*Statistical Modelling* 2005; **5**: 229–242

# Two slice-EM algorithms for fitting generalized linear mixed models with binary response

**Florin Vaida[1] and Xiao-Li Meng[2]**
[1]Division of Biostatistics and Bioinformatics, Department of Family and Preventive Medicine, School of Medicine, University of California at San Diego, La Jolla, CA 92093-0717, USA
[2]Department of Statistics, Harvard University, Cambridge, MA 02138, USA

**Abstract:** The celebrated simplicity of the EM algorithm is somewhat lost in its common use for generalized linear mixed models (GLMMs) because of its analytically intractable E-step. A natural and typical strategy in practice is to implement the E-step via Monte Carlo by drawing the unobserved random effects from their conditional distribution as specified by the E-step. In this paper, we show that further augmenting the missing data (e.g., the random effects) used by the M-step leads to a quite attractive and general slice sampler for implementing the Monte Carlo E-step. The slice sampler scheme is straightforward to implement, and it is neither restricted to the particular choice of the link function (e.g., probit) nor to the distribution of the random effects (e.g., normal). We apply this scheme to the standard EM algorithm as well as to an alternative EM algorithm which treats the variance-standardized random effects, rather than the random effects themselves, as missing data. The alternative EM algorithm does not only have faster convergence, but also leads to generalized linear model-like variance estimation, because it converts the random-effect standard deviations into linear regression parameters. Using the well-known salamander mating problem, we compare these two algorithms with each other, as well as with a variety of methods given in the literature in terms of the resulting point and interval estimates.

**Key words:** auxiliary variables; data augmentation; EM algorithm; Markov chain Monte Carlo; mixed effect; random effect; slice sampler

**Data and software link available from:** http://stat.uibk.ac.at/SMIJ
**Received** November 2004; **revised** May 2005; **accepted** June 2005

## 1 Introduction

The generalized linear mixed model (GLMM) is an extension of the generalized linear model (GLM), which allows for correlated responses through the inclusion of a random-effect term in the linear component. This model has received much attention (e.g., Breslow and Clayton, 1993; Diggle *et al.*, 1994; Lee and Nelder, 1996, 2001; McCulloch and Searle, 2001) due to its wide applicability and ease of interpretation. The computation of the maximum likelihood estimate (MLE) of the parameter vector is a complex task: the likelihood is, in general, an analytically intractable integral. This

Address for correspondence: Florin Vaida, Division of Biostatistics and Bioinformatics, Department of Family and Preventive Medicine, School of Medicine, University of California at San Diego, La Jolla, CA 92093-0717, USA. E-mail: vaida@ucsd.edu

 10.1191/1471082X05st097oa

issue is magnified when the random effects have a crossed design and therefore the data cannot be reduced to small independent clusters. This is the case of interest of this paper.

As with linear mixed effects models (Laird and Ware, 1982), we can treat the random effects in a GLMM as 'missing data'. However, the expectation at the E-step is also analytically intractable in general. For the computation of the E-step, several methods have been proposed: a Metropolis–Hastings algorithm (McCulloch, 1997); an independent sampler based on either multivariate importance sampling or rejection sampling (Booth and Hobert, 1999); Gibbs sampler (Chan and Kuk, 1997, for probit link). Skrondal and Rabe-Hesketh (2004) implement adaptive Gaussian quadrature, for a general class of models (see also Anderson and Hinde, 1988). Bayesian analyses for GLMM include Karim and Zeger (1992), Clayton (1996) and Damien *et al.* (1999).

In this paper, by invoking a data-augmentation scheme larger than the one used by the M-step of EM, we obtain a straightforward slice sampler (Neal, 2003) for the E-step, which naturally accommodates any link function and distribution of random effects. Then, we propose a new EM scheme for fitting GLMM, in which the missing data are the variance-standardized random effects. Using the well-known salamander mating data of McCullagh and Nelder (1989: 439), we compare the two slice-EM algorithms with other methods in the literature.

## 2 Binary regression with random effects

Let $\mathrm{y} = (y_1, \ldots, y_n)$ be the response vector of $n$ observations, where each $y_i$ is binary, with possible values conventionally denoted as 1 or 0. We assume that a binary regression model applies with linear predictor vector

$$\eta = X\beta + Z\mathbf{u} \tag{2.1}$$

with components $\eta_i$ and mean vector $\mathbf{G}$ with components $G_i = G(\eta_i)$. Here $X$ and $Z$ are known $n \times p$ and $n \times q$ matrices of covariates, $\beta$ and $\mathbf{u}$ are, respectively, fixed and random effects and $G$ is a known inverse link function. The $q$-dimensional random effect $\mathbf{u} \sim p(\mathbf{u} \mid \delta)$ follows a known distribution (e.g., multivariate normal or multivariate $t$) with mean zero and covariance modelled by the parameter vector $\delta$.

The general expression (2.1) covers many specific designs. For example, for clustered data, the observations are divided into $r$ independent clusters of sizes $m_1, \ldots, m_r$. The correlation of the observations within a cluster $l$ is modelled by the sharing of a common random-effect subvector $\mathbf{u}_l$ in the linear predictor. A more complex situation occurs in a crossed-design study, as in the salamander mating data. Briefly, 60 females and 60 males of two species of salamander, the Rough Butt ($R$) and the White Side ($W$), were paired following a crossed, blocked and incomplete design in an experiment studying whether the two species have developed genetic mechanisms which would prevent inter-breeding. The response is binary – successful ($y_{ij} = 1$) or unsuccessful ($y_{ij} = 0$) mating between female $i$ and male $j$ (Table 1). We adopt the model

$$\text{logit } \Pr(y_{ij} = 1 | \mathbf{u}, \beta) = \beta_{IJ} + u_i^{\mathrm{F}} + u_j^{\mathrm{M}} \tag{2.2}$$

**Table 1** The salamander mating data: RB, Rough Butt; WS, White Side; M, males; F, females

Males

| Females | | RBM | | | | | WSM | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* |
| | *1* | 1 | | | 1 | 1 | 0 | | | 1 | 1 |
| | *2* | 1 | | 1 | | 1 | | 1 | 1 | | 1 |
| RBF | *3* | 1 | 1 | 1 | | | 0 | | 1 | 1 | |
| | *4* | | 1 | | 1 | 1 | | 1 | | 1 | 0 |
| | *5* | | 1 | 1 | 1 | | 1 | 1 | 1 | | |
| | *6* | 0 | 0 | | 0 | | | 0 | 1 | | 0 |
| | *7* | 1 | 0 | | | 1 | 1 | | | 1 | 1 |
| WSF | *8* | | | 0 | 0 | 0 | 1 | 1 | | | 1 |
| | *9* | 1 | | 1 | 0 | | 1 | | 1 | 1 | |
| | *10* | | 0 | 1 | | 0 | | 1 | 1 | 0 | |

| | RBM | | | | | WSM | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *11* | *12* | *13* | *14* | *15* | *16* | *17* | *18* | *19* | *20* |
| *11* | | 0 | 1 | 1 | | 1 | | | 1 | 1 |
| *12* | 0 | | 0 | 1 | | | 0 | 0 | 0 | |
| *13* | 1 | | | 1 | 0 | 0 | 0 | | | 1 |
| *14* | 1 | 0 | | | 1 | | | 1 | 1 | 0 |
| *15* | | 0 | 0 | | 0 | 1 | 0 | 1 | | |
| *16* | | 1 | | 0 | 0 | | | 1 | 1 | 1 |
| *17* | 0 | 0 | | 0 | | 0 | 1 | | | 0 |
| *18* | 0 | 0 | 0 | | | 1 | 0 | | 0 | |
| *19* | | | 1 | 0 | 0 | | | 1 | 1 | 1 |
| *20* | 0 | | 0 | | 0 | 0 | 0 | 1 | | |

| | | *21* | *22* | *23* | *24* | *25* | *26* | *27* | *28* | *29* | *30* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | *21* | 1 | | | 1 | 1 | 0 | | | 0 | 0 |
| | *22* | 0 | | 1 | | 1 | | 1 | 0 | | 1 |
| RBF | *23* | 1 | 0 | 1 | | | 1 | | 1 | 1 | |
| | *24* | | 0 | | 1 | 0 | | 0 | | 0 | 0 |
| | *25* | | 0 | 1 | 1 | | 0 | 0 | 0 | | |
| | *26* | 1 | 0 | | 1 | | | 1 | 1 | | 1 |
| | *27* | 0 | 0 | | | 0 | 0 | | | 1 | 1 |
| WSF | *28* | | | 0 | 0 | 0 | 0 | 0 | | | 0 |
| | *29* | 1 | | 1 | 1 | | 1 | | 1 | 1 | |
| | *30* | | 0 | 1 | | 0 | | 1 | 1 | 1 | |

| | *31* | *32* | *33* | *34* | *35* | *36* | *37* | *38* | *39* | *40* |
|---|---|---|---|---|---|---|---|---|---|---|
| *31* | | 0 | 0 | 1 | | 0 | | | 1 | 1 |
| *32* | 1 | | 1 | 1 | | | 0 | 1 | 0 | |
| *33* | 0 | | | 0 | 0 | 1 | 0 | | | 1 |
| *34* | 1 | 1 | | | 1 | | | 1 | 0 | 1 |
| *35* | | 0 | 1 | | 0 | 1 | 0 | 1 | | |
| *36* | | 0 | | 0 | 0 | | | 1 | 1 | 0 |
| *37* | 0 | 0 | | 0 | | 1 | 0 | | | 1 |
| *38* | 0 | 0 | 0 | | | 1 | 1 | | 1 | |
| *39* | | | 1 | 0 | 0 | | | 1 | 1 | 0 |
| *40* | 0 | | 0 | | 0 | 0 | 0 | 0 | | |

| | | *41* | *42* | *43* | *44* | *45* | *46* | *47* | *48* | *49* | *50* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | *41* | 1 | | | 1 | 1 | 0 | | | 1 | 1 |
| | *42* | 0 | | 0 | | 0 | | 0 | 1 | | 0 |
| RBF | *43* | 1 | 1 | 1 | | | 1 | | 1 | 0 | |
| | *44* | | 1 | | 1 | 0 | | 0 | | 0 | 1 |
| | *45* | | 1 | 0 | 1 | | 0 | 0 | 1 | | |
| | *46* | 0 | 0 | | 0 | | | 0 | 1 | | 0 |
| | *47* | 0 | 0 | | | 0 | 0 | | | 1 | 1 |
| WSF | *48* | | | 0 | 0 | 0 | 0 | 1 | | | 1 |
| | *49* | 0 | | 0 | 0 | | 0 | | 1 | 1 | |
| | *50* | | 0 | 0 | | 0 | | 0 | 1 | 1 | |

| | *51* | *52* | *53* | *54* | *55* | *56* | *57* | *58* | *59* | *60* |
|---|---|---|---|---|---|---|---|---|---|---|
| *51* | | 1 | 1 | 1 | | 1 | | | 0 | 1 |
| *52* | 0 | | 1 | 0 | | | 0 | 0 | 0 | |
| *53* | 1 | | | 0 | 1 | 1 | 1 | | | 1 |
| *54* | 0 | 0 | | | 1 | | | 1 | 1 | 0 |
| *55* | | 1 | 1 | | 1 | 1 | 0 | 1 | | |
| *56* | | 0 | | 0 | 0 | | | 1 | 0 | 1 |
| *57* | 0 | 0 | | 0 | | 1 | 1 | | | 1 |
| *58* | 1 | 0 | 1 | | | 1 | 0 | | 0 | |
| *59* | | | 1 | 0 | 1 | | | 0 | 0 | 1 |
| *60* | 0 | | 0 | | 1 | 1 | 1 | 1 | | |

where $\beta_{IJ}$ is the fixed effect corresponding to the species combination of the $\{i,j\}$-pair of salamanders with $\beta = (\beta_{\mathrm{RR}}, \beta_{\mathrm{RW}}, \beta_{\mathrm{WR}}, \beta_{\mathrm{WW}})$; $\mathbf{u} = (\mathbf{u}^{\mathrm{F}}, \mathbf{u}^{\mathrm{M}})$ is the vector of female and male random effects, respectively, for which it is assumed that, independently, $u_i^{\mathrm{F}} \sim N(0, \sigma_{\mathrm{F}}^2), u_j^{\mathrm{M}} \sim N(0, \sigma_{\mathrm{M}}^2), i, j = 1, \ldots, 60$. Each animal participates in six matings.

The experiments yielded the female–male mating proportions: $R$–$R = 60/90$, $R$–$W = 50/90$, $W$–$R = 19/90$ and $W$–$W = 60/90$. This is undoubtedly a simplified model in view of the incomplete, blocked, and 'correlated' nature of the design. There were, in fact, three experiments corresponding to the three rows in Table 1, and the first two experiments actually used the same group of salamanders. In each experiment, there were two pairings of 10 males and 10 females, and each salamander was assigned to six matings. In this paper, we only focus on computing MLEs of the parameters $\beta$ and $\delta = (\sigma_{\mathrm{F}}, \sigma_{\mathrm{M}})$ under the simple model (2.2). The proposed methods are applicable to more sophisticated models, such as those that allow correlated, species-specific and/or experiment-specific random effects (e.g., McCullagh and Nelder, 1989; Karim and Zeger, 1992; Chan and Kuk, 1997).

The main interest is on $\theta = (\beta, \delta)$, which determines the probability of a successful mating for each crossing, $\pi_{IJ} = E[G(\beta_{IJ} + u_i^{\mathrm{F}} + u_j^{\mathrm{M}})]$, where $G$ is the inverse logit function, $G(\eta) = (1 + \exp(-\eta))^{-1}$ and the expectation is over $\mathbf{u}$. The random effects may be used to estimate the mating propensity of individual animals. Conditional on $\mathbf{u}$, the likelihood for $\beta$ is

$$p(\mathbf{y}|\beta, \mathbf{u}) = \frac{\exp\left(\sum_{I,J} y_{IJ}\beta_{IJ} + \sum_i y_{i.}u_i^{\mathrm{F}} + \sum_j y_{.j}u_j^{\mathrm{M}}\right)}{\prod_{i,j}\{1 + \exp(\beta_{IJ} + u_i^{\mathrm{F}} + u_j^{\mathrm{M}})\}} \tag{2.3}$$

where $y_{IJ} = \sum y_{ij}$ and the sum extends over all observations from the crossing $(I,J)$; $y_{i.}$ and $y_{.j}$ are the total number of successful matings for the $i$th female, and $j$th male, respectively (between 0 and 6), and $i,j$ extend over all females and males, respectively, in the experiment. The marginal likelihood to maximize is $p(\mathbf{y}|\theta) = E[p(\mathbf{y}|\beta, \mathbf{u})|\delta]$. This 120-dimensional integral can be decomposed into a product of six 20-dimensional integrals, which cannot be reduced any further.

## 3 Slice-EM algorithms for fitting GLMM

The EM algorithm for GLMM solves iteratively Fisher's equation $s(\theta; \mathbf{y}) = E[s(\theta; \mathbf{y}, \mathbf{u})|y, \theta] = 0$, where $s(\theta;\ \mathbf{y})$ and $s(\theta;\ \mathbf{y},\ \mathbf{u})$ are the observed-data and augmented-data score functions, respectively. Operationally, the E-step computes $s(\theta|\theta^{(t)}) \equiv E[s(\theta; \mathbf{y}, \mathbf{u}))|\mathbf{y}, \theta^{(t)}]$ and then the M-step solves $s(\theta|\theta^{(t)}) = 0$, for $\theta$ to determine $\theta^{(t+1)}$. The algorithm is iterated to convergence, although the convergence properties are somewhat tricky, as discussed in Wu (1983) and Vaida (2005).

For a general GLMM with binary response $y_i$, linear predictor $\eta_i = \mathbf{x}_i^\top\beta + \mathbf{z}_i^\top\mathbf{u}$ and mean response $E(y_i) = G(\eta_i) = G_i$, $s(\theta|\theta^{(t)})$ conveniently separates the fixed effect $\beta$ from the variance parameter $\delta$: $s(\theta|\theta^{(t)}) = s(\beta|\theta^{(t)}) + s(\delta|\theta^{(t)})$,

$$s(\beta|\theta^{(t)}) = \sum_{i=1}^{n} \boldsymbol{x}_i, E\left\{\frac{G_i'(\mathbf{y}_i - G_i)}{G_i(1 - G_i)}\middle|\mathbf{y}, \theta^{(t)}\right\} \tag{3.1}$$

$$s(\delta|\theta^{(t)}) = E\left\{\frac{\partial}{\partial\delta}\log p(\mathbf{u}|\delta)\middle|\mathbf{y}, \theta^{(t)}\right\} \tag{3.2}$$

where $G_i' = \mathrm{d}G/\mathrm{d}\eta_i$. Using Monte Carlo simulation to compute (3.1) and (3.2) at the M-step, we solve

$$\hat{s}_m(\beta) = \sum_{i=1}^{n} \mathbf{x}_i \frac{1}{m} \sum_{k=1}^{m} \frac{G'_{ik}(y_i - G_{ik})}{G_{ik}(1 - G_{ik})} = 0 \tag{3.3}$$

and similarly for $s(\delta|\theta^{(t)})$, where $\mathbf{u}_1, \ldots, \mathbf{u}_m$ are draws from $p(\mathbf{u}|\theta^{(t)}, \mathbf{y})$ and $G_{ik}$ and $G'_{ik}$ are obtained by substituting $\eta_{ik} = \mathbf{x}_i^\top \beta + \mathbf{z}_i^\top \mathbf{u}_k$ for $\eta_i$ in $G_i$ and $G_i'$ respectively. In Equation (3.3), $\hat{s}_m(\beta)$ is the score function of a GLM with $m \cdot n$ observations $y_i$ (repeated $m$ times) and linear predictor $\eta_{ik}$, and the M-step for $\beta$ is the estimation of a GLM with offsets $\mathbf{u}_k$. The M-step for $\delta$ is the same as finding MLE for $\delta$ under $p(\mathbf{u}|\delta)$ on the basis of perceived independent observations $\mathbf{u}_1, \ldots, \mathbf{u}_m$. For the salamanders mating data, the solutions were $\{\sigma_{\mathrm{F}}^2\}^{(t+1)} = (Im)^{-1} \sum_{i=1}^{I} \sum_{k=1}^{m} (u_{ik}^{\mathrm{F}})^2$ and $\{\sigma_{\mathrm{M}}^2\}^{(t+1)} = (Jm)^{-1} \sum_{j=1}^{J} \sum_{k=1}^{m} (u_{jk}^{\mathrm{M}})^2$, where $I = J = 60$ and $u_{ik}^{\mathrm{F}}$'s and $u_{jk}^{\mathrm{F}}$'s are the output from the E-step sampler.

We sample from $p(\mathbf{u} \mid \mathbf{y}, \theta)$ via a slice sampler (Neal, 2003), which produces a data-augmentation scheme for the E-step larger than the one for the M-step (Meng and van Dyk, 1997). Specifically, we further augment $\{\mathbf{u}, \mathbf{y}\}$ to $\{\mathbf{u}, \mathbf{y}, \mathbf{v}\}$, where $\mathbf{v} = (v_1, \ldots, v_n)$ are an i.i.d. sample from the uniform distribution on [0,1]; $\mathbf{v}$ is independent of $\mathbf{u}$ and is connected to $\mathbf{y}$ via the threshold representation

$$y_i = I[v_i \le G(\eta_i)], \quad i = 1, \ldots, n \tag{3.4}$$

where $I[\cdot]$ is an indicator function. We sample $\mathbf{v}$ from $p(\mathbf{v}|\mathbf{u}, \mathbf{y})$ and $\mathbf{u}$ from $p(\mathbf{u}|\mathbf{v}, \mathbf{y})$. The two distributions are proportional to the joint distribution restricted by a set of linear inequalities (both also condition on $\theta$): $p(\mathbf{v}|\mathbf{u},\mathbf{y}) \propto I_{\mathcal{R}(\mathbf{y})} p(\mathbf{v})$ and $p(\mathbf{u}|\mathbf{v}, \mathbf{y}) \propto I_{\mathcal{R}(\mathbf{y})} p(\mathbf{u})$, where $\mathcal{R}(\mathbf{y})$ is the set of all vectors $(\mathbf{u}, \mathbf{v})$ for which Equation (3.4) holds. The distribution $p(\mathbf{v}|\mathbf{u},\mathbf{y})$ is truncated uniform on the unit hypercube and $p(\mathbf{u}|\mathbf{v},\mathbf{y})$ is a truncated $p(\mathbf{u})$. Slice sampling is a general method for constructing useful Gibbs samplers (Damien *et al.*, 1999), and it has good convergence properties (Mira and Tierney, 2002).

An alternative slice-EM algorithm is obtained by writing equation (2.2) as

$$\text{logit Pr}(y_{ij} = 1|\beta, w_i^{\mathrm{F}}, w_j^{\mathrm{M}}) = \beta_{IJ} + \sigma_{\mathrm{F}} w_I^{\mathrm{F}} + \sigma_{\mathrm{M}} w_j^{\mathrm{M}} \tag{3.5}$$

with $w_i^{\mathrm{F}}, w_j^{\mathrm{M}} \overset{\text{i.d.}}{\sim} N(0,1)$ for all $i,j$, that is, $\sigma_{\mathrm{F}}, \sigma_{\mathrm{M}}$ become the regression coefficients of the standardized-random effects $w_i^{\mathrm{F}}, w_j^{\mathrm{M}}$. The E-step remains essentially unchanged, but the key difference is that the variance parameters are now part of the mean parameter to be estimated at the M-step, $\theta = (\beta, \sigma_{\mathrm{F}}, \sigma_{\mathrm{M}})$. This leads to faster convergence for $\sigma_{\mathrm{F}}, \sigma_{\mathrm{M}}$, for reasons similar to those given in Meng and van Dyk (1997, 1998). Note that estimation of $\sigma_{\mathrm{F}}$, $\sigma_{\mathrm{M}}$ may in principle, result in either positive or negative values. This will not affect the inference, because in the end only the squares $\sigma_{\mathrm{F}}^2$, $\sigma_{\mathrm{M}}^2$ are reported. (Meng and van Dyk, 1998; van Dyk and Meng, 2001).

The standard errors (SEs) of the estimates are computed from the Fisher information matrix, as a byproduct of the Monte Carlo EM (MCEM) using the formula (Orchard and Woodbury, 1972; Louis, 1972)

$$I_y(\hat{\theta}) = E[-s'(\hat{\theta}; \mathbf{y}, \mathbf{u})|\mathbf{y}, \hat{\theta}] - E[s(\hat{\theta}; \mathbf{y}, \mathbf{u})s(\hat{\theta}; \mathbf{y}, \mathbf{u})^\top|\mathbf{y}, \hat{\theta}] \tag{3.6}$$

where $\hat{\theta}$ is the MLE of $\theta$ and $s' = \mathrm{d}s/\mathrm{d}\theta$. The right-hand side is estimated via Monte Carlo averages of $-s'(\hat{\theta}; \mathbf{y}, \mathbf{u})$ and $s(\hat{\theta}; \mathbf{y}, \mathbf{u})s(\hat{\theta}; \mathbf{y}, \mathbf{u})^\top$. The second slice-EM has the added appeal that the augmented-data score and the Fisher information have standard GLM forms, regardless of the distribution of the random effects.

An alternative to the slice-EM for this situation would be adaptive rejection sampling (Gilks and Wild, 1992). Steele (1996) used a Laplace approximation for estimating the expectations at the E-step.

## 4 Slice-EM implementation and data analysis

In contrast to the standard EM, in an MCEM algorithm such as the slice-EM, the E-step computation is not exact. The EM sequence of likelihood values is no longer guaranteed to be monotone and the convergence to the MLE is stochastic, rather than deterministic. Chan and Ledolter (1995), Biscarat (1994) and Vaida (1998) gave theoretical results concerning the convergence of MCEM under suitable conditions. A number of practical methods for monitoring convergence of MCEM have been proposed: they include graphical monitoring of some sequences of parameters (Chan and Ledolter, 1995), monitoring the likelihood ratio via bridge sampling (Meng and Schilling, 1996) and stopping when the Monte Carlo error is small relative to the statistical error (Booth and Hobert, 1999). Our strategy here is to implement an MCEM in three stages: 1) the burn-in stage, where the starting point is 'forgotten'; we use a small sample size $m$ and our goal is just to approach the region of convergence; 2) the transition stage, $m$ is increased gradually (e.g., linearly); 3) the plateau, where the algorithm is run with large $m$ to achieve small Monte Carlo error. In our experience, these stages are necessary due to the delicate interaction between the deterministic (EM) part and the stochastic (MC) part of the algorithm. Whereas Booth and Hobert's method is seeking a balance between the statistical error of the estimator and the sampling (MCMC) error, our method is balancing the sampling error and the EM convergence.

The starting point for the MCEM was $\beta = 0$ and $\sigma_F^2 = \sigma_M^2 = 1$. The burn-in stage had 50 steps at $m = 100$. Figure 1 shows that this was enough to approach stationarity. The second stage had 20 steps, with $m$ increased linearly to 10 000, ensuring a smooth transition to the plateau. The plateau stage, with $m = 10\,000$ for 50 steps, showed the stationarity of the MCEM process and gave more precise MLE. The MLE was the average of the MCEM iterates from the plateau stage. The total running time was $<$30 min (slice-EM1 and slice-EM2 took about same time per iteration), with 20 s burn-in and 25 min plateau. For practical scope, a satisfactory precision can be achieved with a running time of 3–5 min. All programs were implemented in C and were run on a Sun

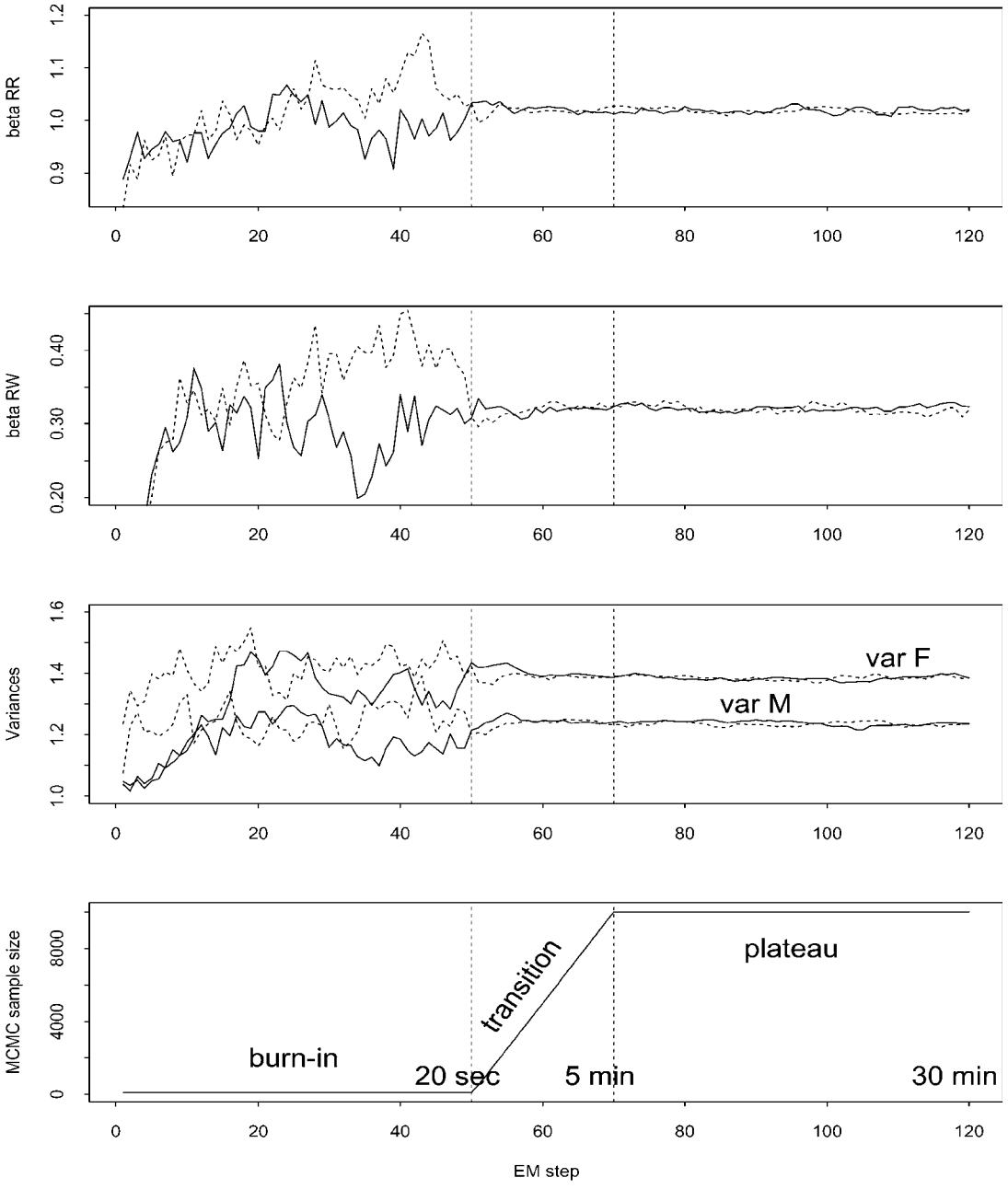


**Figure 1** Convergence and comparison of algorithms for the salamander data: slice-EM1, — and slice-EM2, .... Two mean parameters, $\beta_{RR}$ (top panel), $\beta_{RW}$ (second panel), and the variance parameters $\sigma^2_F$, $\sigma^2_M$ (third panel) are shown. The bottom panel includes the MCMC sample size $m$ as a function of the phase and step of the algorithm and computation time

**Table 2** MLEs for the salamander data from the two slice-EM algorithms, MCEM error, standard error of the MLE and approximate 95% confidence intervals

| | Method | Estimate | MCEM SE | MLE SE | 95% confidence interval |
|---|---|---|---|---|---|
| $\beta_{RR}$ | Slice-EM1 | 1.019 | 0.0016 | 0.415 | (0.19 185) |
| | Slice-EM2 | 1.018 | 0.0013 | 0.407 | (0.20 1.83) |
| $\beta_{RW}$ | Slice-EM1 | 0.321 | 0.0015 | 0.393 | (−0.47 1.11) |
| | Slice-EM2 | 0.320 | 0.0013 | 0.389 | (−0.46 1.10) |
| $\beta_{WR}$ | Slice-EM1 | −1.940 | 0.0019 | 0.475 | (−2.89 −0.99) |
| | Slice EM2 | −1.941 | 0.0013 | 0.473 | (−2.89 −0.99) |
| $\beta_{WW}$ | Slice EM1 | 0.997 | 0.0016 | 0.415 | (0.17 1.83) |
| | Slice EM2 | 0.994 | 0.0014 | 0.417 | (0.16 1.83) |
| $\sigma_F^2$ | Slice-EM1 | 1.384 | 0.0044 | 0.658 | (0.54 3.58) |
| | Slice-EM2 | 1.385 | 0.0016 | 0.626 | (0.56 3.42) |
| $\sigma_M^2$ | Slice-EM1 | 1.238 | 0.0039 | 0.583 | (0.48 3.18) |
| | Slice-EM2 | 1.234 | 0.0016 | 0.580 | (0.48 3.16) |

Ultra 30 workstation. The observed Fisher information matrix was computed using an additional sampling step at convergence.

Table 2 compares the point and interval estimates from the two algorithms and the error due to the simulation. The latter was estimated on the basis of a AR(1) approximation to $\theta^{(t)}$ during the plateau stage (Chan and Ledolter, 1995; Vaida, 1998). This approximation is supported by Figure 2, which shows the plateau stage and the partial autocorrelation function (ACF) for four of the parameter components. Slice-EM2 has uniformly lower MCEM error than Slice-EM1, with small improvements for $\beta$, but 60% reduction for the variance estimates. Moreover, the convergence of the variance components is faster for Slice-EM2 than that for Slice-EM1. This is an important finding, because in MCEM, for mixed models, the variance components are typically slower to converge and therefore dictate the overall convergence of the algorithm. In addition, Slice-EM2 has better mixing than Slice-EM1, as shown by the partial ACF in Figure 2. The MCEM standard error is thus also an indirect measure of the efficiency of the algorithm: for equal MCMC sample size and computation time per iteration, Slice-EM2 leads to more precise estimation. For both implementations, owing to the extensive plateau stage, the MCEM error is negligible when compared with the standard error of the MLE. A shorter plateau would suffice in practice.

The population-level probabilities of mating, $\pi_{IJ}$, are given by

$$\pi_{IJ} = G\left(\frac{\beta_{IJ}}{\sqrt{1 + c^2\sigma^2}}\right) \tag{4.1}$$

where $c^2 = (16\sqrt{3}/15\pi)^2 \approx 0.346$ and $G$ is the inverse logit function (Zeger *et al.*, 1988). They are reported in Table 3. Interval estimates for these were obtained in two ways. The 'naive' ones are computed from Equation (4.1) with $\beta_{IJ}$ inside the approximate 95% confidence interval, $\beta_{IJ} \in \{\hat{\beta}_{IJ} \pm 2\mathrm{SE}(\hat{\beta}_{IJ})\}$ and $\sigma^2 = \sigma_F^2 + \sigma_M^2$ held fixed at the MLE; these ignore the variability in $\sigma^2$. A better choice is the 95% highest posterior density interval for $\pi_{IJ}$ from simulated samples for $(\beta, \log\sigma_F^2, \log\sigma_M^2)$ from their asymptotic normal posterior distribution with mean at the MLE and variance given by the inverse Fisher information. For our data set, the two sets of

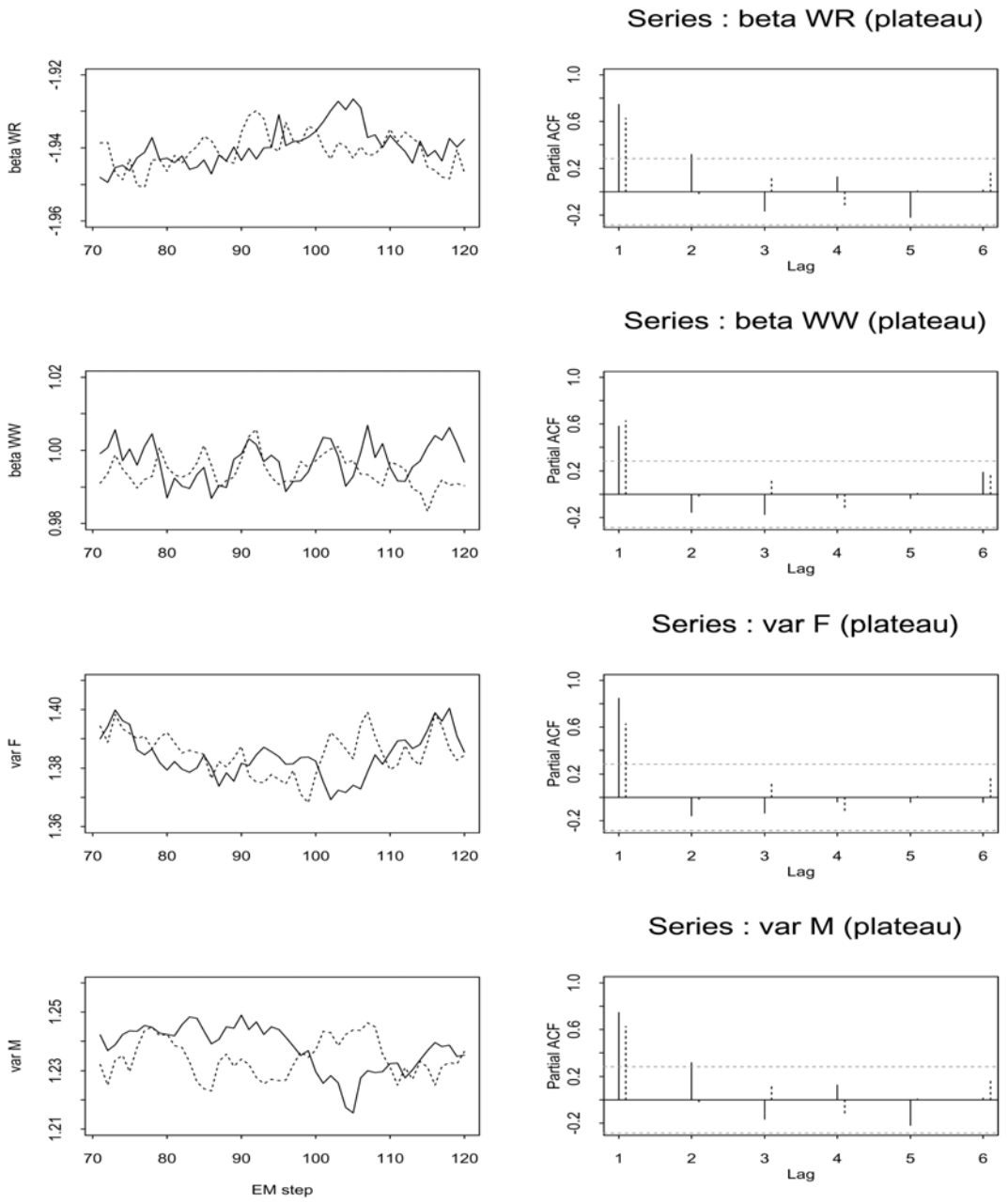


**Figure 2** The plateau stage and partial ACF of the MCEM estimates for the salamander data: slice-EM1, —; slice-EM2, . . .

**Table 3** Marginal probabilities and marginal OR for the salamander data

| | Estimate | 'Naive' 95% confidence interval | MLE 95% confidence interval | Odds ratio | OR 95% confidence interval |
|---|---|---|---|---|---|
| $\pi_{RR}$ | 0.676 | (0.537 0.790) | (0.537 0.784) | 1.00 | |
| $\pi_{RW}$ | 0.558 | (0.418 0.689) | (0.421 0.683) | 0.60 | (0.31 1.18) |
| $\pi_{WR}$ | 0.197 | (0.110 0.327) | (0.119 0.322) | 0.12 | (0.06 0.26) |
| $\pi_{WW}$ | 0.673 | (0.529 0.790) | (0.532 0.784) | 0.99 | (0.44 2.19) |

*Note*: the 'naive' confidence interval ignores the variability in the estimation of variance components. The crossing RR is the reference. The results are based on slice-EM2.

approximations are practically identical (Table 3), indicating that ignoring the statistical error of $\sigma^2$ in Equation (4.1) has little effect on the marginal inference for $\pi_{IJ}$. This is probably due to the complex correlations in the inverse Fisher information matrix; in particular, $\log \sigma_F^2$ and $\log \sigma_M^2$ are negatively correlated, keeping the overall variation of $\sigma^2$ in the posterior sample relatively low.

The mating probabilities are large and very similar for same-species matings, $\pi_{WW} = 0.676$ and $\pi_{RR} = 0.673$, and high for RW, $\pi_{RW} = 0.558$, but very low for WR, $\pi_{WR} = 0.197$. To test whether the mating between species is less probable than within the same species, we calculated the marginal odds ratios (ORs) relative to RR for the three other crossings. There is strong evidence of a smaller probability of mating for WR: OR = 0.12, 95% CI = (0.05, 0.26). For the other three comparisons, the 95% CI includes OR = 1. The White Side females and Rough Butt males clearly do not like each other.

## 5 Comparison with other methods and discussion

Tables 4 and 5 compare our results with those from a variety of methods used in the literature; the blank entries are for results that are not available from the literature. Previous researchers also analysed the data from each of the three experiments separately, so we included results of the first experiment only, conducted in the Summer of 1986 (first two blocks of Table 1). Using an importance sampling-based MCEM, Booth and Hobert (1999) arrived essentially at the same numeric results as

**Table 4** Comparison of methods: the variance components

| | All experiments | | Summer 1986 | |
|---|---|---|---|---|
| Methods | $\sigma_F^2$ | $\sigma_M^2$ | $\sigma_F^2$ | $\sigma_M^2$ |
| Slice-EM2 | 1.39 | 1.23 | 1.74 | 0.23 |
| IS-EM | 1.40 | 1.25 | | |
| Mod Laplace | | | 1.80 | 0.25 |
| Bayes | 1.50 | 1.36 | 2.35 | 0.14 |
| PQL | 0.72 | 0.63 | 1.42 | 0.09 |
| Moments | 0.91 | 0.88 | 1.37 | 0.70 |

**Table 5** Comparison of methods: the fixed effects

| | All experiments | | | | Summer 1986 | | | |
|---|---|---|---|---|---|---|---|---|
| Methods | $\beta_{RR}$ | $\beta_{RW}$ | $\beta_{WR}$ | $\beta_{WW}$ | $\beta_{RR}$ | $\beta_{RW}$ | $\beta_{WR}$ | $\beta_{WW}$ |
| Slice-EM2 | 1.02 | 0.32 | −1.94 | 0.99 | 1.38 | 0.93 | −1.66 | 1.18 |
| IS-EM | 1.03 | 0.32 | −1.95 | 0.99 | | | | |
| Mod Laplace | 1.00 | 0.32 | −1.95 | 1.02 | 1.37 | 0.93 | −1.65 | 1.18 |
| Bayes | 1.03 | 0.34 | −1.98 | 1.07 | 1.48 | 0.98 | −1.77 | 1.35 |
| PQL | 0.79 | 0.25 | −1.50 | 0.78 | | | | |

ours, considering the Monte Carlo errors in both algorithms. A direct comparison of computational efficiency and convergence of the two algorithms is beyond the purposes of this paper; we note that their rejection sampling has the advantage of independent samples, at the price of rejecting a large number of the simulations. Their elegant method for choosing the simulation sample size $m$ at each E-step, valid only for independent E-step sampling, was extended to Markov chain E-step sampling by Levine and Casella (2001).

The modified Laplace approximation of Shun and McCullagh (1995), applied by Shun (1997) to the salamander data, gives results that are also very close to the MLE. The comparisons with non-MLE-based methods are no longer purely computational, and they involve comparisons among different estimation procedures. The Bayesian analysis of Karim and Zeger (1992) yields similar results for the fixed effects. The results for the effects variance parameters appear to be different, but this is probably a reflection of the skewness of the posterior distributions of the random variances as Karim and Zeger (1992) reported posterior medians (and 5th and 95th percentiles), not posterior means. The PQL estimators (Breslow and Clayton, 1993) and the moment estimators (McCullagh and Nelder, 1989, Table 14.10) are rather remote from the MLE. It has been noted in the literature that PQL tends to give biased estimators for binary data GLMM, but with a good mean square error (Neuhaus and Segal, 1997).

In Table 6, we compare the mating probabilities $\pi_{IJ}$ and their 90% confidence intervals estimates from four different models/methods: slice-EM, the Bayesian model, PQL and fixed-effects GLM. The GLM amounts to a separate analysis of the four $2 \times 2$ tables for the female–male crossings. The GLMM and the Bayesian method produce numerically identical results, which are not unexpected because Karim and Zeger

**Table 6** Marginal probabilities and 90% intervals from GLMM, GLM and Karim and Zeger's Bayesian model, and marginal probabilities from PQL

| | GLMM | GLM | Bayes | PQL |
|---|---|---|---|---|
| $\pi_{RR}$ | 0.68 (0.56 0.77) | 0.67 (0.59 0.75) | 0.67 (0.56 0.77) | 0.66 |
| $\pi_{RW}$ | 0.56 (0.44 0.66) | 0.56 (0.47 0.64) | 0.56 (0.44 0.66) | 0.55 |
| $\pi_{WR}$ | 0.20 (0.13 0.30) | 0.21 (0.14 0.28) | 0.20 (0.13 0.30) | 0.22 |
| $\pi_{WW}$ | 0.67 (0.56 0.77) | 0.67 (0.59 0.75) | 0.68 (0.56 0.77) | 0.66 |

(1992) used a constant prior for both the fixed effects and the random-effect variances, and our approach for approximating confidence intervals effectively computes the same posterior intervals as theirs. The remarkable numerical agreement of the two sets of results is a strong validation of the accuracy of both Karim and Zeger's (1992) Gibbs sampler algorithm and our Slice-EM2 algorithm, considering the implementations of the two Monte Carlo algorithms were completely independent and in fact the two algorithms involve different data-augmentation schemes. It is also a validation of the normal approximation we have used for the posterior of $(\beta, \log \sigma_{\mathrm{F}}^2, \log \sigma_{\mathrm{M}}^2)$, as Karim and Zeger's (1992) Gibbs sampler was designed to compute the exact posterior distribution. GLM also provides nearly identical point estimates, but with narrower confidence intervals, owing to the absence of the random effects.

It is also interesting to observe that although PQL gives biased parameter estimates for GLMM, it provides reasonable estimates of the marginal probabilities. Although this behaviour is not completely understood, we like to think of PQL heuristically as an 'intermediate' model between the fixed-effects GLM (or zero random effects) and the GLMM, with the effect of doing a 'partial shrinking' of the random effect $\mathbf{u}$ towards 0. In this context, the interpretation of the parameters of PQL is not the same as for GLMM owing to the bias induced by the PQL approximation. In particular, the variance components tend to underestimate the GLMM variances (Neuhaus and Segal, 1997), and are biased towards the 'zero variances' of the GLM. However, the model predictions are similar (intermediate between GLM and GLMM), and the mean square error for the fixed effects is also generally smaller than for GLMM, but larger than the GLM.

We only focussed here on the MLE. In contrast, PQL computes the variance components using REML, with the purpose of reducing the bias in the variance components due to the estimation of $\beta$. REML estimation for GLMM is discussed by McCulloch and Searle (2001).

The aforementioned methods also differ in their ability to estimate the various relevant quantities. Not surprisingly, the most versatile ones are the MCMC-based methods (e.g., Bayes–Gibbs sampler, importance sampling-EM and slice-EM). In contrast with the PQL or corrected-Laplace methods, the slice-EM, along with the MCMC-based methods allows straightforward estimation of confidence bands for arbitrary functions of the parameters and for individual estimates (or predictors) for the random effects.

We conclude by emphasizing that although we report confidence intervals for the variance components, they cannot be readily used for testing the existence of the random effects within the usual hypotheses testing framework. This is because the null hypothesis lies on the boundary of the parameter space, and thus the standard asymptotic results do not apply. For example, the null distribution of the likelihood ratio statistic is no longer the standard chi-squared, but rather similar to the cases considered by Self and Liang (1987) and exemplified by Stram and Lee (1994) in the linear mixed-effects model. In these cases, the standard chi-square distribution is typically a conservative approximation to the null distribution. Here, the issue of testing goes beyond the usual testing of parameter values, as it covers the fundamental issue of model checking and model selection. More work is needed in this area for GLMM, and effective mode-finding algorithms (together with information

calculation), such as slice-EM algorithms we described, are an integral part of such research.

## Acknowledgements

We would like to thank Mei-Hsiu Chen for valuable computational assistance and David van Dyk for helpful conversations. The research was supported in part by NSF grants DMS 94-03560 and DMS-9626691 and NSA grants MDA 904-96-1-0007 and MDA 904-99-1-0067.

## References

Anderson DA and Hinde JP (1988) Random effects in generalized linear models and the EM algorithm. *Communications in Statistical Theory and Methods* **17**, 3847–56.

Biscarat JC (1994) Almost sure convergence of a class of stochastic algorithms. *Stochastic Processes and their Applications* **50**, 83–9.

Booth JG and Hobert JP (1999) Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society Series B* **61**, 265–85.

Breslow NE and Clayton DG (1993) Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**(421), 9–25.

Chan JSK and Kuk AYC (1997) Estimation for probit-linear mixed models with correlated random effects. *Biometrics* **53**, 86–97.

Chan KS and Ledolter J (1995) Monte Carlo EM estimation of time series models involving counts. *Journal of the American Statistical Association* **90**(429), 242–52.

Clayton DG (1996) Generalized linear mixed models. In Gilks WR, Richardson S and Spiegelhalter DJ, editors, *Markov Chain Monte Carlo in practice*. London: Chapman & Hall, 275–301.

Damien P, Wakefield J and Walker S (1999) Gibbs sampling for Bayesian non-conjugate and hierarchical models using auxilliary variables. *Journal of the Royal Statistical Society Series B* **61**, 331–44.

Diggle PJ, Liang KY and Zeger SL (1994) *Analysis of longitudinal data*. New York: Oxford University Press.

Gilks WR and Wild P (1992) Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* **41**, 337–48.

Karim MR and Zeger SL (1992) Generalized linear models with random effects: salamander mating revisited. *Biometrics* **48**, 631–44.

Laird NM and Ware JH (1982) Random-effects models for longitudinal data. *Biometrics* **38**, 963–74.

Lee Y and Nelder JA (1996) Hierarchical generalized linear models. *Journal of the Royal Statistical Society Series B* **58** (4), 619–78.

Lee Y and Nelder JA (2001) Hierarchical generalised linear models: a synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika* **88** (4), 987–1006.

Levine RA and Casella G (2001) Implementations of the Monte Carlo EM algorithm. *Journal of Computational and Graphical Statistics* **10**, 422–39.

Louis TA (1982) Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society Series B* **44** (2), 190–200.

McCulloch CE (1997) Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association* **92** (437), 162–70.

McCullagh P and Nelder JA (1989) *Generalized linear models*. 2nd Edition. London: Chapman & Hall, 176.

McCulloch CE and Searle SR (2001) *Generalized, linear, and mixed models*. New York: Wiley, 176.

Meng X-L and Schilling S (1996) Fitting full-information item factor models and an empirical investigation of bridge sampling. *Journal of the American Statistical Association* **91** (435), 1254–67.

Meng X-L and van Dyk D (1997) The EM algorithm – an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society Series B* **59** (3), 511–67 (with discussion).

Meng X-L and van Dyk D (1998) Fast EM-type implementations for mixed-effects models. *Journal of the Royal Statistical Society Series B* **60**, 559–78.

Mira A and Tierney L (2002) On the use of auxiliary variables in Markov chain Monte Carlo sampling. *Scandinavian Journal of Statistics* **29**, 1–12.

Neal RM (2003) Slice sampling. *Annals of Statistics* **31**, 705–67.

Neuhaus JM and Segal MR (1997) An assessment of approximate maximum likelihood estimators in generalized linear mixed models. In Gregoire T, Brillinger D, Diggle P, Russek-Cohen E, Warren W and Wolfinger R editors, *Modelling longitudinal and spatially correlated data*. New York: Springer, 11–22.

Orchard T and Woodbury MA (1972) A missing information principle, theory and application. *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability* **1**, 697–715.

Self S and Liang K-Y (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* **82** (398), 605–10.

Shun Z (1997) Another look at the salamander mating data: a modified Laplace approximation approach. *Journal of the American Statistical Association* **92** (437), 341–9.

Shun Z and McCullagh P (1995) Laplace approximation of high-dimensional intergrals. *Journal of the Royal Statistical Society Series B* **57**, 749–60.

Skrondal A and Rabe-Hesketh S (2004) Generalized latent variable modeling: multilevel, longitudinal and structural equation models. Boca Raton, FL: CRC Press.

Steele BM (1996) A modified EM algorithm for estimation in generalized mixed models. *Biometrics* **52**, 1295–1310.

Stram DO and Lee JW (1994) Variance components testing in the longitudinal mixed effects model. *Biometrics* **50**, 1171–7.

Vaida F (1998) *At the confluence of the EM algorithm and Markov chain Monte Carlo: theory and applications*. Ph.D. thesis, The University of Chicago.

Vaida F (2005) Convergence of the EM and MM algorithms. *Statistica Sinica* **15**(3), 831–40.

van Dyk D and Meng X-L (2001) The art of data augmentation. *Journal of Computational and Graphical Statistics* **10**, 1–111 (with discussion).

Wu C-F (1983) On the convergence properties of the EM Algorithm. *Annals of Statistics* **11**, 95–103.

Zeger SL, Liang K-Y and Albert PS (1988) Models for longitudinal data: a generalized estimating equation approach. *Biometrics* **44**, 1049–60.