STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:1685–1698
Published online 11 October 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2318

Model comparison of generalized linear mixed models

Xin-Yuan Song and Sik-Yum Lee*,†

Department of Statistics, The Chinese University of Hong Kong, Hong Kong

SUMMARY

Generalized linear mixed models (GLMMs) have been widely appreciated in biological and medical research. Maximum likelihood estimation has received a great deal of attention. Comparatively, not much has been done on model comparison or hypotheses testing. In this article, we propose a path sampling procedure to compute the observed-data log-likelihood function, so that the Bayesian information criterion (BIC) can be applied to model comparison or hypothesis testing. Advantages of the proposed path sampling procedure are discussed. Two medical data sets are analysed for providing illustrative examples of the proposed methodology. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: latent random effects; maximum likelihood estimation; observed-data log-likelihood; path sampling; Bayesian information criterion

1. INTRODUCTION

Generalized linear mixed models (GLMMs) are natural extensions of the generalized linear models (GLMs) that allow for additional components of variability due to unobservable effects. Typically, the unobserved effects are modelled by the inclusion of random effects in the generalized linear model. This inclusion of random effects substantially increase the usefulness of such models and leads to wide applications to medical research; for example, for family-based analyses of genetic and environmental factors [1], for detecting incident clusters, and for producing shrinkage estimates in construction of maps of disease in small areas [2, 3], and may be more importantly for modelling correlated and discrete outcomes in longitudinal, repeated measures, or cluster designs [4–6]. As the random effects are modelled non-linearly, the observed-data likelihood function involves intractable integrals, even if the random effects are normally distributed. Hence, statistical analysis of GLMMs is not straightforward.

*Correspondence to: Sik-Yum Lee, Department of Statistics, Chinese University of Hong Kong, Shatin, N.T., Hong Kong.
†E-mail: sylee@sparc2.sta.cuhk.edu.hk

Contract/grant sponsor: HKSAR; contract/grant number: CUHK 4243/02H

Received August 2004
Accepted May 2005

In maximum likelihood (ML) estimation, much work has been focused on developing estimation procedures that can avoid the complicated integration. The penalized quasi-likelihood (PQL) approach, which basically uses Laplace approximations to the likelihood function, was proposed [6, 7]. Procedures that are based on the pseudo-likelihood (PL) and the restricted pseudo-likelihood (REPL) approaches have been developed [8]. Some standard software such as GenStat and SAS are capable to produce parameter estimates that are based on the above-mentioned approaches. Recently, a number of ML approaches that are based on the Markov chain Monte Carlo (MCMC) methods, see for example, the automated MCEM algorithm [9], and the stochastic approximation MCMC (SA-MCMC) algorithm [10], have been proposed. These MCMC approaches utilize the idea of data augmentation by augmenting the observed data with some hypothetical missing data, and work with the complete-data likelihood rather than the intractable observed-data likelihood. For analyses beyond ML estimation, inference on the fixed effect parameter is based on the conditional likelihood [11], while local influence measures have been developed on the basis of the conditional expectation of the complete-data log-likelihood [12]. As far as we know, not much attention has been devoted to the computation of the observed-data log-likelihood which is essential in evaluating the well-known statistic, namely the Bayesian information criterion (BIC), for model comparison (or hypothesis testing).

The main objective of this article is to propose a computational procedure to compute the observed-data log-likelihood of GLMMs. The procedure is developed on the basis of path sampling [13], which is a dependable tool for computing ratio of normalizing constants of probability models. Path sampling has been shown to have the following nice features [13]: (i) It is a generalization of the importance sampling, and bridge sampling [14], hence can produce more accurate results. (ii) It computes the logarithm scale of the ratio which is generally more stable. (iii) Its implementation is simple. Recently, path sampling has been applied to compute the observed-data log-likelihood and the Bayes factor for a number of complicated latent variable models, see References [15, 16], among others.

The paper is organized as follows: In Section 2, we introduce the notation and define the GLMMs. A procedure for computing the observed-data log-likelihood of GLMMs is developed in Section 3. Novel applications of the newly developed methodology for analysing two medical data sets are presented in Section 4. In particular, we compare the GLMs and the GLMMs in fitting the data. A discussion is given in Section 5, and some technical details are presented in the Appendices.

2. MODEL COMPARISON OF GLMMs

2.1. The model and the likelihood functions

Consider a data set that is composed of a response y_{ij}, vectors of covariates $\mathbf{x}_{ij}(p_1 \times 1)$ and $\mathbf{z}_{ij}(p_2 \times 1)$, where $j = 1, \ldots, n_i$ represents an observation within the cluster $i = 1, \ldots, N$. For example, a subject (e.g. a patient) can be considered as a cluster, and repeated measurements (e.g. treatments) for this subject i can be obtained at a total of n_i different time points that are represented by $j = 1, \ldots, n_i$. It is assumed that conditional on a vector $\mathbf{b}_i$ $(p_2 \times 1)$ of unobservable random variables, y_{ij} follows an exponential family distribution of the following

form [17]:

$$p(y_{ij}|\mathbf{b}_i) = \exp[\phi\{y_{ij}\theta_{ij} - a(\theta_{ij})\} + c(y_{ij}, \phi)] \tag{1}$$

The conditional mean and variance of y_{ij} given $\mathbf{b}_i$ are $E(y_{ij}|\mathbf{b}_i) = \mu_{ij} = \dot{a}(\theta_{ij})$ and $\text{var}(y_{ij}|\mathbf{b}_i) = \ddot{a}(\theta_{ij})/\phi$, respectively, where $\dot{a}(u) = \mathrm{d}a/\mathrm{d}u$ and $\ddot{a}(u) = \mathrm{d}^2a/\mathrm{d}u^2$. The GLMM is defined by (1) and the following systematic component:

$$\begin{aligned} g(\mu_{ij}) &= \eta_{ij} = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_{ij}^{\mathrm{T}}\mathbf{b}_i \quad \text{or} \\ \theta_{ij} &= k(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_{ij}^{\mathrm{T}}\mathbf{b}_i) = \dot{a}^{-1}[g^{-1}(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_{ij}^{\mathrm{T}}\mathbf{b}_i)] \end{aligned} \tag{2}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{p_1})^{\mathrm{T}}$ is a vector that consists of the regression coefficients, $k(\cdot) = \dot{a}^{-1}(g^{-1}(\cdot))$ is a composite function of $\dot{a}^{-1}$ and g^{-1}, $g(\cdot)$ is a known continuously differentiable function, and $\dot{a}^{-1}(\cdot)$ and $g^{-1}(\cdot)$ are inverse functions of $\dot{a}(\cdot)$ and $g(\cdot)$, respectively. The distribution of $\mathbf{b}_i$ is assumed to be normal, $\mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\gamma})$ depends on a $p_3 \times 1$ vector $\boldsymbol{\gamma}$ of unknown parameters that determine the variance components.

Let $\boldsymbol{\psi} = (\phi, \boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$ be the $p \times 1$ ($p = 1 + p_1 + p_3$) vector of unknown parameters, and $\mathbf{Y}_{\mathrm{o}}$ represent all the observations on y_{ij}. The observed-data log-likelihood function for $\boldsymbol{\psi}$ has the form

$$L_0(\mathbf{Y}_{\mathrm{o}}|\boldsymbol{\psi}) \propto \sum_{i=1}^{n} \log \left[\int \prod_{j=1}^{n_i} p(y_{ij}|\mathbf{b}_i)|\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}\mathbf{b}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{b}_i\right) \mathrm{d}\mathbf{b}_i \right] \tag{3}$$

The integral involved in $L_0(\mathbf{Y}_{\mathrm{o}}|\boldsymbol{\psi})$ usually does not have an analytic solution and its dimension is p_2, that is the dimension of the random effect $\mathbf{b}_i$. As a result, the function is usually intractable, and obtaining the ML estimate or conducting statistical inference directly from this function is difficult. Motivated by the key idea of data augmentation and the EM algorithm, this difficulty is partially alleviated by treating $\mathbf{Y}_{\mathrm{m}} = \{\mathbf{b}_i, i = 1, \ldots, N\}$ as hypothetical missing data that are missing at random (MAR) with an ignorable missing mechanism, and augmenting the observed-data set $\mathbf{Y}_{\mathrm{o}}$ with $\mathbf{Y}_{\mathrm{m}}$. The problem is then formulated as a missing data problem, with the complete-data set $\mathbf{Y}_{\mathrm{c}} = (\mathbf{Y}_{\mathrm{o}}, \mathbf{Y}_{\mathrm{m}})$. Let complete-data likelihood be $p(\mathbf{Y}_{\mathrm{c}}|\boldsymbol{\psi})$, then the complete-data log-likelihood is equal to

$$L_{\mathrm{c}}(\mathbf{Y}_{\mathrm{c}}|\boldsymbol{\psi}) = \sum_{i=1}^{N} \left(\sum_{j=1}^{n_i} [\phi\{y_{ij}\theta_{ij} - a(\theta_{ij})\} + c(y_{ij}, \phi)] - \frac{1}{2}\mathbf{b}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{b}_i - \frac{1}{2}\log|\boldsymbol{\Sigma}| \right) \tag{4}$$

Comparing to $L_0(\mathbf{Y}_{\mathrm{o}}|\boldsymbol{\psi})$, this function is relatively simple. Methods for obtaining ML estimates of the unknown parameters in the model has been developed on the basis of this complete-data log-likelihood. For example, an automated MCEM algorithm have been developed on the basis of an EM formulation [9]. Another attractive algorithm, namely the SA-MCMC algorithm that is applied by Zhu and Lee [10] to GLMMs, is based on the principle of stochastic approximation (SA) and Markov chain Monte Carlo (MCMC) methods. Both algorithms require the simulation of a sufficiently large sample of observations of $\mathbf{Y}_{\mathrm{m}}$ from $p(\mathbf{Y}_{\mathrm{m}}|\mathbf{Y}_{\mathrm{o}}, \boldsymbol{\psi})$, the conditional distribution given the observed data $\mathbf{Y}_{\mathrm{o}}$ and the current parameter value $\boldsymbol{\psi}$.

Because $\mathbf{b}_i$ are independent, $p(\mathbf{Y}_m|\mathbf{Y}_o,\boldsymbol{\psi})$ is a product of $p(\mathbf{b}_i|\mathbf{Y}_o,\boldsymbol{\psi})$ that is proportional to

$$\exp\left\{-\frac{1}{2}\mathbf{b}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{b}_i+\sum_{j=1}^{n_i}\phi\left[y_{ij}k(\mathbf{x}_{ij}^T\boldsymbol{\beta}+\mathbf{z}_{ij}^T\mathbf{b}_i)-a\{k(\mathbf{x}_{ij}^T\boldsymbol{\beta}+\mathbf{z}_{ij}^T\mathbf{b}_i)\}\right]\right\} \tag{5}$$

The target density $p(\mathbf{b}_i|\mathbf{Y}_o,\boldsymbol{\psi})$ is non-standard, the Metropolis–Hasting algorithm was used to simulate $\mathbf{b}_i$ from it. As details about the simulation of $\mathbf{b}_i$ can be obtained in Reference [10], they are not repeated here. Similar to one step Newton–Raphson method, the SA-MCMC algorithm applies the idea of stochastic approximation to update $\boldsymbol{\psi}$ by using inverse of the estimated observed-data information matrix, which involves matrices of the first and the second partial derivatives of $L_c(\mathbf{Y}_c|\boldsymbol{\psi})$ with respect to $\boldsymbol{\psi}$. It can be seen later that the simulated sample $\{\mathbf{Y}_m^{(j)}|j=1,\ldots,J\}$ is also important in computing the observed-data likelihood. As our focus is not in ML estimation, we do not intend to discuss further on the automated MCEM algorithm or the SA-MCMC algorithm, and assume that the ML estimated $\hat{\boldsymbol{\psi}}$ has been obtained. In our numerical illustrations, $\hat{\boldsymbol{\psi}}$ was obtained by the SA-MCMC algorithm.

2.2. Model comparison

By working with $L_c(\mathbf{Y}_c|\boldsymbol{\psi})$, the complicated observed-data log-likelihood function, $L_0(\mathbf{Y}_o|\boldsymbol{\psi})$, can be avoided in the ML estimation. However, to conduct statistical inference after estimation, it is desirable to obtain the values of the observed-data log-likelihood, evaluated at the ML estimates under different models of interest. In this article, we focus on the important issue of model comparison between two completing GLMMs, which could be nested or non-nested. Suppose that the observed data $\mathbf{Y}_o$ have arisen under one of the two competing GLMMs M_1 and M_2. For $h=1,2$, let $p(\mathbf{Y}_o|\hat{\boldsymbol{\psi}}_h,M_h)$ be the observed-data likelihood function corresponding to M_h, evaluated at the ML estimate, $\hat{\boldsymbol{\psi}}_h$, of the unknown parameter vector under M_h. A well-known statistic for comparing M_1 and M_2 is the following Bayesian information criterion (BIC) (see Reference [18]):

$$\mathrm{BIC}_{12}=-2[\log p(\mathbf{Y}_o|\hat{\boldsymbol{\psi}}_1,M_1)-\log p(\mathbf{Y}_o|\hat{\boldsymbol{\psi}}_2,M_2)]+(d_1-d_2)\log N \tag{6}$$

where d_h is the dimension of $\boldsymbol{\psi}_h$. For comparing non-nested models, M_1 should be chosen if $\mathrm{BIC}_{12}<0$; otherwise M_2 should be chosen. The use of BIC_{12} for situations where M_1 is nested in M_2 is based on the following well-recognized criterion [18]:

BIC_{12}	$0<$	0–2	2–6	6–10
	Support M_1	Barely support M_2	Support M_2	Strongly support M_2

By associating the null and alternative hypothesis with M_1 and M_2, respectively, BIC_{12} can be used for hypothesis testing. As pointed out in the literature [18], the BIC approach has at least the following advantages over the significance test that is based on the p-values: (i) tests on the basis of p-values tend to reject the null hypothesis (or choose the null model) too frequently for data sets with large sample sizes; (ii) the p-value is only a measure of evidence against the null model, not a means of supporting the alternative model; hence, no definite conclusion can be drawn if the null hypothesis is not rejected; and (iii) the significance test cannot be applied to test non-nested hypothesis (or to compare non-nested models).

In applying the BIC for comparing GLMMs, we require to compute the complicated observed-data log-likelihoods $\log p(\mathbf{Y}_o|\hat{\boldsymbol{\psi}}_h, M_h)$, $h=1,2$. The well-known importance sampling may be considered as a candidate. However, this method may not be accurate when applying to situations with widely separated competing models [14]. To improve the performance, bridge sampling [14], which utilizes a 'bridge model' between the competing models, has been proposed. Path sampling [13] generalizes bridge sampling by incorporating an arbitrarily large number of bridge models between the competing models in the computation. Hence, we expect that path sampling would achieve better accuracy than importance sampling and bridge sampling. A path sampling procedure for computing $p(\mathbf{Y}_o|\hat{\boldsymbol{\psi}}_h, M_h)$, then the BIC_{12}, is presented in the next section.

3. COMPUTATION OF THE OBSERVED-DATA LIKELIHOOD VIA PATH SAMPLING

Path sampling [13] was originally developed for evaluating (log) ratios of normalizing constants of probability models. Without lost of generality, it is applied here to compute $\log p(\mathbf{Y}_o|\hat{\boldsymbol{\psi}}_1, M_1)$, at the ML estimate $\hat{\boldsymbol{\psi}}_1$. Direct application of this technique to compute the observed-data log-likelihood is difficult. Hence, similar to ML estimation, we utilize the idea of data augmentation by considering the complete-data log-likelihood $L_c(\mathbf{Y}_c|\hat{\boldsymbol{\psi}}_1) = \log p(\mathbf{Y}_o, \mathbf{Y}_m|\hat{\boldsymbol{\psi}}_1, M_1)$, see (4). The key idea of path sampling is coming from the equality $p(\mathbf{Y}_m|\mathbf{Y}_o, \hat{\boldsymbol{\psi}}_1, M_1) = p(\mathbf{Y}_o, \mathbf{Y}_m|\hat{\boldsymbol{\psi}}_1, M_1)/p(\mathbf{Y}_o|\hat{\boldsymbol{\psi}}_1, M_1)$, where the observed-data likelihood $p(\mathbf{Y}_o|\hat{\boldsymbol{\psi}}_1, M_1)$ is regarded as a normalizing constant of $p(\mathbf{Y}_m|\mathbf{Y}_o, \hat{\boldsymbol{\psi}}_1, M_1)$. To facilitate the computation, we make use of an auxiliary submodel of M_1, namely M_0 with a parameter vector $\boldsymbol{\psi}_0$, such that (i) it is simple so that $p(\mathbf{Y}_o|\hat{\boldsymbol{\psi}}_0, M_0)$ can be easily evaluated, where $\hat{\boldsymbol{\psi}}_0$ is a subvector of $\hat{\boldsymbol{\psi}}_1$ (the ML estimate of $\boldsymbol{\psi}_1$ under M_1), and it is not the true ML estimate of $\boldsymbol{\psi}_0$ under M_0; and (ii) it can be linked with M_1 via a linking GLMM M_t with a continuous path t in [0,1]. Let M_t be the linking model that is defined by (1) with $g(\mu_{ij}(t)) = \eta_{ij}(t)$, where $\eta_{ij}(t)$ is a differentiable function with respect to t. A trivial example of M_0, and linking models M_t will be given later. Inspired by Lee and Song [15], we apply the key idea of Gelman and Meng [13] to consider the following class of densities with t in [0,1]:

$$p(\mathbf{Y}_m|\mathbf{Y}_o, \hat{\boldsymbol{\psi}}_1, M_t) = \frac{1}{z(t)} p(\mathbf{Y}_o, \mathbf{Y}_m|\hat{\boldsymbol{\psi}}_1, M_t) \tag{7}$$

where

$$z(t) = p(\mathbf{Y}_o|\hat{\boldsymbol{\psi}}_1, M_t) = \int p(\mathbf{Y}_o, \mathbf{Y}_m|\hat{\boldsymbol{\psi}}_1, M_t)\, \mathrm{d}\mathbf{Y}_m \tag{8}$$

As t is a continuous path in [0,1] to link M_0 and M_1, it follows that $z(0) = p(\mathbf{Y}_o|\hat{\boldsymbol{\psi}}_0, M_0)$ and $z(1) = p(\mathbf{Y}_o|\hat{\boldsymbol{\psi}}_1, M_1)$. The major problem is to compute $\lambda_{10} = \log[z(1)/z(0)]$. Let

$$U(\mathbf{Y}_o, \mathbf{Y}_m; t) = \frac{\mathrm{d}}{\mathrm{d}t} \log p(\mathbf{Y}_o, \mathbf{Y}_m|\hat{\boldsymbol{\psi}}_1, M_t) \tag{9}$$

Following the reasoning in Gelman and Meng [13], it can be shown that

$$\lambda_{10} = \int_0^1 E[U(\mathbf{Y}_o, \mathbf{Y}_m|t)]\,dt \tag{10}$$

where the expectation is taken with respect to the conditional distribution of $\mathbf{Y}_m$ given $\mathbf{Y}_o$ at $\hat{\boldsymbol{\psi}}_1$. The integral in (11) is approximated via the trapezoid rule. More specifically, let $0 = t_0 < t_{(1)} < \cdots < t_{(S)} < t_{(S+1)} = 1$ be fixed grids in [0,1], we approximate λ_{10} by

$$\hat{\lambda}_{10} = \frac{1}{2}\sum_{s=0}^{S}(t_{(s+1)} - t_{(s)})(\bar{U}_{(s+1)} + \bar{U}_{(s)}) \tag{11}$$

where

$$\bar{U}_{(s)} = J^{-1}\sum_{j=1}^{J} U(\mathbf{Y}_o, \mathbf{Y}_m^{(j)}, t_{(s)}) \tag{12}$$

in which $\{\mathbf{Y}_m^{(j)}, j = 1,\ldots,J\}$ are observations simulated from the conditional distribution of $\mathbf{Y}_m$ given $\mathbf{Y}_o$ at $\hat{\boldsymbol{\psi}}_1$. From the definition of λ_{10}, we have

$$\log p(\mathbf{Y}_o|\hat{\boldsymbol{\psi}}_1, M_1) = \log[z(1)] \doteq \hat{\lambda}_{10} + \log p(\mathbf{Y}_o|\hat{\boldsymbol{\psi}}_0, M_0) \tag{13}$$

As $p(\mathbf{Y}_o|\hat{\boldsymbol{\psi}}_0, M_0)$ can be easily computed, we obtain $\log p(\mathbf{Y}_o|\hat{\boldsymbol{\psi}}_1, M_1)$.

Because the program for generating $\{\mathbf{Y}_m^{(j)}, j = 1,\ldots,J\}$ has usually been constructed in the ML estimation, the additional programming of the path sampling procedure is minor and straight-forward. The logarithm scale of $p(\mathbf{Y}_o|\hat{\boldsymbol{\psi}}_1, M_1)$ is computed, which is generally more stable. Although M_2 can be very different from M_1, the corresponding observed-data log-likelihood $\log p(\mathbf{Y}_m|\hat{\boldsymbol{\psi}}_2, M_2)$ can be separately, but similarly computed as above. The BIC_{12} for competing M_1 and M_2 can be computed via (6). The choices of S and J in the procedure depend on the size and the complexity of the model. A choice of $S = 20$ is usually sufficient for most practical problems; but J should not be less than 1000.

To give an example to illustrate the application of the path sampling procedure, we consider a model M_1, which can be a general GLMM that is defined by (1) with an arbitrary but fixed link function $\eta_{ij} = g_1(\mu_{ij})$ that is defined by (2). Let the unknown parameter $\boldsymbol{\psi}_1 = (\phi_1, \boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\gamma}_1^{\mathrm{T}})^{\mathrm{T}}$. An auxiliary submodel M_0 can be chosen as a GLM without random effects in which $g_1(\mu_{ij}) = \eta_{ij} = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}$. As $\mathbf{b}_i$ is not involved in M_0, for (i,j) and (i,j'), $j \neq j'$, y_{ij} is independent of $y_{ij'}$. Hence, the observed-data likelihood under M_0 can be directly computed without difficulty. The linked GLMM M_t is defined via a continuous path t in [0,1]

$$M_t: \quad g_1(\mu_{ij}(t)) = \eta_{ij}(t) = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_1 + t\mathbf{z}_{ij}^{\mathrm{T}}\mathbf{b}_i$$

When $t = 1$, $\eta_{ij}(1) = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_1 + \mathbf{z}_{ij}^{\mathrm{T}}\mathbf{b}_i$, M_t reduces to M_1, and $\hat{\boldsymbol{\psi}}_1 = (\hat{\phi}_1, \hat{\boldsymbol{\beta}}_1^{\mathrm{T}}, \hat{\boldsymbol{\gamma}}_1^{\mathrm{T}})^{\mathrm{T}}$; and when $t = 0$, $\eta_{ij}(0) = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_1$, M_t reduces to M_0, and $\hat{\boldsymbol{\psi}}_0 = (\hat{\phi}_1, \hat{\boldsymbol{\beta}}_1^{\mathrm{T}})^{\mathrm{T}}$. The function $U(\mathbf{Y}_o, \mathbf{Y}_m; t)$ can be obtained by differentiating the complete-data log-likelihood with respect to t. In this example

$$U(\mathbf{Y}_o, \mathbf{Y}_m; t) = \sum_{i=1}^{N}\sum_{j=1}^{n_i}\hat{\phi}_1\{y_{ij} - \dot{a}(k_1(\mathbf{x}_{ij}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_1 + t\mathbf{z}_{ij}^{\mathrm{T}}\mathbf{b}_i))\}\dot{k}_1(\mathbf{x}_{ij}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_1 + t\mathbf{z}_{ij}^{\mathrm{T}}\mathbf{b}_i)\mathbf{z}_{ij}^{\mathrm{T}}\mathbf{b}_i \tag{14}$$

where $k_1(\cdot) = \dot{a}^{-1}(g_1^{-1}(\cdot))$. Given a sufficiently large sample, $\{\mathbf{Y}_{\mathrm{m}}^{(j)}, j = 1, \ldots, J\}$, which is simulated from the conditional distribution of $\mathbf{Y}_{\mathrm{m}}$ given $\mathbf{Y}_{\mathrm{o}}$ at $\hat{\boldsymbol{\psi}}_1$, $\hat{\lambda}_{10}$ can be obtained via (11), (12), and (14); and $\log\, p(\mathbf{Y}_{\mathrm{o}}|\hat{\boldsymbol{\psi}}_1, M_1)$ can be obtained via (13). Note that because $\hat{\boldsymbol{\psi}}_0$ is not the true ML estimate of $\boldsymbol{\psi}_0$ in M_0, $\hat{\lambda}_{10} = \log p(\mathbf{Y}_{\mathrm{o}}|\hat{\boldsymbol{\psi}}_1, M_1) - \log p(\mathbf{Y}_{\mathrm{o}}|\hat{\boldsymbol{\psi}}_0, M_0)$ cannot be directly used to rigorously compare the GLMM M_1 with the GLM M_0; see the numerical illustrations for more details.

The path sampling procedure can also be applied to compute the observed-data log-likelihood corresponding to a model M_2, which could be a GLMM with the same link function but with different parameter vector, or a GLMM with different link function $\eta_{ij} = g_2(\mu_{ij})$, by exactly the same manner as above. Eventually, BIC_{12} can be computed from (6) for comparing M_1 and M_2. Hence, a more appropriate GLMM can be selected to fit the data better. Note that as GLMMs with different link functions are non-nested, they cannot be compared by the classical significance test on the basis of p-values.

4. ILLUSTRATIVE EXAMPLES

In this section, we present novel analyses of two well-known data sets in medical research, by applying the proposed path sampling procedure to compute the BIC for comparing various nested and/or non-nested competing GLMMs. In all examples, we assume that the hypothetical missing data are MAR, and take $S = 20$ in (11) for computing the observed-data log-likelihood.

4.1. Six Cities data

The first example is based on a well-known data set from the Six Cities study, a longitudinal study of the health effects of air pollution. This data set has been analysed via a multivariate logit model [19], a multivariate probit model [20], and a GLMM with more complicated random effect distribution [21]. Whether the child had a respiratory infection in the year prior to each examination was reported by the mother; also available was the mother's baseline smoking status as ascertained at the first interview. The data that were presented in Table I contain repeated binary measures y_{ij} of the wheezing status ($1 =$ yes, $0 =$ no) of 537 children from Stuebenvile, OH, at ages $7, 8, 9$, and 10 years. We first consider formulations of model (1) for the binary response y_{ij} ($1 =$ infection, $0 =$ no infection) for child i at age a_{ij} with different link functions, and the following linear predictor:

$$\eta_{ij} = \beta_0 + \beta_1 s_i + \beta_2 a_{ij} + \beta_3 d_{ij} + u_i + v_{ij} \tag{15}$$

where s_i is a binary indicator that represents the mother's smoking habit ($1 =$ yes, $0 =$ no); a_{ij} is the age of the child, centred at 9 years; d_{ij} is an interaction between smoking habit and age; u_i and v_{ij} are independent random effects that are, respectively, distributed as $\mathrm{N}[0, \sigma_u^2]$ and $\mathrm{N}[0, \sigma_{vj}^2]$. It is assumed that y_{ij} are conditionally independent and that $y_{ij}|u_i, v_{ij} \sim$ Bernoulli(p_{ij}). The Bernoulli density is given by

$$p(y_{ij}|\mathbf{b}_i) = \exp\left\{ y_{ij} \log \frac{p_{ij}}{1 - p_{ij}} + \log(1 - p_{ij}) \right\} \tag{16}$$

and can be formulated from model (1) by taking $\phi = 1$, $\theta_{ij} = \log[p_{ij}/(1 - p_{ij})]$, $a(\theta_{ij}) = -\log(1 - p_{ij}) = \log[1 + \exp(\theta_{ij})]$, and $c(y_{ij}, \phi) = 0$.

Table I. Six Cities data set: child's wheeze frequency.

No maternal smoking					Maternal smoking				
Age of child					Age of child				
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
1	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

From Reference [20].

Three models with different link functions are

M_1: Equation (15), and logit link $g_1(\mu_{ij}) = \eta_{ij} = \log\{p_{ij}/(1 - p_{ij})\}$.
M_2: Equation (15), and probit link $g_2(\mu_{ij}) = \eta_{ij} = \Phi^{-1}(p_{ij})$, where $\Phi(\cdot)$ is the distribution function of N[0,1].
M_3: Equation (15), and complementary log–log link $g_3(\mu_{ij}) = \eta_{ij} = \log\{-\log(1 - p_{ij})\}$.

In the logit link, $\theta_{ij} = k(\eta_{ij}) = \eta_{ij}$, which is an identity link. In the probit link, $\theta_{ij} = k(\eta_{ij}) = \log[\Phi(\eta_{ij})/(1 - \Phi(\eta_{ij}))]$. In the complementary log–log link, $\theta_{ij} = k(\eta_{ij}) = \log[1 - \exp(-\exp(\eta_{ij})] + \exp(\eta_{ij})$. These models are special cases of the GLMM that is defined in (2) with $\mathbf{b}_i = (u_i, v_{i1}, v_{i2}, v_{i3}, v_{i4})^{\mathrm{T}}$, $\mathbf{z}_{i1} = (1,1,0,0,0)^{\mathrm{T}}$, $\mathbf{z}_{i2} = (1,0,1,0,0)^{\mathrm{T}}$, $\mathbf{z}_{i3} = (1,0,0,1,0)^{\mathrm{T}}$, and $\mathbf{z}_{i4} = (1,0,0,0,1)^{\mathrm{T}}$, where the covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\gamma})$ of $\mathbf{b}_i$ is a diagonal matrix with diagonal elements σ_u^2 and σ_{vj}^2, and $\boldsymbol{\gamma} = (\sigma_u^2, \sigma_{v1}^2, \sigma_{v2}^2, \sigma_{v3}^2, \sigma_{v4}^2)^{\mathrm{T}}$. We used the SA-MCMC algorithm to obtain the ML estimates. The first and the second derivatives of $L_{\mathrm{c}}(\mathbf{Y}_{\mathrm{c}}|\boldsymbol{\psi})$ with respect to parameter vector $\boldsymbol{\psi}$ that are involved in the SA-MCMC algorithm can be obtained via similar derivation as in Reference [10]. The ML estimates of the parameters and their associated standard errors estimates are presented in Table II. In applying the path sampling procedure, we took $J = 5000$ after discarding 1000 burn-in iterations for computing $\bar{U}_{(s)}$, see (11) and (12). The observed-data log-likelihood for M_1, M_2, and M_3 that are obtained from the path sampling procedure are equal to -795.97, -799.23, and -798.49, respectively. Hence, $\mathrm{BIC}_{21} = 6.52$, and $\mathrm{BIC}_{31} = 5.04$. According the interpretation of BIC as given in Section 2.2, M_1 with the logit link is selected.

Table II. ML results corresponding to the Six Cities study.

Parameter	ML Estimates and Standard Error Estimates (in parentheses)						
	M_1	M_2	M_3	M_4	M_5	M_6	M_7
β_0	−3.182	−2.094	−3.069	−1.901	−3.143	−2.582	−2.567
	(0.092)	(0.054)	(0.073)	(0.089)	(0.086)	(0.075)	(0.070)
β_1	0.470	0.287	0.378	0.314	0.397	0.339	—
	(0.145)	(0.086)	(0.114)	(0.139)	(0.124)	(0.121)	—
β_2	−0.220	−0.123	−0.164	−0.141	−0.170	—	−0.145
	(0.085)	(0.048)	(0.065)	(0.070)	(0.066)	—	(0.062)
β_3	0.107	0.067	0.084	0.071	—	—	—
	(0.136)	(0.078)	(0.103)	(0.111)	—	—	—
σ_u^2	4.378	1.925	2.956	—	4.272	2.395	2.522
	(0.196)	(0.085)	(0.134)	—	(0.191)	(0.109)	(0.116)
σ_{v1}^2	0.492	0.618	0.553	—	0.612	0.462	0.344
	(0.020)	(0.026)	(0.023)	—	(0.026)	(0.020)	(0.015)
σ_{v2}^2	0.632	0.647	0.457	—	0.539	0.165	0.155
	(0.026)	(0.027)	(0.019)	—	(0.022)	(0.007)	(0.006)
σ_{v3}^2	0.612	0.564	0.504	—	0.562	0.249	0.252
	(0.025)	(0.023)	(0.020)	—	(0.023)	(0.011)	(0.011)
σ_{v4}^2	0.439	0.369	0.288	—	0.389	0.111	0.146
	(0.018)	(0.015)	(0.012)	—	(0.016)	(0.005)	(0.007)
Observed-data log-likelihood	−795.97	−799.23	−798.49	−902.76	−796.08	−806.70	−804.16

An interesting issue is to see whether a GLM is adequate for fitting the data. To address this issue, we compare the M_1 with the following GLM with the logit link and linear predictor

$$M_4\text{: } \eta_{ij} = \beta_0 + \beta_1 s_i + \beta_2 a_{ij} + \beta_3 d_{ij}.$$

In M_4, the random effects do not exist. The ML estimates $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ of the parameters in M_4 are presented in Table II, together with the standard error estimates. The observed-data log-likelihood is equal to −902.76; it follows from (6) that $\mathrm{BIC}_{41} = 182.16$. Hence, the analysis gives a clear indication for selecting M_1, the GLMM with random effects.

Now, the significance of regression coefficients in $\boldsymbol{\beta}$ is assessed by comparing M_1 with the following GLMMs:

M_5: GLMM with logit link and $\eta_{ij} = \beta_0 + \beta_1 s_i + \beta_2 a_{ij} + u_i + v_{ij}$,
M_6: GLMM with logit link and $\eta_{ij} = \beta_0 + \beta_1 s_i + u_i + v_{ij}$,
M_7: GLMM with logit link and $\eta_{ij} = \beta_0 + \beta_2 a_{ij} + u_i + v_{ij}$,

The ML estimates and standard error estimates under M_5, M_6, and M_7 are also reported in Table II. The observed-data log-likelihood for M_4, M_5, and M_6, which are obtained via the path sampling procedure with $J = 10\,000$ after discarding 1000 burn-in iterations, are equal to −796.08, −806.70, and −804.16, respectively. From (6), we get $\mathrm{BIC}_{51} = -6.07$, $\mathrm{BIC}_{65} = 9.89$, $\mathrm{BIC}_{75} = 14.97$. According to the interpretation given in Section 2.2, M_5 is selected. That is, both effects of s_i and a_{ij} in the proposed GLMM are significant, whilst their interaction term d_{ij} is not. Hence, the wheeze status of the child is influenced by his/her age and mother's smoking habit additively but not interactively. Note that the BIC values can be used for testing

hypotheses concerning β_0, β_1, β_2 and/or β_3. Due to the justifications as given in Section 2.2, we prefer the model comparison approach.

As far as we know, the model comparison analyses that are presented in this example have not been reported in the literature.

4.2. Lung cancer data

The second example is to analyse the data set on 14 retrospective studies of the association between smoking and lung cancer [9, 22, 23]. In the ith study, the data set that is presented in Table III gave (i) the number of people with lung cancer y_{i1} within n_{i1} smokers, and (ii) the number of people with lung cancer y_{i2} within n_{i2} non-smokers. It is assumed that the responses y_{ij} are conditionally independent given $\mathbf{b}_i$. The distribution of $y_{ij}|\mathbf{b}_i$ is assumed to be Binomial B(n_{ij}, p_{ij}), with probability density

$$p(y_{ij}|\mathbf{b}_i) = \exp\left\{ y_{ij} \log \frac{p_{ij}}{1-p_{ij}} + n_{ij} \log(1-p_{ij}) \right\} \tag{17}$$

This situation can be formulated via equation (1) by taking $\phi = 1$, $\theta_{ij} = \log[p_{ij}/(1-p_{ij})]$, $a(\theta_{ij}) = -n_{ij}\log[1+\exp(\theta_{ij})]$, and $c(y_{ij}, \phi) = 0$. The systematic component of the GLMM for analysing the data set is given by

$$\eta_{ij} = \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 x_{ij} + \mathbf{z}_{ij}^{\mathrm{T}}\mathbf{b}_i \tag{18}$$

where x_{ij} indicates a smoker ($x_{ij} = 1$ for smokers and $x_{ij} = 0$ for non-smokers), β_0 is the intercept, and β_1 is the regression parameter. As an illustration of the procedure, we consider three models that are defined with different types of random effects. In M_1, $\mathbf{z}_{ij} = 1$, $\mathbf{b}_i = u_i$ is distributed as $\mathrm{N}[0, \sigma_u^2]$, and $\boldsymbol{\gamma} = \sigma_u^2$. In M_2, $\mathbf{z}_{i1} = (1,1,0)$, $\mathbf{z}_{i2} = (1,0,1)$, and $\mathbf{b}_i = (u_i, v_{i1}, v_{i2})^{\mathrm{T}}$ is distributed as $\mathrm{N}[\mathbf{0}, \boldsymbol{\Sigma}]$, where $\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\gamma})$ with $\boldsymbol{\gamma} = (\sigma_u^2, \sigma_{v1}^2, \sigma_{v2}^2)$. In M_3, we consider the

Table III. Lung cancer data.

	Smoker		Non-smoker	
i	y_{i1}	n_{i1}	y_{i2}	n_{i2}
1	83	155	3	17
2	90	317	3	46
3	129	210	7	26
4	70	467	12	137
5	412	711	32	163
6	597	1263	8	122
7	88	262	5	12
8	1350	2646	7	68
9	60	166	3	30
10	459	993	18	99
11	724	970	4	58
12	499	961	19	75
13	451	2180	39	675
14	260	519	5	33

From Reference [9].

Table IV. ML results corresponding to the Lung cancer study.

	ML Estimates and Standard Error Estimates (in parentheses)				
Parameter	M_1	M_2	M_3	M_3(MCEM*)	M_4
β_0	−1.899	−1.931	−1.932	−1.932	−2.139
	(0.083)	(0.066)	(0.063)	—	(0.082)
β_1	1.688	1.695	1.691	1.695	1.922
	(0.086)	(0.068)	(0.065)	—	(0.084)
σ_u^2	0.336	0.195	0.186	0.189	—
	(0.070)	(0.052)	(0.053)	—	—
σ_{v1}^2	—	0.210	0.235	0.232	—
	—	(0.046)	(0.058)	—	—
σ_{v2}^2	—	0.030	—	—	—
	—	(0.008)	—	—	—
Observed-data log-likelihood	−8067.5	−8059.6	−8060.1	—	−8274.02

From Reference [9].

same trivariate normal distribution for $\mathbf{b}_i$, but $\sigma_{v1} = \sigma_{v2} = \sigma_v$, so that $\boldsymbol{\gamma} = (\sigma_u^2, \sigma_v^2, \sigma_v^2)$. By using SA-MCMC algorithm, the ML estimates $(\hat{\beta}_1, \hat{\beta}_2, \hat{\boldsymbol{\gamma}})$ in M_1, M_2, and M_3 are obtained. For completeness, these estimates and their associated standard error estimates are presented in Table IV. Moreover, for model M_3, ML estimates that are obtained by Booth and Hobert [9] with an automated MCEM are also presented, to cross-validate the accuracy of the estimates that are obtained by these two algorithm. The corresponding observed-data log-likelihoods, which are obtained with $J = 10\,000$ after discarding 1000 burn-in iterations, are equal to −8067.49, −8059.59, and −8060.12. As a result, $\mathrm{BIC}_{12} = 10.53$, and $\mathrm{BIC}_{23} = 1.58$. Hence, M_3 is selected.

To see whether the GLMM is better than the GLM in fitting the data, we compare M_3 with following GLM without random effects:

$$M_4\colon\ \eta_{ij} = \beta_0 + \beta_1 x_{ij}.$$

The ML estimates and standard error estimates of β_0 and β_1 are also reported in Table IV. The observed-data log-likelihood of M_4 at its ML estimates is equal to −8274.02. It follows from (6) that $\mathrm{BIC}_{43} = 422.52$. This gives a definite conclusion that the GLMM, M_3, should be selected.

5. DISCUSSION

Recently, GLMMs have been extensively applied to medical research. In this article, we propose a path sampling procedure to compute the complicated observed-data log-likelihood. Given a program that is usually available in ML estimation for simulating observations from the appropriate conditional expectation, the additional programming effort is minor. Hence, the proposed procedure can be easily applied to compute the BIC for model comparison of various forms of GLMMs.

The observed-data log-likelihood is computed at a given ML estimate of the parameter vector that is obtained via the SA-MCMC algorithm [10]. The SA-MCMC algorithm apply the well-known data augmentation strategy [24, 25], which treats $\mathbf{Y}_{\mathrm{m}}$ as hypothetical missing data that are MAR with an ignorable missing mechanism [26], augments the observed data $\mathbf{Y}_{\mathrm{o}}$ with $\mathbf{Y}_{\mathrm{m}}$, and then solves the problem as a missing data problem. It has been shown via a simulation study that this strategy produces accuracy results for estimation in GLMMs [10]. In addition, this strategy produced satisfactory results in the automated MCEM algorithm [9] in analysing GLMMs. Moreover, it has also been widely applied to a large number of statistical models, see Reference [27] and the references therein. Theoretically, the general ideas of the SA-MCMC algorithm and the proposed path sampling procedure can be applied to handle missing data that have a non-ignorable missing mechanism [26]. However, the simulation of $\{\mathbf{Y}_{\mathrm{m}}^{(j)}, j=1,\ldots,J\}$ from the corresponding conditional distribution in the ML estimation and the computation of the observed-data log-likelihood would be more complicated.

Some existing software, for example SAS [28] and GenStat [29], provide estimates of the unknown parameters in a GLMM model. We have used GLIMMIX macro in the standard software SAS [28] for obtaining the estimate of $\boldsymbol{\psi}$ in some of the models in the lung cancer example. The statistical approaches used in this program are the pseudo-likelihood (PL) and the restricted pseudo-likelihood (REPL). The PL and REPL estimates, and their standard error estimates that are available from the SAS program, are reported in Table V, together with the ML estimates and standard error estimates (from Table IV). It can be seen from this table that the different parameter estimates are close (except the estimate of σ_u^2 in M_1), but the standard errors estimates associated with our ML estimates, which were computed through Louis formula [30] with observations simulated by some MCMC methods, are smaller. As the PL and REML approaches are different from our ML approach, this difference in standard errors estimates is reasonable. We use our path sampling procedure to obtain the 'observed-data log-likelihoods' evaluated at the PL and REPL estimates. The results are also reported in Table V. The differences between the different 'observed-data log-likelihoods' are small. However, as the theory of BIC is developed on the basis of the ML estimate, more research

Table V. PL and REPL estimates obtained from SAS.

	ML Estimates and Standard Error Estimates (in parentheses)					
	M_1			M_3		
Parameter	ML	PL	REML	ML	PL	REPL
β_0	−1.899	−1.907	−1.907	−1.932	−1.882	−1.8807
	(0.083)	(0.251)	(0.261)	(0.063)	(0.210)	(0.217)
β_1	1.688	1.682	1.682	1.691	1.645	1.643
	(0.086)	(0.181)	(0.188)	(0.065)	(0.219)	(0.228)
σ_u^2	0.336	0.432	0.466	0.186	0.187	0.198
	(0.070)	(—)	(—)	(0.053)	(—)	(—)
σ_{v1}^2	—	—	—	0.235	0.234	0.259
				(0.058)	(—)	(—)
Observed-data log-likelihood	−8067.5	−8080.2	−8084.1	−8060.1	−8061.3	−8065.5

is required to justify the usage of the 'observed-data log-likelihoods' evaluated at the PL or REPL estimates for model comparison.

In this article, we emphasize the use of the observed-data log-likelihood in model comparison. However, it can be applied to other statistical inferences. For example, it can be applied as follows to discriminant analysis for classifying an observation into one of the possible populations, $h = 1, \ldots, H$. Based on the basic idea of discriminant analysis, an observation $\mathbf{y}$ is classified to population h with a model M_h and ML estimate $\hat{\boldsymbol{\psi}}_h$ if $p(\mathbf{y}|\hat{\boldsymbol{\psi}}_h, M_h) > p(\mathbf{y}|\hat{\boldsymbol{\psi}}_l, M_l)$ for any $l \neq h$. Regarding $\{\mathbf{y}\}$ as a data set, $p(\mathbf{y}|\hat{\boldsymbol{\psi}}_h, M_h)$ can be computed by the proposed path sampling procedure. The probability of misclassification can then be estimated via the jackknife method. Moreover, it can be applied to construct case-deletion measure for identifying influential observations that may be potential outliers. An observation may be regarded as influential if the deletion of it from the data set would greatly change the observed-data log-likelihood. As GLMMs is very useful in medical research, we expect to see many novel applications of the proposed method in the future.

We utilize path sampling [13] to develop a procedure for computing the observed-data log-likelihood that is associated with GLMM. Although path sampling is well known, and has been applied to a number of latent variable models [15, 16], among others, its application to GLMM is novel. As path sampling is a generalization of importance sampling and bridge sampling, we expect that it would produce more accurate results. Another good alternative is the method developed by Chib and Jeliazkov [31]. Comparison of these powerful computing methods, either theoretically or by simulation, represent an interesting topic for further research.

ACKNOWLEDGEMENTS

The work that is described in this paper was fully supported by a Grant from the Research Grants Council of the HKSAR (Project No. CUHK 4243/02H). The authors are thankful to the Editor and reviewers for valuable suggestions for improving the paper.

REFERENCES

1. Pawitan Y, Reilly M, Nilsson E, Cnattinguis S. Estimation of genetic and environmental factors for binary traits using family data. *Statistics in Medicine* 2004; **23**:449–465.
2. Kleinman K, Lazarus R, Platt R. A generalized linear mixed model approach for detecting incident clusters of disease in small areas, with an application to biological terrorism. *American Journal of Epidemiology* 2004; **159**:217–224.
3. Manton KG, Woodbury MA, Stallard E, Riggan WB, Creason JP, Pellom Al. Empirical Bayes procedure for stabilizing maps of U.S. cancer mortality rates. *Journal of the American Statistical Association* 1989; **84**: 637–650.
4. Zeger SL, Liang KY, Albert YS. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 1988; **44**:1049–1060.
5. Zeger SL, Karim MR. Generalized linear model with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association* 1991; **86**:79–86.
6. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* 1993; **88**:9–25.
7. Lin X, Breslow NE. Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association* 1996; **91**:1007–1016.
8. Wolfinger R, O'Connell M. Generalized linear models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation* 1993; **48**:233–234.
9. Booth JG, Hobert JP. Maximum generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society*, *Series B* 1999; **61**:265–285.
10. Zhu HT, Lee SY. Analysis of generalized linear mixed models via a stochastic approximation algorithm with Markov Chain Monte Carlo method. *Statistical Computing* 2002; **12**:175–183.

11. Sarton N, Severini TA. Conditional likelihood inference in generalized linear mixed models. *Statistica Sinica* 2004; **14**:349–360.
12. Zhu HT, Lee SY. Local influence for generalized linear mixed models. *Canadian Journal of Statistics* 2003; **31**:293–309.
13. Gelman A, Meng XL. Simulating normalizing constant: from importance sampling to bridge sampling to path sampling. *Statistical Science* 1998; **13**:163–185.
14. Meng XL, Wong WH. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica* 1996; **6**:831–860.
15. Lee SY, Song XY. Maximum likelihood analysis of a general latent variable model with hierarchically mixed data. *Biometrics* 2004; **60**:624–636.
16. Lee SY, Song XY. Model comparison of nonlinear structural equation models with fixed covariates. *Psychometrika* 2003; **68**:27–47.
17. McCulloch CE. Maximum likelihood algorithm for generalized linear mixed models. *Journal of the American Statistical Association* 1997; **92**:162–170.
18. Kass RE, Raftery AE. Bayes factors. *Journal of the American Statistical Association* 1995; **90**:773–795.
19. Glonke GFV, McCullagh P. Multivariate logistic models. *Journal of the Royal Statistical Society*, *Series B* 1995; **57**:533–546.
20. Chib S, Greenberg E. Analysis of multivariate probit models. *Biometrika* 1998; **85**:347–361.
21. Chen JL, Zhang DW, Davidiad M. A Monte Carlo EM algorithm for generalized linear mixed models with flexible random effects distribution. *Biostatistics* 2002; **3**:347–360.
22. Dorn HF. The relationship of cancer of the lung and the use of tobacco. *American Statistician* 1954; **8**:7–13.
23. Cox DR, Snell EJ. *Analysis of Binary Data* (2nd edn). Chapman & Hall: London, 1989.
24. Tanner MA, Wong WH. The calculation of posterior distribution by data augmentation (with discussion). *Journal of the American Statistical Association* 1987; **86**:79–86.
25. Rubin DB. EM and beyond. *Psychometrika* 1991; **56**:241–254.
26. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. Wiley: New York, 2002.
27. Meng XL, Schilling S. Fitting full-information item factor models and an empirical application of bridge sampling. *Journal of the American Statistical Association* 1996; **91**:1254–1267.
28. Little RC, Millcken GA, Stroap WW, Wolfinger RD. *SAS System for Mixed Models*. SAS Institute: Cary, NC, 1996.
29. Payne RW *et al*. *GenStat Release 8 Reference Manual*. VSN Internation: Oxford, U.K., 2005.
30. Louis TA. Finding the observed information matrix when using EM algorithm. *Journal of the Royal Statistical Society*, *Series B* 1982; **44**:226–233.
31. Chib S, Jeliazkov I. Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association* 2001; **96**:270–281.

