Statistical Science
2013, Vol. 28, No. 2, 168–188
DOI: 10.1214/13-STS418

Variational Inference for Generalized Linear Mixed Models Using Partially Noncentered Parametrizations

Linda S. L. Tan and David J. Nott

Abstract. The effects of different parametrizations on the convergence of Bayesian computational algorithms for hierarchical models are well explored. Techniques such as centering, noncentering and partial noncentering can be used to accelerate convergence in MCMC and EM algorithms but are still not well studied for variational Bayes (VB) methods. As a fast deterministic approach to posterior approximation, VB is attracting increasing interest due to its suitability for large high-dimensional data. Use of different parametrizations for VB has not only computational but also statistical implications, as different parametrizations are associated with different factorized posterior approximations. We examine the use of partially noncentered parametrizations in VB for generalized linear mixed models (GLMMs). Our paper makes four contributions. First, we show how to implement an algorithm called nonconjugate variational message passing for GLMMs. Second, we show that the partially noncentered parametrization can adapt to the quantity of information in the data and determine a parametrization close to optimal. Third, we show that partial noncentering can accelerate convergence and produce more accurate posterior approximations than centering or noncentering. Finally, we demonstrate how the variational lower bound, produced as part of the computation, can be useful for model selection.

Key words and phrases: Variational Bayes, hierarchical centering, variational message passing, nonconjugate models, longitudinal data analysis.

1. INTRODUCTION

The convergence of Markov chain Monte Carlo (MCMC) algorithms depends greatly on the choice of parametrization and simple reparametrizations can often give improved convergence. Here we investigate the use of centered, noncentered and partially noncentered parametrizations of hierarchical models in the context of variational Bayes (VB) (Attias, 1999). As a fast deterministic approach to approximation of the posterior distribution in Bayesian inference, VB is attracting increasing interest due to its suitability for large high-dimensional data (see, e.g., Braun and McAuliffe, 2010; Hoffman et al., 2012). VB methods approximate the intractable posterior by a factorized distribution which can be represented by a directed graph and optimization of the factorized variational posterior can be decomposed into local computations that involve only neighboring nodes. Variational message passing (Winn and Bishop, 2005) is an algorithmic implementation of VB that can be applied to a general class of conjugate-exponential models (Attias, 2000; Ghahramani and Beal, 2001). Knowles and Minka (2011) proposed an algorithm called a nonconjugate variational message passing to extend variational message passing to nonconjugate models.

We examine the use of partially noncentered parametrization in VB for generalized linear mixed models (GLMMs). Our paper makes four contributions.

Linda S. L. Tan is a Ph.D. student and David J. Nott is Associate Professor, Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546, Singapore (e-mail: g0900760@nus.edu.sg; standj@nus.edu.sg).

First, we show how to implement nonconjugate variational message passing for GLMMs. Second, we show that the partially noncentered parametrization is able to adapt to the quantity of information in the data so that it is not necessary to make a choice in advance between centering and noncentering with the data deciding the optimal parametrization. Third, we show that in addition to accelerating convergence, partial noncentering is a good strategy statistically for VB in terms of producing more accurate approximations to the posterior than either centering or noncentering. Finally, we demonstrate how the variational lower bound, which is produced as part of the computation, can be useful for model selection.

GLMMs extend generalized linear models by the inclusion of random effects to account for correlation of observations in grouped data and are of wide applicability. Estimation of GLMMs using maximum likelihood is challenging, as the integral over random effects is intractable. Methods involving numerical quadrature or MCMC to approximate these integrals are computationally intensive. Various approximate methods such as penalized quasi-likelihood (Breslow and Clayton, 1993), Laplace approximation and its extension (Raudenbush, Yang and Yosef, 2000) and Gaussian variational approximation (Ormerod and Wand, 2012) have been developed. Fong, Rue and Wakefield (2010) considered a Bayesian approach using integrated nested Laplace approximations. We show how to fit GLMMs using nonconjugate variational message passing, focusing on Poisson and logistic mixed models and their applications in longitudinal data analysis.

The literature on parametrization of hierarchical models including partial noncentering techniques for accelerating MCMC algorithms is inspired by earlier similar work for the expectation maximization (EM) algorithm (see Meng and van Dyk, 1997, 1999; Liu and Wu, 1999). Gelfand, Sahu and Carlin (1995, 1996) proposed hierarchical centering for normal linear mixed models and GLMMs to improve the slow mixing in MCMC algorithms due to high correlations between model parameters. Papaspiliopoulos, Roberts and Sköld (2003, 2007) demonstrated that centering and noncentering play complementary roles in boosting MCMC efficiency and neither are uniformly effective. They considered the partially noncentered parametrization which is data dependent and lies on the continuum between the centered and noncentered parametrizations. Extending this idea, Christensen, Roberts and Sköld (2006) devised reparametrization techniques to improve performance for Hastings-within Gibbs algorithms for spatial GLMMs. Yu and Meng (2011) introduced a strategy for boosting MCMC efficiency via interweaving the centered and noncentered parametrizations to reduce dependence between draws. Parameter-expanded VB methods were proposed by Qi and Jaakkola (2006) to reduce coupling in updates and speed up VB.

The idea of partial noncentering is to introduce a tuning parameter via reparametrization of the model and then seek its optimal value for fastest convergence. For the normal hierarchical model, Papaspiliopoulos, Roberts and Sköld (2003) showed that the partially noncentered parametrization has convergence properties superior to that of the centered and noncentered parametrizations for the Gibbs sampler. As the rate of convergence of an algorithm based on VB is equal to that of the corresponding Gibbs sampler when the target distribution is Gaussian (Tan and Nott, 2013), partial noncentering will similarly outperform centering and noncentering in the context of VB for the normal hierarchical model. This provides motivation to consider partial noncentering in the VB context. We illustrate this idea with the following example.

1.1 Motivating Example: Linear Mixed Model

Consider the linear mixed model

$$y_i = X_i\beta + X_i u_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2 I), i = 1, \ldots, n, \tag{1}$$

where y_i is a vector of length n_i, β is a vector of length r of fixed effects, X_i is a $n_i \times r$ matrix of covariates and u_i is a vector of length r of random effects independently distributed as $N(0, D)$. For simplicity, we specify a constant prior on β and assume σ^2 and D are known. Let

$$\alpha_i = \beta + u_i \quad \text{and} \quad \tilde{\alpha}_i = \alpha_i - W_i\beta, \quad i = 1, \ldots, n,$$

where W_i is an $r \times r$ tuning matrix to be specified. $W_i = 0$ corresponds to the centered and $W_i = I$ to the noncentered parametrization. For each $i = 1, \ldots, n$,

$$y_i = X_i W_i \beta + X_i\tilde{\alpha}_i + \varepsilon_i$$

and

$$\tilde{\alpha}_i \sim N\big((I - W_i)\beta, D\big).$$

This is the partially noncentered parametrization and the set of unknown parameters is $\theta = \{\beta, \tilde{\alpha}\}$, where $\tilde{\alpha} = [\tilde{\alpha}_1^T, \ldots, \tilde{\alpha}_n^T]^T$. Let $y = [y_1, \ldots, y_n]^T$ denote the observed data. Of interest is the posterior distribution of θ, $p(\theta|y)$.

Initialize $\mu^q_{\tilde{\alpha}_i}$ and $\Sigma^q_{\tilde{\alpha}_i}$ for $i = 1, \ldots, n$.
Cycle:
$\Sigma^q_\beta \leftarrow [\sum_{i=1}^n \{(I - W_i)^T D^{-1}(I - W_i) + \frac{1}{\sigma^2} W_i^T X_i^T X_i W_i\}]^{-1}$
$\mu^q_\beta \leftarrow \Sigma^q_\beta \sum_{i=1}^n [\frac{1}{\sigma^2} W_i^T X_i^T y_i + \{D^{-1}(I - W_i) - \frac{1}{\sigma^2} X_i^T X_i W_i\}^T \mu^q_{\tilde{\alpha}_i}]$
For $i = 1, \ldots, n$,
$\Sigma^q_{\tilde{\alpha}_i} \leftarrow (D^{-1} + \frac{1}{\sigma^2} X_i^T X_i)^{-1}$
$\mu^q_{\tilde{\alpha}_i} \leftarrow \Sigma^q_{\tilde{\alpha}_i} [\frac{1}{\sigma^2} X_i^T y_i + \{D^{-1}(I - W_i) - \frac{1}{\sigma^2} X_i^T X_i W_i\} \mu^q_\beta]$
until convergence.

ALGORITHM 1. *Iterative scheme for obtaining variational parameters in linear mixed model.*

Suppose $p(\theta|y)$ is not analytically tractable. In the variational approach, we approximate $p(\theta|y)$ by a $q(\theta)$ for which inference is more tractable and $q(\theta)$ is chosen to minimize the Kullback–Leibler divergence between $q(\theta)$ and $p(\theta|y)$ given by

$$\int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} \, \mathrm{d}\theta = \int q(\theta) \log \frac{q(\theta)}{p(y, \theta)} \, \mathrm{d}\theta + \log p(y),$$

where $p(y)$ is the marginal likelihood $p(y) = \int p(y|\theta) p(\theta) \, \mathrm{d}\theta$. Since the Kullback–Leibler divergence is nonnegative,

$$\begin{aligned} \log p(y) &\geq \int \log \frac{p(y, \theta)}{q(\theta)} q(\theta) \, \mathrm{d}\theta \\ &= E_q\{\log p(y, \theta)\} - E_q\{\log q(\theta)\} \\ &= \mathcal{L}, \end{aligned} \tag{2}$$

where $\mathcal{L}$ is a lower bound on the log marginal likelihood. Maximization of $\mathcal{L}$ is equivalent to minimization of the Kullback–Leibler divergence between $q(\theta)$ and $p(\theta|y)$. In VB, $q(\theta)$ is assumed to be of a factorized form, say, $q(\theta) = \prod_{i=1}^m q_i(\theta_i)$ for some partition $\{\theta_1, \ldots, \theta_m\}$ of θ. Maximization of $\mathcal{L}$ over each of $q_1, \ldots, q_m$ lead to optimal densities satisfying $q_i(\theta_i) \propto \exp\{E_{-\theta_i} \log p(y, \theta)\}$, $i = 1, \ldots, m$, where $E_{-\theta_i}$ denotes expectation with respect to the density $\prod_{j \neq i} q_j(\theta_j)$. See Ormerod and Wand (2010) for an explanation of variational approximation methods very accessible to statisticians.

If we apply VB to (1) and approximate the posterior $p(\theta|y)$ with $q(\theta) = q(\beta) q(\tilde{\alpha})$, the optimal densities can be derived to be $q(\beta) = N(\mu^q_\beta, \Sigma^q_\beta)$ and $q(\tilde{\alpha}) = \prod_{i=1}^n q(\tilde{\alpha}_i)$, where $q(\tilde{\alpha}_i) = N(\mu^q_{\tilde{\alpha}_i}, \Sigma^q_{\tilde{\alpha}_i})$. The expressions for the variational parameters μ^q_β, Σ^q_β and $\mu^q_{\tilde{\alpha}_i}$, $\Sigma^q_{\tilde{\alpha}_i}$, $i = 1, \ldots, m$, are, however, dependent on each other and can be computed by an iterative scheme such as that given in Algorithm 1.

Observe that Algorithm 1 converges in one iteration if $D^{-1}(I - W_i) = \frac{1}{\sigma^2} X_i^T X_i W_i$ for each i, that is, if

$$W_i = \left(\frac{1}{\sigma^2} X_i^T X_i + D^{-1} \right)^{-1} D^{-1} \quad \text{for } i = 1, \ldots, n. \tag{3}$$

For this specification of the tuning parameters, partial noncentering gives more rapid convergence than centering or noncentering. Moreover, it can be shown that the true posteriors are recovered in this partially noncentered parametrization so that a better fit is achieved than in the centered or noncentered parametrizations. This example suggests that with careful tuning of W_i, $i = 1, \ldots, n$, the partially noncentered parametrization can potentially outperform the centered and noncentered parametrizations in the VB context.

The rest of the paper is organized as follows. Section 2 specifies the GLMM and priors used. Section 3 describes the partially noncentered parametrization for GLMMs. Section 4 describes the nonconjugate variational message passing algorithm for fitting GLMMs. Section 5 discusses briefly the use of the variational lower bound for model selection and Section 6 considers examples including real and simulated data. Section 7 concludes.

2. THE GENERALIZED LINEAR MIXED MODEL

Consider clustered data where y_{ij} denotes the jth response from cluster i, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$. Conditional on the r-dimensional random effects u_i drawn independently from $N(0, D)$, y_{ij} is independently distributed from some exponential family dis-

tribution with density

$$f(y_{ij}|u_i) = \exp\left\{\frac{y_{ij}\zeta_{ij} - b(\zeta_{ij})}{a(\phi)} + c(y_{ij}, \phi)\right\}, \tag{4}$$

where ζ_{ij} is the canonical parameter, ϕ is the dispersion parameter, and $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are functions specific to the family. The conditional mean of y_{ij}, $\mu_{ij} = E(y_{ij}|u_i)$, is assumed to depend on the fixed and random effects through the linear predictor,

$$\eta_{ij} = X_{ij}^{R^T}\beta^R + X_{ij}^{G^T}\beta^G + X_{ij}^{R^T}u_i$$

with $g(\mu_{ij}) = \eta_{ij}$ for some known link function, $g(\cdot)$. Here, X_{ij}^R and $X_{ij} = [X_{ij}^{R^T}, X_{ij}^{G^T}]^T$ are $r \times 1$ and $p \times 1$ vectors of covariates and $\beta = [\beta^{R^T}, \beta^{G^T}]^T$ is a $p \times 1$ vector of fixed effects. We considered the above breakdown (see Zhao et al., 2006) for the linear predictor to allow for centering. For the ith cluster, let $y_i = [y_{i1}, \ldots, y_{in_i}]^T$, $X_i^R = [X_{i1}^R, \ldots, X_{in_i}^R]^T$, $X_i^G = [X_{i1}^G, \ldots, X_{in_i}^G]^T$, $X_i = [X_{i1}, \ldots, X_{in_i}]^T$ and $\eta_i = [\eta_{i1}, \ldots, \eta_{in_i}]^T$. Let 1_{n_i} denote the $n_i \times 1$ column vector with all entries equal to 1. We assume that the first column of X_i^R is 1_{n_i} if X_i^R is not a zero matrix. For Bayesian inference, we specify prior distributions on the fixed effects β and random effects covariance matrix D. In this paper, we focus on responses from the Bernoulli and Poisson families and the dispersion parameter is one in these cases, so we do not consider a prior for ϕ. We assume a diffuse prior, $N(0, \Sigma_\beta)$, for β and an independent inverse Wishart prior, $IW(\nu, S)$, for D. Following the suggestion by Kass and Natarajan (2006), we set $\nu = r$ and let the scale matrix S be determined from first-stage data variability. In particular, $S = r\hat{R}$, where

$$\hat{R} = c\left(\frac{1}{n}\sum_{i=1}^{n} X_i^{R^T} M_i(\hat{\beta}) X_i^R\right)^{-1}, \tag{5}$$

$M_i(\hat{\beta})$ denotes the $n_i \times n_i$ diagonal generalized linear model weight matrix with diagonal elements $[\phi v(\hat{\mu}_{ij}) \cdot g'(\hat{\mu}_{ij})^2]^{-1}$, $v(\cdot)$ is the variance function based on $f(\cdot)$ in (4) and $g(\cdot)$ is the link function. Here, $\hat{\mu}_{ij} = g^{-1}(X_{ij}^T\hat{\beta} + X_{ij}^{R^T}\hat{u}_i)$, where $\hat{u}_i$ is set as 0 for all i and $\hat{\beta}$ is an estimate of the regression coefficients from the generalized linear model obtained by pooling all data and setting $u_i = 0$ for all i. The value of c is an inflation factor representing the amount by which within-cluster variability should be increased in determining $\hat{R}$. We used $c = 1$ in all examples.

3. A PARTIALLY NONCENTERED PARAMETRIZATION FOR THE GENERALIZED LINEAR MIXED MODEL

We introduce the following partially noncentered parametrization for the GLMM. For each $i = 1, \ldots, n$, the linear predictor is $\eta_i = X_i^R\beta^R + X_i^G\beta^G + X_i^R u_i$. Let

$$\begin{aligned} X_i^G\beta^G &= X_i^{G_1}\beta^{G_1} + X_i^{G_2}\beta^{G_2} \\ &= 1_{n_i} x_i^{G_1{}^T}\beta^{G_1} + X_i^{G_2}\beta^{G_2}, \end{aligned}$$

where β^{G_1} is a vector of length g_1 consisting of all parameters corresponding to subject specific covariates (i.e., the rows of $X_i^{G_1}$ are all the same and equal to the vector $x_i^{G_1}$ say). Recall that the first column of X_i^R is 1_{n_i} if X_i^R is not a zero matrix. We have

$$\eta_i = X_i^R(C_i\beta^{RG_1} + u_i) + X_i^{G_2}\beta^{G_2},$$

where

$$C_i = \begin{bmatrix} & x_i^{G_1{}^T} \\ I_r & \\ & 0 \end{bmatrix} \quad \text{and} \quad \beta^{RG_1} = \begin{bmatrix} \beta^R \\ \beta^{G_1} \end{bmatrix}.$$

Let $\alpha_i = C_i\beta^{RG_1} + u_i$ and $\tilde{\alpha}_i = \alpha_i - W_i C_i \beta^{RG_1}$, where W_i is an $r \times r$ matrix to be specified. The proportion of $C_i\beta^{RG_1}$ subtracted from each α_i is allowed to vary with i as in Papaspiliopoulos, Roberts and Sköld (2003) to reflect the varying informativity of each response y_i about the underlying α_i. $W_i = 0$ corresponds to the centered and $W_i = I$ to the noncentered parametrization. Finally,

$$\begin{aligned} \eta_i &= X_i^R(\tilde{\alpha}_i + W_i C_i \beta^{RG_1}) + X_i^{G_2}\beta^{G_2} \\ &= V_i\beta + X_i^R\tilde{\alpha}_i, \end{aligned} \tag{6}$$

where $V_i = [X_i^R W_i C_i \;\; X_i^{G_2}]$ and $\tilde{\alpha}_i \sim N((I - W_i) \cdot C_i\beta^{RG_1}, D)$. We refer to (6) as the partially noncentered parametrization. Let $\tilde{\alpha} = [\tilde{\alpha}_1^T, \ldots, \tilde{\alpha}_n^T]^T$ and $\theta = \{\beta, D, \tilde{\alpha}\}$ denote the set of unknown parameters in the GLMM. The joint distribution of $p(y, \theta)$ is

$$\begin{aligned} p(y, \theta) = &\left\{\prod_{i=1}^{n} p(y_i|\beta, \tilde{\alpha}_i) p(\tilde{\alpha}_i|\beta, D)\right\} \\ &\cdot p(\beta|\Sigma_\beta) p(D|\nu, S). \end{aligned} \tag{7}$$

Figure 1 shows the factor graph for $p(y, \theta)$ where there is a node (circle) for every variable, which is shaded in the case of observed variables, and a node (filled rectangle) for each factor in the joint distribution. Constants or hyperparameters are denoted with smaller

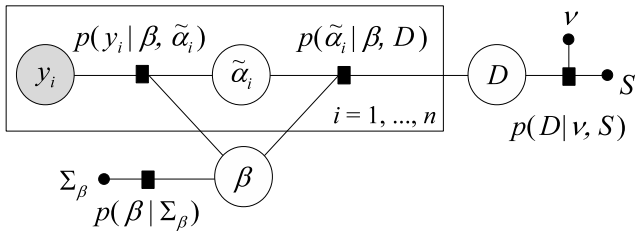

FIG. 1. *Factor graph for $p(y, \theta)$ in (7). Filled rectangles denote factors and circles denote variables (shaded for observed variables). Smaller filled circles denote constants or hyperparameters. The box represents a plate which contains variables and factors to be replicated. Number of repetitions is indicated in the lower right corner.*

filled circles. Each factor node is connected by undirected links to all of the variable nodes on which that factor depends (see Bishop, 2006). Next, we consider specification of the tuning parameter W_i, referring to the linear mixed model example in Section 1.1 which is a special case of the GLMM in (4) with an identity link.

3.1 Specification of Tuning Parameter

It is interesting to note that for the linear mixed model in (1), the expression for W_i leading to optimal performance in VB and the Gibbs sampling algorithm is exactly the same (see Papaspiliopoulos, Roberts and Sköld, 2003). Gelfand, Sahu and Carlin (1995) also observed the importance of W_i in assessing convergence properties of the centered parametrization. They showed that $|W_i| < 1$ for all i and $|W_i|$ is close to zero (centering is more efficient) when the generalized variance $|D|$ is large. On the other hand, $|W_i|$ is close to 1 (noncentering works better) when the error variance is large. Outside the Gaussian context, Papaspiliopoulos, Roberts and Sköld (2003) considered partial noncentering for the spatial GLMM and specified the tuning parameters by using a quadratic expansion of the log-likelihood to obtain an indication of the information present in y_i. Observe that W_i in (3) can be expressed as

$$W_i = \left(\mathcal{I}_f + D^{-1}\right)^{-1} D^{-1}, \tag{8}$$

if $\ell = \log p(y_i|\beta, \alpha_i)$ denotes the log-likelihood and $\mathcal{I}_f = -\frac{\partial^2 \ell}{\partial \alpha_i \, \partial \alpha_i^T}$. We use (8) to extend partially noncentered parametrizations to GLMMs and consider the specification of W_i for responses from the Bernoulli and Poisson families in particular.

Recall that the linear predictor η_i can be expressed as $X_i^R \alpha_i + X_i^{G_2} \beta^{G_2}$. For Poisson responses with the log link function, we allow for an offset $\log E_{ij}$ so that $\log \mu_{ij} = \log E_{ij} + \eta_{ij}$. Let $E_i = [E_{i1}, \ldots, E_{in_i}]^T$. We have

$$\begin{aligned} \ell &= y_i^T (\log E_i + \eta_i) - E_i^T \exp(\eta_i) \\ &\quad - 1_{n_i}^T \log(y_i!) \quad \text{and} \\ \mathcal{I}_f &= \sum_{j=1}^{n_i} E_{ij} \exp(\eta_{ij}) X_{ij}^R X_{ij}^{R^T} \approx \sum_{j=1}^{n_i} y_{ij} X_{ij}^R X_{ij}^{R^T}, \end{aligned} \tag{9}$$

if we approximate the conditional mean μ_{ij} with the response. For Bernoulli responses with the logit link function, we have

$$\begin{aligned} \ell &= y_i^T \eta_i - 1_{n_i}^T \log\{1_{n_i} + \exp(\eta_i)\} \quad \text{and} \\ \mathcal{I}_f &= \sum_{j=1}^{n_i} \frac{\exp(\eta_{ij})}{\{1 + \exp(\eta_{ij})\}^2} X_{ij}^R X_{ij}^{R^T}. \end{aligned} \tag{10}$$

The specification of W_i depends on the random effects covariance D and, for Bernoulli responses, on the linear predictor η_i as well. In Algorithm 3, we initialize W_i by considering $\eta_i = X_i \beta + X_i^R u_i$ and using estimates of D, β and u_i from penalized quasi-likelihood. Subsequently, we can either keep W_i as fixed or update them by replacing D with $\frac{S^q}{\nu^q - r - 1}$, assuming the variational posterior of D is $IW(\nu^q, S^q)$ and η_i with $V_i \mu_\beta^q + X_i^R \mu_{\tilde{\alpha}_i}^q$, where μ_β^q and $\mu_{\tilde{\alpha}_i}^q$ are the variational posterior means of β and $\tilde{\alpha}_i$, respectively. This can be done at the beginning of each iteration after new estimates of μ_β^q, $\mu_{\tilde{\alpha}_i}^q$, ν^q and S^q are obtained (see Algorithm 3 step 1).

4. VARIATIONAL INFERENCE FOR GLMMS

In this section we present the nonconjugate variational message passing algorithm recently developed in machine learning by Knowles and Minka (2011) for fitting GLMMs. Recall that in VB, the posterior distribution $p(\theta|y)$ is approximated by a $q(\theta)$ which is assumed to be of a factorized form, say, $q(\theta) = \prod_{i=1}^m q_i(\theta_i)$ for some partition $\{\theta_1, \ldots, \theta_m\}$ of θ. For conjugate-exponential models, the optimal densities q_i will have the same form as the prior so that it suffices to update the parameters of q_i, such as in Algorithm 1. Variational message passing (Winn and Bishop, 2005) is an algorithm which allows VB to be applied to conjugate-exponential models without having to derive application-specific updates. In the case of GLMMs where the responses are from the Bernoulli or Poisson families, the factor $p(y_i|\beta, \tilde{\alpha}_i)$ of $p(y, \theta)$ in (7) is nonconjugate with respect to the prior distributions over

β and $\tilde{\alpha}_i$ for each $i = 1, \ldots, n$. Therefore, if we apply VB and assume, say, $q(\theta) = q(\beta)q(D)\prod_{i=1}^{n} q(\tilde{\alpha}_i)$, the optimal densities for $q(\beta)$ and $q(\tilde{\alpha}_i)$ will not belong to recognizable density families.

4.1 Nonconjugate Variational Message Passing

In nonconjugate variational message passing, besides assuming that $q(\theta)$ must factorize into $\prod_{i=1}^{m} q_i(\theta_i)$ for some partition $\{\theta_1, \ldots, \theta_m\}$ of θ, we impose another restriction that each q_i must belong to some exponential family. In this way, we only have to find the parameters of each q_i that maximizes the lower bound $\mathcal{L}$. Suppose each q_i can be written in the form

$$q_i(\theta_i) = \exp\{\lambda_i^T t(\theta_i) - h(\lambda_i)\},$$

where λ_i is the vector of natural parameters and $t(\cdot)$ are the sufficient statistics. We wish to maximize $\mathcal{L}$ with respect to the variational parameters $\lambda_1, \ldots, \lambda_m$ which are also natural parameters of $q_1(\theta_1), \ldots, q_m(\theta_m)$, respectively. In the following, we show that nonconjugate variational message passing can be interpreted as a fixed-point iteration where updates are obtained from the condition that the gradient of $\mathcal{L}$ with respect to each λ_i is zero when $\mathcal{L}$ is maximized.

From (2), the gradient of $\mathcal{L}$ with respect to λ_i is

$$\frac{\partial \mathcal{L}}{\partial \lambda_i} = \frac{\partial}{\partial \lambda_i} E_q\{\log p(y, \theta)\} - \frac{\partial}{\partial \lambda_i} E_q\{\log q(\theta)\}. \tag{11}$$

Let us consider the first term in (11). Suppose $p(y, \theta) = \prod_a f_a(y, \theta)$. We have

$$E_q\{\log p(y, \theta)\} = \sum_a S_a,$$

where

$$S_a = E_q\{\log f_a(y, \theta)\}.$$

Note that each S_a is a function of the natural parameters $\lambda_1, \ldots, \lambda_m$. Since we have assumed that θ_i is independent of all θ_j where $j \neq i$ in the variational approximation q, the only terms in $\sum_a S_a$ which depend on λ_i are the factors f_a connected to θ_i in the factor graph of $p(y, \theta)$. Therefore,

$$\frac{\partial}{\partial \lambda_i} E_q\{\log p(y, \theta)\} = \sum_{a \in N(\theta_i)} \frac{\partial S_a}{\partial \lambda_i}, \tag{12}$$

where the summation is over all factors in $N(\theta_i)$, the neighborhood of θ_i in the factor graph. For the second term in (11), we have

$$E_q\{\log q(\theta)\} = \sum_{l=1}^{m} E_q\{\log q_l(\theta_l)\},$$

where the only term in the sum that depends on λ_i is the ith term. Hence,

$$\begin{aligned} \frac{\partial}{\partial \lambda_i} E_q\{\log q(\theta)\} &= \frac{\partial}{\partial \lambda_i}\left\{\lambda_i^T \frac{\partial h(\lambda_i)}{\partial \lambda_i} - h(\lambda_i)\right\} \\ &= \mathcal{V}(\lambda_i)\lambda_i, \end{aligned} \tag{13}$$

where we have used the fact that $E_q\{t(\theta_i)\} = \frac{\partial h(\lambda_i)}{\partial \lambda_i}$ and $\mathcal{V}(\lambda_i) = \frac{\partial^2 h(\lambda_i)}{\partial \lambda_i \partial \lambda_i^T}$ denotes the variance–covariance matrix of $t(\theta_i)$. Note that $\mathcal{V}(\lambda_i)$ is symmetric positive semi-definite. Putting (12) and (13) together, the gradient of the lower bound is

$$\frac{\partial \mathcal{L}}{\partial \lambda_i} = \sum_{a \in N(\theta_i)} \frac{\partial S_a}{\partial \lambda_i} - \mathcal{V}(\lambda_i)\lambda_i$$

and is zero when $\lambda_i = \mathcal{V}(\lambda_i)^{-1} \sum_{a \in N(\theta_i)} \frac{\partial S_a}{\partial \lambda_i}$, provided $\mathcal{V}(\lambda_i)$ is invertible. This condition is used to obtain updates to λ_i in nonconjugate variational message passing (Algorithm 2).

The updates can be simplified when the factor f_a is conjugate to $q_i(\theta_i)$, that is, f_a has the same functional form as $q_i(\theta_i)$ with respect to θ_i. Let $\theta_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_m)$. Suppose

$$f_a(y, \theta) = \exp\{g_a(y, \theta_{-i})^T t(\theta_i) - h_a(y, \theta_{-i})\}.$$

Then $\frac{\partial S_a}{\partial \lambda_i} = \mathcal{V}(\lambda_i) E_q\{g_a(y, \theta_{-i})\}$, where $E_q\{g_a(y, \theta_{-i})\}$ does not depend on λ_i. When every factor in the neighborhood of θ_i is conjugate to $q_i(\theta_i)$, the gradient of the lower bound can be simplified to $\mathcal{V}(\lambda_i)[\sum_{a \in N(\theta_i)} E_q\{g_a(y, \theta_{-i})\} - \lambda_i]$ and the updates in nonconjugate variational message passing reduce to

$$\lambda_i \leftarrow \sum_{a \in N(\theta_i)} E_q\{g_a(y, \theta_{-i})\}. \tag{14}$$

These are precisely the updates in variational message passing. Nonconjugate variational message passing thus reduces to variational message passing for conjugate factors (see also Knowles and Minka, 2011). Unlike variational message passing, however, the

Initialize λ_i for $i = 1, \ldots, m$.
Cycle:
For $i = 1, \ldots, m$,
$\lambda_i \leftarrow \mathcal{V}(\lambda_i)^{-1} \sum_{a \in N(\theta_i)} \frac{\partial S_a}{\partial \lambda_i}$
until convergence.

ALGORITHM 2. *Nonconjugate variational message passing.*

Kullback–Leibler divergence is not guaranteed to decrease at each step and sometimes convergence problems may be encountered. Knowles and Minka (2011) suggested using damping to fix convergence problems. We did not encounter any convergence problems in the examples considered in this paper.

4.2 Updates for Multivariate Gaussian Variational Distribution

Suppose q_i is Gaussian. While the updates in Algorithm 2 are expressed in terms of the natural parameters λ_i, it might be more convenient to express $\frac{\partial S_a}{\partial \lambda_i}$ in terms of the mean and covariance of q_i. Knowles and Minka (2011) have considered the univariate case and Wand (2013) derived fully simplified updates for the multivariate case. However, as Wand (2013) is in preparation, we give enough details of the update so that the derivation can be understood. Magnus and Neudecker (1988) is a good reference for the matrix differential calculus techniques used in the derivation.

Suppose $q_i(\theta_i) = N(\mu^q_{\theta_i}, \Sigma^q_{\theta_i})$ where θ_i is a vector of length d. For a $d \times d$ square matrix A, $\operatorname{vec}(A)$ denotes the $d^2 \times 1$ vector obtained by stacking the columns of A under each other, from left to right in order, and $\operatorname{vech}(A)$ denotes the $\frac{1}{2}d(d+1) \times 1$ vector obtained from $\operatorname{vec}(A)$ by eliminating all supradiagonal elements of A. We can write $q_i(\theta_i)$ as

$$\exp\left\{\lambda_i^T \begin{bmatrix} \operatorname{vech}(\theta_i \theta_i^T) \\ \theta_i \end{bmatrix} - h(\lambda_i)\right\}$$

where

$$\lambda_i = \begin{bmatrix} -\frac{1}{2} D_d^T \operatorname{vec}(\Sigma^{q\,-1}_{\theta_i}) \\ \Sigma^{q\,-1}_{\theta_i} \mu^q_{\theta_i} \end{bmatrix}$$

and $h(\lambda_i) = \frac{1}{2}\mu^{q\,T}_{\theta_i} \Sigma^{q\,-1}_{\theta_i} \mu^q_{\theta_i} + \frac{1}{2}\log|\Sigma^q_{\theta_i}| + \frac{d}{2}\log(2\pi)$. The matrix D_d is a unique $d^2 \times \frac{1}{2}d(d+1)$ matrix that transforms $\operatorname{vech}(A)$ into $\operatorname{vec}(A)$ if A is symmetric, that is, $D_d \operatorname{vech}(A) = \operatorname{vec}(A)$. Let D_d^+ denote the Moore–Penrose inverse of D_d. If we let $\lambda_{i1} = -\frac{1}{2} D_d^T \operatorname{vec}(\Sigma^{q\,-1}_{\theta_i})$ and $\lambda_{i2} = \Sigma^{q\,-1}_{\theta_i} \mu^q_{\theta_i}$, $\frac{\partial S_a}{\partial \lambda_i}$ can be expressed as

$$\begin{aligned} \begin{bmatrix} \frac{\partial S_a}{\partial \lambda_{i1}} \\ \frac{\partial S_a}{\partial \lambda_{i2}} \end{bmatrix} &= \begin{bmatrix} \frac{\partial \operatorname{vec}(\Sigma^q_{\theta_i})}{\partial \lambda_{i1}} & \frac{\partial \mu^q_{\theta_i}}{\partial \lambda_{i1}} \\ \frac{\partial \operatorname{vec}(\Sigma^q_{\theta_i})}{\partial \lambda_{i2}} & \frac{\partial \mu^q_{\theta_i}}{\partial \lambda_{i2}} \end{bmatrix} \begin{bmatrix} \frac{\partial S_a}{\partial \operatorname{vec}(\Sigma^q_{\theta_i})} \\ \frac{\partial S_a}{\partial \mu^q_{\theta_i}} \end{bmatrix} \\ &= U(\lambda_i) \begin{bmatrix} \frac{\partial S_a}{\partial \operatorname{vec}(\Sigma^q_{\theta_i})} \\ \frac{\partial S_a}{\partial \mu^q_{\theta_i}} \end{bmatrix}, \end{aligned}$$

where

$$U(\lambda_i) = \begin{bmatrix} 2D_d^+(\Sigma^q_{\theta_i} \otimes \Sigma^q_{\theta_i}) & 2D_d^+(\mu^q_{\theta_i} \otimes \Sigma^q_{\theta_i}) \\ 0 & \Sigma^q_{\theta_i} \end{bmatrix}$$

and $\otimes$ denotes the Kronecker product. Moreover, $\mathcal{V}(\lambda_i) = \frac{\partial^2 h(\lambda_i)}{\partial \lambda_i \partial \lambda_i^T}$ can be derived to be

$$\begin{bmatrix} 2D_d^+(\mu^q_{\theta_i}\mu^{q\,T}_{\theta_i} \otimes \Sigma^q_{\theta_i} \\ \quad + \Sigma^q_{\theta_i} \otimes \mu^q_{\theta_i}\mu^{q\,T}_{\theta_i} & 2D_d^+(\mu^q_{\theta_i} \otimes \Sigma^q_{\theta_i}) \\ \quad + \Sigma^q_{\theta_i} \otimes \Sigma^q_{\theta_i})D_d^{+T} & \\ \{2D_d^+(\mu^q_{\theta_i} \otimes \Sigma^q_{\theta_i})\}^T & \Sigma^q_{\theta_i} \end{bmatrix}.$$

The update for λ_i can be computed as

$$\lambda_i \leftarrow \mathcal{V}(\lambda_i)^{-1} U(\lambda_i) \sum_{a \in N(\theta_i)} \begin{bmatrix} \frac{\partial S_a}{\partial \operatorname{vec}(\Sigma^q_{\theta_i})} \\ \frac{\partial S_a}{\partial \mu^q_{\theta_i}} \end{bmatrix}$$

and

$$\mathcal{V}(\lambda_i)^{-1} U(\lambda_i) = \begin{bmatrix} D_d^T & 0 \\ -2(\mu^{q\,T}_{\theta_i} \otimes I) D_d^{+T} D_d^T & I \end{bmatrix}.$$

Wand (2013) showed that the updates simplify to

$$\Sigma^q_{\theta_i} \leftarrow -\frac{1}{2}\left[\operatorname{vec}^{-1}\left(\sum_{a \in N(\theta_i)} \frac{\partial S_a}{\partial \operatorname{vec}(\Sigma^q_{\theta_i})}\right)\right]^{-1} \quad \text{and}$$

$$\mu^q_{\theta_i} \leftarrow \mu^q_{\theta_i} + \Sigma^q_{\theta_i} \sum_{a \in N(\theta_i)} \frac{\partial S_a}{\partial \mu^q_{\theta_i}}. \tag{15}$$

A more detailed version of the argument will be given in the forthcoming manuscript of Wand (2013).

4.3 Nonconjugate Variational Message Passing Algorithm for Generalized Linear Mixed Models

For the GLMM, we consider a variational approximation of the form

$$q(\theta) = q(\beta) q(D) \prod_{i=1}^{n} q(\tilde{\alpha}_i), \tag{16}$$

where $q(\beta)$ is $N(\mu^q_\beta, \Sigma^q_\beta)$, $q(D)$ is $IW(\nu^q, S^q)$, and $q(\tilde{\alpha}_i)$ is $N(\mu^q_{\tilde{\alpha}_i}, \Sigma^q_{\tilde{\alpha}_i})$, all belonging to the exponential family. Here, we approximate the posterior distributions of β and $\tilde{\alpha}_i$ by Gaussian distributions which are often reasonable and supported by the asymptotic normality of the posterior. Our results also indicate that Gaussian approximation performs reasonably well as an approximation to the posterior in finite samples. See Gelman et al. (2004) for further discussion and counterexamples. The posterior distribution for D is approximated by an inverse Wishart which can be shown

Initialize μ_β^q, Σ_β^q, S^q and $\mu_{\tilde{\alpha}_i}^q$, $\Sigma_{\tilde{\alpha}_i}^q$, W_i for $i = 1, \ldots, n$ and set $\nu^q = n + \nu$.
Cycle:

1. Update W_i and hence V_i for $i = 1, \ldots, n$. (Optional)
2. $\Sigma_\beta^q \leftarrow (\Sigma_\beta^{-1} + \nu^q \sum_{i=1}^n \tilde{W}_i^T S^{q-1} \tilde{W}_i + \sum_{i=1}^n \sum_{j=1}^{n_i} F_{ij} V_{ij} V_{ij}^T)^{-1}$
   $\mu_\beta^q \leftarrow \mu_\beta^q + \Sigma_\beta^q \{-\Sigma_\beta^{-1} \mu_\beta^q + \nu^q \sum_{i=1}^n \tilde{W}_i^T S^{q-1} (\mu_{\tilde{\alpha}_i}^q - \tilde{W}_i \mu_\beta^q) + \sum_{i=1}^n V_i^T (y_i - G_i)\}$
3. For $i = 1, \ldots, n$,
   $\Sigma_{\tilde{\alpha}_i}^q \leftarrow (\nu^q S^{q-1} + \sum_{j=1}^{n_i} F_{ij} X_{ij}^R X_{ij}^{R\,T})^{-1}$
   $\mu_{\tilde{\alpha}_i}^q \leftarrow \mu_{\tilde{\alpha}_i}^q + \Sigma_{\tilde{\alpha}_i}^q \{-\nu^q S^{q-1} (\mu_{\tilde{\alpha}_i}^q - \tilde{W}_i \mu_\beta^q) + X_i^{R\,T} (y_i - G_i)\}$
4. $S^q \leftarrow S + \sum_{i=1}^n \{(\mu_{\tilde{\alpha}_i}^q - \tilde{W}_i \mu_\beta^q)(\mu_{\tilde{\alpha}_i}^q - \tilde{W}_i \mu_\beta^q)^T + \Sigma_{\tilde{\alpha}_i}^q + \tilde{W}_i \Sigma_\beta^q \tilde{W}_i^T\}$

until the absolute relative change in the lower bound $\mathcal{L}$ is negligible.

ALGORITHM 3. *Nonconjugate variational message passing for fitting GLMMs.*

to be the optimal density under only the VB assumption $q(\theta) = q(\beta)q(D)q(\tilde{\alpha})$. The nonconjugate variational message passing algorithm for GLMMs is outlined in Algorithm 3.

In Algorithm 3, for each $i = 1, \ldots, n$, $j = 1, \ldots, n_i$, $\tilde{W}_i = [(I - W_i)C_i \;\; 0_{r \times (p-r-g_1)}]$, κ_{ij} is the jth component of $\kappa_i = \exp\{V_i \mu_\beta^q + X_i^R \mu_{\tilde{\alpha}_i}^q + \frac{1}{2} \operatorname{diag}(V_i \Sigma_\beta^q V_i^T + X_i^R \Sigma_{\tilde{\alpha}_i}^q X_i^{R\,T})\}$, μ_{ij} is the jth component of $\mu_i = V_i \mu_\beta^q + X_i^R \mu_{\tilde{\alpha}_i}^q$, σ_{ij} is the jth component of $\sigma_i = \sqrt{\operatorname{diag}(V_i \Sigma_\beta^q V_i^T + X_i^R \Sigma_{\tilde{\alpha}_i}^q X_i^{R\,T})}$ and $B^{(r)}(\mu, \sigma) = \int_{-\infty}^{\infty} b^{(r)}(\sigma x + \mu) \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2} \, \mathrm{d}x$ where $b(x) = \log(1 + \mathrm{e}^x)$ and $b^{(r)}(x)$ denotes the rth derivative of $b(\cdot)$ with respect to x. If μ and σ are vectors, say,

$$\mu = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad \text{and} \quad \sigma = \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix},$$

then

$$B^{(r)}(\mu, \sigma) = \begin{bmatrix} B^{(r)}(1, 4) \\ B^{(r)}(2, 5) \\ B^{(r)}(3, 6) \end{bmatrix}.$$

In addition,

$$F_{ij} = \begin{cases} E_{ij} \kappa_{ij}, & \text{if Poisson,} \\ B^{(2)}(\mu_{ij}, \sigma_{ij}), & \text{if logistic,} \end{cases}$$

and

$$G_i = \begin{cases} E_i \odot \kappa_i, & \text{if Poisson,} \\ B^{(1)}(\mu_i, \sigma_i), & \text{if logistic,} \end{cases}$$

where $a \odot b$ denotes the element-wise product of two vectors, a and b.

The updates in Algorithm 3 can be obtained from the formulae in (14) and (15). Consider the parameters ν^q and S^q of $q(D)$. The factors connected to D are $p(D|\nu, S)$ and $p(\tilde{\alpha}_i|\beta, D)$, $i = 1, \ldots, n$, which are all conjugate factors. Therefore, updates for $q(D)$ can be obtained from (14) or by setting $q(D) \propto \exp\{E_{-D} \log p(y, \theta)\}$ as in VB. The shape parameter ν^q can be shown to be deterministic: $\nu^q = n + \nu$ and the update for S^q is given in step 4 of Algorithm 3. The updates of the parameters of $q(\beta)$ and $q(\tilde{\alpha}_i)$, $i = 1, \ldots, n$, have to be computed using (15), as $p(y_i|\beta, \tilde{\alpha}_i)$ connected to β and $\tilde{\alpha}_i$ is a nonconjugate factor. The factors connected to β are $p(\beta|\Sigma_\beta)$, $p(\tilde{\alpha}_i|\beta, D)$ and $p(y_i|\beta, \tilde{\alpha}_i)$, $i = 1, \ldots, n$ (see Figure 1). Let $S_\beta = E_q\{\log p(\beta|\Sigma_\beta)\}$, $S_{\tilde{\alpha}_i} = E_q\{\log p(\tilde{\alpha}_i|\beta, D)\}$ and $S_{y_i} = E_q\{\log p(y_i|\beta, \tilde{\alpha}_i)\}$, $i = 1, \ldots, n$, where E_q denotes expectation with respect to q. We have

$$\sum_{a \in N(\beta)} \frac{\partial S_a}{\partial \operatorname{vec}(\Sigma_\beta^q)} = \frac{\partial S_\beta}{\partial \operatorname{vec}(\Sigma_\beta^q)} + \sum_{i=1}^n \frac{\partial S_{\tilde{\alpha}_i}}{\partial \operatorname{vec}(\Sigma_\beta^q)} + \sum_{i=1}^n \frac{\partial S_{y_i}}{\partial \operatorname{vec}(\Sigma_\beta^q)},$$

$$\sum_{a \in N(\beta)} \frac{\partial S_a}{\partial \mu_\beta^q} = \frac{\partial S_\beta}{\partial \mu_\beta^q} + \sum_{i=1}^n \frac{\partial S_{\tilde{\alpha}_i}}{\partial \mu_\beta^q} + \sum_{i=1}^n \frac{\partial S_{y_i}}{\partial \mu_\beta^q},$$

and the simplified updates for Σ_β^q and μ_β^q are given in step 2 of Algorithm 3. The factors connected to $\tilde{\alpha}_i$ are $p(\tilde{\alpha}_i|\beta, D)$ and $p(y_i|\beta, \tilde{\alpha}_i)$ for $i = 1, \ldots, n$ (see Figure 1). Hence,

$$\sum_{a \in N(\tilde{\alpha}_i)} \frac{\partial S_a}{\partial \operatorname{vec}(\Sigma_{\tilde{\alpha}_i}^q)} = \frac{\partial S_{\tilde{\alpha}_i}}{\partial \operatorname{vec}(\Sigma_{\tilde{\alpha}_i}^q)} + \frac{\partial S_{y_i}}{\partial \operatorname{vec}(\Sigma_{\tilde{\alpha}_i}^q)}$$

and

$$\sum_{a\in N(\tilde{\alpha}_i)} \frac{\partial S_a}{\partial \mu^q_{\tilde{\alpha}_i}} = \frac{\partial S_{\tilde{\alpha}_i}}{\partial \mu^q_{\tilde{\alpha}_i}} + \frac{\partial S_{y_i}}{\partial \mu^q_{\tilde{\alpha}_i}}.$$

The simplified updates for $\Sigma^q_{\tilde{\alpha}_i}$ and $\mu^q_{\tilde{\alpha}_i}$ are given in step 3 of Algorithm 3. See Appendix A for the evaluation of S_β, $S_{\tilde{\alpha}_i}$ and S_{y_i}. All gradients can be computed using vector differential calculus (see Magnus and Neudecker, 1988).

For responses from the Poisson family, S_{y_i} can be evaluated in closed form. However, S_{y_i} cannot be evaluated analytically for Bernoulli responses. Knowles and Minka (2011) discussed several alternatives in handling this integral. One could construct a bound on $\log(1 + e^x)$ such as the "quadratic" bound (Jaakkola and Jordan, 2000) or the "tilted" bound (Saul and Jordan, 1998). We observed a negative bias in the estimates for the random effects variances when using the "tilted bound" in Algorithm 3. This negative bias decreases as the cluster size increases (see also Rijmen and Vomlel, 2008). Hence, we use quadrature to compute the expectation and gradients. Following Ormerod and Wand (2012), we reduce all high-dimensional integrals to univariate ones and evaluate these efficiently using adaptive Gauss–Hermite quadrature (Liu and Pierce, 1994). The details are given in Appendix B.

While the updates in Algorithm 1 can be simplified if $W_i = I$ (noncentered) or 0 (centered) and are more complex in the partially noncentered case, the reduction in efficiency is minimal. Moreover, with a good initialization, it is feasible to keep W_i as fixed throughout the course of running Algorithm 3 so that no additional computation time is used in updating W_i. We use the fit from penalized quasi-likelihood implemented via the function `glmmPQL()` in the `R` package `MASS` (Venables and Ripley, 2002) to initialize Algorithm 3. In our experiments, the lower bound computed at the end of each cycle of updates is usually on an increasing trend although there might be some instability at the beginning. In cases where the algorithm does not converge, we found that changing the initialization can help to alleviate the situation. Although the lower bound is not guaranteed to increase at the end of each cycle, we continue to use it as a means of monitoring convergence and Algorithm 3 is terminated when the absolute relative change in the lower bound is less than 10^{-6}. The lower bounds for the logistic and Poisson GLMMs are presented in Appendix A.

5. MODEL SELECTION BASED ON VARIATIONAL LOWER BOUND

At the point of convergence of Algorithm 3, the lower bound on the log marginal likelihood, $\log p(y)$, is maximized. This variational lower bound is often tight and can be useful for model selection. Bayesian model selection is traditionally based on computation of Bayes factor in which marginal likelihood plays an important role. Suppose there are k candidate models, $M_1, \ldots, M_k$. Let $p(M_j)$ and $p(y|M_j)$ denote the prior probability and marginal likelihood of model M_j, respectively. To compare any two models, say, M_i and M_j, consider the posterior odds in favor of model M_i:

$$\frac{p(M_i|y)}{p(M_j|y)} = \frac{p(M_i)p(y|M_i)}{p(M_j)p(y|M_j)}.$$

The ratio of the marginal likelihoods, $\frac{p(y|M_i)}{p(y|M_j)}$, is the Bayes factor and can be considered as the strength of evidence provided by the data in favor of model M_i over M_j. Therefore, model comparison can be performed using marginal likelihoods once a prior has been specified on the models. See O'Hagan and Forster (2004) for a review of Bayes factors and alternative methods for Bayesian model choice. In Section 6.4, we demonstrate how the variational lower bound, a by-product of Algorithm 3, can be used in place of the log marginal likelihood to obtain approximate posterior model probabilities, assuming all models considered are equally probable. Formerly, Corduneanu and Bishop (2001) verified through experiments and comparisons with cross-validation that the variational lower bound is a good score for model selection in Gaussian mixture models.

We note that standard model selection criteria such as AIC or BIC are difficult to apply to GLMMs, as it is not straightforward to determine the degrees of freedom of a GLMM. Yu and Yau (2012) developed a conditional Akaike information criterion for GLMMs which takes into account estimation uncertainty in variance component parameters. Overstall and Forster (2010) considered a default strategy for Bayesian model selection addressing issues of prior specification and computation. See also Cai and Dunson (2008) for a review of variable selection methods for GLMMs.

6. EXAMPLES

We investigate the performance of Algorithm 3 using different parametrizations by considering a simulation

study and some real data sets. When using partial noncentering, we can either initialize the tuning parameters, W_i for $i = 1, \ldots, n$, and keep them as fixed or update them at the beginning of each cycle (see Algorithm 3, step 1). Such updates are particularly useful when a good initialization is lacking. We present results for both cases. There might not be significant improvement in updating W_i in the examples below, as the initialization using penalized quasi-likelihood is already good.

We assessed the performance of Algorithm 3 using different parametrizations by using MCMC as a "gold standard." Fitting via MCMC was performed in WinBUGS (Lunn et al., 2000) through `R` by using `R2WinBUGS` (Sturtz, Ligges and Gelman, 2005) as an interface. WinBUGS automatically implements a Markov chain simulation for the posterior distribution after the user specifies a model and starting values (see, e.g., Gelman et al., 2004). We used the centered parametrization when specifying the model in WinBUGS, as this produced better mixing than the noncentered parametrization for most of the examples considered (see Brown and Zhou, 2010). The MCMC algorithm was initialized similarly using the fit from penalized quasi-likelihood. In each case, three chains were run simultaneously to assess convergence, each with 50,000 iterations, and the first 5000 iterations were discarded in each chain as burn-in. A thinning factor of 10 was applied to reduce dependence between draws. The posterior means and standard deviations reported were based on the remaining 13,500 iterations. The computation times reported for MCMC are the times taken for updating in WinBUGS. We used the same priors for MCMC and Algorithm 3. For the fixed effects, we used a $N(0, 1000I)$ prior. All code was written in the R language and run on a dual processor Windows PC 3.30 GHz workstation.

6.1 Simulated Data

In this simulation study we consider the Poisson random intercept model

$$y_{ij}|u_i \sim \text{Poisson}\big(\exp(\beta_0 + \beta_1 x_{ij} + u_i)\big)$$

and the logistic random intercept model

$$y_{ij}|u_i \sim \text{Bernoulli}\left(\frac{\exp(\beta_0 + \beta_1 x_{ij} + u_i)}{1 + \exp(\beta_0 + \beta_1 x_{ij} + u_i)}\right),$$

where $u_i \sim N(0, \sigma^2)$. For the Poisson random intercept model, we set $x_{ij} = j - 1$ for $i = 1, \ldots, 100$, $j = 1, 2$, and used $\beta_0 = \beta_1 = -0.5$, $\sigma = 0.1$. For the logistic random intercept model, we set $x_{ij} = \frac{j}{8}$, for $i = 1, \ldots, 50$, $j = 1, \ldots, 8$, and used $\beta_0 = 0$, $\beta_1 = 5$, $\sigma = \sqrt{1.5}$. Similar settings have been considered by Ormerod and Wand (2012). For each model, 100 data sets were generated. No convergence issues were encountered for these simulated data, but experience with other simulated data sets (not shown) indicate that problems may arise when the covariance matrix of the fixed effects estimated from penalized quasi-likelihood is nearly singular or when the standard deviation of the random effects are very close to zero. In such cases, we can use alternative means of initialization such as estimates from the generalized linear model obtained by setting the random effects as zero. The expression in (5) can also serve as a prior guess for D (see Kass and Natarajan, 2006). Table 1 reports the estimates from penalized quasi-likelihood and the posterior means and standard deviations estimated by Algorithm 3 (using different parametrizations) and MCMC. Results are averaged over the 100 sets of simulated data. We have also included root mean squared errors computed as $\sqrt{\frac{1}{100}\sum_{l=1}^{100}(\hat{\vartheta}_l - \vartheta_l^0)^2}$ for an estimate $\hat{\vartheta}_l$ from the lth simulated data set obtained from penalized quasi-likelihood or Algorithm 3 where ϑ_l^0 is the corresponding estimate from MCMC regarded as the "gold standard."

For the Poisson model, the posterior means of the fixed effects and random effects estimated using the centered and noncentered parametrizations are quite close and also close to that of MCMC. However, the posterior standard deviations of the fixed effects are underestimated in the centered parametrization and the noncentered parametrization does better. The average time to convergence was shorter with noncentering and a higher lower bound was attained on average. We observe that the partially noncentered parametrization where tuning parameters were not updated took on average the least time to converge and produced a fit closer to that of the noncentered parametrization but with improvements in the estimation of the posterior means of the random effects. When the tuning parameters were updated, the fit was just as good, although computation time was longer. For the logistic model, centering and noncentering have different merits. While centering produced better estimates of the posterior means, the posterior standard deviations of the fixed effects were underestimated. The partially noncentered parametrization tries to adapt between the centered and noncentered parametrizations, producing better estimates of the posterior means than noncentering and better estimates of the posterior standard deviations than centering. When the tuning parameters were

TABLE 1

Results of simulation study showing initialization values from penalized quasi-likelihood, posterior means and standard deviations estimated by Algorithm 3 (different parametrizations) and MCMC, computation times (seconds) and variational lower bounds ($\mathcal{L}$), averaged over 100 sets of simulated data. Values in () are the corresponding root mean squared errors

Model	Method	β_0	SE(β_0)	β_1	SE(β_1)	σ	SE(σ)	Time	$\mathcal{L}$
Poisson	Penalized quasi-likelihood	−0.54 (0.11)	0.13 (0.02)	−0.48 (0.01)	0.19 (0.03)	0.27 (0.35)	—	0.1	—
	Noncentered	−0.63 (0.01)	0.13 (0.02)	−0.49 (<0.005)	0.21 (<0.005)	0.48 (0.02)	0.03 (0.08)	3.6	−196.0
	Centered	−0.63 (0.01)	0.05 (0.10)	−0.50 (0.01)	0.16 (0.05)	0.50 (0.01)	0.04 (0.07)	4.3	−197.0
	Partially noncentered: W_i fixed	−0.63 (0.01)	0.13 (0.02)	−0.49 (<0.005)	0.20 (0.01)	0.49 (0.01)	0.03 (0.08)	3.5	−196.0
	Partially noncentered: W_i updated	−0.63 (0.01)	0.13 (0.02)	−0.49 (<0.005)	0.19 (0.02)	0.49 (0.01)	0.03 (0.08)	4.0	−196.0
	MCMC	−0.64	0.15	−0.48	0.21	0.50	0.11	60.1	—
Logistic	Penalized quasi-likelihood	−0.10 (0.06)	0.32 (0.07)	5.02 (0.27)	0.63 (0.24)	1.25 (0.16)	—	0.2	—
	Noncentered	−0.07 (0.02)	0.33 (0.06)	5.20 (0.04)	0.77 (0.09)	1.18 (0.06)	0.12 (0.20)	3.2	−140.4
	Centered	−0.07 (0.02)	0.17 (0.21)	5.24 (0.02)	0.41 (0.45)	1.24 (0.03)	0.13 (0.20)	3.1	−141.1
	Partially noncentered: W_i fixed	−0.07 (0.02)	0.30 (0.09)	5.23 (0.02)	0.50 (0.37)	1.22 (0.03)	0.12 (0.20)	2.9	−140.5
	Partially noncentered: W_i updated	−0.07 (0.02)	0.30 (0.08)	5.21 (0.04)	0.50 (0.36)	1.22 (0.04)	0.12 (0.20)	3.9	−140.5
	MCMC	−0.05	0.38	5.23	0.85	1.24	0.32	146.6	—

updated, the results leaned more toward the noncentered parametrization and the algorithm took longer to converge. In both cases, Algorithm 3 using the partially noncentered parametrization was faster than MCMC and provided better estimates of the fixed effects and random effects than penalized quasi-likelihood. There are some difficulties, however, in comparing Algorithm 3 and MCMC in this way, as the time taken for Algorithm 3 to converge depends on the initialization, stopping rule and the rate of convergence also depends on the problem. Similarly, the updating time taken for MCMC is also problem-dependent and depends on the length of burn-in and number of sampling iterations. In addition, we observed (in simulated data sets not shown) that posterior inferences can be sensitive to prior assumptions on the variance components in Poisson models where many of the counts are close to zero or in binary data where the cluster size is small (see Browne and Draper, 2006 and Roos and Held, 2011).

6.2 Epilepsy Data

Here we consider the epilepsy data of Thall and Vail (1990) which has been analyzed by many authors (see, e.g., Breslow and Clayton, 1993; Ormerod and Wand, 2012). In this clinical trial, 59 epileptics were randomized to a new anti-epileptic drug, progabide (Trt = 1) or a placebo (Trt = 0). Before receiving treatment, baseline data on the number of epileptic seizures during the preceding 8-week period were recorded. The logarithm of $\frac{1}{4}$ the number of baseline seizures (Base) and the logarithm of age (Age) were treated as covariates. Counts of epileptic seizures during the 2 weeks before each of four successive clinic visits (Visit, coded as $\text{Visit}_1 = -0.3$, $\text{Visit}_2 = -0.1$, $\text{Visit}_3 = 0.1$ and $\text{Visit}_4 = 0.3$) were recorded. A binary variable (V4 = 1 for fourth visit, 0 otherwise) was also considered as a covariate. We consider models II and IV from Breslow and Clayton (1993). Model II is a Poisson random intercept model where

$$\begin{aligned}\log \mu_{ij} = {} & \beta_0 + \beta_{\text{Base}}\text{Base}_i + \beta_{\text{Trt}}\text{Trt}_i \\ & + \beta_{\text{Base}\times\text{Trt}}\text{Base}_i \times \text{Trt}_i + \beta_{\text{Age}}\text{Age}_i \\ & + \beta_{\text{V4}}\text{V4}_{ij} + u_i\end{aligned}$$

for $i = 1, \ldots, n$, $j = 1, \ldots, 4$ and $u_i \sim N(0, \sigma^2)$. Model IV is a Poisson random intercept and slope

model of the form

$$\log \mu_{ij} = \beta_0 + \beta_{\text{Base}}\text{Base}_i + \beta_{\text{Trt}}\text{Trt}_i + \beta_{\text{Base}\times\text{Trt}}\text{Base}_i \times \text{Trt}_i + \beta_{\text{Age}}\text{Age}_i + \beta_{\text{Visit}}\text{Visit}_{ij} + u_{1i} + u_{2i}\text{Visit}_{ij}$$

for $i = 1, \ldots, n$, $j = 1, \ldots, 4$ and

$$\begin{bmatrix} u_{1i} \\ u_{2i} \end{bmatrix} \sim N\left(0, \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{21} & \sigma_{22}^2 \end{bmatrix}\right).$$

As the MCMC chains for intercept and Age were mixing poorly, we decided to center the covariate Age. In the analysis that follows, we assume Age_i has been replaced by $\text{Age}_i - \text{mean(Age)}$.

Table 2 shows the estimates of the posterior means and standard deviations of the fits from MCMC and Algorithm 3 (using different parametrizations), initialization values from penalized quasi-likelihood and computation times in seconds taken by different methods. All the variational methods are faster than MCMC by an order of magnitude which is especially important in large scale applications. In the noncentered parametrization, the standard deviations of the fixed effects were underestimated and the centered parametrization does better in this aspect. The partially noncentered parametrization produced a fit that is closer to that of the centered parametrization and improved upon it. In both models, the fits produced by partial noncentering are very close to that produced by MCMC and are superior to that of the centered and noncentered parametrizations. The lower bound attained by partial noncentering is also higher than that of centering and noncentering, giving a tighter bound on the log marginal likelihood. It is important to emphasize that the relevant comparison is of the partially noncentered parametrization to the worst of the centered and noncentered parametrizations, since in general we do not know if centering or noncentering is better without running both algorithms. Partial noncentering, on the other hand, automatically chooses a near optimal parametrization. Updating of the tuning parameters helped to improve the fit produced by partial noncentering. Figure 2 shows the marginal posterior distributions for parameters in models II and IV estimated by MCMC (solid line) and Algorithm 3 using

TABLE 2

Results for epilepsy data models II and IV showing initialization values from penalized quasi-likelihood, posterior means and standard deviations (values after ±) estimated by Algorithm 3 (different parametrizations) and MCMC, computation times (seconds) and variational lower bounds ($\mathcal{L}$)

	Penalized quasi-likelihood	**Noncentered**	**Centered**	**Partially noncentered: W_i fixed**	**Partially noncentered: W_i updated**	**MCMC**
Model II						
β_0	0.31 ± 0.26	0.26 ± 0.11	0.27 ± 0.24	0.27 ± 0.26	0.27 ± 0.27	0.26 ± 0.27
β_{Base}	0.88 ± 0.13	0.89 ± 0.04	0.88 ± 0.13	0.88 ± 0.13	0.88 ± 0.14	0.89 ± 0.14
β_{Trt}	-0.91 ± 0.41	-0.94 ± 0.15	-0.94 ± 0.36	-0.94 ± 0.40	-0.94 ± 0.41	-0.94 ± 0.42
$\beta_{\text{Base}\times\text{Trt}}$	0.34 ± 0.20	0.34 ± 0.06	0.34 ± 0.19	0.34 ± 0.21	0.34 ± 0.21	0.34 ± 0.21
β_{Age}	0.54 ± 0.35	0.50 ± 0.12	0.48 ± 0.33	0.48 ± 0.35	0.48 ± 0.36	0.48 ± 0.37
β_{V4}	-0.16 ± 0.08	-0.16 ± 0.05	-0.16 ± 0.05	-0.16 ± 0.05	-0.16 ± 0.05	-0.16 ± 0.05
σ	0.44	0.50 ± 0.05	0.54 ± 0.05	0.53 ± 0.05	0.53 ± 0.05	0.53 ± 0.06
$\mathcal{L}$	—	-707.3	-702.0	-701.6	-701.5	—
Time	0.2	1.1	0.4	0.4	0.6	61
Model IV						
β_0	0.27 ± 0.26	0.21 ± 0.10	0.21 ± 0.24	0.21 ± 0.26	0.21 ± 0.26	0.21 ± 0.27
β_{Base}	0.88 ± 0.13	0.89 ± 0.04	0.88 ± 0.13	0.89 ± 0.13	0.89 ± 0.13	0.88 ± 0.14
β_{Trt}	-0.92 ± 0.41	-0.94 ± 0.15	-0.93 ± 0.36	-0.93 ± 0.40	-0.93 ± 0.40	-0.94 ± 0.42
$\beta_{\text{Base}\times\text{Trt}}$	0.35 ± 0.20	0.34 ± 0.06	0.34 ± 0.19	0.34 ± 0.20	0.34 ± 0.21	0.34 ± 0.22
β_{Age}	0.54 ± 0.35	0.49 ± 0.12	0.47 ± 0.32	0.47 ± 0.35	0.47 ± 0.35	0.47 ± 0.37
β_{Visit}	-0.28 ± 0.16	-0.27 ± 0.10	-0.27 ± 0.10	-0.27 ± 0.14	-0.27 ± 0.15	-0.27 ± 0.17
σ_{11}	0.45	0.50 ± 0.05	0.53 ± 0.05	0.52 ± 0.05	0.53 ± 0.05	0.53 ± 0.06
σ_{22}	0.46	0.75 ± 0.07	0.77 ± 0.07	0.75 ± 0.07	0.76 ± 0.07	0.76 ± 0.15
$\mathcal{L}$	—	-701.4	-696.1	-695.3	-695.1	—
Time	0.5	1.5	1.3	1.2	1.4	122

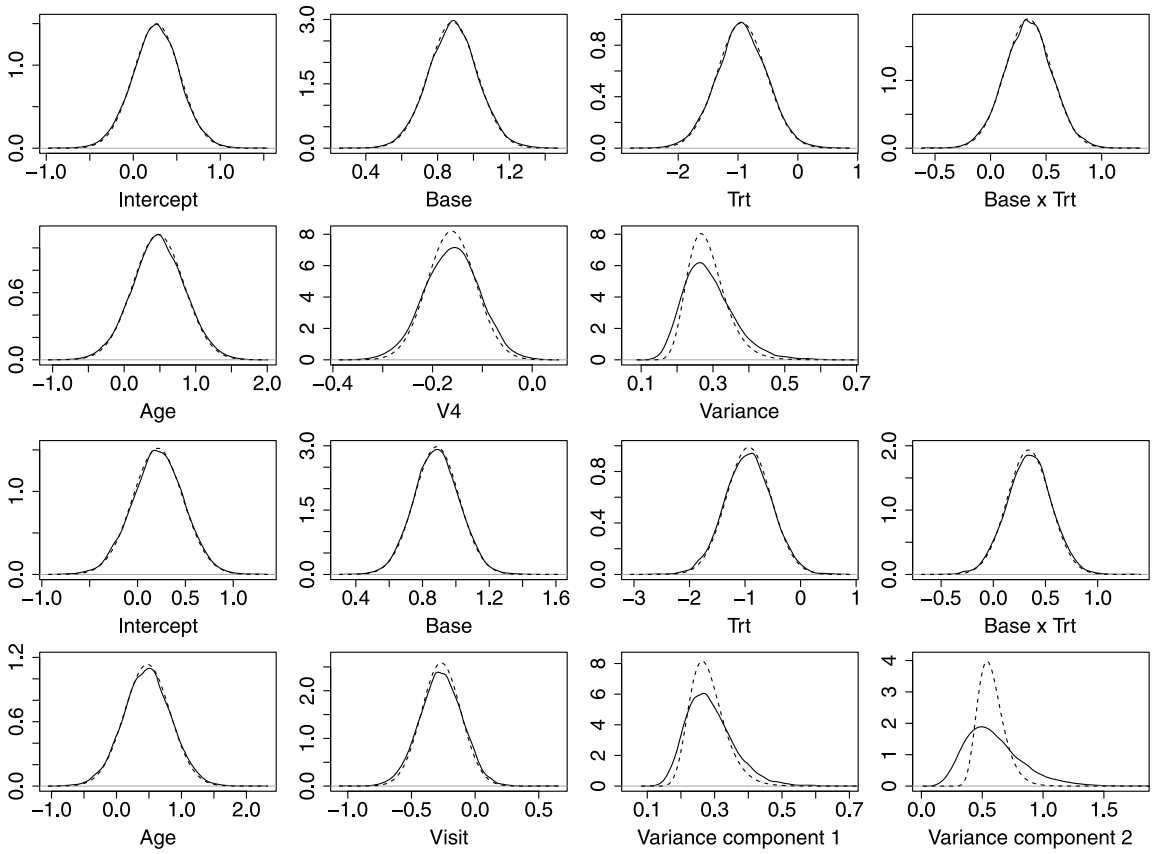

FIG. 2. *Marginal posterior distributions for parameters in model II (first two rows) and model IV (last two rows) of the epilepsy data estimated by MCMC (solid line) and Algorithm 3 using partially noncentered parametrization where tuning parameters are updated (dashed line).*

the partially noncentered parametrization where tuning parameters are updated (dashed line). The variational posterior densities of the fixed effects are very close to those obtained via MCMC. For the variance components, there is still some underestimation of the posterior variance.

6.3 Toenail Data

This data set was obtained from a multicenter study comparing two competing oral antifungal treatments for toenail infection (De Backer et al., 1998), courtesy of Novoartis, Belgium. It contains information for 294 patients to be evaluated at seven visits. Not all patients attended all seven planned visits and there were 1908 measurements in total. The patients were randomized into two treatment groups, one group receiving 250 mg per day of terbinafine (Trt = 1) and the other group 200 mg per day of itraconazole (Trt = 0). Visits were planned at weeks 0, 4, 8, 12, 24, 36 and 48, but patients did not always arrive as scheduled and the exact time in months (t) that they did attend was recorded. The binary response variable (onycholysis) indicates the degree of separation of the nail plate from the nail bed (0 if none or mild, 1 if moderate or severe). We consider the following logistic random intercept model,

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_{\text{Trt}}\text{Trt}_i + \beta_t t_{ij} + \beta_{\text{Trt}\times t}\text{Trt}_i \times t_{ij} + u_i,$$

where $u_i \sim N(0, \sigma^2)$ for $i = 1, \ldots, 294$, $1 \le j \le 7$.

Table 3 shows the posterior means and standard deviations of the fits from MCMC and Algorithm 3 (using different parametrizations), initialization values from penalized quasi-likelihood and computation time in seconds taken by different methods. Again, the VB methods are faster than MCMC by an order of magnitude. In this example, centering produced a better fit than noncentering and partial noncentering produced a fit closer to that of the centered parametrization but improving it. Partial noncentering also took less time to converge and attained a lower bound higher than that of the centered and noncentered parametrizations. Again, we emphasize that it is not easy to know beforehand

TABLE 3

Results for toenail data showing values used for initialization from penalized quasi-likelihood, posterior means and posterior standard deviations (values after ±) from Algorithm 3 (different parametrizations) and MCMC, computation times (seconds) and variational lower bounds ($\mathcal{L}$)

	Penalized quasi-likelihood	**Noncentered**	**Centered**	**Partially noncentered: W_i fixed**	**Partially noncentered: W_i updated**	**MCMC**
β_0	-0.75 ± 0.25	-1.41 ± 0.17	-1.44 ± 0.29	-1.44 ± 0.35	-1.44 ± 0.32	-1.65 ± 0.44
β_{Trt}	-0.04 ± 0.35	-0.13 ± 0.25	-0.13 ± 0.41	-0.13 ± 0.49	-0.13 ± 0.45	-0.17 ± 0.60
β_{t}	-0.30 ± 0.03	-0.38 ± 0.04	-0.38 ± 0.03	-0.38 ± 0.03	-0.38 ± 0.03	-0.40 ± 0.05
$\beta_{\text{Trt}\times\text{Time}}$	-0.10 ± 0.05	-0.13 ± 0.06	-0.13 ± 0.04	-0.13 ± 0.04	-0.13 ± 0.04	-0.14 ± 0.07
σ	2.32	3.52 ± 0.15	3.56 ± 0.15	3.55 ± 0.15	3.55 ± 0.15	4.10 ± 0.39
$\mathcal{L}$	—	-664.1	-663.1	-662.7	-662.9	—
Time	2.8	37.9	27.9	26.0	24.1	1072

which of centering or noncentering will perform better, and a big advantage of partial noncentering is the way that it automatically chooses a good parametrization. In this example, updating the tuning parameters did not result in a better fit although the time to convergence is reduced. The marginal posterior distributions estimated by MCMC (solid line) and Algorithm 3 using the partially noncentered parametrization where tuning parameters were not updated (dashed line) are shown in Figure 3. Compared with the MCMC fit, there is still some underestimation of the variance of the fixed effects particularly for the parameters which could not be centered. Although the partially noncentered parametrization has improved the estimation of the random effects from the initial penalized quasi-likelihood fit, there is still some underestimation of the mean and variance of the random effects when compared to the MCMC fit.

6.4 Six Cities Data

In the previous two real data examples, centering performed better than noncentering and partial noncentering was able to improve on the centering results. While centering often performs better than noncentering, we use this example to show that partial noncentering will automatically tend toward noncentering when noncentering is preferred. We consider the six cities data in Fitzmaurice and Laird (1993), where the binary response variable y_{ij} indicates the wheezing status (1 if wheezing, 0 if not wheezing) of the ith child at time-point j, $i = 1, \ldots, 537$, $j = 1, 2, 3, 4$. We use as covariate the age of the child at time-point j, centered at 9 years (Age), and consider the following random intercept and slope model:

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_{\text{Age}}\text{Age}_i + u_{1i} + u_{2i}\text{Age}_i$$

for $i = 1, \ldots, 537$, $j = 1, \ldots, 4$ and

$$\begin{bmatrix} u_{1i} \\ u_{2i} \end{bmatrix} \sim N\left(0, \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{21} & \sigma_{22}^2 \end{bmatrix}\right).$$

This model has been considered in Overstall and Forster (2010).

Table 4 shows the estimates of the posterior means and standard deviations of the fits from MCMC and Algorithm 3 using different parametrizations, the values from penalized quasi-likelihood used for initialization

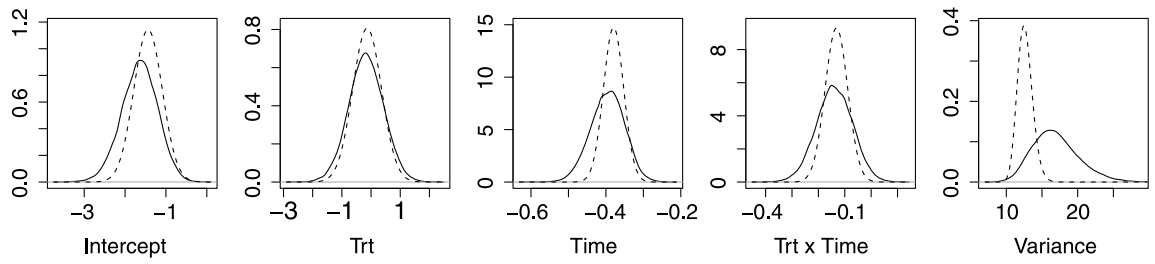

FIG. 3. *Marginal posterior distributions for parameters in toenail data estimated by MCMC (solid line) and Algorithm 3 using partially noncentered parametrization (tuning parameters not updated) (dashed line).*

TABLE 4

Results for six cities data showing values used for initialization from penalized quasi-likelihood, posterior means and posterior standard deviations (values after ±) from Algorithm 3 (different parametrizations) and MCMC, computation times (seconds) and variational lower bounds ($\mathcal{L}$)

	Penalized quasi-likelihood	**Noncentered**	**Centered**	**Partially noncentered: W_i fixed**	**Partially noncentered: W_i updated**	**MCMC**
β_0	-3.12 ± 0.14	-3.05 ± 0.09	-3.05 ± 0.09	-3.05 ± 0.13	-3.05 ± 0.13	-3.29 ± 0.25
β_{Age}	-0.24 ± 0.08	-0.22 ± 0.07	-0.21 ± 0.02	-0.22 ± 0.07	-0.22 ± 0.07	-0.25 ± 0.16
σ_{11}	2.52	2.16 ± 0.07	2.16 ± 0.07	2.16 ± 0.07	2.16 ± 0.07	2.48 ± 0.24
σ_{22}	1.19	0.55 ± 0.02	0.56 ± 0.02	0.55 ± 0.02	0.55 ± 0.02	0.61 ± 0.10
$\mathcal{L}$	—	-833.2	-834.1	-832.8	-832.6	—
Time	3.8	114.7	125.8	110.6	120.6	1010

and the computation times in seconds taken by different methods. Noncentering performed better than centering in this case with a shorter time to convergence, higher lower bound and a better estimate of the posterior standard deviation of β_{Age}. Partial noncentering further improved upon the results of noncentering with an improved estimate of the posterior standard deviation of β_0 and faster convergence. All the variational methods are again faster than MCMC by an order of magnitude.

6.5 Owl Data

In this example we illustrate the use of the variational lower bound, a by-product of Algorithm 3, for model selection. For MCMC, on the other hand, it is not straightforward in general to get a good estimate of the marginal likelihood based on the MCMC output. It is also not always obvious how to apply standard model selection criteria like AIC and BIC to hierarchical models like GLMMs.

Roulin and Bersier (2007) analyzed the begging behavior of nestling barn owls and looked at whether offspring beg for food at different intensities from the mother than father. They sampled $n = 27$ nests and counted the number of calls made by all offspring in the absence of parents. Half of the nests were given extra prey, and from the other half prey were removed. Measurements took place on two nights, and food treatment was swapped the second night. The number of measurements at each nest ranged from 4 to 52 with a total of 599. We use as covariates sex of parent (Sex = 1 if male, 0 if female), the time at which a parent arrived with a prey (t), and food treatment (Trt = 1 if "satiated," 0 if "deprived"). The number of nestlings per nest (broodsize, E) ranged from 1 to 7.

Zuur et al. (2009) modeled the number of calls at nest i for the jth observation as a Poisson distribution with mean μ_{ij} and used log transformed broodsize as an offset with nest as a random effect. The prime aim of their analysis was to find a sex effect and the largest model they considered was the following:

1. $\log(\mu_{ij}) = \log(E_{ij}) + \beta_0 + \beta_{\text{Sex}}\text{Sex}_{ij} + \beta_{\text{Trt}}\text{Trt}_{ij} + \beta_t t_{ij} + \beta_{\text{Sex}\times\text{Trt}}\text{Sex}_{ij} \times \text{Trt}_{ij} + \beta_{\text{Sex}\times t}\text{Sex}_{ij} \times t_{ij} + u_i,$

where $\log(E_{ij})$ is an offset and $u_i \sim N(0, \sigma^2)$ for $i = 1, \ldots, 27$, $j = 1, \ldots, n_i$. At the recommendation of Zuur et al. (2009), we center t to reduce correlation of t with the intercept. Henceforth, we assume t_{ij} has been replaced by $t_{ij} - \text{mean}(t)$. In the first stage, we consider models 1 to 4 and determine if the two interaction terms should be retained. Models 2 to 4 are as follows:

2. $\log(\mu_{ij}) = \log(E_{ij}) + \beta_0 + \beta_{\text{Sex}}\text{Sex}_{ij} + \beta_{\text{Trt}}\text{Trt}_{ij} + \beta_t t_{ij} + \beta_{\text{Sex}\times\text{Trt}}\text{Sex}_{ij} \times \text{Trt}_{ij} + u_i,$
3. $\log(\mu_{ij}) = \log(E_{ij}) + \beta_0 + \beta_{\text{Sex}}\text{Sex}_{ij} + \beta_{\text{Trt}}\text{Trt}_{ij} + \beta_t t_{ij} + \beta_{\text{Sex}\times t}\text{Sex}_{ij} \times t_{ij} + u_i,$
4. $\log(\mu_{ij}) = \log(E_{ij}) + \beta_0 + \beta_{\text{Sex}}\text{Sex}_{ij} + \beta_{\text{Trt}}\text{Trt}_{ij} + \beta_t t_{ij} + u_i.$

From Table 5, the preferred model (with the highest lower bound) is model 4 where both interaction terms have been dropped from model 1. Next, we consider models 5 to 7 where the main terms sex, food treatment and arrival time are each dropped in turn:

5. $\log(\mu_{ij}) = \log(E_{ij}) + \beta_0 + \beta_{\text{Trt}}\text{Trt}_{ij} + \beta_t t_{ij} + u_i,$
6. $\log(\mu_{ij}) = \log(E_{ij}) + \beta_0 + \beta_{\text{Trt}}\text{Trt}_{ij} + \beta_{\text{Sex}}\text{Sex}_{ij} + u_i,$
7. $\log(\mu_{ij}) = \log(E_{ij}) + \beta_0 + \beta_t t_{ij} + \beta_{\text{Sex}}\text{Sex}_{ij} + u_i.$

TABLE 5
Variational lower bounds for owl data models 1 *to* 11 *and computation time in brackets*

	Noncentered	Centered	Partially noncentered: W_i fixed	Partially noncentered: W_i updated
First stage				
Model 1	−2544.6 (0.2)	−2543.7 (0.3)	−2543.6 (0.4)	−2543.7 (0.6)
Model 2	−2537.6 (0.2)	−2536.6 (0.3)	−2536.6 (0.4)	−2536.6 (0.5)
Model 3	−2540.2 (0.2)	−2539.2 (0.3)	−2539.2 (0.3)	−2539.2 (0.5)
Model 4	−2533.2 (0.2)	−2532.1 (0.3)	−2532.1 (0.3)	−2532.1 (0.4)
Second stage				
Model 5	−2527.0 (0.2)	−2525.5 (0.2)	−2525.5 (0.2)	−2525.4 (0.3)
Model 6	−2628.3 (0.2)	−2627.2 (0.3)	−2627.1 (0.3)	−2627.1 (0.5)
Model 7	−2664.0 (0.2)	−2662.9 (0.2)	−2662.8 (0.3)	−2662.8 (0.4)
Third stage				
Model 8	−2621.5 (0.2)	−2620.0 (0.2)	−2620.0 (0.2)	−2620.0 (0.3)
Model 9	−2660.4 (0.2)	−2658.8 (0.2)	−2658.8 (0.2)	−2658.8 (0.2)
Model 10	−2689.4 (<0.05)			
Final stage				
Model 11	−2448.7 (1.1)	−2445.7 (0.4)	−2445.8 (0.3)	−2445.6 (0.4)

Table 5 indicates that model 5 is the preferred model where the term sex of the parent has been dropped from model 4. Now we consider dropping each of the terms food treatment and arrival time in turn or dropping the random effects u_i:

8. $\log(\mu_{ij}) = \log(E_{ij}) + \beta_0 + \beta_{\text{Trt}}\text{Trt}_{ij} + u_i$,
9. $\log(\mu_{ij}) = \log(E_{ij}) + \beta_0 + \beta_t t_{ij} + u_i$,
10. $\log(\mu_{ij}) = \log(E_{ij}) + \beta_0 + \beta_{\text{Trt}}\text{Trt}_{ij} + \beta_t t_{ij}$.

Table 5 indicates that none of the main terms food treatment and arrival time as well as random effects should be dropped from model 5. Finally, we consider adding a random slope for arrival time:

11. $\log(\mu_{ij}) = \log(E_{ij}) + \beta_0 + \beta_{\text{Trt}}\text{Trt}_{ij} + \beta_t t_{ij} + u_{1i} + u_{2i}t_{ij}$,

where

$$\begin{bmatrix} u_{1i} \\ u_{2i} \end{bmatrix} \sim N\left(0, \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{21} & \sigma_{22}^2 \end{bmatrix}\right).$$

From Table 5, the optimal model is model 11. This conclusion is similar to that of Zuur et al. (2009) and is the same regardless of which parametrization was used. It is thus sufficient to consider just the partially noncentered parametrization. The computation time taken by Algorithm 3 for each model fitting is very short and makes this a convenient way of carrying out model selection or for narrowing down the range of likely models. Further model comparisons can be performed using cross-validation or other approaches.

We present the estimated posterior means and standard deviations for the optimal model in Table 6. The marginal posterior distributions estimated by MCMC (solid line) and Algorithm 3 using partially noncentered parametrization where tuning parameters are updated (dashed line) are shown in Figure 4. In this case, centering produced a better fit than noncentering and partial noncentering produced a fit that is close to that of centering. Updating the tuning parameters helped to improve the fit of the partially noncentered parametrization slightly and is closest to the MCMC fit. From the posterior density plots, there is good estimation of the posterior means by Algorithm 3 using partially noncentered parametrization with updated tuning parameters, but there is still some underestimation of the posterior variance.

7. CONCLUSION

In this paper we described a partially noncentered parametrization for GLMMs and compared the performance of different parametrizations using an algorithm called nonconjugate variational message passing developed recently in machine learning. Focusing on Poisson and logistic mixed models, we applied our methods to analysis of longitudinal data sets. For the logistic model, some parameter updates were not available in closed form and we used adaptive Gauss–Hermite quadrature to approximate the intractable inte-

TABLE 6
Results for owl data (model 11) showing values used for initialization from penalized quasi-likelihood, posterior means and standard deviations (values after ±) from Algorithm 3 (different parametrizations) and MCMC and computation times (seconds)

	Penalized quasi-likelihood	**Noncentered**	**Centered**	**Partially noncentered: W_i fixed**	**Partially noncentered: W_i updated**	**MCMC**
β_0	0.60 ± 0.07	0.53 ± 0.02	0.51 ± 0.08	0.51 ± 0.08	0.51 ± 0.09	0.50 ± 0.10
β_{Trt}	-0.55 ± 0.08	-0.57 ± 0.03	-0.57 ± 0.03	-0.57 ± 0.03	-0.57 ± 0.03	-0.57 ± 0.04
β_t	-0.13 ± 0.03	-0.15 ± 0.01	-0.16 ± 0.04	-0.16 ± 0.04	-0.16 ± 0.04	-0.16 ± 0.05
σ_{11}	0.24	0.44 ± 0.06	0.46 ± 0.06	0.45 ± 0.06	0.46 ± 0.06	0.47 ± 0.09
σ_{22}	0.11	0.22 ± 0.03	0.23 ± 0.03	0.22 ± 0.03	0.23 ± 0.03	0.23 ± 0.05
Time	0.4	1.1	0.4	0.3	0.4	255

grals efficiently. Comparing the performance of Algorithm 3 under the partially noncentered parametrization with that of the centered and noncentered parametrizations, we observed that partial noncentering automatically tends toward the better of centering and noncentering so that it is not necessary to choose in advance between the centered and noncentered parametrizations. In many cases, the partially noncentered parametrization was able to improve upon the fit produced by the better of centering and noncentering to produce a fit that was closest to that of MCMC. In terms of computation time, the partially noncentered parametrization can also provide more rapid convergence when centering or noncentering is particularly slow. Very often, the lower bound attained by the partially noncentered parametrization is also higher than that of the centered and noncentered parametrizations, giving a tighter lower bound to the log marginal likelihood. To some degree, the partially noncentered parametrization also alleviates the issue of underestimation of the posterior variance, leading to some improvement in the estimation of the posterior variance, particularly in the fixed effects which could be centered. Algorithm 3 under the partially noncentered parametrization thus offers itself as a fast, deterministic alternative to MCMC methods for fitting GLMMs with improved estimation compared to the centered and noncentered parametrizations. We also demonstrate that the variational lower bound produced as part of the computation in Algorithm 3 can be useful in model selection.

APPENDIX A: EVALUATING THE VARIATIONAL LOWER BOUND

From (2), (7) and (16),

$$\mathcal{L} = \sum_{i=1}^{n} S_{y_i} + \sum_{i=1}^{n} S_{\tilde{\alpha}_i} + S_\beta + E_q\{\log p(D|\nu, S)\} - E_q\{\log q(\beta)\} - \sum_{i}^{n} E_q\{\log q(\tilde{\alpha}_i)\} - E_q\{\log q(D)\}.$$

To evaluate the terms in the lower bound, we use the following two lemmas which we state without proof:

LEMMA 1. *Suppose* $p_1(x) = N(\mu_1, \Sigma_1)$ *and* $p_2(x) = N(\mu_2, \Sigma_2)$ *where* x *is a* p*-dimensional vector, then* $\int p_2(x) \log p_1(x)\, \mathrm{d}x = -\frac{p}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_1| - \frac{1}{2}(\mu_2 - \mu_1)^T \Sigma_1^{-1}(\mu_2 - \mu_1) - \frac{1}{2}\operatorname{tr}(\Sigma_1^{-1}\Sigma_2)$.

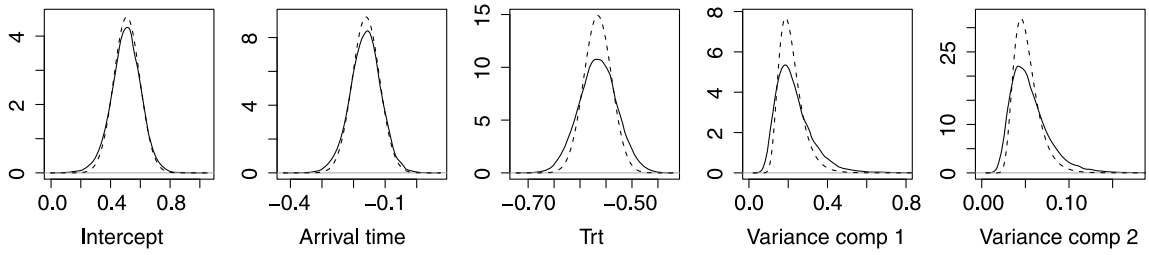

FIG. 4. *Marginal posterior distributions for parameters in model 11 (owl data) estimated by MCMC (solid line) and Algorithm 3 using partially noncentered parametrization where tuning parameters are updated (dashed line).*

LEMMA 2. *Suppose* $p(D) = IW(\nu, S)$ *where* D *is a symmetric, positive definite* $r \times r$ *matrix, then* $\int p(D)\log|D|\,\mathrm{d}D = \log|S| - \sum_{l=1}^{r}\psi(\frac{\nu-l+1}{2}) - r\log 2$ *and* $\int p(D)D^{-1}\,\mathrm{d}D = \nu S^{-1}$ *where* $\psi(\cdot)$ *denotes the digamma function.*

Using these two lemmas, we can compute most of the terms in the lower bound:

$$\begin{aligned}
S_\beta &= \int q(\beta)\log p(\beta|\Sigma_\beta)\,\mathrm{d}\beta \\
&= -\frac{p}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_\beta| \\
&\quad - \frac{1}{2}\mu_\beta^{q\,T}\Sigma_\beta^{-1}\mu_\beta^q - \frac{1}{2}\operatorname{tr}\bigl(\Sigma_\beta^{-1}\Sigma_\beta^q\bigr), \\
S_{\tilde\alpha_i} &= \int q(\beta)q(D)q(\tilde\alpha_i)\log p(\tilde\alpha_i|\beta, D)\,\mathrm{d}\beta\,\mathrm{d}D\,\mathrm{d}\tilde\alpha_i \\
&= -\frac{r}{2}\log(2\pi) \\
&\quad - \frac{1}{2}\biggl\{\log|S^q| - \sum_{l=1}^{r}\psi\biggl(\frac{\nu^q - l + 1}{2}\biggr) - r\log 2\biggr\} \\
&\quad - \frac{\nu^q}{2}\bigl[(\mu_{\tilde\alpha_i}^q - \tilde W_i\mu_\beta^q)^T S^{q\,-1}(\mu_{\tilde\alpha_i}^q - \tilde W_i\mu_\beta^q) \\
&\qquad + \operatorname{tr}\bigl\{S^{q\,-1}(\Sigma_{\tilde\alpha_i}^q + \tilde W_i\Sigma_\beta^q\tilde W_i^T)\bigr\}\bigr], \\
&E_q\{\log p(D|\nu, S)\} \\
&= \int q(D)\log p(D|\nu, S)\,\mathrm{d}D \\
&= -\frac{\nu^q}{2}\operatorname{tr}\bigl(S^{q\,-1}S\bigr) - \frac{r(r-1)}{4}\log(\pi) \\
&\quad - \sum_{l=1}^{r}\log\Gamma\biggl(\frac{\nu + 1 - l}{2}\biggr) + \frac{\nu}{2}\log|S| \\
&\quad - \frac{\nu + r + 1}{2}\biggl\{\log|S^q| - \sum_{l=1}^{r}\psi\biggl(\frac{\nu^q - l + 1}{2}\biggr) - r\log 2\biggr\} \\
&\quad - \frac{\nu r}{2}\log 2, \\
&E_q\{\log q(\beta)\} \\
&= \int q(\beta)\log q(\beta)\,\mathrm{d}\beta \\
&= -\frac{p}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_\beta^q| - \frac{p}{2}, \\
&E_q\{\log q(\tilde\alpha_i)\} \\
&= \int q(\tilde\alpha_i)\log q(\tilde\alpha_i)\,\mathrm{d}\tilde\alpha_i \\
&= -\frac{r}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_{\tilde\alpha_i}^q| - \frac{r}{2}, \\
&E_q\{\log q(D)\} \\
&= \int q(D)\log q(D)\,\mathrm{d}D \\
&= -\frac{\nu^q r}{2}\log 2 - \frac{r(r-1)}{4}\log\pi \\
&\quad - \sum_{l=1}^{r}\log\Gamma\biggl(\frac{\nu^q + 1 - l}{2}\biggr) + \frac{\nu^q}{2}\log|S^q| \\
&\quad - \frac{\nu^q + r + 1}{2}\biggl\{\log|S^q| - \sum_{l=1}^{r}\psi\biggl(\frac{\nu^q - l + 1}{2}\biggr) - r\log 2\biggr\} \\
&\quad - \frac{\nu^q r}{2}.
\end{aligned}$$

The only term left to evaluate is

$$S_{y_i} = \int q(\beta)q(\tilde\alpha_i)\log p(y_i|\beta, \tilde\alpha_i)\,\mathrm{d}\beta\,\mathrm{d}\tilde\alpha_i.$$

For Poisson responses with the log link function [see (9)],

$$\begin{aligned}
S_{y_i} &= y_i^T\{\log(E_i) + V_i\mu_\beta^q + X_i^R\mu_{\tilde\alpha_i}^q\} - E_i^T\kappa_i \\
&\quad - 1_{n_i}^T\log(y_i!),
\end{aligned}$$

where $\kappa_i = \exp\{V_i\mu_\beta^q + X_i^R\mu_{\tilde\alpha_i}^q + \frac{1}{2}\operatorname{diag}(V_i\Sigma_\beta^q V_i^T + X_i^R\Sigma_{\tilde\alpha_i}^q X_i^{R\,T})\}$. For Bernoulli responses with the logit link function [see (10)],

$$\begin{aligned}
S_{y_i} &= y_i^T(V_i\mu_\beta^q + X_i^R\mu_{\tilde\alpha_i}^q) \\
&\quad - \sum_{j=1}^{n_i}E_q\bigl[\log\{1 + \exp(V_{ij}^T\beta + X_{ij}^{R\,T}\tilde\alpha_i)\}\bigr],
\end{aligned}$$

where $E_q[\log\{1 + \exp(V_{ij}^T\beta + X_{ij}^{R\,T}\tilde\alpha_i)\}]$ is evaluated using adaptive Gauss–Hermite quadrature (see Appendix B). The variational lower bound is thus given by

$$\begin{aligned}
\mathcal{L} &= \sum_{i=1}^{n}S_{y_i} + \frac{1}{2}\sum_{i=1}^{n}\log|\Sigma_{\tilde\alpha_i}^q| \\
&\quad + \frac{1}{2}\log|\Sigma_\beta^{-1}\Sigma_\beta^q| - \frac{1}{2}\operatorname{tr}\bigl(\Sigma_\beta^{-1}\Sigma_\beta^q\bigr)
\end{aligned}$$

$$
\begin{aligned}
&- \frac{1}{2}\mu_\beta^{q\,T}\Sigma_\beta^{-1}\mu_\beta^q - \frac{\nu^q}{2}\log|S^q| \\
&+ \frac{\nu}{2}\log|S| - \sum_{l=1}^{r}\log\Gamma\Big(\frac{\nu^q+1-l}{2}\Big) \\
&+ \sum_{l=1}^{r}\log\Gamma\Big(\frac{\nu+1-l}{2}\Big) \\
&+ \frac{p+nr}{2} + \frac{nr}{2}\log 2.
\end{aligned}
$$

Note that this expression is valid only after each of the parameter updates has been made in Algorithm 3.

APPENDIX B: GAUSS–HERMITE QUADRATURE FOR LOGISTIC MIXED MODELS

We want to evaluate $E_q\{b(V_{ij}^T\beta + X_{ij}^{R\,T}\tilde{\alpha}_i)\}$ where $b(x) = \log(1+\mathrm{e}^x)$ for each $i = 1,\dots,n$ and $j = 1,\dots,n_i$. Let $\mu_{ij} = V_{ij}^T\mu_\beta^q + X_{ij}^{R\,T}\mu_{\tilde{\alpha}_i}^q$ and $\sigma_{ij}^2 = V_{ij}^T\Sigma_\beta^q V_{ij} + X_{ij}^{R\,T}\Sigma_{\tilde{\alpha}_i}^q X_{ij}^R$. Following Ormerod and Wand (2012), we reduce $E_q\{b(V_{ij}^T\beta + X_{ij}^{R\,T}\tilde{\alpha}_i)\}$ to a univariate integral such that

$$
\begin{aligned}
&E_q\{b(V_{ij}^T\beta + X_{ij}^{R\,T}\tilde{\alpha}_i)\} \\
&\quad = \int_{-\infty}^{\infty} b(\sigma_{ij}x + \mu_{ij})\phi(x;0,1)\,\mathrm{d}x,
\end{aligned}
$$

where $\phi(x;\mu,\sigma)$ denotes the Gaussian density for a random variable x with mean μ and standard deviation σ. Let $B^{(r)}(\mu,\sigma) = \int_{-\infty}^{\infty} b^{(r)}(\sigma x+\mu)\phi(x;0,1)\,\mathrm{d}x$ where $b^{(r)}(x)$ denotes the rth derivative of $b(\cdot)$ with respect to x. If μ and σ are vectors, say,

$$
\mu = \begin{bmatrix}1\\2\\3\end{bmatrix} \quad \text{and} \quad \sigma = \begin{bmatrix}4\\5\\6\end{bmatrix},
$$

then

$$
B^{(r)}(\mu,\sigma) = \begin{bmatrix}B^{(r)}(1,4)\\B^{(r)}(2,5)\\B^{(r)}(3,6)\end{bmatrix}.
$$

For each cluster i, let $\mu_i = (\mu_{i1},\dots,\mu_{in_i})^T = V_i\mu_\beta^q + X_i^R\mu_{\tilde{\alpha}_i}^q$ and

$$
\begin{aligned}
\sigma_i &= (\sigma_{i1},\dots,\sigma_{in_i})^T \\
&= \sqrt{\operatorname{diag}\big(V_i\Sigma_\beta^q V_i^T + X_i^R\Sigma_{\tilde{\alpha}_i}^q X_i^{R\,T}\big)}.
\end{aligned}
$$

We evaluate $B^{(r)}(\mu_{ij},\sigma_{ij})$ using adaptive Gauss–Hermite quadrature (Liu and Pierce, 1994) for each $i = 1,\dots,n$, $j = 1,\dots,n_i$ and $r = 0,1,2$. Ormerod and Wand (2012) have considered a similar approach. In Gauss–Hermite quadrature, integrals of the form $\int_{-\infty}^{\infty} f(x)\mathrm{e}^{-x^2}\,\mathrm{d}x$ are approximated by $\sum_{k=1}^m w_k f(x_k)$, where m is the number of quadrature points, the nodes x_i are zeros of the mth order Hermite polynomial and w_i are suitably corresponding weights. This approximation is exact for polynomials of degree $2m-1$ or less. For low-order quadrature to be effective, some transformation is usually required so that the integrand is sampled in a suitable range. Following the procedure recommended by Liu and Pierce (1994), we rewrite $B^{(r)}(\mu_{ij},\sigma_{ij})$ as

$$
\begin{aligned}
&B^{(r)}(\mu_{ij},\sigma_{ij}) \\
&\quad = \int_{-\infty}^{\infty} \frac{b^{(r)}(\sigma_{ij}x+\mu_{ij})\phi(x;0,1)}{\phi(x;\hat{\mu}_{ij},\hat{\sigma}_{ij})}\phi(x;\hat{\mu}_{ij},\hat{\sigma}_{ij})\,\mathrm{d}x \\
&\quad = \sqrt{2}\hat{\sigma}_{ij}\int_{-\infty}^{\infty}\big[\mathrm{e}^{x^2}b^{(r)}\big(\sigma_{ij}(\hat{\mu}_{ij}+\sqrt{2}\hat{\sigma}_{ij}x)+\mu_{ij}\big) \\
&\qquad\qquad \cdot\phi(\hat{\mu}_{ij}+\sqrt{2}\hat{\sigma}_{ij}x;0,1)\big] \\
&\qquad\qquad \cdot\mathrm{e}^{-x^2}\,\mathrm{d}x,
\end{aligned}
$$

which can be approximated using Gauss–Hermite quadrature by

$$
\begin{aligned}
&B^{(r)}(\mu_{ij},\sigma_{ij}) \\
&\quad \approx \sqrt{2}\hat{\sigma}_{ij}\sum_{k=1}^{m} w_k\mathrm{e}^{x_k^2}b^{(r)}\big(\sigma_{ij}(\hat{\mu}_{ij}+\sqrt{2}\hat{\sigma}_{ij}x_k)+\mu_{ij}\big) \\
&\qquad\qquad \cdot\phi(\hat{\mu}_{ij}+\sqrt{2}\hat{\sigma}_{ij}x_k;0,1).
\end{aligned}
$$

For the integrand to be sampled in an appropriate region, we take $\hat{\mu}_{ij}$ to be the mode of the integrand and $\hat{\sigma}_{ij}$ to be the standard deviation of the normal density approximating the integrand at the mode, so that

$$
\begin{aligned}
\hat{\mu}_{ij} &= \arg\max_x\{b^{(r)}(\sigma_{ij}x+\mu_{ij})\phi(x;0,1)\}, \\
\hat{\sigma}_{ij} &= \Big[-\frac{\mathrm{d}^2}{\mathrm{d}x^2}\log\{b^{(r)}(\sigma_{ij}x+\mu_{ij}) \\
&\qquad\qquad \cdot\phi(x;0,1)\}\Big|_{x=\hat{\mu}_{ij}}\Big]^{-1/2}
\end{aligned}
$$

for $j = 1,\dots,n_i$ and $i = 1,\dots,n$. For computational efficiency, we evaluate $\hat{\mu}_{ij}$ and $\hat{\sigma}_{ij}$, $i = 1,\dots,n$, $j = 1,\dots,n_i$, for the case $r = 1$ only once in each cycle of updates and use these values for $r = 0,2$. No significant loss of accuracy was observed in doing this. We implement adaptive Gauss–Hermite quadrature in `R` using the `R` package `fastGHQuad` (Blocker,

2011). The quadrature nodes and weights can be obtained via the function `gaussHermiteData()` and the function `aghQuad()` approximates integrals using the method of Liu and Pierce (1994). We used 10 quadrature points in all the examples.

ACKNOWLEDGMENTS

Linda S. L. Tan was partially supported as part of Singapore Delft Water Alliance's tropical reservoir research programme. We thank the Editors and referees for their constructive comments and suggestions which have helped to improve the content and clarity of this paper. We also thank Matt Wand for making available to us his preliminary work on fully simplified multivariate normal nonconjugate variational message passing updates and for his careful reading and comments on an earlier version of this paper.

REFERENCES

ATTIAS, H. (1999). Inferring parameters and structure of latent variable models by variational Bayes. In *Proceedings of the* 15*th Conference on Uncertainty in Artificial Intelligence* 21–30. Morgan Kaufmann, San Francisco, CA.

ATTIAS, H. (2000). A variational Bayesian framework for graphical models. In *Advances in Neural Information Processing Systems* **12** 209–215. MIT Press, Cambridge, MA.

BISHOP, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York. MR2247587

BLOCKER, A. W. (2011). Fast Rcpp implementation of Gauss–Hermite quadrature. R package "fastGHQuad" version 0.1-1. Available at http://cran.r-project.org/.

BRAUN, M. and MCAULIFFE, J. (2010). Variational inference for large-scale models of discrete choice. *J. Amer. Statist. Assoc.* **105** 324–335. MR2757203

BRESLOW, N. E. and CLAYTON, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88** 9–25.

BROWN, P. and ZHOU, L. (2010). MCMC for generalized linear mixed models with glmmBUGS. *The R Journal* **2** 13–16.

BROWNE, W. J. and DRAPER, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Anal.* **1** 473–513 (electronic). MR2221283

CAI, B. and DUNSON, D. B. (2008). Bayesian variable selection in generalized linear mixed models. In *Random Effect and Latent Variable Model Selection. Lecture Notes in Statistics* **192** 63–91. Springer, New York.

CHRISTENSEN, O. F., ROBERTS, G. O. and SKÖLD, M. (2006). Robust Markov chain Monte Carlo methods for spatial generalized linear mixed models. *J. Comput. Graph. Statist.* **15** 1–17. MR2269360

CORDUNEANU, A. and BISHOP, C. M. (2001). Variational Bayesian model selection for mixture distributions. In *Artificial Intelligence and Statistics* 27–34. Morgan Kaufmann, San Francisco, CA.

DE BACKER, M., DE VROEY, C., LESAFFRE, E., SCHEYS, I. and DE KEYSER, P. (1998). Twelve weeks of continuous oral therapy for toenail onychomycosis caused by dermatophytes: A double-blind comparative trial of terbinafine 250 mg/day versus itraconazole 200 mg/day. *Journal of the American Academy of Dermatology* **38** 57–63.

FITZMAURICE, G. and LAIRD, N. (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika* **80** 141–151.

FONG, Y., RUE, H. and WAKEFIELD, J. (2010). Bayesian inference for generalised linear mixed models. *Biostatistics* **11** 397–412.

GELFAND, A. E., SAHU, S. K. and CARLIN, B. P. (1995). Efficient parameterisations for normal linear mixed models. *Biometrika* **82** 479–488. MR1366275

GELFAND, A. E., SAHU, S. K. and CARLIN, B. P. (1996). Efficient parametrizations for generalized linear mixed models. In *Bayesian Statistics* 5 (*Alicante*, 1994) 165–180. Oxford Univ. Press, New York. MR1425405

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2004). *Bayesian Data Analysis*, 2nd ed. Chapman & Hall/CRC, Boca Raton, FL. MR2027492

GHAHRAMANI, Z. and BEAL, M. J. (2001). Propagation algorithms for variational Bayesian learning. In *Advances in Neural Information Processing Systems* **13** 507–513. MIT Press, Cambridge, MA.

HOFFMAN, M. D., BLEI, D. M., WANG, C. and PAISLEY, J. (2012). Stochastic variational inference. Available at arXiv:1206.7051.

JAAKKOLA, T. S. and JORDAN, M. I. (2000). Bayesian parameter estimation via variational methods. *Statist. Comput.* **10** 25–37.

KASS, R. E. and NATARAJAN, R. (2006). A default conjugate prior for variance components in generalized linear mixed models (comment on article by Browne and Draper). *Bayesian Anal.* **1** 535–542 (electronic). MR2221285

KNOWLES, D. A. and MINKA, T. P. (2011). Non-conjugate variational message passing for multinomial and binary regression. In *Advances in Neural Information Processing Systems* **24** 1701–1709. Available at http://books.nips.cc/papers/files/nips24/NIPS2011_0962.pdf.

LIU, Q. and PIERCE, D. A. (1994). A note on Gauss–Hermite quadrature. *Biometrika* **81** 624–629. MR1311107

LIU, J. S. and WU, Y. N. (1999). Parameter expansion for data augmentation. *J. Amer. Statist. Assoc.* **94** 1264–1274. MR1731488

LUNN, D. J., THOMAS, A., BEST, N. and SPIEGELHALTER, D. (2000). WinBUGS—A Bayesian modelling framework: Concepts, structure, and extensibility. *Statist. Comput.* **10** 325–337.

MAGNUS, J. R. and NEUDECKER, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley, Chichester. MR0940471

MENG, X.-L. and VAN DYK, D. (1997). The EM algorithm—An old folk-song sung to a fast new tune (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **59** 511–567. MR1452025

MENG, X.-L. and VAN DYK, D. A. (1999). Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika* **86** 301–320. MR1705351

O'HAGAN, A. and FORSTER, J. (2004). *Kendall's Advanced Theory of Statistics V.* 2*B*: *Bayesian Inference*, 2nd ed. Arnold, London.

ORMEROD, J. T. and WAND, M. P. (2010). Explaining variational approximations. *Amer. Statist.* **64** 140–153. MR2757005

ORMEROD, J. T. and WAND, M. P. (2012). Gaussian variational approximate inference for generalized linear mixed models. *J. Comput. Graph. Statist.* **21** 2–17. MR2913353

OVERSTALL, A. M. and FORSTER, J. J. (2010). Default Bayesian model determination methods for generalised linear mixed models. *Comput. Statist. Data Anal.* **54** 3269–3288. MR2727751

PAPASPILIOPOULOS, O., ROBERTS, G. O. and SKÖLD, M. (2003). Non-centered parameterizations for hierarchical models and data augmentation. In *Bayesian Statistics* 7 (*Tenerife*, 2002) 307–326. Oxford Univ. Press, New York. MR2003180

PAPASPILIOPOULOS, O., ROBERTS, G. O. and SKÖLD, M. (2007). A general framework for the parametrization of hierarchical models. *Statist. Sci.* **22** 59–73. MR2408661

QI, Y. and JAAKKOLA, T. S. (2006). Parameter expanded variational Bayesian methods. In *Advances in Neural Information Processing Systems* **19** 1097–1104. MIT Press, Cambridge, MA.

RAUDENBUSH, S. W., YANG, M.-L. and YOSEF, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *J. Comput. Graph. Statist.* **9** 141–157. MR1826278

RIJMEN, F. and VOMLEL, J. (2008). Assessing the performance of variational methods for mixed logistic regression models. *J. Stat. Comput. Simul.* **78** 765–779. MR2528235

ROOS, M. and HELD, L. (2011). Sensitivity analysis in Bayesian generalized linear mixed models for binary data. *Bayesian Anal.* **6** 259–278. MR2806244

ROULIN, A. and BERSIER, L. F. (2007). Nestling barn owls beg more intensely in the presence of their mother than in the presence of their father. *Animal Behaviour* **74** 1099–1106.

SAUL, L. K. and JORDAN (1998). A mean field learning algorithm for unsupervised neural networks. In *Learning in Graphical Models* 541–554. Kluwer Academic, Dordrecht.

STURTZ, S., LIGGES, U. and GELMAN, A. (2005). R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software* **12** 1–16.

TAN, S. L. and NOTT, D. J. (2013). Variational approximation for mixtures of linear mixed models. *J. Comput. Graph. Statist.* To appear. DOI:10.1080/10618600.2012.761138.

THALL, P. F. and VAIL, S. C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics* **46** 657–671. MR1085814

VENABLES, W. N. and RIPLEY, B. D. (2002). *Modern Applied Statistics with S*, 4th ed. Springer, New York.

WAND, M. P. (2013). Fully simplified multivariate normal updates in non-conjugate variational message passing. Unpublished manuscript. Available at http://www.uow.edu.au/~mwand/fsupap.pdf.

WINN, J. and BISHOP, C. M. (2005). Variational message passing. *J. Mach. Learn. Res.* **6** 661–694. MR2249835

YU, Y. and MENG, X.-L. (2011). To center or not to center: That is not the question—An ancillarity–sufficiency interweaving strategy (ASIS) for boosting MCMC efficiency. *J. Comput. Graph. Statist.* **20** 531–570. MR2878987

YU, D. and YAU, K. K. W. (2012). Conditional Akaike information criterion for generalized linear mixed models. *Comput. Statist. Data Anal.* **56** 629–644. MR2853760

ZHAO, Y., STAUDENMAYER, J., COULL, B. A. and WAND, M. P. (2006). General design Bayesian generalized linear mixed models. *Statist. Sci.* **21** 35–51. MR2275966

ZUUR, A. F., IENO, E. N., WALKER, N. J., SAVELIEV, A. A. and SMITH, G. M. (2009). *Mixed Effects Models and Extensions in Ecology with R*. Springer, New York. MR2722501