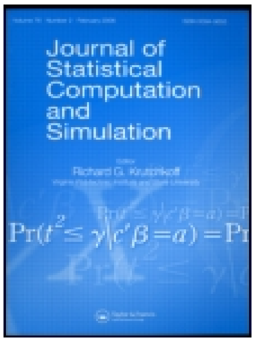

Taylor & Francis
Taylor & Francis Group

Journal of Statistical Computation and Simulation

ISSN: 0094-9655 (Print) 1563-5163 (Online) Journal homepage: https://www.tandfonline.com/loi/gscs20

An efficient MCEM algorithm for fitting generalized linear mixed models for correlated binary data

M. Tan , G.-L. Tian & H.-B. Fang

To cite this article: M. Tan , G.-L. Tian & H.-B. Fang (2007) An efficient MCEM algorithm for fitting generalized linear mixed models for correlated binary data, Journal of Statistical Computation and Simulation, 77:11, 929-943, DOI: 10.1080/10629360600843153

To link to this article: https://doi.org/10.1080/10629360600843153

Published online: 30 Oct 2007.

Journal of Statistical Computation and Simulation
Vol. 77, No. 11, November 2007, 929–943

An efficient MCEM algorithm for fitting generalized linear mixed models for correlated binary data

M. TAN*, G.-L. TIAN and H.-B. FANG

Division of Biostatistics, University of Maryland Greenebaum Cancer Center,
22 South Greene Street, Baltimore, Maryland 21201, USA

(Received 13 May 2005; in final form 6 June 2006)

Generalized linear mixed models have been widely used in the analysis of correlated binary data arisen in many research areas. Maximum likelihood fitting of these models remains to be a challenge because of the complexity of the likelihood function. Current approaches are primarily to either approximate the likelihood or use a sampling method to find the exact likelihood solution. The former results in biased estimates, and the latter uses Monte Carlo EM (MCEM) methods with a Markov chain Monte Carlo algorithm in each E-step, leading to problems of convergence and slow convergence. This paper develops a new MCEM algorithm to maximize the likelihood for generalized linear mixed probit-normal models for correlated binary data. At each E-step, utilizing the inverse Bayes formula, we propose a direct importance sampling approach (*i.e.* weighted Monte Carlo integration) to numerically evaluate the first- and the second-order moments of a truncated multivariate normal distribution, thus eliminating problems of convergence and slow convergence. To monitor the convergence of the proposed MCEM, we again employ importance sampling to directly calculate the log-likelihood values and then to plot the difference of the consecutive log-likelihood values against the MCEM iteration. Two real data sets from the children's wheeze study and a three-period crossover trial are analyzed to illustrate the proposed method and for comparison with existing methods. The results show that the new MCEM algorithm outperformed that of McCulloch [McCulloch, C.E., 1994, Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association*, **89**, 330–335.] substantially.

Keywords: Correlated binary data; Data augmentation; Generalized linear mixed models; Gibbs sampler; Inverse Bayes formula; Monte Carlo EM algorithm; MCMC

1. Introduction

Generalized linear mixed models (GLMMs) are widely used in many areas of research ranging from biomedicine to physical sciences [1, p. 2]. Although likelihood-based methods for the GLMM for Gaussian responses are well developed [2–4], maximum likelihood fitting of the GLMM for correlated binary data remains to be a challenge because of the complexity of the likelihood function. The complexity can be glimpsed through a simple example [5], where the binary outcome Y_{ij} is modeled by a logistic mixed model: $\text{logit}\{\Pr(Y_{ij} = 1|\alpha, \beta)\} = \mu + \alpha_i + \beta_j$, where $\{\alpha_i\}_1^{40} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$, $\{\beta_j\}_1^{40} \overset{\text{i.i.d.}}{\sim} N(0, \tau^2)$, and the α's and β's are

*Corresponding author. Email: mtan@umm.edu

Journal of Statistical Computation and Simulation
ISSN 0094-9655 print/ISSN 1563-5163 online © 2007 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/10629360600843153

independent. The likelihood function of μ, σ^2 and τ^2 involves an 80-dimensional integral whose integrand is a product of 1600 terms. A great deal of recent attention has focused on the development of efficient methods to maximize the likelihood of GLMMs for correlated binary outcomes [6, 7].

To tackle the complicated likelihood, Breslow and Clayton [8] proposed approximate maximum likelihood estimation (MLE) with penalized quasi-likelihood estimates for the fixed effects parameters and restricted MLEs for the variance parameters in the random effects. However, the method is known to yield biased estimates, in particular, to underestimate the variance parameters [5]. To correct the bias, Lin and Breslow [9] used a fourth-order Laplace approximation but it remains to be problematic for correlated binary data. Higher order approximations have been proposed but they are restricted to one random effect per cluster with a large sample size [10]. In addition, these approximate MLEs have been shown to be inconsistent under standard (small domain) asymptotic assumptions and the asymptotic bias can be severe if the variance components are not small [7, 9, 11, 12].

Another approach is to derive the exact MLE in GLMMs by the Monte Carlo EM (MCEM) algorithm [13]. While the M-step is relatively straightforward, the Monte Carlo E-step involves an intractable high-dimensional integral. To fit a probit-normal GLMM for correlated binary data, McCulloch [6] proposed an MCEM algorithm with a Gibbs sampler at each E-step. Later, McCulloch [14] used a Hastings–Metropolis algorithm at each E-step in the MCEM to fit more general models. Both of these two algorithms lead to problems of convergence and slow convergence. Recognizing that MCEM algorithms based on independent samples can be computationally more efficient than Markov chain Monte Carlo (MCMC) EM algorithms based on dependent samples, Booth and Hobert [7] proposed a rejection sampling and a multivariate t importance sampling to produce independent samples at each E-step for GLMMs. However, besides its low acceptance rate, the rejection sampling is difficult to implement in practice because finding the envelope function and the supremum at each E-step is equivalent to finding the MLE of the regression parameter for a GLMM with iteratively reweighted least squares [15, p. 206]. In addition, the multivariate t importance sampling requires to find the second-derivative matrix of a complicated function and to derive the Laplace approximation [7] for the mean and variance in each E-step. A comprehensive surveys of the statistical inferences on the GLMMs is provided recently by McCulloch and Searle [16, Chapter 8].

Other methods for estimation in GLMMs include notably the Bayesian approach where posteriors are approximated under a diffuse prior. For example, Zeger and Karim [17] used MCMC methods in the hierarchical logit-normal model with rejection sampling in each iteration to obtain independent samples of the fixed and random effects. Thus, each iteration involves finding the mode and curvature of a complicated likelihood, increasing computational times exponentially. More recently, for longitudinal binary data, Chib [18] proposed a four-block MCMC sampling for the probit-normal model where each cycle of the MCMC requires another Gibbs sampler to generate truncated multivariate normal distribution, thus, slowing the convergence considerably. In addition, the assessment of convergence to the stationary distribution of a Markov chain required in using an MCMC method such as the Gibbs sampler remains to be problematic [14]. Furthermore, the Bayesian posterior mode resulted from diffuse priors may not always be near the MLE [19, 20].

In this paper, we develop a new and efficient MCEM algorithm to find MLEs of parameters in the probit-normal GLMM for correlated binary data. This model is shown to be advantageous to logit-normal model [14]. Specifically, at each E-step, utilizing the inverse Bayes formula [21], we propose a non-iterative importance sampling approach (*i.e.* weighted Monte carlo integration) to numerically evaluate the first- and the second-order moments of a truncated multivariate normal distribution, thus eliminating problems of convergence and slow convergence associated with MCMC. Thus the method is far more direct than that of McCulloch [6]. To monitor the convergence of the proposed MCEM, we again employ importance sampling to directly calculate the log-likelihood values and then to plot the difference of the consecutive log-likelihood values against the MCEM iteration. The numerical results show that the new MCEM algorithm outperformed that of McCulloch [6] substantially.

The rest of this article is organized as follows. The probit-normal GLMM for correlated binary data is formulated in section 2. A new MCEM algorithm is developed in section 3. In section 4, we analyze two real data sets from the children's wheeze study in six cities and a three-period crossover trial to illustrate the proposed methods and for comparison with existing methods. We conclude with a discussion in section 5.

2. Generalized linear mixed probit-normal models

Let Y_{ij} denote the binary outcome 0 or 1 of the jth measurement and $Y_i = (Y_{i1}, \dots, Y_{in_i})^{\mathrm{T}}$ be the collection of responses from subject i, where $i = 1, \dots, m$ and $j = 1, \dots, n_i$. The generalized linear mixed probit-normal model [6] assumes that given the random effects b_i, the responses $\{Y_{ij}\}_{j=1}^{n_i}$ are conditionally independent with probability

$$\Pr\{Y_{ij} = 1 | b_i, \beta\} = \Phi(\mu_{ij}), \quad \mu_{ij} = x_{ij}^{\mathrm{T}}\beta + w_{ij}^{\mathrm{T}} b_i, \quad b_i | D \overset{\text{i.i.d.}}{\sim} N_q(0, D),$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of $N(0, 1)$, $x_{ij}^{\mathrm{T}} = (x_{ij1}, \dots, x_{ijp})$ and $w_{ij}^{\mathrm{T}} = (w_{ij1}, \dots, w_{ijq})$ are known covariates, β is the $p \times 1$ fixed effect, $\{b_i\}_{i=1}^m$ are the $q \times 1$ random effects, D is unknown $q \times q$ matrix relating to the correlation structure of Y_i. Let $Y_{\mathrm{obs}} = \{Y_i, X_i, W_i\}_1^m$ denote the observed data and $X_i = (x_{i1}, \dots, x_{in_i})$ and $W_i = (w_{i1}, \dots, w_{in_i})$ be two $p \times n_i$ and $q \times n_i$ covariate matrices, then the observed-data likelihood for the unknown parameters $\theta = (\beta, D)$ is given by

$$L(\theta | Y_{\mathrm{obs}}) = \prod_{i=1}^{m} \int N_q(b_i | 0, D) \psi_i(\beta, b_i) \, \mathrm{d}b_i, \tag{1}$$

where $N_q(\cdot | 0, D)$ denotes the normal density with mean 0 and covariance matrix D, $y_i = (y_{i1}, \dots, y_{in_i})^{\mathrm{T}}$ is the realization of Y_i, and

$$\psi_i(\beta, b_i) \equiv \prod_{j=1}^{n_i} [\Phi(\mu_{ij})]^{y_{ij}} [1 - \Phi(\mu_{ij})]^{1-y_{ij}}. \tag{2}$$

Directly maximizing $L(\theta | Y_{\mathrm{obs}})$ is often difficult and this is particularly true when m, n_i and q are very large. Alternatively, we augment the observed data Y_{obs} with the latent data $\{b_i, Z_i\}_1^m$ by defining $Y_{ij} = I_{(Z_{ij}>0)}$ for $i = 1, \dots, m$ and $j = 1, \dots, n_i$, where $I_{(\cdot)}$ denotes the indicator function, $Z_i | (b_i, \beta) \overset{\text{i.i.d.}}{\sim} N_{n_i}(\mu_i, I_{n_i})$ with $Z_i = (Z_{i1}, \dots, Z_{in_i})^{\mathrm{T}}$,

$$\mu_i = (\mu_{i1}, \dots, \mu_{in_i})^{\mathrm{T}} = X_i^{\mathrm{T}}\beta + W_i^{\mathrm{T}} b_i, \tag{3}$$

and I_{n_i} denotes the $n_i \times n_i$ identity matrix. Thus, the conditional probability of Y_i is

$$\Pr\{Y_i = y_i | b_i, \beta\} = \int_{B_i} N_{n_i}(Z_i | \mu_i, I_{n_i}) \, \mathrm{d}Z_i, \quad b_i | D \overset{\text{i.i.d.}}{\sim} N_q(0, D), \tag{4}$$

where $B_i = B_{i1} \times \cdots \times B_{in_i}$ and B_{ij} is the interval $(0, \infty)$ if $y_{ij} = 1$ and the interval $(-\infty, 0]$ if $y_{ij} = 0$. It is important to note that B_i depends only on the value of y_i and not on the parameters.

3. A new MCEM algorithm

In this section, we first derive the M- and E-step for the new MCEM algorithm, and then we show how to use the importance sampling instead of the Gibbs sampling at each E-step. We also provide methods for monitoring the convergence of the MCEM algorithm and for calculating the standard errors.

3.1 *The derivation of M- and E-step*

Let $b = \{b_i\}_1^m$ and $Z = \{Z_i\}_1^m$, we treat both b and Z as missing data. From the definition of the latent data Z, $\{Y_{\text{obs}}, Z\} = Z$, then the joint density for the complete-data $Y_{\text{com}} = \{Y_{\text{obs}}, b, Z\} = \{b, Z\}$ is

$$f(Y_{\text{com}}|\theta) = \prod_{i=1}^{m}\{f(b_i|D)\cdot f(Z_i|b_i,\beta)\} = \prod_{i=1}^{m}\{N_q(b_i|0,D)\cdot N_{n_i}(Z_i|\mu_i, I_{n_i})\}, \tag{5}$$

where $\{\mu_i\}_1^m$ are defined in equation (3). The M-step of the MCEM algorithm is to find the complete-data MLE of θ by maximizing the conditional expectation of the complete-data log-likelihood $\ell(\theta|Y_{\text{obs}}, b, Z) = \log f(Y_{\text{com}}|\theta)$ given the observed data Y_{obs} and the current estimate $\theta^{(t)}$, *i.e.* $Q(\theta|\theta^{(t)}) = E\{\ell(\theta|Y_{\text{obs}}, b, Z)|Y_{\text{obs}}, \theta^{(t)}\}$. We have the following closed-form expressions

$$\hat{\beta} = \left(\sum_{i=1}^{m} X_i X_i^{\mathrm{T}}\right)^{-1} \sum_{i=1}^{m} X_i(Z_i - W_i^{\mathrm{T}} b_i) \quad \text{and} \quad \hat{D} = \frac{1}{m}\sum_{i=1}^{m} b_i b_i^{\mathrm{T}}. \tag{6}$$

The E-step computes conditional expectations of the complete-data sufficient statistics, *i.e.* $E(Z_i|Y_i,\theta)$, $E(b_i|Y_i,\theta)$ and $E(b_i b_i^{\mathrm{T}}|Y_i,\theta)$ for $i = 1,\dots,m$.

To calculate these conditional expectations, we first derive the conditional predictive distribution of the missing data, which is given by

$$f(b, Z|Y_{\text{obs}},\theta) = f(b|Y_{\text{obs}}, Z,\theta)\cdot f(Z|Y_{\text{obs}},\theta) \tag{7}$$

$$= f(Z|Y_{\text{obs}}, b,\theta)\cdot f(b|Y_{\text{obs}},\theta). \tag{8}$$

As $f(b|Y_{\text{obs}}, Z,\theta)$ is proportional to the joint density equation (5), we immediately obtain

$$f(b|Y_{\text{obs}}, Z\theta) = \prod_{i=1}^{m} f(b_i|Y_i, Z_i,\theta) = \prod_{i=1}^{m} f(b_i|Z_i,\theta)$$

$$= \prod_{i=1}^{m} N_q(b_i|\Delta_i(Z_i - X_i^{\mathrm{T}}\beta), \Lambda_i) \tag{9}$$

where $\Delta_i \equiv DW_i\Omega_i^{-1}$, $\Lambda_i \equiv D - DW_i\Omega_i^{-1}W_i^{\mathrm{T}}D$, and $\Omega_i \equiv W_i^{\mathrm{T}}DW_i + I_{n_i}$, $i = 1,\dots,m$. To derive the second term on the right-hand side of equation (7), we use the following result [22, p. 350].

$$\Pr\{Y_i = y_i|b_i, Z_i,\theta\} = I_{(Z_i\in B_i)} = \prod_{j=1}^{n_i}\{I_{(Z_{ij}>0)}I_{(Y_{ij}=1)} + I_{(Z_{ij}\le 0)}I_{(Y_{ij}=0)}\}, \tag{10}$$

which indicates that given Z_i, the conditional probability of Y_i is independent of b_i. Hence equation (10) implies $\Pr\{Y_i = y_i|Z_i,\theta\} = I_{(Z_i\in B_i)}$. As the joint density of (Z_i, b_i) is normally distributed, the marginal distribution of $Z_i|\theta$ follows $N_{n_i}(X_i^{\mathrm{T}}\beta, \Omega_i)$. Moreover, $f(Z_i|Y_i,\theta) \propto f(Z_i, Y_i|\theta) = f(Z_i|\theta) \cdot \Pr\{Y_i = y_i|Z_i,\theta\} = N_{n_i}(Z_i|X_i^{\mathrm{T}}\beta, \Omega_i) \cdot I_{(Z_i \in B_i)}$, yielding

$$f(Z|Y_{\mathrm{obs}},\theta) = \prod_{i=1}^{m} f(Z_i|Y_i,\theta) = \prod_{i=1}^{m} TN_{n_i}(Z_i|X_i^{\mathrm{T}}\beta, \Omega_i; B_i), \tag{11}$$

where $TN_{n_i}(\cdot|X_i^{\mathrm{T}}\beta, \Omega_i; B_i)$ denotes the n_i-dimensional normal density with mean vector $X_i^{\mathrm{T}}\beta$ and covariance matrix Ω_i but truncated to the rectangle B_i.

In section 3.2, we provide an importance sampling approach to evaluate $E(Z_i|Y_i,\theta) \equiv M_1^{(i)}$ and $E(Z_iZ_i^{\mathrm{T}}|Y_i,\theta) \equiv M_2^{(i)}$ based on equation (11). Once both $M_1^{(i)}$ and $M_2^{(i)}$ are available, the equation (9) gives the Rao–Blackwell estimates

$$E(b_i|Y_i,\theta) = E\{E(b_i|Y_i,Z_i,\theta)|Y_i,\theta\} = \Delta_i[M_1^{(i)} - X_i^{\mathrm{T}}\beta], \tag{12}$$

$$\begin{aligned} E(b_ib_i^{\mathrm{T}}|Y_i,\theta) &= E\{E(b_ib_i^{\mathrm{T}}|Y_i,Z_i,\theta)|Y_i,\theta\} \\ &= \Lambda_i + \Delta_i[M_2^{(i)} + \gamma_i\gamma_i^{\mathrm{T}} - M_1^{(i)}\gamma_i^{\mathrm{T}} - (M_1^{(i)}\gamma_i^{\mathrm{T}})^{\mathrm{T}}]\Delta_i^{\mathrm{T}}, \end{aligned} \tag{13}$$

where $\gamma_i \equiv X_i^{\mathrm{T}}\beta$.

3.2 *The use of importance sampling at each E-step*

McCulloch [6] suggested using a Gibbs sampling (*e.g.* [23]) at each E-step to estimate $E(Z_i|Y_i,\theta)$ and $E(Z_iZ_i^{\mathrm{T}}|Y_i,\theta)$ from the truncated multivariate normal distribution (11). Although this Gibbs sampling is easy to implement, it may suffer from convergence and slow convergence owing to the high correlation between $Z_{ij}|(Y_i,\theta)$ and $Z_{ij'}|(Y_i,\theta)$ [24, p. 45]. To overcome these difficulties, using the *inverse Bayes formula* (IBF), we propose a non-iterative importance sampling approach to numerically evaluate $E(Z_i|Y_i,\theta)$ and $E(Z_iZ_i^{\mathrm{T}}|Y_i,\theta)$ at each E-step.

For this purpose, at the beginning, we need to derive the first term on the right-hand side of equation (8). From equations (5) and (10), we have $f(Z_i|Y_i,b_i,\theta) \propto f(Y_i,b_i,Z_i|\theta) = f(b_i|\theta) \cdot f(Z_i|b_i,\theta) \cdot \Pr\{Y_i = y_i|b_i,Z_i,\theta\} \propto N_{n_i}(Z_i|\mu_i, I_{n_i}) \cdot I_{(Z_i \in B_i)}$. Thus,

$$f(Z|Y_{\mathrm{obs}},b,\theta) = \prod_{i=1}^{m} f(Z_i|Y_i,b_i,\theta) = \prod_{i=1}^{m} TN_{n_i}(Z_i|\mu_i, I_{n_i}; B_i), \tag{14}$$

where $\{\mu_i\}_1^m$ are defined in equation (3). Next, from equations (7) and (8), we have $f(Z_i|Y_i,\theta) \propto f(Z_i|Y_i,b_i,\theta)/f(b_i|Y_i,Z_i,\theta)$ for arbitrary b_i. Let b_i^0 be an arbitrary point in the support of b_i, the function-wise IBF [21] gives

$$f(Z_i|Y_i,\theta) = \left\{\int \frac{f(Z_i|Y_i,b_i^0,\theta)}{f(b_i^0|Y_i,Z_i,\theta)}\,\mathrm{d}Z_i\right\}^{-1} \cdot \frac{f(Z_i|Y_i,b_i^0,\theta)}{f(b_i^0|Y_i,Z_i,\theta)}, \quad i = 1,\ldots,m, \tag{15}$$

where the denominator and the numerator are given by equations (9) and (14), respectively. Let $\{Z_i^{(k)}\}_{k=1}^K$ be an i.i.d. sample from the numerator $f(Z_i|Y_i,b_i^0,\theta)$, then the moments can be estimated by the weighted means

$$E(Z_i|Y_i,\theta) \doteq \sum_{k=1}^{K} \omega_i^{(k)} Z_i^{(k)}, \quad E(Z_i Z_i^{\mathrm{T}}|Y_i,\theta) \doteq \sum_{k=1}^{K} \omega_i^{(k)} Z_i^{(k)} Z_i^{(k)\mathrm{T}}, \tag{16}$$

and the weights are given by

$$\omega_i^{(k)} = \frac{\delta_i^{(k)}}{\delta_i^{(1)} + \cdots + \delta_i^{(K)}}, \quad \delta_i^{(k)} \equiv f^{-1}(b_i^0|Y_i, Z_i^{(k)}, \theta), \quad k = 1, \ldots, K. \tag{17}$$

It is worth noting that the weights are numerically stable. Therefore, the proposed approach is different from those associated with the harmonic mean of Newton and Raftery [25] which, as pointed out by Gelfand and Dey [26], is likely to suffer from numeric instability as the reciprocal of a density may approach infinity. In fact, the weights $\{\omega_i^{(k)}\}_{k=1}^K$ in equation (17) depend on the density values $\{f(b_i^0|Y_i, Z_i^{(k)}, \theta)\}_{k=1}^K$ only via the ratio of two such density values, as they can be rewritten as follows

$$\omega_i^{(k)} = \frac{\tau_i^{(k,k_0)}}{1 + \sum_{\ell=1,\ell\neq k_0}^{K} \tau_i^{(\ell,k_0)}}, \quad \tau_i^{(k,k_0)} \equiv \frac{f(b_i^0|Y_i, Z_i^{(k_0)}, \theta)}{f(b_i^0|Y_i, Z_i^{(k)}, \theta)} \in (0, 1], \tag{18}$$

where $f(b_i^0|Y_i, Z_i^{(k_0)}, \theta) \equiv \min_{1\le k\le K}\{f(b_i^0|Y_i, Z_i^{(k)}, \theta)\}$. If $f(b_i^0|Y_i, Z_i^{(k_0)}, \theta) \to 0$, then $\omega_i^{(k_0)} \to 1$ and $\omega_i^{(k)} \to 0$ for $k \neq k_0$. Similar ratios of two densities have been employed by many authors in simulating ratios of normalizing constants via bridge sampling, *e.g.* Meng and Wong [27, p. 837], Chen and Shao [28, 29]. According to our experience, (18) helps enhance numeric accuracy in calculating the weights $\{\omega_i^{(k)}\}_{k=1}^K$.

A natural problem for efficiently computing (16) is how to choose b_i^0 in equation (15). It suffices to select a b_i^0 such that $f(Z_i|Y_i, b_i^0, \theta)$ best approximates $f(Z_i|Y_i, \theta)$. Theorem 1 of Tan *et al.* [21] has shown that the best choice of b_i^0 is the mode of $f(b_i|Y_i, Z_i, \theta)$. As the mode and the mean for a normal distribution is identical, therefore from equation (9), theoretically, we can choose $b_i^0 = \Delta_i(Z_i - X_i^{\mathrm{T}}\beta)$. In practice, as Z_i is unknown, we may replace Z_i by $E(Z_i|Y_i, \theta)(= M_1^{(i)})$ and choose $b_i^0 = b_i^{(t)} = E(b_i|Y_i, \theta^{(t)})$ calculated according to equation (12).

3.3 *Monitoring the convergence of the MCEM algorithm*

An important issue in implementing the MCEM is to assess convergence of the algorithm, which is closely related to the choice of the Monte Carlo sample size K in equation (16). Wei and Tanner [13] suggested plotting individual components of $\theta^{(t)}$ against t and terminating the iteration after a random fluctuation around the $\theta = \hat{\theta}$ line is reached. However, this approach is impractical when the number of parameters is very large. Meng and Schilling [30] proposed to use the bridge sampling for simulating likelihood ratios to monitor the convergence of MCEM. The bridge sampling is suitable to MCEM algorithms where at each E-step an MCMC is utilized to generate dependent samples. By constructing a sandwich variance estimate for the maximizer at each E-step, Booth and Hobert [7] provided an automatic rule to increase the Monte Carlo sample size after iterations in which the change in the parameter value is swamped by Monte Carlo error. However, their rule is based on random samples generated by some non-iterative approach, *i.e.* the rejection sampling rather than dependent samples generated by MCMC.

To monitor the convergence of the new MCEM algorithm, we again employ the importance sampling to directly calculate the log-likelihood values and then to plot $\log L(\theta^{(t)}|Y_{\text{obs}})$ against the MCEM iteration t or the difference of the consecutive log-likelihood values $\log L(\theta^{(t+1)}|Y_{\text{obs}}) - \log L(\theta^{(t)}|Y_{\text{obs}})$ versus t. Specifically, let $b_i^{(\ell)} \overset{\text{i.i.d.}}{\sim} N_q(0, D)$ for $i = 1, \dots, m$ and $\ell = 1, \dots, M$, then from equation (1) the log-likelihood function can be approximated directly by Monte Carlo approach via importance sampling, namely,

$$\log L(\theta|Y_{\text{obs}}) = \sum_{i=1}^{m} \log L(\theta|Y_i) \doteq \sum_{i=1}^{m} \log \left[\frac{1}{M} \sum_{\ell=1}^{M} \psi_i(\beta, b_i^{(\ell)}) \right], \tag{19}$$

where $L(\theta|Y_i)$ is the observed likelihood of θ contributed from subject i $(i = 1, \dots, m)$, and $\psi_i(\beta, b_i)$ is defined in equation (2). Given $\theta^{(t)}$ and $\theta^{(t+1)}$, we plot

$$\log \left[\frac{L(\theta^{(t+1)}|Y_{\text{obs}})}{L(\theta^{(t)}|Y_{\text{obs}})} \right] = \sum_{i=1}^{m} \log \left[\frac{L(\theta^{(t+1)}|Y_i)}{L(\theta^{(t)}|Y_i)} \right] \tag{20}$$

against t. If the plot shows the differences vanish, stablizing around zero, we consider the algorithm achieves approximate convergence, which is all we can obtain with MCEM.

3.4 *The calculation of standard errors*

Denote the MLE from the MCEM algorithm by $\hat{\theta} = (\hat{\beta}, \hat{D})$. Louis [31] showed that the observed information matrix is given by

$$-E \left\{ \frac{\partial^2 \ell(\theta|Y_{\text{obs}}, b, Z)}{\partial\theta\partial\theta^{\text{T}}} \right\}\Bigg|_{\theta=\hat{\theta}} - \text{Var} \left\{ \frac{\partial \ell(\theta|Y_{\text{obs}}, b, Z)}{\partial\theta} \right\}\Bigg|_{\theta=\hat{\theta}} \tag{21}$$

where the expectation and variance are with respect to $f(b, Z|Y_{\text{obs}}, \hat{\theta})$. Based on equations (9) and (14), a data augmentation (DA) algorithm [32] can be used to obtain dependent samples from $f(b, Z|Y_{\text{obs}}, \hat{\theta})$ so that equation (21) can be estimated by Monte Carlo methods. Standard errors are equal to the square roots of the diagonal elements of the inverse of the estimated information matrix.

4. Applications

To illustrate the proposed methods, we analyze two real studies where the random effects $\{b_i\}_1^m$ are two-dimensional in the first one and one-dimensional in the second. The proposed importance sampling is used at each E-step of the MCEM to compute the MLEs of parameters and is also used to monitor the convergence of the MCEM. The standard errors are computed according to equation (21). We compare these results with those from the MCEM with a Gibbs sampler at each E-step [6]. All computations are performed entirely in S-PLUS 6 for Windows on a Pentium IV workstation.

4.1 *Children's wheeze data in six cities*

The longitudinal study is on health effects of air pollution in six cities [22]. The data showed in table 1 consist of an annual binary response indicating the presence or absence of wheeze at ages 7, 8, 9 and 10 years for each of 537 children from one city. Maternal smoking was

Table 1. Children's wheeze data in six cities.

No maternal smoking					Maternal smoking				
Age of child (year)					Age of child (year)				
7	8	9	10	Frequency	7	8	9	10	Frequency
0	0	0	0	237	0	0	0	0	118
0	0	0	1	10	0	0	0	1	6
0	0	1	0	15	0	0	1	0	8
0	0	1	1	4	0	0	1	1	2
0	1	0	0	16	0	1	0	0	11
0	1	0	1	2	0	1	0	1	1
0	1	1	0	7	0	1	1	0	6
0	1	1	1	3	0	1	1	1	4
1	0	0	0	24	1	0	0	0	7
1	0	0	1	3	1	0	0	1	3
1	0	1	0	3	1	0	1	0	3
1	0	1	1	2	1	0	1	1	1
1	1	0	0	6	1	1	0	0	4
1	1	0	1	2	1	1	0	1	2
1	1	1	0	5	1	1	1	0	4
1	1	1	1	11	1	1	1	1	7

Source: Chib and Greenberg [22].

categorized as 1 if the mother smoked regularly and 0 otherwise. The objective of the study is to model the probability of wheeze status as a function of the child's age and the mother's smoking habit. We assume that the probability of wheeze status of the ith child at the jth observation is

$$\Pr\{Y_{ij} = 1 | b_i, \beta\} = \Phi(\beta_1 + \text{age}\ \beta_2 + \text{smoking}\ \beta_3 + (\text{age} \times \text{smoking})\beta_4 + w_{ij}^{\mathrm{T}} b_i),$$

where 'age' is the age of the child centered at nine years, 'smoking' $= 1$ if mother smokes, 0 otherwise, $w_{ij}^{\mathrm{T}} = (1, \text{age})$, implying that both intercept and the age effects are children-specific, and $b_i | D \sim N_2(0, D)$, $i = 1, \ldots, m$ $(m = 537)$, $j = 1, \ldots, n_i$ $(n_i = 4)$.

To find the MLEs $\hat{\beta}$ and $\hat{D}$, we use the importance sampling approach described in section 3.2 at each E-step. Specifically, for fixed i $(i = 1, \ldots, m)$, we generate an i.i.d. sample $\{Z_i^{(k)}\}_{k=1}^{K}$ of size K from $f(Z_i | Y_i, b_i^0, \theta)$ (see equations (15) and (14)) and calculate the weights $\{\omega_i^{(k)}\}_{k=1}^{K}$ according to equation (18). Thus, we can estimate $E(Z_i | Y_i, \theta)$ and $E(Z_i Z_i^{\mathrm{T}} | Y_i, \theta)$ by equation (16) and compute $E(b_i | Y_i, \theta)$ and $E(b_i b_i^{\mathrm{T}} | Y_i, \theta)$ via equations (12) and (13). Because of the iterative nature of the EM algorithm, we also use a K that increases linearly with the number of iterations as McCulloch [6]. Specifically, we choose $K = 50$ for iteration 1 to 20, $K = 100$ for iteration 21 to 40, $K = 200$ for iteration 41 to 60, $K = 500$ for iteration 61 to 80, and $K = 1000$ for iteration 81 and over. The M-step is straightforward as equation (6) has closed-form expressions. Using $\theta^{(0)} = (\beta^{(0)}, D^{(0)})$ with $\beta^{(0)} = (0, 0, 0, 0)^{\mathrm{T}}$ and $D^{(0)} = I_2$ as the initial values, the proposed MCEM algorithm converged at the 100th iteration to the MLEs $\hat{\beta}$ and $\hat{D}$. Their standard errors are calculated according to equation (21). These results listed in the third- and fourth-column of table 2 are very close to the posterior means obtained with flat priors [18, p. 121]. Convergence of the MCEM was determined from figures 1(a), 2 and 3(a). As figure 1(a) shows, the logs of individual components of $\theta^{(t)} = (\beta^{(t)}, D^{(t)})$ appear to have stabilized after about 40 iterations. Similarly, the log-likelihood values $\log L(\theta^{(t)} | Y_{\text{obs}})$ estimated by equation (19) also appear to stabilize after 40 EM cycles (figure 2). Figure 3a shows that the logs of the likelihood ratios estimated from equation (20) decrease to zero after 40 iterations. The computing time for finding the MLEs is 40.82 min (0.68033 h) for 40 iterations or 102.05 min (1.7008 h) for 100 iterations.

Table 2. MLEs and standard errors of β and D in the wheeze data.

Variable	Parameter	MLE[†]	Standard error[†]	MLE[‡]	Standard error[‡]
Intercept	β_1	−1.8245	0.0289	−1.8145	0.0295
Age	β_2	−0.1006	0.0237	−0.1145	0.0239
Smoking	β_3	0.1925	0.0490	0.2742	0.0496
Age × smoking	β_4	0.0368	0.0405	0.1032	0.0403
Covariance	d_{11}	1.7517	0.1080	1.6251	0.1005
	d_{12}	−0.0005	0.0070	0.0005	0.0127
	d_{22}	0.0148	0.0009	0.0525	0.0032

[†]Values obtained by the new MCEM algorithm with a non-iterative importance sampling at each E-step, converged at the 100th iteration.

[‡]Values obtained by the MCEM algorithm with a Gibbs sampling at each E-step [6], nearly converged at 100th iteration.

To compare the proposed MCEM algorithm with the MCEM of McCulloch [6], in the E-step, for a fixed i $(i = 1, \ldots, m)$, we run a Gibbs sampler [23] with a single chain to obtain a dependent sample of size $2K$ from truncated multivariate normal distribution $f(Z_i|Y_i, \theta) = TN_{n_i}(Z_i|X_i^{\mathrm{T}}\beta, \Omega_i; B_i)$ (see equation (11)) and no assessment of convergence of the Markov chain to its stationary distribution is made. Then, we use the second half of the sequence to estimate $E(Z_i|Y_i, \theta)$ and $E(Z_i Z_i^{\mathrm{T}}|Y_i, \theta)$. Note that $E(b_i|Y_i, \theta)$ and $E(b_i b_i^{\mathrm{T}}|Y_i, \theta)$ are still given by equations (12) and (13), respectively, and the M-step is also the same as equation (6). We use the same K and the same initial values $(\beta^{(0)}, D^{(0)})$ as that of the proposed MCEM above. We terminate the MCEM at the 100th iteration. The corresponding MLEs $\hat{\beta}$ and $\hat{D}$ and their standard errors calculated according to equation (21) are listed in the fifth- and sixth-column of table 2. Convergence of the MCEM can be assessed based on figures 1b, 2 and 3b. Although figure 3b shows that the logs of the likelihood ratios estimated from equation (20) decrease to zero after 80 iterations, figure 1b shows that the logs of d_{11} seem to have not stabilized after 100 iterations. From figure 2, the log-likelihood values $\log L(\theta^{(t)}|Y_{\mathrm{obs}})$ estimated by equation (19) do not stabilize after 80 EM cycles and the convergence is thus slower than the proposed MCEM. The computing time for the MLEs is 25.406 h for 100 iterations, which is 25.406/0.68033 = 37.344 (or at least 25.406/1.7008 = 14.938) times slower than the new MCEM.

4.2 *A three-period crossover trial*

This study is a three-period crossover trial comparing three treatment levels (placebo, low and high doses) of an analgesic for the relief of primary dysmenorrhoea (table 1, [33]). The placebo is labeled as treatment A, while treatments B and C are the analgesic with low and high doses, respectively. A total of 86 women were randomized to one of six different groups, corresponding to six possible orders of three treatments over three periods. At the end of each period each subject rated the treatment as giving either no pain relief (coded as response 0) or some relief (coded as response 1). The objective of this study is to assess the period effect, the treatment effect, the carryover effect and/or the treatment-by-period interaction. Let Y_{ij} denotes the response of subject i $(i = 1, \ldots, m$ and $m = 86)$ during period j $(j = 1, \ldots, n_i$ and $n_i = 3)$. The responses $Y_{i1}, \ldots, Y_{in_i}$ from a subject are statistically dependent, but are assumed to be conditionally independent given the random effects b_i. They are modelled by

$$\Pr\{Y_{ij} = 1|b_i, \beta\} = \Phi(\mu_{ij}), \quad \mu_{ij} = x_{ij}^{\mathrm{T}}\beta + b_i, \quad b_i|\sigma^2 \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2),$$

where b_i is the subject-specific tolerance to the drugs, σ^2 is the unknown variance, β is an unknown p-vector $(p = 11)$ representing period, treatment, carryover effect and

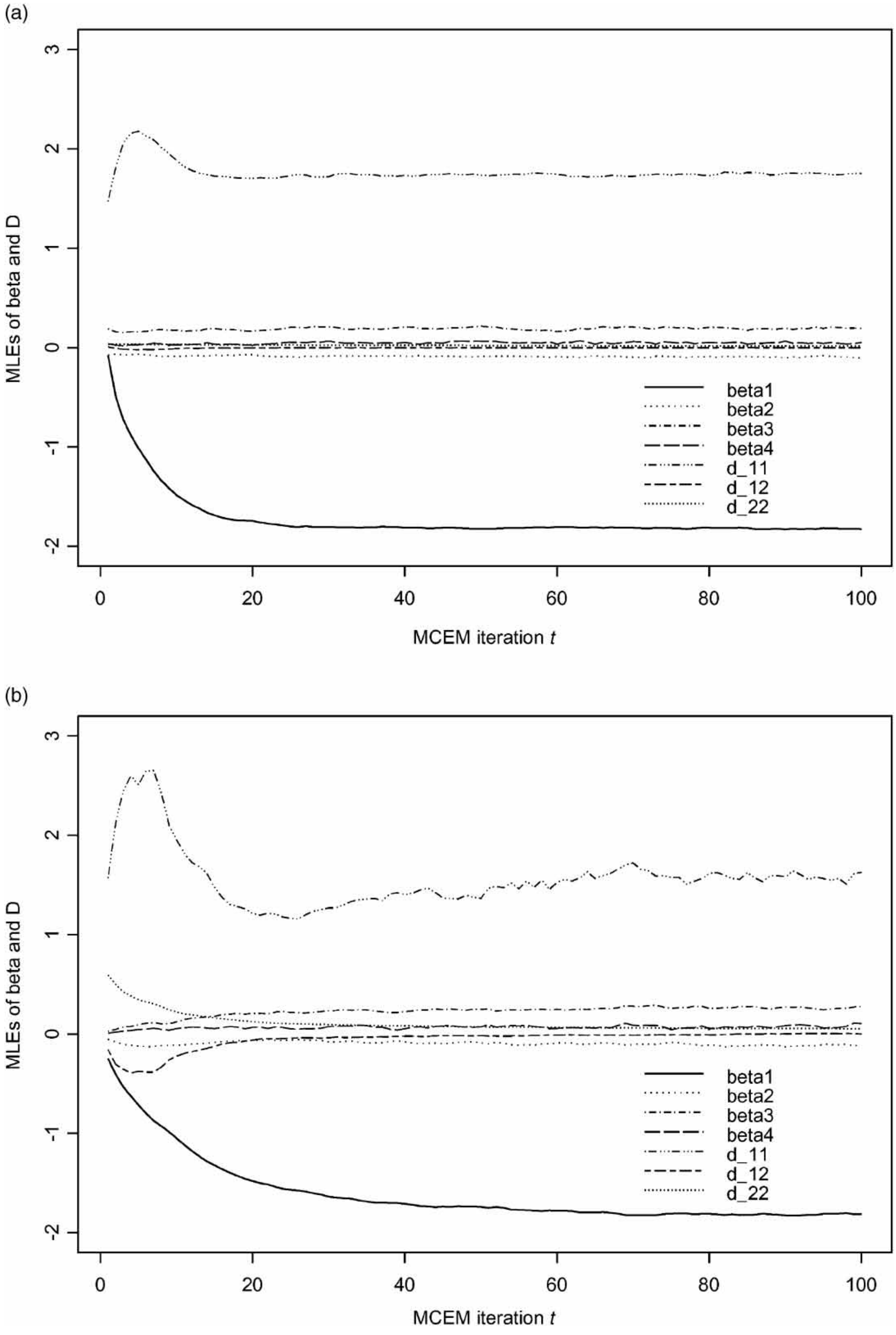

Figure 1. The comparison of convergence between two MCEM algorithms by plotting the MLEs of $\beta = (\beta_1, \ldots, \beta_4)^{\mathrm{T}}$ and D against the MCEM iteration t for the children's wheeze data in six cities: (a) MCEM with importance sampling at each E-step; (b) MCEM with Gibbs sampling at each E-step.

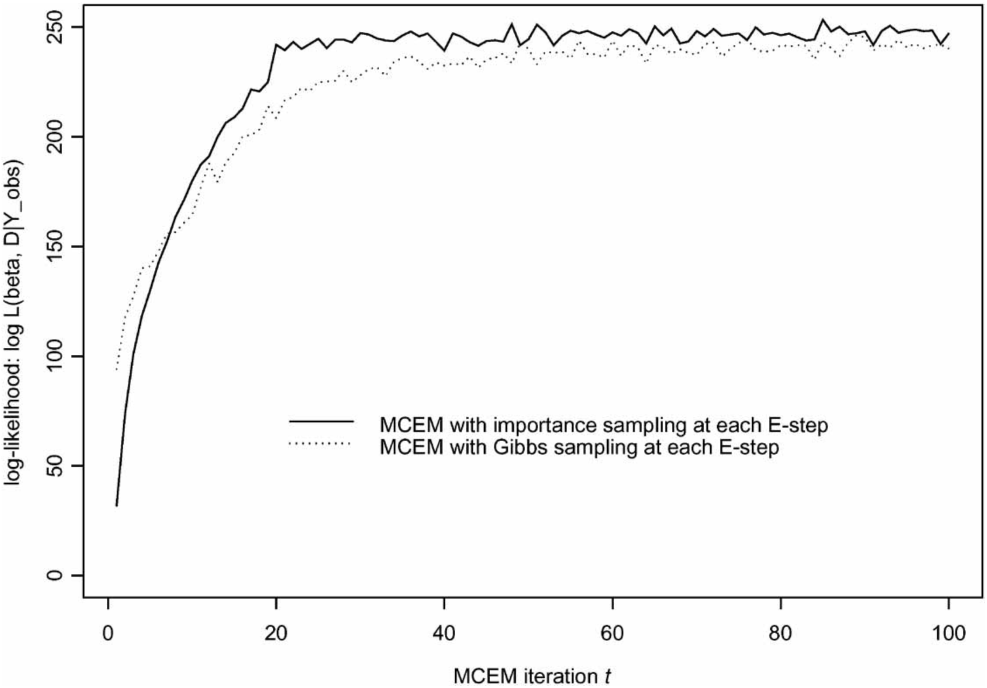

Figure 2. The comparison of convergence between two MCEM algorithms with importance/Gibbs sampling at each E-step by plotting the log-likelihood values against the MCEM iteration t for the children's wheeze data in six cities.

treatment-by-period interaction, and $x_{ij}^{\mathrm{T}} = (x_{ij,1}, \ldots, x_{ij,11})$ is the corresponding covariate vector with $x_1 = 1$, $x_8 = x_2x_4$, $x_9 = x_2x_5$, $x_{10} = x_3x_4$ and $x_{11} = x_3x_5$, where

$$x_2(x_3) = \begin{cases} 1 & \text{period } 2(3) \\ 0 & \text{otherwise,} \end{cases} \qquad x_4(x_5) = \begin{cases} 1 & \text{treatment } B(C) \\ 0 & \text{otherwise,} \end{cases}$$

$$x_6(x_7) = \begin{cases} 1 & \text{the previous assignment is } B(C) \\ 0 & \text{otherwise.} \end{cases}$$

To calculate the MLEs $\hat{\beta}$ and $\hat{\sigma}^2$, we use the importance sampling (section 3.2) at each E-step. For fixed i, we generate an i.i.d. sample $\{Z_i^{(k)}\}_{k=1}^{K}$ of size K from $f(Z_i|Y_i, b_i^0, \theta)$ (see equations (15) and (14)) and calculate the weights $\{\omega_i^{(k)}\}_{k=1}^{K}$ based on equation (18). Thus, we can estimate $E(Z_i|Y_i, \theta)$ and $E(Z_iZ_i^{\mathrm{T}}|Y_i, \theta)$ by equation (16) and compute $E(b_i|Y_i, \theta)$ and $E(b_ib_i^{\mathrm{T}}|Y_i, \theta)$ via equations (12) and (13). Similarly, we choose $K = 50$ for the EM iteration 1 to 50, $K = 100$ for iteration 51 to 100, $K = 150$ for iteration 101 to 150, $K = 175$ for iteration 151 to 175, and $K = 500$ for iteration 176 and over. The M-step is straightforward as equation (6) has closed-form expressions. Using $\theta^{(0)} = (\beta^{(0)}, \sigma^{2(0)})$ with $\beta^{(0)} = (0, \ldots, 0)^{\mathrm{T}}$ and $\sigma^{2(0)} = 1$ as the initial values, the MCEM algorithm converged at the 195th iteration to the MLEs $\hat{\beta}$ and $\hat{\sigma}^2$ (table 3). The computing time for finding the MLEs is 33.24 min. The corresponding standard errors are calculated according to equation (21). The convergence of the MCEM was determined by plotting the MLEs of β and σ^2 against the MCEM iteration (here the plot is omitted).

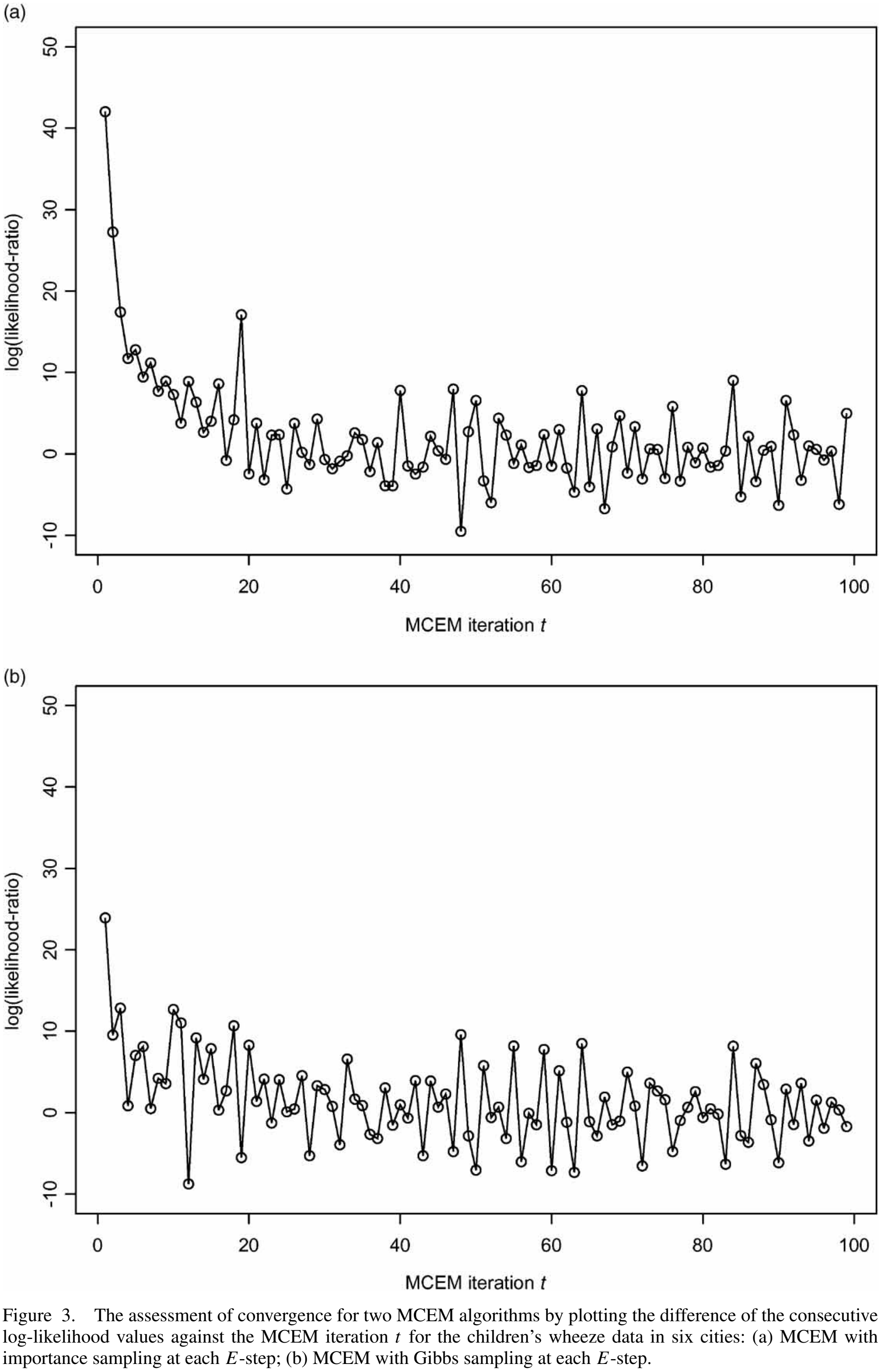

Figure 3. The assessment of convergence for two MCEM algorithms by plotting the difference of the consecutive log-likelihood values against the MCEM iteration t for the children's wheeze data in six cities: (a) MCEM with importance sampling at each E-step; (b) MCEM with Gibbs sampling at each E-step.

Table 3. MLEs and standard errors of β and σ^2 in the crossover trial.

Variable	Parameter	MLE[†]	Standard error
Intercept(x_1)	β_1	2.5253	0.1801
Period$_2$(x_2)	β_2	0.6712	0.3169
Period$_3$(x_3)	β_3	0.3312	0.3245
Treatment$_\mathrm{B}$(x_4)	β_4	0.7303	0.2652
Treatment$_\mathrm{C}$(x_5)	β_5	0.5715	0.2584
Carryover$_\mathrm{B}$(x_6)	β_6	−0.1820	0.2177
Carryover$_\mathrm{C}$(x_7)	β_7	−0.3363	0.2164
Interaction(x_2x_4)	β_8	−0.5327	0.3889
Interaction(x_2x_5)	β_9	−0.2230	0.3874
Interaction(x_3x_4)	β_{10}	−0.0945	0.3913
Interaction(x_3x_5)	β_{11}	0.0368	0.3874
Variance	σ^2	0.0614	0.0097

[†]Values obtained by the new MCEM algorithm with a non-iterative conditional sampling at each E-step, converged at the 195th iteration.

5. Discussion

We developed an efficient MCEM algorithm to find the MLEs of the generalized linear mixed probit-normal model for correlated binary data. At each E-step, we replaced the Gibbs sampler by a non-iterative importance sampling approach to directly evaluate the first- and the second-order moments of a truncated multivariate normal distribution (TMVND), thus eliminating problems of convergence and slow convergence associated with MCMC methods. We combined three graphical methods by plotting individual components of $\theta^{(t)}$, log-likelihood values and log likelihood-ratios against iteration t to monitor the convergence of MCEM. As expected, computation results from the wheeze example show that the proposed MCEM algorithm outperformed the MCEM of McCulloch [6] substantially (about 37 times faster).

Why is the Gibbs sampler in calculating the moments of a TMVND so slow when comparing it with some non-iterative sampling approaches such as importance sampling? We try to give several reasons to explain it. First, as pointed out by McFadden [24, p. 45], the Gibbs sampler often suffers from painful slow convergence owing to the high correlation between components in TMVND. From equation (11) we know that the covariance matrix $\Omega_i^{(t)} = W_i^{\mathrm{T}} D^{(t)} W_i + I_{n_i}$ will affect the speed of convergence of Gibbs sampler, where t is the MCEM iteration. Let $R_i^{(t)}$ denotes the corresponding correlation matrix associated with $\Omega_i^{(t)}$. For the wheeze data, as W_i is equal for all i ($i = 1, \ldots, m$), we have $R_i^{(t)} = R^{(t)}$. When $t = 5$, 10, 20, 40, 60 and 100, the maximal values of correlation coefficients in $R^{(t)}$ are 0.80, 0.74, 0.6, 0.612, 0.614 and 0.627, respectively. Next, Gibbs sampler with a single chain having length 2K will yield only K available samples because of discarding the first half of the sequence. Therefore, its efficiency drop at least a half when comparing with a non-iterative sampling approach. Finally, to generate an n_i-dimensional TMVND, Gibbs sampler needs to employ a loop statement with n_i circle (*e.g.* `FOR(i in 1:`n_i`){ }` in SPLUS) to generate n_i truncated univariate normal distributions because of updating (see equation (11)), while the proposed importance sampling in section 3.2 can simultaneously generate n_i truncated univariate normal distributions (see equations (14)–(16)) in software with vectorized environment such as SPLUS. Note that the `FOR` statement in SPLUS is very slow.

Another potential improvement on the MCEM of McCulloch [6] is to use the DA algorithm instead of the Gibbs sampler. In fact, based on equations (9) and (14), the DA can be used to generate dependent samples $\{Z_i^{[k]}\}_{k=1}^{K}$ from the TMVND (11). Thus, the moments can be estimated by $E(Z_i|Y_i,\theta) \doteq K^{-1}\sum_{k=1}^{K} Z_i^{[k]}$ and $E(Z_i Z_i^{\mathrm{T}}|Y_i,\theta) \doteq K^{-1}\sum_{k=1}^{K} Z_i^{[k]} Z_i^{[k]^{\mathrm{T}}}$. However, as the native Gibbs sampler, the DA algorithm also requires extra steps to assess its convergence to the stationary distribution.

Acknowledgements

We are grateful to the editor, an associate editor and one referee for their very constructive comments and suggestions. This research was supported in part by US National Institute of Health grant CA 106767.

References

[1] Raftery, A.E., Tanner, M.A. and Wells, M.T., 2001, *Statistics in the 21st Century* (Boca Raton: Chapman & Hall/CRC).

[2] Laird, N.M. and Ware, J.H., 1982, Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.

[3] Robinson, G.K., 1991, That BLUP is a good thing: the estimation of random effects (with discussion). *Statistical Science*, **6**, 15–51.

[4] 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**, 559–578.

[5] Jiang, J., 1998, Consist estimators in generalized linear mixed models. *Journal of the American Statistical Association*, **93**, 720–729.

[6] McCulloch, C.E., 1994, Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association*, **89**, 330–335.

[7] Booth, J.G. and Hobert, J.P., 1999, Maximizing generalized linear model likelihood with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical*, *Society, Series B*, **61**, 265–285.

[8] Breslow, N.E. and Clayton, D.G., 1993, Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

[9] Lin, X. and Breslow, N.E., 1996, Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, **91**, 1007–1016.

[10] Raudenbush, S.W., Yang, M.L. and Yosef, M., 2000, Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics*, **9**, 141–157.

[11] Kuk, A.Y.C., 1995, Asymptotically unbiased estimation in generalized linear models with random effects. *Journal of the Royal Statistical Society, Series B*, **57**, 395–407.

[12] Breslow, N.E. and Lin, X., 1995, Bias correction in generalized linear mixed models with single component of dispersion. *Biometrika*, **82**, 81–91.

[13] Wei, G.C.G. and Tanner, M.A., 1990, A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *Journal of the American Statistical Association*, **85**, 699–704.

[14] McCulloch, C.E., 1997, Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, **92**, 162–170.

[15] McCullagh, P. and Nelder, J.A., 1989, *Generalized Linear Models* (2nd edn) (London: Chapman and Hall).

[16] McCulloch, C.E. and Searle, S.R., 2001, *Generalized, Linear, and Mixed Models* (New York: Wiley).

[17] Zeger, S.L. and Karim, M.R., 1991, Generalized linear models with random effects: A Gibbs sampling approach. *Journal of the American Statistical Association*, **86**, 79–86.

[18] Chib, S., 2000, Bayesian methods for correlated binary data. In: D.K. Dey, S.K. Ghosh and B.K. Mallick (Eds) *Generalized Linear Models: A Bayesian Perspective*, pp. 113–131 (New York: Marcel Dekker).

[19] Kass, R.E. and Wasserman, L., 1996, The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, **91**, 1343–1370.

[20] Natarajan, R. and McCulloch, C.E., 1995, A note on the existence of the posterior distribution for a class of mixed models for binomial responses. *Biometrika*, **82**, 639–643.

[21] Tan, M., Tian, G.L. and Ng, K.W., 2003, A non-iterative sampling method for computing posteriors in the structure of EM-type algorithms. *Statistica Sinica*, **13**, 625–639.

[22] Chib, S. and Greenberg, E., 1998, Analysis of multivariate probit models. *Biometrika*, **85**, 347–361.

[23] Robert, C.P., 1995, Simulation of truncated normal variables. *Statistics and Computing*, **5**, 121–125.

[24] McFadden, D.L., 1996, Lectures on simulation-assisted statistical inference. EC-squared Conference, Florence, Italy, 12–14 December, Available online at: http://elsa.berkeley.edu/mcfadden.

[25] Newton, M.A. and Raftery, A.E., 1994, Approximate Bayesian inference with the weighted likelihood bootstrap (with discussions). *Journal of the Royal Statistical Society, Series B*, **56**, 3–48.

[26] Gelfand, A.E. and Dey, D.K., 1994, Bayesian model choice: asymptotic and exact calculations. *Journal of the Royal Statistical Society, Series B*, **56**, 501–514.

[27] Meng, X.L. and Wong, W.H., 1996, Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, **6**, 831–860.

[28] Chen, M.H. and Shao, Q.M., 1997a, Estimating ratios of normalizing constants for densities with different dimensions. *Statistica Sinica*, **7**, 607–630.

[29] Chen, M.H. and Shao, Q.M., 1997b, On Monte Carlo methods for estimating ratios of normalizing constants. *The Annals of Statistics*, **25**, 1563–1594.

[30] Meng, X.L. and Schilling, S., 1996, Fitting full-information factor models and an empirical investigation of bridge sampling. *Journal of the American Statistical Association*, **91**, 1254–1267.

[31] Louis, T.A., 1982, Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **44**, 226–233.

[32] Tanner, M.A. and Wong, W.H., 1987, The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, **82**, 528–540.

[33] Jones, B. and Kenward, M.G., 1987, Modelling binary data from a three-period crossover trial. *Statistics in Medicine*, **6**, 555–564.