Gaussian Variational Approximate Inference for Generalized Linear Mixed Models

BY J.T. ORMEROD AND M.P. WAND

Centre for Statistical and Survey Methodology,

School of Mathematics and Applied Statistics,

University of Wollongong, Wollongong 2522, Australia.

7th July, 2009

Abstract:

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 (GLMM) 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.

Key Words: Best prediction; Longitudinal data analysis; Likelihood-based inference; 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 dif-

ficult inferential questions arising in, for example, Bayesian models (e.g. Gelman, Carlin, Stern & Rubin, 2004; Marin & Robert, 2007, Carlin & Louis, 2008), mixed and latent variable models (e.g. Skrondal & Rabe-Hesketh, 2004; McCulloch, Searle & Neuhaus, 2008) and missing data models (e.g. Little & Rubin, 2002). 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 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 Jordan *et al.* (1999), Jordan (2004) and Bishop (2006). An introduction to variational approximation from a statistical perspective is provided by Ormerod & Wand (2009).

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 (GLMM). In the interest of conciseness, we focus on the commonest type of GLMM – that arising in the analysis of grouped data with Gaussian random effects. General design GLMMs, as described in Zhao, Coull, Staudenmayer & Wand (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 & Clayton, 1993). However, with the aid of the formulae presented in Appendix A, 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 inte-

gration. 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. Raudenbush, Yang & Yosef, 2000; Lee & Nelder, 1996; Rue, Martino & Chopin, 2009), also offer accuracy improvements over PQL but, like GVA, have their own costs in terms of implementability.

Recently, Hall, Ormerod & Wand (2009) 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-Liebler distance (e.g. Bishop, 2006, Section 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 give rise to explicit updating formulae. In addition, mean field approximation has a tendency to markedly underestimate the variability of parameter estimates (Wang & Titterington, 2005; Rue *et al.*, 2009).

Another variant of variational approximations is the tangent transform approach of Jaakkola & 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 under-estimation problems with the Jaakkola and Jordan variational approximation (Ormerod & Wand, 2008).

The use of Gaussian densities in variational approximation has a small and recent literature in Machine Learning. Gaussian variational approximations have been considered in the context of neural networks (Barber & Bishop, 1998; Honkela & Valpola, 2004), Support Vector Machines (Seeger, 2000), stochastic differential equations (Archambeah, Cornford, Opper &

Shawe-Taylor, 2007) and robust regression (Opper & Archambeau, 2009). None of this research is directly connected with GLMMs.

Section 2 describes the type of data and the types GLMMs that we consider. In Section 3 we explain the use of Gaussian variational approximation for avoiding multivariate intractable integrals when making inference about model parameters. Connections with Kullback-Liebler divergence theory are described in Section 4. Section 5 deals with the problem of prediction of random effects and explains how Gaussian variational approximation provides an attractive solution. Theoretical properties of Gaussian variational approximations for grouped data GLMMs are the subject of Section 6. Examples are given in Section 7 and concluding remarks are made in Section 8.

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 jth predictor/response pair for the ith group is denoted by

$$(\boldsymbol{x}_{ij}, y_{ij}), \quad 1 \leq j \leq n_i, \quad 1 \leq i \leq m.$$

Here the entries of the predictor vectors $\boldsymbol{x}_{ij}$ are unrestricted, while the y_{ij} are subject to restrictions such as being binary or non-negative 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:

$$\boldsymbol{y}_i \equiv \begin{bmatrix} y_{i1} \\ \vdots \\ y_{in_i} \end{bmatrix} \qquad \text{and} \qquad \mathbf{1}_i \equiv \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}.$$

The first of these is the vector of responses for the ith group. It is reasonable to assume that the vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ 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

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

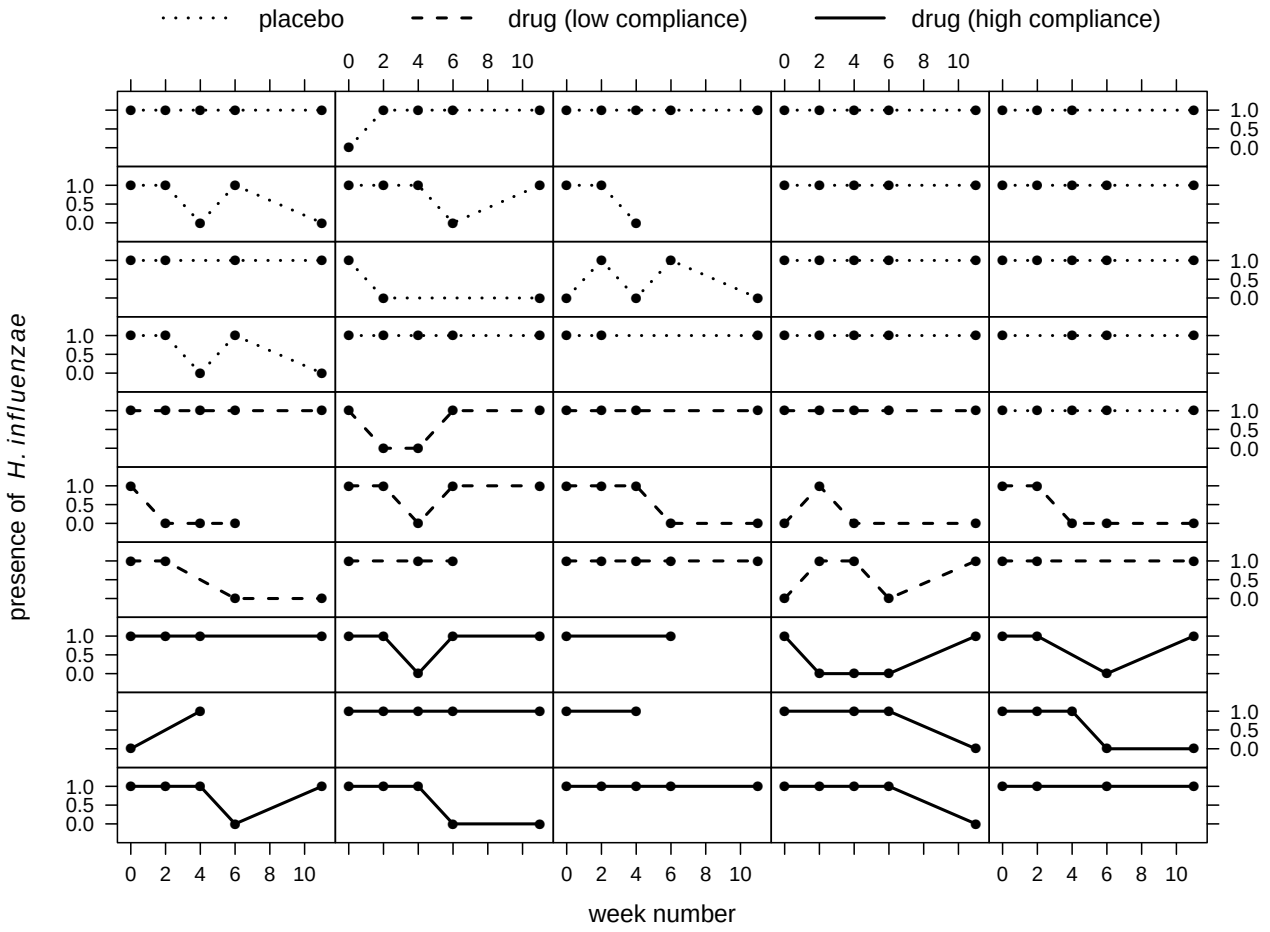

Figure 1: *Example of binary response grouped regression data. Each panel corresponds to longitudinal measurements on a child with a history of otitis media who participated in a clinical trial in Northern Territory, Australia. The horizontal axis (x) is week number since the start of the trial. The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

where $\boldsymbol{u}_i$, $1 \leq i \leq m$, are $K \times 1$ random effects vectors and $\boldsymbol{\Sigma}$ ($K \times K$) 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 $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are design matrices dependent on the $\boldsymbol{x}_{ij}$, and are assumed to be fixed. Digestion of the design structure is aided by the following two examples:

EXAMPLE 1. *Suppose* $(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{u}_i)_j = \beta_0 + u_i + \beta_1\, x_{ij}$ *where* $u_i \sim N(0, \sigma_u^2)$. *In this case* $\boldsymbol{\beta} = [\beta_0, \beta_1]^T$, $\boldsymbol{u}_i = u_i$, $K = 1$ *and* $\boldsymbol{\Sigma} = \sigma^2$, *The design matrices are*

$$\boldsymbol{X}_i = [1\ x_{ij}]_{1\le j\le n_i} \quad \textit{and} \quad \boldsymbol{Z}_i = [1]_{1\le j\le n_i} \quad \textit{for } 1 \le i \le m.$$

□

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

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

In this case $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2]^T$, $\boldsymbol{u}_i = [u_{0i}, u_{1i}]^T$, $K = 2$ *and*

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_0^2 & \rho\sigma_0\sigma_1 \\ \rho\sigma_0\sigma_1 & \sigma_1^2 \end{bmatrix}.$$

The design matrices are

$$\boldsymbol{X}_i = [1, x_{1ij}, x_{2ij}]_{1\le j\le n_i} \quad \textit{and} \quad \boldsymbol{Z}_i = [1, x_{1ij}]_{1\le j\le n_i} \quad \textit{for } 1 \le i \le m.$$

□

Model (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 (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 (1) a *grouped data* GLMM.

The class of GLMMs is much more general than (1), as explained in Zhao *et al.* (2006). Staying within the one-parameter exponential family, a more general class of models is

$$\boldsymbol{y}|\boldsymbol{u} \sim \exp\{\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}) - \mathbf{1}^T b(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}) + \mathbf{1}^T c(\boldsymbol{y})\}, \quad \boldsymbol{u} \sim N(\mathbf{0}, \boldsymbol{G}), \tag{2}$$

where the design matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ and covariance matrix $\boldsymbol{G}$ are quite general. In the special case of (1) we have

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_m \end{bmatrix}, \ \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_m \end{bmatrix}, \ \boldsymbol{u} = \begin{bmatrix} \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_m \end{bmatrix}, \ \boldsymbol{Z} = \underset{1 \le i \le m}{\text{blockdiag}}(\boldsymbol{Z}_i) \text{ and } \boldsymbol{G} = \boldsymbol{I}_m \otimes \boldsymbol{\Sigma}.$$

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

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

and the maximum likelihood estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are

$$(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}}) = \underset{\boldsymbol{\beta}, \boldsymbol{\Sigma}}{\text{argmax}} \, \ell(\boldsymbol{\beta}, \boldsymbol{\Sigma}). \tag{3}$$

Evaluation of (3) 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 $1 \le i \le m$, 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} \{\boldsymbol{y}_i^T \boldsymbol{X}_i \boldsymbol{\beta} + \mathbf{1}_i^T c(\boldsymbol{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\{ \boldsymbol{y}_i^T \boldsymbol{Z}_i \boldsymbol{u} - \mathbf{1}_i^T b(\boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{u}) - \frac{1}{2} \boldsymbol{u}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{u} \right\} \\
&\qquad \times \frac{(2\pi)^{-m/2} |\boldsymbol{\Lambda}_i|^{-1/2} \exp\{-\frac{1}{2}(\boldsymbol{u} - \boldsymbol{\mu}_i)^T \boldsymbol{\Lambda}_i^{-1} (\boldsymbol{u} - \boldsymbol{\mu}_i)\}}{(2\pi)^{-m/2} |\boldsymbol{\Lambda}_i|^{-1/2} \exp\{-\frac{1}{2}(\boldsymbol{u} - \boldsymbol{\mu}_i)^T \boldsymbol{\Lambda}_i^{-1} (\boldsymbol{u} - \boldsymbol{\mu}_i)\}} \, d\boldsymbol{u} \\
&= \sum_{i=1}^{m} \{\boldsymbol{y}_i^T \boldsymbol{X}_i \boldsymbol{\beta} + \mathbf{1}_i^T c(\boldsymbol{y}_i)\} - \frac{m}{2} \log |\boldsymbol{\Sigma}| - \frac{mK}{2} \log(2\pi) \\
&\quad + \sum_{i=1}^{m} \log E_{\boldsymbol{u} \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)} \left[\frac{\exp\left\{ \boldsymbol{y}_i^T \boldsymbol{Z}_i \boldsymbol{u} - \mathbf{1}_i^T b(\boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{u}) - \frac{1}{2} \boldsymbol{u}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{u} \right\}}{(2\pi)^{-m/2} |\boldsymbol{\Lambda}_i|^{-1/2} \exp\{-\frac{1}{2}(\boldsymbol{u} - \boldsymbol{\mu}_i)^T \boldsymbol{\Lambda}_i^{-1} (\boldsymbol{u} - \boldsymbol{\mu}_i)\}} \right] \\
&\geq \sum_{i=1}^{m} \{\boldsymbol{y}_i^T \boldsymbol{X}_i \boldsymbol{\beta} + \mathbf{1}_i^T c(\boldsymbol{y}_i)\} - \frac{m}{2} \log |\boldsymbol{\Sigma}| - \frac{mK}{2} \log(2\pi) \\
&\quad + \sum_{i=1}^{m} E_{\boldsymbol{u} \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)} \left(\log \left[\frac{\exp\left\{ \boldsymbol{y}_i^T \boldsymbol{Z}_i \boldsymbol{u} - \mathbf{1}_i^T b(\boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{u}) - \frac{1}{2} \boldsymbol{u}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{u} \right\}}{(2\pi)^{-m/2} |\boldsymbol{\Lambda}_i|^{-1/2} \exp\{-\frac{1}{2}(\boldsymbol{u} - \boldsymbol{\mu}_i)^T \boldsymbol{\Lambda}_i^{-1} (\boldsymbol{u} - \boldsymbol{\mu}_i)\}} \right] \right) \\
&\equiv \underline{\ell}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Lambda})
\end{aligned}$$

where

$$(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \equiv (\boldsymbol{\mu}_1, \boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m).$$

The *variational lower bound* on the log-likelihood simplifies to:

$$\begin{aligned}
\underline{\ell}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) &= \tfrac{mK}{2} + \textstyle\sum_{i=1}^{m} \mathbf{1}_i^T c(\boldsymbol{y}_i) - \frac{m}{2} \log |\boldsymbol{\Sigma}| \\
&\quad + \textstyle\sum_{i=1}^{m} [\boldsymbol{y}_i^T (\boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{\mu}_i) - \mathbf{1}_i^T B(\boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{\mu}_i, \text{diagonal}(\boldsymbol{Z}_i \boldsymbol{\Lambda}_i \boldsymbol{Z}_i^T)) \\
&\qquad + \tfrac{1}{2} \{\log |\boldsymbol{\Lambda}_i| - \boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i - \text{tr}(\boldsymbol{\Sigma}^{-1} \boldsymbol{\Lambda}_i)\}]
\end{aligned} \tag{4}$$

where

$$B(\mu, \sigma^2) \equiv \int_{-\infty}^{\infty} b(\sigma x + \mu) \phi(x) \, dx,$$

ϕ is the $N(0,1)$ density function and, for a square matrix $\boldsymbol{A}$, diagonal($\boldsymbol{A}$) is the column vector containing the diagonal entries of $\boldsymbol{A}$. As with b and c, evaluations of B for vector arguments are applied in an element-wise fashion. An explicit example is:

$$B\left(\begin{bmatrix} 3 \\ 5 \\ 1 \end{bmatrix}, \begin{bmatrix} 6 \\ 7 \\ 4 \end{bmatrix} \right) = \begin{bmatrix} B(3,6) \\ B(5,7) \\ B(1,4) \end{bmatrix}.$$

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 K-dimensional integration. 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

$$B(\mu, \sigma^2) = \int_{-\infty}^{\infty} \log\{1 + \exp(\sigma x + \mu)\}\phi(x)\, dx,$$

which does not have an analytic solution. However, adaptive Gauss-Hermite quadrature (Liu & Pearce, 1994) is well-suited to efficient and very accurate evaluation of $B(\mu, \sigma^2)$ in this case. The details are given in Appendix A.2.

Given the lower-bound result,

$$\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma}) \geq \underline{\ell}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \quad \text{for all} \quad (\boldsymbol{\mu}, \boldsymbol{\Lambda})$$

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 $\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma})$. This leads to the *Gaussian variational approximate* maximum likelihood estimators:

$$(\widehat{\underline{\boldsymbol{\beta}}}, \widehat{\underline{\boldsymbol{\Sigma}}}) = (\boldsymbol{\beta}, \boldsymbol{\Sigma}) \text{ component of } \underset{\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Lambda}}{\operatorname{argmax}} \underline{\ell}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Lambda}). \tag{5}$$

Appendix A provides efficient computational formulas for solving this maximization problem.

3.1 Approximate Standard Errors

Define

$$\boldsymbol{\theta} \equiv \begin{bmatrix} \boldsymbol{\beta} \\ \operatorname{vech}(\boldsymbol{\Sigma}) \end{bmatrix} \quad \text{and} \quad \boldsymbol{\xi} \equiv \begin{bmatrix} \boldsymbol{\mu}_1 \\ \operatorname{vech}(\boldsymbol{\Lambda}_1) \\ \vdots \\ \boldsymbol{\mu}_m \\ \operatorname{vech}(\boldsymbol{\Lambda}_m) \end{bmatrix} \tag{6}$$

to be the vectors containing the unique model and variational parameters, respectively. If $\underline{\ell}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ is treated as a log-likelihood function and $\boldsymbol{\xi}$ is treated as a set of nuisance parameters then, according to standard likelihood theory, the asymptotic covariance matrix of $\widehat{\underline{\boldsymbol{\theta}}}$ is

$$\widehat{\text{Asy. Cov}}(\widehat{\underline{\boldsymbol{\theta}}}) \equiv \boldsymbol{\theta} \text{ block of } \underline{I}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Lambda})^{-1} \tag{7}$$

where

$$\underline{I}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \equiv E\{-\mathsf{H}\underline{\ell}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Lambda})\}$$

is the *variational approximate Fisher information* matrix and H is the Hessian matrix operator with respect to $(\boldsymbol{\beta}, \boldsymbol{\xi})$. Approximate standard errors are given by the square-roots of the diagonal entries of $\widehat{\text{Asy. Cov}}(\widehat{\underline{\boldsymbol{\theta}}})$, with all parameters set to their converged values.

Efficient calculation of $\widehat{\text{Asy. Cov}}(\widehat{\underline{\boldsymbol{\theta}}})$ is described in Appendix A.6.

4 Relationship with Kullback-Liebler Divergence

The lower bound expression can also be derived using the ideas of Kullback-Leibler divergence, which underpins much of variational approximation methodology (e.g. Titterington, 2004; Bishop, 2006, Chapter 10). In this section, we first work with the general form of the GLMM given by (2). We also use p to denote density or probability mass functions of random vectors according to the model. For example, $p(\boldsymbol{u}|\boldsymbol{y})$ is the conditional density function of $\boldsymbol{u}$ given $\boldsymbol{y}$. The log-likelihood can be written in terms of the joint density of $\boldsymbol{y}$:

$$\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \log p(\boldsymbol{y}; \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \log p(\boldsymbol{y}).$$

Let M be the dimension of the $\boldsymbol{u}$ vector. For an arbitrary density functions q on $\mathbb{R}^M$

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

The second term is the Kullback-Leibler distance between $q(\boldsymbol{u})$ and $p(\boldsymbol{u}|\boldsymbol{y})$. Since this is always non-negative (Kullback & Leibler, 1951) we get

$$\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma}) \geq \int_{\mathbb{R}^M} q(\boldsymbol{u}) \log \left\{ \frac{p(\boldsymbol{y}, \boldsymbol{u})}{q(\boldsymbol{u})} \right\} d\boldsymbol{u}. \tag{8}$$

Substitution of $q(\boldsymbol{u}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ into this expression 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 (8) to be a $N(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ density. For the special case of (2) corresponding to (1) the optimal $\boldsymbol{\Lambda}$ is of the form* 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 (4).*

A proof of Theorem 1 is given in Appendix B. It tells us that, in the case of grouped data GLMMs (Section 2), there is nothing to be gained from taking $\boldsymbol{\Lambda}$ to be a general $(mK) \times (mK)$ positive definite matrix. Rather, one can work with the smaller class of block-diagonal matrices, where the blocks are of dimension $K \times K$. Such a result is in keeping with the fact that, for grouped data GLMMs, the log-likelihood reduces to a set of $K \times K$ integrals.

5 Approximate Best Prediction of Random Effects

Prediction of the random effects vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ is often of interest. For example, it is needed for residual-based model diagnostics. As before, let

$$\boldsymbol{u} = \begin{bmatrix} \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_m \end{bmatrix}$$

be the full vector of random effects. The *best predictor* of $\boldsymbol{u}$ is

$$\text{BP}(\boldsymbol{u}) = E(\boldsymbol{u}|\boldsymbol{y}) = \int_{\mathbb{R}^{mK}} \boldsymbol{u}\, p(\boldsymbol{u}|\boldsymbol{y})\, d\boldsymbol{u}.$$

This integral is also intractable for the class of GLMMs being considered in this article. However, maximizing over the variational parameters coincides with minimizing the Kullback-Leibler distance between $q(\boldsymbol{u})$ and $p(\boldsymbol{u}|\boldsymbol{y})$. For Gaussian variational approximation $q = q(\cdot; \boldsymbol{\mu}, \boldsymbol{\Lambda})$ is the $N(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ density function, so an appropriate approximation to BP$(\boldsymbol{u})$ is

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

where

$$(\underline{\widehat{\boldsymbol{\mu}}}, \underline{\widehat{\boldsymbol{\Lambda}}}) = (\boldsymbol{\mu}, \boldsymbol{\Lambda}) \text{ component of } \operatorname*{argmax}_{\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Lambda}} \underline{\ell}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Lambda}).$$

Next we address the question of variability of $\underline{\text{BP}}(\boldsymbol{u})$. From best prediction theory (e.g. Chapter 13 of McCulloch *et al.* 2008) we have the result

$$\text{Cov}\{\text{BP}(\boldsymbol{u}) - \boldsymbol{u}\} = E_{\boldsymbol{y}}\{\text{Cov}(\boldsymbol{u}|\boldsymbol{y})\}.$$

Replacement of $p(\boldsymbol{u}|\boldsymbol{y})$ by $q(\boldsymbol{u}; \widehat{\underline{\boldsymbol{\mu}}}, \widehat{\underline{\boldsymbol{\Lambda}}})$ then leads to the estimated asymptotic covariance matrix:

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

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

6 Theoretical Properties

Hall, Ormerod & Wand (2009) studied the theoretical properties of GVA in the Poisson case with $(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{u}_i)_j = \beta_0 + u_i + \beta_1 x_{ij}$ (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 $m, n \to \infty$ and under relatively mild regularity assumptions,

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

where $\boldsymbol{\beta}^0$ and $\boldsymbol{\Sigma}^0$ are the true parameters. 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). We have performed some heuristic arguments that support such a result although we do not yet have a rigorous proof.

7 Examples

We will examine the effectiveness of GVA based on the well-examined *Epilepsy* dataset first presented by Thall & Vail (1990), the *Bacteria* dataset (Leach, 2000), the *Toenail* dataset (De Backer, De Vroey, Lesaffre, Scheys & De Keyser, 1998; Lesaffre & Spiessens, 2001) and a simulation study similar to one used by Zeger & Karim (1991) and Breslow & Clayton (1993).

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 & Ripley, 2009) via the function `glmmPQL()`, Adaptive

Gauss-Hermite Quadrature (AGHQ) (Liu & Pierce, 1994; see also Pinheiro & Bates, 1995) in the package `lme4` (Bates & Maechler, 2009) via the function `glmer()`. AGHQ 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. Note that the `lme4` package does not report standard errors for variance components.

7.1 Simulated Data

We conducted a simulation study similar to that described in Zeger & Karim (1991). We generated data according to:

$$\text{logit}\left\{P(y_{ij}=1)\right\} = \boldsymbol{\beta}^T \boldsymbol{x}_{ij} + u_i \quad \text{and} \quad u_i \sim N(0, \sigma^2) \tag{9}$$

with $\boldsymbol{\beta} = [-2.5, 1, -1, 0.5]^T$, $\sigma = 1$, $\boldsymbol{x}_{ij} = [1, t_i, x_j, t_i x_j]^T$ for $1 \leq i \leq 100$, $1 \leq j \leq 7$. The t_is take the value 0 for $1 \leq i \leq 50$ and 1 otherwise and the x_js take the values $-3, -2, \ldots, 2, 3$. We simulated 200 datasets based on (9). The results are summarized in Figure 2 where we see that the estimates for $\boldsymbol{\beta}$ are almost identical for each method, but that PQL underestimates σ. In addition, with the exception of β_0, the approximate 95% confidence intervals for the β_j are overly narrow for PQL, but close to exact for GVA.

7.2 Epilepsy Dataset

The *Epilepsy* dataset represents data collected from a clinical trial of 59 epileptics. Each patient was randomly selected to either be administered a new drug ($\texttt{trt}_i = 1$) or a placebo ($\texttt{trt}_i = 0$). The number of seizures in the 8 weeks before the trial period was recorded as $\texttt{base}_i$ as well as the age of the patient, recorded as $\texttt{age}_i$. Counts for the number of seizures were recorded during the two weeks before each clinical visit. Visits are recorded as $\texttt{visit}_1 = -3$, $\texttt{visit}_2 = -1$, $\texttt{visit}_3 = 1$ and $\texttt{visit}_4 = 3$. Finally, previous analyses (Thall & 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 (9) with $\log(\texttt{base}_i/4)$, $\texttt{trt}_i$, $\texttt{trt}_i \times \log(\texttt{base}_i/4)$, $\log(\texttt{age}_i)$ and $\mathcal{I}(j = 4)$ as covariates for $1 \leq i \leq 59$,

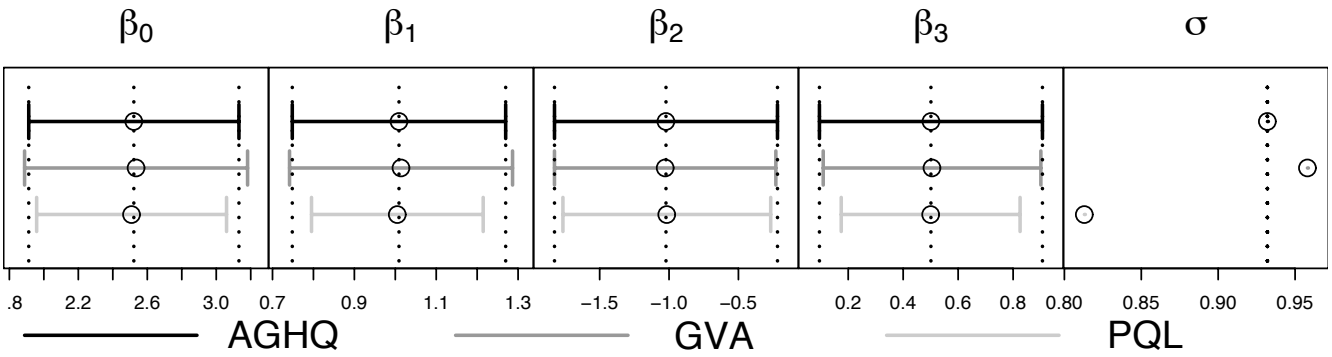

Figure 2: *Point estimates and approximate 95% confidence intervals based on each of AGHQ, GVA and PQL for the simulated logistic dataset. The vertical dotted lines correspond to the AGHQ values. Note that confidence intervals for σ are not available for AGHQ in* R*, so are not compared for this parameter.*

$1 \leq j \leq 4$; where $\mathcal{I}(\mathcal{P})$ is an indicator of $\mathcal{P}$ being true. This corresponds to Model II of Breslow & Clayton (1993). The parameter estimates and approximate standard errors for this model are summarized in Figure 3.

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

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

where

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

This corresponds to Model IV of Breslow & Clayton (1993). The parameter estimates and approximate standard errors for this model are summarized in Figure 4.

From Figures 3 and 4 we see that all of the approximation methods compare favourably with AGHQ, although the estimate of the variance component for PQL is slightly too small and the estimated correlation parameter ρ is too large.

7.3 Bacteria Dataset

The *bacteria* datasets records tests of the presence of the bacteria H. influenzae in children with a history of otitis media in Northern Territory, Australia (Leach, 2000). The children were

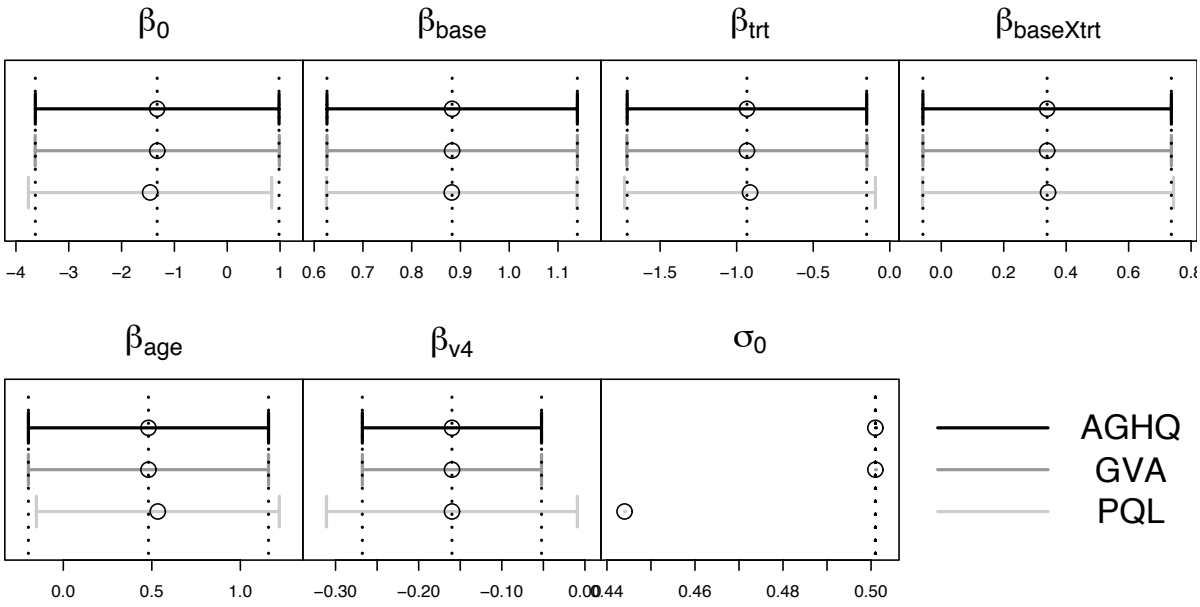

Figure 3: *Point estimates and approximate 95% confidence intervals based on each of AGHQ, GVA and PQL for the Epilepsy data random intercept model. The vertical dotted lines correspond to the AGHQ values. Note that confidence intervals for σ_0 are not available for AGHQ in* `R`*, so are not compared for this parameter.*

randomized into three groups: placebo, drug, and drug with active encouragement to comply. The presence of H. influenzae was checked at weeks 0, 2, 4, 6 and 11 to 30 and recorded as $\texttt{week}_{ij}$. If a particular check was missed no data was recorded for that week. High or low compliance of the patient in taking the treatment are indicated by the variables $\texttt{drugHi}_{ij}$ and $\texttt{drugLo}_{ij}$ respectively. The data are shown in Figure 1.

Using the *Bacteria* dataset we first considered the model

$$\text{logit}\left\{P(y_{ij}=1)\right\} = \beta_0 + \beta_{\text{drugLo}}\texttt{drugLo}_{ij} + \beta_{\text{drugHi}}\texttt{drugHi}_{ij} + \beta_{\text{week}}\texttt{week}_{ij} + u_i$$

where $u_i \sim N(0, \sigma_0^2)$ for $1 \leq i \leq 50$ and n_i takes values from 2 to 5. The parameter estimates and approximate standard errors for these models are summarized in Figure 5. Here we see that the estimates for $\boldsymbol{\beta}$ are quite similar for each method, although PQL underestimates σ_0. In addition, the approximate 95% confidence intervals for the βs are overly narrow for PQL, but close to exact for GVA.

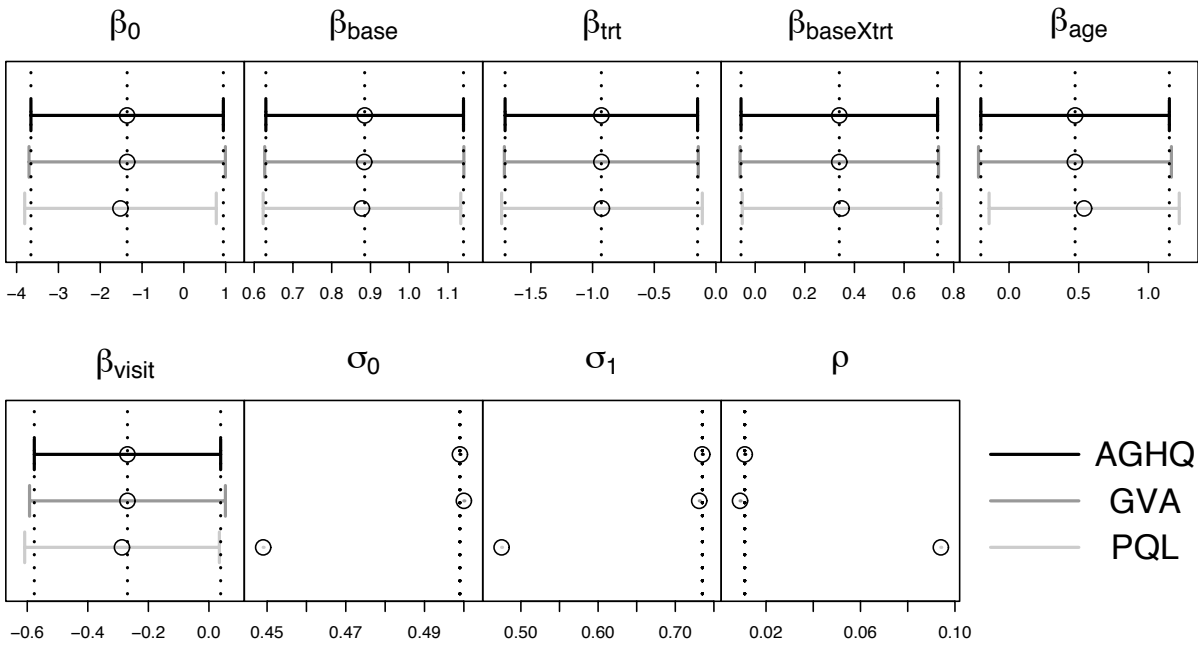

Figure 4: *Point estimates and approximate 95% confidence intervals based on each of AGHQ, GVA and PQL for the Epilepsy data random intercept and slope model. The vertical dotted lines correspond to the AGHQ values. Note that confidence intervals for* σ_0, σ_1 *and* ρ *are not available for AGHQ in* R*, so are not compared for these parameters.*

7.4 Toenail Dataset

This data set 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 trail was randomly either administered a treatment $\texttt{trt}_{ij} = 1$ or a placebo $\texttt{trt}_{ij} = 0$ and was evaluated at seven visits (approximately on weeks $\texttt{time}_{ij} = 0, 4, 8, 12, 24, 36$ 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 (9) with $\texttt{trt}_{ij}$, $\texttt{time}_{ij}$ and $\texttt{trt}_{ij} \times \texttt{time}_{ij}$ as covariates for $1 \leq i \leq 294$ and n_i taking values from 1 to 7. The results for each GLMM approximation is summarized in Figure 6. From this figure we see that the GVA estimates of

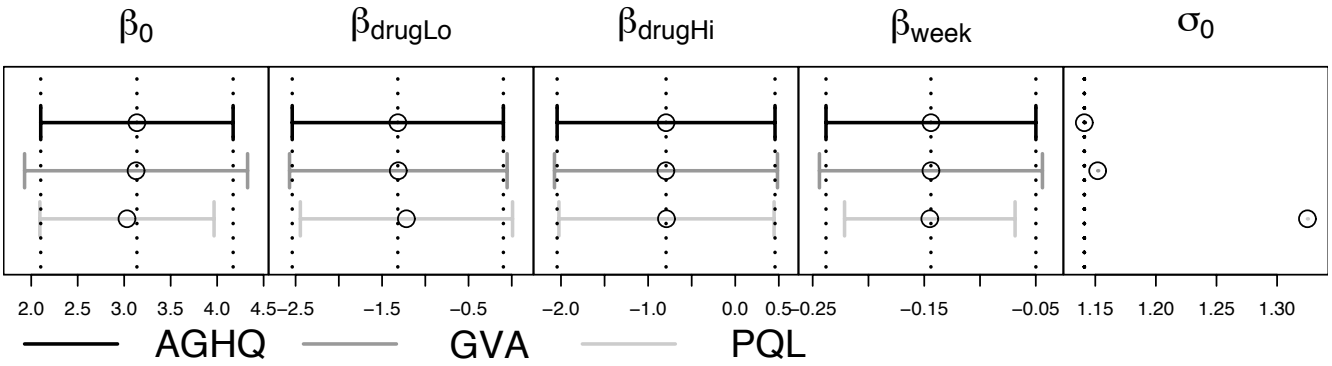

Figure 5: *Point estimates and approximate 95% confidence intervals based on each of AGHQ, GVA and PQL for the model fitted to the Bacteria dataset. The vertical dotted lines correspond to the AGHQ values. Note that confidence intervals for σ_0 are not available for AGHQ in* R*, so are not compared for this parameter.*

all parameters are better than those based on PQL. For this example we see severe bias of PQL for the parameters β_0, β_{time} and σ_0. The GVA method shows some bias for estimation of β_0 and σ_0 but is less severe than that of PQL.

7.5 Approximating $p(\boldsymbol{u}|\boldsymbol{y})$

The analysis of Lesaffre & Spiessens (2001) on the Toenail dataset found that parameter estimates could vary significantly even amongst several Gauss-Hermite quadrature approximations. Hence it is not surprising that there are large discrepancies between AGHQ, GVA and PQL approximations for this dataset. On the other hand PQL and GVA approximations for the bacteria data are, in comparison, reasonable. Both PQL and GVA methods are based on Gaussian approximations of $p(\boldsymbol{u}|\boldsymbol{y})$. In particular for fixed $\boldsymbol{\beta}$ and σ_0 PQL is equivalent to the Laplace approximation where $E(\boldsymbol{u}|\boldsymbol{y})$ is approximated by the mode of $p(\boldsymbol{y}, \boldsymbol{u})$ and the covariance is approximated by the inverse negative Hessian of $\log p(\boldsymbol{y}, \boldsymbol{u})$ with respect to $\boldsymbol{u}$. We will now illustrate that when $p(\boldsymbol{u}|\boldsymbol{y})$ is not Gaussian, as occurs for the Toenail dataset, then both PQL and GVA methods have reduced accuracy.

The standard definition of skewness for the univariate conditional density function $p(u_i|\boldsymbol{y})$

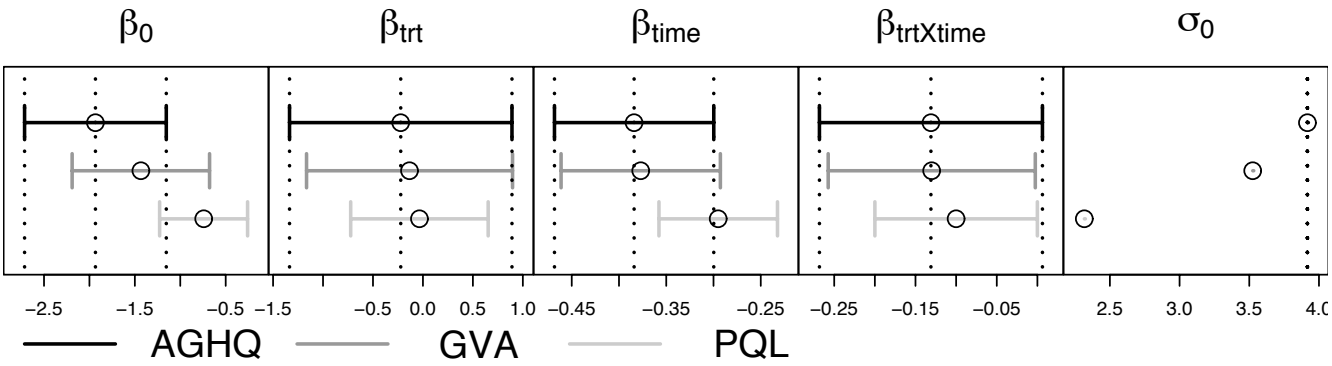

Figure 6: *Point estimates and approximate 95% confidence intervals based on each of AGHQ, GVA and PQL for the model fitted to the Toenail dataset. The vertical dotted lines correspond to the AGHQ values. Note that confidence intervals for σ_0 are not available for AGHQ in* R*, so are not compared for this parameter.*

is

$$\text{Skew}(u_i|\boldsymbol{y}) = \frac{\int_{-\infty}^{\infty}\{u_i - E(u_i|\boldsymbol{y})\}^3 p(u_i|\boldsymbol{y})du_i}{\left[\int_{-\infty}^{\infty}\{u_i - E(u_i|\boldsymbol{y})\}^2 p(u_i|\boldsymbol{y})du_i\right]^{3/2}}$$

where $E(u_i|\boldsymbol{y}) = \int_{-\infty}^{\infty} u_i p(u_i|\boldsymbol{y})du_i$. Each of the integrals is analytically intractable and so we approximate them using numerical quadrature.

Figure 7 illustrates the kernel density estimates of the approximate $\text{Skew}(u_i|\boldsymbol{y})$ values and density approximations of the $p(u_i|\boldsymbol{y})$ along with the $p(u_i|\boldsymbol{y})$ with largest absolute skewness for the Bacteria and Toenail datasets. We note that for the Bacteria dataset almost no skewness is evident for each of the $p(u_i|\boldsymbol{y})$s and that even the most skewed $p(u_i|\boldsymbol{y})$ is nearly Gaussian. However, for the Toenail dataset many of the $p(u_i|\boldsymbol{y})$s have large negative values $\text{Skew}(u_i|\boldsymbol{y})$ values and the $p(u_i|\boldsymbol{y})$s are clearly non-Gaussian for these cases.

Based on the rightmost panel in Figure 7 GVA appears to approximate the $E(u_i|\boldsymbol{y})$ better than the Laplace approximation. Let $\mu_{\text{Laplace},i}$ and $\mu_{\text{GVA},i}$ be the Laplace and GVA approximations of $E(u_i|\boldsymbol{y})$ respectively. For the *Bacteria* dataset

$$\frac{\sum_{i=1}^{m}(\mu_{\text{Laplace},i} - E(u_i|\boldsymbol{y}))^2}{\sum_{i=1}^{m}(\mu_{\text{GVA},i} - E(u_i|\boldsymbol{y}))^2} \approx 7787.8$$

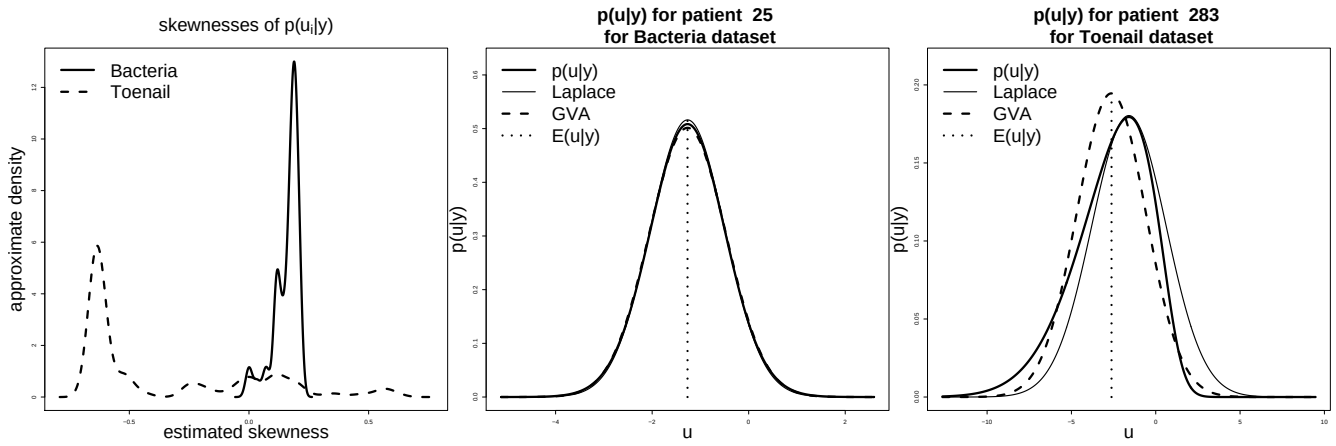

Figure 7: *Left panel: Kernel density estimates of approximated skewnesses for each $p(u_i|\boldsymbol{y})$ for the Bacteria and Toenail datasets. Middle panel: Approximations of the most skewed $p(u_i|\boldsymbol{y})$ for the Bacteria dataset. Right panel: Approximations of the most skewed $p(u_i|\boldsymbol{y})$ for the Toenail dataset.*

and for the *Toenail* dataset

$$\frac{\sum_{i=1}^{m}(\mu_{\text{Laplace},i} - E(u_i|\boldsymbol{y}))^2}{\sum_{i=1}^{m}(\mu_{\text{GVA},i} - E(u_i|\boldsymbol{y}))^2} \approx 3029.3$$

and so we speculate that GVA is a better approximation of $E(u_i|\boldsymbol{y})$ than the Laplace approximation since it attempts to approximate $E(\boldsymbol{u}|\boldsymbol{y})$ more directly. While we do not have theoretical evidence to support this we believe that this is a matter for further investigation.

8 Concluding Remarks

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. One example is model selection in the presence of a high number of candidate predictors. In such a situation, it is often too expensive to fit several models via MCMC. Gaussian variational approximation offers itself as an attractive alternative to PQL for fast GLMM approximate inference. In this paper we have focussed on the grouped data version of GLMMs and mostly resolved issues regarding their implementation. 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.

Appendix A: Computational Details

A.1 Notation Useful for Derivative and Hessian Expressions

Let f be a real-valued function in the $d \times 1$ vector $\boldsymbol{x} = [x_1, \ldots, x_d]^T$. Then the derivative vector $\mathsf{D}_{\boldsymbol{x}} f(\boldsymbol{x})$, is the $1 \times d$ with ith entry $\partial f(\boldsymbol{x})/\partial x_i$. The corresponding Hessian matrix is given by $\mathsf{H}_{\boldsymbol{x}} f(\boldsymbol{x}) = \mathsf{D}_{\boldsymbol{x}}\{\mathsf{D}_{\boldsymbol{x}} f(\boldsymbol{x})\}^T$.

We extend the B notation to higher derivatives as follows:

$$B^{(r)}(\mu, \sigma^2) \equiv \int_{-\infty}^{\infty} b^{(r)}(\sigma x + \mu)\phi(x)\, dx.$$

Define

$$\mathcal{Q}(\boldsymbol{A}) \equiv (\boldsymbol{A} \otimes \boldsymbol{1}^T) \odot (\boldsymbol{1}^T \otimes \boldsymbol{A})$$

where $\boldsymbol{A} \odot \boldsymbol{B}$ is the element-wise product of two equi-sized matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. Next, we let $\boldsymbol{D}_p$ denote the *duplication matrix* of order p. This matrix is defined through the relationship

$$\text{vec}(\boldsymbol{A}) = \boldsymbol{D}_p \text{vech}(\boldsymbol{A})$$

for a symmetric $p \times p$ matrix $\boldsymbol{A}$. Lastly, for each $1 \leq i \leq m$, let

$$\mathcal{B}^{(r)}(\boldsymbol{\beta}, \boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i) \equiv B^{(r)}(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\mu}_i, \text{diagonal}(\boldsymbol{Z}_i\boldsymbol{\Lambda}_i\boldsymbol{Z}_i^T)).$$

A.2 Gauss-Hermite Quadrature for Evaluation of $B^{(r)}$

Here we briefly describe numerical evaluation of $B^{(r)}$ using the adaptive Gauss-Hermite quadrature procedure developed by Liu & Pierce (1994) for approximating positive Gaussian integrals.

For $r = 1, 2, \ldots$ we have

$$\begin{aligned} B^{(r)}(\mu, \sigma) &= \int_{-\infty}^{\infty} \frac{b^{(r)}(\mu+\sigma x)\phi(x)}{\phi_{\sigma^*}(x-\mu^*)} \phi_{\sigma^*}(x - \mu^*) dx \\ &= \int_{-\infty}^{\infty} \left[\sqrt{2}\sigma^* b^{(r)}(\mu + \sigma(\mu^* + \sqrt{2}\sigma^* x))\phi(\mu^* + \sqrt{2}\sigma^* x) e^{x^2}\right] e^{-x^2} dx \end{aligned}$$

for any μ^* and σ^*. We choose μ^* and σ^* so that the integrand of $B(\mu, \sigma)$ is "most Gaussian" in shape so that

$$\begin{aligned} \mu^* &= \text{argmax}_x \left\{ b(\mu + \sigma x)\phi(x) \right\} \\ \text{and} \quad \sigma^* &= -\left\{ \left[\tfrac{d^2}{dx^2} \log \left\{ b(\mu + \sigma x)\phi(x) \right\} \right]_{x=\mu^*} \right\}^{-1/2} \end{aligned}$$

We use the values of μ^* and σ^* corresponding to $r = 0$ because it is both computationally cheaper to evaluate μ^* and σ^* once and because $b^{(r)}$ may not be positive everywhere (potentially making the corresponding evaluation of σ^* problematic). We then apply Gauss-Hermite quadrature which uses

$$\int_{-\infty}^{\infty} g(x)e^{-x^2}dx = \sum_{k=1}^{N} w_k g(x_k) + \frac{\sqrt{\pi}N!}{2^N(2N)!} g^{(2N)}(\xi), \quad \text{for some} \quad -\infty \leq \xi \leq \infty.$$

for some integer N and is exact for polynomials of degree less than $2N$. Hence we may approximate $B^{(r)}(\mu, \sigma)$ by

$$B^{(r)}(\mu, \sigma) \approx \sum_{k=1}^{N} w_k^* b^{(r)}(\mu + \sigma x_k^*)\phi(x_k^*)$$

where $w_k^* = \sqrt{2}\sigma^* w_k e^{x_k^2}$ and $x_k^* = \mu^* + \sqrt{2}\sigma^* x_k$ and the w_k and x_k values are the weights and abscissae of standard (or non-adaptive) Gauss-Hermite quadrature respectively.

There are several ways of obtaining w_j and x_j in practice. Tables for these values can be obtained form Abramowitz & Stegun (1972, Chapter 25). Computational details for calculating these values be found form Golub & Welsch (1969) or Press, Teukolsky, Vetterling & Flannery (2007). In the `R` package `statmod` (Smyth, 2009) the function `gauss.quad()` implements the method outlined in Golub & Welsch (1969) and may also be used to find w_j and x_j.

Finally, it is worth noting the number of quadrature points needed to calculate $\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ and $\underline{\ell}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ using adaptive Gauss-Hermite quadrature. Suppose that N points are used in each dimension via a tensor product method then the calculation $\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ and its derivatives uses $O(mN^K)$ points. In comparison the calculation of $\underline{\ell}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ and its derivative uses $O(N \sum_{i=1} n_i)$ points if we use N quadrature points to evaluate the $B^{(r)}$s. Since the number of quadrature points for GVA is independent of K the relative computational efficiency of GVA over adaptive Gauss-Hermite quadrature can be substantial when K is large.

A.3 Derivative Vector of Lower Bound on Log-Likelihood

The derivative vector of

$$\underline{\ell} \equiv \ell(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$$

with respect to

$$(\boldsymbol{\beta}, \text{vech}(\boldsymbol{\Sigma}), \boldsymbol{\mu}_1, \text{vech}(\boldsymbol{\Lambda}_1), \ldots, \boldsymbol{\mu}_m, \text{vech}(\boldsymbol{\Lambda}_m))$$

is

$$\mathsf{D}\,\underline{\ell} = [\mathsf{D}_{\boldsymbol{\beta}}\,\underline{\ell},\ \mathsf{D}_{\text{vech}(\boldsymbol{\Sigma})}\,\underline{\ell},\ \mathsf{D}_{\boldsymbol{\mu}_1}\,\underline{\ell},\ \mathsf{D}_{\text{vech}(\boldsymbol{\Lambda}_1)}\,\underline{\ell}, \cdots,\ \mathsf{D}_{\boldsymbol{\mu}_m}\,\underline{\ell},\ \mathsf{D}_{\text{vech}(\boldsymbol{\Lambda}_m)}\,\underline{\ell}].$$

We now give matrix algebraic expressions for each of these components:

$$\begin{aligned}
\mathsf{D}_{\boldsymbol{\beta}}\,\underline{\ell} &= \sum_{i=1}^{m} \{\boldsymbol{y}_i - \mathcal{B}^{(1)}(\boldsymbol{\beta}, \boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)\}^T \boldsymbol{X}_i, \\
\mathsf{D}_{\text{vech}(\boldsymbol{\Sigma})}\,\underline{\ell} &= \tfrac{1}{2} \sum_{i=1}^{m} \text{vec}\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i \boldsymbol{\mu}_i^T + \boldsymbol{\Lambda}_i)\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\}^T \boldsymbol{D}_K,
\end{aligned}$$

and, for $1 \leq i \leq m$,

$$\begin{aligned}
\mathsf{D}_{\boldsymbol{\mu}_i}\,\underline{\ell} &= \{\boldsymbol{y}_i - \mathcal{B}^{(1)}(\boldsymbol{\beta}, \boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)\}^T \boldsymbol{Z}_i - \boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1}, \\
\mathsf{D}_{\text{vech}(\boldsymbol{\Lambda}_i)}\,\underline{\ell} &= \tfrac{1}{2} \text{vec}[\boldsymbol{\Lambda}_i^{-1} - \boldsymbol{\Sigma}^{-1} - \boldsymbol{Z}_i^T \text{diag}\{\mathcal{B}^{(2)}(\boldsymbol{\beta}, \boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)\}\boldsymbol{Z}_i]^T \boldsymbol{D}_K.
\end{aligned}$$

A.4 Hessian Matrix of Lower Bound on Log-Likelihood

The Hessian matrix of

$$\underline{\ell} \equiv \ell(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$$

with respect to

$$(\boldsymbol{\beta}, \text{vech}(\boldsymbol{\Sigma}), \boldsymbol{\mu}_1, \text{vech}(\boldsymbol{\Lambda}_1), \ldots, \boldsymbol{\mu}_m, \text{vech}(\boldsymbol{\Lambda}_m))$$

is

$$\mathsf{H}\,\underline{\ell} =$$

$$\begin{bmatrix}
\mathsf{H}_{\boldsymbol{\beta}\boldsymbol{\beta}}\,\underline{\ell} & \mathbf{0} & \mathsf{H}_{\boldsymbol{\beta}\boldsymbol{\mu}_1}\,\underline{\ell} & \mathsf{H}_{\boldsymbol{\beta}\mathrm{vech}(\boldsymbol{\Lambda}_1)}\,\underline{\ell} & \cdots & \mathsf{H}_{\boldsymbol{\beta}\boldsymbol{\mu}_m}\,\underline{\ell} & \mathsf{H}_{\boldsymbol{\beta}\mathrm{vech}(\boldsymbol{\Lambda}_m)}\,\underline{\ell} \\
\mathbf{0} & \mathsf{H}_{\mathrm{vech}(\boldsymbol{\Sigma})\mathrm{vech}(\boldsymbol{\Sigma})}\,\underline{\ell} & \mathsf{H}_{\mathrm{vech}(\boldsymbol{\Sigma})\boldsymbol{\mu}_1}\,\underline{\ell} & \mathsf{H}_{\mathrm{vech}(\boldsymbol{\Sigma})\mathrm{vech}(\boldsymbol{\Lambda}_1)}\,\underline{\ell} & \cdots & \mathsf{H}_{\mathrm{vech}(\boldsymbol{\Sigma})\boldsymbol{\mu}_m}\,\underline{\ell} & \mathsf{H}_{\mathrm{vech}(\boldsymbol{\Sigma})\mathrm{vech}(\boldsymbol{\Lambda}_m)}\,\underline{\ell} \\
\mathsf{H}_{\boldsymbol{\mu}_1\boldsymbol{\beta}}\,\underline{\ell} & \mathsf{H}_{\boldsymbol{\mu}_1\mathrm{vech}(\boldsymbol{\Sigma})}\,\underline{\ell} & \mathsf{H}_{\boldsymbol{\mu}_1\boldsymbol{\mu}_1}\,\underline{\ell} & \mathsf{H}_{\boldsymbol{\mu}_1\mathrm{vech}(\boldsymbol{\Lambda}_1)}\,\underline{\ell} & \cdots & \mathbf{0} & \mathbf{0} \\
\mathsf{H}_{\mathrm{vech}(\boldsymbol{\Lambda}_1)\boldsymbol{\beta}}\,\underline{\ell} & \mathsf{H}_{\mathrm{vech}(\boldsymbol{\Lambda}_1)\mathrm{vech}(\boldsymbol{\Sigma})}\,\underline{\ell} & \mathsf{H}_{\mathrm{vech}(\boldsymbol{\Lambda}_1)\boldsymbol{\mu}_1}\,\underline{\ell} & \mathsf{H}_{\mathrm{vech}(\boldsymbol{\Lambda}_1)\mathrm{vech}(\boldsymbol{\Lambda}_1)}\,\underline{\ell} & \cdots & \mathbf{0} & \mathbf{0} \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
\mathsf{H}_{\boldsymbol{\mu}_m\boldsymbol{\beta}}\,\underline{\ell} & \mathsf{H}_{\boldsymbol{\mu}_m\mathrm{vech}(\boldsymbol{\Sigma})}\,\underline{\ell} & \mathbf{0} & \mathbf{0} & \cdots & \mathsf{H}_{\boldsymbol{\mu}_m\boldsymbol{\mu}_m}\,\underline{\ell} & \mathsf{H}_{\boldsymbol{\mu}_m\mathrm{vech}(\boldsymbol{\Lambda}_m)}\,\underline{\ell} \\
\mathsf{H}_{\mathrm{vech}(\boldsymbol{\Lambda}_m)\boldsymbol{\beta}}\,\underline{\ell} & \mathsf{H}_{\mathrm{vech}(\boldsymbol{\Lambda}_m)\mathrm{vech}(\boldsymbol{\Sigma})}\,\underline{\ell} & \mathbf{0} & \mathbf{0} & \cdots & \mathsf{H}_{\mathrm{vech}(\boldsymbol{\Lambda}_m)\boldsymbol{\mu}_m}\,\underline{\ell} & \mathsf{H}_{\mathrm{vech}(\boldsymbol{\Lambda}_m)\mathrm{vech}(\boldsymbol{\Lambda}_m)}\,\underline{\ell}
\end{bmatrix}$$

where

$$\begin{aligned}
\mathsf{H}_{\boldsymbol{\beta}\boldsymbol{\beta}}\underline{\ell} &= -\sum_{i=1}^{m} \boldsymbol{X}_i^T \mathrm{diag}\{\mathcal{B}^{(2)}(\boldsymbol{\beta}, \boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)\}\boldsymbol{X}_i, \\
\mathsf{H}_{\mathrm{vech}(\boldsymbol{\Sigma})\mathrm{vech}(\boldsymbol{\Sigma})}\underline{\ell} &= \tfrac{1}{2}\boldsymbol{D}_K^T\Big(m(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}) - \sum_{i=1}^{m}[\boldsymbol{\Sigma}^{-1}\otimes\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i\boldsymbol{\mu}_i^T+\boldsymbol{\Lambda}_i)\boldsymbol{\Sigma}^{-1}\} \\
&\qquad +\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i\boldsymbol{\mu}_i^T+\boldsymbol{\Lambda}_i)\boldsymbol{\Sigma}^{-1}\}\otimes\boldsymbol{\Sigma}^{-1}]\Big)\boldsymbol{D}_K,
\end{aligned}$$

and, for $1 \le i \le m$,

$$\begin{aligned}
\mathsf{H}_{\boldsymbol{\beta}\boldsymbol{\mu}_i}\underline{\ell} &= -\boldsymbol{X}_i^T \mathrm{diag}\{\mathcal{B}^{(2)}(\boldsymbol{\beta}, \boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)\}\boldsymbol{Z}_i, \\
\mathsf{H}_{\boldsymbol{\beta}\mathrm{vech}(\boldsymbol{\Lambda}_i)}\,\underline{\ell} &= -\tfrac{1}{2}\boldsymbol{X}_i^T \mathrm{diag}\{\mathcal{B}^{(3)}(\boldsymbol{\beta}, \boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)\}\mathcal{Q}(\boldsymbol{Z}_i)\boldsymbol{D}_K, \\
\mathsf{H}_{\mathrm{vech}(\boldsymbol{\Sigma})\boldsymbol{\mu}_i}\,\underline{\ell} &= \boldsymbol{D}_K^T\{(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i)\otimes\boldsymbol{\Sigma}^{-1}\}, \\
\mathsf{H}_{\mathrm{vech}(\boldsymbol{\Sigma})\mathrm{vech}(\boldsymbol{\Lambda}_i)}\,\underline{\ell} &= -\tfrac{1}{2}\boldsymbol{D}_K^T(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})\boldsymbol{D}_K, \\
\mathsf{H}_{\boldsymbol{\mu}_i\boldsymbol{\mu}_i}\,\underline{\ell} &= -\boldsymbol{Z}_i^T \mathrm{diag}\{\mathcal{B}^{(2)}(\boldsymbol{\beta}, \boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)\}\boldsymbol{Z}_i - \boldsymbol{\Sigma}^{-1}, \\
\mathsf{H}_{\boldsymbol{\mu}_i\mathrm{vech}(\boldsymbol{\Lambda}_i)}\,\underline{\ell} &= -\tfrac{1}{2}\boldsymbol{Z}_i^T \mathrm{diag}\{\mathcal{B}^{(3)}(\boldsymbol{\beta}, \boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)\}\mathcal{Q}(\boldsymbol{Z}_i)\boldsymbol{D}_K, \\
\mathsf{H}_{\mathrm{vech}(\boldsymbol{\Lambda}_i)\mathrm{vech}(\boldsymbol{\Lambda}_i)}\,\underline{\ell} &= -\tfrac{1}{4}\boldsymbol{D}_K^T[\mathcal{Q}(\boldsymbol{Z}_i)^T \mathrm{diag}\{\mathcal{B}^{(4)}(\boldsymbol{\beta}, \boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)\}\mathcal{Q}(\boldsymbol{Z}_i) + 2(\boldsymbol{\Lambda}_i^{-1}\otimes\boldsymbol{\Lambda}_i^{-1})]\boldsymbol{D}_K.
\end{aligned}$$

A.5 Newton-Raphson Scheme

We are a now in a position to describe an efficient Newton-Raphson scheme for solving the maximization problem (5). In particular, we make use of the block-diagonal structure in the $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ section of $\mathsf{H}_{\boldsymbol{\beta}\boldsymbol{\beta}}\underline{\ell}$ to reduce the number of operations to $O(m)$.

Recall the notation given in (6) for the $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$. In keeping with this notation, define the gradient vectors

$$\boldsymbol{g}_{\boldsymbol{\theta}} \equiv \begin{bmatrix} (\mathsf{D}_{\boldsymbol{\beta}}\, \underline{\ell})^T \\ \{\mathsf{D}_{\mathrm{vech}(\boldsymbol{\Sigma})}\, \underline{\ell}\}^T \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{g}_{\boldsymbol{\xi}_i} \equiv \begin{bmatrix} (\mathsf{D}_{\boldsymbol{\mu}_i}\, \underline{\ell})^T \\ \{\mathsf{D}_{\mathrm{vech}(\boldsymbol{\Lambda}_i)}\, \underline{\ell}\}^T \end{bmatrix}$$

and the Hessian matrix components

$$\boldsymbol{H}_{\boldsymbol{\theta}\boldsymbol{\theta}} \equiv \begin{bmatrix} \mathsf{H}_{\boldsymbol{\theta}\boldsymbol{\theta}}\, \underline{\ell} & \mathbf{0} \\ \mathbf{0} & \mathsf{H}_{\mathrm{vech}(\boldsymbol{\Sigma})\mathrm{vech}(\boldsymbol{\Sigma})}\, \underline{\ell} \end{bmatrix}, \qquad \boldsymbol{H}_{\boldsymbol{\xi}_i\boldsymbol{\xi}_i} \equiv \begin{bmatrix} \mathsf{H}_{\boldsymbol{\mu}_i\boldsymbol{\mu}_i}\, \underline{\ell} & \mathsf{H}_{\boldsymbol{\mu}_i\mathrm{vech}(\boldsymbol{\Lambda}_i)}\, \underline{\ell} \\ \mathsf{H}_{\mathrm{vech}(\boldsymbol{\Lambda}_i)\boldsymbol{\mu}_i}\, \underline{\ell} & \mathsf{H}_{\mathrm{vech}(\boldsymbol{\Lambda}_i)\mathrm{vech}(\boldsymbol{\Lambda}_i)}\, \underline{\ell} \end{bmatrix},$$

and

$$\boldsymbol{H}_{\boldsymbol{\theta}\boldsymbol{\xi}_i} \equiv \begin{bmatrix} \mathsf{H}_{\boldsymbol{\beta}\boldsymbol{\mu}_i}\, \underline{\ell} & \mathsf{H}_{\boldsymbol{\beta}\mathrm{vech}(\boldsymbol{\Lambda}_i)}\, \underline{\ell} \\ \mathsf{H}_{\mathrm{vech}(\boldsymbol{\Sigma})\boldsymbol{\mu}_i}\, \underline{\ell} & \mathsf{H}_{\mathrm{vech}(\boldsymbol{\Sigma})\mathrm{vech}(\boldsymbol{\Lambda}_i)}\, \underline{\ell} \end{bmatrix}$$

for $1 \leq i \leq m$. Finally, define

$$\boldsymbol{s}_{\boldsymbol{\theta}\boldsymbol{\xi}} \equiv \left(\boldsymbol{H}_{\boldsymbol{\theta}\boldsymbol{\theta}} - \sum_{i=1}^{m} \boldsymbol{H}_{\boldsymbol{\theta}\boldsymbol{\xi}_i} \boldsymbol{H}_{\boldsymbol{\xi}_i\boldsymbol{\xi}_i}^{-1} \boldsymbol{H}_{\boldsymbol{\theta}\boldsymbol{\xi}_i}^{T} \right)^{-1} \left(\boldsymbol{g}_{\boldsymbol{\theta}} - \sum_{i=1}^{m} \boldsymbol{H}_{\boldsymbol{\theta}\boldsymbol{\xi}_i} \boldsymbol{H}_{\boldsymbol{\xi}_i\boldsymbol{\xi}_i}^{-1} \boldsymbol{g}_{\boldsymbol{\xi}} \right).$$

Let

$$\begin{bmatrix} \boldsymbol{\beta} \\ \mathrm{vech}(\boldsymbol{\Sigma}) \end{bmatrix}^{(0)} \quad \text{and} \quad \begin{bmatrix} \boldsymbol{\xi}_i \\ \mathrm{vech}(\boldsymbol{\Lambda}_i) \end{bmatrix}^{(0)}, \quad 1 \leq i \leq m,$$

be starting values of the relevant parameter vectors and let a superscript of (t) denote the same vectors after t iterations of the Newton-Raphson algorithm. Then the updates are given by:

$$\begin{bmatrix} \boldsymbol{\beta} \\ \mathrm{vech}(\boldsymbol{\Sigma}) \end{bmatrix}^{(t+1)} = \begin{bmatrix} \boldsymbol{\beta} \\ \mathrm{vech}(\boldsymbol{\Sigma}) \end{bmatrix}^{(t)} - \boldsymbol{s}_{\boldsymbol{\theta}\boldsymbol{\xi}}^{(t)} \tag{10}$$

and

$$\begin{bmatrix} \boldsymbol{\xi}_i \\ \mathrm{vech}(\boldsymbol{\Lambda}_i) \end{bmatrix}^{(t+1)} = \begin{bmatrix} \boldsymbol{\xi}_i \\ \mathrm{vech}(\boldsymbol{\Lambda}_i) \end{bmatrix}^{(t)} - (\boldsymbol{H}_{\boldsymbol{\xi}_i\boldsymbol{\xi}_i}^{(t)})^{-1} (\boldsymbol{g}_{\boldsymbol{\xi}_i}^{(t)} - \boldsymbol{H}_{\boldsymbol{\xi}_i\boldsymbol{\theta}}^{(t)}\, \boldsymbol{s}_{\boldsymbol{\theta}\boldsymbol{\xi}}^{(t)}) \tag{11}$$

Note that these updates involves inversion of 'small' matrices – i.e. those of dimension similar to $\boldsymbol{\beta}$ and the $\boldsymbol{u}_i$. The Newton-Raphson scheme then involves cycling through (10) and (11) until convergence.

To increase the robustness of the Newton-Raphson algorithm we: incorporated step-halving to ensure that $\boldsymbol{\Sigma}$ and $\boldsymbol{\Lambda}_i$s remained positive definite and that the approximate likelihood increased at each step of the algorithm. Laplace approximation was used to choose, for fixed $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, the initial values for $\boldsymbol{\mu}_i$s and $\boldsymbol{\Lambda}_i$s.

A.6 Asymptotic Covariance Matrix

Results used in Section A.5 that take advantage of diagonal structure in $\mathsf{H}\underline{\ell}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ can also be used to obtain a streamlined expression for the asymptotic covariance matrix of the model parameters. These lead to

$$\widehat{\text{Asy. Cov}}(\widehat{\boldsymbol{\theta}}) = \left(\boldsymbol{H}_{\boldsymbol{\theta}\boldsymbol{\theta}} - \sum_{i=1}^{m} \boldsymbol{H}_{\boldsymbol{\theta}\boldsymbol{\xi}_i} \boldsymbol{H}_{\boldsymbol{\xi}_i\boldsymbol{\xi}_i}^{-1} \boldsymbol{H}_{\boldsymbol{\theta}\boldsymbol{\xi}_i}^{T} \right)^{-1}$$

and allow standard errors to be computed with $O(m)$ operations.

Appendix B: Proof of Theorem 1

The proof relies on straightforward algebra and the following lemmas:

LEMMA 1: *Let $\boldsymbol{A}$ be a positive definite matrix and $\boldsymbol{B}$ be a positive semidefinite matrix of the same dimensions as $\boldsymbol{A}$. Then*

$$|\boldsymbol{A} + \boldsymbol{B}| \geq |\boldsymbol{A}|$$

with equality if and only if $\boldsymbol{B} = \boldsymbol{0}$.

Proof of Lemma 1.

Lemma 1 corresponds to Theorem 22 of Magnus & Neudecker (1988) and a prove is given there.

□

LEMMA 2. Let the symmetric matrix $\boldsymbol{A}$ be partitioned as

$$\begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{12}^{T} & \boldsymbol{A}_{22} \end{bmatrix}$$

where $\boldsymbol{A}_{11}$ is square and invertible. Then $\boldsymbol{A}$ is positive definite if any only if both $\boldsymbol{A}_{11}$ and $\boldsymbol{A}_{22} - \boldsymbol{A}_{12}^{T}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}$ are positive definite.

Proof of Lemma 2.

Lemma 2 is a special case of Theorem 7.7.6 of Horn & Johnson (1985).

□

LEMMA 3. *Let $\boldsymbol{\Psi}$ be a symmetric, positive definite $mK \times mK$ matrix. Given the $K \times K$ blocks $\boldsymbol{\Psi}_i$, $1 \leq i \leq m$, down the main diagonal of $\boldsymbol{\Psi}$, the determinant $|\boldsymbol{\Psi}|$ is uniquely maximized by setting all other entries of $\boldsymbol{\Psi}$ equal to zero.*

Proof of Lemma 3.

Consider the partition of $\boldsymbol{\Psi}$:

$$\boldsymbol{\Psi} = \begin{bmatrix} \widetilde{\boldsymbol{\Psi}} & \boldsymbol{C} \\ \boldsymbol{C}^T & \boldsymbol{D} \end{bmatrix}.$$

where $\widetilde{\boldsymbol{\Psi}}$ is an $m(K-1) \times m(K-1)$ matrix, $\boldsymbol{C}$ is an $m(K-1) \times K$ matrix and $\boldsymbol{D}$ is a $K \times K$ matrix. Then, from a standard result on determinants of partitioned matrices,

$$|\boldsymbol{\Psi}| = |\widetilde{\boldsymbol{\Psi}}||\boldsymbol{D} - \boldsymbol{C}^T \widetilde{\boldsymbol{\Psi}}^{-1} \boldsymbol{C}|.$$

Since $\boldsymbol{\Psi}$ is positive definite then, from Lemma 2, the matrices $\widetilde{\boldsymbol{\Psi}}$, $\boldsymbol{D} - \boldsymbol{C}^T \widetilde{\boldsymbol{\Psi}}^{-1} \boldsymbol{C}$ must also be positive definite. Also $\boldsymbol{\Psi}^{-1}$ is positive definite, which implies that $\boldsymbol{C}^T \widetilde{\boldsymbol{\Psi}}^{-1} \boldsymbol{C}$ is positive semidefinite.

We shall prove the lemma by induction over m. The lemma holds trivially when $m = 1$. By the induction hypothesis, we may assume that $|\widetilde{\boldsymbol{\Psi}}|$ is uniquely maximized by taking the off block-diagonal components of $\widetilde{\boldsymbol{\Psi}}$ to vanish. For $\widetilde{\boldsymbol{\Psi}}$ and $\boldsymbol{D}$ fixed, we can use Lemma 1 with $\boldsymbol{A} = \boldsymbol{D} - \boldsymbol{C}^T \widetilde{\boldsymbol{\Psi}}^{-1} \boldsymbol{C}$ and $\boldsymbol{B} = \boldsymbol{C}^T \widetilde{\boldsymbol{\Psi}}^{-1} \boldsymbol{C}$ to show that $|\boldsymbol{D} - \boldsymbol{C}^T \widetilde{\boldsymbol{\Psi}} \boldsymbol{C}|$ is uniquely maximized by taking $\boldsymbol{C} = \boldsymbol{0}$. The lemma then follows by induction.

□

The right-hand side of (8), with q set to the $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ density and $p(\boldsymbol{y}|\boldsymbol{u})$ given by (2), is

$$\begin{aligned} \underline{\ell}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \;\; = \;\; & \tfrac{mK}{2} + \boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\mu}) - \boldsymbol{1}^T B(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\mu}, \text{diagonal}(\boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{Z}^T)) + \boldsymbol{1}^T c(\boldsymbol{y}) \\ & -\tfrac{1}{2}\{\boldsymbol{\mu}^T \boldsymbol{G}^{-1} \boldsymbol{\mu} + \text{tr}(\boldsymbol{G}^{-1}\boldsymbol{\Lambda})\} + \tfrac{1}{2} \log |\boldsymbol{G}^{-1}\boldsymbol{\Lambda}|. \end{aligned}$$

Now consider the special case of the grouped data GLMM (1). Applying the definitions of $\boldsymbol{y}_i$,

$\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ given in Section 2 and setting $\boldsymbol{Z} = \text{blockdiag}_{1\le i\le m}(\boldsymbol{Z}_i)$ and $\boldsymbol{G} = \boldsymbol{I} \otimes \boldsymbol{\Sigma}$ we obtain

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

By Lemma 3, $|\boldsymbol{\Lambda}|$ is maximal for $\boldsymbol{\Lambda} = \text{blockdiag}_{1\le i\le m}(\boldsymbol{\Lambda}_i)$. Hence there is no loss from replacement of $\frac{1}{2}\log|\boldsymbol{\Lambda}|$ by $\frac{1}{2}\sum_{i=1}^m \log(\boldsymbol{\Lambda}_i)$, and this leads to the expression for $\underline{\ell}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ given at (4).

Acknowledgement

This research was partially supported by Australian Research Council Discovery Project DP0877055.

References

Abramowitz, M. & Stegun, I. (1972). *Handbook of Mathematical Functions, with Formulas, Graphs, and Mathematical Tables.* New York: Dover Publications.

Archambeau, C., Cornford, D., Opper, M. & Shawe-Taylor, J. (2007). Gaussian process approximations of stochastic differential equations. *Journal of Machine Learning Research: Workshop and Conference Proceedings*, **1**, 1–16.

Bates, D. & Maechler, M. (2009). `lme4 0.999375`. Linear mixed-effects models using S4 classes. R package. `http://cran.r-project.org`.

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

Barber, D. & Bishop, C. M. (1998) Ensemble learning for multi-layer networks. In Jordan, M. I. Kearns, K. J. and Solla, S. A. (Eds.), Advances in Neural Information Processing Systems, **10**, 395-401.

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

Carlin,B. P. & Louis, T.A. (2008). *Bayes and Empirical Bayes Methods for Data Analysis (Third Edition).* New York: Chapman and Hall.

De Backer, M., De Vroey, C., Lesaffre, E., Scheys, I. & 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)**, S57–63.

Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (2004). *Bayesian Data Analysis*. Boca Raton, Florida: Chapman and Hall.

Golub, G.H. & Welsch, J.H. (1969). Calculation of Gauss quadrature rules. *Mathematics of Computation*, **23**, 221–230.

Hall, P., Ormerod, J.T. & Wand, M.P. (2009). Theory of Gaussian variational approximation for a generalised linear mixed model. Submitted to *Statistica Sinica*.

Honkela, A. & Valpola, H. (2005). Unsupervised variational Bayesian learning of nonlinear models. *Advances in Neural Information Processing Systems 17*, 593–600.

Horn, R.A. & Johnson, C.R. (1985). *Matrix Analysis*, Cambridge, UK:Cambridge University Press.

Jaakkola, T.S. & Jordan, M.I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, **10**, 25–37.

Jordan, M.I. (2004). Graphical models. *Statistical Science*, **19**, 140–155.

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

Kullback, S. & Leibler, R.A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, **22**, 79–86.

Leach, A. (2000). Menzies School of Health Research 1999-2000 Annual Report, pp. 18–21.

Lee, Y. & Nelder, J.A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society, Series B*, **58**, 619–656.

Lesaffre, E. & 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.

Little, R.J. & Rubin, D.B. (2002). *Statistical Analysis with Missing Data, Second Edition*. New York: John Wiley & Sons.

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

Magnus, J.R. & Neudecker, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics.* Chichester: John Wiley & Sons.

Marin, J.-M. & Robert, C.P. (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*, New York: Springer.

McCulloch, C.E., Searle, S.R. & Neuhaus, J.M. (2008). *Generalized, Linear, and Mixed Models, Second Edition*. New York: John Wiley & Sons.

Opper, M. & Archambeau, C. (2009). Variational Gaussian approximation revisited. Unpublished manuscript.

Ormerod, J.T. (2008). *On Semiparametric Regression and Data Mining.* PhD Thesis. School of Mathematics and Statistics, The University of New South Wales, Sydney, Australia.

Ormerod, J.T. & Wand, M.P. (2008). Variational approximations for logistic mixed models. *Proceedings of the Ninth Iranian Statistics Conference, Department of Statistics, University of Isfahan, Isfahan, Iran*, pp. 450–467.

Ormerod, J.T. & Wand, M.P. (2009). Explaining variational approximation. Submitted to *The American Statistician*.

Parisi, G. (1988). *Statistical Field Theory*. Redwood City, California: Addison-Wesley.

Pinheiro, J.C. & Bates, D.M. (1995). Approximations to the log-likelihood function in the non-

linear mixed-effects model. *Journal of Computational and Graphical Statistics*, **4**, 12–35.

Press, W.H, Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P. (2007). *Numerical Recipes: The Art of Scientific Computing (Third Edition)*, New York: Cambridge University Press.

Raudenbush, S.W., Yang, M.-L. & 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.

Rue, H., Martino, S. & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B*, **7**, 319–392.

Seeger, M. (2000) Bayesian Model Selection for Support Vector Machines, Gaussian Processes and Other Kernel Classifiers. *Neural Information Processing Systems 12*, 603–609.

Skrondal, A. & Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models.* Boca Raton, Florida: Chapman & Hall.

Titterington, D.M. (2004). Bayesian methods for neural networks and related models. *Statistical Science* **19**, 128–139.

R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. `www.R-project.org`.

Smyth, G. (2009). `statmod 1.4.0` Statistical modeling. R package. `http://cran.r-project.org`.

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

Venables, W.N. & Ripley, B.D. `VR 7.2` Functions and datasets to support Venables and Ripley 'Modern Applied Statistics with S' (4th Edition). R package. `http://cran.r-project.org`.

Wang, B. & Titterington, 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), pp. 373–380. Society for Artificial Intelligence and Statistics.

Zhao, Y., Staudenmayer, J., Coull, B.A. & Wand, M.P. (2006). General design Bayesian generalized linear mixed models. *Statistical Science*, **21**, 35–51.

Zeger, S.L. & Karim, M.R. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association*, **86**, 79–86.