Streamlined mean field variational Bayes for longitudinal and multilevel data analysis

Cathy Yuen Yi Lee* and **Matt P. Wand**

School of Mathematical and Physical Sciences, University of Technology Sydney, P.O. Box 123, Broadway, New South Wales 2007, Australia

Received 11 January 2015; revised 6 December 2015; accepted 18 December 2015

Streamlined mean field variational Bayes algorithms for efficient fitting and inference in large models for longitudinal and multilevel data analysis are obtained. The number of operations is linear in the number of groups at each level, which represents a two orders of magnitude improvement over the naïve approach. Storage requirements are also lessened considerably. We treat models for the Gaussian and binary response situations. Our algorithms allow the fastest ever approximate Bayesian analyses of arbitrarily large longitudinal and multilevel datasets, with little degradation in accuracy compared with Markov chain Monte Carlo. The modularity of mean field variational Bayes allows relatively simple extension to more complicated scenarios.

Keywords: Bayesian computing; Longitudinal data; Matrix decomposition; Multilevel model; Variational approximations.

Additional supporting information including source code to reproduce the results may be found in the online version of this article at the publisher's web-site

1 Introduction

We develop mean field variational Bayes (MFVB) algorithms for fitting and inference in large longitudinal and multilevel models that are streamlined in terms of number of operations and storage. The number of operations is linear in the number of groups at each level. This constitutes a major improvement over naïve implementations, which have cubic dependence on the number of groups. Our algorithms are also optimal in terms of storage. The inferential accuracy is shown to be very good to excellent. The upshot is the ability to perform high quality Bayesian inference for arbitrarily large longitudinal and multilevel datasets faster than ever before.

MFVB (e.g. Wainwright and Jordan, 2008) is a deterministic alternative to Markov chain Monte Carlo (MCMC) for approximate Bayesian inference. Its utility is speed for high volume data and easy extendability to online processing for high velocity data. This entails some inferential inaccuracy, although this is relatively minor for the models considered in the present article. The modularity of MFVB means that it can accommodate various complications, such as missing data and the need for robustness, within the attractive graphical models framework that has become popular for Bayesian inference (e.g. Bishop, 2008). The software package `Infer.NET` (Minka et al., 2013) has MFVB as one of its main inference engines. `Infer.NET` accommodates the longitudinal and multilevel models treated here, but its nonstreamlined implementation becomes quite slow for large numbers of groups.

*Corresponding author: e-mail: Yuen.Y.Lee@student.uts.edu.au

Variational methods (e.g. Bishop, 2006; Ormerod and Wand, 2010) have evolved mainly in areas such as Machine Learning and Pattern Recognition. However, since the early 2000s there has been an increasing recognition of their usefulness in mainstream Statistics (e.g. Titterington, 2004). An early contribution is Teschendorff et al. (2005), who applied MFVB to mixture modeling for gene expression data. Longitudinal and multilevel models (e.g. Diggle et al., 2002; Gelman and Hill, 2007; Goldstein, 2010; Fitzmaurice et al., 2011) represent a major branch of Statistics which, to date, has had relatively little overlap with the graphical models viewpoint and variational methods. Algorithm 3 of Ormerod and Wand (2010) and Algorithms 3 and 5 of Luts et al. (2014) support MFVB fitting of longitudinal and multilevel data but use naïve matrix inversion for the effects covariance matrix parameter updates.

The essence of our approach is to take advantage of the sparseness of the matrix requiring inversion and streamline its inversion. This involves omitting correlations of estimated effects between groups, which are rarely of interest. Our resultant streamlined algorithms are shown, via simulation, to have inferential accuracy that is comparable with MCMC. It is likely that they provide the fastest ever means of approximate Bayesian analyses of very large longitudinal and multilevel datasets. We restrict attention to two-level Gaussian and binary response models.

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

The next section presents a brief overview of variational approximations. Sections 3 treats Gaussian response longitudinal and multilevel models. Details on the direct implementation of variational algorithms and inference for models are given. Section 4 gives details of our streamlined approach to sparse covariance estimation vis matrix permutation and decomposition. The binary response case is treated in Section 5. In Section 6, we provide numerical evidence of the efficacy of the new methodology, in terms of both inferential accuracy and computation speed. Illustration of two real data examples is presented in Section 7 and concluding remarks are given in Section 8.

2 A brief overview of variational approximations

Consider a generic Bayesian model with observed vector $\boldsymbol{y}$ and parameter vector $\boldsymbol{\theta}$ that is continuous over the parameter space Θ. Bayesian inference is based on finding the joint posterior distribution of parameters of interest given the observed data

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

Let q be an arbitrary density function over Θ. Then the logarithm of the marginal likelihood satisfies $p(\boldsymbol{y}) \geq \underline{p}(\boldsymbol{y}; q)$ is

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

with $\mathrm{KL}\{q(\boldsymbol{\theta}) \,\|\, p(\boldsymbol{\theta}|\boldsymbol{y})\} \geq 0$ for all densities q. The gap between $\log p(\boldsymbol{y})$ and the lower bound $\log \underline{p}(\boldsymbol{y}; q)$ is known as the *Kullback–Leibler divergence* and is minimized by $q_{\text{exact}}(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{y})$, the exact posterior density function. Typically, direct minimization of Kullback–Leibler divergence is intractable since it depends on the true posterior whose normalization constant is difficult to calculate.

For many models of practical interest, exact inference is typically infeasible due to the intractable posterior distributions, requiring approximate inference for use in general. Markov chain Monte Carlo has been the most widely used approximate inference method in this setting, but can be computationally intensive and often suffers from poor mixing that leads to slow convergence in large and complex statistical models. We therefore opt for a fast and analytical alternative approach, namely *variational approximations*.

Variational approximations are a family of approximate inference techniques that aim to recast the problem of computing posterior probabilities as a functional optimization problem, by finding the distribution within a manageable class of distributions that best matches the posterior through optimization of some criterion. We provide here a brief overview of variational approximations and highlight the key concepts. Detailed expositions on the method can be found in Bishop (2006) (Sections 10.1– 10.4) and Ormerod and Wand (2010).

The central idea of variational methods is to approximate the true posterior density function $p(\boldsymbol{\theta}|\boldsymbol{y})$ with a so-called *approximating density function* $q(\boldsymbol{\theta})$ through minimization of a measure of dissimilarity between the two density functions. A natural choice for the dissimilarity measure is the Kullback–Leibler divergence

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

for some set $\mathcal{Q}$ of density functions. Tractability is achieved by restricting $q(\boldsymbol{\theta})$ to some product density form:

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

known as the *mean field restriction*. The resultant $q^*(\boldsymbol{\theta})$ is known as a *mean field variational Bayes approximation* to $p(\boldsymbol{\theta}|\boldsymbol{y})$, and from now on we refer to it as an *optimal q-density*. The tractability afforded by (1) comes with a trade-off in accuracy, because it imposes posterior independence between partition elements $\boldsymbol{\theta}_i$ that may not be present in the true posterior. Depending on the amounts of true posterior dependence, the MFVB approximations can range from excellent to poor. For example, regression coefficients and the error variance in linear regression models tend to have a weak posterior correlation and as a result the assumption of posterior independence has a negligible effect. In contrast, in univariate asymmetric Laplace models, the variance and auxiliary parameters tend to have a strong posterior correlation and so the assumption of posterior independence leads to inaccurate approximate inference (Wand et al., 2011).

Under restriction (1), the optimal q-density functions can be shown to satisfy (Ormerod and Wand, 2010)

$$q_i^*(\boldsymbol{\theta}_i) \propto \exp\left[E_{q(-\boldsymbol{\theta}_i)}\{\log p(\boldsymbol{\theta}_i|\text{rest})\}\right], \quad 1 \leq i \leq M, \tag{2}$$

where $E_{q(-\boldsymbol{\theta}_i)}$ denotes expectation with respect to the density $\prod_{j \neq i} q_j(\boldsymbol{\theta}_j)$, the notation $\text{rest} \equiv \{\boldsymbol{y}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1}, \boldsymbol{\theta}_{i+1}, \ldots, \boldsymbol{\theta}_M\}$ is the set containing the rest of the random vectors in the model, except $\boldsymbol{\theta}_i$, and the distributions $\boldsymbol{\theta}_i|\text{rest}$, $1 \leq i \leq M$ are known as the *full conditionals* in the MCMC literature. We proceed by initializing each of the optimal q-density factors $q_i^*(\boldsymbol{\theta}_i)$ and updating each factor in turn by replacing the current estimate with a revised estimate given in (2). Each update uniquely minimizes the Kullback–Leibler divergence, or equivalently, maximizes the lower bound, with respect to the

parameters of $q_i^*(\boldsymbol{\theta}_i)$. This forms an iterative *coordinate ascent algorithm*, Algorithm 1. Convergence of such an algorithm to at least local optima is guaranteed based on convexity properties (Luenberger and Ye, 2008, p. 253). If all the parameters in a model are conditionally conjugate, then the optimal q-density functions in Algorithm 1 are available in closed form. Otherwise, numerical quadrature techniques are required to estimate the marginal likelihood, which are computationally more demanding.

Algorithm 1. Iterative scheme for obtaining the optimal q-density functions under product restriction (1).

Initialize: $q_1^*(\boldsymbol{\theta}_1), \cdots, q_M^*(\boldsymbol{\theta}_M)$

Cycle:

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

$$\vdots$$

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

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

3 Gaussian response models

The Gaussian response models that we consider all take the following *linear-mixed model* form

$$\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2 \sim N(X\boldsymbol{\beta} + Z\boldsymbol{u}, \sigma_\varepsilon^2 I), \quad \boldsymbol{u} \mid G \sim N(\boldsymbol{0}, G), \tag{3}$$

where X and Z are fixed effects and random effects design matrices, $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are the so-called fixed effects and random effects vectors and G is the random effects covariance matrix. A common extension of (3) involves replacement of $\sigma_\varepsilon^2 I$ by a general covariance matrix R (e.g. Robinson, 1991), but is not treated here.

A very wide range of models can be obtained through various choices of Z and G. We focus on those corresponding to two-level hierarchical structures, which are often treated separately within the broad branches of Statistics known as *longitudinal data analysis* (e.g. Diggle et al., 2002; Fitzmaurice et al., 2011) and *multilevel modeling* (e.g. Gelman and Hill, 2007; Goldstein, 2010). Since around 2000, there has been a major interplay between longitudinal data analysis and semiparametric regression. An overview is given in Fitzmaurice et al. (2008). Our algorithms accommodate many of the models of this type, using the general design matrix set-up of Zhao et al. (2006).

3.1 Sample size and subscript notation

Many of the models used in longitudinal and multilevel data analysis are mathematically equivalent, but use different sample size and subscript notation. It is prudent to delineate the various notational conventions commonly in use so that methodology developed using one set of notation can be transferred to models that use other notational systems. We achieve that in this section by way of a concrete example of (3).

The special case of (3) with

$$X\boldsymbol{\beta} = \mathbf{0}, \quad Z = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad G = \sigma_{\mathrm{R}}^2 I \tag{4}$$

imposes the following covariance matrix on the response vector:

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

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

There are at least two established notations for conveying the hierarchical structure imposed by (4). The first is

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

where

$$\begin{aligned} &y_{ij} \equiv j\text{-th measurement in the } i\text{th group,} \\ &u_i^{\mathrm{R}} \equiv \text{random effect in the } i\text{th group,} \\ &\qquad \text{for } 1 \le i \le m,\ 1 \le j \le n_i. \end{aligned} \tag{5}$$

In the current example, $m = 3$ and $n_1 = 2$, $n_2 = 3$, $n_3 = 2$. This notation is used by prominent longitudinal data analysis textbooks (Diggle et al., 2002; Fitzmaurice et al., 2011), as well as the semiparametric regression book (Ruppert et al., 2003).

An alternative notation is

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

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

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

and

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

This is used in the multilevel model textbook by Gelman and Hill (2007), which is a rehash of the "classification notation" introduced in Browne et al. (2001). Goldstein (2010) uses

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

for two-level models with m groups.

For the remainder of this article we will use the sample size and subscript notation adopted by the above-mentioned longitudinal data analysis textbooks. This involves subscript notation corresponding to (5) and

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

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

3.2 Predictor structure and matrix notation

Our predictor structure corresponds to the set-up notation of Section 2 of Zhao et al. (2006) with the spatial correlation structure omitted. We first partition $\boldsymbol{\beta}$, X, $\boldsymbol{u}$, and Z into random group effects (superscript R) and general components (superscript G) as follows:

$$\boldsymbol{\beta} \equiv \begin{bmatrix} \boldsymbol{\beta}^{\mathrm{R}} \\ \boldsymbol{\beta}^{\mathrm{G}} \end{bmatrix}, \quad X \equiv [X^{\mathrm{R}}\ X^{\mathrm{G}}], \quad \boldsymbol{u} \equiv \begin{bmatrix} \boldsymbol{u}^{\mathrm{R}} \\ \boldsymbol{u}^{\mathrm{G}} \end{bmatrix} \quad \text{and} \quad Z \equiv [Z^{\mathrm{R}}\ Z^{\mathrm{G}}], \tag{6}$$

which leads to

$$X\boldsymbol{\beta} + Z\boldsymbol{u} = X^{\mathrm{R}}\boldsymbol{\beta}^{\mathrm{R}} + X^{\mathrm{G}}\boldsymbol{\beta}^{\mathrm{G}} + Z^{\mathrm{R}}\boldsymbol{u}^{\mathrm{R}} + Z^{\mathrm{G}}\boldsymbol{u}^{\mathrm{G}}. \tag{7}$$

The random group effects design matrices are of the form

$$X^{\mathrm{R}} \equiv \begin{bmatrix} X_1^{\mathrm{R}} \\ \vdots \\ X_m^{\mathrm{R}} \end{bmatrix} \quad \text{and} \quad Z^{\mathrm{R}} \equiv \underset{1\le i\le m}{\text{blockdiag}}(X_i^{\mathrm{R}}), \tag{8}$$

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

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

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

The matrices X^{G} and Z^{G} are general design matrices corresponding to the fixed effects vector $\boldsymbol{\beta}^{\mathrm{G}}$ and random effects vector $\boldsymbol{u}^{\mathrm{G}}$, respectively. Typically, X^{G} contains predictors that are not already included in X^{R}. As explained in Zhao et al. (2006), X^{G} may also contain polynomial basis functions of a continuous predictor that enter the model as a penalized spline. The Z^{G} matrix would then contain spline basis functions of the same predictor. Mixed model-based penalized spline fitting of (7) (e.g. Ruppert et al., 2003) involves modeling a smooth, but otherwise unspecified, function f according to:

$$f(x) = \beta_x x + \sum_{k=1}^{K} u_k^{\mathrm{G}} z_k(x), \quad u_k^{\mathrm{G}} \overset{\text{ind.}}{\sim} N(0, \sigma_u^2), \tag{9}$$

where $\{z_k : 1 \le k \le K\}$ are spline bases of size K and σ_u^2 is the penalized parameter for the spline coefficients $u_1^{\mathrm{G}}, \cdots, u_K^{\mathrm{G}}$. A common choice of the z_k are suitably linearly transformed cubic O'Sullivan splines, as described in Section 4 of Wand and Ormerod (2008). However, as illustrated by Example 3

of Zhao et al. (2006), the $Z^{\rm G}$ may be used for other purposes such as handling crossed random effects. The $Z^{\rm G}\boldsymbol{u}^{\rm G}$ term may be further decomposed according to

$$Z^{\rm G}\boldsymbol{u}^{\rm G} \equiv \sum_{\ell=1}^{L} Z^{\rm G}_{\ell}\boldsymbol{u}^{\rm G}_{\ell} \quad \text{with} \quad \text{Cov}(\boldsymbol{u}^{\rm G}) = \underset{1\le\ell\le L}{\text{blockdiag}}(\sigma^2_{u\ell}\, I_{q^{\rm G}_{\ell}}). \tag{10}$$

Here the $\boldsymbol{u}^{\rm G}_{\ell}$ are $q^{\rm G}_{\ell} \times 1$ random spline coefficient vectors for $1 \le \ell \le L$, so the blocks of $\text{Cov}(\boldsymbol{u}^{\rm G})$ correspond to the decomposition of $\boldsymbol{u}^{\rm G}$. The expression in (10) is in keeping with spline penalization in multipredictor semiparametric regression models (e.g. Ruppert et al., 2003). The random effects covariance matrix, denoted by G in (3), takes the following form for the current model:

$$G = \begin{bmatrix} I_m \otimes \Sigma^{\rm R} & 0 \\ 0 & \underset{1\le\ell\le L}{\text{blockdiag}}\left(\sigma^2_{u\ell} I_{q^{\rm G}_{\ell}}\right) \end{bmatrix}. \tag{11}$$

The examples of Section 2 of Zhao et al. (2006) show the wide variety of models that are encompassed by (7), (8), and (10). Included are the random intercept and slope models commonly used in longitudinal data analysis and the multilevel studies in Chapter 2 of Goldstein (2010). However, various semiparametric regression extensions are also covered – including additive models, varying coefficient models and extensions involving interaction terms and higher dimensional smooth functions.

Our presentation of streamlined MFVB algorithms in Section 4 for models with this predictor structure benefits from some additional notation. It is useful to define

$$C^{\rm G} \equiv [X\ Z^{\rm G}].$$

Lastly, we partition $\boldsymbol{y}$, X, Z, and $C^{\rm G}$ row-wise corresponding to the groups in $X^{\rm R}$:

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_m \end{bmatrix}, \quad X = \begin{bmatrix} X_1 \\ \vdots \\ X_m \end{bmatrix}, \quad Z = \begin{bmatrix} Z_1 \\ \vdots \\ Z_m \end{bmatrix} \quad \text{and} \quad C^{\rm G} = \begin{bmatrix} C^{\rm G}_1 \\ \vdots \\ C^{\rm G}_m \end{bmatrix}. \tag{12}$$

Here $\boldsymbol{y}_i \equiv [y_{i1} \cdots y_{in_i}]^T$ denotes the $n_i \times 1$ vector of responses for the ith group. The matrices X_i, Z_i, and $C^{\rm G}_i$ are defined in the same fashion.

3.3 Prior distributions

In this article, we take a Bayesian approach to model fitting and inference. The prior distribution on the fixed effects vector $\boldsymbol{\beta}$ is taken to be of the form

$$\boldsymbol{\beta} \sim N(\boldsymbol{0}, \sigma^2_{\boldsymbol{\beta}} I_P) \quad \text{for some } \sigma^2_{\boldsymbol{\beta}} > 0, \tag{13}$$

where P is the length of the vector $\boldsymbol{\beta}$ and $\sigma^2_{\boldsymbol{\beta}}$ is a hyperparameter to be specified by the analyst. The extension to general mean and covariance matrices is relatively simple, but we use (13) in our algorithms for simplicity of exposition.

The standard deviation parameters have independent Half–Cauchy priors:

$$\sigma_{u\ell} \overset{\text{ind.}}{\sim} \text{Half}-\text{Cauchy}(A_{u\ell}), \quad 1 \le \ell \le L, \tag{14}$$

where $A_{u\ell} > 0$ are hyperparameters and $\overset{\text{ind.}}{\sim}$ stands for "independently distributed as". Here the $\text{Half}-\text{Cauchy}(A)$ density function is $p(x; A) = 2A/\{\pi(x^2 + A^2)\}$, $x > 0$.

As explained in Gelman (2006), (14) is an attractive means by which weakly informative priors can be imposed on the $\sigma_{u\ell}$s. Result 5 of Wand et al. (2011) leads to the following equivalent distributional statement:

$$\sigma_{u\ell}^2 \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\ (\tfrac{1}{2}, 1/a_{u\ell}), \quad a_{u\ell} \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{1}{2}, 1/A_{u\ell}^2), \tag{15}$$

where, for $A, B > 0$, the Inverse-Gamma (A, B) density function is

$$p(x; A, B) = B^A \Gamma(A)^{-1}\, x^{-A-1} \exp(-B/x), \quad x > 0.$$

We use representation (15) rather than (14) since it is more conducive to MFVB. Similar arguments lead to the prior distribution of σ_ε^2 being

$$\sigma_\varepsilon^2 \sim \text{Inverse-Gamma}\ (\tfrac{1}{2}, 1/a_\varepsilon), \quad a_\varepsilon \sim \text{Inverse-Gamma}\ (\tfrac{1}{2}, 1/A_\varepsilon^2)$$

for $A_\varepsilon > 0$.

For Σ^{R} we use the following extension of (15) (Huang and Wand, 2013):

$$\begin{aligned} \Sigma^{\text{R}} \mid a_1^{\text{R}}, \ldots, a_{q^{\text{R}}}^{\text{R}} &\sim \text{Inverse-Wishart}\left(\nu + q^{\text{R}} - 1,\ 2\,\nu\,\text{diag}(1/a_1^{\text{R}}, \ldots, 1/a_{q^{\text{R}}}^{\text{R}})\right), \\ a_r^{\text{R}} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\ \left(\tfrac{1}{2}, 1/A_{\text{R}r}^2\right), \quad 1 \le r \le q^{\text{R}}, \end{aligned} \tag{16}$$

where $\nu,\ A_{\text{R}r} > 0$ and the Inverse-Wishart(A, B) density function in the $d \times d$ argument X is

$$p(X; A, B) = \mathcal{C}_{A,d}^{-1} |B|^{A/2} |X|^{-(A+d+1)/2} \exp\left\{-\text{tr}\left(BX^{-1}\right)\right\}, \quad X \text{ positive definite},$$

where $A > 0$ and the $d \times d$ matrix B is positive definite. The normalizing factor is

$$\mathcal{C}_{d,A} \equiv 2^{Ad/2} \pi^{d(d-1)/4} \prod_{i=1}^{d} \Gamma\left(\frac{A+1-i}{2}\right). \tag{17}$$

As shown in Huang and Wand (2013), (16) imposes prior distributions on the entries of Σ^{R} that are analogous to (15). The choice $\nu = 2$ corresponds to the standard deviation parameters in Σ^{R} having Half-t distributions with 2 degrees of freedom and the correlation parameters having uniform distributions over $(-1, 1)$.

3.4 Full Bayesian two-level Gaussian response model

Our Bayesian two-level Gaussian response model can now be stated in full:

$$\begin{aligned} \boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2 &\sim N(X\boldsymbol{\beta} + Z\boldsymbol{u}, \sigma_\varepsilon^2 I), \quad \boldsymbol{\beta} \sim N(\boldsymbol{0}, \sigma_{\boldsymbol{\beta}}^2 I_P), \\ \boldsymbol{u} \mid \Sigma^{\text{R}}, \sigma_{u\ell}^2 &\sim N\left(0, \begin{bmatrix} I_m \otimes \Sigma^{\text{R}} & 0 \\ \boldsymbol{0} & \underset{1\le\ell\le L}{\text{blockdiag}}(\sigma_{u\ell}^2 I_{q_\ell^{\text{G}}}) \end{bmatrix}\right), \\ \Sigma^{\text{R}} \mid a_1^{\text{R}}, \ldots, a_{q^{\text{R}}}^{\text{R}} &\sim \text{Inverse-Wishart}\left(\nu + q^{\text{R}} - 1,\ 2\,\nu\,\text{diag}(1/a_1^{\text{R}}, \ldots, 1/a_{q^{\text{R}}}^{\text{R}})\right), \\ a_r^{\text{R}} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\ (\tfrac{1}{2}, A_{\text{R}r}^{-2}), \quad 1 \le r \le q^{\text{R}}, \\ \sigma_\varepsilon^2 \mid a_\varepsilon &\sim \text{Inverse-Gamma}\ (\tfrac{1}{2}, 1/a_\varepsilon), \quad a_\varepsilon \sim \text{Inverse-Gamma}\ (\tfrac{1}{2}, A_\varepsilon^{-2}), \\ \sigma_{u\ell}^2 \mid a_{u\ell} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\ (\tfrac{1}{2}, 1/a_{u\ell}), \\ a_{u\ell} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\ (\tfrac{1}{2}, A_{u\ell}^{-2}), \quad 1 \le \ell \le L. \end{aligned} \tag{18}$$

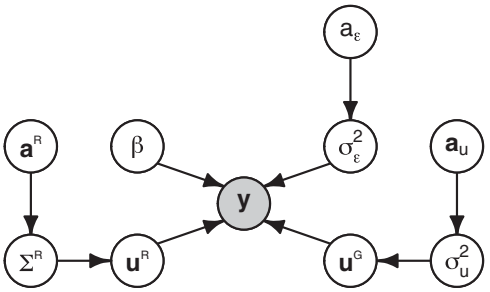

Figure 1 Directed acyclic graph for model (18). The shaded node corresponds to the observed data vector. Random effects and auxiliary variables are referred to as hidden nodes.

Figure 1 shows the directed acyclic graph, as a graph theoretic representation of model (18). The advantages to such graphical representation are twofolds: firstly, it provides a visualization of the hierarchical structure of its corresponding Bayesian model; and secondly, the graph theoretic results can be used to determine probabilistic relationships between nodes. The node $\boldsymbol{a}^{\mathrm{R}}$ corresponds to the random vector $[a_1^{\mathrm{R}} \cdots a_{q^{\mathrm{R}}}^{\mathrm{R}}]^{\mathsf{T}}$. The nodes $\boldsymbol{\sigma}_u^2$ and $\boldsymbol{a}_u$ are defined analogously, that is $\boldsymbol{\sigma}_u^2 \equiv [\sigma_{u1}^2 \cdots \sigma_{uL}^2]^{\mathsf{T}}$ and $\boldsymbol{a}_u \equiv [a_{u1} \cdots a_{uL}]^{\mathsf{T}}$, respectively. The node $\boldsymbol{u}$ is separated into two nodes $\boldsymbol{u}^{\mathrm{R}}$ and $\boldsymbol{u}^{\mathrm{G}}$.

3.5 Mean field variational Bayesian approximate inference

Consider the problem of Bayesian inference for the parameters and random effects in model (18), depicted in Fig. 1. The essence of our MFVB approach involves approximation of the full joint posterior density function of the form

$$p(\boldsymbol{\beta}, \boldsymbol{u}^{\mathrm{R}}, \boldsymbol{u}^{\mathrm{G}}, \boldsymbol{a}^{\mathrm{R}}, \boldsymbol{a}_u, a_\varepsilon, \Sigma^{\mathrm{R}}, \boldsymbol{\sigma}_u^2, \sigma_\varepsilon^2 | \boldsymbol{y}) \approx q(\boldsymbol{\beta}, \boldsymbol{u}^{\mathrm{R}}, \boldsymbol{u}^{\mathrm{G}}, \boldsymbol{a}^{\mathrm{R}}, \boldsymbol{a}_u, a_\varepsilon, \Sigma^{\mathrm{R}}, \boldsymbol{\sigma}_u^2, \sigma_\varepsilon^2) \tag{19}$$

subject to the q-density product density restriction

$$q(\boldsymbol{\beta}, \boldsymbol{u}^{\mathrm{R}}, \boldsymbol{u}^{\mathrm{G}}, \boldsymbol{a}^{\mathrm{R}}, \boldsymbol{a}_u, a_\varepsilon, \Sigma^{\mathrm{R}}, \boldsymbol{\sigma}_u^2, \sigma_\varepsilon^2) = q(\boldsymbol{\beta}, \boldsymbol{u}^{\mathrm{R}}, \boldsymbol{u}^{\mathrm{G}}, \boldsymbol{a}^{\mathrm{R}}, \boldsymbol{a}_u, a_\varepsilon)\, q(\Sigma^{\mathrm{R}}, \boldsymbol{\sigma}_u^2, \sigma_\varepsilon^2). \tag{20}$$

Restriction (20) is the minimal factorization in our MFVB approximation. *Induced product results* (Bishop, 2006, Section 10.2.5) lead to additional factorizations in the q-density product restriction form. Such induced factorizations can be easily detected using simple graphical tests such as *d-separation theory* (e.g. Bishop, 2006, Section 8.2). For example,

$$\Sigma^{\mathrm{R}} \perp\!\!\!\perp \left\{\sigma_u^2, \sigma_\varepsilon^2\right\} \mid \{\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{u}^{\mathrm{R}}, \boldsymbol{u}^{\mathrm{G}}\},$$

where $\boldsymbol{a} \perp\!\!\!\perp \boldsymbol{b} \mid \boldsymbol{c}$ means that $\boldsymbol{a}$ and $\boldsymbol{b}$ are conditionally independent given $\boldsymbol{c}$. Similar arguments lead to the q-densities having the product form:

$$\begin{aligned} &q(\boldsymbol{\beta}, \boldsymbol{u}^{\mathrm{R}}, \boldsymbol{u}^{\mathrm{G}}, \boldsymbol{a}^{\mathrm{R}}, \boldsymbol{a}_u, a_\varepsilon, \Sigma^{\mathrm{R}}, \boldsymbol{\sigma}_u^2, \sigma_\varepsilon^2) = \\ &= q(\boldsymbol{\beta}, \boldsymbol{u}^{\mathrm{R}}, \boldsymbol{u}^{\mathrm{G}})\, q(\Sigma^{\mathrm{R}})\, q(\sigma_\varepsilon^2)\, q(a_\varepsilon) \left\{\prod_{r=1}^{q^{\mathrm{R}}} q(a_r^{\mathrm{R}})\right\} \left\{\prod_{\ell=1}^{L} q(a_{u\ell})\right\} \left\{\prod_{\ell=1}^{L} q(\sigma_{u\ell}^2)\right\}. \end{aligned} \tag{21}$$

As explained in Section 2, the optimal q-densities are chosen to minimize Kullback–Leibler divergence between the right-hand side of (20) and the full joint posterior density function (e.g. Wainwright and Jordan, 2008). Calculations similar to those in Appendix B of Wand and Ormerod (2011) show

that the optimal q-densities admit the following forms:

$$\begin{aligned}
&q^*(\boldsymbol{\beta}, \boldsymbol{u}) \text{ is the } N(\boldsymbol{\mu}_{q(\boldsymbol{\beta},u)}, \Sigma_{q(\boldsymbol{\beta},u)}) \text{ density function,}\\
&q^*(\sigma_\varepsilon^2) \text{ is the Inverse-Gamma}(\tfrac{1}{2}(\textstyle\sum_{i=1}^m n_i + 1), B_{q(\sigma_\varepsilon^2)}) \text{ density function,}\\
&q^*(a_\varepsilon) \text{ is the Inverse-Gamma}(1, B_{q(a_\varepsilon)}) \text{ density function,}\\
&q^*(\sigma_{u\ell}^2) \text{ is the Inverse-Gamma}(\tfrac{1}{2}(q_\ell^{\mathrm{G}} + 1), B_{q(\sigma_{u\ell}^2)}) \text{ density function,}\\
&q^*(a_{u\ell}) \text{ is the Inverse-Gamma}(1, B_{q(a_{u\ell})}) \text{ density function,}\\
&q^*(a_r^{\mathrm{R}}) \text{ is the Inverse-Gamma}(\tfrac{1}{2}(\nu + q^{\mathrm{R}}), B_{q(a_r^{\mathrm{R}})}) \text{ density function and}\\
&q^*(\Sigma^{\mathrm{R}}) \text{ is the Inverse-Wishart}(\nu + m + q^{\mathrm{R}} - 1, B_{q(\Sigma^{\mathrm{R}})}) \text{ density function,}
\end{aligned} \tag{22}$$

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

The q-density parameters are interrelated and their optimal values are obtained via coordinate ascent in Algorithm 2 with notation $C \equiv [X\ Z]$. The stopping criterion is based on the variational lower bound on the marginal likelihood (e.g. Ormerod and Wand, 2010) and denoted here by $\underline{p}(\boldsymbol{y}; q)$. Its logarithm can be obtained by

$$\begin{aligned}
\log \underline{p}(\boldsymbol{y}; q) = E_q\{&\log\, p(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{a}^{\mathrm{R}}, \boldsymbol{a}_u, a_\varepsilon, \Sigma^{\mathrm{R}}, \boldsymbol{\sigma}_u^2, \sigma_\varepsilon^2) + \\
&- \log\, q(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{a}^{\mathrm{R}}, \boldsymbol{a}_u, a_\varepsilon, \Sigma^{\mathrm{R}}, \boldsymbol{\sigma}_u^2, \sigma_\varepsilon^2)\},
\end{aligned} \tag{23}$$

and is presented in Appendix A.

Algorithm 2. Naïve mean field variational Bayes algorithm for the two-level Gaussian response model according to product density restriction (21).

Initialize: $\mu_{q(1/\sigma_\varepsilon^2)} > 0$, $\mu_{q(1/a_\varepsilon)} > 0$, $\mu_{q(1/\sigma_{u\ell}^2)} > 0$, $\mu_{q(1/a_{u\ell})} > 0$, $1 \le \ell \le L$, $\mu_{q(1/a_r^{\mathrm{R}})} > 0$, $1 \le r \le q^{\mathrm{R}}$, $M_{q((\Sigma^R)^{-1})}$ positive definite.

Cycle through updates:

$$\Sigma_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \left(\mu_{q(1/\sigma_\varepsilon^2)}\, C^{\mathsf{T}}C + \begin{bmatrix} \sigma_\beta^{-2} I_P & 0 & 0 \\ 0 & \underset{1\le\ell\le L}{\mathrm{blockdiag}}(\mu_{q(1/\sigma_{u\ell}^2)} I_{q_\ell^{\mathrm{G}}}) & 0 \\ 0 & 0 & I_m \otimes M_{q((\Sigma^R)^{-1})} \end{bmatrix}\right)^{-1}$$

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

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

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

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

$$B_{q(a_r^{\mathrm{R}})} \leftarrow \nu\left(M_{q((\Sigma^R)^{-1})}\right)_{rr} + A_{\mathrm{R}r}^{-2}\ ;\ \mu_{q(1/a_r^{\mathrm{R}})} \leftarrow \tfrac{1}{2}\Big(\nu + q^{\mathrm{R}}\Big)/B_{q(a_r^{\mathrm{R}})}$$

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

$$M_{q((\Sigma^R)^{-1})} \leftarrow (\nu + m + q^{\mathrm{R}} - 1)\, B_{q(\Sigma^{\mathrm{R}})}^{-1}$$

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

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

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

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

4 Streamlining mean field variational Bayes algorithms

A major computational problem arising in Algorithm 2 lies in the update for the covariance matrix of the coefficient estimates:

$$\Sigma_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \left(\mu_{q(1/\sigma_\varepsilon^2)}\, C^{\mathsf{T}}C + \begin{bmatrix} \sigma_\beta^{-2} I_P & 0 & 0 \\ 0 & \underset{1\le\ell\le L}{\text{blockdiag}}\left(\mu_{q(1/\sigma_{u\ell}^2)} I_{q_\ell^{\mathrm{G}}}\right) & 0 \\ 0 & 0 & I_m \otimes M_{q((\Sigma^{\mathrm{R}})^{-1})} \end{bmatrix} \right)^{-1}.$$

This update expression requires storage and inversion of a large sparse matrix of dimension

$$\left(P + q^{\mathrm{R}}\, m + \sum_{\ell=1}^{L} q_\ell^{\mathrm{G}}\right) \times \left(P + q^{\mathrm{R}}\, m + \sum_{\ell=1}^{L} q_\ell^{\mathrm{G}}\right),$$

where P is the number of predictors, q_ℓ^{G} $(1 \le \ell \le L)$ is the size of the ℓ-th spline basis function, typically in the range $15 - 40$ regardless of sample size and m is the number of groups that can be arbitrarily large. For example, the *Six Cities Study of Air Pollution and Health* described in Fitzmaurice et al. (2011) has $m = 13{,}379$. Without doubt the number of groups m dominates the dimension of $\Sigma_{q(\boldsymbol{\beta},\boldsymbol{u})}$ in many practical situations. Henceforth, if $P = 10$, $q^{\mathrm{R}} = 2$, $m = 10{,}000$, $\ell = 1$ and $q_1^{\mathrm{G}} = 25$, then the dimension of the matrix requiring storage and inversion would exceed $20{,}000 \times 20{,}000$. In addition, it is well known from numerical linear algebra that *naïve* computation of $\Sigma_{q(\boldsymbol{\beta},\boldsymbol{u})}$ is $O(m^3)$, that is cubic dependence on the number of groups. Hence, direct implementation of Algorithm 2 can be very costly, or even infeasible, in large longitudinal and multilevel studies. Our aim in the next few sections is to *streamline* the MFVB algorithms by removing computational obstacles involving update expressions such as above for most practical values of m and, thus, maximizing the benefits of variational methods for Bayesian hierarchical models.

To get around the computational problem, we exploit the fact that most of the matrix requiring inversion is block-diagonal. Hence, we propose a streamlined approach-based around block decomposition of a matrix:

$$\begin{aligned} \mu_{q(1/\sigma_\varepsilon^2)}\, C^{\mathsf{T}}C + \begin{bmatrix} \sigma_\beta^{-2} I_P & 0 & 0 \\ 0 & \underset{1\le\ell\le L}{\text{blockdiag}}(\mu_{q(1/\sigma_{u\ell}^2)} I_{q_\ell^{\mathrm{G}}}) & 0 \\ 0 & 0 & I_m \otimes M_{q((\Sigma^{\mathrm{R}})^{-1})} \end{bmatrix} = \\ = \begin{bmatrix} F & G_1 & G_2 & \cdots & G_m \\ G_1^{\mathsf{T}} & H_1^{-1} & 0 & \cdots & 0 \\ G_2^{\mathsf{T}} & 0 & H_2^{-1} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ G_m^{\mathsf{T}} & 0 & 0 & \cdots & H_m^{-1} \end{bmatrix}, \end{aligned} \tag{24}$$

where the submatrices F, G_i, and H_i are defined as follows:

$$\begin{aligned} F &\equiv \mu_{q(1/\sigma_\varepsilon^2)}\, (C^{\mathrm{G}})^{\mathsf{T}} C^{\mathrm{G}} + \begin{bmatrix} \sigma_\beta^{-2} I_P & 0 \\ 0 & \underset{1\le\ell\le L}{\text{blockdiag}}\left(\mu_{q(1/\sigma_{u\ell}^2)} I_{q_\ell^{\mathrm{G}}}\right) \end{bmatrix}, \\ G_i &\equiv \mu_{q(1/\sigma_\varepsilon^2)}\, (C_i^{\mathrm{G}})^{\mathsf{T}} X_i^{\mathrm{R}}, \\ \text{and} \quad H_i &\equiv \left\{ \mu_{q(1/\sigma_\varepsilon^2)}\, (X_i^{\mathrm{R}})^{\mathsf{T}} X_i^{\mathrm{R}} + M_{q((\Sigma^{\mathrm{R}})^{-1})} \right\}^{-1}. \end{aligned}$$

The block-diagonal structure in the bottom right submatrix of (24), the contribution from the random group effects, is crucial as it enables us to reduce the inversion to $O(m)$ operations.

A standard result (e.g. Harville, 2008, Corollary 8.5.12; Smith, 2008) on inversion of a block-partitioned matrix leads to the inverse of (24) equalling

$$\Sigma_{q(\boldsymbol{\beta},\boldsymbol{u})} \equiv \begin{bmatrix} \Sigma_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})} & \Lambda_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}},\boldsymbol{u}_1^{\mathrm{R}}\cdots\boldsymbol{u}_m^{\mathrm{R}})} \\ \Lambda^{\mathsf{T}}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}},\boldsymbol{u}_1^{\mathrm{R}}\cdots\boldsymbol{u}_m^{\mathrm{R}})} & \Sigma_{q(\boldsymbol{u}^{\mathrm{R}})} \end{bmatrix}, \tag{25}$$

where

$$\Sigma_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})} \equiv \left(F - \sum_{i=1}^{m} G_i\, H_i\, G_i^{\mathsf{T}}\right)^{-1}$$

$$\text{and} \quad \Lambda_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}},\boldsymbol{u}_1^{\mathrm{R}}\cdots\boldsymbol{u}_m^{\mathrm{R}})} \equiv \left[-\Sigma_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})}\, G_1\, H_1 \quad \cdots \quad -\Sigma_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})}\, G_m\, H_m\right].$$

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

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

$$\Sigma_{q(\boldsymbol{u}_i^{\mathrm{R}})} \equiv H_i + H_i\, G_i^{\mathsf{T}}\, \Sigma_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})}\, G_i\, H_i, \quad 1 \le i \le m. \tag{26}$$

Since $\Sigma_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})}$, G_i and H_i have dimensions much smaller than m, the complexity of the matrix calculations required in these submatrices does not increase as m increases. Therefore, the calculations with the highest order of complexity are the calculations of the m relevant submatrices of $\Lambda_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}},\boldsymbol{u}_i^{\mathrm{R}})}$ and $\Sigma_{q(\boldsymbol{u}_i^{\mathrm{R}})}$. This renders the whole process as $O(m)$, representing an improvement of order m^2 over the naïve approach to matrix inversion. As the number of groups increases, the improvements due to streamlining become enormous.

4.1 Streamlining update expressions involving $\Sigma_{q(\boldsymbol{\beta},\boldsymbol{u})}$

We see in Algorithm 2 that the update expressions for the mean vector of the coefficient estimates and the scale parameter of the error variance involve the matrix $\Sigma_{q(\boldsymbol{\beta},\boldsymbol{u})}$. Henceforth we now work toward a streamlined alternative to these updates.

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

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

By replacing $\Sigma_{q(\boldsymbol{\beta},\boldsymbol{u})}$ with (25) and partitioning C into $[\,(C^{\mathrm{G}})^{\mathsf{T}} \quad (Z^{\mathrm{R}})^{\mathsf{T}}\,]^{\mathsf{T}}$, we can separate the above update into two expressions corresponding to $(\boldsymbol{\beta}, \boldsymbol{u}^{\mathrm{G}})$ and $\boldsymbol{u}^{\mathrm{R}}$, respectively:

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

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

$$\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})} \leftarrow \mu_{q(1/\sigma_\varepsilon^2)}\, \Sigma_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})} \left\{ (C^{\mathrm{G}})^{\mathsf{T}} \boldsymbol{y} - \sum_{i=1}^{m} G_i\, H_i\, (X_i^{\mathrm{R}})^{\mathsf{T}} \boldsymbol{y}_i \right\} \tag{27}$$

$$\text{and} \quad \boldsymbol{\mu}_{q(\boldsymbol{u}^{\mathrm{R}})} \leftarrow \mu_{q(1/\sigma_\varepsilon^2)} \left\{ \Lambda^{\mathsf{T}}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}},\boldsymbol{u}_1^{\mathrm{R}} \cdots \boldsymbol{u}_m^{\mathrm{R}})}\, (C^{\mathrm{G}})^{\mathsf{T}} \boldsymbol{y} + \Sigma_{q(\boldsymbol{u}^{\mathrm{R}})}\, (Z^{\mathrm{R}})^{\mathsf{T}} \boldsymbol{y} \right\}.$$

Using (25) and (26), we can further break down $\boldsymbol{\mu}_{q(\boldsymbol{u}^{\mathrm{R}})}$ into components corresponding to the i-th group only:

$$\boldsymbol{\mu}_{q(\boldsymbol{u}_i^{\mathrm{R}})} \leftarrow H_i \left\{ \mu_{q(1/\sigma_\varepsilon^2)}\, (X_i^{\mathrm{R}})^{\mathsf{T}} \boldsymbol{y}_i - G_i^{\mathsf{T}} \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})} \right\}.$$

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

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

By replacing $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ with (27) and partitioning C into $[\,(C^{\mathrm{G}})^{\mathsf{T}}\ (Z^{\mathrm{R}})^{\mathsf{T}}\,]^{\mathsf{T}}$, we can rewrite $C\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ and $\mathrm{tr}(C^{\mathsf{T}} C \Sigma_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})})$ in the following equivalent streamlined form:

$$C\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} = C^{\mathrm{G}} \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})} + \begin{bmatrix} X_1^{\mathrm{R}} \boldsymbol{\mu}_{q(\boldsymbol{u}_1^{\mathrm{R}})} \\ \vdots \\ X_m^{\mathrm{R}} \boldsymbol{\mu}_{q(\boldsymbol{u}_m^{\mathrm{R}})} \end{bmatrix}$$

$$\text{and} \quad \mathrm{tr}\left(C^{\mathsf{T}} C \Sigma_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})} \right) = \mathrm{tr}\left\{ (C^{\mathrm{G}})^{\mathsf{T}} C^{\mathrm{G}} \Sigma_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})} \right\} - 2\mu^{-1}_{q(1/\sigma_\varepsilon^2)} \sum_{i=1}^{m} \mathrm{tr}\left(G_i H_i G_i^{\mathsf{T}} \Sigma_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})} \right) +$$

$$+ \sum_{i=1}^{m} \mathrm{tr}\left\{ (X_i^{\mathrm{R}})^{\mathsf{T}} X_i^{\mathrm{R}} \Sigma_{q(\boldsymbol{u}_i^{\mathrm{R}})} \right\},$$

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

We are now ready to present the streamlined algorithm, Algorithm 3, for fast MFVB-approximate fitting and inference in longitudinal and multilevel models with the Gaussian response. We note that the q-density covariance matrix of the vector of coefficients for the i-th group is

$$\mathrm{Cov}_q \left(\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u}^{\mathrm{G}} \\ \boldsymbol{u}_i^{\mathrm{R}} \end{bmatrix} \right) = \begin{bmatrix} \Sigma_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})} & \Lambda_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}},\boldsymbol{u}_i^{\mathrm{R}})} \\ \Lambda^{\mathsf{T}}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}},\boldsymbol{u}_i^{\mathrm{R}})} & \Sigma_{q(\boldsymbol{u}_i^{\mathrm{R}})} \end{bmatrix}. \tag{28}$$

This matrix is needed for variability estimates to accompany group-specific mean estimates. Figure 2 provides an illustration of the fast approximate inference produced by Algorithm 3. The data consist of mathematics test scores of 728 students from 48 schools in inner London, United Kingdom. Details of the data are described in Goldstein (2010). The figure demonstrates the school-specific estimates and pointwise 95% credible sets for the mean 11-year-old mathematics test score by the 8-year-old test score.

Algorithm 3 is an alternative to (restricted) maximum likelihood and best prediction-based fitting of large longitudinal and multilevel models, the latter being employed by popular software such as

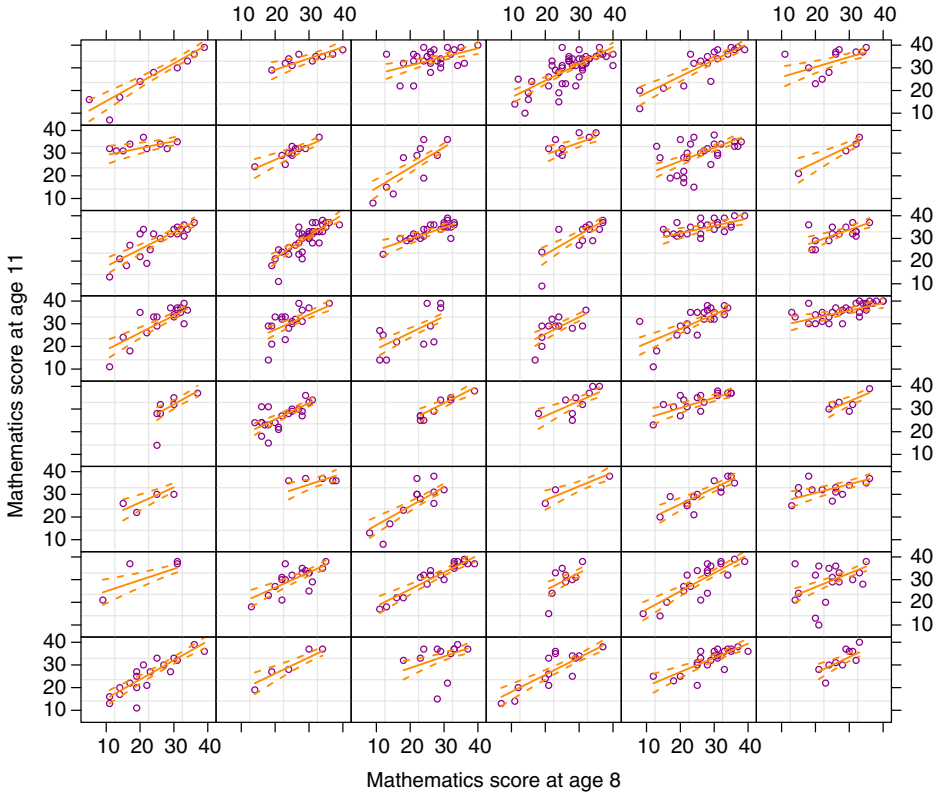

Figure 2 School-specific estimates (solid orange lines) and pointwise 95% credible sets (dotted orange lines) for the mean 11-year-old mathematics test score by the 8-year-old test score. The purple open circles are the raw data.

`PROC MIXED` in SAS (SAS Institute Inc. 2013) and `lmer()` in the R package `lme4` (Douglas et al., 2014). Leaving aside the fact that the latter are based on frequentist inference paradigms, whilst Algorithm 3 is intrinsically Bayesian, it is worth pointing out that Algorithm 3 is purely matrix algebraic that has advantages such as stability, speed, parallelizability, and online fitting. On the other hand, (restricted) maximum likelihood estimation of covariance matrices involves multidimensional, nonlinear optimization (e.g. Wolfinger et al., 1994). Such optimization procedures become computationally burdensome for very large longitudinal and multilevel datasets, as exemplified by the computing times given in Table 1 as part of the speed comparisons given later in Section 6. Earlier covariance estimation

Table 1 Average (standard error) elapsed of the computing times in seconds for fitting of the naïve Algorithm 3, streamlined Algorithm 4, and `gam()` function in the simulation described in the text. The ratio of naïve over streamlined and the ratio of `gamm()` over streamlined are also presented.

m	Naïve	Streamlined	`gamm`	Naïve/Stream.	`gamm`/Stream.
100	0.198 (0.023)	0.027 (0.003)	0.17 (0.010)	7.4	5.8
500	14.482 (1.704)	0.106 (0.021)	7.075 (0.165)	136.8	76.2
2500	1488.757 (108.311)	0.419 (0.056)	739.687 (17.793)	3550.4	1765.4
12, 500	Failed	1.766 (0.096)	Failed	N/A	N/A

procedures, such as the minimum norm quadratic unbiased estimator method of Rao (1972) and the moment-based approach described in Section 3 of Goldstein (1986), are faster but have drawbacks such as positive definiteness not being guaranteed.

Algorithm 3. Streamlined mean field variational Bayes algorithm for the two-level Gaussian response model according to product density restriction (21).

Initialize: $\mu_{q(1/\sigma_\varepsilon^2)} > 0$, $\mu_{q(1/a_\varepsilon)} > 0$, $\mu_{q(1/\sigma_{u\ell}^2)} > 0$, $\mu_{q(1/a_{u\ell})} > 0$, $1 \le \ell \le L$, $\mu_{q(1/a_r^{\rm R})} > 0$, $1 \le r \le q^{\rm R}$, $M_{q((\Sigma^R)^{-1})}$ positive definite.

Cycle through updates:

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

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

$G_i \leftarrow \mu_{q(1/\sigma_\varepsilon^2)} (C_i^{\rm G})^{\mathsf T} X_i^{\rm R}$; $H_i \leftarrow \left\{ \mu_{q(1/\sigma_\varepsilon^2)} (X_i^{\rm R})^{\mathsf T} X_i^{\rm R} + M_{q((\Sigma^R)^{-1})} \right\}^{-1}$

$S \leftarrow S + G_i H_i G_i^{\mathsf T}$; $\boldsymbol{s} \leftarrow \boldsymbol{s} + G_i H_i (X_i^{\rm R})^{\mathsf T} \boldsymbol{y}_i$

$$\Sigma_{q(\boldsymbol{\beta},\boldsymbol{u}^{\rm G})} \leftarrow \left\{ \mu_{q(1/\sigma_\varepsilon^2)} (C^{\rm G})^{\mathsf T} C^{\rm G} + \begin{bmatrix} \sigma_\beta^{-2} I_P & 0 \\ 0 & \underset{1\le\ell\le L}{\rm blockdiag}(\mu_{q(1/\sigma_{u\ell}^2)} I_{q_\ell^{\rm G}}) \end{bmatrix} - S \right\}^{-1}$$

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

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

$\Sigma_{q(\boldsymbol{u}_i^{\rm R})} \leftarrow H_i + H_i\, G_i^{\mathsf T}\, \Sigma_{q(\boldsymbol{\beta},\boldsymbol{u}^{\rm G})}\, G_i\, H_i$

$\boldsymbol{\mu}_{q(\boldsymbol{u}_i^{\rm R})} \leftarrow H_i \left\{ \mu_{q(1/\sigma_\varepsilon^2)} (X_i^{\rm R})^{\mathsf T} \boldsymbol{y}_i - G_i^{\mathsf T} \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\rm G})} \right\}$

$$B_{q(\sigma_\varepsilon^2)} \leftarrow \mu_{q(1/a_\varepsilon)} + \tfrac{1}{2} \left[\left\| \boldsymbol{y} - C^{\rm G} \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\rm G})} - \begin{bmatrix} X_1^{\rm R} \boldsymbol{\mu}_{q(\boldsymbol{u}_1^{\rm R})} \\ \vdots \\ X_m^{\rm R} \boldsymbol{\mu}_{q(\boldsymbol{u}_m^{\rm R})} \end{bmatrix} \right\|^2 + \right.$$
$$+ {\rm tr}\{(C^{\rm G})^{\mathsf T} C^{\rm G} \Sigma_{q(\boldsymbol{\beta},\boldsymbol{u}^{\rm G})}\} + \textstyle\sum_{i=1}^m {\rm tr}\{(X_i^{\rm R})^{\mathsf T} X_i^{\rm R} \Sigma_{q(\boldsymbol{u}_i^{\rm R})}\} +$$
$$\left. - 2\mu_{q(1/\sigma_\varepsilon^2)}^{-1} \textstyle\sum_{i=1}^m {\rm tr}(G_i H_i G_i^{\mathsf T} \Sigma_{q(\boldsymbol{\beta},\boldsymbol{u}^{\rm G})}) \right]$$

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

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

$B_{q(a_r^{\rm R})} \leftarrow \nu (M_{q((\Sigma^R)^{-1})})_{rr} + A_{{\rm R}r}^{-2}$; $\mu_{q(1/a_r^{\rm R})} \leftarrow \frac{1}{2}(\nu + q^{\rm R})/B_{q(a_r^{\rm R})}$

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

$$M_{q((\Sigma^R)^{-1})} \leftarrow (\nu + m + q^{\rm R} - 1)\, B_{q(\Sigma^{\rm R})}^{-1}$$

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

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

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

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

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

$$\Lambda_{q(\boldsymbol{\beta},\boldsymbol{u}^{\rm G},\boldsymbol{u}_i^{\rm R})} \equiv E_q \left[\left(\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u}^{\rm G} \end{bmatrix} - \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\rm G})} \right) \left(\boldsymbol{u}_i^{\rm R} - \boldsymbol{\mu}_{q(\boldsymbol{u}_i^{\rm R})} \right)^{\mathsf T} \right] \leftarrow -\Sigma_{q(\boldsymbol{\beta},\boldsymbol{u}^{\rm G})} G_i\, H_i$$

5 Binary response models

We now consider the case where the entries of $\boldsymbol{y}$ are binary and, instead of (3), the *logistic-mixed model*

$$\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{u} \sim \text{Bernoulli}\Big(\text{logit}^{-1}(X\boldsymbol{\beta} + Z\boldsymbol{u})\Big), \quad \boldsymbol{u} \mid G \sim N(\boldsymbol{0}, G) \tag{29}$$

is appropriate. For a general random vector $\boldsymbol{v}$, the notation $\boldsymbol{v} \sim \text{Bernoulli}(\boldsymbol{p})$ used in (29) is shorthand for the entries of $\boldsymbol{v}$ having independent Bernoulli distributions with parameters corresponding to the entries of $\boldsymbol{p}$.

The notational infrastructure described in Sections 3.1–3.3 for the Gaussian response model can be transferred to the binary response case. The only change is the response distribution specification and the omission of σ_ε^2 and a_ε. The full Bayesian two-level binary response model is then

$$\begin{aligned}
\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{u} &\sim \text{Bernoulli}\Big(\text{logit}^{-1}(X\boldsymbol{\beta} + Z\boldsymbol{u})\Big), \quad \boldsymbol{\beta} \sim N(\boldsymbol{0}, \sigma_{\boldsymbol{\beta}}^2 I_P), \\
\boldsymbol{u} \mid \Sigma^{\text{R}}, \sigma_{u\ell}^2 &\sim N\left(\boldsymbol{0}, \begin{bmatrix} I_m \otimes \Sigma^{\text{R}} & 0 \\ 0 & \underset{1\le\ell\le L}{\text{blockdiag}}(\sigma_{u\ell}^2 I_{q_\ell^{\text{G}}}) \end{bmatrix}\right), \\
\Sigma^{\text{R}} \mid a_1^{\text{R}}, \ldots, a_{q^{\text{R}}}^{\text{R}} &\sim \text{Inverse-Wishart}\left(\nu + q^{\text{R}} - 1,\ 2\,\nu\,\text{diag}(1/a_1^{\text{R}}, \ldots, 1/a_{q^{\text{R}}}^{\text{R}})\right), \\
a_r^{\text{R}} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\ (\tfrac{1}{2}, A_{\text{R}r}^{-2}), \quad 1 \le r \le q^{\text{R}}, \\
\sigma_{u\ell}^2 \mid a_{u\ell} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\ (\tfrac{1}{2}, 1/a_{u\ell}), \\
a_{u\ell} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\ (\tfrac{1}{2}, A_{u\ell}^{-2}), \quad 1 \le \ell \le L.
\end{aligned} \tag{30}$$

The journey toward a practical MFVB algorithm commences with the q-density product density restriction

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

analogous to that given by (20). As in the Gaussian case, induced product results lead to the q-densities satisfying:

$$q(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{a}^{\text{R}}, \boldsymbol{a}_u, \Sigma^{\text{R}}, \boldsymbol{\sigma}_u^2) = q(\boldsymbol{\beta}, \boldsymbol{u})\, q(\Sigma^{\text{R}}) \left\{\prod_{r=1}^{q^{\text{R}}} q(a_r^{\text{R}})\right\} \left\{\prod_{\ell=1}^{L} q(\sigma_{u\ell}^2)\right\} \left\{\prod_{\ell=1}^{L} q(a_{u\ell})\right\}. \tag{31}$$

The optimal q-densities for Σ^{R}, $\boldsymbol{a}^{\text{R}}$, $\boldsymbol{\sigma}_u^2$, and $\boldsymbol{a}_u$ take the same form as the Gaussian response case given in (22). However, the optimal q-density for $(\boldsymbol{\beta}, \boldsymbol{u})$ satisfies

$$q^*(\boldsymbol{\beta}, \boldsymbol{u}) \propto \exp\left\{\boldsymbol{y}^{\mathsf{T}}(X\boldsymbol{\beta} + Z\boldsymbol{u}) - \boldsymbol{1}^{\mathsf{T}}\log(\boldsymbol{1} + e^{X\boldsymbol{\beta}+Z\boldsymbol{u}}) - \frac{1}{2\sigma_{\boldsymbol{\beta}}^2}\|\boldsymbol{\beta}\|^2 - \tfrac{1}{2}\boldsymbol{u}^{\mathsf{T}}G^{-1}\boldsymbol{u}\right\}, \tag{32}$$

where G is given by (11). Unfortunately, (32) does not lend itself to convenient approximate inference for $(\boldsymbol{\beta}, \boldsymbol{u})$ since, for example, the mean and covariance matrix of $q^*(\boldsymbol{\beta}, \boldsymbol{u})$ are not available in closed form. Instead, we replace (32) by a member of the following family of Multivariate Normal approximations:

$$\underline{q}^*(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\xi}) \sim N(\boldsymbol{\mu}_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}, \Sigma_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}),$$

where

$$\Sigma_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})} \equiv \Big[2C^{\mathsf{T}}\text{diag}\{\lambda(\boldsymbol{\xi})\}\, C + \text{blockdiag}\{\sigma_{\boldsymbol{\beta}}^{-2} I_P, G^{-1}\}\Big]^{-1}, \tag{33}$$

where $C \equiv [X \;\; Z]$, $\boldsymbol{\xi}$ is an $\left(\sum_{i=1}^m n_i\right) \times 1$ vector of positive *variational* parameters, $\lambda(x) \equiv \tanh(x/2)/(4\,x)$ and

$$\boldsymbol{\mu}_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})} \equiv \Sigma_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}\, C^{\mathsf{T}} \left(\boldsymbol{y} - \frac{1}{2}\mathbf{1}\right).$$

Scalar functions applied to vectors are evaluated element-wise.

The justification for this family of approximations is given in Tipping (1999b), Jaakkola and Jordan (2000), and Section 3.1 of Ormerod and Wand (2010). Jaakkola and Jordan (2000) also show that the optimal variational parameter vector can be obtained via the update

$$\boldsymbol{\xi} \leftarrow \sqrt{\text{diagonal}[C\,\{\Sigma_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u};\,\xi))} + \boldsymbol{\mu}_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u};\xi)}\,\boldsymbol{\mu}^{T}_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u};\,\xi)}\}\,C^{\mathsf{T}}]},$$

where

$$\text{diagonal}(M) \equiv \text{vector containing the diagonal entries of } M$$

for any square matrix M.

As in the Gaussian case, (33) becomes infeasible for very large longitudinal and multilevel problems, so there is a strong imperative for streamlining the corresponding update. Algorithm 4 achieves this and requires an additional matrix notation:

$$\boldsymbol{\xi}_i = \text{subvector of } \boldsymbol{\xi} \text{ corresponding to the } i\text{-th group}, \quad 1 \le i \le m,$$

which is analogous to the $\boldsymbol{y}_i$ notation given by (12). Again, the stopping criterion for Algorithm 4 is based on the variational lower bound on the marginal likelihood that is presented in Appendix B.

Algorithm 4. Streamlined MFVB algorithm for the two-level binary response model according to product density restriction (31).

Initialize: $\mu_{q(1/\sigma^2_{u\ell})} > 0, \mu_{q(1/a_{u\ell})} > 0, 1 \le \ell \le L, \mu_{q(1/a^{\mathrm{R}}_r)} > 0, 1 \le r \le q^{\mathrm{R}}, M_{q((\Sigma^R)^{-1})}$ positive definite, $\boldsymbol{\xi}$ $(\sum_{i=1}^m n_i) \times 1$ vector of positive entries.

Cycle through updates:

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

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

$G_i \leftarrow 2(C^{\mathrm{G}}_i)^{\mathsf{T}}\text{diag}\{\lambda(\boldsymbol{\xi}_i)\}X^{\mathrm{R}}_i$

$H_i \leftarrow \left\{2(X^{\mathrm{R}}_i)^{\mathsf{T}}\text{diag}\{\lambda(\boldsymbol{\xi}_i)\}X^{\mathrm{R}}_i + M_{q((\Sigma^R)^{-1})}\right\}^{-1}$

$S \leftarrow S + G_i H_i G_i^{\mathsf{T}}\,;\, \boldsymbol{s} \leftarrow \boldsymbol{s} + G_i H_i \left(X^{\mathrm{R}}_i\right)^{\mathsf{T}} (\boldsymbol{y}_i - \frac{1}{2}\mathbf{1})$

$$\Sigma_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}};\boldsymbol{\xi})} \leftarrow \left\{2(C^{\mathrm{G}})^{\mathsf{T}}\text{diag}\{\lambda(\boldsymbol{\xi})\}C^{\mathrm{G}} + \begin{bmatrix} \sigma_\beta^{-2} I_P & 0 \\ 0 & \underset{1\le\ell\le L}{\text{blockdiag}}(\mu_{q(1/\sigma^2_{u\ell})} I_{q^{\mathrm{G}}_\ell}) \end{bmatrix} - S\right\}^{-1}$$

$$\boldsymbol{\mu}_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}};\boldsymbol{\xi})} \leftarrow \Sigma_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}};\boldsymbol{\xi})} \left\{(C^{\mathrm{G}})^{\mathsf{T}}(\boldsymbol{y} - \tfrac{1}{2}\mathbf{1}) - \boldsymbol{s}\right\}$$

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

$$\Sigma_{\underline{q}(\boldsymbol{u}^{\mathrm{R}}_i;\boldsymbol{\xi})} \leftarrow H_i + H_i\, G_i^{\mathsf{T}}\, \Sigma_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}};\boldsymbol{\xi})}\, G_i\, H_i$$

$$\boldsymbol{\mu}_{\underline{q}(\boldsymbol{u}^{\mathrm{R}}_i;\boldsymbol{\xi})} \leftarrow H_i \left\{(X^{\mathrm{R}}_i)^{\mathsf{T}}(\boldsymbol{y}_i - \tfrac{1}{2}\mathbf{1}) - G_i^{\mathsf{T}} \boldsymbol{\mu}_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}};\boldsymbol{\xi})}\right\}$$

$$\boldsymbol{\xi}^2 \leftarrow \text{diagonal}\{C^{\mathrm{G}}(\Sigma_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}};\boldsymbol{\xi})} + \boldsymbol{\mu}_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}};\boldsymbol{\xi})}\boldsymbol{\mu}^{\mathsf{T}}_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}};\boldsymbol{\xi})})(C^{\mathrm{G}})^{\mathsf{T}}\}$$

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

$$\boldsymbol{\xi}^2_i \leftarrow \boldsymbol{\xi}^2_i + 2\,\text{diagonal}\{C^{\mathrm{G}}(-\Sigma_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}};\boldsymbol{\xi})} G_i H_i + \boldsymbol{\mu}_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}};\boldsymbol{\xi})}\boldsymbol{\mu}^{\mathsf{T}}_{\underline{q}(\boldsymbol{u}^{\mathrm{R}}_i;\boldsymbol{\xi})})(X^{\mathrm{R}}_i)^{\mathsf{T}}\}$$

$\boldsymbol{\xi}_i^2 \leftarrow \boldsymbol{\xi}_i^2 + \text{diagonal}\{(X_i^{\text{R}}(\Sigma_{\underline{q}(\boldsymbol{u}_i^{\text{R}};\boldsymbol{\xi})} + \boldsymbol{\mu}_{\underline{q}(\boldsymbol{u}_i^{\text{R}};\boldsymbol{\xi})}\boldsymbol{\mu}_{\underline{q}(\boldsymbol{u}_i^{\text{R}};\boldsymbol{\xi})}^{\mathsf{T}})(X_i^{\text{R}})^{\mathsf{T}}\}$

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

$B_{q(a_r^{\text{R}})} \leftarrow \nu(M_{q((\Sigma^R)^{-1})})_{rr} + A_{\text{R}r}^{-2}$; $\mu_{q(1/a_r^{\text{R}})} \leftarrow \frac{1}{2}(\nu + q^{\text{R}})/B_{q(a_r^{\text{R}})}$

$B_{q(\Sigma^{\text{R}})} \leftarrow \sum_{i=1}^{m}(\boldsymbol{\mu}_{\underline{q}(\boldsymbol{u}_i^{\text{R}};\boldsymbol{\xi})}\boldsymbol{\mu}_{\underline{q}(\boldsymbol{u}_i^{\text{R}};\boldsymbol{\xi})}^{\mathsf{T}} + \Sigma_{\underline{q}(\boldsymbol{u}_i^{\text{R}};\boldsymbol{\xi})}) + 2\,\nu\,\text{diag}\left(\mu_{q(1/a_1^{\text{R}})}, \ldots, \mu_{q(1/a_{q^{\text{R}}}^{\text{R}})}\right)$

$M_{q((\Sigma^{\text{R}})^{-1})} \leftarrow (\nu + m + q^{\text{R}} - 1)\, B_{q(\Sigma^{\text{R}})}^{-1}$

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

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

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

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

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

$\Lambda_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u}^{\text{G}},\boldsymbol{u}_i^{\text{R}};\boldsymbol{\xi})} \equiv E_{\underline{q}}\left[\left(\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}^{\text{G}}\end{bmatrix} - \boldsymbol{\mu}_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u}^{\text{G}};\boldsymbol{\xi})}\right)(\boldsymbol{u}_i^{\text{R}} - \boldsymbol{\mu}_{\underline{q}(\boldsymbol{u}_i^{\text{R}};\boldsymbol{\xi})})^{\mathsf{T}}\right] \leftarrow -\Sigma_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u}^{\text{G}};\boldsymbol{\xi})} G_i\, H_i$

6 Numerical evaluation

We conducted a comprehensive simulation study to assess the performance of Algorithm 3 in terms of Bayesian inferential accuracy and computational speed. We generated 100 datasets according to the following simulation setting, which corresponds to the special case of a Bayesian semiparametric random intercept and slope model.

$$y_{ij} \,|\, \beta_0, \beta_x, u_{0i}^{\text{R}}, u_{1i}^{\text{R}}, \sigma_\varepsilon^2 \overset{\text{ind.}}{\sim} N\left(\beta_0 + u_{0i}^{\text{R}} + (\beta_x + u_{1i}^{\text{R}})\, x_{ij} + f(s_{ij}),\ \sigma_\varepsilon^2\right), \tag{34}$$

where

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

and $s_{ij} \overset{\text{ind.}}{\sim} \text{Uniform}\,(0,\ 1)$, $x_{ij} \overset{\text{ind.}}{\sim} \text{Uniform}\,(0,\ 1)$ and $[u_{0i}^{\text{R}}\ u_{1i}^{\text{R}}]^{\mathsf{T}}\big|\Sigma^{\text{R}} \overset{\text{ind.}}{\sim} N\,(\mathbf{0},\, \Sigma^{\text{R}})$ for $1 \le i \le m$, $1 \le j \le n_i$. The true parameter values are:

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

The number of groups is varied, $m \in \{100,\ 500,\ 2500,\ 12,\ 500\}$, with the within group sample sizes n_i ranged between 10 and 20. Source code to reproduce the results is available as Supporting Information on the journal's web page (http://onlinelibrary.wiley.com/doi/xxx/suppinfo).

6.1 Assessment of accuracy

We fitted each replicated dataset using both the MFVB and MCMC. The MFVB fits were obtained using Algorithm 3 with the iterations terminated when the relative increase in $\log \underline{p}(\boldsymbol{y};\, q)$ fell below 10^{-7}. The MCMC samples were obtained using Stan (Stan Development Team, 2014) with R (R Development Core Team, 2014) interfacing via the `RStan` package (Stan Development Team, 2014). Stan is a probabilistic programming language, written in C++, for implementing full Bayesian statistical

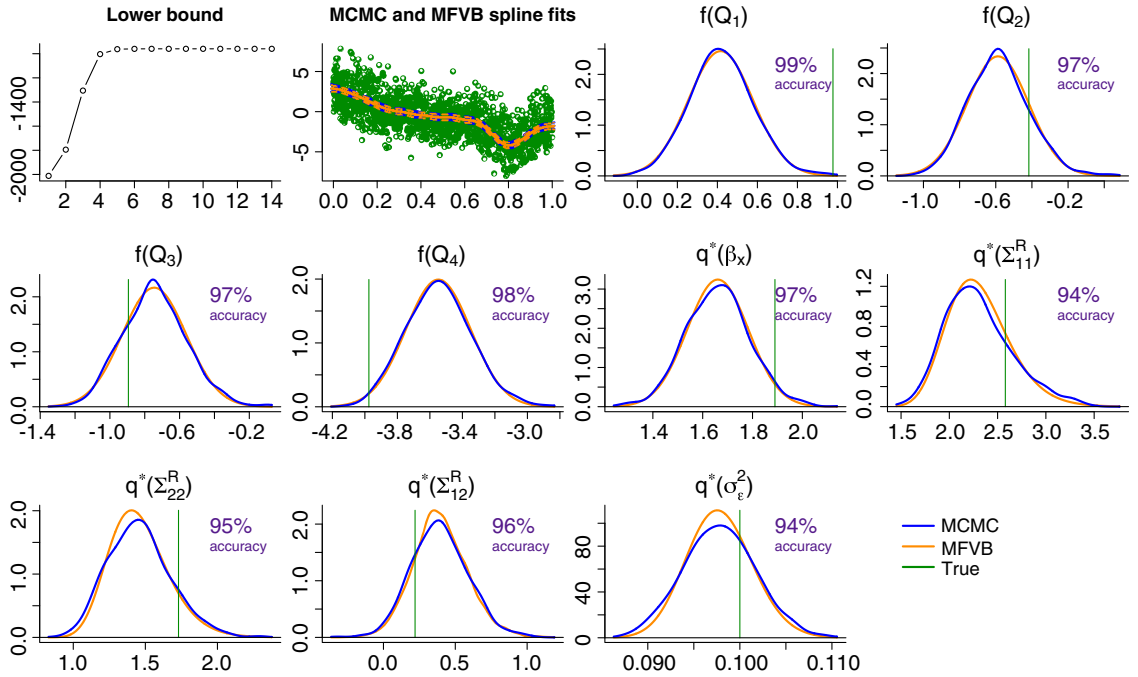

Figure 3 Top left panel: successive value of log $\underline{p}(\boldsymbol{y}; q)$ to monitor the convergence of the mean field variational Bayes algorithm. Second panel: Fitted function estimates and pointwise 95% credible sets for both mean field variational Bayes and Markov chain Monte Carlo. Approximate posterior density functions obtained via mean field variational Bayes and Markov chain Monte Carlo for a single replication of the simulation study described in the text with $m = 100$. Each pair of density function corresponds to a model parameter in (34). The green lines represent the true parameter values. The accuracy scores on the top right of each plot show the accuracy of mean field variational Bayes approximation compared against a Markov chain Monte Carlo benchmark.

inference. In each case, MCMC samples of size 10,000 were generated. The first 5000 values of each sample were discarded as burn-in and the remaining 5000 values were thinned by a factor of 5.

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

$$\text{accuracy}(q^*(\theta)) \equiv 100\left(1 - \frac{1}{2}\int_{-\infty}^{\infty} |\, q^*(\theta) - p_{\,\text{MCMC}}(\theta|\boldsymbol{y})|d\theta\right)\%, \tag{35}$$

where $p_{\text{MCMC}}(\theta|\boldsymbol{y})$ is an accurate MCMC-based approximation to $p(\theta|\boldsymbol{y})$. The $p_{\text{MCMC}}(\theta|\boldsymbol{y})$ is based on the binned kernel density estimate with the direct plug-in bandwidth, obtained via the R package `KernSmooth` (Wand and Ripley, 2010).

Figures 3 and 4 display approximate posterior density functions and side-by-side boxplots of the accuracy scores for the model parameters β_0, β_x, Σ^{R}_{11}, Σ^{R}_{12}, Σ^{R}_{21} , and σ^2_ε and $f(Q_k)$, $1 \le k \le 4$, where Q_k are the k-th sample quintiles of the s_{ij}s, with $m = 100$. The boxplots show that the majority of the accuracy scores exceed 95% and they rarely drop below 90%. These very good accuracy results are consistent with the heuristic justification given in Section 3.1 of Menictas and Wand (2013). In addition, the first panel of Fig. 3 indicates that the MFVB algorithm converged quickly, after 14 iterations. The second panel indicates that the MFVB and MCMC penalized splines fits are virtually identical.

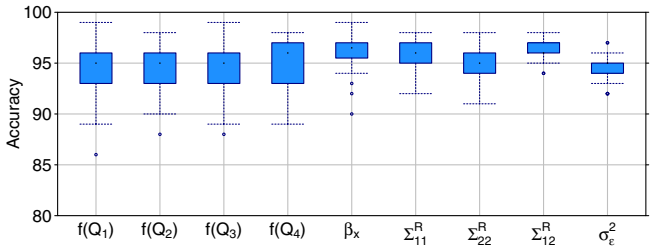

Figure 4 Side-by-side boxplots of accuracy scores for mean field variational Bayes approximations compared against Markov chain Monte Carlo over 100 runs with $m = 100$. Each boxplot corresponds to model parameters in (34).

6.2 Assessment of speed

We now turn our attention to quantification of the speed gains afforded by the streamlined MFVB algorithm. Both the naïve Algorithm 2 and streamlined Algorithm 3 were implemented in Fortran77, and the simulation described in Section 6.1 was rerun using both versions of MFVB. We also compared Algorithm 3 with contemporary software for fitting the frequentist version of (34). This was achieved using the function `gamm()` in the R package mgcv (Wood, 2015). Note that `gamm()` uses the function `lme()` in the R package nlme (Pinheiro et al., 2014) as a fitting engine.

All computations were performed on a Mac OS X laptop with a 2.6 GHz Intel Core i5 processor and 8 GBytes of random access memory. Table 1 summarizes the average (standard error) computing times over 100 runs and shows the practical benefits of the streamlined MFVB approach. As m increases the average computing time for the naïve approach increases rapidly from 0.2 seconds to almost 25 min, compared with about a quarter to half of a second for the streamlined approach. For $m = 12{,}500$, the naïve approach failed due to required storage of $\Sigma_{q(\boldsymbol{\beta},\boldsymbol{u})}$ exceeding memory restrictions for typical 2014 personal computing environments. However, the streamlined approach took just over 1.8 seconds on average to compute. The impressive speed gains in the streamlined approach are clearly reflected in the ratios of naïve over streamlined and the ratios of `gamm()` over streamlined, with respect to the average computing times. In situations where a dataset has a large number of groups, the streamlined approach is more than 3500 times faster than the naïve approach and 1700 times faster than the `gamm()` approach.

It is well-established that MCMC is computationally intensive in situations where complex models are applied to large datasets that lead to very slow run times. For $m = 12{,}500$, we expect the MCMC fitting takes hours to days, and therefore a similar timing comparison between MFVB and MCMC is not practical.

7 Applications

We now provide illustration of our streamlined MFVB methodology by reanalyzing two real data examples. Throughout this section, we standardize the response and continuous variables to have zero means and unit standard deviations. We use a normal prior for $\boldsymbol{\beta}$ and a Half–Cauchy prior for σ_{R}^2, σ_{ε}^2, σ_{u}^2, with the values of the hyperparameters all set to 100,000.

7.1 Application to smoking data

The first analysis is based on the perinatal health data from the United States National Center for Health Statistics, which are presented in Abrevaya (2006). The full data are not publicly available. However, a 10% random subsample is provided by Rabe-Hesketh and Skrondal (2008) and our analysis is of this subset. The data have a two-level structure with 8604 births as units at level 1 and

Table 2 Description of the United States National Center for Health Statistics perinatal health data presented in Abrevaya (2006).

Variable	Description
`momid`	Mother identifier
`birwt`	Birthweight in grams
`gestat`	Infant's gestational age in weeks
`mage`	Mother's age at the birth of the infant in years
`smoke`	Indicator for mother smoking during pregnancy
`male`	Indicator for infant being male
`married`	Indicator for mother being married
`hsgrad`	Indicator for mother having some college education, but not degree
`somecoll`	Indicator for mother having graduated from college
`black`	Indicator of mother being black
`kessner2`	Indicator for Kessner index equalling 2
`kessner3`	Indicator for Kessner index equalling 3
`novisit`	Indicator for no prenatal care visit
`pretri2`	Indicator for first prenatal care visit having occurred in second trimester
`pretri3`	Indicator for first prenatal care visit having occurred in third trimester

3978 mothers as groups at level 2. There are an average of 2.2 births per mother. The following variables are given in Table 2.

In this example, study variables can either vary at the birth level (therefore also at the mother level) or at the mother level only. For instance, `smoke` is a level-1 variable as maternal smoking status can change from one pregnancy to the next, whereas `black` is a level-2 variable as mother's ethnicity remains unchanged between pregnancies.

Motivated by a report from the United States Surgeon General: "Infants born to women who smoke during pregnancy have a lower average birthweight and are more likely to be small for gestational age than infants born to women who do not smoke ...." Abrevaya (2006) examined the effect of smoking on birthweight for women in the United States between 1990 and 1998 using a matched panel data approach. Here, we analyzed the data using a different approach via streamlined variational Bayes. The main study factor is maternal smoking status and the response of interest is infant's birthweight. The following Bayesian Gaussian semiparametric regression model with a random intercept for each mother was fitted via Algorithm 3:

$$\begin{aligned}
\texttt{birwt}_{ij} \mid \boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2 &\overset{\text{ind.}}{\sim} N\,(\,\beta_0 + u_i^{\mathrm{R}} + \beta_1\,\texttt{smoke}_{ij} + \beta_2\,\texttt{mage}_{ij} + \beta_3\,\texttt{male}_{ij} + \beta_4\,\texttt{married}_{ij} + \\
&\quad + \beta_5\,\texttt{hsgrad}_{ij} + \beta_6\,\texttt{somecoll}_{ij} + \beta_7\,\texttt{collgrad}_{ij} + \beta_8\,\texttt{black}_{ij} + \\
&\quad + \beta_9\,\texttt{kessner2}_{ij} + \beta_{10}\,\texttt{kessner3}_{ij} + \beta_{11}\,\texttt{novisit}_{ij} + \\
&\quad + \beta_{12}\,\texttt{pretri2}_{ij} + \beta_{13}\,\texttt{pretri3}_{ij} + f\,(\texttt{gestat}_{ij}),\ \sigma_\varepsilon^2\,)\,, \\
\boldsymbol{\beta} \sim N\,(\mathbf{0}, \sigma_\beta^2 I_P), \quad u_i^{\mathrm{R}} | \sigma_{\mathrm{R}}^2 &\overset{\text{ind.}}{\sim} N\,(0, \sigma_{\mathrm{R}}^2), \quad \sigma_{\mathrm{R}}^2 \sim \text{Half-Cauchy}\,(0,\ A_{\mathrm{R}}), \\
\sigma_\varepsilon^2 \sim \text{Half-Cauchy}\,(0,\ A_\varepsilon), \quad u_k^{\mathrm{G}} | \sigma_u^2 &\overset{\text{ind.}}{\sim} N\,(0, \sigma_u^2), \quad \sigma_u^2 \sim \text{Half-Cauchy}\,(0, A_u)
\end{aligned} \tag{36}$$

where

$$f(\texttt{gestat}) = \beta_{14}\,\texttt{gestat} + \sum_{k=1}^{K} u_k^{\mathrm{G}}\, z_k(\texttt{gestat}) \tag{37}$$

Table 3 Description of the 2000 Program for International Student Assessment conducted by the Organisation for Economic Cooperation and Development.

Variable	Description
`idschool`	School identifier
`passread`	Indicator for being proficient in reading
`isei`	International Socioeconomic Index
`college`	Indicator for highest education level by either parent being college
`oneforeign`	Indicator for one parent being foreign born
`twoforeign`	Indicator for both parents being foreign born
`language`	Indicator for test language (English) being spoken at home

is a penalized spline function for gestational age. Here $z_1(\cdot), \ldots, z_K(\cdot)$ is a set of O'Sullivan spline basis functions and σ_u^2 represents the amount of penalization of the spline coefficients $u_1^{\mathrm{G}}, \ldots u_{q^{\mathrm{G}}}^{\mathrm{G}}$ as described in (9).

Figures 5 and 7 allow a visual assessment of the MFVB-based approximate posterior density functions against a MCMC benchmark for (36). MCMC was fitted using Stan with a burn-in of size 5000, a post burn-in of size 5000 with a thinning factor of 5. The accuracy of approximate posterior density functions of $\boldsymbol{\beta}$ is excellent, ranging from 95% to 98%, while it is about 75% to 82% for the variance parameters σ_u^2 and σ_ε^2 , respectively. Inspection of Fig. 7 shows that the MFVB fits and pointwise 95% credible sets are very close to those obtained using MCMC. The MCMC fits took 4.1 days, while the MFVB fits took 1.4 seconds. This represents more than a two hundred thousand-fold improvement in computing time.

7.2 Application to student assessment data

The second example is based on a dataset from the 2000 Program for International Student Assessment conducted by the Organisation for Economic Cooperation and Development. The survey assessed education attainment of 15 years old students in 43 counties, with an emphasis on reading proficiency. Our analysis is of the United States sample of the full data, provided by Rabe-Hesketh and Skrondal (2008). The data have a two-level structure with 2069 observations as units at level 1 and 148 schools as groups at level 2. The following variables are given in Table 3.

We treat reading proficiency as the response variable, an indicator taking the value 1 to indicate proficient and 0 otherwise. The effect of socio-economic status has considerable interest in education, we therefore model its effect flexibly via a penalized regression spline. The following Bayesian semiparametric random intercept logistic regression model for student i in school j was fitted via Algorithm 4:

$$\begin{aligned}
\texttt{passread}_{ij}|\boldsymbol{\beta}, \boldsymbol{u} &\overset{\text{ind.}}{\sim} \text{Bernoulli}\left(\text{logit}^{-1}\left\{\beta_0 + u_i^{\mathrm{R}} + \beta_1\,\texttt{eduction}_{ij} + \beta_2\,\texttt{oneforeign}_{ij} + \right.\right.\\
&\left.\left. + \beta_3\,\texttt{bothforeign}_{ij} + \beta_4\,\texttt{language}_{ij} + f(\texttt{isei}_{ij})\right\}\right),\\
\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma_\beta^2 I_P), \quad u_i^{\mathrm{R}}|\sigma_{\mathrm{R}}^2 &\overset{\text{ind.}}{\sim} N(0, \sigma_{\mathrm{R}}^2), \quad \sigma_{\mathrm{R}}^2 \sim \text{Half-Cauchy}(0,\ A_{\mathrm{R}}),\\
u_k^{\mathrm{G}}|\sigma_u^2 &\overset{\text{ind.}}{\sim} N(0, \sigma_u^2), \quad \text{and} \quad \sigma_u^2 \sim \text{Half-Cauchy}(0, A_u).
\end{aligned} \tag{38}$$

The function f is a penalized spline, analogous to (37). Figures 6 and 7 allow a visual assessment of the MFVB-based approximate posterior density functions against a MCMC benchmark for (38). MCMC was fitted using Stan with a burn-in of size 5000, a post burn-in of size 5000 with a thinning factor of 5. The accuracy of the approximate posterior density functions of $\boldsymbol{\beta}$ and $f(Q_k)$ is good, ranging from

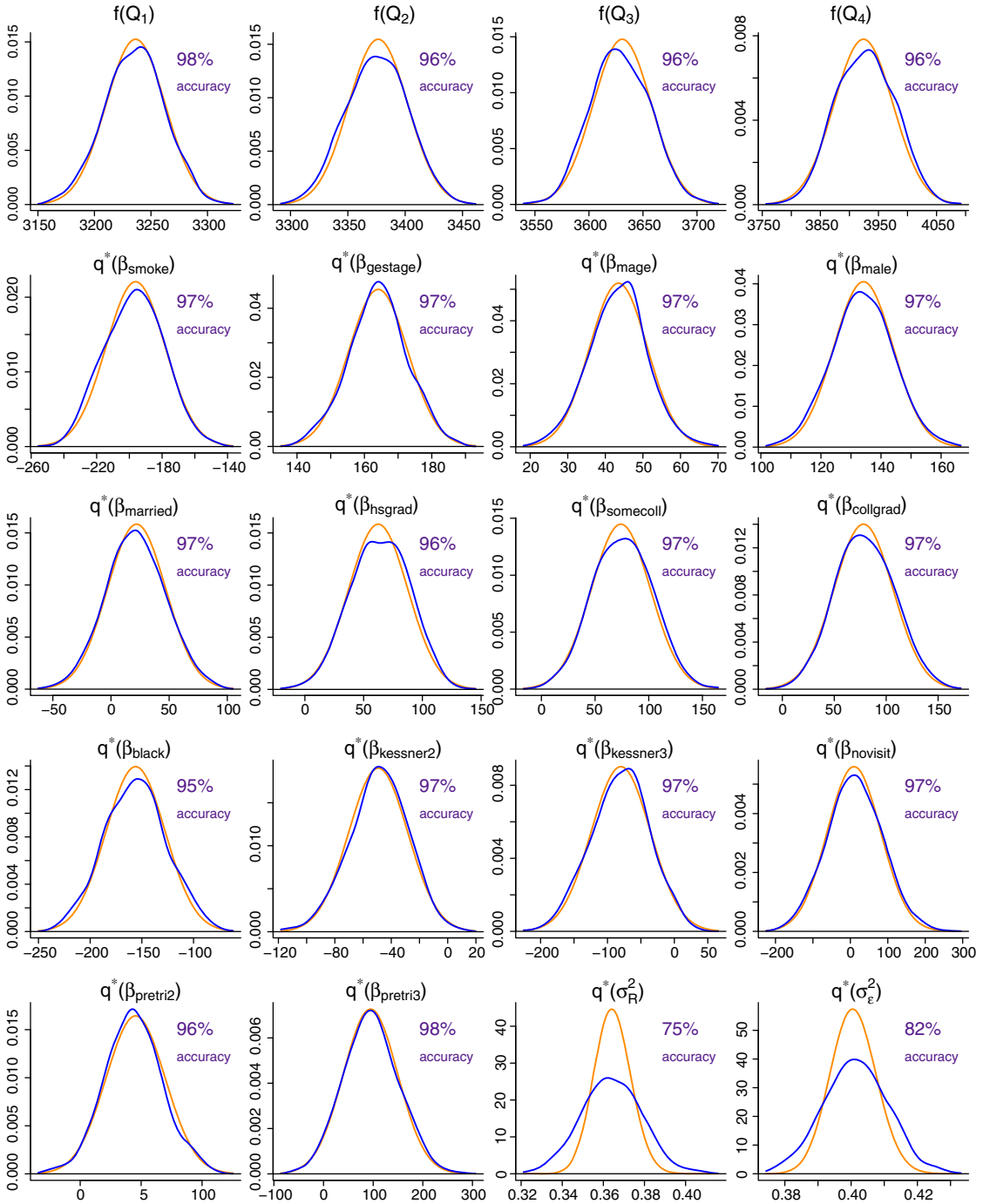

Figure 5 Approximate posterior density functions obtained via mean field variational Bayes (orange curves) and Markov chain Monte Carlo (blue curves) for fitting the Gaussian random intercept model corresponding to (36). Each pair of density function corresponds to a model parameter in (36). The accuracy scores on the top right of on each plot show the accuracy of mean field variational Bayes approximation compared against a Markov chain Monte Carlo benchmark.

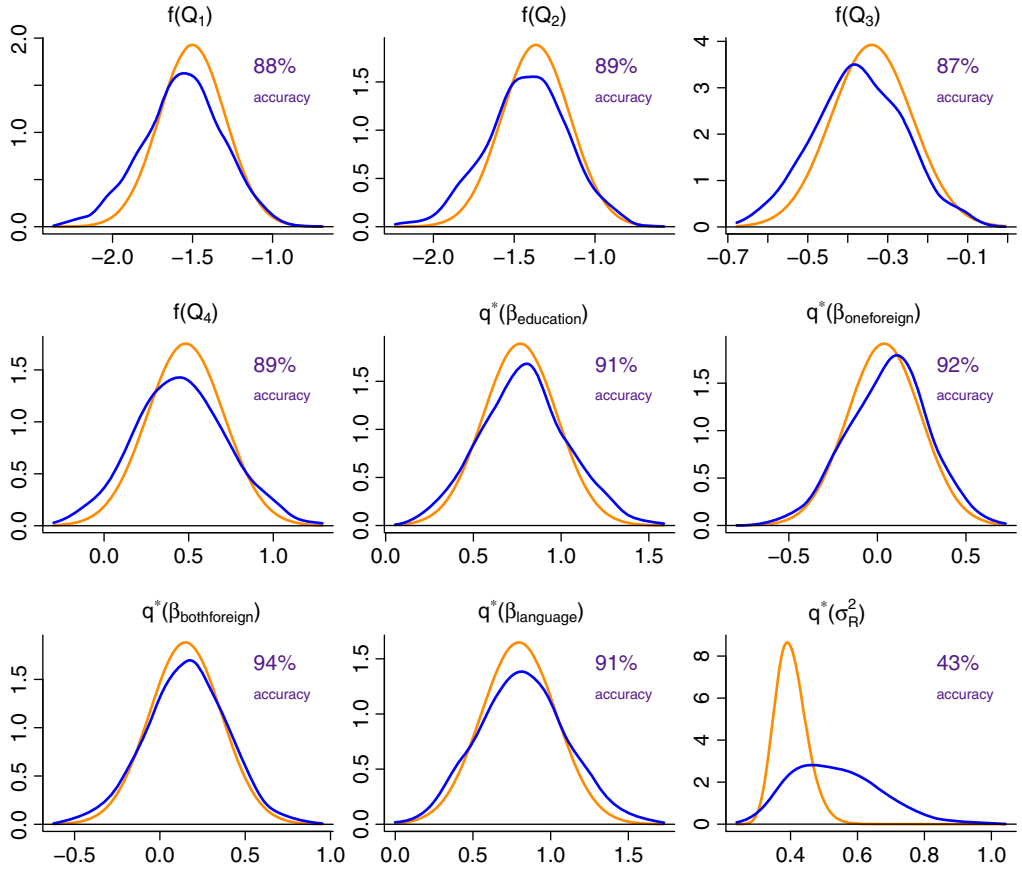

Figure 6 Approximate posterior density functions obtained via mean field variational Bayes (orange curves) and Markov chain Monte Carlo (blue curves) for fitting the Bernoulli random intercept model corresponding to (38). Each pair of density function corresponds to a model parameter in (38). The accuracy scores on the top right of on each plot show the accuracy of mean field variational Bayes approximation compared against a Markov chain Monte Carlo benchmark.

87% to 94%. In contrast, the MFVB approximation for the variance parameter σ_R^2 is mediocre. This might be due to the limitations of the Jaakkola and Jordan's approximation described in Section 5. Nonetheless, inspection of Fig. 7 shows that the MFVB fits and pointwise 95% credible sets are close to those obtained using MCMC. The MCMC fits took 57 min, while the MFVB fits took 1.2 min.

8 Concluding remarks

Mean field variational Bayes with streamlining, as exemplified by Algorithms 3 and 4, is a useful addition to the longitudinal and multilevel data analysis arsenal. It allows rapid and accurate approximate Bayesian inference for very large data-sets with computational times that increase only linearly in the number of groups. Extension of streamlined MFVB to higher level longitudinal and multilevel models is also of interest, although this requires a significant amount of additional mathematical analysis. Another extension is real-time processing that is achieved in Luts et al. (2014) using online versions

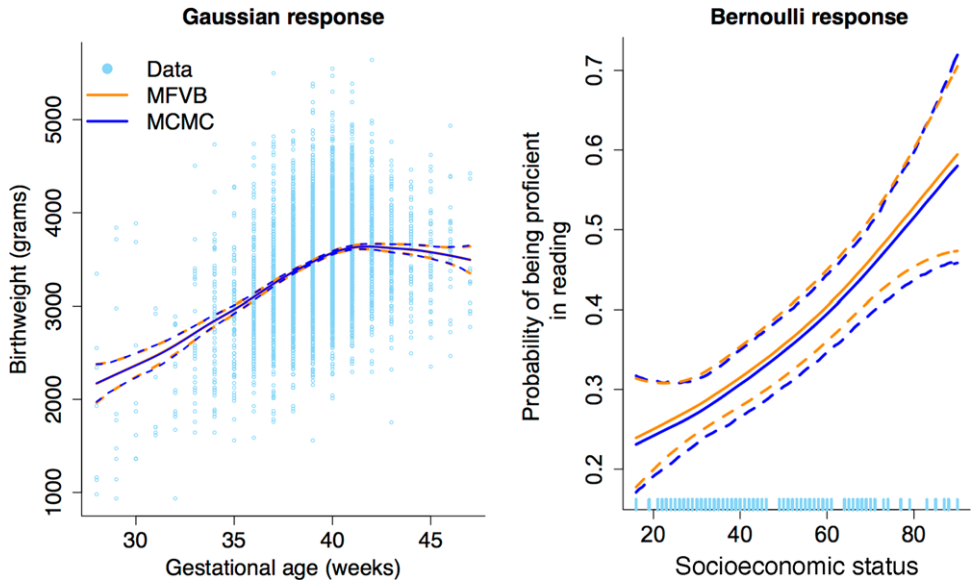

Figure 7 Comparison of mean field variational Bayes and Markov chain Monte Carlo penalized spline fits for the Gaussian and Bernoulli response models. The solid curves are approximate posterior means of the regression function and the dashed curves are pointwise 95% credible sets. The blue open circles are the raw data.

of MFVB. This article has helped open up a new branch of methodology for fitting and inference for grouped data in the high volume/velocity era.

Acknowledgment We are grateful to Chris Oates, Louise Ryan, and Brandon Stewart for their comments on this research. This research was partially supported by an Australian Postgraduate Award, a University of Technology Sydney Chancellor's Research Award, and Australian Research Council Discovery Project DP110100061.

Conflict of interest
The authors have declared no conflict of interest.

Appendix

A. The variational lower bound on the marginal log-likelihood for Algorithm 3

The marginal log-likelihood lower bound $\log \underline{p}(\boldsymbol{y}; q)$ can be obtained from (23) via straightforward, albeit long-winded, algebra. In addition, it uses a streamlined computation of $\log|\Sigma_{q(\boldsymbol{\beta},\boldsymbol{u})}|$, which we now justify. It depends on the following result concerning determinant of block-diagonal matrices:

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix} = |D|\,|A - BD^{-1}C|, \tag{1}$$

where D is a square matrix and is invertible (e.g. Harville, 2008, Theorem 13.3.8). From (1) we then get

$$\log|\Sigma_{q(\boldsymbol{\beta},\boldsymbol{u})}| = -\log\begin{vmatrix} F & G_1 & G_2 & \cdots & G_m \\ G_1^\mathsf{T} & H_1^{-1} & 0 & \cdots & 0 \\ G_2^\mathsf{T} & 0 & H_2^{-1} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ G_m^\mathsf{T} & 0 & 0 & \cdots & H_m^{-1} \end{vmatrix} =$$

$$= -\sum_{i=1}^{m}\log|H_i^{-1}| - \log|\Sigma^{-1}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}})}|.$$

Convergence in Algorithm 3 is assessed using the variational lower bound on the marginal log-likelihood and admits the following explicit expression:

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

where $\mathcal{C}_{q^{\mathrm{R}},\,\nu+q^{\mathrm{R}}-1}$ and $\mathcal{C}_{q^{\mathrm{R}},\,\nu+m+q^{\mathrm{R}}-1}$ are defined by (17).

B. The variational lower bound on the marginal log-likelihood for Algorithm 4

Convergence in Algorithm 4 is assessed using the variational lower bound on the marginal log-likelihood and admits the following explicit expression:

$$\begin{aligned}\log \underline{p}(\boldsymbol{y};\underline{q}) = {} & \tfrac{1}{2}\,q^{\mathrm{R}}(\nu+q^{\mathrm{R}}-1)\,\log(2\nu) - \left(\tfrac{1}{2}q^{\mathrm{R}}+L\right)\,\log(\pi) - \lambda(\boldsymbol{\xi})^\mathsf{T}(\boldsymbol{\xi}^2) + \mathbf{1}^\mathsf{T}\zeta(\boldsymbol{\xi})+\\ & +\left(\boldsymbol{y}-\tfrac{1}{2}\mathbf{1}\right)^\mathsf{T}\left\{C^{\mathrm{G}}\boldsymbol{\mu}_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}};\boldsymbol{\xi})} + \begin{bmatrix} X_1^{\mathrm{R}}\boldsymbol{\mu}_{\underline{q}(\boldsymbol{u}_1^{\mathrm{R}};\boldsymbol{\xi})} \\ \vdots \\ X_m^{\mathrm{R}}\boldsymbol{\mu}_{\underline{q}(\boldsymbol{u}_m^{\mathrm{R}};\boldsymbol{\xi})} \end{bmatrix}\right\}+\\ & -\tfrac{1}{2}P\,\log(\sigma^2_{\boldsymbol{\beta}}) - \tfrac{1}{2}\sigma_{\beta}^{-2}\left\{\|\boldsymbol{\mu}_{\underline{q}(\boldsymbol{\beta};\boldsymbol{\xi})}\|^2 + \mathrm{tr}(\Sigma_{\underline{q}(\boldsymbol{\beta};\boldsymbol{\xi})})\right\} + \tfrac{1}{2}\left(\textstyle\sum_{\ell=1}^{L} q_\ell^{\mathrm{G}} + P + m\right)+\\ & -\tfrac{1}{2}\textstyle\sum_{i=1}^{m}\log\left|(X_i^{\mathrm{R}})^\mathsf{T}X_i^{\mathrm{R}} + M_{q((\Sigma^R)^{-1})}\right| - \tfrac{1}{2}\log|\Sigma^{-1}_{\underline{q}(\boldsymbol{\beta},\boldsymbol{u}^{\mathrm{G}};\boldsymbol{\xi})}|+\end{aligned}$$

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

where $\zeta(x) \equiv x/2 - \log(1+e^x) + x\tanh(x/2)/4$.

References

Abrevaya, J. (2006). Estimating the effect of smoking on birth outcomes using a matched panel data approach. *Journal of Applied Statistics* **21**, 489–519.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York, NY.

Bishop, C. M. (2008). A new framework for machine learning. *World Congress on Computational Intelligence, 2008 Plenary/Invited Lectures, Lecture Notes in Computer Science 5050*. Springer-Verlag, Germany.

Browne, W. J., Goldstein, H. and Rasbash, J. (2001). Multiple membership multiple classification (MMMC) models. *Statistical Modelling* **1**, 103–124.

Diggle, P., Heagerty, P., Liang, K. L. and Zeger, S. (2002). *Analysis of Longitudinal Data* (2nd edn.). Oxford University Press, Oxford, UK.

Douglas, B., Maechler, M., Bolker, B. and Walker, S. (2014). `lme4` 1.1-5: linear mixed-effects models using Eigen and S4. R package version 3.1. http://CRAN.R-project.org/package=lme4.

Faes, C., Ormerod, J. and Wand, M. P. (2011). Variational Bayesian inference for parametric and non-parametric regression with missing predictor data. *Journal of the American Statistical Association* **106**, 495.

Fitzmaurice, G., Davidian, M., Verbeke, G. and Molenberghs, G. (2008). *Longitudinal Data Analysis: A Handbook of Modern Statistical Methods*. Chapman and Hall/CRC, FL.

Fitzmaurice, G. M., Laird, N. M. and Ware, J. H. (2011). *Applied Longitudinal Analysis* (2nd edn.). John Wiley and Sons, New York, NY.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* **1**, 515–533.

Gelman, A. and Hill, J. (2007). *Data Analysis using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York, NY.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika* **73**, 43–56.

Goldstein, H. (2010). *Multilevel Statistical Models* (4th edn.). Wiley, Chichester, UK.

Harville, D. A. (2008). *Matrix Algebra from a Statistician's Perspective*. Springer, New York, NY.

Huang, A. and Wand, M. P. (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis* **2**, 439–452.

Jaakkola, T. S. and Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing* **10**, 25–37.

Knowles, D. A and Minka, T. P. (2011). Non-conjugate variational message passing for multinomial and binary regression. *Advances in Neural Information Processing Systems* **2** 4, 1701–1709.

Luenberger, D. G. and Ye, Y. (2008). *Linear and Nonlinear Programming* (3rd edn.). Springer, New York, NY.

Luts, J., Broderick, T. and Wand, M. P. (2014). Real-time semiparametric regression. *Journal of Computational and Graphical Statistics* **23**, 589–615.

Menictas, M. and Wand, M. P. (2013). Variational inference for marginal longitudinal semiparametric regression. *Statistical* **2**, 61–71.

Minka, T., Winn, J., Guiver, J. and Knowles, D. (2013). `Infer.NET 2.5` Microsoft Research Cambridge, UK.

Ormerod, J. T. and Wand, M. P. (2010). Explaining variational approximations. *American Statistical Association* **64**, 140–153.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., EISPACK authors and R Core Team. (2014). `nlme`: linear and nonlinear mixed effects models. R package version 3.1. http://www.r-project.org.

R Development Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.

Rao, C. R. (1972). Estimation of variance and covariance components in linear models. *Journal of the American Statistical Association* **67**, 112–115.

Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science* **6**, 15–51.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press, New York, NY.

SAS Institute Inc. (2013). SAS/STAT *13.1 User's Guide*. SAS Institute Inc., North Carolina.

Stan Development Team. (2014). Stan: *A C++ Library for Probability and Sampling Version 2.2*. http://mc-stan.org.

Smith, A. D. A. C. and Wand, M. P. (2008). Streamlined variance calculations for semiparametric mixed models. *Statistics in Medicine* **29**, 435–448.

Tan, S. L. and Nott, D. J. (2013). Variational inference for generalized linear mixed models using partially noncentered parametrizations. *Statistical Science* **28**, 167–188.

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.

Tipping, M. E. (1999b). Probabilistic visualisation of high-dimensional binary data. In: M. S. Kearns, S. A. Solla, and D. A. Cohn (Eds.), *Advances in Neural Information Processing Systems 11*. MIT Press, Cambridge, MA, pp. 592–598.

Titterington, D. M. (2004). Bayesian methods for neural networks and related models. *Statistical Science* **19**, 128–139.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundation and Trends in Machine Learning* **1**, 1–305.

Wand, M. P. and Ormerod, J. T. (2008). On semiparametric regression with O'Sullivan penalized splines. *Australian and New Zealand Journal of Statistics* **50**, 179–198.

Wand, M. P. and Ormerod, J. T. (2011). Penalized wavelets: embedding wavelets into semiparametric regression. *Electronic Journal of Statistics* **5**, 1654–1717.

Wand, M. P., Ormerod, J. T., Padoan, S. A. and Frühwirth, R. (2011). Mean field variational Bayes for elaborate distributions. *Bayesian Analysis* **6**, 847–900.

Wand, M. P. and Ripley, B. D. (2009). KernSmooth 2.23: Functions for Kernel Smoothing Corresponding to the Book: Wand, M. P. and Jones, M. C. (1995), *Kernel Smoothing*. R package http://cran.r-project.org.

Wolfinger, R., Tobias, R. and Sall, J. (1994). Computing Gaussian likelihoods and their derivatives for general linear mixed models. *SIAM Journal on Scientific Computing* **15**, 1294–1310.

Wood, S. N. (2014). `mgcv`: mixed GAM computation vehicle with GCV/AIC/REML smoothness estimation. R package version 3.1. http://cran.r-project.org.

Zhao, Y., Staudenmayer, J., Coull, B. A. and Wand, M. P. (2006). General design Bayesian generalized linear mixed models. *Statistical Science* **21**, 35–51.

