

Closed form GLM cumulants and GLMM fitting with a SQUAR-EM-LA₂ algorithm.

Vadim Zipunnikov and James G. Booth*

November 14, 2011

Abstract

We find closed form expressions for the standardized cumulants of generalized linear models. This reduces the complexity of their calculation from $O(p^6)$ to $O(p^2)$ operations which allows efficient construction of second-order saddle-point approximations to the pdf of sufficient statistics. We adapt the result to obtain a closed form expression for the second-order Laplace approximation for a GLMM likelihood. Using this approximation, we develop a computationally highly efficient accelerated EM procedure, SQUAR-EM-LA₂. The procedure is illustrated by fitting a GLMM to a well-known data set. Extensive simulations show the phenomenal performance of the approach. Matlab software is provided for implementing the proposed algorithm.

Key Words: Second-order Laplace approximation, EM algorithm, GLM cumulants, GLMM.

1 Introduction

The class of generalized linear models (GLM), introduced by Nelder and Wedderburn [1972], includes many popular statistical models [McCullagh and Nelder, 1989]. In many applications one needs to know the finite-sample distribution of sufficient statistic in GLMs. However, this distribution often can not be expressed explicitly and has to be approximated. Saddlepoint approximation (SA) is a popular approach and gives quite accurate results [Butler, 2007]. To further increase the accuracy of SA, second-order terms may be included. These terms are based on the standardized GLM cumulants and require calculation of some crossed sums involving $O(p^6)$ terms with p being a dimension of the sufficient statistics. The first contribution of this paper is a closed form expression for these standardized cumulants. Our result dramatically reduces the complexity of required calculations from $O(p^6)$ to $O(p^2)$.

*Vadim Zipunnikov is Postdoctoral Fellow, Department of Biostatistics, Johns Hopkins University, Baltimore, MD, 21205 (email: vzipunni@jhsp.h.edu), James G. Booth is Professor, Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853

Generalized linear mixed models (GLMMs) extend the GLM class by including random effects in their linear predictor (Hobert, 2000, McCulloch and Searle, 2001, Demidenko, 2004, Jiang, 2006). The likelihood function for a GLMM involves an integral over the distribution of the random effects. The integral is generally intractable analytically and hence some form of approximation must be used in practice to enable likelihood-based inference. The existing approaches may be divided in three types. The first type uses Monte Carlo or Quasi Monte Carlo samples to approximate intractable integral. These methods include Simulated Maximum Likelihood (Geyer and Thompson, 1992, Durbin and Koopman, 1997), Markov Chain Monte Carlo (McCulloch, 1994, 1997, Booth and Hobert, 1999), and Quasi-Monte Carlo (Kuo et al., 2008) and some others (Jiang, 2006).

The second type involves analytical methods using the Laplace approximation (LA) (Tierney and Kadane, 1986), and variants such as Penalized-Quasi Likelihood (PQL) (Schall, 1991, Breslow and Clayton, 1993, Wolfinger and O'Connell, 1993), and H-likelihood (Lee et al., 2006, Noh and Lee, 2007). All these approaches deliver a fixed error of approximation and in a lot of practical cases result in biased estimators of the parameters (Breslow and Lin, 1995, Lin and Breslow, 1996, Shun and McCullagh, 1995, Raudenbush et al., 2000, Noh and Lee, 2007). The third type uses numerical integral approximations such as adaptive Gaussian quadrature and tends to be limited to one or two dimensional integrals.

To address the bias problem several papers have studied higher-order Laplace approximation for GLMMs. Shun and McCullagh [1995] introduced a modified second-order Laplace approximation with exponentiated correction term. Laplace 6 approximation was developed in Raudenbush et al. [2000] for nested designs and was implemented in the HLM package (Raudenbush et al. [2004] and Diaz [2007]). The correction terms used in Laplace 6 are different from those in the second-order Laplace approximation. The Laplace 6 approximation was later modified and implemented via an EM algorithm, providing a more reliable convergence. However, the details of this particular modification are not available in the literature (see Raudenbush et al. [2004] and Ng et al. [2006]). Steele [1996] proposed an EM algorithm for GLMMs with independent random effects. To improve the accuracy of the first-order Laplace approximation, Steele [1996] included some higher order adjustments which, however, were different from those in the second-order Laplace approximation. Noh and Lee [2007] considered the second-order Laplace approximation in the context of restricted maximum likelihood for binary data. Evangelou et al. [2008] adapted Shun and McCullagh [1995] results for estimation and prediction of spatial GLMMs. Shun [1997] and following him Evangelou et al. [2008] showed how to reduce the computational burden by ignoring terms with decaying contribution. Noh and Lee [2007] suggested a procedure for calculating the higher-order terms using sparsity of the binary design matrix. In contrast, we find a closed form expression for the higher-order terms which makes the calculation problem a non-issue. In particular, the complexity of all necessary calculations is reduced from $O(q^6)$ to $O(q^2)$ where q is a dimension of the

approximated integral.

Based on second-order Laplace approximations, we develop a highly stable and very accurate EM approach (Dempster et al. [1977]). We then accelerate the EM algorithm using the remarkably efficient SQUAR method developed in Roland and Varadhan [2005] and Varadhan and Roland [2008] and call our procedure SQUAR-EM-LA₂. It is applied to Minnesota Health Plan data (Waller and Zelterman, 1997), compared to the other available methods, and evaluated in an extensive simulation study.

The paper proceeds as follows. In Section 2 we derive the closed form expression for standardized GLM cumulants. We then propose a second-order LA for GLMM likelihood integrals in Section 3, and show how these approximations can be used in the EM-LA₂ algorithm. In Section 4 we use the developed methods for fitting a GLMM and apply them to the Minnesota Health Plan Data. The simulation study is in Section 5, and we conclude with a discussion.

2 Second-Order Saddlepoint Approximation for GLM

We start with a brief review of GLM. Let $\{y_i\}_{i=1}^n$'s be observable response variables coming from an exponential dispersion distribution

$$f(y_i; \theta_i, \phi) = \exp \left\{ \frac{w_i}{\phi} [y_i \theta_i - b(\theta_i)] + c(y_i; \phi) \right\} \quad (1)$$

with means $\{\mu_i\}_{i=1}^n$, where $\mu_i = b'(\theta_i)$, and let $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T$ be a vector of explanatory variables associated with the i th response. For the models considered in this paper the dispersion parameter ϕ is a known constant which, without loss of generality, we set to be one. In addition, we consider only canonical link models, that is, $\theta_i = \mathbf{x}_i^T \boldsymbol{\beta} = (b')^{-1}(\mu_i) = g(\mu_i)$ for the i th response, where $\boldsymbol{\beta}$ is a parameter of interest. Under these assumptions, the log-likelihood function for the parameter $\boldsymbol{\beta}$ is

$$l(\boldsymbol{\beta}; \mathbf{y}) = \sum_{i=1}^n w_i (y_i \mathbf{x}_i^T \boldsymbol{\beta} - b(\mathbf{x}_i^T \boldsymbol{\beta})) + \sum_{i=1}^n c(y_i). \quad (2)$$

Note that the last term in (2) does not depend on parameter $\boldsymbol{\beta}$ and can be ignored. From (2) it immediately follows that $\mathbf{t} = \sum_{i=1}^n w_i y_i \mathbf{x}_i^T$ is the sufficient statistic for parameter $\boldsymbol{\beta}$. In most cases, the finite sample distribution of sufficient statistic $\mathbf{t}$ is complicated and has to be approximated.

Let $\mathbf{u}$ be a p -dimensional random vector. Then the saddlepoint approximation to the distribution of $\mathbf{u}$ is defined as

$$\hat{f}(\mathbf{u}) = \frac{e^{K(\hat{\mathbf{s}}) - \hat{\mathbf{s}}^T \mathbf{u}}}{(2\pi)^{p/2} \det[K''(\hat{\mathbf{s}})]^{1/2}}, \quad \mathbf{u} \in \mathbb{R}^p \quad (3)$$

where $K(\mathbf{s}) = \log E e^{\mathbf{u}^T \mathbf{s}}$, $\mathbf{s} \in \mathbb{R}^p$ is the cumulant generating function (CGF) and $\hat{\mathbf{s}}$

is the unique solution to the p -dimensional saddlepoint equation

$$K'(\hat{\mathbf{s}}) = \mathbf{u}, \quad (4)$$

(Butler, 2007). In particular, since the CGF for the distribution of $\mathbf{t}$ is given by

$$K(\mathbf{s}) = \sum_{i=1}^n w_i b(\mathbf{x}_i^T \mathbf{s}), \quad \mathbf{s} \in \mathbb{R}^p, \quad (5)$$

the saddlepoint density for the sufficient statistic $\mathbf{t}$ is

$$\hat{f}(\mathbf{t}) = \frac{\exp\{\sum_{i=1}^n w_i [b(\mathbf{x}_i^T \hat{\mathbf{s}}) - y_i \mathbf{x}_i^T \hat{\mathbf{s}}]\}}{(2\pi)^{p/2} \det[\sum_{i=1}^n w_i b''(\mathbf{x}_i^T \hat{\mathbf{s}}) \mathbf{x}_i \mathbf{x}_i^T]^{1/2}}, \quad \mathbf{t} \in \mathbb{R}^p. \quad (6)$$

where, $\hat{\mathbf{s}}$, satisfies $\sum_{i=1}^n w_i b'(\mathbf{x}_i^T \hat{\mathbf{s}}) = \mathbf{t}$. Notice that the saddlepoint equation is exactly the score equation of GLM likelihood. Therefore, the saddlepoint is just the maximum likelihood estimator of parameter $\boldsymbol{\beta}$. In addition, the derivatives of the CGF (5) are given by $K^{(k)}(\mathbf{s}) = -l^{(k)}(\mathbf{s})$, for any $k \geq 2$ where the equality means the equality of the corresponding partial derivatives of $K(\mathbf{s})$ and $-l(\mathbf{s})$. In particular, $\hat{K}''$ is equal to the estimated information matrix $\hat{\mathbf{I}} = \mathbf{X}^T \hat{\mathbf{W}} \mathbf{X}$, where $\hat{\mathbf{W}}$ is an $n \times n$ diagonal matrix with positive diagonal elements $\hat{\mathbf{w}}_i = w_i b''(\mathbf{x}_i^T \hat{\mathbf{s}})$.

Adding higher order terms to the saddlepoint approximation improves its accuracy, at least asymptotically (Butler, 2007). We consider two ways of correcting (3), both discussed in Butler [2007, Chapter 3]. The first is an additive correction and the second is its exponential counterpart suggested by McCullagh [1987, Section 6.3],

$$\tilde{f}_1(\mathbf{u}) = \hat{f}(\mathbf{u})(1 + O) \quad \text{and} \quad \tilde{f}_2(\mathbf{u}) = \hat{f}(\mathbf{u})e^O. \quad (7)$$

The correction term, O , is given by the formula (Butler, 2007, Section 3.2.2)

$$O = \frac{1}{8} \hat{\kappa}_4 - \frac{1}{24} (2\hat{\kappa}_{23}^2 + 3\hat{\kappa}_{13}^2) \quad (8)$$

where $\hat{\kappa}_4, \hat{\kappa}_{13}^2, \hat{\kappa}_{23}^2$ are standardized cumulants given by

$$\begin{aligned} \hat{\kappa}_4 &= \sum_{t_1, t_2, t_3, t_4} \hat{K}_{t_1 t_2 t_3 t_4} \hat{K}^{t_1 t_2} \hat{K}^{t_3 t_4} \\ \hat{\kappa}_{13}^2 &= \sum_{t_1, t_2, t_3, t_4, t_5, t_6} \hat{K}_{t_1 t_2 t_3} \hat{K}_{t_4 t_5 t_6} \hat{K}^{t_1 t_2} \hat{K}^{t_3 t_4} \hat{K}^{t_5 t_6} \\ \hat{\kappa}_{23}^2 &= \sum_{t_1, t_2, t_3, t_4, t_5, t_6} \hat{K}_{t_1 t_2 t_3} \hat{K}_{t_4 t_5 t_6} \hat{K}^{t_1 t_4} \hat{K}^{t_2 t_5} \hat{K}^{t_3 t_6} \end{aligned} \quad (9)$$

with

$$\hat{K}^{t_1 t_2} = (\hat{\mathbf{I}}^{-1})_{t_1 t_2}, \quad \hat{K}_{t_1 t_2 t_3} = -\frac{\partial^3 l(\hat{\boldsymbol{\beta}})}{\partial \beta_{t_1} \partial \beta_{t_2} \partial \beta_{t_3}}, \quad \hat{K}_{ijkl} = -\frac{\partial^4 l(\hat{\boldsymbol{\beta}})}{\partial \beta_{t_1} \partial \beta_{t_2} \partial \beta_{t_3} \partial \beta_{t_4}}.$$

The major obstacle for using second-order correction is that calculation of the standardized cumulants in (8) is very computationally demanding due to the crossed sums

involving $O(p^6)$ terms. It quickly becomes prohibitive when p increases. For instance, if p is equal to 40 this results in $4.09 \cdot 10^9$ terms to calculate and sum up which is enormously challenging.

However, we now show that for GLM models the sums are separable and the standardized cumulants have a simple form which can be calculated with $O(p^2)$ computational effort.

Theorem 1. *The standardized cumulants in (8) are given by*

$$\begin{aligned}\hat{\kappa}_4 &= \sum_{i=1}^n w_i \hat{b}_i^{(4)} \hat{\gamma}_{ii}^2 \\ \hat{\kappa}_{13}^2 &= \sum_{i_1=1}^n \sum_{i_2=1}^n w_{i_1} w_{i_2} \hat{b}_{i_1}^{(3)} \hat{b}_{i_2}^{(3)} \hat{\gamma}_{i_1 i_1} \hat{\gamma}_{i_1 i_2} \hat{\gamma}_{i_2 i_2} \\ \hat{\kappa}_{23}^2 &= \sum_{i_1=1}^n \sum_{i_2=1}^n w_{i_1} w_{i_2} \hat{b}_{i_1}^{(3)} \hat{b}_{i_2}^{(3)} \hat{\gamma}_{i_1 i_2}^3\end{aligned}\tag{10}$$

where $\hat{\gamma}_{i_1 i_2} = \mathbf{x}_{i_1}^T \hat{\mathbf{I}}^{-1} \mathbf{x}_{i_2}$, $i_1, i_2 = 1, \dots, p$ and $\hat{b}_i^{(k)} = b^{(k)}(\mathbf{x}_i^T \hat{\boldsymbol{\beta}})$, $k = 3, 4$.

Theorem 1 allows us to construct second-order saddlepoint approximations for the sufficient statistic even in very high-dimensional cases. The Proof of Theorem 1 is given in the Appendix.

3 Second-order Laplace approximation for GLMM

Generalized linear mixed models are a natural extension of GLMs that are often used to model longitudinal and clustered observations (see McCulloch and Searle, 2001, Demidenko, 2004, Jiang, 2006 for recent reviews). The likelihood function for a GLMM involves an integral over the distribution of the random effects. In this section we derive a closed form expression for the second-order Laplace approximation (LA₂) to GLMM likelihoods.

Let $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})^T$, $i = 1, \dots, n$, be independent random response vectors. Let $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ denote known p - and q -dimensional covariate vectors associated with the j^{th} component of $\mathbf{y}_i$. Dependence between the components of the $\mathbf{y}_i$'s is induced by unobservable q -dimensional random effects vectors,

$$\mathbf{u}_i^\Sigma = (u_{i1}^\Sigma, \dots, u_{iq}^\Sigma)^T \sim \text{i.i.d. } N_q(\mathbf{0}, \boldsymbol{\Sigma}), \quad i = 1, \dots, n,$$

where $\boldsymbol{\Sigma}$ is assumed to be positive definite. Conditionally on the random effect $\mathbf{u}_i^\Sigma$, the univariate components, y_{ij} , $j = 1, \dots, n_i$ are independent with means, $\mu_{ij} = E(y_{ij} | \mathbf{u}_i^\Sigma)$, satisfying

$$g(\mu_{ij}) = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{z}_{ij}^T \mathbf{u}_i^\Sigma, \tag{11}$$

where $\boldsymbol{\beta}$ is a p -dimensional parameter and $g(\cdot)$ is a link function. Since $\boldsymbol{\Sigma}$ is positive definite there exists a unique $q \times q$ lower-triangular matrix $\mathbf{D}$ with positive diagonal entries such that $\boldsymbol{\Sigma} = \mathbf{D}\mathbf{D}^T$, and hence $\mathbf{u}_i^\Sigma \stackrel{d}{=} \mathbf{D}\mathbf{u}_i$, where $\mathbf{u}_i \sim \text{i.i.d. } N_q(0, \mathbf{I}_q)$, $i = 1, \dots, n$. Therefore, without loss of generality, we may consider the distributionally equivalent form (Demidenko, 2004, page 411), $g(\mu_{ij}) = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{z}_{ij}^T \mathbf{D}\mathbf{u}_i$ in place of (11).

The matrix $\mathbf{D}$ usually can be characterized by a few non-zero entries. Let $\boldsymbol{\sigma}$ be a q^* -dimensional vector containing these unique non-zero elements. It will sometimes be more convenient to use an alternative representation $\mathbf{z}_{ij}^T \mathbf{D} = \boldsymbol{\sigma}^T \mathbf{P}(\mathbf{z}_{ij})$ where $\mathbf{P}(\mathbf{z}_{ij})$ is $q^* \times q$ matrix based on $\mathbf{z}_{ij}$. Then $g(\mu_{ij}) = \hat{\mathbf{x}}_{ij}^T \boldsymbol{\psi}$ where $\hat{\mathbf{x}}_{ij}^T = (\mathbf{x}_{ij}^T, \mathbf{u}_i^T \mathbf{P}^T(\mathbf{z}_{ij}))$ and $\boldsymbol{\psi} = (\boldsymbol{\beta}^T, \boldsymbol{\sigma}^T)^T$ is the $p + q^*$ dimensional parameter of interest. Specification of a GLMM is completed by describing variability in the response, y_{ij} , about its conditional mean, μ_{ij} , using an exponential model of the form $f(y_{ij}|\mu_{ij}) = \exp\{w_{ij}[\theta_{ij}y_{ij} - b(\theta_{ij})] + c(y_{ij})\}$ for some function $c(\cdot)$, canonical parameter $\theta_{ij} = (b')^{-1}(\mu_{ij})$, and known weights w_{ij} .

The observable likelihood function for the parameter $\boldsymbol{\psi}$ is therefore

$$L(\boldsymbol{\psi}; \mathbf{y}) = \int_{\mathbb{R}^q} f(\mathbf{y}|\mathbf{u}; \boldsymbol{\psi}) \phi(\mathbf{u}, \mathbf{I}_{nq}) d\mathbf{u}, \quad (12)$$

where $\mathbf{y} = (\mathbf{y}_1^T, \dots, \mathbf{y}_n^T)^T$ and $\mathbf{u} = (\mathbf{u}_1^T, \dots, \mathbf{u}_n^T)^T$, $\phi(\mathbf{u}, \mathbf{I}_{nq}) = \prod_{i=1}^n \prod_{r=1}^q \phi(u_{ir})$, where $\phi(\cdot)$ is the standard normal density, and $f(\mathbf{y}|\mathbf{u}; \boldsymbol{\psi}) = \prod_{i=1}^n f(\mathbf{y}_i|\mathbf{u}_i; \boldsymbol{\psi}) = \prod_{i=1}^n \prod_{j=1}^{n_i} \exp\{w_{ij}[\theta_{ij}y_{ij} - b(\theta_{ij})] + c(y_{ij})\}$. The log-likelihood is therefore

$$l(\boldsymbol{\psi}; \mathbf{y}) = \sum_{i=1}^n \log \int_{\mathbb{R}^q} f(\mathbf{y}_i|\mathbf{u}_i; \boldsymbol{\psi}) \phi(\mathbf{u}_i, \mathbf{I}_q) d\mathbf{u}_i = \sum_{i=1}^n \log \mathbf{G}_i, \quad (13)$$

with

$$\mathbf{G}_i = \int_{\mathbb{R}^q} f(\mathbf{y}_i|\mathbf{u}_i; \boldsymbol{\psi}) \phi(\mathbf{u}_i, \mathbf{I}_q) d\mathbf{u}_i, \quad i = 1, \dots, n. \quad (14)$$

The integral in (14) in most practical cases, cannot be evaluated explicitly. We now develop the second-order Laplace approximation to the GLMM likelihood. Below we consider only the i^{th} integral and so the index i will be omitted whenever it does not interfere with clarity of the exposition.

Suppose the function $h(\mathbf{u})$ is a $\mathcal{C}^6(\mathbb{R}^q)$ convex function with a global minimum at $\hat{\mathbf{u}}$ and $h(\mathbf{u}) = O(m)$. Then (see, for example, Kolassa, 2006, Section 6.5) we have

$$\int_{\mathbb{R}^q} e^{-h(\mathbf{u})} d\mathbf{u} = \frac{(2\pi)^{q/2} e^{-h(\hat{\mathbf{u}})}}{\sqrt{\det[h''(\hat{\mathbf{u}})]}} (1 - \tau_1(\hat{\mathbf{u}}) + O(m^{-2})), \quad (15)$$

where the $O(m^{-1})$ correction term $\tau_1(\hat{\mathbf{u}})$ carries information about higher-order derivatives of the function $h(\mathbf{u})$ and is given by (8) with the standardized cumulants having the same form as (9) with h in place of K . To get LA₂ for (14), we set

$$h(\mathbf{u}_i) = - \left(\sum_{j=1}^{n_i} \{w_{ij}[y_{ij}\theta_{ij} - b(\theta_{ij})] + c(y_{ij})\} - \frac{1}{2} \mathbf{u}_i^T \mathbf{u}_i \right) \quad (16)$$

with $\theta_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{z}_{ij}^T \mathbf{D} \mathbf{u}_i$. It is easy to see that $h(\mathbf{u}_i)$ is $O(n_i)$. Applying (15) to each

of the integrals (14) and taking logarithm, we get the LA_2 log-likelihood

$$l_{LA_2}(\boldsymbol{\psi}; \mathbf{y}) = - \sum_{i=1}^n h(\hat{\mathbf{u}}_i) - \frac{1}{2} \sum_{i=1}^n \log \det[h''(\hat{\mathbf{u}}_i)] + \sum_{i=1}^n \log(1 - \tau_1(\hat{\mathbf{u}}_i)). \quad (17)$$

The last term in (17) represents a higher order correction to the standard (first-order) Laplace approximation. Hence,

$$l_{LA_2}^1(\boldsymbol{\psi}; \mathbf{y}) = l_{LA_1}(\boldsymbol{\psi}; \mathbf{y}) + \sum_{i=1}^n \log(1 - \tau_1(\hat{\mathbf{u}}_i)) \quad (18)$$

Similarly to (7), the exponential correction may be more accurate in some cases. In particular, Shun and McCullagh [1995] assumed that $q = O(m^{1/2})$ in (15) and gave an argument for why the correction $\tau_1(\hat{\mathbf{u}}_i)$ must be exponentiated for crossed designs. With the correction term in the exponential form, the LA_2 log-likelihood takes form

$$l_{LA_2}^2(\boldsymbol{\psi}; \mathbf{y}) = l_{LA_1}(\boldsymbol{\psi}; \mathbf{y}) - \sum_{i=1}^n \tau_1(\hat{\mathbf{u}}_i). \quad (19)$$

The main result of this section reduces the computational burden of calculating higher-order corrections $\tau_1(\hat{\mathbf{u}}_i)$ from $O(q^6)$ to $O(q^2)$. It is based on the observation that the function $h(\mathbf{u}_i)$ in (16) is a sum of a GLM log-likelihood and a quadratic term in $\mathbf{u}_i$. Therefore, all results for the GLM log-likelihood based solely on derivatives of order three and higher can be rewritten for the function $h(\mathbf{u}_i)$ with some necessary corrections for second-order derivative of $h(\mathbf{u}_i)$.

Theorem 2. *The standardized cumulants $\hat{\kappa}_4, \hat{\kappa}_{13}^2, \hat{\kappa}_{23}^2$ in (15) are given by*

$$\begin{aligned} \hat{\kappa}_4 &= \sum_{j=1}^{n_i} w_{ij} \hat{b}_{ij}^{(4)} \hat{\gamma}_{ijj}^2 \\ \hat{\kappa}_{13}^2 &= \sum_{j_1=1}^{n_i} \sum_{j_2=1}^{n_i} w_{ij_1} w_{ij_2} \hat{b}_{ij_1}^{(3)} \hat{b}_{ij_2}^{(3)} \hat{\gamma}_{ij_1 j_1} \hat{\gamma}_{ij_1 j_2} \hat{\gamma}_{ij_2 j_2} \\ \hat{\kappa}_{23}^2 &= \sum_{j_1=1}^{n_i} \sum_{j_2=1}^{n_i} w_{ij_1} w_{ij_2} \hat{b}_{ij_1}^{(3)} \hat{b}_{ij_2}^{(3)} \hat{\gamma}_{ij_1 j_2}^3 \end{aligned} \quad (20)$$

where $\hat{\gamma}_{ij_1 j_2} = \mathbf{z}_{ij_1}^T \mathbf{D}[h''(\hat{\mathbf{u}}_i)]^{-1} \mathbf{D}^T \mathbf{z}_{ij_2}$, $j_1, j_2 = 1, \dots, n_i$ with

$$h''(\hat{\mathbf{u}}_i) = \mathbf{I}_q + \mathbf{D}^T \left(\sum_{j=1}^{n_i} w_{ij} b''(\hat{\theta}_{ij}) \mathbf{z}_{ij} \mathbf{z}_{ij}^T \right) \mathbf{D} \quad (21)$$

Theorem 2 may be immediately adapted to many estimating procedures requiring approximation of GLMM likelihoods. One of the applications we see is h-likelihood in the context of estimating hierarchical GLMs (Lee et al., 2006). Lee et al. [2006, page 187] write “... when the number of random components is greater than two this second-order method is computationally too expensive”. Theorem 2 provides a way of doing these computations very efficiently. Another application which has a great potential is integrated nested Laplace approximations in the Bayesian framework

(Rue et al., 2009). Finally, one can directly maximize the LA_2 approximated GLMM likelihoods (18) or (19) to get ML estimates. We, however, propose an alternate approach using the EM algorithm for the reasons outlined in the next section.

3.1 SQUAR-EM- LA_2 for GLMM

The EM algorithm (Dempster et al., 1977) can be applied in the GLMM context by treating random effects as missing data. The algorithm includes two steps at each iteration, an expectation E-step and a maximization M-step. Let $\boldsymbol{\psi}^{(s)}$ denote the value of the parameter after iteration s . Then the E-step at iteration $s + 1$ involves the computation of the so-called Q -function, $Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(s)}) = E[l(\boldsymbol{\psi}; \mathbf{y}, \mathbf{u})|\mathbf{y}; \boldsymbol{\psi}^{(s)}]$, where $l(\boldsymbol{\psi}; \mathbf{y}, \mathbf{u}) = \log f(\mathbf{y}, \mathbf{u}; \boldsymbol{\psi})$ is the complete data loglikelihood for parameter $\boldsymbol{\psi}$. The M-step consists of finding $\boldsymbol{\psi}^{(s+1)}$ which maximizes the Q -function; that is $\boldsymbol{\psi}^{(s+1)} = \arg \max_{\boldsymbol{\psi} \in \Psi} Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(s)})$. Under mild regularity conditions the observable likelihood function (12) is non-decreasing when evaluated along the EM sequence $\{\boldsymbol{\psi}^{(s)}\}_{s=0}^{\infty}$ [see e.g. Wu, 1983]. Hence, the sequence converges to a local maximum of the likelihood surface.

As with direct likelihood approaches the E-step involves some intractable integrals. The approach we will describe in this section, EM- LA_2 , replaces those integrals with second-order Laplace approximations. There are a few reasons why we prefer the EM approach over direct maximization in this setting. First, maximization of the LA_2 approximated direct likelihood would require differentiating terms already involving higher-order derivatives of the link function. Our EM- LA_2 avoids this problem. Secondly, the EM algorithm is known to be very stable in a broad range of problems, and the numerical examples and simulations discussed later in this paper appear to substantiate this in the GLMM context. Thirdly, the M-step of EM in the GLMM context is equivalent to fitting a GLM, and can therefore be accomplished using a procedure similar to the standard iteratively reweighted least squares (IRLS) algorithm.

Let us recall that the complete data loglikelihood is given by $-\sum_{i=1}^n h(\mathbf{u}_i)$ where function $h(\mathbf{u}_i)$ is defined in (16). Hence, the Q -function calculated at iteration $s + 1$ is $Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(s)}) = \sum_{i=1}^n E\left[\sum_{j=1}^{n_i}\{w_{ij}[\theta_{ij}y_{ij} - b(\theta_{ij})] + c(y_{ij})\} - \frac{1}{2}\mathbf{u}_i^T \mathbf{u}_i | \mathbf{y}_i; \boldsymbol{\psi}^{(s)}\right]$. However, part of this expression, $\sum_{i=1}^n E\left[\sum_{j=1}^{n_i} w_{ij}c(y_{ij}) - \frac{1}{2}\mathbf{u}_i^T \mathbf{u}_i | \mathbf{y}_i; \boldsymbol{\psi}^{(s)}\right]$ can be eliminated because it does not depend on the parameter $\boldsymbol{\psi}$, and has no effect on the M-step. Therefore, without loss of generality, we shall consider the reduced Q -function,

$$Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(s)}) = \sum_{i=1}^n E\left[a(\mathbf{y}_i, \mathbf{u}_i; \boldsymbol{\psi}) | \mathbf{y}_i; \boldsymbol{\psi}^{(s)}\right] = \sum_{i=1}^n \int_{\mathbb{R}^q} a(\mathbf{y}_i, \mathbf{u}_i; \boldsymbol{\psi}) f(\mathbf{u}_i | \mathbf{y}_i; \boldsymbol{\psi}^{(s)}) d\mathbf{u}_i \quad (22)$$

where $a(\mathbf{y}_i, \mathbf{u}_i; \boldsymbol{\psi}) = \sum_{j=1}^{n_i} w_{ij}[\theta_{ij}y_{ij} - b(\theta_{ij})]$ and

$$f(\mathbf{u}_i | \mathbf{y}_i; \boldsymbol{\psi}^{(s)}) = \frac{f(\mathbf{y}_i, \mathbf{u}_i; \boldsymbol{\psi}^{(s)})}{f(\mathbf{y}_i; \boldsymbol{\psi}^{(s)})} = \frac{\exp \left\{ a(\mathbf{y}_i, \mathbf{u}_i; \boldsymbol{\psi}^{(s)}) - \frac{1}{2} \mathbf{u}_i^T \mathbf{u}_i \right\}}{\int_{\mathbb{R}^q} \exp \left\{ a(\mathbf{y}_i, \mathbf{u}_i; \boldsymbol{\psi}^{(s)}) - \frac{1}{2} \mathbf{u}_i^T \mathbf{u}_i \right\} d\mathbf{u}_i} \quad (23)$$

As noted earlier the denominator in (23) is generally analytically intractable in the GLMM context. We now show how (22) may be approximated with LA_2 .

Suppose function $h(\mathbf{u})$ is a $\mathcal{C}^6(\mathbb{R}^q)$ convex function with a global minimum at $\hat{\mathbf{u}}$ and $h(\mathbf{u}) = O(m)$. Suppose that the $\mathcal{C}^5(\mathbb{R}^q)$ function $\zeta(\mathbf{u})$ is such that $|\zeta_{t_1 t_2 t_3 t_4}(\mathbf{u})| \leq C_1 \exp(C_2 |\mathbf{u}|^2)$, $\forall \mathbf{u} \in \mathbb{R}^q$, for some constants C_1 and C_2 and all fourth degree partial derivatives $\zeta_{t_1 t_2 t_3 t_4}(\mathbf{u})$. Then due to Kolassa [2006, Section 6.5], as well as personal communication with John Kolassa correcting the original formula, we have

$$\int_{\mathbb{R}^q} e^{-h(\mathbf{u})} \zeta(\mathbf{u}) d\mathbf{u} = \frac{(2\pi)^{q/2} e^{-h(\hat{\mathbf{u}})}}{\sqrt{\det[h''(\hat{\mathbf{u}})]}} \left(\zeta(\hat{\mathbf{u}})(1 - \tau_1(\hat{\mathbf{u}})) + \tau_2(\hat{\mathbf{u}}) + O(m^{-2}) \right), \quad (24)$$

where τ_1 is defined in Section 3 and τ_2 is $O(m^{-1})$ and it is defined below. The correction term τ_2 contains information about an interaction of the first two derivatives of $\zeta(\mathbf{u})$ with the second and third derivatives of $h(\mathbf{u})$, correspondingly. It is given by

$$\tau_2 = -\frac{1}{2} \sum_{t_1, t_2, t_3, t_4} \hat{\zeta}^{t_1} \hat{h}^{t_2 t_3 t_4} \hat{h}_{t_1 t_2} \hat{h}_{t_3 t_4} + \frac{1}{2} \sum_{t_1, t_2} \hat{\zeta}^{t_1 t_2} \hat{h}_{t_1 t_2}, \quad (25)$$

where $\hat{\zeta}^{t_1} = \frac{\partial \zeta(\hat{\mathbf{u}})}{\partial u_{t_1}}$ and $\hat{\zeta}^{t_1 t_2} = \frac{\partial^2 \zeta(\hat{\mathbf{u}})}{\partial u_{t_1} \partial u_{t_2}}$.

Formally applying (24) with

$$h(\mathbf{u}_i; \boldsymbol{\psi}^{(s)}) = \frac{1}{2} \mathbf{u}_i^T \mathbf{u}_i - a(\mathbf{y}_i, \mathbf{u}_i; \boldsymbol{\psi}^{(s)}) \quad \text{and} \quad \zeta(\mathbf{u}_i; \boldsymbol{\psi}) = a(\mathbf{y}_i, \mathbf{u}_i; \boldsymbol{\psi}) \quad (26)$$

to the numerator and denominator integrals in (22) and (23) results in the LA_2 approximation of the Q function given by $\tilde{Q}_{LA_2}(\boldsymbol{\psi} | \boldsymbol{\psi}^{(s)}) = \sum_{i=1}^n (\zeta(\hat{\mathbf{u}}_i, \boldsymbol{\psi})(1 - \tau_1(\hat{\mathbf{u}}_i, \boldsymbol{\psi}^{(s)})) + \tau_2(\hat{\mathbf{u}}_i, \boldsymbol{\psi}, \boldsymbol{\psi}^{(s)}) / (1 - \tau_1(\hat{\mathbf{u}}_i, \boldsymbol{\psi}^{(s)})))$. Following Butler [2007, Section 4.1] and Kolassa [2006, Section 7.1], we expand the denominators in a Taylor series as $(1 - \tau_1)^{-1} = 1 + \tau_1 + O(\tau_1^2)$. Dropping higher order terms, this results in the approximation for function $\tilde{Q}(\boldsymbol{\psi} | \boldsymbol{\psi}^{(s)})$ given by

$$\tilde{Q}_{LA_2}(\boldsymbol{\psi} | \boldsymbol{\psi}^{(s)}) = \sum_{i=1}^n \left(\zeta(\hat{\mathbf{u}}_i, \boldsymbol{\psi}) + \tau_2(\hat{\mathbf{u}}_i, \boldsymbol{\psi}, \boldsymbol{\psi}^{(s)}) \right). \quad (27)$$

Note that (24) assumes that $\zeta(\mathbf{u})$ is $O(1)$. However, $\zeta(\hat{\mathbf{u}}_i, \boldsymbol{\psi})$ in (26) is a sum of n_i functions and each of them is $O(1)$. Therefore, (24) is only $O(n_i^{-1})$ correct. This suggests that higher-order LA terms cannot be ignored when (24) is applied to $\zeta(\hat{\mathbf{u}}_i, \boldsymbol{\psi})$. Maximization of the $\tilde{Q}_{LA_2}(\boldsymbol{\psi} | \boldsymbol{\psi}^{(s)})$'s at step s can be done using a Newton-Raphson

algorithm which is described in the Appendix. Note that the saddlepoints $\hat{\mathbf{u}}_i$ depend only on $\boldsymbol{\psi}^{(s)}$ and do not need to be recalculated during maximization of $\tilde{Q}_{LA_2}(\boldsymbol{\psi}|\boldsymbol{\psi}^{(s)})$. The terms $\zeta(\hat{\mathbf{u}}_i, \boldsymbol{\psi})$ and $\tau_2(\hat{\mathbf{u}}_i, \boldsymbol{\psi}, \boldsymbol{\psi}^{(s)})$ involve only first and second derivatives of the link function. As mentioned earlier, this is one of the main advantages of combining EM with LA_2 .

Approximation (24) with $\zeta(\mathbf{u}_i)$ equal to 1 is equivalent to having a linear second-order correction as in (18). Therefore, one should expect EM- LA_2 procedure based on (27) will not perform well for situations described in Shun and McCullagh [1995], particularly, crossed designs. An exponentially corrected equivalent of (24) assuming $q = O(m^{1/2})$ for a non-constant function $\zeta(\mathbf{u})$ should be adapted in those cases. For a positive $\zeta(\mathbf{u})$ one may consider using results from Evangelou et al. [2008, Section 3.1]. To the best of our knowledge, finding the exponential correction for the integral in (24) with a general function $\zeta(\mathbf{u})$ remains an open problem. For those cases, we suggest using the exponentiated LA_2 approximation (19).

We now find a closed form expression for $\tau_2(\hat{\mathbf{u}}_i, \boldsymbol{\psi}, \boldsymbol{\psi}^{(s)})$ in (27). For notational convenience, we split it into two terms as $\tau_2 = -\frac{1}{2}\tau_{21} + \frac{1}{2}\tau_{22}$ with

$$\tau_{21} = \sum_{t_1, t_2, t_3, t_4} \hat{\zeta}^{t_1} \hat{h}^{t_2 t_3 t_4} \hat{h}_{t_1 t_2} \hat{h}_{t_3 t_4} \quad \text{and} \quad \tau_{22} = \sum_{t_1, t_2} \hat{\zeta}^{t_1 t_2} \hat{h}_{t_1 t_2}. \quad (28)$$

As with $\tau_1(\hat{\mathbf{u}}_i, \boldsymbol{\psi}^{(s)})$ given in Theorem 2 there exist closed form expressions for $\tau_{21}(\hat{\mathbf{u}}_i, \boldsymbol{\psi}, \boldsymbol{\psi}^{(s)})$ and $\tau_{22}(\hat{\mathbf{u}}_i, \boldsymbol{\psi}, \boldsymbol{\psi}^{(s)})$. Next result determines the LA_2 approximation to (27).

Theorem 3. *The second-order correction terms (28) are given by*

$$\begin{aligned} \tau_{21}(\hat{\mathbf{u}}_i, \boldsymbol{\psi}, \boldsymbol{\psi}^{(s)}) &= \sum_{j=1}^{n_i} w_{ij} [y_{ij} - b'(\theta_{ij})] \lambda_j(\boldsymbol{\psi}, \boldsymbol{\psi}^{(s)}) \\ \tau_{22}(\hat{\mathbf{u}}_i, \boldsymbol{\psi}, \boldsymbol{\psi}^{(s)}) &= - \sum_{j=1}^{n_i} w_{ij} b''(\theta_{ij}) \gamma_{jj}(\boldsymbol{\psi}, \boldsymbol{\psi}^{(s)}), \end{aligned} \quad (29)$$

with

$$\begin{aligned} \lambda_j(\boldsymbol{\psi}, \boldsymbol{\psi}^{(s)}) &= \mathbf{z}_{ij}^T \mathbf{D}[h''(\hat{\mathbf{u}}_i, \boldsymbol{\psi}^{(s)})]^{-1} \sum_{l=1}^{n_i} w_{il} b_{il}^{(3)} \gamma_{il}^{(s)} \mathbf{D}^{(s)T} \mathbf{z}_{il}, \\ \gamma_{jj}(\boldsymbol{\psi}, \boldsymbol{\psi}^{(s)}) &= \mathbf{z}_{ij}^T \mathbf{D}[h''(\hat{\mathbf{u}}_i, \boldsymbol{\psi}^{(s)})]^{-1} \mathbf{D}^T \mathbf{z}_{ij}, \quad j = 1, \dots, n_i. \end{aligned}$$

Recall that the (generalized) ascent property of EM algorithm (see, for example, Caffo et al., 2005) states that if $\Delta Q^{(s+1)} = Q(\boldsymbol{\psi}^{(s+1)}|\boldsymbol{\psi}^{(s)}) - Q(\boldsymbol{\psi}^{(s)}|\boldsymbol{\psi}^{(s)}) \geq 0$ then, via an application of Jensen's inequality, $l(\boldsymbol{\psi}^{(s+1)}|\mathbf{y}) \geq l(\boldsymbol{\psi}^{(s)}|\mathbf{y})$. However, if $\Delta Q^{(s+1)}$ is approximated with LA_2 the inequality, $\Delta \hat{Q}_{LA_2}^{(s+1)} \geq 0$ no longer guarantees increasing of the log-likelihood. We study the convergence of the algorithm for a range of examples in simulation studies of Section 5.

The EM algorithm is known for its slow (linear) convergence, especially in high-dimensional problems. To address this, several accelerating schemes have been suggested in the literature (Meng and Rubin, 1997, Liu et al., 1998 and many others). Roland and Varadhan [2005] and Varadhan and Roland [2008] proposed a simple but remarkably efficient “off-the-shelf” EM accelerator, SQUAR-EM. Since this accelerator does not require any knowledge about underlying model and uses only EM param-

eter updates it quickly gained its popularity (Schifano et al. 2010). We incorporate it into our approach as follows. Let $\boldsymbol{\psi}^{(s-2)}$, $\boldsymbol{\psi}^{(s-1)}$, and $\boldsymbol{\psi}^{(s)}$ be three sequential EM parameter updates. Then SQUAR-EM substitute $\boldsymbol{\psi}^{(s)}$ with $\boldsymbol{\psi}^{(s)} = \boldsymbol{\psi}^{(s-2)} - 2\alpha\mathbf{r} + \alpha^2\mathbf{v}$, where $\mathbf{r} = \boldsymbol{\psi}^{(s-1)} - \boldsymbol{\psi}^{(s-2)}$, $\mathbf{v} = (\boldsymbol{\psi}^{(s)} - \boldsymbol{\psi}^{(s-1)}) - \mathbf{r}$, and $\alpha = -\|\mathbf{r}\|/\|\mathbf{v}\|$. Using the SQUAR-EM updated $\boldsymbol{\psi}^{(s)}$ the next two EM updates are calculated and a new SQUAR-EM update is obtained based on the three. The procedure is repeated until convergence is declared.

There are multiple potential stopping rules for the EM algorithm (Caffo et al., 2005). We require the relative change in the parameter estimates at the $(s+1)$ th iteration to be sufficiently small; that is

$$\max_{1 \leq i \leq p+q^*} \frac{|\psi_i^{(s+1)} - \psi_i^{(s)}|}{|\psi_i^{(s)}| + \delta_1} \leq \delta_2 \quad (30)$$

for some pre-defined small positive δ_1 and δ_2 .

We obtain standard errors of ML estimates using the observable log-likelihood (13). The information matrix of the parameter $\boldsymbol{\psi}$ can be calculated as

$$\mathbf{I}(\boldsymbol{\psi}) = \sum_{i=1}^n \frac{(\mathbf{G}_i)' \boldsymbol{\psi} (\mathbf{G}_i)'^T}{\mathbf{G}_i^2}, \quad (31)$$

where $\mathbf{G}_i$ is defined in (14). The denominators in (31) can be approximated by either (18) or its exponential counterpart (19). The numerators can be obtained by differentiating (18) (or (19)) with respect to $\boldsymbol{\psi}$. However, it would require calculation of some complicated higher-order derivatives. Instead, we suggest applying (24) to approximate each term in (31) which results in much simpler expressions. Indeed, note that the denominators $\mathbf{G}_i$ are exactly the denominators we approximated in (23). Applying (24) coordinatewise to $\zeta(\mathbf{u}_i) = -h'_{\boldsymbol{\psi}}(\mathbf{u}_i)$ and using the same argument as in (27), we get an LA_2 approximated information matrix given by

$$\hat{\mathbf{I}}_{LA_2}(\boldsymbol{\psi}) = \sum_{i=1}^n \left(-h'_{\boldsymbol{\psi}}(\hat{\mathbf{u}}_i) + \tau_2(h', \hat{\mathbf{u}}_i) \right) \left(-h'_{\boldsymbol{\psi}}(\hat{\mathbf{u}}_i) + \tau_2(h', \hat{\mathbf{u}}_i) \right)^T \quad (32)$$

where $\tau_2(h', \hat{\mathbf{u}}_i)$ is defined in (28) for each coordinate of the vector function $-h'_{\boldsymbol{\psi}}(\mathbf{u}_i)$. A formal expression for $t_2(\hat{\mathbf{u}}_i)$ as well as its derivation are provided in Appendix.

4 Minnesota Health Plan Data

Waller and Zelterman [1997] reported data (MHP) from longitudinal records on 121 senior citizens enrolled in a health plan in Minnesota. The data consist of the number of times each subject visited or called the medical clinic in each of four 6-month periods. Let y_{ikl} denote the count for subject i , event k (visit or call), and period l .

It is natural to consider subject as a random factor, but event and period as fixed. Hence we consider a Poisson loglinear model with $y_{ikl}|\mathbf{u}_i^\Sigma \sim \text{Poisson}(\mu_{ikl})$, and

$$\log \mu_{ikl} = a_0 + a_k + b_l + c_{kl} + \gamma_i + v_{ik} + \omega_{il}, \quad k = 1, 2, \text{ and } l = 1, 2, 3, 4, \quad (33)$$

where a_0 is an intercept, a_k is the fixed effect of event k , b_l is the fixed effect of period l , c_{kl} is fixed event $\times$ period interaction, γ_i is a random effect associated with subject i , v_{ik} is a random subject $\times$ event interaction, and ω_{il} is a random subject $\times$ period interaction. The model therefore involves a 7-dimensional random effect $\mathbf{u}_i^\Sigma = (\gamma_i, v_{i1}, v_{i2}, \omega_{i1}, \omega_{i2}, \omega_{i3}, \omega_{i4})$, $i = 1, \dots, 121$, associated with the subject i . We suppose that $\mathbf{u}_i^\Sigma \sim \text{i.i.d. } N_7(\mathbf{0}, \Sigma)$, $i = 1, \dots, 121$ where Σ is 7×7 diagonal matrix with $\Sigma_{11} = \sigma_\gamma^2$, $\Sigma_{ii} = \sigma_v^2$, $i = 2, 3$ and $\Sigma_{ii} = \sigma_\omega^2$, $i = 4, 5, 6, 7$. We achieve identifiability by setting $a_2 = b_4 = c_{14} = c_{21} = c_{22} = c_{23} = c_{24} = 0$. The fixed effects parameter is then $\beta = (a_0, a_1, b_1, b_2, b_3, c_{11}, c_{12}, c_{13})$.

To eliminate the double index kl , and express the model in the form in (11), we consider a new index $j = 4(k - 1) + l$. Accordingly, $(y_{i1}, \dots, y_{i4}, y_{i5}, \dots, y_{i8}) = (y_{i11}, \dots, y_{i14}, y_{i21}, \dots, y_{i24})$ and $(\mu_{i1}, \dots, \mu_{i4}, \mu_{i5}, \dots, \mu_{i8}) = (\mu_{i11}, \dots, \mu_{i14}, \mu_{i21}, \dots, \mu_{i24})$, for each $i = 1, \dots, 121$. In addition, we introduce

$$\begin{aligned} \mathbf{x}_{ij} &= (1, I_{\{1 \leq j \leq 4\}}, I_{\{j=1 \text{ or } 5\}}, I_{\{j=2 \text{ or } 6\}}, I_{\{j=3 \text{ or } 7\}}, I_{\{j=1\}}, I_{\{j=2\}}, I_{\{j=3\}})^T \\ \mathbf{z}_{ij} &= (1, I_{\{1 \leq j \leq 4\}}, I_{\{5 \leq j \leq 8\}}, I_{\{j=1 \text{ or } 5\}}, I_{\{j=2 \text{ or } 6\}}, I_{\{j=3 \text{ or } 7\}}, I_{\{j=4 \text{ or } 8\}})^T \end{aligned}$$

where $I_{\{A\}}$ is the indicator of event A . Note that any multi-index model can be reduced to the form (12), in a similar manner, by appropriate re-indexing of variables.

A similar model was proposed by Booth et al. [2003] for this data, the difference being that the event by period interaction term was not included in their analysis. We applied non-accelerated EM-LA₂ with stopping rule parameters δ_1 and δ_2 in (30) set equal to 10^{-4} . Convergence was declared at the 155th iteration of the algorithm. We started with a randomly chosen starting point and it took 4.5 seconds for the algorithm to converge in Matlab 2010a on a PC with a quad core i7-2.67Gz processor and 6Gb of RAM memory. Then we ran SQUAR-EM-LA₂ with δ_2 equal to a much more conservative 10^{-6} . It converged in 58 iterations and took roughly 2 seconds. Table 1 gives the ML estimates and their standard errors found by SQUAR-EM-LA₂.

The speed of our algorithm allows us to perform bootstrap procedures. To illustrate this, we report bootstrap standard errors based on 1000 bootstrap samples in the last column of Table 1. For one of the bootstrap samples the algorithm did not converge due to the random choice of starting value. For comparison, we fit the same model using the SAS/GLIMMIX [SAS, 2005] procedure, which employs a restricted pseudo-likelihood method by default. The other estimates reported were obtained by using the Bayesian software package WinBUGS (D.J.Spiegelhalter et al., 1999, Crainiceanu et al., 2005, Crainiceanu and Goldsmith, 2010). The values given for WinBUGS are medians and standard deviations of the marginal posterior distributions obtained using the weakly-informative priors $a_0, a_1, b_1, b_2, b_3, c_{11}, c_{12}, c_{13} \sim N(0, 10^6)$

and $1/\sigma_\gamma^2, 1/\sigma_\nu^2, 1/\sigma_\omega^2 \sim U[0, 10^3]$. The estimates of all parameters except that of the constant agree. The SQUAR-EM-LA₂ estimate of a_0 is close to that of WinBUGS. Also, based on the estimates and their standard errors, it appears that there is a significant event by period interaction. One of the previously mentioned advantages of SQUAR-EM-LA₂ is the robustness to the choice of starting point. We tried a few randomly chosen ones and still the algorithm always converged to the reported estimate.

To briefly examine accuracy of EM-LA₂ for crossed designs, we consider the Salamander data from McCullagh and Nelder [1989, pages 439-450] (see Jiang, 2006, Section 4 for a comprehensive review). Here we consider the logit-normal GLMM described by Booth and Hobert [1999]. The estimation procedure in this case boils down to the approximation of six 20-dimensional integrals, and estimating the four-dimensional fixed-effect parameter $\boldsymbol{\beta} = (\boldsymbol{\beta}_{R/R}, \boldsymbol{\beta}_{R/W}, \boldsymbol{\beta}_{W/R}, \boldsymbol{\beta}_{W/W})$ as well as variance components $\boldsymbol{\sigma}^T = (\sigma_m, \sigma_f)^T$. There is an established agreement in literature about the ML estimates for this model (Booth and Hobert, 1999) as well as its equivalent fixed effects parametrization (Lee et al., 2006, Section 6.6.1 and Jiang, 2006, Section 4.4.3). We compare those ML estimates with EM-LA₂ estimates in Table 2. There is a small bias in the EM-LA₂ estimates. It is interesting to notice that the bias is not as severe as one might expect. We speculate that the exponential version of (27) would further reduce even this small bias.

5 Simulations

One of the key advantages of the proposed EM-LA₂ algorithm is its speed. This allows us to conduct extensive simulations to study its accuracy and convergence.

Inspired by Minnesota Health Plan Data from Section 4 we ran simulations under the following scenarios. Consider a Poisson loglinear model for counts y_{ij} modeled as $y_{ij}|\mathbf{u}_i^\Sigma \sim \text{Poisson}(\mu_{ij})$, where $\mu_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{z}_{ij}^T \mathbf{u}_i^\Sigma$. Our main interest lies in exploring the dependence between the dimensionality of the approximated integrals and accuracy of the EM-LA₂ estimates. Therefore, we define the design matrix for subject i as $\mathbf{Z}_i = [\mathbf{1}_{KL}, \mathbf{I}_K \otimes \mathbf{1}_L, \mathbf{1}_K \otimes \mathbf{I}_L]$ where $\mathbf{1}_m$ denotes the m -dimensional vector of ones and $\mathbf{I}_m$ denotes the $m \times m$ identity matrix. Defining $\mathbf{Z}_i$ in this way is equivalent to having K events during L periods for each subject i . Thus, the dimension of random vector, q , is equal to $(1 + K + L)$. The vector of random effects $\mathbf{u}_i^\Sigma$ is modeled as $N_q(\mathbf{0}, \boldsymbol{\Sigma})$ with a $q \times q$ diagonal matrix $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_{11} = \sigma_\gamma^2$, $\boldsymbol{\Sigma}_{ll} = \sigma_v^2$, $l = 2, \dots, K+1$ and $\boldsymbol{\Sigma}_{ll} = \sigma_\omega^2$, $l = K+2, \dots, K+L$. The design matrix $\mathbf{X}_i$ is defined as $[\mathbf{1}_{KL}, \mathbf{A}_K \otimes \mathbf{1}_L]$ with $\mathbf{A}_K = [\mathbf{I}_{K-1}, \mathbf{0}_{K-1}]^T$ and $\mathbf{0}_{K-1}$ being the $K-1$ dimensional vector of zeros. The matrix $\mathbf{X}_i$ does not have the same interpretation as in MHP data example. Instead, this choice allows us to set side by side the results for the estimated variance components from scenarios with different dimensions of the random effect. To stay consistent with MHP data example we set the number of subjects, I , to 121 with the

number of observations per subject $m = KL$ and the total number of observations $n = 121KL$.

We begin with the non-accelerated EM-LA₂ algorithm and consider a few scenarios: $K = 2, 4, 8$ and $L = 4, 8, 16$. The dimension of the fixed effect parameter β depends on K . So, given K , we set β be equal to the first K coordinates of vector $\beta^0 = (0.85, -0.15, -0.1, 0.4, 0.5, 0.25, 0.1, -0.1)$. Note that vector β^0 is the estimated vector of fixed effects (with rounded coordinates) for MHP data. The variance components $(\sigma_\gamma^2, \sigma_v^2, \sigma_\omega^2)$ are set to be close to their ML estimates for MHP data $(\sigma_\gamma^2, \sigma_v^2, \sigma_\omega^2) = (0.5^2, 0.6^2, 0.7^2)$. We simulated the data according to the above model $M = 100$ times for different combinations of K and L . Each coordinate of the starting point of EM-LA₂, $\psi^{(0)}$, was randomly generated from the uniform distribution on $[-1, 1]$. The stopping parameters δ_1 and δ_2 in (30) were chosen to be 10^{-4} . If the convergence was not declared after 1200 EM-steps we terminated the algorithm.

Table 3 reports the Monte Carlo means, N , of the termination step for each (K, L) combination. The main observation is that the dimension of random effect, q , has a strong impact on the speed of convergence of EM-LA₂. The larger the dimensionality of the random effect, the slower the convergence of the EM algorithm.

Next we ran SQUAR-EM-LA₂ under the same settings as above adding three additional scenarios: $K = 2, 4, 8$ and $L = 32$. For the accelerated version we set stopping parameter δ_2 to be a very conservative 10^{-6} . If the convergence was not declared after 2000 EM-steps we terminated the algorithm. We simulated the data according to the chosen scenarios $M = 1000$ times. Table 4 report the Monte Carlo means of the estimators, $(1/M) \sum_{m=1}^M \hat{\psi}_i(m)$, $i = 1, K$, where $\hat{\psi}(m)$ is the SQUAR-EM-LA₂ estimator from the m^{th} run. The Monte Carlo averages, N , for EM termination step are given in column three. MC average times (in seconds), T , are provided in column four. The results highlight impressive performance of SQUAR accelerator. Recall that the stopping parameter δ_2 was set to be two orders smaller for SQUAR-EM compared to the non-accelerated EM. Despite this, SQUAR-EM was faster by a factor ranging between five to ten, and this factor tends to be larger exactly in high-dimensional cases when the non-accelerated EM algorithm is extremely slow.

In terms of the accuracy of the estimates, we see that there is no major bias. Table 5 gives MC standard deviations for (27). We see that standard errors decrease with L increasing with fixed K , as one would expect. The dimension of the random effect, $q = 1 + K + L$, increases with K and L . Even though the information for each parameter grows with K and L , one might expect a growing error of approximation of the q -dimensional integrals. However, this does not seem to be the case for (27) which we find to be very encouraging.

Our simulations indicate that the SQUAR-EM-LA₂ algorithm is very stable. Even with random starting values our algorithm converged in all cases except for one run (out 1000) with $(K, L) = (2, 4)$ and $(K, L) = (8, 8)$. Overall, we see that SQUAR-EM-LA₂ gives strikingly accurate estimators.

6 Discussion

In this paper, closed form expressions for the standardized cumulants of GLMs are derived. In addition to their theoretical novelty, these expressions reduce the burden of computation by four orders of magnitude making it a quadratic complexity problem. This enables the calculation of second-order saddlepoint approximations in the cases previously thought to have a prohibitively large dimensionality. Using the similarity between the integrand in a GLMM likelihood and GLM loglikelihood we extended the result and obtained a closed form expression for the second-order Laplace approximation to a GLMM likelihood. This allowed us to develop a computationally highly efficient SQUAR-EM-LA₂ algorithm for fitting a GLMM with multivariate normal random effects. We applied SQUAR-EM-LA₂ to the Minnesota Health Plan data. The estimates produced by SQUAR-EM-LA₂ for this data set and in extensive simulation studies of Section 5 were very accurate highlighting the great potential of the approach. A Matlab implementation of SQUAR-EM-LA₂ algorithm can be found on the web-pages of the authors.

Our approach can be further extended to include more higher-order terms. The original second-order LA was obtained by ignoring expansion terms with a faster decay rate. However, we have shown that from computational perspective it is not a problem to keep them to further increase the accuracy of LA approximation. Another promising application is in adapting our techniques for h-likelihood developed for Hierarchical and Double Hierarchical GLMs (Lee et al., 2006). This rich class of models inherits the exact same dependence structure of random effects. The integrated nested Laplace approximations approach (Rue et al., 2009) could also considerably benefit from including a more accurate LA₂ approximation of Section 3.

The main limitation of SQUAR-EM-LA₂ approach is that it is not accurate for the data with crossed designs, and further research is needed to directly adapt the approach to these settings. As an alternative, one can directly maximize the LA₂ approximation of a GLMM likelihood developed in Section 3.

Acknowledgement

The authors would like to thank Ciprian Crainiceanu of Johns Hopkins Bloomberg School of Public Health for some valuable suggestions and his help with fitting the Minnesota data model in WinBUGS. Special thanks to Brian Caffo for giving useful insights into SQUAR-EM algorithm. Both authors received support from NSF-DMS grant 0805865 while working on this project. Vadim Zipunnikov was also partially supported by grant R01NS060910 from the National Institute of Neurological Disorders and Stroke.

Supplementary Materials

em_la_2_code: Matlab functions to perform GLMM fitting with SQUAR-EM-LA₂.

The archive also contains all data sets used as examples in the article. (GNU zipped file)

usage_of_em_la_2: Usage instructions for the Matlab package. (pdf)

Appendix

Here we provide the proofs for Theorems 1, 2, and 3, technical details of implementation of Newton-Raphson algorithm as well as some tables from Sections 4 and 5.

Proof of Theorem 1: We start with $\hat{\kappa}_4$. Recall that

$$\hat{K}_{t_1 t_2 t_3 t_4} = -l^{(4)}(\hat{\beta})_{t_1 t_2 t_3 t_4} = \sum_{i=1}^n w_i \hat{b}_i^{(4)} x_{it_1} x_{it_2} x_{it_3} x_{it_4}. \quad (34)$$

Let $\{\mathbf{e}_1, \dots, \mathbf{e}_p\}$ be the standard basis of $\mathbb{R}^p$ where p is the dimension of vectors $\{\mathbf{x}_i\}_{i=1}^n$. It is easy to see that

$$\hat{K}^{t_1 t_2} = (\hat{K}'')_{t_1 t_2}^{-1} = \hat{\mathbf{I}}_{t_1 t_2}^{-1} = \mathbf{e}_{t_1}^T \hat{\mathbf{I}}^{-1} \mathbf{e}_{t_2} \quad \text{and} \quad x_{it_1} = \mathbf{e}_{t_1}^T \mathbf{x}_i. \quad (35)$$

Combining (34) and (35) the standardized cumulant $\hat{\kappa}_4$ can be written as

$$\begin{aligned} \hat{\kappa}_4 &= \sum_{t_1, t_2, t_3, t_4} \hat{K}_{t_1 t_2 t_3 t_4} \hat{K}^{t_1 t_2} \hat{K}^{t_3 t_4} = \sum_{t_1, t_2, t_3, t_4} \sum_{i=1}^n w_i \hat{b}_i^{(4)} x_{it_1} x_{it_2} x_{it_3} x_{it_4} \hat{\mathbf{I}}_{t_1 t_2}^{-1} \hat{\mathbf{I}}_{t_3 t_4}^{-1} = \\ &= \sum_{i=1}^n w_i \hat{b}_i^{(4)} \sum_{t_1, t_2, t_3, t_4} \mathbf{x}_i^T \mathbf{e}_{t_1} \mathbf{e}_{t_1}^T \hat{\mathbf{I}}^{-1} \mathbf{e}_{t_2} \mathbf{e}_{t_2}^T \mathbf{x}_i \mathbf{x}_i^T \mathbf{e}_{t_3} \mathbf{e}_{t_3}^T \hat{\mathbf{I}}^{-1} \mathbf{e}_{t_4} \mathbf{e}_{t_4}^T \mathbf{x}_i. \end{aligned}$$

Notice that now we can separate the sums over t_1, t_2, t_3, t_4 and each sum of the form $\sum_t \mathbf{e}_t \mathbf{e}_t^T$ is the identity matrix. Therefore, we obtain $\hat{\kappa}_4 = \sum_{i=1}^n w_i \hat{b}_i^{(4)} \hat{\gamma}_{ii}^2$. Using the same separability argument, the result of the theorem can be shown for $\hat{\kappa}_{13}^2$ and $\hat{\kappa}_{23}^2$.

Proof of Theorem 2: Using the exact same separability argument as in the proof of Theorem 1 we get the result stated in Theorem 2.

Proof of Theorem 3: Let $\{\mathbf{e}_1, \dots, \mathbf{e}_q\}$ be the standard basis of $\mathbb{R}^q$. Let $\boldsymbol{\xi}_{ij}^T = \mathbf{z}_{ij}^T \mathbf{D}$. We can write

$$\begin{aligned} \hat{\zeta}^{t_1} &= \sum_{j=1}^{n_i} w_{ij} [y_{ij} - b'(\theta_{ij})] \boldsymbol{\xi}_{ij}^T \mathbf{e}_{t_1}, \quad \hat{h}_{t_1 t_2} = \mathbf{e}_{t_1}^T [h''(\hat{\mathbf{u}}_i, \boldsymbol{\psi}^{(s)})]^{-1} \mathbf{e}_{t_2}, \\ \hat{h}^{t_2 t_3 t_4} &= \sum_{j=1}^{n_i} w_{ij} b^{(3)}(\theta_{ij}) \boldsymbol{\xi}_{ij}^T \mathbf{e}_{t_2} \boldsymbol{\xi}_{ij}^T \mathbf{e}_{t_3} \boldsymbol{\xi}_{ij}^T \mathbf{e}_{t_4}, \quad \hat{h}_{t_3 t_4} = \mathbf{e}_{t_3}^T [h''(\hat{\mathbf{u}}_i, \boldsymbol{\psi}^{(s)})]^{-1} \mathbf{e}_{t_4}. \end{aligned} \quad (36)$$

Putting (36) in (28) and summing over t_1, t_2, t_3, t_4 , we get the expression for $\tau_{21}(\hat{\mathbf{u}}_i, \boldsymbol{\psi}, \boldsymbol{\psi}^{(s)})$. Similarly, we get the result for $\tau_{22}(\hat{\mathbf{u}}_i, \boldsymbol{\psi}, \boldsymbol{\psi}^{(s)})$.

Details on the Newton-Raphson algorithm for EM-LA₂: Maximization of

the $\tilde{Q}(\boldsymbol{\psi}|\boldsymbol{\psi}^{(s)})$ at step s is done using Newton-Raphson algorithm: $\boldsymbol{\psi}^{(k+1)} = \boldsymbol{\psi}^{(k)} - [\tilde{Q}''_{\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\psi}^{(k)}|\boldsymbol{\psi}^{(s)})]^{-1}\tilde{Q}'_{\boldsymbol{\psi}}(\boldsymbol{\psi}^{(k)}|\boldsymbol{\psi}^{(s)})$. Below we will find all the components involved in it. We begin with the function $\tilde{Q}'_{\boldsymbol{\psi}}(\boldsymbol{\psi}|\boldsymbol{\psi}^{(s)})$

$$\begin{aligned}\tilde{Q}'_{\boldsymbol{\psi}}(\boldsymbol{\psi}|\boldsymbol{\psi}^{(s)}) &= \sum_{i=1}^n \left(\zeta'_{\boldsymbol{\psi}}(\hat{\mathbf{u}}_i, \boldsymbol{\psi}) + (\tau_2)'_{\boldsymbol{\psi}}(\hat{\mathbf{u}}_i, \boldsymbol{\psi}, \boldsymbol{\psi}^{(s)}) \right), \\ \tilde{Q}''_{\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\psi}|\boldsymbol{\psi}^{(s)}) &= \sum_{i=1}^n \left(\zeta''_{\boldsymbol{\psi}\boldsymbol{\psi}}(\hat{\mathbf{u}}_i, \boldsymbol{\psi}) + (\tau_2)''_{\boldsymbol{\psi}\boldsymbol{\psi}}(\mathbf{u}_i, \boldsymbol{\psi}, \boldsymbol{\psi}^{(s)}) \right).\end{aligned}$$

For the function $\zeta(\hat{\mathbf{u}}_i, \boldsymbol{\psi})$ the first and the second derivatives with respect to $\boldsymbol{\psi}$ are given by

$$\zeta'_{\boldsymbol{\psi}}(\hat{\mathbf{u}}_i, \boldsymbol{\psi}) = \sum_{j=1}^{n_i} w_{ij} [y_{ij} - b'(\theta_{ij})] \hat{\mathbf{x}}_{ij}, \quad \zeta''_{\boldsymbol{\psi}\boldsymbol{\psi}}(\hat{\mathbf{u}}_i, \boldsymbol{\psi}) = - \sum_{j=1}^{n_i} w_{ij} b''(\theta_{ij}) \hat{\mathbf{x}}_{ij} \hat{\mathbf{x}}_{ij}^T.$$

Omitting the dependence on $\hat{\mathbf{u}}_i$ and $\boldsymbol{\psi}^{(s)}$, the first derivatives of $\tau_2(\hat{\mathbf{u}}_i, \boldsymbol{\psi}, \boldsymbol{\psi}^{(s)})$ terms with respect to $\boldsymbol{\psi}$ are given by

$$\begin{aligned}(\tau_{22})'_{\boldsymbol{\psi}}(\boldsymbol{\psi}) &= - \sum_{j=1}^{n_i} w_{ij} b^{(3)}(\theta_{ij}) \gamma_{jj}(\boldsymbol{\psi}) \hat{\mathbf{x}}_{ij} - \sum_{j=1}^{n_i} w_{ij} b''(\theta_{ij}) [\gamma_{jj}(\boldsymbol{\psi})]'_{\boldsymbol{\psi}} \\ (\tau_{21})'_{\boldsymbol{\psi}}(\boldsymbol{\psi}) &= - \sum_{j=1}^{n_i} w_{ij} b''(\theta_{ij}) \lambda_j(\boldsymbol{\psi}) \hat{\mathbf{x}}_{ij} + \sum_{j=1}^{n_i} w_{ij} [y_{ij} - b'(\theta_{ij})] [\lambda_j(\boldsymbol{\psi})]'_{\boldsymbol{\psi}},\end{aligned}$$

with

$$[\gamma_{jj}(\boldsymbol{\psi})]'_{\boldsymbol{\psi}} = \begin{bmatrix} \mathbf{0}_{p \times 1} \\ 2\mathbf{P}(\mathbf{z}_{ij})[\hat{h}'']^{-1}\mathbf{P}^T(\mathbf{z}_{ij})\boldsymbol{\sigma} \end{bmatrix}, \quad [\lambda_j(\boldsymbol{\psi})]'_{\boldsymbol{\psi}} = \begin{bmatrix} \mathbf{0}_{p \times 1} \\ \mathbf{P}(\mathbf{z}_{ij})[\hat{h}'']^{-1}\boldsymbol{\Gamma}_i^{(s)} \end{bmatrix}.$$

and $\boldsymbol{\Gamma}_i^{(s)} = \sum_{l=1}^{n_i} w_{il} b_{il}^{(3)} \gamma_{ll}^{(s)} \mathbf{D}^{(s)T} \mathbf{z}_{il}$. The second derivatives are therefore

$$\begin{aligned}(\tau_{22})''_{\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\psi}) &= - \sum_{j=1}^{n_i} w_{ij} b^{(4)}(\theta_{ij}) \gamma_{jj}(\boldsymbol{\psi}) \hat{\mathbf{x}}_{ij} \hat{\mathbf{x}}_{ij}^T - \sum_{j=1}^{n_i} w_{ij} b^{(3)}(\theta_{ij}) (\hat{\mathbf{x}}_{ij} [\gamma_{jj}(\boldsymbol{\psi})]'_{\boldsymbol{\psi}} + [\gamma_{jj}(\boldsymbol{\psi})]''_{\boldsymbol{\psi}} \hat{\mathbf{x}}_{ij}^T) \\ &\quad - \sum_{j=1}^{n_i} w_{ij} b^{(2)}(\theta_{ij}) [\gamma_{jj}(\boldsymbol{\psi})]''_{\boldsymbol{\psi}\boldsymbol{\psi}} \\ (\tau_{21})''_{\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\psi}) &= - \sum_{j=1}^{n_i} w_{ij} b^{(3)}(\theta_{ij}) \lambda_j(\boldsymbol{\psi}) \hat{\mathbf{x}}_{ij} \hat{\mathbf{x}}_{ij}^T - \sum_{j=1}^{n_i} w_{ij} b''(\theta_{ij}) (\hat{\mathbf{x}}_{ij} [\lambda_j(\boldsymbol{\psi})]'_{\boldsymbol{\psi}}^T + [\lambda_j(\boldsymbol{\psi})]''_{\boldsymbol{\psi}} \hat{\mathbf{x}}_{ij}^T)\end{aligned}$$

with

$$[\gamma_{jj}(\boldsymbol{\psi})]''_{\boldsymbol{\psi}\boldsymbol{\psi}} = \begin{bmatrix} \mathbf{0}_{p \times p} & \mathbf{0}_{p \times q^*} \\ \mathbf{0}_{q^* \times p} & 2\mathbf{P}(\mathbf{z}_{ij})[\hat{h}'']^{-1}\mathbf{P}^T(\mathbf{z}_{ij}) \end{bmatrix}.$$

This completes the construction of the NR algorithm.

Details on the information matrix for EM-LA₂: Here we give details on calculation of information matrix (31). For vector function $\zeta(\mathbf{u}_i) = -h'_{\boldsymbol{\psi}}(\mathbf{u}_i)$ in the formulas below, the first line corresponds to fixed effect parameters $\boldsymbol{\psi}_r, r = 1, \dots, p$ and the second to variance components $\boldsymbol{\psi}_r, r = p+1, \dots, p+q^*$. Function $\zeta(\mathbf{u}_i)$ are given by

$$\zeta_r(\mathbf{u}_i) = \begin{cases} \sum_{j=1}^{n_i} w_{ij} (y_{ij} - b'(\theta_{ij})) x_{ijr}, \\ \sum_{j=1}^{n_i} w_{ij} (y_{ij} - b'(\theta_{ij})) P_r(\mathbf{z}_{ij}) \mathbf{u}_i. \end{cases}$$

	With Interaction					
	GLIMMIX		WinBUGS		SQUAR-EM-LA ₂	
a_0	0.961	(0.104)	0.844	(0.109)	0.858	(0.096) (0.134)
a_1	-0.164	(0.106)	-0.160	(0.110)	-0.166	(0.124) (0.124)
b_1	-0.089	(0.109)	-0.085	(0.111)	-0.091	(0.094) (0.144)
b_2	0.394	(0.104)	0.422	(0.110)	0.414	(0.097) (0.133)
b_3	0.468	(0.103)	0.498	(0.110)	0.491	(0.108) (0.124)
c_{11}	0.240	(0.103)	0.243	(0.103)	0.246	(0.097) (0.131)
c_{12}	0.101	(0.095)	0.102	(0.097)	0.104	(0.080) (0.124)
c_{13}	-0.084	(0.096)	-0.088	(0.097)	-0.085	(0.099) (0.098)
σ_γ	0.491	(0.081)	0.511	(0.078)	0.529	(0.083) (0.092)
σ_ν	0.578	(0.048)	0.605	(0.053)	0.593	(0.054) (0.059)
σ_ω	0.593	(0.034)	0.627	(0.038)	0.618	(0.040) (0.038)

Table 1: Parameter estimates for the Poisson linear mixed effects model (33) obtained by SQUAR-EM-LA₂ and using the SAS/GLIMMIX and WinBUGS software packages. Last panel reports SQUAR-EM-LA₂ estimates along with ML standard errors (first value in parenthesis) and bootstrap standard errors(second value in parenthesis).

	$\beta_{R/R}$	$\beta_{R/W}$	$\beta_{W/R}$	$\beta_{W/W}$	σ_f	σ_m
EM-LA ₂	0.986	0.305	-1.833	0.979	1.026	0.948
ML	1.030	0.320	-1.950	0.990	1.183	1.118

Table 2: Maximum likelihood estimates for the logit-normal model for Salamander data obtained using EM-LA₂.

where $P_r(\mathbf{z}_{ij})$ is the r^{th} row of matrix $P_r(\mathbf{z}_{ij})$. Hence, the first derivatives with respect to $\mathbf{u}_{it_1}$ are

$$\zeta_r^{t_1}(\mathbf{u}_i) = \begin{cases} -\sum_{j=1}^{n_i} w_{ij} b^{(2)}(\theta_{ij}) x_{ijr} \boldsymbol{\xi}_{ijt_1}, \\ -\sum_{j=1}^{n_i} w_{ij} b^{(2)}(\theta_{ij}) P_r(\mathbf{z}_{ij}) \mathbf{u}_i \boldsymbol{\xi}_{ijt_1} + \sum_{j=1}^{n_i} w_{ij} (y_{ij} - b'(\theta_{ij})) P_{rt_1}(\mathbf{z}_{ij}). \end{cases}$$

Combining $\zeta_r^{t_1}(\mathbf{u}_i)$ above and (28) we get

$$\tau_{21r} = \begin{cases} -\sum_{j=1}^{n_i} w_{ij} b^{(2)}(\theta_{ij}) x_{ijr} \lambda_j, \\ -\sum_{j=1}^{n_i} w_{ij} b^{(2)}(\theta_{ij}) P_r(\mathbf{z}_{ij}) \mathbf{u}_i \lambda_j + \sum_{j=1}^{n_i} w_{ij} (y_{ij} - b'(\theta_{ij})) P_r(\mathbf{z}_{ij}) \Gamma_i. \end{cases}$$

The second derivatives with respect to $\mathbf{u}_{it_1}$ and $\mathbf{u}_{it_2}$ are given by

$$\zeta_r^{t_1 t_2}(\mathbf{u}_i) = \begin{cases} -\sum_{j=1}^{n_i} w_{ij} b^{(3)}(\theta_{ij}) x_{ijr} \boldsymbol{\xi}_{ijt_1} \boldsymbol{\xi}_{ijt_2}, \\ -\sum_{j=1}^{n_i} w_{ij} b^{(3)}(\theta_{ij}) P_r(\mathbf{z}_{ij}) \mathbf{u}_i \boldsymbol{\xi}_{ijt_1} \boldsymbol{\xi}_{ijt_2} - \sum_{j=1}^{n_i} w_{ij} b^{(2)}(\theta_{ij}) (P_{rt_1}(\mathbf{z}_{ij}) \boldsymbol{\xi}_{ijt_2} + P_{rt_2}(\mathbf{z}_{ij}) \boldsymbol{\xi}_{ijt_1}). \end{cases}$$

where $P_{rt_1}(\mathbf{z}_{ij})$ is the (r, t_1) element of the matrix $P(\mathbf{z}_{ij})$. Using (28) we obtain

$$\tau_{22r} = \begin{cases} -\sum_{j=1}^{n_i} w_{ij} b^{(3)}(\theta_{ij}) \gamma_{jj}, \\ -\sum_{j=1}^{n_i} w_{ij} b^{(3)}(\theta_{ij}) \gamma_{jj} P_r(\mathbf{z}_{ij}) \mathbf{u}_i - 2 \sum_{j=1}^{n_i} w_{ij} b^{(2)}(\theta_{ij}) P_r(\mathbf{z}_{ij}) [\hat{h}'']^{-1} \mathbf{D} \mathbf{z}_{ij}. \end{cases}$$

(K, L)	q	N	(K, L)	q	N	(K, L)	q	N
(2,4)	7	193	(4,4)	9	276	(8,4)	13	438
(2,8)	11	267	(4,8)	13	367	(8,8)	17	574
(2,16)	19	406	(4,16)	21	557	(8,16)	25	806

Table 3: The results of running the non-accelerated EM-LA₂ algorithm on 100 simulated Poisson loglinear data sets. For each (K, L) combination: q is the dimension of random effect, N is the Monte Carlo average for the EM termination step.

(K, L)	q	N	T	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	σ_γ	σ_v	σ_ω
(2, 4)	7	36	1.1	0.850	-0.150	-0.100	0.400	0.500	0.250	0.100	-0.100	0.500	0.600	0.700
(2, 8)	11	44	2.2	0.857	-0.152	-	-	-	-	-	-	0.475	0.594	0.693
(2, 16)	19	53	3.3	0.851	-0.149	-	-	-	-	-	-	0.479	0.597	0.693
(2, 32)	35	61	9.2	0.851	-0.151	-	-	-	-	-	-	0.485	0.598	0.695
(4, 4)	9	39	1.8	0.852	-0.152	-	-	-	-	-	-	0.488	0.594	0.693
(4, 8)	13	44	2.2	0.852	-0.149	-0.101	0.398	-	-	-	-	0.489	0.595	0.696
(4, 16)	21	55	5.2	0.848	-0.148	-0.097	0.401	-	-	-	-	0.488	0.595	0.698
(4, 32)	37	67	15.6	0.852	-0.152	-0.101	0.396	-	-	-	-	0.489	0.597	0.697
(8, 4)	13	52	2.5	0.849	-0.146	-0.100	0.402	-	-	-	-	0.494	0.596	0.697
(8, 8)	17	50	4.6	0.849	-0.147	-0.097	0.400	0.502	0.251	0.099	-0.101	0.489	0.595	0.698
(8, 16)	25	58	11.3	0.847	-0.151	-0.098	0.401	0.500	0.257	0.101	-0.098	0.491	0.597	0.699
(8, 32)	41	73	31.5	0.854	-0.154	-0.102	0.396	0.494	0.243	0.095	-0.105	0.495	0.597	0.699
(8, 32)	41	73	31.5	0.849	-0.151	-0.098	0.400	0.504	0.254	0.100	-0.100	0.497	0.597	0.699

Table 4: Means of SQUAR-EM-LA₂ parameter estimates in 1000 simulated Poisson log-linear data sets. The dimension of random effect, q , is given in the second column. The third column reports the means of EM termination step. The average time (in seconds), T , is provided in column four.

(K, L)	q	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	σ_γ	σ_v	σ_ω
(2, 4)	7	0.850	-0.150	-0.100	0.400	0.500	0.250	0.100	-0.100	0.500	0.600	0.700
(2, 8)	11	0.082	0.090	-	-	-	-	-	-	0.091	0.051	0.038
(2, 16)	19	0.078	0.085	-	-	-	-	-	-	0.077	0.047	0.026
(2, 32)	35	0.077	0.079	-	-	-	-	-	-	0.074	0.040	0.018
(4, 4)	9	0.072	0.080	-	-	-	-	-	-	0.063	0.039	0.012
(4, 8)	13	0.083	0.085	0.086	0.082	-	-	-	-	0.067	0.030	0.032
(4, 16)	21	0.077	0.082	0.081	0.083	-	-	-	-	0.056	0.026	0.021
(4, 32)	37	0.076	0.081	0.080	0.083	-	-	-	-	0.051	0.024	0.014
(8, 4)	13	0.073	0.078	0.079	0.076	-	-	-	-	0.047	0.023	0.010
(8, 8)	17	0.084	0.089	0.091	0.086	0.086	0.085	0.088	0.088	0.057	0.019	0.030
(8, 16)	25	0.078	0.083	0.084	0.081	0.081	0.081	0.081	0.083	0.049	0.017	0.019
(8, 32)	41	0.075	0.080	0.081	0.081	0.079	0.081	0.078	0.082	0.042	0.016	0.013
(8, 32)	41	0.074	0.078	0.079	0.079	0.077	0.079	0.081	0.080	0.040	0.015	0.009

Table 5: Standard errors of SQUAR-EM-LA₂ parameter estimates in 1000 simulated Poisson loglinear data sets. The dimension of random effect, q , is given in the second column.

References

- James G. Booth and James P. Hobert. Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. *Journal of the Royal Statistical Society*, B 61:265–285, 1999.
- James G. Booth, George Casella, Herwig Friedl, and James P. Hobert. Negative binomial loglinear mixed models. *Statistical Modelling*, 3:179–191, 2003.
- N.E. Breslow and D.G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25, 1993.
- N.E. Breslow and X. Lin. Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika*, 82:81–91, 1995.
- Ronald W. Butler. *Saddlepoint Approximations with Applications*. Number 22 in Cambridge Series on Statistical and Probabilistic Mathematics. Cambridge University Press, 2007.
- Brian S. Caffo, Wolfgang Jank, and Galin L. Jones. Ascent-based monte carlo expectation-maximization. *Journal of the Royal Statistical Society*, B 67:235–251, 2005.
- Ciprian M. Crainiceanu and Jeffrey Goldsmith. Bayesian functional data analysis using winbugs. *Journal of Statistical Software*, 32(11):1–33, 2010.
- Ciprian M. Crainiceanu, David Ruppert, and M.P. Wand. Bayesian analysis for penalized spline regression using winbugs. *Journal of Statistical Software*, 14(14):1–24, 2005.
- E. Z. Demidenko. *Mixed Models: Theory and Applications*. Wiley-Interscience, 2004.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society (with discussion)*, B 39:1–39, 1977.
- Rafael E. Diaz. Comparison of pql and laplace 6 estimates of hierarchical linear models when comparing groups of small incident rates in cluster randomised trials. *Computational Statistics and Data Analysis*, 51:2871–2888, 2007.
- D.J.Spiegelhalter, A.Thomas, and N.G.Best. Winbugs version 1.2 user manual. *MRC Biostatistics Unit*, 1999.
- J. Durbin and S.J. Koopman. Monte carlo maximum likelihood estimation for non-gaussian state space models. *Biometrika*, 84:669–684, 1997.
- Evangelos Evangelou, Zhengyuan Zhu, and Richard L. Smith. Asymptotic inference for spatial glmm using high-order laplace approximation. *submitted*, 2008.

- C.J. Geyer and E.A. Thompson. Constrained monte carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society, Series B*, 54:657–699, 1992.
- James P. Hobert. Hierarchical models: a current computational perspective. *Journal of the American Statistical Association*, 95(542):1312–1316, 2000.
- Jiming Jiang. *Linear and Generalized Linear Mixed Models and Their Applications*. Springer, 2006.
- John E. Kolassa. *Series Approximation Methods in Statistics*. Springer, 2006.
- F. Kuo, W. Dunsmuir, I. Sloan, M. Wand, and R. Womersley. Quasi-monte carlo for highly structured generalised response models. *Methodology and Computing in Applied Probability*, 10:239–275, 2008.
- Youngjo Lee, John Nelder, and Yudi Pawitan. *Generalized Linear Models with Random Effects*. Chapman and Hall, 1 edition, 2006.
- X. Lin and N.E. Breslow. Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, 91:1007–1016, 1996.
- C. Liu, D.B. Rubin, and Y.N. Wu. Parameter expansion to accelerate the em: Px-em algorithm. *Biometrika*, 85:755–770, 1998.
- Peter McCullagh. *Tensor Methods in Statistics*. Chapman and Hall, 1987.
- Peter McCullagh and John A. Nelder. *Generalized Linear Models*. Chapman and Hall, 2 edition, 1989.
- Charles E. McCulloch. Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association*, 89:330–335, 1994.
- Charles E. McCulloch. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92:162–170, 1997.
- Charles E. McCulloch and Shayle R. Searle. *Generalized, Linear and Mixed Models*. John Wiley & Sons, 2001.
- X.L. Meng and D.B. Rubin. The em algorithm - an old folk-song sung to a fast new tune. *Journal of Royal Statistical Society, Series B*, 59:511–567, 1997.
- J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society, A* 135:370–384, 1972.
- Edmond Ng, James Carpenter, Harvey Goldstein, and Joh Rasbash. Estimation in generalized linear mixed models with binary outcomes by simulated maximum likelihood. *Statistical Modelling*, 6:23–42, 2006.

- Maengseok Noh and Youngjo Lee. Reml estimation for binary data in glmms. *Journal of Multivariate Analysis*, 98:896–915, 2007.
- Stephen Raudenbush, M. Yang, and Matheos Yosef. 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, 2000.
- Stephen Raudenbush, A. Bryk, Y. Cheong, R. Congdon, and M. du Toit. *HLM 6: Hierarchical Linear and Nonlinear Modelling*. Scientific Software International, 2 edition, 2004.
- Christophe Roland and Ravi Varadhan. New iterative schemes for nonlinear fixed point problems, with applications to problems with bifurcations and incomplete-data problems. *Applied Numerical Mathematics*, 55:215–226, 2005.
- Havard Rue, Sara Martino, and Nicolas Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society, B*, 71(2):319–392, 2009.
- SAS. Sas/stat software: The glimmix procedure, documentation. *SAS Institute Inc.*, 2005.
- R. Schall. Estimation in generalized linear models with random effects. *Biometrika*, 78:719–727, 1991.
- Elizabeth D. Schifano, Robert L. Strawderman, and Martin T. Wells. Majorization-minimization algorithms for nonsmoothly penalized objective functions. *Electronic Journal of Statistics*, 4:1258–1299, 2010.
- Zhenming Shun. Another look at the salamander mating data: A modified laplace approximation approach. *Journal of the American Statistical Association*, 92:341–349, 1997.
- Zhenming Shun and Peter McCullagh. Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society, B*, 57, N4:749–760, 1995.
- Brian M. Steele. A modified em algorithm for estimation in generalized mixed models. *Biometrics*, 52:1295–1310, 1996.
- L. Tierney and J.B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81:82–86, 1986.
- Ravi Varadhan and Christophe Roland. Simple and globally convergent methods for accelerating the convergence of any em algorithm. *Scandinavian Journal of Statistics*, 35:335–353, 2008.
- L. A. Waller and D. Zelterman. Loglinear modeling with the negative multinomial distribution. *Biometrics*, 53:971–982, 1997.

- R. Wolfinger and M. O'Connell. Generalized linear mixed models: A pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, 48:233–243, 1993.
- C.F.J. Wu. On convergence properties of the em algorithm. *Annals of Statistics*, 11: 95–103, 1983.