A stochastic variational framework for fitting and diagnosing generalized linear mixed models

Siew Li Tan
National University of Singapore, Singapore

David J. Nott
National University of Singapore, Singapore

Summary. Variational Bayes computational methods are attracting increasing interest because of their ability to scale to large data sets. Here, application of the non-conjugate variational message passing (NCVMP) algorithm to generalized linear mixed models (GLMMs) is considered. Extending recently developed methods in stochastic variational inference to the non-conjugate case, we combine NCVMP updates with stochastic natural gradient optimization of the variational lower bound to derive a stochastic NCVMP algorithm for fitting GLMMs, scalable to large data sets. We demonstrate that convergence for moderate-sized data sets can be accelerated by using stochastic NCVMP initially before switching to standard NCVMP. Finally, we show that the way variational message passing updates separate into messages from above and below a node in a hierarchical model facilitates an automatic computation of diagnostics for prior-likelihood conflict in the NCVMP algorithm which are very useful for model criticism.

Keywords: Hierarchical model; Identify divergent units; Large longitudinal data; Non-conjugate model; Stochastic approximation; Variational Bayes

1. Introduction

Variational Bayes (VB) (Attias, 2000) computational methods are very attractive for Bayesian inference with large data sets. Nevertheless, for very large data sets, batch variational algorithms for many common models become increasingly inefficient as local variational parameters associated to each unit in the data set have to be updated at every iteration (Hoffman et al., 2010). They are also unsuitable in online settings where data arrive continuously as the algorithm can never complete one iteration. On the other hand, stochastic approximation (SA) methods (Robbins and Monro, 1951) are able to make use of gradient estimates from mini-batches of data so that computational cost is reduced. Recently, Tan and Nott (2012) showed how to implement an algorithm called non-conjugate
variational message passing (NCVMP) (Knowles and Minka, 2011) for inference in generalized linear mixed models (GLMMs). Similar to other batch VB algorithms for models with observation specific latent variables, the NCVMP algorithm for GLMMs has to iterate between updating local variational parameters associated with individual observations and global variational parameters. We propose to combine NCVMP updates with stochastic natural gradient optimization of the variational lower bound when fitting GLMMs to large data sets. In addition, we show that diagnostics for prior-likelihood conflict can be obtained as an automatic by-product of NCVMP. Our paper makes three contributions. First, we extend stochastic variational inference to non-conjugate models and derive a stochastic NCVMP algorithm for fitting GLMMs, scalable to large data sets. Second, we show that convergence for moderate-sized data sets can be accelerated by using stochastic NCVMP in the first few iterations before switching to standard NCVMP. Third, we show that the way variational message passing (VMP) updates separate into messages from above and below a node in a hierarchical model facilitates an automatic computation of diagnostics for prior-likelihood conflict useful for model criticism. Our “mixed message” diagnostics can be shown to approximate existing diagnostics in the statistical literature, namely the conflict diagnostics of Marshall and Spiegelhalter (2007).

VMP is an algorithmic implementation of VB developed by Winn and Bishop (2005) which can be applied to a very general class of conjugate-exponential models. VMP proceeds by passing messages between nodes in a directed graph and the posterior distribution associated with a particular node can be updated using only local operations at the node. Knowles and Minka (2011) developed NCVMP to extend VMP to non-conjugate models, allowing more choice in the computation of expectations in the variational lower bound. Tan and Nott (2012) recently applied NCVMP to GLMMs using a partially non-centered parametrization (PNCP). The PNCP lies on the continuum between the centered parametrization (CP) and non-centered parametrization (NCP) and is data-dependent (Papaspiliopoulos et al., 2003, 2007). The PNCP has previously been used to accelerate MCMC (Markov Chain Monte Carlo) and EM (expectation maximization) algorithms for hierarchical models. In the VB context, Tan and Nott (2012) showed that the PNCP is able to automatically determine a parametrization that is close to optimal and improve convergence while resulting in more accurate approximations statistically. In this paper, we derive a stochastic version of NCVMP for GLMMs employing the PNCP.

Recent development in VB methodology have branched out to online inference and stochastic optimization so that VB has become a viable approach for handling data sets which are too large for standard batch VB algorithms. For computational efficiency, VB methods often rely on analytic solutions to integrals and conjugacy in the posterior. This limits the type of approximations and posteriors VB can handle. To overcome this restriction, Salimans and Knowles (2012) proposed a SA algorithm that does not require analytic evaluation of integrals, extending the VB approach to any posterior that is available in closed form.
up to the proportionality constant. Hierarchical extensions of the basic approach allow the method to be made arbitrarily precise. Paisley et al. (2012) proposed a stochastic optimization algorithm using control variates that allows direct maximization of the variational lower bound involving intractable integrals. Similar algorithms were considered by Ji et al. (2010) and Nott et al. (2012). Hoffman et al. (2010) and Wang et al. (2011) developed online VB algorithms for latent Dirichlet allocation and the hierarchical Dirichlet process respectively using stochastic natural gradient optimization of the variational lower bound. Hoffman et al. (2012) generalized these methods to derive stochastic variational inference for conjugate-exponential family models and demonstrated that for large data sets, stochastic variational inference converges faster than batch VB and to a better model. The key idea in stochastic gradient optimization is to use only a random subset of the data at each iteration to approximate the true gradient over the whole data. This can result in significant reduction in computational costs for large data sets. For instance, Bottou and Cun (2005) and Bottou and Bousquet (2008) showed that well-designed stochastic gradient algorithms outperform batch algorithms in large-scale learning problems. Welling and Teh (2011) combined stochastic gradient optimization with Langevin dynamics for Bayesian learning of large-scale data sets and this algorithm was extended to the stochastic gradient Fisher scoring algorithm by Ahn et al. (2012).

Amari (1998) discussed the use of natural gradient in stochastic gradient optimization and pointed out that the steepest direction is given by the natural gradient when the parameter space is not Euclidean but has a Riemannian metric structure. Natural gradient online learning was also shown to be statistically efficient. However, the use of the Riemannian structure of the predictive distribution in natural gradient learning is often complicated and Honkela et al. (2008) proposed using the Riemannian structure of the factorized variational posterior assumed in VB instead. This is very useful for models not in the conjugate-exponential family. Honkela et al. (2008) also showed that replacing the ordinary gradient in the conjugate gradient algorithm with the natural gradient can speed up variational learning. The VB algorithm was shown to be a type of natural gradient method in Sato (2001) and an online VB algorithm with a model selection mechanism was derived using SA for Gaussian mixture models. In the optimization of the variational lower bound, Hoffman et al. (2012) justify that the Riemannian metric is more appropriate than Euclidean distance since a large change in the parameter can lead to a small change in Kullback-Leibler (KL) divergence and vice versa. Therefore, we consider the natural gradient instead of the ordinary gradient of the variational lower bound in the stochastic optimization. For non-conjugate models, we will have to pre-multiply the gradient with the Fisher information matrix to obtain the natural gradient, unlike the conjugate-exponential case (Hoffman et al. 2012).

GLMMs extend generalized linear models (GLMs) via the introduction of random effects to account for within-subject association and have wide applications. Estimation of GLMMs using maximum likelihood is challenging as
the integral over the random effects is intractable. Computationally intensive methods such as numerical quadrature or MCMC can be used to approximate the intractable integrals. Other approximate methods include penalized quasi-likelihood (Breslow et al. 1993) and Laplace approximation and its extensions (Raudenbush et al. 2000). Ormerod and Wand (2012) introduced Gaussian variational approximation for GLMMs and Fong et al. (2010) illustrated the use of integrated nested Laplace approximations for fitting GLMMs using a Bayesian approach. SA has also been used in combination with MCMC (Zhu et al. 2002) and the EM algorithm (Jank 2006) for fitting GLMMs. In this paper, we focus on Poisson and logistic mixed models and their applications in longitudinal data analysis (Fitzmaurice et al. 2004).

Model checking is an important part of statistical analyses. In the Bayesian approach, assumptions are made in the form of a sampling model and prior, and prior-likelihood conflict arises when the observed data are very unlikely under the prior model. A discussion of how to assess whether there is prior-data conflict and whether its effects can be ignored can be found in Evans and Moshonov (2006). Recently, 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 recent review of other methods in Bayesian model criticism. Marshall and Spiegelhalter (2007) proposed a diagnostic test for identifying divergent units (units that do not appear to be drawn from assumed underlying distributions) 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 VMP framework and that VMP facilitates an automatic computation of diagnostics for prior-likelihood conflict. These diagnostics are useful for model criticism and we focus on NCVMP for GLMMs.

The paper is organized as follows. Section 2 specifies the model and describes a PNCP for the GLMM. A stochastic version of the NCVMP algorithm is derived in Section 3 and Section 4 describes how VMP facilitates automatic computation of diagnostics for prior-likelihood conflict. Section 5 considers examples including real and simulated data and Section 6 concludes.

2. The GLMM with a PNCP

In this paper, we concentrate on one-parameter exponential family models and consider a PNCP for the GLMM. Let $y_{ij}$ denote the $j$th response in cluster $i$, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$. We assume that conditional on a vector of random effects $u_i$, $y_{ij}$ is independently distributed as

$$y_{ij}|u_i \sim \exp \{y_{ij}\zeta_{ij} - b(\zeta_{ij}) + c(y_{ij})\}$$
where \( \zeta_{ij} \) 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 \( y_{ij} \), 
\[ \mu_{ij} = b(g(y_{ij})|u_i) \]

with 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, \( \beta \) is a \( p \times 1 \) vector of unknown fixed regression parameters and the \( r \times 1 \) random effects \( u_i \), \( i = 1, \ldots, n \) are independently distributed as \( N(0, D) \). We focus on logistic and Poisson models. For example, if \( y_{ij} \sim \text{Bernoulli}(\mu_{ij}) \), then \( b(x) = \log(1 + \exp(x)) \), \( c(x) = 0 \) and we let \( g(\mu_{ij}) = \log(\mu_{ij}/(1 - \mu_{ij})) \). For Poisson models, we allow for an offset so that if \( y_{ij} \sim \text{Poisson}(\mu_{ij}) \) where \( \mu_{ij} = E_{ij} \lambda_{ij} \) and \( \log \lambda_{ij} = \eta_{ij} \), we let \( g(\mu_{ij}) = \log(\mu_{ij}/E_{ij}) \) with \( b(x) = \exp(x) \) and \( c(x) = -\log(x!) \).

For the \( i \)-th cluster, let \( y_i = (y_{i1}, \ldots, y_{in_i})^T \), \( X_i = (X_{i1}^T, \ldots, X_{in_i}^T)^T \), \( Z_i = (Z_{i1}^T, \ldots, Z_{in_i}^T)^T \) and \( \eta_i = (\eta_{i1}, \ldots, \eta_{in_i})^T \). We assume that the first columns of \( X_i \) and \( Z_i \) (if non-null) are formed from a vector of ones of length \( n_i \), and the columns of \( Z_i \) are a subset of the columns of \( X_i \). We partition \( X_i \) as \( [X_i^R \quad 1_{n_i} x_i^{ST} \quad X_i^G] \) with \( Z_i = X_i^R \) and \( \beta \) as \( (\beta^T, \beta^ST, \beta^G)^T \) accordingly. Note that \( \beta^S \) is a vector of length \( s \) consisting of all the parameters corresponding to subject specific covariates. Then 

\[
\eta_i = X_i^R(\beta^R + u_i) + 1_{n_i} x_i^{ST} \beta^S + X_i^G \beta^G
\]

and \( \beta^RS = (\beta^R, \beta^ST)^T \). We introduce \( \alpha_i = C_i \beta^RS + u_i \) and \( \tilde{\alpha}_i = u_i - W_i C_i \beta^RS \), where \( W_i \) is an \( r \times r \) matrix controlling the degree of centering. \( W_i = 0 \) corresponds to the CP while \( W_i = I \) corresponds to the NCP. The PNCP is 

\[
\eta_i = \tilde{C}_i \beta + X_i^R \tilde{\alpha}_i
\]

where \( \tilde{C}_i = [X_i^R W_i C_i \quad X_i^G] \). Let \( \hat{W}_i = [(I_r - W_i) C_i \quad 0_{r \times (p-r)}] \) so that \( \tilde{\alpha}_i \sim N(\hat{W}_i \beta, D) \). From Tan and Nott (2012), we set \( W_i = (I_f + D^{-1})^{-1} D^{-1} \) where 
\( I_f = \sum_{j=1}^{n_i} y_{ij} X_{ij}^R X_{ij}^R \) for Poisson models and 
\( I_f = \sum_{j=1}^{n_i} \exp(\eta_{ij}) X_{ij}^R X_{ij}^R \) for logistic models. We approximate \( D \) and \( \eta_{ij} \) by their posterior means in the actual implementation.

For Bayesian inference, we specify a diffuse prior, \( N(0, \Sigma_\beta) \), on \( \beta \) where \( \Sigma_\beta \) is large and an independent inverse Wishart prior, \( IW(\nu, S) \) on \( D \). Following Kass and Natarajan (2006), we set \( \nu = r \) and \( S = r \hat{R} \) with 

\[
\hat{R} = c \left\{ \frac{1}{n} \sum_{i=1}^{n} Z_i^T M_i(\hat{\beta}) Z_i \right\}^{-1},
\]

where \( M_i(\hat{\beta}) \) denotes the usual \( n_i \times n_i \) diagonal GLM weight matrix with diagonal elements \( \{v(\hat{\mu}_{ij}) g'(\hat{\mu}_{ij})^2\}^{-1} \), \( v(\cdot) \) is the variance function and \( g(\cdot) \) is the link function.
function. Here, \( \hat{\mu}_{ij} = g^{-1}(X_{ij}^T \hat{\beta} + Z_{ij}^T \hat{u}_i) \) where \( \hat{u}_i \) is set as 0 for \( i = 1, ..., n \) and \( \hat{\beta} \) is an estimate of the regression coefficients from the GLM obtained by pooling all the data and setting \( u_i = 0 \) for all \( i \). The value of \( c \) is an inflation factor representing how much within-cluster variability should be increased in determining \( \hat{R} \). We used \( c = 1 \) for the examples in this paper.

3. Stochastic variational inference for GLMMs

In the GLMM, the fixed effects \( \beta \) and the random effects covariance matrix \( D \) can be regarded as “global” variables which are common across clusters while the partially centered random effects \( \tilde{\alpha}_i, i = 1, ..., n \), can be thought of as “local” variables associated only with the individual units. Tan and Nott [2012] proposed fitting GLMMs with a NCVMP algorithm that iterates between updating variational parameters of the local variables and re-estimating variational parameters of the global variables. For large data sets, this procedure is inefficient as every unit in the data set has to be analysed before variational parameters of the global variables can be updated each time. Hoffman et al. [2012] suggested using SA methods (Robbins and Monro, 1951) to derive more efficient algorithms. The key idea is to use stochastic gradients computed on mini-batches to approximate the true gradient over the whole dataset. They considered conjugate-exponential family models with applications on latent Dirichlet allocation and the hierarchical Dirichlet process. We extend SA methods to non-conjugate models, focusing on the NCVMP algorithm for GLMMs.

3.1. Natural gradient of the variational lower bound

The stochastic gradient form of SA (see Spall [2003]) aims to optimize an objective function based on noisy but unbiased estimates of its gradients. In variational approximation (VA), we wish to minimize the KL divergence between the true posterior and the approximating variational distribution which is equivalent to maximizing the variational lower bound. Following Hoffman et al. [2012], we use the natural gradient of the lower bound instead of the usual gradient in the SA. In this section, we review VA briefly and derive the natural gradient of the variational lower bound in the non-conjugate case.

Let \( y \) denote the observed data and \( \theta \) the unknown model parameters. In variational inference (Jordan et al. [1999]), we approximate the true posterior \( p(\theta|y) \) by a more tractable distribution \( q(\theta) \) which is optimized to be close to \( p(\theta|y) \) in terms of the KL divergence. The KL divergence between \( q(\theta) \) and \( p(\theta|y) \), \( KL(q||p) \), can be written as

\[
\int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} \, d\theta = \int q(\theta) \log \frac{q(\theta)}{p(y, \theta)} \, d\theta + \log p(y)
\]
where \( p(y) = \int p(y|\theta)p(\theta)\ d\theta \) is the marginal likelihood. As \( KL(q\|p) \geq 0 \),

\[
\log p(y) \geq \int q(\theta) \log \frac{p(y, \theta)}{q(\theta)}\ d\theta \\
= E_q \{\log p(y, \theta)\} - E_q \{\log q(\theta)\} \\
= \mathcal{L}
\]

which gives a lower bound \( \mathcal{L} \) on the log marginal likelihood.

In NCVMP, the VA \( q(\theta) \) is of a factorized form \( \prod_{i=1}^{m} q_i(\theta_i) \) for some partition \( \{\theta_1, ..., \theta_m\} \) of \( \theta \), and each \( q_i \) belongs to some exponential family. Let

\[
q_i(\theta_i) = \exp \{\lambda_i^T t(\theta_i) - h(\lambda_i)\}
\]

where \( \lambda_i \) is the vector of natural parameters and \( t(\cdot) \) are the sufficient statistics. To minimize the KL divergence between \( q \) and the true posterior \( p(\theta|y) \), we maximize the variational lower bound \( \mathcal{L} \) with respect to the variational parameters \( \lambda_1, ..., \lambda_m \) which are the natural parameters of \( q_1(\theta_1), ..., q_m(\theta_m) \) respectively. The gradient of \( \mathcal{L} \) with respect to \( \lambda_i \), \( i = 1, ..., m \), is given by

\[
\frac{\partial \mathcal{L}}{\partial \lambda_i} = \frac{\partial}{\partial \lambda_i} E_q \{\log p(y, \theta)\} - \frac{\partial}{\partial \lambda_i} E_q \{\log q(\theta)\}. \tag{2}
\]

To evaluate the first term in (2), suppose that \( p(y, \theta) = \prod_a f_a(y, \theta) \) so that \( E_q \{\log p(y, \theta)\} = \sum_a E_q \{\log f_a(y, \theta)\} \). Since we have assumed in the VA that \( \theta_i \) is independent of all \( \theta_j \) where \( j \neq i \), the only terms in \( \sum_a E_q \{\log f_a(y, \theta)\} \) which depend on \( \lambda_i \) come from those factors \( f_a \) connected to \( \theta_i \) in the factor graph of \( p(y, \theta) \) (see Bishop (2006) and Fig. 1). Therefore,

\[
\frac{\partial}{\partial \lambda_i} E_q \{\log p(y, \theta)\} = \sum_{a \in N(\theta_i)} \frac{\partial S_a(\lambda_i)}{\partial \lambda_i} \tag{3}
\]

where the summation is over all factors in \( N(\theta_i) \), the neighbourhood of \( \theta_i \) in the factor graph and \( S_a(\lambda_i) = \int q_j(\theta_j|\lambda_i) E_{\theta_j} \{\log f_a(y, \theta)\}\ d\theta_i \) where \( E_{\theta_j} \) denotes expectation with respect to the density \( \prod_{j \neq i} q_j(\theta_j) \). Note that \( S_a(\lambda_i) \) also depends on variational parameters of other variables connected to \( f_a \) in the factor graph of \( p(y, \theta) \). For the second term in (2), \( E_q \{\log q(\theta)\} = \sum_{i=1}^{m} E_q \{\log q_i(\theta_i)\} \) of which only the \( i \)th term depends on \( \lambda_i \). Hence,

\[
\frac{\partial}{\partial \lambda_i} E_q \{\log q(\theta)\} = \frac{\partial}{\partial \lambda_i} \left\{ \lambda_i^T \frac{\partial h(\lambda_i)}{\partial \lambda_i} - h(\lambda_i) \right\} \\
= V(\lambda_i)\lambda_i, \tag{4}
\]

where we have used the fact that \( E_q \{t(\theta_i)\} = \frac{\partial h(\lambda_i)}{\partial \lambda_i} \) and \( V(\lambda_i) = \frac{\partial^2 h(\lambda_i)}{\partial \lambda_i \partial \lambda_i} \) denotes the variance-covariance matrix of \( t(\theta_i) \). Putting (4) and (3) together, the gradient of the lower bound is

\[
\frac{\partial \mathcal{L}}{\partial \lambda_i} = \sum_{a \in N(\theta_i)} \frac{\partial S_a(\lambda_i)}{\partial \lambda_i} - V(\lambda_i)\lambda_i.
\]
To obtain the natural gradient of the variational lower bound, we multiply the inverse of the Riemannian metric with the gradient of $\mathcal{L}$ (Honkela et al. 2008; Hoffman et al. 2012). For the variational parameter $\lambda_i$ of $q_i(\theta_t)$, the Riemannian metric is $E_q\{\frac{\partial \log q_i(\theta_t|\lambda_i)}{\partial \lambda_i} (\frac{\partial \log q_i(\theta_t|\lambda_i)}{\partial \lambda_i})^T\}$. Since $\frac{\partial \log q_i(\theta_t|\lambda_i)}{\partial \lambda_i} = t(\theta_t) - \frac{\partial h(\lambda_i)}{\partial \lambda_i}$, the Riemannian metric is simply $V(\lambda_i)$. The natural gradient $\nabla \mathcal{L}(\lambda_i)$ is thus

$$\nabla \mathcal{L}(\lambda_i) = V(\lambda_i)^{-1} \sum_{a \in N(\theta_i)} \frac{\partial S_a(\lambda_i)}{\partial \lambda_i} - \lambda_i,$$

where $L(\lambda_i)$ denotes the lower bound considered as a function of the subset of the parameters, $\lambda_i$, with the other values held fixed. The natural gradient is zero when $\lambda_i = V(\lambda_i)^{-1} \sum_{a \in N(\theta_i)} \frac{\partial S_a(\lambda_i)}{\partial \lambda_i}$. In NCVMP, this optimality condition is used to obtain updates to $\lambda_i$ in a fixed-point iterations algorithm where

$$\lambda_i^{(t)} = V\left(\lambda_i^{(t-1)}\right)^{-1} \sum_{a \in N(\theta_i)} \left. \frac{\partial S_a(\lambda_i)}{\partial \lambda_i} \right|_{\lambda_i = \lambda_i^{(t-1)}}$$

at the $t$th iteration for $i = 1, \ldots, m$. The lower bound is not guaranteed to increase at every step in NCVMP and sometimes convergence issues may be encountered which may require damping to fix (see Knowles and Minka 2011).

When the factor $f_a$ is conjugate to $q_i(\theta_t)$, that is, $f_a$ has the same functional form as $q_i(\theta_t)$ with respect to $\theta_t$, the gradient of the lower bound can be simplified. Suppose

$$f_a(y, \theta) = \exp\{g_a(y, \theta-1)^T t(\theta_t) - h_a(y, \theta-1)\}$$

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

$$\frac{\partial}{\partial \lambda_i} E_q\{\log f_a(y, \theta)\} = V(\lambda_i) E_q\{g_a(y, \theta-1)\},$$

where $E_q\{g_a(y, \theta-1)\}$ does not depend on $\lambda_i$. When this holds for each factor in the neighbourhood of $\theta_i$ in the factor graph, the gradient of the lower bound is $V(\lambda_i) \sum_{a \in N(\theta_i)} E_q\{g_a(y, \theta-1)\} - \lambda_i$ and the natural gradient is simply $\nabla \mathcal{L}(\lambda_i) = \sum_{a \in N(\theta_i)} E_q\{g_a(y, \theta-1)\} - \lambda_i$. The updates in the NCVMP algorithm reduce to

$$\lambda_i^{(t+1)} = \sum_{a \in N(\theta_i)} E_q\{g_a(y, \theta-1)\} \big|_{\lambda_{-i} = \lambda_{-i}^{(t)}}$$

where $\lambda_{-i} = (\lambda_1, \ldots, \lambda_{i-1}, \lambda_{i+1}, \ldots, \lambda_m)$ and $\lambda_i^{(t)} = (\lambda_1^{(t)}, \ldots, \lambda_{i-1}^{(t)}, \lambda_{i+1}^{(t)}, \ldots, \lambda_m^{(t)})$. This is the same as updates in the VMP algorithm. NCVMP thus reduces to VMP for conjugate factors (Knowles and Minka 2011).
3.2. Stochastic variational algorithm for the GLMM

For the GLMM, \( \theta = (\tilde{\alpha}, \beta, D) \) where \( \tilde{\alpha} = (\tilde{\alpha}_1^T, ..., \tilde{\alpha}_n^T)^T \) and

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

This factorized distribution can be expressed as the factor graph in Fig. 1. We consider a VA to the posterior of the form

\[
q(\theta) = q(\beta) q(D) \prod_{i=1}^{n} q(\tilde{\alpha}_i)
\]

where \( q(\beta) \) is \( N(\mu_\beta^q, \Sigma_\beta^q) \), \( q(D) \) is \( IW(\nu^q, S^q) \), and \( q(\tilde{\alpha}_i) \) is \( N(\mu_{\tilde{\alpha}_i}^q, \Sigma_{\tilde{\alpha}_i}^q) \). Let \( \lambda_\beta, \lambda_D \) and \( \lambda_{\tilde{\alpha}_i} \) denote the natural parameter vectors of \( q(\beta) \), \( q(D) \) and \( q(\tilde{\alpha}_i) \) respectively for \( i = 1, ..., n \). In the standard NCVMP algorithm, we would iterate between updating the local variational parameters \( \lambda_{\tilde{\alpha}_i} \) for each unit \( i, i = 1, ..., n \), and updating the global variational parameters \( \lambda_\beta \) and \( \lambda_D \). This can be impractical for large data sets and impossible to accomplish for streaming data. Instead, we randomly select a mini-batch, \( S \), of units (with or without replacement), say of size \( |S| \geq 1 \) at each iteration and compute NCVMP updates for \( \lambda_{\tilde{\alpha}_i}, i \in S \) repeatedly until convergence. Using these optimized local variational parameters, we can compute unbiased estimates of the natural gradients \( \nabla L(\lambda_\beta) \) and \( \nabla L(\lambda_D) \) and estimate \( \lambda_\beta \) and \( \lambda_D \) using SA. Similar approaches to online variational inference have been developed for latent Dirichlet allocation by [Hoffman et al. (2010)] and for the hierarchical Dirichlet process by [Wang et al. (2011)]. [Hoffman et al. (2012)] developed stochastic variational inference for conjugate-exponential family models and showed that this type of algorithm is a form of stochastic natural gradient ascent of the global variational parameters. The idea is to find a setting of the global variational parameters that maximizes

![Fig. 1. Factor graph for GLMM model. Dotted box indicates that the contained node and its edges are duplicated n times.](image)
the lower bound by considering the variational lower bound as a function of the global variational parameters with the local parameters optimized as a function of these global parameters. We have

$$\nabla L(\lambda_\beta) = V(\lambda_\beta)^{-1} \frac{\partial}{\partial \lambda_\beta} \left[ E_q \{ \log p(\beta | \Sigma_\beta) \} + \sum_{i=1}^{n} A_i \right] - \lambda_\beta,$$

where $A_i = E_q \{ \log p(y_i | \beta, \hat{\alpha}_i) \} + E_q \{ \log p(\hat{\alpha}_i | \beta, D) \}$. An unbiased estimate of $\nabla L(\lambda_\beta)$ using the mini-batch $S$ is

$$\tilde{\nabla} L(\lambda_\beta) = \frac{1}{|S|} \sum_{i \in S} A_i.$$

As the factors in the neighbourhood of $D$ are all conjugate factors,

$$\nabla L(\lambda_D) = \left( -\frac{1}{2} \text{vec}(S) \right) + \sum_{i=1}^{n} \left( -\frac{1}{2} B_i \right) - \lambda_D$$

where $B_i = \text{vec} \left[ \mu^q_{\alpha_i} - \tilde{W}_i \mu^q_{\beta} \right] (\mu^q_{\alpha_i} - \tilde{W}_i \mu^q_{\beta})^T + \Sigma^q_{\alpha_i} + \tilde{W}_i \Sigma^q_{\beta} \tilde{W}_i^T$. An unbiased estimate of $\nabla L(\lambda_D)$ using mini-batch $S$ is

$$\tilde{\nabla} L(\lambda_D) = \left( -\frac{1}{2} \text{vec}(S) \right) + \frac{1}{|S|} \sum_{i \in S} \left( -\frac{1}{2} B_i \right).$$

When $S$ is the entire data set, $\tilde{\lambda}_\beta$ and $\tilde{\lambda}_D$ are the updates of $\lambda_\beta$ and $\lambda_D$ in the standard NCVMP algorithm. Readers may refer to Tan and Nott (2012) where this variational posterior was considered, for the NCVMP update expressions from which the natural gradients may be inferred. The natural gradients are available in closed form for the Poisson GLMM, while adaptive Gauss-Hermite quadrature (Liu and Pierce, 1994) is required for evaluation of the natural gradients in the logistic case. In this paper, we focus on presenting a stochastic variational algorithm for the GLMM suitable for analyzing moderate to large datasets or even data arriving continuously in an infinite data stream.

In the stochastic NCVMP algorithm, we combine NCVMP updates for the local variational parameters $\lambda_{\tilde{\alpha}, i}, i \in S$ with SA estimation of the global parameters $\lambda_\beta$ and $\lambda_D$ as outlined in Table 1. The SA steps in lines 10 and 11 were introduced by Robbins and Monro (1951) for optimizing an objective function which in our case is the lower bound $L$, with local variational parameters optimized as a function of the global ones. Hoffman et al. (2012) note that the gradient of this function is just the gradient of $L$ with the local parameters fixed at their optimized values (see Hoffman et al., 2012, equation (39)). Under certain regularity conditions (see Spall 2003), the iterates will converge to a local
As step sizes can strongly influence performance of the algorithm in practice (Jank, 2006). To avoid false convergence (Spall, 2003), we consider step sizes of the form $a_t \rightarrow 0$, $\sum_t a_t = \infty$, and $\sum_t a_t^2 < \infty$. Intuitively, the condition $(a_t \rightarrow 0, \sum_t a_t^2 < \infty)$ ensures that the step size goes to zero sufficiently fast so that the iterates will converge while $(\sum_t a_t = \infty)$ ensures that the rate at which the step sizes approach zero is slow enough to avoid false convergence (Spall, 2003). We consider step sizes of the form $\frac{1}{(t+K)^\gamma}$ where $0.5 < \gamma \leq 1$ and $K \geq 0$ is a stability constant that helps to avoid unstable behaviour in the early iterations. In principle, the iterates will converge to a local optimum as long as conditions in (5) are satisfied. However, choices of the step sizes can strongly influence performance of the algorithm in practice (Jank, 2006). As $\nabla L(\lambda_\beta) = \hat{\lambda}_\beta - \lambda_\beta$, we have $\lambda_\beta(t) = (1 - a_t)\lambda_\beta(t-1) + a_t\tilde{\lambda}_\beta$ from line 10. Thus the $t$-iterate can be interpreted as a weighted average of the previous iterate and the NCVMP update of $\lambda_\beta$ estimated using mini-batch $S$. The same holds for $\lambda_D$. Standard NCVMP is recovered from stochastic NCVMP when the update for the local parameters in line 7 is performed only once and $\gamma = 0$ in the SA steps in lines 10 and 11.

To initialize the variational parameters, we use the GLM obtained by pooling all the data and setting the random effects $u_i = 0$ for all $i$. This GLM was used to compute the prior of $D$, $IW(\nu, S)$. We set $\mu_0^\beta$ and $\Sigma_0^\beta$ as the estimate of the regression coefficients and its covariance matrix from the GLM respectively, $\nu^\beta = r$, $S^\beta = S$, $\mu_0^\beta = \hat{W}_i\mu_0^\beta$ and $\Sigma_0^\beta = \hat{R}$ where $\hat{W}_i$ is initialized by setting $D = \hat{R}$ and $\eta_i$ by $\hat{C}_i\mu_0^\beta + X_i^R\mu_0^\beta$ for $i = 1, ..., n$. Kass and Natarajan (2006) contain a justification of $\hat{R}$ being a reasonable guess for $D$ in the absence of any other prior knowledge. Care should be taken in initializing the variational parameters as the NCVMP update steps in line 7 are not guaranteed to converge. We used

| Step | Description |
|------|-------------|
| 1    | Initialize variational parameters $\lambda_\beta$, $\lambda_D$, $\lambda_{\tilde{a}}$ for $i = 1, ..., n$ |
| 2    | and the partial centering parameters $W_i$ for $i = 1, ..., n$. |
| 3    | At the $t$th iteration, |
| 4    | randomly select a set $S$ of $|S|$ units. |
| 5    | Update local variational parameters $\lambda_{\tilde{a}}$ for $i \in S$ |
| 6    | repeatedly using NCVMP update: |
| 7    | $\lambda_{\tilde{a}}^{(t)} \leftarrow V \left( \lambda_{\tilde{a}}^{(t-1)} \right)^{-1} \frac{\partial}{\partial \lambda_{\tilde{a}}} \left[ E_q \{ \log p(y_i|\beta, \tilde{a}_i) \} + E_q \{ \log p(\tilde{a}_i|\beta, D) \} \right] \bigg|_{\tilde{a}_i = \tilde{a}_i^{(t-1)}}$ |
| 8    | until convergence. |
| 9    | Next, update the global variational parameters $\lambda_\beta$ and $\lambda_D$ using |
| 10   | $\lambda_\beta^{(t)} = \lambda_\beta^{(t-1)} + a_t \nabla L \left( \lambda_\beta^{(t-1)} \right)$ and |
| 11   | $\lambda_D^{(t)} = \lambda_D^{(t-1)} + a_t \nabla L \left( \lambda_D^{(t-1)} \right)$. |
| 12   | Repeat steps 3 to 11 until convergence. |
the initialization suggested above in all our examples and did not experience any convergence issues. The mean parameters of $\tilde{\alpha}_i$ were used to test for convergence in line 8 and we terminate the fixed-point iterations in line 7 when

$$\frac{\|\mu_q^{\tilde{\alpha}_i}(t) - \mu_q^{\tilde{\alpha}_i}(t-1)\|}{\|\mu_q^{\tilde{\alpha}_i}(t)\|} < 0.01,$$

where $\| \cdot \|$ represents the Euclidean norm.

### 3.3. Switching from Stochastic to Batch VA

Determining an appropriate stopping criterion for a SA algorithm can be a very challenging task. Some commonly used stopping 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) gives an illustrative example to show that even this is not enough safeguard. Moreover, the performance of SA can become excruciatingly slow in the later iterations due to the small step sizes.

Through our experimentations, we observe that gains made by stochastic NCVMP are usually largest in the first few iterations. However, beyond a certain point, it can become even slower than standard NCVMP if the step sizes are too small or the iterates simply bounce around if the step sizes are still too big. We therefore suggest switching from stochastic to standard NCVMP when the stochastic NCVMP shows signs of slowing down. Using the lower bound both as a switching and stopping criterion, we propose switching from stochastic to standard NCVMP when the relative increase in the lower bound is less than $10^{-3}$ and terminating the standard NCVMP when the absolute relative change in the lower bound is less than $10^{-6}$. This proposal may be more applicable to moderate-sized data sets such as those considered in this paper. For large datasets or streaming data, it might be more practical to terminate the algorithm beyond a certain period of available runtime.

The mini-batches in line 4 of the stochastic NCVMP algorithm were chosen by random-partitioning of the data set. In this paper, the mini-batch sizes considered were such that different batches differ in size by at most one when $n$ is not divisible by $|S|$. For greater efficiency, the lower bound is computed only after a complete sweep has been made through the data set. We replace $t$ by $s_w + \frac{m}{M}$ in the step size where $s_w$ indicates the number of sweeps that has been made through the data, $M$ denotes the number of partitions of the data and $0 \leq m \leq M - 1$ denotes the number of batches that has been analysed. It is possible to include an update of $W_i, i = 1, \ldots, n$ after each complete sweep. However,
preliminary investigation suggested no significant improvement in results when \( W_i \) is updated and hence, for the examples in this paper, we did not update \( W_i \) beyond the initialization.

4. Automatic diagnostics of prior-likelihood conflict as a by-product of VMP

Marshall and Spiegelhalter (2007) considered hierarchical models where observations are grouped within units and parameters of these units are assumed to be drawn from some population model. They investigated a diagnostic test for identifying divergent units based on measuring the conflict between likelihood of a parameter and its predictive prior given the remaining data. A simulation-based approach was adopted and diagnostic tests were performed using MCMC. As full cross-validation methods are computationally demanding, they also considered full data approximations which were shown to demonstrate only moderate conservatism. We show that the approach of Marshall and Spiegelhalter (2007) can be approximated in the VMP framework and that VMP facilitates an automatic computation of diagnostics for prior-likelihood conflict. Intuitively, the idea is that we can separate the messages coming from above and below a node in a hierarchical model, and when there are “mixed messages” this is indicative of conflict. These diagnostics are very useful for model criticism. We focus on NCVMP for GLMMs.

First, we review briefly the diagnostic test proposed by Marshall and Spiegelhalter (2007). In the context of GLMMs with a PNCP, the parameter of interest for identifying divergent units is \( \tilde{\alpha}_i \), \( i = 1, ..., n \). For \( \tilde{\alpha}_i \), Marshall and Spiegelhalter (2007) suggest generating a predictive prior replicate \( \tilde{\alpha}_i^{\text{rep}} \sim p(\tilde{\alpha}_i|y_{-i}) \) where \( y_{-i} \) denotes the observed data \( y \) with unit \( i \) left out and

\[
p(\tilde{\alpha}_i|y_{-i}) = \int p(\tilde{\alpha}_i|\beta, D)p(\beta, D|y_{-i}) \text{ d}\beta \text{d}D. \tag{6}
\]

In the simulation approach, \( \beta^{\text{rep}}, D^{\text{rep}} \) would be generated from \( p(\beta, D|y_{-i}) \) using MCMC followed by simulation of \( \tilde{\alpha}_i^{\text{rep}}|\beta^{\text{rep}}, D^{\text{rep}} \). This is compared with a likelihood replicate \( \tilde{\alpha}_i^{\text{fix}} \sim p(\tilde{\alpha}_i|y_{i}) \) generated using only data from the unit \( y_i \) being tested and a non-informative prior, \( p(\tilde{\alpha}_i) \), for \( \tilde{\alpha}_i \) since \( p(\tilde{\alpha}_i|y_i) \propto p(y_i|\tilde{\alpha}_i)p(\tilde{\alpha}_i) \). These prior and likelihood replications represent two independent sources of evidence about \( \tilde{\alpha}_i \) and conflict between them suggests discrepancies in the model.

The above discussion ignores nuisance parameters. In our case, we need to regard \( \beta \) as a nuisance parameter. As \( p(\tilde{\alpha}_i|y_i) \propto p(\tilde{\alpha}_i) \int p(y_i|\beta, \tilde{\alpha}_i)p(\beta|\tilde{\alpha}_i) \text{ d}\beta \) and \( \beta \) is not estimable from individual unit \( i \), Marshall and Spiegelhalter (2007) (pg. 420) recommend generating \( \tilde{\alpha}_i^{\text{fix}} \) from \( f(\alpha_i|y) \) where

\[
f(\alpha_i|y) \propto p(\tilde{\alpha}_i) \int p(y_i|\tilde{\alpha}_i, \beta)p(\beta|y_{-i}) \text{ d}\beta.
\]
Note that the two replications $\tilde{\alpha}_i^{\text{rep}}$ and $\tilde{\alpha}_i^{\text{fix}}$ are no longer entirely independent as $y_{-i}$ will slightly influence $\tilde{\alpha}_i^{\text{fix}}$ through $\beta$.

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

\[ p_i^{\text{L,con}} = P(\tilde{\alpha}_i^{\text{diff}} \leq 0 | y) \]

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 $p_i^{\text{U,con}} = 1 - p_i^{\text{L,con}}$ or the 2-sided $p$-value $2 \times \min(p_i^{\text{L,con}}, p_i^{\text{U,con}})$ may be of interest instead. If $\tilde{\alpha}_i^{\text{diff}}$ is not a scalar, $E(\tilde{\alpha}_i^{\text{diff}} | y) = \frac{1}{2} \text{vec}(\Sigma_{\beta,D}^{-1})^{-1} \text{vec}(\Sigma_{\text{lik}}) \text{Cov}(\tilde{\alpha}_i^{\text{diff}} | y)$ can be used as a standardised discrepancy measure. An alternative to this cross-validation approach is to simulate $\tilde{\alpha}_i^{\text{rep}} | \beta^{\text{rep}}, D^{\text{rep}}$ using $\beta^{\text{rep}}, D^{\text{rep}}$ generated from $p(\beta, D | y)$ without leaving out $y_i$. This introduces only mild conservatism in normal hierarchical models as $y_i$ influences $\tilde{\alpha}_i^{\text{rep}}$ through $\beta$ and $D$ (Marshall and Spiegelhalter, 2007).

We attempt to approximate the approach of Marshall and Spiegelhalter (2007) in the VMP framework. Recall that the variational posterior for $\tilde{\alpha}_i$ is Gaussian with natural parameter $\lambda_{\tilde{\alpha}_i} = \left( -\frac{1}{2} \text{vec}(\Sigma_{\beta,D}^{-1})^{-1} \Sigma_{\text{lik}}^{-1} \mu_{\text{lik}} \right)$. The NCVMP update for $\lambda_{\tilde{\alpha}_i}$ is given by

\[
V(\lambda_{\tilde{\alpha}_i})^{-1} \frac{\partial}{\partial \lambda_{\tilde{\alpha}_i}} [E_q \{ \log p(\tilde{\alpha}_i | \beta, D) \} + E_q \{ \log p(y_i | \tilde{\alpha}_i, \beta) \}] = \left( -\frac{1}{2} \text{vec}(\Sigma_{\beta,D}^{-1})^{-1} \Sigma_{\text{lik}}^{-1} \mu_{\text{lik}} \right)
\]

where we let $V(\lambda_{\tilde{\alpha}_i})^{-1} \frac{\partial}{\partial \lambda_{\tilde{\alpha}_i}} E_q \{ \log p(y_i | \tilde{\alpha}_i, \beta) \} = \left( -\frac{1}{2} \text{vec}(\Sigma_{\text{lik}}^{-1}) \right)$. The first term can be considered a message from the prior $p(\tilde{\alpha}_i | \beta, D)$ and the second term a message from the likelihood of unit $y_i$, $p(y_i | \tilde{\alpha}_i, \beta)$. We argue that the first message from the prior can be interpreted as natural parameter of a Gaussian approximation to $p(\tilde{\alpha}_i | y_{-i})$ while the second message from the likelihood can be interpreted as natural parameter of a Gaussian approximation to $f(\tilde{\alpha}_i | y)$. This implies that $\tilde{\alpha}_i^{\text{rep}} \sim N(\tilde{W}_i \mu_{\beta}, \frac{1}{\nu} S^\gamma)$ and $\tilde{\alpha}_i^{\text{fix}} \sim N(\mu_{\text{lik}}, \Sigma_{\text{lik}})$ so that $\tilde{\alpha}_i^{\text{diff}} \sim N(\tilde{W}_i \mu_{\beta} - \mu_{\text{lik}}, \frac{1}{\nu} S^\gamma + \Sigma_{\text{lik}})$ assuming $\tilde{\alpha}_i^{\text{rep}}$ and $\tilde{\alpha}_i^{\text{fix}}$ are considered independent.

Since these messages are computed in the NCVMP algorithm, conflict $p$-values can be calculated easily at convergence for identification of divergent units.

For moderate to large data sets, the difference between $p(\beta, D | y_{-i})$ and $p(\beta, D | y_i)$ is small and we approximate $p(\beta, D | y_{-i})$ in (6) by the variational posterior $q(\beta) q(D)$. This combined with Jensen’s inequality gives

\[
\log p(\tilde{\alpha}_i | y_{-i}) \approx \log E_{-\tilde{\alpha}_i} \{ p(\tilde{\alpha}_i | \beta, D) \} \\
\geq E_{-\tilde{\alpha}_i} \{ \log p(\tilde{\alpha}_i | \beta, D) \}.
\]

Approximating $p(\tilde{\alpha}_i | y_{-i})$ by $\text{exp} [ E_{-\tilde{\alpha}_i} \{ \log p(\tilde{\alpha}_i | \beta, D) \} ]$, we then have $\tilde{\alpha}_i^{\text{rep}} \sim N(\tilde{W}_i \mu_{\beta}, \frac{1}{\nu} S^\gamma)$. On the other hand, the total message gives us the natural parameter of $q(\tilde{\alpha}_i)$ which is an approximation of $p(\tilde{\alpha}_i | y)$. If we think of $p(\tilde{\alpha}_i | y_{-i})$
as the ‘prior’ to be updated when \(y_i\) becomes available, we have

\[
p(\tilde{\alpha}_i | y) \propto p(\tilde{\alpha}_i | y_{i-1}) p(y_i | \tilde{\alpha}_i, y_{i-1})
\]

which implies that

\[
\frac{p(\tilde{\alpha}_i | y)}{p(\tilde{\alpha}_i | y_{i-1})} \propto p(y_i | \tilde{\alpha}_i, y_{i-1}).
\]

Thus, interpreting the first message as natural parameter of a Gaussian approximation to \(p(\tilde{\alpha}_i | y_{i-1})\) and the sum of the two messages as natural parameter of 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 \(p(y_i | \tilde{\alpha}_i, y_{i-1})\). As a function of \(\tilde{\alpha}_i\), the ratio of the two normal distributions is proportional to

\[
\frac{\exp\{-\frac{1}{2} (\tilde{\alpha}_i - \mu_{\tilde{\alpha}_i})^T \Sigma_{\tilde{\alpha}_i}^{-1} (\tilde{\alpha}_i - \mu_{\tilde{\alpha}_i})\}}{\exp\{-\frac{1}{2} (\tilde{\alpha}_i - \mu_{\tilde{\alpha}_i})^T \nu S^{-1} (\tilde{\alpha}_i - \mu_{\tilde{\alpha}_i})\}}
\]

which gives a normal distribution with natural parameters

\[
\left( -\frac{1}{2} \text{vec}(\Sigma_{\tilde{\alpha}_i}^{-1} - \nu S^{-1}) \right) = \left( -\frac{1}{2} \text{vec}(\Sigma_{\text{lik}}^{-1}) \right)
\]

precisely that given by the second message. As

\[
p(y_i | \tilde{\alpha}_i, y_{i-1}) = \int p(y_i | \beta, \tilde{\alpha}_i) p(\beta | \tilde{\alpha}_i, y_{i-1}) \, d\beta
\]

and \(p(\beta | \tilde{\alpha}_i, y_{i-1})\) is close to \(p(\beta | y_{i-1})\) when the number of clusters is large, the second message can be considered as giving the natural parameter of a Gaussian approximation to \(f(\tilde{\alpha}_i | y_{i-1})\) if we assume a uniform prior for \(p(\tilde{\alpha}_i)\). Finally, even though \(\tilde{\alpha}_i^{\text{rep}}\) and \(\tilde{\alpha}_i^{\text{fix}}\) are not entirely independent, for large data sets, the dependence between \(\tilde{\alpha}_i^{\text{rep}}\) and \(\tilde{\alpha}_i^{\text{fix}}\) will be increasingly weak as the number of clusters increases so that \(\tilde{\alpha}_i^{\text{diff}} \sim N(\tilde{W}_i \mu_{\beta}^{\text{lik}} - \mu_{\text{lik}}, \frac{1}{\nu S} + \Sigma_{\text{lik}})\) approximately.

This approximate distribution for \(\tilde{\alpha}_i^{\text{diff}}\) can be used to calculate conflict p-values when the NCVMP algorithm has converged. We illustrate this method with an example in Section 5.1 and the results show very good agreement between the cross-validatory conflict p-values computed using the approach of Marshall and Spiegelhalter (2007) and those obtained from NCVMP. For large data sets, automatic computation of diagnostics for prior-likelihood conflict can be an attractive alternative to the simulation-based approach using MCMC methods. While the approximations made in our derivation of the diagnostic are crude, the diagnostics can be computed automatically as part of the NCVMP algorithm and can be regarded as a handy screening tool with the clusters flagged as divergent studied more closely and possibly conflict p-values recomputed by Monte Carlo. The arguments above generalize to detecting conflict for other parameters of the model also.
5. Examples

In Section 5.1, we use the Bristol infirmary inquiry data to compare the conflict $p$-values computed using the NCVMP algorithm with those obtained using the approach of Marshall and Spiegelhalter (2007). In Sections 5.2 and 5.3, we apply the stochastic version of NCVMP to a real data set and a simulated data set respectively. In all the examples, the PNCP was used and we considered a $N(0,1000)$ prior for $\beta$ and a inverse Wishart default conjugate prior (Kass and Natarajan 2006) for $D$. We also experimented with various settings of $K$ and $\gamma$. The Muscatine coronary risk factor study data set and the skin cancer prevention study data set can be found at http://www.biostat.harvard.edu/~fitzmaur/ala2e/.

5.1. Bristol infirmary 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 from 1984 to 1995. The outcomes of paediatric cardiac surgical services at Bristol relative to other specialist centres was a key issue. We consider a subset of the data presented to the Inquiry recorded by Hospital Episode Statistics on the mortality rates in open surgeries for 12 hospitals including Bristol (hospital 1), for children under 1 year old, from 1991–March 1995. This data can be found in Marshall and Spiegelhalter (2007) Table 1. Spiegelhalter et al. (2002) and Marshall and Spiegelhalter (2007) modelled this data using a logistic GLMM. 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 calculating prior-likelihood conflict diagnostics from the NCVMP algorithm.

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

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

To assess the accuracy of the approximate conflict $p$-values obtained from the standard NCVMP algorithm, we use the cross-validatory conflict $p$-values obtained using the simulation-based approach of Marshall and Spiegelhalter (2007) as a ‘gold-standard’ and compute these for comparison. In the cross-validatory approach, each hospital $i$ is removed in turn from the analysis, and the parameters $\beta^{\text{rep}}, D^{\text{rep}}|y_{-i}$ are generated using MCMC followed by a simulated $\pi_i^{\text{rep}}|\beta^{\text{rep}}, D^{\text{rep}}$. Assuming a Jeffrey’s prior for $\pi_i$, a $\pi_i^{\text{fix}}$ is then simulated from 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 $p_{i,\text{con}} = P(\pi_i^{\text{rep}} \geq \pi_i^{\text{fix}})$. 100 000 simulations were used in calculating the cross-validatory conflict $p$-values. Fitting via MCMC was performed in WinBUGS (Lunn et al., 2000) through R by using R2WinBUGS (Sturtz et al., 2005) as an interface. The CPU time taken to compute the cross-validatory conflict $p$-values using the simulation-based approach

16
Table 2. Cross-validatory conflict $p$-values ($p_{CV_i,con}$) and approximate conflict $p$-values from NCVMP ($p_{NCVMP_i,con}$).

| hospital | $p_{CV_i,con}$ | $p_{NCVMP_i,con}$ |
|----------|----------------|------------------|
| 1        | 0.001          | 0.004            |
| 2        | 0.434          | 0.448            |
| 3        | 0.935          | 0.934            |
| 4        | 0.125          | 0.131            |
| 5        | 0.298          | 0.305            |
| 6        | 0.721          | 0.730            |
| 7        | 0.740          | 0.751            |
| 8        | 0.660          | 0.672            |
| 9        | 0.438          | 0.450            |
| 10       | 0.377          | 0.386            |
| 11       | 0.766          | 0.771            |
| 12       | 0.723          | 0.733            |

was 465 seconds.

The cross-validatory conflict $p$-values ($p_{CV_i,con}$) and approximate conflict $p$-values from NCVMP ($p_{NCVMP_i,con}$) for all hospitals are shown in Table 2. These two sets of $p$-values are plotted in Fig. 2 which indicates very good agreement between the two. To reflect the importance of good agreement at the extremes, [Marshall and Spiegelhalter (2007)] computed the relative agreement between $p$-values as

\[
\left| \frac{\Phi^{-1}(p_{CV_i,con}) - \Phi^{-1}(p_{NCVMP_i,con})}{\Phi^{-1}(p_{NCVMP_i,con})} \right| \times 100%
\]

where $\Phi^{-1}$ denotes the inverse cumulative distribution function of the standard normal. The relative error between $p_{CV_i,con}$ and $p_{NCVMP_i,con}$ is 9% which is close to the relative error of 7% between cross-validatory and full data conflict $p$-values reported in [Marshall and Spiegelhalter (2007)]. In this case, the NCVMP algorithm yields very good results and takes only 7 seconds. For moderate to large data sets, the VMP approach will be an extremely attractive alternative to computationally intensive MCMC methods for obtaining prior-likelihood conflict diagnostics.

5.2. Muscatine coronary risk factor study

A total of 4856 children took part in the Muscatine coronary risk factor study [Woolson and Clarke (1984)] which was undertaken to examine the development and persistence of risk factors for coronary disease in children. Over the period 1977–1981, weight and height data were collected biennially from five cohorts of children, aged 5–7, 7–9, 9–11, 11–13 and 13–15 at the beginning of the study. The data is incomplete with less than 40% of the children surveyed on all three
occasions. In previous analyses, some authors treated this data as potentially missing not at random (for instance, Zhou et al. (2010)) while others assumed the data are missing at random (MAR) (Fitzmaurice et al. 1994; Kenward and Molenberghs 1998). We assume the data are MAR and focus on computational comparisons between standard and stochastic NCVMP. The binary response, $y_{ij}$, is an indicator of whether the $i$th child is obese at the $j$th occasion. For the $i$th child, we consider the covariates, gender$_{i} = 1$ if female, 0 if male and age$_{ij} =$ midpoint of age cohort at $j$th occasion $- 12$. Fitzmaurice et al. (2004) modelled the marginal probability of obesity as a logistic function of gender and linear and quadratic age. We consider the following logistic random intercept model,

$$\logit(\mu_{ij}) = \beta_0 + \beta_1 \text{gender}_i + \beta_2 \text{age}_{ij} + \beta_3 \text{age}_{ij}^2 + u_i,$$

where $u_i \sim N(0, \sigma^2)$ for $i = 1, ..., 4856$, $1 \leq j \leq 3$. The standard NCVMP algorithm took 823 seconds to converge for this moderately large data set. The performance of stochastic NCVMP was investigated using different mini-batch sizes and various parameter settings for the step sizes. We considered $|S| \in \{1, 50, 99, 242\}$ where the mini-batch sizes were chosen to correspond to the online setting and approximately 1%, 2% and 5% of $n = 4856$. We let the stability constant $K$ take values 0, 1 and 5 and $\gamma$ be 0.5, 0.75 or 1. In the online setting, the stochastic VA diverges for each $K \in \{0, 1, 5\}$ and $\gamma \in \{0.5, 0.75, 1\}$. We thus considered larger stability constants $K \in \{250, 500, 1000\}$ for $|S| = 1$. For each mini-batch size and parameter setting for the step-size, we did five runs of the stochastic NCVMP switching to standard NCVMP each time the relative increment in the lower bound after a complete sweep through the data is less than $10^{-3}$. The average time taken for the algorithm to converge in each case is
shown in Fig. 3. The solid lines, dashed lines and dot-dashed lines correspond to $\gamma = 1$, 0.75 and 0.5 respectively. The best parameter settings and average time to convergence for each mini-batch size are summarized in Table 3. From these results, we observed that as the mini-batch size increases, smaller values of $\gamma$ and $K$, that is, a slower rate of decrease in step-size and larger step-sizes lead to faster convergence. However, a significantly larger stability constant and smaller step sizes are required in the online setting to prevent unstable behaviour in the early iterations. The mini-batch size of 50 (approximately 1% of $n$) performed well across a wide range of step-sizes with the average time to convergence ranging from 310 to 413 seconds. The shortest average time to convergence is 271 seconds for the mini-batch of size 99 with $K = 0$ and $\gamma = 0.75$. This implies a reduction in convergence time of around 67% from standard NCVMP. Fig. 4 tracks the average lower bound attained at the end of each sweep through the data for the different batch sizes corresponding to the best parameter settings listed in Table 3. Only the first ten sweeps are shown. This figure shows that with appropriately chosen step-sizes, stochastic NCVMP is able to make much bigger gains than standard NCVMP particularly in the first few sweeps. Thus, even for moderate-sized data sets, significant gains can be made by making use of stochastic NCVMP.
5.3. Skin cancer prevention study

In a clinical trial conducted 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. Subjects were biopsied once a year to ascertain the number of new skin cancers since the last examination. The response $y_{ij}$ is a count of the number of new skin cancers in year $j$ for the $i$th subject. Covariate information for the $i$th subject include age$_{i}$, the age in years at the beginning of the study, gender$_{i}$ = 1 if male and 0 if female, exposure$_{i}$, a count of the number of previous skin cancers, skin$_{i}$ = 1 if skin has burns and 0 otherwise, treatment$_{i}$ = 1 if the $i$th subject receives beta-carotene and 0 if placebo and year$_{ij}$, the year of follow-up. We consider $n = 1683$ subjects with complete covariate information. Using conditional Akaike information to perform model selection, Donohue et al. (2011) fitted different Poisson GLMMs to this data and arrived at the model

$$\log(\mu_{ij}) = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{skin}_i + \beta_3 \text{gender}_i + \beta_4 \text{exposure}_i + u_i,$$

where $u_i \sim N(0, \sigma^2)$ for $i = 1, ..., 1683$, $1 \leq j \leq 5$. The treatment and year effects did not prove to be significant in their analyses. Using this model, we investigate the performance of standard and stochastic NCVMP algorithms. As this data set is small, preliminary investigation shows that the time to convergence of the standard and stochastic NCVMP algorithms are close and stochastic NCVMP did not provide any significant gains over standard NCVMP. We thus simulated a data set comprising of $n = 1683 \times 6 = 10098$ subjects by using the posterior means of the unknown parameters from the standard NCVMP fit to the original data set. Thus, we replicate the design matrices for each cluster 6 times. For this
simulated data, standard NCVMP took 415 seconds to converge.

We considered mini-batch sizes corresponding to the online setting and approximately 1%, 2% and 5% of \(n = 10098\), that is, \(|S| \in \{1, 100, 198, 504\}\). We let \(\gamma\) be 0.5, 0.75 or 1 and the stability constant \(K\) take values 0, 1 and 5 for \(|S| \in \{100, 198, 504\}\) and values 250, 500, 1000 for \(|S| = 1\). For each mini-batch size and parameter setting for the step-size, we did five runs of the stochastic NCVMP, switching to standard NCVMP each time the relative increment in the lower bound after a complete sweep through the data is less than \(10^{-3}\). The average time taken for the algorithm to converge in each case is shown in Fig. 5. The solid lines, dashed lines and dot-dashed lines correspond to \(\gamma = 1, 0.75\) and 0.5 respectively. The best parameter settings and average time to convergence for each mini-batch size are summarized in Table 4.

In the online setting, stochastic NCVMP diverges for each \(\gamma \in \{0.5, 0.75, 1\}\) when \(K = 250\). As in Example 5.2, larger stability constants are preferred when \(|S| = 1\). For this simulated data, a higher rate of decrease in step-size is desirable with \(\gamma = 1\) yielding the best performance across different mini-batch sizes. Larger batch sizes also seem to lead to faster convergence. Fig. 6 compares the rate of convergence of standard and stochastic NCVMP for one of the runs where \(|S| = 504\), \(K = 0\) and \(\gamma = 1\). The variational lower bound \(\mathcal{L}\) is \(-23617.3\) at convergence and we have plotted \(\log(-23617 - \mathcal{L})\) against time. Stochastic NCVMP took just 7 sweeps to converge in 216 seconds while standard NCVMP

---

**Table 4.** Skin cancer study. Best parameter settings and average time to convergence (in seconds) for different mini-batch sizes

| \(|S|\) | 1   | 100 | 198 | 504 |
|-------|-----|-----|-----|-----|
| \(K\) | 1000| 1   | 1   | 0   |
| \(\gamma\) | 1   | 1   | 1   | 1   |
| time  | 278 | 233 | 219 | 216 |

**Fig. 5.** Skin cancer study. Plot of average time to convergence against the stability constant \(K\) for different mini-batch sizes. The solid, dashed and dot-dashed lines correspond to \(\gamma = 1, 0.75, 0.5\) respectively.
Fig. 6. Plot of $\log(-23617 - L)$ against time for the mini-batch of size 504, $K = 0$ and $\gamma = 1$. took 22 sweeps and converged in 415 seconds. This represents a reduction of about 50% in convergence time.

6. Conclusion

In this paper, we have extended stochastic variational inference to non-conjugate models and derived a stochastic version of the NCVMP algorithm, scalable to large data sets. The data sets that we have considered in this paper were only of moderate size. Nevertheless, by applying the stochastic version of the NCVMP algorithm in the first few iterations, the time to convergence for these data sets was reduced by half or more. We would 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 $K$ and $\gamma$ suggest that $\gamma$ close to 1 and a large stability constant $K$ is preferred in the online setting while mini-batches larger in size perform better with larger step-sizes. Comparison of the conflict $p$-values obtained from the NCVMP algorithm with those computed using the approach of [Marshall and Spiegelhalter, 2007] suggest very good agreement. For large data sets, the VMP approach will be an extremely attractive alternative to computationally intensive MCMC methods in obtaining prior-likelihood diagnostics. All code was written in the R language and run on a dual processor Window PC 3GHz workstation.

Acknowledgements

Siew Li Tan was partially supported as part of the Singapore-Delft Water Alliance (SDWA)’s tropical reservoir research programme.
References

Ahn, S., Korattikara, A. and Welling, M. (2012) Bayesian posterior sampling via stochastic gradient Fisher scoring. Appearing in Proceedings of the 29th International Conference on Machine Learning.

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

Attias, H. (2000) A variational Bayesian framework for graphical models. Advances in Neural Information Processing Systems, 12, 209–215.

Bishop, C.M. (2006) Pattern recognition and machine learning. Springer.

Bretz, S. (1998) Natural gradient works efficiently in learning. Neural Computation, 10, 251–276.

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.

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

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

Bottou, L. and Bousquet, O. (2008) The trade-offs of large scale learning. Advances in Neural Information Processing Systems, 0, 161–168.

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.

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

Fitzmaurice, G.M., Laird, N.M. and Lipsitz, S.R. (1994) Analysing incomplete longitudinal binary responses: a likelihood-based approach. Biometrics, 50, 601–612.

Fitzmaurice, G.M., Laird, N.M. and Ware, J.H. (2004) Applied Longitudinal Analysis. Wiley-Interscience.

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

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.

Hoffman, M.D., Blei, D.M. and Bach, F. (2010) Online learning for latent Dirichlet allocation. Advances in Neural Information Processing Systems, 23, 856–864.

Hoffman, M.D., Blei, D.M., Wang, C. and Paisley, J. (2012) Stochastic variational inference. Available at http://arxiv.org/abs/1206.7051

Honkela, A., Tornio, M., Raiko, T. and Karhunen, J. (2008) Natural conjugate gradient in variational inference. Neural Information Processing, 305–314.

Ji, C., Shen, H. and West, C. (2010) Bounded approximations for marginal likelihoods. Available at http://ftp.stat.duke.edu/WorkingPapers/10-05.pdf
Jank, W. (2006) Implementing and Diagnosing the Stochastic Approximation EM Algorithm. *Journal of Computational and Graphical Statistics*, 15, 803–829.

Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., Saul, L.K. (1999) An introduction to variational methods for graphical models. *Machine Learning*, 37, 183–233.

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.

Kenward, M.G. and Molenberghs, G. (1998) Likelihood based frequentist inference when data are missing at random *Statistical Science*, 13, 236–247.

Knowles, D.A., Minka, T.P. (2011) Non-conjugate variational message passing for multinomial and binary regression. *Advances in Neural Information Processing Systems*, 25, 1701–1709.

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

Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000) WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10, 325–337.

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

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.

Ormerod, J.T. and Wand, M.P. (2012) Gaussian variational approximate inference for generalized linear mixed models. *Journal of Computational and Graphical Statistics*, 21, 2–17.

Paisley, J., Blei, D.M. and Jordan, M.I. (2012) Variational Bayesian inference with stochastic search. Appearing in *Proceedings of the 29th International Conference on Machine Learning*.

Papaspiliopoulos, O., Roberts, G.O. and Sköld, M. (2003) Non-centered parametrizations for hierarchical models and data augmentation. *Bayesian Statistics 7*, 307–326.

Papaspiliopoulos, O., Roberts, G.O. and Sköld, M. (2007) A general framework for the parametrization of hierarchical models. *Statistical Science*, 22, 59–73.

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.

Robbins, H. and Monro, S. (1951) A Stochastic Approximation Method. *Annals of Mathematical Statistics* 22, 400–407.

Salimans, T. and Knowles, D.A. (2012) Fixed-form variational posterior approximation through stochastic linear regression. Available at [http://arxiv.org/abs/1206.6679](http://arxiv.org/abs/1206.6679).

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

Spall, J.C. (2003) *Introduction to Stochastic Search and Optimization: Estimation, Simulation and Control*. New Jersey: Wiley.

Spiegelhalter, D.J., Aylin, P., Best, N.G., Evans, S.J.W. and Murray, G.D. (2002) Commissioned Analysis of Surgical Performance Using Routine Data: Lessons from the Bristol Inquiry. *Journal of the Royal Statistical Society. Series A*, 165, 191–231.

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

Tan, S.L. and Nott, D.J. (2012) Variational inference for generalized linear mixed models using partially non-centered parametrizations. Available at [http://arxiv.org/abs/1205.3906](http://arxiv.org/abs/1205.3906).

Wang, C., Paisley, J. and Blei, D.M. (2011) Online variational inference for the hierarchical Dirichlet process. *Journal of Machine Learning Research - Proceedings Track*, 15, 752–760.

Welling, M. and Teh, Y.W. (2011) Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, 681–688.

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

Woolson, R.F. and Clarke, W.R. (1984) Analysis of categorical incomplete longitudinal data. *Journal of the Royal Statistical Society. Series A*, 147, 87–99.

Zhou, Y., Little, R.J.A. and Kalbfleisch, J.D. (2010) Block-conditional missing at random models for missing data. *Statistical Science*, 25, 517-532.

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.