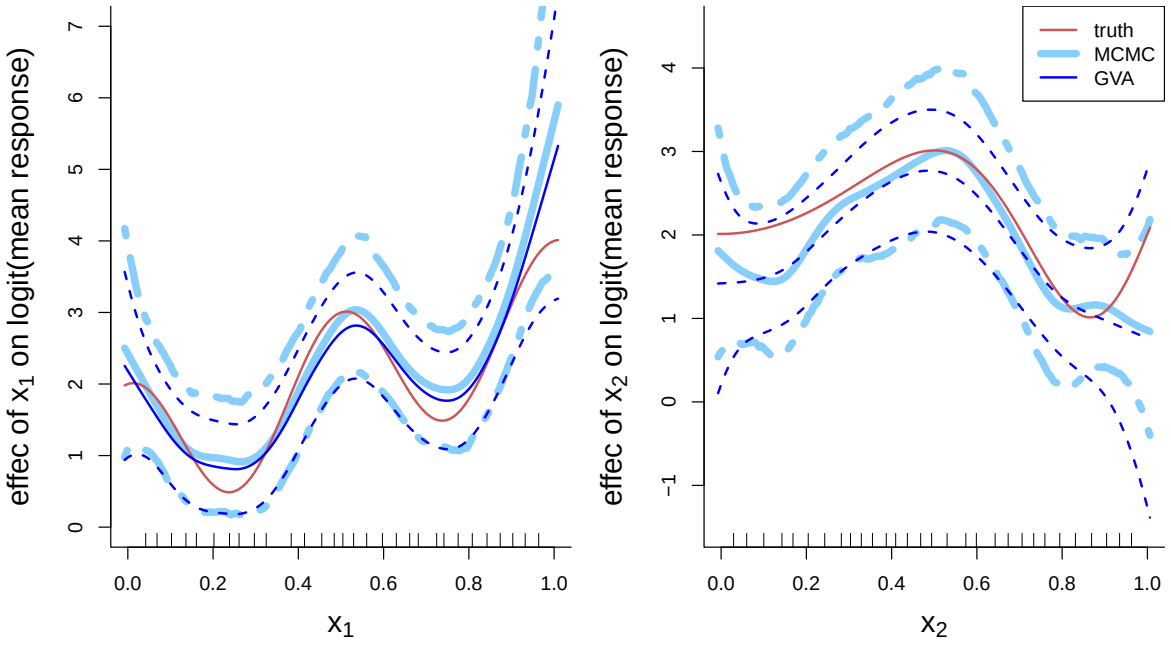

FIGURE 2.3: Left panel: mean estimate of f_1 (solid line) and pointwise 95% credible intervals (dashed lines) with respect to $\boldsymbol{x}_1$, keeping $\boldsymbol{x}_2$ fixed to its mean, obtained via MCMC and GVA for the logistic additive regression model (2.10). The interior knots are drawn on the x axis. The true function from which the data were generated is shown as a red solid line line. Right panel: As for the left panel, but for f_2 plotted against $\boldsymbol{x}_2$, keeping $\boldsymbol{x}_1$ fixed to its mean.

2.6.4 Generalized geoadditive model

Another type of extension is to allow for bivariate functions of pairs of continuous predictors to be included in additive models such as the one of Subsection 2.6.3. In a spatial data context, these models may be referred to as *geoadditive* models (Kammann and Wand, 2003) and can be included in the framework given by (2.2). We exemplify this concept with the following illustrative example.

We simulate a dataset from the model

$$y_i \overset{\text{ind.}}{\sim} \text{Poisson}\left(\exp\left\{f_1\left(x_{1i}\right) + f_2\left(x_{2i}\right) + g\left(\boldsymbol{x}_i^{\text{geo}}\right)\right\}\right), \quad 1 \leq i \leq n, \tag{2.11}$$

with $n = 100$, $x_{1i} \overset{\text{ind.}}{\sim} \text{Uniform}\left(0,1\right)$, $x_{2i} \overset{\text{ind.}}{\sim} \text{Uniform}\left(0,1\right)$, $x_{1i}^{\text{geo}} \overset{\text{ind.}}{\sim} \text{Uniform}\left(0,1\right)$, $x_{2i}^{\text{geo}} \overset{\text{ind.}}{\sim} \text{Uniform}\left(0,1\right)$ and

$$f_1\left(x\right) = \sin\left(2\pi x^2\right), \quad f_2\left(x\right) = \cos\left(3\pi x^3\right), \quad g\left(x,y\right) = \sin\left(3\pi xy\right).$$

In this context, x_{1i} and x_{2i} are predictor variables, $\boldsymbol{x}_i^{\text{geo}}$ is a 2×1 vector containing spatial

coordinates, while f_1, f_2 and g are functions of the predictors. Then, we estimate the Poisson geoadditive model

$$y_i \overset{\text{ind.}}{\sim} \text{Poisson}\left(\exp\left\{\beta_0 + \beta_1 x_{1i} + \sum_{k_1=1}^{K_1} u_{1k_1} z_{1k_1}(x_{1i}) + \beta_2 x_{2i} + \sum_{k_2=1}^{K_2} u_{2k_2} z_{2k_2}(x_{2i}) + \boldsymbol{\beta}^{\text{geo}T}\boldsymbol{x}_i^{\text{geo}} + \sum_{k^{\text{geo}}=1}^{K^{\text{geo}}} u_{k^{\text{geo}}}^{\text{geo}} z_{k^{\text{geo}}}^{\text{geo}}(\boldsymbol{x}_i^{\text{geo}})\right\}\right), \quad 1 \le i \le n, \tag{2.12}$$

$$u_{1k_1} \sim N\left(0, \sigma_1^2\right), \quad 1 \le k_1 \le K_1, \quad u_{1k_2} \sim N\left(0, \sigma_2^2\right), \quad 1 \le k_2 \le K_2,$$

$$\boldsymbol{u}^{\text{geo}} = N\left(\boldsymbol{0}, \sigma_{\text{geo}}^2 \boldsymbol{\Omega}^{-1}\right).$$

Here

$$\{z_{1k_1}(\cdot) : 1 \le k_1 \le K_1\} \quad \text{and} \quad \{z_{2k_2}(\cdot) : 1 \le k_2 \le K_2\}$$

are O'Sullivan spline functions over the range of the x_{1i}s and x_{2i}s, respectively and

$$\{z_{k^{\text{geo}}}^{\text{geo}}(\cdot) : 1 \le k^{\text{geo}} \le K^{\text{geo}}\}$$

give rise to a matrix $\boldsymbol{Z}^{\text{geo}}$ which reflects the covariance structure of the spatial coordinates $\boldsymbol{x}^{\text{geo}}$s. In applications, K^{geo} may be a particularly large number giving rise to a Gaussian approximating density with large covariance matrix $\boldsymbol{\Lambda}$. One possibility to overcome this problem is to select or propose a set of representative knots using the *space filling* design (Johnson *et al.*, 1990; Nychka and Saltzman, 1998). The `cover.design()` function from the `R` package `Fields` (Nychka *et al.*, 2018) provides software for space filling knot selection. We employ these tools to select a set of $\boldsymbol{\kappa}_{k^{\text{geo}}}$, $1 \le k^{\text{geo}} \le K^{\text{geo}} \le n$ representative points. Following Kammann and Wand (2003), we define

$$\boldsymbol{Z}^{\text{geo}} = \left[C_0\left(\underset{1\le k^{\text{geo}}\le K^{\text{geo}}}{\left\|\boldsymbol{x}_i^{\text{geo}} - \boldsymbol{\kappa}_{k^{\text{geo}}}\right\|/\rho}\right)\right]_{1\le i\le n}, \quad \boldsymbol{\Omega} = \left[C_0\left(\underset{1\le k^{\text{geo}}, k^{\text{geo}'}\le K^{\text{geo}}}{\left\|\boldsymbol{\kappa}_{k^{\text{geo}}} - \boldsymbol{\kappa}_{k^{\text{geo}'}}\right\|/\rho}\right)\right]_{1\le i\le n},$$

where $C_0(r) = (1 + |r|)\exp(-|r|)$ is the underlying covariance structure. We choose $\rho = \max_{1\le i,j\le n}\left\|\boldsymbol{x}_i^{\text{geo}} - \boldsymbol{x}_j^{\text{geo}}\right\|$ and find the singular value decomposition of $\boldsymbol{\Omega}$, that is, $\boldsymbol{\Omega} = \boldsymbol{U}\text{diag}(\boldsymbol{d})\boldsymbol{V}^T$, to obtain the matrix square root of $\boldsymbol{\Omega}$, $\boldsymbol{\Omega}^{1/2} = \boldsymbol{U}\text{diag}\left(\sqrt{\boldsymbol{d}}\right)\boldsymbol{V}^T$. For fitting purposes, we compute the reparameterization $\tilde{\boldsymbol{Z}}^{\text{geo}} = \boldsymbol{Z}^{\text{geo}}\boldsymbol{\Omega}^{-1/2}$, whose associate random effect vector $\tilde{\boldsymbol{u}}^{\text{geo}}$ has covariance $\sigma_{\text{geo}}^2\boldsymbol{I}$. Referring to the GLMM formulation

(2.2), the following matrices are involved:

$$\boldsymbol{X} = \left[\begin{array}{cccc} 1 & x_{1i} & x_{2i} & \boldsymbol{x}_i^{\text{geo}} \end{array} \right]_{1 \leq i \leq n}, \quad \boldsymbol{Z} = \left[\begin{array}{ccc} z_{1k_1}(x_{1i}) & z_{2k_2}(x_{2i}) & \tilde{\boldsymbol{Z}}^{\text{geo}} \\ {\scriptstyle 1 \leq k_1 \leq K_1} & {\scriptstyle 1 \leq k_2 \leq K_2} & \end{array} \right]_{1 \leq i \leq n},$$

$$\boldsymbol{\beta} = \left[\begin{array}{cccc} \beta_0 & \beta_{x_1} & \beta_{x_2} & \boldsymbol{\beta}^{\text{geo}} \end{array} \right]^T, \quad \boldsymbol{u} = \left[\begin{array}{ccc} u_{1k} & u_{2k} & \tilde{\boldsymbol{u}}^{\text{geo}} \\ {\scriptstyle 1 \leq k \leq K_1} & {\scriptstyle 1 \leq k \leq K_2} & \end{array} \right]^T,$$

$$\boldsymbol{G} = \left[\begin{array}{ccc} \sigma_1^2 \boldsymbol{I}_{K_1} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_2^2 \boldsymbol{I}_{K_2} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \sigma_{\text{geo}}^2 \boldsymbol{I} \end{array} \right].$$

GVA can be applied in a similar way as done in Subsection 2.6.3, setting

$$\boldsymbol{\Lambda} = \left[\begin{array}{ccc} \boldsymbol{\Lambda}_1 & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Lambda}_2 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{\Lambda}^{\text{geo}} \end{array} \right],$$

where $\boldsymbol{\Lambda}_1$, $\boldsymbol{\Lambda}_2$ and $\boldsymbol{\Lambda}^{\text{geo}}$ are symmetric positive definite matrices of size $K_1 \times K_1$, $K_2 \times K_2$ and $K^{\text{geo}} \times K^{\text{geo}}$, respectively. We set $K_1 = K_2 = 30$ and, according to step (1) of Ruppert *et al.* (2003, Section 13.5), we select $K^{\text{geo}} = 25$.

Figure 2.4 displays the simulated data highlighting the selected representative knots. Figure 2.5 shows the univariate functions resulting from the fitting. In a similar manner, a fitted surface can be produced through kriging prediction, accounting for the covariates effect.

2.7 Simulation study

We here include a first performance investigation consisting in a simulation study that involves a Poisson non-parametric model and a logistic additive model with canonical link.

Data were simulated using the same settings of Subsection 2.6.1, for the Poisson case. The linear predictor was scaled so that there was about 50% unexplained variance in each replicate dataset, similarly to what suggested in Wood (2011). The GLMs of Subsection 2.6.1 were fitted to each replicate dataset, using the correct distribution and link, via GVA, penalized quasi likelihood (PQL), generalized cross-validation (GCV) and MCMC with the already mentioned uninformative priors. We used the `R` function `glmmPQL` of libary `MASS` for PQL and function `gam`, and then `gamm` for the logistic model, from package `mgcv` for GCV. Boxplots in the left panel of Figure 2.6 show the distributions, over 100 replicates, of differences in mean square error (MSE) between each alternative

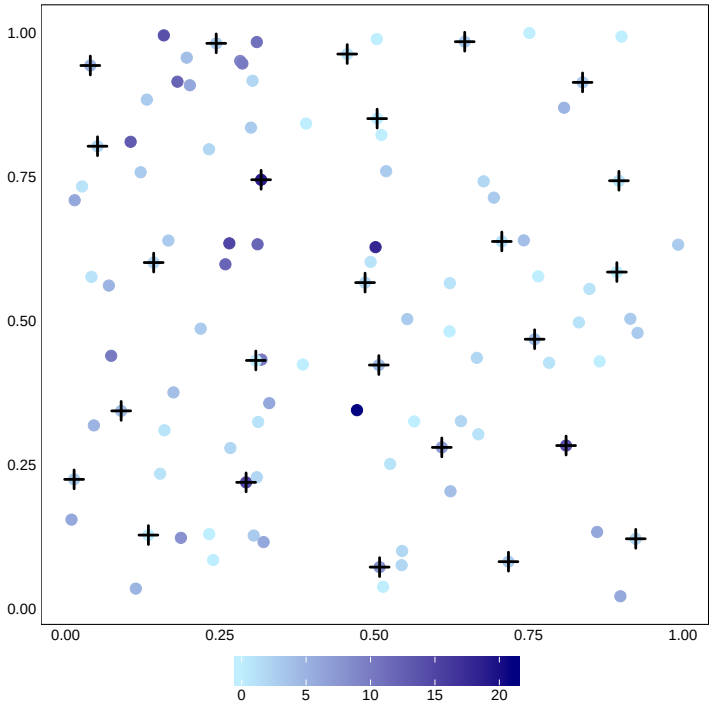

FIGURE 2.4: Plot of the data simulated according to (2.11). The symbol "+" indicates the representative knots.

method and GCV, divided by the average MSE of GCV. Indeed, GCV showed the lowest average MSE, measured on the scale of the linear predictor. If compared to GCV, PQL performs very similarly in terms of MSE, followed by GVA. However, it should be pointed out that GVA sometimes performed better or slightly worse when varying the number of nodes or generating functions in simulations studies that have not been included here. The initialization of the optimization procedure for GVA is also crucial and requires further investigations to face the presence of local optima and improve GVA performances.

In a similar way, we performed the simulation study for the logistic case with 100 replicates, using the settings and model of Subsection 2.6.3. However, the MSE was measured on the probability scale and not on the linear predictor. This time, MCMC showed better performances in terms of MSE, but the boxplot of GVA in the right panel of Figure 2.6 highlights a significant difference with the MCMC benchmark, which was less evident for the Poisson nonparametric case.

Nevertheless, what is apparent in both the simulation studies is that the boxplots associated with GVA are more concentrated, if compared to the others, excluding the boxplot associated with PQL in the Poisson nonparametric case.

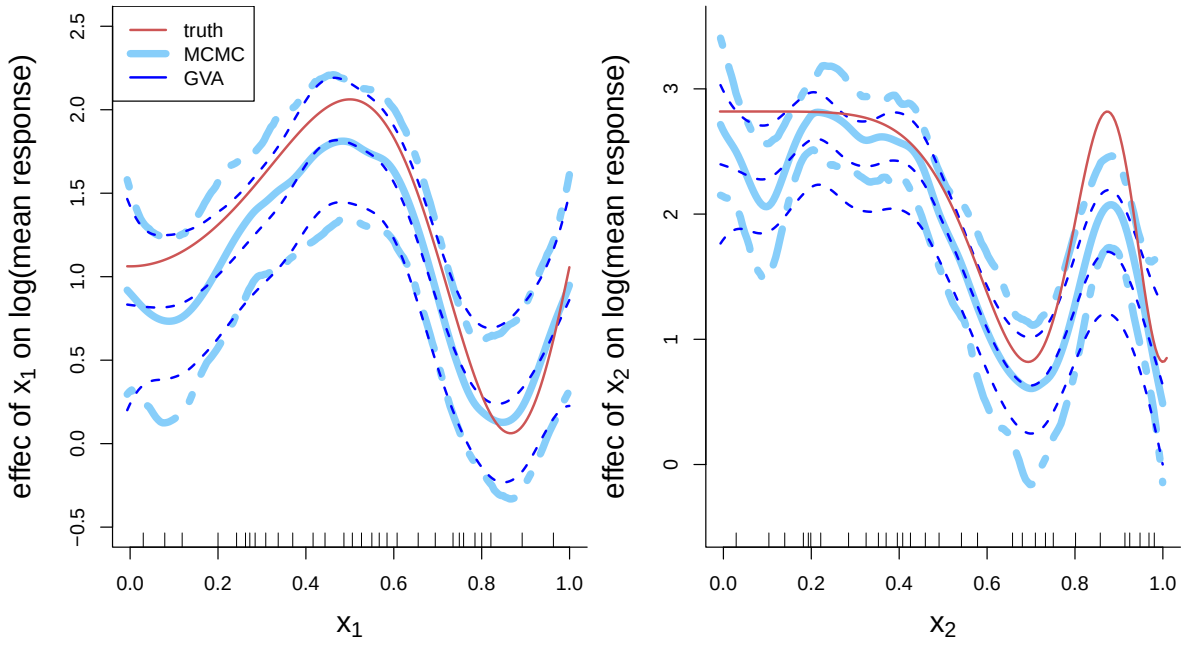

FIGURE 2.5: Left panel: mean estimate of f_1 (solid line) and pointwise 95% credible intervals (dashed lines) with respect to $\boldsymbol{x}_1$, keeping $\boldsymbol{x}_2$ and spatial coordinates fixed to their mean, obtained via MCMC and GVA for the Poisson geoadditive model (2.12). The interior knots are drawn on the x axis. The true function from which the data were generated is shown as a red solid line line. Right panel: As for the left panel, but for f_2 plotted against $\boldsymbol{x}_2$, keeping $\boldsymbol{x}_1$ and spatial coordinates fixed to their mean.

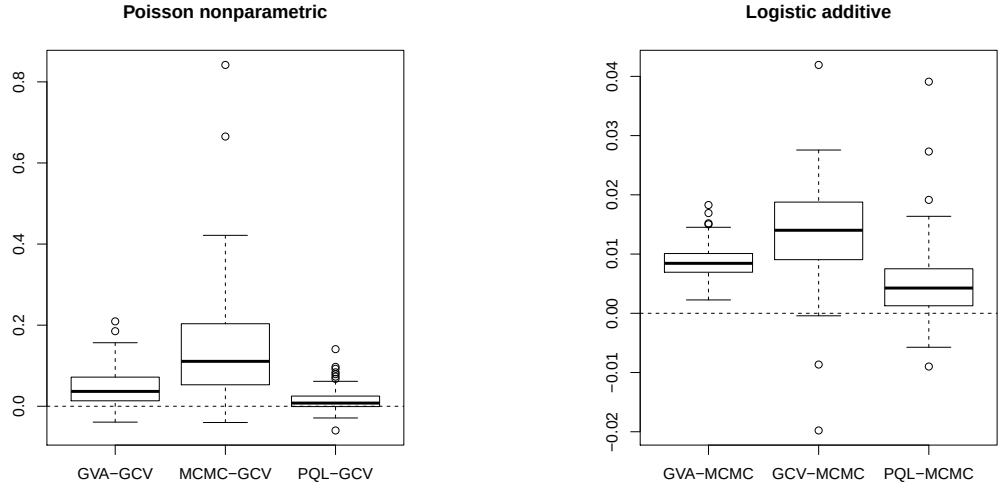

FIGURE 2.6: Mean square error comparisons. Left panel: comparison between GCV and other methods involving Poisson non parametric models. Right panel: comparison between MCMC and other methods involving logistic additive models.

2.8 Concluding remarks

Gaussian variational approximations are an approximate inference tool that easily applies to a GLMM context, covering a wide range of models with peculiar random effects structures. Numerous other cases not listed here could be contemplated. A type of extension is to allow for models having both random intercept and slopes, which we could name generalized additive semiparametric mixed models. One more obvious extension is the treatment of additional distributional families for the response variable. Examples are normal, gamma and negative binomial distributions. All of these extensions are relatively straightforward with respect to the framework of general GLMMs we have presented. However, more complete numerical studies should be implemented to partially fill the gap in theoretical results and asymptotic properties for GVA. Further simulation studies should involve more generating functions for each model under exam or a larger number of nodes in models with spline components. Fitting of other types of models should be tested, covering distributions other then the Poisson and Bernoulli ones and also using alternative competing methods. Simulation studies may be complemented by tracking means and standard errors of estimated model parameters.

Chapter 3

Variational inference for elaborate response models

3.1 Introduction

We extend recent work concerning variational approximations via message passing to accommodate approximate fitting and inference for some elaborate response models. Derivation of variational message passing is challenging owing to the presence of non-standard exponential families and numerical integration being needed. Nevertheless the factor graph fragment approach means that algorithm updates only need to be derived once for a particular response model, which can be integrated in an arbitrarily complex model. Another advantage of this approach is that the VMP framework is such that arbitrarily large semiparametric regression models can be handled using factor graph fragments.

Wand (2017) introduced the notion of *factor graph fragments* to design a general framework for variational Bayes approximate inference, considering situations that are fundamental to semiparametric regression analysis via VMP: Gaussian prior, inverse Wishart prior, iterated inverse G-Wishart, Gaussian penalization and Gaussian likelihood. Additional cases are examined to handle logistic, probit and Poisson regression models. The idea of message passing on factor graph fragments implicitly indicates the possibility of compartmentalize the algebraic derivations to single parts of the model at hand, or *likelihood fragments*. Such an approach is extendible to models with more elaborate likelihood structures. Nolan and Wand (2017) provide accurate algebraic and numerical details for fitting logistic likelihood regression via VMP. McLean and Wand (2018) consider six other likelihood families: negative binomial, Student's t, asymmetric Laplace, skew normal, finite normal mixture and support vector machine. We add to

this recent body of work and derive VMP updates for approximate fitting and inference for the Pareto random sample, support vector regression (SVR) and skew t responses[1]. Furthermore, we investigate how various auxiliary random variable representations of the likelihood impact the variational approximating results. The response likelihoods are re-expressed in terms of auxiliary variables and more common distributions since the route without auxiliary variables is usually numerically complex or intractable. The use of auxiliary variables has the practical advantage of producing algorithm message updates which are derivable in closed form or requiring only univariate numerical integration. On the other hand, such a representation may introduce strong posterior dependence which is hard to capture with simple forms of approximating densities.

Section 1.3 provides a brief description of VMP and its implementation via the notion of message passing on a factor graph. An introduction to the parallel methodology of mean field variational Bayes (MFVB) on directed acyclic graphs is also included. All the VMP algorithms proposed in the following sections have been checked comparing the parameter estimates with those from the corresponding MFVB algorithms, that have been derived but not all included in this thesis. In fact, our VMP and MFVB algorithms converge to the same posterior density function approximations, since they are each based on the same optimization problem.

When performing variational inference based on a mean field restriction as in the case of VMP and MFVB, the variational lower bound on the marginal loglikelihood is commonly used to assess convergence. However, the algebra required to obtain the lower bound expression is involved for large models since it includes the calculation of entropy components, which is non-standard for the exponential families arising in the models considered here.

Only the MFVB algorithm for SVR and related derivation details are included here as an illustration, despite all the MFVB algorithms have been derived and checked with the corresponding VMP versions. The reason is that VMP provides a more flexible and scalable framework for variational inference in which iterative updates are amenable to modularization and extendible to arbitrarily large models via the notion of factor graph fragments, as shown in the application of Subsection 3.4.2. For the sake of conciseness and clarity, explanations are provided in this chapter for the following points:

1. In Section 3.2, the Pareto random sample is proposed as an introductory example to illustrate the steps for the complete derivation of a VMP algorithm;

[1]The paper Maestrini and Wand (2018) is based on the work concerning the skew t likelihood fragment presented in this chapter.

2. A MFVB algorithm for SVR is derived in Section 3.3 in addition to the corresponding VMP alternative to highlight the major differences between the two perspectives;

3. Two different VMP algorithms are displayed for the skew t regression likelihood fragment in Section 3.4, according to two alternative product density restrictions on the approximating density; we prove with a numerical study and a theoretical result that the more computationally convenient one has a serious pitfall and show that posterior dependence arising from an auxiliary variable representation of a skew t model may lead to poor performances in terms of variational message passing approximation; this happens if using simple auxiliary variable representations of the likelihood fragment and convenient factorizations of the approximating densities; we conclude with an illustration.

3.1.1 Notation

Before treating variational algorithms for the response models of interest, we define some relevant additional notation.

For a $d \times 1$ vector $\boldsymbol{v}_1$ and a $d^2 \times 1$ vector $\boldsymbol{v}_2$ such that $\text{vec}^{-1}(\boldsymbol{v}_2)$ is symmetric and given a $d \times d$ matrix $\boldsymbol{Q}$, a $d \times 1$ vector $\boldsymbol{r}$ and $s \in \mathbb{R}$ we define

$$G_{\text{VMP}}\left(\begin{bmatrix} \boldsymbol{v}_1 \\ \boldsymbol{v}_2 \end{bmatrix}; \boldsymbol{Q}, \boldsymbol{r}, s\right) = -\frac{1}{8}\text{tr}\left(\boldsymbol{Q}\left\{\text{vec}^{-1}(\boldsymbol{v}_2)\right\}^{-1}\left[\boldsymbol{v}_1\boldsymbol{v}_1^T\left\{\text{vec}^{-1}(\boldsymbol{v}_2)\right\}^{-1} - 2\boldsymbol{I}\right]\right) - \frac{1}{2}\boldsymbol{r}^T\left\{\text{vec}^{-1}(\boldsymbol{v}_2)\right\}^{-1}\boldsymbol{v}_1 - \frac{1}{2}s.$$

The G_{VMP} function originates from the fact that

$$E_{\boldsymbol{\theta}}\left\{-\frac{1}{2}\left(\boldsymbol{\theta}^T\boldsymbol{Q}\boldsymbol{\theta} - 2\boldsymbol{r}^T\boldsymbol{\theta} + s\right)\right\} = G_{\text{VMP}}(\boldsymbol{\eta}; \boldsymbol{Q}, \boldsymbol{r}, s)$$

when $\boldsymbol{\theta}$ is a $d \times 1$ multivariate normal random vector with natural parameter vector $\boldsymbol{\eta}$. We introduce the notations $(E\boldsymbol{T})_2^{\text{ISRN}}$, $(E\boldsymbol{T})_3^{\text{ISRN}}$, $(E\boldsymbol{T})_2^{\text{MR}}$, $(E\boldsymbol{T})_2^{\text{SS}}$ and $(E\boldsymbol{T})_3^{\text{SS}}$, as expressed in (A.4)–(A.8), referring to the expected value of the sufficient statistic of particular exponential families that are defined in Appendix A: the inverse square root Nadarajah, Sea Sponge and Moon Rock distributions.

3.1.2 A note on the inverse chi-squared prior

The inverse chi-squared distribution is the conjugate family for variance parameters in normal mean-scale models written in auxiliary variable form and Bayesian semi-parametric regression. Since the models we examine involve normal distributions the conjugacy property helps reduce the number of non-analytic forms but alternative scale parameter priors may be also convenient. Gelman (2006) explains that approximate non-informativeness of scale parameters can be achieved via half t distribution priors and pays particular attention to half Cauchy priors.

VMP and MFVB algorithm derivations benefit from the following auxiliary variable result. Let x and a be random variables such that

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

Then

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

The conjugacy relationship between the Gaussian and inverse gamma or, equivalently, inverse chi-squared families arising from a half Cauchy prior specification reduces the number of intractable integrals in mean field variational inference.

An extension of this result to covariance matrices via the inverse G-Wishart distribution described in Subsection A.2.13 of Appendix A supports algorithm derivations in Chapter 4.

3.2 The Pareto likelihood fragment

The Pareto distribution is a skewed distribution with "heavy" tails named after the Italian civil engineer, economist, and sociologist Vilfredo Pareto. It finds main applications is social sciences to model the distribution of incomes or populations but also in the fields of engineering and actuarial sciences. Among several alternative definitions of Pareto distributions we consider the one known as *Pareto distribution of II type*.

Consider the model

$$y_i \overset{\text{ind.}}{\sim} \text{Pareto}\,(\mu, \alpha, \beta)\,, \quad 1 \leq i \leq n, \tag{3.1}$$

where $y_i \geq \mu$, $\alpha > 0$ is the Pareto exponent and $\beta > 0$ is the scale parameter, according to the density specification in Subsection A.2.19 of Appendix A. A Pareto exponent value of $\alpha = \log_4 5 \approx 1.16$ is associated with a famous result known as *Pareto principle*,

or "80-20" rule. A notorious exemplification of this rule states that the 80% of the wealth of a society is held by 20% of its population.

As previously affirmed, an auxiliary variable representation of the likelihood fragment may simplify the derivation of VMP algorithms. Then, if we introduce a random variable a_i, $1 \leq i \leq n$, such that $a_i \,|\, \alpha, \beta \overset{\text{ind.}}{\sim} \text{Gamma}(\alpha, \beta)$ with α and β shape and scale parameters respectively, we can write model (3.1) as

$$y_i \,|\, \mu, a_i \overset{\text{ind.}}{\sim} \text{Exp}(\mu, a_i), \quad a_i \,|\, \alpha, \beta \overset{\text{ind.}}{\sim} \text{Gamma}(\alpha, \beta), \tag{3.2}$$

where $p(y_i \,|\, \mu, a_i) = a_i \exp\{-a_i (y_i - \mu)\}$ is a shifted exponential distribution.

In accordance with the theory of mean field variational approximations, a product density restriction as in (1.6) is required. We assume the q-density admits the simplest product density restriction, that is,

$$\begin{aligned} q(\mu, \alpha, \beta, \boldsymbol{a}) &= q(\mu)\, q(\alpha)\, q(\beta)\, q(\boldsymbol{a}) \\ &= q(\mu)\, q(\alpha)\, q(\beta) \prod_{i=1}^{n} q(a_i), \end{aligned} \tag{3.3}$$

where $\boldsymbol{a}$ is the vector containing the a_i, $1 \leq i \leq n$, auxiliary variables. Similarly to $\boldsymbol{a}$, we define the vector of observations $\boldsymbol{y}$. The factor graph representation in Figure 3.1 is designed around model (3.2) and restriction (3.3) to support the derivation of a VMP algorithm, following steps (1.14)–(1.16). The procedure basically requires to obtain

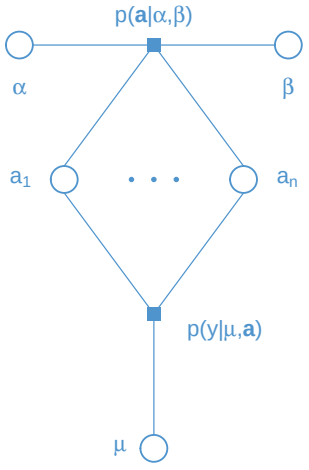

FIGURE 3.1: Factor graph for the Pareto likelihood specification in (3.2) under the assumption in (3.3) with independent auxiliary variables $a_1 \ldots a_n$.

expressions for messages passed from each stochastic node to connected factors and vice versa. Messages from factors to stochastic nodes produce the optimal approximating densities at convergence.

As a starting point, we consider messages involving the parameter μ, which only depend on the logarithm of the exponential density term in the likelihood factor, the only part of the likelihood fragment expressing dependence on μ:

$$\log p\left(\boldsymbol{y} \,|\, \mu, \boldsymbol{a}\right) = \sum_{i=1}^{n} \log\left(a_i I\left(y_i > \mu\right)\right) - \sum_{i=1}^{n} a_i \left(y_i - \mu\right), \tag{3.4}$$

where $\log\left(I\left(y_i > \mu\right)\right) = -\infty$ for $I\left(y_i > \mu\right) = 0$. It follows from (1.15) that messages from factor $p\left(\boldsymbol{y} \,|\, \mu, \boldsymbol{a}\right)$ to μ take form

$$m_{p(\boldsymbol{y} \,|\, \mu, \boldsymbol{a}) \to \mu}\left(\mu\right) = \exp\left(\mu \mathbf{1}_n^T E_{q(\boldsymbol{a})}\left(\boldsymbol{a}\right)\right) I\left(\mu < \min_{1 \leq i \leq n}\left(y_i\right)\right), \tag{3.5}$$

where, according to (1.16), $E_{q(\boldsymbol{a})}$ denotes expectation with respect to the density function arising from the normalization of

$$m_{p(\boldsymbol{y} \,|\, \mu, \boldsymbol{a}) \to \boldsymbol{a}}\left(\boldsymbol{a}\right) m_{\boldsymbol{a} \to p(\boldsymbol{y} \,|\, \mu, \boldsymbol{a})}\left(\boldsymbol{a}\right) = m_{p(\boldsymbol{y} \,|\, \mu, \boldsymbol{a}) \to \boldsymbol{a}}\left(\boldsymbol{a}\right) m_{p(\boldsymbol{a} \,|\, \alpha, \beta) \to \boldsymbol{a}}\left(\boldsymbol{a}\right). \tag{3.6}$$

It is apparent that messages $m_{p(\boldsymbol{y} \,|\, \mu, \boldsymbol{a}) \to \mu}\left(\mu\right)$ have the form of a truncated distribution. Such information helps defining a model in which the messages entering the main factor graph fragment ensure conjugacy with the outgoing ones and facilitates the derivation of a VMP scheme. If, for instance, the only message that μ receives is a prior within the same family of truncated distribution with the same truncation value, then $m_{\mu \to p(\boldsymbol{y} \,|\, \mu, \boldsymbol{a})}\left(\mu\right)$ is conjugate to $m_{p(\boldsymbol{y} \,|\, \mu, \boldsymbol{a}) \to \mu}\left(\mu\right)$.

Now, consider messages from factor $p\left(\boldsymbol{y} \,|\, \mu, \boldsymbol{a}\right)$ to a single auxiliary variable a_i. Applying formula (1.15) to the log-likelihood component (3.4) and expectation with respect to (1.16), these messages take the form

$$m_{p(\boldsymbol{y} \,|\, \mu, \boldsymbol{a}) \to a_i}\left(a_i\right) = \exp\left\{ \begin{bmatrix} \log\left(a_i\right) \\ a_i \end{bmatrix}^T \boldsymbol{\eta}_{p(\boldsymbol{y} \,|\, \mu, \boldsymbol{a}) \to a_i} \right\}, \tag{3.7}$$

with natural parameter update

$$\boldsymbol{\eta}_{p(\boldsymbol{y} \,|\, \mu, \boldsymbol{a}) \to a_i} \longleftarrow \begin{bmatrix} 1 \\ \mu_{q(\mu)} - y_i \end{bmatrix},$$

with $\mu_{q(\mu)}$ denoting expectation of μ with respect to the normalization of

$$m_{p(\boldsymbol{y}\,|\,\mu,\boldsymbol{a})\to\mu}(\mu)\, m_{\mu\to p(\boldsymbol{y}\,|\,\mu,\boldsymbol{a})}(\mu).$$

Supposing, for example, that the only message entering μ is a prior $p(\mu)$ with hyperparameter $\lambda_\mu > 0$ such that

$$p(\mu) \propto \exp(\lambda_\mu \mu)\, I\left(\mu < \min_{1\le i\le n}(y_i)\right), \tag{3.8}$$

then

$$\begin{aligned}\mu_{q(\mu)} &= \frac{\int_{-\infty}^{y_{min}} e^{\boldsymbol{\eta}_{p(\boldsymbol{y}\,|\,\boldsymbol{a},\mu)\leftrightarrow\mu}x}dx}{\int_{-\infty}^{y_{min}} x e^{\boldsymbol{\eta}_{p(\boldsymbol{y}\,|\,\boldsymbol{a},\mu)\leftrightarrow\mu}x}dx}\\ &= y_{min} - \frac{1}{\eta_{p(\boldsymbol{y}\,|\,\boldsymbol{a},\mu)\leftrightarrow\mu}},\end{aligned}$$

with $y_{min} = \min_{1\le i\le n}(y_i)$ and $\eta_{p(\boldsymbol{y}\,|\,\boldsymbol{a},\mu)\leftrightarrow\mu} = \mathbf{1}_n^T E_{q(\boldsymbol{a})}(\boldsymbol{a}) + \lambda_\mu$. It follows that $\mu_{q(\mu)}$ is updated according to

$$\mu_{q(\mu)} \longleftarrow y_{min} - \left(\mathbf{1}_n^T E_{q(\boldsymbol{a})}(\boldsymbol{a}) + \lambda_\mu\right)^{-1}.$$

Observing the sufficient statistic component of $m_{p(\boldsymbol{y}\,|\,\mu,\boldsymbol{a})\to a_i}(a_i)$ as in (3.7) it is apparent that this message is within the family of the gamma distribution.

The auxiliary random variable a_i also appears in the gamma density term of the likelihood factor, therefore also messages from $p(\boldsymbol{a}\,|\,\alpha,\beta)$ to a_i have to be obtained. The logarithm of the auxiliary random variable a_i likelihood component is

$$\log p(a_i\,|\,\alpha,\beta) = \alpha\log\beta - \log\Gamma(\alpha) + (\alpha-1)\log(a_i) - \beta a_i.$$

This log-likelihood expression also produces the messages from factor $p(a_i\,|\,\alpha,\beta)$ to variables α and β. It follows from application of (1.15) that

$$m_{p(\boldsymbol{a}\,|\,\alpha,\beta)\to a_i}(a_i) = \exp\left\{\begin{bmatrix}\log(a_i)\\ a_i\end{bmatrix}^T \boldsymbol{\eta}_{p(\boldsymbol{a}\,|\,\alpha,\beta)\to a_i}\right\}, \tag{3.9}$$

$$m_{p(\boldsymbol{a}\,|\,\alpha,\beta)\to\alpha}(\alpha) = \exp\left\{\begin{bmatrix}\log\{\Gamma(\alpha)\}\\ \alpha\end{bmatrix}^T \boldsymbol{\eta}_{p(\boldsymbol{a}\,|\,\alpha,\beta)\to\alpha}\right\} \tag{3.10}$$

$$\text{and}\quad m_{p(\boldsymbol{a}\,|\,\alpha,\beta)\to\beta}(\beta) = \exp\left\{\begin{bmatrix}\log(\beta)\\ \beta\end{bmatrix}^T \boldsymbol{\eta}_{p(\boldsymbol{a}\,|\,\alpha,\beta)\to\beta}\right\}. \tag{3.11}$$

Note that the sufficient statistic vectors of messages $m_{p(\boldsymbol{a}\,|\,\alpha,\beta)\to a_i}(a_i)$ and $m_{p(\boldsymbol{a}\,|\,\alpha,\beta)\to\beta}(\beta)$ are those of a gamma density function. Expectation with respect to (1.16) indicates that the messages from $p(\boldsymbol{a}\,|\,\alpha,\beta)$ to β and from $p(\boldsymbol{a}\,|\,\alpha,\beta)$ to a_i are proportional to a gamma density function with natural parameter updates which are respectively

$$\boldsymbol{\eta}_{p(\boldsymbol{a}\,|\,\alpha,\beta)\to a_i} \longleftarrow \begin{bmatrix} \mu_{q(\alpha)} - 1 \\ -\mu_{q(\beta)} \end{bmatrix}$$

$$\text{and} \quad \boldsymbol{\eta}_{p(\boldsymbol{a}\,|\,\alpha,\beta)\to\beta} \longleftarrow \begin{bmatrix} n\mu_{q(\alpha)} \\ -\mathbf{1}_n^T E_{q(\boldsymbol{a})}(\boldsymbol{a}) \end{bmatrix},$$

where

$$\mu_{q(\alpha)} = \int_0^\infty \alpha q^*(\alpha)\, d\alpha,$$

$$\text{and} \quad \mu_{q(\beta)} = \int_0^\infty \beta q^*(\beta)\, d\beta,$$

with $q^*(\alpha)$ proportional to

$$m_{p(\boldsymbol{a}\,|\,\alpha,\beta)\to\alpha}(\alpha)\, m_{\alpha\to p(\boldsymbol{a}\,|\,\alpha,\beta)}(\alpha)$$

and $q^*(\beta)$ defined in an analogous way. An appropriate choice of messages entering the β node may facilitate the VMP algorithm derivation and computation. If, for example, $m_{\beta\to p(\boldsymbol{a}\,|\,,\alpha,\beta)}$ only receives a gamma prior on β then conjugacy will be ensured.

The sufficient statistic vector of messages $m_{p(\boldsymbol{a}\,|\,\alpha,\beta)\to\alpha}(\alpha)$ is not associable to any notorious exponential family instead. Expectation with respect to (1.16) indicates that the messages from $p(\boldsymbol{a}\,|\,\alpha,\beta)$ to α are proportional to a non-standard density function with natural parameter

$$\boldsymbol{\eta}_{p(\boldsymbol{a}\,|\,\alpha,\beta)\to\alpha} \longleftarrow \begin{bmatrix} -n \\ n\mu_{q(\log(\beta))} + \mathbf{1}_n^T E_{q(\boldsymbol{a})}(\log(\boldsymbol{a})) \end{bmatrix},$$

where

$$\mu_{q(\log(\beta))} = \int_0^\infty \log(\beta)\, q^*(\beta)\, d\beta.$$

However, if for instance $m_{\alpha\to p(\boldsymbol{a}\,|\,\alpha,\beta)}(\alpha)$ is an exponential density prior on α conjugacy is ensured and we can compute $\mu_{q(\alpha)}$ via numerical integration involving integrals of the form

$$\mathcal{J}(r,s) = \int_0^\infty \Gamma(x)^r\, x^s dx.$$

Then

$$\mu_{q(\alpha)} = \frac{\mathcal{J}\left(\left(\boldsymbol{\eta}_{p(\boldsymbol{a}\,|\,\alpha,\beta)\leftrightarrow\alpha}\right)_1, \left(\boldsymbol{\eta}_{p(\boldsymbol{a}\,|\,\alpha,\beta)\leftrightarrow\alpha}\right)_2 + 1\right)}{\mathcal{J}\left(\left(\boldsymbol{\eta}_{p(\boldsymbol{a}\,|\,\alpha,\beta)\leftrightarrow\alpha}\right)_1, \left(\boldsymbol{\eta}_{p(\boldsymbol{a}\,|\,\alpha,\beta)\leftrightarrow\alpha}\right)_2\right)}.$$

Working on the log-scale is recommended, as described in Section C.2.3 of the Appendix concerning numerical integration steps for the skew t fragment, paying attention to restrictions on coefficients r and s.

Now, noting that $m_{p(\boldsymbol{a}\,|\,\alpha,\beta)\to a_i}(a_i)$ is the only message that a_i receives and then passes to factor $p(\boldsymbol{y}\,|\,\mu,\boldsymbol{a})$, from $m_{p(\boldsymbol{y}\,|\,\mu,\boldsymbol{a})\to a_i}(a_i)$ and $m_{p(\boldsymbol{a}\,|\,\alpha,\beta)\to a_i}(a_i)$, we get the natural parameter vector of the density obtained normalizing (3.6),

$$\begin{aligned}\boldsymbol{\eta}_{q(\boldsymbol{a})} = &\begin{bmatrix} \left(\boldsymbol{\eta}_{q(\boldsymbol{a})}\right)_1 \\ \left(\boldsymbol{\eta}_{q(\boldsymbol{a})}\right)_2 \end{bmatrix} \\ =&\boldsymbol{\eta}_{p(\boldsymbol{y}\,|\,\mu,\boldsymbol{a})\to a_i} + \boldsymbol{\eta}_{p(\boldsymbol{a}\,|\,\alpha,\beta)\to a_i} \\ =&\begin{bmatrix} \mu_{q(\alpha)} \\ \mu_{q(\mu)} - \boldsymbol{y} - \mu_{q(\beta)} \end{bmatrix},\end{aligned}$$

which belongs to a gamma distribution, as both the messages that generate it.

Consequently, by applying result (A.2) in Appendix about the expectation of the sufficient statistic of a gamma distribution we get

$$\begin{aligned} E_{q(\boldsymbol{a})}\left(\log\left(\boldsymbol{a}\right)\right) =&\psi\left\{\left(\boldsymbol{\eta}_{q(\boldsymbol{a})}\right)_1 + 1\right\} - \log\left\{-\left(\boldsymbol{\eta}_{q(\boldsymbol{a})}\right)_2\right\} \\ =&\psi\left(\mu_{q(\alpha)} + 1\right) - \log\left(\boldsymbol{y} + \mu_{q(\beta)} - \mu_{q(\mu)}\right) \\ \text{and}\quad E_{q(\boldsymbol{a})}\left(\boldsymbol{a}\right) =&-\left\{\left(\boldsymbol{\eta}_{q(\boldsymbol{a})}\right)_1 + 1\right\} / \left(\boldsymbol{\eta}_{q(\boldsymbol{a})}\right)_2 \\ =&\left(\mu_{q(\alpha)} + 1\right) / \left(\boldsymbol{y} + \mu_{q(\beta)} - \mu_{q(\mu)}\right). \end{aligned}$$

In a similar way, under conjugacy, we get

$$\begin{aligned}\mu_{q(\log(\beta))} &= \psi\left\{\left(\boldsymbol{\eta}_{p(\boldsymbol{a}\,|\,\alpha,\beta)\leftrightarrow\beta}\right)_1 + 1\right\} - \log\left\{-\left(\boldsymbol{\eta}_{p(\boldsymbol{a}\,|\,\alpha,\beta)\leftrightarrow\beta}\right)_2\right\}, \\ \mu_{q(\beta)} &= \left\{\left(\boldsymbol{\eta}_{p(\boldsymbol{a}\,|\,\alpha,\beta)\leftrightarrow\beta}\right)_1 + 1\right\} / \left(\boldsymbol{\eta}_{p(\boldsymbol{a}\,|\,\alpha,\beta)\leftrightarrow\beta}\right)_2.\end{aligned}$$

Finally, the optimal approximating densities as in (1.17) for the variables of interest μ, α and β and the auxiliary variables a_is are given by messages (3.5), (3.7) and (3.9)–(3.11). Algorithm 3.1 summarizes the previous results and provides a VMP scheme for the Pareto random sample.

As an example, Algorithm 3.1 is run on a simulated dataset of size $n = 500$ with

Algorithm 3.1 The VMP inputs, updates and outputs of the Pareto random sample likelihood fragment assuming $q(\mu, \alpha, \beta, \boldsymbol{a}) = q(\mu)\, q(\alpha)\, q(\beta) \prod_{i=1}^{n} q(a_i)$.

Data Inputs: $\boldsymbol{y}$.

Parameter Inputs: $\eta_{p(\boldsymbol{y} \mid \mu, \boldsymbol{a}) \to \mu}$, $\eta_{\mu \to p(\boldsymbol{y} \mid \mu, \boldsymbol{a})}$, $\boldsymbol{\eta}_{p(\boldsymbol{a} \mid \alpha, \beta) \to \alpha}$, $\boldsymbol{\eta}_{\alpha \to p(\boldsymbol{a} \mid \alpha, \beta)}$, $\boldsymbol{\eta}_{p(\boldsymbol{a} \mid \alpha, \beta) \to \beta}$, $\boldsymbol{\eta}_{\beta \to p(\boldsymbol{a} \mid \alpha, \beta)}$.

Updates:

$$
\begin{aligned}
\mu_{q(\mu)} &\longleftarrow \min_{1 \leq i \leq n} (y_i) - \left(\eta_{p(y \mid \boldsymbol{a}, \mu) \leftrightarrow \mu}\right)^{-1} \\
\mu_{q(\alpha)} &\longleftarrow \frac{\mathcal{J}\left(\left(\boldsymbol{\eta}_{p(\boldsymbol{a} \mid \alpha, \beta) \leftrightarrow \alpha}\right)_1, \left(\boldsymbol{\eta}_{p(\boldsymbol{a} \mid \alpha, \beta) \leftrightarrow \alpha}\right)_2 + 1\right)}{\mathcal{J}\left(\left(\boldsymbol{\eta}_{p(\boldsymbol{a} \mid \alpha, \beta) \leftrightarrow \alpha}\right)_1, \left(\boldsymbol{\eta}_{p(\boldsymbol{a} \mid \alpha, \beta) \leftrightarrow \alpha}\right)_2\right)} \\
\mu_{q(\log(\beta))} &\longleftarrow \psi\left\{\left(\boldsymbol{\eta}_{p(\boldsymbol{a} \mid \alpha, \beta) \leftrightarrow \beta}\right)_1 + 1\right\} - \log\left\{-\left(\boldsymbol{\eta}_{p(\boldsymbol{a} \mid \alpha, \beta) \leftrightarrow \beta}\right)_2\right\} \\
\mu_{q(\beta)} &\longleftarrow \left\{\left(\boldsymbol{\eta}_{p(\boldsymbol{a} \mid \alpha, \beta) \leftrightarrow \beta}\right)_1 + 1\right\} / \left(\boldsymbol{\eta}_{p(\boldsymbol{a} \mid \alpha, \beta) \leftrightarrow \beta}\right)_2 \\
E_{q(\boldsymbol{a})}(\log(\boldsymbol{a})) &\longleftarrow \psi\left(\mu_{q(\alpha)} + 1\right) - \log\left(\boldsymbol{y} + \mu_{q(\beta)} - \mu_{q(\mu)}\right) \\
E_{q(\boldsymbol{a})}(\boldsymbol{a}) &\longleftarrow \left(\mu_{q(\alpha)} + 1\right) / \left\{\boldsymbol{y} + \mu_{q(\beta)} - \mu_{q(\mu)}\right\} \\
\eta_{p(y \mid \boldsymbol{a}, \mu) \to \mu} &\longleftarrow \mathbf{1}_n^T E_{q(\boldsymbol{a})}(\boldsymbol{a}) \\
\boldsymbol{\eta}_{p(\boldsymbol{a} \mid \alpha, \beta) \to \alpha} &\longleftarrow \begin{bmatrix} -n \\ n\mu_{q(\log(\beta))} + \mathbf{1}_n^T E_{q(\boldsymbol{a})}(\log(\boldsymbol{a})) \end{bmatrix} \\
\boldsymbol{\eta}_{p(\boldsymbol{a} \mid \alpha, \beta) \to \beta} &\longleftarrow \begin{bmatrix} n\mu_{q(\alpha)} \\ -\mathbf{1}_n^T E_{q(\boldsymbol{a})}(\boldsymbol{a}) \end{bmatrix}.
\end{aligned}
$$

Parameter Outputs: $\eta_{p(\boldsymbol{x} \mid \mu, \boldsymbol{a}) \to \mu}$, $\boldsymbol{\eta}_{p(\boldsymbol{a} \mid \alpha, \beta) \to \alpha}$, $\boldsymbol{\eta}_{p(\boldsymbol{a} \mid \alpha, \beta) \to \beta}$.

parameters $\mu = 2$, $\alpha = 2$ and $\beta = 1$. We integrate the Pareto random sample model (3.2) with some prior specifications. The prior distribution on μ is defined as (3.8) with hyperparameter $\lambda_\mu = 0.01$ while for α and β we use priors Gamma$(\alpha_\alpha, \beta_\alpha)$ and Gamma$(\alpha_\beta, \beta_\beta)$ respectively, with $\alpha_\alpha = \alpha_\beta = 1$ and $\beta_\alpha = \beta_\beta = 0.01$. We compare VMP approximate densities with the posterior densities of single parameters obtainable via Markov chain Monte Carlo (MCMC) through `rstan` with the `R` computing environment (R Core Team, 2018) interfacing via the `rstan` package (Stan Development Team, 2018). MCMC samples of size 10,000 were drawn setting a burn-in of 5000 values and thinning the remaining 5000 by a factor of 5. Figure 3.2 shows both the VMP and MCMC results. The density curves produced by Algorithm 3.1 seem to capture the modes of MCMC posterior densities and the true generating parameters. However, note the lower variance of variational approximating densities which corresponds with the theoretical results in Wang and Blei (2018) concerning variance underestimation of variational Bayes. A more structured and appropriate assessment of VMP performances is proposed for the skew t likelihood fragment.

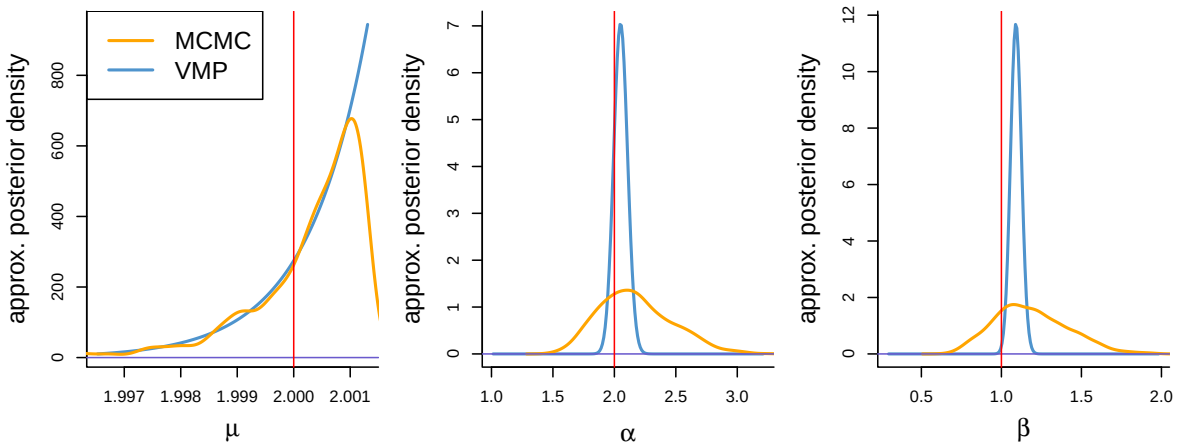

FIGURE 3.2: VMP-approximate and MCMC posterior density functions for a dataset simulated with parameters $\mu = 2$, $\alpha = 1$, $\beta = 2$. Vertical lines indicate the true values.

3.3 The support vector regression likelihood fragment

Support vector machine (SVM) tools, including *support vector regression* (SVR), are a class of popular supervised learning techniques used for classification and regression analysis. However they scale relatively badly with increasing sample size, due to their quadratic optimization algorithm, the use of kernel transformations for linear learning machine mapping and the choice of kernel parameters (e.g. Bennett and Campbell, 2000). Similarly to the classification approach, SVR seeks the optimal regression func-

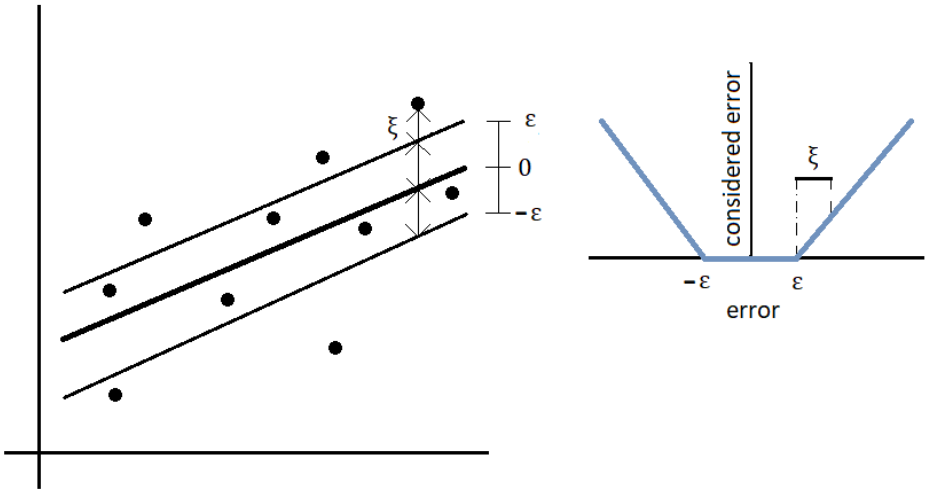

FIGURE 3.3: Loss function in support vector regression.

tion using a loss function that ignores errors which are situated within a certain distance from the true value. This type of function is often called *epsilon intensive loss function*.

Figure 3.3 shows an example of one-dimensional linear regression function with *epsilon intensive band*. The model variables measure the cost of the errors on the training points. These are zero for all points which fall inside the band.

We apply the result in Lemma 2 in Zhu *et al.* (2014) to write the pseudo-likelihood specification at the base of SVR:

$$\exp\left\{-2\left(|x|-\varepsilon\right)_+\right\} = \int_0^\infty (2\pi a_1)^{-1/2} \exp\left\{-\frac{(a_1+x-\varepsilon)^2}{2a_1}\right\} da_1 \times \int_0^\infty (2\pi a_2)^{-1/2} \exp\left\{-\frac{(a_2-x-\varepsilon)^2}{2a_2}\right\} da_2,$$

where $u_+ = \max(0,u)$ for any $u \in \mathbb{R}$. Indicating with $\check{p}$ an improper density function which does not integrate to 1, the previous expression can be rewritten as follows:

$$\check{p}(x) = \int_{-\infty}^\infty \int_{-\infty}^\infty p_1(x \mid a_1)\, p_2(x \mid a_2)\, \check{p}(a_1)\, \check{p}(a_2)\, da_1 da_2 = \exp\left\{-2\left(|x|-\varepsilon\right)_+\right\}, \quad (3.12)$$

where $p_1(x \mid a_1)$ is the $N(\varepsilon - a_1, a_1)$ density function in x, $p_2(x \mid a_2)$ is the $N(a_2 - \varepsilon, a_2)$ density function in x and $\check{p}(a_j) = I(a_j > 0)$, $j = 1, 2$. In other terms, the pseudo-density function $\check{p}(x)$ is represented as a mixture of specific normal density functions and two auxiliary variables pseudo-density functions $\check{p}(a_1)$ and $\check{p}(a_2)$.

The Support Vector Regression pseudo-likelihood fragments are concerned with the pseudo-likelihood specification

$$\check{p}(\boldsymbol{y} \mid \boldsymbol{\theta}) = \prod_{i=1}^n \exp\left\{-2\left(|\boldsymbol{y} - (\boldsymbol{A}\boldsymbol{\theta})_i| - \varepsilon\right)_+\right\}, \quad (3.13)$$

with $\boldsymbol{y}$ vector in $\mathbb{R}$. We now introduce two auxiliary variable vectors $\boldsymbol{a}_j = (a_{j1}, \ldots, a_{jn})$, $j = 1, 2$, with entries a_{ji}, $1 \le i \le n$, each independently having the pseudo-density function $\check{p}(a_{ji}) = I(a_{ji} > 0)$. Then, using (3.12), (3.13) is equivalent to

$$\begin{aligned} \check{p}(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{a}_1, \boldsymbol{a}_2) = &\prod_{i=1}^n (2\pi a_{1i})^{-1/2} \exp\left[-\frac{\{a_{1i} + y_i - (\boldsymbol{A}\boldsymbol{\theta})_i - \varepsilon\}^2}{2a_{1i}}\right] \\ &\times (2\pi a_{2i})^{-1/2} \exp\left[-\frac{\{a_{2i} - y_i + (\boldsymbol{A}\boldsymbol{\theta})_i - \varepsilon\}^2}{2a_{2i}}\right], \quad (3.14) \\ \check{p}(\boldsymbol{a}_1) = \prod_{i=1}^n I(a_{1i} > 0), \quad &\check{p}(\boldsymbol{a}_2) = \prod_{i=1}^n I(a_{2i} > 0). \end{aligned}$$

To produce one of the simplest and tractable VMP schemes, we propose the approximation of the full joint posterior density function of the form

$$p\left(\boldsymbol{\theta}, \boldsymbol{a}_1, \boldsymbol{a}_2 \,|\, \boldsymbol{y}\right) \approx q\left(\boldsymbol{\theta}, \boldsymbol{a}_1, \boldsymbol{a}_2\right)$$

subject to the q-density product restriction

$$\begin{aligned} q\left(\boldsymbol{\theta}, \boldsymbol{a}_1, \boldsymbol{a}_2\right) &= q\left(\boldsymbol{\theta}\right) q\left(\boldsymbol{a}_1\right) q\left(\boldsymbol{a}_2\right) \\ &= q\left(\boldsymbol{\theta}\right) \prod_{i=1}^{n} q\left(a_{1i}\right) q\left(a_{2i}\right). \end{aligned} \tag{3.15}$$

The likelihood specification (3.14) and the product density restriction (3.15) produce the factor graph representation in Figure 3.4. Messages from the factors to node appearing

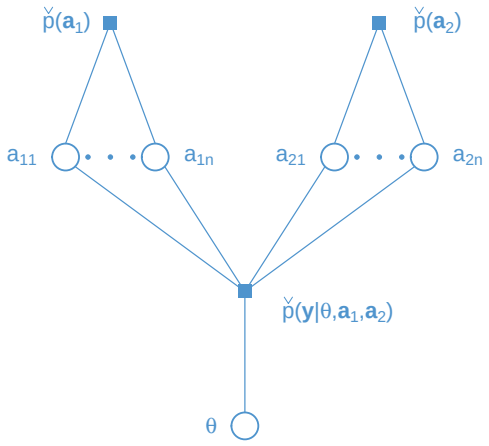

FIGURE 3.4: Factor graph for the support vector regression likelihood specification in (3.14) under the assumption in (3.15) with independent auxiliary variables $a_{11} \ldots a_{1n}$ and $a_{21} \ldots a_{2n}$.

in the factor graph representation are obtained by manipulation of the log-likelihood components as a function of the node of interest, applying steps (1.14)–(1.16). The messages passed from the pseudo-likelihood factor to the parameter vector of interest, $\boldsymbol{\theta}$, have the form

$$m_{\check{p}\left(\boldsymbol{y} \,|\, \boldsymbol{\theta}, \boldsymbol{a}_1, \boldsymbol{a}_2\right) \rightarrow \boldsymbol{\theta}}\left(\boldsymbol{\theta}\right) = \exp\left\{ \begin{bmatrix} \boldsymbol{\theta} \\ \operatorname{vec}\left(\boldsymbol{\theta}\boldsymbol{\theta}^T\right) \end{bmatrix}^T \boldsymbol{\eta}_{\check{p}\left(\boldsymbol{y} \,|\, \boldsymbol{\theta}, \boldsymbol{a}_1, \boldsymbol{a}_2\right) \rightarrow \boldsymbol{\theta}} \right\},$$

which has a multivariate normal structure. Therefore, to ensure conjugacy, messages that $\boldsymbol{\theta}$ receives from factors outside of the SVR likelihood fragment, such as a prior on $\boldsymbol{\theta}$, have to be proportional to a multivariate normal density. The structures of these messages together with those involving the auxiliary random variables derived in Section C.1.1 of the Appendix produce the VMP scheme listed as Algorithm 3.2.

Algorithm 3.2 The VMP inputs, updates and outputs of the support vector regression likelihood fragment assuming $q(\boldsymbol{\theta}, \boldsymbol{a}_1, \boldsymbol{a}_2) = q(\boldsymbol{\theta}) \prod_{i=1}^{n} q(a_{1i}) q(a_{2i})$.

Data Inputs: $\boldsymbol{y}$, $\boldsymbol{A}$.

Parameter Inputs: $\boldsymbol{\eta}_{\check{p}(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{a}_1, \boldsymbol{a}_2) \to \boldsymbol{\theta}}$, $\boldsymbol{\eta}_{\boldsymbol{\theta} \to \check{p}(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{a}_1, \boldsymbol{a}_2)}$.

Updates:

$$
\begin{aligned}
\boldsymbol{\upsilon}_1 &\longleftarrow -\tfrac{1}{2}\boldsymbol{A}\left\{\mathrm{vec}^{-1}\left(\left(\boldsymbol{\eta}_{\check{p}(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{a}_1, \boldsymbol{a}_2) \leftrightarrow \boldsymbol{\theta}}\right)_2\right)\right\}^{-1}\left(\boldsymbol{\eta}_{\check{p}(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{a}_1, \boldsymbol{a}_2) \leftrightarrow \boldsymbol{\theta}}\right)_1 \\
\boldsymbol{\upsilon}_2 &\longleftarrow -\tfrac{1}{2}\mathrm{dg}\left[\boldsymbol{A}\left\{\mathrm{vec}^{-1}\left(\left(\boldsymbol{\eta}_{\check{p}(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{a}_1, \boldsymbol{a}_2) \leftrightarrow \boldsymbol{\theta}}\right)_2\right)\right\}^{-1}\boldsymbol{A}^T\right] \\
\boldsymbol{\upsilon}_3 &\longleftarrow \left\{(\boldsymbol{\upsilon}_1 + \varepsilon \mathbf{1}_n - \boldsymbol{y})^2 + \boldsymbol{\upsilon}_2\right\}^{-1/2} \\
\boldsymbol{\upsilon}_4 &\longleftarrow \left\{(\boldsymbol{y} - \boldsymbol{\upsilon}_1 + \varepsilon \mathbf{1}_n)^2 + \boldsymbol{\upsilon}_2\right\}^{-1/2} \\
\boldsymbol{\eta}_{\check{p}(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{a}_1, \boldsymbol{a}_2) \to \boldsymbol{\theta}} &\longleftarrow \begin{bmatrix} \boldsymbol{A}^T\{\boldsymbol{\upsilon}_3 \odot (\boldsymbol{y} - \mathbf{1}_n\varepsilon) + \boldsymbol{\upsilon}_4 \odot (\boldsymbol{y} + \mathbf{1}_n\varepsilon)\} \\ -\tfrac{1}{2}\mathrm{vec}\left\{\boldsymbol{A}^T \mathrm{diag}(\boldsymbol{\upsilon}_3 + \boldsymbol{\upsilon}_4)\boldsymbol{A}\right\} \end{bmatrix}.
\end{aligned}
$$

Parameter Outputs: $\boldsymbol{\eta}_{\check{p}(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{a}_1, \boldsymbol{a}_2) \to \boldsymbol{\theta}}$.

3.3.1 Approximate inference via mean field variational Bayes

Mean field variational Bayes does not benefit of the fragmentation property of VMP. For this reason additional model details such as parameter priors are required to derive a complete variational algorithm. We add to the Bayesian model (3.14) a prior specification $p(\boldsymbol{\theta})$ on the vector $\boldsymbol{\theta}$ such that

$$
\boldsymbol{\theta} \sim N(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}), \tag{3.16}
$$

where $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ is a vector of length d and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ is a positive semidefinite matrix of size $d \times d$, to derive the corresponding MFVB algorithm.

As a starting point, consider the DAG for model (3.14) under the prior specification (3.16) in Figure 3.5. Once again, to achieve tractability, we approximate the full joint posterior density function with an approximating density q which is subject to the density product restriction (3.15).

The q-densities are chosen to minimise the Kullback-Leibler divergence between the full joint posterior density function and (3.15).

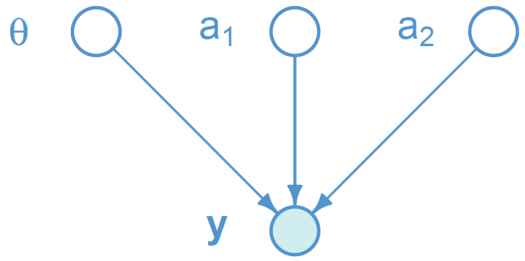

FIGURE 3.5: Directed acyclic graph for the support vector regression likelihood specification in (3.14).

First note that the full conditional of $\boldsymbol{\theta}$ satisfies

$$p\left(\boldsymbol{\theta} \,|\, \text{rest}\right) \propto \check{p}\left(\boldsymbol{y} \,|\, \boldsymbol{\theta}, \boldsymbol{a}_1, \boldsymbol{a}_2\right) p\left(\boldsymbol{\theta}\right),$$

where "rest" denotes all of the random variables included in the Markov blanket of the variable of interest, according to Figure 3.5. Taking logarithms on both sides gives

$$\begin{aligned}
\log p\left(\boldsymbol{\theta} \,|\, \text{rest}\right) &= \log \check{p}\left(\boldsymbol{y} \,|\, \boldsymbol{\theta}, \boldsymbol{a}_1, \boldsymbol{a}_2\right) + \log p\left(\boldsymbol{\theta}\right) + \text{const} \\
&= -\frac{1}{2}\Big\{\left(\boldsymbol{a}_1 + \boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta} - \varepsilon\right)^T \text{diag}\left(\boldsymbol{a}_1\right)^{-1}\left(\boldsymbol{a}_1 + \boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta} - \varepsilon\right) \\
&\quad + \left(\boldsymbol{a}_2 - \boldsymbol{y} + \boldsymbol{A}\boldsymbol{\theta} - \varepsilon\right)^T \text{diag}\left(\boldsymbol{a}_2\right)^{-1}\left(\boldsymbol{a}_2 - \boldsymbol{y} + \boldsymbol{A}\boldsymbol{\theta} - \varepsilon\right)\Big\} \\
&\quad - \frac{1}{2}\left(\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}}\right)^T \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\left(\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}}\right) + \text{const} \\
&= \begin{bmatrix} \boldsymbol{\theta} \\ \text{vec}\left(\boldsymbol{\theta}\boldsymbol{\theta}^T\right) \end{bmatrix}^T \begin{bmatrix} \boldsymbol{A}^T\left\{\frac{1}{\boldsymbol{a}_1} \odot \left(\boldsymbol{y} - \varepsilon\right) + \frac{1}{\boldsymbol{a}_2} \odot \left(\boldsymbol{y} + \varepsilon\right)\right\} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\theta}} \\ -\frac{1}{2}\text{vec}\left\{\boldsymbol{A}^T \text{diag}\left(\frac{1}{\boldsymbol{a}_1} + \frac{1}{\boldsymbol{a}_2}\right)\boldsymbol{A}\right\} - \frac{1}{2}\text{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right) \end{bmatrix} + \text{const}.
\end{aligned}$$

Taking expectations with respect to all parameters except $\boldsymbol{\theta}$ it follows that the optimal q-density for $\boldsymbol{\theta}$ is

$$\begin{aligned}
q^*\left(\boldsymbol{\theta}\right) &\propto \exp\left\{E_{q(\boldsymbol{a}_1, \boldsymbol{a}_2)} \log p\left(\boldsymbol{\theta} \,|\, \text{rest}\,\right)\right\} \\
&= \exp\left(\begin{bmatrix} \boldsymbol{\theta} \\ \text{vec}\left(\boldsymbol{\theta}\boldsymbol{\theta}^T\right) \end{bmatrix}^T \begin{bmatrix} \boldsymbol{A}^T\left\{\mu_{q(1/\boldsymbol{a}_1)} \odot \left(\boldsymbol{y} - \varepsilon\right) \right. \\ \left. + \mu_{q(1/\boldsymbol{a}_2)} \odot \left(\boldsymbol{y} + \varepsilon\right)\right\} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\theta}} \\ -\frac{1}{2}\left[\text{vec}\left\{\boldsymbol{A}^T \text{diag}\left(\mu_{q(1/\boldsymbol{a}_1)} + \mu_{q(1/\boldsymbol{a}_2)}\right)\boldsymbol{A} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right\}\right] \end{bmatrix}\right),
\end{aligned}$$

with $\mu_{q(1/\boldsymbol{a}_1)} = \left[\mu_{q(1/a_{11})}, \ldots, \mu_{q(1/a_{1n})}\right]^T$ and $\mu_{q(1/\boldsymbol{a}_2)} = \left[\mu_{q(1/a_{21})}, \ldots, \mu_{q(1/a_{2n})}\right]^T$. The optimal density $q^*\left(\boldsymbol{\theta}\right)$ is then a multivariate normal $N\left(\boldsymbol{\mu}_{q(\boldsymbol{\theta})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}\right)$ density function

with

$$\begin{aligned}\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})} &= \left(\text{vec}^{-1}\left[\text{vec}\left\{\boldsymbol{A}^T \text{diag}\left(\mu_{q(1/\boldsymbol{a}_1)} + \mu_{q(1/\boldsymbol{a}_2)}\right)\boldsymbol{A}\right\} + \text{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right)\right]\right)^{-1}, \\ \boldsymbol{\mu}_{q(\boldsymbol{\theta})} &= \boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}^{-1}\left[\boldsymbol{A}^T\left\{\mu_{q(1/\boldsymbol{a}_1)} \odot (\boldsymbol{y} - \varepsilon) + \mu_{q(1/\boldsymbol{a}_2)} \odot (\boldsymbol{y} + \varepsilon)\right\} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\theta}}\right].\end{aligned} \tag{3.17}$$

Expressions for $\mu_{q(1/\boldsymbol{a}_1)}$ and $\mu_{q(1/\boldsymbol{a}_2)}$ are derived from the full conditionals for $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$. The full conditional of $\boldsymbol{a}_1$ satisfies

$$\begin{aligned}\log p\left(\boldsymbol{a}_1 \,|\, \text{rest}\right) &= \log \check{p}\left(\boldsymbol{y} \,|\, \boldsymbol{\theta}, \boldsymbol{a}_1, \boldsymbol{a}_2\right) + \log p\left(\boldsymbol{a}_1\right) + \text{const} \\ &= -\frac{1}{2}\sum_{i=1}^{n}\log\left(a_{1i}\right) - \frac{1}{2}\sum_{i=1}^{n}\frac{1}{a_{1i}}\left[a_{1i} - \left\{\left(\boldsymbol{A}\boldsymbol{\theta}\right)_i + \varepsilon + y_i\right\}\right]^2 + \text{const} \\ &= \sum_{i=1}^{n}\log\left[a_{1i}^{-1/2}\exp\left\{\begin{bmatrix} a_{1i} \\ 1/a_{1i}\end{bmatrix}^T\begin{bmatrix} -\frac{1}{2} \\ -\frac{1}{2}\left\{\left(\boldsymbol{A}\boldsymbol{\theta}\right)_i + \varepsilon - y_i\right\}^2\end{bmatrix}\right\}\right] + \text{const},\end{aligned}$$

from which we derive $q^*\left(\boldsymbol{a}_1\right)$ as the product of the $q\left(a_{1i}\right)$ densities, $1 \leq i \leq n$, taking expectation with respect to $\boldsymbol{\theta}$ and $\boldsymbol{a}_2$. It follows that

$$\begin{aligned}q^*\left(a_{1i}\right) &\propto \exp\left\{E_{q(\boldsymbol{\theta})}\log p\left(\boldsymbol{a}_1 \,|\, \text{rest}\right)\right\} \\ &= a_{1i}^{-1/2}\exp\left\{\begin{bmatrix} a_{1i} \\ 1/a_{1i}\end{bmatrix}^T\begin{bmatrix} -\frac{1}{2} \\ -\frac{1}{2}E_{q(\boldsymbol{\theta})}\left\{\left(\boldsymbol{A}\boldsymbol{\theta}\right)_i + \varepsilon - y_i\right\}^2\end{bmatrix}\right\}.\end{aligned}$$

Using result (A.3) concerning the expectation of the sufficient statistic of the generalized inverse Gaussian distribution we have

$$\mu_{q(1/a_{1i})} = \left[\left\{\left(\boldsymbol{A}\boldsymbol{\mu}_{q(\boldsymbol{\theta})}\right)_i + \varepsilon - y_i\right\}^2 + \left(\boldsymbol{A}\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}\boldsymbol{A}^T\right)_{ii}\right]^{-1/2}. \tag{3.18}$$

Similarly, for $\boldsymbol{a}_2$ and $1 \leq i \leq n$,

$$\mu_{q(1/a_{2i})} = \left[\left\{y_i - \left(\boldsymbol{A}\boldsymbol{\mu}_{q(\boldsymbol{\theta})}\right)_i + \varepsilon\right\}^2 + \left(\boldsymbol{A}\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}\boldsymbol{A}^T\right)_{ii}\right]^{-1/2}. \tag{3.19}$$

Expressions (3.17), (3.18) and (3.19) provide the MFVB coordinate ascent procedure to obtain the optimal variational parameters for the SVR problem, which is here proposed as Algorithm 3.3 .

Algorithm 3.3 MFVB coordinate ascent procedure to obtain the parameters in the optimal densities $q^*(\boldsymbol{\theta})$, $q^*(\boldsymbol{a}_1)$ and $q^*(\boldsymbol{a}_2)$ for the support vector regression model assuming $q(\boldsymbol{\theta}, \boldsymbol{a}_1, \boldsymbol{a}_2) = q(\boldsymbol{\theta}) \prod_{i=1}^{n} q(a_{1i}) q(a_{2i})$.

Initialize: $\boldsymbol{\mu}_{q(\boldsymbol{\theta})}$ $(d \times 1)$, $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ $(d \times d)$ symmetric and positive definite.

Cycle:

For $1 \leq i \leq n$:

$$\mu_{q(1/a_{1i})} \longleftarrow \left[\left\{ \left(\boldsymbol{A}\boldsymbol{\mu}_{q(\boldsymbol{\theta})} \right)_i + \varepsilon - y_i \right\}^2 + \left(\boldsymbol{A}\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}\boldsymbol{A}^T \right)_{ii} \right]^{-1/2}$$

$$\mu_{q(1/a_{2i})} \longleftarrow \left[\left\{ y_i - \left(\boldsymbol{A}\boldsymbol{\mu}_{q(\boldsymbol{\theta})} \right)_i + \varepsilon \right\}^2 + \left(\boldsymbol{A}\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}\boldsymbol{A}^T \right)_{ii} \right]^{-1/2}.$$

Update:

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})} \longleftarrow \left(\text{vec}^{-1} \left[\text{vec} \left\{ \boldsymbol{A}^T \text{diag} \left(\mu_{q(1/\boldsymbol{a}_1)} + \mu_{q(1/\boldsymbol{a}_2)} \right) \boldsymbol{A} \right\} + \text{vec} \left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \right) \right] \right)^{-1}$$

$$\boldsymbol{\mu}_{q(\boldsymbol{\theta})} \longleftarrow \boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}^{-1} \left[\boldsymbol{A}^T \left\{ \mu_{q(1/\boldsymbol{a}_1)} \odot (\boldsymbol{y} - \varepsilon) + \mu_{q(1/\boldsymbol{a}_2)} \odot (\boldsymbol{y} + \varepsilon) \right\} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\theta}} \right]$$

until convergence.

3.4 The skew t likelihood fragment

Skew distributions have emerged as a popular tool in modelling heterogeneous data with asymmetric features. The well-know skew normal distribution certainly plays a major role. However, as pointed out in Frühwirth-Schnatter and Pyne (2010), the kurtosis coefficient of a skew normal distribution is restricted to the interval $[3, 3.8692]$. To achieve a higher degree of excess kurtosis, *skew t* distributions have been introduced.

The skew t likelihood fragment corresponds to the likelihood specification

$$y_i \,|\, \boldsymbol{\theta}, \sigma^2, \lambda, \nu \overset{\text{ind.}}{\sim} \text{Skew-t} \left((\boldsymbol{A}\boldsymbol{\theta})_i, \sigma^2, \lambda, \nu \right), \quad 1 \leq i \leq n, \tag{3.20}$$

where $\boldsymbol{A}$ is a generic design matrix, $\boldsymbol{\theta}$ is a generic vector of coefficients, $\sigma^2 > 0$, $\lambda \in \mathbb{R}$ and $\nu > 0$. Among the possible definitions of the skew t distribution, we consider the one described in Azzalini and Capitanio (2003) and recalled in Subsection A.2.20 of Appendix A. Their skew t distribution becomes a symmetric Student's t distribution when $\lambda = 0$, a conditional normal distribution as $\nu \to \infty$ and allows the inclusion of left-tailed or negative skewness when $\lambda < 0$ and right-tailed or positive skewness when $\lambda > 0$. One of the advantages of treating the skew t fragment under the VMP framework is that all the parameters can be inferred, rather than being held fixed.

The response likelihood can be conveniently re-expressed in terms of auxiliary variables and more common distributions to aid the construction of a tractable VMP algorithm. The introduction of auxiliary variables has the practical advantage of reducing

the complexity of message updates which either can be expressed in closed form or require only univariate numerical integration.

Equation (25) of Azzalini and Capitanio (2003) or Proposition 2.1 of Parisi and Liseo (2018) suggest a useful auxiliary variable representation for the skew t distribution. If we introduce two auxiliary random variables a_{1i} and a_{2i}, $1 \leq i \leq n$, such that

$$a_{1i} \overset{\text{ind.}}{\sim} N(0,1) \quad \text{and} \quad a_{2i} \overset{\text{ind.}}{\sim} \text{Inverse-}\chi^2(\nu,\nu),$$

then, with standard distribution manipulations, the model in (3.20) can be alternatively written as

$$\begin{aligned} y_i \,|\, \boldsymbol{\theta}, \sigma^2, \lambda, a_{1i}, a_{2i} &\overset{\text{ind.}}{\sim} N\left((\boldsymbol{A}\boldsymbol{\theta})_i + \frac{\sigma\lambda\,|a_{1i}|\,\sqrt{a_{2i}}}{\sqrt{1+\lambda^2}}, \frac{a_{2i}\sigma^2}{1+\lambda^2} \right), \\ a_{1i} &\overset{\text{ind.}}{\sim} N(0,1), \quad a_{2i} \,|\, \nu \overset{\text{ind.}}{\sim} \text{Inverse-}\chi^2(\nu,\nu). \end{aligned} \tag{3.21}$$

Considering this last model specification, we provide fragment updates that allow for the skew t distribution to be handled within the VMP framework. An assumption on the optimal q-density product restriction is required to produce the VMP solution in (1.17). An assumption producing one of the simplest VMP schemes is

$$\begin{aligned} q\left(\boldsymbol{\theta}, \sigma^2, \lambda, \nu, \boldsymbol{a}_1, \boldsymbol{a}_2\right) &= q(\boldsymbol{\theta})\, q\left(\sigma^2\right) q(\lambda)\, q(\nu)\, q(\boldsymbol{a}_1)\, q(\boldsymbol{a}_2) \\ &= q(\boldsymbol{\theta})\, q\left(\sigma^2\right) q(\lambda)\, q(\nu) \prod_{i=1}^{n} q(a_{1i})\, q(a_{2i}). \end{aligned} \tag{3.22}$$

Combining this product density restriction with the likelihood model in (3.21), we obtain the factor graph representation in the left panel of Figure 3.6. The structure of messages from the likelihood factor to each node is obtained by manipulation of the log-likelihood factor as a function of the node of interest, according to the VMP equations (1.14)–(1.16).

The messages passed from the likelihood factor to $\boldsymbol{\theta}$ take the form

$$m_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\boldsymbol{\theta}}(\boldsymbol{\theta}) = \exp\left\{ \begin{bmatrix} \boldsymbol{\theta} \\ \text{vec}\left(\boldsymbol{\theta}\boldsymbol{\theta}^T\right) \end{bmatrix}^T \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\boldsymbol{\theta}} \right\},$$

which has multivariate normal structure. Therefore, to ensure conjugacy, messages that $\boldsymbol{\theta}$ receives from factors outside of the skew t likelihood fragment, such as a prior on $\boldsymbol{\theta}$, have to be proportional to a multivariate normal density.

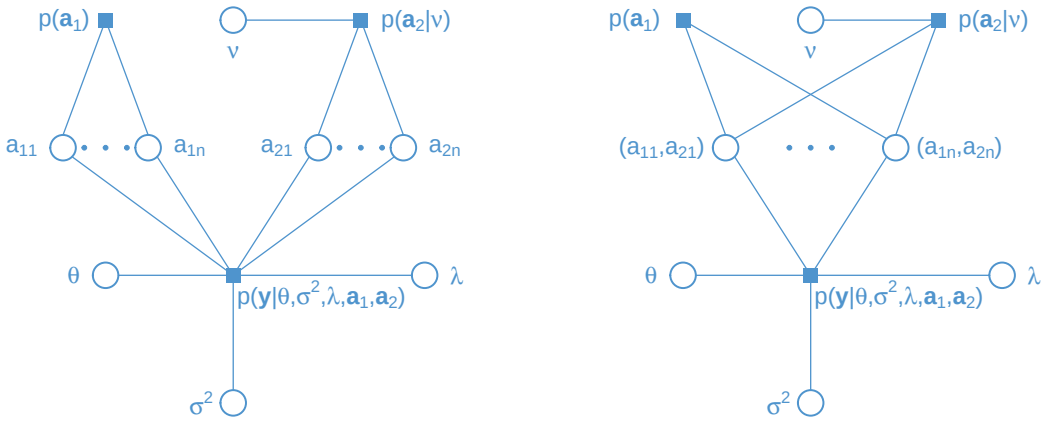

FIGURE 3.6: Factor graph for the skew t likelihood specification in (3.21) with independent $N(0,1)$ auxiliary variables $a_{11} \ldots a_{1n}$ and independent Inverse-$\chi^2(\nu,\nu)$ auxiliary variables $a_{21} \ldots a_{2n}$ under the assumption in (3.22) (left panel) and (3.24) (right panel).

The messages passed from the likelihood factor to σ^2 have the form

$$m_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\sigma^2}\left(\sigma^2\right) = \exp\left\{\begin{bmatrix}\log\left(\sigma^2\right)\\ 1/\sigma\\ 1/\sigma^2\end{bmatrix}^T \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\sigma^2}\right\},$$

which is within the *inverse square root Nadarajah* family described in McLean and Wand (2018, Section S.2.3). The imposition of conjugacy means that we assume that all messages passed to σ^2 from factors outside of the skew t fragment are also proportional to inverse square root Nadarajah density functions. For instance, an Inverse-χ^2 prior on σ^2 is suitable to ensure conjugacy.

The message from the likelihood factor to λ has the exponential family form

$$m_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\lambda}(\lambda) = \exp\left\{\begin{bmatrix}\log\left(1+\lambda^2\right)\\ \lambda^2\\ \lambda\sqrt{1+\lambda^2}\end{bmatrix}^T \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\lambda}\right\},$$

which is within the *Sea Sponge* family identified in McLean and Wand (2018, Section S.2.5). We assume that each of the messages that λ receives from factors outside of this fragment are conjugate to Sea Sponge density functions. If, for instance, the only factor that sends a message to λ is the prior density function $p(\lambda)$, then $m_{p(\lambda)\to\lambda}(\lambda) = p(\lambda)$

and, under conjugacy, $p(\lambda)$ must be of the form

$$p(\lambda) \propto \exp\left\{ \begin{bmatrix} \log(1+\lambda^2) \\ \lambda^2 \\ \lambda\sqrt{1+\lambda^2} \end{bmatrix}^T \boldsymbol{\eta}_\lambda \right\}, \tag{3.23}$$

for some 3×1 vector $\boldsymbol{\eta}_\lambda$. A special case of (3.23) is priors of the form $\lambda \sim N(0, \sigma_\lambda^2)$, having $\boldsymbol{\eta}_\lambda = [0, -1/(2\sigma_\lambda^2), 0]$.

As a function of ν we have

$$\log p(\boldsymbol{a}_2|\nu) = \begin{bmatrix} (\nu/2)\log(\nu/2) - \log\{\Gamma(\nu/2)\} \\ (\nu/2) \end{bmatrix}^T \begin{bmatrix} n \\ -\mathbf{1}_n^T\{\log(\boldsymbol{a}_2) + \mathbf{1}_n/\boldsymbol{a}_2\} \end{bmatrix} + \text{const},$$

indicating that messages from $p(\boldsymbol{a}_2|\nu)$ to a_{2i}, $1 \leq i \leq n$, are within the *Moon Rock* family defined McLean and Wand (2018, Section S.2.4). We assume messages passed to ν from factors outside the skew t likelihood fragments are conjugate with the Moon Rock family. For example, if the only other factor passing messages to ν is its prior density function $p(\nu)$ then we require that $p(\nu)$ is a Moon Rock density function or conjugate with one, such as an exponential density function.

The structures of these messages serve as a base to build a VMP algorithm on assumption (3.22). Algorithm 3.4 contains a listing of such a VMP scheme while algebraic derivations are given in Section C.2.1 of the Appendix.

However, the implementation of such an algorithm in simulation studies reveals poor performances of VMP if roughly compared with the posterior densities of single parameters obtainable via MCMC. The cause of this discrepancy is the strong posterior dependence between the two auxiliary variables a_{1i} and a_{2i}, whereas the product density factorization in (3.22) ignores such a dependence. Figure 3.7 provides some insight into why VMP developed according to assumption (3.22) is prone to inaccuracy, showing pairwise scatterplots of series $|a_1|$ and $1/\sqrt{a_2}$ from an MCMC fitting output of a skew t random sample of size $n = 1000$ as in the figure description. As for the Pareto likelihood fragment, MCMC draws are obtained through `rstan`. Expectations of these series involving the auxiliary random variables appear when deriving message updates. It is apparent that the posterior correlation between the auxiliary variables increases as the value of λ increases.

The following result confirms this posterior correlation problem.

Algorithm 3.4 The VMP inputs, updates and outputs of the skew t likelihood fragment assuming $q\left(\boldsymbol{\theta},\sigma^2,\lambda,\nu,\boldsymbol{a}_1,\boldsymbol{a}_2\right)=q\left(\boldsymbol{\theta}\right)q\left(\sigma^2\right)q\left(\lambda\right)q\left(\nu\right)\prod_{i=1}^{n}q\left(a_{1i}\right)q\left(a_{2i}\right)$.

Data Inputs: $\boldsymbol{y}$, $\boldsymbol{A}$.

Parameter Inputs: $\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\rightarrow\boldsymbol{\theta}}$, $\boldsymbol{\eta}_{\boldsymbol{\theta}\rightarrow p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)}$, $\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\rightarrow\sigma^2}$, $\boldsymbol{\eta}_{\sigma^2\rightarrow p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)}$, $\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\rightarrow\lambda}$, $\boldsymbol{\eta}_{\lambda\rightarrow p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)}$, $\boldsymbol{\eta}_{p(\boldsymbol{a}_2|\nu)\rightarrow\nu}$, $\boldsymbol{\eta}_{\nu\rightarrow p(\boldsymbol{a}_2|\nu)}$, $E_{q(\boldsymbol{a}_2)}\left(\boldsymbol{1}_n/\sqrt{\boldsymbol{a}_2}\right)$.

Updates:

$$\mu_{q(1/\sigma)}\longleftarrow\left(E\boldsymbol{T}\right)_2^{\text{ISRN}}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\leftrightarrow\sigma^2}\right)$$

$$\mu_{q(1/\sigma^2)}\longleftarrow\left(E\boldsymbol{T}\right)_3^{\text{ISRN}}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\leftrightarrow\sigma^2}\right)$$

$$\mu_{q(\lambda^2)}\longleftarrow\left(E\boldsymbol{T}\right)_2^{\text{SS}}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\leftrightarrow\lambda}\right)$$

$$\mu_{q\left(\lambda\sqrt{1+\lambda^2}\right)}\longleftarrow\left(E\boldsymbol{T}\right)_3^{\text{SS}}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\leftrightarrow\lambda}\right)$$

$$\mu_{q(\nu)}\longleftarrow 2\left(E\boldsymbol{T}\right)_2^{\text{MR}}\left(\boldsymbol{\eta}_{p(\boldsymbol{a}_2|\nu)\leftrightarrow\nu}\right)$$

$$\boldsymbol{\omega}_1\longleftarrow\boldsymbol{y}+\tfrac{1}{2}\boldsymbol{A}\left\{\text{vec}^{-1}\left(\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\leftrightarrow\boldsymbol{\theta}}\right)_2\right)\right\}^{-1}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\leftrightarrow\boldsymbol{\theta}}\right)_1$$

$$\boldsymbol{\omega}_2\longleftarrow E_{q(\boldsymbol{a}_2)}\left(\boldsymbol{1}_n/\sqrt{\boldsymbol{a}_2}\right)\odot\boldsymbol{\omega}_1$$

$$\boldsymbol{\omega}_3\longleftarrow\frac{\mu_{q(1/\sigma)}\mu_{q\left(\lambda\sqrt{1+\lambda^2}\right)}\boldsymbol{\omega}_2}{\sqrt{1+\mu_{q(\lambda^2)}}}$$

$$E_{q(\boldsymbol{a}_1)}\left|\boldsymbol{a}_1\right|\longleftarrow\frac{\boldsymbol{\omega}_3+\zeta^{'}\left(\boldsymbol{\omega}_3\right)}{\sqrt{1+\mu_{q(\lambda^2)}}}$$

$$E_{q(\boldsymbol{a}_1)}\left\|\boldsymbol{a}_1\right\|^2\longleftarrow\frac{n+\boldsymbol{1}_n^T\left[\boldsymbol{\omega}_3\odot\left\{\boldsymbol{\omega}_3+\zeta^{'}\left(\boldsymbol{\omega}_3\right)\right\}\right]}{1+\mu_{q(\lambda^2)}}$$

$$\boldsymbol{\omega}_4\longleftarrow\left[G_{\text{VMP}}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\leftrightarrow\boldsymbol{\theta}};\boldsymbol{A}^T\boldsymbol{e}_i\boldsymbol{e}_i^T\boldsymbol{A},\boldsymbol{A}^T\boldsymbol{e}_i\boldsymbol{e}_i^T\boldsymbol{y},y_i^2\right)\right]_{1\leq i\leq n}$$

$$\boldsymbol{\eta}_{q(\boldsymbol{a}_2)}\longleftarrow\begin{bmatrix}-\frac{1}{2}\mu_{q(\nu)}-\frac{3}{2}\\ \mu_{q(1/\sigma)}\mu_{q\left(\lambda\sqrt{1+\lambda^2}\right)}\boldsymbol{\omega}_1\odot E_{q(\boldsymbol{a}_1)}\left|\boldsymbol{a}_1\right|\\ \mu_{q(1/\sigma^2)}\left(1+\mu_{q(\lambda^2)}\right)\boldsymbol{\omega}_4-\frac{1}{2}\mu_{q(\nu)}\end{bmatrix}$$

$$E_{q(\boldsymbol{a}_2)}\left(\log\left(\boldsymbol{a}_2\right)\right)\longleftarrow\left(E\boldsymbol{T}\right)_1^{\text{ISRN}}\left(\boldsymbol{\eta}_{q(\boldsymbol{a}_2)}\right)$$

$$E_{q(\boldsymbol{a}_2)}\left(\boldsymbol{1}_n/\sqrt{\boldsymbol{a}_2}\right)\longleftarrow\left(E\boldsymbol{T}\right)_2^{\text{ISRN}}\left(\boldsymbol{\eta}_{q(\boldsymbol{a}_2)}\right)$$

$$E_{q(\boldsymbol{a}_2)}\left(\boldsymbol{1}_n/\boldsymbol{a}_2\right)\longleftarrow\left(E\boldsymbol{T}\right)_3^{\text{ISRN}}\left(\boldsymbol{\eta}_{q(\boldsymbol{a}_2)}\right)$$

$$\boldsymbol{\omega}_5\longleftarrow G_{\text{VMP}}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\leftrightarrow\boldsymbol{\theta}};\boldsymbol{A}^T\text{diag}\left\{E_{q(\boldsymbol{a}_2)}\left(\boldsymbol{1}_n/\boldsymbol{a}_2\right)\right\}\boldsymbol{A},,\right.$$
$$\left.\boldsymbol{A}^T\text{diag}\left\{E_{q(\boldsymbol{a}_2)}\left(\boldsymbol{1}_n/\boldsymbol{a}_2\right)\right\}\boldsymbol{y}\boldsymbol{y}^T\text{diag}\left\{E_{q(\boldsymbol{a}_2)}\left(\boldsymbol{1}_n/\boldsymbol{a}_2\right)\right\}\boldsymbol{y}\right)$$

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\rightarrow\boldsymbol{\theta}}\longleftarrow\left(1+\mu_{q(\lambda^2)}\right)\mu_{q(1/\sigma^2)}\begin{bmatrix}\boldsymbol{A}^T\text{diag}\left\{E_{q(\boldsymbol{a}_2)}\left(\boldsymbol{1}_n/\boldsymbol{a}_2\right)\right\}\boldsymbol{y}\\ -\frac{1}{2}\text{vec}\left(\boldsymbol{A}^T\text{diag}\left\{E_{q(\boldsymbol{a}_2)}\left(\boldsymbol{1}_n/\boldsymbol{a}_2\right)\right\}\boldsymbol{A}\right)\end{bmatrix}$$
$$-\mu_{q\left(\lambda\sqrt{1+\lambda^2}\right)}\mu_{q(1/\sigma)}\begin{bmatrix}\boldsymbol{A}^T\text{diag}\left\{E_{q(\boldsymbol{a}_2)}\left(\boldsymbol{1}_n/\sqrt{\boldsymbol{a}_2}\right)\right\}E_{q(\boldsymbol{a}_1)}\left|\boldsymbol{a}_1\right|\\ \boldsymbol{0}\end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\rightarrow\sigma^2}\longleftarrow\begin{bmatrix}-n/2\\ \mu_{q\left(\lambda\sqrt{1+\lambda^2}\right)}\boldsymbol{\omega}_2^TE_{q(\boldsymbol{a}_1)}\left|\boldsymbol{a}_1\right|\\ \left(1+\mu_{q(\lambda^2)}\right)\boldsymbol{\omega}_5\end{bmatrix}$$

Continued overleaf...

Continuing...

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\lambda} \longleftarrow \begin{bmatrix} n/2 \\ \mu_{q(1/\sigma^2)}\boldsymbol{\omega}_5 - \frac{1}{2}E_{q(\boldsymbol{a}_1)}\|\boldsymbol{a}_1\|^2 \\ \mu_{q(1/\sigma)}\boldsymbol{\omega}_2^T E_{q(\boldsymbol{a}_1)}|\boldsymbol{a}_1| \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{a}_2|\nu)\to\nu} \longleftarrow \begin{bmatrix} n \\ -\mathbf{1}_n^T E_{q(\boldsymbol{a}_2)}\{\log(\boldsymbol{a}_2) + \mathbf{1}_n/\boldsymbol{a}_2\} \end{bmatrix}.$$

Parameter Outputs: $\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\boldsymbol{\theta}}$, $\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\sigma^2}$, $\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\lambda}$, $\boldsymbol{\eta}_{p(\boldsymbol{a}_2|\nu)\to\nu}$.

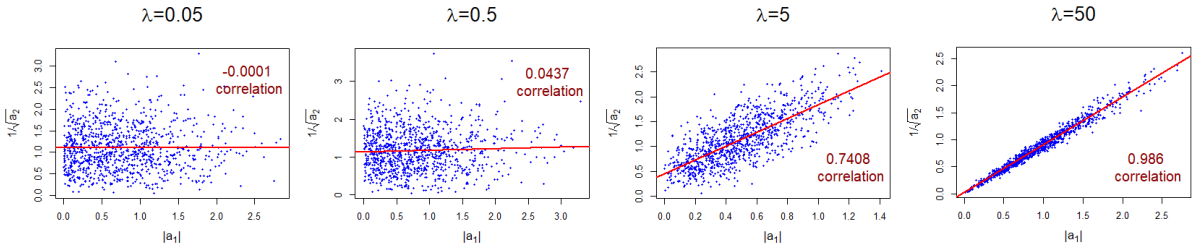

FIGURE 3.7: Markov chain Monte Carlo samples ($n = 1000$) drawn via `rstan` from the distribution $\{|a_1|, 1/\sqrt{a_2} \,|\, \text{rest}\}$ for a skew t random sample with $\boldsymbol{\theta} = \mu = 0$, $\sigma = 1$, $\nu = 1.5$ and $\lambda = (0.05, 0.5, 5, 50)$, using the hyperparameters specified in Section 3.4.1. Sample correlations are also shown.

Theorem 3.1. *Consider random variables satisfying*

$$x \,|\, a_1, a_2 \sim N\left(\mu_0 + \frac{\sigma_0\lambda_0\,|a_1|\,\sqrt{a_2}}{\sqrt{1+\lambda_0^2}}, \frac{a_2\sigma_0^2}{1+\lambda_0^2}\right),$$
$$\textit{where} \quad a_1 \sim N(0,1) \quad \textit{and} \quad a_2 \sim \textit{Inverse-}\chi^2(\nu_0, \nu_0),$$

with $\mu_0, \lambda_0 \in \mathbb{R}$ *and* $\sigma_0, \nu_0 > 0$. *Then for any* $x_0 \in \mathbb{R}$ *and* μ_0, σ_0, ν_0

$$\lim_{|\lambda_0|\to\infty} Corr\left(|a_1|, 1/\sqrt{a_2} \,|\, x = x_0\right) = 1.$$

A proof is given in Section C.2.2 of the Appendix. As the q densities are assumed to approximate the posterior density structure, Theorem 3.1 suggests a modification on our previous assumption to a less simplistic product density restriction. At the cost of further algebra, we propose the replacement of the assumption in (3.22) with

$$\begin{aligned} q\left(\boldsymbol{\theta}, \sigma^2, \lambda, \nu, \boldsymbol{a}_1, \boldsymbol{a}_2\right) &= q(\boldsymbol{\theta})\, q\left(\sigma^2\right) q(\lambda)\, q(\nu)\, q(\boldsymbol{a}_1, \boldsymbol{a}_2) \\ &= q(\boldsymbol{\theta})\, q\left(\sigma^2\right) q(\lambda)\, q(\nu) \prod_{i=1}^{n} q(a_{1i}, a_{2i}). \end{aligned} \tag{3.24}$$

This gives rise to the factor graph representation in the right panel of Figure 3.6 and Algorithm 3.5, whose output at convergence provides the optimal approximating densities

according to (1.17), without alteration of previous message structures. Further details about derivations and notation $(E\boldsymbol{T})_j^{\text{MW}}$, $j = 1, \ldots, 4$, are displayed in section C.2.3 of the Appendix. In particular, under assumption (3.24), the moments with respect to $q^*(a_{1i}, a_{2i})$ are expressible in a closed form. However, further numerical integration may be required when the arguments of the Gaussian hypergeometric functions appearing in moment expressions are close to 1.

3.4.1 Simulation study

We performed a simulation study to compare the performances of the VMP algorithm designed around the assumption in (3.22), Algorithm 3.4, and Algorithm 3.5, which is based on the assumption in (3.24). We generated 100 datasets of size $n = 500$ setting two regression parameters to be $\theta_0 = 1$ and $\theta_1 = 2$, scale parameter $\sigma = 1$ and shape parameters $\lambda = 5$ and $\nu = 1.5$. The hyperparameters for $\boldsymbol{\theta}$ were fixed to $\boldsymbol{\mu_\theta} = \mathbf{0}$ and $\boldsymbol{\Sigma_\theta} = 10^{10}\boldsymbol{I}$ over a prior $N(\boldsymbol{\mu_\theta}, \boldsymbol{\Sigma_\theta})$. We used an Inverse-$\chi^2(A, B)$ prior on the squared scale with $A = B = 0.01$. The prior for the parameter of skewness λ is assumed to be $N(\mu_\lambda, \sigma_\lambda^2)$, with $\mu_\lambda = 0$ and $\sigma_\lambda^2 = 10^{10}$ and that for the degrees of freedom ν to be a Gamma(α_ν, β_ν) with $\alpha_\nu = 1$ and $\beta_\nu = 0.01$.

Let ξ be a generic parameter. The accuracy of each VMP approximation $q^*(\xi)$ as from (1.17) can be assessed using the L_1 error, or *integrated absolute error (IAE)* of q^*, given by

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

As pointed out in Wand *et al.* (2011), the L_1 error is a scale independent number that is invariant to monotone transformation on the parameter ξ. This implies, for instance, that the IAE values for $q^*(\sigma)$ and $q^*(\sigma^2)$ coincide. Note that the L_1 error is a number between 0 and 2. To express this measure as a percentage we can then define the accuracy as

$$\text{accuracy}(q^*) = 1 - \left\{ \text{IAE}(q^*) \Big/ \sup_{q \text{ a density}} \text{IAE}(q) \right\} = 1 - \frac{1}{2}\text{IAE}(q^*),$$

so that $0 \leq \text{accuracy}(q^*) \leq 1$, with 1 reflecting perfect correspondence between VMP approximations and posterior densities. The computation of $p(\xi \mid \boldsymbol{y})$ is complex, so we worked with MCMC samples obtained using `rstan`. MCMC samples of size 10,000 were generated setting a burn-in of 5000 values and thinning the remaining 5000 by a factor of 5. Subsection C.2.4 of Appendix C displays the `rstan` code for fitting skew t regression.

Table 3.1 includes the accuracy values from the simulation study. As expected,

Algorithm 3.5 The VMP inputs, updates and outputs of the skew t likelihood fragment assuming $q(\boldsymbol{\theta}, \sigma^2, \lambda, \nu, \boldsymbol{a}_1, \boldsymbol{a}_2) = q(\boldsymbol{\theta})\, q(\sigma^2)\, q(\lambda)\, q(\nu) \prod_{i=1}^{n} q(a_{1i}, a_{2i})$.

Data Inputs: $\boldsymbol{y}$, $\boldsymbol{A}$.

Parameter Inputs: $\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\boldsymbol{\theta}}$, $\boldsymbol{\eta}_{\boldsymbol{\theta}\to p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)}$, $\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\sigma^2}$, $\boldsymbol{\eta}_{\sigma^2\to p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)}$, $\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\lambda}$, $\boldsymbol{\eta}_{\lambda\to p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)}$, $\boldsymbol{\eta}_{p(\boldsymbol{a}_2|\nu)\to\nu}$, $\boldsymbol{\eta}_{\nu\to p(\boldsymbol{a}_2|\nu)}$.

Updates:

$$\mu_{q(1/\sigma)} \longleftarrow (E\boldsymbol{T})_2^{\text{ISRN}}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\leftrightarrow\sigma^2}\right)$$

$$\mu_{q(1/\sigma^2)} \longleftarrow (E\boldsymbol{T})_3^{\text{ISRN}}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\leftrightarrow\sigma^2}\right)$$

$$\mu_{q(\lambda^2)} \longleftarrow (E\boldsymbol{T})_2^{\text{SS}}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\leftrightarrow\lambda}\right)$$

$$\mu_{q\left(\lambda\sqrt{1+\lambda^2}\right)} \longleftarrow (E\boldsymbol{T})_3^{\text{SS}}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\leftrightarrow\lambda}\right)$$

$$\mu_{q(\nu)} \longleftarrow 2\,(E\boldsymbol{T})_2^{\text{MR}}\left(\boldsymbol{\eta}_{p(\boldsymbol{a}_2|\nu)\leftrightarrow\nu}\right)$$

$$\boldsymbol{\tau}_1 \longleftarrow \boldsymbol{y} + \tfrac{1}{2}\boldsymbol{A}\left\{\text{vec}^{-1}\left(\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\leftrightarrow\boldsymbol{\theta}}\right)_2\right)\right\}^{-1}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\leftrightarrow\boldsymbol{\theta}}\right)_1$$

$$\boldsymbol{\tau}_2 \longleftarrow \left[G_{\text{VMP}}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\leftrightarrow\boldsymbol{\theta}}; \boldsymbol{A}^T\boldsymbol{e}_i\boldsymbol{e}_i^T\boldsymbol{A}, \boldsymbol{A}^T\boldsymbol{e}_i\boldsymbol{e}_i^T\boldsymbol{y}, y_i^2\right)\right]_{1\le i\le n}$$

$$\boldsymbol{\eta}_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)} \longleftarrow \begin{bmatrix} -\frac{1}{2}\left(1+\mu_{q(\lambda^2)}\right) \\ \mu_{q\left(\lambda\sqrt{1+\lambda}\right)}\mu_{q(1/\sigma)}\boldsymbol{\tau}_1 \\ \left(1+\mu_{q(\lambda^2)}\right)\mu_{q(1/\sigma^2)}\boldsymbol{\tau}_2 - \frac{1}{2}\mu_{q(\nu)} \\ -\frac{1}{2}\left(3+\mu_{q(\nu)}\right) \end{bmatrix}$$

$$E_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(\boldsymbol{a}_1^2\right) \longleftarrow (E\boldsymbol{T})_1^{\text{MW}}\left(\boldsymbol{\eta}_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\right)$$

$$E_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(|\boldsymbol{a}_1|/\sqrt{\boldsymbol{a}_2}\right) \longleftarrow (E\boldsymbol{T})_2^{\text{MW}}\left(\boldsymbol{\eta}_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\right)$$

$$E_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(\mathbf{1}_n/\boldsymbol{a}_2\right) \longleftarrow (E\boldsymbol{T})_3^{\text{MW}}\left(\boldsymbol{\eta}_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\right)$$

$$E_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\left\{\log(\boldsymbol{a}_2)\right\} \longleftarrow (E\boldsymbol{T})_4^{\text{MW}}\left(\boldsymbol{\eta}_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\right)$$

$$\boldsymbol{\tau}_3 \longleftarrow G_{\text{VMP}}\Big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\leftrightarrow\boldsymbol{\theta}}; \boldsymbol{A}^T\text{diag}\left\{E_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(\mathbf{1}_n/\boldsymbol{a}_2\right)\right\}\boldsymbol{A},$$
$$\boldsymbol{A}^T\text{diag}\left\{E_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(\mathbf{1}_n/\boldsymbol{a}_2\right)\right\}\boldsymbol{y}, \boldsymbol{y}^T\text{diag}\left\{E_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(\mathbf{1}_n/\boldsymbol{a}_2\right)\right\}\boldsymbol{y}\Big)$$

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\boldsymbol{\theta}} \longleftarrow \left(1+\mu_{q(\lambda^2)}\right)\mu_{q(1/\sigma^2)}\begin{bmatrix} \boldsymbol{A}^T\text{diag}\left\{E_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(\mathbf{1}_n/\boldsymbol{a}_2\right)\right\}\boldsymbol{y} \\ -\frac{1}{2}\text{vec}\left(\boldsymbol{A}^T\text{diag}\left\{E_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(\mathbf{1}_n/\boldsymbol{a}_2\right)\right\}\boldsymbol{A}\right) \end{bmatrix}$$
$$-\mu_{q\left(\lambda\sqrt{1+\lambda^2}\right)}\mu_{q(1/\sigma)}\begin{bmatrix} \boldsymbol{A}^T E_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(|\boldsymbol{a}_1|/\sqrt{\boldsymbol{a}_2}\right) \\ \mathbf{0} \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\sigma^2} \longleftarrow \begin{bmatrix} -n/2 \\ \mu_{q\left(\lambda\sqrt{1+\lambda^2}\right)}\boldsymbol{\tau}_1^T E_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(|\boldsymbol{a}_1|/\sqrt{\boldsymbol{a}_2}\right) \\ \left(1+\mu_{q(\lambda^2)}\right)\boldsymbol{\tau}_3 \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\lambda} \longleftarrow \begin{bmatrix} n/2 \\ \mu_{q(1/\sigma^2)}\boldsymbol{\tau}_3 - \frac{1}{2}\mathbf{1}_n^T E_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(\boldsymbol{a}_1^2\right) \\ \mu_{q(1/\sigma)}\boldsymbol{\tau}_1^T E_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(|\boldsymbol{a}_1|/\sqrt{\boldsymbol{a}_2}\right) \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{a}_2|\nu)\to\nu} \longleftarrow \begin{bmatrix} n \\ -\mathbf{1}_n^T E_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\left\{\log(\boldsymbol{a}_2) + \mathbf{1}_n/\boldsymbol{a}_2\right\} \end{bmatrix}.$$

Parameter Outputs: $\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\boldsymbol{\theta}}$, $\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\sigma^2}$, $\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\lambda}$, $\boldsymbol{\eta}_{p(\boldsymbol{a}_2|\nu)\to\nu}$.

TABLE 3.1: Average (standard deviation) accuracy from the simulation study. "VMP 1" and "VMP 2" refer to Algorithms 3.4 and 3.5 respectively.

	Accuracy			
Parameter	VMP 1		VMP 2	
β_0	0.0	(0.0)	37.7	(6.2)
β_1	51.2	(17.7)	57.1	(4.5)
σ^2	0.0	(0.0)	11.0	(3.5)
λ	0.0	(0.0)	10.0	(3.4)
ν	0.0	(0.0)	56.6	(6.1)

the algorithm based on assumption (3.24) is seen to provide more accurate inference. However, accuracy and percentage of coverage of σ^2 and λ are particularly low for both the algorithms. This might suggest the application of less generic product density restriction to take into account other possible posterior dependence among variables. Nonetheless, this choice would imply more involved message update derivations and further numerical integration. Figure 3.8 permits visualization of these results with the plot of approximate and MCMC posterior densities from a single simulation. The density curves produced by Algorithm 3.5 are sensibly closer to the modes of MCMC posterior densities than those from the other VMP algorithm.

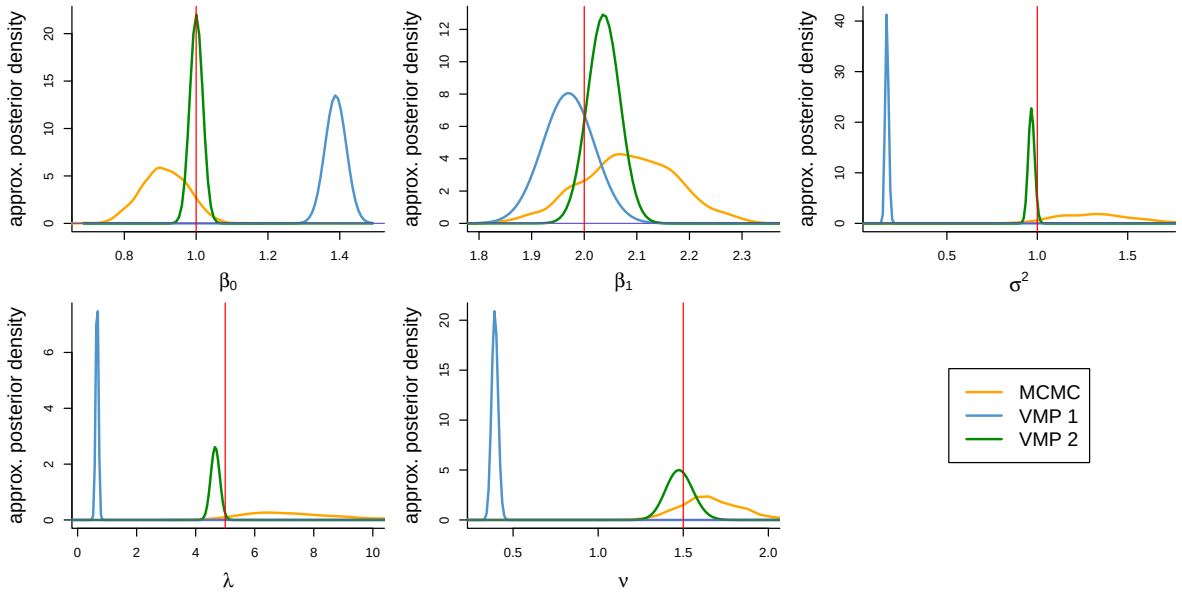

FIGURE 3.8: VMP-approximate and MCMC posterior density functions from a single dataset of the simulation study. "VMP 1" and "VMP 2" respectively refer to Algorithms 3.4 and 3.5. VMP, variational message passing; MCMC, Markov chain Monte Carlo. Vertical lines indicate the true values.

3.4.2 Applications

Variational message passing is a flexible instrument for inference and prediction in a number of applications. With the intent to illustrate VMP performances and advantages, we first provide the study of the Martin Marietta dataset via the simple regression model presented in Azzalini and Capitanio (2003). On the other hand, the second application is intended to show how the VMP methodology easily adapts to models that extend beyond the original likelihood fragment, such as the skew t nonparametric regression model we propose for the `Workinghours` dataset.

3.4.2.1 Martin Marietta data

We illustrate the parameter estimation of a skew t regression model via VMP.

Consider the Martin Marietta dataset examined in Azzalini and Capitanio (2003) with the linear model

$$y_i = \beta_0 + \beta_1 \text{CRSP}_i + \varepsilon_i, \quad \varepsilon_i \sim \text{Skew-t}\left(0, \sigma^2, \lambda, \nu\right), \quad 1 \leq i \leq 60.$$

The variables y_i and CRSP_i denote the Martin Marietta company excess rate and the return excess index for the whole New York Stock Exchange respectively. Data over a period of $n = 60$ consecutive months from January 1982 to December 1986 are available.

According to the properties of the skew t distribution described in Section 4.2.3 of Azzalini and Capitanio (2003), the response y_i will be skew t distributed as the error term, but with mean $\beta_0 + \beta_1 \text{CRSP}_i$. As before, we write the skew t distribution in terms

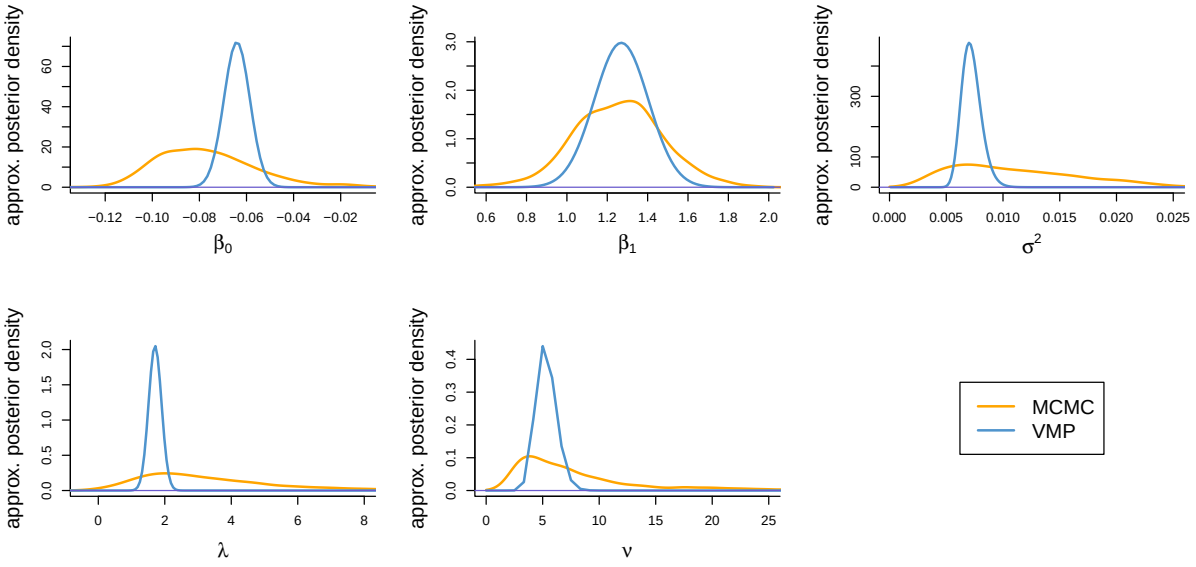

FIGURE 3.9: Martin Marietta data: posterior density plots via MCMC and VMP.

of standard normal and inverse χ^2 auxiliary variables. We adopt the q-density product restriction (3.24) to approximate the parameter posterior densities with VMP. As a check, we compare them to MCMC density estimation via `rstan`. The hyperparameters for $\boldsymbol{\beta}$ are fixed to $\boldsymbol{\mu}_{\boldsymbol{\beta}} = \mathbf{0}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = 10^5 \boldsymbol{I}$ over a prior $N(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$ while those on the shape parameters are Inverse-$\chi^2(0.01, 0.01)$ on the squared scale, $N(0, 10^5)$ on λ and $\Gamma(1, 0.01)$ on ν. Posterior density plots are shown in Fig. 3.9. VMP curves apparently underestimate the variance of MCMC posterior densities but locate around their modes. Differently from Azzalini and Capitanio (2003), we did not set $\sigma^2 = 1$, therefore their estimates are not directly comparable with the results from MCMC and VMP.

3.4.2.2 `Workinghours` dataset

Here we provide an application that illustrates how the derivations in Section 3.4 can be integrated to perform variational inference on extensions of the skew t likelihood fragment without deriving a VMP scheme from scratch. We consider the dataset `Workinghours` from the `R` package `Ecdat` (Croissant, 2016) which contains a cross-section study of 3,382 observations. The response variable is *income* divided by a factor of 10 (the other household income in thousands of dollars) versus the variable *age* (age of the wife). The pairs of predictors and responses (x_i, y_i), $1 \leq i \leq n$, are analyzed via nonparametric regression and the following penalized spline model in Bayesian mixed model form:

$$y_i \,|\, f, \sigma_\varepsilon^2, \lambda, \nu \overset{\text{ind.}}{\sim} \text{Skew-}t\left(f(x_i), \sigma_\varepsilon^2, \lambda, \nu\right),$$

with function f structured as $f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^K u_k z_k(x)$, with $u_k \,|\, \sigma_u^2 \overset{\text{ind.}}{\sim} N(0, \sigma_u^2)$ and $\{z_k : 1 \leq k \leq K\}$ suitable spline basis. The full model with auxiliary variable representation is

$$\begin{aligned}
&y_i \,|\, \boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2, \lambda, a_{1i}, a_{2i} \overset{\text{ind.}}{\sim} N\left((\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u})_i + \frac{\sigma_\varepsilon \lambda \,|a_{1i}| \sqrt{a_{2i}}}{\sqrt{1+\lambda^2}}, \frac{a_{2i}\sigma_\varepsilon^2}{1+\lambda^2}\right) \\
&a_{1i} \overset{\text{ind.}}{\sim} N(0,1), \quad a_{2i} \,|\, \nu \overset{\text{ind.}}{\sim} \text{Inverse-}\chi^2(\nu, \nu), \\
&\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} \,\Big|\, \sigma_u^2 \sim N\left(\begin{bmatrix} \boldsymbol{\mu}_{\boldsymbol{\beta}} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}} & \mathbf{0} \\ \mathbf{0} & \sigma_u^2 \boldsymbol{I} \end{bmatrix}\right), \\
&\sigma_u^2 \sim \text{Inverse-}\chi^2\left(A_{\sigma_u^2}, B_{\sigma_u^2}\right), \quad \sigma_\varepsilon^2 \sim \text{Inverse-}\chi^2\left(A_{\sigma_\varepsilon^2}, B_{\sigma_\varepsilon^2}\right), \\
&\lambda \sim N\left(\mu_\lambda, \sigma_\lambda^2\right), \quad \nu \sim \text{Gamma}(\alpha_\nu, \beta_\nu),
\end{aligned} \tag{3.25}$$

where

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \text{and} \quad \boldsymbol{Z} = \begin{bmatrix} z_1(x_1) & \cdots & z_K(x_1) \\ \vdots & \ddots & \vdots \\ z_1(x_n) & \cdots & z_K(x_n) \end{bmatrix}.$$

The 2×1 vector $\boldsymbol{\mu}_{\boldsymbol{\beta}}$, 2×2 symmetric positive definite matrix $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$, positive numbers $A_{\sigma_u^2}$, $B_{\sigma_u^2}$, $A_{\sigma_\varepsilon^2}$, $B_{\sigma_\varepsilon^2}$, σ_λ^2, α_ν and β_ν and number μ_λ are user-specified hyperparameters that we choose to be ideally uninformative as for the study of Martin Marietta data. We adopt canonical cubic O'Sullivan splines described in Section 1.6.1 with $K = 14$, according to the rule of thumb (1.21).

Assuming that the joint posterior density approximation admits the product density approximation

$$p\left(\boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2, \sigma_u^2, \lambda, \nu \,|\, \boldsymbol{y}\right) \approx q\left(\boldsymbol{\beta}, \boldsymbol{u}\right) q\left(\sigma_\varepsilon^2\right) q\left(\sigma_u^2\right) q\left(\lambda\right) q\left(\nu\right) \prod_{i=1}^{n} q\left(a_{1i}, a_{2i}\right), \tag{3.26}$$

Algorithm 3.5 can be integrated with updates involving the blue nodes in the factor graph in the left panel of Figure 3.10 to fit the regression model (3.25) via VMP. The estimated nonparametric regression function and corresponding pointwise 95% credible set are shown in the right panel of Figure 3.10. The results show higher mean of the other household income for average-age wives which tends to decrease more remarkably around the age of 43 and 57.

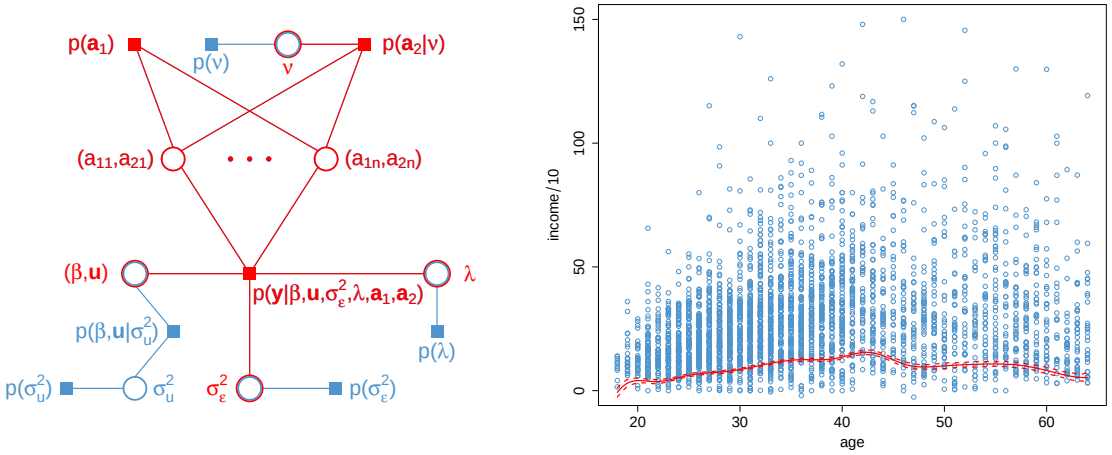

FIGURE 3.10: Study of data from `Workinghours` dataset. Left panel: factor graph corresponding to the model in (3.25) under the product density restriction in (3.26). Right panel: approximate posterior mean (solid line) and pointwise 95% credible sets (dashed line) obtained via VMP, integrating Algorithm 3.5; 20 observations whose "income/10" value exceeds 150 have been excluded from the plot.

3.5 Concluding remarks

Variational message passing offers a flexible framework to standardize variational Bayes algorithm derivations and numerical integration steps. Motivated by the desire to have fast approximate inference methods for additional notable likelihood models, we have developed VMP algorithms for fitting and inference for Pareto random samples, support vector regression and skew t regression likelihood fragments. As indicated by the simulation study in Section 3.4.1, the performance of variational Bayes is not always satisfactory, especially in presence of a high number of parameters and strong correlation among model parameters. The VMP algorithms we propose are designed around a choice of the mean field restriction which is a compromise among algebraic complexity, feasibility and quality of the approximation. As revealed by the application on a real dataset, VMP allows one to integrate several fragments and compose algorithms for more complex models without significant additional effort.

Chapter 4

Streamlined variational message passing

4.1 Introduction

In this chapter we present streamlined MFVB and VMP algorithms for models containing higher level random effects. The estimation of multilevel models via standard approaches may be too slow, or even computationally infeasible, when the number of groups or the dimension of possible spline basis functions become large.

Consider an educational study predicting for each classroom and school the grades of students on a standardized test given their scores on a pre-test and other information. Imagine we would like to study intra and inter-level dependence. Here the students, classrooms and schools are the three levels around which a multilevel model can be designed. Following, for instance, Gelman and Hill (2014) we name these *three-level* data, while other references such as Pinheiro and Bates (2000) use the term two-level for analogous scenarios, considering the two levels of nesting. Nolan *et al.* (2018) point out that if the dataset is such to include, for instance, 500 groups, each containing 60 second level groups with 1000 observations, then the combined fixed and random effects design matrices have $1.83 \cdot 10^{12}$ entries of which at least the 99.99% are zeroes. These issues motivate the development of fast and scalable variational methods where updates can be potentially streamlined in terms of number of operations and storage. Variational inference is able to perform model fitting employing only the about 0.01% non-zero design matrix components, with algorithm updates that are linear in the numbers of groups.

The present chapter is an attempt to extend the class of streamlined variational inference algorithms for higher level random effects models with preliminary results, relying

on recent work in Nolan *et al.* (2018) concerning linear system solutions and sub-blocks of matrix inverses. Our aim is to provide some directions for future developments. In particular, *two-level* models for non-Gaussian responses are considered. In the variational Bayesian framework that we adopt, the linear unbiased prediction is replaced by variational approximate posterior means and confidence intervals are replaced by variational approximate credible intervals. We denote with p the number of columns of the fixed effect design matrix $\boldsymbol{X}$, with q that of the random effects design matrix $\boldsymbol{Z}$, with m the number of groups in a two-level model and n_i, $1 \leq i \leq m$, the number of observations per group.

Section 4.2 presents two algorithms for solving two-level sparse matrix problems which support variational inference for non-Gaussian response models with hierarchical random effects structure. We examine the use of variational approximations, specifically MFVB and VMP methodologies, for fitting and inference in Bayesian GLMMs, with an emphasis on the Bernoulli and Poisson response cases. The resulting MFVB schemes are Algorithms 4.3 and 4.4, while the corresponding VMP versions are listed as Algorithms 4.6 and 4.7. A very simple illustration on a simuluated dataset is included in Section 4.5. As for Chapter 3, both the MFVB and VMP approaches produce the same results at convergence, but we highlight once again that VMP algorithms easily adapt to arbitrarily large models through the notion of factor graph fragment.

4.2 Two-level sparse matrix problem algorithms

This section describes two algorithms, named SolveTwoLevelSparseMatrix and SolveTwoLevelSparseLeastSquares by Nolan *et al.* (2018), which serve as the base for streamlined variational inference for two-level models. Two-level sparse matrix problems are treated in Nolan *et al.* (2018). Here, their results are recalled preserving the same notation.

The SolveTwoLevelSparseMatrix algorithm is conceived to solve the general two-level sparse linear system problem

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{a}, \tag{4.1}$$

where

$$\boldsymbol{A} = \left[\begin{array}{c|c|c|c|c} \boldsymbol{A}_{11} & \boldsymbol{A}_{12,1} & \boldsymbol{A}_{12,2} & \cdots & \boldsymbol{A}_{12,m} \\ \hline \boldsymbol{A}_{12,1}^T & \boldsymbol{A}_{22,1} & \boldsymbol{O} & \cdots & \boldsymbol{O} \\ \hline \boldsymbol{A}_{12,2}^T & \boldsymbol{O} & \boldsymbol{A}_{22,2} & \cdots & \boldsymbol{O} \\ \hline \vdots & \vdots & \vdots & \ddots & \vdots \\ \hline \boldsymbol{A}_{12,m}^T & \boldsymbol{O} & \boldsymbol{O} & \cdots & \boldsymbol{A}_{22,m} \end{array}\right], \quad \boldsymbol{a} = \left[\begin{array}{c} \boldsymbol{a}_1 \\ \hline \boldsymbol{a}_{2,1} \\ \hline \boldsymbol{a}_{2,2} \\ \hline \vdots \\ \hline \boldsymbol{a}_{2,m} \end{array}\right] \quad \text{and} \quad \boldsymbol{x} = \left[\begin{array}{c} \boldsymbol{x}_1 \\ \hline \boldsymbol{x}_{2,1} \\ \hline \boldsymbol{x}_{2,2} \\ \hline \vdots \\ \hline \boldsymbol{x}_{2,m} \end{array}\right] \tag{4.2}$$

and obtain the sub-matrices of $\boldsymbol{A}^{-1}$ corresponding to the non-zero blocks of $\boldsymbol{A}$:

$$\boldsymbol{A}^{-1} = \left[\begin{array}{c|c|c|c|c} \boldsymbol{A}^{11} & \boldsymbol{A}^{12,1} & \boldsymbol{A}^{12,2} & \cdots & \boldsymbol{A}^{12,m} \\ \hline \boldsymbol{A}^{12,1\,T} & \boldsymbol{A}^{22,1} & \times & \cdots & \times \\ \hline \boldsymbol{A}^{12,2\,T} & \times & \boldsymbol{A}^{22,2} & \cdots & \times \\ \hline \vdots & \vdots & \vdots & \ddots & \vdots \\ \hline \boldsymbol{A}^{12,m\,T} & \times & \times & \cdots & \boldsymbol{A}^{22,m} \end{array}\right]. \tag{4.3}$$

The sub-blocks represented by the $\times$ symbol are not of interest when deriving streamlined variational algorithms since they correspond to between group covariances. On the opposite side, the remaining sub-blocks, which are in the same position as the non-zero blocks of $\boldsymbol{A}$, are sufficient for both coordinate ascent and message passing parameter optimization with minimal product density restrictions to obtain within-group standard errors. The sub-matrices of $\boldsymbol{A}$ have dimensions:

$$\boldsymbol{A}_{11} \text{ is } p \times p \quad \text{and,} \quad \text{for each } 1 \leq i \leq m, \quad \boldsymbol{A}_{12,i} \text{ is } p \times q \quad \text{and} \quad \boldsymbol{A}_{22,i} \text{ is } q \times q.$$

The dimensions of the sub-vectors of $\boldsymbol{a}$ and $\boldsymbol{x}$ are:

$$\text{both } \boldsymbol{a}_1 \text{ and } \boldsymbol{x}_1 \text{ are } p \times 1 \quad \text{and,} \quad \text{for each } 1 \leq i \leq m, \quad \text{both } \boldsymbol{a}_{2,i} \text{ and } \boldsymbol{x}_{2,i} \text{ are } q \times 1.$$

Algorithm 4.1 lists the SolveTwoLevelSparseMatrix algorithm.

Algorithm 4.1 (Nolan *et al.*, 2018) *The* SolveTwoLevelSparseMatrix *algorithm for solving the two-level sparse matrix problem* $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{a}$ *and sub-blocks of* $\boldsymbol{A}^{-1}$ *corresponding to the non-zero sub-blocks of* $\boldsymbol{A}$*. The sub-block notation is given by* (4.2) *and* (4.3)*.*

Inputs: $\Big(\boldsymbol{a}_1(p\times 1), \boldsymbol{A}_{11}(p\times p), \big\{\big(\boldsymbol{a}_{2,i}(q\times 1), \boldsymbol{A}_{22,i}(q\times q), \boldsymbol{A}_{12,i}(p\times q)\big) : 1 \le i \le m\big\}\Big)$.

$\boldsymbol{\omega}_1 \longleftarrow \boldsymbol{a}_1 \quad ; \quad \boldsymbol{\Omega}_2 \longleftarrow \boldsymbol{A}_{11}$

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

$$\boldsymbol{\omega}_1 \longleftarrow \boldsymbol{\omega}_1 - \boldsymbol{A}_{12,i}\boldsymbol{A}_{22,i}^{-1}\boldsymbol{a}_{2,i} \quad ; \quad \boldsymbol{\Omega}_2 \longleftarrow \boldsymbol{\Omega}_2 - \boldsymbol{A}_{12,i}\boldsymbol{A}_{22,i}^{-1}\boldsymbol{A}_{12,i}^T$$

$\boldsymbol{A}^{11} \longleftarrow \boldsymbol{\Omega}_2^{-1} \quad ; \quad \boldsymbol{x}_1 \longleftarrow \boldsymbol{A}^{11}\boldsymbol{\omega}_1$

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

$$\boldsymbol{x}_{2,i} \longleftarrow \boldsymbol{A}_{22,i}^{-1}(\boldsymbol{a}_{2,i} - \boldsymbol{A}_{12,i}^T\boldsymbol{x}_1) \quad ; \quad \boldsymbol{A}^{12,i} \longleftarrow -(\boldsymbol{A}_{22,i}^{-1}\boldsymbol{A}_{12,i}^T\boldsymbol{A}^{11})^T$$

$$\boldsymbol{A}^{22,i} \longleftarrow \boldsymbol{A}_{22,i}^{-1}\big(\boldsymbol{I} - \boldsymbol{A}_{12,i}^T\boldsymbol{A}^{12,i}\big).$$

Outputs: $\Big(\boldsymbol{x}_1, \boldsymbol{A}^{11}, \big\{\big(\boldsymbol{x}_{2,i}, \boldsymbol{A}^{22,i}, \boldsymbol{A}^{12,i}\big) : 1 \le i \le m\big\}\Big)$.

The SolveTwoLevelSparseLeastSquares algorithm arises in the special case where $\boldsymbol{x}$ is the minimizer of the least squares problem

$$\|\boldsymbol{b} - \boldsymbol{B}\boldsymbol{x}\|^2 = (\boldsymbol{b} - \boldsymbol{B}\boldsymbol{x})^T(\boldsymbol{b} - \boldsymbol{B}\boldsymbol{x}), \tag{4.4}$$

where

$$\boldsymbol{B} = \left[\begin{array}{c|c|c|c|c} \boldsymbol{B}_1 & \dot{\boldsymbol{B}}_1 & \boldsymbol{O} & \cdots & \boldsymbol{O} \\ \hline \boldsymbol{B}_2 & \boldsymbol{O} & \dot{\boldsymbol{B}}_2 & \cdots & \boldsymbol{O} \\ \hline \vdots & \vdots & \vdots & \ddots & \vdots \\ \hline \boldsymbol{B}_m & \boldsymbol{O} & \boldsymbol{O} & \cdots & \dot{\boldsymbol{B}}_m \end{array}\right] \quad \text{and} \quad \boldsymbol{b} = \left[\begin{array}{c} \boldsymbol{b}_1 \\ \hline \boldsymbol{b}_2 \\ \hline \vdots \\ \hline \boldsymbol{b}_m \end{array}\right], \tag{4.5}$$

with sub-blocks and sub-vectors such that

$$\boldsymbol{B}_i \text{ is } n_i \times p, \quad \dot{\boldsymbol{B}}_i \text{ is } n_i \times q \quad \text{and} \quad \boldsymbol{b}_i \text{ is } n_i \times 1 \quad \text{for} \quad 1 \le i \le m.$$

Then the $\boldsymbol{x}$ that minimizes (4.4) is the solution to the two-level sparse linear system (4.1) with

$$\boldsymbol{A} = \boldsymbol{B}^T\boldsymbol{B} \quad \text{and} \quad \boldsymbol{a} = \boldsymbol{B}^T\boldsymbol{b},$$

so that the sub-blocks of $\boldsymbol{A}$ and the sub-vectors of $\boldsymbol{a}$ take the forms

$$\boldsymbol{A}_{11} = \sum_{i=1}^{m} \boldsymbol{B}_i^T \boldsymbol{B}_i, \quad \boldsymbol{A}_{12,i} = \boldsymbol{B}_i^T \dot{\boldsymbol{B}}_i, \quad \boldsymbol{A}_{22,i} = \dot{\boldsymbol{B}}_i^T \dot{\boldsymbol{B}}_i, \quad \boldsymbol{a}_1 = \sum_{i=1}^{m} \boldsymbol{B}_i^T \boldsymbol{b}_i \text{ and } \boldsymbol{a}_{2,i} = \dot{\boldsymbol{B}}_i^T \boldsymbol{b}_i.$$

These forms arise in two-level random effects models and Nolan *et al.* (2018) show that they admit a computationally fast and stable QR-decomposition based solution (e.g. Gentle, 2007, Section 6.7.2). A QR-decomposition of a rectangular $n \times p\,(n \geq p)$ matrix $\boldsymbol{X}$ is based on the representation

$$\boldsymbol{X} = \boldsymbol{Q} \begin{bmatrix} \boldsymbol{R} \\ \boldsymbol{O} \end{bmatrix},$$

where $\boldsymbol{Q}$ is a $n \times n$ orthogonal matrix and $\boldsymbol{R}$ is a $p \times p$ upper-triangular matrix.

The SolveTwoLevelSparseLeastSquares algorithm is listed as Algorithm 4.2. Some of its steps require to implement the QR-decomposition, which is a standard procedure within most computing environments.

An important result at the base of the streamlined expressions appearing in the next sections is that for any symmetric $d \times d$ matrix $\boldsymbol{M}$

$$\text{vech}\,(\boldsymbol{M}) = \boldsymbol{D}_d^{+} \text{vec}\,(\boldsymbol{M}), \tag{4.6}$$

where $\boldsymbol{D}_d^{+}$ is the Moore–Penrose inverse matrix of dimension d of the duplication matrix $\boldsymbol{D}_d$ (e.g. Magnus and Neudecker, 2007, p. 57). The reduced expressions (A.10) for natural parameter vector and inverse mapping of the multivariate normal distribution in Subsection A.2.12 of Appendix A are also relevant for streamlined variational inference.

4.3 MFVB for two-level random effects models

This section presents streamlined MFVB algorithms for two-level linear mixed models with Poisson and binomial responses, taking advantage of Algorithm 4.1.

Algorithm 4.2 (Nolan *et al.*, 2018) SolveTwoLevelSparseLeastSquares *for solving the two-level sparse matrix least squares problem: minimise* $\|\boldsymbol{b} - \boldsymbol{B}\,\boldsymbol{x}\|^2$ *in* $\boldsymbol{x}$ *and sub-blocks of* $\boldsymbol{A}^{-1}$ *corresponding to the non-zero sub-blocks of* $\boldsymbol{A} = \boldsymbol{B}^T\boldsymbol{B}$*. The sub-block notation is given by* (4.2)*,* (4.3) *and* (4.5)*.*

Inputs: $\big\{\big(\boldsymbol{b}_i(n_i \times 1),\ \boldsymbol{B}_i(n_i \times p),\ \dot{\boldsymbol{B}}_i(n_i \times q)\big) :\ 1 \le i \le m\big\}$.

$\boldsymbol{\omega}_3 \longleftarrow$ NULL ; $\boldsymbol{\Omega}_4 \longleftarrow$ NULL

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

Decompose $\dot{\boldsymbol{B}}_i = \boldsymbol{Q}_i \begin{bmatrix} \boldsymbol{R}_i \\ \boldsymbol{0} \end{bmatrix}$ such that $\boldsymbol{Q}_i^{-1} = \boldsymbol{Q}_i^T$ and $\boldsymbol{R}_i$ is upper-triangular.

$\boldsymbol{c}_{0i} \longleftarrow \boldsymbol{Q}_i^T\boldsymbol{b}_i$; $\boldsymbol{C}_{0i} \longleftarrow \boldsymbol{Q}_i^T\boldsymbol{B}_i$

$\boldsymbol{c}_{1i} \longleftarrow$ first q rows of $\boldsymbol{c}_{0i}$; $\boldsymbol{c}_{2i} \longleftarrow$ remaining rows of $\boldsymbol{c}_{0i}$; $\boldsymbol{\omega}_3 \longleftarrow \begin{bmatrix} \boldsymbol{\omega}_3 \\ \boldsymbol{c}_{2i} \end{bmatrix}$

$\boldsymbol{C}_{1i} \longleftarrow$ first q rows of $\boldsymbol{C}_{0i}$; $\boldsymbol{C}_{2i} \longleftarrow$ remaining rows of $\boldsymbol{C}_{0i}$

$\boldsymbol{\Omega}_4 \longleftarrow \begin{bmatrix} \boldsymbol{\Omega}_4 \\ \boldsymbol{C}_{2i} \end{bmatrix}$

Decompose $\boldsymbol{\Omega}_4 = \boldsymbol{Q} \begin{bmatrix} \boldsymbol{R} \\ \boldsymbol{0} \end{bmatrix}$ such that $\boldsymbol{Q}^{-1} = \boldsymbol{Q}^T$ and $\boldsymbol{R}$ is upper-triangular

$\boldsymbol{c} \longleftarrow$ first p rows of $\boldsymbol{Q}^T\boldsymbol{\omega}_3$; $\boldsymbol{x}_1 \longleftarrow \boldsymbol{R}^{-1}\boldsymbol{c}$; $\boldsymbol{A}^{11} \longleftarrow \boldsymbol{R}^{-1}\boldsymbol{R}^{-T}$

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

$\boldsymbol{x}_{2,i} \longleftarrow \boldsymbol{R}_i^{-1}(\boldsymbol{c}_{1i} - \boldsymbol{C}_{1i}\boldsymbol{x}_1)$; $\boldsymbol{A}^{12,i} \longleftarrow -\boldsymbol{A}^{11}(\boldsymbol{R}_i^{-1}\boldsymbol{C}_{1i})^T$

$\boldsymbol{A}^{22,i} \longleftarrow \boldsymbol{R}_i^{-1}(\boldsymbol{R}_i^{-T} - \boldsymbol{C}_{1i}\boldsymbol{A}^{12,i})$.

Output: $\Big(\boldsymbol{x}_1, \boldsymbol{A}^{11}, \big\{\big(\boldsymbol{x}_{2,i}, \boldsymbol{A}^{22,i}, \boldsymbol{A}^{12,i}\big) :\ 1 \le i \le m\big\}\Big)$.

4.3.1 Streamlined MFVB for Poisson response models

Consider the two-level Poisson mixed model, for $1 \le i \le m$:

$$\begin{aligned}
&\boldsymbol{y}_i \,|\, \boldsymbol{\beta}, \boldsymbol{u}_i \overset{\text{ind.}}{\sim} \text{Poisson}\left(\exp\left(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{u}_i\right)\right), \quad \boldsymbol{u}_i \,|\, \boldsymbol{\Sigma} \sim N\left(\boldsymbol{0}, \boldsymbol{\Sigma}\right), \\
&\boldsymbol{\beta} \sim N\left(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right), \quad \boldsymbol{\Sigma} | \boldsymbol{A}_{\boldsymbol{\Sigma}} \sim \text{Inverse-G-Wishart}\big(G_{\text{full}}, \nu_{\boldsymbol{\Sigma}} + 2q - 2, \boldsymbol{A}_{\boldsymbol{\Sigma}}^{-1}\big), \\
&\boldsymbol{A}_{\boldsymbol{\Sigma}} \sim \text{Inverse-G-Wishart}(G_{\text{diag}}, 1, \boldsymbol{\Lambda}_{\boldsymbol{A}_{\boldsymbol{\Sigma}}}), \quad \boldsymbol{\Lambda}_{\boldsymbol{A}_{\boldsymbol{\Sigma}}} = \{\nu_{\boldsymbol{\Sigma}}\text{diag}(s^2_{\boldsymbol{\Sigma},1}, \ldots, s^2_{\boldsymbol{\Sigma},q})\}^{-1},
\end{aligned} \tag{4.7}$$

The dimensions of vectors and matrices are, for $1 \le i \le m$:

$$\begin{aligned}
&\boldsymbol{y}_i \text{ is } n_i \times 1, \quad \boldsymbol{X}_i \text{ is } n_i \times p, \quad \boldsymbol{\beta} \text{ is } p \times 1, \\
&\boldsymbol{Z}_i \text{ is } n_i \times q, \quad \boldsymbol{u}_i \text{ is } q \times 1 \quad \text{and} \quad \boldsymbol{\Sigma} \text{ is } q \times q.
\end{aligned} \tag{4.8}$$

The hyperparameters are the $\boldsymbol{\mu}_{\boldsymbol{\beta}}$ vector of length p, the symmetric and positive definite matrix $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ of size $p \times p$ and $\nu_{\boldsymbol{\Sigma}}, s_{\boldsymbol{\Sigma},1}, \ldots, s_{\boldsymbol{\Sigma},q} > 0$. In the previous model, the prior on $\boldsymbol{\Sigma}$ is within the class described in Huang and Wand (2013), which generalizes the

univariate case treated in Subsection 3.1.2. As for the half Cauchy prior specification for the univariate case, such priors allow standard deviation and correlation parameters to be arbitrary non-informative.

Next, define the matrices

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_m \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_m \end{bmatrix}, \quad \boldsymbol{Z} = \underset{1 \leq i \leq m}{\text{blockdiag}}(\boldsymbol{Z}_i), \quad \boldsymbol{u} = \begin{bmatrix} \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_m \end{bmatrix}. \tag{4.9}$$

Derivation of MFVB and VMP algorithms requires the usual mean field restriction on the joint posterior density function of all parameters in (4.7). In this case we assume

$$p\left(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{A}_{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma} | \boldsymbol{y}\right) \approx q\left(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{A}_{\boldsymbol{\Sigma}}\right) q\left(\boldsymbol{\Sigma}\right). \tag{4.10}$$

The corresponding DAG representation is given in Figure 4.1.

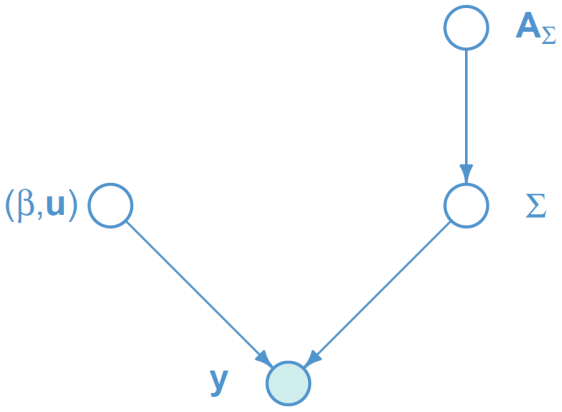

FIGURE 4.1: Directed acyclic graph for the two-level Poisson and logistic response mixed model in (4.7) and (4.17).

Evaluation of the optimal q-density as a function of $\boldsymbol{\beta}$ and $\boldsymbol{u}$ according to the canonical MFVB formula (1.8) involves multivariate integrals that are not available in closed form. The non-conjugate variational message passing solution proposed in Knowles and Minka (2011) is one that instead works also for the MFVB case and leads to the optimal $q\left(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{A}_{\boldsymbol{\Sigma}}\right)$ being

$$q^*\left(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{A}_{\boldsymbol{\Sigma}}\right) = q^*\left(\boldsymbol{\beta}, \boldsymbol{u}\right) q^*\left(\boldsymbol{A}_{\boldsymbol{\Sigma}}\right)$$

and to the following optimal q-density functions for the parameters of interest contained in $\boldsymbol{\beta}$, $\boldsymbol{u}$ and $\boldsymbol{\Sigma}$:

$$\begin{aligned} &q^*(\boldsymbol{\beta},\boldsymbol{u}) \text{ has a } N\big(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\big) \text{ density function;} \\ &q^*(\boldsymbol{\Sigma}) \text{ has an Inverse-G-Wishart}(G_{\text{full}}, \xi_{q(\boldsymbol{\Sigma})}, \boldsymbol{\Lambda}_{q(\boldsymbol{\Sigma})}) \text{ density function.} \end{aligned} \tag{4.11}$$

Section D.1.1 of Appendix D provides details about the derivation of the optimal q-density parameters via an iterative coordinate ascent algorithm. In particular, the updates for $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ can be written as

$$\begin{aligned} \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} &\leftarrow \left(\boldsymbol{C}^T\boldsymbol{R}^{-1}_{\text{2PMFVB}}\boldsymbol{C} + \boldsymbol{D}_{\text{2PMFVB}}\right)^{-1}\left(\boldsymbol{C}^T\boldsymbol{z}_{\text{2PMFVB}} + \boldsymbol{o}_{\text{2PMFVB}}\right) \\ \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} &\leftarrow \left(\boldsymbol{C}^T\boldsymbol{R}^{-1}_{\text{2PMFVB}}\boldsymbol{C} + \boldsymbol{D}_{\text{2PMFVB}}\right)^{-1}, \end{aligned} \tag{4.12}$$

where $\boldsymbol{C} = [\boldsymbol{X}\ \boldsymbol{Z}]$,

$$\begin{aligned} &\boldsymbol{R}^{-1}_{\text{2PMFVB}} = \text{diag}\left(\boldsymbol{\omega}_{\text{2PMFVB}}\right), \quad \boldsymbol{D}_{\text{2PMFVB}} = \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I}_m \otimes \boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1})} \end{bmatrix}, \\ &\boldsymbol{o}_{\text{2PMFVB}} = \left(\boldsymbol{\Sigma}^{-1}_{q(\boldsymbol{\beta},\boldsymbol{u})} - \boldsymbol{D}_{\text{2PMFVB}}\right)^{-1}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}} \\ \boldsymbol{0} \end{bmatrix}, \\ &\boldsymbol{z}_{\text{2PMFVB}} = \boldsymbol{y} - \boldsymbol{\omega}_{\text{2PMFVB}} \end{aligned} \tag{4.13}$$

and

$$\boldsymbol{\omega}_{\text{2PMFVB}} = \exp\left\{\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \frac{1}{2}\text{dg}\left(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T\right)\right\}.$$

Details about the moment $\boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1})}$ are given in Section D.1.1 of Appendix D. The updates expressed in form of (4.12) somehow reflect the algebraic forms of results on *best linear unbiased prediction* for mixed models (e.g. Ruppert *et al.*, 2003, Subsection 4.5.3).

The matrix $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ can have massive computational and storage costs when the number of groups is particularly large. However, only the following sub-blocks of $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ are required for variational inference concerning $\boldsymbol{\Sigma}$:

$$\begin{aligned} &\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} = \text{top left-hand } p \times p \text{ sub-block of } \left(\boldsymbol{C}^T\boldsymbol{R}^{-1}_{\text{2PMFVB}}\boldsymbol{C} + \boldsymbol{D}_{\text{2PMFVB}}\right)^{-1}; \\ &\boldsymbol{\Sigma}_{q(\boldsymbol{u}_i)} = \text{subsequent } q \times q \text{ diagonal sub-blocks of} \\ &\qquad\qquad \left(\boldsymbol{C}^T\boldsymbol{R}^{-1}_{\text{2PMFVB}}\boldsymbol{C} + \boldsymbol{D}_{\text{2PMFVB}}\right)^{-1} \text{below } \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})},\ 1 \le i \le m; \\ &E_q\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{q(\beta)})(\boldsymbol{u}_i - \boldsymbol{\mu}_{q(\boldsymbol{u}_i)})^T\} = \text{subsequent } p \times q \text{ sub-blocks of} \\ &\qquad\qquad \left(\boldsymbol{C}^T\boldsymbol{R}^{-1}_{\text{2PMFVB}}\boldsymbol{C} + \boldsymbol{D}_{\text{2PMFVB}}\right)^{-1} \text{ to the right of } \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})},\ 1 \le i \le m. \end{aligned} \tag{4.14}$$

The following result supports the use of Algorithm 4.2 to obtain $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ and the relevant sub-blocks of $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ in (4.14).

Result 4.1. *The mean field variational Bayes updates of* $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ *and each sub-block of* $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ *in* (4.14) *are expressible as a two-level sparse matrix least squares problem of the form:*

$$\left\|\boldsymbol{b} - \boldsymbol{B}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\right\|^2,$$

where $\boldsymbol{b}$ *and the non-zero sub-blocks of* $\boldsymbol{B}$*, according to the notation in (4.5), are, for* $1 \leq i \leq m$*,*

$$\begin{aligned}
\boldsymbol{b}_i &= \begin{bmatrix} diag\left\{(\boldsymbol{\omega}_{2PMFVB})_i^{-1/2}\right\}\{\boldsymbol{y}_i - (\boldsymbol{\omega}_{2PMFVB})_i\} \\ m^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2}\left(\boldsymbol{\mu}_{\boldsymbol{\beta}} - \boldsymbol{\mu}_{q(\boldsymbol{\beta})}\right) \\ -\boldsymbol{M}^{1/2}_{q\left(\boldsymbol{\Sigma}^{-1}\right)}\boldsymbol{\mu}_{q(\boldsymbol{u}_i)} \end{bmatrix}, \\
\boldsymbol{B}_i &= \begin{bmatrix} diag\left\{(\boldsymbol{\omega}_{2PMFVB})_i^{1/2}\right\}\boldsymbol{X}_i \\ m^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2} \\ \boldsymbol{O} \end{bmatrix}, \quad \dot{\boldsymbol{B}}_i = \begin{bmatrix} diag\left\{(\boldsymbol{\omega}_{2PMFVB})_i^{1/2}\right\}\boldsymbol{Z}_i \\ \boldsymbol{O} \\ \boldsymbol{M}^{1/2}_{q(\boldsymbol{\Sigma}^{-1})} \end{bmatrix}.
\end{aligned} \tag{4.15}$$

The solutions can be obtained via the SolveTwoLevelSparseLeastSquares *Algorithm 4.2 and are* $\boldsymbol{\mu}_{q(\boldsymbol{\beta})} = \boldsymbol{x}_1$*,* $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} = \boldsymbol{A}^{11}$ *and*

$$\boldsymbol{\mu}_{q(\boldsymbol{u}_i)} = \boldsymbol{x}_{2,i}, \quad \boldsymbol{\Sigma}_{q(\boldsymbol{u}_i)} = \boldsymbol{A}^{22,i}, \quad E_q\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{q(\beta)})(\boldsymbol{u}_i - \boldsymbol{\mu}_{q(\boldsymbol{u}_i)})^T\} = \boldsymbol{A}^{12,i}, \quad 1 \leq i \leq m.$$

Result 4.1 produces the streamlined MFVB scheme listed as Algorithm 4.3 for the two-level Poisson response model. Its derivation is given in Subsection D.1.1 of Appendix D. It employs the matrix square root of a symmetric positive definite matrix $\boldsymbol{M}$ given by $\boldsymbol{M}^{1/2} = \boldsymbol{U}\text{diag}(\sqrt{\boldsymbol{d}})\boldsymbol{U}^T$, where $\boldsymbol{M} = \boldsymbol{U}\text{diag}(\boldsymbol{d})\boldsymbol{U}^T$ is the singular value decomposition of $\boldsymbol{M}$.

The MFVB approximate marginal log-likelihood

$$\log\{\underline{p}(\boldsymbol{y}; q)\} = E_q\{\log\, p(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\Sigma}, \boldsymbol{A}_{\boldsymbol{\Sigma}}) - q(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\Sigma}, \boldsymbol{A}_{\boldsymbol{\Sigma}})\}. \tag{4.16}$$

can be used as a stopping criterion for Algorithm 4.3. However, an explicit streamlined expression for $\log\{\underline{p}(\boldsymbol{y}; q)\}$ is not provided here and we define convergence by monitoring parameters at each iteration.

4.3.2 Streamlined MFVB for logistic models

Consider the two-level logistic mixed model, for $1 \leq i \leq m$:

$$\begin{aligned}
&\boldsymbol{y}_i \,|\, \boldsymbol{\beta}, \boldsymbol{u}_i \overset{\text{ind.}}{\sim} \text{Bernoulli}\left(\text{logit}^{-1}\left(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{u}_i\right)\right), \quad \boldsymbol{u}_i \,|\, \boldsymbol{\Sigma} \sim N\left(\boldsymbol{0}, \boldsymbol{\Sigma}\right), \\
&\boldsymbol{\beta} \sim N\left(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right), \quad \boldsymbol{\Sigma} | \boldsymbol{A}_{\boldsymbol{\Sigma}} \sim \text{Inverse-G-Wishart}\left(G_{\text{full}}, \nu_{\boldsymbol{\Sigma}} + 2q - 2, \boldsymbol{A}_{\boldsymbol{\Sigma}}^{-1}\right), \\
&\boldsymbol{A}_{\boldsymbol{\Sigma}} \sim \text{Inverse-G-Wishart}(G_{\text{diag}}, 1, \boldsymbol{\Lambda}_{\boldsymbol{A}_{\boldsymbol{\Sigma}}}), \quad \boldsymbol{\Lambda}_{\boldsymbol{A}_{\boldsymbol{\Sigma}}} = \{\nu_{\boldsymbol{\Sigma}} \text{diag}(s^2_{\boldsymbol{\Sigma},1}, \ldots, s^2_{\boldsymbol{\Sigma},q})\}^{-1}.
\end{aligned} \tag{4.17}$$

Matrix and vector dimensions are the same as in (4.8). Analogous arguments to the Poisson case about prior distributions specification hold and we define matrices $\boldsymbol{y}$, $\boldsymbol{X}$, $\boldsymbol{Z}$ and $\boldsymbol{u}$ as in (4.9). Also, we assume the same q-density restriction (4.10). Then, the DAG representation is still the one depicted in Figure 4.1.

As for the Poisson model, evaluation of the optimal q-density as a function of $\boldsymbol{\beta}$ and $\boldsymbol{u}$ according to the canonical MFVB formula (1.8) involves multivariate integrals that are not available in closed form. The non-conjugate variational message passing approach of Knowles and Minka (2011) can be applied also in this case. Nevertheless, the resulting optimal q-density in $\boldsymbol{\beta}$ and $\boldsymbol{u}$ introduces further numerical integration, which is related to the log-partition function of the Bernoulli distribution. Nolan and Wand (2017) provide a detailed illustration on variational inference for logistic regression and overcome this issue by performing the normal scale mixture uniform approximation by Monahan and Stefanski (1989). Table 4.1 displays the vectors $\boldsymbol{p}$ and $\boldsymbol{s}$ containing the coefficients for the most accurate approximation achievable with the values provided Monahan and Stefanski (1989), that is, the 8 normal scale mixture uniform approximation. This

TABLE 4.1: Vectors p and s corresponding to the $k = 8$ normal scale mixture uniform approximation of Monahan and Stefanski (1989).

$\boldsymbol{p}$	$\boldsymbol{s}$
0.003246343272134	1.365340806296348
0.051517477033972	1.059523971016916
0.195077912673858	0.830791313765644
0.315569823632818	0.650732166639391
0.274149576158423	0.508135425366489
0.131076880695470	0.396313345166341
0.027912418727972	0.308904252267995
0.001449567805354	0.238212616409306

solution leads to the following optimal q-density functions for the parameters of interest,

which is similar to that of the Poisson case:

$$\begin{aligned} &q^*(\boldsymbol{\beta}, \boldsymbol{u}) \text{ has a } N\big(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\big) \text{ density function;} \\ &q^*(\boldsymbol{\Sigma}) \text{ has an Inverse-G-Wishart}(G_{\text{full}}, \xi_{q(\boldsymbol{\Sigma})}, \boldsymbol{\Lambda}_{q(\boldsymbol{\Sigma})}) \text{ density function.} \end{aligned} \tag{4.18}$$

The updates for $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ can be written as

$$\begin{aligned} \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} &\leftarrow \left(\boldsymbol{C}^T \boldsymbol{R}^{-1}_{\text{2LMFVB}} \boldsymbol{C} + \boldsymbol{D}_{\text{2LMFVB}}\right)^{-1} \left(\boldsymbol{C}^T \boldsymbol{z}_{\text{2LMFVB}} + \boldsymbol{o}_{\text{2LMFVB}}\right) \\ \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} &\leftarrow \left(\boldsymbol{C}^T \boldsymbol{R}^{-1}_{\text{2LMFVB}} \boldsymbol{C} + \boldsymbol{D}_{\text{2LMFVB}}\right)^{-1}, \end{aligned} \tag{4.19}$$

where $\boldsymbol{C} = [\boldsymbol{X} \ \boldsymbol{Z}]$,

$$\begin{aligned} &\boldsymbol{R}^{-1}_{\text{2LMFVB}} = \text{diag}\left(\boldsymbol{\omega}_{\text{2LMFVB2}}\right), \quad \boldsymbol{D}_{\text{2LMFVB}} = \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I}_m \otimes \boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1})} \end{bmatrix}, \\ &\boldsymbol{o}_{\text{2LMFVB}} = \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\beta}} \\ \boldsymbol{0} \end{bmatrix}, \quad \boldsymbol{z}_{\text{2LMFVB}} = \boldsymbol{y} - \boldsymbol{\omega}_{\text{2LMFVB1}} + \boldsymbol{\omega}_{\text{2LMFVB2}} \odot \boldsymbol{\mu}, \end{aligned} \tag{4.20}$$

and

$$\begin{aligned} &\boldsymbol{\omega}_{\text{2LMFVB1}} = \Phi\left(\left(\boldsymbol{\mu}\boldsymbol{s}^T\right)/\boldsymbol{\Omega}\right)\boldsymbol{p}, \quad \boldsymbol{\omega}_{\text{2LMFVB2}} = \left\{\phi\left(\left(\boldsymbol{\mu}\boldsymbol{s}^T\right)/\boldsymbol{\Omega}\right)/\boldsymbol{\Omega}\right\}(\boldsymbol{p} \odot \boldsymbol{s}), \\ &\boldsymbol{\mu} = \boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}, \quad \boldsymbol{\sigma}^2 = \text{dg}\left(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T\right), \quad \boldsymbol{\Omega} = \sqrt{\boldsymbol{1}_n \boldsymbol{1}_8^T + \boldsymbol{\sigma}^2 \left(\boldsymbol{s}^2\right)^T}, \end{aligned}$$

with $\boldsymbol{p}$ and $\boldsymbol{s}$ defined in Table 4.1. These results can be derived considering the non-streamlined VMP scheme for the logistic likelihood fragment listed as Algorithm 2 of Nolan and Wand (2017) and following the derivation of Algorithm 4.3.

As for the Poisson model, only the sub-blocks of $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ indicated in (4.14) are relevant for streamlined MFVB, with $\boldsymbol{R}^{-1}_{\text{2PMFVB}}$ and $\boldsymbol{D}_{\text{2PMFVB}}$ replaced by $\boldsymbol{R}^{-1}_{\text{2LMFVB}}$ and $\boldsymbol{D}_{\text{2LMFVB}}$, respectively. The following result allows to take advantage of Algorithm 4.2 to obtain $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ and the relevant sub-blocks of $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$.

Result 4.2. *The mean field variational Bayes updates of* $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ *and each sub-block of* $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ *in* (4.14) *are expressible as a two-level sparse matrix least squares problem of the form:*

$$\left\|\boldsymbol{b} - \boldsymbol{B}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\right\|^2,$$

where $\boldsymbol{b}$ *and the non-zero sub-blocks of* $\boldsymbol{B}$*, according to the notation in (4.5), are, for* $1 \leq i \leq m$*,*

$$\boldsymbol{b}_i = \begin{bmatrix} diag\left\{(\boldsymbol{\omega}_{2LMFVB2})_i^{-1/2}\right\}\{\boldsymbol{y}_i - (\boldsymbol{\omega}_{2LMFVB1})_i + (\boldsymbol{\omega}_{2LMFVB2})_i \odot (\boldsymbol{\mu})_i\} \\ m^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2}\boldsymbol{\mu}_{\boldsymbol{\beta}} \\ \mathbf{0} \end{bmatrix},$$

$$\boldsymbol{B}_i = \begin{bmatrix} diag\left\{(\boldsymbol{\omega}_{2LMFVB2})_i^{1/2}\right\}\boldsymbol{X}_i \\ m^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2} \\ \boldsymbol{O} \end{bmatrix}, \quad \dot{\boldsymbol{B}}_i = \begin{bmatrix} diag\left\{(\boldsymbol{\omega}_{2LMFVB2})_i^{1/2}\right\}\boldsymbol{Z}_i \\ \boldsymbol{O} \\ \boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1})}^{1/2} \end{bmatrix}.$$

The solutions can be obtained via the SolveTwoLevelSparseLeastSquares *Algorithm 4.2 and are* $\boldsymbol{\mu}_{q(\boldsymbol{\beta})} = \boldsymbol{x}_1$*,* $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} = \boldsymbol{A}^{11}$ *and*

$$\boldsymbol{\mu}_{q(\boldsymbol{u}_i)} = \boldsymbol{x}_{2,i}, \quad \boldsymbol{\Sigma}_{q(\boldsymbol{u}_i)} = \boldsymbol{A}^{22,i}, \quad E_q\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{q(\beta)})(\boldsymbol{u}_i - \boldsymbol{\mu}_{q(\boldsymbol{u}_i)})^T\} = \boldsymbol{A}^{12,i}, \quad 1 \leq i \leq m.$$

The derivation of Result 4.2 is similar to that of Result 4.1. Result 4.2 produces the streamlined MFVB scheme listed as Algorithm 4.4 for the two-level logistic model, via the use of the SolveTwoLevelSparseLeastSquares Algorithm 4.2. Algorithm 4.4 can be derived similarly to Algorithm 4.3. The MFVB approximate marginal log-likelihood $\log\{\underline{p}(\boldsymbol{y}; q)\}$ expressed in (4.16) can be used as a stopping criterion.

4.4 VMP for two-level random effects models

This section presents streamlined VMP algorithms for two-level linear mixed models with Poisson and binomial responses.

First, note that both the Poisson and binomial models (4.7) and (4.17) admit the following factorization:

$$p(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\Sigma}, \boldsymbol{A}_{\boldsymbol{\Sigma}}) = p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u})\, p(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma})\, p(\boldsymbol{\Sigma}|\boldsymbol{A}_{\boldsymbol{\Sigma}})\, p(\boldsymbol{A}_{\boldsymbol{\Sigma}}). \tag{4.21}$$

Figure 4.2 shows a factor graph representation of (4.21). Colors differentiate the various fragments appearing in the factor graph. The main likelihood fragment in sky blue given by factor $p(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \boldsymbol{u})$ and stochastic node $(\boldsymbol{\beta}, \boldsymbol{u})$ is treated in Wand (2017, Section 5.3) for the Poisson case and Nolan and Wand (2017) for the logistic model. The other fragments are treated in Wand (2017, Section 4.1).

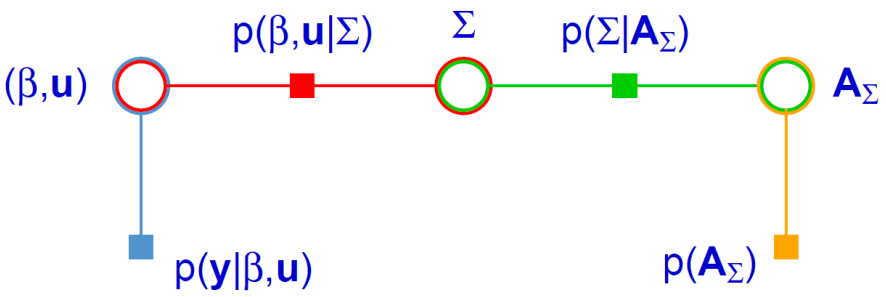

FIGURE 4.2: Factor graph for the two-level Poisson and logistic response mixed model in (4.7) and (4.17).

Nolan *et al.* (2018) provide streamlined VMP updates for the Gaussian penalization fragment in red given by factor $p(\boldsymbol{\beta}, \boldsymbol{u} \mid \boldsymbol{\Sigma})$ and stochastic nodes $(\boldsymbol{\beta}, \boldsymbol{u})$ and $\boldsymbol{\Sigma}$, when the random effects vector has a two-level structure. The corresponding algorithm can be used also for the non-Gaussian response models considered here, thanks to the notions of message passing on factor graph fragments.

4.4.1 Streamlined Poisson and logistic likelihood fragments updates

We now focus on the Poisson and logistic likelihood fragments, which correspond to the sky blue factor graph fragment in Figure 4.2. Similarly to Wand (2017, Section 5.3) and Algorithm 2 of Nolan and Wand (2017), in both cases the messages passed between $p(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{u})$ and $(\boldsymbol{\beta}, \boldsymbol{u})$ involve multivariate normal distributions with natural parameter vectors containing

$$p + mq + \frac{1}{2}(p + mq)(p + mq + 1)$$

unique entries. Therefore the sizes of these vectors grow quadratically with the number of groups m, making message passing computationally demanding. However, these messages are within *reduced* multivariate normal families, where the relevant components of the sufficient statistic vector are those associated with the relevant sub-blocks of $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$, as for MFVB.

By imposing a conjugacy constraint, all messages passed to $(\boldsymbol{\beta}, \boldsymbol{u})$ from factors outside of the Poisson or logistic two-level likelihood fragment are within the same reduced multivariate normal family. Under such a constraint, the natural parameter vectors of messages passed to and from $(\boldsymbol{\beta}, \boldsymbol{u})$ have length

$$p + \frac{1}{2}p(p+1) + m\left\{q + \frac{1}{2}q(q+1) + pq\right\}, \tag{4.22}$$

which is linear in the number of groups.

The message from $p(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{u})$ to $(\boldsymbol{\beta}, \boldsymbol{u})$ is

$$m_{p(\boldsymbol{y} \mid \boldsymbol{\beta},\boldsymbol{u}) \to (\boldsymbol{\beta},\boldsymbol{u})} (\boldsymbol{\beta}, \boldsymbol{u}) = \exp \left\{ \begin{bmatrix} \boldsymbol{\beta} \\ \operatorname{vech}\left(\boldsymbol{\beta}\boldsymbol{\beta}^T\right) \\ \underset{1 \le i \le m}{\operatorname{stack}} \begin{bmatrix} \boldsymbol{u}_i \\ \operatorname{vech}\left(\boldsymbol{u}_i \boldsymbol{u}_i^T\right) \\ \operatorname{vec}\left(\boldsymbol{\beta}\boldsymbol{u}_i^T\right) \end{bmatrix} \end{bmatrix}^T \boldsymbol{\eta}_{p(\boldsymbol{y} \mid \boldsymbol{\beta},\boldsymbol{u}) \to (\boldsymbol{\beta},\boldsymbol{u})} \right\}, \quad (4.23)$$

with natural parameter vector $\boldsymbol{\eta}_{p(\boldsymbol{y} \mid \boldsymbol{\beta},\boldsymbol{u}) \to (\boldsymbol{\beta},\boldsymbol{u})}$ of length (4.22). Under the conjugacy constraint, also the message $m_{(\boldsymbol{\beta},\boldsymbol{u}) \to p(\boldsymbol{y} \mid \boldsymbol{\beta},\boldsymbol{u})}$ has the form (4.23) with natural parameter vector also of length (4.22).

Result 4.3. *The variational message passing updates of the quantities* $\boldsymbol{\mu}_{q(\boldsymbol{\beta})}$, $\boldsymbol{\mu}_{q(\boldsymbol{u}_i)}$, $1 \le i \le m$, *and the sub-blocks of* $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ *listed in* (4.14), *with q-density expectations with respect to the normalization of*

$$m_{p(\boldsymbol{y} \mid \boldsymbol{\beta},\boldsymbol{u}) \to (\boldsymbol{\beta},\boldsymbol{u})} (\boldsymbol{\beta}, \boldsymbol{u})\, m_{(\boldsymbol{\beta},\boldsymbol{u}) \to p(\boldsymbol{y} \mid \boldsymbol{\beta},\boldsymbol{u})} (\boldsymbol{\beta}, \boldsymbol{u}),$$

are expressible as a two-level sparse matrix problem with

$$\boldsymbol{A} = -2 \begin{bmatrix} vec^{-1}\left(\boldsymbol{D}_p^{+T} \boldsymbol{\eta}_{1,2}\right) & \left[\frac{1}{2} \underset{1 \le i \le m}{\operatorname{stack}} \left\{ vec^{-1}\left(\boldsymbol{\eta}_{2,3,i}\right)^T \right\}\right]^T \\ \frac{1}{2} \underset{1 \le i \le m}{\operatorname{stack}} \left\{ vec^{-1}\left(\boldsymbol{\eta}_{2,3,i}\right)^T \right\} & \underset{1 \le i \le m}{\operatorname{blockdiag}} \left\{ vec^{-1}\left(\boldsymbol{D}_q^{+T} \boldsymbol{\eta}_{2,2,i}\right) \right\} \end{bmatrix}$$

and

$$\boldsymbol{a} = \begin{bmatrix} \boldsymbol{\eta}_{1,1} \\ \\ \underset{1 \le i \le m}{\operatorname{stack}} \left(\boldsymbol{\eta}_{2,1,i}\right) \end{bmatrix}, \quad \textit{where} \quad \begin{bmatrix} \boldsymbol{\eta}_{1,1} \ \ (p \times 1) \\ \boldsymbol{\eta}_{1,2} \ \ \left(\frac{1}{2} p\,(p+1) \times 1\right) \\ \underset{1 \le i \le m}{\operatorname{stack}} \begin{bmatrix} \boldsymbol{\eta}_{2,1,i} & (q \times 1) \\ \boldsymbol{\eta}_{2,2,i} & \left(\frac{1}{2} q\,(q+1) \times 1\right) \\ \boldsymbol{\eta}_{2,3,i} & (pq \times 1) \end{bmatrix} \end{bmatrix}$$

is the partitioning of $\boldsymbol{\eta}_{p(\boldsymbol{y} \mid \boldsymbol{\beta},\boldsymbol{u}) \leftrightarrow (\boldsymbol{\beta},\boldsymbol{u})}$ *that defines*

$$\boldsymbol{\eta}_{1,1}, \quad \boldsymbol{\eta}_{1,2}, \quad \left\{ \left(\boldsymbol{\eta}_{2,1,i}, \boldsymbol{\eta}_{2,2,i}, \boldsymbol{\eta}_{2,3,i}\right) : 1 \le i \le m \right\}.$$

The solutions are $\boldsymbol{\mu}_{q(\boldsymbol{\beta})} = \boldsymbol{x}_1$, $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} = \boldsymbol{A}^{11}$ *and*

$$\boldsymbol{\mu}_{q(\boldsymbol{u}_i)} = \boldsymbol{x}_{2,i}, \quad \boldsymbol{\Sigma}_{q(\boldsymbol{u}_i)} = \boldsymbol{A}^{22,i}, \quad E_q\left\{\left(\boldsymbol{\beta} - \boldsymbol{\mu}_{q(\beta)}\right)\left(\boldsymbol{u}_i - \boldsymbol{\mu}_{q(\boldsymbol{u}_i)}\right)^T\right\} = \boldsymbol{A}^{12,i}, \quad 1 \leq i \leq m.$$

Result 4.3 is derived in Subsection D.2.1 of Appendix D As illustrated by Result 4.3, the process of converting a generic reduced natural parameter vector $\boldsymbol{\eta}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ to $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ and relevant sub-blocks of $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ is an important step to streamlining variational message passing. This procedure is formalized as the TwoLevelNaturalToCommonParameters algorithm, which is listed as Algorithm 4.5.

Result 4.3 gives rise to the streamlined VMP schemes listed as Algorithms 4.6 and 4.7 for the two-level Poisson response and logistic models respectively, which can be derived considering the VMP scheme for the Poisson and logistic likelihood fragments listed in Wand (2017, Subsection 5.3) and as Algorithm 2 of Nolan and Wand (2017), respectively.

4.5 Illustrative examples

Consider a fictional educational experiment involving data from a set of schools in which some students were given the possibility to participate in supplementary tutorial activities. Suppose the number of hours of extra activities were recorded for each student as a response variable. Also, scores of a preliminary test and covariates characterizing each school are available.

We then simulate a dataset from

$$y_{ij} \,|\, \boldsymbol{\beta}, \boldsymbol{u}_i \overset{\text{ind.}}{\sim} \text{Poisson}\left(\exp\left(\beta_0 + u_{0i} + (\beta_1 + u_{1i})\, x_{ij}\right)\right), \quad \boldsymbol{u}_i \,|\, \boldsymbol{\Sigma} \sim N\left(\boldsymbol{0}, \boldsymbol{\Sigma}\right),$$

with $1 \leq i \leq 36$, $1 \leq j \leq 200$, $x_{ij} \sim \text{Uniform}\,(0,1)$ and

$$\boldsymbol{\beta} = (\beta_0, \beta_1) = (0.5, -0.9), \quad \boldsymbol{u}_i = (u_{01}, u_{1i}), \quad \boldsymbol{\Sigma} = \begin{bmatrix} 0.25 & 0.15 \\ 0.15 & 0.15 \end{bmatrix}.$$

Therefore, $p = q = 2$, $m = 36$ and $n_i = n = 200$, for $1 \leq i \leq m$, that is, 36 schools with 200 students each are monitored. Next, we fit the model (4.7) via the VMP Algorithm 4.6 for Poisson two-level linear mixed models, or equivalently the MFVB Algorithm 4.3. We choose the hyperparameters $\boldsymbol{\mu}_{\boldsymbol{\beta}} = \boldsymbol{0}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = 10^5 \boldsymbol{I}$ are chosen. Results of the fitting procedure are available in a fraction of second with the algorithms coded in `R`

(R Core Team, 2018). Convergence has been considered achieved when the absolute relative error for the optimal q-density parameters computed at the current iteration with respect to the previous one is less than 10^{-10}. In each panel of Figure 4.3, the approximate posterior mean corresponding to the best linear unbiased prediction fit and the pointwise 95% credible sets are displayed.

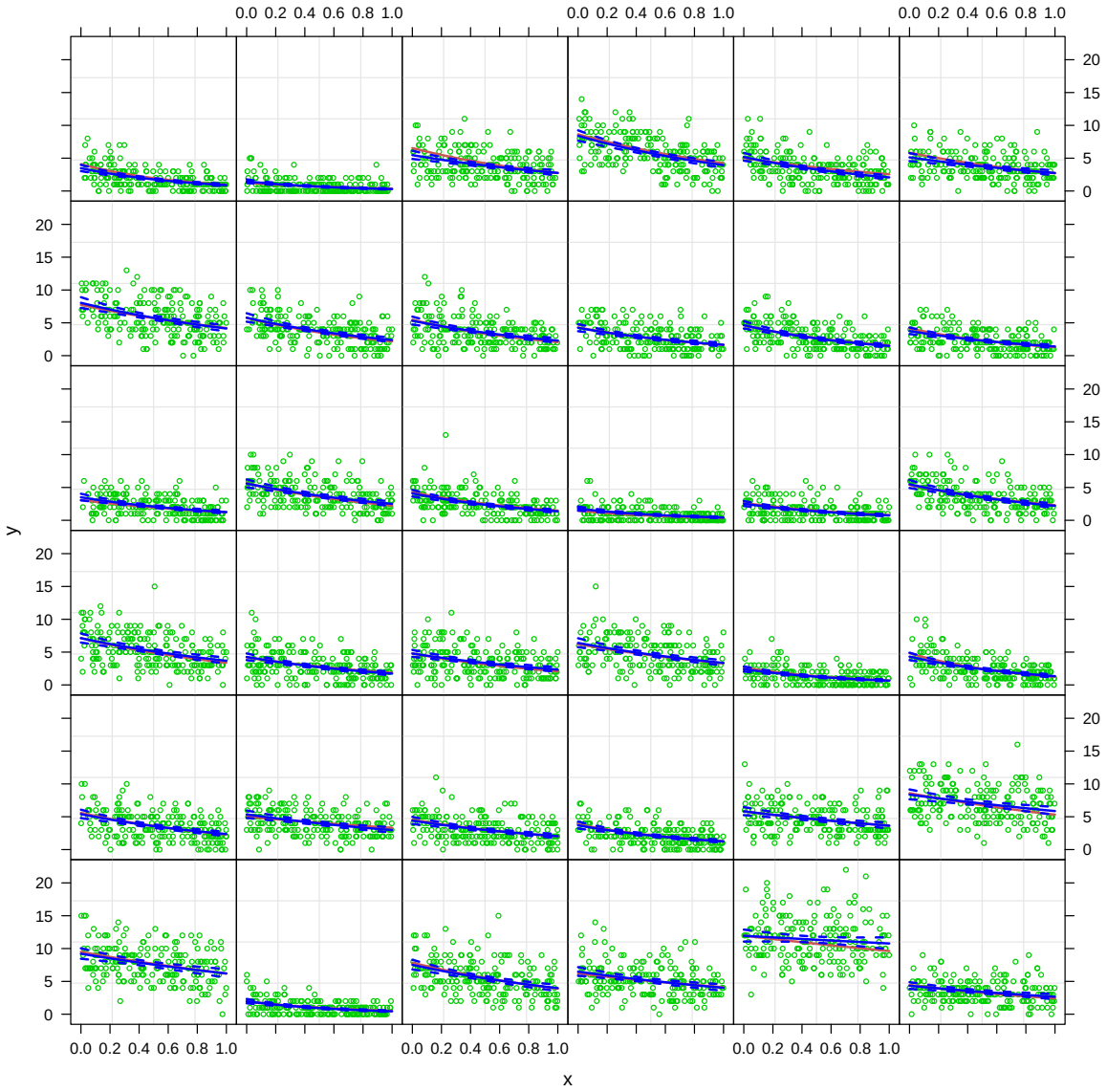

FIGURE 4.3: Simulated two-level data with 36 schools, each having 200 students, according to the Poisson multilevel model described in Section 4.5. Each panel contains the approximate posterior mean (solid line) and pointwise 95% credible intervals for the mean response. The true function from which the data were generated is shown as a red solid line.

In a similar fashion, we generate data from the model

$$y_{ij} \,|\, \boldsymbol{\beta}, \boldsymbol{u}_i \overset{\text{ind.}}{\sim} \text{Bernoulli}\left(\text{logit}^{-1}\left(\beta_0 + u_{0i} + (\beta_1 + u_{1i})\, x_{ij}\right)\right), \quad \boldsymbol{u}_i \,|\, \boldsymbol{\Sigma} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}\right),$$

with $1 \leq i \leq 100$, $1 \leq j \leq 500$, $x_{ij} \sim \text{Uniform}\,(0,1)$ and

$$\boldsymbol{\beta} = (\beta_0, \beta_1) = (0.58, 1.89)\,, \quad \boldsymbol{u}_i = (u_{01}, u_{1i})\,, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 0.05 & 0.005 \\ 0.005 & 0.35 \end{bmatrix}.$$

Hence, $p = q = 2$, $m = 100$ and $n_i = n = 500$, for $1 \leq i \leq m$. We then perform variational inference under the same specification of prior distributions and convergence assessment to obtain the logistic counterpart of the simulation study, whose results are displayed in Figure 4.4. The estimation in `R` was performed in less than 10 seconds in a standard working laptop, without performing parallel computing or defining any accurate parameters initialization. A more structured simulation study for higher dimensional datasets could also be implemented to quantify the effective benefits of the streamlined algorithms in terms of computing time.

4.6 Concluding remarks

Sparsity is one of the keys for big model inference as datasets and models continue to grow in size. Matrix algebraic streamlining in the form of sparse matrix-type refinements are then fundamental for efficient handling of large longitudinal and multilevel datasets via MFVB or VMP. These variational approximation methods are relatively simple approaches whose algorithms benefit of streamlined computing of the fragment updates. Taking advantage of recent results in Nolan *et al.* (2018), we have derived explicit algorithms that facilitate streamlined variational inference for two-level models with the two common Poisson and binomial responses.

These first achievements are very important for future extensions to higher level models and especially for approximate inference with some very large semiparametric models. Much has to be done to cover a variety of models and applications. Additional tedious effort will be necessary to cover a larger variety of responses and produce wide spectrum software implementations with low level programming languages.

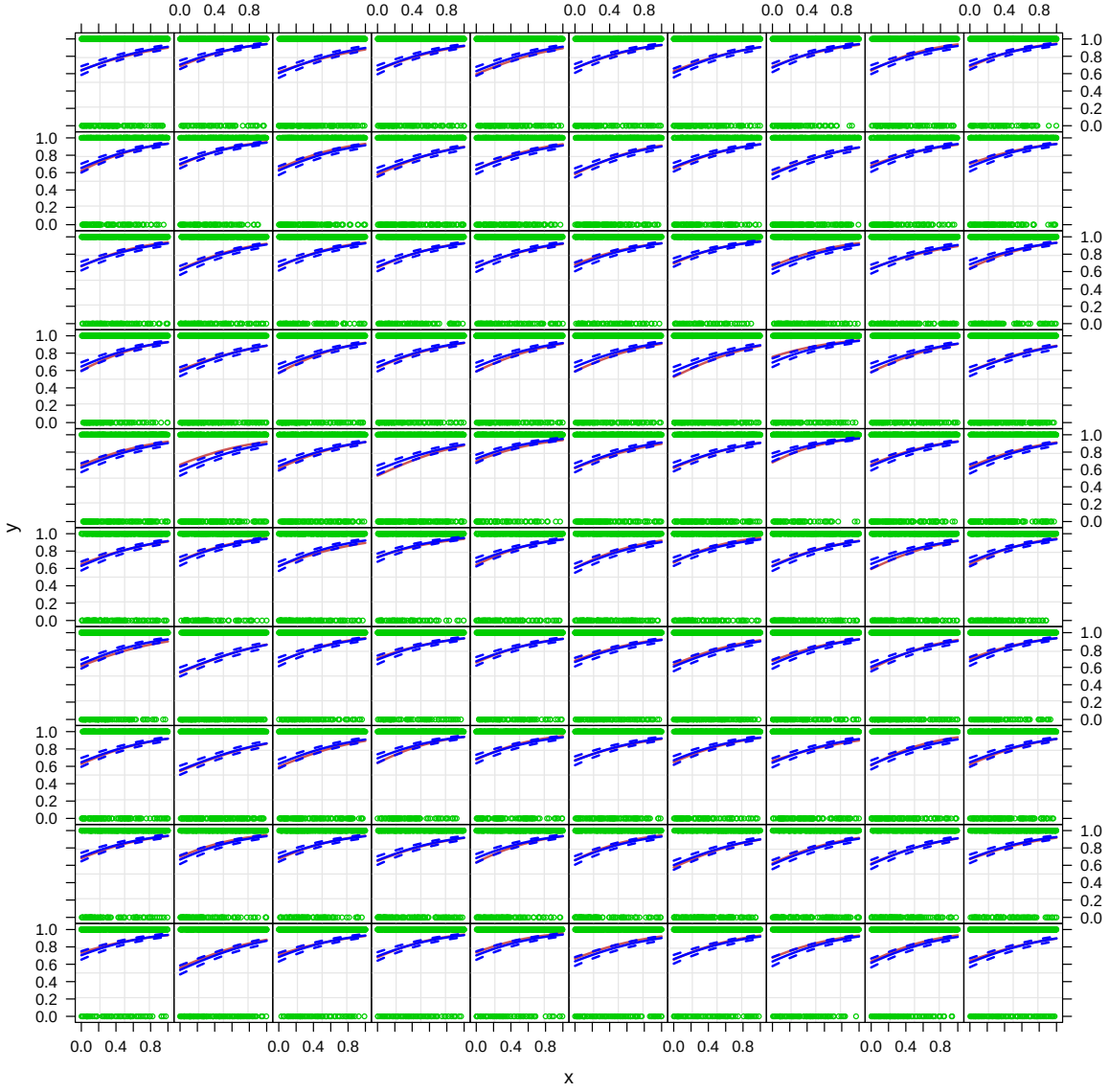

FIGURE 4.4: Simulated two-level data with 100 schools, each having 500 students, according to the logistic multilevel model described in Section 4.5. Each panel contains the approximate posterior mean (solid line) and pointwise 95% credible intervals for the mean response. The true function from which the data were generated is shown as a red solid line.

Algorithm 4.3 *QR-decomposition-based streamlined algorithm for obtaining mean field variational Bayes approximate posterior density functions for the parameters in the two-level Poisson mixed model* (4.7) *with product density restriction* (4.10).

Data Inputs: $\boldsymbol{y}_i(n_i \times 1),\ \boldsymbol{X}_i(n_i \times p),\ \boldsymbol{Z}_i(n_i \times q),\ 1 \leq i \leq m.$

Hyperparameter Inputs: $\boldsymbol{\mu}_{\boldsymbol{\beta}}(p \times 1)$, $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}(p \times p)$ symmetric and positive definite, $s_{\boldsymbol{\Sigma},1}, \ldots, s_{\boldsymbol{\Sigma},q}, \nu_{\boldsymbol{\Sigma}} > 0.$

Initialize: $\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\,(p \times 1)$, $E_q\left\{\left(\boldsymbol{\beta} - \boldsymbol{\mu}_{q(\boldsymbol{\beta})}\right)\left(\boldsymbol{u}_i - \boldsymbol{\mu}_{q(\boldsymbol{u}_i)}\right)^T\right\}(p \times q)$, $\boldsymbol{\mu}_{q(\boldsymbol{u}_i)}\,(q \times q)$;

$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\,(p \times p)$, $\boldsymbol{\Sigma}_{q(\boldsymbol{u}_i)}\,(q \times q)$, $M_{q(\boldsymbol{\Sigma}^{-1})}(q \times q),\ M_{q(\boldsymbol{A}_{\boldsymbol{\Sigma}}^{-1})}(q \times q)$ all symmetric and positive definite.

$\xi_{q(\boldsymbol{\Sigma})} \longleftarrow \nu_{\boldsymbol{\Sigma}} + 2q - 2 + m \quad ; \quad \xi_{q(\boldsymbol{A}_{\boldsymbol{\Sigma}})} \longleftarrow \nu_{\boldsymbol{\Sigma}} + q$

Cycle:

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

$$\boldsymbol{\omega}_{\text{2PMFVB}i} \longleftarrow \exp\left\{\boldsymbol{X}_i\boldsymbol{\mu}_{q(\boldsymbol{\beta})} + \boldsymbol{Z}_i\boldsymbol{\mu}_{q(\boldsymbol{u}_i)} + \tfrac{1}{2}\text{dg}\left(\boldsymbol{X}_i\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\boldsymbol{X}_i^T + \boldsymbol{Z}_i\boldsymbol{\Sigma}_{q(\boldsymbol{u}_i)}\boldsymbol{Z}_i^T + 2\boldsymbol{X}_i E_q\left\{\left(\boldsymbol{\beta} - \boldsymbol{\mu}_{q(\boldsymbol{\beta})}\right)\left(\boldsymbol{u}_i - \boldsymbol{\mu}_{q(\boldsymbol{u}_i)}\right)^T\right\}\boldsymbol{Z}_i^T\right)\right\}$$

$$\boldsymbol{b}_i \longleftarrow \begin{bmatrix} \text{diag}\left(\boldsymbol{\omega}_{\text{2PMFVB}i}^{-1/2}\right)(\boldsymbol{y}_i - \boldsymbol{\omega}_{\text{2PMFVB}i}) \\ m^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2}\left(\boldsymbol{\mu}_{\boldsymbol{\beta}} - \boldsymbol{\mu}_{q(\boldsymbol{\beta})}\right) \\ -\boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1})}^{1/2}\boldsymbol{\mu}_{q(\boldsymbol{u}_i)} \end{bmatrix}$$

$$\boldsymbol{B}_i \longleftarrow \begin{bmatrix} \text{diag}\left(\boldsymbol{\omega}_{\text{2PMFVB}i}^{1/2}\right)\boldsymbol{X}_i \\ m^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2} \\ \boldsymbol{O} \end{bmatrix} \ ; \ \dot{\boldsymbol{B}}_i \longleftarrow \begin{bmatrix} \text{diag}\left(\boldsymbol{\omega}_{\text{2PMFVB}i}^{1/2}\right)\boldsymbol{Z}_i \\ \boldsymbol{0} \\ \boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1})}^{1/2} \end{bmatrix}$$

$$S_{\text{2PMFVB}} \longleftarrow \text{SolveTwoLevelSparseLeastSquares}\left(\left\{\left(\boldsymbol{b}_i, \boldsymbol{B}_i, \dot{\boldsymbol{B}}_i\right) : 1 \leq i \leq m\right\}\right)$$

$\boldsymbol{\mu}_{q(\boldsymbol{\beta})} \longleftarrow \boldsymbol{x}_1$ component of $S_{\text{2PMFVB}} \quad ; \quad \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} \longleftarrow \boldsymbol{A}^{11}$ component of S_{2PMFVB}

$\Lambda_{q(\boldsymbol{\Sigma})} \longleftarrow \boldsymbol{M}_{q(\boldsymbol{A}_{\boldsymbol{\Sigma}}^{-1})}$

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

$\boldsymbol{\mu}_{q(\boldsymbol{u}_i)} \longleftarrow \boldsymbol{x}_{2,i}$ component of S_{2PMFVB}

$\boldsymbol{\Sigma}_{q(\boldsymbol{u}_i)} \longleftarrow \boldsymbol{A}^{22,i}$ component of S_{2PMFVB}

$E_q\left\{\left(\boldsymbol{\beta} - \boldsymbol{\mu}_{q(\boldsymbol{\beta})}\right)\left(\boldsymbol{u}_i - \boldsymbol{\mu}_{q(\boldsymbol{u}_i)}\right)^T\right\} \longleftarrow \boldsymbol{A}^{12,i}$ component of S_{2PMFVB}

$\boldsymbol{\Lambda}_{q(\boldsymbol{\Sigma})} \longleftarrow \boldsymbol{\Lambda}_{q(\boldsymbol{\Sigma})} + \boldsymbol{\mu}_{q(\boldsymbol{u}_i)}\boldsymbol{\mu}_{q(\boldsymbol{u}_i)}^T + \boldsymbol{\Sigma}_{q(\boldsymbol{u}_i)}$

$\boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1})} \longleftarrow (\xi_{q(\boldsymbol{\Sigma})} - q + 1)\,\boldsymbol{\Lambda}_{q(\boldsymbol{\Sigma})}^{-1}$

$\boldsymbol{\Lambda}_{q(\boldsymbol{A}_{\boldsymbol{\Sigma}})} \longleftarrow \text{diag}\left\{\text{dg}\left(\boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1})}\right)\right\} + \{\nu_{\boldsymbol{\Sigma}}\text{diag}(s_{\boldsymbol{\Sigma},1}^2, \ldots, s_{\boldsymbol{\Sigma},q}^2)\}^{-1}$

$\boldsymbol{M}_{q(\boldsymbol{A}_{\boldsymbol{\Sigma}}^{-1})} \longleftarrow \xi_{q(\boldsymbol{A}_{\boldsymbol{\Sigma}})}\boldsymbol{\Lambda}_{q(\boldsymbol{A}_{\boldsymbol{\Sigma}})}^{-1}$

until convergence.

Output: $\Big(\boldsymbol{\mu}_{q(\boldsymbol{\beta})},\ \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})},\ \left\{\left(\boldsymbol{\mu}_{q(\boldsymbol{u}_i)}, \boldsymbol{\Sigma}_{q(\boldsymbol{u}_i)}, E_q\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{q(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{q(\boldsymbol{u}_i)})^T\}\right) : 1 \leq i \leq m\right\},$ $\xi_{q(\boldsymbol{\Sigma})}, \boldsymbol{\Lambda}_{q(\boldsymbol{\Sigma})}\Big).$

Algorithm 4.4 *QR-decomposition-based streamlined algorithm for obtaining mean field variational Bayes approximate posterior density functions for the parameters in the two-level logistic mixed model* (4.17) *with product density restriction* (4.10).

Data Inputs: $\boldsymbol{y}_i(n_i \times 1),\ \boldsymbol{X}_i(n_i \times p),\ \boldsymbol{Z}_i(n_i \times q),\ 1 \leq i \leq m.$

Hyperparameter Inputs: $\boldsymbol{\mu}_{\boldsymbol{\beta}}(p \times 1)$, $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}(p \times p)$ symmetric and positive definite, $s_{\boldsymbol{\Sigma},1}, \ldots, s_{\boldsymbol{\Sigma},q}, \nu_{\boldsymbol{\Sigma}} > 0.$

Constant inputs: $\boldsymbol{p}$, $\boldsymbol{s}$ from Table 4.1

Initialize: $\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\,(p \times 1)$, $E_q\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{q(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{q(\boldsymbol{u}_i)})^T\}\,(p \times q)$, $\boldsymbol{\mu}_{q(\boldsymbol{u}_i)}\,(q \times q)$;
$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\,(p \times p)$, $\boldsymbol{\Sigma}_{q(\boldsymbol{u}_i)}\,(q \times q)$, $M_{q(\boldsymbol{\Sigma}^{-1})}(q \times q),\ M_{q(\boldsymbol{A}_{\boldsymbol{\Sigma}}^{-1})}(q \times q)$ all symmetric and positive definite.

$\xi_{q(\boldsymbol{\Sigma})} \longleftarrow \nu_{\boldsymbol{\Sigma}} + 2q - 2 + m \ ; \ \xi_{q(\boldsymbol{A}_{\boldsymbol{\Sigma}})} \longleftarrow \nu_{\boldsymbol{\Sigma}} + q$

Cycle:

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

$$\boldsymbol{\mu}_i \longleftarrow \boldsymbol{X}_i\boldsymbol{\mu}_{q(\boldsymbol{\beta})} + \boldsymbol{Z}_i\boldsymbol{\mu}_{q(\boldsymbol{u}_i)}$$

$$\boldsymbol{\sigma}_i^2 \longleftarrow \mathrm{dg}\left(\boldsymbol{X}_i\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\boldsymbol{X}_i^T + \boldsymbol{Z}_i\boldsymbol{\Sigma}_{q(\boldsymbol{u}_i)}\boldsymbol{Z}_i^T + 2\boldsymbol{X}_i E_q\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{q(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{q(\boldsymbol{u}_i)})^T\}\boldsymbol{Z}_i^T\right)$$

$$\boldsymbol{\Omega}_i \longleftarrow \sqrt{\mathbf{1}_{n_i}\mathbf{1}_8^T + \boldsymbol{\sigma}_i^2\,(\boldsymbol{s}^2)^T} \ ; \ \boldsymbol{\omega}_{\text{2LMFVB1}i} \longleftarrow \Phi\left(\left(\boldsymbol{\mu}_i\boldsymbol{s}^T\right)/\boldsymbol{\Omega}_i\right)\boldsymbol{p}$$

$$\boldsymbol{\omega}_{\text{2LMFVB2}i} \longleftarrow \left\{\phi\left(\left(\boldsymbol{\mu}_i\boldsymbol{s}^T\right)/\boldsymbol{\Omega}_i\right)/\boldsymbol{\Omega}_i\right\}(\boldsymbol{p} \odot \boldsymbol{s})$$

$$\boldsymbol{b}_i \longleftarrow \begin{bmatrix} \mathrm{diag}\left(\boldsymbol{\omega}_{\text{2LMFVB2}i}^{-1/2}\right)(\boldsymbol{y}_i - \boldsymbol{\omega}_{\text{2LMFVB1}i} + \boldsymbol{\omega}_{\text{2LMFVB2}i} \odot \boldsymbol{\mu}_i) \\ m^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2}\boldsymbol{\mu}_{\boldsymbol{\beta}} \\ \mathbf{0} \end{bmatrix}$$

$$\boldsymbol{B}_i \longleftarrow \begin{bmatrix} \mathrm{diag}\left(\boldsymbol{\omega}_{\text{2LMFVB2}i}^{-1/2}\right)\boldsymbol{X}_i \\ m^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2} \\ \boldsymbol{O} \end{bmatrix} \ ; \ \dot{\boldsymbol{B}}_i \longleftarrow \begin{bmatrix} \mathrm{diag}\left(\boldsymbol{\omega}_{\text{2LMFVB2}i}^{1/2}\right)\boldsymbol{Z}_i \\ \mathbf{0} \\ \boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1})}^{1/2} \end{bmatrix}.$$

$$S_{\text{2LMFVB}} \longleftarrow \textsc{SolveTwoLevelSparseLeastSquares}\left(\left\{\left(\boldsymbol{b}_i, \boldsymbol{B}_i, \dot{\boldsymbol{B}}_i\right) : 1 \leq i \leq m\right\}\right)$$

$$\boldsymbol{\mu}_{q(\boldsymbol{\beta})} \longleftarrow \boldsymbol{x}_1 \text{ component of } S_{\text{2LMFVB}} \ ; \ \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} \longleftarrow \boldsymbol{A}^{11} \text{ component of } S_{\text{2LMFVB}}$$

$$\Lambda_{q(\boldsymbol{\Sigma})} \longleftarrow \boldsymbol{M}_{q(\boldsymbol{A}_{\boldsymbol{\Sigma}}^{-1})}$$

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

$$\boldsymbol{\mu}_{q(\boldsymbol{u}_i)} \longleftarrow \boldsymbol{x}_{2,i} \text{ component of } S_{\text{2LMFVB}} \ ; \ \boldsymbol{\Sigma}_{q(\boldsymbol{u}_i)} \longleftarrow \boldsymbol{A}^{22,i} \text{ component of } S_{\text{2LMFVB}}$$

$$E_q\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{q(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{q(\boldsymbol{u}_i)})^T\} \longleftarrow \boldsymbol{A}^{12,i} \text{ component of } S_{\text{2LMFVB}}$$

$$\boldsymbol{\Lambda}_{q(\boldsymbol{\Sigma})} \longleftarrow \boldsymbol{\Lambda}_{q(\boldsymbol{\Sigma})} + \boldsymbol{\mu}_{q(\boldsymbol{u}_i)}\boldsymbol{\mu}_{q(\boldsymbol{u}_i)}^T + \boldsymbol{\Sigma}_{q(\boldsymbol{u}_i)}$$

$$\boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1})} \longleftarrow (\xi_{q(\boldsymbol{\Sigma})} - q + 1)\,\boldsymbol{\Lambda}_{q(\boldsymbol{\Sigma})}^{-1}$$

$$\boldsymbol{\Lambda}_{q(\boldsymbol{A}_{\boldsymbol{\Sigma}})} \longleftarrow \mathrm{diag}\left\{\mathrm{dg}\left(\boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1})}\right)\right\} + \{\nu_{\boldsymbol{\Sigma}}\mathrm{diag}(s_{\boldsymbol{\Sigma},1}^2, \ldots, s_{\boldsymbol{\Sigma},q}^2)\}^{-1}$$

$$\boldsymbol{M}_{q(\boldsymbol{A}_{\boldsymbol{\Sigma}}^{-1})} \longleftarrow \xi_{q(\boldsymbol{A}_{\boldsymbol{\Sigma}})}\boldsymbol{\Lambda}_{q(\boldsymbol{A}_{\boldsymbol{\Sigma}})}^{-1}$$

until convergence.

Output: $\Big(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}, \left\{\left(\boldsymbol{\mu}_{q(\boldsymbol{u}_i)}, \boldsymbol{\Sigma}_{q(\boldsymbol{u}_i)}, E_q\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{q(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{q(\boldsymbol{u}_i)})^T\}\right) : 1 \leq i \leq m\right\}$, $\xi_{q(\boldsymbol{\Sigma})}, \boldsymbol{\Lambda}_{q(\boldsymbol{\Sigma})}\Big)$.

Algorithm 4.5 (Nolan *et al.*, 2018) *The* TwoLevelNaturalToCommonParameters *algorithm for conversion of a two-level reduced natural parameter vector to its corresponding common parameters.*

Inputs: $p, q, m, \boldsymbol{\eta}_{q(\boldsymbol{\beta},\boldsymbol{u})}$

$\boldsymbol{\omega}_{\text{NCP1}} \longleftarrow$ first p entries of $\boldsymbol{\eta}_{q(\boldsymbol{\beta},\boldsymbol{u})}$

$\boldsymbol{\omega}_{\text{NCP2}} \longleftarrow$ next $p(p+1)$ entries of $\boldsymbol{\eta}_{q(\boldsymbol{\beta},\boldsymbol{u})}$; $\boldsymbol{\Omega}_{\text{NCP3}} \longleftarrow -2\text{vec}^{-1}(\boldsymbol{D}_p^{+T}\boldsymbol{\omega}_{\text{NCP2}})$

$i_{\text{stt}} \longleftarrow p + p(p+1) + 1$; $i_{\text{end}} \longleftarrow i_{\text{stt}} + q - 1$

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

- $\boldsymbol{\omega}_{\text{NCP4}i} \longleftarrow$ sub-vector of $\boldsymbol{\eta}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ with entries i_{stt} to i_{end} inclusive
- $i_{\text{stt}} \longleftarrow i_{\text{end}} + 1$; $i_{\text{end}} \longleftarrow i_{\text{stt}} + q(q+1) - 1$
- $\boldsymbol{\omega}_{\text{NCP5}} \longleftarrow$ sub-vector of $\boldsymbol{\eta}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ with entries i_{stt} to i_{end} inclusive
- $i_{\text{stt}} \longleftarrow i_{\text{end}} + 1$; $i_{\text{end}} \longleftarrow i_{\text{stt}} + pq - 1$
- $\boldsymbol{\omega}_{\text{NCP6}} \longleftarrow$ sub-vector of $\boldsymbol{\eta}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ with entries i_{stt} to i_{end} inclusive
- $i_{\text{stt}} \longleftarrow i_{\text{end}} + 1$; $i_{\text{end}} \longleftarrow i_{\text{stt}} + q - 1$
- $\boldsymbol{\Omega}_{\text{NCP7}i} \longleftarrow -2\,\text{vec}^{-1}(\boldsymbol{D}_q^{+T}\boldsymbol{\omega}_{\text{NCP5}})$; $\boldsymbol{\Omega}_{\text{NCP8}i} \longleftarrow -\text{vec}^{-1}_{p\times q}(\boldsymbol{\omega}_{\text{NCP6}})$

$\mathcal{S}_{\text{NCP}} \longleftarrow$ SolveTwoLevelSparseMatrix$\Big(\boldsymbol{\omega}_{\text{NCP1}}, \boldsymbol{\Omega}_{\text{NCP3}}, \big\{(\boldsymbol{\omega}_{\text{NCP4}i}, \boldsymbol{\Omega}_{\text{NCP7}i}, \boldsymbol{\Omega}_{\text{NCP8}i}) : 1 \leq i \leq m\big\}\Big)$

$\boldsymbol{\mu}_{q(\boldsymbol{\beta})} \longleftarrow \boldsymbol{x}_1$ component of $\mathcal{S}_{\text{NCP}}$; $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} \longleftarrow \boldsymbol{A}^{11}$ component of $\mathcal{S}_{\text{NCP}}$

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

- $\boldsymbol{\mu}_{q(\boldsymbol{u}_i)} \longleftarrow \boldsymbol{x}_{2,i}$ component of $\mathcal{S}_{\text{NCP}}$; $\boldsymbol{\Sigma}_{q(\boldsymbol{u}_i)} \longleftarrow \boldsymbol{A}^{22,i}$ component of $\mathcal{S}_{\text{NCP}}$
- $E_q\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{q(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{q(\boldsymbol{u}_i)})^T\} \longleftarrow \boldsymbol{A}^{12,i}$ component of $\mathcal{S}_{\text{NCP}}$

Output: $\Big(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}, \big\{\big(\boldsymbol{\mu}_{q(\boldsymbol{u}_i)}, \boldsymbol{\Sigma}_{q(\boldsymbol{u}_i)}, E_q\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{q(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{q(\boldsymbol{u}_i)})^T\}\big) : 1 \leq i \leq m\big\}\Big)$

Algorithm 4.6 *The inputs, updates and outputs of the matrix algebraic streamlined Poisson likelihood fragment for two-level models.*

Data Inputs: $\boldsymbol{y}_i$ $(n_i \times 1)$, $\boldsymbol{X}_i$ $(n_i \times p)$, $\boldsymbol{Z}_i$ $(n_i \times q)$, $1 \leq i \leq m$.

Parameter Inputs: $\boldsymbol{\eta}_{p(\boldsymbol{y}\,|\,\boldsymbol{\beta},\boldsymbol{u})\to(\boldsymbol{\beta},\boldsymbol{u})}$, $\boldsymbol{\eta}_{(\boldsymbol{\beta},\boldsymbol{u})\to p(\boldsymbol{y}\,|\,\boldsymbol{\beta},\boldsymbol{u})}$.

Updates:

$$S_{\text{2PVMP}} \longleftarrow \textsc{TwoLevelNaturalToCommonParameters}\left(p, q, m, \boldsymbol{\eta}_{p(\boldsymbol{y}\,|\,\boldsymbol{\beta},\boldsymbol{u})\leftrightarrow(\boldsymbol{\beta},\boldsymbol{u})}\right)$$

$$\boldsymbol{\mu}_{q(\boldsymbol{\beta})} \longleftarrow \boldsymbol{\mu}_{q(\boldsymbol{\beta})} \text{ component of } S_{\text{2PVMP}} \;;\; \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} \longleftarrow \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} \text{ component of } S_{\text{2PVMP}}$$

$$\boldsymbol{\omega}_{\text{2PVMP3}} \longleftarrow \boldsymbol{0}_p \;;\; \boldsymbol{\omega}_{\text{2PVMP4}} \longleftarrow \boldsymbol{0}_{\frac{1}{2}p(p+1)}$$

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

$$\boldsymbol{\mu}_{q(\boldsymbol{u}_i)} \longleftarrow \boldsymbol{\mu}_{q(\boldsymbol{u}_i)} \text{ component of } S_{\text{2PVMP}}$$

$$\boldsymbol{\Sigma}_{q(\boldsymbol{u}_i)} \longleftarrow \boldsymbol{\Sigma}_{q(\boldsymbol{u}_i)} \text{ component of } S_{\text{2PVMP}}$$

$$E_q\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{q(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{q(\boldsymbol{u}_i)})^T\} \longleftarrow E_q\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{q(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{q(\boldsymbol{u}_i)})^T\}$$
$$\text{component of } S_{\text{2PVMP}}$$

$$\boldsymbol{\omega}_{\text{2PVMP1}i} \longleftarrow \exp\Big\{\boldsymbol{X}_i\boldsymbol{\mu}_{q(\boldsymbol{\beta})} + \boldsymbol{Z}_i\boldsymbol{\mu}_{q(\boldsymbol{u}_i)} + \tfrac{1}{2}\text{dg}\Big(\boldsymbol{X}_i\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\boldsymbol{X}_i^T + \boldsymbol{Z}_i\boldsymbol{\Sigma}_{q(\boldsymbol{u}_i)}\boldsymbol{Z}_i^T + 2\boldsymbol{X}_i E_q\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{q(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{q(\boldsymbol{u}_i)})^T\}\boldsymbol{Z}_i^T\Big)\Big\}$$

$$\boldsymbol{\omega}_{\text{2PVMP2}i} \longleftarrow \boldsymbol{y}_i - \boldsymbol{\omega}_{\text{2PVMP1}i} \odot \left(\boldsymbol{1}_{n_i} - \boldsymbol{X}_i\boldsymbol{\mu}_{q(\boldsymbol{\beta})} + \boldsymbol{Z}_i\boldsymbol{\mu}_{q(\boldsymbol{u}_i)}\right)$$

$$\boldsymbol{\omega}_{\text{2PVMP3}} \longleftarrow \boldsymbol{\omega}_{\text{2PVMP3}} + \boldsymbol{X}_i^T\boldsymbol{\omega}_{\text{2PVMP2}i}$$

$$\boldsymbol{\omega}_{\text{2PVMP4}} \longleftarrow \boldsymbol{\omega}_{\text{2PVMP4}} - \tfrac{1}{2}\boldsymbol{D}_p^T\text{vec}\left(\boldsymbol{X}_i^T\,\text{diag}\left(\boldsymbol{\omega}_{\text{2PVMP1}i}\right)\boldsymbol{X}_i\right)$$

$$\boldsymbol{\eta}_{p(\boldsymbol{y}\,|\,\boldsymbol{\beta},\boldsymbol{u})\to(\boldsymbol{\beta},\boldsymbol{u})} \longleftarrow \begin{bmatrix} \boldsymbol{\omega}_{\text{2PVMP3}} \\ \boldsymbol{\omega}_{\text{2PVMP4}} \\ \underset{1\leq i\leq m}{\text{stack}} \begin{bmatrix} \boldsymbol{Z}_i^T\boldsymbol{\omega}_{\text{2PVMP2}i} \\ -\frac{1}{2}\boldsymbol{D}_q^T\text{vec}\left(\boldsymbol{Z}_i^T\,\text{diag}\left(\boldsymbol{\omega}_{\text{2PVMP1}i}\right)\boldsymbol{Z}_i\right) \\ -\text{vec}\left(\boldsymbol{X}_i^T\,\text{diag}\left(\boldsymbol{\omega}_{\text{2PVMP1}i}\right)\boldsymbol{Z}_i\right) \end{bmatrix} \end{bmatrix}.$$

Parameter Output: $\boldsymbol{\eta}_{p(\boldsymbol{y}\,|\,\boldsymbol{\beta},\boldsymbol{u})\to(\boldsymbol{\beta},\boldsymbol{u})}$.

Algorithm 4.7 *The inputs, updates and outputs of the matrix algebraic streamlined logistic likelihood fragment for two-level models.*

Data Inputs: $\boldsymbol{y}_i$ $(n_i \times 1)$, $\boldsymbol{X}_i$ $(n_i \times p)$, $\boldsymbol{Z}_i$ $(n_i \times q)$, $1 \leq i \leq m$.

Constant Inputs: $\boldsymbol{p}$, $\boldsymbol{s}$

Parameter Inputs: $\boldsymbol{\eta}_{p(\boldsymbol{y} \mid \boldsymbol{\beta},\boldsymbol{u}) \to (\boldsymbol{\beta},\boldsymbol{u})}$, $\boldsymbol{\eta}_{(\boldsymbol{\beta},\boldsymbol{u}) \to p(\boldsymbol{y} \mid \boldsymbol{\beta},\boldsymbol{u})}$

Updates:

$$S_{\text{2LVMP}} \longleftarrow \textsc{TwoLevelNaturalToCommonParameters}\left(p, q, m, \boldsymbol{\eta}_{p(\boldsymbol{y} \mid \boldsymbol{\beta},\boldsymbol{u}) \leftrightarrow (\boldsymbol{\beta},\boldsymbol{u})}\right)$$

$$\boldsymbol{\mu}_{q(\boldsymbol{\beta})} \longleftarrow \boldsymbol{\mu}_{q(\boldsymbol{\beta})} \text{ component of } S_{\text{2LVMP}} \;\; ; \;\; \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} \longleftarrow \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} \text{ component of } S_{\text{2LVMP}}$$

$$\boldsymbol{\omega}_{\text{2LVMP4}} \longleftarrow \boldsymbol{0}_p \;\; ; \;\; \boldsymbol{\omega}_{\text{2LVMP5}} \longleftarrow \boldsymbol{0}_{\frac{1}{2}p(p+1)}$$

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

$$\boldsymbol{\mu}_{q(\boldsymbol{u}_i)} \longleftarrow \boldsymbol{\mu}_{q(\boldsymbol{u}_i)} \text{ component of } S_{\text{2LVMP}}$$

$$\boldsymbol{\Sigma}_{q(\boldsymbol{u}_i)} \longleftarrow \boldsymbol{\Sigma}_{q(\boldsymbol{u}_i)} \text{ component of } S_{\text{2LVMP}}$$

$$E_q\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{q(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{q(\boldsymbol{u}_i)})^T\} \longleftarrow E_q\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{q(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{q(\boldsymbol{u}_i)})^T\} \text{ component of } S_{\text{2LVMP}}$$

$$\boldsymbol{\mu} \longleftarrow \boldsymbol{X}_i\boldsymbol{\mu}_{q(\boldsymbol{\beta})} + \boldsymbol{Z}_i\boldsymbol{\mu}_{q(\boldsymbol{u}_i)}$$

$$\boldsymbol{\sigma}^2 \longleftarrow \operatorname{dg}\left(\boldsymbol{X}_i\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\boldsymbol{X}_i^T + 2\boldsymbol{X}_i E_q\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{q(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{q(\boldsymbol{u}_i)})^T\}\boldsymbol{Z}_i^T + \boldsymbol{Z}_i\boldsymbol{\Sigma}_{q(\boldsymbol{u}_i)}\boldsymbol{Z}_i^T\right)$$

$$\boldsymbol{T} \longleftarrow \sqrt{\boldsymbol{1}_{n_i}\boldsymbol{1}_8^T + (\boldsymbol{\sigma}^2)\,(\boldsymbol{s}^2)^T}$$

$$\boldsymbol{\omega}_{\text{2LVMP1}i} \longleftarrow \Phi\left(\left(\boldsymbol{\mu}\boldsymbol{s}^T\right)/\boldsymbol{T}\right)\boldsymbol{p} \;\; ; \;\; \boldsymbol{\omega}_{\text{2LVMP2}i} \longleftarrow \left\{\phi\left(\left(\boldsymbol{\mu}\boldsymbol{s}^T\right)/\boldsymbol{T}\right)/\boldsymbol{T}\right\}(\boldsymbol{p} \odot \boldsymbol{s})$$

$$\boldsymbol{\omega}_{\text{2LVMP3}i} \longleftarrow \boldsymbol{y}_i - \boldsymbol{\omega}_{\text{2LVMP1}i} + \boldsymbol{\omega}_{\text{2LVMP2}i} \odot \boldsymbol{\mu} \;\; ; \;\; \boldsymbol{\omega}_{\text{2LVMP4}} \longleftarrow \boldsymbol{\omega}_{\text{2LVMP4}} + \boldsymbol{X}_i^T\boldsymbol{\omega}_{\text{2LVMP3}i}$$

$$\boldsymbol{\omega}_{\text{2LVMP5}} \longleftarrow \boldsymbol{\omega}_{\text{2LVMP5}} - \tfrac{1}{2}\boldsymbol{D}_p^T\operatorname{vec}\left(\boldsymbol{X}_i^T\operatorname{diag}\left(\boldsymbol{\omega}_{\text{2LVMP2}i}\right)\boldsymbol{X}_i\right)$$

$$\boldsymbol{\eta}_{p(\boldsymbol{y} \mid \boldsymbol{\beta},\boldsymbol{u}) \to (\boldsymbol{\beta},\boldsymbol{u})} \longleftarrow \begin{bmatrix} \boldsymbol{\omega}_{\text{2LVMP4}} \\ \boldsymbol{\omega}_{\text{2LVMP5}} \\ \underset{1 \leq i \leq m}{\operatorname{stack}} \begin{bmatrix} \boldsymbol{Z}_i^T\boldsymbol{\omega}_{\text{2LVMP3}i} \\ -\frac{1}{2}\boldsymbol{D}_q^T\operatorname{vec}\left(\boldsymbol{Z}_i^T\operatorname{diag}\left(\boldsymbol{\omega}_{\text{2LVMP2}i}\right)\boldsymbol{Z}_i\right) \\ -\operatorname{vec}\left(\boldsymbol{X}_i^T\operatorname{diag}\left(\boldsymbol{\omega}_{\text{2LVMP2}i}\right)\boldsymbol{Z}_i\right) \end{bmatrix} \end{bmatrix}.$$

Parameter Output: $\boldsymbol{\eta}_{p(\boldsymbol{y} \mid \boldsymbol{\beta},\boldsymbol{u}) \to (\boldsymbol{\beta},\boldsymbol{u})}$.

Conclusions and future directions

Discussion

In this thesis we applied variational inference techniques for frequentist and Bayesian fitting and inference in a variety of models. Even though the numerical studies considered here did not involve very large datasets, the variational inference framework offers the opportunity to develop many fast algorithms that can be applied to datasets whose storage requirements exceed what is available on a standard computer.

In frequentist settings, we worked with Gaussian density families to approximate, in terms of Kullback–Leibler divergence, the distribution of random effects vectors, given the output (e.g. Ormerod and Wand, 2012). An assortment of models contextualized in a general design GLMMs infrastructure were covered. What we presented is a new framework for inference in cases that can be treated through GLMs with random intercept and slope and models with additive non-linear components or spatial correlation structures, for example. We showed that GVA performs similarly to and sometimes better than currently available software for fitting the considered models.

From a Bayesian perspective, we extended recent achievements concerning mean field variational Bayes algorithms, adopting the approach of variational message passing on factor graph fragment. Three elaborate likelihood fragments, including Pareto random samples, support vector regression and skew t regression, were explored. We also presented explicit algorithms to perform streamlined computing variational inference for GLMMs containing two-level random effects. Recent results in Nolan *et al.* (2018) include streamlined MFVB and VMP algorithms for Gaussian response two-level and three-level models. Our extensions now cover Poisson and logistic two-level mixed models. If compared to standard MCMC methods, variational techniques for Bayesian inference may suffer some accuracy loss. In several instances, variational mean field approximations can be shown to underestimate posterior variances (e.g. Wang and Blei, 2018). Another type of loss, in terms of bias, may be the consequence of simple auxiliary variable representations of the likelihood fragment, associated with a convenient factorization of the approximating density, as shown for the skew t likelihood fragment.

However, putting accuracy evidences apart, mean field approximations can always be used to complement more traditional MCMC methods or other methods which tend to asymptotically recover the actual posterior, as well as to quickly explore the data.

Future directions of research

Variational inference is prone to the exploration of several promising research directions, especially in a context in which datasets are continuously growing in size, while modern applications require inference tools with reduced computational times.

In particular, the interesting performances in terms of accuracy of GVA motivate further exploration of such a technique applied to GLMMs. Efficient and general optimization procedures have to be considered, if the intention is to provide an effective inference tool which can be implemented in standard computing environments. In parallel, reasonable parametrizations of the approximating Gaussian density covariance matrix are fundamental for efficient computational strategies, especially for large semi-parametric regression models. Also for the frequentist case, sparsity may be a key to fast and robust optimization. Useful insights are described in Tan and Nott (2018), who take advantage of the conditional independence structure which may characterize a model to allow parsimonious parametrization of precision matrices.

The research on mean field algorithms can be now oriented to widen the class of streamlined variational algorithms. General results allowing the treatment of higher than three-level random effects models are desirable. Once sparse matrix-type refinements of fragment updates relevant to prominent models are obtained, software implementation in low-level programming language will be necessary, also to diffuse these useful results, however lacking of statistical glamour.

Future work can move beyond mere algorithm derivations and application of variational inference tools for fast approximate inference. Indeed, while a comparison with a MCMC benchmark highlights some guarantees of accuracy, Gaussian variational methods lack of general theoretical results. The fascinating proves and results exposed in Hall *et al.* (2011b), for instance, motivate further research endeavor.

On the other side, the application of frequentist results on model misspecification (Pace and Salvan, 1997) may be explored to address the problem of variance underestimation in mean field variational applications. Such a perspective has not been considered in the literature yet but would be worthwhile to investigate in the future. Intuitively, variational versions of sandwich estimators and sandwich asymptotic variance can be proposed and derived from the lower bound expressions.

Appendix A

A.1 Vector differential calculus

Let f be a scalar-valued function with argument $\boldsymbol{x} \in \mathbb{R}^d$. The *derivative vector* of f, $\mathrm{D}f(\boldsymbol{x})$, is the $1 \times d$ vector whose ith entry is

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

The *Hessian matrix* of f is

$$\mathrm{H}f(\boldsymbol{x}) = \mathrm{D}\left\{\mathrm{D}f(\boldsymbol{x})\right\}^T$$

and is, alternatively, the $d \times d$ matrix with (i, j) entry equal to

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

Useful related results are provided in Magnus and Neudecker (2007) and Wand (2002).

A.2 Distributions and special functions

Here we describe the distribution functions mentioned in this PhD thesis and include in each subsection the relevant results. The Inverse Square Root Nadarajah, Moon Rock and Sea Sponge distributions are defined in the supplementary material of McLean and Wand (2018). The remaining distributions follow the convention of Wand (2017) and related supplementary material, when present. For the exponential family densities appearing in the derivation of variational algorithms further results such as the expectation of the sufficient statistic are included.

A.2.1 Exponential families

A univariate exponential family density or probability mass function can be written as

$$p(x) = \exp\left\{\boldsymbol{T}(x)^T \boldsymbol{\eta} - A(\boldsymbol{\eta})\right\} h(x)$$

where $\boldsymbol{T}(x)$ is the *sufficient statistic*, $\boldsymbol{\eta}$ is the *natural parameter*, $A(\boldsymbol{\eta})$ is the *log-partition function* and $h(x)$ is the *base measure*. The sufficient statistic is not unique but it is commonly chosen to be the simplest algebraic form given $p(x)$.
The following results link the sufficient statistic and log-partition function:

$$E\{\boldsymbol{T}(x)\} = \mathrm{D}_{\boldsymbol{\eta}} A(\boldsymbol{\eta})^T \quad \text{and} \quad \mathrm{Cov}\{\boldsymbol{T}(x)\} = \mathrm{D}_{\boldsymbol{\eta}}\left\{\mathrm{D}_{\boldsymbol{\eta}} A(\boldsymbol{\eta})^T\right\}$$

where $\mathrm{Cov}\{\boldsymbol{T}(x)\}$ is the covariance matrix of $\boldsymbol{T}(x)$. Similar results hold for multivariate distributions.
The expression involving the expectation of the sufficient statistics is a relevant result for variational message passing since the messages from factors to stochastic nodes of conjugate factor graphs reduce to sufficient statistic expectations.
In frequentist settings, alternative notation in use in this thesis for one-parameter exponential families is

$$p(x) = \exp\{\eta T(x) - b(\eta) + c(x)\},$$

where $e^{c(x)}$ corresponds to the base measure and $b(x)$ is the log-partition.

A.2.2 Digamma function

The digamma function ψ is

$$\psi(x) = \frac{d}{dx}\log\Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}.$$

Evaluation of ψ is supported, for instance, in the `R` computing environment (R Core Team, 2018) via the function `digamma()`.

A.2.3 Modified Bessel functions of the second kind

The *Modified Bessel function of the second kind* of order $p \in \mathbb{R}$ is denoted by K_p and its argument can be and arbitrary complex number. Restricting attention to real positive arguments, the modified Bessel functions of the second kind have the integral

representation

$$K_p(x) = \frac{\Gamma(|p| + 1/2)(2x)^{|p|}}{\sqrt{\pi}} \int_0^\infty \frac{\cos(t)}{(x^2 + t^2)^{|p|+1/2}} dt, \quad x > 0.$$

Evaluation of $K_p(x)$ for $p \in \mathbb{R}$ and $x > 0$ is supported, for instance, in the `R` computing environment (R Core Team, 2018) via the function `besselK(x,p)`, where `p` and `x` denote the values of p and x respectively. If $p = (1/2)(2k+1)$ for some $k \in \mathbb{Z}$ then K_p admits explicit expressions. For our purposes, a useful formula is expression S.1.3 in the supplementary material of McLean and Wand (2018):

$$\frac{K_{3/2}(x)}{K_{1/2}(x)} = 1 + \frac{1}{x}, \quad x > 0. \tag{A.1}$$

Further details about efficient computation of modified Bessel functions of the second kind and useful related formulae are provided in Section S.1.1 of the supplementary material of McLean and Wand (2018).

A.2.4 Parabolic cylinder functions

The *parabolic cylinder function* of order $\nu \in \mathbb{R}$ is denoted by D_ν. If of *negative order* it can be expressed in terms of a simple integral:

$$D_\nu(x) = \Gamma(-\nu)^{-1} \exp\left(-x^2/4\right) \int_0^\infty t^{-\nu-1} \exp\left(-xt - \frac{1}{2}t^2\right) dt, \quad \nu < 0, \quad x \in \mathbb{R}.$$

Further details about efficient computation of parabolic cylinder functions and useful related formulae are provided in Section S.1.2 of the supplementary material of McLean and Wand (2018).

Distributions with exponential family theory results

A.2.5 Univariate normal distribution

A random variable x has a *univariate normal distribution* with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$, written $x \sim N(\mu, \sigma^2)$, if its density function is

$$p(x) = \left(2\pi\sigma^2\right)^{-1/2} \exp\left\{-(x-\mu)^2 / \left(2\sigma^2\right)\right\}, \quad x \in \mathbb{R}.$$

The sufficient statistic and base measure are

$$\boldsymbol{T}(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix} \quad \text{and} \quad h(x) = (2\pi)^{-1/2}.$$

The natural parameter vector and its inverse mapping are

$$\boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} \mu/\sigma^2 \\ -1/(2\sigma^2) \end{bmatrix}, \quad \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} = \begin{bmatrix} -\eta_1/(2\eta_2) \\ -1/(2\eta_2) \end{bmatrix}$$

and the log-partition function is

$$A(\boldsymbol{\eta}) = -\frac{1}{4}\left(\eta_1^2/\eta_2\right) - \frac{1}{2}\log(-2\eta_2).$$

The expectation of the sufficient statistic is

$$E\{\boldsymbol{T}(x)\} = \begin{bmatrix} -\eta_1/(2\eta_2) \\ (\eta_1^2 - 2\eta_2)/(4\eta_2^2) \end{bmatrix}.$$

A.2.6 Gamma, chi-squared and exponential distributions

A random variable x has a *gamma distribution* with shape $\alpha > 0$ and scale $\beta > 0$, written $x \sim \Gamma(\alpha, \beta)$, if the density function of x is

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

The sufficient statistic and base measure are

$$\boldsymbol{T}(x) = \begin{bmatrix} \log(x) \\ x \end{bmatrix} \quad \text{and} \quad h(x) = 1.$$

The natural parameter vector and its inverse mapping are

$$\boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} \alpha - 1 \\ -\beta \end{bmatrix}, \quad \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \eta_1 + 1 \\ -\eta_2 \end{bmatrix}$$

and the log-partition function is

$$A(\boldsymbol{\eta}) = \log\{\Gamma(\eta_1 + 1)\} - (\eta_1 + 1)\log(-\eta_2).$$

The expectation of the sufficient statistic is

$$E\left\{\boldsymbol{T}\left(x\right)\right\} = \left[\begin{array}{c} \psi\left(\eta_1+1\right)-\log\left(-\eta_2\right) \\ -\left(\eta_1+1\right)/\eta_2. \end{array}\right]. \tag{A.2}$$

A random variable x has a *chi-squared distribution* with degrees of freedom $\nu \in \mathbb{N}_{>0}$, written $x \sim \chi^2\left(\nu\right)$, if the density function of x is

$$p\left(x\right) = 2^{-\nu/2}\Gamma\left(\nu/2\right)^{-1} x^{\nu/2-1}\exp\left(-x/2\right), \quad x \geq 0$$

With a reparametrization, we obtain

$$x \sim \chi^2\left(\nu\right) \quad \text{if and only if} \quad x \sim \Gamma\left(\nu/2, 1/2\right).$$

A random variable x has an *exponential distribution* with rate $\lambda > 0$, written $x \sim \text{Exp}\left(\lambda\right)$, if the density function of x is

$$p\left(x\right) = \lambda\exp(-\lambda x), \quad x \geq 0.$$

With a reparametrization, we obtain

$$x \sim \text{Exp}\left(\lambda\right) \quad \text{if and only if} \quad x \sim \Gamma\left(1, \lambda^{-1}\right).$$

A.2.7 Inverse chi-squared and inverse gamma distributions

A random variable x has an *inverse chi-squared distribution* with shape parameter $\kappa > 0$ and scale parameter $\lambda > 0$, written $x \sim \text{Inverse-}\chi^2\left(\kappa, \lambda\right)$, if the density function of x is

$$p\left(x\right) = \left\{\left(\lambda/2\right)^{\kappa/2}/\Gamma\left(\kappa/2\right)\right\} x^{-(\kappa/2)-1}\exp\left\{-\left(\lambda/2\right)/x\right\}, \quad x > 0.$$

A random variable x has an *inverse gamma distribution* with shape parameter $\tilde{\kappa} > 0$ and scale parameter $\tilde{\lambda} > 0$, written $x \sim \text{Inverse-Gamma}\left(\tilde{\kappa}, \tilde{\lambda}\right)$, if the density function of x is

$$p\left(x\right) = \left\{\tilde{\kappa}^{\tilde{\lambda}}/\Gamma\left(\tilde{\kappa}\right)\right\} x^{-\tilde{\kappa}-1}\exp\left\{-\tilde{\lambda}/x\right\}, \quad x > 0.$$

With a reparametrization, we obtain

$$x \sim \text{Inverse -}\chi^2\left(\kappa, \lambda\right) \quad \text{if and only if} \quad x \sim \text{Inverse-Gamma}\left(\kappa/2, \lambda/2\right).$$

The sufficient statistic and base measure are

$$\boldsymbol{T}(x) = \begin{bmatrix} \log(x) \\ 1/x \end{bmatrix} \quad \text{and} \quad h(x) = I(x > 0).$$

The natural parameter vector and its inverse mappings are

$$\boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} -\frac{1}{2}(\kappa + 2) \\ -\frac{1}{2}\lambda \end{bmatrix} = \begin{bmatrix} -(\tilde{\kappa} + 1) \\ -\tilde{\lambda} \end{bmatrix},$$
$$\begin{bmatrix} \kappa \\ \lambda \end{bmatrix} = \begin{bmatrix} -2 - 2\eta_1 \\ -2\eta_2 \end{bmatrix}, \quad \begin{bmatrix} \tilde{\kappa} \\ \tilde{\lambda} \end{bmatrix} = \begin{bmatrix} -1 - \eta_1 \\ -\eta_2 \end{bmatrix}$$

and the log-partition function is

$$A(\boldsymbol{\eta}) = (\eta_1 + 1)\log(-\eta_2) + \log\Gamma(-\eta_1 - 1).$$

The expectation of the sufficient statistic is

$$E\{\boldsymbol{T}(x)\} = \begin{bmatrix} \log(-\eta_2) - \psi(-\eta_1 - 1) \\ (\eta_1 + 1)/\eta_2 \end{bmatrix}.$$

Note also that the *inverse Wishart distribution* for random matrices described in Subsection A.2.13 reduces to the inverse chi-squared distribution in the 1×1 case.

A.2.8 Generalized inverse Gaussian distribution

A random variable x has an *generalized inverse Gaussian distribution* with parameters $\alpha, \beta > 0$, for any fixed $p \in \mathbb{R}$, written $x \sim \text{Generalized-Inverse-Gaussian}(\alpha, \beta; p)$, if its density function is

$$p(x) = \frac{(\alpha/\beta)^{p/2} x^{p-1}}{2K_p\left(\sqrt{\alpha\beta}\right)} \exp\left\{-\frac{1}{2}(\alpha x + \beta/x)\right\}, \quad x > 0,$$

where K_p is the modified Bessel function of second kind. The sufficient statistic and base measure are

$$\boldsymbol{T}(x) = \begin{bmatrix} x \\ 1/x \end{bmatrix} \quad \text{and} \quad h(x) = \frac{1}{2}x^{p-1} I(x > 0).$$

The natural parameter vector and its inverse mappings are

$$\boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} -\alpha/2 \\ -\beta/2 \end{bmatrix}, \quad \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} -2\eta_1 \\ -2\eta_2 \end{bmatrix}$$

and the log-partition function is

$$A(\boldsymbol{\eta}) = \frac{1}{2} p \log(\eta_1/\eta_2) - \log K_p\left(2(\eta_1\eta_2)^{1/2}\right).$$

The expectation of the sufficient statistic is

$$E\{\boldsymbol{T}(x)\} = \begin{bmatrix} \frac{(\eta_2/\eta_1)^{1/2} K_{p+1}\left(2(\eta_1\eta_2)^{1/2}\right)}{K_p\left(2(\eta_1\eta_2)^{1/2}\right)} \\ \frac{(\eta_1/\eta_2)^{1/2} K_{p+1}\left(2(\eta_1\eta_2)^{1/2}\right)}{K_p\left(2(\eta_1\eta_2)^{1/2}\right)} + \frac{p}{\eta_2} \end{bmatrix}.$$

Applying A.1, it follows that for the special case $p = 1/2$

$$E\{\boldsymbol{T}(x)\} = \begin{bmatrix} \{\eta_1/(2\eta_2)\}^{1/2} - 1/(2\eta_2) \\ (\eta_1/\eta_2)^{1/2} \end{bmatrix}. \tag{A.3}$$

A.2.9 Inverse square root Nadarajah distribution

A random variable x has an *inverse square root Nadarajah distribution* with parameters $\alpha, \beta > 0$ and $\gamma \in \mathbb{R}$, written $x \sim$ Inverse-Square-Root-Nadarajah(α, β, γ), if the corresponding density function is

$$p(x) = (2\beta)^{\alpha/2} / \left[2\exp\left\{\gamma^2/(8\beta)\right\}\Gamma(\alpha) D_{-\alpha}\left(\gamma/\sqrt{2\beta}\right)\right] x^{-(\alpha/2)-1} \exp\left(-\beta/x - \gamma/\sqrt{x}\right),$$

$x > 0$.

The sufficient statistic and base measure are

$$\boldsymbol{T}(x) = \begin{bmatrix} \log(x) \\ 1/\sqrt{x} \\ 1/x \end{bmatrix} \quad \text{and} \quad h(x) = I(x > 0).$$

The natural parameter vector and its inverse mapping are

$$\boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix} = \begin{bmatrix} -(\alpha/2) - 1 \\ -\gamma \\ -\beta \end{bmatrix}, \quad \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} = \begin{bmatrix} -2(\eta_1 + 1) \\ -\eta_3 \\ -\eta_2 \end{bmatrix}$$

and the log-partition function is

$$A(\boldsymbol{\eta}) = -\frac{1}{2}(\eta_1+1)\log(-2\eta_3) - \log(2) - \frac{1}{8}\left(\eta_2^2/\eta_3\right) + \log\{\Gamma(\eta_1+1)\} + \log\left\{D_{-\eta_1-1}\left(-\eta_2/\sqrt{-2\eta_3}\right)\right\}.$$

The expectation of the sufficient statistic is

$$E\{\boldsymbol{T}(x)\} = \begin{bmatrix} \int_0^\infty \log(x)\, x^{\eta_1} \exp(\eta_2/\sqrt{x} + \eta_3/x)\, dx \\ \frac{-2(\eta_1+1)D_{2\eta_1+1}\left(-\eta_2/\sqrt{-2\eta_3}\right)}{\sqrt{-2\eta_3}D_{2\eta_1+2}\left(-\eta_2/\sqrt{-2\eta_3}\right)} \\ \frac{-(\eta_1+1)(2\eta_1+1)D_{2\eta_1+1}\left(-\eta_2/\sqrt{-2\eta_3}\right)}{\eta_3 D_{2\eta_1+2}\left(-\eta_2/\sqrt{-2\eta_3}\right)} \end{bmatrix},$$

from which we define the notation

$$(E\boldsymbol{T})_2^{\text{ISRN}} = \frac{-2(\eta_1+1)D_{2\eta_1+1}\left(-\eta_2/\sqrt{-2\eta_3}\right)}{\sqrt{-2\eta_3}D_{2\eta_1+2}\left(-\eta_2/\sqrt{-2\eta_3}\right)}, \tag{A.4}$$

$$(E\boldsymbol{T})_3^{\text{ISRN}} = \frac{-(\eta_1+1)(2\eta_1+1)D_{2\eta_1+1}\left(-\eta_2/\sqrt{-2\eta_3}\right)}{\eta_3 D_{2\eta_1+2}\left(-\eta_2/\sqrt{-2\eta_3}\right)}. \tag{A.5}$$

Section S.1.2 in the supplementary material of McLean and Wand (2018) suggests how to perform stable and efficient computation of $(E\boldsymbol{T})_2^{\text{ISRN}}$ and $(E\boldsymbol{T})_3^{\text{ISRN}}$.

A.2.10 Moon Rock distribution

A random variable x has a *Moon Rock distribution* with parameters $\alpha > 0$ and $\beta > \alpha$, written $x \sim \text{Moon-Rook}(\alpha, \beta)$, if the density function of x is

$$p(x) = \left[\int_0^\infty \left\{t^t/\Gamma(t)\right\}^\alpha \exp(-\beta t)\, dt\right]^{-1} \left\{x^x/\Gamma(x)\right\}^\alpha \exp(-\beta x), \quad x > 0.$$

The sufficient statistic and base measure are

$$\boldsymbol{T}(x) = \begin{bmatrix} x\log(x) - \log\{\Gamma(x)\} \\ x \end{bmatrix} \quad \text{and} \quad h(x) = I(x > 0).$$

The natural parameter vector and its inverse mapping are

$$\boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} \alpha \\ -\beta \end{bmatrix}, \quad \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \eta_1 \\ -\eta_2 \end{bmatrix}$$

and the log-partition function is

$$A(\boldsymbol{\eta}) = \log\left\{\int_0^\infty \left\{t^t/\Gamma(t)\right\}^{\eta_1} \exp(\eta_2 t)\, dt\right\}.$$

The expectation of the sufficient statistic is

$$E\left\{\boldsymbol{T}(x)\right\} = \exp\left\{-A(\boldsymbol{\eta})\right\} \begin{bmatrix} \int_0^\infty \left[x\log(x) - \log\left\{\Gamma(x)\right\}\right] \left\{x^x/\Gamma(x)\right\}^{\eta_1} \exp(\eta_2 x)\, dx \\ \int_0^\infty x \left\{x^x/\Gamma(x)\right\}^{\eta_1} \exp(\eta_2 x)\, dx \end{bmatrix},$$

from which we define the notation

$$(E\boldsymbol{T})_2^{\text{MR}} = \exp\left\{-A(\boldsymbol{\eta})\right\} \int_0^\infty x \left\{x^x/\Gamma(x)\right\}^{\eta_1} \exp(\eta_2 x)\, dx. \tag{A.6}$$

The integrals here appearing are not expressible in terms of established special functions. See sections S.1.3 and S.2.4 in the supplementary material of McLean and Wand (2018) for further details about stable and efficient integral computation.

A.2.11 Sea Sponge distribution

The random variable x has a *Sea Sponge distribution* with parameters $\alpha > 0$, $\beta > 0$ and $|\gamma| < \beta$, written $x \sim \text{Sea-Sponge}(\alpha, \beta, \gamma)$, if the density function of x is

$$\begin{aligned} p(x) = &\left\{\int_{-\infty}^\infty \left(1+t^2\right)^\alpha \exp\left(-\beta t^2 + \gamma t\sqrt{1+t^2}\right) dt\right\}^{-1} \left(1+x^2\right)^\alpha \\ &\times \exp\left(-\beta x^2 + \gamma x\sqrt{1+x^2}\right). \end{aligned}$$

The sufficient statistic and base measure are

$$\boldsymbol{T}(x) = \begin{bmatrix} \log\left(1+x^2\right) \\ x^2 \\ x\sqrt{1+x^2} \end{bmatrix} \quad \text{and} \quad h(x) = 1.$$

The natural parameter vector and its inverse mapping are

$$\boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix} = \begin{bmatrix} \alpha \\ -\beta \\ \gamma \end{bmatrix}, \quad \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} = \begin{bmatrix} \eta_1 \\ -\eta_2 \\ \eta_3 \end{bmatrix}$$

and the log-partition function is

$$A(\boldsymbol{\eta}) = \log\left\{\int_{-\infty}^{\infty}\left(1+t^2\right)^{\eta_1}\exp\left(\eta_2 t^2+\eta_3 t\sqrt{1+t^2}\right)dt\right\}.$$

The expectation of the sufficient statistic is

$$E\left\{\boldsymbol{T}(x)\right\} = \exp\left\{-A(\boldsymbol{\eta})\right\}\begin{bmatrix}\int_{-\infty}^{\infty}\log\left(1+x^2\right)\left(1+x^2\right)^{\eta_1}\exp\left(\eta_2 x^2+\eta_3 x\sqrt{1+x^2}\right)dx\\ \int_{-\infty}^{\infty}x^2\left(1+x^2\right)^{\eta_1}\exp\left(\eta_2 x^2+\eta_3 x\sqrt{1+x^2}\right)dx\\ \int_{-\infty}^{\infty}x\sqrt{1+x^2}\left(1+x^2\right)^{\eta_1}\exp\left(\eta_2 x^2+\eta_3 x\sqrt{1+x^2}\right)dx\end{bmatrix},$$

from which we define the notation

$$(E\boldsymbol{T})_2^{\text{SS}} = \exp\left\{-A(\boldsymbol{\eta})\right\}\int_{-\infty}^{\infty}x^2\left(1+x^2\right)^{\eta_1}\exp\left(\eta_2 x^2+\eta_3 x\sqrt{1+x^2}\right)dx, \tag{A.7}$$

$$(E\boldsymbol{T})_3^{\text{SS}} = \exp\left\{-A(\boldsymbol{\eta})\right\}\int_{-\infty}^{\infty}x\sqrt{1+x^2}\left(1+x^2\right)^{\eta_1}\exp\left(\eta_2 x^2+\eta_3 x\sqrt{1+x^2}\right)dx. \tag{A.8}$$

The integrals here appearing are not expressible in terms of established special functions. See sections S.1.3 and S.2.5 in the supplementary material of McLean and Wand (2018) for further details about stable and efficient integral computation.

A.2.12 Multivariate normal distribution

A $d \times 1$ random vector $\boldsymbol{x}$ has a *multivariate normal distribution* with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, a symmetric positive definite $d \times d$ matrix, written $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if its density function is

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

The sufficient statistic and base measure are

$$\boldsymbol{T}(\boldsymbol{x}) = \begin{bmatrix}\boldsymbol{x}\\ \text{vec}\left(\boldsymbol{x}\boldsymbol{x}^T\right)\end{bmatrix} \quad \text{and} \quad h(\boldsymbol{x}) = (2\pi)^{-d/2}.$$

The natural parameter vector and inverse mapping are

$$\boldsymbol{\eta} = \begin{bmatrix}\boldsymbol{\eta}_1\\ \boldsymbol{\eta}_2\end{bmatrix} = \begin{bmatrix}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\\ -\frac{1}{2}\text{vec}\left(\boldsymbol{\Sigma}^{-1}\right)\end{bmatrix}, \quad \begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\Sigma}\end{bmatrix} = \begin{bmatrix}-\frac{1}{2}\left\{\text{vec}^{-1}(\boldsymbol{\eta_2})\right\}^{-1}\boldsymbol{\eta_1}\\ -\frac{1}{2}\left\{\text{vec}^{-1}(\boldsymbol{\eta_2})\right\}^{-1}\end{bmatrix} \tag{A.9}$$

and the log-partition function is

$$A(\boldsymbol{\eta}) = -\frac{1}{4}\boldsymbol{\eta_1}^T \left\{\text{vec}^{-1}(\boldsymbol{\eta_2})\right\}^{-1} \boldsymbol{\eta}_1 - \frac{1}{2}\log\left|-2\text{vec}^{-1}(\boldsymbol{\eta}_2)\right|.$$

The expectation of the sufficient statistic is

$$E\{\boldsymbol{T}(\boldsymbol{x})\} = \begin{bmatrix} -\frac{1}{2}\left\{\text{vec}^{-1}(\boldsymbol{\eta}_2)\right\}^{-1}\boldsymbol{\eta}_1 \\ \frac{1}{4}\text{vec}\left(\left\{\text{vec}^{-1}(\boldsymbol{\eta}_2)\right\}^{-1}\right)\left[\boldsymbol{\eta}_1\boldsymbol{\eta}_1^T\left\{\text{vec}^{-1}(\boldsymbol{\eta}_2)\right\}^{-1} - 2\boldsymbol{I}\right] \end{bmatrix}.$$

Reduced expressions

The sufficient statistic can be written in a more efficient way as

$$\boldsymbol{T}(\boldsymbol{x}) = \begin{bmatrix} \boldsymbol{x} \\ \text{vech}\left(\boldsymbol{x}\boldsymbol{x}^T\right) \end{bmatrix}.$$

More compact expressions for the natural parameter vector ad inverse mapping are

$$\boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\boldsymbol{D}_d^T\text{vec}\left(\boldsymbol{\Sigma}^{-1}\right) \end{bmatrix}, \quad \begin{bmatrix} -\frac{1}{2}\left\{\text{vec}^{-1}\left(\boldsymbol{D}_d^{+T}\boldsymbol{\eta_2}\right)\right\}^{-1}\boldsymbol{\eta_1} \\ -\frac{1}{2}\left\{\text{vec}^{-1}\left(\boldsymbol{D}_d^{+T}\boldsymbol{\eta_2}\right)\right\}^{-1} \end{bmatrix} \tag{A.10}$$

and for the log-partition function is

$$A(\boldsymbol{\eta}) = -\frac{1}{4}\boldsymbol{\eta_1}^T \left\{\text{vec}^{-1}\left(\boldsymbol{D}_d^{+T}\boldsymbol{\eta_2}\right)\right\}^{-1} \boldsymbol{\eta}_1 - \frac{1}{2}\log\left|-2\text{vec}^{-1}\left(\boldsymbol{D}_d^{+T}\boldsymbol{\eta}_2\right)\right|,$$

where $\boldsymbol{D}_d$ is the duplication matrix of size d and $\boldsymbol{D}_d^{+T}$ is the corresponding Moore–Penrose inverse matrix.

A.2.13 Inverse G-Wishart distribution

Let G be an undirected graph with d nodes labeled $1 \ldots d$ and E the sets of pairs of nodes that are connected by an edge. The symmetric $d \times d$ matrix $\boldsymbol{M}$ *respects* G if

$$\boldsymbol{M}_{ij} = 0 \quad \text{for all} \quad \{i,j\} \notin E.$$

A $d \times d$ random matrix $\boldsymbol{X}$ has an *inverse G-Wishart* distribution with graph G, parameters $\xi > 0$ and symmetric $d \times d$ matrix $\boldsymbol{\Lambda}$, written $\boldsymbol{X} \sim \text{Inverse-G-Wishart}(G, \xi, \boldsymbol{\Lambda})$ if and only if the density function of $\boldsymbol{X}$ satisfies

$$p(\boldsymbol{X}) \propto |\boldsymbol{X}|^{-(\xi+2)/2} \exp\left\{-\frac{1}{2}\text{tr}\left(\boldsymbol{\Lambda}\boldsymbol{X}^{-1}\right)\right\},$$

where $\boldsymbol{X}$ is symmetric and positive definite and $\boldsymbol{X}^{-1}$ respects G. A special case arises when G is a totally connected d-node graph, G_{full}, that is, $\boldsymbol{X}^{-1}$ is a full matrix, for which the inverse G-Wishart distribution coincides with the inverse Wishart distribution. A second special case is when G is a totally disconnected d-node graph, G_{diag}, that is, $\boldsymbol{X}^{-1}$ is a diagonal matrix, for which the inverse G-Wishart distribution coincides with a product of independent inverse chi-squared random variables. If $d = 1$, $G = G_{\text{full}} = G_{\text{diag}}$ and the inverse G-Wishart distribution reduces to the inverse chi-squared distribution. The sufficient statistic is

$$\boldsymbol{T}(\boldsymbol{x}) = \begin{bmatrix} \log|\boldsymbol{X}| \\ \operatorname{vech}\left(\boldsymbol{X}^{-1}\right) \end{bmatrix}.$$

The natural parameter vector and inverse mapping are

$$\boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \boldsymbol{\eta}_2 \end{bmatrix} = \begin{bmatrix} -\frac{1}{2}(\xi + 2) \\ -\frac{1}{2}\boldsymbol{D}_d^T \operatorname{vec}(\boldsymbol{\Lambda}) \end{bmatrix}, \quad \begin{bmatrix} \xi \\ \boldsymbol{\Lambda} \end{bmatrix} = \begin{bmatrix} -2\eta_1 - 2 \\ -2\operatorname{vec}^{-1}\left(\boldsymbol{D}_d^{+T}\boldsymbol{\eta}_2\right) \end{bmatrix},$$

where $\boldsymbol{D}_d$ is the duplication matrix of size d and $\boldsymbol{D}_d^{+T}$ is the corresponding Moore–Penrose inverse matrix. For the aforementioned special cases, expectations of $\boldsymbol{X}^{-1}$ are

$$E\left(\boldsymbol{X}^{-1}\right) = \left\{\eta_1 + \frac{1}{2}(d+1)\right\}\left\{\operatorname{vec}^{-1}\left(\boldsymbol{D}_d^{+T}\boldsymbol{\eta}_2\right)\right\}^{-1}, \quad \text{if} \quad G = G_{\text{full}},$$
$$E\left(\boldsymbol{X}^{-1}\right) = (\eta_1 + 1)\left\{\operatorname{vec}^{-1}\left(\boldsymbol{D}_d^{+T}\boldsymbol{\eta}_2\right)\right\}^{-1}, \quad \text{if} \quad G = G_{\text{diag}}.$$

The inverse G-Wishart distribution generalizes the inverse Wishart distribution and corresponds to the matrix inverses of random matrices that have a G-Wishart distribution (e.g. Atay-Kayis and Massam, 2005).

Other distributions

A.2.14 Bernoulli distribution

A random variable x has a *Bernoulli distribution* with probability $p \in [0, 1]$, written $x \sim \text{Bernoulli}(p)$, if its probability mass function is

$$p(x) = \begin{cases} p & \text{for} \quad x = 1, \\ 1 - p & \text{for} \quad x = 0. \end{cases}$$

A.2.15 Poisson distribution

A random variable x has a *Poisson distribution* with rate parameter $\lambda > 0$, written $x \sim \text{Poisson}(\lambda)$, if its probability mass function is

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x \in \mathbb{N}.$$

A.2.16 Uniform distribution

A random variable x has a *uniform distribution* defined over the interval $[a, b]$ such that $-\infty < a < b < \infty$, written $x \sim \text{Uniform}(a, b)$, if its probability density function is

$$p(x) = \begin{cases} \frac{1}{b-a} & \text{for} \quad x \in [a, b], \\ 0 & \text{otherwise.} \end{cases}$$

A.2.17 Student's t distribution

A random variable x has a *Student's t distribution* or, shortly, *t distribution* with $\nu > 0$ degrees of freedom, written $x \sim \text{t}(\nu)$ if the corresponding probability density function is such that

$$p(x) = \frac{\Gamma\{(\nu+1)/2\}}{\sqrt{\nu\pi}\Gamma(\nu/2)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad x \in \mathbb{R}.$$

A.2.18 Half Cauchy distribution

A random variable x has a *half Cauchy distribution* with scale parameter $\sigma > 0$, written $x \sim \text{Half-Cauchy}(\sigma)$, if its probability density function is

$$p(x) = \frac{2}{\pi\sigma\left\{1 + (x/\sigma)^2\right\}}, \quad x > 0.$$

A.2.19 Pareto distribution of II type

A random variable x has a *Pareto distribution of II type* with location parameter $\mu \in \mathbb{R}$, Pareto exponent $\alpha > 0$ and scale parameter $\beta > 0$, written $x \sim \text{Pareto}(\mu, \alpha, \beta)$, if its probability density function is

$$p(x) = \alpha\beta^{-1}\left(1 + \frac{x-\mu}{\beta}\right)^{-(\alpha+1)}, \quad x \geq \mu.$$

A.2.20 Univariate and multivariate skew t distribution

According to the formulation of Azzalini and Capitanio (2003), a $d \times 1$ random vector $\boldsymbol{x}$ is distributed as a *d-variate skew t distribution*, written $\boldsymbol{x} \sim \text{Skew-t}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu)$, if its probability density function is

$$p(\boldsymbol{x}) = 2t_d(\boldsymbol{x}; \nu)\, T_1\left\{\boldsymbol{\lambda}^T \boldsymbol{\Omega}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \left(\frac{\nu + d}{Q_{\boldsymbol{x}} + \nu}\right)^{1/2}; \nu + d\right\}$$

where

$$Q_{\boldsymbol{x}} = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}), \quad t_d(\boldsymbol{x}; \nu) = \frac{\Gamma\{(\nu + d)/2\}}{|\boldsymbol{\Sigma}|^{1/2}(\pi\nu)^{d/2}\Gamma(\nu/2)} \left(1 + \frac{Q_{\boldsymbol{x}}}{\nu}\right)^{-(\nu+d)/2}$$

is a d-dimensional t-variate density function with ν degrees of freedom and $T_1(y; \nu + d)$ indicates the scalar t distribution function with $\nu + d$ degrees of freedom. The vectors $\boldsymbol{\mu}, \boldsymbol{\lambda} \in \mathbb{R}^d$ are location and shape (skewness) parameter vectors respectively, while ν is the number of degrees of freedom. Also, $\boldsymbol{\Omega}$ is the diagonal matrix having the square root of the diagonal elements of the $d \times d$ full rank covariance matrix $\boldsymbol{\Sigma}$ on its main diagonal such that $\mathbf{R} = \boldsymbol{\Omega}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Omega}^{-1}$ is the correlation matrix associated with $\boldsymbol{\Sigma}$. The skew t distribution approaches the skew normal distribution as $\nu \to \infty$.
In our *univariate skew t* formulation, written as $\text{Skew-t}(\mu, \sigma^2, \lambda, \nu)$, the multivariate parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\lambda}$ respectively correspond to the univariate parameters μ, σ^2 and λ with $\boldsymbol{\Omega}$ corresponding to $\sqrt{\sigma^2}$.

A.3 The sufficient statistic expectation of the auxiliary variables arising in the skew normal and skew t VMP calculations

The following results were formalized by Wand as personal notes to derive the skew normal VMP algorithm. They are here included to support the VMP calculations for the skew t likelihood fragment and related theoretical results. Notations $\phi(\cdot)$ and $\Phi(\cdot)$ respectively refer to the standard normal density and distribution functions.
Consider

$$p(x) \propto \exp\left\{\begin{bmatrix} |x| \\ x^2 \end{bmatrix}^T \boldsymbol{\eta}\right\}, \quad x \in \mathbb{R}.$$

We express $E|x|$ and $E(x^2)$ in terms of the entries of

$$\boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix}.$$

Define

$$\mathcal{R}(p,q) = \int_{-\infty}^{\infty} |x|^p \left(x^2\right)^q \exp\left(\eta_1 |x| + \eta_2 x^2\right) dx,$$

where $p, q \in \{0,1\}$ and $p+q<1$. It follows that

$$E\left(\begin{bmatrix} |x| \\ x^2 \end{bmatrix}\right) = \begin{bmatrix} \mathcal{R}(1,0)/\mathcal{R}(0,0) \\ \mathcal{R}(0,1)/\mathcal{R}(0,0) \end{bmatrix}.$$

Then note that

$$\begin{aligned}
\mathcal{R}(p,q) &= \int_{-\infty}^{0} (-x)^p \left(x^2\right)^q \exp\left\{\eta_1(-x) + \eta_2 x^2\right\} dx + \int_0^{\infty} x^p \left(x^2\right)^q \exp\left(\eta_1 x + \eta_2 x^2\right) dx \\
&= (-1)^p \int_{-\infty}^{0} x^{p+2q} \exp\left\{\eta_1(-x) + \eta_2 x^2\right\} dx + \int_0^{\infty} x^{p+2q} \exp\left(\eta_1 x + \eta_2 x^2\right) dx \\
&= (-1)^p \int_{-\infty}^{0} (-u)^{p+2q} \exp\left(\eta_1 u + \eta_2 u^2\right)(-du) + \int_0^{\infty} x^{p+2q} \exp\left(\eta_1 x + \eta_2 x^2\right) dx \\
&= (-1)^{2(p+q)} \int_0^{\infty} u^{p+2q} \exp\left(\eta_1 u + \eta_2 u^2\right) du + \int_0^{\infty} x^{p+2q} \exp\left(\eta_1 x + \eta_2 x^2\right) dx \\
&= 2\int_0^{\infty} x^{p+2q} \exp\left(\eta_1 x + \eta_2 x^2\right) dx \\
&= 2\int_0^{\infty} x^{p+2q} \exp\left\{\begin{bmatrix} x \\ x^2 \end{bmatrix}^T \boldsymbol{\eta} - A_N(\boldsymbol{\eta})\right\} (2\pi)^{-1/2} dx \left[\exp\left\{A_N(\boldsymbol{\eta})\right\}(2\pi)^{1/2}\right] \\
&= \mathcal{Z}\int_0^{\infty} x^{p+2q} \exp\left\{\begin{bmatrix} x \\ x^2 \end{bmatrix}^T \boldsymbol{\eta} - A_N(\boldsymbol{\eta})\right\} (2\pi)^{-1/2} dx
\end{aligned}$$

where $\mathcal{Z} = 2\exp\left\{A_N(\boldsymbol{\eta})\right\}(2\pi)^{-1/2}$ and A_N is the log-partition of the normal distribution. Let

$$\mu = -\eta_1/(2\eta_2) \quad \text{and} \quad \sigma^2 = -1/(2\eta_2).$$

Then

$$\mathcal{R}(p,q) = \mathcal{Z}\int_0^{\infty} x^{p+2q} \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) dx,$$

where $\phi(z) = (2\pi)^{-1/2} \exp(-z^2/2)$ is the $N(0,1)$ density function. Introducing the change of variables $z = (x-\mu)/\sigma$ we get

$$\mathcal{R}(p,q) = \mathcal{Z} \int_{-\mu/\sigma}^{\infty} (\mu + \sigma z)^{p+2q} \phi(z)\, dz.$$

Expression for $\mathcal{R}(0,0)$

We have that

$$\begin{aligned}
\mathcal{R}(0,0) &= \mathcal{Z} \int_{-\mu/\sigma}^{\infty} \phi(z)\, dz \\
&= \mathcal{Z} \left\{ 1 - \int_{-\infty}^{-\mu/\sigma} \phi(z)\, dz \right\} \\
&= \mathcal{Z} \{1 - \Phi(-\mu/\sigma)\} \\
&= \mathcal{Z}\Phi(\mu/\sigma) \\
&= \mathcal{Z}\Phi\left(\eta_1/\sqrt{-2\eta_2}\right)
\end{aligned}$$

since

$$\mu/\sigma = \eta_1/\sqrt{-2\eta_2}.$$

Hence,

$$\mathcal{R}(0,0) = \mathcal{Z}\Phi\left(\eta_1/\sqrt{-2\eta_2}\right). \tag{A.11}$$

Expression for $\mathcal{R}(1,0)$

We have that

$$\begin{aligned}
\mathcal{R}(1,0) &= \int_{-\mu/\sigma}^{\infty} (\mu + \sigma z)\, \phi(z)\, dz \\
&= \mu\mathcal{R}(0,0) + \mathcal{Z}\sigma \int_{-\mu/\sigma}^{\infty} z\phi(z)\, dz \\
&= \mu\mathcal{R}(0,0) - \mathcal{Z}\sigma \int_{-\mu/\sigma}^{\infty} \phi'(z)\, dz \\
&= \mu\mathcal{R}(0,0) - \mathcal{Z}\sigma \left[\phi(z)\right]_{-\mu/\sigma}^{\infty} \\
&= \mu\mathcal{R}(0,0) + \mathcal{Z}\sigma\phi(\mu/\sigma) \\
&= -\eta_1\mathcal{R}(0,0)/(2\eta_2) + \mathcal{Z}\phi\left(\eta_1/\sqrt{-2\eta_2}\right)/\sqrt{-2\eta_2}.
\end{aligned}$$

Hence,

$$\mathcal{R}(1,0) = \mathcal{Z}\left\{\frac{\phi\left(\eta_1/\sqrt{-2\eta_2}\right)}{\sqrt{-2\eta_2}} - \frac{\eta_1\Phi\left(\eta_1/\sqrt{-2\eta_2}\right)}{2\eta_2}\right\}. \tag{A.12}$$

Expression for $\mathcal{R}(0,1)$

We have that

$$\begin{aligned}
\mathcal{R}(0,1) &= \mathcal{Z}\int_{-\mu/\sigma}^{\infty}(\mu+\sigma z)^2\phi(z)\,dz \\
&= \mu^2\mathcal{R}(0,0) + 2\mu\sigma\mathcal{Z}\int_{-\mu/\sigma}^{\infty} z\phi(z)\,dz + \sigma^2\mathcal{Z}\int_{-\mu/\sigma}^{\infty} z^2\phi(z)\,dz \\
&= \mu^2\mathcal{R}(0,0) + 2\mu\sigma\mathcal{Z}\phi(\mu/\sigma) + \sigma^2\mathcal{Z}\int_{-\mu/\sigma}^{\infty}\left(z^2-1\right)\phi(z)\,dz + \sigma^2\mathcal{Z}\int_{-\mu/\sigma}^{\infty}\phi(z)\,dz \\
&= \left(\mu^2+\sigma^2\right)\mathcal{R}(0,0) + 2\mu\sigma\mathcal{Z}\phi(\mu/\sigma) + \mathcal{Z}\sigma^2\int_{-\mu/\sigma}^{\infty}\phi''(z)\,dz \\
&= \left(\mu^2+\sigma^2\right)\mathcal{R}(0,0) + 2\mu\sigma\mathcal{Z}\phi(\mu/\sigma) - \mathcal{Z}\sigma^2\left[z\phi(z)\right]_{-\mu/\sigma}^{\infty} \\
&= \left(\mu^2+\sigma^2\right)\mathcal{R}(0,0) + 2\mu\sigma\mathcal{Z}\phi(\mu/\sigma) - \mathcal{Z}\mu\sigma\phi(\mu/\sigma) \\
&= \mathcal{Z}\left\{\left(\mu^2+\sigma^2\right)\Phi(\mu/\sigma) + \mu\sigma\phi(\mu/\sigma)\right\} \\
&= \mathcal{Z}\left\{\frac{\left(\eta_1^2-2\eta_2\right)\Phi\left(\eta_1/\sqrt{-2\eta_2}\right)}{4\eta_2^2} + \frac{\eta_1\phi\left(\eta_1/\sqrt{-2\eta_2}\right)}{\left(-2\eta_2\right)^{3/2}}\right\}.
\end{aligned}$$

Hence,

$$\mathcal{R}(0,1) = \mathcal{Z}\left\{\frac{\eta_1\phi\left(\eta_1/\sqrt{-2\eta_2}\right)}{\left(-2\eta_2\right)^{3/2}} + \frac{\left(\eta_1^2-2\eta_2\right)\Phi\left(\eta_1/\sqrt{-2\eta_2}\right)}{4\eta_2^2}\right\}. \tag{A.13}$$

Expression for $E|x|$

Combining (A.11) and (A.12) we get

$$E|x| = \frac{\mathcal{R}(1,0)}{\mathcal{R}(0,0)} \tag{A.14}$$

$$\begin{aligned}
&= \frac{(\phi/\Phi)\left(\eta_1/\sqrt{-2\eta_2}\right)}{\sqrt{-2\eta_2}} - \frac{\eta_1}{2\eta_2} \\
&= \frac{\zeta'\left(\eta_1/\sqrt{-2\eta_2}\right)}{\sqrt{-2\eta_2}} - \frac{\eta_1}{2\eta_2} \\
&= \frac{\sqrt{-2\eta_2}\zeta'\left(\eta_1/\sqrt{-2\eta_2}\right) + \eta_1}{(-2\eta_2)}.
\end{aligned} \tag{A.15}$$

Expression for $E\left(x^2\right)$

Combining (A.11) and (A.13) we get

$$\begin{aligned} E\left(x^2\right) &= \frac{\mathcal{R}\left(0,1\right)}{\mathcal{R}\left(0,0\right)} \\ &= \frac{\eta_1 \zeta^{'}\left(\eta_1/\sqrt{-2\eta_2}\right)}{\left(-2\eta_2\right)^{3/2}} + \frac{\left(\eta_1^2 - 2\eta_2\right)}{4\eta_2^2} \\ &= \frac{\eta_1\sqrt{-2\eta_2}\zeta^{'}\left(\eta_1/\sqrt{-2\eta_2}\right) + \eta_1^2 - 2\eta_2}{4\eta_2^2}. \end{aligned} \tag{A.16}$$

Appendix B

B.1 Derivations concerning Gaussian variational approximations for general design GLMMs

We here provide details about the derivations concerning Gaussian variational inference for general design GLMMs that can be easily adapted to further GLMMs and response distributions not treated here.

B.1.1 Proof of Proposition 2.1

Write the log-likelihood lower bound (2.3) as

$$\underline{\ell}\left(\boldsymbol{\beta}, \boldsymbol{G}, \boldsymbol{\mu}, \boldsymbol{\Lambda}\right) = \int q\left(\boldsymbol{u}; \boldsymbol{\xi}\right) \log\left\{p\left(\boldsymbol{y}, \boldsymbol{u}; \boldsymbol{\beta}, \boldsymbol{G}\right)\right\} d\boldsymbol{u} - \int q\left(\boldsymbol{u}; \boldsymbol{\xi}\right) \log\left\{q\left(\boldsymbol{u}; \boldsymbol{\xi}\right)\right\} d\boldsymbol{u}. \quad \text{(B.1)}$$

By applying properties of the expected value of a multivariate normal density and the formula to derive the expected value of the square from expected value and variance, the first component at the right hand side of (B.1) gives

$$\begin{aligned}
&\int q\left(\boldsymbol{u}; \boldsymbol{\xi}\right) \left\{\boldsymbol{y}^{\mathrm{T}}\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}\right) - \mathbf{1}^{\mathrm{T}} b\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}\right) + \mathbf{1}^{\mathrm{T}} c\left(\boldsymbol{y}\right)\right\} d\boldsymbol{u} \\
&+ \int q\left(\boldsymbol{u}; \boldsymbol{\xi}\right) \log\left\{\left(2\pi\right)^{-\frac{K}{2}} \left|\boldsymbol{G}\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{G}^{-1}\boldsymbol{u}\right)\right\} d\boldsymbol{u} \\
=&\boldsymbol{y}^{\mathrm{T}}\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\mu}\right) - \mathbf{1}^{\mathrm{T}} B\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\mu}, \mathrm{dg}\left(\boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{Z}^{\mathrm{T}}\right)\right) + \mathbf{1}^{\mathrm{T}} c\left(\boldsymbol{y}\right) \\
&- \frac{K}{2}\log\left(2\pi\right) - \frac{1}{2}\log\left|\boldsymbol{G}\right| - \frac{1}{2}\left\{\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{G}^{-1}\boldsymbol{\mu} + \mathrm{tr}\left(\boldsymbol{G}^{-1}\boldsymbol{\Lambda}\right)\right\}.
\end{aligned}$$

The second component at the right hand side of (B.1) is the negative entropy of a $N\left(\boldsymbol{\mu}, \boldsymbol{\Lambda}\right)$, that is,

$$-\frac{K}{2} + \frac{K}{2}\log\left(2\pi\right) + \frac{1}{2}\log\left|\boldsymbol{\Lambda}\right|.$$

The sum of the two previous results gives the final lower bound expression.

B.1.2 First and second order derivatives of the Gaussian variational lower bound

The first and second order derivatives of the lowerbound (2.4) presented here are similar to the results of Ormerod and Wand (2012) and can be derived using the rules indicated in Wand (2002) and the definitions of Section A.1 in Appendix A.
Define $\mathcal{B}^{(r)}(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = B^{(r)}\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\mu}, \mathrm{dg}\left(\boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{Z}^T\right)\right)$, where

$$B^{(r)}\left(\mu, \sigma^2\right) = \int_{-\infty}^{+\infty} b^{(r)}(\mu + \sigma x)\,\phi(x)\,dx,$$

and $\boldsymbol{Q}(\boldsymbol{A}) = \left(\boldsymbol{A} \otimes \boldsymbol{1}^T\right) \odot \left(\boldsymbol{1}^T \otimes \boldsymbol{A}\right)$. Let $\boldsymbol{M} = \boldsymbol{G}^{-1}\left(\boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Lambda}\right)\boldsymbol{G}^{-1}$. Shortly, we indicate the lower bound $\underline{\ell}(\boldsymbol{\beta}, \boldsymbol{G}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ with $\underline{\ell}$.
The first order derivatives of (2.4) with respect to the model parameters of interest and the variational parameters are

$$\begin{aligned}
\mathrm{D}_{\boldsymbol{\beta}}\underline{\ell} &= \left\{\boldsymbol{y} - \mathcal{B}^{(1)}(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Lambda})\right\}^T \boldsymbol{X}, \\
\mathrm{D}_{\mathrm{vech}(\boldsymbol{G})}\underline{\ell} &= \frac{1}{2}\mathrm{vec}\left(\boldsymbol{M} - \boldsymbol{G}^{-1}\right)\boldsymbol{D}_K, \\
\mathrm{D}_{\boldsymbol{\mu}}\underline{\ell} &= \left\{\boldsymbol{y} - \mathcal{B}^{(1)}(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Lambda})\right\}^T \boldsymbol{Z} - \boldsymbol{\mu}^T\boldsymbol{G}^{-1}, \\
\mathrm{D}_{\mathrm{vech}(\boldsymbol{\Lambda})}\underline{\ell} &= \frac{1}{2}\mathrm{vec}\left\{\boldsymbol{\Lambda}^{-1} - \boldsymbol{G}^{-1} - \boldsymbol{Z}^T\mathrm{diag}\left(\mathcal{B}^{(2)}(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Lambda})\right)\boldsymbol{Z}\right\}^T \boldsymbol{D}_K,
\end{aligned} \tag{B.2}$$

where $\boldsymbol{D}_K$ indicates the duplication matrix of order K.
The second order derivatives of (2.4) are

$$\begin{aligned}
\mathrm{H}_{\boldsymbol{\beta}\boldsymbol{\beta}}\underline{\ell} &= -\boldsymbol{X}^T\mathrm{diag}\left(\mathcal{B}^{(2)}(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Lambda})\right)\boldsymbol{X}, \\
\mathrm{H}_{\mathrm{vech}(\boldsymbol{G})\mathrm{vech}(\boldsymbol{G})}\underline{\ell} &= \frac{1}{2}\boldsymbol{D}_K^T\left(\boldsymbol{G}^{-1} \otimes \boldsymbol{G}^{-1} - \boldsymbol{G}^{-1} \otimes \boldsymbol{M} + \boldsymbol{M} \otimes \boldsymbol{G}^{-1}\right)\boldsymbol{D}_K, \\
\mathrm{H}_{\boldsymbol{\beta}\boldsymbol{\mu}}\underline{\ell} &= -\boldsymbol{X}^T\mathrm{diag}\left(\mathcal{B}^{(2)}(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Lambda})\right)\boldsymbol{Z}, \\
\mathrm{H}_{\boldsymbol{\beta}\mathrm{vech}(\boldsymbol{\Lambda})}\underline{\ell} &= -\frac{1}{2}\boldsymbol{X}^T\mathrm{diag}\left(\mathcal{B}^{(3)}(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Lambda})\right)\boldsymbol{Q}(\boldsymbol{Z})\,\boldsymbol{D}_K, \\
\mathrm{H}_{\mathrm{vech}(\boldsymbol{G})\boldsymbol{\mu}}\underline{\ell} &= \boldsymbol{D}_K^T\left\{\left(\boldsymbol{G}^{-1}\boldsymbol{\mu}\right) \otimes \boldsymbol{G}^{-1}\right\}, \\
\mathrm{H}_{\mathrm{vech}(\boldsymbol{G})\mathrm{vech}(\boldsymbol{\Lambda})}\underline{\ell} &= \frac{1}{2}\boldsymbol{D}_K^T\left(\boldsymbol{G}^{-1} \otimes \boldsymbol{G}^{-1}\right)\boldsymbol{D}_K, \\
\mathrm{H}_{\boldsymbol{\mu}\boldsymbol{\mu}}\underline{\ell} &= -\boldsymbol{Z}^T\mathrm{diag}\left(\mathcal{B}^{(2)}(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Lambda})\right)\boldsymbol{Z} - \boldsymbol{G}^{-1}, \\
\mathrm{H}_{\boldsymbol{\mu}\mathrm{vech}(\boldsymbol{\Lambda})}\underline{\ell} &= -\frac{1}{2}\boldsymbol{Z}^T\mathrm{diag}\left(\mathcal{B}^{(3)}(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Lambda})\right)\boldsymbol{Q}(\boldsymbol{Z})\,D_K, \\
\mathrm{H}_{\mathrm{vech}(\boldsymbol{\Lambda})\mathrm{vech}(\boldsymbol{\Lambda})}\underline{\ell} &= -\frac{1}{4}\boldsymbol{D}_K^T\left\{\boldsymbol{Q}(\boldsymbol{Z})^T\mathrm{diag}\left(\mathcal{B}^{(4)}(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Lambda})\right)\boldsymbol{Q}(\boldsymbol{Z}) + 2\left(\boldsymbol{\Lambda}^{-1} \otimes \boldsymbol{\Lambda}^{-1}\right)\right\}\boldsymbol{D}_K.
\end{aligned}$$

B.1.3 Proof of Proposition 2.2

The matrix $\widehat{\underline{\boldsymbol{\Lambda}}}$ solves the score equation $\mathrm{D}_{\mathrm{vech}(\boldsymbol{\Lambda})}\underline{\ell} = \mathbf{0}$, with $\mathrm{D}_{\mathrm{vech}(\boldsymbol{\Lambda})}$ expressed as in B.2, from which we get

$$\begin{aligned}\mathrm{vec}\left\{\boldsymbol{\Lambda}^{-1} - \boldsymbol{G}^{-1} - \boldsymbol{Z}^T\mathrm{diag}\left(\mathcal{B}^{(2)}\left(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Lambda}\right)\right)\boldsymbol{Z}\right\}^T \boldsymbol{D}_K &= \mathbf{0},\\ \boldsymbol{\Lambda}^{-1} - \boldsymbol{G}^{-1} - \boldsymbol{Z}^T\mathrm{diag}\left(\mathcal{B}^{(2)}\left(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Lambda}\right)\right)\boldsymbol{Z} &= \mathbf{0}.\end{aligned}$$

From the last expression it follows that at convergence of the optimization procedure

$$\widehat{\underline{\boldsymbol{\Lambda}}} = \left(\boldsymbol{G}^{-1} + \boldsymbol{Z}^T\mathrm{diag}\left(\mathcal{B}^{(2)}\left(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Lambda}\right)\right)\boldsymbol{Z}\right)^{-1}.$$

Noting that $\mathrm{H}_{\boldsymbol{\mu}\boldsymbol{\mu}}\underline{\ell} = -\boldsymbol{G}^{-1} - \boldsymbol{Z}^T\mathrm{diag}\left(\mathcal{B}^{(2)}\left(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Lambda}\right)\right)\boldsymbol{Z}$ the thesis follows.

B.1.4 `rstan` code for fitting Poisson nonparametric regression via MCMC

`Stan` (Carpenter *et al.*, 2017) is a probabilistic programming language, written in C++, to perform Bayesian statistical inference through a particular form of MCMC sampling (Hamiltonian Monte Carlo no U-turn sampling). The `R` computing environment interfaces with `Stan` via `rstan`. We here provide the `rstan` code for fitting Poisson nonparametric regression via MCMC. Similar scripts were employed for the other GLMMs. The first step consists in setting the values for burn-in, kept sample size and thinning factor.

```
nWarm <- 10000       # Length of burn-in.
nKept <- 5000        # Size of the kept sample.
nThin <- 5           # Thinning factor.
```

Second, specify the input data: the number of observations, n, the design matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$, the response vector, y and the hyperparameters $\sigma_{\boldsymbol{\beta}}$ and A, which are associated with the prior distributions defined in the `model` environment. Constrains can also be included. For instance, in our case hyperparameters must be positive.

```
PoissRespNPregn <-
'data
{
  int<lower=1> ncZ;           int<lower=1> n;
  int<lower=0> y[n];
```

```
  matrix[n,2] X;              matrix[n,ncZ] Z;
  real<lower=0> sigmaBeta;    real<lower=0> A;
}
```

Then, specify parameters and transformed parameters.

```
parameters
{
  vector[2] beta;    real<lower=0> sigma;
  vector[ncZ] u;
}
transformed parameters
{
  vector[n] etaVec;
  etaVec = X*beta + Z*u;
}
```

Finally, define the Bayesian model, including prior distributions.

```
model
{
  for (i in 1:n)
  {
    y[i] ~ poisson(exp(etaVec[i]));
  }
  beta ~ normal(0,sigmaBeta);   u ~ normal(0,sigma);
  sigma ~ cauchy(0,A);
}'
```

The `rstan` code is now ready to be compiled and fit the model on real data.

```
allData <- list(n=n,ncZ=ncZ,y=y,X=X,Z=Z,sigmaBeta=sigmaBeta,A=A)


stanCompilObj <- stan(model_code=PoissRespNPregn,data=allData,
                      iter=1,chains=1)
```

Last, obtain the MCMC samples for each parameter and save the `Stan` output. In our case the $\boldsymbol{\beta}$ vector is of dimension 1×2.

```
stanObj <-  stan(model_code=PoissRespNPregn,data=allData,
                 warmup=nBurnin,iter=(nBurnin+nIter),
```

```
                chains=1,thin=nThin,refresh=100,fit=stanCompilObj)

betaMCMC <- NULL
for (j in 1:2)
{
  charVar <- paste("beta[",as.character(j),"]",sep="")
  betaMCMC <- rbind(betaMCMC,extract(stanObj,charVar,permuted=FALSE))
}
beta0MCMCvec <- betaMCMC[1,]
beta1MCMCvec <- betaMCMC[2,]

uMCMC <- NULL
for (k in 1:ncZ)
{
  charVar <- paste("u[",as.character(k),"]",sep="")
  uMCMC <- rbind(uMCMC,extract(stanObj,charVar,permuted=FALSE))
}

sigmaMCMC <- as.vector(extract(stanObj,"sigma",permuted=FALSE))
```

Appendix C

C.1 Derivations concerning the SVR likelihood fragment

We here provide details about the derivation of Algorithm 3.2 concerning VMP for the SVR likelihood fragment.

C.1.1 Derivation of Algorithm 3.2

According to (1.15)

$$m_{\check{p}(a_1)\to a_{1i}} = \check{p}\left(a_{1i}\right) = I\left(a_{1i} > 0\right), \quad 1 \le i \le n \tag{C.1}$$

$$m_{\check{p}(a_2)\to a_{2i}} = \check{p}\left(a_{2i}\right) = I\left(a_{2i} > 0\right), \quad 1 \le i \le n. \tag{C.2}$$

Messages from $\check{p}\left(\boldsymbol{y} \,|\, \boldsymbol{\theta}, \boldsymbol{a}_1, \boldsymbol{a}_2\right)$ to each of the a_{1i}, for $1 \le i \le n$, can be easily obtained writing $\log \check{p}\left(\boldsymbol{y} \,|\, \boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{\theta}\right)$ as a function of a_{1i}, indicating with "const" terms which are independent of a_{1i}:

$$\begin{aligned}
\log \check{p}\left(\boldsymbol{y} \,|\, \boldsymbol{\theta}, \boldsymbol{a}_1, \boldsymbol{a}_2\right) =& -\frac{1}{2}\log\left(a_{1i}\right) - \frac{1}{2a_{1i}}\left(a_{1i}^2 - 2a_{1i}\left\{\left(\boldsymbol{A}\boldsymbol{\theta}\right)_i + \varepsilon - y_i\right\}\right. \\
&\left. + \left\{\left(\boldsymbol{A}\boldsymbol{\theta}\right)_i + \varepsilon - y_i\right\}^2\right) + \text{const} \\
=& -\frac{1}{2}\log\left(a_{1i}\right) - \frac{1}{2}a_{1i} - \frac{1}{2a_{1i}}\left\{\left(\boldsymbol{A}\boldsymbol{\theta}\right)_i + \varepsilon - y_i\right\}^2 + \text{const} \\
=& \log\left[a_{1i}^{-1/2}\exp\left\{\begin{bmatrix} a_{1i} \\ 1/a_{1i} \end{bmatrix}^T \begin{bmatrix} -\frac{1}{2} \\ -\frac{1}{2}\left\{\left(\boldsymbol{A}\boldsymbol{\theta}\right)_i + \varepsilon - y_i\right\}^2 \end{bmatrix}\right\}\right] + \text{const}.
\end{aligned}$$

Therefore

$$m_{\check{p}(\boldsymbol{y}\,|\,\boldsymbol{\theta},\boldsymbol{a}_1,\boldsymbol{a}_2)\to a_{1i}}(a_{1i}) = a_{1i}^{-1/2}\exp\left\{\begin{bmatrix} a_{1i} \\ 1/a_{1i}\end{bmatrix}^T \begin{bmatrix} -\frac{1}{2} \\ -\frac{1}{2}E_{q(\boldsymbol{\theta})}\left[\{(\boldsymbol{A}\boldsymbol{\theta})_i + \varepsilon - y_i\}^2\right]\end{bmatrix}\right\}, \tag{C.3}$$

where $E_{q(\boldsymbol{\theta})}$ denotes expectation with respect to the normalization of

$$m_{\check{p}(\boldsymbol{y}\,|\,\boldsymbol{\theta},\boldsymbol{a}_1,\boldsymbol{a}_2)\to\boldsymbol{\theta}}(\boldsymbol{\theta})\, m_{\boldsymbol{\theta}\to\check{p}(\boldsymbol{y}\,|\,\boldsymbol{\theta},\boldsymbol{a}_1,\boldsymbol{a}_2)}(\boldsymbol{\theta}). \tag{C.4}$$

Similarly

$$m_{\check{p}(\boldsymbol{y}\,|\,\boldsymbol{\theta},\boldsymbol{a}_1,\boldsymbol{a}_2)\to a_{2i}}(a_{2i}) = a_{2i}^{-1/2}\exp\left\{\begin{bmatrix} a_{2i} \\ 1/a_{2i}\end{bmatrix}^T \begin{bmatrix} -\frac{1}{2} \\ -\frac{1}{2}E_{q(\boldsymbol{\theta})}\left[\{y_i - (\boldsymbol{A}\boldsymbol{\theta})_i + \varepsilon\}^2\right]\end{bmatrix}\right\}. \tag{C.5}$$

Messages from $\check{p}(\boldsymbol{y}\,|\,\boldsymbol{a}_1,\boldsymbol{a}_2,\boldsymbol{\theta})$ to $\boldsymbol{\theta}$ can be easily obtained writing $\log\check{p}(\boldsymbol{y}\,|\,\boldsymbol{\theta},\boldsymbol{a}_1,\boldsymbol{a}_2)$ as a function of $\boldsymbol{\theta}$, indicating with "const" terms which are independent of $\boldsymbol{\theta}$. Recall also the trace trick

$$\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a} = \mathrm{tr}\left(\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}\right) = \mathrm{tr}\left(\boldsymbol{a}\boldsymbol{a}^T\boldsymbol{A}\right) = \mathrm{tr}\left(\boldsymbol{A}\boldsymbol{a}\boldsymbol{a}^T\right), \tag{C.6}$$

for any vector $\boldsymbol{a}$ and any matrix $\boldsymbol{A}$ and that

$$\mathrm{tr}\left(\boldsymbol{X}^T\boldsymbol{Y}\right) = \mathrm{vec}\,(\boldsymbol{Y})^T\,\mathrm{vec}\,(\boldsymbol{X}), \tag{C.7}$$

for any matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$. We get

$$\begin{aligned}
\log\check{p}(\boldsymbol{y}\,|\,\boldsymbol{\theta},\boldsymbol{a}_1,\boldsymbol{a}_2) = &-\frac{1}{2}\Big\{(\boldsymbol{a}_1+\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\theta}-\varepsilon)^T\,\mathrm{diag}\,(\boldsymbol{a}_1)^{-1}(\boldsymbol{a}_1+\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\theta}-\varepsilon) \\
&+(\boldsymbol{a}_2-\boldsymbol{y}+\boldsymbol{A}\boldsymbol{\theta}-\varepsilon)^T\,\mathrm{diag}\,(\boldsymbol{a}_2)^{-1}(\boldsymbol{a}_2-\boldsymbol{y}+\boldsymbol{A}\boldsymbol{\theta}-\varepsilon)\Big\} + \mathrm{const} \\
= &-\frac{1}{2}\left[-2\boldsymbol{\theta}^T\boldsymbol{A}^T\mathrm{diag}\,(\boldsymbol{a}_1)^{-1}(\boldsymbol{a}_1+\boldsymbol{y}-\varepsilon) + \mathrm{tr}\left\{\boldsymbol{A}^T\mathrm{diag}\,(\boldsymbol{a}_1)^{-1}\boldsymbol{A}\boldsymbol{\theta}\boldsymbol{\theta}^T\right\}\right. \\
&\left.+2\boldsymbol{\theta}^T\boldsymbol{A}^T\mathrm{diag}\,(\boldsymbol{a}_2)^{-1}(\boldsymbol{a}_2-\boldsymbol{y}-\varepsilon) + \mathrm{tr}\left\{\boldsymbol{A}^T\mathrm{diag}\,(\boldsymbol{a}_2)^{-1}\boldsymbol{A}\boldsymbol{\theta}\boldsymbol{\theta}^T\right\}\right] \\
&+\mathrm{const} \\
= &\begin{bmatrix}\boldsymbol{\theta} \\ \mathrm{vec}\left(\boldsymbol{\theta}\boldsymbol{\theta}^T\right)\end{bmatrix}^T \begin{bmatrix}\boldsymbol{A}^T\left\{\frac{1}{\boldsymbol{a}_1}\odot(\boldsymbol{y}-\varepsilon)+\frac{1}{\boldsymbol{a}_2}\odot(\boldsymbol{y}+\varepsilon)\right\} \\ -\frac{1}{2}\mathrm{vec}\left\{\boldsymbol{A}^T\mathrm{diag}\left(\frac{1}{\boldsymbol{a}_1}+\frac{1}{\boldsymbol{a}_2}\right)\boldsymbol{A}\right\}\end{bmatrix} + \mathrm{const}.
\end{aligned}$$

Then the message from $\check{p}(\boldsymbol{y} \,|\, \boldsymbol{\theta}, \boldsymbol{a}_1, \boldsymbol{a}_2)$ to $\boldsymbol{\theta}$ is proportional to a multivariate normal density function with natural parameter update

$$\boldsymbol{\eta}_{\check{p}(\boldsymbol{y} \,|\, \boldsymbol{\theta}, \boldsymbol{a}_1, \boldsymbol{a}_2) \to \boldsymbol{\theta}} \longleftarrow \begin{bmatrix} \boldsymbol{A}^T \left\{ E_{q(\boldsymbol{a}_1)} \left(1/\boldsymbol{a}_1\right) \odot \left(\boldsymbol{y} - \mathbf{1}_n \varepsilon\right) \right. \\ \qquad \left. + E_{q(\boldsymbol{a}_2)} \left(1/\boldsymbol{a}_2\right) \odot \left(\boldsymbol{y} + \mathbf{1}_n \varepsilon\right) \right\} \\ \\ -\frac{1}{2} \mathrm{vec} \left\{ \boldsymbol{A}^T \mathrm{diag} \left(E_{q(\boldsymbol{a}_1)} \left(1/\boldsymbol{a}_1\right) + E_{q(\boldsymbol{a}_2)} \left(1/\boldsymbol{a}_2\right) \right) \boldsymbol{A} \right\} \end{bmatrix}, \tag{C.8}$$

where

$$E_{q(\boldsymbol{a}_1)} \left(1/\boldsymbol{a}_1\right) = \left[E_{q(a_{11})} \left(1/a_{11}\right), \ldots, E_{q(a_{1n})} \left(1/a_{1n}\right) \right]^T$$

and $E_{q(a_{1i})}$ denotes expectation with respect to the normalized

$$q^* \left(a_{1i}\right) \propto m_{\check{p}(\boldsymbol{y} \,|\, \boldsymbol{\theta}, \boldsymbol{a}_1, \boldsymbol{a}_2) \to a_{1i}} \left(a_{1i}\right) m_{\check{p}(\boldsymbol{a}_1) \to a_{1i}} \left(a_{1i}\right), \quad 1 \leq i \leq n. \tag{C.9}$$

$E_{q(\boldsymbol{a}_2)} \left(1/\boldsymbol{a}_2\right)$ is similarly defined with $E_{q(a_{2i})}$ denoting expectation with respect to the normalized

$$q^* \left(a_{2i}\right) \propto m_{\check{p}(\boldsymbol{y} \,|\, \boldsymbol{\theta}, \boldsymbol{a}_1, \boldsymbol{a}_2) \to a_{2i}} \left(a_{2i}\right) m_{\check{p}(\boldsymbol{a}_2) \to a_{2i}} \left(a_{2i}\right), \quad 1 \leq i \leq n. \tag{C.10}$$

Combining (C.1) and (C.3) with (C.9), and (C.2) and (C.5) with (C.10), it is evident that both $E_{q(a_{1i})}$ and $E_{q(a_{2i})}$ denote expectation with respect to a generalized inverse Gaussian distribution with $p = \frac{1}{2}$ and natural parameter vectors

$$\begin{bmatrix} -\frac{1}{2} \\ -\frac{1}{2} E_{q(\boldsymbol{\theta})} \left[\left\{ \left(\boldsymbol{A}\boldsymbol{\theta}\right)_i + \varepsilon - y_i \right\}^2 \right] \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} -\frac{1}{2} \\ -\frac{1}{2} E_{q(\boldsymbol{\theta})} \left[\left\{ y_i - \left(\boldsymbol{A}\boldsymbol{\theta}\right)_i + \varepsilon \right\}^2 \right] \end{bmatrix}$$

respectively. Then, from (A.3)

$$\begin{aligned} E_{q(a_{1i})} \left(1/a_{1i}\right) &= \left(E_{q(\boldsymbol{\theta})} \left[\left\{ \left(\boldsymbol{A}\boldsymbol{\theta}\right)_i + \varepsilon - y_i \right\}^2 \right] \right)^{-1/2} \\ &= \left[E_{q(\boldsymbol{\theta})} \left\{ \left(\boldsymbol{A}\boldsymbol{\theta}\right)_i^2 \right\} + 2 \left(\varepsilon - y_i\right) E_{q(\boldsymbol{\theta})} \left\{ \left(\boldsymbol{A}\boldsymbol{\theta}\right)_i \right\} + \left(\varepsilon - y_i\right)^2 \right]^{-1/2} \\ &= \Big(\left[E_{q(\boldsymbol{\theta})} \left\{ \left(\boldsymbol{A}\boldsymbol{\theta}\right)_i \right\} \right]^2 + \mathrm{Var}_{q(\boldsymbol{\theta})} \left\{ \left(\boldsymbol{A}\boldsymbol{\theta}\right)_i \right\} \\ &\quad + 2 \left(\varepsilon - y_i\right) E_{q(\boldsymbol{\theta})} \left\{ \left(\boldsymbol{A}\boldsymbol{\theta}\right)_i \right\} + \left(\varepsilon - y_i\right)^2 \Big)^{-1/2} \\ &= \left\{ \left(\boldsymbol{A}\boldsymbol{\mu}_{q(\boldsymbol{\theta})} \right)_i^2 + \left(\boldsymbol{A}\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})} \boldsymbol{A}^T \right)_{ii} + 2 \left(\varepsilon - y_i\right) \left(\boldsymbol{A}\boldsymbol{\mu}_{q(\boldsymbol{\theta})} \right)_i + \left(\varepsilon - y_i\right)^2 \right\}^{-1/2} \\ &= \left[\left\{ \left(\boldsymbol{A}\boldsymbol{\mu}_{q(\boldsymbol{\theta})} \right)_i + \varepsilon - y_i \right\}^2 + \left(\boldsymbol{A}\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})} \boldsymbol{A}^T \right)_{ii} \right]^{-1/2}, \end{aligned}$$

and

$$
\begin{aligned}
E_{q(a_{2i})}\left(1/a_{2i}\right) &= \left(E_{q(\boldsymbol{\theta})}\left[\left\{y_i - \left(\boldsymbol{A}\boldsymbol{\theta}\right)_i + \varepsilon\right\}^2\right]\right)^{-1/2} \\
&= \left[\left(y_i+\varepsilon\right)^2 - 2\left(y_i+\varepsilon\right) E_{q(\boldsymbol{\theta})}\left\{\left(\boldsymbol{A}\boldsymbol{\theta}\right)_i\right\} + E_{q(\boldsymbol{\theta})}\left\{\left(\boldsymbol{A}\boldsymbol{\theta}\right)_i^2\right\}\right]^{-1/2} \\
&= \Big(\left(y_i+\varepsilon\right)^2 - 2\left(y_i+\varepsilon\right) E_{q(\boldsymbol{\theta})}\left\{\left(\boldsymbol{A}\boldsymbol{\theta}\right)_i\right\} \\
&\quad + \left[E_{q(\boldsymbol{\theta})}\left\{\left(\boldsymbol{A}\boldsymbol{\theta}\right)_i\right\}\right]^2 + \mathrm{Var}_{q(\boldsymbol{\theta})}\left\{\left(\boldsymbol{A}\boldsymbol{\theta}\right)_i\right\}\Big)^{-1/2} \\
&= \left\{\left(y_i+\varepsilon\right)^2 - 2\left(y_i+\varepsilon\right)\left(\boldsymbol{A}\boldsymbol{\mu}_{q(\boldsymbol{\theta})}\right)_i + \left(\boldsymbol{A}\boldsymbol{\mu}_{q(\boldsymbol{\theta})}\right)_i^2 + \left(\boldsymbol{A}\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}\boldsymbol{A}^T\right)_{ii}\right\}^{-1/2} \\
&= \left[\left\{y_i - \left(\boldsymbol{A}\boldsymbol{\mu}_{q(\boldsymbol{\theta})}\right)_i + \varepsilon\right\}^2 + \left(\boldsymbol{A}\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}\boldsymbol{A}^T\right)_{ii}\right]^{-1/2},
\end{aligned}
$$

where $\boldsymbol{\mu}_{q(\boldsymbol{\theta})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}$ are the common parameters of the multivariate normal density that arises normalizing (C.4), if we assume that $m_{\boldsymbol{\theta}\to\check{p}(\boldsymbol{y}\,|\,\boldsymbol{\theta},\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(\boldsymbol{\theta}\right)$ is conjugate to $m_{\check{p}(\boldsymbol{y}\,|\,\boldsymbol{\theta},\boldsymbol{a}_1,\boldsymbol{a}_2)\to\boldsymbol{\theta}}\left(\boldsymbol{\theta}\right)$.
Define now the updates

$$
\begin{aligned}
&\boldsymbol{v}_1 \longleftarrow \boldsymbol{A}\boldsymbol{\mu}_{q(\boldsymbol{\theta})}, \quad \boldsymbol{v}_2 \longleftarrow \mathrm{diagonal}\left(\boldsymbol{A}\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}\boldsymbol{A}^T\right) \\
&\boldsymbol{v}_3 \longleftarrow \left\{\left(\boldsymbol{v}_1 + \varepsilon\mathbf{1}_n - \boldsymbol{y}\right)^2 + \boldsymbol{v}_2\right\}^{-1/2}, \\
&\boldsymbol{v}_4 \longleftarrow \left\{\left(\boldsymbol{y} - \boldsymbol{v}_1 + \varepsilon\mathbf{1}_n\right)^2 + \boldsymbol{v}_2\right\}^{-1/2}.
\end{aligned}
\tag{C.11}
$$

Then Algorithm 3.2 follows from updates (C.8) and (C.11), with $\boldsymbol{\mu}_{q(\boldsymbol{\theta})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}$ replaced by their natural parameter counterparts according to (A.9).

C.2 Derivations concerning the skew t likelihood fragment

We here provide details about the derivation of Algorithm 3.4 and 3.5 concerning VMP skew t likelihood fragment derived according to the two alternative approximating density specifications. Moreover, we illustrate the related theoretical results and provide a code for MCMC fitting.

C.2.1 Derivation of Algorithm 3.4

We here derive the algorithm based on the product density restriction (3.22) to which we refer as Algorithm 3.4.

It follows from (8) that the logarithm of the normal density term in the likelihood factor is

$$\begin{aligned}\log p\left(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2\right) = &-\frac{1+\lambda^2}{2\sigma^2}\left\{\left(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\theta}-\frac{\sigma\lambda\left|\boldsymbol{a}_1\right|\odot\sqrt{\boldsymbol{a}_2}}{\sqrt{1+\lambda^2}}\right)^T\operatorname{diag}\left(\boldsymbol{a}_2\right)^{-1}\right.\\ &\left.\times\left(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\theta}-\frac{\sigma\lambda\left|\boldsymbol{a}_1\right|\odot\sqrt{\boldsymbol{a}_2}}{\sqrt{1+\lambda^2}}\right)\right\}-\frac{1}{2}\mathbf{1}_n\log\left(\boldsymbol{a_2}\right)\\ &+\frac{n}{2}\left\{\log\left(1+\lambda^2\right)-\log\left(\sigma^2\right)\right\}+\text{const}\end{aligned}$$

where "const" indicates terms not depending on the likelihood parameters.
With simple applications of formulae (1.14)–(1.17) and steps similar to those given in Sections 4.1.5 and S.2.5.5 of Wand (2017) for the Gaussian likelihood fragment and in McLean and Wand (2018, Sections S.3.2 and S.3.4) for the t likelihood and skew normal fragment updates we derive the message updates of the VMP algorithm. From hereafter we denote with "const" terms that do not depend on the variable(s) of interest.

Applying (1.16), the message from factor $p\left(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2\right)$ to node $\boldsymbol{\theta}$ update involves expectation with respect to

$$\begin{aligned}&m_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\sigma^2}\left(\sigma^2\right)m_{\sigma^2\to p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(\sigma^2\right)m_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\lambda}\left(\lambda\right)\\ &\times m_{\lambda\to p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(\lambda\right)\prod_{i=1}^{n}\left\{m_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to a_{1i}}\left(a_{1i}\right)\right.\\ &\left.\times\ m_{a_{1i}\to p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(a_{1i}\right)m_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to a_{2i}}\left(a_{2i}\right)m_{a_{2i}\to p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(a_{2i}\right)\right\}.\end{aligned}\tag{C.12}$$

We write $\log p\left(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2\right)$ as a function of $\boldsymbol{\theta}$ in terms of the sufficient statistic vector. Using (C.6) and (C.7) we have

$$\begin{aligned}\log p\left(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2\right) =& -\frac{1+\lambda^2}{2\sigma^2}\left(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\theta}-\frac{\sigma\lambda\left|\boldsymbol{a}_1\right|\odot\sqrt{\boldsymbol{a}_2}}{\sqrt{1+\lambda^2}}\right)^T\operatorname{diag}\left(\boldsymbol{a}_2\right)^{-1}\\ &\times\left(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\theta}-\frac{\sigma\lambda\left|\boldsymbol{a}_1\right|\odot\sqrt{\boldsymbol{a}_2}}{\sqrt{1+\lambda^2}}\right)+\text{const}\\ =& -\frac{1+\lambda^2}{2\sigma^2}\left[2\boldsymbol{\theta}^T\boldsymbol{A}^T\operatorname{diag}\left(\boldsymbol{a}_2\right)^{-1}\frac{\sigma\lambda\left|\boldsymbol{a}_1\right|\odot\sqrt{\boldsymbol{a}_2}}{\sqrt{1+\lambda^2}}\right.\\ &\left.-2\boldsymbol{\theta}^T\boldsymbol{A}^T\operatorname{diag}\left(\boldsymbol{a}_2\right)^{-1}\boldsymbol{y}+\operatorname{tr}\left\{\boldsymbol{A}^T\operatorname{diag}\left(\boldsymbol{a}_2\right)^{-1}\boldsymbol{A}\boldsymbol{\theta}\boldsymbol{\theta}^T\right\}\right]\\ =&\begin{bmatrix}\boldsymbol{\theta}\\ \operatorname{vec}\left(\boldsymbol{\theta}\boldsymbol{\theta}^T\right)\end{bmatrix}^T\begin{bmatrix}\frac{1+\lambda^2}{\sigma^2}\boldsymbol{A}^T\operatorname{diag}\left(\boldsymbol{a}_2\right)^{-1}\boldsymbol{y}\\ -\frac{\lambda\sqrt{1+\lambda^2}}{\sigma}\boldsymbol{A}^T\operatorname{diag}\left(\sqrt{\boldsymbol{a}_2}\right)^{-1}\left|\boldsymbol{a}_1\right|\\ -\frac{1+\lambda^2}{2\sigma^2}\operatorname{vec}\left\{\boldsymbol{A}^T\operatorname{diag}\left(\boldsymbol{a}_2\right)^{-1}\boldsymbol{A}\right\}\end{bmatrix}+\text{const.}\end{aligned}$$

Hence, applying (1.15) and expectation with respect to (C.12), the message $m_{p(y|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\boldsymbol{\theta}}$ is proportional to a multivariate normal density function with natural parameter update

$$\begin{aligned}\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\boldsymbol{\theta}}\longleftarrow&\left(1+\mu_{q(\lambda^2)}\right)\mu_{q(1/\sigma^2)}\begin{bmatrix}\boldsymbol{A}^T\operatorname{diag}\left\{E_{q(\boldsymbol{a}_2)}\left(\boldsymbol{1}_n/\boldsymbol{a}_2\right)\right\}\boldsymbol{y}\\ -\frac{1}{2}\operatorname{vec}\left(\boldsymbol{A}^T\operatorname{diag}\left\{E_{q(\boldsymbol{a}_2)}\left(\boldsymbol{1}_n/\boldsymbol{a}_2\right)\right\}\boldsymbol{A}\right)\end{bmatrix}\\ &-\mu_{q\left(\lambda\sqrt{1+\lambda^2}\right)}\mu_{q(1/\sigma)}\begin{bmatrix}\boldsymbol{A}^T\operatorname{diag}\left\{E_{q(\boldsymbol{a}_2)}\left(\boldsymbol{1}_n/\sqrt{\boldsymbol{a}_2}\right)\right\}E_{q(\boldsymbol{a}_1)}\left|\boldsymbol{a}_1\right|\\ \boldsymbol{0}\end{bmatrix},\end{aligned}$$

where

$$\begin{aligned}\mu_{q\left(1/\sigma^k\right)}&=\int_0^\infty\left(1/\sigma^k\right)q^*\left(\sigma^2\right)d\sigma^2\quad\text{for}\quad k=1,2,\\ \mu_{q(\lambda^2)}&=\int_{-\infty}^\infty\lambda^2q^*\left(\lambda\right)d\lambda,\\ \mu_{q\left(\lambda\sqrt{1+\lambda^2}\right)}&=\int_{-\infty}^\infty\lambda\sqrt{1+\lambda^2}q^*\left(\lambda\right)d\lambda,\end{aligned}$$

with $q^*\left(\sigma^2\right)$ proportional to

$$m_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\sigma^2}\left(\sigma^2\right)m_{\sigma^2\to p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(\sigma^2\right)$$

and $q^*\left(\lambda\right)$ similarly defined. $E_{q(\boldsymbol{a}_1)}$ denotes expectation with respect to $q^*\left(\boldsymbol{a}_1\right)=\prod_{i=1}^n q^*\left(a_{1i}\right)$, where $q^*\left(a_{1i}\right)$ is proportional to

$$m_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to a_{1i}}\left(a_{1i}\right)m_{a_{1i}\to p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(a_{1i}\right)=m_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to a_{1i}}\left(a_{1i}\right)m_{p(\boldsymbol{a}_1)\to a_{1i}}\left(a_{1i}\right)$$

and $E_{q(\boldsymbol{a}_2)}$ is similarly defined with $q^*(a_{2i})$ being proportional to

$$m_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to a_{2i}}(a_{2i})\, m_{a_{2i}\to p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)} = m_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to a_{2i}}(a_{2i})\, m_{p(\boldsymbol{a}_2|\nu)\to a_{2i}}(a_{2i}).$$

Applying (1.16), the message from factor $p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)$ to node σ^2 update involves expectation with respect to the normalization of

$$\begin{aligned} &m_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\boldsymbol{\theta}}(\boldsymbol{\theta})\, m_{\boldsymbol{\theta}\to p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)}(\boldsymbol{\theta})\, m_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\lambda}(\lambda) \\ &\times m_{\lambda\to p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)}(\lambda) \times \prod_{i=1}^{n}\left\{ m_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to a_{1i}}(a_{1i}) \right. \\ &\left. \times\ m_{a_{1i}\to p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)}(a_{1i})\, m_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to a_{2i}}(a_{2i})\, m_{a_{2i}\to p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)}(a_{2i}) \right\}. \end{aligned} \tag{C.13}$$

Writing $\log p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)$ as a function of σ^2 in terms of the sufficient statistic vector we have

$$\begin{aligned} \log p\left(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2\right) =& -\frac{n}{2}\log\left(\sigma^2\right) - \frac{1+\lambda^2}{2\sigma^2}\left(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\theta}-\frac{\sigma\lambda\left|\boldsymbol{a}_1\right|\odot\sqrt{\boldsymbol{a}_2}}{\sqrt{1+\lambda^2}}\right)^T \\ &\times \operatorname{diag}\left(\boldsymbol{a}_2\right)^{-1}\left(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\theta}-\frac{\sigma\lambda\left|\boldsymbol{a}_1\right|\odot\sqrt{\boldsymbol{a}_2}}{\sqrt{1+\lambda^2}}\right)+\text{const} \\ =& -\frac{n}{2}\log\left(\sigma^2\right) - \frac{1+\lambda^2}{2\sigma^2}\left(\left(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\theta}\right)^T\operatorname{diag}\left(\boldsymbol{a}_2\right)^{-1}\left(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\theta}\right)\right. \\ &\left.-2\left(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\theta}\right)^T\operatorname{diag}\left(\boldsymbol{a}_2\right)^{-1}\frac{\sigma\lambda\left|\boldsymbol{a}_1\right|\odot\sqrt{\boldsymbol{a}_2}}{\sqrt{1+\lambda^2}}\right)+\text{const} \\ =& \begin{bmatrix} \log\left(\sigma^2\right) \\ 1/\sigma \\ 1/\sigma^2 \end{bmatrix}^T \begin{bmatrix} -n/2 \\ \\ \lambda\sqrt{1+\lambda^2}\left(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\theta}\right)^T \\ \times\operatorname{diag}\left(\sqrt{\boldsymbol{a}_2}\right)^{-1}\left|\boldsymbol{a}_1\right| \\ \\ -\frac{1}{2}\left(1+\lambda^2\right)\left(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\theta}\right)^T \\ \times\operatorname{diag}\left(\boldsymbol{a}_2\right)^{-1}\left(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\theta}\right) \end{bmatrix} \\ &+\text{const}. \end{aligned}$$

Therefore, application of formula (1.15) and expectation with respect to (C.13) show that the message $m_{p(y|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\sigma^2}$ is proportional to an inverse square root Nadarajah

density function with natural parameter update

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\sigma^2} \longleftarrow \begin{bmatrix} -n/2 \\ \mu_{q\left(\lambda\sqrt{1+\lambda^2}\right)} \left\{\boldsymbol{y} - \boldsymbol{A}E_{q(\boldsymbol{\theta})}\left(\boldsymbol{\theta}\right)\right\}^T \\ \times \mathrm{diag}\left\{E_{q(\boldsymbol{a}_2)}\left(\mathbf{1}_n/\sqrt{\boldsymbol{a}_2}\right)\right\} E_{q(\boldsymbol{a}_1)}\left|\boldsymbol{a}_1\right| \\ \left(1+\mu_{q(\lambda^2)}\right) G_{\mathrm{VMP}}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\leftrightarrow\boldsymbol{\theta}}; \right. \\ \boldsymbol{A}^T\mathrm{diag}\left\{E_{q(\boldsymbol{a}_2)}\left(\mathbf{1}_n/\boldsymbol{a}_2\right)\right\}\boldsymbol{A}, \boldsymbol{A}^T\mathrm{diag}\left\{E_{q(\boldsymbol{a}_2)}\left(\mathbf{1}_n/\boldsymbol{a}_2\right)\right\}\boldsymbol{y}, \\ \left.\boldsymbol{y}^T\mathrm{diag}\left\{E_{q(\boldsymbol{a}_2)}\left(\mathbf{1}_n/\boldsymbol{a}_2\right)\right\}\boldsymbol{y}\right) \end{bmatrix},$$

where $E_{q(\boldsymbol{\theta})}$ denotes expectation with respect to the normalization of

$$m_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\boldsymbol{\theta}}\left(\boldsymbol{\theta}\right) m_{\boldsymbol{\theta}\to p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(\boldsymbol{\theta}\right).$$

The treatment of $m_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\sigma^2}\left(\sigma^2\right)$ is analogous to that for the messages from the likelihood factor to σ^2 for the asymmetric Laplace distribution in McLean and Wand (2018, Section S.3.3).

Applying (1.16), the message from factor $p\left(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2\right)$ to node λ update involves expectation with respect to the normalization of

$$\begin{aligned} &m_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\boldsymbol{\theta}}\left(\boldsymbol{\theta}\right) m_{\boldsymbol{\theta}\to p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(\boldsymbol{\theta}\right) m_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\sigma^2}\left(\sigma^2\right) \\ &\times m_{\sigma^2\to p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(\sigma^2\right) \prod_{i=1}^{n} \left\{m_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to a_{1i}}\left(a_{1i}\right)\right. \\ &\left.\times\ m_{a_{1i}\to p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(a_{1i}\right) m_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to a_{2i}}\left(a_{2i}\right) m_{a_{2i}\to p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(a_{2i}\right)\right\}. \end{aligned} \tag{C.14}$$

Writing $\log p\left(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2\right)$ as a function of λ in terms of the sufficient statistic vector we have

$$\begin{aligned}
\log p\left(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2\right) =& \frac{n}{2}\log\left(1+\lambda^2\right) - \frac{1+\lambda^2}{2\sigma^2}\left(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\theta}-\frac{\sigma\lambda\left|\boldsymbol{a}_1\right|\odot\sqrt{\boldsymbol{a}_2}}{\sqrt{1+\lambda^2}}\right)^T \\
& \times \operatorname{diag}\left(\boldsymbol{a}_2\right)^{-1}\left(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\theta}-\frac{\sigma\lambda\left|\boldsymbol{a}_1\right|\odot\sqrt{\boldsymbol{a}_2}}{\sqrt{1+\lambda^2}}\right)+\text{const} \\
=& \frac{n}{2}\log\left(1+\lambda^2\right) - \frac{\lambda^2}{2\sigma^2}\left(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\theta}\right)^T\operatorname{diag}\left(\boldsymbol{a}_2\right)^{-1}\left(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\theta}\right) \\
& -\frac{1}{2}\lambda^2\left\|\boldsymbol{a}_1\right\|^2+\frac{\lambda\sqrt{1+\lambda^2}}{\sigma}\left(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\theta}\right)^T\operatorname{diag}\left(\boldsymbol{a}_2\right)^{-1}\left|\boldsymbol{a}_1\right|+\text{const} \\
=& \begin{bmatrix} \log\left(1+\lambda^2\right) \\ \lambda^2 \\ \lambda\sqrt{1+\lambda^2} \end{bmatrix}^T \begin{bmatrix} n/2 \\ -\frac{1}{2}\left\{\frac{1}{\sigma^2}\left(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\theta}\right)^T\operatorname{diag}\left(\boldsymbol{a}_2\right)^{-1} \\ \times\left(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\theta}\right)+\left\|\boldsymbol{a}_1\right\|^2\right\} \\ \frac{1}{\sigma}\left(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\theta}\right)^T\operatorname{diag}\left(\sqrt{\boldsymbol{a}_2}\right)^{-1}\left|\boldsymbol{a}_1\right| \end{bmatrix} \\
& +\text{const}
\end{aligned}$$

where $\|\boldsymbol{v}\|=\sqrt{\boldsymbol{v}^T\boldsymbol{v}}$ for any vector $\boldsymbol{v}$.

It follows from application of (1.15) and expectation with respect to (C.14) that the message $m_{p(y|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\lambda}$ is proportional to density functions within the Sea Sponge exponential family with natural parameter update

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\lambda} \longleftarrow \begin{bmatrix} n/2 \\ \mu_{q(1/\sigma^2)}G_{\text{VMP}}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\leftrightarrow\boldsymbol{\theta}};\right. \\ \boldsymbol{A}^T\operatorname{diag}\left\{E_{q(\boldsymbol{a}_2)}\left(\boldsymbol{1}_n/\boldsymbol{a}_2\right)\right\}\boldsymbol{A},\boldsymbol{A}^T\operatorname{diag}\left\{E_{q(\boldsymbol{a}_2)}\left(\boldsymbol{1}_n/\boldsymbol{a}_2\right)\right\}\boldsymbol{y}, \\ \left.\boldsymbol{y}^T\operatorname{diag}\left\{E_{q(\boldsymbol{a}_2)}\left(\boldsymbol{1}_n/\boldsymbol{a}_2\right)\right\}\boldsymbol{y}\right)-\frac{1}{2}E_{q(\boldsymbol{a}_1)}\left\|\boldsymbol{a}_1\right\|^2 \\ \mu_{q(1/\sigma)}\left\{\boldsymbol{y}-\boldsymbol{A}E_{q(\boldsymbol{\theta})}\left(\boldsymbol{\theta}\right)\right\}^T \\ \times\operatorname{diag}\left\{E_{q(\boldsymbol{a}_2)}\left(\boldsymbol{1}_n/\sqrt{\boldsymbol{a}_2}\right)\right\}E_{q(\boldsymbol{a}_1)}\left|\boldsymbol{a}_1\right| \end{bmatrix},$$

where $\mu_{q\left(1/\boldsymbol{a}_2^k\right)}=\int_0^\infty\left(1/\boldsymbol{a}_2^k\right)q^*\left(\boldsymbol{a}_2\right)d\boldsymbol{a}_2$, for $k=1/2,1$.

Applying (1.16), the message from factor $p\left(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2\right)$ to node a_{1i} update involves expectation with respect to the normalization of

$$\begin{aligned}
&m_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\boldsymbol{\theta}}\left(\boldsymbol{\theta}\right) m_{\boldsymbol{\theta}\to p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(\boldsymbol{\theta}\right) m_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\sigma^2}\left(\sigma^2\right) \\
&\times m_{\sigma^2\to p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(\sigma^2\right) m_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\lambda}\left(\lambda\right) m_{\lambda\to p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(\lambda\right) \\
&\times \prod_{i=1}^{n} m_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to a_{2i}}\left(a_{2i}\right) m_{a_{2i}\to p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(a_{2i}\right).
\end{aligned} \tag{C.15}$$

Writing $\log p\left(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2\right)$ as a function of a_{1i} in terms of the sufficient statistic vector we have

$$\begin{aligned}
\log p\left(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2\right) &= -\frac{1+\lambda^2}{2\sigma^2 a_{2i}}\left(\left(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\theta}\right)_i - \frac{\sigma\lambda\left|a_{1i}\right|\sqrt{a_{2i}}}{\sqrt{1+\lambda^2}}\right)^2 + \text{const} \\
&= \frac{\lambda\sqrt{1+\lambda^2}}{\sigma\sqrt{a_{2i}}}\left(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\theta}\right)_i\left|a_{1i}\right| - \frac{1}{2}\lambda^2 a_{1i}^2 + \text{const} \\
&= \begin{bmatrix} \left|a_{1i}\right| \\ a_{1i}^2 \end{bmatrix}^T \begin{bmatrix} \frac{\lambda\sqrt{1+\lambda^2}}{\sigma\sqrt{a_{2i}}}\left(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\theta}\right)_i \\ -\frac{1}{2}\lambda^2 \end{bmatrix} + \text{const}.
\end{aligned}$$

Application of formula (1.15) and expectation with respect to (C.15) indicates that the natural parameter for the messages from $p\left(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2\right)$ to the a_{1i}, $1\le i\le n$, variables are

$$m_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to a_{1i}}\left(a_{1i}\right) = \exp\left\{\begin{bmatrix} \left|a_{1i}\right| \\ a_{1i}^2 \end{bmatrix}^T \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to a_{1i}}\right\}$$

with natural parameter update

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to a_{1i}} \longleftarrow \begin{bmatrix} \mu_{q(1/\sigma)}\mu_{q\left(\lambda\sqrt{1+\lambda^2}\right)}\left\{E_{q(\boldsymbol{a}_2)}\left(1/\sqrt{\boldsymbol{a}_2}\right)\right\}_i\left\{\boldsymbol{y}-\boldsymbol{A}E_{q(\boldsymbol{\theta})}\left(\boldsymbol{\theta}\right)\right\}_i \\ -\frac{1}{2}\mu_{q(\lambda^2)} \end{bmatrix}.$$

Messages from $p\left(\boldsymbol{a}_1\right)$ to a_{1i}, $1\le i\le n$, are

$$m_{p(\boldsymbol{a}_1)\to a_{1i}}\left(a_{1i}\right) = \exp\left(-\frac{1}{2}a_{1i}^2\right),$$

hence

$$q^*\left(a_{1i}\right) \propto \exp\left\{\begin{bmatrix} \left|a_{1i}\right| \\ a_{1i}^2 \end{bmatrix}^T \begin{bmatrix} \mu_{q(1/\sigma)}\mu_{q\left(\lambda\sqrt{1+\lambda^2}\right)}\left\{E_{q(\boldsymbol{a}_2)}\left(1/\sqrt{\boldsymbol{a}_2}\right)\right\}_i\left\{\boldsymbol{y}-\boldsymbol{A}E_{q(\boldsymbol{\theta})}\left(\boldsymbol{\theta}\right)\right\}_i \\ -\frac{1}{2}\left(1+\mu_{q(\lambda^2)}\right) \end{bmatrix}\right\}.$$

Standard manipulations involving the standard normal density and, in particular, results (A.14) and (A.16) provide expressions for the expectations with respect to $E_{q(\boldsymbol{a}_1)}$

$$E_{q(\boldsymbol{a}_1)}|\boldsymbol{a}_1| = \frac{\boldsymbol{\omega}_3 + \zeta'(\boldsymbol{\omega}_3)}{\sqrt{1+\mu_{q(\lambda^2)}}} \quad \text{and} \quad E_{q(\boldsymbol{a}_1)}\|\boldsymbol{a}_1\|^2 = \frac{n + \mathbf{1}_n^T\left[\boldsymbol{\omega}_3 \odot \left\{\boldsymbol{\omega}_3 + \zeta'(\boldsymbol{\omega}_3)\right\}\right]}{1+\mu_{q(\lambda^2)}},$$

where $\boldsymbol{\omega}_3$ is defined in Algorithm 3.4. The previous expressions involve the first derivative of $\zeta(x) = \log(2\Phi(x))$, that is, $\zeta'(x) = \phi(x)/\Phi(x)$ with ϕ and Φ distribution density function and cumulative distribution function of the standard normal respectively. The function `zeta()` within the `R` package `sn` (Azzalini, 2017) supports stable computation of ζ'.

Applying (1.16), the message from factor $p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)$ to node a_{2i} update involves expectation with respect to the normalization of

$$\begin{aligned}
&m_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\boldsymbol{\theta}}(\boldsymbol{\theta})\, m_{\boldsymbol{\theta}\to p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)}(\boldsymbol{\theta})\, m_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\sigma^2}\left(\sigma^2\right)\\
&\times m_{\sigma^2\to p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(\sigma^2\right) m_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\lambda}(\lambda)\, m_{\lambda\to p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)}(\lambda)\\
&\times \prod_{i=1}^{n} m_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to a_{1i}}(a_{1i})\, m_{a_{1i}\to p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)}(a_{1i}).
\end{aligned} \tag{C.16}$$

Writing $\log p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)$ as a function of a_{2i} in terms of the sufficient statistic vector we have

$$\begin{aligned}
\log p\left(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2\right) &= -\frac{1}{2}\log(a_{2i}) - \frac{1+\lambda^2}{2\sigma^2 a_{2i}}\left((\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\theta})_i - \frac{\sigma\lambda|a_{1i}|\sqrt{a_{2i}}}{\sqrt{1+\lambda^2}}\right)^2 + \text{const}\\
&= -\frac{1}{2}\log(a_{2i}) - \frac{1+\lambda^2}{2\sigma^2 a_{2i}}\left\{(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\theta})_i\right\}^2\\
&\quad + \frac{\lambda\sqrt{1+\lambda^2}}{\sigma\sqrt{a_{2i}}}(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\theta})_i|a_{1i}| + \text{const}\\
&= \begin{bmatrix} \log(a_{2i}) \\ 1/\sqrt{a_{2i}} \\ 1/a_{2i} \end{bmatrix}^T \begin{bmatrix} -1/2 \\ \frac{\lambda\sqrt{1+\lambda^2}}{\sigma}(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\theta})_i|a_{1i}| \\ -\frac{1+\lambda^2}{2\sigma^2}\left\{(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\theta})_i\right\}^2 \end{bmatrix} + \text{const}.
\end{aligned}$$

Application of formula (1.15) and expectation with respect to (C.16) indicates that the messages from $p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)$ to a_{2i}, $1 \le i \le n$, are

$$m_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to a_{2i}}(a_{2i}) = \exp\left\{\begin{bmatrix} \log(a_{2i}) \\ 1/\sqrt{a_{2i}} \\ 1/a_{2i} \end{bmatrix}^T \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to a_{2i}}\right\},$$

with natural parameter update

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to a_{2i}} \longleftarrow \begin{bmatrix} -1/2 \\ \mu_{q(1/\sigma)}\mu_{q\left(\lambda\sqrt{1+\lambda^2}\right)}\left\{\boldsymbol{y}-\boldsymbol{A}E_{q(\boldsymbol{\theta})}\left(\boldsymbol{\theta}\right)\right\}_i\left(E_{q(\boldsymbol{a}_1)}\left|\boldsymbol{a}_1\right|\right)_i \\ -\frac{1}{2}\mu_{q(1/\sigma^2)}\left(1+\mu_{q(\lambda^2)}\right)E_{q(\boldsymbol{\theta})}\left\{\left(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\theta}\right)_i\right\}^2 \end{bmatrix},$$

which is within the inverse square root Nadarajah family, that provides expressions for expectations with respect to $E_{q(\boldsymbol{a}_2)}$ included in the algorithm by conjugacy with

$$m_{p(\boldsymbol{a}_{2i}|\nu)\to a_{2i}}\left(a_{2i}\right) = \exp\left\{\begin{bmatrix}\log\left(a_{2i}\right)\\ 1/a_{2i}\end{bmatrix}^T\begin{bmatrix}-\frac{1}{2}\mu_{q(\nu)}-1\\ -\frac{1}{2}\mu_{q(\nu)}\end{bmatrix}\right\}$$

from the Inverse-χ^2 family.

As a function of ν we have

$$\begin{aligned}\log p\left(\boldsymbol{a}_2|\nu\right) &= n\left\{\left(\nu/2\right)\log\left(\nu/2\right)-\log\left\{\Gamma\left(\nu/2\right)\right\}\right\}-\left(\nu/2\right)\mathbf{1}_n^T\left\{\log\left(\boldsymbol{a}_2\right)+1/\boldsymbol{a}_2\right\}+\text{const}\\ &= \begin{bmatrix}\left(\nu/2\right)\log\left(\nu/2\right)-\log\left\{\Gamma\left(\nu/2\right)\right\}\\ \left(\nu/2\right)\end{bmatrix}^T\begin{bmatrix}n\\ -\mathbf{1}_n^T\left\{\log\left(\boldsymbol{a}_2\right)+1/\boldsymbol{a}_2\right\}\end{bmatrix}+\text{const.}\end{aligned}$$

Therefore the natural parameter vector of the message $m_{p(\boldsymbol{a}_2|\nu)\to\nu}$ involves expectation with respect to $\boldsymbol{a}_2$ only and according to formula (1.15) has form

$$\boldsymbol{\eta}_{p(\boldsymbol{a}_2|\nu)\to\nu} \longleftarrow \begin{bmatrix} n \\ -\mathbf{1}_n^T E_{q(\boldsymbol{a}_2)}\left\{\log\left(\boldsymbol{a}_2\right)+\mathbf{1}_n/\boldsymbol{a}_2\right\}\end{bmatrix},$$

which is proportional to a factor of 2 rescaling of a Moon Rock density function. Imposing conjugacy, the message $m_{\nu\to p(\boldsymbol{a}_2|\nu)}\left(\nu\right)$ is proportional to the same exponential family. It follows that

$$q^*\left(\nu\right)\propto\exp\left\{\begin{bmatrix}\left(\nu/2\right)\log\left(\nu/2\right)-\log\left(\Gamma\left(\nu/2\right)\right)\\ \left(\nu/2\right)\end{bmatrix}^T\boldsymbol{\eta}_{p(\boldsymbol{a}_2|\nu)\leftrightarrow\nu}\right\},$$

which leads to

$$\mu_{q(\nu)} = \int_0^\infty \nu q^*\left(\nu\right)d\nu.$$

Note that the algorithm requires initialization of one of the vectors of expectations involving the two auxiliary variables $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$. In Algorithm 3.4 we choose to initialize

$E_{q(\boldsymbol{a}_2)}\left(\mathbf{1}_n/\sqrt{\boldsymbol{a}_2}\right)$.

C.2.2 Proof of Theorem 3.1

First note that

$$\operatorname{Corr}\left(|a_1|, 1/\sqrt{a_2}\,|\,x\right) = \frac{E\left(|a_1|/\sqrt{a_2}\,|\,x\right) - E\left(|a_1|\,|\,x\right) E\left(1/\sqrt{a_2}\,|\,x\right)}{\sqrt{E\left(a_1^2\,|\,x\right) - E\left(|a_1|\,|\,x\right)^2}\sqrt{E\left(1/a_2\,|\,x\right) - E\left(1/\sqrt{a_2}\,|\,x\right)^2}}. \tag{C.17}$$

We then study single components of the previous expression.

Term $E\left(|a_1|\,|\,x\right)$

Note that

$$E\left(|a_1|\,|\,x\right) = \int_{-\infty}^{\infty} |a_1|\, p\left(a_1\,|\,x\right) da_1 = \frac{1}{p\left(x\right)} \int_{-\infty}^{\infty} p\left(a_2\right) \int_{0}^{\infty} |a_1|\, p\left(x\,|\,a_1, a_2\right) p\left(a_1\right) da_1 da_2$$

and consider the inner integral. With standard manipulations involving the standard normal distribution density function and the cumulative distribution function and using (A.12) we can write

$$\int_{-\infty}^{\infty} |a_1|\, p\left(x\,|\,a_1, a_2\right) p\left(a_1\right) da_1 = \frac{1}{\sqrt{\pi}\sigma_0\sqrt{1+\lambda_0^2}\sqrt{a_2}} \left\{ \frac{1}{\sqrt{\pi}} K_1 + \frac{\sqrt{2}\lambda_0\left(x-\mu_0\right)}{\sigma_0\sqrt{a_2}} K_2 \right\} \tag{C.18}$$

where

$$K_1 = \exp\left\{ -\frac{\left(1+\lambda_0^2\right)\left(x-\mu_0\right)^2}{2\sigma_0^2 a_2} \right\} \quad \text{and} \quad K_2 = \exp\left\{ -\frac{\left(x-\mu_0\right)^2}{2\sigma_0^2 a_2} \right\} \Phi\left\{ \frac{\lambda_0\left(x-\mu_0\right)}{\sigma_0\sqrt{a_2}} \right\}.$$

Hence,

$$\begin{aligned} E\left(|a_1|\,|\,x\right) = & \frac{1}{\sqrt{\pi}\sigma_0\sqrt{1+\lambda_0^2}p\left(x\right)} \left\{ \frac{1}{\sqrt{\pi}} \int_0^{\infty} \frac{1}{\sqrt{a_2}} p\left(a_2\right) K_1 da_2 \right. \\ & \left. + \frac{\sqrt{2}\lambda_0\left(x-\mu_0\right)}{\sigma_0} \int_0^{\infty} \frac{1}{a_2} p\left(a_2\right) K_2 da_2 \right\}. \end{aligned} \tag{C.19}$$

Term $E\left(|a_1|/\sqrt{a_2}\,|\,x\right)$

Note that

$$E\left(\frac{|a_1|}{\sqrt{a_2}}\,\middle|\,x\right) = \int_0^\infty \int_{-\infty}^\infty \frac{|a_1|}{\sqrt{a_2}} p\left(a_1, a_2\,|\,x\right) da_1 da_2$$
$$= \frac{1}{p\left(x\right)} \int_0^\infty \frac{1}{\sqrt{a_2}} p\left(a_2\right) \int_{-\infty}^\infty |a_1|\, p\left(x\,|\,a_1, a_2\right) p\left(a_1\right) da_1 da_2.$$

Using result (C.18) we get

$$E\left(\frac{|a_1|}{\sqrt{a_2}}\,\middle|\,x\right) = \frac{1}{\sqrt{\pi}\sigma_0\sqrt{1+\lambda_0^2 p\left(x\right)}} \left\{ \frac{1}{\sqrt{\pi}} \int_0^\infty \frac{1}{a_2} p\left(a_2\right) K_1 da_2 + \frac{\sqrt{2}\lambda_0\left(x-\mu_0\right)}{\sigma_0} \int_0^\infty \frac{1}{a_2^{3/2}} p\left(a_2\right) K_2 da_2 \right\}. \tag{C.20}$$

Term $E\left(a_1^2\,|\,x\right)$

Note that

$$E\left(a_1^2\,|\,x\right) = \int_{-\infty}^\infty a_1^2 p\left(a_1\,|\,x\right) da_1 = \frac{1}{p\left(x\right)} \int_{-\infty}^\infty p\left(a_2\right) \int_0^\infty a_1^2 p\left(x\,|\,a_1, a_2\right) p\left(a_1\right) da_1 da_2.$$

With standard manipulations involving the standard normal distribution density and cumulative distribution functions and using (A.13) the inner integral in the previous expression becomes

$$\int_{-\infty}^\infty a_1^2 p\left(x\,|\,a_1, a_2\right) p\left(a_1\right) da_1 = \frac{1}{\sqrt{\pi}\sigma_0^2\left(1+\lambda_0^2\right)} \left[\frac{\lambda_0\left(x-\mu_0\right)}{\sqrt{\pi} a_2} K_1 + \frac{\sqrt{2}\left\{\lambda_0^2\left(x-\mu_0\right)^2 + \sigma_0^2 a_2\right\}}{\sigma_0 a_2^{3/2}} K_2 \right].$$

Finally,

$$E\left(a_1^2\,|\,x\right) = \frac{1}{\sqrt{\pi}\sigma_0^2\left(1+\lambda_0^2\right) p\left(x\right)} \left[\frac{\lambda_0\left(x-\mu_0\right)}{\sqrt{\pi}} \int_0^\infty \frac{1}{a_2} p\left(a_2\right) K_1 da_2 + \frac{\sqrt{2}}{\sigma_0} \int_0^\infty \left\{ \frac{\lambda_0^2\left(x-\mu_0\right)^2}{a_2^{3/2}} + \frac{\sigma_0^2}{\sqrt{a_2}} \right\} p\left(a_2\right) K_2 da_2 \right]. \tag{C.21}$$

Term $E\left(1/\sqrt{a_2}\,|\,x\right)$

Note that

$$E\left(\frac{1}{\sqrt{a_2}}\,\middle|\,x\right)=\int_0^\infty\frac{1}{\sqrt{a_2}}p\left(a_2\,|\,x\right)da_2=\frac{1}{p\left(x\right)}\int_{-\infty}^\infty\frac{1}{\sqrt{a_2}}p\left(a_2\right)\int_0^\infty p\left(x\,|\,a_1,a_2\right)p\left(a_1\right)da_1da_2.$$

With standard manipulations involving the standard normal distribution density and cumulative distribution functions and using (A.11) the inner integral in the previous expression becomes

$$\int_0^\infty p\left(x\,|\,a_1,a_2\right)p\left(a_1\right)da_1=\frac{\sqrt{2}}{\sqrt{\pi}\sigma_0\sqrt{a_2}}K_2.$$

It follows that

$$E\left(\frac{1}{\sqrt{a_2}}\,\middle|\,x\right)=\frac{\sqrt{2}}{\sqrt{\pi}\sigma_0p\left(x\right)}\int_0^\infty\frac{1}{a_2}p\left(a_2\right)K_2da_2. \tag{C.22}$$

Term $E\left(1/a_2\,|\,x\right)$

Similarly to term $E\left(1/\sqrt{a_2}\,|\,x\right)$ we get

$$E\left(\frac{1}{a_2}\,\middle|\,x\right)=\frac{\sqrt{2}}{\sqrt{\pi}\sigma_0p\left(x\right)}\int_0^\infty\frac{1}{a_2^{3/2}}p\left(a_2\right)K_2da_2. \tag{C.23}$$

Consider expression (C.17) again. Substituting expressions (C.20)–(C.23) in it and dividing numerator and denominator by $\left\{\sqrt{2}\lambda_0\left(x-\mu_0\right)\right\}\left\{\sqrt{\pi}\sigma_0p\left(x\right)\right\}^{-1}\int_0^\infty\frac{1}{a_2^{3/2}}p\left(a_2\right)K_2da_2$ we get

$$\begin{aligned}&\mathrm{Corr}\left(\left|a_1\right|,1/\sqrt{a_2}\,|\,x=x_0\right)=\\&\left[1+\frac{\sigma_0}{\sqrt{2\pi}\lambda_0\left(x-\mu_0\right)}\frac{C_2}{G_3}-\frac{1}{p\left(x\right)}\left\{\frac{1}{\pi\lambda_0\left(x-\mu_0\right)}\frac{C_1G_2}{G_3}+\frac{\sqrt{2}}{\sqrt{\pi}\sigma_0}\frac{G_2^2}{G_3}\right\}\right]\\&\times\left\{1-\frac{\sqrt{2}}{\sqrt{\pi}\sigma_0p\left(x\right)}\frac{G_2^2}{G_3}\right\}^{-1/2}\times\left[1+\frac{\sigma_0}{\sqrt{2\pi}\lambda_0\left(x-\mu_0\right)}\frac{C_2}{G_3}+\frac{\sigma_0^2}{\lambda_0^2\left(x-\mu_0\right)^2}\frac{G_1}{G_3}\right.\\&\left.-\frac{1}{p\left(x\right)}\left\{\frac{\sigma_0}{\sqrt{2}\pi^{3/2}\lambda_0^2\left(x-\mu_0\right)^2}\frac{C_1^2}{G_3}+\frac{2}{\pi\lambda_0\left(x-\mu_0\right)}\frac{C_1G_2}{G_3}+\frac{\sqrt{2}}{\sqrt{\pi}\sigma_0}\frac{G_2^2}{G_3}\right\}\right]^{-1/2}\end{aligned} \tag{C.24}$$

with

$$C_1 = \int_0^\infty \frac{1}{\sqrt{a_2}} p\left(a_2\right) K_1 da_2 = \left(\frac{\nu_0}{2}\right)^{-\frac{1}{2}} \frac{\Gamma\left\{\left(\nu_0+1\right)/2\right\}}{\Gamma\left(\nu_0/2\right)} \times \left\{1+\frac{\left(1+\lambda_0^2\right)\left(x-\mu_0\right)^2}{\nu_0\sigma_0^2}\right\}^{-\frac{\nu_0+1}{2}}, \tag{C.25}$$

$$C_2 = \int_0^\infty \frac{1}{a_2} p\left(a_2\right) K_1 da_2 = \left\{1+\frac{\left(1+\lambda_0^2\right)\left(x-\mu_0\right)^2}{\nu_0\sigma_0^2}\right\}^{-\frac{\nu_0}{2}-1}, \tag{C.26}$$

$$G_1 = \int_0^\infty \frac{1}{\sqrt{a_2}} p\left(a_2\right) K_2 da_2 < \left(\frac{\nu_0}{2}\right)^{-\frac{1}{2}} \frac{\Gamma\left\{\left(\nu_0+1\right)/2\right\}}{\Gamma\left(\nu_0/2\right)} \left\{1+\frac{\left(x-\mu_0\right)^2}{\nu_0\sigma_0^2}\right\}^{-\frac{\nu_0+1}{2}}, \tag{C.27}$$

$$G_2 = \int_0^\infty \frac{1}{a_2} p\left(a_2\right) K_2 da_2 < \left\{\left(\frac{\nu_0}{2}\right)^{\frac{\nu_0}{2}} \Big/ \Gamma\left(\frac{\nu_0}{2}\right)\right\} \left\{1+\frac{\left(x-\mu_0\right)^2}{\nu_0\sigma_0^2}\right\}^{-\frac{\nu_0}{2}-1}, \tag{C.28}$$

and

$$G_3 = \int_0^\infty \frac{1}{a_2^{3/2}} p\left(a_2\right) K_2 da_2 > \left(\frac{\nu_0}{2}\right)^{-\frac{3}{2}} \frac{\Gamma\left\{\left(\nu_0+3\right)/2\right\}}{\Gamma\left(\nu_0/2\right)} \left\{1+\frac{\left(x-\mu_0\right)^2}{\nu_0\sigma_0^2}\right\}^{-\frac{\nu_0+3}{2}} - \frac{\sigma_0}{\sqrt{2\pi}\lambda_0\left(x-\mu_0\right)} \left\{1+\frac{\left(1+\lambda_0^2\right)\left(x-\mu_0\right)^2}{\nu_0\sigma_0^2}\right\}^{-\frac{\nu_0}{2}-1}. \tag{C.29}$$

The expressions and inequalities in (C.25)–(C.29) are obtained with standard algebra and integration involving the gamma function. In particular, the upper bound for G_1 and G_2 in (C.27) and (C.28) are derived using the fact that $\Phi\left(t\right) > 1-\left(2\pi\right)^{-1/2} t^{-1} e^{-t/2}$, $\forall t \in \mathbb{R}$. The lower bound for G_3 in (C.29) is derived making use of $\Phi\left(t\right) < 1$, $\forall t \in \mathbb{R}$. We can then study the behavior of single components of (C.24) when $|\lambda_0| \to \infty$ using simplifications (C.25)–(C.29) and setting $x = x_0 \in \mathbb{R}$. We then have, for instance,

$$\begin{aligned}\lim_{|\lambda_0|\to\infty} \frac{\sigma_0}{\sqrt{2\pi}\lambda_0\left(x_0-\mu_0\right)} \frac{C_2}{G_3} \leq & \lim_{|\lambda_0|\to\infty} \frac{\sigma_0}{\sqrt{2\pi}\lambda_0\left(x_0-\mu_0\right)} \left\{1+\frac{\left(1+\lambda_0^2\right)\left(x_0-\mu_0\right)^2}{\nu_0\sigma_0^2}\right\}^{-\frac{\nu_0}{2}-1} \\ & \times \left[\left(\frac{\nu_0}{2}\right)^{-\frac{3}{2}} \frac{\Gamma\left\{\left(\nu_0+3\right)/2\right\}}{\Gamma\left(\nu_0/2\right)} \left\{1+\frac{\left(x_0-\mu_0\right)^2}{\nu_0\sigma_0^2}\right\}^{-\frac{\nu_0+3}{2}} \right. \\ & \left. - \frac{\sigma_0}{\sqrt{2\pi}\lambda_0\left(x_0-\mu_0\right)} \left\{1+\frac{\left(1+\lambda_0^2\right)\left(x_0-\mu_0\right)^2}{\nu_0\sigma_0^2}\right\}^{-\frac{\nu_0}{2}-1}\right]^{-1} \\ = & \, 0.\end{aligned}$$

Similar arguments lead to the final expression

$$\lim_{|\lambda_0|\to\infty} \mathrm{Corr}\left(|a_1|, 1/\sqrt{a_2}\,|\, x = x_0\right) = \lim_{|\lambda_0|\to\infty} \frac{1 - \frac{\sqrt{2}}{\sqrt{\pi}\sigma_0 p(x_0)}\frac{G_2^2}{G_3}}{\sqrt{1 - \frac{\sqrt{2}}{\sqrt{\pi}\sigma_0 p(x_0)}\frac{G_2^2}{G_3}}\sqrt{1 - \frac{\sqrt{2}}{\sqrt{\pi}\sigma_0 p(x_0)}\frac{G_2^2}{G_3}}} = 1.$$

C.2.3 Derivation of Algorithm 3.5

We now derive the algorithm based on product density restriction (3.24). The main implications in terms of algebra passing from assumption (3.22) to assumption (3.24) concern the auxiliary variables. Expectations with respect to $E_{q(\boldsymbol{a}_1)}$ and $E_{q(\boldsymbol{a}_2)}$ are replaced by the joint expectation $E_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}$. Following steps similar to those for Algorithm 3.4 we obtain

$$q^*\left(a_{1i}, a_{2i}\right) \propto \exp\left\{\begin{bmatrix} a_{1i}^2 \\ |a_{1i}|/\sqrt{a_{2i}} \\ 1/a_{2i} \\ \log\left(a_{2i}\right) \end{bmatrix}^T \boldsymbol{\eta}_{q(a_{1i},a_{2i})}\right\} = \exp\left\{\begin{bmatrix} a_{1i}^2 \\ |a_{1i}|/\sqrt{a_{2i}} \\ 1/a_{2i} \\ \log\left(a_{2i}\right) \end{bmatrix}^T \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \end{bmatrix}\right\},$$

where we use the shorthand:

$$\begin{aligned}
\eta_1 &= \left(\boldsymbol{\eta}_{q(a_{1i},a_{2i})}\right)_1 = -\frac{1}{2}\left(1 + \mu_{q(\lambda^2)}\right), \\
\eta_2 &= \left(\boldsymbol{\eta}_{q(a_{1i},a_{2i})}\right)_2 = \mu_{q\left(\lambda\sqrt{1+\lambda}\right)}\mu_{q(1/\sigma)}\boldsymbol{\tau}_1, \\
\eta_3 &= \left(\boldsymbol{\eta}_{q(a_{1i},a_{2i})}\right)_3 = \left(1 + \mu_{q(\lambda^2)}\right)\mu_{q(1/\sigma^2)}\boldsymbol{\tau}_2 - \frac{1}{2}\mu_{q(\nu)}, \\
\eta_4 &= \left(\boldsymbol{\eta}_{q(a_{1i},a_{2i})}\right)_4 = -\frac{1}{2}\left(3 + \mu_{q(\nu)}\right).
\end{aligned}$$

Expressions for $\boldsymbol{\tau}_1$ and $\boldsymbol{\tau}_2$ are given in Algorithm 3.5.
Algorithm 3.5 updates include the following sufficient statistic expectations of a bivariate exponential family identified by the superscript MW

$$\begin{aligned}
\left(E\boldsymbol{T}\right)_1^{\mathrm{MW}}\left(\boldsymbol{\eta}_{q(a_{1i},a_{2i})}\right) &= E_{q(a_{1i},a_{2i})}\left(a_{1i}^2\right) = \frac{N_1}{D}, \\
\left(E\boldsymbol{T}\right)_2^{\mathrm{MW}}\left(\boldsymbol{\eta}_{q(a_{1i},a_{2i})}\right) &= E_{q(a_{1i},a_{2i})}\left(|a_{1i}|/\sqrt{a_{2i}}\right) = \frac{N_2}{D}, \\
\left(E\boldsymbol{T}\right)_3^{\mathrm{MW}}\left(\boldsymbol{\eta}_{q(a_{1i},a_{2i})}\right) &= E_{q(a_{1i},a_{2i})}\left(1/a_{2i}\right) = \frac{N_3}{D}, \\
\left(E\boldsymbol{T}\right)_4^{\mathrm{MW}}\left(\boldsymbol{\eta}_{q(a_{1i},a_{2i})}\right) &= E_{q(a_{1i},a_{2i})}\left\{\log\left(a_{2i}\right)\right\} = \frac{N_4}{D}
\end{aligned}$$

where

$$
\begin{aligned}
D &= \int_0^\infty \int_{-\infty}^\infty a_{2i}^{\eta_4} \exp\left(\eta_1 a_{1i}^2 + \eta_2 \frac{|a_{1i}|}{\sqrt{a_{2i}}} + \frac{\eta_3}{a_{2i}}\right) da_{1i} da_{2i}, \\
N_1 &= \int_0^\infty \int_{-\infty}^\infty a_{1i}^2 a_{2i}^{\eta_4} \exp\left(\eta_1 a_{1i}^2 + \eta_2 \frac{|a_{1i}|}{\sqrt{a_{2i}}} + \frac{\eta_3}{a_{2i}}\right) da_{1i} da_{2i}, \\
N_2 &= \int_0^\infty \int_{-\infty}^\infty |a_{1i}|\, a_{2i}^{\eta_4 - \frac{1}{2}} \exp\left(\eta_1 a_{1i}^2 + \eta_2 \frac{|a_{1i}|}{\sqrt{a_{2i}}} + \frac{\eta_3}{a_{2i}}\right) da_{1i} da_{2i}, \\
N_3 &= \int_0^\infty \int_{-\infty}^\infty a_{2i}^{\eta_4 - 1} \exp\left(\eta_1 a_{1i}^2 + \eta_2 \frac{|a_{1i}|}{\sqrt{a_{2i}}} + \frac{\eta_3}{a_{2i}}\right) da_{1i} da_{2i}, \\
N_4 &= \int_0^\infty \int_{-\infty}^\infty a_{2i}^{\eta_4} \log(a_{2i}) \exp\left(\eta_1 a_{1i}^2 + \eta_2 \frac{|a_{1i}|}{\sqrt{a_{2i}}} + \frac{\eta_3}{a_{2i}}\right) da_{1i} da_{2i}.
\end{aligned}
$$

With standard manipulations involving the standard normal distribution density and cumulative distribution functions and, in particular, results (A.11)–(A.13) the previous expressions simplify as follows:

$$
\begin{aligned}
(E\boldsymbol{T})_1^{\mathrm{MW}}\left(\boldsymbol{\eta}_{q(a_{1i},a_{2i})}\right) &= \frac{\eta_2}{4I_1}\left\{\frac{I_2}{\sqrt{\pi}\,(-\eta_1)^{3/2}} + \frac{\eta_2 I_3}{\eta_1^2}\right\} - \frac{1}{2\eta_1}, \\
(E\boldsymbol{T})_2^{\mathrm{MW}}\left(\boldsymbol{\eta}_{q(a_{1i},a_{2i})}\right) &= \frac{1}{2I_1}\left(\frac{I_2}{\sqrt{-\pi\eta_1}} - \frac{\eta_2 I_3}{\eta_1}\right), \\
(E\boldsymbol{T})_3^{\mathrm{MW}}\left(\boldsymbol{\eta}_{q(a_{1i},a_{2i})}\right) &= \frac{I_3}{I_1}, \\
(E\boldsymbol{T})_4^{\mathrm{MW}}\left(\boldsymbol{\eta}_{q(a_{1i},a_{2i})}\right) &= \frac{I_4}{I_1}.
\end{aligned}
$$

where

$$
I_i = \int_0^\infty \{\log(x)\}^{p_i} x^{q_i} e^{r_i/x} \Phi\left(\frac{s_i}{\sqrt{x}}\right) dx, \quad i = 1, \ldots, 4
$$

and

$$
\begin{aligned}
&p_1 = 0, \quad q_1 = \eta_4 < -\frac{3}{2}, \quad r_1 = \eta_3 - \frac{\eta_2^2}{4\eta_1} < 0, \quad s_1 = \frac{\eta_2}{\sqrt{-2\eta_1}} \in \mathbb{R}, \\
&p_2 = 0, \quad q_2 = \eta_4 - \frac{1}{2} < -2, \quad r_2 = \eta_3 < 0, \quad s_2 = \infty, \\
&p_3 = 0, \quad q_3 = \eta_4 - 1 < -\frac{5}{2}, \quad r_3 = \eta_3 - \frac{\eta_2^2}{4\eta_1} < 0, \quad s_3 = \frac{\eta_2}{\sqrt{-2\eta_1}} \in \mathbb{R}, \\
&p_4 = 1, \quad q_4 = \eta_4 < -\frac{3}{2}, \quad r_4 = \eta_3 - \frac{\eta_2^2}{4\eta_1} < 0 \quad \text{and} \quad s_4 = \frac{\eta_2}{\sqrt{-2\eta_1}} \in \mathbb{R}.
\end{aligned}
$$

Integral I_2 has the following simple closed form:

$$
I_2 = (-r_2)^{q_2+1}\,\Gamma(-q_2 - 1).
$$

The integrals I_1, I_3 and I_4 are expressible in closed form in terms of Gaussian hypergeometric functions ${}_2F_1(a, b; c; z)$, making use of results 4.3.8 and 4.3.9 in Ng and Geller (1969). For $i = 1, 3, 4$,

$$\begin{aligned} I_i =& \frac{1}{2}(-r_i)^{q_i+1}\Gamma(-q_i-1) + \frac{s}{\sqrt{2\pi}}(-r_i)^{q_i+\frac{1}{2}}\Gamma\left(-q_i-\frac{1}{2}\right) \\ & \times {}_2F_1\left(\frac{1}{2}, -q_i-\frac{1}{2}; \frac{3}{2}; \frac{s_i^2}{2r_i}\right) \quad \text{if} \quad \left|\frac{s_i^2}{2r_i}\right| < 1, \\ I_i =& \frac{1}{2}(-r_i)^{q_i+1}\Gamma(-q_i-1) + \frac{s}{\sqrt{2\pi}}(-r_i)^{q_i+\frac{1}{2}}\Gamma\left(-q_i-\frac{1}{2}\right) \\ & \times {}_2F_1\left(-q_i-1, -q_i-\frac{1}{2}; -q_i; \frac{2r_i}{s_i^2}\right) \quad \text{if} \quad \left|\frac{s_i^2}{2r_i}\right| > 1. \end{aligned}$$

Evaluation of the Gaussian hypergeometric function is supported by the function `hyperg_2F1` in the `R` package `gsl` (Hankin, 2006) (Hankin, 2006). However, evaluation of ${}_2F_1(a, b; c; z)$ for argument values close to 1 is cumbersome in practical implementations. Therefore numerical integration is necessary when the argument $z = |s_i^2/(2r_i)|$ of the Gaussian hypergeometric function is close to 1.

Efficient numerical integration can be performed via the simple trapezoidal rule (see Appendix B of (Wand *et al.*, 2011)). Working on the log-scale is strongly recommended to avoid underflow and overflow. Integrals I_1, I_3 and I_4 are basically concerned with computation of integrals of the form

$$\mathcal{I} = \int_0^\infty \log(x)^p x^q e^{r/x + \zeta\left(s/\sqrt{x}\right)} dx \tag{C.30}$$

Setting $u = \log x$, expression (C.30)

$$\begin{aligned} \mathcal{I} &= \int_{-\infty}^\infty e^{h(u)} du \\ &= e^{h(u_{\max})} \int_{-\infty}^\infty e^{h(u) - h(u_{\max})} du, \end{aligned}$$

where $h(u) = p\log u + (q+1)u + re^{-u} + \zeta\left(se^{-u/2}\right)$ and $u_{\max} = \max_u h(u)$. First and second order derivatives are

$$\begin{aligned} h'(u) &= \frac{p}{u} + q + 1 - re^{-u} - \frac{s}{2}\zeta'\left(se^{-u/2}\right)e^{-u/2} \\ h''(u) &= -\frac{p}{u^2} + re^{-u} + \frac{s}{4}e^{-u/2}\zeta'\left(se^{-u/2}\right) + \frac{s^2}{4}e^{-u}\zeta''\left(se^{-u/2}\right). \end{aligned}$$

Study of the second order derivative can take advantage of expression

$$\zeta^{''}(x) = -\zeta^{'}(x)\left\{x + \zeta^{'}(x)\right\}$$

appearing in Section 4 of Azzalini and Capitanio (1999) and reveals integrand functions are *log-concave* if $s < 0$ for all other parameter values, which aids numerical integration strategies. If s is positive, log-concavity is guaranteed for certain values of r.
Derivation of Algorithm 3.5 is then analogous to that of Algorithm 3.4 apart for replacement of the components involving auxiliary variables with the previous results on the joint distribution of auxiliary variables.

C.2.4 `rstan` code for fitting skew t regression via MCMC

A description of `Stan` is provided in Subsection B.1.4 of Appendix B. We here describe the `rstan` code for fitting of skew t regression, taking advantage of the formulation 3.21. Similar scripts were employed to test the algorithms for the Pareto and SVR likelihood fragments.
The first step consists in setting the values for burn-in, kept sample size and thinning factor.

```
nWarm <- 10000       # Length of burn-in.
nKept <- 5000        # Size of the kept sample.
nThin <- 5           # Thinning factor.
```

Second, specify the input data: the number of observations, n, the design matrix, $\boldsymbol{A}$, the response vector, y and the hyperparameters, $\sigma_{\boldsymbol{\theta}}$, A_σ, σ_λ, which are associated with the prior distributions defined in the `model` environment. Constrains can also be included. For instance, in our case all hyperparameters must be positive.

```
SkewtRegModel <-
'data
{
  int<lower=1> n;
  vector[n] y;                  matrix[n,2] A;
  real<lower=0> sigmaTheta;     real<lower=0> AsigmaHYP;
  real<lower=0> sigmaLambda;    real<lower=0> nuHYP;
}
```

Then, specify parameters and transformed parameters. The `transformed parameters` environment allows to include the auxiliary variable formulation of the model.

```
parameters
{
  vector[2] theta;      real<lower=0> sigma;
  vector[n] aux1Vec;    vector<lower=0>[n] aux2Vec;
  real lambda;          real<lower=0> nu;
}
transformed parameters
{
  vector[n] meanShift;       vector<lower=0>[n] sigmaAdj;
  for (i in 1:n)
  {
    meanShift[i] = sigma*lambda*fabs(aux1Vec[i])*
                      sqrt(aux2Vec[i])/sqrt(1 + lambda^2);
    sigmaAdj[i] = sigma*sqrt(aux2Vec[i])/sqrt(1 + lambda^2);
  }
}
```

Finally, define the Bayesian model, including prior distributions.

```
  model
{
  for (i in 1:n)
    {
      y[i] ~ poisson(exp(etaVec[i]));
    }
  beta ~ normal(0,sigmaBeta);   u ~ normal(0,sigma);
  sigma ~ cauchy(0,A);
}'
```

The `rstan` code is now ready to be compiled and fit the model on real data.

```
allData <- list(n=n,ncZ=ncZ,y=y,X=X,Z=Z,sigmaBeta=sigmaBeta,A=A)


stanCompilObj <- stan(model_code=PoissRespNPregn,data=allData,
                      iter=1,chains=1)
```

Last, obtain the MCMC samples for each parameter and save the `Stan` output. In our case the $\boldsymbol{\beta}$ vector is of dimension 1×2.

```
stanObj <-  stan(model_code=PoissRespNPregn,data=allData,warmup=nBurnin,
```

```
                iter=(nWarm+nKept),chains=1,thin=nThin,refresh=100,
                fit=stanCompilObj)

betaMCMC <- NULL
for (j in 1:2)
  {
    charVar <- paste("beta[",as.character(j),"]",sep="")
    betaMCMC <- rbind(betaMCMC,extract(stanObj,charVar,permuted=FALSE))
  }

uMCMC <- NULL
for (k in 1:ncZ)
  {
    charVar <- paste("u[",as.character(k),"]",sep="")
    uMCMC <- rbind(uMCMC,extract(stanObj,charVar,permuted=FALSE))
  }
sigmaMCMC <- as.vector(extract(stanObj,"sigma",permuted=FALSE))
```

Appendix D

D.1 Derivations concerning MFVB for Poisson and logistic two-level random effects models

We here provide details about the derivation of Algorithm 4.3 and Result 4.1 concerning MFVB for the Poisson two-level random effects models. Derivations of Algorithm 4.4 and Result 4.2 concerning MFVB for the logistic two-level random effects models can be obtained in a similar way.

D.1.1 Derivation of Algorithm 4.3

Arguments analogous to those given for the Poisson case treated in Algorithm 1 of Luts and Wand (2015) and in the appendix of Nolan *et al.* (2018) for the two-level Gaussian mixed model lead to:

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

where

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} = \left(\boldsymbol{C}^T \boldsymbol{R}^{-1}_{\text{2PMFVB}} \boldsymbol{C} + \boldsymbol{D}_{\text{2PMFVB}}\right)^{-1} \quad \text{and} \quad \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} = \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \left(\boldsymbol{C}^T \boldsymbol{z}_{\text{2PMFVB}} + \boldsymbol{o}_{\text{2PMFVB}}\right),$$

with $\boldsymbol{R}_{\text{2PMFVB}}$, $\boldsymbol{D}_{\text{2PMFVB}}$, $\boldsymbol{z}_{\text{2PMFVB}}$ and $\boldsymbol{o}_{\text{2PMFVB}}$ defined as in (4.13);

$$q(\boldsymbol{\Sigma}) \text{ is an Inverse-G-Wishart } \left(G_{\text{full}}, \xi_{q(\boldsymbol{\Sigma})}, \boldsymbol{\Lambda}_{q(\boldsymbol{\Sigma})}\right) \text{ density function,}$$

where $\xi_{q(\boldsymbol{\Sigma})} = \nu_{\boldsymbol{\Sigma}} + 2q - 2 + m$ and

$$\boldsymbol{\Lambda}_{q(\boldsymbol{\Sigma})} = \boldsymbol{M}_{q\left(\boldsymbol{A}^{-1}_{\boldsymbol{\Sigma}}\right)} + \sum_{i=1}^{m} \left(\boldsymbol{\mu}_{q(\boldsymbol{u}_i)} \boldsymbol{\mu}^T_{q(\boldsymbol{u}_i)} + \boldsymbol{\Sigma}_{q(\boldsymbol{u}_i)}\right),$$

with $\boldsymbol{M}_{q\left(\boldsymbol{\Sigma}^{-1}\right)} = \left(\xi_{q(\boldsymbol{\Sigma})} - q + 1\right)\boldsymbol{\Lambda}^{-1}_{q(\boldsymbol{\Sigma})}$;

$$q\left(\boldsymbol{A}_{\boldsymbol{\Sigma}}\right) \text{ is an Inverse-G-Wishart } \left(G_{\text{diag}}, \xi_{q(\boldsymbol{A}_{\boldsymbol{\Sigma}})}, \boldsymbol{\Lambda}_{q(\boldsymbol{A}_{\boldsymbol{\Sigma}})}\right) \text{ density function,}$$

where $\xi_{q(\boldsymbol{A}_{\boldsymbol{\Sigma}})} = \nu_{\boldsymbol{\Sigma}} + q$ and

$$\boldsymbol{\Lambda}_{q(\boldsymbol{A}_{\boldsymbol{\Sigma}})} = \text{diag}\left\{\text{dg}\left(\boldsymbol{M}_{q\left(\boldsymbol{\Sigma}^{-1}\right)}\right)\right\} + \boldsymbol{\Lambda}_{\boldsymbol{A}_{\boldsymbol{\Sigma}}},$$

with inverse moment $\boldsymbol{M}_{q\left(\boldsymbol{A}^{-1}_{\boldsymbol{\Sigma}}\right)} = \xi_{q(\boldsymbol{A}_{\boldsymbol{\Sigma}})}\boldsymbol{\Lambda}^{-1}_{q(\boldsymbol{A}_{\boldsymbol{\Sigma}})}$. The previous results give rise to a coordinate ascent iterative algorithm, which includes, for instance, Algorithm 1 of Luts and Wand (2015) as a special case of model (4.7). A streamlined version of such an algorithm can be obtained taking advantage of 4.1 and using SolveTwoLevelSparseLeastSquares (Algorithm 4.2) to compute $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}$.

D.1.2 Derivation of Result 4.1

It is easy to see that the $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}$ updates in (4.12) may be written as

$$\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \longleftarrow \left(\boldsymbol{B}^T\boldsymbol{B}\right)^{-1}\boldsymbol{B}^T\boldsymbol{b} = \boldsymbol{A}^{-1}\boldsymbol{a}, \quad \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} \longleftarrow \left(\boldsymbol{B}^T\boldsymbol{B}\right)^{-1} = \boldsymbol{A}^{-1},$$

where $\boldsymbol{B}$ and $\boldsymbol{b}$ are specified as in (4.5), with $\boldsymbol{b}_i$, $\boldsymbol{B}_i$ and $\dot{\boldsymbol{B}}_i$ from (4.15).

D.2 Derivations concerning VMP for Poisson and logistic two-level random effects models

We here provide details about the derivation of Result 4.3.

D.2.1 Derivation of Result 4.3

First note that

$$
\begin{aligned}
q\left(\boldsymbol{\beta}, \boldsymbol{u}\right) &\propto m_{p(\boldsymbol{y}\,|\,\boldsymbol{\beta},\boldsymbol{u})\rightarrow(\boldsymbol{\beta},\boldsymbol{u})}\left(\boldsymbol{\beta}, \boldsymbol{u}\right) m_{(\boldsymbol{\beta},\boldsymbol{u})\rightarrow p(\boldsymbol{y}\,|\,\boldsymbol{\beta},\boldsymbol{u})}\left(\boldsymbol{\beta}, \boldsymbol{u}\right) \\
&= \exp\left\{ \begin{bmatrix} \boldsymbol{\beta} \\ \operatorname{vech}\left(\boldsymbol{\beta}\boldsymbol{\beta}^T\right) \\ \underset{1\leq i\leq m}{\operatorname{stack}} \begin{bmatrix} \boldsymbol{u}_i \\ \operatorname{vech}\left(\boldsymbol{u}_i\boldsymbol{u}_i^T\right) \\ \operatorname{vec}\left(\boldsymbol{\beta}\boldsymbol{u}_i^T\right) \end{bmatrix} \end{bmatrix}^T \boldsymbol{\eta}_{p(\boldsymbol{y}\,|\,\boldsymbol{\beta},\boldsymbol{u})\leftrightarrow(\boldsymbol{\beta},\boldsymbol{u})} \right\} \\
&= \exp\left\{ \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix}^T \boldsymbol{a} - \frac{1}{2} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix}^T \boldsymbol{A} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} \right\},
\end{aligned}
$$

where $\boldsymbol{a}$ and $\boldsymbol{A}$ are given in Result 4.3 and the last step makes use of (4.6). With standard manipulations we obtain

$$
\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} = \boldsymbol{A}^{-1}\boldsymbol{a}, \quad \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} = \boldsymbol{A}^{-1}.
$$

Extraction of the sub-blocks of $\boldsymbol{A}^{-1}\boldsymbol{a}$ and the important sub-blocks of $\boldsymbol{A}^{-1}$ according to (4.14) gives Result 4.3.

Bibliography

Atay-Kayis, A. and Massam, H. (2005) A monte carlo method for computing marginal likelihood in nondecomposable gaussian graphical models. *Biometrika* **92**, 317–335.

Attias, H. (1999) Independent factor analysis. *Neural computation* **11**, 803–851.

Azzalini, A. (2017) The `R` *package 'sn': the skew-normal and related distributions such as the skew-t.* `R` *package version 1.5.1.* `http://azzalini.stat.unipd.it/SN`.

Azzalini, A. and Capitanio, A. (1999) Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61**, 579–602.

Azzalini, A. and Capitanio, A. (2003) Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65**, 367–389.

Bates, D., Maechler, M. and Walker, S. (2018) *Linear mixed-effects models using* `Eigen` *and* `S4`. `R` *package version 1.1.18.1.* `https://cran.r-project.org/web/packages/lme4/lme4.pdf`.

Bennett, K. P. and Campbell, C. (2000) Support vector machines: hype or hallelujah? *Acm Sigkdd Explorations Newsletter* **2**, 1–13.

Bishop, C. M. (2006) *Pattern Recognition and Machine Learning*. New York: Springer.

Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2017) Variational inference: a review for statisticians. *Journal of the American Statistical Association* **112**, 859–877.

Boyd, S. and Vandenberghe, L. (2004) *Convex Optimization*. Cambridge University Press.

Broström, G. (2018) `glmmML`*: generalized linerar models with clustering.* `R` *package version 1.0.3.* `https://cran.r-project.org/web/packages/glmmML/index.html`.

Brown, P. and Zhou, L. (2018) *Generalized linear mixed models with BUGS and JAGS. R package version 2.4.2.* https://cran.r-project.org/web/packages/glmmBUGS/index.html.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. and Riddell, A. (2017) Stan: a probabilistic programming language. *Journal of statistical software* **76**.

Chappell, M., Groves, A. R., Whitcher, B. and Woolrich, M. W. (2009) Variational bayesian inference for a nonlinear forward model. *IEEE Transactions on Signal Processing* **57**, 223–236.

Croissant, Y. (2016) *Ecdat: data sets for econometrics. R package version 0.3.1.* https://CRAN.R-project.org/package=Ecdat.

Dechter, R. and Pearl, J. (1988) Network-based heuristics for constraint-satisfaction problems. *Artificial Intelligence* pp. 370–425.

Frey, B. J., Kschischang, F. R., Loeliger, H. A. and Wiberg, N. (1998) Factor graphs and algorithms. *Proceedings of the 35th Allerton Conference on Communication, Control and Computing 1997* pp. 666–680.

Frey, B. J. and MacKay, D. J. (1998) A revolution: belief propagation in graphs with cycles. *Advances in Neural Information Processing Systems* pp. 479–485.

Frühwirth-Schnatter, S. and Pyne, S. (2010) Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. *Biostatistics* **11**, 317–336.

Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* **1**, 515–533.

Gelman, A. and Hill, J. (2014) *Data Analysis Using Regression and Multilevelhierarchical Models.* New York: Cambridge University Press.

Gentle, J. E. (2007) *Matrix Algebra.* New York: Springer.

Green, P. and Silverman, B. (1994) *Nonparametric Regression and Generalized Linear Models.* New York: Chapman & Hall/CRC Monographs on Statistics & Applied Probability, vol. 58.

Hadfield, J. (2018) *MCMC generalized linear mixed models. R package version 2.26.* https://cran.r-project.org/web/packages/MCMCglmm/MCMCglmm.pdf.

Hall, P., Humphreys, K. and Titterington, D. M. (2002) On the adequacy of variational lower bound functions for likelihood-based inference in markovian models with missing values. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **64**, 549–564.

Hall, P., Ormerod, J. T. and Wand, M. P. (2011a) Theory of gaussian variational approximation for a poisson mixed model. *Statistica Sinica* **21**, 369–389.

Hall, P., Pham, T., Wand, M. P. and Wang, S. S. J. (2011b) Asymptotic normality and valid inference for gaussian variational approximation. *The Annals of Statistics* **39**, 2502–2532.

Hankin, R. K. S. (2006) Special functions in R: introducing the gsl package. *R News* **6**, 24–26.

Hastie, T. and Tibshirani, R. (1990) *Generalized Addititive Models.* London: Chapman and Hall.

Hastie, T., Tibshirani, R., Friedman, J. and Franklin, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* New York: Springer.

Heskes, T. (2003) Stable fixed points of loopy belief propagation are local minima of the bethe free energy. *Advances in Neural Information Processing Systems* pp. 359–366.

Hinton, G. and Van Camp, D. (1993) Keeping the neural networks simple by minimizing the description length of the weights. *Proceedings of the 6th annual conference on Computational Learning Theory* pp. 5–13.

Hoffman, M. D., Blei, D., Wang, C. and Paisley, J. (2013) Stochastic variational inference. *Journal of Machine Learning Research* **14**, 1303–1347.

Huang, A. and Wand, M. P. (2013) Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis* **8**, 439–452.

Hui, F. K. C., You, C., Shang, H. L. and Müller, S. (2018) Semiparametric regression using variational approximations. *Journal of the American Statistical Association (accepted),* pp. xxx–xxx.

Johnson, M. E., Moore, L. M. and Ylvisaker, D. (1990) Minimax and maximin distance designs. *Journal of statistical planning and inference* **26**, 131–148.

Jordan, M. I. (2004) Graphical models. *Statistical Science* **19**, 140–155.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. and Saul, L. K. (1999) An introduction to variational methods for graphical models. *Machine Learning* **37**, 183–233.

Kammann, E. E. and Wand, M. P. (2003) Geoadditive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **52**, 1–18.

Knowles, D. A. and Minka, T. P. (2011) Non-conjugate message passing for multinomial and binary regression. *Advances in Neural Information Processing Systems* pp. 1701–1709.

Kristensen, K. (2018) *TMB: template model builder: a general random effect tool inspired by ADMB. R package version 1.7.14.* `https://cran.r-project.org/web/packages/TMB/index.html`.

Kschischang, F. R., Frey, B. J. and Loeliger, H. A. (2001) Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory* **47**, 498–519.

Kullback, S. and Leibler, R. A. (1951) On information and sufficiency. *The Annals of Mathematical Statistics* **22**, 79–86.

Lasserre, J. B. (2001) Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization* **11**, 796–817.

Li, Y. and Turner, R. E. (2016) Rényi divergence variational inference. *Advances in Neural Information Processing Systems* pp. 1073–1081.

Luts, J. and Ormerod, J. T. (2014) Mean field variational bayesian inference for support vector machine classification. *Computational Stastistics and Data Analysis* **73**, 163–176.

Luts, J. and Wand, M. P. (2015) Variational inference for count response semiparametric regression. *Bayesian Analysis* **10**, 991–1023.

Maestrini, L. and Wand, M. P. (2018) Variational message passing for skew t regression. *Stat* **7**, e196.

Magnus, J. R. and Neudecker, H. (2007) *Matrix Differential Calculus with Applications in Statistics and Econometrics*. New York: John Wiley & Sons.

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models (2nd edition)*. London: Chapman and Hall.

McCulloch, C. E., Searle, S. R. and Neuhaus, J. M. (2008) *Generalized, Linear, and Mixed Models*. New York: Wiley.

McGrory, C. A. and Titterington, D. M. (2007) Variational approximations in bayesian model selection for finite mixture distributions. *Computation Statistics and Data Analysis* **51**, 5352–5367.

McGrory, C. A., Titterington, D. M., Reeves, R. and Pettitt, A. N. (2009) Variational bayes for estimating the parameters of a hidden potts model. *Statistics and Computing* **19**, 329–340.

McLean, M. W. and Wand, M. P. (2018) Variational message passing for elaborate response regression models. *Bayesian Analysis* **13**, 1–28.

Minka, T. P. (2001) Expectation propagation for approximate bayesian inference. *Uncertainty in Artificial Intelligence* pp. 362–369.

Minka, T. P. (2004) Power ep. *Microsoft Research Technical Report Series* pp. 1–6.

Minka, T. P. (2005) Divergence measures and message passing. *Microsoft Research Technical Report Series* **173**, 1–17.

Minka, T. P. and Winn, J. (2008) Gates: a graphical notation for mixture models. *Microsoft Research Technical Report Series* **185**, 1–16.

Monahan, J. F. and Stefanski, L. A. (1989) *Normal Scale Mixture Approximations to $F^*(z)$ and Computation of the Logistic-Normal Integral*. New York: Marcel Dekker.

Nelder, J. A. and Wedderburn, R. W. M. (1972) Generalized linear models. *Journal of the Royal Statistical Society: Series A* **135**, 370–384.

Ng, E. W. and Geller, M. (1969) A table of integrals of the error functions. *Journal of Research of the National Bureau of Standards B* **73**, 1–20.

Nolan, T. H., Menictas, M. and Wand, M. P. (2018) Streamlined computing for variational inference with higher-level random effects. *In preparation,* pp. xxx–xxx.

Nolan, T. H. and Wand, M. P. (2017) Accurate logistic variational message passing: algebraic and numerical details. *Stat* **6**, 102–112.

Nychka, D., Furrer, R., Paige, J. and Sain, S. (2018) *`fields`: tools for spatial data. `R` package version 9.6.* `https://cran.r-project.org/web/packages/fields/index.html`.

Nychka, D. and Saltzman, N. (1998) Design of air-quality monitoring networks. In *Case studies in environmental statistics*, pp. 51–76. New York: Springer.

Opper, M. and Archambeau, C. (2009) Variational gaussian approximation revisited. *Neural Computation* **21**, 786–792.

Ormerod, J. T. and Wand, M. P. (2010) Explaining variational approximations. *The American Statistician* **64**, 140–153.

Ormerod, J. T. and Wand, M. P. (2012) Gaussian variational approximate inference for generalized linear mixed models. *Journal of Computational and Graphical Statistics* **21**, 2–17.

O'Sullivan, F. (1986) A statistical perspective on ill-posed inverse problems. *Statistical Science* **1**, 502–527.

Pace, L. and Salvan, A. (1997) *Principles of Statistical Inference from a Neo-Fisherian Perspective*. Singapore, New Jersey, London, Hong Kong: World Scientific Publishing Company.

Parisi, A. and Liseo, B. (2018) Objective bayesian analysis for the multivariate skew-t model. *Statistical Methods & Applications* **27**, 277–295.

Parisi, G. (1988) *Statistical Field Theory*. Redwood City, CA: Addison-Wesley.

Peterson, C. and Anderson, J. (1987) A mean field theory learning algorithm for neural networks. *Complex Systems* **1**, 995–1019.

Pham, T. H. and Wand, M. P. (2018) Generalized additive mixed model analysis via `gammSlice`. *Australian and New Zealand Journal of Statistics* **60**, 279–300.

Pinheiro, J. C. and Bates, D. M. (2000) *Mixed-Effects Models in S and S-PLUS*. New York: Springer-Verlag.

R Core Team (2018) `R`*: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ripley, B. D. (2012) `MASS`*: functions and datasets to support Venables and Ripley, 'Modern Applied Statistics with S' (4th edition, 2002).* `R` *package version 7.3.50.* `http://CRAN.R-project.org/package=MASS`.

Robbins, H. and Monro, S. (1951) A stochastic approximation method. *The Annals of Mathematical Statistics* **22**, 400–407.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003) *Semiparametric Regression*. New York: Cambridge University Press.

Scheipl, F. and Gruen, B. (2017) *spikeSlabGAM: Bayesian variable selection and model choice for generalized additive mixed models. R package version 1.1.14*. https://cran.r-project.org/web/packages/spikeSlabGAM/index.html.

Schoenberg, I. (1964) Spline functions and the problem of gradation. *Proceedings of the National Academy of Sciences of the United States of America* **52**, 947–950.

Stan Development Team (2018) rstan*: the* R *interface to* Stan*. R package version 2.17.3*. http://mc-stan.org/.

Tan, L. S. L. and Nott, D. J. (2018) Gaussian variational approximation with sparse precision matrices. *Statistics and Computing* **28**, 259–275.

Teschendorff, A. E., Wang, Y., Barbosa-Morais, N. L., Brenton, J. D. and Caldas, C. (2005) A variational bayesian mixture modelling framework for cluster analysis of gene-expression data. *Bioinformatics* **21**, 3025–3033.

Titterington, D. M. (2004) Bayesian methods for neural networks and related models. *Statistical Science* **19**, 128–139.

Tukey, J. W. (1962) The future of data analysis. *The Annals of Mathematical Statistics* **33**, 1–67.

Umlauf, N., Kneib, T., Lang, S. and Zeileis, A. (2017) *R2BayesX: estimate structured additive regression models with BayesX. R package version 1.1.1*. https://cran.r-project.org/web/packages/R2BayesX/index.html.

Wainwright, M. J., Jaakkola, T. S. and Willsky, A. S. (2005) A new class of upper bounds on the log partition function. *IEEE Transactions on Information Theory* **51**, 2313–2335.

Wainwright, M. J. and Jordan, M. I. (2008) Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning* **1**, 1–305.

Wand, M. P. (2002) Vector differential calculus in statistics. *The American Statistician* **56**, 55–62.

Wand, M. P. (2017) Fast approximate inference for arbitrarily large semiparametric regression models via message passing (with discussion). *Journal of the American Statistical Association* **112**, 137–168.

Wand, M. P. and Ormerod, J. T. (2008) On semiparameteric regression with O'Sullivan penalized splines. *Australian & New Zealand Journal of Statistics* **50**, 179–198.

Wand, M. P., Ormerod, J. T., Padoan, S. A. and Frühwirth, R. F. (2011) Mean field variational bayes for elaborate distribution. *Bayesian Analysis* **6**, 847–900.

Wang, B. and Titterington, D. (2003) Local convergence of variational bayes estimators for mixing coefficients. *Preprint,* pp. 1–12.

Wang, B. and Titterington, D. M. (2006) Convergence properties of a general algorithm for calculating variational bayesian estimates for a normal mixture model. *Bayesian Analysis* **1**, 625–650.

Wang, Y. and Blei, D. M. (2018) Frequentist consistency of variational bayes (accepted). *Jornal of the American Statistical Association,* pp. xxx–xxx.

Wiegerinck, W. and Heskes, T. (2003) Fractional belief propagation. *Advances in Neural Information Processing Systems* pp. 438–445.

Winn, J. M. and Bishop, C. M. (2005) Variational message passing. *Journal of Machine Learning Research* **6**, 661–694.

Wood, S. (2018) *Mixed GAM computation vehicle with automatic smoothness estimation. R package version 1.8.24.* `https://cran.r-project.org/web/packages/mgcv/mgcv.pdf`.

Wood, S. and Scheipl, F. (2017) *gamm4: generalized additive mixed models using mgcv and lme4. R package version 0.2.5.* `https://cran.r-project.org/web/packages/gamm4/index.html`.

Wood, S. N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **73**, 3–36.

Yedidia, J. S., Freeman, W. T. and Weiss, Y. (2001) Generalized belief propagation. In *Advances in neural information processing systems*, pp. 689–695.

Yedidia, J. S., Freeman, W. T. and Weiss, Y. (2005) Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory* **51**, 2282–2312.

Zhang, F. and Gao, C. (2018) Convergence rates of variational posterior distributions. *arXiv:1712.02519v3,* pp. 1–58.

Zhao, Y., Staudenmayer, J., Coull, B. A. and Wand, M. P. (2006) General design bayesian generalized linear mixed models. *Statistical Science* **21**, 35–51.

Zhu, J., Chen, N., Perkins, H. and Zhang, B. (2014) Gibbs max-margin topic models with data augmentation. *Journal of Machine Learning Research* **15**, 1073–1110.

Luca Maestrini

CURRICULUM VITAE

Contact Information

University of Padova
Department of Statistical Sciences
Via Cesare Battisti, 241-243
35121 Padova (PD), Italy.
Tel. +39 049 827 4174
e-mail: luca.maestrini@phd.unipd.it

Current Position

Since October 2018
Postdoctoral Research Fellow, School of Mathematical and Physical Sciences, University of Technology Sydney
Supervisor: Prof. Matt P. Wand

Since October 2015
PhD Student in Statistical Sciences, University of Padova
Thesis title: "On Variational Approximations for Frequentist and Bayesian Inference"
Supervisor: Prof. Nicola Sartori
Co-supervisors: Prof. Alessandra Salvan and Prof. Matt P. Wand

Research interests

- Computational statistics
- Theory and methods of inference
- Time series analysis
- Variational approximations

Education

October 2013 – September 2015
Master (*laurea specialistica/magistrale*) degree in Management Engineering and Finance
Polytechnic University of Milan, Milan, Italy
Title of dissertation: "Pairs trading: a new approach based on the BeveridgeNelson decomposition"
Supervisor: Prof. Rocco R. Mosconi
Final mark: 110 cum Laude / 110

March 2015
Athens Programme in Non-linear Mathematical Models and Applications
Universidad Politecnica de Madrid, Madrid, Spain

August 2014 – January 2015
Erasmus Exchange Programme in Industrial Engineering and Management
Royal Institute of Technology, Stockholm, Sweden

September 2010 – July 2013
Bachelor degree (*laurea triennale*) in Management and Production Engineering
Polytechnic University of Milan, Milan, Italy
Final mark: 110 cum Laude / 110

September 2005 – July 2010
High School Diploma
Bramante–Genga Technical Institute, Pesaro, Italy
Final mark: 100 cum Laude / 100

Visiting periods

April 2017 – November 2017 and January 2018 – June 2018
University of Technology Sydney
Sydney, New South Wales, Australia
Supervisor: Prof. Matt P. Wand

Awards and Scholarship

2017
Grant for international cooperation activities, 2000€, University of Padova

2015
Grant for "cum Laude" master graduation, 400€, Banca di Credito Cooperativo del Metauro

2014–2015
Erasmus Programme 2015 grant, European Union, about 1200€

2011–2015
Grant for out of town students with GPA higher than 27/30, about 4000€ per academic year (four academic years), Polytechnic University of Milan

2010–2015
Tuition fees exemption for students with GPA higher than 29/30, about 2500€ per academic year (five academic years), Polytechnic University of Milan

2011
Grant for the best first-year students (students with the highest GPAs), 4000€, Polytechnic University of Milan

2011
Grant for the best high-school student (student with the highest GPA) at Bramante–Genga Technical Institute, 500€, Banca di Gradara

2010
Grant for "cum Laude" high-school diploma, 500€, Italian Ministry of Education

Computer skills

- Good knowledge of R and Stan
- Good knowledge of LaTeX and LyX
- Good knowledge of Microsoft Office
- Basic knowledge of C++
- Basic knowledge of Unix
- Basic knowledge of Matlab, WinBugs, Gretl, Minitab, Lindo and Webratio

Language skills

Italian: native; English: fluent; French: fluent; Mandarin Chinese: basic; Spanish: basic; Swedish: basic

Publications

Articles in journals

Maestrini, L. and Wand, M.P. (2018). Variational message passing for skew t regression. *Stat* **7**, e196

Articles in conference proceedings

Maestrini, L. and Wand, M. P. (2018). Variational approximations for frequentist and Bayesian inference. *Book of short Papers SIS 2018* (Abbruzzo, A., Piacentino, D., Chiodi, M., and Brentari, E., editors). ISBN: 9788891910233

Maestrini, L. and Wand, M. P. (2018). Variational message passing for skew t regression. *Proceedings of the 33rd International Workshop on Statistical Modelling*, Bristol, 204–208

Conference posters

Maestrini, L. and Wand, M. P. (2018). Variational mean field approximations: general principles and pitfalls. *Workshop on Advanced Statistics for Physics Discovery*, Padova

Maestrini, L. and Wand, M. P. (2018). Variational message passing for regression models. *2018 International Society of Bayesian Analysis World Meeting*, Edinburgh

Working papers

Maestrini, L., Ormerod, J. T. and Wand, M. P. Gaussian variational approximate inference for general design generalized linear mixed models. *In preparation*

Maestrini, L., Aykroyd, R. G. and Wand, M. P. Variational inference for inverse problems. *In preparation*

Conference presentations

Maestrini, L. (2018). Variational approximations for frequentist and Bayesian inference (oral presentation). *Australian Research Council Centre of Excellence for Mathematical and Statistical frontiers*

2018 Students and Early Career Researchers Retreat, Torquay, Victoria, Australia, October 29–31

Maestrini, L. and Wand, M. P. (2018). Variational mean field approximations: general principles and pitfalls (poster and three minutes oral presentation). *Workshop on Advanced Statistics for Physics Discovery*, Padova, Italy, September 24–25

Maestrini, L. and Wand, M. P. (2018). Variational message passing for skew t regression (oral presentation). *33rd International Workshop on Statistical Modelling*, Bristol, United Kingdom, July 15–20

Maestrini, L. and Wand, M.P. (2018). Variational message passing for regression models (poster). *2018 International Society of Bayesian Analysis World Meeting*, Edinburgh, United Kingdom, June 24–29

Maestrini, L. and Wand, M. P. (2018). Variational approximations for frequentist and Bayesian inference (oral presentation). *49th Scientific Meeting of the Italian Statistical Society*, Palermo, Italy, June 20–22

Other Interests

Volunteer for AVIS, Italian association active in support of blood donation, and AVO, Italian association coordinating volunteering activities in hospitals

References

Prof. Alessandra Salvan
University of Padova
Department of Statistical Sciences
Via Cesare Battisti, 241-243
35121 Padova (PD), Italy
Phone: +39 049 8274166
e-mail: alessandra.salvan@unipd.it

Prof. Nicola Sartori
University of Padova
Department of Statistical Sciences
Via Cesare Battisti, 241-243
35121 Padova (PD), Italy
Phone: +39 049 8274127
e-mail: nicola.sartori@unipd.it

Prof. Matt P. Wand
University of Technology Sydney
School of Mathematical and Physical Sciences,
Sydney, Australia
PO Box 123 Broadway, NSW 2007
Phone: +61 2 9514 2240
e-mail: matt.wand@uts.edu.au