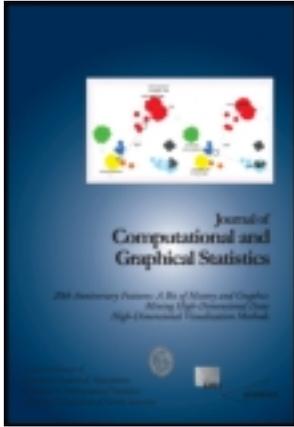


This article was downloaded by: [138.25.250.243]

On: 21 May 2012, At: 18:57

Publisher: Taylor & Francis

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



## Journal of Computational and Graphical Statistics

Publication details, including instructions for authors and subscription information:

<http://amstat.tandfonline.com/loi/ucgs20>

### Gaussian Variational Approximate Inference for Generalized Linear Mixed Models

J. T. Ormerod<sup>a</sup> & M. P. Wand<sup>b</sup>

<sup>a</sup> School of Mathematics and Statistics, University of Sydney, Sydney, 2006, Australia

<sup>b</sup> School of Mathematical Sciences, University of Technology, Sydney, Broadway, 2007, Australia

Available online: 10 Apr 2012

To cite this article: J. T. Ormerod & M. P. Wand (2012): Gaussian Variational Approximate Inference for Generalized Linear Mixed Models, Journal of Computational and Graphical Statistics, 21:1, 2-17

To link to this article: <http://dx.doi.org/10.1198/jcgs.2011.09118>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://amstat.tandfonline.com/page/terms-and-conditions>

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

# Gaussian Variational Approximate Inference for Generalized Linear Mixed Models

J. T. ORMEROD and M. P. WAND

Variational approximation methods have become a mainstay of contemporary machine learning methodology, but currently have little presence in statistics. We devise an effective variational approximation strategy for fitting generalized linear mixed models (GLMMs) appropriate for grouped data. It involves Gaussian approximation to the distributions of random effects vectors, conditional on the responses. We show that *Gaussian variational approximation* is a relatively simple and natural alternative to Laplace approximation for fast, non-Monte Carlo, GLMM analysis. Numerical studies show Gaussian variational approximation to be very accurate in grouped data GLMM contexts. Finally, we point to some recent theory on consistency of Gaussian variational approximation in this context. Supplemental materials are available online.

**Key Words:** Best prediction; Likelihood-based inference; Longitudinal data analysis; Machine learning; Variance components.

## 1. INTRODUCTION

Statistical and probabilistic models continue to grow in complexity in response to the demands of modern applications. Fitting and inference for such models is an ongoing issue and new sectors of research have emerged to meet this challenge. In statistics, the most prominent of these is Markov chain Monte Carlo (MCMC), which is continually being tailored to handle difficult inferential questions arising in, for example, Bayesian models (e.g., Gelman et al. 2004; Marin and Robert 2007; Carlin and Louis 2008), mixed and latent variable models (e.g., Skrondal and Rabe-Hesketh 2004; McCulloch, Searle, and Neuhaus 2008), and missing data models (e.g., Little and Rubin 2004). The main difficulty addressed by MCMC is the presence of intractable multivariate integrals in likelihood and posterior density expressions.

In parallel to these developments in statistics, the machine learning community has been developing approximate solutions to inferential problems using the notion of variational

---

J. T. Ormerod is Lecturer, School of Mathematics and Statistics, University of Sydney, Sydney 2006, Australia (E-mail: [jormerod@sydney.edu.au](mailto:jormerod@sydney.edu.au)). M. P. Wand is Distinguished Professor, School of Mathematical Sciences, University of Technology, Sydney, Broadway 2007, Australia.

© 2012 American Statistical Association, Institute of Mathematical Statistics,  
and Interface Foundation of North America

*Journal of Computational and Graphical Statistics*, Volume 21, Number 1, Pages 2–17  
DOI: 10.1198/jcgs.2011.09118

bounds. These *variational approximations* sacrifice some of the accuracy of MCMC by solving perturbed versions of the problems at hand, but offer vast improvements in terms of computational speed. Motivating settings include probabilistic graphical models, hidden Markov models, and phylogenetic trees. Summaries of recent variational approximation research may be found in the works of Jordan et al. (1999), Jordan (2004), and Bishop (2006). An introduction to variational approximation from a statistical perspective was provided by the article by Ormerod and Wand (2010).

In this article we help bring variational approximation into mainstream statistics by tailoring it to the most popular current setting for which integration difficulties arise: generalized linear mixed models (GLMMs). In the interest of conciseness, we focus on the most common type of GLMM—that arising in the analysis of grouped data with Gaussian random effects. General design GLMMs, as described by Zhao et al. (2006), are not treated here.

We identify a particular type of variational approximation that is well-suited to grouped data GLMMs. It involves approximation of the distributions of random effects vectors, given the responses, by Gaussian distributions. The resulting *Gaussian variational approximation (GVA)* approach emerges as a new alternative to Laplace approximation for fast, deterministic fitting of grouped data GLMMs. Conceptually, the approach is very simple: its derivation requires application of Jensen's inequality to the log-likelihood to obtain a variational lower bound. Maximization is then carried out over the original parameters and the introduced *variational* parameters. GVA involves a little more algebra and calculus to implement compared with some of the simpler versions of Laplace approximation such as penalized quasi-likelihood (PQL) (Breslow and Clayton 1993). However, with the aid of the formulas presented in Appendix A of the supplemental material, effective computation can be achieved in order- $m$  operations, where  $m$  is the number of groups. For some GLMMs, such as Poisson GLMMs, the GVA completely eradicates the need for integration. In others, such as logistic GLMMs, only *univariate* numerical integration is required on well-behaved integrands.

Standard errors for fixed effect and covariance parameter estimates are a by-product of the fitting algorithm. Approximate best predictions for the random effects, along with prediction variances, also arise quite simply from the Gaussian approximation. Moreover, numerical studies show GVA to be very accurate, often almost as good as MCMC and a significant improvement on PQL. Other varieties of Laplace approximation (e.g., Lee and Nelder 1996; Raudenbush, Yang, and Yosef 2000; Rue, Martino, and Chopin 2009) also offer accuracy improvements over PQL but, like GVA, have their own costs in terms of implementability.

Recently, Hall, Ormerod, and Wand (2011) investigated the theoretical properties of GVA for a simple special case of the grouped data GLMMs considered here. They established root- $m$  consistency of Gaussian variational approximate maximum likelihood estimators under relatively mild assumptions.

A significant portion of variational approximation methodology is based on the notion of factorized density approximations to key conditional densities with respect to Kullback–Leibler distance (e.g., Bishop 2006, sec. 10.1). This general strategy is sometimes called

*mean field* approximation, and has its roots in 1980s statistical physics research (Parisi 1988). However, mean field approximation is not well-suited to GLMMs since they lack the conjugacy that normally gives rise to explicit updating formulas. In addition, mean field approximation has a tendency to markedly underestimate the variability of parameter estimates (Wang and Titterton 2005; Rue, Martino, and Chopin 2009).

Another variant of variational approximations is the tangent transform approach of Jaakkola and Jordan (2000). It may be applied to logistic GLMMs (Ormerod 2008) but does not extend to other situations such as Poisson response models. We have also encountered variance underestimation problems with the Jaakkola and Jordan variational approximation (Ormerod and Wand 2008).

The use of Gaussian densities in variational approximation has a small and recent literature in machine learning (Barber and Bishop 1998; Seeger 2000; Honkela and Valpola 2005; Archambeau et al. 2007; Opper and Archambeau 2009). None of this research is directly connected with GLMMs, although the work of Opper and Archambeau (2009) is the most closely related work to ours.

Section 2 describes the type of data and the types GLMMs that we consider. Section 3 explains the use of Gaussian variational approximation for avoiding multivariate intractable integrals when making inference about model parameters and describes connections with Kullback–Leibler divergence theory. Section 3 also deals with the approximation of standard errors and prediction of random effects and explains how Gaussian variational approximation provides an attractive solution to these problems. Theoretical properties of Gaussian variational approximations for grouped data GLMMs are the subject of Section 4 including conditions where the Laplace approximation and GVA may be quite different. Examples are given in Section 5 and concluding remarks are made in Section 6. Appendixes in the supplemental material deal with various details which arise.

## 2. DATA AND MODEL

We consider regression-type data collected repeatedly within  $m$  groups. Examples of *groups* in applications are *humans* in a medical study, *counties* in a sample survey, and *animals* in a breeding experiment. The  $j$ th predictor/response pair for the  $i$ th group is denoted by  $(\mathbf{x}_{ij}, y_{ij})$ ,  $1 \leq j \leq n_i$ ,  $1 \leq i \leq m$ . Here the entries of the predictor vectors  $\mathbf{x}_{ij}$  are unrestricted, while the  $y_{ij}$  are subject to restrictions such as being binary or nonnegative integers. A specific example is provided by Figure 1. The response data are disease indicators from a clinical trial involving longitudinal checks on the participants.

For each  $1 \leq i \leq m$  define the  $n_i \times 1$  vectors  $\mathbf{y}_i = [y_{i1}, \dots, y_{in_i}]^T$  and  $\mathbf{1}_i \equiv [1, \dots, 1]^T$ . The first of these is the vector of responses for the  $i$ th group. It is reasonable to assume that the vectors  $\mathbf{y}_1, \dots, \mathbf{y}_m$  are independent of each other. However, within-group measurements may be dependent and we use random effects to model this dependence. Specifically, we consider one-parameter exponential family models of the form

$$\begin{aligned} \mathbf{y}_i | \mathbf{u}_i &\stackrel{\text{ind.}}{\sim} \exp\{\mathbf{y}_i^T (\mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{u}_i) - \mathbf{1}_i^T b(\mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{u}_i) + \mathbf{1}_i^T c(\mathbf{y}_i)\}, \\ \mathbf{u}_i &\stackrel{\text{ind.}}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma}), \end{aligned} \tag{2.1}$$

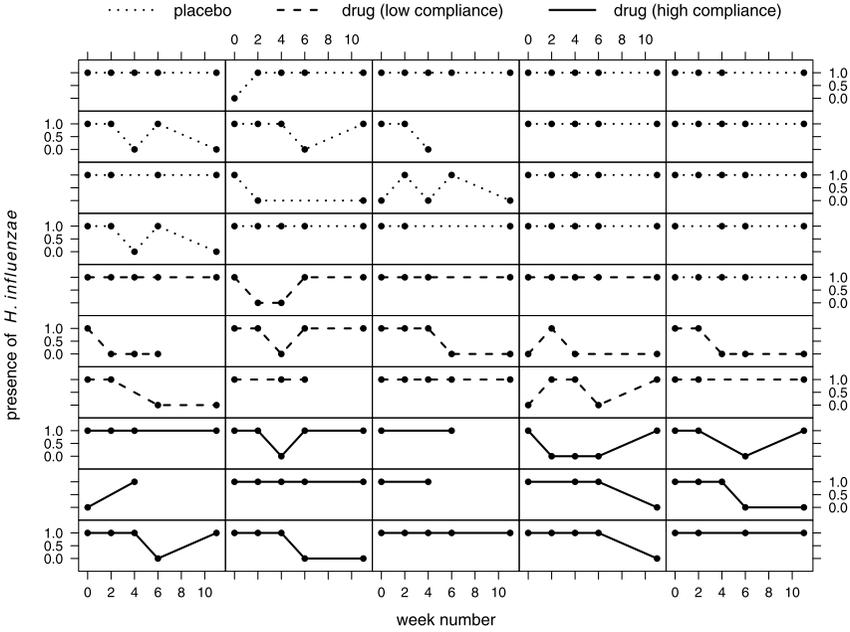


Figure 1. Example of binary response grouped regression data. Each panel corresponds to longitudinal measurements on a child with a history of otitis media who participated in a clinical trial in Northern Territory, Australia. The horizontal axis ( $x$ ) is week number since the start of the trial. The vertical axis ( $y$ ) is presence ( $y = 1$ ) or absence ( $y = 0$ ) of *H. influenzae* (source: Leach 2000).

where  $\mathbf{u}_i$ ,  $1 \leq i \leq m$ , are  $K \times 1$  random effects vectors and  $\Sigma$  ( $K \times K$ ) is their common covariance matrix. The functions  $b$  and  $c$  are specific to members of the family. The most common examples are Bernoulli for which  $b(x) = \log(1 + e^x)$  and  $c(x) = 0$  and Poisson for which  $b(x) = e^x$  and  $c(x) = -\log(x!)$ . Note that operations of  $b$  and  $c$  on a vector are applied element-wise. For example,  $b\left(\begin{bmatrix} 3 \\ 8 \end{bmatrix}\right) = \begin{bmatrix} b(3) \\ b(8) \end{bmatrix}$ . The matrices  $\mathbf{X}_i$  and  $\mathbf{Z}_i$  are design matrices dependent on the  $\mathbf{x}_{ij}$ , and are assumed to be fixed. Digestion of the design structure is aided by the following examples.

**Example 1:** Suppose  $(\mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{u}_i)_j = \beta_0 + u_i + \beta_1 x_{ij}$  where  $u_i \sim N(0, \sigma_u^2)$ . In this case  $\boldsymbol{\beta} = [\beta_0, \beta_1]^T$ ,  $\mathbf{u}_i = u_i$ ,  $K = 1$ ,  $\Sigma = \sigma^2$ ,  $\mathbf{X}_i = [1, x_{ij}]_{1 \leq j \leq n_i}$ , and  $\mathbf{Z}_i = [1]_{1 \leq j \leq n_i}$  for  $1 \leq i \leq m$ .

**Example 2:** Suppose  $(\mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{u}_i)_j = \beta_0 + u_{0i} + (\beta_1 + u_{1i})x_{1ij} + \beta_2 x_{2ij}$  where

$$\begin{bmatrix} u_{0i} \\ u_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \rho\sigma_0\sigma_1 \\ \rho\sigma_0\sigma_1 & \sigma_1^2 \end{bmatrix}\right). \tag{2.2}$$

In this case  $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2]^T$ ,  $\mathbf{u}_i = [u_{0i}, u_{1i}]^T$ ,  $K = 2$ ,  $\Sigma = \begin{bmatrix} \sigma_0^2 & \rho\sigma_0\sigma_1 \\ \rho\sigma_0\sigma_1 & \sigma_1^2 \end{bmatrix}$ ,  $\mathbf{X}_i = [1, x_{1ij}, x_{2ij}]_{1 \leq j \leq n_i}$ , and  $\mathbf{Z}_i = [1, x_{1ij}]_{1 \leq j \leq n_i}$  for  $1 \leq i \leq m$ .

Model (2.1) is a generalized linear mixed model (GLMM) suited to grouped data. In many applications of interest, the data are collected longitudinally in which case (2.1)

might be called a *longitudinal data* GLMM. But to cater for other areas of application, such as complex sample surveys, we will simply call (2.1) a *grouped data* GLMM.

The class of GLMMs is much more general than (2.1), as explained by Zhao et al. (2006). Staying within the one-parameter exponential family, a more general class of models is

$$\mathbf{y}|\mathbf{u} \sim \exp\{\mathbf{y}^T(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) - \mathbf{1}^T b(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) + \mathbf{1}^T c(\mathbf{y})\}, \quad \mathbf{u} \sim N(\mathbf{0}, \mathbf{G}), \quad (2.3)$$

where the design matrices  $\mathbf{X}$  and  $\mathbf{Z}$  and covariance matrix  $\mathbf{G}$  are quite general. In the special case of (2.1) we have  $\mathbf{y} = [\mathbf{y}_1^T, \dots, \mathbf{y}_m^T]$ ,  $\mathbf{X} = [\mathbf{X}_1^T, \dots, \mathbf{X}_m^T]^T$ ,  $\mathbf{u} = [\mathbf{u}_1^T, \dots, \mathbf{u}_m^T]$ ,  $\mathbf{G} = \mathbf{I}_m \otimes \boldsymbol{\Sigma}$ , and  $\mathbf{Z} = \text{blockdiag}_{1 \leq i \leq m}(\mathbf{Z}_i)$ . We return to general design GLMMs in Section 4 since connections with variational approximation theory are better elucidated at that level of generality.

### 3. GAUSSIAN VARIATIONAL APPROXIMATE INFERENCE

The parameters in model (2.1) are the fixed effects vector  $\boldsymbol{\beta}$  and the random effects covariance matrix  $\boldsymbol{\Sigma}$ . Their log-likelihood is

$$\begin{aligned} \ell(\boldsymbol{\beta}, \boldsymbol{\Sigma}) &= \sum_{i=1}^m \{\mathbf{y}_i^T \mathbf{X}_i \boldsymbol{\beta} + \mathbf{1}_i^T c(\mathbf{y}_i)\} - \frac{m}{2} \log |\boldsymbol{\Sigma}| - \frac{mK}{2} \log(2\pi) \\ &\quad + \sum_{i=1}^m \log \int_{\mathbb{R}^K} \exp\left\{\mathbf{y}_i^T \mathbf{Z}_i \mathbf{u} - \mathbf{1}_i^T b(\mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{u}) - \frac{1}{2} \mathbf{u}^T \boldsymbol{\Sigma}^{-1} \mathbf{u}\right\} d\mathbf{u} \end{aligned}$$

and the maximum likelihood estimators of  $\boldsymbol{\beta}$  and  $\boldsymbol{\Sigma}$  are  $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}}) = \text{argmax}_{\boldsymbol{\beta}, \boldsymbol{\Sigma}} \ell(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ . Evaluation of the maximum likelihood estimators and associated standard error calculations are hindered by the fact that the  $K$ -dimensional integral in  $\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma})$  cannot be solved analytically.

We can get around this by introducing a pair of *variational* parameters  $(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)$  for each  $1 \leq i \leq m$ , where the  $\boldsymbol{\mu}_i$  are  $K \times 1$  vectors and the  $\boldsymbol{\Lambda}_i$  are  $K \times K$  positive definite matrices. By Jensen's inequality and concavity of the logarithm function:

$$\begin{aligned} \ell(\boldsymbol{\beta}, \boldsymbol{\Sigma}) &= \sum_{i=1}^m \{\mathbf{y}_i^T \mathbf{X}_i \boldsymbol{\beta} + \mathbf{1}_i^T c(\mathbf{y}_i)\} - \frac{m}{2} \log |\boldsymbol{\Sigma}| - \frac{mK}{2} \log(2\pi) \\ &\quad + \sum_{i=1}^m \log \int_{\mathbb{R}^K} \exp\left\{\mathbf{y}_i^T \mathbf{Z}_i \mathbf{u} - \mathbf{1}_i^T b(\mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{u}) - \frac{1}{2} \mathbf{u}^T \boldsymbol{\Sigma}^{-1} \mathbf{u}\right\} \frac{\phi_{\boldsymbol{\Lambda}_i}(\mathbf{u} - \boldsymbol{\mu}_i)}{\phi_{\boldsymbol{\Lambda}_i}(\mathbf{u} - \boldsymbol{\mu}_i)} d\mathbf{u} \\ &\geq \sum_{i=1}^m \{\mathbf{y}_i^T \mathbf{X}_i \boldsymbol{\beta} + \mathbf{1}_i^T c(\mathbf{y}_i)\} - \frac{m}{2} \log |\boldsymbol{\Sigma}| - \frac{mK}{2} \log(2\pi) \\ &\quad + \sum_{i=1}^m E_{\mathbf{u} \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)} \left( \mathbf{y}_i^T \mathbf{Z}_i \mathbf{u} - \mathbf{1}_i^T b(\mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{u}) \right. \\ &\quad \left. - \frac{1}{2} \mathbf{u}^T \boldsymbol{\Sigma}^{-1} \mathbf{u} - \log(\phi_{\boldsymbol{\Lambda}_i}(\mathbf{u} - \boldsymbol{\mu}_i)) \right) \\ &\equiv \underline{\ell}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Lambda}), \end{aligned}$$

where  $(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \equiv (\boldsymbol{\mu}_1, \boldsymbol{\Lambda}_1, \dots, \boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m)$  and  $\phi_{\boldsymbol{\Lambda}_i}(\mathbf{u})$  denotes the multivariate Gaussian density function with covariance  $\boldsymbol{\Lambda}_i$ . The *variational lower bound* on the log-likelihood simplifies to

$$\begin{aligned} \underline{\ell}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) &= \sum_{i=1}^m \left[ \mathbf{y}_i^T (\mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \boldsymbol{\mu}_i) - \mathbf{1}_i^T B(\mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \boldsymbol{\mu}_i, \text{dg}(\mathbf{Z}_i \boldsymbol{\Lambda}_i \mathbf{Z}_i^T)) + \mathbf{1}_i^T c(\mathbf{y}_i) \right. \\ &\quad \left. + \frac{1}{2} \{ \log |\boldsymbol{\Sigma}^{-1} \boldsymbol{\Lambda}_i| - \boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i - \text{tr}(\boldsymbol{\Sigma}^{-1} \boldsymbol{\Lambda}_i) \} + \frac{K}{2} \right], \end{aligned} \quad (3.1)$$

where  $B(\mu, \sigma^2) \equiv \int_{-\infty}^{\infty} b(\sigma x + \mu) \phi(x) dx$ ,  $\phi$  is the  $N(0, 1)$  density function and, for a square matrix  $\mathbf{A}$ ,  $\text{dg}(\mathbf{A})$  is the column vector containing the diagonal entries of  $\mathbf{A}$ . Evaluations of  $B$  for vector arguments are applied in an element-wise fashion, for example,  $B\left(\begin{bmatrix} 3 \\ 5 \end{bmatrix}, \begin{bmatrix} 6 \\ 7 \end{bmatrix}\right) = \begin{bmatrix} B(3,6) \\ B(5,7) \end{bmatrix}$ .

The advantage of  $\underline{\ell}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$  over  $\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma})$  is that the former no longer involves integrals of size  $K$ . For Poisson mixed models all integrals in the lower-bound expression disappear since  $B(\mu, \sigma^2) = \exp(\mu + \frac{1}{2}\sigma^2)$ . In the Bernoulli case  $B(\mu, \sigma^2) = \int_{-\infty}^{\infty} \log\{1 + \exp(\mu + \sigma x)\} \phi(x) dx$ , which does not have an analytic solution. However, adaptive Gauss–Hermite quadrature (Liu and Pearce 1994) is well-suited to efficient and very accurate evaluation of  $B(\mu, \sigma^2)$  in this case. The details are given in the supplemental material.

It is also worth noting the number of quadrature points needed to approximate  $\ell$  and  $\underline{\ell}$  using adaptive Gauss–Hermite quadrature (AGHQ) (noting quadrature is only required for GVA in the logistic case). Suppose that we approximate  $\ell$  using AGHQ with tensor product of  $N$  quadrature points over each of the  $K$  dimensions. Then the calculation of  $\ell$ ,  $\underline{\ell}$  and their derivatives requires  $O(mN^K)$  and  $O(N \sum_{i=1}^m n_i)$  quadrature points, respectively. Since the number of quadrature points for GVA is independent of  $K$ , the relative computational efficiency of GVA over AGHQ can be substantial when  $K$  is large.

Given the lower-bound result,  $\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma}) \geq \underline{\ell}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$  for all  $(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ , it is clear that maximizing over the variational parameters  $(\boldsymbol{\mu}, \boldsymbol{\Lambda})$  narrows the gap between  $\underline{\ell}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$  and  $\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ . Let

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Lambda}}) = \underset{\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Lambda}}{\text{argmax}} \underline{\ell}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Lambda}). \quad (3.2)$$

Then the *Gaussian variational approximate* maximum likelihood estimators for  $\boldsymbol{\beta}$  and  $\boldsymbol{\Sigma}$  are  $\hat{\boldsymbol{\beta}}$  and  $\hat{\boldsymbol{\Sigma}}$ , respectively. Appendix A in the supplemental material provides an efficient computational formula for solving this maximization problem. Alternatively an approximate expectation maximization approach could also be considered (Neal and Hinton 1999). We note that for other response types and noncanonical links  $\underline{\ell}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$  can be calculated with little modification (Ormerod 2008).

### 3.1 RELATIONSHIP WITH KULLBACK–LEIBLER DIVERGENCE

The lower-bound expression (3.1) can also be derived using the ideas of Kullback–Leibler divergence, which underpins much of variational approximation methodology (e.g., Titterton 2004; Bishop 2006, chap. 10). In this section, we first work with the

general form of the GLMM given by (2.3). We also use  $p$  to denote density or probability mass functions of random vectors according to the model. For example,  $p(\mathbf{u}|\mathbf{y})$  is the conditional density function of  $\mathbf{u}$  given  $\mathbf{y}$ . Furthermore, the log-likelihood can be written in terms of the joint density of  $\mathbf{y}$ :  $\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \log p(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \log p(\mathbf{y})$ . For an arbitrary density functions  $q$  on  $\mathbb{R}^M$ ,

$$\begin{aligned}\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma}) &= \int_{\mathbb{R}^M} \log p(\mathbf{y})q(\mathbf{u}) d\mathbf{u} = \int_{\mathbb{R}^M} \log \left\{ \frac{p(\mathbf{y}, \mathbf{u})/q(\mathbf{u})}{p(\mathbf{u}|\mathbf{y})/q(\mathbf{u})} \right\} q(\mathbf{u}) d\mathbf{u} \\ &= \int_{\mathbb{R}^M} q(\mathbf{u}) \log \left\{ \frac{p(\mathbf{y}, \mathbf{u})}{q(\mathbf{u})} \right\} d\mathbf{u} + \int_{\mathbb{R}^M} q(\mathbf{u}) \log \left\{ \frac{q(\mathbf{u})}{p(\mathbf{u}|\mathbf{y})} \right\} d\mathbf{u},\end{aligned}$$

where  $M$  is the dimension of the  $\mathbf{u}$  vector. The last term is the Kullback–Leibler distance between  $q(\mathbf{u})$  and  $p(\mathbf{u}|\mathbf{y})$ . Since this is always nonnegative (Kullback and Leibler 1951) we get

$$\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma}) \geq \int_{\mathbb{R}^M} q(\mathbf{u}) \log \left\{ \frac{p(\mathbf{y}, \mathbf{u})}{q(\mathbf{u})} \right\} d\mathbf{u}. \quad (3.3)$$

Substitution of  $q(\mathbf{u}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Lambda})$  into (3.3) gives a closed form lower bound on the log-likelihood.

**Theorem 1.** *Consider the family of lower bounds on  $\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma})$  obtained by taking  $q$  in (3.3) to be a  $N(\boldsymbol{\mu}, \boldsymbol{\Lambda})$  density. For the special case (2.3) of (2.1) the optimal  $\boldsymbol{\Lambda}$  is of the form  $\text{blockdiag}_{1 \leq i \leq m}(\boldsymbol{\Lambda}_i)$ , where each  $\boldsymbol{\Lambda}_i$  is a  $K \times K$  positive definite matrix, and the lower bound reduces to (3.1).*

A proof of Theorem 1 is given in the supplemental material. In the context of grouped data GLMMs (Section 2) this tells us that nothing is gained from taking  $\boldsymbol{\Lambda}$  to be a  $(mK) \times (mK)$  matrix. Rather, one can work with block-diagonal matrices, where the blocks are of dimension  $K \times K$  in keeping with the fact that, in this case,  $\ell$  can be expressed in terms of  $K$ -dimensional integrals.

### 3.2 APPROXIMATE STANDARD ERRORS

Define  $\boldsymbol{\theta} \equiv [\boldsymbol{\beta}^T, \text{vech}(\boldsymbol{\Sigma})^T]^T$ . If we treat  $\underline{\ell}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$  as a log-likelihood and  $[\boldsymbol{\mu}^T, \text{vech}(\boldsymbol{\Lambda})^T]^T$  as a vector of nuisance parameters, then, according to standard likelihood theory,

$$\widehat{\text{Asy.Cov}}(\widehat{\boldsymbol{\theta}}) \equiv \boldsymbol{\theta} \quad \text{block of } \underline{I}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Lambda})^{-1} \quad (3.4)$$

is the asymptotic covariance matrix of  $\widehat{\boldsymbol{\theta}}$  where  $\underline{I}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \equiv E\{-H\underline{\ell}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Lambda})\}$  is the *variational approximate Fisher information* matrix and  $H$  is the Hessian matrix operator with respect to  $(\boldsymbol{\theta}, \boldsymbol{\mu}, \text{vech}(\boldsymbol{\Lambda}))$ . Approximate standard errors are given by the square roots of the diagonal entries of  $\widehat{\text{Asy.Cov}}(\widehat{\boldsymbol{\theta}})$ , with all parameters set to their converged values. Efficient calculation of  $\widehat{\text{Asy.Cov}}(\widehat{\boldsymbol{\theta}})$  is described in Appendix A.6 in the supplemental material.

### 3.3 APPROXIMATE BEST PREDICTION OF RANDOM EFFECTS

Prediction of the random effects vectors  $\mathbf{u}_1, \dots, \mathbf{u}_m$  is often of interest, for example in residual-based model diagnostics. As before, let  $\mathbf{u}$  be the full vector of random effects. The *best predictor* of  $\mathbf{u}$  is  $\text{BP}(\mathbf{u}) = E(\mathbf{u}|\mathbf{y}) = \int_{\mathbb{R}^{mK}} \mathbf{u} p(\mathbf{u}|\mathbf{y}) d\mathbf{u}$ . This integral is intractable for the class of GLMMs being considered here. However, maximizing over the variational parameters coincides with minimizing the Kullback–Leibler distance between  $q(\mathbf{u})$  and  $p(\mathbf{u}|\mathbf{y})$ . For GVA the  $q(\mathbf{u}) = q(\cdot; \boldsymbol{\mu}, \boldsymbol{\Lambda})$  is the  $N(\boldsymbol{\mu}, \boldsymbol{\Lambda})$  density function, so that

$$\underline{\text{BP}}(\mathbf{u}) = \int_{\mathbb{R}^{mK}} \mathbf{u} q(\mathbf{u}; \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Lambda}}) d\mathbf{u} = \hat{\boldsymbol{\mu}}$$

is an appropriate approximation to  $\text{BP}(\mathbf{u})$  where  $\hat{\boldsymbol{\mu}}$  and  $\hat{\boldsymbol{\Lambda}}$  are defined by (3.2).

Next we address the question of variability of  $\underline{\text{BP}}(\mathbf{u})$ . From best prediction theory (e.g., McCulloch, Searle, and Neuhaus 2008, chap. 13) we have the result  $\text{Cov}\{\text{BP}(\mathbf{u}) - \mathbf{u}\} = E_{\mathbf{y}}\{\text{Cov}(\mathbf{u}|\mathbf{y})\}$ . Replacement of  $p(\mathbf{u}|\mathbf{y})$  by  $q(\mathbf{u}; \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Lambda}})$  then leads to the estimated asymptotic covariance matrix:

$$\widehat{\text{Asy.Cov}}\{\underline{\text{BP}}(\mathbf{u}) - \mathbf{u}\} = \hat{\boldsymbol{\Lambda}}.$$

So, in summary, the maximizing variational parameters,  $\hat{\boldsymbol{\mu}}$  and  $\hat{\boldsymbol{\Lambda}}$ , can be used for predicting the random effects and measuring their variability.

### 3.4 MARGINAL INFERENCE

Some of the calculations arising in Gaussian variational approximations also arise in marginal inference. Using notation defined in Appendix A.2 in the supplemental material, the marginal mean of  $\mathbf{y}_i$  is given by  $E(\mathbf{y}_i) = B^{(1)}(\mathbf{X}_i \boldsymbol{\beta} + \text{dg}(\mathbf{Z}_i \boldsymbol{\Sigma} \mathbf{Z}_i^T))$ . This can be approximated either by the adaptive Gauss–Hermite quadrature described in the supplemental material or by simple approximation methods such as those used by Zeger, Liang, and Albert (1988).

## 4. THEORETICAL PROPERTIES

Hall, Ormerod, and Wand (2011) studied the theoretical properties of GVA in the Poisson case with  $(\mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{u}_i)_j = \beta_0 + u_i + \beta_1 x_{ij}$  (the design structure of Example 1 in Section 2) and  $n_i = n$  for all  $1 \leq i \leq m$ . For this special case only, they proved that, as  $m, n \rightarrow \infty$  and under relatively mild regularity assumptions,

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^0 + O_p(m^{-1/2} + n^{-1}) \quad \text{and} \quad \hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}^0 + O_p(m^{-1/2} + n^{-1}),$$

where  $\boldsymbol{\beta}^0$  and  $\boldsymbol{\Sigma}^0$  are the true parameters. For simple generalized linear mixed models we provide heuristic arguments showing GVA provides similar rates of convergence to those shown earlier in the supplementary material. This suggests that GVA is root- $m$  consistent for the general grouped data GLMM setting of Section 2 (provided number of repeated measurements to be at least as large as the square root of  $m$ ).

Opper and Archambeau (2009) showed that the Laplace approximation and the GVA are closely related. Using arguments similar to those used by Opper and Archambeau (2009)

in the current setting, using notation defined in Appendix A.2 of the supplemental material, the following equations hold:

$$\begin{aligned} D_{\boldsymbol{\mu}} \underline{\ell}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) &= \int_{\mathbb{R}^K} q(\mathbf{u}; \boldsymbol{\mu}, \boldsymbol{\Lambda}) [D_{\mathbf{u}} \log p(\mathbf{y}, \mathbf{u}; \boldsymbol{\beta}, \boldsymbol{\Sigma})] d\mathbf{u}, \\ D_{\text{vech}(\boldsymbol{\Lambda})} \underline{\ell}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) &= \text{vec} \left[ \boldsymbol{\Lambda} + \int_{\mathbb{R}^K} q(\mathbf{u}; \boldsymbol{\mu}, \boldsymbol{\Lambda}) (\mathbf{H}_{\mathbf{u}\mathbf{u}} \log p(\mathbf{y}, \mathbf{u}; \boldsymbol{\beta}, \boldsymbol{\Sigma})) d\mathbf{u} \right]^T \mathbf{D}_K. \end{aligned} \quad (4.1)$$

Setting these to zero and solving, we can see that GVA can be interpreted as being the Laplace approximation averaged over  $q(\mathbf{u}; \boldsymbol{\mu}, \boldsymbol{\Lambda})$ . Since  $\lim_{\boldsymbol{\Lambda} \rightarrow \mathbf{0}} \int_{\mathbb{R}^K} q(\mathbf{u}; \boldsymbol{\mu}, \boldsymbol{\Lambda}) f(\mathbf{u}) d\mathbf{u} = f(\boldsymbol{\mu})$ , Equations (4.1) can be used to show that as  $\boldsymbol{\Lambda} \rightarrow \mathbf{0}$  the GVA and Laplace approximations coincide.

In the context of the design structure of Example 1 in Section 2 we have  $\boldsymbol{\mu}_i \equiv \mu_i$  and  $\boldsymbol{\Lambda}_i \equiv \lambda_i$  for  $1 \leq i \leq m$  and the first-order optimality conditions for the  $\lambda_i$ 's are satisfied when  $\lambda_i = [\sigma^{-2} + \sum_{j=1}^{n_i} B^{(2)}(\beta_0 + \mu_i + \beta_1 x_{ij}, \lambda_i)]^{-1}$ . Hence, if  $\sum_{j=1}^{n_i} B^{(2)}(\beta_0 + \mu_i + \beta_1 x_{ij}, \lambda_i)$  is large and  $\sigma^2$  is small, then the GVA and Laplace approximations should be similar. However, if  $\sum_{j=1}^{n_i} B^{(2)}(\beta_0 + \mu_i + \beta_1 x_{ij}, \lambda_i)$  is small and  $\sigma^2$  large, then the two approximations may be quite different.

## 5. EXAMPLES

We will examine the effectiveness of GVA based on a simulation study and based on the work of Noh and Lee (2007), who took some settings from the work of McCulloch (1997), and some real datasets. Our real datasets include the well-examined *Epilepsy* dataset first presented by Thall and Vail (1990), and the *Toenail* dataset (De Backer et al. 1998; Lesaffre and Spiessens 2001). We compared the fits obtained using GVA with several alternative approximations implemented in R (R Core Development Team 2009). These approximations include PQL as implemented in the VR bundle (Venables and Ripley 2009) via the function `glmPQL()`, AGHQ (Liu and Pierce 1994; see also Pinheiro and Bates 1995) in the package `lme4` (Bates and Maechler 2009) via the function `glmer()`, and Laplace's approximation via the function `glmer()`. Note that the `lme4` package does not report standard errors for variance components.

The AGHQ method can be made arbitrarily accurate by increasing the number of quadrature points. We adopted the strategy of doubling the number of quadrature points until there was negligible difference in the values of the estimators. This means that the AGHQ results are exact, and hence AGHQ is the "gold standard" against which GVA and PQL may be compared when the true values of the parameters are not known.

### 5.1 SIMULATED DATA

We first consider a simulation study with settings similar to those described by McCulloch (1997) and Noh and Lee (2007). Consider the Poisson random intercept model

$$y_{ij} | u_i \sim \text{Poisson}(\exp(\boldsymbol{\beta}^T \mathbf{x}_{ij} + u_i)) \quad (5.1)$$

and logistic random intercept model

$$y_{ij} = 1 | u_i = \text{Bernoulli}\left[\frac{1}{1 + \exp(-\boldsymbol{\beta}^T \mathbf{x}_{ij} - u_i)}\right] \quad (5.2)$$

with  $u_i \sim N(0, \sigma^2)$ . We considered the following settings:

1. (5.1) using  $\boldsymbol{\beta} = [-2, -2]^T$ ,  $\mathbf{x}_{ij} = [1, j - 1]^T$ ,  $\sigma = 1\frac{1}{4}$ ,  $n_i = 2$ , and  $m \in \{100, 500\}$ .
2. (5.2) using  $\boldsymbol{\beta} = [1, 1]^T$ ,  $\mathbf{x}_{ij} = [1, j - 1]^T$ ,  $\sigma = 2$ ,  $n_i = 2$ , and  $m \in \{100, 500\}$ .
3. (5.2) using  $\boldsymbol{\beta} = [0, 5]^T$ ,  $\mathbf{x}_{ij} = [1, j/8]^T$ ,  $\sigma = \sqrt{1.5}$ ,  $n_i = 8$ , and  $m \in \{15, 50\}$ .

The above settings are a subset of those considered by [Noh and Lee \(2007\)](#). This subset was chosen, based on the discussion in Section 4, on cases where there could be a difference between GVA and the PQL, Laplace, and AGHQ methods. We also compared smaller values of  $\sigma$ . However, the results from these settings were not particularly informative as to the differences between the various methods and were omitted from the article.

We will use these settings to compare the performance of parameter estimates, standard error estimates, prediction of random effects, and computational efficiency of the PQL, Laplace, AGHQ, and GVA methods. For 2000 simulations of the data the parameter estimates, standard error estimates, predicted random effects, and times were recorded.

### 5.1.1 Assessment of Parameter and Standard Error Estimates

Table 1 compares means and standard deviations of parameter estimates  $\hat{\theta}_i$  where  $\hat{\theta}_i$  are estimates of  $\theta_i$  for the PQL, Laplace, AGHQ, and GVA methods. It also compares mean estimated standard errors (MESE) and root mean squared errors (RMSE) defined by

$$\text{MESE} = \frac{1}{2000} \sum_{i=1}^{2000} \text{se}(\hat{\theta}_i) \quad \text{and} \quad \text{RMSE} = \sqrt{\frac{1}{2000} \sum_{i=1}^{2000} (\hat{\theta}_i - \theta_i^0)^2}$$

respectively where  $\text{se}(\hat{\theta}_i)$  is the estimated standard error reported by each method and  $\theta_i^0$  is the true value of  $\theta$ . Note that the PQL, Laplace, and AGHQ methods had rare convergence problems resulting in diverging  $\sigma$  estimates. We assumed that such estimates were due to the implementation of these methods rather than properties of these methods themselves. Hence, such cases were not included in the final results.

We note from Table 1 that GVA performed well on each of the settings. While GVA may not have the smallest bias for each of the simulation settings, GVA achieved either the smallest RMSE or was comparable in RMSE for each of the simulation settings. Furthermore, the MESEs for GVA compare favorably with SD values for all parameters, which cannot be said for some of the other methods where SD is notably underestimated by MESE.

[Noh and Lee \(2007\)](#) also compared a wide range of estimators for binary response mixed models including PQL, corrected PQL, various estimators based on  $h$ -likelihoods, and Gauss–Hermite quadrature. We refer the readers to this article for further comparisons with other methods for settings 2 and 3. We see from tables 5 and 6 of the article by [Noh](#)

Table 1. Means, standard deviations, mean estimated standard errors (MESE), and root mean squared errors (RMSE) of each model parameter for the simulation study of Section 5.1.

Setting	Method	$\beta_0$			$\beta_1$			$\sigma$	
		Mean (SD)	MESE	RMSE	Mean (SD)	MESE	RMSE	Mean (SD)	RMSE
1. $m = 100$	PQL	-2.72 (0.65)	0.28	0.97	-2.09 (0.58)	0.24	0.59	5.03 (2.36)	4.45
	Laplace	-2.22 (0.50)	0.30	0.55	-2.09 (0.58)	0.68	0.59	1.43 (0.42)	0.46
	AGHQ	-2.14 (0.43)	0.29	0.45	-2.09 (0.58)	0.68	0.59	1.34 (0.36)	0.37
	GVA	-1.86 (0.31)	0.35	0.34	-2.09 (0.58)	0.59	0.59	1.03 (0.30)	0.37
1. $m = 500$	PQL	-2.72 (0.29)	0.13	0.77	-2.02 (0.24)	0.08	0.24	4.84 (0.97)	3.72
	Laplace	-2.20 (0.22)	0.13	0.30	-2.02 (0.24)	0.28	0.24	1.45 (0.17)	0.26
	AGHQ	-2.14 (0.20)	0.13	0.24	-2.02 (0.24)	0.28	0.24	1.38 (0.15)	0.20
	GVA	-1.89 (0.15)	0.15	0.19	-2.02 (0.24)	0.24	0.24	1.11 (0.12)	0.19
2. $m = 100$	PQL	0.78 (0.30)	0.28	0.37	1.07 (0.53)	0.29	0.53	2.92 (0.90)	1.29
	Laplace	0.99 (0.40)	0.30	0.40	0.95 (0.41)	0.38	0.41	1.72 (0.55)	0.62
	AGHQ	0.95 (0.35)	0.30	0.35	0.94 (0.40)	0.38	0.40	1.66 (0.46)	0.57
	GVA	0.91 (0.31)	0.35	0.32	0.98 (0.42)	0.43	0.42	1.78 (0.41)	0.46
2. $m = 500$	PQL	0.80 (0.15)	0.12	0.24	1.02 (0.23)	0.13	0.23	2.86 (0.42)	0.95
	Laplace	0.98 (0.20)	0.13	0.20	0.93 (0.18)	0.17	0.19	1.70 (0.27)	0.40
	AGHQ	0.96 (0.18)	0.13	0.18	0.93 (0.18)	0.17	0.19	1.67 (0.23)	0.40
	GVA	0.93 (0.15)	0.16	0.17	0.96 (0.19)	0.17	0.19	1.80 (0.19)	0.27
3. $m = 15$	PQL	-0.10 (0.71)	0.79	0.72	5.43 (1.87)	1.52	1.92	1.84 (1.22)	1.36
	Laplace	-0.04 (0.73)	0.69	0.73	5.29 (1.60)	1.59	1.62	1.06 (0.63)	0.65
	AGHQ	-0.05 (0.72)	0.69	0.72	5.29 (1.60)	1.59	1.62	1.05 (0.61)	0.64
	GVA	-0.08 (0.70)	0.70	0.70	5.32 (1.61)	1.65	1.64	1.05 (0.60)	0.62
3. $m = 50$	PQL	-0.08 (0.37)	0.32	0.38	5.10 (0.94)	0.63	0.94	1.72 (0.59)	0.77
	Laplace	-0.01 (0.38)	0.37	0.39	5.09 (0.87)	0.82	0.88	1.16 (0.33)	0.33
	AGHQ	-0.02 (0.38)	0.37	0.39	5.10 (0.88)	0.82	0.88	1.16 (0.32)	0.33
	GVA	-0.04 (0.39)	0.38	0.38	5.13 (0.89)	0.85	0.90	1.17 (0.32)	0.32

and Lee (2007) that GVA achieved the smallest RMSE or had comparable RMSE for all parameters in table 5 (our setting 2) and table 6 (our  $\beta_1$  for our setting 3).

The mean times for the PQL, Laplace, AGHQ, and GVA methods for the most computationally demanding case (setting 2 with  $m = 500$ ) were 1.69, 0.52, 1.05, and 1.49 seconds, respectively. Hence, the GVA method appears to be on par in terms of speed with these other methods despite being programmed entirely in R and having the additional burden of calculating hundreds of one-dimensional integrals each iteration.

### 5.1.2 Assessment of Random Effect Predictions

The PQL, Laplace, and GVA methods are based on Gaussian approximations of  $p(\mathbf{u}|\mathbf{y})$ . In particular, for fixed  $\boldsymbol{\beta}$  and  $\boldsymbol{\Sigma}$  the PQL method is equivalent to the Laplace approximation for the prediction of random effects where  $E(\mathbf{u}|\mathbf{y})$  is approximated by the mode of  $p(\mathbf{y}, \mathbf{u})$  and the covariance is approximated by the inverse negative Hessian of  $\log p(\mathbf{y}, \mathbf{u})$  with respect to  $\mathbf{u}$ .

For the first 200 simulations of the simulation study described in Section 5 we recorded  $\|\boldsymbol{\mu} - E(\mathbf{u}|\mathbf{y})\|$  where  $\boldsymbol{\mu}$  are the estimated values (using either the Laplace or GVA method)

of  $E(\mathbf{u}|\mathbf{y})$  (approximated to high precision using numerical quadrature) when the likelihood parameters were fixed to their true values. The mean values of  $\|\boldsymbol{\mu} - E(\mathbf{u}|\mathbf{y})\|$  (averaged over values of  $m$ ) for settings 1–3 were 0.154, 0.238, and 0.098, respectively, for the Laplace approximation and 0.003, 0.028, and 0.001, respectively, for GVA.

These values suggest that GVA gives smaller values of  $\|\boldsymbol{\mu} - E(\mathbf{u}|\mathbf{y})\|$ , and hence possibly better predictions, than the Laplace approximation. We speculate that GVA is a better approximation of  $E(\mathbf{u}|\mathbf{y})$  than the Laplace approximation since it attempts to approximate  $E(\mathbf{u}|\mathbf{y})$  more directly. These results are consistent with the findings of Bishop (2006, chap. 10). While we do not have theoretical evidence to support this, we believe that this is a matter for further investigation.

### 5.2 EPILEPSY DATA

The *Epilepsy* dataset represents data collected from a clinical trial of 59 epileptics. Each patient was randomly selected to be administered either a new drug ( $\text{trt}_i = 1$ ) or a placebo ( $\text{trt}_i = 0$ ). The number of seizures in the eight weeks before the trial period was recorded as  $\text{base}_i$  as well as the age of the patient, recorded as  $\text{age}_i$ . Counts for the number of seizures were recorded during the two weeks before each clinical visit. Visits are recorded as  $\text{visit}_1 = -3$ ,  $\text{visit}_2 = -1$ ,  $\text{visit}_3 = 1$ , and  $\text{visit}_4 = 3$ . Finally, previous analyses (Thall and Vail 1990) have shown that the mean number of seizure counts was substantially lower for the fourth visit.

Using the *Epilepsy* dataset we first considered the Poisson random intercept model (5.1) with  $\log(\text{base}_i/4)$ ,  $\text{trt}_i$ ,  $\text{trt}_i \times \log(\text{base}_i/4)$ ,  $\log(\text{age}_i)$ , and  $\mathcal{I}(j = 4)$  as covariates for  $1 \leq i \leq 59$ ,  $1 \leq j \leq 4$ , where  $\mathcal{I}(P)$  is an indicator of  $P$  being true. This corresponds to Model II from the article of Breslow and Clayton (1993). The parameter estimates and approximate standard errors for this model are summarized in Figure 2.

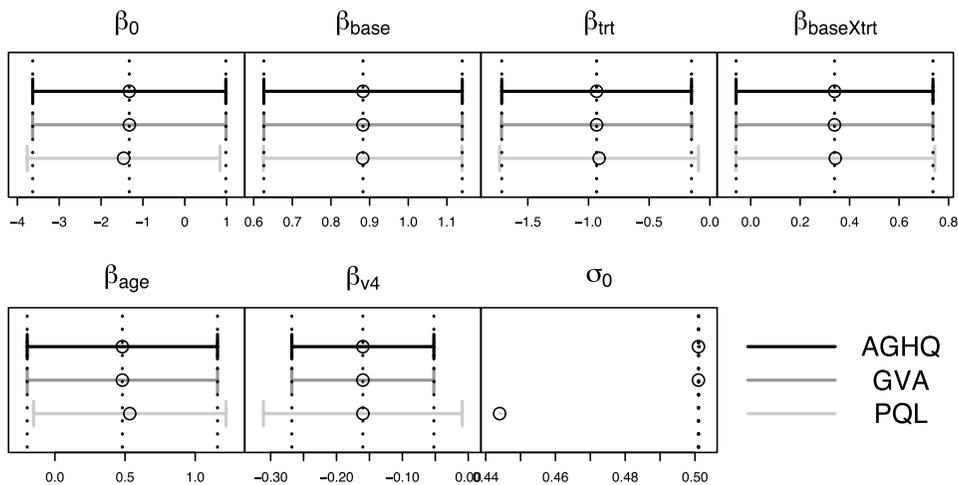


Figure 2. Point estimates and approximate 95% confidence intervals based on each of AGHQ, GVA, and PQL for the *Epilepsy* data random intercept model. The vertical dotted lines correspond to the AGHQ values.

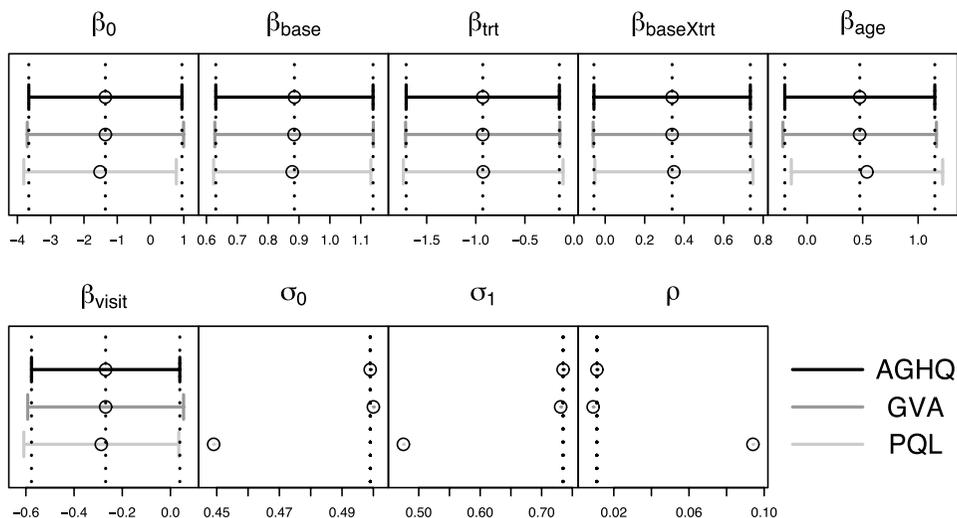


Figure 3. Point estimates and approximate 95% confidence intervals based on each of AGHQ, GVA, and PQL for the *Epilepsy* data random intercept and slope model. The vertical dotted lines correspond to the AGHQ values.

We also considered the Poisson random intercept and slope models of the form

$$y_{ij}|u_{0i}, u_{1i} \sim \text{Poisson}[\exp\{(\beta_0 + u_{0i}) + (\beta_{\text{visit}} + u_{1i})\text{visit}_j + \beta_{\text{base}} \log(\text{base}_i/4) + \beta_{\text{trt}} \text{trt}_i + \beta_{\text{baseXtrt}} \log(\text{base}_i/4) \times \text{trt}_i + \beta_{\text{age}} \log(\text{age}_i)\}],$$

where the distribution of  $[u_{0i}, u_{1i}]^T$  is the same as in Example 2 in Section 2. This model corresponds to Model IV from the article of [Breslow and Clayton \(1993\)](#). The parameter estimates and approximate standard errors for this model are summarized in Figure 3. From Figures 2 and 3 we see that all of the approximation methods compare favorably with AGHQ, although the estimates of the covariance parameters for PQL are poor. Last, fitting the above random slope model on the first author's laptop took around 2 seconds to fit using GVA while AGHQ took around 50 seconds.

### 5.3 TOENAIL DATA

This dataset considers information gathered from a longitudinal dermatological clinical trial. The aim of the trials was to compare the effectiveness of two oral treatments for a particular type of toenail infection ([De Backer et al. 1998](#)). In total, 1908 measurements from 294 patients are recorded in the dataset. Each participant in the trial was randomly administered either a treatment  $\text{trt}_{ij} = 1$  or a placebo  $\text{trt}_{ij} = 0$  and was evaluated at seven visits (approximately on weeks  $\text{time}_{ij} = 0, 4, 8, 12, 24, 36,$  and  $48$ ). The degree of separation of the nail plate from the nail bed (0, absent; 1, mild; 2, moderate; 3, severe) at each visit was recorded. In this dataset, only a dichotomized response of onycholysis (0, absent or mild; 1, moderate or severe) is available.

We considered the logistic random intercept model (5.2) with  $\text{trt}_{ij}$ ,  $\text{time}_{ij}$ , and  $\text{trt}_{ij} \times \text{time}_{ij}$  as covariates for  $1 \leq i \leq 294$  and  $n_i$  taking values from 1 to 7. The results for each GLMM approximation are summarized in Figure 4. From this figure we

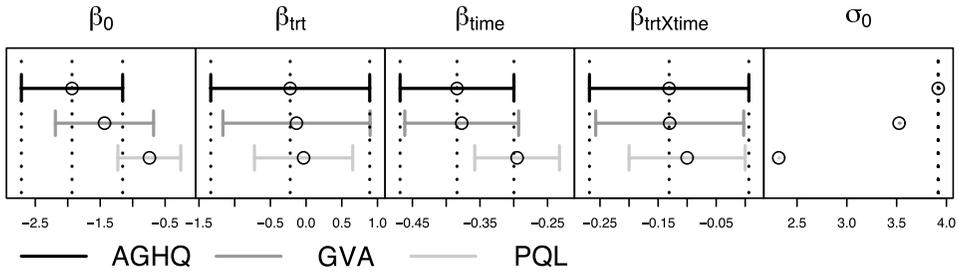


Figure 4. Point estimates and approximate 95% confidence intervals based on each of AGHQ, GVA, and PQL for the model fitted to the Toenail dataset. The vertical dotted lines correspond to the AGHQ values.

see that the GVA estimates of all parameters are better than those based on PQL. For this example we see severe bias of PQL for the parameters  $\beta_0$ ,  $\beta_{\text{time}}$ , and  $\sigma_0$ . The GVA method shows some bias for estimation of  $\beta_0$  and  $\sigma_0$ , but this bias is less severe than that of PQL.

The analysis of [Lesaffre and Spiessens \(2001\)](#) on the Toenail dataset found that parameter estimates could vary significantly even among several Gauss–Hermite quadrature and adaptive Gauss–Hermite quadrature approximations. Hence it is not surprising that there are large discrepancies between AGHQ, GVA, and PQL approximations for this dataset.

## 6. DISCUSSION

As data become cheaper to collect and store, both the number and size of data analyses will continue to grow. Therefore, it is important that statistical methodology adapts accordingly. MCMC provides an effective means of analysis for many grouped data GLMM applications. However, there are numerous situations where faster approximate methods are desirable.

Gaussian variational approximation offers itself as an attractive alternative to the PQL, Laplace, and AGHQ methods for fast GLMM approximate inference. It is more accurate than PQL for parameter estimation, a better predictor of random effects than Laplace’s method, and, unlike AGHQ, it is scalable to high-dimensional integrals.

Our methodology provides an efficient and attractive solution to the problems of parameter and standard error estimation, and prediction of random effects in a seamless manner while avoiding multivariate intractable integrals. It is competitive with a variety of other methods, especially in terms of RMSE of parameter estimates. A future challenge is to handle more general GLMMs, such as those with spatial correlation structures or containing spline basis functions in the random effects design matrix.

## SUPPLEMENTARY MATERIALS

**Appendixes:** Appendix A contains computational details for implementing Gaussian variational inference for generalized linear mixed models. Appendix B contains proofs of results in the article. (gvapAppendix.pdf)

**R Code:** The R code for reproducing the results corresponding to Table 1 and Figures 2–4, and a “readme” file describing each of the files are contained in the zip file. (GVA.zip)

## ACKNOWLEDGMENTS

This research was supported in part by Australian Research Council Discovery Project DP0877055. The authors are grateful to Dobrin Marchev and John Robinson for their helpful comments.

[Received July 2009. Revised February 2011.]

## REFERENCES

- Archambeau, C., Cornford, D., Opper, M., and Shawe-Taylor, J. (2007), “Gaussian Process Approximations of Stochastic Differential Equations,” *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 1, 1–16. [4]
- Barber, D., and Bishop, C. M. (1998), “Ensemble Learning for Multi-Layer Networks,” in *Advances in Neural Information Processing Systems*, eds. M. I. Jordan, K. J. Kearns, and S. A. Solla, Vol. 10, Cambridge, MA: MIT Press, pp. 395–401. [4]
- Bates, D., and Maechler, M. (2009), “lme4 0.999375. Linear Mixed-Effects Models Using S4 Classes,” R package, available at <http://cran.r-project.org>. [10]
- Bishop, C. M. (2006), *Pattern Recognition and Machine Learning*, New York: Springer. [3,7,13]
- Breslow, N. E., and Clayton, D. G. (1993), “Approximate Inference in Generalized Linear Mixed Models,” *Journal of the American Statistical Association*, 88, 9–25. [3,13,14]
- Carlin, B. P., and Louis, T. A. (2008), *Bayes and Empirical Bayes Methods for Data Analysis* (3rd ed.), New York: Chapman & Hall. [2]
- De Backer, M., De Vroey, C., Lesaffre, E., Scheys, I., and De Keyser, P. (1998), “Twelve Weeks of Continuous Oral Therapy for Toenail Onychomycosis Caused by Dermatophytes: A Double-Blind Comparative Trial of Terbinafine 250 mg/Day versus Itraconazole 200 mg/Day,” *Journal of the American Academy of Dermatology*, 38 (5), S57–S63. [10,14]
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis*, Boca Raton, FL: Chapman & Hall. [2]
- Hall, P., Ormerod, J. T., and Wand, M. P. (2011), “Theory of Gaussian Variational Approximation for a Poisson Linear Mixed Model,” *Statistica Sinica*, 21, 369–389. [3,9]
- Honkela, A., and Valpola, H. (2005), “Unsupervised Variational Bayesian Learning of Nonlinear Models,” in *Advances in Neural Information Processing Systems*, Vol. 17, Cambridge, MA: MIT Press, pp. 593–600. [4]
- Jaakkola, T. S., and Jordan, M. I. (2000), “Bayesian Parameter Estimation via Variational Methods,” *Statistics and Computing*, 10, 25–37. [4]
- Jordan, M. I. (2004), “Graphical Models,” *Statistical Science*, 19, 140–155. [3]
- Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999), “An Introduction to Variational Methods for Graphical Models,” *Machine Learning*, 37, 183–233. [3]
- Kullback, S., and Leibler, R. A. (1951), “On Information and Sufficiency,” *The Annals of Mathematical Statistics*, 22, 79–86. [8]
- Leach, A. (2000), “Menzie's School of Health Research 1999–2000 Annual Report,” Menzie's School of Health Research, pp. 18–21. [5]
- Lee, Y., and Nelder, J. A. (1996), “Hierarchical Generalized Linear Models,” *Journal of the Royal Statistical Society, Ser. B*, 58, 619–656. [3]
- Lesaffre, E., and Spiessens, B. (2001), “On the Effect of the Number of Quadrature Points in a Logistic Random-Effects Model: An Example,” *Applied Statistics*, 50, 325–335. [10,15]

- Little, R. J., and Rubin, D. B. (2004), *Statistical Analysis With Missing Data* (2nd ed.), New York: Wiley. [2]
- Liu, Q., and Pierce, D. A. (1994), “A Note on Gauss–Hermite Quadrature,” *Biometrika*, 81, 624–629. [7,10]
- Marin, J.-M., and Robert, C. P. (2007), *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*, New York: Springer. [2]
- McCulloch, C. E. (1997), “Maximum Likelihood for Algorithms for Generalized Linear Mixed Models,” *Journal of the American Statistical Association*, 92, 162–170. [10]
- McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008), *Generalized, Linear, and Mixed Models* (2nd ed.), New York: Wiley. [2,9]
- Neal, R. M., and Hinton, G. E. (1999), “A New View of the EM Algorithm That Justifies Incremental and Other Variants,” in *Learning in Graphical Models*, ed. M. I. Jordan, Cambridge, MA: MIT Press, pp. 355–368. [7]
- Noh, M., and Lee, Y. (2007), “REML Estimation for Binary Data in GLMMs,” *Journal of Multivariate Analysis*, 98, 896–915. [10-12]
- Opper, M., and Archambeau, C. (2009), “Variational Gaussian Approximation Revisited,” *Neural Computation*, 21, 786–792. [4,9]
- Ormerod, J. T. (2008), “On Semiparametric Regression and Data Mining.” Ph.D. thesis, School of Mathematics and Statistics, The University of New South Wales, Sydney, Australia. [4,7]
- Ormerod, J. T., and Wand, M. P. (2008), “Variational Approximations for Logistic Mixed Models,” in *Proceedings of the Ninth Iranian Statistics Conference*, Department of Statistics, University of Isfahan, Isfahan, Iran, pp. 450–467. [4]
- (2010), “Explaining Variational Approximation,” *The American Statistician*, 64, 140–153. [3]
- Parisi, G. (1988), *Statistical Field Theory*, Redwood City, CA: Addison-Wesley. [4]
- Pinheiro, J. C., and Bates, D. M. (1995), “Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model,” *Journal of Computational and Graphical Statistics*, 4, 12–35. [10]
- R Development Core Team (2009), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. Available at [www.R-project.org](http://www.R-project.org). [10]
- 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. [3]
- Rue, H., Martino, S., and Chopin, N. (2009), “Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations,” *Journal of the Royal Statistical Society, Ser. B*, 7, 319–392. [3,4]
- Seeger, M. (2000), “Bayesian Model Selection for Support Vector Machines, Gaussian Processes and Other Kernel Classifiers,” in *Advances in Neural Information Processing Systems*, Vol. 12, Cambridge, MA: MIT Press, pp. 603–609. [4]
- Skrondal, A., and Rabe-Hesketh, S. (2004), *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*, Boca Raton, FL: Chapman & Hall. [2]
- Thall, P. F., and Vail, S. C. (1990), “Some Covariance Models for Longitudinal Count Data With Overdispersion,” *Biometrics*, 46, 657–671. [10,13]
- Titterton, D. M. (2004), “Bayesian Methods for Neural Networks and Related Models,” *Statistical Science*, 19, 128–139. [7]
- Venables, W. N., and Ripley, B. D. (2009), “vr 7.2. Functions and Datasets to Support Venables and Ripley ‘Modern Applied Statistics With S’ (4th ed.),” R package, available at <http://cran.r-project.org>. [10]
- Wang, B., and Titterton, D. M. (2005), “Inadequacy of Interval Estimates Corresponding to Variational Bayesian Approximations,” in *Proceedings of the 10th International Workshop on Artificial Intelligence*, eds. R. G. Cowell and Z. Ghahramani, Barbados: Society for Artificial Intelligence and Statistics, pp. 373–380. [4]
- Zeger, S. L., Liang, K. Y., and Albert, P. S. (1988), “Models for Longitudinal Data: A Generalized Estimating Equation Approach,” *Biometrics*, 44, 1049–1060. [9]
- Zhao, Y., Staudenmayer, J., Coull, B. A., and Wand, M. P. (2006), “General Design Bayesian Generalized Linear Mixed Models,” *Statistical Science*, 21, 35–51. [3,6]