Featured Article

Statistics in Medicine

Received 7 May 2015, Accepted 27 August 2015 Published online 28 September 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6737

Variational methods for fitting complex Bayesian mixed effects models to health data

Cathy Yuen Yi Lee*† and Matt P. Wand

We consider approximate inference methods for Bayesian inference to longitudinal and multilevel data within the context of health science studies. The complexity of these grouped data often necessitates the use of sophisticated statistical models. However, the large size of these data can pose significant challenges for model fitting in terms of computational speed and memory storage. Our methodology is motivated by a study that examines trends in cesarean section rates in the largest state of Australia, New South Wales, between 1994 and 2010. We propose a group-specific curve model that encapsulates the complex nonlinear features of the overall and hospital-specific trends in cesarean section rates while taking into account hospital variability over time. We use penalized spline-based smooth functions that represent trends and implement a fully mean field variational Bayes approach to model fitting. Our mean field variational Bayes algorithms allow a fast (up to the order of thousands) and streamlined analytical approximate inference for complex mixed effects models, with minor degradation in accuracy compared with the standard Markov chain Monte Carlo methods. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords: Bayesian inference; group-specific curves; longitudinal and multilevel data; Markov chain Monte Carlo; mean field variational Bayes approximation; semiparametric regression

1. Introduction

Longitudinal and multilevel data are common in many applied areas such as medicine, epidemiology, and social science. Such data often have a grouped or hierarchical structure. Examples include human growth studies in which individuals' body measurements are collected at multiple follow-up times, medical studies in which patients are grouped within hospitals and hospital-specific disease rates are examined, and sample surveys in which respondents to questionnaires are grouped within households and geographical districts. Parametric regression methods for these types of data have been well developed in the last two decades. Such methods can be broadly classified into estimating equations-based methods (e.g., [1] and [2]) and mixed effects models (e.g., [3–6]).

A parametric regression model assumes the relationship between the mean of a response outcome and predictor variables to be of a known functional form. Although such parametric assumption offers simplicity, it is inappropriate for situations when the relationship between the mean response and predictors is unknown. As an illustration, Figure 1 shows the overall and hospital-specific trends in cesarean section rates for low-risk nulliparous women in the largest state of Australia, New South Wales (NSW), between 1994 and 2010. The trends are clearly nonlinear and show substantial within-hospital and between-hospital variability over time. In order to capture these features simultaneously, a more complicated and flexible model is required. For example, a more flexible model is a generalization to additive models [7], where the overall mean and deviation of each group (in this case hospital) from that overall mean are modeled nonparametrically with arbitrary smooth functions of time. We refer to such model as a group-specific curve model.

Early work on group-specific curve models used different smoothing techniques (e.g., kernels and smoothing splines) to estimate both overall mean curve and group-specific curves, with random effects

School of Mathematical and Physical Sciences, University of Technology Sydney, Ultimo, New South Wales 2007, Australia
*Correspondence to: Cathy Yuen Yi Lee, School of Mathematical and Physical Sciences, University of Technology Sydney, 638 Jones Street Broadway, Ultimo, New South Wales 2007, Australia.
†E-mail: Yuen.Y.Lee@student.uts.edu.au

Copyright © 2015 John Wiley & Sons, Ltd.

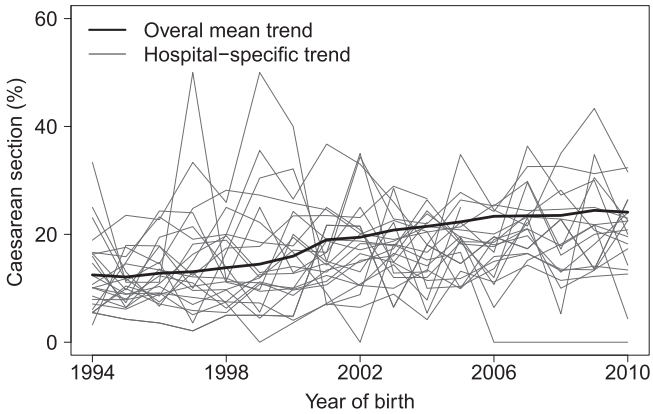

Figure 1. Trends in cesarean section rates for low-risk nulliparous women in the largest state of Australia, New South Wales, between 1994 and 2010. The darker line is the overall mean trend, and the lighter lines are the selected hospital-specific trends.

modeled either by a parametric function or by a Gaussian process [8–10]. Brumback and Rice [11] embedded smoothing splines within the mixed model framework. However, they ran into computational problems in the estimation because they assumed fixed intercepts and slopes for the group-specific curves, thus leading to the number of fixed effects being at least twice as large as the number of curves. Rice and Wu [12] alleviated computational problems by modeling group-specific curves as spline basis functions with random coefficients. However, their approach requires a careful selection of the number and location of knots. Durban *et al*. [13] relaxed the importance of this selection by using low-rank smoothing splines with a penalty approach, which expresses the group-specific curves as a linear combination of truncated polynomial spline bases with random coefficients. Their proposed covariance matrix for the random basis coefficients was modeled parametrically with independence assumptions. Chen and Wang [14] and Ryu, Li, and Mallick [15] extended this work by allowing a more general covariance matrix structure for the random basis coefficients, although the former was confined to functional data analysis.

When the dimension of the spline basis functions and the number of groups become large, many of the aforementioned approaches become too slow, or even computationally infeasible. These problems motivate the development of fast and scalable methods for fitting group-specific curve models to large longitudinal and multilevel datasets. In Bayesian statistics, exact inference for nonparametric or semiparametric regression models that use penalized spline basis functions is typically intractable, requiring approximate inference methods for use in practice. Markov chain Monte Carlo (MCMC) is the most commonly used approximate inference method in this setting but can be computationally intensive and often suffers from poor convergence in large and complex models.

A faster, deterministic alternative to MCMC is variational approximations (e.g., [16] and [17]). The basic idea behind variational approximations is to recast the problem of computing posterior probabilities as an optimization problem by introducing a class of more manageable approximating distributions, and then optimizing some criterion to find the distribution within that class that best matches the posterior. Recently, there has been a growing interest in variational methods for longitudinal and multilevel models. Ormerod and Wand [17] and Luts, Broderick, and Wand [18] developed variational algorithms for fitting and inference for grouped data. However, their algorithms are infeasible for datasets with a large number of groups because of naïve inversion of the random effects covariance matrix. Armagan and Dunson [19] developed a fast remedy for sparse covariance estimation relying on a decomposition but is restricted to linear response. Tan and Nott [20] extended their approach and introduced a partially non-centered nonparametrization strategy for generalized linear mixed models, allowing the random effects for each group to be independent. Such restriction was shown to improve efficiency of the variational algorithms. Lee and Wand [21] did not impose such restriction but rather took advantage of the inherent blocked structure and sparseness of the random effects covariance matrix and developed algorithms that streamline its inversion and estimation. Their streamlined algorithms for the longitudinal and multilevel models result in impressive speed improvements (up to the order of thousands) and currently represent

Copyright © 2015 John Wiley & Sons, Ltd.

the state-of-the-art in this area. Stewart [22] applied their streamlined algorithms to research studies in social sciences.

In this article, we present a fully variational approach to fitting a series of Bayesian logistic mixed effects models in order to characterize trends in cesarean section rates in NSW, Australia. We begin with the standard random intercept and slope model and then generalize the model by replacing linear models for the overall mean and hospital-specific trends with arbitrary smooth functions that are nonparametrically estimated from the data. The next section presents a brief overview of variational methods. Section 3 reviews the random intercept and slope model and presents various extensions of the standard model including nonparametric functions and factor-by-curve interactions. A basic framework of approximate inference for a simple Bayesian logistic mixed effects model is outlined, serving as a building block for the more complicated models. Details on variational fitting and inference for models are also given. In Section 4, we present numerical evidence of the efficacy of our developed variational algorithms in terms of inferential accuracy and computational speed. The final proposed model is illustrated in Section 5 with the analysis of cesarean section data. The paper concludes with a brief discussion in Section 6.

2. A brief introduction to variational approximations

Bayesian inference involves finding the joint posterior distribution of parameters of interest given the observed data. In many instances, exact inference is infeasible because of the posterior distributions being intractable, requiring approximate inference methods for use in practice. In recent years, Bayesian inference engines have emerged for approximate inference for general classes of mixed effects models. Examples include BUGS [23] and Stan [24], based on MCMC methods. MCMC is the current standard method for estimating Bayesian statistical models and has proven useful in a wide range of problems. However, for large models with complex posteriors, MCMC can be computationally intensive and suffers from poor mixing that leads to slow convergence. We therefore opt for an alternative approach, namely variational approximations.

Variational approximations are a fast, versatile, and deterministic class of approximations with origins in statistical physics and computer science literature (e.g., [25–28]). Since about 2005, variational methods have been increasingly explored in the statistical literature. For example, McGory and Titterington [29], Wand *et al*. [30], and Wand and Ormerod [31] present variational methods for a diversified range of applications, from finite mixture models, complex models with elaborate distributions (such as asymmetric Laplace and skew normal) to spline and wavelet regression models. In addition, Wang and Titterington [32] proved convergence of variational algorithms for normal mixture models, and You *et al*. [33] propose several information criteria that are useful for model selection. We provide here a brief overview of variational approximations and highlight the key concepts. Comprehensive summaries can be found in the work by Bishop [28] and by Ormerod and Wand [17].

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

$$q^*(\boldsymbol{\theta}) = \underset{q\in\mathcal{Q}}{\text{argmin}}\, \text{KL}\left\{q(\boldsymbol{\theta}) \,\|\, p(\boldsymbol{\theta}|\boldsymbol{y})\right\} \quad \text{where} \quad \text{KL}\left\{q(\boldsymbol{\theta}) \,\|\, p(\boldsymbol{\theta}|\boldsymbol{y})\right\} = \int_{\boldsymbol{\Theta}} q(\boldsymbol{\theta}) \log\left\{\frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{y})}\right\} d\boldsymbol{\theta}. \quad (1)$$

In order to enhance tractability, $q(\boldsymbol{\theta})$ is subject to some form of factorization or product density restriction [28]:

$$\mathcal{Q} = \left\{ q(\boldsymbol{\theta}) \;:\; q(\boldsymbol{\theta}) = \prod_{i=1}^{M} q_i(\boldsymbol{\theta}_i) \;\text{ for some partition }\; \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\} \text{ of } \boldsymbol{\theta} \right\}, \quad (2)$$

known as mean field restriction. The resultant $q^*(\boldsymbol{\theta})$ is known as mean field variational Bayes (MFVB) approximation to the actual posterior, and from now on, we refer to it as optimal q-density. Essentially, it assumes posterior independence among parameters that may not be present in the actual posterior.

Copyright © 2015 John Wiley & Sons, Ltd.

Depending on the chosen partition of $\boldsymbol{\theta}$, this independence assumption may be rather unrealistic in settings where there exist high posterior correlations among parameters, thus leading to poor approximations. The particular parametric families that constitute each of the approximating q-density factors $q_i(\boldsymbol{\theta}_i)$ will be derived through the variational methods. Indeed, minimizing the Kullback–Liebler divergence between $p(\boldsymbol{\theta}|\boldsymbol{y})$ and $q(\boldsymbol{\theta})$ in (1) is equivalent to maximizing the lower bound $\underline{p}(\boldsymbol{y}; q)$ because the marginal log-likelihood can be expressed as [17]

$$\log p(\boldsymbol{y}) = \log \underline{p}(\boldsymbol{y}; q) + \mathrm{KL}\left\{q(\boldsymbol{\theta}) \,\|\, p(\boldsymbol{\theta}|\boldsymbol{y})\right\},$$

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

The optimal q-density functions under product restriction (2) can be obtained via an iterative algorithm that is analogous to the expectation–maximization algorithm (e.g., [17] and [28]). Define

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

to be the log posterior averaged over the current estimates of the approximating density functions for all but the i-th parameter vector. The term $E_{q(-\boldsymbol{\theta}_i)}$ denotes expectation with respect to the q-densities of all parameters except $\boldsymbol{\theta}_i$. Variational approximations is now reduced to solving an optimization problem in the form of (1). We proceed by initializing each of the optimal q-density factors $q_i^*(\boldsymbol{\theta}_i)$ and updating each factor successively using the current estimates of the other factors. At the end of each iteration, an updated value of the lower bound is computed,

$$\log \underline{p}(\boldsymbol{y}; q^*) = E_q\{\log p(\boldsymbol{y}, \boldsymbol{\theta}) - \log q^*(\boldsymbol{\theta})\},$$

and the algorithm is iterated until convergence of the lower bound to its maximum (Algorithm 1). While the distributional forms of the approximating density functions $q_i(\boldsymbol{\theta}_i)$ are unspecified, the structure of the statistical model itself lends to a solution that lies in a particular parametric family for each $q_i(\boldsymbol{\theta}_i)$. For example, when all the parameters in a model are conditionally conjugate, the optimal q-density functions in Algorithm 1 are available in closed form. If $q_i(\boldsymbol{\theta}_i)$ cannot be recognized as a standard distribution, then numerical integration methods are required to estimate the marginal likelihood, which is computationally more demanding.

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.

Algorithm 1: Iterative scheme for obtaining the optimal q-density functions under product restriction (2).

3. Mean field variational Bayes inference of cesarean section data

In this section, we present an epidemiology study that motivates our methodological development. Cesarean section rates are increasing worldwide, and among countries of Organisation for Economic Co-operation and Development, the average rate increased from 20% in 2000 to 27% in 2011 [34]. It is recognized that cesarean section rates vary considerably across regions and hospitals in several countries. For example, in the USA, a fourfold variation was found between low-use and high-use areas [35]. In the UK, rates of emergency cesarean section among National Health Service trusts ranged from 15% to

Copyright © 2015 John Wiley & Sons, Ltd.

32% [36]. In the largest state of Australia, NSW, there were 1,500,964 deliveries from 1994 to 2010, with cesarean section rate increased from 17.4% to 30.6% over this 17-year period. While the statewide trend in cesarean section rate is well reported, little is known about the trend in cesarean section rate for each hospital, especially for small hospitals. Previous work by the author addressed this gap by examining hospital-specific trends in cesarean section rates for low-risk nulliparous women in NSW between 1994 and 2010 [37]. Here, we extend the model applied previously, using the MFVB approximation, to compare hospital-specific trends between subpopulations of women.

Data were obtained from the NSW Perinatal Data Collection. The Perinatal Data Collection is a legislated population-based surveillance system covering all live births, and stillbirths of at least 20 weeks gestation or at least 400 g birth weight. Information is recorded by the attending midwife or medical practitioner providing maternity care and includes maternal demographic, medical and obstetric information of the mother, and details of labor, birth, and condition of the infant. Data for this analysis were de-identified. Permission for use of data was approved by the NSW Ministry of Health. The study population consists of low-risk women giving birth for the first time in a NSW public or private hospital between 1994 and 2010. Low-risk women were defined as those women giving birth for the first time (nulliparous women), aged 20–34 years who did not smoke in pregnancy and did not have any preexisting or gestational medical conditions, such as diabetes and hypertension, and gave birth to singleton cephalic (head-down position) live infants at term (37 weeks gestation).

The primary outcome is the annual rate of cesarean section for each hospital in NSW over a 17-year period. Public and private hospitals with $\geqslant 50$ births per annum for more than half of the 17-year study period were included. The data have a two-level hierarchical structure with women nested within hospitals. To define notation, we use a double subscript convention with i denoting hospital ($i = 1, \ldots, m$) and j denoting woman ($j = 1, \ldots n_i$). For example, y_{ij} denote the binary indicator of cesarean section for woman j in hospital i and x_{ij} the corresponding year of birth. The total number of low-risk nulliparous women in the study population is $\sum_{i=1}^{m} n_i$. To examine the underlying trends in cesarean section rates overall and for each hospital, we consider a series of Bayesian logistic mixed effects models with increasing complexity. For each model, the year of birth is modeled by adding together two arbitrary unspecified functions as covariates, one representing the overall mean $f(x)$ and the other being hospital-specific departures from that overall mean $g_i(x)$. To keep notation simple, we consider only a single predictor in the model. Extension to models with more than one predictor is straightforward. More precisely, we assume

$$y_{ij} \overset{\text{ind.}}{\sim} \text{Bernoulli}\left(\text{logit}^{-1}\left\{f\left(x_{ij}\right) + g_i\left(x_{ij}\right)\right\}\right), \quad 1 \leqslant j \leqslant n_i,\ 1 \leqslant i \leqslant m, \tag{3}$$

where the functional forms of f and g_i will be reviewed in the later subsections. The function g_i controls for hospital-to-hospital variation, both in magnitude and across the 17-year trend. The sum of f and g_i provides the trend in cesarean section rate for each hospital, which can have a unique intercept, slope, and even shape depending upon whether they are of a parametric or nonparametric form. The term $f(x_{ij}) + g_i(x_{ij})$ represents the log odds of cesarean section for woman j in hospital i and will be the focus in the following subsections.

The Bayesian hierarchical model corresponding to (3) is

$$\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{u} \sim \text{Bernoulli}\left(\text{logit}^{-1}\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}\right)\right), \quad \boldsymbol{u} \mid \boldsymbol{G} \sim N(\boldsymbol{0}, \boldsymbol{G}). \tag{4}$$

The notation $\boldsymbol{v} \sim \text{Bernoulli}(\boldsymbol{p})$ used in (4) is shorthand for the entries of the random vector $\boldsymbol{v}$ having independent Bernoulli distributions with parameters corresponding to the entries of $\boldsymbol{p}$. Scalar functions applied to vectors are evaluated element-wise. For example,

$$\text{logit}^{-1}\left(\begin{bmatrix} v_1 \\ \vdots \\ v_d \end{bmatrix}\right) = \begin{bmatrix} \text{logit}^{-1}(v_1) \\ \vdots \\ \text{logit}^{-1}(v_d) \end{bmatrix}.$$

The matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the respective $\sum_{i=1}^{m} n_i \times P$ fixed effects design matrix and $\sum_{i=1}^{m} n_i \times d$ random effects design matrix, and $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are the so-called $P \times 1$ fixed effects vector and $d \times 1$ random effects vector. The random effects vectors are given a multivariate normal prior with zero mean and covariance matrix $\boldsymbol{G}$. Throughout, we take the prior distribution of the fixed effects vector $\boldsymbol{\beta}$ to be of the form $\boldsymbol{\beta} \sim N(0, \sigma_\beta^2 \boldsymbol{I})$, where σ_β^2 is to be chosen by the analyst. In logistic regression, situations where a shift

Copyright © 2015 John Wiley & Sons, Ltd.

in the predictor x corresponds to the probability of response y changing from 0.01 to 0.99 are rare. Hence, a prior distribution that assigns low probabilities to changes of 10 on the logistic scale would be appropriate [38]. Further, we use proper but 'diffuse' conditionally conjugate priors for the random effects; this corresponds to a half-Cauchy distribution for a single variance component (Result 5 of [30])

$$\sigma^2 \,|\, a \sim \text{inverse-gamma}\left(\tfrac{1}{2}, 1/a\right), \quad a \sim \text{inverse-gamma}\left(\tfrac{1}{2}, 1/A^2\right).$$

The multivariate extension of the half-Cauchy distribution is a scaled inverse-Wishart distribution for an unstructured $q^{\text{R}} \times q^{\text{R}}$ random effects covariance matrix [39]

$$\begin{aligned} \boldsymbol{\Sigma}_{\text{R}} \,|\, a_{\text{R},1}, \ldots, a_{\text{R},q^{\text{R}}} &\sim \text{inverse-Wishart}\left(\nu + q^{\text{R}} - 1,\ 2\,\nu\,\text{diag}\left(1/a_{\text{R},1} \ldots, 1/a_{\text{R},q^{\text{R}}}\right)\right), \\ a_{\text{R},r} &\overset{\text{ind.}}{\sim} \text{inverse-gamma}\left(\tfrac{1}{2}, 1/A_{\text{R}}^2\right), \quad 1 \leqslant r \leqslant q^{\text{R}}, \end{aligned} \tag{5}$$

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

It will soon become apparent that the model framework introduced in (4) encompasses a wide range of logistic mixed effects models with different types of group structures by simply changing $\boldsymbol{G}$.

3.1. Random intercept and slope model

A natural starting point for these cesarean section data is to assume that the overall mean and deviation of the i-th hospital from that overall mean are simply straight lines. This leads to the standard random intercept and slope model [40], with $f(x)$ and $g_i(x)$ modeling through linear functions:

$$\begin{aligned} f(x) = \beta_0 + \beta_1 x_{ij} \ ;\ g_i(x) = U_{0i} + U_{1i} x_{ij} \\ \text{and} \quad \boldsymbol{G} = \boldsymbol{I}_m \otimes \boldsymbol{\Sigma}_{\text{R}} = \underset{1 \leqslant i \leqslant m}{\text{blockdiag}} \begin{bmatrix} \sigma^2_{u_0} & \rho_{u_0,u_1} \\ \rho_{u_0,u_1} & \sigma^2_{u_1} \end{bmatrix}, \end{aligned} \tag{6}$$

where β_0 and β_1 are the respective overall intercept and slope, and U_{0i} and U_{1i} are the hospital-specific deviations from that overall intercept and slope, being treated as a random sample from a bivariate normal distribution with an unstructured 2×2 covariance matrix $\boldsymbol{\Sigma}_{\text{R}}$. The model accounts for possible variability in the intercepts $\sigma^2_{u_0}$ and slopes $\sigma^2_{u_1}$ of hospitals and allows for an intercept–slope correlation ρ_{u_0,u_1}. In common practice, we generalize the fixed and random components of the mixed effects models to arbitrary general design matrices (Section 2 of [41]). This allows one to take advantage of the ever-expanding methods and software for inference in these models. Henceforth, we rewrite model (6) in matrix notation as

$$\begin{aligned} \boldsymbol{X}_i \equiv \begin{bmatrix} 1 & x_{i1} \\ \vdots & \vdots \\ 1 & x_{in_i} \end{bmatrix}, \quad \boldsymbol{X} \equiv \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_m \end{bmatrix}, \quad \boldsymbol{\beta} \equiv \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \\ \boldsymbol{Z} \equiv \begin{bmatrix} \boldsymbol{X}_1 & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \boldsymbol{X}_m \end{bmatrix} \quad \text{and} \quad \boldsymbol{u} \equiv \begin{bmatrix} u_{01} \\ u_{11} \\ \vdots \\ u_{0m} \\ u_{1m} \end{bmatrix}. \end{aligned} \tag{7}$$

Note that the random effects design matrix $\boldsymbol{Z}$ has a block-diagonal form, where each block corresponds to the i-th hospital-specific deviations from the overall intercept and slope.

We now combine the ideas introduced in Section 2 to develop a scalable iterative algorithm for approximate Bayesian inference in model (6). We first briefly introduce the following useful notation: for a scalar random variable θ, let $\boldsymbol{\mu}_{q(\theta)} \equiv E_q(\theta)$ and $\sigma_{q(\theta)} \equiv \text{Var}_q(\theta)$ be the mean and variance with respect to the approximating q-distribution. For a random vector parameter $\boldsymbol{\theta}$, we use the analogously defined $\boldsymbol{\mu}_{q(\boldsymbol{\theta})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}$. In addition, we define $\boldsymbol{M}_{q(\boldsymbol{\Theta})}$ to be the mean with respect to the approximating q-distribution for a random matrix $\boldsymbol{\Theta}$. The journey towards a practical MFVB algorithm commences with a q-density product restriction

Copyright © 2015 John Wiley & Sons, Ltd.

$$p(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{G}, a_{\mathrm{R},1}, a_{\mathrm{R},2} \mid \boldsymbol{y}) \approx q(\boldsymbol{\beta}, \boldsymbol{u})\, q(\boldsymbol{G})\, q(a_{\mathrm{R},1})\, q(a_{\mathrm{R},2}),$$

where $a_{\mathrm{R},1}$ and $a_{\mathrm{R},2}$ are auxiliary parameters defined in (5). Such restriction arises from an interaction between the initial factorization assumed in the approximating posterior and the underlying conditional independence properties of the true joint posterior [28].

To provide an example of how optimal densities are constructed, we derive the optimal density $q^*(\boldsymbol{\beta}, \boldsymbol{u})$. According to Algorithm 1, the optimal q-density for $(\boldsymbol{\beta}, \boldsymbol{u})$ satisfies

$$\begin{aligned} q^*(\boldsymbol{\beta}, \boldsymbol{u}) &\propto E_q\{\log p(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{y}, \boldsymbol{G}, a_{\mathrm{R},1}, a_{\mathrm{R},2})\} \\ &\propto E_q\{\log p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}) \times \log p(\boldsymbol{u}|\boldsymbol{G}) \times \log p(\boldsymbol{\beta})\} \\ &\propto E_q\left[\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}) - \mathbf{1}^T \log\{\mathbf{1} + \exp(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u})\} - \tfrac{1}{2}\boldsymbol{u}^T\boldsymbol{G}^{-1}\boldsymbol{u} - \tfrac{1}{2\sigma_\beta^2}\boldsymbol{\beta}^T\boldsymbol{\beta}\right]. \end{aligned} \tag{8}$$

We see that the non-quadratic convex term $-\log(1+e^x)$ in the likelihood poses a multivariate intractability problem with regard to approximate inference for $(\boldsymbol{\beta}, \boldsymbol{u})$. To get around this, we transform this convex term to a simple quadratic function $f(x; \xi)$, a trick first introduced by Jaakkola and Jordan [42]. Different values of ξ correspond to different parabolas, all of which are smaller than $-\log(1 + e^x)$. Thus, we can simplify the convex term to be the maxima of a family of parabolas [42]

$$-\log(1 + e^x) = \max_{\xi \in \mathbb{R}} \left\{ \lambda(\xi) x^2 - \tfrac{1}{2}x + \psi(\xi) \right\} \quad \text{for all} \quad x \in \mathbb{R}, \tag{9}$$

where $\lambda(\xi) \equiv -\tanh(\xi/2)/(4\xi)$ and $\psi(\xi) \equiv \xi/2 - \log(1 + e^\xi) + \xi \tanh(\xi/2)/4$. Let $\boldsymbol{C} \equiv [\boldsymbol{X}\ \boldsymbol{Z}]$ and $\boldsymbol{v} \equiv [\boldsymbol{\beta}^T\ \boldsymbol{u}^T]^T$, and substitute (9) into (8) to give the following lower bound on $q^*(\boldsymbol{\beta}, \boldsymbol{u})$:

$$\begin{aligned} q^*(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\xi}) &\geqslant \boldsymbol{y}^T\boldsymbol{C}\boldsymbol{v} - \left[\boldsymbol{v}^T\boldsymbol{C}^T \mathrm{diag}\{\lambda(\boldsymbol{\xi})\}\boldsymbol{C}\boldsymbol{v} - \frac{1}{2}\mathbf{1}^T\boldsymbol{C}\boldsymbol{v}\right] - \frac{1}{2}\boldsymbol{v}^T \begin{bmatrix} \sigma_\beta^{-2}\boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & E_q\left(\boldsymbol{G}^{-1}\right) \end{bmatrix} \boldsymbol{v} \\ &= -\frac{1}{2}\left\{ \boldsymbol{v}^T \left(2\boldsymbol{C}^T \mathrm{diag}\{\lambda(\boldsymbol{\xi})\}\boldsymbol{C} + \begin{bmatrix} \sigma_\beta^{-2}\boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & E_q\left(\boldsymbol{G}^{-1}\right) \end{bmatrix} \right) \boldsymbol{v} - \boldsymbol{C}^T\left(\boldsymbol{y} - \frac{1}{2}\mathbf{1}\right) \right\} \\ &= \log \underline{q}^*(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\xi}), \end{aligned} \tag{10}$$

where $E_{q(\boldsymbol{G}^{-1})} \equiv \boldsymbol{I}_m \otimes \boldsymbol{M}_{q(\boldsymbol{\Sigma}_{\mathrm{R}}^{-1})}$ and $\boldsymbol{\xi}$ is being introduced as a $\sum_{i=1}^m n_i \times 1$ vector of variational parameters. Scalar functions applied to vectors are evaluated element-wise. Noting that because the right-hand side of (10) is a quadratic form and by completing the square in the usual way to identify the mean and covariance, we approximate the posterior of $(\boldsymbol{\beta}, \boldsymbol{u})$ by a multivariate normal distribution

$$\begin{aligned} & q^*(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\xi}) \sim N(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}), \\ \text{where} \quad & \boldsymbol{\Sigma}_{q\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi}} \equiv \left(2\,\boldsymbol{C}^T \mathrm{diag}\{\lambda(\boldsymbol{\xi})\}\boldsymbol{C} + \begin{bmatrix} \sigma_\beta^{-1}\boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & E_q(\boldsymbol{G}^{-1}) \end{bmatrix} \right)^{-1} \\ \text{and} \quad & \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})} \equiv \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}\boldsymbol{C}^T\left(\boldsymbol{y} - \tfrac{1}{2}\mathbf{1}\right). \end{aligned}$$

The remaining q-densities involve similar adaptation of the previous derivation, and hence, we omit the details and state their functional forms directly:

$$\begin{aligned} & \boldsymbol{\xi} \leftarrow \sqrt{\mathrm{diagonal}\left\{ \boldsymbol{C}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})} + \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}^T)\boldsymbol{C}^T \right\}}, \\ & q^*(\boldsymbol{\Sigma}_{\mathrm{R}}) \text{ is the inverse-Wishart } \left(\nu + m + 1, \boldsymbol{B}_{q(\boldsymbol{\Sigma}_{\mathrm{R}})}\right) \text{ density function}, \\ & q^*(a_{\mathrm{R},r}) \text{ is the inverse-gamma } \left(1, B_{q(a_{\mathrm{R},r})}\right) \text{ density function}, \ 1 \leqslant r \leqslant 2, \end{aligned}$$

where the parameter $\boldsymbol{B}_{q(\boldsymbol{\Sigma}_{\mathrm{R}})}$ is the scale matrix of the inverse-Wishart q-density and $B_{q(a_{\mathrm{R},r})}$ is the scale parameter of the inverse-gamma q-density. Thus, all optimal densities belong to parametric families with the parameters explicitly determined by the distributions of the remaining model parameters and observed data. Taken together, these solutions lead to Algorithm 2 for approximate Bayesian inference in model (6).

Copyright © 2015 John Wiley & Sons, Ltd.

Initialize: $\boldsymbol{\xi}$ ($\sum_{i=1}^{m} n_i \times 1$; all entries positive), $\boldsymbol{M}_{q(\boldsymbol{\Sigma}_{\rm R}^{-1})}$ positive definite and $\mu_{q(a_{{\rm R},r})} > 0$, $r = 1, 2$

Cycle through updates:

Define: $E_{q(\boldsymbol{G}^{-1})} \equiv \boldsymbol{I}_m \otimes \boldsymbol{M}_{q(\boldsymbol{\Sigma}_{\rm R}^{-1})}$

Update mean and covariance matrix of Multivariate Normal $q^*(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\xi})$**:**

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})} \leftarrow \left(2\, \boldsymbol{C}^T \text{diag}\left\{\lambda(\boldsymbol{\xi})\right\} \boldsymbol{C} + \begin{bmatrix} \sigma_\beta^{-2} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & E_{q(\boldsymbol{G}^{-1})} \end{bmatrix} \right)^{-1}$$

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

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

Update mean and scale matrix of Inverse-Wishart $q^*(\boldsymbol{\Sigma}_{\rm R})$**:**

$$\boldsymbol{B}_{q(\boldsymbol{\Sigma}_{\rm R})} \leftarrow \sum_{i=1}^{m} \left(\boldsymbol{\mu}_{q(\boldsymbol{u}_i)} \boldsymbol{\mu}_{q(\boldsymbol{u}_i)} + \boldsymbol{\Sigma}_{q(\boldsymbol{u}_i)} \right) + 2\,\nu\, \text{diag}(\mu_{q(1/a_{{\rm R},1})}, \mu_{q(a_{{\rm R},2})})$$

$$\boldsymbol{M}_{q(\boldsymbol{\Sigma}_{\rm R}^{-1})} \leftarrow \tfrac{1}{2}(\nu + m + 1) \boldsymbol{B}_{q(\boldsymbol{\Sigma}_{\rm R})}^{-1}$$

Update mean and scale parameter of Inverse-Gamma $q^*(a_{{\rm R},r})$**:**

For $r = 1, 2$:

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

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

Algorithm 2: Iterative scheme for obtaining the optimal q-density functions in the Bayesian logistic mixed effects model (6).

Algorithm 2 is a basic framework for approximate inference for a simple Bayesian logistic mixed effects model that serves as a building block for the more complicated models. In what follows, we will gradually increase the complexity of model (6) and demonstrate a great advantage of the MFVB algorithms, modularity. By this we mean that concepts like main effects, interaction effects, higher-order random effects, and spline regression can be viewed as modules that can be put together into an almost endless variety of statistical models. All we require are relatively straightforward modifications on the structures of the general design matrices and random effects covariance matrix in order to accommodate larger and more complicated mixed effects models.

The Jaakkola and Jordan [42] trick is one of a few approaches that have been proposed in the variational approximation literature for handling binary response regression models. Others include the use of the probit auxiliary variable trick of Albert and Chib [43] (e.g., [44]) and Gaussian variational approximation (e.g., [45]). Various trade-offs are involved with the choice among these options. For example, Gaussian variational approximation requires numerical integration, whereas Algorithm 2 has purely algebraic updates.

3.2. Group-specific curve model

Close inspection of Figure 1 shows that the straight line assumption imposed by the random intercept and slope model is unreasonable. We seek to relax this linearity assumption by replacing the linear mean and hospital-specific functions with arbitrary smooth functions, hereafter group-specific curve model, taking the form (3), but with

$$f(x) \;=\; \beta_0 + \beta_1 x + \sum_{k=1}^{K_{\rm gbl}} u_{{\rm gbl},k} z_{{\rm gbl},k}(x) \;\; ; \;\; g_i(x) = U_{0i} + U_{1i} x + \sum_{k=1}^{K_{\rm grp}} u_{{\rm grp},ik} z_{{\rm grp},k}(x)$$

$$\text{and} \quad \boldsymbol{G} = \begin{bmatrix} \overbrace{\sigma_{\rm gbl}^2 \boldsymbol{I}_{K_{\rm gbl}}}^{\text{Overall mean}} & \overbrace{\boldsymbol{0}}^{\text{Hospital-specific}} \\ \boldsymbol{0} & \underset{1 \le i \le m}{\text{blockdiag}} \left(\begin{bmatrix} \boldsymbol{\Sigma}_{\rm R} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_{\rm grp}^2 \boldsymbol{I}_{K_{\rm grp}} \end{bmatrix} \right) \end{bmatrix}. \tag{11}$$

Copyright © 2015 John Wiley & Sons, Ltd.

This model is a direct extension of the random intercept and slope model because in (11), the overall mean curve and each hospital-specific curve have two components: a linear part (analogous to model (6)) and a nonlinear part, which allows more flexibility.

There are numerous approaches to modeling and estimating f and g_i nonparametrically. The one that is most conducive to inference via variational approximations is penalized regression splines with mixed model representation. An advantage of this approach is that one can think in terms of constructing a series of basis functions that can be used as covariates, such as $z_{\text{gbl},k}$ and $z_{\text{grp},k}$ being the respective spline bases of size K_{gbl} and K_{grp}. Our preference of $z_{\text{gbl},k}$ and $z_{\text{grp},k}$ is suitably linearly transformed cubic O'Sullivan penalized splines (Section 4 of [46]), because this leads to approximate smoothing splines with good boundary and extrapolation properties. In practice, $K_{\text{gbl}} = 25$ is a sufficient choice for most spline basis functions [47], and typically, K_{grp} is smaller than K_{gbl} because fewer basis functions are needed to handle group-specific deviations. The coefficients $u_{\text{gbl},k}$ and $u_{\text{grp},ik}$ can be considered as a measure of the basis amplitude because they regulate the roughness of the time curves. In order to avoid overfitting the data, we impose a penalty on the basis coefficients by treating them as a random sample from a normal distribution with mean 0 and variance σ^2, that is, $u_{\text{gbl},k} \overset{\text{ind.}}{\sim} N(0, \sigma^2_{\text{gbl}})$ and $u_{\text{grp},ik} \overset{\text{ind.}}{\sim} N(0, \sigma^2_{\text{grp}})$, respectively. The variances σ^2_{gbl} and σ^2_{grp} are commonly known as the smoothing parameters; thus, selection of smoothing parameters in model (11) simply reduces to variance component estimation in a mixed effect model.

From an extendability standpoint, the general design framework for mixed effect models is a particularly attractive approach that allows smoothing-type models such as group-specific curve models to be fitted as generalized linear mixed models. Moving from parametric regression to nonparametric regression using mixed model-based penalized splines involves replacing $\boldsymbol{Z}$

$$\text{from} \quad \begin{array}{c} \begin{matrix} \text{Hospital 1} & & \text{Hospital } m \end{matrix} \\ \begin{bmatrix} \boldsymbol{X}_1 & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \boldsymbol{X}_m \end{bmatrix} \end{array} \quad \text{to} \quad \begin{array}{c} \begin{matrix} \text{Overall mean} & \text{Hospital 1} & & \text{Hospital } m \end{matrix} \\ \begin{bmatrix} & \boldsymbol{X}_1 \mid \boldsymbol{Z}_{\text{grp},1} & \cdots & \boldsymbol{0} \\ \boldsymbol{Z}_{\text{gbl}} & \vdots & \ddots & \vdots \\ & \boldsymbol{0} & \cdots & \boldsymbol{X}_m \mid \boldsymbol{Z}_{\text{grp},m} \end{bmatrix} \end{array}, \tag{12}$$

where the notation $(\boldsymbol{A} \mid \boldsymbol{B})$ denotes concatenation (by columns) of matrices $\boldsymbol{A}_{(k\times m)}$ and $\boldsymbol{B}_{(k\times n)}$, $\boldsymbol{Z}_{\text{gbl}}$ is a $\sum_{i=1}^m n_i \times K_{\text{gbl}}$ random spline basis matrix for the overall mean curve, and $\boldsymbol{Z}_{\text{grp},i}$ is an $n_i \times K_{\text{grp}}$ random spline basis matrix for the i-th hospital-specific deviations. Further, even with the additional $K_{\text{gbl}} + mK_{\text{grp}}$ random effects parameters, the covariance matrix $\boldsymbol{G}$ shown in (11) remains to be a block-diagonal, but certainly larger, matrix with different entries for the variance components corresponding to the random basis coefficients for the overall mean and random effects for each hospital. Taking these into consideration, it then follows that the MFVB algorithm for model (11) simply involves modifications of Algorithm 2 to incorporate the aforementioned structural changes in the $\boldsymbol{Z}$ and $\boldsymbol{G}$ matrices, as well as updates for the additional model parameters as follows:

$$\begin{aligned}
&q^*(a_{\text{gbl}}) \text{ is the inverse-gamma } \left(1, B_{q(a_{\text{gbl}})}\right) \text{ density function,} \\
&q^*(\sigma^2_{\text{gbl}}) \text{ is the inverse-gamma } \left(\tfrac{1}{2}\left(K_{\text{gbl}} + 1\right), B_{q\left(\sigma^2_{\text{gbl}}\right)}\right) \text{ density function,} \\
&q^*(a_{\text{grp}}) \text{ is the inverse-gamma } \left(1, B_{q(a_{\text{grp}})}\right) \text{ density function, and} \\
&q^*(\sigma^2_{\text{grp}}) \text{ is the inverse-gamma } \left(\tfrac{1}{2}\left(m\,K_{\text{grp}} + 1\right), B_{q\left(\sigma^2_{\text{grp}}\right)}\right) \text{ density function,}
\end{aligned}$$

where $B_{q(a_{\text{gbl}})}, B_{q(a_{\text{grp}})}, B_{q(\sigma^2_{\text{gbl}})}$, and $B_{q(\sigma^2_{\text{grp}})}$ are the scale parameters with respect to the corresponding q-density functions.

3.3. *Factor-by-curve interactions*

The second objective of this cesarean section analysis is to examine differences in trends in cesarean section rates between low-risk nulliparous women aged less than 25 years and those aged greater than or equal to 25 years. We are therefore interested in fitting a separate time curve for each maternal age group (Figure 2). This leads to a group-specific curve model with a factor-by-curve interaction, where one can view the effect of maternal age on the probability of cesarean section varies smoothly over time. From here onwards, we refer low-risk nulliparous women aged less than 25 years as group A (younger group) and those greater or equal to 25 years as group B (older group).

Copyright © 2015 John Wiley & Sons, Ltd.

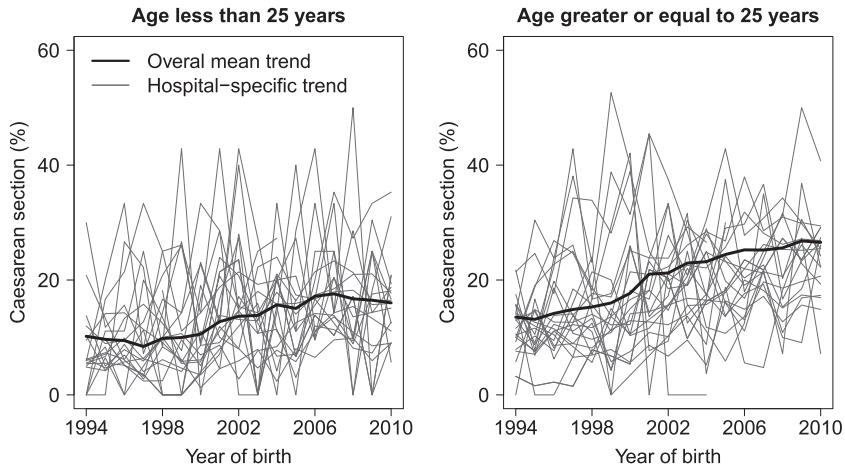

Figure 2. Trends in cesarean section rates for low-risk nulliparous women aged less than 25 years and those aged greater or equal to 25 years in New South Wales, Australia, from 1994 to 2010. The darker line is the overall mean trend, and the lighter lines are the selected hospital-specific trends.

Additive models that include factor-by-curve interactions are known to be useful but can be computationally very demanding [48, 49]. Coull, Catalaon, and Godleski [50] used a backfitting algorithm, in which, for each backfit iteration corresponding to a curve interaction term, one splits the data into subsets and fits a smooth function to each subset. However, this backfitting approach can be computationally impractical when the number of interaction terms or the number of data subsets becomes large. Maringwa *et al*. [51] used penalized regression splines in a mixed model framework to compare population-averaged profiles but only focused on the random intercept models without any group-specific curves. The penalized spline approach is preferable because the former data subsetting approach must be nested within a backfitting algorithm. We follow on Maringwa *et al.* [51] and incorporate a factor-by-curve interaction into the group-specific curve model (11). Let $I_{ij}^{\text{A}} = 1$ if (x_{ij}, y_{ij}) belongs to a low-risk nulliparous woman of aged less than 25 years and zero otherwise; we propose a model of the following form:

$$y_{ij} \overset{\text{ind.}}{\sim} \text{Bernoulli}\left(\text{logit}^{-1}\left[I_{ij}^{\text{A}}\left\{f^{\text{A}}\left(x_{ij}\right) + g_i^{\text{A}}\left(x_{ij}\right)\right\} + \left(1 - I_{ij}^{\text{A}}\right)\left\{f^{\text{B}}\left(x_{ij}\right) + g_i^{\text{B}}\left(x_{ij}\right)\right\}\right]\right),$$

where

$$\begin{aligned}
f^{\text{A}}(x) &= \beta_0^{\text{A}} + \beta_1^{\text{A}}\, x + \sum_{k=1}^{K_{\text{gbl}}} u_{\text{gbl},k}^{\text{A}}\, z_{\text{gbl},k}(x), \\
g_i^{\text{A}}(x) &= U_{0i}^{\text{A}} + U_{1i}^{\text{A}}\, x \sum_{k=1}^{K_{\text{grp}}} u_{\text{grp},ik}^{\text{A}}\, z_{\text{grp},k}(x), \\
f^{\text{B}}(x) &= \beta_0^{\text{A}} + \beta_0^{\text{BvsA}} + (\beta_1^{\text{A}} + \beta_1^{\text{BvsA}})\, x + \sum_{k=1}^{K_{\text{gbl}}} u_{\text{gbl},k}^{\text{B}}\, z_{\text{gbl},k}(x), \\
g_i^{\text{B}}(x) &= U_{0i}^{\text{B}} + U_{1i}^{\text{B}}\, x \sum_{k=1}^{K_{\text{grp}}} u_{\text{grp},ik}^{\text{B}}\, z_{\text{grp},k}(x)
\end{aligned} \tag{13}$$

$$\text{and} \quad \boldsymbol{G} = \begin{bmatrix} \overbrace{\begin{matrix} (\sigma_{\text{gbl}}^{\text{A}})^2 \boldsymbol{I}_{K_{\text{gbl}}} & \boldsymbol{0} \\ \boldsymbol{0} & (\sigma_{\text{gbl}}^{\text{B}})^2 \boldsymbol{I}_{K_{\text{gbl}}} \end{matrix}}^{\text{Overall mean}} & \overbrace{\begin{matrix} \boldsymbol{0} \\ \boldsymbol{0} \end{matrix}}^{\text{Hospital-specific}} \\ \boldsymbol{0} & \underset{1\leqslant i\leqslant m}{\text{blockdiag}}\left(\begin{bmatrix} \boldsymbol{\Sigma}_{\text{R}} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_{\text{grp}}^2 \boldsymbol{I}_{K_{\text{grp}}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \sigma_{\text{grp}}^2 \boldsymbol{I}_{K_{\text{grp}}} \end{bmatrix}\right) \end{bmatrix}.$$

Copyright © 2015 John Wiley & Sons, Ltd.

We consider low-risk women aged less than 25 years as our reference group and define the corresponding f^{A} and g_i^{A} functions the same way as in model (11). For low-risk women aged greater or equal to 25 years, the f^{B} function is defined slightly differently by introducing the parameters β_0^{BvsA} and β_1^{BvsA} that represent the differences in overall intercepts and slopes between low-risk nulliparous women of younger and older age. This gives an asymmetrical structure in the fixed effects formulation. While asymmetrical formulation of fixed effects is common in the presence of a reference group, we do not recommend such structure imposed on the random spline coefficients because it would induce restriction on the smoothness of a particular curve. We impose a different variance parameter for each overall mean curve, allowing the level of smoothing to differ between curves. A simpler model would be to assume a common variance parameter for all overall mean curves, that is, $\sigma^{\text{A}}_{\text{gbl}} = \sigma^{\text{B}}_{\text{gbl}} = \sigma_{\text{gbl}}$, meaning that all curves have equivalent smoothness, but with different forms of shape. In addition, we account for the potential correlation between the random intercepts and slopes of each of the women age groups within the same hospital using an unstructured 4×4 covariance matrix, that is, $\left[U^{\text{A}}_{0i}\ U^{\text{A}}_{1i}\ U^{\text{B}}_{0i}\ U^{\text{A}}_{1i}\right]^T \overset{\text{ind.}}{\sim} N(\mathbf{0}, \mathbf{\Sigma}_{\text{R}})$.

The differences in trends in cesarean section rates for low-risk nulliparous women aged greater or equal to 25 years versus those aged less than 25 years can be quantified via estimation of a so-called contrast function. At the population level, this is simply obtained by subtracting the two overall mean curves:

$$c^{\text{BvsA}}(x) \equiv f^{\text{B}}(x) - f^{\text{A}}(x) = \beta_0^{\text{BvsA}} + \beta_1^{\text{BvsA}}\, x + \sum_{k=1}^{K_{\text{gbl}}} \left(u^{\text{B}}_{\text{gbl},k} - u^{\text{A}}_{\text{gbl},k}\right) z_{\text{gbl},k}(x). \tag{14}$$

This contrast function can be interpreted as the log odds of cesarean section (averaged across hospitals) for low-risk nulliparous women aged greater or equal to 25 years, as a function of time, compared with those aged less than 25 years.

The structural changes in the random effects matrices $\boldsymbol{Z}$ and $\boldsymbol{G}$ underpin the complexity of a mixed effect model being transitioned from the simplest random intercept and slope model to a more complicated model including a factor-by-curve interaction. For example, Figure 3 shows that the $\boldsymbol{Z}$ matrix starts off as having a simple block-diagonal structure and, as the model increases in complexity, it grows into a 'nested' block-diagonal structure (each large block is itself block-diagonal, with one or more small blocks on the diagonal). As the number of random spline basis functions increases in the model, the dimensions of the $\boldsymbol{Z}$ and $\boldsymbol{G}$ matrices increase accordingly. While these changes do not directly affect the algebraic aspect of Algorithm 2 because of the advantage of modularity, one must be aware that the update of the covariance matrix $\mathbf{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}$ now requires storage and inversion of a large sparse matrix. Specifically, the number of columns in $\mathbf{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}$ for each considered model is

1. Random intercepts and slopes $\quad P + m\,q^{\text{R}}$
2. Group-specific curves $\quad P + K_{\text{gbl}} + m\,\left(q^{\text{R}} + K_{\text{grp}}\right)$
3. Factor-by-curve interactions $\quad P + 2K_{\text{gbl}} + 2\,m\,\left(q^{\text{R}} + K_{\text{grp}}\right)$

Without doubt, the number of groups m dominates the dimension of $\mathbf{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}$. For example, if $P = 10, m = 1000, K_{\text{gbl}} = 25,\ q^{\text{R}} = 2$, and $K_{\text{grp}} = 10$, then the dimension of the matrix requiring storage and inversion for the most complicated model would be $24\,060 \times 24\,060$. This aspect renders Algorithm 2 infeasible for very large and complex longitudinal and multilevel models.

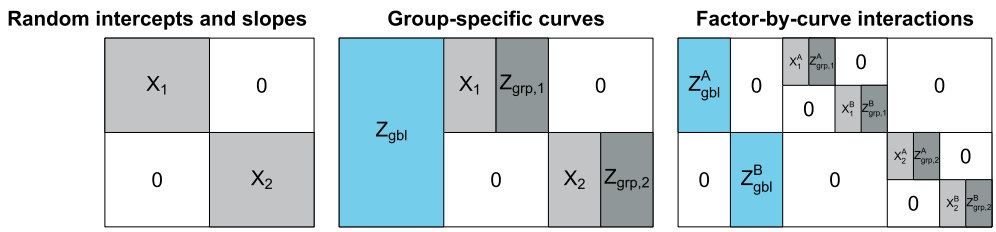

Figure 3. Various structures of the $\boldsymbol{Z}$ matrix across logistic mixed effects models (6), (11), and (13), with the number of groups $m = 2$. The $\boldsymbol{Z}$ matrix starts off as having a simple block-diagonal structure, and as the model increases in complexity, it grows into a 'nested' block-diagonal structure. The definitions of matrices are described in Subsection 3.2.

Copyright © 2015 John Wiley & Sons, Ltd.

As pointed out by Lee and Wand [21], the naïve implementation of MFVB algorithms for arbitrarily large grouped data is extremely inefficient in terms of speed and storage. The time complexity can be as high as $O(m^3)$, making variational inference impractical for large and complex mixed effects models. Through exploiting the inherent block-diagonal structure of the random effects covariance matrix, the authors developed fast and memory-efficient MFVB algorithms that streamline inversion and update for $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}$. Here, we briefly reiterate their streamlined approach in more general terms. Recall that the general update expression for $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}$ in Algorithm 2 involves inversion of the following matrix:

$$2\,\boldsymbol{C}^T \mathrm{diag}\left\{\lambda(\boldsymbol{\xi})\right\}\boldsymbol{C} + \begin{bmatrix} \sigma_\beta^{-2}\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & E_{q(\boldsymbol{G}^{-1})} \end{bmatrix} = \left[\begin{array}{c|cccc} \boldsymbol{U} & \boldsymbol{V}_1 & \boldsymbol{V}_2 & \cdots & \boldsymbol{V}_m \\ \hline \boldsymbol{V}_1^T & \boldsymbol{W}_1^{-1} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{V}_2^T & \boldsymbol{0} & \boldsymbol{W}_2^{-1} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{V}_m^T & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{W}_m^{-1} \end{array}\right]. \tag{15}$$

The central idea of the streamlined approach requires permuting the previous matrix into an approximate block-diagonal form for decomposition, as shown in (15). The matrices $\boldsymbol{U}, \boldsymbol{V}_i$, and $\boldsymbol{W}_i^{-1}$ are usually of general forms with small dimension (Appendix of [21]); each can be easily derived via straightforward matrix manipulation. With this transformation, we can efficiently invert the block-partitioned form of (15) by solving a system of simultaneous linear equations in matrix form [52]

$$\overbrace{\begin{bmatrix} \boldsymbol{U} & \boldsymbol{V} \\ \boldsymbol{V}^T & \boldsymbol{W} \end{bmatrix}}^{\boldsymbol{\Sigma}^{-1}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}} \overbrace{\begin{bmatrix} \boldsymbol{X} & \boldsymbol{Y} \\ \boldsymbol{Y}^T & \boldsymbol{Z} \end{bmatrix}}^{\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix}.$$

Standard calculations lead to the following forms of submatrices:

$$\begin{aligned} \boldsymbol{X} &\equiv \left(\boldsymbol{U} - \boldsymbol{V}\boldsymbol{W}^{-1}\boldsymbol{V}^T\right)^{-1} \\ \boldsymbol{Y} &\equiv -\boldsymbol{X}\boldsymbol{V}\boldsymbol{W}^{-1} \\ \text{and} \quad \boldsymbol{Z} &\equiv \boldsymbol{W}^{-1} + \boldsymbol{W}^{-1}\boldsymbol{V}^T\boldsymbol{X}\boldsymbol{V}\boldsymbol{W}^{-1}, \end{aligned}$$

assuming the inverse matrices $\boldsymbol{W}^{-1}$ and $(\boldsymbol{U} - \boldsymbol{V}\boldsymbol{W}^{-1}\boldsymbol{V}^T)^{-1}$ in the preceding text exist. It is worth noting that the matrix $\boldsymbol{Z}$ is not a block-diagonal matrix. However, because the covariance between the fitted values of two different groups is rarely of interest, it suffices to compute and store the diagonal blocks. Appendix A provides details of the streamlined MFVB algorithm for model (13).

4. Simulation study

In this section, we conducted a comprehensive simulation study to evaluate the performance of Algorithm 2 in terms of Bayesian inferential accuracy and computational speed. We simulated 30 independent datasets from model (11) and used 25 knots for the overall mean function and 10 knots for the group-specific deviation functions:

$$\begin{aligned} y_{ij} &\overset{\text{ind.}}{\sim} \text{Bernoulli}\left(\text{logit}^{-1}\left\{f\left(x_{ij}\right) + g\left(x_{ij}\right)\right\}\right), \\ \text{where} \quad f(x) &= 10\,\phi(x; 1.5, 0.3) + 6\,\phi(x; 4, 0.6), \\ g_i(x) &= \alpha \sin\left(2\pi x^\beta\right) \quad ; \quad x_{ij} \sim \text{Uniform}\,(0, 1/n), \\ \alpha &\sim N\,(0, 0.5) \quad \text{and} \quad \beta \in \{1,\ 2,\ 3\}. \end{aligned} \tag{16}$$

Each dataset consists of $m = 50$ groups, and the number of observations per group is balanced with $n_i = n = 50$. The term $\phi(x; a, b)$ is an univariate normal density with mean a and standard deviation b.

We fitted each replicated dataset using both MFVB approximation via a streamlined version of Algorithm 2 and MCMC sampling. The MFVB algorithm was implemented in the R computing environment [53], and its stopping criterion is when the relative change in the lower bound $\underline{p}(\boldsymbol{y}; q)$ falls below

Copyright © 2015 John Wiley & Sons, Ltd.

10^{-8}. The MCMC samples were obtained using the R package `RStan` [24]. In each dataset, MCMC samples of size 10,000 were generated, with the first 5000 values of each sample discarding as burn-in and the remaining 5000 values thinning by a factor of 5. Trace plots and the autocorrelation functions for all model parameters were used to assess MCMC convergence.

4.1. Assessment of inferential accuracy

For a generic parameter θ, we assessed the accuracy of the MFVB approximation by comparing the optimal q-density function $q^*(\theta)$ with a highly accurate MCMC-based posterior approximation $p_{\text{MCMC}}(\theta|\boldsymbol{y})$. While there are numerous means of measuring accuracy, we recommend working with a measure that is based on the L_1 distance, also known as the integrated absolute error [54] of $q^*(\theta)$, given by

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

This accuracy measure has the attraction of being invariant to monotone transformations on the parameter θ. Exact computation of $p_{\text{MCMC}}(\theta|\boldsymbol{y})$ is numerically challenging, and hence, we used binned kernel density estimation with direct plug-in bandwidth selection, as facilitated in the R package `KernSmooth` [55].

Figures 4 and 5 display the approximate posterior density functions and side-by-side boxplots of accuracy scores for the model parameters $f(Q_k)$, $1 \leqslant k \leqslant 3$ and $g_i(Q_k)$, $k = 2$, where Q_k is the k-th

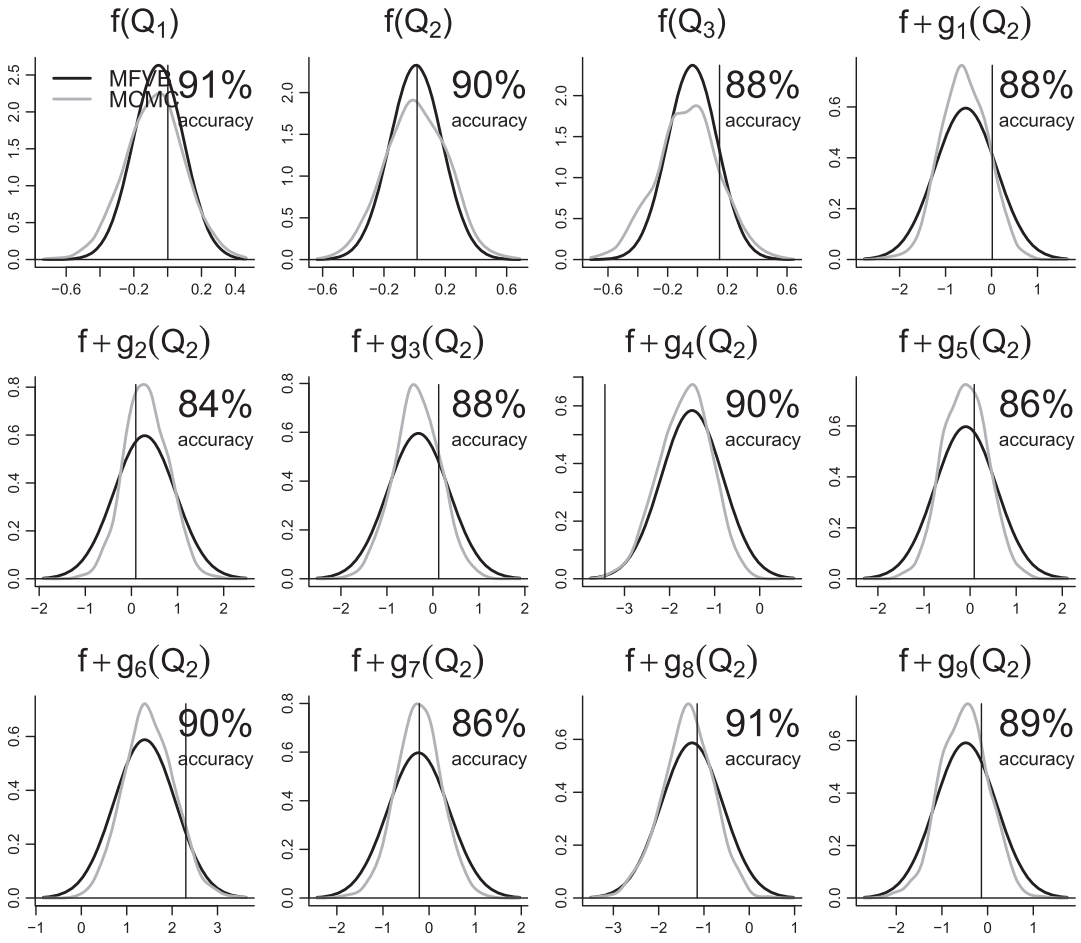

Figure 4. The approximate posterior density functions obtained from mean field variational Bayes (MFVB) and MCMC for a single replication of the simulation study described in the text. Each pair of density function corresponds to a model parameter $f(Q_k)$, $1 \leqslant k \leqslant 3$ and $g_i(Q_k)$, $k = 2$, where Q_k is the k-th sample quintile of the x's. The vertical lines represent the true parameter values. The accuracy scores on the top right of on each plot show the accuracy of MFVB approximation compared against an MCMC benchmark.

Copyright © 2015 John Wiley & Sons, Ltd.

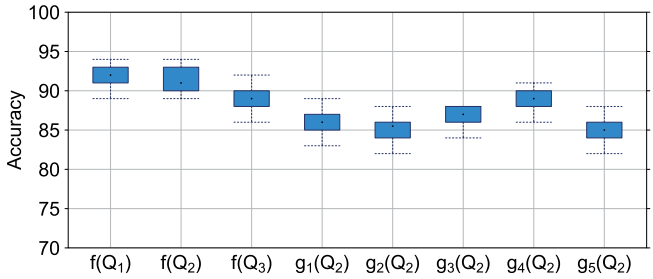

Figure 5. Side-by-side boxplots of accuracy scores for the mean field variational Bayes approximation compared against MCMC over 30 runs. Each boxplot corresponds to a model parameter $f(Q_k)$, $1 \leqslant k \leqslant 3$ and $g_i(Q_k)$, $k = 2$, where Q_k is the k-th sample quintile of the x's.

Table I. Average (standard error) elapsed of the computing times in seconds for the streamlined MFVB Algorithm 2 in the simulation setting described in (16).

Number of groups (m)	100	500	1000
Within-group size (n_i)			
10	65.1 (0.86)	351.6 (5.49)	709.4 (4.42)
50	127.1 (2.57)	826.7 (9.63)	1859.2 (6.15)
100	106.5 (1.51)	929.3 (10.77)	2870.3 (13.62)

MFVB, mean field variational Bayes.

sample quintile of the x's. The boxplots show that the majority of the accuracy scores exceed 80%, with some over 90%. As indicated by Figure 4, the MFVB credible intervals generally cover the true parameter values. The average elapsed time for the MCMC fits is 7.1 h (standard error 3.0 h) while 1.9 min (standard error 4.7 s) for the MFVB fits. This corresponds to a speedup in the order of several hundreds.

4.2. Assessment of computational speed

We now turn our attention to quantification of the speed gains afforded by the streamlined MFVB algorithm. The simulation described in Section 4 was rerun using MFVB with a combination of m and n values and MCMC omitted. All computations were performed on a Mac OS X laptop (Apple Pty. Limited, Sydney, Australia) with a 2.6 GHz Intel Core i7 processor and 4 GB of random access memory. Table I summarizes the average (standard error) computing times over 30 runs and highlights the practical and scalability benefits of the streamlined MFVB approach.

It is well-established that MCMC can be very slow in situations where complex models are applied to large datasets. For the setting in Table I, we expect the MCMC fitting takes days to weeks to run, and therefore, a similar timing comparison between MFVB and MCMC is not practical.

5. Application to cesarean section data

As described in Section 3, we applied our streamlined MFVB algorithms to the cesarean section data, with the aim of characterizing trends in cesarean section rates for low-risk nulliparous women aged less than 25 years and those aged greater or equal to 25 years in NSW hospitals between 1994 and 2010. A group-specific curve model with a factor-by-curve interaction was fitted to data, allowing for nonlinear estimation of time courses as seen in Figure 2. The `RStan` code for the model specification is given in Appendix B.

From 1994 to 2010, there were 295,340 low-risk nulliparous women giving birth for the first time in 99 NSW public or private hospitals. Of these, 73,795 women (24.5%) aged less than 25 years and 221,545 women (75.0%) aged greater or equal to 25 years. The annual rate of cesarean section among low-risk nulliparous women has increased radically from 12.5% to 24.1% in NSW over this 17-year period; however, rates for women of older age were consistently higher than those for women of younger

Copyright © 2015 John Wiley & Sons, Ltd.

age (younger group: 10.2% to 16.0%; older group: 13.5% to 26.6%). Figure 6 shows the fitted probability function of cesarean section for each hospital between 1994 and 2010. Compared with low-risk women aged less than 25 years, those aged greater or equal to 25 years are more likely to have a cesarean section, rather than a vaginal delivery. The fitted probabilities are seen to vary markedly among hospitals and over time, with some hospitals exhibiting a consistently high probability and others exhibiting an upside-down U shape. It is evident that the hospital-specific probability functions of cesarean section cannot be adequately modeled by linear functions.

Figure 7 shows the estimated overall and selected hospital-specific contrast curves defined by (14), corresponding to the odds of cesarean section for low-risk nulliparous women aged greater or equal to 25 years, as a function of time, compared with those aged less than 25 years. The pointwise 95% credible sets, shown in the first panel by the gray-shaded region, suggest that there are significant differences in trends in cesarean section rates between the two maternal age groups over the year of birth. Low-risk nulliparous women of older age are on average 1.3 times more likely to have a cesarean section compared with those of younger age in the year 1994. This odds ratio increases as the year of birth increases.

To examine hospital differences in odds of cesarean section for low-risk nulliparous women of older age compared with those of younger age, we calculated and plotted the hospital-specific odds ratios for the latest year of birth (Table II). In 2010, the statewide odds ratio is 1.74, meaning that, on average, a low-risk nulliparous woman of older age giving birth by cesarean section is about 1.74 times more likely than that of a woman of younger age. There is a wide variation in odds ratios among hospitals, ranging from 1.21 to 2.24 (Figure 8).

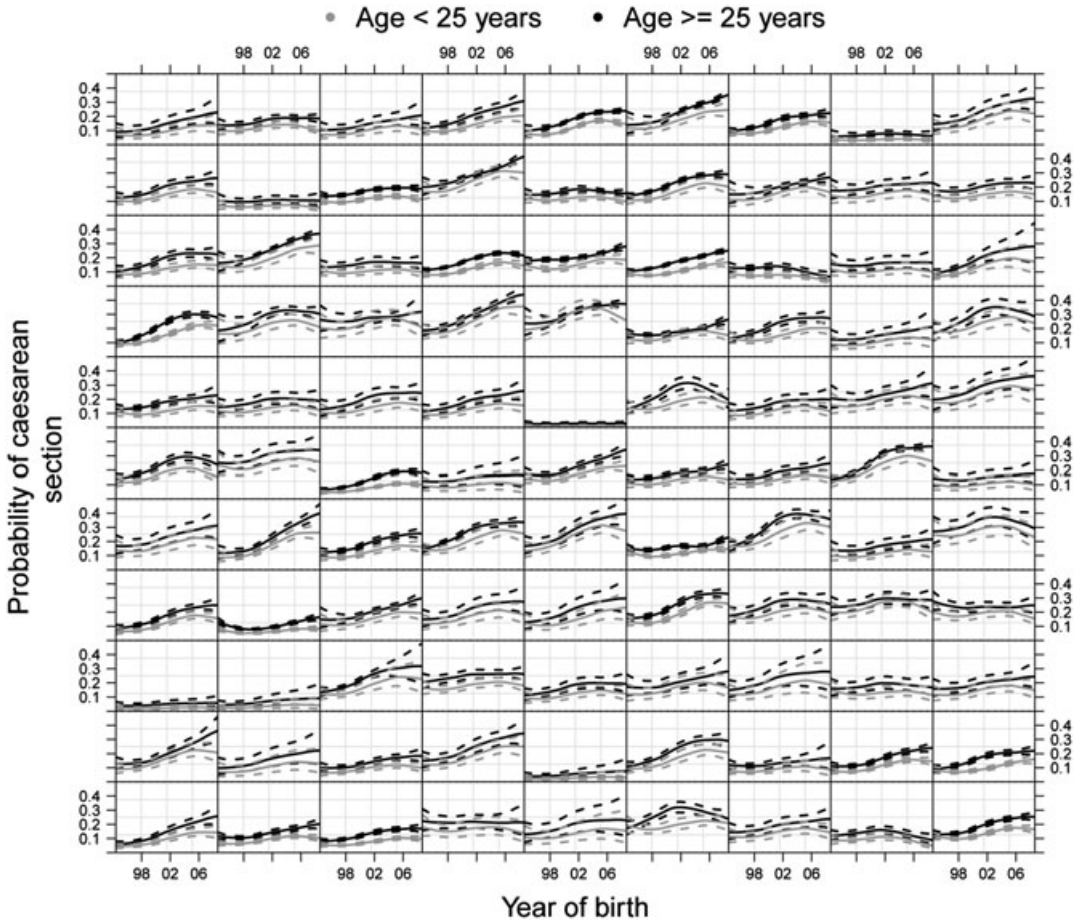

Figure 6. The mean field variational Bayes fitted hospital-specific probability functions of cesarean section for low-risk nulliparous women of aged less than 25 years and those of aged greater or equal to 25 years, as a function of time for each hospital. The dashed curves represent pointwise 95% credible sets. Each panel corresponds to a different hospital.

Copyright © 2015 John Wiley & Sons, Ltd.

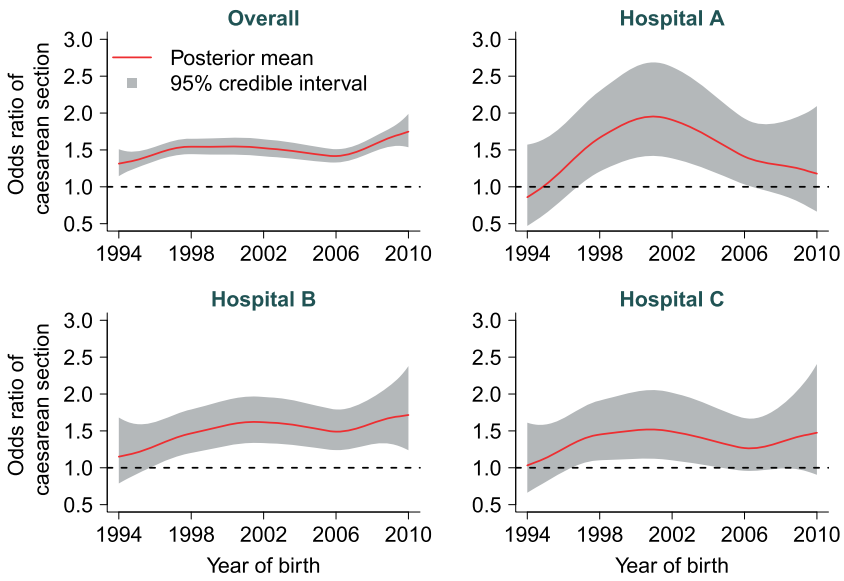

Figure 7. The mean field variational Bayes estimated overall and selected hospital-specific contrast curves defined by (14), corresponding to the odds of cesarean section for low-risk nulliparous women aged greater or equal to 25 years, as a function of time, compared with those aged less than 25 years. The shaded regions correspond to pointwise 95% credible sets.

Table II. The MFVB estimated odds ratios of cesarean section (averaged across hospitals) for low-risk nulliparous women aged greater than or equal to 25 years compared with those aged less than 25 years for selected years of birth.

Year	Odds ratio	95% Credible interval
1994	1.31	(1.14, 1.50)
1998	1.42	(1.24, 1.62)
2002	1.52	(1.34, 1.74)
2006	1.63	(1.44, 1.86)
2010	1.74	(1.53, 1.98)

MFVB, mean field variational Bayes.

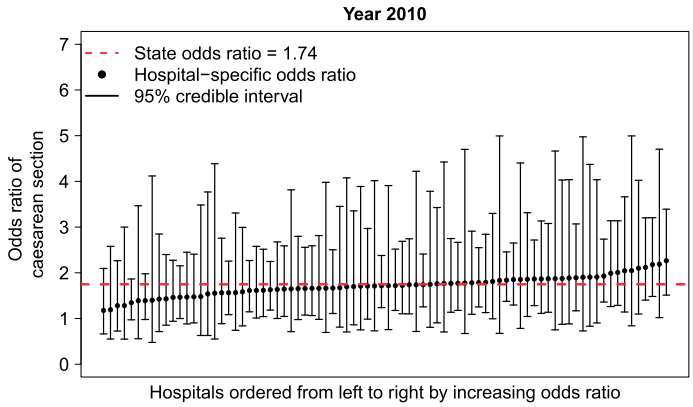

Figure 8. The mean field variational Bayes estimated hospital-specific odds ratios of cesarean section for low-risk nulliparous women of older age compared with those of younger age in the year 2010.

Copyright © 2015 John Wiley & Sons, Ltd.

6. Discussion

We have described MFVB methods for fast approximate inference pertaining to three Bayesian logistic mixed effects models. Their utility to describe the overall mean and hospital-specific trends in cesarean section rates in NSW between 1994 and 2010 has been demonstrated. Using MCMC methods as a benchmark, we have evaluated the inferential accuracy and computation speed of our streamlined MFVB algorithms.

Compared with MCMC methods, which directly generate Markov chain samples from the target posterior distribution, MFVB methods seek an approximating distribution in a factorized form that has minimum KL distance to the target posterior. Although both methods are an approximation to the true posterior, MCMC methods are theoretically guaranteed to converge asymptotically to the true posterior (but come at great expense), while MFVB methods are known to underestimate the true variance because of the factorization assumption [56]. Our simulation study shows moderately good to excellent accuracy of the MFVB approximation for all posterior densities of trend parameters. There is some underestimation in the posterior standard deviations, which may also be attributable to the Jaakkola and Jordan's approach described in (9). Despite minor degradation in accuracy, we believe that the MFVB approximation is a strong complement to the standard MCMC methods, as a way to efficiently explore possible models for revealing the complex nonlinear temporal evolution of disease rates and identifying regions with unusual patterns at any point in time. Perhaps more practically, the MFVB parameter estimates can be used as starting points for MCMC to eliminate the need for a burn-in. In addition, our simulation results highlight the key advantage of the MFVB methods – computational efficiency. It serves as a great tool for iterative model building/construction. Even with concerns over accuracy in mind, MFVB methods provide an easy way to experiment with a range of models of different specifications, predictors, and random effects covariance structures. As illustrated, the computational time for the MFVB algorithms is significantly shorter than that required for MCMC, with answers delivered in minutes rather than hours, although considerations such as computing environment and convergence criterion need to be taken into account to allow a fairer speed comparison. One of the more computationally expensive steps in the MFVB algorithms is the matrix inversion associated with the update of the covariance matrix of $q^*(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\xi})$, a cost that grows cubically in the number of groups. However, through matrix permutation and block decomposition, we derived a streamlined approach that reduces the number of operations to be linear in the number of groups and is memory efficient.

The MFVB methods we considered here are fast and versatile and can be easily extended to more complicated scenarios. For example, the methods allow arbitrary priors for the hyperparameters [57] and similar types of model with Gaussian responses [18, 20, 21]. Stewart [22] provides great examples of more elaborate models within the context of social sciences. The ability to handle binary outcomes is particularly useful in health science studies, where binary outcomes are common. For variational inference in logistic regression, we followed on Jaakkola and Jordan [42] justifying based on constructing a lower bound for the marginal likelihood using convex duality [42] and derived closed-form expressions that approximate the posterior distributions for the parameters; thus, numerical quadrature techniques are not required.

In this article, we presented a flexible estimation of population-level and hospital-level trends in cesarean section rates using a penalized spline basis function approach. Apart from being conceptually appealing, the approach allows the resultant models to be couched within a Bayesian mixed effects model framework for approximate inference. The presented group-specific curve models incorporate both temporal and hospital variabilities into a cohesive modeling framework, enabling one to visually describe the dynamic evolution of outcome rates across hospitals and over time. Our proposed contrast functions are useful in identifying particular regions of the temporal profiles that show significant differences in odds of cesarean section between low-risk nulliparous women of older and younger age. These regions are potential priority targets for public health interventions that aim at reducing temporal and hospital variations in cesarean section rates.

Appendix A. Derivation of the mean field variational Bayes algorithm for model (13)

The full Bayesian representation for model (13) is

Copyright © 2015 John Wiley & Sons, Ltd.

$$\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{u} \sim \text{Bernoulli}\left(\text{logit}^{-1}\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}\right)\right), \quad \boldsymbol{\beta} \sim N\left(\boldsymbol{0}, \sigma_\beta^2 \boldsymbol{I}\right),$$

$$\boldsymbol{u} \mid \sigma^{\text{A}}_{\text{gbl}}, \sigma^{\text{B}}_{\text{gbl}}, \boldsymbol{\Sigma}_{\text{R}}, \sigma_{\text{grp}} \sim N\left(\boldsymbol{0}, \begin{bmatrix} \left(\sigma^{\text{A}}_{\text{gbl}}\right)^2 \boldsymbol{I}_{K_{\text{gbl}}} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \left(\sigma^{\text{B}}_{\text{gbl}}\right)^2 \boldsymbol{I}_{K_{\text{gbl}}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \underset{1 \leqslant i \leqslant m}{\text{blockdiag}}\left(\begin{bmatrix} \boldsymbol{\Sigma}_{\text{R}} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma^2_{\text{grp}} \boldsymbol{I}_{K_{\text{grp}}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \sigma^2_{\text{grp}} \boldsymbol{I}_{K_{\text{grp}}} \end{bmatrix}\right) \end{bmatrix}\right),$$

$$\boldsymbol{\Sigma}_{\text{R}} \mid a_{\text{R},1}, a_{\text{R},2}, a_{\text{R},3}, a_{\text{R},4} \sim \text{inverse-Wishart}\left(\nu + 3, 2\nu\, \text{diag}\left(1/a_{\text{R},1}, \ldots, 1/a_{\text{R},4}\right)\right),$$

$$a_{\text{R},1}, a_{\text{R},2}, a_{\text{R},3}, a_{\text{R},4} \overset{\text{ind.}}{\sim} \text{inverse-gamma}\left(\tfrac{1}{2}, A_{\text{R}}^{-2}\right),$$

$$\left(\sigma^{\text{A}}_{\text{gbl}}\right)^2 \mid a^{\text{A}}_{\text{gbl}} \sim \text{inverse-gamma}\left(\tfrac{1}{2}, 1/a^{\text{A}}_{\text{gbl}}\right), \quad a^{\text{A}}_{\text{gbl}} \sim \text{inverse-gamma}\left(\tfrac{1}{2}, A^{-2}_{\text{gbl}}\right),$$

$$\left(\sigma^{\text{B}}_{\text{gbl}}\right)^2 \mid a^{\text{B}}_{\text{gbl}} \sim \text{inverse-gamma}\left(\tfrac{1}{2}, 1/a^{\text{B}}_{\text{gbl}}\right), \quad a^{\text{B}}_{\text{gbl}} \sim \text{inverse-gamma,}\left(\tfrac{1}{2}, A^{-2}_{\text{gbl}}\right),$$

$$\sigma^2_{\text{grp}} \mid a_{\text{grp}} \sim \text{inverse-gamma}\left(\tfrac{1}{2}, 1/a_{\text{grp}}\right), \quad a_{\text{grp}} \sim \text{inverse-gamma,}\left(\tfrac{1}{2}, A^{-2}_{\text{grp}}\right).$$

Calculations similar to those in Subsection 3.1 are used to approximate the full joint posterior density function p by a product of approximating q-density functions:

$$p\left(\boldsymbol{\beta}, \boldsymbol{u}, a^{\text{A}}_{\text{gbl}}, a^{\text{B}}_{\text{gbl}}, a^{\text{A}}_{\text{grp}}, a^{\text{B}}_{\text{grp}}, \boldsymbol{a}_{\text{R}}, \sigma^{\text{A}}_{\text{gbl}}, \sigma^{\text{B}}_{\text{gbl}}, \sigma_{\text{grp}}, \boldsymbol{\Sigma}_{\text{R}}\right)$$
$$\approx q(\boldsymbol{\beta}, \boldsymbol{u})\, q\left(a^{\text{A}}_{\text{gbl}}\right)\, q\left(a^{\text{B}}_{\text{gbl}}\right)\, q\left(a_{\text{grp}}\right)\; q\left(\boldsymbol{a}_{\text{R}}\right)\; q\left(\sigma^{\text{A}}_{\text{gbl}}\right)\, q\left(\sigma^{\text{B}}_{\text{gbl}}\right)\, q\left(\sigma_{\text{grp}}\right)\; q(\boldsymbol{\Sigma}_{\text{R}}),$$

with the optimal q-density functions admitting the following forms:

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

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

$q^*\left(a^{\text{A}}_{\text{gbl}}\right)$ is the inverse-gamma $\left(1, B_{q\left(a^{\text{A}}_{\text{gbl}}\right)}\right)$ density function,

$q^*\left(a^{\text{B}}_{\text{gbl}}\right)$ is the inverse-gamma $\left(1, B_{q\left(a^{\text{B}}_{\text{gbl}}\right)}\right)$ density function,

$q^*(a_{\text{R},r})$ is the inverse-gamma $\left(1, B_{q(a_{\text{R},r})}\right)$ density function, $1 \leqslant r \leqslant 4$,

$q^*\left(\left(\sigma^{\text{A}}_{\text{gbl}}\right)^2\right)$ is the inverse-gamma $\left(\tfrac{1}{2}(K_{\text{gbl}} + 1), B_{q\left(\left(\sigma^{\text{A}}_{\text{gbl}}\right)^2\right)}\right)$ density function,

$q^*\left(\left(\sigma^{\text{B}}_{\text{gbl}}\right)^2\right)$ is the inverse-gamma $\left(\tfrac{1}{2}\left(K_{\text{gbl}} + 1\right), B_{q\left(\left(\sigma^{\text{B}}_{\text{gbl}}\right)^2\right)}\right)$ density function,

$q^*\left(\sigma^2_{\text{grp}}\right)$ is the inverse-gamma $\left(\tfrac{1}{2}\left(m\, K_{\text{grp}} + 1\right), B_{q\left(\sigma^2_{\text{grp}}\right)}\right)$ density function, and

$q^*\left(\boldsymbol{\Sigma}_{\text{R}}\right)$ is the inverse-Wishart $\left(\nu + m + 1, \boldsymbol{B}_{q(\boldsymbol{\Sigma}_{\text{R}})}\right)$ density function.

The parameters $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}$ are the respective mean vector and covariance matrix of $q^*(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\xi})$. The parameter $B_{q(\cdot)}$ is the scale parameter of the inverse-gamma q-density. Appendix B of Wand and Ormerod [31] provides a step-by step derivation for the optimal q-densities of similar type.

Mixed model representation

Using the equivalence between penalized splines and mixed models, we now aim to express model (13) within the mixed model framework. Define the following linear and nonlinear predictor vectors and random effects vectors corresponding to the i-th group to be as follows:

Copyright © 2015 John Wiley & Sons, Ltd.

$$\boldsymbol{X}^*_i \equiv \left[\mathbf{1}\ \boldsymbol{x}_i\right]\ ;\ \boldsymbol{Z}_{\text{gbl},i} \equiv \left[z_{\text{gbl},1}\left(\boldsymbol{x}_i\right)\ \cdots\ z_{\text{gbl},K_{\text{gbl}}}\left(\boldsymbol{x}_i\right)\right],$$

$$\boldsymbol{Z}_{\text{grp},i} \equiv \left[z_{\text{grp},1}\left(\boldsymbol{x}_i\right)\ \cdots\ z_{\text{grp},K_{\text{grp}}}\left(\boldsymbol{x}_i\right)\right],$$

$$\boldsymbol{u}^{\text{A}}_{\text{grp},i} \equiv \left[u^{\text{A}}_{\text{grp},i1}\ \cdots\ u^{\text{A}}_{\text{grp},iK_{\text{grp}}}\right]^T\ ;\ \boldsymbol{u}^{\text{B}}_{\text{grp},i} \equiv \left[u^{\text{B}}_{\text{grp},i1}\ \cdots\ u^{\text{B}}_{\text{grp},iK_{\text{grp}}}\right]^T,$$

$$\text{and}\quad \boldsymbol{u}_{\text{R},i} \equiv \left[U^{\text{A}}_{0i}\ U^{\text{A}}_{1i}\ U^{\text{B}}_{0i}\ U^{\text{B}}_{1i}\right]^T,$$

where $\boldsymbol{x}_i$ is an $n_i \times 1$ vector containing the x_{ij}.

In Section 2 of Zhao *et al*., [41] notation, the compact matrix form of model (13) is

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

which leads to

$$\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} = \boldsymbol{X}^{\text{R}}\boldsymbol{\beta}^{\text{R}} + \boldsymbol{X}^{\text{G}}\boldsymbol{\beta}^{\text{G}} + \boldsymbol{Z}^{\text{R}}\boldsymbol{u}^{\text{R}} + \boldsymbol{Z}^{\text{G}}\boldsymbol{u}^{\text{G}}.$$

Define $\boldsymbol{I}^{\text{A}}_i = \mathbf{1}$ if $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ is of type A and zero otherwise. The matrices $\boldsymbol{X}^{\text{R}}$ and $\boldsymbol{Z}^{\text{R}}$ are random design matrices corresponding to the fixed effects vector $\boldsymbol{\beta}^{\text{R}}$ and random group effects vector $\boldsymbol{u}^{\text{R}}$ (superscript R), respectively. They are defined as

$$\boldsymbol{X}^{\text{R}} \equiv \varnothing,\quad \boldsymbol{Z}^{\text{R}} \equiv \underset{1\leqslant i\leqslant m}{\text{blockdiag}}\left(\boldsymbol{X}^*_i\ \boldsymbol{I}^{\text{A}} \odot \boldsymbol{Z}_{\text{grp},i}\ (\mathbf{1} - \boldsymbol{I}^{\text{A}}) \odot \boldsymbol{Z}_{\text{grp},i}\right),$$

$$\boldsymbol{\beta}^{\text{R}} \equiv \varnothing,\quad \boldsymbol{u}^{\text{R}} \equiv \left[\left(\boldsymbol{u}^T_{\text{R},1}\ \left(\boldsymbol{u}^{\text{A}}_{\text{grp},1}\right)^T\ \left(\boldsymbol{u}^{\text{B}}_{\text{grp},1}\right)^T\ \cdots\ \boldsymbol{u}^T_{\text{R},m}\ \left(\boldsymbol{u}^{\text{A}}_{\text{grp},m}\right)^T\ \left(\boldsymbol{u}^{\text{B}}_{\text{grp},m}\right)^T\right]^T.$$

The matrices $\boldsymbol{X}^{\text{G}}$ and $\boldsymbol{Z}^{\text{G}}$ are general design matrices corresponding to the fixed effects vector $\boldsymbol{\beta}^{\text{G}}$ and random spline coefficients vector $\boldsymbol{u}^{\text{G}}$ (superscript G), respectively. Typically, $\boldsymbol{X}^{\text{G}}$ contains polynomial functions of continuous predictors that are modeled as penalized splines, and $\boldsymbol{Z}^{\text{G}}$ would then contain random spline basis functions of the same predictors. They are defined as

$$\boldsymbol{\beta}^{\text{G}} \equiv \begin{bmatrix}\beta^{\text{A}}_0\\ \beta^{\text{A}}_1\\ \beta^{\text{BvsA}}_0\\ \beta^{\text{BvsA}}_1\end{bmatrix},\quad \boldsymbol{X}^{\text{G}} \equiv \begin{bmatrix}\boldsymbol{X}^*_i & \left(\mathbf{1} - \boldsymbol{I}^{\text{A}}_i\right) \odot \boldsymbol{X}^*_i\\ \vdots & \vdots\\ \boldsymbol{X}^*_m & \left(\mathbf{1} - \boldsymbol{I}^{\text{A}}_m\right) \odot \boldsymbol{X}^*_m\end{bmatrix},$$

$$\boldsymbol{Z}^{\text{G}} \equiv \begin{bmatrix}\boldsymbol{I}^{\text{A}}_i \odot \boldsymbol{Z}_{\text{gbl},i} & \left(\mathbf{1} - \boldsymbol{I}^{\text{A}}_i\right) \odot \boldsymbol{Z}_{\text{gbl},i}\\ \vdots & \vdots\\ \boldsymbol{I}^{\text{A}}_m \odot \boldsymbol{Z}_{\text{gbl},m} & \left(\mathbf{1} - \boldsymbol{I}^{\text{A}}_m\right) \odot \boldsymbol{Z}_{\text{gbl},m}\end{bmatrix},$$

$$\text{and}\quad \boldsymbol{u}^{\text{G}} \equiv \left[u^{\text{A}}_{\text{gbl},1}\ \cdots\ u^{\text{A}}_{\text{gbl},K_{\text{gbl}}}\ u^{\text{B}}_{\text{gbl},1}\ \cdots\ u^{\text{B}}_{\text{gbl},K_{\text{gbl}}}\right]^T,$$

where $\odot$ denotes the element-wise product of matrices.

Building on Lee and Wand [21], we derive a streamlined iterative scheme for obtaining optimal moments for all model parameters of the group-specific curve model with a factor-by-curve interaction. Our presentation of the streamlined MFVB algorithm benefits from the following notation, where $\boldsymbol{y}$, $\boldsymbol{X}^{\text{G}}$, $\boldsymbol{Z}^{\text{R}}$, and $\boldsymbol{Z}^{\text{G}}$ are partitioned row-wise corresponding to the i-th group:

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

Here, $\boldsymbol{y}_i \equiv [y_{i1}\cdots y_{in_i}]^T$ denotes the $n_i \times 1$ vector of responses for the i-th group. The matrices $\boldsymbol{X}^{\text{G}}_i$, $\boldsymbol{Z}^{\text{R}}_i$, and $\boldsymbol{Z}^{\text{G}}_i$ are defined in the same fashion. In addition, it is useful to define

$$\boldsymbol{C}^{\text{G}} \equiv [\boldsymbol{X}\ \boldsymbol{Z}^{\text{G}}]\quad\text{and}\quad \boldsymbol{C}^{\text{R}} \equiv [\boldsymbol{X}^{\text{R}}\ \boldsymbol{Z}^{\text{G}}].$$

Copyright © 2015 John Wiley & Sons, Ltd.

A.1. *Streamlined mean field variational Bayes algorithm*

We are now ready to present the streamlined algorithm for fast MFVB approximate fitting and inference for model (13). The R scripts for this algorithm are provided as an online supplementary material.

Initialize: $\nu = 4$, $\sigma_\beta^2 = 10^5$, $\mu_{q(1/(\sigma^{\rm A}_{\rm gbl})^2)} > 0$, $\mu_{q(1/(\sigma^{\rm B}_{\rm gbl})^2)} > 0$, $\mu_{q(1/a^{\rm A}_{\rm gbl})} > 0$, $\mu_{q(1/a^{\rm B}_{\rm gbl})} > 0$, $\mu_{q(1/\sigma^2_{\rm grp})} > 0$, $\mu_{q(1/a_{\rm grp})} > 0$, $\mu_{q(1/a_{{\rm R},r})} > 0$, $\boldsymbol{M}_{q(\boldsymbol{\Sigma}_{\rm R}^{-1})}$, a 4×4 positive definite matrix and $\boldsymbol{\xi}$, a $(\sum_{i=1}^m n_i) \times 1$ vector of positive entries.

Cycle through updates:

$$\textbf{Define: } \boldsymbol{M}_{q(\boldsymbol{\Omega})} \equiv \begin{bmatrix} \sigma_\beta^{-2}\boldsymbol{I}_P & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \mu_{q(1/(\sigma^{\rm A}_{\rm gbl})^2)}\boldsymbol{I}_{K_{\rm gbl}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \mu_{q(1/(\sigma^{\rm B}_{\rm gbl})^2)}\boldsymbol{I}_{K_{\rm gbl}} \end{bmatrix}$$

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

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

$\boldsymbol{G}_i \leftarrow 2(\boldsymbol{C}_i^{\rm G})^T \text{diag}\{\lambda(\boldsymbol{\xi}_i)\}\boldsymbol{C}_i^{\rm R}$

$$\boldsymbol{H}_i \leftarrow \left\{2(\boldsymbol{C}_i^{\rm R})^T \text{diag}\{\lambda(\boldsymbol{\xi}_i)\}\boldsymbol{C}_i^{\rm R} + \begin{bmatrix} \boldsymbol{M}_{q(\boldsymbol{\Sigma}_{\rm R}^{-1})} & \boldsymbol{0} \\ \boldsymbol{0} & \mu_{q(1/\sigma^2_{\rm grp})}\boldsymbol{I}_{2K_{\rm grp}} \end{bmatrix}\right\}^{-1}$$

$\boldsymbol{S} \leftarrow \boldsymbol{S} + \boldsymbol{G}_i\boldsymbol{H}_i\boldsymbol{G}_i^T$; $\boldsymbol{s} \leftarrow \boldsymbol{s} + \boldsymbol{G}_i\boldsymbol{H}_i(\boldsymbol{C}_i^{\rm R})^T(\boldsymbol{y}_i - \frac{1}{2}\boldsymbol{1})$

Update Multivariate Normal $q^*(\boldsymbol{\beta}, \boldsymbol{u}^{\rm A}_{\rm gbl}, \boldsymbol{u}^{\rm B}_{\rm gbl})$ parameters:

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\rm A}_{\rm gbl},\boldsymbol{u}^{\rm B}_{\rm gbl})} \leftarrow \left\{2(\boldsymbol{C}^{\rm G})^T \text{diag}\{\lambda(\boldsymbol{\xi}_i)\}\boldsymbol{C}^{\rm G} + \boldsymbol{M}_{q(\boldsymbol{\Omega})} - \boldsymbol{S}\right\}^{-1}$$

$$\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\rm A}_{\rm gbl},\boldsymbol{u}^{\rm B}_{\rm gbl})} \leftarrow \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\rm A}_{\rm gbl},\boldsymbol{u}^{\rm B}_{\rm gbl})}\{(\boldsymbol{C}^{\rm G})^T(\boldsymbol{y} - \tfrac{1}{2}\boldsymbol{1}) - \boldsymbol{s}\}$$

Update Multivariate Normal $q^*(\boldsymbol{u}_{{\rm R},i}, \boldsymbol{u}^{\rm A}_{{\rm grp},i}, \boldsymbol{u}^{\rm B}_{{\rm grp},i})$ parameters:

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

$$\boldsymbol{\Sigma}_{q(\boldsymbol{u}_{{\rm R},i},\boldsymbol{u}^{\rm A}_{{\rm grp},i},\boldsymbol{u}^{\rm B}_{{\rm grp},i})} \leftarrow \boldsymbol{H}_i + \boldsymbol{H}_i\,\boldsymbol{G}_i^T\,\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\rm A}_{\rm gbl},\boldsymbol{u}^{\rm B}_{\rm gbl})}\,\boldsymbol{G}_i\,\boldsymbol{H}_i$$

$$\boldsymbol{\mu}_{q(\boldsymbol{u}_{{\rm R},i},\boldsymbol{u}^{\rm A}_{{\rm grp},i},\boldsymbol{u}^{\rm B}_{{\rm grp},i})} \leftarrow \boldsymbol{H}_i\{(\boldsymbol{X}_i^{\rm R})^T\boldsymbol{y}_i - \boldsymbol{G}_i^T\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\rm A}_{\rm gbl},\boldsymbol{u}^{\rm B}_{\rm gbl})}\}$$

Update $\boldsymbol{\xi}$ parameters:

$$\boldsymbol{\xi}^2 \leftarrow \text{diagonal}\{\boldsymbol{C}^{\rm G}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\rm A}_{\rm gbl},\boldsymbol{u}^{\rm B}_{\rm gbl})} + \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\rm A}_{\rm gbl},\boldsymbol{u}^{\rm B}_{\rm gbl})}\boldsymbol{\mu}^T_{q(\boldsymbol{\beta},\boldsymbol{u}^{\rm A}_{\rm gbl},\boldsymbol{u}^{\rm B}_{\rm gbl})})(\boldsymbol{C}^{\rm G})^T\}$$

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

$$\boldsymbol{\xi}_i^2 \leftarrow \boldsymbol{\xi}_i^2 + 2\,\text{diagonal}\{\boldsymbol{C}^{\rm G}(-\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\rm A}_{\rm gbl},\boldsymbol{u}^{\rm B}_{\rm gbl})}\boldsymbol{G}_i\boldsymbol{H}_i + \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\rm A}_{\rm gbl},\boldsymbol{u}^{\rm B}_{\rm gbl})}\boldsymbol{\mu}^T_{q(\boldsymbol{u}_{{\rm R},i},\boldsymbol{u}^{\rm A}_{{\rm grp},i},\boldsymbol{u}^{\rm B}_{{\rm grp},i})})(\boldsymbol{C}_i^{\rm R})^T\}$$

$$\boldsymbol{\xi}_i^2 \leftarrow \boldsymbol{\xi}_i^2 + \text{diagonal}\{(\boldsymbol{C}_i^{\rm R}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_{{\rm R},i},\boldsymbol{u}^{\rm A}_{{\rm grp},i},\boldsymbol{u}^{\rm B}_{{\rm grp},i})} + \boldsymbol{\mu}_{q(\boldsymbol{u}_{{\rm R},i},\boldsymbol{u}^{\rm A}_{{\rm grp},i},\boldsymbol{u}^{\rm B}_{{\rm grp},i})}\boldsymbol{\mu}^T_{q(\boldsymbol{u}_{{\rm R},i},\boldsymbol{u}^{\rm A}_{{\rm grp},i},\boldsymbol{u}^{\rm B}_{{\rm grp},i})})(\boldsymbol{C}_i^{\rm R})^T\}$$

Update Inverse Gamma $q^*(\sigma^{\rm A}_{\rm gbl})$ and Inverse Gamma $q^*(a^{\rm A}_{\rm gbl})$ parameters:

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

$$\mu_{q(1/(\sigma^{\rm A}_{\rm gbl})^2)} \leftarrow \tfrac{1}{2}(K_{\rm gbl} + 1)/B_{q((\sigma^{\rm A}_{\rm gbl})^2)} \quad ; \quad \mu_{q(1/a^{\rm A}_{\rm gbl})} \leftarrow 1/\{\mu_{q(1/(\sigma^{\rm A}_{\rm gbl})^2)} + A_{\rm gbl}^{-2}\}$$

Update Inverse Gamma $q^*(\sigma^{\rm B}_{\rm gbl})$ and Inverse Gamma $q^*(a^{\rm B}_{\rm gbl})$ parameters:

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

$$\mu_{q(1/(\sigma^{\rm B}_{\rm gbl})^2)} \leftarrow \tfrac{1}{2}(K_{\rm gbl} + 1)/B_{q((\sigma^{\rm B}_{\rm gbl})^2)} \quad ; \quad \mu_{q(1/a^{\rm B}_{\rm gbl})} \leftarrow 1/\{\mu_{q(1/(\sigma^{\rm B}_{\rm gbl})^2)} + A_{\rm gbl}^{-2}\}$$

Update Inverse Gamma $q^*(\sigma_{\rm grp})$ and Inverse Gamma $q^*(a_{\rm grp})$ parameters:

$$B_{q(\sigma^2_{\rm grp})} \leftarrow \mu_{q(1/a_{\rm grp})} + \tfrac{1}{2}\sum_{i=1}^m\{\|\boldsymbol{\mu}_{q(\boldsymbol{u}^{\rm A}_{{\rm grp},i},\boldsymbol{u}^{\rm B}_{{\rm grp},i})}\|^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}^{\rm A}_{{\rm grp},i},\boldsymbol{u}^{\rm B}_{{\rm grp},i})})\}$$

$$\mu_{q(1/\sigma^2_{\rm grp})} \leftarrow \tfrac{1}{2}(m\,K_{\rm grp} + 1)/B_{q(\sigma^2_{\rm grp})} \quad ; \quad \mu_{q(1/a_{\rm grp})} \leftarrow 1/\{\mu_{q(1/\sigma^2_{\rm grp})} + A_{\rm grp}^{-2}\}$$

Copyright © 2015 John Wiley & Sons, Ltd.

Update Inverse Gamma $q^*(a_{\text{R},r})$ and Inverse Gmma $q^*(\boldsymbol{\Sigma}_{\text{R}})$ parameters:

For $r = 1, \ldots, 4$:

$$B_{q(a_{\text{R},r})} \leftarrow \nu \left(M_{q(\boldsymbol{\Sigma}_{\text{R}}^{-1})}\right)_{rr} + A_{\text{R}}^{-2} \;;\; \mu_{q(1/a_{\text{R},r})} \leftarrow \tfrac{1}{2}(\nu+3)/B_{q(a_{\text{R},r})}$$

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

$$\boldsymbol{B}_{q(\boldsymbol{\Sigma}_{\text{R}})} \leftarrow \textstyle\sum_{i=1}^{m}(\boldsymbol{\mu}_{q(\boldsymbol{u}_{\text{R},i})}\boldsymbol{\mu}^T_{q(\boldsymbol{u}_{\text{R},i})} + \boldsymbol{\Sigma}_{q(\boldsymbol{u}_{\text{R},i})}) + 2\,\nu\,\text{diag}\left(\mu_{q(1/a_{\text{R},1})}, \ldots, \mu_{q(1/a_{\text{R},4})}\right)$$

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

Update the q-density covariance matrix for the ith group:

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

$$\boldsymbol{\Lambda}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\text{A}}_{\text{gbl}},\boldsymbol{u}^{\text{B}}_{\text{gbl}},\boldsymbol{u}_{\text{R},i},\boldsymbol{u}_{\text{grp},i})} \equiv E_q\left[\left(\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}^{\text{A}}_{\text{gbl}}\\ \boldsymbol{u}^{\text{B}}_{\text{gbl}}\end{bmatrix} - \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\text{A}}_{\text{gbl}},\boldsymbol{u}^{\text{B}}_{\text{gbl}})}\right)\left(\begin{bmatrix}\boldsymbol{u}_{\text{R},i}\\ \boldsymbol{u}^{\text{A}}_{\text{grp},i}\\ \boldsymbol{u}^{\text{B}}_{\text{grp},i}\end{bmatrix} - \boldsymbol{\mu}_{q(\boldsymbol{u}_{\text{R},i},\boldsymbol{u}^{\text{A}}_{\text{grp},i},\boldsymbol{u}^{\text{B}}_{\text{grp},i})}\right)^T\right]$$

$$\leftarrow -\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}^{\text{A}}_{\text{gbl}},\boldsymbol{u}^{\text{B}}_{\text{gbl}})}\boldsymbol{G}_i\,\boldsymbol{H}_i$$

Algorithm 3: *Streamlined mean field variational Bayes algorithm for the group-specific curve model (13).*

Appendix B. Fitting the group-specific curve model (13) in `Rstan`

Stan is a probabilistic programming language, written in C++ for implementing full Bayesian statistical inference through Hamiltonian Monte Carlo sampling and No U-Turn sampling [58], a form of MCMC sampling. We now provide the `Rstan` code for fitting of model (13).

Specify the data: the number of observation, $\sum_{i=1}^{m} n_i$; the number of hospitals, m; the hospital identification number, `idnum`; the response vector, $\boldsymbol{y}$; the respective fixed and random effects design matrices, $\boldsymbol{X}$ and $\boldsymbol{Z}$; and the hyperparameters, σ^2_β, A_{R}, A_{gbl}, and A_{grp}. Data are labeled as integer or real and can be vectors if dimensions are specified. Data can also be constrained; for example, all hyperparameters must be positive.

```
grpSpecCurveModel <-
'data
{
   int<lower=1> numObs;              int<lower=1> m;
   int<lower=1> idnum[numObs];       int<lower=1> ncXR;
   int<lower=1> ncZ;                 int<lower=1> ncZgrp;
   real<lower=0> sigmaBeta;          real<lower=0> AR;
   real<lower=0> AuGbl;              real<lower=0> AuGrp;
   matrix[numObs,ncXR] Xbase;        matrix[numObs,ncXR] XTypA;
   matrix[numObs,ncXR] XTypB;        int<lower=0,upper=1> y[numObs];
   real x[numObs];                   matrix[numObs,ncZ] ZTypA;
   matrix[numObs,ncZ] ZTypB;         real ZgrpTypA[numObs,ncZgrp];
   real ZgrpTypB[numObs,ncZgrp];     vector[2*ncXR] zeroVec;
}
```

Specify the model parameters: the unknown to be estimated in the model fit. These are the fixed effects vector, $\boldsymbol{\beta}$; the random effects vectors, $\boldsymbol{u}_{\text{gbl}}$, $\boldsymbol{u}_{\text{grp}}$, and $\boldsymbol{u}_{\text{R}}$; and the covariance vectors and matrix, $\boldsymbol{\sigma}^2_{\text{gbl}}$, $\boldsymbol{\sigma}^2_{\text{grp}}$, and $\boldsymbol{\Sigma}_{\text{R}}$. In addition, we parametrize $\mu \equiv \boldsymbol{X\beta} + \boldsymbol{Zu}$ to be a transformation parameter in order to ensure the sampler runs more efficiently.

```
parameters
{
   vector[ncXR] beta;                vector[ncXR] betaTypB;
   vector[ncZ] uGblTypA;             vector[ncZ] uGblTypB;
```


Copyright © 2015 John Wiley & Sons, Ltd.

```
   vector[2*ncXR] uR[m];              real uGrpTypA[m,ncZgrp];
   real uGrpTypB[m,ncZgrp];           cov_matrix[2*ncXR] SigmaR;
   vector[2*ncXR] aR;                 real<lower=0> sigmauGblTypA;
   real<lower=0> sigmauGblTypB;       real<lower=0> sigmauGrp;
}
transformed parameters
{
   vector[numObs] fmean;              vector[numObs] fullMean;
   fmean <- (Xbase*beta + XTypB*betaTypB
             + ZTypA*uGblTypA + ZTypB*uGblTypB);
   for (iAll in 1:numObs)
      fullMean[iAll] <- (fmean[iAll]
             + uR[idnum[iAll],3]*XTypA[iAll,1]
             + uR[idnum[iAll],4]*XTypA[iAll,2]
             + uR[idnum[iAll],1]*XTypB[iAll,1]
             + uR[idnum[iAll],2]*XTypB[iAll,2]
             + dot_product(uGrpTypA[idnum[iAll]],ZgrpTypA[iAll])
             + dot_product(uGrpTypB[idnum[iAll]],ZgrpTypB[iAll]));
}
```

Specify the model statement: The `Bernoulli` specifies that the response vector $\boldsymbol{y}$ has a Bernoulli distribution with mean $\text{logit}^{-1}(\mu)$, where the mean is specified to be the sum of variables for the overall mean, maternal age deviation from that overall mean, and hospital deviations.

```
model
{
   matrix[2*ncXR,2*ncXR] rateForWish;

   y ~ bernoulli_logit(fullMean);

   for (i in 1:m)
      uR[i] ~ multi_normal(zeroVec,SigmaR);

   uGblTypA ~ normal(0,sigmauGblTypA);
   uGblTypB ~ normal(0,sigmauGblTypB);

   for (i in 1:m)
   {
      for (k in 1:ncZgrp)
      {
         uGrpTypA[i,k] ~ normal(0,sigmauGrp);
         uGrpTypB[i,k] ~ normal(0,sigmauGrp);
      }
   }

   rateForWish <- rep_matrix(0,4,4);
   for (r in 1:(2*ncXR))
   {
      aR[r] ~ inv_gamma(0.5,pow(AR,-2));
      rateForWish[r,r] <- 4/aR[r];
   }
   SigmaR ~ inv_wishart(5,rateForWish);

   beta ~ normal(0,sigmaBeta);
   betaTypB ~ normal(0,sigmaBeta);
   sigmauGblTypA ~ cauchy(0,AuGbl);
   sigmauGblTypB ~ cauchy(0,AuGbl);
```


Copyright © 2015 John Wiley & Sons, Ltd.

```
   sigmauGrp ~ cauchy(0,AuGrp);
}'
```

Acknowledgements

We are grateful to Chris Oates, Craig Anderson, David Rhode, Joanna Wang, and Stephen Wright for their comments on this research. This research was partially supported by an Australian Postgraduate Award, a University of Technology Sydney Chancellors Research Award, and Australian Research Council Discovery Project DP110100061.

References

1. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 1986; **42**:121–130.
2. Diggle P, Heagerty P, Liang KY, Zeger S. *Analysis of Longitudinal Data* 2nd ed. Oxford University Press: Oxford, 2002.
3. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982; **38**:963–974.
4. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* 1993; **88**:9–25.
5. Gelman A, Hill J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press: New York, 2007.
6. Goldstein H. *Multilevel Statistical Models* 4th ed. Wiley: Chichester, 2010.
7. Hastie TJ, Tibshirani RJ. *Generalized Additive Models*. CRC Press: Florida, 1990.
8. Zeger SL, Diggle PJ. Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics* 1994; **50**:689–699.
9. Verbyla AP, Cullis BR, Kenward MG, Welham SJ. The analysis of designed experiments and longitudinal data by using smoothing splines. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 1999; **48**:269–311.
10. Zhang D, Lin X, Raz J, Sowers M. Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association* 1998; **93**:710–719.
11. Brumback BA, Rice JA. Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association* 1998; **93**:961–976.
12. Rice JA, Wu CO. Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* 2001; **57**:253–259.
13. Durban M, Harezlak J, Wand MP, Carroll RJ. Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine* 2005; **24**:1153–1167.
14. Chen H, Wang Y. A penalized spline approach to functional mixed effects model analysis. *Biometrics* 2011; **67**:861–870.
15. Ryu D, Li E, Mallick BK. Bayesian nonparametric regression analysis of data with random effects covariates from longitudinal measurements. *Biometrics* 2011; **67**:454–466.
16. Bishop CM. *Pattern Recognition and Machine Learning*. Springer: New York, 2006.
17. Ormerod JT, Wand MP. Explaining variational approximations. *The American Statistician* 2010; **64**:140–153.
18. Luts J, Broderick T, Wand MP. Real-time semiparametric regression. *Journal of Computational and Graphical Statistics* 2014; **23**:589–615.
19. Armagan A, Dunson D. Sparse variational analysis of linear mixed models for large data sets. *Statistics & Probability Letters* 2011; **81**:1056–1062.
20. Tan SLL, Nott DJ. Variational inference for generalized linear mixed models using partially noncentered parametrizations. *Statistic Science* 2013; **28**:167–188.
21. Lee CYY, Wand MP. Streamlined mean field variational Bayes for longitudinal and multilevel data analysis. *Under review by the Biometrical Journal* Submitted.
22. Stewart B. Latent factor regressions for the social sciences, 2015. http://scholar.harvard.edu/files/bstewart/files/tensorreg.pdf [Accessed on 1 May 2015].
23. Spiegelhalter DJ, Thomas A, Best NG, Gilks WR, Lunn D. BUGS: Bayesian Inference Using Gibbs Sampling. Medical Research Council Biostatistics Unit: Cambridge, 2003. http://www.mrc-bsu.cam.ac.uk/bugs.
24. Stan Development Team. *Stan: A C++ Library for Probability and Sampling*, Version 2.6, 2015. http://mc-stan.org.
25. Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK. An introduction to variational methods for graphical models. *Machine learning* 1999; **37**:183–233.
26. Wainwright MJ, Jordan MI. Graphical models, exponential families, and variational inference. *ıFoundations and Trends in Machine Learning* 2008; **1**:1–305.
27. Titterington DM. Bayesian methods for neural networks and related models. *Statistical Science* 2004; **19**:128–139.
28. Bishop CM. *Pattern Recognition and Machine Learning*. Springer: New York, 2006.
29. McGrory CA, Titterington DM. Variational approximations in Bayesian model selection for finite mixture distributions. *Computational Statistics & Data Analysis* 2007; **51**:5352–5367.
30. Wand MP, Ormerod JT, Padoan SA, Frühwirth R. Mean field variational Bayes for elaborate distributions. *Bayesian Analysis* 2011; **6**:847–900.
31. Wand MP, Ormerod JT. Penalized wavelets: embedding wavelets into semiparametric regression. *Electronic Journal of Statistics* 2011; **5**:1654–1717.
32. Wang B, Titterington DM. Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. *Bayesian Analysis* 2006; **1**:625–650.

Copyright © 2015 John Wiley & Sons, Ltd.

33. You C, Ormerod JT, Müller S. On variational Bayes estimation and variational information criteria for linear regression models. *Australian & New Zealand Journal of Statistics* 2014; **56**:73–87.
34. Organisation for Economic Co-operation and Development. Caesarean sections. In *Health at a Glance 2011: OECD Indicators*. OECD Publishing: Paris, 2011.
35. Baicker K, Buckles KS, Chandra A. Geographic variation in the appropriate use of cesarean delivery. *Health Affairs* 2006; **25**:355–367.
36. Bragg F, Cromwell DA, Edozien LC, Gurol-Urganci I, Mahmood TA, Templeton A, van der Meulen JH. Variation in rates of caesarean section among English NHS trusts after accounting for maternal and clinical risk: cross sectional study. *BMJ* 2010; **c5065**:341.
37. Lee CYY, Homer C, Bisits A, Ryan L. *Increasing variation in hospital caesarean section rates among low-risk nulliparous women in Australia, from 1994 to 2010*. Under review by Birth.
38. Gelman A, Jakulin A, Pittau MG, Su YS. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics* 2008; **2**:1360–1383.
39. Huang A, Wand MP. Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis* 2013; **2**:439–452.
40. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982; **38**:963–974.
41. Zhao Y, Staudenmayer J, Coull BA, Wand MP. General design Bayesian generalized linear mixed models. *Statistical Science* 2006; **21**:35–51.
42. Jaakkola T, Jordan MI. A variational approach to Bayesian logistic regression models and their extensions. In *Sixth International Workshop on Artificial Intelligence and Statistics*, Florida, 1997.
43. Albert JH, Chib S. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* 1993; **88**:669–679.
44. Girolami M, Rogers S. Variational Bayesian multinomial probit regression. *Neural Computation* 2006; **18**:1790–1817.
45. Ormerod JT, Wand MP. Gaussian variational approximate inference for generalized linear mixed models. *Journal of Computational and Graphical Statistics* 2012; **21**:2–17.
46. Wand MP, Ormerod JT. On semiparametric regression with O'Sullivan penalized splines. *Australian and New Zealand Journal of Statistics* 2008; **50**:179–198.
47. Li Y, Ruppert D. On the asymptotics of penalized splines. *Biometrika* 2008; **95**:415–436.
48. Ruppert D, Wand MP, Carroll RJ. *Semiparametric Regression*. Cambridge University Press: New York, 2003.
49. Coull BA, Ruppert D, Wand MP. Simple incorporation of interactions into additive models. *Biometrics* 2001; **57**: 539–545.
50. Coull BA, Catalano PJ, Godleski JJ. Semiparametric analyses of cross-over data with repeated measures. *Journal of Agricultural, Biological, and Environmental Statistics* 2000; **5**:417–429.
51. Maringwa JT, Geys H, Shkedy Z, Faes C, Molenberghs G, Aerts M, Bijnens L. Application of semiparametric mixed models and simultaneous confidence bands in a cardiovascular safety experiment with longitudinal data. *Journal of Biopharmaceutical Statistics* 2008; **18**:1043–1062.
52. Smith A DAC, Wand MP. Streamlined variance calculations for semiparametric mixed models. *Statistics in Medicine* 2008; **29**:435–448.
53. R Development Core Team. *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing: Vienna, Austria, 2015. http://www.R-project.org/.
54. Faes C, Ormerod J, Wand MP. Variational Bayesian inference for parametric and non-parametric regression with missing predictor data. *Journal of the American Statistical Association* 2011; **106**:959–971.
55. Wand MP, Ripley BD. KernSmooth 2.23. Functions for kernel smoothing corresponding to the book: Wand, M. P. & Jones, M. C. (1995) "Kernel Smoothing", 2015. R package. http://cran.r-project.org.
56. Wainwright MJ, Jordan MI. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning* 2008; **1**:1–305.
57. Neville SE, Ormerod JT, Wand MP. Mean field variational Bayes for continuous sparse signal shrinkage: pitfalls and remedies. *Electronic Journal of Statistics* 2014; **8**:1113–1151.
58. Homan MD, Gelman A. The no-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *The Journal of Machine Learning Research* 2014; **1**:1593–1623.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.

Copyright © 2015 John Wiley & Sons, Ltd.