Model reparametrization for improving variational inference

Linda S. L. Tan[1]

18 May 2018

Abstract

In this article, we propose a strategy to improve variational Bayes inference for a class of models whose variables can be classified as global (common across all observations) or local (observation specific) by using a model reparametrization. In particular, an invertible affine transformation is applied on the local variables so that their posterior dependency on the global variables is minimized. The functional form of this transformation is deduced by approximating the conditional posterior distribution of each local variable given the global variables by a Gaussian distribution via a second order Taylor expansion. Variational inference for the reparametrized model is then obtained using stochastic approximation techniques. Our approach can be readily extended to large datasets via a divide and recombine strategy. Application of the methods is illustrated using generalized linear mixed models.

Keywords: Model reparametrization; variational approximation; stochastic approximation; generalized linear mixed models; partially non-centered

1 Introduction

It is well known that the parametrization of a model can have a huge impact on the performance of the methods used for obtaining statistical inference. In classical statistics, *data augmentation* strategies for speeding up EM algorithms (Dempster et al., 1977) have been developed for a wide range of models with missing data, such as multivariate t-models (Meng and van Dyk, 1997), mixed effects models (Meng and van Dyk, 1998) and Gaussian state space models (Tan, 2017). As the rate of convergence of EM algorithms depends on the proportion of missing information in the augmented data, these strategies often involve a transformation of the missing data so that the missing information is minimized. On the other hand, Markov chain Monte Carlo (MCMC) methods are often used in Bayesian statistics to obtain inference from models with intractable posterior distributions. The Gibbs sampler, for instance, has been found to converge slowly for normal and generalized linear mixed models when there is weak identifiability or high correlation among the variables (Gelfand et al., 1995, 1996). To improve the convergence of MCMC samplers, strategies similar to the data augmentation techniques for EM algorithms have also been considered and the main goal here is to minimize the correlation among different

[1]Department of Statistics and Applied Probability, National University of Singapore, 6 Science Drive 2, Singapore 117546. Email: statsll@nus.edu.sg.

groups of variables. However, as the "missing data" are regarded as part of the variables in the Bayesian context, such strategies are often referred to as a reparametrization of the model (Browne et al., 2009).

Two popular parametrizations for hierarchical models are the *centered* and *noncentered* parametrizations. A simple illustration of such parametrizations is as follows. Suppose y, x and θ represent the observed data, missing data and model parameters respectively, and

$$y|x \sim N(x, 1),$$
$$x|\theta \sim N(\theta, \sigma^2),$$

where σ^2 is known. In this case, we can transform the missing data by introducing $\tilde{x} = x - \theta$ and reparametrize the model as $y|\tilde{x} \sim N(\tilde{x} + \theta, 1)$ and $\tilde{x} \sim N(0, \sigma^2)$. The former is termed the centered parametrization since the missing data is "centered" about θ while the latter is the noncentered one. We can even consider transforming the missing data as $\tilde{x} = x - w\theta$, where $0 \leq w \leq 1$ is some working parameter that can be optimized to maximize the rate of convergence in say a Gibbs sampler. Papaspiliopoulos et al. (2003, 2007) refer to this as a partially noncentered parametrization. Note that the centered and noncentered parametrizations are obtained when $w = 0$ and $w = 1$ respectively. They demonstrate that centering and noncentering are complementary to each other in the sense that if the Gibbs sampler converges very slowly under one parametrization, it will converge much faster under the other. Partial noncentering, which lies on the continuum between centering and noncentering, also has the potential to perform better than both in some cases. Yu and Meng (2011) introduce an ancillarity-sufficiency strategy which interweaves the centered and noncentered parametrizations to improve the efficiency of MCMC algorithms.

In recent years, variational approximation methods which originate from machine learning (Jacobs et al., 1991) have become an increasingly popular alternative to MCMC methods for estimating posterior densities due to its ability to scale up to high-dimensional data. Suppose y is the observed data and θ is the set of variables in a model. In variational approximation, some restriction is placed on the approximating density $q(\theta)$ so that it is more tractable and the Kullback-Leibler divergence between $q(\theta)$ and $p(\theta|y)$ is then minimized subjected to these restrictions (see Ormerod and Wand, 2010). Common restrictions include assumption of a certain parametric form such as a Gaussian density or a product of densities form, where $q(\theta) = \prod_{i=1}^{n} q_i(\theta_i)$ for some partition $\theta = \{\theta_1, \ldots, \theta_n\}$ of θ. The latter is also known as mean-field variational Bayes (Attias, 1999). While a variational Bayes approximation can have many advantages such as being low-dimensional, quick to converge, having closed form updates (for conditionally conjugate models) and scalability to large datasets (Hoffman et al., 2013), the resulting approximation can be poor if there are strong posterior dependencies between the θ_i's which are not being

captured (Wang and Titterington, 2005).

Tan and Nott (2013, 2014) demonstrate that partially noncentered parametrizations can also be useful in the context of variational Bayes inference for generalized linear mixed models (GLMMs). To illustrate their approach, suppose β, Ω and b_i denote the regression coefficients, random effects covariance matrix and random effects for subject i for $i = 1, \ldots, n$. For simplicity, suppose there are random effects corresponding to each regression coefficient as well. By transforming the random effects so that $\tilde{b}_i = b_i + w_i\beta$ for some working parameter w_i, it is possible to tune w_i so that the posterior dependency between $\tilde{b}_i$ and β is minimized. A variational approximation of the form $q(\Omega)q(\beta)\prod_{i=1}^n q(\tilde{b}_i)$ is thus less restrictive and a better approximation of the posterior than $q(\Omega)q(\beta)\prod_{i=1}^n q(b_i)$. Tan and Nott (2013) demonstrate that in the context of variational Bayes, partial noncentering not only leads to improvements in the quality of posterior approximation but also in the rate of convergence as well. However, as the transformation does not take into account posterior dependencies between b_i and Ω, there remains difficulties in estimating the posteriors of Ω as well as regression coefficients which cannot be centered accurately using their approach.

In this article, we consider variational Bayesian inference for a class of models where the variables can be classified as *global* variables which are common across across all observations and *local* variables which are observation specific. Our goal is to obtain a low-dimensional approximation to $p(\theta|y)$ that is scalable to large datasets using variational Bayes. To achieve this aim, we propose using reparametrization techniques to transform the local variables so that the posterior dependency between the local and global variables is minimized. In particular, we consider an invertible affine transformation that is a function of the global parameters. The functional form of the affine transformation is model dependent and is deduced by considering a second order Taylor approximation to the conditional distribution of the local variables given the global variables. We demonstrate that this transformation can be considered as a generalization of the partially noncentered parametrization when the location and scale parameters are both taken into account. Christensen et al. (2006) consider a similar approach of approximating the random effects conditional posteriors in spatial GLMMs using Taylor approximation for improving the efficiency of MCMC methods. We then consider independent Gaussian approximations to the posteriors of the global parameters and individual random effects. As it is not possible to obtain closed form updates of the variational parameters, optimization is performed using stochastic gradient ascent via the doubly stochastic variational algorithm of Titsias and Lázaro-Gredilla (2014). We also demonstrate how our variational approximation can be easily extended to large datasets by using a divide and recombine strategy (Broderick et al., 2013) and Tran et al. (2016). We illustrate the application of this approach using GLMMs. However, our approach can be extended to a wider class of models including, for instance, the mixed logit mod-

els of discrete choice. We refer to our approach as reparametrized variational inference (RVI) where a model reparametrization via an affine transformation is first undertaken to minimize posterior dependencies between global and local variables before variational inference is obtained.

Other approaches to relax the independence assumption in variational Bayes exist in the literature (Gershman et al., 2012; Salimans and Knowles, 2013; Rezende and Mohamed, 2015) and we discuss some closely related work in detail below. Hoffman and Blei (2015) consider models with global and local variables as well and they develop an approach known as structured stochastic variational inference, which allows the local variables to depend explicitly on the global variables in their variational posterior through some function $\gamma(\cdot)$. However, their approach is limited to conditionally conjugate models and γ is optimized using numerical methods by maximizing a local lower bound. Titsias (2017) also propose model reparametrization techniques using affine transformations for improving variational inference. However, the parameters of the affine transformation parameters are optimized by minimizing the Kullback Leibler divergence between the variational approximation and a density obtained after application of a variable number of MCMC steps. Our approach differs from Titsias (2017) as the functional form of our affine transformation is deduced from a second-order Taylor approximation to the conditional posteriors of the local variables and remain fixed during optimization of the variational parameters. Kucukelbir et al. (2016) develop an automatic differentiation variational inference algorithm in Stan, where the approximating density is allowed to be a full Gaussian approximation. However, there may be difficulties in inferring such a high-dimensional approximation for complex models such as GLMMs where there is a local variable for every observation. Tan and Nott (2018) also consider a Gaussian approximation to the posterior but posterior dependency among variables is captured through sparse precision matrices and the resulting approximation is lower in dimension.

Our article is organized as follows. Section 2 specifies the model and defines the affine transformation. The variational algorithm is described in Section 4 and Section 5 describes how the gradients are computed. Methods are extending the variational algorithm to large datasets are described in Section 6. The application to GLMMs is described in Section 7 and the results are presented in Section 8. Section 9 concludes.

2 Reparametrization using affine transformations

Suppose y_i is the ith observation and b_i is the $r \times 1$ vector of variables specific to observation i for $i = 1, \ldots, n$. Let $y = [y_1^T, \ldots, y_n^T]^T$, $b = [b_1^T, \ldots, b_n^T]^T$ denote the vector of local variables, θ_G denote the $g \times 1$ vector of global variables, $\theta = [\theta_G^T, b^T]^T$ and $d = nr + g$ be

the length of θ. We assume that the joint density of the model is of the form

$$p(y, \theta) = p(\theta_G) \prod_{i=1}^{n} p(y_i, b_i, \theta_G)$$

The posterior distribution of θ thus has the structure

$$p(\theta|y) = p(\theta_G|y) \prod_{i=1}^{n} p(b_i|\theta_G, y_i),$$

where the conditional posteriors of the local variables are independent given the global variables. Suppose we assume

$$q(\theta) = q(\theta_G) \prod_{i=1}^{n} q(b_i). \tag{1}$$

In this case, $q(\theta)$ can be a poor approximation to $p(\theta|y)$ if there are strong dependencies between the global variables θ_G and local variables $\{b_i\}$. To improve the variational approximation, we reparametrize the model by applying an invertible affine transformation to each local variable, $\tilde{b}_i = f(b_i|\theta_G)$, so that the posterior dependency of the transformed local variables $\{\tilde{b}_i\}$ on the global variables is minimized. Subsequently, we estimate the posterior of $\tilde{\theta} = [\theta_G^T, \tilde{b}^T]^T$ using a variational approximation of the form

$$q(\tilde{\theta}) = q(\theta_G) \prod_{i=1}^{n} q(\tilde{b}_i), \tag{2}$$

where $\tilde{b} = [\tilde{b}_1^T, \ldots, \tilde{b}_n^T]^T$. The product density assumption in (2) is less restrictive than in (1) as the posterior dependency of $\tilde{b}_i$ on θ_G has been minimized. It is a more accurate reflection of the dependency structure in the true posterior and hence, the variational approximation in (2) is expected to be much better than that in (1).

To minimize the dependency of local variables on the global variables, we propose an invertible affine transformation of each local variable b_i of the form

$$\tilde{b}_i = f(b_i|\theta_G) = L_i^{-1}(b_i - \lambda_i), \tag{3}$$

where λ_i (a vector of length r) and L_i (an $r \times r$ lower triangular matrix) are functions of θ_G. The inverse transformation is $b_i = f^{-1}(b_i|\theta_G) = L_i\tilde{b}_i + \lambda_i$. The functional forms of $\{\lambda_i, L_i\}$ are to be deduced from the conditional posterior distribution of $p(b_i|\theta_G, y_i)$. The motivation is as follows. Suppose $p(b_i|\theta_G, y_i)$ can be approximated by a normal distribution with mean λ_i and covariance matrix Λ_i, and $L_iL_i^T$ is the unique Cholesky decomposition of Λ_i where the diagonal elements of L_i are positive. Then $\tilde{b}_i|\theta_G, y \sim N(0, I_r)$ approximately, which implies that $\tilde{b}_i$ is almost independent of θ_G in the posterior.

2.1 Motivating example: linear mixed model

We illustrate how to deduce the functional forms of $\{\lambda_i, L_i\}$ using the linear mixed model. We also draw parallels between the affine transformation and the partially noncentered parametrization (Papaspiliopoulos et al., 2003, 2007). We show that the affine transformation is a generalization of the partially noncentered parametrization for GLMMs introduced in Tan and Nott (2013) when the variance components are unknown.

Suppose $y_i = [y_{i1}, \ldots, y_{in_i}]^T$ is the vector of responses for the ith subject. For $i = 1, \ldots, n$, $j = 1, \ldots, n_i$,

$$y_{ij} = X_{ij}^T \beta + Z_{ij}^T b_i + \sigma^2,$$

where β is a $p \times 1$ vector of fixed effects, $b_i \sim N(0, \Omega)$ is a $r \times 1$ vector of random effects, X_{ij} and Z_{ij} are covariates of length p and r respectively, and σ^2 is assumed to be known for simplicity. Then

$$p(b_i|\beta, \Omega, y_i) \propto \exp\left[-\frac{1}{2}\{b_i^T(\Omega^{-1} + Z_i^T Z_i/\sigma^2)b_i - 2b_i^T Z_i^T(y_i - X_i\beta)/\sigma^2\}\right].$$

Hence $b_i|\beta, \Omega, y_i \sim N(\lambda_i, \Lambda_i)$ where $\Lambda_i^{-1} = \Omega^{-1} + Z_i^T Z_i/\sigma^2$ and $\lambda_i = \Lambda_i Z_i^T(y_i - X_i\beta)/\sigma^2$. We seek a transformation such that the posterior dependence of $\tilde{b}_i$ on $\{\beta, \Omega\}$ is minimized.

Suppose Ω is known and $X_i = Z_i$ as in Tan and Nott (2013). Then $b_i|\beta, y_i$ depends on β only through its mean and it suffices to consider the transformation,

$$\tilde{b}_i = b_i + \Lambda_i X_i^T X_i \beta/\sigma^2, \tag{4}$$

where $L_i = I_r$ and $\lambda_i = -\Lambda_i X_i^T X_i \beta/\sigma^2$ in (3). Then $\tilde{b}_i$ is independent of β in the posterior since $\tilde{b}_i|\beta, y \sim N(\Lambda_i X_i^T y_i/\sigma^2, \Lambda_i)$. Now even if we assume $q(\tilde{\theta}) = q(\beta)\prod_{i=1}^n q(\tilde{b}_i)$, the resulting $q(\tilde{\theta})$ can be shown to be equal to the true posterior as this product density structure is obeyed in the true posterior. Tan and Nott (2013) introduced a partially noncentered parametrization to improve the quality and rate of convergence of variational Bayes inference for generalized linear mixed models which sets

$$\tilde{b}_i = b_i + (I_r - W_i)\beta.$$

Here W_i is a partial noncentering parameter that can be tuned. When $W_i = 0$, the parametrization is *centered* in the sense that $\tilde{b}_i$ has a normal prior distribution centered around the fixed effects β. When $W_i = I_r$, $\tilde{b}_i$ has zero mean in the normal prior and the parametrization is *noncentered*. For the linear mixed model, the optimal value of W_i can be shown to be $\Lambda_i\Omega^{-1}$, which leads to instant convergence and recovery of the true posterior in the variational Bayes algorithm. This optimal reparametrization is equivalent to (4) as $I_r - W_i = \Lambda_i(\Lambda_i^{-1} - \Omega^{-1}) = \Lambda_i X_i X_i^T/\sigma^2$. The centered and noncentered parametrizations correspond to setting $L_i = I_r$ and $\lambda_i = -\beta$ and 0 respectively. Thus the

optimal partially noncentered parametrization can be interpreted as a transformation of b_i so that it becomes independent of β in the posterior.

Now consider the general case where Ω is unknown and X_i may not be equal to Z_i. We can consider $\tilde{b}_i = L_i^{-1}(b_i - \lambda_i)$, where$L_i L_i^T$ is the unique Cholesky decomposition of Λ_i and L_i is lower triangular with positive diagonal elements. Then $\tilde{b}_i|\beta, \Omega, y_i \sim N(0, I_r)$ and hence $\tilde{b}_i$ is independent of $\{\beta, \Omega\}$ in the posterior.

For linear mixed models, it is easy to identify the forms of λ_i and L_i as $p(b_i|\beta, \Omega, y_i)$ is a normal distribution. More generally, we can use second-order Taylor expansions to obtain a Gaussian approximation to $p(b_i|\theta_G, y_i)$.

3 Notation

For any $r \times r$ matrix A, let $\text{diag}(A)$ denote the diagonal of A, $\text{dg}(A)$ denote the diagonal matrix derived from A by setting all non-diagonal elements to zero and $\bar{A}$ denote the lower triangular matrix derived from A by setting all superdiagonal elements to zero. Let $\text{vec}(A)$ denote the vector of length r^2 obtained by stacking the columns of A under each other from left to right and $v(A)$ denote the vector of length $r(r+1)/2$ obtained from $\text{vec}(A)$ by eliminating all superdiagonal elements of A. Let E_r denote the $r(r+1)/2 \times r^2$ elimination matrix such that $E_r\text{vec}(A) = v(A)$ and K_r denote the $r^2 \times r^2$ commutation matrix such that $K_r\text{vec}(A) = \text{vec}(A^T)$. If A is lower triangular, then $E_r^T v(A) = \text{vec}(A)$. If A is symmetric, then $D_r v(A) = \text{vec}(A)$, where D_r is the $r^2 \times r(r+1)/2$ duplication matrix. The Kronecker product between any two matrices is denoted by $\otimes$. Scalar functions applied to vector arguments are evaluated element by element.

4 Variational inference

We consider a variational approximation for the reparametrized model of the form in (2), where $q(\theta_G)$ and $q(\tilde{b}_i)$ are both assumed to be Gaussian distributions. This implies that $q(\tilde{\theta})$ is a joint Gaussian distribution, $N(\mu, \Sigma)$, where Σ is a block diagonal matrix with $n+1$ blocks. The first block is of order g and each of the subsequent n blocks is of order r. Let CC^T be the unique Cholesky factorization of Σ where C is lower triangular with positive diagonal elements. Then $q(\tilde{\theta}) = q(\tilde{\theta}|\mu, C)$ and the variational parameters $\{\mu, C\}$ are optimized so that the Kullback-Leibler divergence between $q(\tilde{\theta}|\mu, C)$ and the true posterior $p(\tilde{\theta}|y)$,

$$D_{\text{KL}}(q||p) = \int q(\tilde{\theta}|\mu, C) \log\{q(\tilde{\theta}|\mu, C)/p(\tilde{\theta}|y)\} \text{d}\tilde{\theta},$$

is minimized. By Jensen's inequality, this is equivalent to maximizing a lower bound $\mathcal{L}$ on the marginal likelihood $p(y)$, where

$$\mathcal{L} = \mathcal{L}(\mu, C^q) = E_q\{\log p(y, \tilde{\theta}) - \log q(\tilde{\theta}|\mu, C)\}, \tag{5}$$

and E_q denotes expectation with respect to $q(\tilde{\theta}|\mu, C)$.

For non-conjugate models such as generalized linear mixed models or mixed logit models, $E_q\{\log p(y, \tilde{\theta})\}$ cannot be evaluated in closed form and hence the lower bound is intractable. To overcome this limitation, we maximize $\mathcal{L}$ with respect to $\{\mu, C\}$ using doubly stochastic variational inference (Titsias and Lázaro-Gredilla, 2014). This approach is based on stochastic approximation (Robbins and Monro, 1951), where at each iteration t, the variational parameters are updated by taking a small step ρ_t in the direction of the stochastic gradients,

$$\mu^{(t)} = \mu^{(t-1)} + \rho_t \widehat{\nabla}_\mu \mathcal{L}, \quad v(C^{(t)}) = v(C^{(t-1)}) + \rho_t \widehat{\nabla}_{v(C)} \mathcal{L}. \tag{6}$$

The stochastic gradients $\widehat{\nabla}_\mu \mathcal{L}$ and $\widehat{\nabla}_{v(C)} \mathcal{L}$ are unbiased estimates of the true gradients $\nabla_\mu \mathcal{L}$ and $\nabla_{v(C)} \mathcal{L}$ respectively. Under mild regularity conditions, the stochastic approximation algorithm will converge to a local maximum if the stepsize $\{\rho_t\}$ satisfy $\sum_t \rho_t = \infty$, $\sum_t \rho_t^2 < \infty$ and the lower bound is concave (Spall, 2003). Doubly stochastic variational inference is so named as the stochasticity in the gradients can arise from two sources, one due to analyzing the full data set in mini-batches and the second due to sampling from the variational distribution.

From (5), unbiased estimates of the true gradients can be constructed by sampling $\tilde{\theta}$ from $q(\tilde{\theta}|\mu, C)$ directly. However, this approach often results in estimators with very high variance (Paisley et al., 2012). Hence we employ the "reparametrization trick" (Kingma and Welling, 2014; Rezende et al., 2014; Titsias and Lázaro-Gredilla, 2014), which introduces an invertible transformation $s = C^{-1}(\tilde{\theta} - \mu)$. This implies that the density of s, denoted by $\phi(s)$, is that of $N(0, I_d)$. Let $\ell(\tilde{\theta}) = \log p(y, \tilde{\theta})$. From (5),

$$\mathcal{L}(\mu, C) = E_\phi\{\ell(\tilde{\theta}) - \log q(\tilde{\theta}|\mu, C)\}. \tag{7}$$

where E_ϕ denotes expectation with respect to $\phi(s)$ and $\tilde{\theta} = Cs + \mu$. Unbiased estimates of the true gradients can thus be constructed by sampling s from $\phi(s)$ instead of $\tilde{\theta}$ from $q(\tilde{\theta}, \mu, C)$. The advantage of this reparametrization is that the log joint density is now a function of the variational parameters $\{\mu, C\}$, which enables gradient information from the log joint density to be utilized effectively.

4.1 Unbiased estimates of stochastic gradients

In this section, we derive unbiased estimators of the true gradients $\nabla_\mu \mathcal{L}$ and $\nabla_{v(C)} \mathcal{L}$. We show that several unbiased estimators can be constructed from (7) but some has nicer properties at convergence than others. The estimators presented below are based on a single sample s generated from $\phi(s)$, as it has been observed in many cases that there one sample alone provides sufficient gradient information (Titsias and Lázaro-Gredilla, 2014). It is, however, straightforward to extend the estimators to a larger number of samples by averaging if necessary. For example, if $\hat{e}(s)$ is an unbiased estimator, then $\sum_{m=1}^{M} \hat{e}(s_m)/M$ is also an unbiased estimator if $s_m \sim N(0, I_d)$ independently for $m = 1, \ldots, M$.

First, if we evaluate the second term in (7) analytically, then

$$\mathcal{L} = E_\phi\{\ell(\tilde{\theta})\} + \log|C| + C',$$

where C' is a constant that does not depend on $\{\mu, C\}$. The gradients of $\mathcal{L}$ with respect to μ and $v(C)$ are then given by

$$\nabla_\mu \mathcal{L} = E_\phi\{\nabla_{\tilde{\theta}} \ell(\tilde{\theta})\}, \quad \nabla_{v(C)} \mathcal{L} = E_\phi[v\{\nabla_{\tilde{\theta}} \ell(\tilde{\theta}) s^T + C^{-T}\}].$$

Details of the derivation are given in Appendix A. This leads to the unbiased estimators,

$$\widehat{\nabla}_\mu \mathcal{L}_1 = \nabla_{\tilde{\theta}} \ell(\tilde{\theta}), \quad \widehat{\nabla}_{v(C)} \mathcal{L}_1 = v\{\nabla_{\tilde{\theta}} \ell(\tilde{\theta}) s^T + C^{-T}\}. \tag{8}$$

Alternatively, we can choose not to evaluate $E_\phi\{\log q(\tilde{\theta}|\mu, C)\}$ analytically but to approximate both terms in (7) using the same samples. As $\log q(\tilde{\theta}|\mu, C)$ depends on $\{\mu, C\}$ directly as well as through $\tilde{\theta}$, we apply chain rule to obtain

$$\nabla_\mu \mathcal{L} = E_\phi\{\nabla_{\tilde{\theta}} \ell(\tilde{\theta}) - \nabla_{\tilde{\theta}} \log q(\tilde{\theta}|\mu, C) - \nabla_\mu \log q(\tilde{\theta}|\mu, C)\}, \tag{9}$$

$$\nabla_{v(C)} \mathcal{L} = E_\phi[v\{\nabla_{\tilde{\theta}} \ell(\tilde{\theta}) s^T - \nabla_{\tilde{\theta}} \log q(\tilde{\theta}|\mu, C) s^T\} - \nabla_{v(C)} \log q(\tilde{\theta}|\mu, C)], \tag{10}$$

where $\nabla_{\tilde{\theta}} \log q(\tilde{\theta}|\mu, C) = -\nabla_\mu \log q(\tilde{\theta}|\mu, C)\} = -C^{-T} s$ and

$$\nabla_{v(C)} \log q(\tilde{\theta}|\mu, C) = v\{C^{-T}(ss^T - I_d)\}.$$

If we use all the terms in (9) and (10) to construct the gradient estimators, we will obtain the same estimators in (8) after simplification. However, as $E_\phi(s) = 0$ and $E_\phi(ss^T) = I_d$,

$$E_\phi\{\nabla_{\tilde{\theta}} \log q(\tilde{\theta}|\mu, C)\} = E_\phi\{\nabla_\mu \log q(\tilde{\theta}|\mu, C)\} = 0, \quad E_\phi\{\nabla_{v(C)} \log q(\tilde{\theta}|\mu, C)\} = 0.$$

This implies that unbiased estimators can be constructed by omitting the last term in

(9) and (10). We thus obtain the second set of estimators,

$$\widehat{\nabla}_{\mu}\mathcal{L}_2 = \nabla_{\tilde{\theta}}\ell(\tilde{\theta}) + C^{-T}s, \quad \widehat{\nabla}_{v(C)}\mathcal{L}_2 = v\{\nabla_{\tilde{\theta}}\ell(\tilde{\theta})s^T + C^{-T}ss^T\}. \tag{11}$$

A third possible unbiased estimator is $\widehat{\nabla}_{\mu}\mathcal{L}_3 = \nabla_{\tilde{\theta}}\ell(\tilde{\theta}) - C^{-T}s$, which is obtained by omitting the second term in (9). We show below that the estimators $\widehat{\nabla}_{\mu}\mathcal{L}_2$ and $\widehat{\nabla}_{v(C)}\mathcal{L}_2$ are preferred as they have smaller variation at convergence by using an argument similar to that in Tan and Nott (2018).

Consider a second-order Taylor approximation to $\ell(\tilde{\theta})$ at the posterior mode $\tilde{\theta}^*$,

$$\ell(\tilde{\theta}) \approx \ell(\tilde{\theta}^*) + (\tilde{\theta} - \tilde{\theta}^*)^T\nabla^2_{\tilde{\theta}}\ell(\tilde{\theta}^*)(\tilde{\theta} - \tilde{\theta}^*)/2,$$

where $\nabla^2_{\tilde{\theta}}\ell(\tilde{\theta}^*)$ denotes the Hessian of ℓ evaluated at the mode $\tilde{\theta}^*$. This implies that $p(\tilde{\theta}|y)$ is approximately $N(\tilde{\theta}^*, -\{\nabla^2_{\tilde{\theta}}\ell(\tilde{\theta}^*)\}^{-1})$. Differentiating the Taylor approximation with respect to $\tilde{\theta}$ yields

$$\nabla_{\tilde{\theta}}\ell(\tilde{\theta}) \approx \nabla^2_{\tilde{\theta}}\ell(\tilde{\theta}^*)(\tilde{\theta} - \tilde{\theta}^*). \tag{12}$$

Since $q(\tilde{\theta}|\mu, \Sigma)$ provides a Gaussian approximation to $p(\tilde{\theta}|y)$, $\mu \approx \tilde{\theta}^*$, $\Sigma \approx -\{\nabla^2_{\tilde{\theta}}\ell(\tilde{\theta}^*)\}^{-1}$ at convergence. Thus, $\nabla_{\tilde{\theta}}\ell(\tilde{\theta}) \approx -\Sigma^{-1}(\tilde{\theta} - \mu) = -C^{-T}s$ and

$$\begin{aligned} &\widehat{\nabla}_{\mu}\mathcal{L}_1 \approx -C^{-T}s, && \widehat{\nabla}_{v(C)}\mathcal{L}_1 \approx v\{-C^{-T}ss^T + C^{-T}\}, \\ &\widehat{\nabla}_{\mu}\mathcal{L}_2 \approx 0, && \widehat{\nabla}_{v(C)}\mathcal{L}_2 \approx 0. \end{aligned}$$

The estimators $\widehat{\nabla}_{\mu}\mathcal{L}_2$ and $\widehat{\nabla}_{v(C)}\mathcal{L}_2$ are close to zero as the contributions from $\ell(\tilde{\theta})$ and $\log q(\tilde{\theta}|\mu, C)$ cancel out each other when the stochastic approximation algorithm is close to convergence. However $\widehat{\nabla}_{\mu}\mathcal{L}_1$ and $\widehat{\nabla}_{v(C)}\mathcal{L}_1$ still contain a certain amount of noise and $\widehat{\nabla}_{\mu}\mathcal{L}_3 \approx -2C^{-T}s$ is even noisier than $\widehat{\nabla}_{\mu}\mathcal{L}_2$. In addition, from (12), we have

$$\begin{aligned} &\mathrm{cov}_{\phi}(\widehat{\nabla}_{\mu}\mathcal{L}_1) = \mathrm{cov}_q(\nabla_{\tilde{\theta}}\ell(\tilde{\theta})) \approx \nabla^2_{\tilde{\theta}}\ell(\tilde{\theta}^*)\Sigma\{\nabla^2_{\tilde{\theta}}\ell(\tilde{\theta}^*)\}^T \approx \Sigma^{-1}. \\ &\mathrm{cov}_{\phi}(\widehat{\nabla}_{\mu}\mathcal{L}_2) = \mathrm{cov}_q(\nabla_{\tilde{\theta}}\ell(\tilde{\theta}) + \Sigma^{-1}\tilde{\theta}) \approx \{\nabla^2_{\tilde{\theta}}\ell(\tilde{\theta}^*) + \Sigma^{-1}\}\Sigma\{\nabla^2_{\tilde{\theta}}\ell(\tilde{\theta}^*) + \Sigma^{-1}\}^T \approx 0. \end{aligned}$$

which supports the claim that $\widehat{\nabla}_{\mu}\mathcal{L}_2$ is less noisy than $\widehat{\nabla}_{\mu}\mathcal{L}_1$ at convergence.

4.2 Algorithm implementation

As the update for $v(C)$ in (6) do not ensure that the diagonal elements of C remain positive, we introduce the lower triangular matrix C^* such that $C^*_{ii} = \log(C_{ii})$ and $C^*_{ij} = C_{ij}$ if $i \neq j$ and the stochastic gradient updates are applied to C^* instead. Let $D_C = \mathrm{diag}\{v(\mathrm{dg}(C) + J_d - I_d)\}$ where J_d is a matrix of ones of order d. Then $\nabla_{v(C')}\mathcal{L} = D_C\nabla_{v(C)}\mathcal{L}$. The stochastic variational algorithm using the estimators $\widehat{\nabla}_{\mu}\mathcal{L}_2$ and $\widehat{\nabla}_{v(C)}\mathcal{L}_2$ is summarized in Algorithm 1. For the stepsize, we will use the ADADELTA method of

Initialize $\mu^{(1)} = 0$ and $C^{(1)} = I_d$. For $t = 1, \ldots, T$,

1. Generate $s \sim N(0, I_d)$.
2. Compute $\tilde{\theta}^{(t)} = C^{(t)}s + \mu^{(t)}$ and $\mathcal{G}^{(t)} = \nabla_{\tilde{\theta}}\ell(\tilde{\theta}^{(t)}) + {C^{(t)}}^{-T}s$.
3. Update $\mu^{(t+1)} = \mu^{(t)} + \rho_t \mathcal{G}^{(t)}$.
4. Update $v(C'^{(t+1)}) = v(C'^{(t)}) + \rho_t D_C v\{\mathcal{G}^{(t)}s^T\}$.
5. Compute $C^{(t+1)}$ from $C'^{(t+1)}$.

Algorithm 1: RVI algorithm.

Zeiler (2012) which is adaptive and automatically tuned to different parameters.

5 Gradient of the log joint density

To implement Algorithm 1, we require $\nabla_{\tilde{\theta}}\ell(\tilde{\theta}) = [\nabla_{\theta_G}\ell(\tilde{\theta}), \nabla_{\tilde{b}_1}\ell(\tilde{\theta}), \ldots, \nabla_{\tilde{b}_n}\ell(\tilde{\theta})]$. As $b_i = L_i\tilde{b}_i + \lambda_i$, $p(\tilde{b}_i|\omega) = p(b_i|\omega)|L_i|$. Hence $p(y, \tilde{\theta}) = p(y, \theta)\prod_{i=1}^n |L_i|$ and the log joint density is

$$\ell(\tilde{\theta}) = \log p(\theta_G) + \sum_{i=1}^n \{\log p(y_i, b_i|\theta_G) + \log|L_i|\}.$$

Let d denote the differential operator (see Magnus and Neudecker, 1980) and $a_i = \nabla_{b_i} \log p(y_i, b_i|\theta_G)$. Differentiating $\ell(\tilde{\theta})$ with respect to $\tilde{b}_i$,

$$\mathrm{d}\ell(\tilde{\theta}) = a_i^T L_i \mathrm{d}\tilde{b}_i.$$

Hence $\nabla_{\tilde{b}_i}\ell(\tilde{\theta}) = L_i^T a_i$. Suppose θ_G is partitioned as $[\theta_{G_1}^T, \ldots, \theta_{G_M}^T]^T$. Differentiating $\ell(\tilde{\theta})$ with respect to θ_{G_m} for $m = 1, \ldots, M$,

$$\begin{aligned}\mathrm{d}\ell(\tilde{\theta}) = \sum_{i=1}^n [a_i^T \mathrm{d}b_i + \{\nabla_{\theta_{G_m}} \log p(y_i, b_i|\theta_G)\}^T \mathrm{d}\theta_{G_m} + \mathrm{tr}(L_i^{-1}\mathrm{d}L_i)] \\ + \{\nabla_{\theta_{G_m}} \log p(\theta_G)\}^T \mathrm{d}\theta_{G_m}.\end{aligned}$$

To find the differential of the Cholesky factor, we can differentiate $L_i L_i^T = \Lambda_i$ and then multiply by L_i^{-1} on the left and L_i^{-T} on the right (Murray, 2016):

$$\begin{aligned}(\mathrm{d}L_i)L_i^T + L_i(\mathrm{d}L_i)^T &= \mathrm{d}\Lambda_i, \\ L_i^{-1}(\mathrm{d}L_i) + (\mathrm{d}L_i)^T L_i^{-T} &= L_i^{-1}\mathrm{d}\Lambda_i L_i^{-T}.\end{aligned}$$

On the left-hand side, the first term is lower triangular and the second term is upper triangular (transpose of first term). If we define a function $k(\cdot)$ such that for any square

matrix A, $k(A) = \bar{A} - \mathrm{dg}(A)/2$, then

$$\mathrm{d}L_i = L_i k(A_i). \tag{13}$$

where $A_i = L_i^{-1}\mathrm{d}\Lambda_i L_i^{-T}$. Thus $\mathrm{tr}(L_i^{-1}\mathrm{d}L_i) = \mathrm{tr}(k(A_i)) = \mathrm{tr}(A_i)/2 = \mathrm{vec}(\Lambda_i^{-1})^T\mathrm{dvec}(\Lambda_i)/2$. By Lemma 3.3, 3.4, 3.6 and 4.4 of Magnus and Neudecker (1980),

$$\begin{aligned}
\mathrm{vec}\{k(A_i)\} &= \mathrm{vec}(\bar{A}_i) - \mathrm{vec}\{\mathrm{dg}(A_i)\}/2 \\
&= E_r^T E_r \mathrm{vec}(A_i) - E_r^T E_r K_r E_r^T E_r \mathrm{vec}(A_i)/2 \\
&= E_r^T (2I_{r(r+1)/2} - E_r K_r E_r^T) E_r (L_i^{-1} \otimes L_i^{-1}) \mathrm{dvec}(\Lambda_i)/2 \\
&= E_r^T D_r^T D_r E_r (L_i^{-1} \otimes L_i^{-1}) D_r \mathrm{d}v(\Lambda_i)/2 \\
&= E_r^T D_r^T (L_i^{-1} \otimes L_i^{-1}) \mathrm{dvec}(\Lambda_i)/2.
\end{aligned}$$

Thus

$$\begin{aligned}
a_i^T \mathrm{d}b_i &= a_i^T (\mathrm{d}L_i)\tilde{b}_i + a_i^T \mathrm{d}\lambda_i \\
&= a_i^T L_i k(A_i)\tilde{b}_i + a_i^T \mathrm{d}\lambda_i \\
&= \mathrm{vec}(B_i)^T \mathrm{vec}(k(A_i)) + a_i^T \mathrm{d}\lambda_i \\
&= \{(L_i^{-T} \otimes L_i^{-T}) D_r v(B_i)\}^T \mathrm{dvec}(\Lambda_i)/2 + a_i^T \mathrm{d}\lambda_i, \\
&= \mathrm{vec}(L_i^{-T} \tilde{B}_i L_i^{-1}/2)^T \mathrm{dvec}(\Lambda_i) + a_i^T \mathrm{d}\lambda_i,
\end{aligned}$$

where $B_i = L_i^T a_i \tilde{b}_i^T$ and $\tilde{B}_i = \bar{B}_i + \bar{B}_i^T - \mathrm{dg}(B_i)$. Hence

$$\begin{aligned}
\nabla_{\theta_{G_m}} \ell(\tilde{\theta}) = &\sum_{i=1}^{n} (\nabla_{\theta_{G_m}} \lambda_i) a_i + \sum_{i=1}^{n} \{\nabla_{\theta_{G_m}} \mathrm{vec}(\Lambda_i)\} \mathrm{vec}(\Lambda_i^{-1} + L_i^{-T} \tilde{B}_i L_i^{-1})/2) \\
&+ \sum_{i=1}^{n} \nabla_{\theta_{G_m}} \log p(y_i, b_i | \theta_G) + \nabla_{\theta_{G_m}} \log p(\theta_G).
\end{aligned}$$

6 Extension to large data sets

In this section, we discuss how the variational inference for this class of models can be applied to large data sets using a divide and recombine strategy (Broderick et al., 2013; Tran et al., 2016). Suppose the n observations in the data are partitioned into S parts such that $y = (y^1, \ldots, y^S)^T$ and let $\tilde{b}^s$ be the set of random effects corresponding to y^s.

The true posterior distribution can be written as

$$\begin{aligned} p(\tilde{\theta}|y) = p(\tilde{b}, \theta_G|y) &\propto p(\theta_G) \prod_{s=1}^{S} p(y^s|\tilde{b}^s, \theta_G) \\ &\propto \frac{\prod_{s=1}^{S} \{p(y^s|\tilde{b}^s, \theta_G) p(\theta_G)\}}{p(\theta_G)^{s-1}} \\ &\propto \frac{\prod_{s=1}^{S} p(\theta_G, \tilde{b}^s|y^s)}{p(\theta_G)^{s-1}}. \end{aligned}$$

If we replace $p(\theta_G, \tilde{b}^s|y^s)$ by our variational approximation $q^s(\theta_G)q^s(\tilde{b}^s)$, which is obtained using the portion of the data y^s only, then we have

$$p(\tilde{b}, \theta_G|y) \propto \frac{\prod_{s=1}^{S} \{q^s(\theta_G)q^s(\tilde{b}^s)\}}{p(\theta_G)^{s-1}}.$$

approximately. Thus if we integrate out the random effects $\tilde{b}$ on both sides, we have

$$p(\theta_G|y) \propto \frac{\prod_{s=1}^{S} q^s(\theta_G)}{p(\theta_G)^{s-1}}.$$

Suppose that $p(\theta_G)$ is Gaussian $N(\mu_0, \Sigma_0)$. Then since $q(\theta_G)$ is Gaussian say $N(\mu_s, \Sigma_s)$, then

$$p(\theta_G|y) \propto \exp\left[-\frac{1}{2}\exp\left\{\sum_{s=1}^{S}(\theta_G - \mu_s)^T\Sigma_s^{-1}(\theta_G - \mu_s) - (s-1)(\theta_G - \mu_0)^T\Sigma_0^{-1}(\theta_G - \mu_0)\right\}\right].$$

Thus the distribution of $p(\theta_G|y)$ can be approximated by $N(\mu, \Sigma)$, where

$$\Sigma = \left\{\sum_{s=1}^{S}\Sigma_s^{-1} - (s-1)\Sigma_0^{-1}\right\}^{-1}, \mu = \Sigma\left\{\sum_{s=1}^{S}\Sigma_s^{-1}\mu_s - (s-1)\Sigma_0^{-1}\mu_0\right\}.$$

7 Application to generalized linear mixed models

Let $y_i = [y_{i1}, \ldots, y_{in_i}]^T$ denote the vector of responses for the ith subject for $i = 1, \ldots, n$. We consider generalized linear mixed models of the form $g(\mu_{ij}) = \eta_{ij}$, where $\mu_{ij} = E(y_{ij})$, $g(\cdot)$ is a link function and the linear predictor,

$$\eta_{ij} = X_{ij}^T\beta + Z_{ij}^T b_i,$$

for $i = 1, \ldots, n$, $j = 1, \ldots, n_i$. Let β be a vector of fixed effects of length p, $b_i \sim N(0, \Omega)$ be a vector of random effects of length r for the ith subject, and X_{ij} and Z_{ij} be covariates of length p and r respectively for y_{ij}. Let WW^T be the unique Cholesky factorization of Ω^{-1}, where W is lower triangular with positive diagonal elements. The $r \times r$ matrix W^*

is defined such that $W^*_{ii} = \log W_{ii}$ and $W^*_{ij} = W_{ij}$ if $i \neq j$. We consider normal priors $N(0, \sigma^2_\beta I_p)$ for β and $N(0, \sigma^2_\omega I_{r(r+1)/2})$ for $\omega = v(W^*)$.

We focus on the Poisson mixed model where the responses are counts $y_{ij} \sim \text{Poisson}(\mu_{ij})$ with the log link, $g(\mu_{ij}) = \log(\mu_{ij})$, and the logistic mixed model with binary responses $y_{ij} \sim \text{Bernoulli}(\mu_{ij})$ with the logit link, $g(\mu_{ij}) = \text{logit}(\mu_{ij})$. Let $X_i = [X_{i1}, \ldots, X_{in_i}]^T$, $Z_i = [Z_{i1}, \ldots, Z_{in_i}]^T$ and $\eta_i = X_i\beta + Z_i b_i$. The vector of global variables is $\theta_G = [\beta^T, \omega^T]^T$, where the length of θ_G is $g = p + r(r+1)/2$. The joint distribution is $p(y, \theta) = p(\beta)p(\omega)\prod_{i=1}^n \{p(y_i|\eta_i)p(b_i|\omega)\}$ and

$$\log p(y_i, \theta) = \sum_{i=1}^n \{y_i^T \eta_i - 1^T h(\eta_i)\} - \frac{1}{2}\{\log|\Omega| + b_i^T \Omega^{-1} b_i + \beta^T\beta/\sigma^2_\beta + \omega^T\omega/\sigma^2_\omega\} + C'',$$

where the constant C'' is independent of θ. For the Poisson mixed model, $h(x) = \exp(x)$ and for the logistic mixed model, $h(x) = \log\{1 + \exp(x)\}$. Let $h^{(k)}(\cdot)$ denote the kth derivative of $h(\cdot)$.

As $p(b_i|\omega)$ is Gaussian but the likelihood $p(y_i|\eta_i)$ is not, we will first find a Gaussian approximation to the likelihood by using a second-order Taylor expansion about some estimate $\hat{\eta}_i$ of η_i and then combine this expression with $p(b_i|\omega)$ to obtain a Gaussian approximation to $p(b_i|\theta_G, y_i)$. Let $g_i = y_i - h'(\eta_i)$ and $H_i = \text{diag}\{h''(\eta_i)\}$ denote the gradient and negative Hessian of $\log p(y_i|\eta_i)$ with respect to η_i, and $\hat{g}_i$ and $\hat{H}_i$ denote the values of g_i and H_i evaluated at $\hat{\eta}_i$ respectively. Then

$$\begin{aligned} p(b_i|\theta_G, y_i) &\propto \exp\{\log p(y_i|\eta_i) + \log p(b_i|\omega)\} \\ &\propto \exp\{\eta_i^T \hat{g}_i - (\eta_i - \hat{\eta}_i)^T \hat{H}_i(\eta_i - \hat{\eta}_i)/2 - b_i^T\Omega^{-1}b_i/2\} \\ &\propto \exp\{-[b_i^T(\Omega^{-1} + Z_i^T\hat{H}_i Z_i)b_i - 2b_i^T Z_i^T\{\hat{g}_i + \hat{H}_i(\hat{\eta}_i - X_i\beta)\}]/2\} \end{aligned}$$

Thus, $b_i|\theta_G, y_i \sim N(\lambda_i, \Lambda_i)$ approximately, where

$$\Lambda_i = (\Omega^{-1} + Z_i^T\hat{H}_i Z_i)^{-1}, \quad \lambda_i = \Lambda_i Z_i^T\{\hat{g}_i + \hat{H}_i(\hat{\eta}_i - X_i\beta)\}.$$

If $\hat{\eta}_i$ is the mode of $\log p(y_i|\eta_i)$, then $\hat{g}_i = 0$, which means that $h'(\hat{\eta}_i) = y_i$. For the Poisson model, this implies that $\exp(\hat{\eta}_i) = y_i$ and $\hat{H}_i = \text{diag}(y_i)$. For the Bernoulli model, $\exp(\hat{\eta}_i)/(1 + \exp(\hat{\eta}_i)) = y_i$ and $\hat{H}_i = \text{diag}\{y_i(1 - y_i)\} = 0$. Alternatively, we can consider a direct Gaussian approximation of $p(b_i|\theta_G, y_i)$ by using a Taylor expansion about some estimate $\hat{b}_i$ of b_i. We have experimented with this option but found that it leads to a highly unstable algorithm.

We reparametrize the model by transforming each random effect b_i so that $b_i = L_i\tilde{b}_i + \lambda_i$, where $L_i L_i^T = \Lambda_i$. The gradient $\nabla_{\tilde{\theta}}\ell(\tilde{\theta})$ is $[\nabla_\beta\ell(\tilde{\theta})^T, \nabla_\omega\ell(\tilde{\theta})^T, \nabla_{\tilde{b}_1}\ell(\tilde{\theta})^T, \ldots, \nabla_{\tilde{b}_n}\ell(\tilde{\theta})^T]^T$,

where the components can be evaluated as follows. For $i = 1, \ldots, n$,

$$\nabla_{\tilde{b}_i} \ell(\tilde{\theta}) = L_i^T a_i, \quad a_i = Z_i^T g_i - \Omega^{-1} b_i,$$

As $\nabla_\beta \log p(\theta_G) = -\beta/\sigma_\beta^2$, $\nabla_\beta \log p(y_i, b_i|\theta_G) = X_i^T g_i$, $\nabla_\beta \text{vec}(\Lambda_i) = 0$ and $\nabla_\beta \lambda_i = -X_i^T \hat{H}_i Z_i \Lambda_i$, we have

$$\nabla_\beta \ell(\tilde{\theta}) = \sum_{i=1}^{n} X_i^T (g_i - \hat{H}_i Z_i \Lambda_i a_i) - \beta/\sigma_\beta^2.$$

Let $D^* = \text{diag}\{v(\text{dg}(W) + \bar{J}_r - I_r)\}$, where J_r is a matrix of ones. We have $\text{dvec}(W) = E_r^T \text{d}v(W) = E_r^T D^* \text{d}\omega$. Since $\Omega^{-1} = WW^T$, $\text{d}\Omega^{-1} = (\text{d}W)W^T + W(\text{d}W)^T$ and $\text{d}\Lambda_i = -\Lambda_i \text{d}\Omega^{-1} \Lambda_i$

$$\text{dvec}(\Lambda_i) = -\{(\Lambda_i W \otimes \Lambda_i) + (\Lambda_i \otimes \Lambda_i W) K_r\} \text{dvec}(W) = -2N_r(\Lambda_i W \otimes \Lambda_i) E_r^T D^* \text{d}\omega,$$

where $N_r = (K_r + I_{r^2})/2$. We have

$$\nabla_\omega \log p(\theta_G) = -\omega/\sigma_\omega^2, \quad \nabla_\omega \log p(y_i, b_i|\theta_G) = D^* v(b_i b_i^T W + W^{-T}),$$
$$\nabla_\omega \lambda_i = -D^* E_r \{W^T \Lambda_i \otimes \lambda_i + W^T \lambda_i \otimes \Lambda_i\}, \quad \nabla_\omega \text{vec}(\Lambda_i) = -2D^* E_r (\Lambda_i W^T \otimes \Lambda_i) N_r.$$

Thus

$$\nabla_\omega \ell(\tilde{\theta}) = D^* \sum_{i=1}^{n} v\{W^{-T} - (\Lambda_i + \lambda_i a_i^T \Lambda_i + \Lambda_i a_i \lambda_i^T + b_i b_i^T + L_i \tilde{B}_i L_i^T) W\} - \omega/\sigma_\omega^2. \tag{14}$$

8 Experimental results

In this section, we present the results of fitting the RVI algorithm to Poisson, binomial and Bernoulli GLMMs. We also use some larger datasets to investigate the quality of the results obtained via dataset partitioning and parallel processing. We compare the results of the RVI algorithm with the Gaussian variational approximation (GVA) proposed by Tan and Nott (2018), which is also based on the doubly stochastic variational algorithm of Titsias and Lázaro-Gredilla (2014). Suppose the variational approximation for θ is $N(\mu, \Sigma)$ and TT^T is a Cholesky decomposition of Σ^{-1}. Tan and Nott (2018) assume that

T and Σ^{-1} are sparse matrices of the form

$$T = \begin{bmatrix} T_{11} & \dots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & T_{nn} & 0 \\ T_{n+1,1} & \dots & T_{n+1,n} & T_{n+1,n+1} \end{bmatrix}, \quad \Sigma^{-1} = \begin{bmatrix} \Sigma_{11}^{-1} & \dots & 0 & \Sigma_{1,n+1}^{-1} \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & \Sigma_{nn}^{-1} & \Sigma_{n,n+1}^{-1} \\ \Sigma_{n+1,1}^{-1} & \dots & \Sigma_{n+1,n}^{-1} & \Sigma_{n+1,n+1}^{-1} \end{bmatrix}$$

in order to capture the dependency structure in the posterior distribution. The matrices $\Sigma_{11}^{-1}, \dots, \Sigma_{nn}^{-1}, \Sigma_{n+1,n+1}^{-1}$ correspond to the precision matrices of the local variables $b_1, \dots b_n$ and θ_G respectively while $\Sigma_{n+1,i}^{-1}$ for $i = 1, \dots, n$ help to captures the dependency between the local variables and global variables. On the other hand, RVI considers a block diagonal structure for the covariance matrix of the Gaussian approximation to $p(\tilde{\theta}|y)$ and a direct Cholesky decomposition of this covariance matrix. The number of variational parameters is thus reduced by nrg (corresponding to the off-diagonal blocks $T_{n+1}, 1, \dots, T_{n+1,n}$). Such a reduction can be significant when n, r and g are large. Posterior dependency between local and global variables is minimized via model reparametrization instead. The RVI approach does not return the posterior distributions of the random effects directly however. It considers a Gaussian approximation for the transformed random effects $\{\tilde{b}_i\}$ and the posterior distributions of $\{b_i\}$ have to be computed using simulation techniques. While this takes more work, the advantage is that the posterior distributions of $\{b_i\}$ are not restricted to be Gaussian. While GVA and RVI account for the posterior dependency between local and global variables in different manners, we find that RVI can often achieve a better posterior approximation and higher lower bound than GVA when the data is highly informative. This is because the model reparametrization is data dependent. Code for both variational algorithms are written in Julia (`https://julialang.org/`) and run on a Intel Core i5-2500 CPU @ 3.30GHz 8.0GB RAM machine. The posterior distributions estimated using MCMC via RStan(`http://mc-stan.org/interfaces/rstan`) using the same priors are regarded as the ground truth. We set $\sigma_\beta^2 = \sigma_\omega^2 = 100$ in all experiments below, representing a vague prior.

8.1 Epilepsy data

In the epilepsy data of Thall and Vail (1990), which is available in the `R` package `MASS` using `data(epil)`, $n = 59$ epileptics were assigned either a new drug Progabide or a placebo randomly. The response variable y_{ij} is the number of epileptic seizures of patient i in the two weeks before clinic visit j for $j = 1, \dots, 4$. We consider the Poisson random intercept and slope model of Breslow and Clayton (1993) where

$$\begin{aligned} \log \mu_{ij} = \beta_0 &+ \beta_{\text{Base}}\text{Base}_i + \beta_{\text{Trt}}\text{Trt}_i + \beta_{\text{Age}}\text{Age}_i \\ &+ \beta_{\text{Base}\times\text{Trt}}\text{Base}_i \times \text{Trt}_i + \beta_{\text{Visit}}\text{Visit}_{ij} + b_{i1} + b_{i2}\text{Visit}_{ij}, \end{aligned}$$

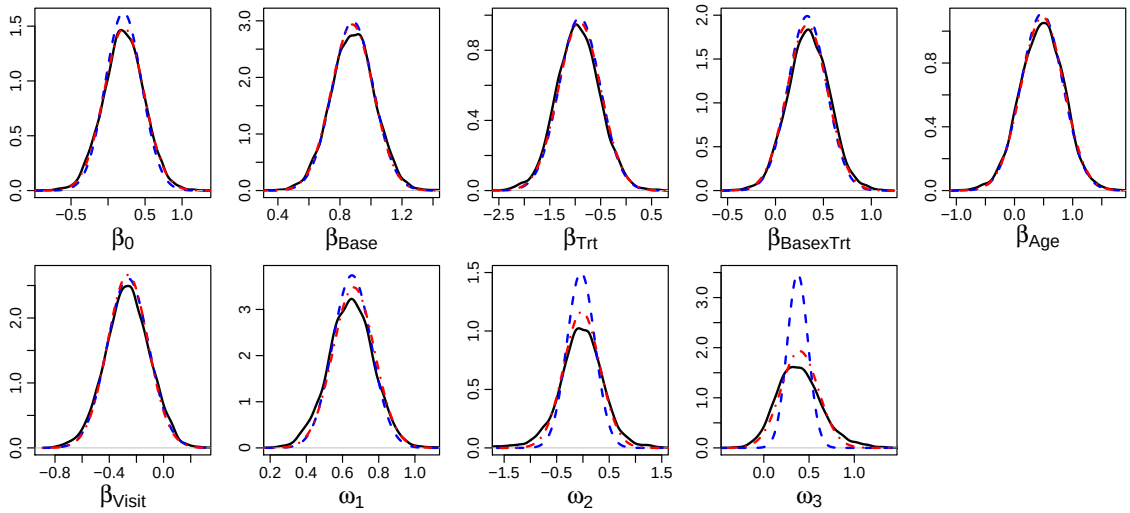

Figure 1: Epilepsy data: posterior distributions of global parameters obtained using GVA (blue dashed line), RVI (red dot-dashed line) and MCMC (black solid line).

where $i = 1, \ldots, 59$, $j = 1, \ldots, 4$. The covariates for patient i are Base_i (log of 1/4 the number of baseline seizures), Trt_i (1 for drug treatment and 0 for placebo), Age_i (log of age of patient at baseline centered at zero) and Vistit_{ij} which is coded as -0.3, -0.1, -0.1 and 0.3 for $j = 1, \ldots, 4$ respectively. The value of y_{ij} range from 0 to 102 with a mean of 8.26 and median of 4. We set $\hat{\eta}_{ij} = \log(y_{ij})$ if $y_{ij} > 0$ and $\hat{\eta}_{ij} = \log(y_{ij} + 0.1)$ if $y_{ij} = 0$. We also explored estimating $\hat{\eta}_{ij}$ using its posterior mean. The lower bound attained using each method is summarized in Table 1. RVI achieves a higher lower bound than

	GVA	RVI	RVI ($\hat{\eta}$ updated)
Lower bound	3139.0	3140.9	3141.2

Table 1: Epilepsy data: Lower bound (excluding constants independent of the variational parameters) for each approach.

GVA (representing a better variational approximation) and does slightly better when $\hat{\eta}$ is updated with its posterior mean. This result, together with our experiments on Poisson GLMMs fitted to other datasets, indicate that data with large counts are generally very informative and a good model reparametrization can often be deduced based on the count data. Thus the estimate of the linear predictor, $\hat{\eta}$, can usually be constructed from the data alone without any further updates.

We compare the posterior distributions of the global parameters obtained using GVA and RVI with MCMC in Figure 1. The posteriors of RVI ($\hat{\eta}$ updated) are very similar to that of RVI and are not shown for simplicity. While the posterior distributions of the β coefficients returned by both methods are quite similar, RVI performs better in approximating the posteriors of the precision parameters ω while GVA is over-confident.

We also observe that RVI is able to converge faster towards the (local) mode of the lower bound than GVA. Figure 2 shows the estimate of the lower bound obtained by

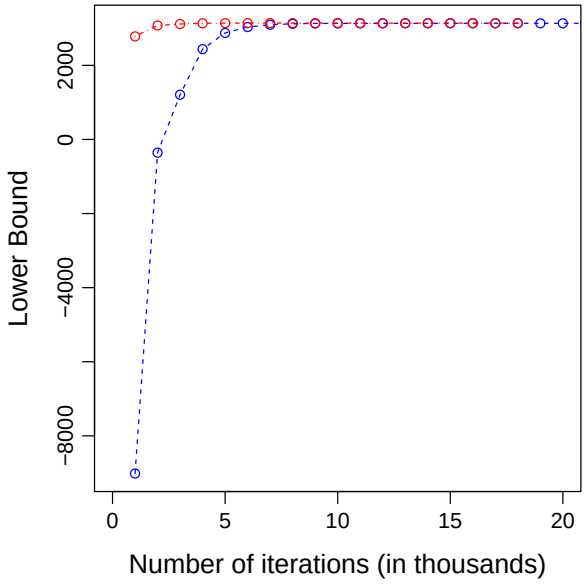

Figure 2: Epilepsy data: Lower bound attained using GVA (blue dashed line) and RVI (red dot-dashed line). Each point is the average value over the past 1000 iterations.

averaging over the past 1000 iterations. We see that RVI is able to attain a lower bound of over 2000 within the first 1000 iterations while GVA only reaches a lower bound of less than -8000. GVA manages to catch up with RVI only after about 8000 iterations. Thus the model reparametrization serves not only to reduce the posterior dependency between global and local variables but it also helps to improve convergence. This can be especially useful in high dimensional problems where the number of variational parameters is large.

8.2 Seeds data

Next, we fit a binomial GLMM to the seeds germination data (Crowder, 1978), which is available from the `R` package `hglm` using `data(seeds)`. The response is the number of seeds y_i that germinated out of m_i that were brushed on plate i for $i = 1, \ldots, 21$. This data arise from a 2×2 factorial experiment and the two factors are seed type (O. aegyptica 75 and O. aegyptica 73) and type of root extract (bean and cucumber). We consider the binomial GLMM introduced by Breslow and Clayton (1993) for handling overdispersion, where $y_i \sim \text{binomial}(m_i, p_i)$,

$$\text{logit}(p_i) = \beta_0 + \beta_{\text{seed}}\text{seed}_i + \beta_{\text{extract}}\text{extract}_i + b_i,$$

and b_i is the random effect for plate i for $i = 1, \ldots, 21$. We define $\text{seed}_i = 1$ if O. aegyptica 75 and 0 if O. aegyptica 73 and $\text{extract}_i = 1$ if bean and 0 if cucumber. The value of m_i ranges from 5 to 81 while y_i ranges from 0 to 55.

Fitting this model using GVA and RVI, we set $\hat{\eta}_i = \text{logit}(y_i/m_i)$ if $y_i > 0$ and

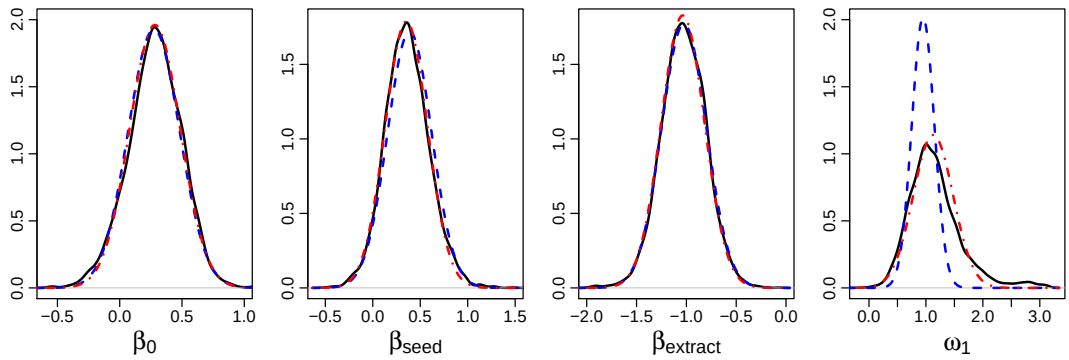

Figure 3: Seeds data: posterior distributions of global parameters obtained using GVA (blue dashed line), RVI (red dot-dashed line) and MCMC (black solid line).

$\hat{\eta}_i = \text{logit}\{(y_i + 0.1)/m_i\}$ if $y_i > 0$. The lower bounds attained are given in Table 2. RVI is able to achieve a higher lower bound than GVA again. As binomial type data is also very informative, we see that it is not necessary to update $\hat{\eta}$ and the model reparametrization deduced from the data itself is sufficient. Figure 3 shows the estimated

	GVA	RVI	RVI ($\hat{\eta}$ updated)
Lower bound	-551.0	-550.4	-550.4

Table 2: Seeds data: Lower bound (excluding constants independent of the variational parameters) for each approach.

posterior distributions. We note that the posterior distributions of the β coefficients estimated by GVA and RVI are again similar but GVA underestimates the variance of the precision parameter ω_1 quite severely. RVI also converges to the local mode of the lower bound faster than GVA and the trend is similar to that in Figure 2. Within the first 1000 iterations, RVI has attained a lower bound of -566.8 as compared to -657.1 for GVA. Thus, the model reparametrization seems to work very well for both Poisson and binomial mixed models.

8.3 HERS data

We consider a large longitudinal data set derived from the Heart and Estrogen/Progestin Study (HERS, Hulley et al., 1998) available at `www.biostat.ucsf.edu/vgsm/data.html`. The aim of the study was to determine if estrogen plus Progestin therapy reduces the risk of coronary heart disease (CHD) events for post-menopausal women with existing CHD. In this clinical trial, 2763 women were randomly assigned to a hormone group or a placebo group, and they were followed up for the next five years with an annual clinic visit where data are collected on their health status. Some patients did not turn up for all 5 subsequent visits and data for certain covariates are also missing. Here we only use the data for 2031 women where data on all covariates concerned are available. We

consider the response $y_{ij} \sim \text{Bernoulli}(p_{ij})$ as the binary indicator of whether the systolic blood pressure of patient i is higher than 140 at the jth visit, and the logit mixed model

$$\text{logit}(p_{ij}) = \beta_0 + \beta_{\text{visit}}\text{visit}_{ij} + \beta_{\text{BMI}}\text{BMI}_{ij} + \beta_{\text{HTN}}\text{HTN}_{ij} + \beta_{\text{age}}\text{age}_i + b_{i1} + b_{i2}\text{visit}_{ij}.$$

for $i = 1, \ldots, 2031$ and $0 \leq j \leq 5$. For patient i, visit_{ij} is coded as -1, -0.6, -0.2, 0.2, 0.6, 1 for $j = 0, 1, \ldots, 5$ respectively, BMI_{ij} is the body mass index at the jth visit, HTN_{ij} is a binary indicator for whether the patient is taking high blood pressure medication at the jth visit and age_i is the age of patient at baseline. The covariates BMI and age were normalized before fitting the model. As the responses are binary, we first obtain an estimate of $\hat{\eta}$ by using the `glm` function in `Julia` to fit a GLM to the dataset.

Table 3 shows the lower bound obtained by using GVA and RVI to fit the mixed logit model. The lower bound attained by RVI is higher than that of GVA and and a further improvement is obtained if $\hat{\eta}$ is updated. In contrast to the Poisson and binomial data, binary data is less informative and it is more difficult to obtain a good estimate of $\hat{\eta}$. Hence it may be useful to update $\hat{\eta}$ using its posterior mean.

	GVA	RVI	RVI ($\hat{\eta}$ updated)
Lower bound	-5029.9	-5027.9	-5026.5

Table 3: HERS data: Lower bound (excluding constants independent of the variational parameters) for each approach.

Besides using the GVA and RVI algorithms to fit the Bernoulli mixed model, we also investigate the performance of parallel processing using RVI. We partition the subjects randomly into three parts each with 677 subjects and run the three partial datasets in parallel using RVI before using the techniques in Section 6 to combine the results together. Figure 4 shows the resulting estimates of the posterior distributions. We note that the posteriors of the regression coefficients are reasonably well estimated by GVA and RVI but there is overestimation of the precision parameters ω_1 and ω_3 in both methods. RVI does better than GVA in estimating the posterior variance of the precision parameters. The posteriors obtained using the divide and recombine strategy are also very similar to that obtained using batch processing.

9 Conclusion

In this article, we have proposed an approach for improving inference from variational Bayes through model reparametrization. A class of models with both global and local variables is considered and the local variables are transformed via an affine transformation so as to minimize their posterior dependency on the global variables. The resulting Gaussian variational approximation, which is obtained using stochastic gradient ascent

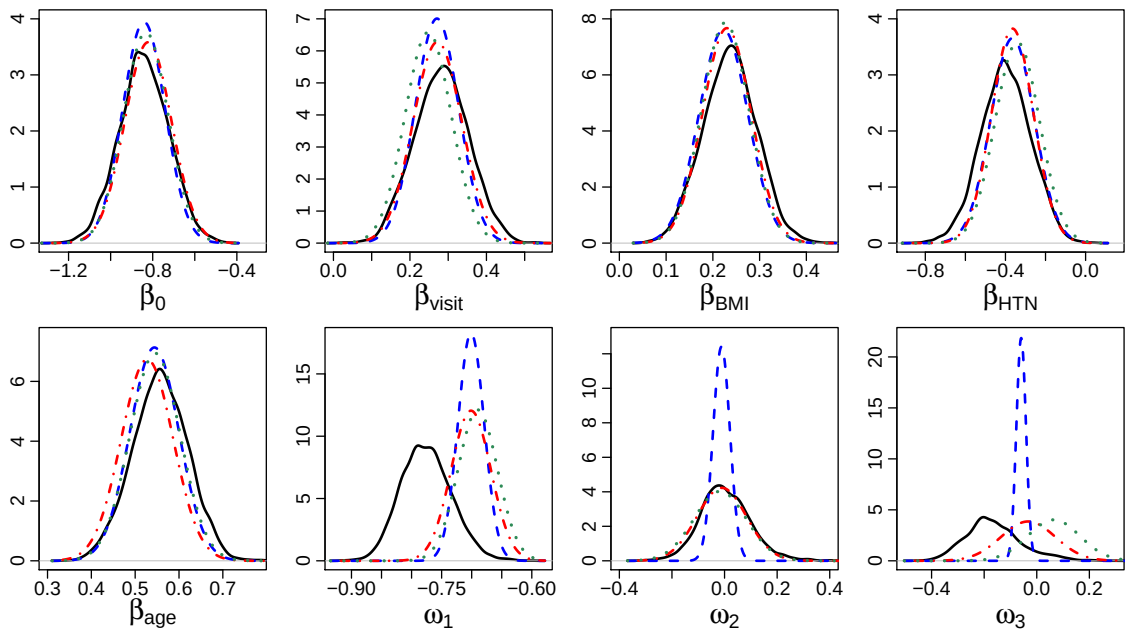

Figure 4: HERS data: posterior distributions of global parameters obtained using GVA (blue dashed line), RVI ($\hat{\eta}$ updated) (red dot-dashed line), RVI ($\hat{\eta}$ updated, parallel) (green dotted line) and MCMC (black solid line).

methods, is low-dimensional and the approach can be readily extended to large datasets by using a "divide and recombine" method and parallel processing. In the application to GLMMs, we find that the method works very well especially for the Poisson and binomial mixed models where the data is usually more informative and it is easy to obtain a good reparametrization based on the raw data alone. For Bernoulli mixed models, the data is not as informative and it is more difficult to deduce a "good" reparametrization. However, we find that as compared to GVA, RVI is still often able to yield better estimates of the posterior variance of the precision parameters. The results obtained from GLMMs is promising and it will be interesting to investigate the performance of RVI for other complex models.

Acknowledgments

Linda Tan was supported by the start-up grant R-155-000-190-133.

References

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

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

Broderick, T., Boyd, N., Wibisono, A., Wilson, A. C., and Jordan, M. I. (2013). Streaming variational bayes. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 1727–1735, USA. Curran Associates Inc.

Browne, W. J., Steele, F., Golalizadeh, M., and Green, M. J. (2009). The use of simple reparameterizations to improve the efficiency of markov chain monte carlo estimation for multilevel models with applications to discrete time survival models. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 172(3):579–598.

Christensen, O. F., Roberts, G. O., and Skld, M. (2006). Robust markov chain monte carlo methods for spatial generalized linear mixed models. *Journal of Computational and Graphical Statistics*, 15(1):1–17.

Crowder, M. J. (1978). Beta-binomial anova for proportions. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 27(1):34–37.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist.Soc. B*, 39:1–38.

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

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1996). Efficient parametrisations for generalized linear mixed models. In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics 5*.

Gershman, S., Hoffman, M., and Blei, D. (2012). Nonparametric variational inference. In Langford, J. and Pineau, J., editors, *Proceedings of the 29th International Conference on Machine Learning*, pages 663–670.

Hoffman, M. and Blei, D. (2015). Stochastic structured variational inference. In Lebanon, G. and Vishwanathan, S., editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38, pages 361–369. JMLR Workshop and Conference Proceedings.

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

Hulley, S., Grady, D., Bush, T., and et al. (1998). Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. *JAMA*, 280(7):605–613.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Comput.*, 3(1):79–87.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2016). Automatic differentiation variational inference. arXiv: 1603.00788.

Magnus, J. R. and Neudecker, H. (1980). The elimination matrix: Some lemmas and applications. *SIAM Journal on Algebraic Discrete Methods*, 1(4):422–449.

Meng, X.-L. and van Dyk, D. (1997). The EM algorithm–an old folk-song sung to a fast new tune. *J. R. Statist.Soc. B*, 59:511–567.

Meng, X.-L. and van Dyk, D. (1998). Fast EM-type implementations for mixed effects models. *J. R. Statist.Soc. B*, 60:559–578.

Murray, I. (2016). Differentiation of the Cholesky decomposition. arXiv:1602.07527.

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

Paisley, J. W., Blei, D. M., and Jordan, M. I. (2012). Variational Bayesian inference with stochastic search. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*.

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

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

Rezende, D. J. and Mohamed, S. (2015). Variational inference with normalizing flows. In Bach, F. and Blei, D., editors, *Proceedings of The 32nd International Conference on Machine Learning*, pages 1530–1538. JMLR Workshop and Conference Proceedings.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In Xing, E. P. and Jebara, T., editors, *Proceedings of The 31st International Conference on Machine Learning*, pages 1278–1286. JMLR Workshop and Conference Proceedings.

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

Salimans, T. and Knowles, D. A. (2013). Fixed-form variational posterior approximation through stochastic linear regression. *Bayesian Analysis*, 8:837–882.

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

Tan, L. S. L. and Nott, D. J. (2013). Variational inference for generalized linear mixed models using partially non-centered parametrizations. *Statistical Science*, 28:168–188.

Tan, L. S. L. and Nott, D. J. (2014). A stochastic variational framework for fitting and diagnosing generalized linear mixed models. *Bayesian Analysis*, 9:963–1004.

Tan, L. S. L. and Nott, D. J. (2018). Gaussian variational approximation with sparse precision matrices. *Statistics and Computing*, 28:259–275.

Tan, S. L. L. (2017). Efficient data augmentation techniques for gaussian state space models. arXiv:1712.08887.

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

Titsias, M. and Lázaro-Gredilla, M. (2014). Doubly stochastic variational Bayes for non-conjugate inference. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1971–1979.

Titsias, M. K. (2017). Learning model reparametrizations: Implicit variational inference by

fitting mcmc distributions. arXiv:1708.01529.

Tran, M.-N., Nott, D. J., Kuk, A. Y. C., and Kohn, R. (2016). Parallel variational bayes for large datasets with an application to generalized linear mixed models. *Journal of Computational and Graphical Statistics*, 25(2):626–646.

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

Yu, Y. and Meng, X.-L. (2011). To center or not to center: That is not the question—An ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC efficiency. *Journal of Computational and Graphical Statistics*, 20:531–570.

Zeiler, M. D. (2012). Adadelta: An adaptive learning rate method. arXiv: 1212.5701.

A Unbiased estimates of stochastic gradients

As $\tilde{\theta} = Cs + \mu$, differentiating $\ell(\tilde{\theta})$ with respect to μ and $v(C)$ separately,

$$\mathrm{d}\{\ell(\tilde{\theta})\} = \nabla_{\tilde{\theta}}\ell(\tilde{\theta})^T\mathrm{d}\mu, \qquad \begin{aligned}\mathrm{d}\{\ell(\tilde{\theta})\} &= \nabla_{\tilde{\theta}}\ell(\tilde{\theta})^T(\mathrm{d}C)s \\ &= (s^T \otimes \nabla_{\tilde{\theta}}\ell(\tilde{\theta})^T)\mathrm{dvec}(C) \\ &= \mathrm{vec}(\nabla_{\tilde{\theta}}\ell(\tilde{\theta})s^T)^T E_d^T \mathrm{d}v(C) \\ &= v(\nabla_{\tilde{\theta}}\ell(\tilde{\theta})s^T)^T\mathrm{d}v(C).\end{aligned}$$

Therefore $\nabla_{\mu}\ell(\tilde{\theta}) = \nabla_{\tilde{\theta}}\ell(\tilde{\theta})$ and $\nabla_{v(C)}\ell(\tilde{\theta}) = v(\nabla_{\tilde{\theta}}\ell(\tilde{\theta})s^T)$. In addition,

$$\mathrm{d}\log|C| = \mathrm{tr}(C^{-1}\mathrm{d}C) = \mathrm{vec}(C^{-T})^T E_d^T \mathrm{d}v(C) = v(C^{-T})^T\mathrm{d}v(C).$$

Hence $\nabla_{v(C)}\log|C| = v(C^{-T})$. For the second estimator, differentiating $\log q(\tilde{\theta}|\mu, C)$ with respect to $\tilde{\theta}$,

$$\mathrm{d}\log q(\tilde{\theta}|\mu, C) = -(\tilde{\theta}-\mu)^T C^{-T}C^{-1}\mathrm{d}\tilde{\theta} = -s^T C^{-1}\mathrm{d}\tilde{\theta},$$

Hence $\nabla_{\tilde{\theta}}\log q(\tilde{\theta}|\mu, C) = -C^{-T}s$. We also have

$$\begin{aligned}-\mathrm{d}(\tilde{\theta}-\mu)^T C^{-T}C^{-1}(\tilde{\theta}-\mu) &= s^T\{(\mathrm{d}C^T)C^{-T} + C^{-1}(\mathrm{d}C)\}s \\ &= \{(s^T C^{-1}\otimes s^T)K_d + (s^T \otimes s^T C^{-1})\}\mathrm{dvec}(C) \\ &= \{\mathrm{vec}(ss^T C^{-1})^T K_d + \mathrm{vec}(C^{-T}ss^T)^T\}\mathrm{dvec}(C) \\ &= 2\mathrm{vec}(C^{-T}ss^T)^T E_d^T \mathrm{d}v(C) \\ &= 2v(C^{-T}ss^T)^T\mathrm{d}v(C)\end{aligned}$$

Hence $\nabla_{v(C)}\{\log q(\tilde{\theta}|\mu, C)\} = v(C^{-T}ss^T - C^{-T})$.