Bayesian Analysis (2014) **9**, Number 4, pp. 963–1004

A Stochastic Variational Framework for Fitting and Diagnosing Generalized Linear Mixed Models

Linda S. L. Tan * and David J. Nott †

Abstract. In stochastic variational inference, the variational Bayes objective function is optimized using stochastic gradient approximation, where gradients computed on small random subsets of data are used to approximate the true gradient over the whole data set. This enables complex models to be fit to large data sets as data can be processed in mini-batches. In this article, we extend stochastic variational inference for conjugate-exponential models to nonconjugate models and present a stochastic nonconjugate variational message passing algorithm for fitting generalized linear mixed models that is scalable to large data sets. In addition, we show that diagnostics for prior-likelihood conflict, which are useful for Bayesian model criticism, can be obtained from nonconjugate variational message passing automatically, as an alternative to simulation-based Markov chain Monte Carlo methods. Finally, we demonstrate that for moderate-sized data sets, convergence can be accelerated by using the stochastic version of nonconjugate variational message passing in the initial stage of optimization before switching to the standard version.

Keywords: Variational Bayes, stochastic approximation, nonconjugate variational message passing, conflict diagnostics, hierarchical models, identify divergent units

1 Introduction

Generalized linear mixed models (GLMMs) extend generalized linear models (GLMs) by introducing random effects to account for within-subject association and have wide applications. Estimation of GLMMs using maximum likelihood is, however, challenging as the integrals over random effects are intractable and have to be approximated using computationally intensive methods such as numerical quadrature or Markov chain Monte Carlo (MCMC). Various approximate methods for fitting GLMMs have been proposed, such as penalized quasi-likelihood (Breslow *et al.* 1993), Laplace approximation and its extensions (Raudenbush *et al.* 2000), Gaussian variational approximation (Ormerod and Wand 2012) and integrated nested Laplace approximations (Fong *et al.* 2010). Stochastic approximation has also been used in conjunction with the expectation maximization (EM) algorithm (Jank 2006) and MCMC (Zhu *et al.* 2002) to fit GLMMs.

Recently, Tan and Nott (2013) demonstrated how GLMMs can be fitted using varia-

*Department of Statistics and Applied Probability, National University of Singapore, statsll@nus.edu.sg

†Department of Statistics and Applied Probability, National University of Singapore, standj@nus.edu.sg

 DOI:10.1214/14-BA885

tional Bayes (VB, Attias 1999) via an algorithm called nonconjugate variational message passing (Knowles and Minka 2011). A popular method of approximation, VB is deterministic and requires much less computation time than MCMC methods. In VB, the intractable true posterior is approximated by a factorized distribution, which is optimized to be close to the true posterior in terms of Kullback-Leibler divergence. Variational message passing (Winn and Bishop 2005) is an algorithmic implementation of VB for conjugate-exponential models (Ghahramani and Beal 2001). Knowles and Minka (2011) extended variational message passing to nonconjugate models by assuming that the factors in VB belong to some exponential family.

The nonconjugate variational message passing algorithm for GLMMs (Tan and Nott 2013) has to update local variational parameters associated with every unit before re-estimating the global variational parameters at each iteration. This algorithm is inefficient for large data sets and is unsuitable for streaming data as it can never complete one iteration. To address these issues, Hoffman *et al.* (2013) proposed optimizing the VB objective function using stochastic gradient approximation (Robbins and Monro 1951), where gradients computed on small random subsets of data are used to approximate the true gradient over the whole data set. This approach reduces the computational cost for large data sets significantly (Bottou and Cun 2005). Hoffman *et al.* (2013) focused on developing stochastic variational inference for conjugate-exponential models.

In this article, we extend stochastic variational inference to nonconjugate models and develop a stochastic nonconjugate variational message passing algorithm for fitting GLMMs that is scalable to large data sets. A strong motivation for developing stochastic gradient optimization algorithms is their efficiency in terms of memory. As data are processed in mini-batches, analysis of data sets too large to fit into memory can still be contemplated. We focus on Poisson and logistic GLMMs, and applications in longitudinal data analysis. Our paper makes three contributions. First, we show how updates in nonconjugate variational message passing can be used in stochastic natural gradient optimization of the variational lower bound. Second, we show that variational message passing facilitates an automatic computation of diagnostics for prior-likelihood conflict (useful for Bayesian model criticism) and provides an attractive alternative to simulation-based MCMC methods. Third, we demonstrate that for moderate-sized data sets, convergence can be accelerated by using the stochastic version of nonconjugate variational message passing in the initial stage of optimization before switching to the standard version.

Recently, there is increasing interest in developing VB algorithms capable of handling large data sets or streaming data (e.g. Luts *et al.* 2013; Broderick *et al.* 2013). Stochastic optimization is an important tool in parameter estimation for large data sets (e.g. Bottou and Bousquet 2008; Liang *et al.* 2013) and has been considered in the context of VB. For example, the online VB algorithms for latent Dirichlet allocation (Hoffman *et al.* 2010) and the hierarchical Dirichlet process (Wang *et al.* 2011) are based on stochastic natural gradient optimization of the VB objective function, with data processed one at a time or in mini-batches. Hoffman *et al.* (2013) generalized these methods to derive stochastic variational inference for conjugate-exponential family models. Stochastic approximation methods have also been considered by Ji *et al.*

(2010), Nott *et al.* (2012) and Paisley *et al.* (2012) for optimization of VB objective functions containing intractable integrals. Salimans and Knowles (2013) proposed a stochastic approximation algorithm that does not require analytic evaluation of integrals and allows fixed-form VB to be applied to any posterior available in closed form up to the proportionality constant. Nott *et al.* (2013) consider the approach of Salimans and Knowles (2013) for fitting GLMMs, and analyze large data sets by combining variational approximations learned in parallel on smaller partitions. Random effects in each partition were treated as a single block. In this paper, we consider a different approach of fitting GLMMs to large data sets by using nonconjugate variational message passing within stochastic variational inference. Variational posteriors of random effects from different clusters are assumed to be independent and partial noncentering (Tan and Nott 2013) is used to improve posterior approximation. Global variational parameters are updated using stochastic gradient approximation based on mini-batches of optimized local variational parameters.

Model checking is an integral part of statistical analyses. In the Bayesian approach, assumptions are made about the sampling model and prior, and prior-likelihood conflict arises when observed data are very unlikely under the prior model. Evans and Moshonov (2006) discuss how to assess whether there is prior-data conflict and Scheel *et al.* (2011) proposed a graphical diagnostic, the local critique plot, for identifying influential statistical modelling choices at the node level. See also Scheel *et al.* (2011) for a review of other methods in Bayesian model criticism. Marshall and Spiegelhalter (2007) proposed a diagnostic test for identifying divergent units in hierarchical models, based on measuring the conflict between the likelihood of a parameter and its predictive prior given the remaining data. A simulation-based approach was adopted and diagnostic tests were carried out using MCMC. We show that the approach of Marshall and Spiegelhalter (2007) can be approximated in the variational message passing framework.

Section 2 introduces some notation. Section 3 specifies the model and motivates partial noncentering for GLMMs. The stochastic nonconjugate variational message passing algorithm is developed in Section 4. Section 5 describes how variational message passing facilitates computation of prior-likelihood conflict diagnostics. Section 6 considers examples including real and simulated data and Section 7 concludes.

2 Notation

We use 1_d to denote the $d \times 1$ column vector with all entries equal to 1 and I_d to denote the $d \times d$ identity matrix. Scalar functions such as $\exp(\cdot)$ applied to vector arguments are evaluated element by element. We use $\odot$ to denote element by element multiplication of two vectors. If a is a $d \times 1$ vector, we use $\text{diag}(a)$ to denote the $d \times d$ diagonal matrix with diagonal entries given by a. If A is a $d \times d$ square matrix, we use $\text{diag}(A)$ to denote the $d \times 1$ vector containing the diagonal entries of A.

3 Generalized linear mixed models

We consider one-parameter exponential family models which are specified as follows. Let y_{ij} denote the jth response in cluster i, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$. Conditional on a vector of length r of random effects u_i, independently distributed as $N(0, D)$, y_{ij} is independently distributed as

$$y_{ij}|u_i \sim \exp\left\{y_{ij}\zeta_{ij} - b(\zeta_{ij}) + c(y_{ij})\right\},$$

where ζ_{ij} is the canonical parameter and $b(\cdot)$ and $c(\cdot)$ are functions specific to the exponential family. The link function g relates the conditional mean of y_{ij}, $\mu_{ij} = E(y_{ij}|u_i)$ to the linear predictor $\eta_{ij} = x_{ij}^T\beta + z_{ij}^T u_i$ as $g(\mu_{ij}) = \eta_{ij}$. Here, x_{ij} and z_{ij} are $p \times 1$ and $r \times 1$ vectors of covariates and β is a $p \times 1$ vector of unknown fixed regression parameters. We consider responses from the Bernoulli and Poisson families. If $y_{ij} \sim$ Bernoulli(μ_{ij}), then $b(x) = \log\{1 + \exp(x)\}$, $c(x) = 0$ and $\text{logit}(\mu_{ij}) = \eta_{ij}$. For Poisson responses, we allow for an offset $\log E_{ij}$. If $y_{ij} \sim$ Poisson(μ_{ij}), then $b(x) = \exp(x)$, $c(x) = -\log(x!)$ and $\log \mu_{ij} = \log E_{ij} + \eta_{ij}$. For the ith cluster, let

$$y_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{in_i} \end{bmatrix}, \ \eta_i = \begin{bmatrix} \eta_{i1} \\ \ldots \\ \eta_{in_i} \end{bmatrix}, \ X_i = \begin{bmatrix} x_{i1}^T \\ \vdots \\ x_{in_i}^T \end{bmatrix}, \ Z_i = \begin{bmatrix} z_{i1}^T \\ \vdots \\ z_{in_i}^T \end{bmatrix} \text{ and } E_i = \begin{bmatrix} E_{i1} \\ \vdots \\ E_{in_i} \end{bmatrix},$$

We assume that the first column of Z_i is 1_{n_i} if Z_i is not a zero matrix and that the columns of Z_i are a subset of the columns of X_i.

For Bayesian inference, we specify a diffuse prior $N(0, \Sigma_\beta)$ on β where Σ_β is large and an independent inverse-Wishart prior, $IW(\nu, S)$ on D. We use the default conjugate prior proposed in Kass and Natarajan (2006), which is based on a prior guess for D determined from first-stage data variability. For this prior, $\nu = r$ and $S = r\hat{R}$ where

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

Here, $M_i(\hat{\beta})$ denotes the $n_i \times n_i$ diagonal GLM weight matrix with diagonal elements $[v(\hat{\mu}_{ij})\, g'(\hat{\mu}_{ij})^2]^{-1}$, where $v(\cdot)$ is the variance function and $g(\cdot)$ is the link function. We let $\hat{\mu}_{ij} = g^{-1}(x_{ij}^T\hat{\beta} + z_{ij}^T\hat{u}_i)$ where $\hat{u}_i$ is set as 0 for all i and $\hat{\beta}$ is an estimate of the regression coefficients from the GLM obtained by pooling all data and setting $u_i = 0$ for all i. The constant c is an inflation factor representing the amount in which within-cluster variability can be increased. We use $c = 1$ in all examples. Some heuristic justifications for $\hat{R}$ are given in Kass and Natarajan (2006). A similar prior was used in Overstall and Forster (2010). Alternatively, one may consider marginally noninformative priors for covariance matrices (Huang and Wand 2013). Methods in this paper can be extended to these priors easily.

3.1 A partially noncentered parametrization for the GLMM

Reparametrization techniques such as centering, noncentering and partial noncentering have been used in hierarchical models to boost efficiency in MCMC and EM algorithms (e.g. Gelfand *et al.* 1995, 1996; Papaspiliopoulos *et al.* 2003, 2007). Recently, Tan and Nott (2013) introduced a partially noncentered parametrization for GLMMs and studied its performance in VB. We introduce the idea of partial noncentering by considering the following linear mixed model (see Tan and Nott 2013). Suppose

$$y_i = X_i\beta + Z_i u_i + \epsilon_i \text{ where } \epsilon_i \sim N(0, \sigma^2) \text{ for } i = 1, \ldots, n, \tag{2}$$

and y_i, X_i, Z_i, u_i and β are as defined previously. Let us specify a constant prior on β and assume that σ^2 and D are known. Suppose $X_i = Z_i$. In this case, we can introduce $\alpha_i = \beta + u_i$ so that $\alpha_i \sim N(\beta, D)$ is "centered" about β. We can also obtain a partially noncentered parametrization by introducing $\tilde{\alpha}_i = \alpha_i - W_i\beta$, where W_i is an $r \times r$ tuning matrix to be specified. The proportion of β subtracted from α_i is allowed to vary with i as each y_i carries a different amount of information about the underlying α_i. The centered ($W_i = 0$) and noncentered ($W_i = I_r$) parametrizations are special cases of the partially noncentered parametrization. Rewriting (2) as

$$y_i = Z_i W_i \beta + Z_i \tilde{\alpha}_i + \epsilon_i,$$

we can apply VB to the reparametrized model and assume that $q(\beta, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_n) = q(\beta)\prod_{i=1}^n q(\tilde{\alpha}_i)$. Tan and Nott (2013) showed that the resulting VB algorithm converges in one iteration when

$$W_i = (Z_i^T Q_i Z_i + D^{-1})^{-1} D^{-1}, \tag{3}$$

where $Q_i = \frac{1}{\sigma^2} I_r$. This result implies that partial noncentering can yield more rapid convergence than centering or noncentering. More importantly, the true posteriors are recovered in (3) but not in the centered or noncentered parametrizations. Even though the assumption of a factorized posterior in VB tends to result in underestimation of posterior variance, partial noncentering was (in this case) able to capture dependence between fixed and random effects via tuning parameters W_i so that the true posterior can be recovered.

The above result is particularly useful in the context of stochastic variational inference for GLMMs. To implement stochastic variational inference, we need to assume that variational posteriors of random effects associated with each unit are independent of each other and of the global variables β and D. However, correlation between fixed and random effects is often strong and partial noncentering allows some of this dependence to be captured via the tuning matrices W_i. This leads to more accurate posterior approximations of the fixed and random effects. In particular, estimation of the posterior variance of fixed effects which can be centered is improved greatly. Partial noncentering can also give more rapid convergence than centering or noncentering. This is desirable in the analysis of large data sets and is particularly useful when the convergence of one of the centered or noncentered parametrizations is especially slow. We emphasize that it is not easy to tell beforehand which of centering or noncentering will perform better, and partial noncentering automatically chooses a parametrization close to optimal.

We adopt the partially noncentered parametrization introduced by Tan and Nott (2013) for the GLMM, which is explained below. First, partition X_i as $[Z_i \quad X_{si} \quad X_{gi}]$ and β as $[\beta_z^T, \beta_s^T, \beta_g^T]^T$ accordingly, where X_{si} is an $n_i \times s$ matrix consisting of "subject specific" covariates and X_{gi} is an $n_i \times g$ matrix consisting of "general" covariates (i.e. not subject specific). All the rows of X_{si} are the same and equal to say x_{si}^T. We have

$$\begin{aligned}\eta_i &= Z_i(\beta_z + u_i) + 1_{n_i} x_{si}^T \beta_s + X_{gi}\beta_g \\ &= Z_i(C_i\beta_c + u_i) + X_{gi}\beta_g, \quad \text{where } C_i = \begin{bmatrix} I_r & x_{si}^T \\ & 0_{(r-1)\times s} \end{bmatrix} \text{ and } \beta_c = \begin{bmatrix} \beta_z \\ \beta_s \end{bmatrix}.\end{aligned}$$

Note that C_i is an $r \times (r+s)$ matrix. We introduce

$$\alpha_i = C_i\beta_c + u_i \text{ and } \tilde{\alpha}_i = \alpha_i - W_i C_i \beta_c,$$

where W_i is an $r \times r$ tuning matrix. $W_i = 0$ corresponds to the centered and $W_i = I_r$ to the noncentered parametrization. Letting $\tilde{W}_i = [(I_r - W_i)C_i \quad 0_{r\times g}]$ be an $r \times p$ matrix, $\tilde{\alpha}_i \sim N(\tilde{W}_i\beta, D)$. The partially noncentered parametrization is thus

$$\eta_i = V_i\beta + Z_i\tilde{\alpha}_i,$$

where $V_i = [Z_i W_i C_i \quad X_{gi}]$ is an $n_i \times p$ matrix. Following Tan and Nott (2013), W_i can be specified as in (3) with $Q_i = \text{diag}\left(\frac{\exp(\eta_i)}{\{1+\exp(\eta_i)\}^2}\right)$ for logistic GLMMs and $Q_i = \text{diag}(E_i \odot \eta_i) \approx \text{diag}(y_i)$ for Poisson GLMMs.

Let $y = [y_1^T, \ldots, y_n^T]^T$ and $\tilde{\alpha} = [\tilde{\alpha}_1^T, \ldots, \tilde{\alpha}_n^T]^T$. The set of unknown parameters in the GLMM is $\theta = \{\beta, D, \tilde{\alpha}\}$ and

$$p(y, \theta) = \left\{\prod_{i=1}^{n} p(y_i|\beta, \tilde{\alpha}_i) p(\tilde{\alpha}_i|\beta, D)\right\} p(\beta|\Sigma_\beta) p(D|\nu, S). \tag{4}$$

The fixed effects β and random effects covariance D can be regarded as "global" variables which are common across clusters, while the partially noncentered random effects $\tilde{\alpha}_i$ can be thought of as "local" variables associated only with the individual units.

4 Stochastic variational inference for generalized linear mixed models

In this section, we derive and present the stochastic nonconjugate variational message passing algorithm for fitting GLMMs, scalable to large data sets. We start with a brief introduction to variational approximation methods (see, e.g. Ormerod and Wand 2010) and review of nonconjugate variational message passing (Knowles and Minka 2011).

In variational approximation, the true posterior $p(\theta|y)$ is approximated by a more tractable distribution $q(\theta|\lambda)$, where λ denotes the set of parameters of q. We attempt

to make $q(\theta|\lambda)$ a good approximation to $p(\theta|y)$ by minimizing the Kullback-Leibler divergence between $q(\theta|\lambda)$ and $p(\theta|y)$. This is given by

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

where $p(y) = \int p(y,\theta)\, d\theta$ is the marginal likelihood. As the Kullback-Leibler divergence is nonnegative, we have

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

where E_q denotes expectation with respect to $q(\theta|\lambda)$ and $\mathcal{L}$ is a lower bound on the log marginal likelihood. Maximization of $\mathcal{L}$ is thus equivalent to minimization of the Kullback-Leibler divergence between $q(\theta|\lambda)$ and $p(\theta|y)$. In some cases, $\mathcal{L}$ is used as an approximation to the log marginal likelihood for performing model selection. Tan and Nott (2013) illustrate how $\mathcal{L}$ can be used for model selection in GLMMs.

4.1 Nonconjugate variational message passing

In VB, $q(\theta|\lambda)$ is assumed to factorize into $\prod_{l=1}^{m} q_l(\theta_l|\lambda_l)$ for some partition $\{\theta_1, \ldots, \theta_m\}$ of θ and λ_l denotes variational parameters associated with each factor. Optimization of $\mathcal{L}$ with respect to $q_1, \ldots, q_m$ leads to optimal densities satisfying

$$q_l(\theta_l) \propto \exp[E_{-\theta_l}\{\log p(y,\theta)\}], \tag{5}$$

where $E_{-\theta_l}$ denotes expectation with respect to $\prod_{j \neq l} q_j(\theta_j|\lambda_j)$. When conjugate priors are used, the optimal densities have the same form as the priors and it suffices to update the parameters of q_l. However, for non-conjugate priors, the optimal densities may not belong to recognizable density families. To address this issue, Knowles and Minka (2011) imposed a further restriction that each q_l must belong to some exponential family. Let

$$q_l(\theta_l|\lambda_l) = \exp\{\lambda_l^T t_l(\theta_l) - h_l(\lambda_l)\},$$

where λ_l is the vector of natural parameters and $t_l(\cdot)$ are the sufficient statistics. Updates in nonconjugate variational message passing can be derived by maximizing $\mathcal{L}$ with respect to each λ_l and setting $\nabla_{\lambda_l}\mathcal{L} = 0$. Let $\mathcal{V}_l(\lambda_l)$ denote the covariance matrix of $t_l(\theta_l)$. It can be shown that

$$\nabla_{\lambda_l}\mathcal{L} = \nabla_{\lambda_l} E_q\{\log p(y,\theta)\} - \mathcal{V}_l(\lambda_l)\lambda_l. \tag{6}$$

Updates in nonconjugate variational message passing are thus given by

$$\lambda_l \leftarrow \mathcal{V}(\lambda_l)^{-1} \nabla_{\lambda_l} E_q\{\log p(y,\theta)\} \tag{7}$$

for $l = 1, \ldots, m$. As nonconjugate variational message passing is a type of fixed-point iterations algorithm, the lower bound is not guaranteed to increase after each update.

Sometimes, convergence issues may be encountered which may require damping to fix (see Knowles and Minka 2011). For conjugate factors, the update in (7) can be simplified and details are given in Appendix A.

The nonconjugate variational message passing algorithm for GLMMs (Tan and Nott 2013) considers a variational approximation of the form

$$q(\theta|\lambda) = q(\beta|\lambda_\beta)q(D|\lambda_D)\prod_{i=1}^{n} q(\tilde{\alpha}_i|\lambda_{\tilde{\alpha}_i}), \tag{8}$$

where

$$q(\beta|\lambda_\beta) = N(\mu_{q(\beta)}, \Sigma_{q(\beta)}),\ q(D|\lambda_D) = IW(\nu_{q(D)}, S_{q(D)}),\ q(\tilde{\alpha}_i|\lambda_{\tilde{\alpha}_i}) = N(\mu_{q(\tilde{\alpha}_i)}, \Sigma_{q(\tilde{\alpha}_i)})$$

and λ_β, λ_D, $\lambda_{\tilde{\alpha}_i}$ are the respective natural parameter vectors. For Bernoulli or Poisson responses, $p(y_i|\beta, \tilde{\alpha}_i)$ is nonconjugate with respect to the priors over β and $\tilde{\alpha}_i$. Applying nonconjugate variational message passing and approximating the posteriors of β and $\tilde{\alpha}_i$ by Gaussian distributions, parameter updates for $q(\beta)$ and $q(\tilde{\alpha}_i)$ can be derived using (7). The variational posterior for D is optimal under (8) and parameter updates can be derived using (5). The main steps are given in Algorithm 1 below.

Initialize $\mu_{q(\beta)}$, $\Sigma_{q(\beta)}$, $\nu_{q(D)}$, $S_{q(D)}$, $\mu_{q(\tilde{\alpha}_i)}$, $\Sigma_{q(\tilde{\alpha}_i)}$ and tuning parameters W_i for $i = 1, \ldots, n$.

Cycle:

1. Update local variational parameters $\mu_{q(\tilde{\alpha}_i)}$ and $\Sigma_{q(\tilde{\alpha}_i)}$ for $i = 1, \ldots, n$.
2. Update global variational parameters $\mu_{q(\beta)}$, $\Sigma_{q(\beta)}$, $\nu_{q(D)}$ and $S_{q(D)}$.

until the lower bound converges.

Algorithm 1: Nonconjugate variational message passing for GLMMs.

Algorithm 1 iterates repeatedly between updating local variational parameters for each unit $i = 1, \ldots, n$, and re-estimating the global variational parameters. This procedure is inefficient for large data sets and impossible to accomplish for streaming data or data sets too massive to fit into memory. Using ideas in stochastic variational inference (Hoffman *et al.* 2013), we develop a stochastic nonconjugate variational message passing algorithm for fitting GLMMs that is more efficient at handling large data.

4.2 Natural gradient of the variational lower bound

In stochastic variational inference, the global variational parameters are optimized using stochastic gradient ascent. Updates of the form

$$\lambda^{(t+1)} = \lambda^{(t)} + a_t \nabla_\lambda \mathcal{L}(\lambda^{(t)})$$

are considered, where a_t denotes a small step taken in the direction of steepest ascent at the tth iteration. Under the Euclidean metric, the direction of steepest ascent is given by the regular gradient $\nabla_\lambda \mathcal{L}(\lambda^{(t)})$. In stochastic gradient ascent, a noisy estimate of $\nabla_\lambda \mathcal{L}(\lambda^{(t)})$ is used in its place. Hoffman *et al.* (2013) propose using natural gradients instead of regular gradients in this optimization step. Their motivation is that the Euclidean distance between two parameter settings λ and λ' is often a poor measure of how dissimilar two distributions $q(\theta|\lambda)$ and $q(\theta|\lambda')$ are. A more intuitive measure of dissimilarity between two probability distributions is given by the symmetrized Kullback-Leibler divergence, which is invariant to parameter transformations. Under this measure, Hoffman *et al.* (2013) showed that the direction of steepest ascent is given by the natural gradient (Amari 1998). Previously, Honkela *et al.* (2008) also showed that replacing regular gradients with natural gradients in the conjugate gradient algorithm can speed up variational learning.

Following Hoffman *et al.* (2013), we use natural gradients instead of regular gradients in the stochastic approximation. To obtain the natural gradient of $\mathcal{L}$ with respect to λ_l, we premultiply $\nabla_{\lambda_l}\mathcal{L}$ with the inverse of the Fisher information matrix of $q_l(\theta_l|\lambda_l)$ (see, e.g. Amari 1998). In nonconjugate variational message passing, the Fisher information matrix is given by

$$E_q\left[\{\nabla_{\lambda_l}\log q_l(\theta_l|\lambda_l)\}\{\nabla_{\lambda_l}\log q_l(\theta_l|\lambda_l)\}^T\right] = E_q\left[\{t_l(\theta_l)-\nabla_{\lambda_l}h_l(\lambda_l)\}\{t_l(\theta_l)-\nabla_{\lambda_l}h_l(\lambda_l)\}^T\right] = \mathcal{V}_l(\lambda_l).$$

From (6), the natural gradient denoted by $\widetilde{\nabla}_{\lambda_l}\mathcal{L}$ is thus given by

$$\widetilde{\nabla}_{\lambda_l}\mathcal{L} = \mathcal{V}_l(\lambda_l)^{-1}\nabla_{\lambda_l}E_q\{\log p(y,\theta)\} - \lambda_l. \tag{9}$$

4.3 Stochastic variational inference

In this section, we review the key ideas in stochastic variational inference (Hoffman *et al.* 2013) and discuss how they can be extended to nonconjugate models via nonconjugate variational message passing. The following steps are carried out in each iteration of stochastic variational inference until convergence is reached.

1. Randomly select a mini-batch B of $|B| \geq 1$ units from the whole data set.
2. Optimize local variational parameters of units in mini-batch B (as a function of the global variational parameters at their current setting).
3. Update global variational parameters using stochastic natural gradient ascent. Noisy gradients are computed based on optimized local variational parameters of units in mini-batch B.

The main difficulty in extending stochastic variational inference to nonconjugate models lies in step 2. For conjugate models, the local variational parameters can be

optimized as a function of the global variational parameters in a single update (see Appendix A) but the same is not true for nonconjugate models. In nonconjugate variational message passing, the update equation for the local variational parameters is recursive (they depend on the current setting of the local variational parameters) and has to be applied repeatedly until convergence is reached [see (7)]. This incurs a higher computational cost. We have tried performing the update for local variational parameters only once but this further slowed down convergence of the global variational parameters. We have also tried using a loose criterion for assessing convergence. This approach yielded much better results. Choosing a good initialization is also important as convergence problems can be encountered in recursive updates if the starting point is poor.

The other main difference is that for conjugate models, the update equations and natural gradients are easier to compute as the Fisher information matrix $\mathcal{V}_i(\lambda_i)$ does not have to be evaluated (see Appendix A). Fortunately, nonconjugate variational message passing updates can be simplified considerably when the variational posteriors are multivariate Gaussian (see Wand 2013) and the Fisher information matrix does not have to be computed explicitly as well.

The extension of stochastic variational inference to nonconjugate models greatly widens the scope of models to which stochastic variational inference can be applied. We think that nonconjugate variational message passing is an important tool in facilitating this extension as it allows for efficient closed-form updates in some cases (e.g. Poisson GLMMs) and there is a lot of flexibility in the evaluation of expectations (using bounds or quadrature). While convergence issues remain in fixed-point iterations algorithms, these can usually be mitigated by good initializations. We later show that nonconjugate variational message passing, like VB, is a type of natural gradient method (see Sato 2001). With this interpretation, some convergence issues might be resolved by taking adaptive steps in the direction of the natural gradient.

Let λ_{global} and λ_{local} denote the global and local variational parameters respectively. The lower bound $\mathcal{L}$ is a function of $\lambda = (\lambda_{\text{global}}, \lambda_{\text{local}})$, i.e. $\mathcal{L} = \mathcal{L}(\lambda) = \mathcal{L}(\lambda_{\text{global}}, \lambda_{\text{local}})$. Hoffman *et al.* (2013) showed that to find a setting of λ_{global} that maximizes $\mathcal{L}$ using stochastic natural gradient ascent, we can first optimize λ_{local} as a function of λ_{global} so that $\lambda_{\text{local}} = k(\lambda_{\text{global}})$ for some function k. In nonconjugate variational message passing, this is done by computing the update in (7) repeatedly until convergence, starting with some current setting of λ_{local} and keeping λ_{global} fixed. This implies that $\nabla_k \mathcal{L}(\lambda_{\text{global}}, k(\lambda_{\text{global}})) = 0$ since $k(\lambda_{\text{global}})$ is a local optimum of the local variational parameters. The current value of the lower bound is $\mathcal{L}(\lambda_{\text{global}}, k(\lambda_{\text{global}}))$ which is a function of λ_{global} only. Let us define $\mathcal{L}(\lambda_{\text{global}}) = \mathcal{L}(\lambda_{\text{global}}, k(\lambda_{\text{global}}))$. To optimize $\mathcal{L}(\lambda_{\text{global}})$ with respect to λ_{global}, we have

$$\begin{aligned}
&\nabla_{\lambda_{\text{global}}} \mathcal{L}(\lambda_{\text{global}}) \\
&\quad = \nabla_{\lambda_{\text{global}}} \mathcal{L}(\lambda_{\text{global}}, k(\lambda_{\text{global}})) + \{\nabla_{\lambda_{\text{global}}} k(\lambda_{\text{global}})\}^T \nabla_k \mathcal{L}(\lambda_{\text{global}}, k(\lambda_{\text{global}})) \\
&\qquad = \nabla_{\lambda_{\text{global}}} \mathcal{L}(\lambda_{\text{global}}, k(\lambda_{\text{global}})).
\end{aligned}$$

Therefore, $\nabla_{\lambda_{\text{global}}} \mathcal{L}(\lambda_{\text{global}})$ can be computed by finding the optimized local variational parameters $k(\lambda_{\text{global}})$ and then computing the gradient of $\mathcal{L}(\lambda_{\text{global}}, k(\lambda_{\text{global}}))$ with

respect to λ_{global} by keeping $k(\lambda_{\text{global}})$ fixed. The corresponding natural gradient can be obtained as discussed in Section 4.2.

In stochastic variational inference, noisy estimates of the natural gradients are used in stochastic optimization of the global variational parameters. The idea is to approximate true gradients over the whole data with gradients computed on mini-batches of data. For large data sets, this can lead to significant reductions in computation time. For the GLMM, $\lambda_{\text{global}} = (\lambda_\beta, \lambda_D)$ and $\lambda_{\text{local}} = (\lambda_{\tilde{\alpha}_1}, \ldots, \lambda_{\tilde{\alpha}_n})$. As β and D are independent in the variational posterior, stochastic gradient ascent for λ_β and λ_D can be done separately. From (4) and (9), the natural gradient of $\mathcal{L}$ with respect to λ_β, $\widetilde{\nabla}_{\lambda_\beta}\mathcal{L}$ is given by

$$\mathcal{V}_\beta(\lambda_\beta)^{-1}\nabla_{\lambda_\beta}\left\{\sum_{i=1}^{n} E_q\{\log p(y_i|\beta,\tilde{\alpha}_i)+\log p(\tilde{\alpha}_i|\beta,D)\}|_{\lambda_{\tilde{\alpha}_i}=\lambda^{\text{opt}}_{\tilde{\alpha}_i}} + E_q\{\log p(\beta|\Sigma_\beta)\}\right\} - \lambda_\beta, \tag{10}$$

where $\lambda^{\text{opt}}_{\tilde{\alpha}_i}$ denotes $\lambda_{\tilde{\alpha}_i}$ optimized as a function of the global variational parameters. If B is a mini-batch of $|B|$ units randomly sampled from the whole data set (with or without replacement), then an unbiased estimate of $\widetilde{\nabla}_{\lambda_\beta}\mathcal{L}$ is $\hat{\lambda}_\beta - \lambda_\beta$, where $\hat{\lambda}_\beta$ is

$$\mathcal{V}_\beta(\lambda_\beta)^{-1}\nabla_{\lambda_\beta}\left\{\frac{n}{|B|}\sum_{i\in B} E_q\{\log p(y_i|\beta,\tilde{\alpha}_i)+\log p(\tilde{\alpha}_i|\beta,D)\}|_{\lambda_{\tilde{\alpha}_i}=\lambda^{\text{opt}}_{\tilde{\alpha}_i}} + E_q\{\log p(\beta|\Sigma_\beta)\}\right\}.$$

Note that each of the n units in the whole data set has a probability $\frac{|B|}{n}$ of being selected and hence the expectation of $\hat{\lambda}_\beta - \lambda_\beta$ is equal to (10) (Hoffman *et al.* 2013, pg. 18 – 19). Similarly, an unbiased estimate of $\widetilde{\nabla}_{\lambda_D}\mathcal{L}$ is $\hat{\lambda}_D - \lambda_D$, where $\hat{\lambda}_D$ is

$$\mathcal{V}_D(\lambda_D)^{-1}\nabla_{\lambda_D}\left\{\frac{n}{|B|}\sum_{i\in B} E_q\{\log p(\tilde{\alpha}_i|\beta,D)\}|_{\lambda_{\tilde{\alpha}_i}=\lambda^{\text{opt}}_{\tilde{\alpha}_i}} + E_q\{\log p(D|\nu,B)\}\right\}.$$

When B is the whole data set, $\hat{\lambda}_\beta$ and $\hat{\lambda}_D$ are respectively the updates of λ_β and λ_D in nonconjugate variational message passing.

With these unbiased estimates of the natural gradients, λ_β and λ_D can be updated using stochastic gradient approximation (Robbins and Monro 1951). At the tth iteration,

$$\lambda_\beta^{(t+1)} = \lambda_\beta^{(t)} + a_t\,(\hat{\lambda}_\beta - \lambda_\beta^{(t)}) \quad \text{and} \quad \lambda_D^{(t+1)} = \lambda_D^{(t)} + a_t\,(\hat{\lambda}_D - \lambda_D^{(t)}), \tag{11}$$

where $\hat{\lambda}_\beta$ and $\hat{\lambda}_D$ are evaluated using the current settings of λ_β and λ_D. Under certain regularity conditions (see Spall 2003), the iterates will converge to a local maximum of the lower bound. The gain sequence a_t, $t \geq 0$ should satisfy

$$a_t \to 0, \quad \sum_{t=0}^{\infty} a_t = \infty, \quad \text{and} \quad \sum_{t=0}^{\infty} a_t^2 < \infty.$$

The condition $(a_t \to 0, \sum_{t=0}^{\infty} a_t^2 < \infty)$ ensures that the step size goes to zero sufficiently fast so that iterates will converge while $(\sum_{t=0}^{\infty} a_t = \infty)$ ensures that the rate at

which step sizes approach zero is slow enough to avoid false convergence. Spall (2003) recommends

$$a_t = \frac{a}{(t+A)^\alpha}, \tag{12}$$

where $0.5 < \alpha \leq 1$, $A \geq 0$ is a stability constant that helps to avoid unstable behaviour in the early iterations and a keeps step sizes nonnegligible in later iterations. Note that updates in (11) can be rewritten as

$$\lambda_\beta^{(t+1)} = (1-a_t)\,\lambda_\beta^{(t)} + a_t\,\hat{\lambda}_\beta \quad \text{and} \quad \lambda_D^{(t+1)} = (1-a_t)\,\lambda_D^{(t)} + a_t\,\hat{\lambda}_D. \tag{13}$$

The tth iterate is thus a weighted average of the previous iterate and the nonconjugate variational message passing update estimated using mini-batch B. When $a_t = 1$ and B is the whole data set, $\lambda_\beta^{(t)}$ is precisely the update in nonconjugate variational message passing. This shows that nonconjugate variational message passing is a type of natural gradient method with step size 1 and other schedules are equivalent to damping.

4.4 Stochastic nonconjugate variational message passing algorithm

The stochastic nonconjugate variational message passing algorithm for fitting Poisson and logistic GLMMs is presented in Algorithm 2. Derivation of updates and definitions of F_i and g_i (appearing in Algorithm 2 and which differ according to whether logistic or Poisson GLMMs are fitted) are given in Appendix B. Algorithm 2 reduces to Algorithm 1 when mini-batch B is the entire data set, $a_t = 1$, and updates for local variational parameters are performed only once.

To initialize Algorithm 2, we suggest using the fit from penalized quasi-likelihood. This can be implemented in `R` via the `glmmPQL` function in `MASS` (Venables and Ripley 2002). Alternatively, for large data sets where penalized quasi-likelihood converges too slowly, we can use the fit from the GLM (obtained by pooling all data and setting random effects as zero) for initialization. For instance, we can set $\mu_{q(\beta)}$ and $\Sigma_{q(\beta)}$ respectively as estimates of the regression coefficients and their covariances from the GLM, $S_{q(D)} = (\nu_{q(D)} - r - 1)\hat{R}$ where $\nu_{q(D)} = \nu + n$, $\mu_{q(\tilde{\alpha}_i)} = \tilde{W}_i \mu_{q(\beta)}$ and $\Sigma_{q(\tilde{\alpha}_i)} = \hat{R}$. Kass and Natarajan (2006) gave a justification of $\hat{R}$ [defined in (1)] being a reasonable guess for D in the absence of any other prior knowledge. The tuning parameters W_i can be initialized by setting $D = \hat{R}$ and $\eta_i = X_i \mu_{q(\beta)}$ (for logistic GLMMs).

In step 1, mini-batches may be selected with or without replacement from the whole data set. Here, we consider sampling randomly at each iteration without replacement. Suppose the data set consist of n clusters and we randomly select $|B|$ clusters at the first iteration. At the second iteration, we will randomly sample $|B|$ clusters from the remaining $n - |B|$ clusters and so on. Algorithm 2 is considered to have made a sweep through the data when all clusters have been used once. This process is then repeated. Mini-batches in each sweep are sampled randomly and do not depend on those in previous sweeps. We allow mini-batches in each sweep to differ in size by one when n is not divisible by $|B|$. The advantage of sampling without replacement is that this scheme ensures all clusters (and local variational parameters) have been used or updated once in each sweep.

Initialize $\mu_{q(\beta)}$, $\Sigma_{q(\beta)}$, $S_{q(D)}$, $\mu_{q(\tilde{\alpha}_i)}$, $\Sigma_{q(\tilde{\alpha}_i)}$ and W_i for $i = 1, \ldots, n$. Set $\nu_{q(D)} = \nu + n$. For $t = 0, 1, 2, \ldots,$

1. Draw a mini-batch B of $|B| \geq 1$ units from the whole data set at random and without replacement.
2. Update local variational parameters $\mu_{q(\tilde{\alpha}_i)}$ and $\Sigma_{q(\tilde{\alpha}_i)}$ for $i \in B$ repeatedly:
   $\Sigma_{q(\tilde{\alpha}_i)} \leftarrow \left(\nu_{q(D)} S_{q(D)}^{-1} + Z_i^T F_i Z_i\right)^{-1}$
   $\mu_{q(\tilde{\alpha}_i)} \leftarrow \mu_{q(\tilde{\alpha}_i)} + \Sigma_{q(\tilde{\alpha}_i)} \left\{ Z_i^T (y_i - g_i) - \nu_{q(D)} S_{q(D)}^{-1} (\mu_{q(\tilde{\alpha}_i)} - \tilde{W}_i \mu_{q(\beta)}) \right\}$
   until convergence is reached.
3. Update global variational parameters $\mu_{q(\beta)}$, Σ_β^q and $S_{q(D)}$:
   $\Sigma_{q(\beta)} \leftarrow \left[(1 - a_t) \Sigma_{q(\beta)}^{-1} + a_t \left\{ \Sigma_\beta^{-1} + \frac{n}{|B|} \sum_{i \in B} \left(\nu_{q(D)} \tilde{W}_i^T S_{q(D)}^{-1} \tilde{W}_i + V_i^T F_i V_i \right) \right\} \right]^{-1}$
   $\mu_{q(\beta)} \leftarrow \mu_{q(\beta)} + a_t \Sigma_{q(\beta)} \left[\frac{n}{|B|} \sum_{i \in B} \left\{ \nu_{q(D)} \tilde{W}_i^T S_{q(D)}^{-1} (\mu_{q(\tilde{\alpha}_i)} - \tilde{W}_i \mu_{q(\beta)}) + V_i^T (y_i - g_i) \right\} - \Sigma_\beta^{-1} \mu_{q(\beta)} \right]$
   $S_{q(D)} \leftarrow (1 - a_t) S_{q(D)} + a_t \left[\frac{n}{|B|} \sum_{i \in B} \left\{ (\mu_{q(\tilde{\alpha}_i)} - \tilde{W}_i \mu_{q(\beta)})(\mu_{q(\tilde{\alpha}_i)} - \tilde{W}_i \mu_{q(\beta)})^T + \Sigma_{q(\tilde{\alpha}_i)} + \tilde{W}_i \Sigma_{q(\beta)} \tilde{W}_i^T \right\} + S \right]$

Algorithm 2: Stochastic nonconjugate variational message passing for GLMMs.

In step 2, we consider a loose criterion for assessing convergence to reduce computational overhead. Suppose mini-batch B consist of units $\{j_1, \ldots, j_{|B|}\}$. We define $\mu^B_{q(\tilde{\alpha})} = [\mu^T_{q(\tilde{\alpha}_{j_1})}, \ldots, \mu^T_{q(\tilde{\alpha}_{j_{|B|}})}]^T$ and terminate repetitions in step 2 when $\frac{\|\mu^{B}_{q(\tilde{\alpha})}{}^{(t)} - \mu^{B}_{q(\tilde{\alpha})}{}^{(t-1)}\|}{\|\mu^{B}_{q(\tilde{\alpha})}{}^{(t)}\|} < 0.05$, where $\|\cdot\|$ represents the Euclidean norm. Typically 3–7 repetitions are required for each mini-batch in the first sweep. The number of repetitions reduces steadily with the number of sweeps and usually just a single update is required by the third sweep.

For the examples in this paper, we did not update the tuning parameters W_i beyond initialization when the partially noncentered parametrization was used. While updating tuning parameters (at the end of each cycle in Algorithm 1 or at the end of each sweep in Algorithm 2) can lead to further improvements, more computation is required and this can be time-consuming for large data sets. A good initialization of the tuning parameters based on say penalized quasi-likelihood usually suffices.

The choice of step sizes a_t can strongly influence the performance of a stochastic approximation algorithm (Jank 2006). We discuss the choice of a gain sequence for Algorithm 2 in the next section.

4.5 Switching from stochastic to standard version

Determining an appropriate stopping criterion for a stochastic approximation algorithm can be challenging. Some commonly used criteria include stopping when the relative change in parameter values or objective function is sufficiently small or when the gradient of the objective function is sufficiently close to zero (Spall 2003). Such criteria do not provide any guarantees of the terminal iterate being close to the optimum however, and may be satisfied by random chance. Booth *et al.* (1999) recommend applying such rules for several consecutive iterations to minimize chances of a premature stop. However, Jank (2006) gave an illustrative example to show that even this may not be enough of a safeguard. Moreover, stochastic approximation can become excruciatingly slow in later iterations due to small step sizes.

Our experimentations with moderate-sized data sets indicate that gains made by Algorithm 2 are usually largest in the first few sweeps. However, beyond a certain point, it can become slower than Algorithm 1 if step sizes are too small or iterates simply bounce around if step sizes are still too big. An example is shown in Figure 1 where global variational parameters $\mu_{q(\beta)}$ and $S_{q(D)}$ are plotted against iterations t (or number of sweeps). Here Algorithm 2 is applied to a simulated data set of size $n = 10000$ (details in Example 6.3) and mini-batches of size $|B| = 100$ are used with step size $a_t = 1/(t+1)^\alpha$. Blue trajectories correspond to $\alpha = 0.55$ and black to $\alpha = 0.65$. Red dotted lines represent values obtained using Algorithm 1. Figure 1 shows that the blue and black trajectories converge towards the red dotted lines quickly in the first few sweeps. However, full convergence takes much longer. A larger step size ($\alpha = 0.55$) implies faster convergence at first but the iterates bounce around the optimum eventually if step sizes are still too large. A possible remedy to this is iterate averaging (Polyak and Juditsky 1992).

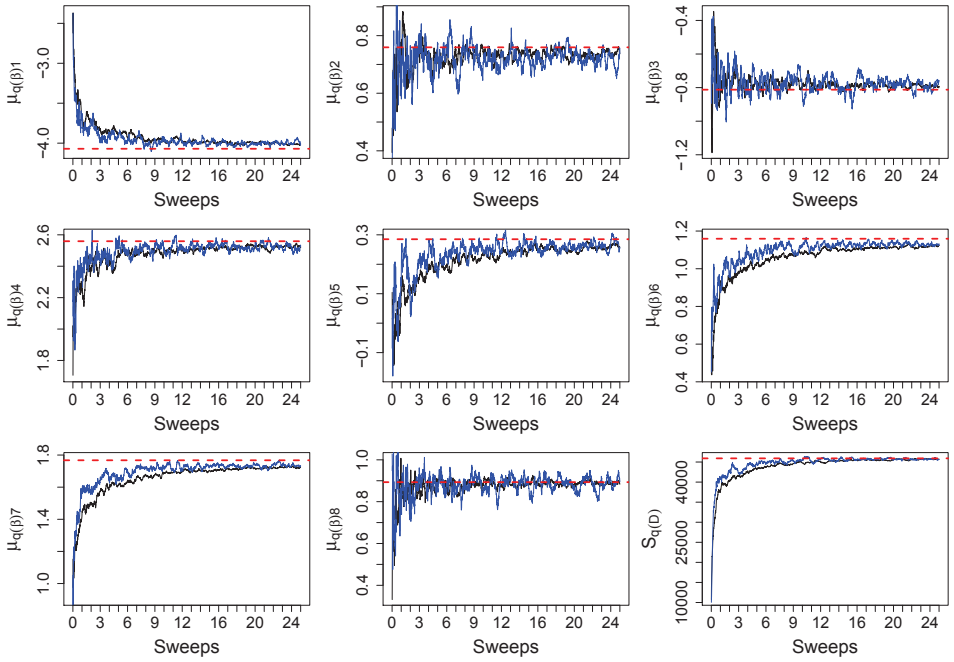

Figure 1: Polypharmacy simulated data ($n = 10000$). Global variational parameters $\mu_{q(\beta)}$ and $S_{q(D)}$ fitted using Algorithm 2 plotted against number of sweeps. Mini-batches $|B| = 100$ and step size $a_t = 1/(t+1)^{\alpha}$. Blue trajectories correspond to $\alpha = 0.55$ and black to $\alpha = 0.65$. Red dotted line denotes values obtained using Algorithm 1.

We suggest switching to Algorithm 1 when Algorithm 2 shows signs of slowing down. Using the lower bound both as a switching and stopping criterion, we propose switching from stochastic to standard nonconjugate variational message passing when the relative increase in the lower bound after a sweep is less than 10^{-3}, and terminating Algorithm 1 when the relative increase in the lower bound is less than 10^{-6}. For large datasets or streaming data, it might be more practical to terminate Algorithm 2 beyond a certain period of available runtime. To switch from Algorithm 2 to 1, the final setting of local and global variational parameters computed using Algorithm 2 is used as initialization of Algorithm 1.

Let M denote the number of iterations required to make a sweep through the data set. Following Spall (2003), we consider step sizes of the form $a_t = \frac{1}{t+A}$ setting $a = 1$ and $\alpha = 1$ in (12). We let $t = s_w + \frac{m}{M}$ where $0 \leq m \leq M-1$ denotes the number of mini-batches that has been analysed at the s_wth sweep. This specification slows down the rate of decrease in step size within each sweep and the larger step sizes help iterates move faster towards the optimum. We investigate performance of different stability constants A for various mini-batch sizes. Smaller values of α correspond to a slower decrease in step size and are desirable in some cases as they provide bigger step sizes in

later iterations. For our proposed strategy, we observed that smaller mini-batch sizes generally performed better. Since smaller step sizes are preferred for smaller mini-batch sizes (see Hoffman *et al.* 2010), we set $\alpha = 1$ for simplicity and report results only for this case.

Recently, Ranganath *et al.* (2013) developed an adaptive learning rate for stochastic variational inference, which is designed to minimize the expected distance between stochastic and optimal updates of the global variational parameters. They showed that adaptive step sizes led to improved convergence for the latent Dirichlet allocation model in topic modelling. It might be possible to extend this adaptive learning rate to nonconjugate models and we are working on this area. A wide variety of approaches have been developed to enhance the rate of convergence of stochastic approximation algorithms, and examples include iterate averaging (Polyak and Juditsky 1992), the momentum method (Tseng 1998) and gradient averaging (Xiao 2010). See Roux *et al.* (2012) for the stochastic average gradient method as well as a review of other approaches.

5 Prior-likelihood conflict diagnostics

In this section, we consider diagnostic tests for identifying divergent units in GLMMs. Such diagnostics are useful for detecting institutions (e.g. hospitals, trusts or schools) which deviate from the rest in a certain outcome. In healthcare for instance, it may be of interest to identify hospitals which are divergent in terms of quality of care provided or choice of surgical procedure for treating a cancer (Farrell *et al.* 2010). We demonstrate how prior-likelihood conflict diagnostics for identifying divergent units can be obtained as a by-product of nonconjugate variational message passing. The intuitive idea is that messages coming from above and below a node in a hierarchical model can be separated and "mixed messages" indicate conflict. Our "mixed messages" diagnostics can be shown to approximate the conflict diagnostics of Marshall and Spiegelhalter (2007). We start with a review of the simulation-based diagnostic test (Marshall and Spiegelhalter 2007), which is based on measuring the conflict between likelihood of a parameter and its predictive prior given the remaining data. Subsequently, we show how their approach can be approximated in the variational message passing framework.

5.1 Conflict p-values from simulation-based approach

For GLMMs with a partially noncentered parametrization, the linear predictor is

$$\eta_i = V_i\beta + Z_i\tilde{\alpha}_i \;\text{ where }\; \tilde{\alpha}_i \sim N(\tilde{W}_i\beta, D) \;\text{ for }\; i = 1, \ldots, n.$$

To identify units that do not appear to be drawn from the assumed random effects distributions, Marshall and Spiegelhalter (2007) suggest comparing replicates of $\tilde{\alpha}_i$ from its likelihood and predictive prior. A predictive prior replicate $\tilde{\alpha}_i^{\text{rep}}$ is first generated from

$$p_r(\tilde{\alpha}_i|y_{-i}) = \int p(\tilde{\alpha}_i|\beta, D)\, p(\beta, D|y_{-i})\, d\beta\, dD \tag{14}$$

where y_{-i} denotes observed data y with unit i left out. This replicate can be obtained by generating β^{rep} and D^{rep} from $p(\beta, D|y_{-i})$ using MCMC, followed by simulation of $\tilde{\alpha}_i^{\text{rep}}|\beta^{\text{rep}}, D^{\text{rep}}$. A likelihood replicate $\tilde{\alpha}_i^{\text{lik}} \sim p(\tilde{\alpha}_i|y_i)$ is then generated using only the unit y_i being tested and a non-informative prior $p(\tilde{\alpha}_i)$ for $\tilde{\alpha}_i$. Marshall and Spiegelhalter (2007) recommend the Jeffreys's prior as a noninformative prior for $\tilde{\alpha}_i$ (see Box and Tiao 1973). These prior and likelihood replications represent independent sources of evidence about $\tilde{\alpha}_i$ and conflict between them suggests discrepancies in the model.

The discussion above ignores nuisance parameters. For GLMMs, we need to regard β as a nuisance parameter. As $p(\tilde{\alpha}_i|y_i) \propto p(\tilde{\alpha}_i) \int p(y_i|\beta, \tilde{\alpha}_i)\, p(\beta|\tilde{\alpha}_i)\, d\beta$ and β is not estimable from individual unit i, Marshall and Spiegelhalter (2007)[pg. 420] recommend generating $\tilde{\alpha}_i^{\text{lik}}$ from

$$p_l(\alpha_i|y) \propto p(\tilde{\alpha}_i) \int p(y_i|\tilde{\alpha}_i, \beta) p(\beta|y_{-i})\, d\beta.$$

Note that the two replications $\tilde{\alpha}_i^{\text{rep}}$ and $\tilde{\alpha}_i^{\text{lik}}$ are no longer entirely independent as y_{-i} will slightly influence $\tilde{\alpha}_i^{\text{lik}}$ through β.

To compare prior and likelihood replicates, Marshall and Spiegelhalter (2007) considered $\tilde{\alpha}_i^{\text{diff}} = \tilde{\alpha}_i^{\text{rep}} - \tilde{\alpha}_i^{\text{lik}}$ and calculated a conflict p-value,

$$p_{i,\text{con}}^l = P(\tilde{\alpha}_i^{\text{diff}} \leq 0|y)$$

as the proportion of times simulated values of $\tilde{\alpha}_i^{\text{diff}}$ are less than or equal to zero for scalar $\tilde{\alpha}_i$. Depending on the context, the upper tail area $p_{i,\text{con}}^u = 1 - p_{i,\text{con}}^l$ or two-sided p-value $2 \times \min(p_{i,\text{con}}^l, p_{i,\text{con}}^u)$ may be of interest instead. If $\tilde{\alpha}_i^{\text{diff}}$ is not a scalar,

$$\Delta = \text{E}(\tilde{\alpha}_i^{\text{diff}}|y)^T \text{Cov}(\tilde{\alpha}_i^{\text{diff}}|y)^{-1} \text{E}(\tilde{\alpha}_i^{\text{diff}}|y)$$

can be used as a standardized discrepancy measure. If we further assume a multivariate normal distribution for $\tilde{\alpha}_i^{\text{diff}}$, then a conflict p-value for testing $\tilde{\alpha}_i^{\text{diff}} = 0$ can be calculated as $P(\chi_r^2 > \Delta)$, where χ_r^2 denotes a Chi-square random variable with r degrees of freedom. Further discussion on p-values in multivariate case can be found in Presanis *et al.* (2013).

MCMC methods are not well-suited to cross-validation approaches and an alternative simulation-based full-data approach was proposed by Marshall and Spiegelhalter (2007). The procedure is the same as before except that $\tilde{\alpha}_i^{\text{rep}}|\beta^{\text{rep}}, D^{\text{rep}}$ is simulated using β^{rep}, D^{rep} generated from $p(\beta, D|y)$, without leaving out y_i. Mild conservatism is introduced as y_i will influence $\tilde{\alpha}_i^{\text{rep}}$ slightly through β and D.

5.2 Conflict p-values from nonconjugate variational message passing

Next, we show how approximate conflict p-values can be calculated within nonconjugate variational message passing. From (7), the update for $\lambda_{\tilde{\alpha}_i}$ is

$$\mathcal{V}_{\tilde{\alpha}_i}(\lambda_{\tilde{\alpha}_i})^{-1} \nabla_{\lambda_{\tilde{\alpha}_i}} E_q\{\log p(\tilde{\alpha}_i|\beta, D)\} + \mathcal{V}_{\tilde{\alpha}_i}(\lambda_{\tilde{\alpha}_i})^{-1} \nabla_{\lambda_{\tilde{\alpha}_i}} E_q\{\log p(y_i|\tilde{\alpha}_i, \beta)\}.$$

The first term can be considered as a message from the prior $p(\tilde{\alpha}_i|\beta, D)$ and the second term a message from the likelihood $p(y_i|\tilde{\alpha}_i, \beta)$ of unit y_i. We argue below that the first message from the prior can be interpreted as the natural parameter of a Gaussian approximation say $N(\mu_{\text{rep}}, \Sigma_{\text{rep}})$ to $p_r(\tilde{\alpha}_i|y_{-i})$. On the other hand, the second message from the likelihood can be interpreted as the natural parameter of a Gaussian approximation say $N(\mu_{\text{lik}}, \Sigma_{\text{lik}})$ to $p_l(\tilde{\alpha}_i|y)$. This implies that $\tilde{\alpha}_i^{\text{rep}} \sim N(\mu_{\text{rep}}, \Sigma_{\text{rep}})$ and $\tilde{\alpha}_i^{\text{lik}} \sim N(\mu_{\text{lik}}, \Sigma_{\text{lik}})$. If we further assume $\tilde{\alpha}_i^{\text{rep}}$ and $\tilde{\alpha}_i^{\text{lik}}$ are independent, then $\tilde{\alpha}_i^{\text{diff}} \sim N(\mu_{\text{rep}} - \mu_{\text{lik}}, \Sigma_{\text{rep}} + \Sigma_{\text{lik}})$. Even though $\tilde{\alpha}_i^{\text{rep}}$ and $\tilde{\alpha}_i^{\text{lik}}$ are not entirely independent, the dependence between $\tilde{\alpha}_i^{\text{rep}}$ and $\tilde{\alpha}_i^{\text{lik}}$ will be increasingly weak as the number of clusters increases. Since these messages are computed in the nonconjugate variational message passing algorithm, conflict p-values can be calculated easily at convergence for identifying divergent units.

The arguments presented below are by no means rigorous. However, they lend some insight into how conflict p-values can be approximated from nonconjugate variational message passing and experimental results suggest the approximations work well in practice. For large data sets, automatic computation of diagnostics for prior-likelihood conflict can be an attractive alternative to simulation-based MCMC approaches. They are also useful generally as initial screening tools and clusters flagged as divergent can be studied more closely and possibly conflict p-values recomputed by Monte Carlo.

First, consider the message from the prior. If we treat the message as the natural parameter of a normal distribution, we get $\mu_{\text{rep}} = \tilde{W}_i \mu_{q(\beta)}$ and $\Sigma_{\text{rep}} = S_{q(D)}/\nu_{q(D)}$. For large data sets, $p(\beta, D|y_{-i})$ is close to $p(\beta, D|y)$ and we approximate $p(\beta, D|y_{-i})$ in (14) by the variational posterior $q(\beta|\lambda_\beta)q(D|\lambda_D)$. This combined with Jensen's inequality gives

$$\begin{aligned} p_r(\tilde{\alpha}_i|y_{-i}) &\approx \int p(\tilde{\alpha}_i|\beta, D)\, q(\beta|\lambda_\beta)q(D|\lambda_D)\, d\beta\, dD \\ &\geq \exp[E_{-\tilde{\alpha}_i}\{\log p(\tilde{\alpha}_i|\beta, D)\}]. \end{aligned}$$

While $\exp[E_{-\tilde{\alpha}_i}\{\log p(\tilde{\alpha}_i|\beta, D)\}]$ is only a lower bound to $p_r(\tilde{\alpha}_i|y_{-i})$, we find that by using it as an approximation to $p_r(\tilde{\alpha}_i|y_{-i})$, we get $p_r(\tilde{\alpha}_i|y_{-i}) \propto \exp[E_{-\tilde{\alpha}_i}\{\log p(\tilde{\alpha}_i|\beta, D)\}]$. This gives $\tilde{\alpha}_i^{\text{rep}} \sim N(\tilde{W}_i \mu_{q(\beta)}, S_{q(D)}/\nu_{q(D)})$, which is what we would get if we interpret the first message as being the natural parameter of a Gaussian approximation to $p_r(\tilde{\alpha}_i|y_{-i})$.

Next, consider the second message from the likelihood. If we treat the message as the natural parameter of a normal distribution, it can be shown that $\Sigma_{\text{lik}}^{-1} = Z_i^T F_i Z_i$ and $\mu_{\text{lik}} = \mu_{q(\tilde{\alpha}_i)} + \Sigma_{\text{lik}} Z_i^T (y_i - g_i)$. Now consider the sum of the two messages. This gives us the natural parameter of $q(\tilde{\alpha}_i|\lambda_{\tilde{\alpha}_i})$ which is an approximation of $p(\tilde{\alpha}_i|y)$. Note that

$$\Sigma_{\text{rep}}^{-1} + \Sigma_{\text{lik}}^{-1} = \Sigma_{q(\tilde{\alpha}_i)}^{-1} \text{ and } \Sigma_{\text{rep}}^{-1}\mu_{\text{rep}} + \Sigma_{\text{lik}}^{-1}\mu_{\text{lik}} = \Sigma_{q(\tilde{\alpha}_i)}^{-1}\mu_{q(\tilde{\alpha}_i)}.$$

If we think of $p(\tilde{\alpha}_i|y_{-i})$ as the 'prior' to be updated when y_i becomes available, we have

$$p(\tilde{\alpha}_i|y) \propto p(\tilde{\alpha}_i|y_{-i})p(y_i|\tilde{\alpha}_i, y_{-i}) \Rightarrow \frac{p(\tilde{\alpha}_i|y)}{p(\tilde{\alpha}_i|y_{-i})} \propto p(y_i|\tilde{\alpha}_i, y_{-i}).$$

Interpreting the first message as a Gaussian approximation to $p(\tilde{\alpha}_i|y_{-i})$ and the sum of the two messages as a Gaussian approximation to $p(\tilde{\alpha}_i|y)$, the ratio of these two normal distributions gives an approximation (up to a proportionality constant) of $p(y_i|\tilde{\alpha}_i, y_{-i})$. As a function of $\tilde{\alpha}_i$, the ratio of the two normal distributions is proportional to

$$\frac{\exp\{-\frac{1}{2}(\tilde{\alpha}_i - \mu_{q(\tilde{\alpha}_i)})^T \Sigma^{-1}_{q(\tilde{\alpha}_i)}(\tilde{\alpha}_i - \mu_{q(\tilde{\alpha}_i)})\}}{\exp\{-\frac{1}{2}(\tilde{\alpha}_i - \mu_{\text{rep}})^T \Sigma^{-1}_{\text{rep}}(\tilde{\alpha}_i - \mu_{\text{rep}})\}},$$

which gives a normal distribution with mean μ_{lik} and covariance Σ_{lik}, precisely that given by the second message. As

$$p(y_i|\tilde{\alpha}_i, y_{-i}) = \int p(y_i|\beta, \tilde{\alpha}_i) p(\beta|\tilde{\alpha}_i, y_{-i})\, d\beta$$

and $p(\beta|\tilde{\alpha}_i, y_{-i})$ is close to $p(\beta|y_{-i})$ when the number of clusters is large (in the sense that dependence of β on $\tilde{\alpha}_i$ is reduced), the second message can be considered as the natural parameter of a Gaussian approximation to $p_l(\tilde{\alpha}_i|y)$ if we assume a uniform prior for $p(\tilde{\alpha}_i)$. The arguments above generalize to detecting conflict for other parameters of the model as well.

While the discussion here uses the partially noncentered parametrization, conclusions hold for the centered and noncentered parametrizations as well. We observed small differences in conflict p-values computed using different parametrizations, which is due likely to varying accuracy of approximations to the true posterior. To compare the accuracy of different approaches, we first transform the conflict p-values to z-scores to reflect the importance of good agreement at the extremes (Marshall and Spiegelhalter 2007). Using the cross-validatory conflict p-values as a "gold-standard", we use the mean absolute difference in z-scores,

$$\frac{1}{n}\sum_{i=1}^{n} |\Phi^{-1}(p^{\text{CV}}_{i,\text{con}}) - \Phi^{-1}(p^{\text{method}}_{i,\text{con}})|,$$

as a measure of the degree of agreement between the cross-validatory conflict p-values ($p^{\text{CV}}_{i,\text{con}}$) and conflict p-values computed from the method we are trying to assess ($p^{\text{method}}_{i,\text{con}}$).

To compute conflict p-values for large data sets, one needs to ensure that local variational parameters for every unit are optimized. As Algorithm 2 focuses on optimization of global variational parameters using stochastic approximation, not all local variational parameters may have been fully optimized when the global variational parameters have converged. This can be resolved by performing an additional step of optimizing local variational parameters for every unit as a function of the converged global variational parameters. Alternatively, our proposed strategy of switching from Algorithm 2 to 1 also ensures that local variational parameters for every unit are optimized. However, due to the difficulty in computing conflict p-values for large data sets using cross-validatory or even full-data approaches with MCMC, we focus on comparisons with nonconjugate variational message passing using only small data problems in the examples.

6 Examples

In Sections 6.1 and 6.2, we use the Bristol inquiry data and epilepsy data to compare conflict p-values computed using nonconjugate variational message passing with those obtained using the simulation-based cross-validatory approach (Marshall and Spiegelhalter 2007). An additional example on Madras schizophrenia data can be found in Appendix D. These data sets are relatively small and we only use Algorithm 1 for fitting.

In Sections 6.3 and 6.4, we use moderately large simulated data sets to illustrate the improvements in efficiency that can be obtained by using stochastic nonconjugate variational message passing in the initial stage of optimization. We compare performances of Algorithms 1 and 2 for the simulated data sets using only the partially noncentered parametrization. Algorithms 1 and 2 were initialized using penalized quasi-likelihood in all examples except for the large simulated data set in Section 6.4, where penalized quasi-likelihood converges too slowly. The GLM fit was used instead for initialization.

In all examples, fitting via MCMC was performed in OpenBUGS (Lunn *et al.* 2009) through `R` by using `R2OpenBUGS` as an interface. `R2OpenBUGS` was adapted by Neal Thomas from `R2WinBUGS` (Sturtz *et al.* 2005). The MCMC algorithm was initialized using penalized quasi-likelihood and the same priors were used in MCMC and nonconjugate variational message passing. We consider a vague $N(0, 1000)$ prior for β in each case. All code was written in `R` and run on a dual processor Windows PC 3.30 GHz workstation. Computation times reported are in seconds (s).

In some examples below, the variational posterior approximations are biased as compared to results from MCMC. This is due to the assumption of a factorized variational posterior and the impact of this restriction depends on how strong posterior dependence is among the factored variables. In VB, the posterior variance tends to be underestimated and this issue has been noted by Wang and Titterington (2005) and Bishop (2006). Recently, Zhao and Marriott (2013) proposed some diagnostics for assessing how well VB approximates the true posterior as well as correction measures that can be undertaken when the approximation error is large. Salimans and Knowles (2013) developed stochastic approximation methods for hierarchical approximations that allow independence assumptions in VB to be relaxed.

6.1 Bristol inquiry data

In 1998, a public inquiry was set up to look into the management of children receiving complex cardiac surgical services at the Bristol Royal Infirmary. The outcomes of surgical services at Bristol, UK, relative to other specialist centres was a key issue. We consider a subset of the data recorded by Hospital Episode Statistics on mortality rates in open surgeries for 12 hospitals including Bristol (hospital 1), for children under 1 year old, from 1991 to 1995 (see Marshall and Spiegelhalter 2007, Table 1). Although the number of clusters is small in this example whereas our methodology is motivated by applications to large data sets, this example is interesting as a benchmark data set in

the literature for computing conflict diagnostics using nonconjugate variational message passing.

Let $y_{ij} \sim \text{Bernoulli}(\pi_i)$ where $y_{ij} = 1$ if patient j at hospital i died and 0 otherwise. We use $Y_i = \sum_{j=1}^{n_i} y_{ij}$ to denote the number of deaths at hospital i, $i = 1, \ldots, 12$. Let

$$\text{logit}(\pi_i) = \beta + u_i \quad \text{where} \quad u_i \sim N(0, D).$$

In the cross-validatory approach, each hospital i was removed in turn from the analysis, and $\beta^{\text{rep}}, D^{\text{rep}}|y_{-i}$ were generated using MCMC followed by a simulated $\pi_i^{\text{rep}}|\beta^{\text{rep}}, D^{\text{rep}}$. Assuming a Jeffreys's prior for π_i, a π_i^{lik} was simulated from $p(\pi_i|y_i) = \text{Beta}(Y_i+0.5, n_i - Y_i + 0.5)$. Excess mortality is of concern and the upper-tail area is used as a 1-sided p-value so that $p_{i,\text{con}} = P(\pi_i^{\text{rep}} \geq \pi_i^{\text{lik}})$. For each fitting via MCMC, two chains were run simultaneously to assess convergence, each with 51,000 iterations, and the first 1000 iterations were discarded in each chain as burn-in. Cross-validatory conflict p-values were calculated based on the remaining 100,000 simulations. The total time taken for model updating in OpenBUGS is 5 s $\times$ 12 = 60 s for the cross-validatory approach.

The variational lower bounds and CPU times taken for model fitting and computation of conflict p-values by Algorithm 1 (via different parametrizations) and MCMC (full-data approach) are shown in Table 1. Figure 2 shows the marginal posteriors of β and D estimated using MCMC and Algorithm 1. The partially noncentered parametrization attained the highest lower bound, was quick to converge and produced posterior approximations very close to that of MCMC.

Figure 3 compares conflict p-values computed using the cross-validatory approach and nonconjugate variational message passing using partial noncentering. The plot indicates very good agreement between the two sets of p-values. Both approaches suggest hospital 1 (Bristol) is discrepant. The mean absolute difference in z-scores for nonconjugate variational message passing and the simulation-based full-data approach relative to the cross-validatory approach are given in Table 1. Nonconjugate variational message passing does better than the simulation-based full-data approach both in terms of z-scores and computation time. The difference in conflict p-values computed using different parametrizations is small.

For this example, nonconjugate variational message passing is of an order of magnitude faster than the cross-validatory approach. We will see in the next two examples that the reduction in computation time is even greater for larger data sets. There are some difficulties in comparing nonconjugate variational message passing and MCMC in this way as the time taken for the variational algorithm to converge depends on the initialization, stopping rule and the rate of convergence is problem-dependent. The updating time for MCMC is also problem-dependent and depends on the length of burn-in and number of sampling iterations. It is clear, however, that for large data sets, the variational approach is attractive as an alternative to MCMC methods for obtaining prior-likelihood conflict diagnostics or as an initial screening tool.

	noncentered	centered	partially noncentered	MCMC (full-data)
Lower bound ($\mathcal{L}$)	-1213.7	-1213.0	-1212.9	–
Time (model fitting)	7.6	3.7	3.8	5
Time (compute conflict p-values)	0.3	0.3	0.3	14.4
Mean abs difference in z-scores	0.087	0.086	0.083	0.125

Table 1: Bristol data. Variational lower bounds (first row), CPU times (s) for model fitting (second row) and computing conflict p-values (third row) and mean absolute difference in z-scores relative to cross-validatory approach (last row) for Algorithm 1 (different parametrizations) and MCMC (full-data).

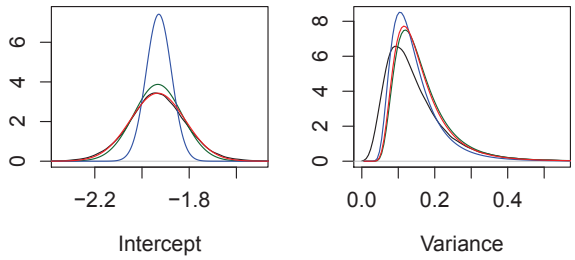

Figure 2: Bristol data. Marginal posteriors estimated by MCMC (black) and Algorithm 1 using the centered (green), noncentered (blue) and partially noncentered (red) parametrizations.

6.2 Epilepsy data

The epilepsy data set (Thall and Vail 1990) contains records from a clinical trial of 59 patients with epilepsy. Each patient was randomly administered a new anti-epileptic drug, progabide, (Trt=1) or a placebo (Trt=0) and the number of seizures during the two 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 $\text{Visit}_4 = 0.3$) was recorded. The number of seizures during the 8-week period prior to randomization was also noted. We consider the logarithm of $\frac{1}{4}$ the number of baseline seizures (Base) and the logarithm of the age of patient (Age) as covariates. We center Age at its mean to improve mixing in MCMC methods. Breslow *et al.* (1993) considered a Poisson random intercept and slope model:

$$\log \mu_{ij} = \beta_0 + \beta_1 \text{Base}_i + \beta_2 \text{Trt}_i + \beta_3 \text{Base}_i \times \text{Trt}_i + \beta_4 \text{Age}_i + \beta_5 \text{Visit}_{ij} + u_{1i} + u_{2i} \text{Visit}_{ij}, \tag{15}$$

for $i = 1, \ldots, 59$, $j = 1, \ldots, 4$ and $\left[\begin{smallmatrix} u_{1i} \\ u_{2i} \end{smallmatrix}\right] \sim N\left(0, \left[\begin{smallmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{21} & \sigma_{22}^2 \end{smallmatrix}\right]\right)$. We compare conflict p-values computed using the cross-validatory approach and nonconjugate variational message passing for two models. Model I is a random intercept model where the random slope is dropped from (15). Model II is the random intercept and slope model in (15).

hospital	$p_{i,\text{con}}^{\text{CV}}$	$p_{i,\text{con}}^{\text{NCVMP}}$
1	0.001	0.005
2	0.436	0.450
3	0.935	0.928
4	0.125	0.138
5	0.298	0.311
6	0.720	0.725
7	0.737	0.745
8	0.661	0.667
9	0.440	0.453
10	0.380	0.390
11	0.763	0.764
12	0.721	0.727

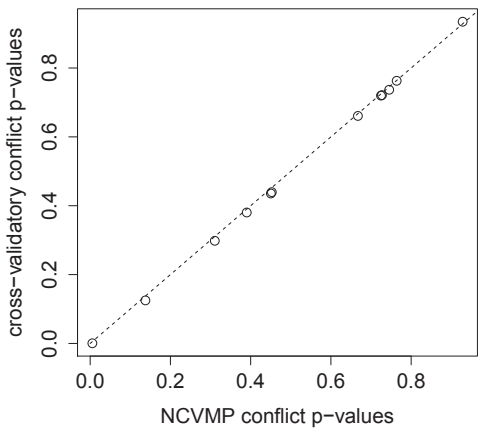

Figure 3: Bristol data. Cross-validatory conflict p-values ($p_{i,\text{con}}^{\text{CV}}$) and conflict p-values from nonconjugate variational message passing ($p_{i,\text{con}}^{\text{NCVMP}}$) using partial noncentering.

We examine suitability of the assumed random effects distribution and report two-sided conflict p-values for both models.

For simulation-based approaches, it is easier to work with the centered parametrization as handling of nuisance parameters is minimized (see details in Appendix C). Under this parametrization, there are no nuisance parameters in Model II and only β_5 needs to be regarded as a nuisance parameter in Model I. Each patient was removed in turn from the analysis in the cross-validatory approach. For each model fitting via MCMC, two chains were run simultaneously to assess convergence, each with 26,000 iterations, and the first 1000 iterations were discarded in each chain as burn-in. Cross-validatory conflict p-values were calculated based on the remaining 50,000 simulations. The total time taken for model updating in OpenBUGS is 61 s $\times$ 59 = 3599 s for Model I and 54 s $\times$ 59 = 3186 s for Model II. Simulation of prior and likelihood replicates of the centered random effects α_i was performed in `R`. To simulate likelihood replicates, we assume Jeffreys's prior for α_i and use adaptive rejection Metropolis sampling via the `arms` function in the `HI` package (Petris and Tardella 2003).

Variational lower bounds and CPU times taken for model fitting and computation of conflict p-values by Algorithm 1 (via different parametrizations) and MCMC (full-data approach) are given in Table 2. Marginal posteriors of parameters in Model I estimated using MCMC and Algorithm 1 are given in Figure 4. Comparison of parameter estimates for Model II can be found in Tan and Nott (2013). Partial noncentering performed very well in posterior approximations and was quick to converge.

Cross-validatory conflict p-values are plotted against conflict p-values from nonconjugate variational message passing using the partially noncentered parametrization in Figure 5, for Model I (left) and Model II (right). The mean absolute difference in z-scores for nonconjugate variational message passing and the simulation-based full-

	noncentered	centered	partially noncentered	MCMC (full-data)
Model I				
Lower bounds ($\mathcal{L}$)	-707.0	-701.5	-701.1	–
Time (model fitting)	1.4	0.2	0.2	62
Time (compute conflict p-values)	< 0.05	< 0.05	< 0.05	4278.2
Mean abs difference in z-scores	0.167	0.159	0.155	0.103
Model II				
Lower bounds ($\mathcal{L}$)	-701.4	-696.1	-695.3	–
Time (model fitting)	1.3	0.5	0.5	55
Time (compute conflict p-values)	< 0.05	< 0.05	< 0.05	3109.6
Mean abs difference in z-scores	0.105	0.107	0.101	0.116

Table 2: Epilepsy data. Variational lower bounds (first row), CPU times (s) for model fitting (second row) and computing conflict p-values (third row), and mean absolute difference in z-scores relative to cross-validatory approach (fourth row) for Algorithm 1 (different parametrizations) and MCMC (full-data).

data approach relative to the cross-validatory approach are given in Table 2. Figure 5 shows good agreement between cross-validatory conflict p-values and conflict p-values computed using nonconjugate variational message passing. The agreement is better in Model II and this is reflected in the z-scores in Table 2. Nonconjugate variational message passing compares well with the simulation-based full-data approach in terms of z-scores and is faster than both simulation-based approaches by an order of magnitude.

At the 0.05 level, outliers identified by the cross-validatory approach are patients 10, 25, 35, 56 and 58 for Model I and patients 10, 25 and 56 for Model II. Table 3 shows the cross-validatory conflict p-values for these patients. The corresponding conflict p-values computed using nonconjugate variational message passing with partial noncentering are shown for comparison. While p-values from the two approaches are close, some of the outliers identified by the cross-validatory approach are not detected using nonconjugate variational message passing. One way to resolve this issue is to flag all patients with conflict p-values < 0.1 say as possible outliers and recompute conflict p-values for this smaller group using the cross-validatory approach. In this way, nonconjugate variational message passing can be regarded as a screening tool which will be very useful for large data sets.

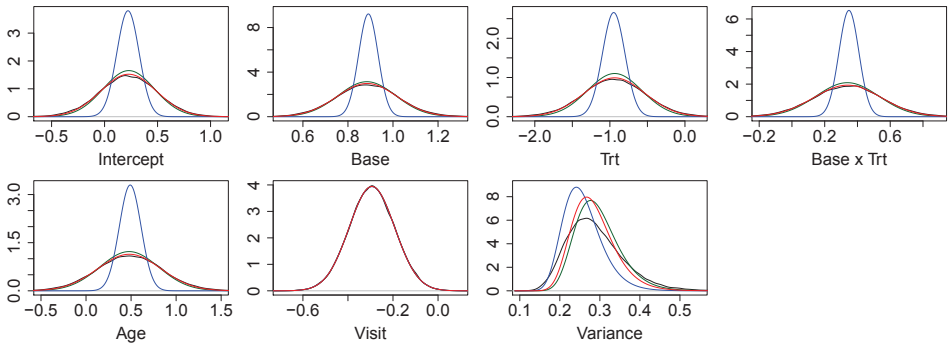

Figure 4: Epilepsy data Model I. Marginal posteriors estimated by MCMC (black) and Algorithm 1 using the centered (green), noncentered (blue) and partially noncentered (red) parametrizations.

Model I

Patient	$p_{i,\text{con}}^{\text{CV}}$	$p_{i,\text{con}}^{\text{NCVMP}}$
10	0.047	0.056
25	0.048	0.062
35	0.038	0.044
56	0.023	0.028
58	0.002	0.006

Model II

Patient	$p_{i,\text{con}}^{\text{CV}}$	$p_{i,\text{con}}^{\text{NCVMP}}$
10	0.001	0.005
25	0.024	0.049
56	0.038	0.051

Table 3: Epilepsy data. Conflict p-values for outliers in models I and II from cross-validatory approach and nonconjugate variational message passing using partial noncentering.

6.3 Polypharmacy data

The polypharmacy data set (Hosmer *et al.* 2013) contains data on 500 subjects studied over a period of seven years. The outcome of interest is whether the subject is taking drugs from 3 or more different groups. The number of outpatient mental health visits (MHV) and inpatient mental health visits made by each subject were recorded each year. We consider the dummy variables MHV_1=1 if $1 \leq \text{MHV} \leq 5$ and 0 otherwise, MHV_2=1 if if $6 \leq \text{MHV} \leq 14$ and MHV_3=1 if $\text{MHV} \geq 15$ and 0 otherwise. Let INPTMHV = 0 if there were no inpatient mental health visits and 1 otherwise. Other covariates include Age, Gender = 1 if male and 0 if female and Race = 0 if subject is White and 1 otherwise. The data set is avaliable at http://www.umass.edu/statdata/statdata/stat-logistic.html. Following Hosmer *et al.* (2013), we consider a logistic random intercept model of the form

$$\begin{aligned}\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 \text{Gender}_i + \beta_2 \text{Race}_i + \beta_3 \text{Age}_{ij} + \beta_4 \text{MHV\_1}_{ij} \\ + \beta_5 \text{MHV\_2}_{ij} + \beta_6 \text{MHV\_3}_{ij} + \beta_7 \text{INPTMHV}_{ij} + u_i, \quad (16)\end{aligned}$$

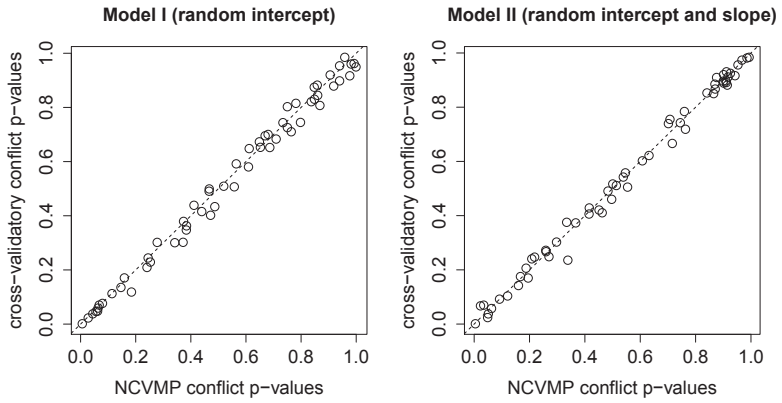

Figure 5: Epilepsy data. Cross-validatory conflict p-values plotted against conflict p-values from nonconjugate variational message passing using partially noncentered parametrization, for Model I (left) and Model II (right).

where $u_i \sim N(0, \sigma^2)$ for $i = 1, \ldots, 500$, $j = 1, \ldots, 7$.

This model was fitted using Algorithm 1 and MCMC. Variational lower bounds and CPU times for model fitting are shown in Table 4. For MCMC, two chains were run simultaneously to assess convergence, each with 11,000 iterations, and the first 1000 iterations were discarded in each chain as burn-in. Algorithm 1 is of an order of magnitude faster than MCMC. Figure 6 shows the marginal posterior distributions of parameters estimated using MCMC and Algorithm 1. The partially noncentered parametrization attained the highest lower bound and took much less time to converge than the noncentered parametrization. Posterior approximations for β_0, β_1 and β_2 from partial noncentering were better than that of centering and noncentering. While posterior variances of β_4, β_5 and β_6 were underestimated by partial noncentering, the estimated posterior means were close to that of MCMC. As this data set is relatively small, using Algorithm 2 in the initial stage of optimization did not lead to significant reductions in computation times.

	noncentered	centered	partially noncentered	MCMC
Lower bound ($\mathcal{L}$)	-1414.9	-1414.4	-1414.0	–
Time (model fitting)	109.0	38.8	65.0	4320

Table 4: Polypharmacy data. Variational lower bounds (first row) and CPU times (s) for model fitting (second row), for Algorithm 1 (different parametrizations) and MCMC.

To illustrate the improvements in efficiency that can be obtained from stochastic nonconjugate variational message passing, we simulated a larger data set comprised of $n = 500 \times 20 = 10,000$ subjects from the model fitted by Algorithm 1 (using partial noncentering). The design matrices for each cluster were replicated 20 times and re-

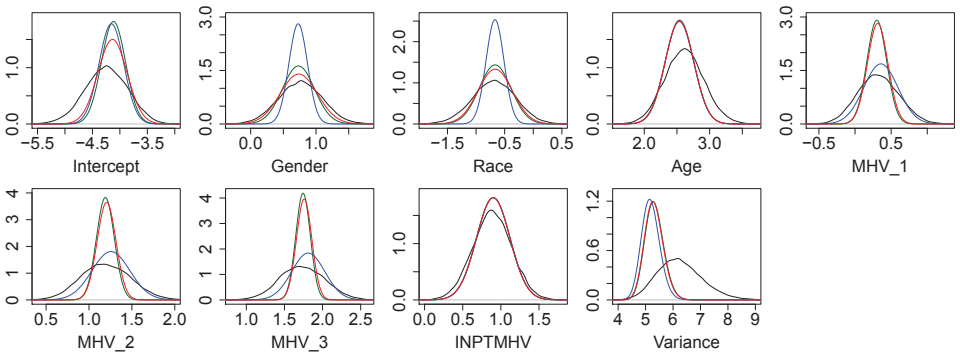

Figure 6: Polypharmacy data. Marginal posteriors estimated by MCMC (black) and Algorithm 1 using the centered (green), noncentered (blue) and partially noncentered (red) parametrizations.

sponses were generated from the model in (16), using as parameters variational posterior means from the fitted model. For this simulated data, Algorithm 1 using the partially noncentered parametrization took 656.6 s to converge.

For Algorithm 2, we considered mini-batch sizes $|B| \in \{50, 100, 200, 400\}$ (corresponding to 0.05%, 1%, 2% and 4% of $n = 10,000$) and stability constants $A \in \{1, 2, 4, 8, 16, 32, 64\}$. Larger constants were used for smaller mini-batch sizes. For each setting, we performed ten runs of Algorithm 2 switching to Algorithm 1 when the relative increment in the lower bound after a sweep is less than 10^{-3}. Computation times for the four mini-batch sizes corresponding to different stability constants are displayed in boxplots in Figure 7. The shortest average time to convergence for the different mini-batch sizes are given in Table 5 together with corresponding stability constants A. From Figure 7, computation times were reduced by a factor of close to 2 or more across different mini-batch sizes and stability constants considered. Table 5 showed that larger stability constants A are preferred for smaller mini-batch sizes. The shortest average time to convergence of 236.7 s was achieved by mini-batches of size 100 with $A = 16$. This represents a reduction in computation time from Algorithm 1 by a factor of 2.8.

$\|B\|$	50	100	200	400
A	32	16	8	2
time	239.6	236.7	246.0	251.9

Table 5: Polypharmacy simulated data. Shortest average time to convergence (s) for different mini-batch sizes together with corresponding stability constant A.

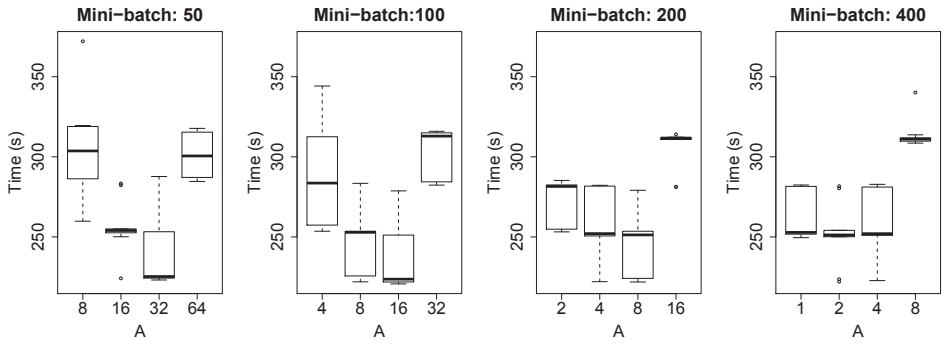

Figure 7: Polypharmacy simulated data. CPU times (s) for mini-batch sizes 50, 100, 200 and 400 corresponding to different stability constants displayed in boxplots.

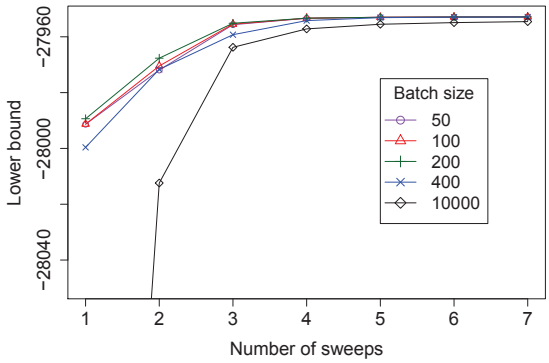

Figure 8: Polypharmacy simulated data. Plot of average lower bound against number of sweeps for different batch sizes, with stability constants A given in Table 5.

Figure 8 tracks the average lower bound attained at the end of each sweep for different mini-batch sizes, with stability constants A given in Table 5. Only the first seven sweeps are shown. Figure 8 shows that with appropriate step sizes, stochastic nonconjugate variational message passing is able to make much bigger gains than the standard version, particularly in the first few sweeps. Thus, for moderate-sized data sets, gains in computation times can be obtained by using Algorithm 2 in the initial stage of optimization.

6.4 Skin cancer prevention study

In a clinical trial to test the effectiveness of beta-carotene in preventing non-melanoma skin cancer (Greenberg *et al.* 1989), 1805 high risk patients were randomly assigned to receive either a placebo or 50 mg of beta-carotene per day for five years. The response

y_{ij} is a count of the number of new skin cancers in year j for the ith subject. Covariate information for the ith subject includes Age_i, the age in years at the beginning of the study, $\text{Gender}_i = 1$ if male and 0 if female, $\text{Skin}_i = 1$ if skin has burns and 0 otherwise, Exposure_i, a count of the number of previous skin cancers, and Year_{ij}, the year of follow-up. The treatment effect has been shown to be insignificant in previous analyses. We consider $n = 1683$ subjects with complete covariate information (data set available at http://www.biostat.harvard.edu/~fitzmaur/ala2e/). Following Donohue *et al.* (2011), we consider the random intercept and slope model

$$\log(\mu_{ij}) = \beta_0 + \beta_1 \text{Year}_{ij} + \beta_2 \text{Age}_i + \beta_3 \text{Gender}_i + \beta_4 \text{Skin}_i + \beta_5 \text{Exposure}_i + u_{1i} + u_{2i} \text{Year}_{ij}, \tag{17}$$

where $\left[\begin{smallmatrix} u_{1i} \\ u_{2i} \end{smallmatrix}\right] \sim N\left(0, \left[\begin{smallmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{21} & \sigma_{22}^2 \end{smallmatrix}\right]\right)$ for $i = 1, \ldots, 1683$, $1 \leq j \leq 5$. The covariates Year, Age and Skin were standardized to have mean 0 and variance 1.

Fitting this model using Algorithm 1 and MCMC, the estimated marginal posterior distributions of model parameters are shown in Figure 9 and computation times and variational lower bounds are given in Table 6. For MCMC, two chains were run simultaneously to assess convergence, each with 11,000 iterations, and the first 1000 iterations were discarded in each chain as burn-in. Partial noncentering performed very well as compared to centering and noncentering, producing posterior approximations that were closest to that of MCMC and converging in the shortest time.

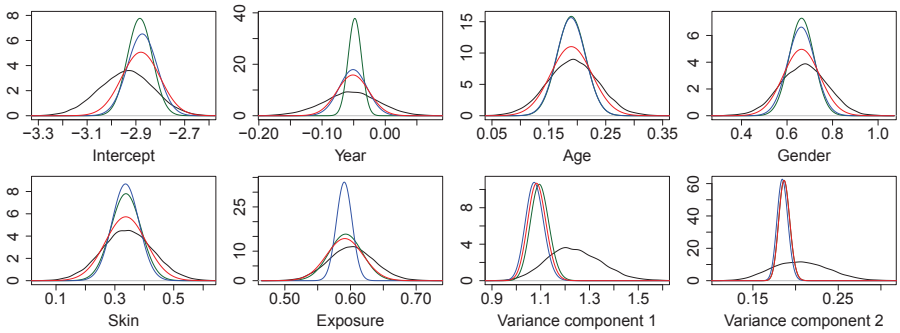

Figure 9: Skin cancer data. Marginal posteriors estimated by MCMC (black) and Algorithm 1 using the centered (green), noncentered (blue) and partially noncentered (red) parametrizations.

	noncentered	centered	partially noncentered	MCMC
Lower bound ($\mathcal{L}$)	-4054.1	-4054.1	-4051.7	–
Time (model fitting)	46.6	42.6	42.0	11113

Table 6: Skin cancer data. Lower bounds (first row) and CPU times (s) for model fitting (second row), for Algorithm 1 (different parametrizations) and MCMC.

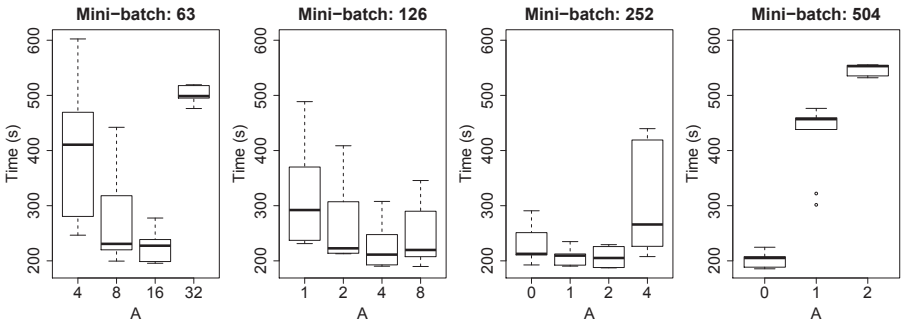

Figure 10: Skin cancer simulated data. CPU times (s) for mini-batch sizes 63, 126, 252 and 504 corresponding to different stability constants displayed in boxplots.

$\lvert B\rvert$	63	126	252	504
A	8	4	2	0
time	266.3	224.4	205.3	200.8

Table 7: Skin cancer simulated data. Shortest average time to convergence (s) for different mini-batch sizes together with corresponding stability constant A.

To investigate the performance of stochastic nonconjugate variational message passing, we simulated a much larger data set (comprised of $n = 1683 \times 15 = 25245$ subjects) from the model fitted by Algorithm 1 (using the partially noncentered parametrization). The design matrices for each cluster were replicated 15 times and responses were generated from model (17) using as parameters variational posterior means from the fitted model. For large data sets, penalized quasi-likelihood may not be feasible for use as initialization as they converge too slowly (e.g. penalized quasi-likelihhood took more than 9 mins to converge for this simulated data set). Using the fit from GLM as initialization, Algorithm 1 (using partial noncentering) took 1230.9 s to converge.

We consider mini-batch sizes $|B| \in \{63, 126, 252, 504\}$ (corresponding to 0.025%, 0.05%, 1%, and 2% of $n = 25245$) and stability constants $A \in \{0, 1, 2, 4, 8, 16, 32\}$. Larger stability constants were used for smaller mini-batch sizes. For each setting, we performed ten runs of Algorithm 2, switching to Algorithm 1 when the relative increment in the lower bound after a sweep is less than 10^{-3}. Computation times for the four mini-batch sizes corresponding to different stability constants are displayed in boxplots in Figure 10. The shortest average time to convergence for different mini-batch sizes are given in Table 7 together with corresponding stability constants A. From Figure 10, computation times were reduced by a factor of 2 or more across the different mini-batch sizes and stability constants considered. As in the previous example, Table 7 showed that larger stability constants A are preferred for smaller mini-batch sizes. The shortest average time to convergence of 200.8 s was achieved by mini-batches of size 504 with $A = 0$. This represents a reduction in computation time from Algorithm 1 by a factor of 6. Similar results can be achieved by smaller mini-batch sizes with appropriately

chosen step sizes.

Figure 11 compares the rate of convergence of standard and stochastic nonconjugate variational message passing for one of the runs where $|B| = 504$ and $A = 0$. The variational lower bound $\mathcal{L}$ is -57958 at convergence and $\log(-57957 - \mathcal{L})$ is plotted against time. Stochastic nonconjugate variational message passing took just 8 sweeps to converge in 208.0 s while standard nonconjugate variational message passing took 62 sweeps and converged in 1230.9 seconds. This represents a reduction in computation time by a factor of close to 6.

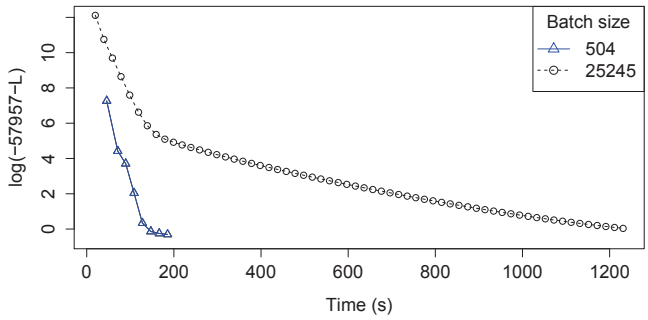

Figure 11: Skin cancer simulated data. Plot of $\log(-57957 - \mathcal{L})$ against time for mini-batch size 504 ($A = 0$) fitted using Algorithm 2 in the initial stage followed by Algorithm 1 and the whole data set fitted using Algorithm 1.

7 Conclusion

In this paper, we have extended stochastic variational inference to nonconjugate models and derived a stochastic nonconjugate variational message passing algorithm that is scalable to large data sets. The data sets that we have considered in this paper were only of moderate size. Nevertheless, we show that computation times can be reduced by applying stochastic nonconjugate variational message passing in the initial stage of optimization. The stochastic version seems computationally preferable once the number of clusters is of the order of ten thousand and above. We imagine the gain to be bigger for larger data sets and more work remains to be done in that aspect. Experimentation with various settings of stability constants A suggest that larger A is preferred for smaller mini-batch sizes. To avoid hand-tuning of step sizes, it will be useful to develop adaptive step sizes for stochastic nonconjugate variational message passing and we are currently working on extending the work of Ranganath *et al.* (2013) to nonconjuagte models. We have also shown that conflict diagnostics for identifying divergent units can be obtained as a by-product of nonconjugate variational message passing. Our diagnostics approximate the approach of Marshall and Spiegelhalter (2007) and experiments suggest relatively good agreement between the two methods. For large data sets, computation of conflict p-values using simulation-based approaches is very computationally

intensive and nonconjugate variational message passing is attractive as an alternative for obtaining prior-likelihood diagnostics or for use as an initial screening tool.

References

Amari, S. (1998). "Natural gradient works efficiently in learning." *Neural Computation*, 10: 251–276. 971

Attias, H. (1999). "Inferring parameters and structure of latent variable models by variational Bayes." In Laskey, K. and Prade, H. (eds.), *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, 21–30. San Francisco, CA: Morgan Kaufmann. 964

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer. 982

Booth, J. G. and Hobert, J. P. (1999). "Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm." *Journal of the Royal Statistical Society: Series B*, 61: 265–285. 976

Bottou, L. and Le Cun, Y. (2005). "On-line learning for very large data sets." *Applied stochastic models in business and industry*, 21: 137–151. 964

Bottou, L. and Bousquet, O. (2008). "The trade-offs of large scale learning." In Platt, J. C., Koller, D., Singer, Y. and Roweis, S. (eds.), *Advances in Neural Information Processing Systems 20*, 161–168. Red Hook, NY: Curran Associates, Inc. 964

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. MA: Addison-Wesley. 979

Breslow, N. E. and Clayton, D. G. (1993). "Approximate inference in generalized linear mixed models." *Journal of the American Statistical Association*, 88, 9–25. 963, 984

Broderick, T., Boyd, N., Wibisono, A., Wilson, A. C. and Jordan, M. I. (2013). "Streaming variational Bayes." In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z. and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 26*, 1727–1735. Red Hook, NY: Curran Associates, Inc. 964

Diggle, P. J., Heagerty, P., Liang, K. and Zeger, S. L. (2002). *Analysis of longitudinal data*. UK: Oxford University Press, 2nd edition. 1002

Donohue, M. C., Overholser, R., Xu, R. and Vaida, F. (2011). "Conditional Akaike information under generalized linear and proportional hazards mixed models." *Biometrika*, 98: 685–700. 991

Evans, M. and Moshonov, H. (2006). "Checking for prior-data conflict." *Bayesian Analysis*, 4: 893–914. 965

Farrell, P. J., Groshen, S., MacGibbon, B. and Tomberlin, T. J. (2010). "Outlier detection for a hierarchical Bayes model in a study of hospital variation in surgical procedures." *Statistical Methods in Medical Research*, 19: 601–619. 978

Fong, Y., Rue, H. and Wakefield, J. (2010). "Bayesian inference for generalised linear mixed models." *Biostatistics*, 11: 397–412. 963

Gelfand, A. E., Sahu, S. K. and Carlin, B. P. (1995). "Efficient parametrisations for normal linear mixed models." *Biometrika*, 82: 479–488. 967

—— (1996). "Efficient parametrizations for generalized linear mixed models." In Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. (eds.), *Bayesian Statistics 5*, 165–180. Oxford: Clarendon Press. 967

Ghahramani, Z. and Beal, M. J. (2001). "Propagation algorithms for variational Bayesian learning." In Leen, T. K., Dietterich, T. G. and Tresp, V. (eds.), *Advances in Neural Information Processing Systems 13*, 507–513. Cambridge, MA: MIT Press. 964

Greenberg, E. R., Baron, J. A., Stevens, M. M., Stukel, T. A., Mandel, J. S., Spencer, S. K., Elias, P. M., Lowe, N., Nierenberg, D. N., Bayrd G. and Vance, J. C. (1989). "The skin cancer prevention study: design of a clinical trial of beta-carotene among persons at high risk for nonmelanoma skin cancer." *Controlled Clinical Trials*, 10: 153–166. 990

Hoffman, M. D., Blei, D. M. and Bach, F. (2010). "Online learning for latent Dirichlet allocation." In Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R. and Culotta, A. (eds.), *Advances in Neural Information Processing Systems 23*, 856–864. Red Hook, NY: Curran Associates, Inc. 964, 978

Hoffman, M. D., Blei, D. M., Wang, C. and Paisley, J. (2013). "Stochastic variational inference." *Journal of Machine Learning Research*, 14: 1303–1347. 964, 970, 971, 972, 973

Honkela, A., Tornio, M., Raiko, T. and Karhunen, J. (2008). "Natural conjugate gradient in variational inference." In Ishikawa, M., Doya, K., Miyamoto, H. and Yamakawa, T. (eds.), *Neural Information Processing*, 305–314. Berlin: Springer-Verlag. 971

Hosmer, D. W., Lemeshow, S. and Sturdivant, R. X. (2013). *Applied Logistic Regression*. Hoboken, New Jersey: John Wiley & Sons Inc., 3rd edition. 987

Huang, A. and Wand, M. P. (2013). "Simple Marginally Noninformative Prior Distributions for Covariance Matrices." *Bayesian Analysis*, 8: 439–452. 966

Ibrahim, J. G. and Laud, P. W. (1991). "On Bayesian analysis of generalized linear models using Jeffreys's prior." *Journal of the American Statistical Association*, 86: 981–986. 1001

Jank, W. (2006). "Implementing and diagnosing the stochastic approximation EM algorithm." *Journal of Computational and Graphical Statistics*, 15: 803–829. 963, 976

Ji, C., Shen, H. and West, M. (2010). "Bounded approximations for marginal likelihoods." Available at http://ftp.stat.duke.edu/WorkingPapers/10-05.pdf. 964

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 Analysis*, 1: 535–542. 966, 974

Knowles, D. A., Minka, T. P. (2011). "Non-conjugate variational message passing for multinomial and binary regression." In Shawe-Taylor, J., Zemel, R. S., Bartlett, P., Pereira, F. and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 24*, 1701–1709. Red Hook, NY: Curran Associates, Inc. 964, 968, 969, 970

Liang, F., Cheng, Y., Song, Q., Park, J. and Yang, P. (2013). "A resampling-based stochastic approximation method for analysis of large geostatistical data." *Journal of the American Statistical Association*, 108: 325–339. 964

Liu, Q. and Pierce, D. A. (1994). "A note on Gauss-Hermite quadrature." *Biometrika*, 81: 624–629. 1000

Lunn, D., Spiegelhalter, D., Thomas, A. and Best, N. (2009). "The BUGS project: Evolution, critique and future directions." *Statistics in Medicine*, 28: 3049–3067. 982

Luts, J., Broderick, T. and Wand, M. P. (2013). "Real-time semiparametric regression." *Journal of Computational and Graphical Statistics*, (to appear). 964

Magnus, J. R. and Neudecker, H. (1988). *Matrix differential calculus with applications in statistics and econometrics*. Chichester, UK: Wiley. 1000

Marshall, E. C. and Spiegelhalter, D. J. (2007). "Identifying outliers in Bayesian hierarchical models: a simulation-based approach." *Bayesian Analysis*, 2: 409-444. 965, 978, 979, 981, 982, 993

Nott, D. J., Tan, S. L., Villani, M. and Kohn, R. (2012). "Regression density estimation with variational methods and stochastic approximation." *Journal of Computational and Graphical Statistics*, 21: 797–820. 965

Nott, D. J., Tran, M.-N., Kuk, A. Y. C., Kohn, R. (2013). "Efficient variational inference for generalized linear mixed models with large datasets." arXiv: 1307.7963. 965

Ormerod, J. T. and Wand, M. P. (2010). "Explaining variational approximations." *The American Statistician*, 64: 140–153. 968

—— (2012). "Gaussian variational approximate inference for generalized linear mixed models." *Journal of Computational and Graphical Statistics*, 21: 2–17. 963

Overstall, A. M. and Forster, J. J. (2010). "Default Bayesian model determination methods for generalised linear mixed models." *Computational Statistics and Data Analysis*, 54: 3269–3288. 966

Paisley, J., Blei, D. M. and Jordan, M. I. (2012). "Variational Bayesian inference with stochastic search." In Langford, J. and Pineau, J. (eds.), *Proceedings of the 29th International Conference on Machine Learning*, 1367–1374. Madison, WI: Omnipress. 965

Papaspiliopoulos, O., Roberts, G. O. and Sköld, M. (2003). "Non-centered parametrizations for hierarchical models and data augmentation." In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D. A., Smith, F. M. and West, M. (eds.), *Bayesian Statistics 7*, 307–326. New York: Oxford University Press. 967

—— (2007). "A general framework for the parametrization of hierarchical models." *Statistical Science*, 22: 59–73. 967

Petris, G. and Tardella, L. (2003). "A geometric approach to transdimensional Markov chain Monte Carlo." *The Canadian Journal of Statistics*, 31: 469–482. 985

Polyak, B. T. and Juditsky, A. B. (1992). "Acceleration of stochastic approximation by averaging." *SIAM Journal on Control and Optimization*, 30: 838–855. 976, 978

Presanis, A. M., Ohlssen, D., Spiegelhalter, D. J. and De Angelis, D. (2013). "Conflict diagnostics in directed acyclic graphs, with applications in Bayesian evidence synthesis." *Statistical Science*, 28: 376–397. 979

Ranganath, R., Wang, C., Blei, D. M. and Xing, E. P. (2013). "An adaptive learning rate for stochastic variational inference." In Dasgupta, S. and McAllester, D. (eds.) *JMLR W&CP: Proceedings of the 30th International Conference on Machine Learning*, 28: 298–306. 978, 993

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." *Journal of Computational and Graphical Statistics*, 9: 141–157. 963

Robbins, H. and Monro, S. (1951). "A stochastic approximation method." *Annals of Mathematical Statistics* 22: 400–407. 964, 973

Roux, N. L., Schmidt, M. and Bach, F. (2012). "A stochastic gradient method with an exponential convergence rate for finite training sets." In Pereira, F., Burges, C. J. C., Bottou, L. and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 25*, 2663–2671. Red Hook, NY: Curran Associates, Inc. 978

Salimans, T. and Knowles, D. A. (2013). "Fixed-form variational posterior approximation through stochastic linear regression." *Bayesian Analysis*, 4: 837–882. 965, 982

Sato, M. (2001). "Online model selection based on the variational Bayes." *Neural Computation*, 13: 1649–1681. 972

Scheel, I., Green, P. J. and Rougier, J. C. (2011). "A graphical diagnostic for identifying influential model choices in Bayesian hierarchical models." *Scandinavian Journal of Statistics*, 38: 529–550. 965

Spall, J. C. (2003). *Introduction to stochastic search and optimization: estimation, simulation and control*. New Jersey: Wiley. 973, 974, 976, 977

Sturtz, S., Ligges, U., and Gelman, A. (2005). "R2WinBUGS: A package for running WinBUGS from R." *Journal of Statistical Software*, 12: 1–16. 982

Tan, L. S. L. and Nott, D. J. (2013). "Variational inference for generalized linear mixed models using partially noncentered parametrizations." *Statistical Science*, 28: 168–188. 963, 964, 965, 967, 968, 969, 970, 985, 999, 1000, 1001

Thall, P. F. and Vail, S. C. (1990). "Some covariance models for longitudinal count data with overdispersion." *Biometrics*, 46: 657–671. 984

Thara, R., Henrietta, M., Joseph, A., Rajkumar, S. and Eaton, W. (1994). "Ten year course of schizophrenia - the Madras longitudinal study." *Acta Psychiatrica Scandinavica*, 90: 329–336. 1002

Tseng, P. (1998). *An incremental gradient(-projection) method with momentum term and adaptive stepsize rule. SIAM Journal on Optimization*, 8: 506–531. 978

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. NY: Springer, 4th edition. 974

Wand, M. P. (2013). "Fully simplified multivariate normal updates in non-conjugate variational message passing." Available at `http://www.uow.edu.au/~mwand/fsupap.pdf`. 972

Wang, C., Paisley, J. and Blei, D. M. (2011). "Online variational inference for the hierarchical Dirichlet process." In Gordon, G., Dunson, D. and Dudik, M. (eds.) *JMLR W&CP: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, 15: 752–760. 964

Wang, B. and Titterington, D. M. (2005). "Inadequacy of interval estimates corresponding to variational Bayesian approximations." In Cowell, R. G. and Ghahramani, Z. (eds.), *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*, 373–380. Society for Artificial Intelligence and Statistics. 982

Winn, J. and Bishop, C. M. (2005). "Variational message passing." *Journal of Machine Learning Research*, 6: 661–694. 964

Xiao, L. (2010). "Dual averaging methods for regularized stochastic learning and online optimization." *Journal of Machine Learning Research*, 11: 2543–2596. 978

Zhao, H. and Marriott, P. (2013). "Diagnostics for variational Bayes approximations." arXiv:1309.5117. 982

Zhu, H. T. and Lee, S. Y. (2002). "Analysis of generalized linear mixed models via a stochastic approximation algorithm with Markov chain Monte Carlo method." *Statistics and Computing*, 12: 175–183.

Appendix A: Natural gradients for conjugate factors

In this section, we present the simplified updates and natural gradients for conjugate factors in nonconjugate variational message passing. Let $N(\theta_i)$ denote the neighbourhood of θ_i in the factor graph of $p(y,\theta)$ (see Tan and Nott 2013). Suppose $p(y,\theta) = \prod_a f_a(y,\theta)$ and each factor f_a in $N(\theta_i)$ is conjugate to $q_i(\theta_i|\lambda_i)$, say

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

where $\theta_{-i} = (\theta_1,\ldots,\theta_{i-1},\theta_{i+1},\ldots,\theta_m)$. Then

$$\nabla_{\lambda_i}\mathcal{L} = \mathcal{V}_i(\lambda_i)\left[\sum\nolimits_{a\in N(\theta_i)} E_q\{g_a(y,\theta_{-i})\} - \lambda_i\right]$$

and the nonconjugate variational message passing update in (7) reduces to

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

Note that $E_q\{g_a(y,\theta_{-i})\}$ does not depend on λ_i. The natural gradient in (9) can also be simplified as

$$\widetilde{\nabla}_{\lambda_i}\mathcal{L} = \sum_{a\in N(\theta_i)} E_q\{g_a(y,\theta_{-i})\} - \lambda_i.$$

Appendix B: Notation and Updates in Algorithm 2

For Poisson responses,

$$g_i = E_i \odot \exp\left\{V_i\mu_{q(\beta)} + Z_i\mu_{q(\tilde{\alpha}_i)} + \tfrac{1}{2}\mathrm{diag}(V_i\Sigma_{q(\beta)}V_i^T + Z_i\Sigma_{q(\tilde{\alpha}_i)}Z_i^T)\right\} \text{ and } F_i = \mathrm{diag}(g_i)$$

for $i = 1,\ldots,n$. For Bernoulli responses,

$$g_i = B^{(1)}(\mu_i^q,\sigma_i^q) \text{ and } F_i = \mathrm{diag}(B^{(2)}(\mu_i^q,\sigma_i^q))$$

for $i = 1,\ldots,n$, where $\mu_i^q = V_i\mu_{q(\beta)} + Z_i\mu_{q(\tilde{\alpha}_i)}$ and $\sigma_i^q = \sqrt{\mathrm{diag}(V_i\Sigma_{q(\beta)}V_i^T + Z_i\Sigma_{q(\tilde{\alpha}_i)}Z_i^T)}$. We have

$$B^{(r)}(\mu,\sigma) = \int_{-\infty}^{\infty} b^{(r)}(\sigma x+\mu)\frac{1}{\sqrt{2\pi}}\exp(-x^2)\,dx,$$

where $b(x) = \log\{1+\exp(x)\}$ and $b^{(r)}(x)$ denotes the rth derivative of $b(\cdot)$ with respect to x. If μ and σ are vectors, say $\mu = \begin{bmatrix}1\\2\\3\end{bmatrix}$ and $\sigma = \begin{bmatrix}4\\5\\6\end{bmatrix}$, then $B^{(r)}(\mu,\sigma) = \begin{bmatrix}B^{(r)}(1,4)\\B^{(r)}(2,5)\\B^{(r)}(3,6)\end{bmatrix}$. The terms $B^{(r)}(\mu,\sigma)$, $r = 0,1,2$ may be evaluated efficiently using adaptive Gauss-Hermite quadrature (Liu and Pierce 1994). More details can be found in Tan and Nott (2013).

The updates in step 2 of Algorithm 2 are taken directly from the nonconjugate variational message passing algorithm for GLMMs in Tan and Nott (2013). To derive the updates in step 3, let us first introduce the following notation for specification of the natural parameter vectors λ_β and λ_D. For a $d\times d$ square matrix A, let $\mathrm{vec}(A)$ denote the $d^2\times 1$ vector obtained by stacking the columns of A under each other, from left to right in order and $\mathrm{vech}(A)$ denotes the $\frac{1}{2}d(d+1)\times 1$ vector obtained from $\mathrm{vec}(A)$ by eliminating all supradiagonal elements of A. The matrix D_d is a unique $d^2\times\frac{1}{2}d(d+1)$ matrix that transforms $\mathrm{vech}(A)$ into $\mathrm{vec}(A)$ if A is symmetric, that is, $D_d\mathrm{vech}(A) = \mathrm{vec}(A)$. See Magnus and Neudecker (1988) for more details. We have

$$\lambda_\beta = \begin{bmatrix}-\frac{1}{2}D_p^T\mathrm{vec}(\Sigma_{q(\beta)}^{-1})\\ \Sigma_{q(\beta)}^{-1}\mu_{q(\beta)}\end{bmatrix} \quad\text{and}\quad \lambda_D = \begin{bmatrix}-\frac{1}{2}\mathrm{vec}(S_{q(D)})\\ -\frac{1}{2}(\nu_{q(D)}+r+1)\end{bmatrix}.$$

From (13),

$$\begin{bmatrix}-\frac{1}{2}D_p^T\mathrm{vec}({\Sigma_{q(\beta)}^{(t+1)}}^{-1})\\ {\Sigma_{q(\beta)}^{(t+1)}}^{-1}\mu_{q(\beta)}^{(t+1)}\end{bmatrix} = (1-a_t)\begin{bmatrix}-\frac{1}{2}D_p^T\mathrm{vec}({\Sigma_{q(\beta)}^{(t)}}^{-1})\\ {\Sigma_{q(\beta)}^{(t)}}^{-1}\mu_{q(\beta)}^{(t)}\end{bmatrix} + a_t\begin{bmatrix}-\frac{1}{2}D_p^T\mathrm{vec}(\hat{\Sigma}_{q(\beta)}^{-1})\\ \hat{\Sigma}_{q(\beta)}^{-1}\hat{\mu}_{q(\beta)}\end{bmatrix}, \quad (18)$$

where

$$\hat{\Sigma}_{q(\beta)} = \left[\Sigma_\beta^{-1} + \frac{n}{|B|}\sum_{i\in B}\{\nu_{q(D)}\tilde{W}_i^T S_{q(D)}^{-1}\tilde{W}_i + V_i^T F_i V_i\}\right]^{-1} \text{ and } \hat{\mu}_{q(\beta)} = \mu_{q(\beta)}^{(t)}$$
$$+\hat{\Sigma}_{q(\beta)}\left[\frac{n}{|B|}\sum_{i\in B}\{\nu_{q(D)}\tilde{W}_i^T S_{q(D)}^{-1}(\mu_{q(\tilde{\alpha}_i)} - \tilde{W}_i\mu_{q(\beta)}^{(t)}) + V_i^T(y_i - G_i)\} - \Sigma_\beta^{-1}\mu_{q(\beta)}^{(t)}\right].$$

Expressions for $\hat{\Sigma}_{q(\beta)}$ and $\hat{\mu}_{q(\beta)}$ can be deduced from Algorithm 3 of Tan and Nott (2013). The first line in (18) gives

$$\Sigma_{q(\beta)}^{(t+1)} = \left\{(1-a_t){\Sigma_{q(\beta)}^{(t)}}^{-1} + a_t\hat{\Sigma}_{q(\beta)}^{-1}\right\}^{-1},$$

which is the update for $\Sigma_{q(\beta)}$ in Algorithm 2. The second line in (18) gives

$$\begin{aligned}\mu_{q(\beta)}^{(t+1)} &= \Sigma_{q(\beta)}^{(t+1)}\left\{(1-a_t){\Sigma_{q(\beta)}^{(t)}}^{-1}\mu_{q(\beta)}^{(t)} + a_t\hat{\Sigma}_{q(\beta)}^{-1}\hat{\mu}_{q(\beta)}\right\}\\ &= \Sigma_{q(\beta)}^{(t+1)}\left\{\left({\Sigma_{q(\beta)}^{(t+1)}}^{-1} - a_t\hat{\Sigma}_{q(\beta)}^{-1}\right)\mu_{q(\beta)}^{(t)} + a_t\hat{\Sigma}_{q(\beta)}^{-1}\hat{\mu}_{q(\beta)}\right\}\\ &= \mu_{q(\beta)}^{(t)} + a_t\Sigma_{q(\beta)}^{(t+1)}\hat{\Sigma}_{q(\beta)}^{-1}\left(\hat{\mu}_{q(\beta)} - \mu_{q(\beta)}^{(t)}\right)\\ &= \mu_{q(\beta)}^{(t)} + a_t\Sigma_{q(\beta)}^{(t+1)}\left[\frac{n}{|B|}\sum_{i\in B}\left\{\nu_{q(D)}\tilde{W}_i^T S_{q(D)}^{-1}(\mu_{q(\tilde{\alpha}_i)} - \tilde{W}_i\mu_{q(\beta)}^{(t)}) + V_i^T(y_i - G_i)\right\}\right.\\ &\qquad \left.-\Sigma_\beta^{-1}\mu_{q(\beta)}^{(t)}\right],\end{aligned}$$

which is the update for $\mu_{q(\beta)}$ in Algorithm 2.

Similarly, from (13),

$$\begin{bmatrix}-\frac{1}{2}\mathrm{vec}(S_{q(D)}^{(t+1)})\\ -\frac{1}{2}(\nu_{q(D)}^{(t+1)}+r+1)\end{bmatrix} = (1-a_t)\begin{bmatrix}-\frac{1}{2}\mathrm{vec}(S_{q(D)}^{(t)})\\ -\frac{1}{2}(\nu_{q(D)}^{(t)}+r+1)\end{bmatrix} + a_t\begin{bmatrix}-\frac{1}{2}\mathrm{vec}(\hat{S}_{q(D)})\\ -\frac{1}{2}(\hat{\nu}_{q(D)}+r+1)\end{bmatrix}, \quad (19)$$

where $\hat{S}_{q(D)}$ and $\hat{\nu}_{q(D)}$ can be deduced from Tan and Nott (2013) as $\hat{\nu}_{q(D)} = \nu + n$ and $\hat{S}_{q(D)} = S + \frac{n}{|B|}\sum_{i\in B}\{(\mu_{q(\tilde{\alpha}_i)} - \tilde{W}_i\mu_{q(\beta)})(\mu_{q(\tilde{\alpha}_i)} - \tilde{W}_i\mu_{q(\beta)})^T + \Sigma_{q(\tilde{\alpha}_i)} + \tilde{W}_i\Sigma_{q(\beta)}\tilde{W}_i^T\}$. The updates for $S_{q(D)}$ and $\nu_{q(D)}$ in Algorithm 2 can be obtained by simplifying (19).

Appendix C: Generating likelihood replicates

In the centered parametrization,

$$\eta_i = Z_i\alpha_i + X_{gi}\beta_g \quad \text{where} \quad \alpha_i = C_i\beta_c + u_i \sim N(C_i\beta_c, D)$$

for $i = 1,\ldots,n$. To generate likelihood replicates α_i^{lik} from $p(\alpha_i|y_i)$ in the cross-validatory approach, we consider Jeffreys's prior for the centered random effects α_i. Jeffreys's prior is defined as $p(\alpha_i) \propto \sqrt{|I(\alpha_i)|}$, where $I(\alpha_i)$ is the Fisher information matrix of α_i. For Poisson and logistic GLMMs, it can be shown that $p(\alpha_i) \propto |Z_i^T Q_i Z_i|^{\frac{1}{2}}$, where Q_i is an $n_i \times n_i$ diagonal matrix (see, e.g. Ibrahim and Laud 1991). Definitions of Q_i are given in Section 3.1. In general, we will need to consider β_g as a nuisance parameter. Following the discussion in Section 5.1, we generate a β_g from $p(\beta_g|y_{-i})$ and simulate α_i^{lik} from $p(y_i|\alpha_i,\beta^g)p(\alpha_i)$ where $p(\alpha_i)$ is Jeffreys's prior. For Poisson GLMMs,

$$p(y_i|\alpha_i,\beta^g)p(\alpha_i) \propto \exp\{y_i^T(\log E_i + Z_i\alpha_i + X_{gi}\beta_g) - E_i^T\exp(Z_i\alpha_i + X_{gi}\beta_g)\}|Z_i^T Q_i Z_i|^{\frac{1}{2}}.$$

For logistic GLMMs,

$$p(y_i|\alpha_i,\beta^g)p(\alpha_i) \propto \exp[y_i^T(Z_i\alpha_i + X_{gi}\beta_g) - 1_{n_i}^T\log\{1_{n_i} + \exp(Z_i\alpha_i + X_{gi}\beta_g)\}]|Z_i^T Q_i Z_i|^{\frac{1}{2}}.$$

Appendix D: Madras schizophrenia example

The Madras schizophrenia study (Thara *et al.* 1994) contains records of the psychiatric symptoms of 86 patients in the first year after initial hospitalization. This data set has been analyzed by Diggle *et al.* (2002) and is available at `http://faculty.washington.edu/heagerty/Books/AnalysisLongitudinal/datasets.html`. The reponse y_{ij} is 1 if the symptom "thought disorder" is present and 0 otherwise. We consider the covariates, age at onset of disease (Age = 1 if patient is at least 20 years old and 0 otherwise), sex of patient (Gender = 1 if female and 0 otherwise) and number of months since hospitalization when symptom was recorded (t). We consider the logistic random effects model:

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{Gender}_i + \beta_3 t_{ij} + \beta_4 \text{Age}_i \times t_{ij} + \beta_5 \text{Gender}_i \times t_{ij} + u_i,$$

where $u_i \sim N(0, \sigma^2)$ for $i = 1, \ldots, 86$, $1 \leq j \leq 12$. We report both one-sided (upper-tail) and two-sided conflict p-values for this example. The upper-tail one-sided conflict p-values are useful for identifying patients with an unusually large number of "thought disorders" while the two-sided conflict p-values can be used to detect patients with either more or less than the expected number of "thought disorders".

In the simulation-based approaches, β_3, β_4 and β_5 have to be regarded as nuisance parameters under the centered parametrization (see Appendix C). For each model fitting via MCMC, two chains were run simultaneously to assess convergence, each with 26,000 iterations, and the first 1000 iterations were discarded in each chain as burn-in. Simulation-based conflict p-values were calculated based on the remaining 50,000 simulations. For the cross-validatory approach, model refitting took a total of 372 s $\times$ 86 (more than 8 hours) to complete in OpenBUGS. Simulation of prior and likelihood replicates of the centered random effects α_i was performed in `R`. Assuming Jeffreys's prior for α_i, likelihood replicates were simulated using adaptive rejection Metropolis sampling.

Variational lower bounds and CPU times taken for model fitting and computing conflict p-values by Algorithm 1 (different parametrizations) and MCMC (full-data approach) are given in Table 8. Figure 12 shows the marginal posteriors of parameters estimated using MCMC and Algorithm 1. The partially noncentered parametrization took the shortest time to converge and attained the highest lower bound. From Figure 12, partial noncentering produced better posterior approximations for β_0, β_1 and β_2 than both centering and noncentering. For β_3, β_4, β_5, partial centering performed better than centering but did not do as well as noncentering.

Cross-validatory conflict p-values are plotted against conflict p-values from nonconjugate variational message passing using the partially noncentered parametrization in Figure 13. The left plot shows the upper-tail one-sided p-values while the right plot shows the two-sided p-values. The mean absolute difference in z-scores for nonconjugate variational message passing and the simulation-based full-data approach relative to the cross-validatory approach are given in Table 8. Figure 13 shows that the agreement between the cross-validatory approach and nonconjugate variational message passing is better for the one-sided p-values than in the two-sided case. This is expected as any

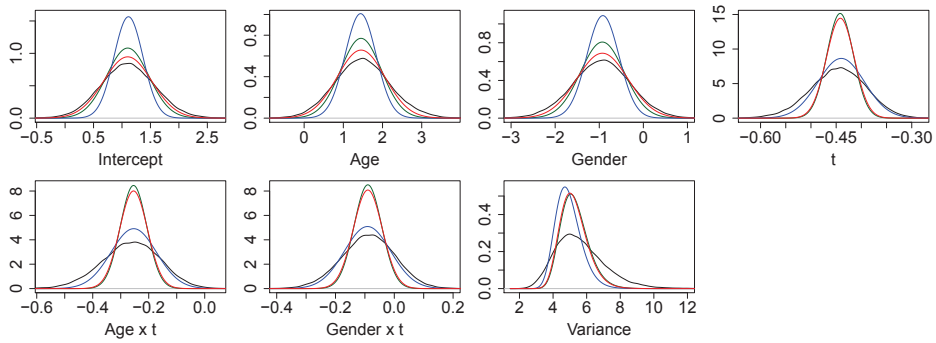

Figure 12: Madras data. Marginal posteriors estimated by MCMC (black) and Algorithm 1 using the centered (green), noncentered (blue) and partially noncentered (red) parametrizations.

	non-centered	centered	partially non-centered	MCMC (full-data)
Lower bound ($\mathcal{L}$)	-407.9	-407.1	-406.6	–
Time (model fitting)	6.4	6.0	5.0	372
Time (compute conflict p-values)	0.1	0.1	0.1	16266.9
Mean abs difference in z-scores (1-sided)	0.115	0.102	0.104	0.040
Mean abs difference in z-scores (2-sided)	0.227	0.201	0.204	0.069

Table 8: Madras data. Variational lower bounds (first row), CPU times (s) for model fitting (second row) and calculating conflict p-values (third row) and mean absolute difference in z-scores (relative to cross-validatory approach) for one-sided (fourth row) and two-sided (fifth row) p-values, for Algorithm 1 (different parametrizations) and MCMC (full-data).

discrepancy between the two sets of p-values will be doubled in the two-sided case. However, we note that agreement at the extremes is still relatively good. For this example, the simulation-based full-data approach performed better in terms of z-scores than nonconjugate variational message passing. This is likely due to the fact that in this case, the variational posterior does not provide as good an approximation to the true posterior as in Examples 6.1 and 6.2. However, nonconjugate variational message passing remains useful as a screening tool as the computation time required to compute conflict p-values even in the simulation-based full-data approach is quite significant. Finally, outliers (at the 0.05 level) identified by the cross-validatory approach and nonconjugate variational message passing using the partially noncentered parametrization are identical in this example. Conflict p-values for these outliers are shown in Table 9.

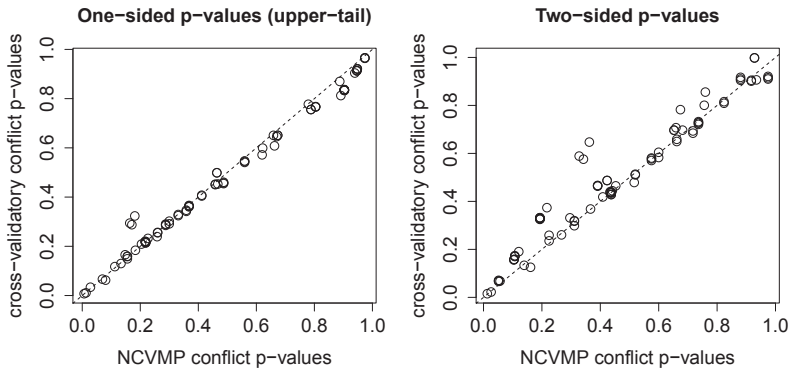

Figure 13: Madras data. Cross-validatory conflict p-values plotted against conflict p-values from nonconjugate variational message passing with partial noncentering.

One-sided p-values (upper-tail)		
Patient	$p_{i,\text{con}}^{\text{CV}}$	$p_{i,\text{con}}^{\text{NCVMP}}$
14	0.034	0.028
27	0.011	0.013
68	0.008	0.007

Two-sided p-values		
Patient	$p_{i,\text{con}}^{\text{CV}}$	$p_{i,\text{con}}^{\text{NCVMP}}$
25	0.023	0.026
56	0.016	0.013

Table 9: Madras data. Conflict p-values for outliers from cross-validatory approach and nonconjugate variational message passing using partially noncentered parametrization.

Acknowledgments

Linda Tan was partially supported by the Singapore-Delft Water Alliance's tropical reservoir research programme and David Nott's research was supported by a Singapore Ministry of Education Academic Research Fund Tier 2 grant (R-155-000-143-112). The authors would like to thank Matt Wand for making available his work on fully simplified multivariate normal nonconjugate variational message passing updates, and the referees, associate editor and editor for their suggestions which have helped improve the manuscript.