Monte Carlo EM for Generalized Linear Mixed Models using Randomized Spherical Radial Integration

Vadim V. Zipunnikov* and James G. Booth†

October 19, 2006

Abstract

The expectation-maximization algorithm has been advocated recently by a number of authors for fitting generalized linear mixed models. Since the E-step typically involves analytically intractable integrals, one approach is to approximate them by Monte Carlo methods. However, in practice, the Monte Carlo sample sizes required for convergence are often prohibitive. In this paper we show how randomized spherical-radial integration (Genz and Monahan, 1997) can be implemented in such cases, and can dramatically reduce the computational burden of implementing EM. After a standardizing transformation, a change to polar coordinates results in a double integral consisting of a one dimensional integral on the real line and a multivariate integral on the surface of a unit sphere. Randomized quadratures are used to approximate both of them. An attractive feature of the randomized spherical-radial rule is that its implementation only involves generating from standard probability distributions. The resulting approximation at the E-step has the form of a fixed effects generalized linear model likelihood and so a standard iteratively reweighted least squares procedure may be utilized for the M-step. We illustrate the method by fitting models to two well-known data sets, and compare our results with those of other authors.

Key Words: Fisher Scoring; Randomized Quadratures; Salamander data.

*Vadim Zipunnikov is Ph.D. candidate, Field of Statistics, Cornell University (email: vvz2@cornell.edu)

†James Booth is Professor, Biological Statistics and Computational Biology, Cornell University, Ithaca, NY, 14853

1 Introduction

The class of generalized linear models (GLM), introduced by Nelder and Wedderburn (1972), includes many popular statistical methods as special cases, such as logistic regression for binary responses, loglinear models for counts, as well as normal theory linear models. McCullagh and Nelder (1989) provide an extensive introduction to the topic. A restriction is that the GLM assumes that the observations are independent of one another, which is not the case, for instance, in longitudinal studies, or if the observations are clustered. Generalized linear mixed models (GLMMs) extend the GLM class by including random effects in their linear predictor. The result is a mixed model containing both fixed effects and random effects. Recent reviews of generalized linear mixed models and related techniques may be found in McCulloch and Searle (2001), Demidenko (2004), Hobert (2000), and Agresti et al. (2000).

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. This paper concerns the use of an approximation at the E-step of the expectation-maximization (EM) algorithm (Dempster et al., 1977). As with the likelihood the E-step involves an intractable integral, and while standard numerical integration techniques can be utilized in low dimensions problems, it is common in practice for the dimension to be too large for such methods. One solution is to use Monte Carlo approximation, as proposed by Wei and Tanner (1990). This approach, which is known as Monte Carlo EM (MCEM), has been applied in the GLMM context in several recent papers including McCulloch (1994,1997), Booth and Hobert (1999) and Caffo et al. (2005).

The main contribution of this paper is the use of randomized spherical-radial (SR) integration rules, developed in a series of paper by Genz and Monahan (1998,1998,1999), at the E-step of the EM algorithm in the GLMM context. These rules have been shown to dramatically outperform standard integration rules in many situations, resulting in remarkably accurate approximations even in relatively high dimensional problems.

The implementation of MCEM using SR rules described here is an alternative to their use to directly approximate the likelihood function, as proposed by Clarkson and Zhan (2002). The issue of which approach is to be preferred boils down to the pros and cons of EM versus direct maximization. For example, the EM algorithm is known to be very stable in a broad range of problems, and the numerical examples discussed later in this paper appear to substantiate this in the GLMM context. Also, the M-step of EM in the GLMM context is equivalent to fitting a GLM, and can therefore be accomplished using the standard iteratively reweighted least squares

(IRLS) algorithm.

The use of SR rules at the E-step of the MCEM algorithm substantially expands the applicability of the method by reducing the sample size required for accurate Monte Carlo approximation. Furthermore, the randomized SR rules are simpler to apply than competing methods cited above in that they only involve simulation from two standard distributions. The end result is an algorithm that is relatively simple, generally applicable, and practical to implement.

The structure of the article is as follows. In the next section we give a general description of the GLMM and introduce an example which we use to illustrate notation and methodology throughout the paper. Section 3 outlines the expectation-maximization algorithm in a GLMM setting. Spherical-radial integration rules are explained in Section 4. Section 5 contains a simulation study comparing MCEM with the direct maximization approach of Clarkson and Zhan (2002). In section 6, we illustrate the proposed algorithm by fitting GLMMs to three well-known datasets. We conclude with some discussion in Section 7.

2 Generalized Linear Mixed Models

2.1 The model

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

$$\mathbf{u}_i^{\Sigma} = (u_{i1}^{\Sigma}, \ldots, u_{iq}^{\Sigma})^T \sim \text{i.i.d. N}_q(\mathbf{0}, \mathbf{\Sigma}), \qquad i = 1, \ldots, n\,,$$

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

$$g(\mu_{ij}) = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{u}_i^{\Sigma}\,, \tag{2.1}$$

where $\boldsymbol{\beta}$ is a p-dimensional parameter and $g(\cdot)$ is a link function. Since $\mathbf{\Sigma}$ is positive definite there exists a unique $q \times q$ lower-triangular matrix $\mathbf{D}$ with positive diagonal entries such that $\mathbf{\Sigma} = \mathbf{D}\mathbf{D}^T$, and hence

$$\mathbf{u}_i^{\Sigma} \overset{d}{=} \mathbf{D}\mathbf{u}_i, \quad \text{where} \quad \mathbf{u}_i \sim \text{i.i.d. N}_q(0, \mathbf{I}_q), \quad i = 1, \ldots, n\,.$$

Therefore, without loss of generality, we may consider the distributionally equivalent form,

$$g(\mu_{ij}) = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{D}\mathbf{u}_i \tag{2.2}$$

in place of (2.1) (Demidenko, 2004, page 411). Notice that we may write

$$\mathbf{z}_{ij}^T \mathbf{D} \mathbf{u}_i = vech(\mathbf{z}_{ij} \mathbf{u}_i^T)^T vech(\mathbf{D}) \, ,$$

where $vech$ is vectorization operation. However, the matrix $\mathbf{D}$ is often characterized by a few non-zero entries. Let $\boldsymbol{\sigma}$ be a q_*-dimensional vector containing these elements. Then, there exists a $q(q+1)/2 \times q_*$ matrix $\mathbf{G}$ of rank q_* such that

$$vech(\mathbf{D}) = \mathbf{G} \boldsymbol{\sigma}$$

If $\boldsymbol{\xi}_{ij}$ is such that

$$\boldsymbol{\xi}_{ij} = \mathbf{G}^T \; vech(\mathbf{z}_{ij} \mathbf{u}_i^T) \, ,$$

then (2.2) can be rewritten as

$$g(\mu_{ij}) = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \boldsymbol{\xi}_{ij}^T \boldsymbol{\sigma} \, . \tag{2.3}$$

It is sometimes more convenient to use a shorter form

$$g(\mu_{ij}) = \tilde{\mathbf{x}}_{ij}^T \boldsymbol{\psi} \, ,$$

where $\tilde{\mathbf{x}}_{ij} = (\mathbf{x}_{ij}^T, \boldsymbol{\xi}_{ij}^T)^T$ and $\boldsymbol{\psi} = (\boldsymbol{\beta}^T, \boldsymbol{\sigma}^T)^T$ is a $(p + q_*)$-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 parameter $\boldsymbol{\psi}$ is therefore

$$L(\boldsymbol{\psi}; \mathbf{y}) = \int_{\mathbb{R}^q} f(\mathbf{y}|\mathbf{u}; \boldsymbol{\psi}) \phi(\mathbf{u}, \mathbf{I}_{nq}) d\mathbf{u} \tag{2.4}$$

where $\mathbf{y} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_n^T)^T$ and $\mathbf{u} = (\mathbf{u}_1^T, \ldots, \mathbf{u}_n^T)^T$, $\phi(\mathbf{u}, \mathbf{I}_{nq}) = \prod_{i=1}^n \prod_{r=1}^q \phi(u_{ir})$, where $\phi(\cdot)$ is the standard normal density, and

$$f(\mathbf{y}|\mathbf{u}; \boldsymbol{\psi}) = \prod_{i=1}^n f(\mathbf{y}_i|\mathbf{u}_i; \boldsymbol{\psi}) = \prod_{i=1}^n \prod_{j=1}^{n_i} \exp\{w_{ij}[\theta_{ij} y_{ij} - b(\theta_{ij})] + c(y_{ij})\}$$

Since θ_{ij} is usually a nonlinear function of $\mathbf{u}_i$, in most practical cases the integral in (2.4) cannot be evaluated explicitly. Therefore, the maximization of (2.4) cannot be accomplished without an approximation of the integral.

The follow subsection describes a specific application involving a GLMM with a multivariate random effect.

2.2 Minnesota health plan data

Waller and Zelterman (1997) reported data 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 $y_{ikl}|\mathbf{u}_i^\Sigma \sim Poisson(\mu_{ikl})$, and

$$\log \mu_{ikl} = a_0 + a_k + b_l + c_{kl} + \gamma_i + \upsilon_{ik} + \omega_{il}, \quad k = 1, 2, \text{ and } \quad l = 1, 2, 3, 4\,, \tag{2.5}$$

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×period interaction, γ_i is a random effect associated with subject i, υ_{ik} is a random subject×event interaction, and ω_{il} is a random subject×period interaction. The model therefore involves a 7-dimensional random effect

$$\mathbf{u}_i^\Sigma = (\gamma_i, \upsilon_{i1}, \upsilon_{i2}, \omega_{i1}, \omega_{i2}, \omega_{i3}, \omega_{i4}), \quad i = 1, \ldots, 121\,,$$

associated with the subject i. We suppose that

$$\mathbf{u}_i^\Sigma \sim \text{i.i.d. } N_7(\mathbf{0}, \boldsymbol{\Sigma}),\ i = 1, \ldots, 121$$

where

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_\gamma^2 & 0 & 0 \\ 0 & \sigma_\upsilon^2 \mathbf{I}_2 & 0 \\ 0 & 0 & \sigma_\omega^2 \mathbf{I}_4 \end{pmatrix}$$

We achieve identifiability by setting $a_2 = b_4 = c_{14} = c_{21} = c_{22} = c_{23} = c_{24} = 0$. The fixed effects parameter in (2.3) is then

$$\boldsymbol{\beta} = (a_0, a_1, b_1, b_2, b_3, c_{11}, c_{12}, c_{13})\,.$$

To eliminate the double index kl, and express the model in the form in (2.3), we consider a new index $j = 4(k-1)+l$. Accordingly, $(y_{i1}, \ldots, y_{i4}, y_{i5}, \ldots, y_{i8}) = (y_{i11}, \ldots, y_{i14}, y_{i21}, \ldots, y_{i24})$ and $(\mu_{i1}, \ldots, \mu_{i4}, \mu_{i5}, \ldots, \mu_{i8}) = (\mu_{i11}, \ldots, \mu_{i14}, \mu_{i21}, \ldots, \mu_{i24})$, for each $i = 1, \ldots, 121$. In addition, we introduce

$$\mathbf{x}_{ij} = (1, I_{\{1 \le j \le 4\}}, I_{\{j=1 \text{ or } 5\}}, I_{\{j=2 \text{ or } 6\}}, I_{\{j=3 \text{ or } 7\}}, I_{\{j=1\}}, I_{\{j=2\}}, I_{\{j=3\}})^T$$

and

$$\mathbf{z}_{ij} = (1, I_{\{1 \le j \le 4\}}, I_{\{5 \le j \le 8\}}, I_{\{j=1 \text{ or } 5\}}, I_{\{j=2 \text{ or } 6\}}, I_{\{j=3 \text{ or } 7\}}, I_{\{j=4 \text{ or } 8\}})^T$$

where $I_{\{\mathrm{A}\}}$ is the indicator of event A. With these definitions (2.3) becomes

$$g(\mu_{ij}) = \log(\mu_{ij}) = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{\xi}_{ij}^T\boldsymbol{\sigma} = \tilde{\mathbf{x}}_{ij}^T\boldsymbol{\psi}$$

where $\boldsymbol{\sigma} = (\sigma_\gamma, \sigma_\upsilon, \sigma_\omega)^T$, $\boldsymbol{\xi}_{ij}^T = (z_{ij1}u_{i1}, z_{ij2}u_{i2} + z_{ij3}u_{i3}, z_{ij4}u_{i4} + z_{ij5}u_{i5} + z_{ij6}u_{i6} + z_{ij7}u_{i7})$, and $\mathbf{u}_i \sim$ i.i.d. $N_7(0, \mathbf{I}_7)$.

The observable likelihood for the model is

$$L(\boldsymbol{\psi}; \mathbf{y}) = \prod_{i=1}^{121} \int_{\mathbb{R}^7} f(\mathbf{y}_i|\mathbf{u}_i; \boldsymbol{\psi})\phi(\mathbf{u}_i, \mathbf{I}_7)\mathrm{d}\mathbf{u}_i \,,$$

where the ith integral in the product is equal to

$$\Big(\frac{1}{2\pi}\Big)^{7/2}\Big(\prod_{j=1}^{8}\frac{1}{y_{ij}!}\Big)\int_{\mathbb{R}^7} \exp\Big(-\sum_{j=1}^{8}\exp\{\tilde{\mathbf{x}}_{ij}^T\boldsymbol{\psi}\} + \sum_{j=1}^{8} y_{ij}\tilde{\mathbf{x}}_{ij}^T\boldsymbol{\psi} - \frac{1}{2}\mathbf{u}_i^T\mathbf{u}_i\Big)\mathrm{d}\mathbf{u}_i \,,$$

and cannot be evaluated analytically.

Any multi-index model can be reduced to the form (2.4), in a similar manner, by appropriate re-indexing of variables.

3 Monte Carlo expectation maximization

The expectation maximization (EM) algorithm introduced in the seminal work of Dempster et al. (1977) is a widely-used iterative method for finding maximum likelihood estimates when there is missing or unobserved data. The EM algorithm can be applied in the GLMM context because the random effects are unobserved. The algorithm includes two steps at each iteration, an E-step and an 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,

$$Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(s)}) = E\Big[l(\boldsymbol{\psi}; \mathbf{y}, \mathbf{u})|\mathbf{y}; \boldsymbol{\psi}^{(s)}\Big] \,,$$

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

$$\boldsymbol{\psi}^{(s+1)} = \arg\max_{\boldsymbol{\psi}\in\boldsymbol{\Psi}} Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(s)})$$

Under mild regularity conditions the observable likelihood function (2.4) 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.

In the GLMM setting, the complete data loglikelihood is given by

$$l(\boldsymbol{\psi};\mathbf{y},\mathbf{u}) = \sum_{i=1}^{n}\Big(\sum_{j=1}^{n_i}\{w_{ij}[\theta_{ij}y_{ij} - b(\theta_{ij})] + c(y_{ij})\} - \frac{1}{2}\mathbf{u}_i^T\mathbf{u}_i\Big)$$

Hence, the Q-function calculated at the iteration $s+1$ is

$$Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(s)}) = \sum_{i=1}^{n} E\Big[\sum_{j=1}^{n_i}\{w_{ij}[\theta_{ij}y_{ij} - b(\theta_{ij})] + c(y_{ij})\} - \frac{1}{2}\mathbf{u}_i^T\mathbf{u}_i|\mathbf{y}_i;\boldsymbol{\psi}^{(s)}\Big]$$

However, part of this expression,

$$\sum_{i=1}^{n} E\Big[\sum_{j=1}^{n_i} w_{ij}c(y_{ij}) - \frac{1}{2}\mathbf{u}_i^T\mathbf{u}_i|\mathbf{y}_i;\boldsymbol{\psi}^{(s)}\Big]\,,$$

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,

$$Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(s)}) = \sum_{i=1}^{n} E\Big[\sum_{j=1}^{n_i} w_{ij}[\theta_{ij}y_{ij} - b(\theta_{ij})]|\mathbf{y}_i;\boldsymbol{\psi}^{(s)}\Big]\,,$$

in what follows.

Notice that

$$Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(s)}) = \sum_{i=1}^{n} E\Big[a(\mathbf{y}_i,\mathbf{u}_i;\boldsymbol{\psi})|\mathbf{y}_i;\boldsymbol{\psi}^{(s)}\Big] \tag{3.1}$$

where

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

Hence, the ith term in the Q-function is given by

$$E\Big[a(\mathbf{y}_i,\mathbf{u}_i;\boldsymbol{\psi})|\mathbf{y}_i;\boldsymbol{\psi}^{(s)}\Big] = \int_{\mathbb{R}^q} a(\mathbf{y}_i,\mathbf{u}_i;\boldsymbol{\psi}) f(\mathbf{u}_i|\mathbf{y}_i;\boldsymbol{\psi}^{(s)})\mathbf{d}\mathbf{u}_i\,, \tag{3.2}$$

where

$$f(\mathbf{u}_i|\mathbf{y}_i;\boldsymbol{\psi}^{(s)}) = \frac{f(\mathbf{y}_i,\mathbf{u}_i;\boldsymbol{\psi}^{(s)})}{f(\mathbf{y}_i;\boldsymbol{\psi}^{(s)})} = \frac{\exp\Big\{a(\mathbf{y}_i,\mathbf{u}_i;\boldsymbol{\psi}^{(s)}) - \frac{1}{2}\mathbf{u}_i^T\mathbf{u}_i\Big\}}{\int_{\mathbb{R}^q}\exp\Big\{a(\mathbf{y}_i,\mathbf{u}_i;\boldsymbol{\psi}^{(s)}) - \frac{1}{2}\mathbf{u}_i^T\mathbf{u}_i\Big\}\mathbf{d}\mathbf{u}_i} \tag{3.3}$$

As noted earlier the denominator in (3.3) is generally analytically intractable in the GLMM context. In such cases Wei and Tanner (1990) suggested approximating the expectations in the Q-function by Monte Carlo averages, resulting in the so-called MCEM algorithm. For example, if it is possible to generate i.i.d. vectors $\{\mathbf{u}_i^{(1)}, \ldots, \mathbf{u}_i^{(M)}\}$ from (3.3), a Monte Carlo approximation to Q is given by

$$\hat{Q}(\boldsymbol{\psi}|\boldsymbol{\psi}^{(s)}) = \frac{1}{M}\sum_{i=1}^{n}\sum_{k=1}^{M} a(\mathbf{y}_i, \mathbf{u}_i^{(k)}; \boldsymbol{\psi}) = \frac{1}{M}\sum_{i=1}^{n}\sum_{k=1}^{M}\sum_{j=1}^{n_i} w_{ij}[\theta_{ij}^{(k)} y_{ij} - b(\theta_{ij}^{(k)})] \tag{3.4}$$

where $\theta_{ij}^{(k)}$ involves the parameter vector, $\boldsymbol{\psi}$, via the identities,

$$\theta_{ij}^{(k)} = (b^{'})^{-1}[\mu_{ij}^{(k)}]\,, \quad \mu_{ij}^{(k)} = g^{-1}(\eta_{ij}^{(k)})\,, \quad \text{and} \quad \eta_{ij}^{(k)} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{\xi}_{ij}^{(k)T}\boldsymbol{\sigma} = \tilde{\mathbf{x}}_{ij}^{(k)T}\boldsymbol{\psi}\,.$$

Notice that $\hat{Q}(\boldsymbol{\psi}|\boldsymbol{\psi}^{(s)})$ has the form of the loglikelihood of a GLM, and hence the M-step can be performed by using the standard IRLS fitting algorithm (McCullagh and Nelder, 1989, Section 2.5).

However, direct i.i.d. sampling from (3.3) is usually not feasible. To overcome this, McCulloch (1994,1997), suggested using MCMC with stationary distribution (3.3) to approximate the E-step. In contrast, two i.i.d sampling schemes, rejection sampling and importance sampling, were proposed by Booth and Hobert (1999) to generate a Monte Carlo sample following (3.3).

It should be noted that MCEM is not deterministic. One must increase MC sample size to decrease Monte Carlo error and to ensure convergence. An automated rule was described in Booth and Hobert (1999) where estimates of Monte Carlo error were used to determine if the Monte Carlo sample size is sufficient. Caffo et al. (2005) suggested a data-driven algorithm based on the EM ascent property. The algorithm not only determines the sample size for each iteration of MCEM but also provides a convenient stopping rule by monitoring the change in the Q-function.

A serious drawback of implementations of MCEM to date, in the GLMM context, is that the MC sample size required for convergence can be so large as to make the algorithm impractical. In the next section, we describe a new implementation of MCEM using the spherical-radial integration rule (Genz and Monahan (1998,1998,1999), which can dramatically reduce the MC sample size required for convergence.

4 Spherical-radial rule implementation of MCEM

4.1 The E-step

Recall that we need to approximate the integral (3.2). In view of (3.3) the integral has the form,

$$I(c;p) = \frac{\int_{\mathbb{R}^q} c(\mathbf{u})p(\mathbf{u})\mathrm{d}\mathbf{u}}{\int_{\mathbb{R}^q} p(\mathbf{u})\mathrm{d}\mathbf{u}}$$

where $p(\mathbf{u})$ is an unnormalized probability density and $c(\mathbf{u})$ is a matrix-valued function with elements integrable with respect to $p(\mathbf{u})$. The SR approximation method can be described in four steps.

1. *Standardization of the density.*

   Let $\mathbf{u}^*$ denote the mode of the unnormalized density $p(\mathbf{u})$, and let $\mathbf{H} = -\partial^2 \log p(\mathbf{u}^*)/\partial\mathbf{u}\partial\mathbf{u}^T$ be the negative of its Hessian matrix evaluated at the mode. We suppose that $\mathbf{H}$ is positive definite, and denote its Cholesky decomposition by $\mathbf{H}^{1/2}(\mathbf{H}^{1/2})^T$. After changing the variable of integration from $\mathbf{u}$ to $\tilde{\mathbf{u}} = \mathbf{H}^{1/2}(\mathbf{u} - \mathbf{u}^*)$ the integral becomes

   $$I(c;p) = \frac{\det(\mathbf{H}^{1/2}) \int_{\mathbb{R}^q} \tilde{c}(\tilde{\mathbf{u}})\tilde{p}(\tilde{\mathbf{u}})\mathrm{d}\tilde{\mathbf{u}}}{\det(\mathbf{H}^{1/2}) \int_{\mathbb{R}^q} \tilde{p}(\tilde{\mathbf{u}})\mathrm{d}\tilde{\mathbf{u}}} = \frac{\int_{\mathbb{R}^q} \tilde{c}(\tilde{\mathbf{u}})\tilde{p}(\tilde{\mathbf{u}})\mathrm{d}\tilde{\mathbf{u}}}{\int_{\mathbb{R}^q} \tilde{p}(\tilde{\mathbf{u}})\mathrm{d}\tilde{\mathbf{u}}}\,,$$

   where $\tilde{c}(\tilde{\mathbf{u}}) = c(\mathbf{u}^* + \mathbf{H}^{-1/2}\tilde{\mathbf{u}})$ and $\tilde{p}(\tilde{\mathbf{u}}) = p(\mathbf{u}^* + \mathbf{H}^{-1/2}\tilde{\mathbf{u}})$. The density $\tilde{p}$ is standardized in the sense that it attains its maximum at $\mathbf{0}$ and $-\tilde{\mathbf{H}}(\mathbf{0}) = -\partial^2 \log \tilde{p}(\mathbf{0})/\partial\tilde{\mathbf{u}}\partial\tilde{\mathbf{u}}^T = \mathbf{I}_q$.

2. *The spherical-radial transformation.*

   At this step we change the variables of integration from $\tilde{\mathbf{u}}$ to $(r, \mathbf{s})$, where r is the radius, and $\mathbf{s}$ is a point on the surface of the unit sphere U_q; that is, $\tilde{\mathbf{u}} = r\mathbf{s}$, and $\mathbf{s}^T\mathbf{s} = 1$. The integral now becomes

   $$I(c;p) = \frac{\int_0^\infty \int_{U_q} \tilde{c}(r\mathbf{s})\tilde{p}(r\mathbf{s})r^{q-1}\mathrm{d}\mathbf{s}\mathrm{d}r}{\int_0^\infty \int_{U_q} \tilde{p}(r\mathbf{s})r^{q-1}\mathrm{d}\mathbf{s}\mathrm{d}r}$$

   According to Genz and Monahan (1997) "the value of changing to $(r, \mathbf{s})$ is that the most common failure of the normal approximation to the posterior appears in the tails, goes after the SR transformation to the radius r". Notice that, if we denote

   $$G_{\text{num}}(r) = \int_{U_q} \tilde{c}(r\mathbf{s})\tilde{p}(r\mathbf{s})\mathrm{d}\mathbf{s} \quad \text{and} \quad G_{\text{den}}(r) = \int_{U_q} \tilde{p}(r\mathbf{s})\mathrm{d}\mathbf{s}\,, \tag{4.1}$$

   then

   $$I(c;p) = \frac{\int_0^\infty G_{\text{num}}(r)r^{q-1}\mathrm{d}r}{\int_0^\infty G_{\text{den}}(r)r^{q-1}\mathrm{d}r}\,.$$

3. *Approximation of the spherical integral.*

   Given r, the inner spherical integral $G(r)$ may be approximated by

   $$\hat{G}(r) = \sum_{j=1}^{N^*}\sum_{k=1}^{s} \nu_{jk}\tilde{c}(r\mathbf{Q}_j\mathbf{v}_k)\tilde{p}(r\mathbf{Q}_j\mathbf{v}_k)\,, \tag{4.2}$$

where $\mathbf{Q}_1, \ldots, \mathbf{Q}_N{}^*$ are i.i.d. random orthogonal matrices, $\mathbf{v}_1, \ldots, \mathbf{v}_s$ are points on the q-dimensional unit sphere, and $\{\nu_{jk}\}$ are weights chosen such that $E\hat{G}(r) = G(r)$. A particular choice is a simplex rule with $s = q+1$, $\nu_{jk} = 1/N^*(q+1)$, and $\mathbf{v}_1, \ldots, \mathbf{v}_{q+1}$ the vertices of the regular q-dimensional simplex with coordinates given by

$$v_{ij} = \begin{cases} 0 & \text{for } 1 \le j < i < q+1 \\ \Big(\dfrac{(q+1)(q-i+1)}{q(q-i+2)}\Big)^{1/2} & \text{for } i = j \\ -\Big(\dfrac{q+1}{(q-i+1)q(q-i+2)}\Big)^{1/2} & \text{for } 1 \le i < j \le q+1 \end{cases} .$$

Some other possible rules are described in Genz and Monahan (1997).

4. *Approximation of the radial integral.*

The remaining one dimensional radial integral,

$$\int_0^\infty \hat{G}(r) r^{q-1} \mathrm{d}r = \int_0^\infty \hat{G}(r) r^{q-1} \exp(r^2/2) \exp(-r^2/2) \mathrm{d}r \,, \tag{4.3}$$

can be approximated in a variety of ways. For example, the third-order rule

$$\gamma_1(R)\hat{G}(0) + \gamma_2(R)\hat{G}(R) \exp(R^2/2) \,, \tag{4.4}$$

where $R \sim \chi_{q+2}$, $\gamma_1(R) = 1 - q/R^2$, and $\gamma_2(R) = q/R^2$, gives an unbiased estimate of the radial integral (4.3) and it is exact for integrating cubic functions with respect to the kernel $r^{q-1}e^{-r^2/2}$. A general method of constructing an unbiased degree $2n+1$ rule is given in Genz and Monahan (1997).

The final approximations of the integrals in (4.1) are i.i.d. averages of approximations of the form (4.4). Specifically, if $R_1, \ldots, R_{M^*}$ are i.i.d. χ_{q+2}, then

$$\int_0^\infty G(r) r^{q-1} \mathrm{d}r \approx \frac{1}{M^*} \sum_{i=1}^{M^*} \Big\{ \gamma_1(R_i)\hat{G}_i(0) + \gamma_2(R_i)\hat{G}_i(R_i) \exp(R_i^2/2) \Big\} \tag{4.5}$$

where $\hat{G}_i(r)$ is of the form (4.2) with i.i.d. random orthogonal matrices, $\mathbf{Q}_{i1}, \ldots, \mathbf{Q}_{iN^*}$. Combining approximations to all the integrals in (3.1) results in a Monte Carlo SR approximation to the Q-function. Notice, that since the Q-function involves a ratio of integrals, the approximation is not unbiased. However, the Law of Large Numbers ensures asymptotic unbiasedness as M^* goes to infinity. In particular, the approximation converges as $M^* \to \infty$ with $N^* = 1$.

4.2 M-Step.

The randomized SR rule approximation to the Q-function is of the form (3.4). However, in the SR case the subscript k is an index for different combinations of independent χ_{q+2} variables, random orthogonal matrices, vertices of the the q-dimensional simplex, and terms in the radial rule approximation. Thus, the value of M in (3.4) is proportional to the product, M^*N^*, of the two samples sizes defined in the previous section. Notice that $\hat{Q}$ need only be determined up to a constant of proportionality, since the constant has no impact on the maximization step. In particular, it is not necessary to divide by the Monte Carlo sample size in (3.4). It is also important to recognize that in the SR rule approximation the weights, $w_{ij}^{(k)}$, are (random) functions of the current parameter estimate $\boldsymbol{\psi}^{(s)}$. In this section we will discuss the maximization of $\hat{Q}$ for a generic iteration, and so the dependence on s will be suppressed.

The right side of (3.4) has the form of a GLM loglikelihood in which the response, y_{ij}, occurs M times with associated pseudo-covariate vectors, $\tilde{\mathbf{x}}_{ij}^{(1)}, \ldots, \tilde{\mathbf{x}}_{ij}^{(M)}$. Let $\mathbf{X}_i^{(k)} = (\tilde{\mathbf{x}}_{i1}^{(k)}, \ldots, \tilde{\mathbf{x}}_{in_i}^{(k)})^T$, and $\mathbf{X}_i = (\mathbf{X}_i^{(1)T}, \ldots, \mathbf{X}_i^{(M)T})^T$. That is, $\mathbf{X}_i$ is the (pseudo) covariate matrix associated with the ith response vector. Similarly, let $\mathbf{W}_i^{(k)} = \text{diag}\{\omega_{ij}^{(k)}\}_{j=1}^{n_i}$, and $\mathbf{W}_i = \text{blockdiag}\{\mathbf{W}_i^{(k)}\}_{k=1}^{M}$, where

$$\omega_{ij}^{(k)} = \frac{w_{ij}^{(k)}}{g'(\mu_{ij}^{(k)})^2 V(\mu_{ij}^{(k)})},$$

and $V = b''(b')^{-1}$ is the GLM variance function. Finally, let $\mathbf{X} = (\mathbf{X}_1^T, \ldots, \mathbf{X}_n^T)^T$ and $\mathbf{W} = \text{blockdiag}\{\mathbf{W}_i\}_{i=1}^{n}$. Then, the IRLS algorithm for maximizing (3.4) involves iteratively solving the weighted least squares equations

$$(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\boldsymbol{\psi} = \mathbf{X}^T\mathbf{W}\tilde{\mathbf{y}} \tag{4.6}$$

where $\tilde{\mathbf{y}}$ is a working response vector with components, $\tilde{y}_{ij}^{(k)} = \eta_{ij}^{(k)} + g'(\mu_{ij}^{(k)})(y_{ij} - \mu_{ij}^{(k)})$ (McCullagh and Nelder, 1989, Section 2.5). After each iteraction the values of $\tilde{\mathbf{y}}$, $\mathbf{W}$, and $\mathbf{X}$, must be updated to reflect the new value of $\boldsymbol{\psi}$.

The dimensions of the matrices, $\mathbf{W}$ and $\mathbf{X}$ are $NM \times NM$ and $NM \times (p+q^*)$ respectively, where $N = \sum_i n_i$. The value of NM can be very large in practice. For example, in the Minnesota clinic data, $N = 121 \cdot 8 = 968$, and hence, even with relative small M values, the size of the pseudo data set can easily be in the tens or even hundreds of thousands. A key attraction of randomized SR approximation rules is that their accuracy results in a dramatic reduction in the value of M that is necessary compared to less sophisticated Monte Carlo approximation methods. However, it is not necessary to store the entire $\mathbf{W}$ and $\mathbf{X}$ matrices to carry out the

IRLS update in (4.6), since each side of the equation can be decomposed into computations involving the individual pseudo-covariate vectors. Specifically

$$\mathbf{X}^T\mathbf{W}\mathbf{X} = \sum_{i=1}^{n}\sum_{k=1}^{M}\mathbf{X}_i^{(k)T}\mathbf{W}_i^{(k)}\mathbf{X}_i^{(k)} = \sum_{i=1}^{n}\sum_{k=1}^{M}\sum_{j=1}^{n_i}\omega_{ij}^{(k)}\tilde{\mathbf{x}}_{ij}^{(k)}\tilde{\mathbf{x}}_{ij}^{(k)T},$$

and

$$\mathbf{X}^T\mathbf{W}\tilde{\mathbf{y}} = \sum_{i=1}^{n}\sum_{k=1}^{M}\mathbf{X}_i^{(k)T}\mathbf{W}_i^{(k)}\tilde{\mathbf{y}}_i^{(k)} = \sum_{i=1}^{n}\sum_{k=1}^{M}\sum_{j=1}^{n_i}\omega_{ij}^{(k)}\tilde{\mathbf{x}}_{ij}^{(k)}\tilde{y}_{ij}^{(k)}.$$

4.3 Ascent-based MCEM-SR and stopping rule

Booth and Hobert (1999) and Caffo et al. (2005) propose methods for controlling Monte Carlo sample size when implementing MCEM. The approach of Caffo et al. is based on the ascent property of the EM algorithm, that the loglikelihood increases at each iteration. More specifically,

$$\Delta Q^{(s+1)} = Q(\boldsymbol{\psi}^{(s+1)}|\boldsymbol{\psi}^{(s)}) - Q(\boldsymbol{\psi}^{(s)}|\boldsymbol{\psi}^{(s)}) \geq 0$$

implies

$$l(\boldsymbol{\psi}^{(s+1)}|\mathbf{y}) \geq l(\boldsymbol{\psi}^{(s)}|\mathbf{y}) \tag{4.7}$$

However, in MCEM $\Delta Q^{(s+1)}$ is approximated by

$$\Delta\hat{Q}^{(s+1)} = \hat{Q}(\boldsymbol{\psi}^{(s+1)}|\boldsymbol{\psi}^{(s)}) - \hat{Q}(\boldsymbol{\psi}^{(s)}|\boldsymbol{\psi}^{(s)})$$

and the inequality, $\Delta\hat{Q}^{(s+1)} \geq 0$ no longer guarantees (4.7).

In fact, since the value of $\Delta\hat{Q}^{(s+1)}$ is a a ratio of two Monte Carlo means, its standard error, $\sigma_{\Delta\hat{Q}}$, can be estimated using the delta method (Stuart and Ord, 1994, 10.5-7), and this can be used to construct a lower confidence limit for $\Delta Q^{(s+1)}$ of the form

$$\Delta\hat{Q}^{(s+1)} - z_{\gamma_1}\hat{\sigma}_{\Delta\hat{Q}}\,. \tag{4.8}$$

The approach advocated by Caffo et al. is to compute a lower bound of the form (4.8) after each iteration. If the lower bound is positive, the algorithm continues as usual. However, if the lower bound is negative, $\boldsymbol{\psi}^{(s+1)}$ calculated with Monte Carlo sample size m is rejected and the MCEM iteration is repeated with an increased Monte Carlo sample size $m + m/k$, for some k. Caffo et al. (2005, equation 15) suggest that the increase should be determined by the standard sample size formula for a formal test of $\Delta Q^{(s+1)} = 0$ versus $\Delta Q^{(s+1)} > 0$ with type 1 error equal to α and type 2 error equal to β, using estimates of ΔQ and $\sigma_{\Delta Q}$ from the previous iteration

$$m_{s+1,start} = \max\{m_{s,start}, \sigma^2_{\Delta\hat{Q}}(z_\alpha + z_\beta)^2/(\Delta\hat{Q}^{(s)})^2\} \tag{4.9}$$

The deterministic EM algorithm is usually terminated when changes in the Q-function (and hence in the loglikelihood) are negligible. Even though the Q-function is not observed directly in implementations of the MCEM algorithm, one can calculate an upper confidence limit for ΔQ after each iteration, in a similar manner to the lower limit (Caffo et al., 2005, equation 13). The algorithm may then be judged to have converged if the upper bound is negligibly small (but non-negative), that is,

$$\Delta\hat{Q}^{(s+1)} + z_{\gamma_2}\hat{\sigma}_{\Delta Q} \leq \epsilon. \tag{4.10}$$

In addition to (4.10) we require the relative change in the parameter estimates at the $(s+1)$th iteration to be sufficiently small; that is

$$\max_{1\leq i\leq p+q^*} \frac{|\psi_i^{(s+1)} - \psi_i^{(s)}|}{|\psi_i^{(s)}| + \delta_1} \leq \delta_2 \tag{4.11}$$

Hence, the convergence is declared if both (4.10) and (4.11) hold.

5 A simulation study

Clarkson and Zhan (2002) proposed the use of SR rules to directly approximate GLMM likelihood functions. They illustrated their approach with a simulation study involving the following logit-binomial model with random effects. Let y_{ij} denote the jth binary observation on subject i, for $j = 1, \ldots, 7$ and $i = 1, \ldots, 100$. Suppose that, observations on different subjects are independent, but the repeated binary outcomes on a given individual share a subject specific random effects vector, $\mathbf{u}_i^{\Sigma}$, where $\mathbf{u}_i^{\Sigma} \sim$ i.i.d. $N_5(\mathbf{0}, \mathbf{\Sigma})$, with $\mathbf{\Sigma} =$ diag$\{\sigma_k^2\}_{k=1}^5$. Conditional on the subject specific effects, the binary outcomes are independent Bernoulli variates with success probabilities, π_{ij}, satisfying

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{u}_i^{\Sigma}\,,$$

where $\mathbf{x}_{ij} \equiv \mathbf{z}_{ij}$, $j = 1, \ldots, 7$, are columns of the matrix

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ I_{\{i\leq 50\}} & I_{\{i\leq 50\}} & I_{\{i\leq 50\}} & I_{\{i\leq 50\}} & I_{\{i\leq 50\}} & I_{\{i\leq 50\}} & I_{\{i\leq 50\}} \\ -3 & -2 & -1 & 0 & 1 & 2 & 3 \\ -3I_{\{i\leq 50\}} & -2I_{\{i\leq 50\}} & -1I_{\{i\leq 50\}} & 0 & I_{\{i\leq 50\}} & 2I_{\{i\leq 50\}} & 3I_{\{i\leq 50\}} \\ \zeta_{i1} & \zeta_{i2} & \zeta_{i3} & \zeta_{i4} & \zeta_{i5} & \zeta_{i6} & \zeta_{i7} \end{pmatrix}$$

with $\zeta_{ij} \sim$ i.i.d.$N(0,1)$. One hundred data sets were generated according to this scheme with $\boldsymbol{\beta} = (-2.5, 1, -1, 0.5, -0.5)^T$ and $\mathbf{\Sigma} = \mathbf{I}_5$.

For each dataset we applied the MCEM-SR algorithm utilizing the randomized third-order radial rule approximation, and a starting Monte Carlo sample size of $M^* = 20$. The algorithm was initially run with the number of orthogonal rotations $N^* = 10$, and then repeated with $N^* = 1$ with the results essentially identical for the two values of N^*. This is consistent with the findings of Genz and Monahan (1997), that the main source of variability in (4.5) is in the approximation of the (outer) radial, as opposed to the (inner) spherical integral. We set α and β in (4.9) and γ_1 and γ_2 to 0.05. For (4.11) δ_1 was chosen to be equal to 0.001 and $\delta_2 = 0.005$.

The maximum sample size M^*_{max} for MCEM-SR ranged from 820 to 2680 with an average of 1310. The average parameter estimates are given in Table 5, along with the average of their estimated standard errors, and their empirical standard errors. Table 5 gives the corresponding results obtained using a direct SR rule approximation to the likelihood with $M^* = 1500$ and $N^* = 1$, followed by Gauss-Newton maximization to obtain the MLE. Convergence was declared if (4.11) with $\delta_1 = 0.001$ and $\delta = 0.005$ held for three consecutive iterations. This approach is similar to that used by Clarkson and Zhan (2002) in their simulation study, except that they used a fixed quadrature rule to approximate the radial integral rather than a randomized third-order rule. As can be seen from the tables, direct maximization of the loglikelihood and indirect maximization via the EM algorithm give essentially identically results in this simulation study.

An advantage of the iterative MCEM approach is that the Monte Carlo sample size is automatically calibrated to the specific dataset and model. In principle, the direct maximization approach could be modified to include this adaptive feature. However, the EM algorithm also exploits the exponential family structure of the conditional model (given the random effects), resulting in an algorithm which is potentially more stable in complex settings. In the next section we illustrate the use of our MCEM-SR approach in two well-known examples in which GLMM fitting has proven to be problematic.

6 Examples

For the two examples considered in this section we use the third-order rule for the radial integral and the simplex rule to approximate the spherical integral with N^* fixed at 1. Clarkson and Zhan (2002) provide some explanation on why one rotation may be sufficient for Spherical-Radial approximations in GLMM settings. The MCEM-SR algorithm was run with an initial Monte Carlo sample size of $M^* = 20$. We set $\alpha = \beta = \gamma_1 = \gamma_2 = 0.05$ and $\delta_1 = 0.001$ and $\delta_2 = 0.005$, $k = 5$.

	β_1	β_2	β_3	β_4	β_5
True β	-2.5000	1.0000	-1.0000	0.5000	-0.5000
Average $\hat{\beta}$	-2.5511	1.0239	-1.0092	0.4701	-0.4846
Average $\hat{\text{se}}(\hat{\beta})$	0.3855	0.4835	0.2250	0.3308	0.1908
Empirical sd$(\hat{\beta})$	0.4551	0.4658	0.2252	0.3546	0.1984

	σ_1	σ_2	σ_3	σ_4	σ_5
True σ	1.0000	1.0000	1.0000	1.0000	1.0000
Average $\hat{\sigma}$	0.8734	0.9250	0.9511	0.9914	0.9647
Average $\hat{\text{se}}(\hat{\sigma})$	0.8215	1.2613	0.2478	0.5427	0.3313
Empirical sd$(\hat{\sigma})$	0.3517	0.5521	0.2238	0.4645	0.3487

Table 1: MCEM-SR estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}$. The values given are the average estimates, their average estimated standard errors based on the observed Fisher information matrix, and their empirical standard errors over 100 simulated datasets.

	β_1	β_2	β_3	β_4	β_5
True β	-2.5000	1.0000	-1.0000	0.5000	-0.5000
Average $\hat{\beta}$	-2.5499	1.0192	-1.0087	0.4633	-0.4843
Average $\hat{\text{se}}(\hat{\beta})$	0.4615	0.5371	0.2853	0.3947	0.2303
Empirical sd$(\hat{\beta})$	0.4540	0.4685	0.2249	0.3551	0.1986

	σ_1	σ_2	σ_3	σ_4	σ_5
True σ	1.0000	1.0000	1.0000	1.0000	1.0000
Average $\hat{\sigma}$	0.8670	0.9050	0.9472	1.0073	0.9663
Average $\hat{\text{se}}(\hat{\sigma})$	0.5472	0.8004	0.2811	0.5496	0.3174
Empirical sd$(\hat{\sigma})$	0.3663	0.5954	0.2252	0.4617	0.3517

Table 2: Direct MC-SR estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}$. The values given are the average estimates, their average estimated standard errors based on the observed Fisher information matrix, and their empirical standard errors over 100 simulated datasets.

6.1 Minnesota Health Plan Data

First we present the results for the Minnesota Health Plan data (Waller and Zelterman, 1997), and the Poisson linear mixed model described in Section 2.2. 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. Table 3 gives the ML estimates and their standard errors. Convergence was declared after 69 iterations with $M^*_{69} = 820$ and $M^*_{max} = 1370$. 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 are obtained by using the Bayesian software package WinBUGS (D.J.Spiegelhalter et al., 1999). The values given for WinBUGS are medians and standard deviations of the marginal posterior distributions obtained using the following non-informative priors $a_0, a_1, b_1, b_2, b_3, c_{11}, c_{12}, c_{13} \sim N(0, 10^6)$ and $1/\sigma^2_\gamma, 1/\sigma^2_\nu, 1/\sigma^2_\omega \sim U[0, 10^3]$. As we can see the estimates of all parameters except that of the constant agree with each other. The MCEM-SR estimate of a_0 is close to that of WinBUGS. Also, based on the ML estimates and their standard errors it appears that there is a significant event by period interaction. To compare our results to those of Booth et al. (2003) we refit the model without the interaction term. Table 4 gives the results for this model. In this case the MCEM-SR algorithm converged at the 76th iteration with $M^*_{76} = 720$ and $M^*_{max} = 1130$. Our results are in agreement with the estimates obtained using WinBUGS and the SAS/GLIMMIX procedures. However, the estimates reported by Booth et al. (2003) appear to be incorrect.

6.2 Salamander Mating Data

The salamander data from McCullagh and Nelder (1989, pages 439-450) have been analyzed by numerous authors using linear mixed effects models for binary responses (Booth and Hobert, 1999; Karim and Zeger, 1992; Lee and Nelder, 1996; McCulloch, 1994; Sung and Geyer, 2006). Here we consider the logit-normal GLMM described by Booth and Hobert, which is a frequentist version of the Bayesian model proposed by Karim and Zeger. As noted by Booth and Hobert, Sung and Geyer, and others, maximum likelihood estimation for this model is quite challenging.

The data, as described in McCullagh and Nelder (1989), arise from three experiments, each involving two groups consisting of twenty salamanders, 10 Roughbutt (R) and 10 Whiteside (W), with 5 males and 5 females in each case. Thus, there are 100 possible hetersexual crosses in each group. However, due to time constraints, only 60 crosses were permitted in each group.

	With Interaction		
	MCEM-SR	GLIMMIX	WinBUGS
a_0	0.868 (0.096)	0.961 (0.104)	0.844 (0.109)
a_1	-0.165 (0.127)	-0.164 (0.106)	-0.160 (0.110)
b_1	-0.091 (0.095)	-0.089 (0.109)	-0.085 (0.111)
b_2	0.414 (0.098)	0.394 (0.104)	0.422 (0.110)
b_3	0.491 (0.109)	0.468 (0.103)	0.498 (0.110)
c_{11}	0.246 (0.097)	0.240 (0.103)	0.243 (0.103)
c_{12}	0.104 (0.080)	0.101 (0.095)	0.102 (0.097)
c_{13}	-0.085 (0.099)	-0.084 (0.096)	-0.088 (0.097)
σ_γ	0.493 (0.082)	0.491 (0.081)	0.511 (0.078)
σ_ν	0.608 (0.056)	0.578 (0.048)	0.605 (0.053)
σ_ω	0.625 (0.040)	0.593 (0.034)	0.627 (0.038)

Table 3: Parameter estimates for the Poisson linear mixed effects model (2.5) obtained by maximum likelihood and using the SAS/GLIMMIX and WinBUGS software packages.

	Without Interaction			
	MCEM-SR	GLIMMIX	WinBUGS	BCFH
a_0	0.763 (0.109)	0.854 (0.099)	0.744 (0.107)	1.64 (0.001)
a_2	0.109 (0.109)	0.110 (0.083)	0.111 (0.087)	-0.12 (0.001)
b_2	0.435 (0.108)	0.414 (0.093)	0.481 (0.176)	0.35 (0.001)
b_3	0.425 (0.106)	0.402 (0.093)	0.470 (0.176)	0.23 (0.001)
b_4	-0.028 (0.904)	-0.026 (0.096)	-0.021 (0.102)	0.17 (0.001)
σ_γ	0.499 (0.084)	0.493 (0.080)	0.510 (0.082)	1.04 (0.091)
σ_ν	0.604 (0.059)	0.574 (0.048)	0.598 (0.052)	0.60 (0.053)
σ_ω	0.623 (0.043)	0.591 (0.034)	0.624 (0.038)	0.60 (0.036)

Table 4: Parameter estimates for the Poisson linear mixed effects model (2.5) obtained by maximum likelihood and using the SAS/GLIMMIX and WinBUGS software packages.

Two of the experiments involved the same set of 40 salamanders. However, following McCullagh and Nelder (1989, page 441) and Booth and Hobert (1999, Section 7.3) we shall analyze the study as though it consisted of 6 independent groups of 20 salamanders, each resulting in 60 binary indicators of successful mating.

Let π_{ij} denote the probability of successful mating for pair j in group i, $j = 1, \ldots, 60$, $i = 1, \ldots, 6$. Let $\mathbf{u}_i^f$ and $\mathbf{u}_i^m$ denote random effect vectors associated with the 10 female and 10 male salamanders in group i, and suppose that $(\mathbf{u}_i^{fT}, \mathbf{u}_i^{mT})^T \sim \mathbf{D}\mathbf{u}_i$, where $\mathbf{u}_i \sim N_{20}(\mathbf{0}, \mathbf{I})$, and

$$\mathbf{D} = \begin{pmatrix} \sigma_f \mathbf{I}_{10} & \mathbf{0}_{10} \\ \mathbf{0}_{10} & \sigma_m \mathbf{I}_{10} \end{pmatrix}$$

Booth and Hobert (1999) consider a logit model of the form

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{D}\mathbf{u}_i \,, \tag{6.1}$$

where $\mathbf{x}_{ij}$ is a 4×1 vector indicating the type of cross, and $\mathbf{z}_{ij}$ is a 20×1 vector with 1's at the coordinates corresponding to pair j, and 0's otherwise. The parameter vector

$$\boldsymbol{\beta} = (\boldsymbol{\beta}_{R/R}, \boldsymbol{\beta}_{R/W}, \boldsymbol{\beta}_{W/R}, \boldsymbol{\beta}_{W/W})^T$$

consists of unknown fixed coefficients associated with the four types of cross, with subscripts indicating the species of the female and male respectively.

The likelihood in this example involves six intractable 20-dimensional integrals. Maximum likelihood estimates of the parameter vector $\boldsymbol{\psi} = (\boldsymbol{\beta}^T, \boldsymbol{\sigma}^T)^T$ obtained using the MCEM-SR algorithm are displayed in Table 5. To compare our results with those of Booth and Hobert (1999) we started with $\theta^{(0)} = (0, 0, 0, 0, 1, 1)$. The algorithm converged after 48 iterations with $M^*_{48} = 365$ and $M^*_{max} = 840$. The MCEM-SR estimates agree with the ones obtained by Booth and Hobert (1999) who used an MCEM algorithm involving importance sampling at E-step. Booth and Hobert (1999) reported convergence in 51 iterations. Their Monte Carlo sample size increased from 1000 at the beginning to 66,169 at the end of the MCEM algorithm. Hence, much less computational effort was required in MCEM-SR to reach the same level of accuracy.

In addition, Table 5 contains the Bayesian estimates based on non-informative priors of Karim and Zeger (1992) (KZ). Booth and Hobert (1999) also reported the estimates produced by SAS%GLIMMIX macro (GLIMMIX(BH)) which was not the part of the SAS/STAT package at that time. We refitted the model using the current version of SAS/GLIMMIX with default settings which estimates a model using restricted maximum pseudo-likelihood (GLIMMIX). Finally, we fitted the model running SAS/GLIMMIX with the other available pseudo-likelihood

	$\boldsymbol{\beta}_{R/R}$	$\boldsymbol{\beta}_{R/W}$	$\boldsymbol{\beta}_{W/R}$	$\boldsymbol{\beta}_{W/W}$	σ_f	σ_m
MCEM-SR	1.022	0.325	-1.944	0.999	1.180	1.116
	(0.224)	(0.241)	(0.274)	(0.240)	(0.152)	(0.159)
BH	1.030	0.320	-1.950	0.990	1.183	1.118
SG	1.004	0.534	-1.783	1.268	1.099	1.167
	(0.161)	(0.271)	(0.101)	(0.606)	(0.149)	(0.237)
KZ	1.03	0.34	-1.98	1.07	1.50	1.36
GLIMMIX	0.787	0.247	-1.500	0.777	0.848	0.797
	(0.320)	(0.311)	(0.352)	(0.320)	(0.194)	(0.193)
GLIMMIX(BH)	0.87	0.28	-1.69	0.95	1.16	0.96

Table 5: Maximum likelihood estimates for the logit-normal model (6.1) obtained using the the MCEM-SR algorithm along with their standard errors. Maximum likelihood estimates reported by Booth and Hobert (1999), and by Sung and Geyer (2006) (http://www.stat.umn.edu/geyer/bernor/), as well as posterior means obtained from a Bayesian analysis of the same model in Karim and Zeger (1992) are given for comparison.

estimation techniques (not reported here) such as MSPL, RMPL, and MMPL (see SAS (2005)). The results were far from ours and those of Booth and Hobert (1999). Therefore, it appears that SAS/GLIMMIX cannot handle the estimation of a GLMM involving high-dimensional integrals as in the Salamander data case.

7 Discussion

In this paper, we have proposed a computationally feasible MCEM algorithm for fitting a GLMM with multivariate normal random effects. Our MCEM-SR algorithm can be generalized to GLMMs with another symmetric distribution for random effects such as the multivariate t-distribution. In our computations we found the 3rd order rule for radial integral and the simplex rule with one rotation for the spherical part were quite adequate. However, one could further refine the method by using the 5th order or more general $2n+1$ order rule for the radial part. In addition, there are other rules available to approximate the multivariate integral over the surface of the unit q-dimensional sphere, such as the antipodal and extended simplex rules.

The results show that MCEM-SR performs very well both in terms of the accuracy the estimates and the Monte Carlo sample size to attain this accuracy. It should not be a surprise that we needed a Monte Carlo sample size of 1370 for Minnesota data with a 7-dimensional random effect, and only 840 for Salamander data involving a 20-dimensional random effect. The Monte Carlo sample size in MCEM-SR is determined not only by the dimension of the random effect but also by the number of independent subjects observed. This follows from the fact that the variance of a MC approximation of a sum of n q-dimensional integrals is proportional to n.

For another example of the acccurary of the SR rule consider the following. In the salamander example, when we ran our algorithm with the Monte Carlo sample size M^* fixed at 2, MCEM converged to the MLE from Table 5 and then oscillated around it with a MC standard error of approximately 0.1. This is quite impressive considering the challenges reported by Sung and Geyer (2006) for this model.

In conclusion, the use of randomized spherical radial integration at the E-step of the EM algorithm leads to a computational feasible algorithm for fitting GLMMs. We have illustrated the power of the method with some challenging examples. The method is also relatively simple to program, and we are in the process of developing a R package to implement it.

Acknowledgement

The authors would like to thank Ciprian M. Crainiceanu of Bloomberg School of Public Health, Johns Hopkins University for his help with fitting the Minnesota data model in WinBUGS. Both authors received support from NSF DMS grant 00051086 while working on this project.

References

AGRESTI, A., BOOTH, J. G., HOBERT, J. P. and CAFFO, B. (2000). Random effects modeling of categorical response data. *Sociological Methodology*, **30** 27–80.

BOOTH, J. G., CASELLA, G., FRIEDL, H. and HOBERT, J. P. (2003). Negative binomial loglinear mixed models. *Statistical Modelling*, **3** 179–191.

BOOTH, J. G. and HOBERT, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. *Journal of the Royal Statistical Society*, **B 61** 265–285.

CAFFO, B. S., JANK, W. and JONES, G. L. (2005). Ascent-based monte carlo expectation-maximization. *Journal of the Royal Statistical Society*, **B 67** 235–251.

CLARKSON, D. B. and ZHAN, Y. (2002). Using spherical-radial quadrature to fit generalized linear mixed effects models. *Journal of Computational and Graphical Statistics*, **11** 639–659.

DEMIDENKO, E. Z. (2004). *Mixed Models: Theory and Applications*. Wiley-Interscience.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society (with discussion)*, **B 39** 1–39.

D.J.SPIEGELHALTER, A.THOMAS and N.G.BEST (1999). Winbugs version 1.2 user manual. *MRC Biostatistics Unit*.

GENZ, A. and MONAHAN, J. (1997). Spherical-radial integration rules for bayesian computation. *Journal of the American Statistical Association*, **93**.

GENZ, A. and MONAHAN, J. (1998). Stochastic integration rules for infinite regions. *SIAM Journal on Scientific Computing*, **19** 426–439.

GENZ, A. and MONAHAN, J. (1999). A stochastic algorithm for high-dimensional integrals over unbounded regions with gaussian weight. *Journal of Computational and Applied Mathematics*, **112** 71–81.

HOBERT, J. P. (2000). Hierarchical models: a current computational perspective. *Journal of the American Statistical Association*, **95** 1312–1316.

KARIM, M. R. and ZEGER, S. L. (1992). Generalized linear models with random effects; salamander data revisited. *Biometrics*, **48** 631–644.

LEE, Y. and NELDER, J. A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society*, **B 58** 619–678.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*. 2nd ed. Chapman and Hall.

MCCULLOCH, C. E. (1994). Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association*, **89** 330–335.

MCCULLOCH, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, **92** 162–170.

McCulloch, C. E. and Searle, S. R. (2001). *Generalized, Linear and Mixed Models*. John Wiley & Sons.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, **A 135** 370–384.

SAS (2005). Sas/stat software: The glimmix procedure, documentation. *SAS Institute Inc.*

Stuart, A. and Ord, K. (1994). *Kendall's Advanced Theory of Statistics: Distribution Theory*, vol. 1. 6th ed. Edward-Arnold.

Sung, Y. J. and Geyer, C. J. (2006). Monte carlo likelihood inference for missing data models. Tech. rep., University of Minnesota, School of Statistics.

Waller, L. A. and Zelterman, D. (1997). Loglinear modeling with the negative multinomial distribution. *Biometrics*, **53** 971–982.

Wei, G. C. G. and Tanner, M. A. (1990). A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, **85** 699–704.

Wu, C. (1983). On convergence properties of the em algorithm. *Annals of Statistics*, **11** 95–103.