



ELSEVIER

Available online at [www.sciencedirect.com](http://www.sciencedirect.com)

SCIENCE @ DIRECT®

Journal of  
Multivariate  
Analysis

Journal of Multivariate Analysis 94 (2005) 300–312

[www.elsevier.com/locate/jmva](http://www.elsevier.com/locate/jmva)

# Monte Carlo approximation through Gibbs output in generalized linear mixed models

Jennifer S.K. Chan<sup>a,\*</sup>, Anthony Y.C. Kuk<sup>b</sup>, Carrie H.K. Yam<sup>c</sup>

<sup>a</sup>*Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong*

<sup>b</sup>*Department of Statistics and Applied Probability, National University of Singapore, Block S16, 3 Science Drive 2, Singapore 117543, Singapore*

<sup>c</sup>*Department of Community Medicine, The University of Hong Kong, Hong Kong*

Received 5 March 2003

Available online 13 July 2004

## Abstract

Geyer (J. Roy. Statist. Soc. 56 (1994) 291) proposed Monte Carlo method to approximate the whole likelihood function. His method is limited to choosing a proper reference point. We attempt to improve the method by assigning some prior information to the parameters and using the Gibbs output to evaluate the marginal likelihood and its derivatives through a Monte Carlo approximation. Vague priors are assigned to the parameters as well as the random effects within the Bayesian framework to represent a non-informative setting. Then the maximum likelihood estimates are obtained through the Newton Raphson method. Thus, our method serves as a bridge between Bayesian and classical approaches. The method is illustrated by analyzing the famous salamander mating data by generalized linear mixed models.

© 2004 Elsevier Inc. All rights reserved.

AMS 1991 subject classification: 62J12; 62F10; 62H12; 62M10

Keywords: Generalized linear mixed model; Monte Carlo Newton Raphson; Monte Carlo relative likelihood; Gibbs sampler; Metropolis–Hastings algorithm

## 1. Introduction

The generalized linear models (GLMs) extend the classical linear models to the exponential family of sampling distributions. GLMs have an immense impact on both theoretical and

\* Corresponding author. Fax: +852-2858-9041.

E-mail addresses: [jchan@hkustasc.hku.hk](mailto:jchan@hkustasc.hku.hk) (J.S.K. Chan), [kuk@stat.nus.edu.sg](mailto:kuk@stat.nus.edu.sg) (A.Y.C. Kuk), [cyam@graduate.hku.hk](mailto:cyam@graduate.hku.hk) (C.H.K. Yam).

practical aspects in statistics. Inclusion of random effects in the GLMs defines the class of generalized linear mixed model (GLMM) which overcome the problem of over-dispersion and accommodate population heterogeneity. These models are applicable in many practical situations. However, the presence of random effects in the model complicates the computation of marginal likelihood and hence the maximum likelihood estimates considerably, as the likelihood function may involve high-dimensional integrals. Diverse methodologies, both Bayesian or classical approaches, arise in the implementation and estimation of the GLMMs.

In Bayesian perspective, Zeger and Karim [24] investigated GLMM by Gibbs sampling approach. They analyzed the famous salamander mating data which has crossed random effects [11]. Gibbs sampler is used to draw samples from the full conditional density. This method becomes computationally very intensive when the full conditional density is not in a standard form. Apart from the Bayesian approach, there are methodologies that adopt a classical approach. McCullagh and Nelder [17] used the estimating equation approach in GLMM using Taylor series expansion to approximate the integrands. This approach is not efficient when the integrand is high-dimensional. Breslow and Clayton [1] proposed the penalized quasi-likelihood (PQL). PQL estimates are biased towards zero for some variance components. Breslow and Lin [2] and Lin and Breslow [16] revised the methodology by a bias-corrected PQL for GLMM with single and multiple components of dispersion, respectively. This method improves the asymptotic performance of PQL estimates, but inflates the sampling variance. The efficiency of the estimates also depends on the sample size. McCulloch [18] investigated GLMM with a probit link using Monte Carlo EM (MCEM) method. He extended MCEM to the logit model and introduced the Monte Carlo Newton Raphson (MCNR) and simulated maximum likelihood (SML) methods [19]. However, iterations of MCEM and MCNR do not always converge to the global maximum. The importance function used in SML may be far away from the true function and this will impose difficulties in the estimation. Thus, Kuk [12] suggested the Laplace importance sampling in the SML and MCNR methods. He chose a normal importance function for the random effects with mean as the maximizer of the joint density and variance as the corresponding information matrix. Kuk and Cheng [15] also suggested a functional approach called Monte Carlo relative likelihood (MCRL) and a pointwise approach using the MCNR procedures to approximate the likelihood function and obtain maximum likelihood (ML) estimates in GLMMs. However, the functional approach to calculating the relative likelihood requires a proper reference point [9]. This is difficult to choose. One remedy is to update the reference point several times.

Apart from using Bayesian or classical approach separately, some researchers suggested methodologies that combine the two approaches. Chib [5] suggested the Gibbs output in calculating the marginal likelihood which is the normalizing constant of the posterior density. To obtain Gibbs output, the full conditional densities are required but they may not be in a standard form. Chib and Jeliazkov [6] further investigated the use of Metropolis–Hastings output when the full conditional densities are not standard. In this paper, we propose a new methodology. It also uses the Gibbs output to calculate the marginal likelihood, but instead of choosing a reference point for the parameters in calculating the relative likelihood [15], we assign prior distribution to the parameters and sample random effects as well as parameters from the joint posterior density using the Gibbs sampler. Then, based on the

Gibbs output, we use the MCRL and MCNR, which adopt a Monte Carlo approximation to the relative marginal likelihood function and its derivatives in order to obtain the maximum likelihood estimates by the Newton Raphson method. This provides an alternative method of evaluating the marginal likelihood in the classical approach using Gibbs sampling outputs in the Bayesian approach. Thus the same Gibbs output can be used to conduct both Bayesian and Frequentist inference, bridging the two approaches. We illustrate this method using the famous salamander mating data analysed by McCullagh and Nelder [17].

The paper will be presented as follows. Section 2 presents the evolution and introduction of our method. We will illustrate this method by using the famous salamander mating data in Section 3. Section 4 reports the numerical results with comments. A conclusion is given in Section 5.

## 2. The Monte Carlo approximation through Gibbs output

We define the marginal likelihood based on the observed data  $\mathbf{y}$  over the random effects  $\mathbf{z}$  as

$$L(\boldsymbol{\theta}; \mathbf{y}) = \int f(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}) d\mathbf{z}$$

where  $\boldsymbol{\theta}$  is a vector of model parameters. Geyer and Thompson [10] proposed to calculate the marginal relative likelihood using a Monte Carlo approximation as follows:

$$\begin{aligned} \frac{L(\boldsymbol{\theta}; \mathbf{y})}{L(\boldsymbol{\theta}_0; \mathbf{y})} &= \int \frac{f(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta})}{f(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}_0)} f(\mathbf{z}|\mathbf{y}; \boldsymbol{\theta}_0) d\mathbf{z} \\ &\approx \frac{1}{M} \sum_{i=1}^M \frac{f(\mathbf{y}, \mathbf{z}_i|\boldsymbol{\theta})}{f(\mathbf{y}, \mathbf{z}_i|\boldsymbol{\theta}_0)} \end{aligned} \quad (1)$$

where the random effects  $\mathbf{z}_i$  are drawn from the conditional density  $f(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_0)$  based on a given reference point  $\boldsymbol{\theta}_0$ , and  $i$  indexes simulations used in the approximation. Since the conditional density  $f(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_0)$  in (1) is only used as an importance sampling function, it will not affect the unbiasedness of the approximation but only its efficiency [12]. However, the local approximation may be good only if the reference point  $\boldsymbol{\theta}_0$  is close to the true ML estimates. Kuk and Cheng [15] demonstrated that the resulting maximizer may differ substantially from the true ML estimates if an inappropriate reference point is chosen. One remedy is to update the reference point  $\boldsymbol{\theta}_0$  to the current  $\hat{\boldsymbol{\theta}}$  after each updating and then simulate a new vector of  $\mathbf{z}$  [10]. This can solve the problem of choosing an appropriate reference point in simulating a vector of random effects to approximate the likelihood function by Monte Carlo method. However, this method becomes computationally very intensive as it requires nested iterations and the simulation of a new set of random effects based on each update of current  $\boldsymbol{\theta}$ .

Apart from the Monte Carlo relative likelihood approach, McCulloch [19] suggested a similar simulated maximum likelihood (SML) approach. This method requires an optimal importance sampling function to draw the random effects of Monte Carlo approximation. It performs poorly when the choice of importance sampling distribution is far away from the true distribution of the random effects.

Our method is an extension of the Monte Carlo relative likelihood (MCRL) approach proposed by Kuk and Cheng [15]. Its advantage is that we do not need to specify a proper reference point in the estimation. Instead, we adopt any conveniently chosen prior density of  $\theta^*$ , say  $h(\theta^*)$ , as the prior information for the parameters and sample random effects  $z$  as well as the parameters  $\theta^*$  from the joint posterior distribution using Gibbs sampler. The method does not rely on a single specified reference point  $\theta_0$  and hence avoids iterations. Based on the Gibbs output, we calculate the marginal relative likelihood and its derivatives by Monte Carlo method using the expression:

$$\begin{aligned}
 L(\theta) &= \int \frac{f(\mathbf{y}, z|\theta)}{f(\mathbf{y}, z|\theta^*)} f(\mathbf{y}, z|\theta^*) dz \quad \text{for any } \theta^* \\
 &= \int \int \frac{f(\mathbf{y}, z|\theta)}{f(\mathbf{y}, z|\theta^*)} f(\mathbf{y}, z|\theta^*) dz h(\theta^*) d\theta^* \\
 \frac{L(\theta)}{f(\mathbf{y})} &= \int \int \frac{f(\mathbf{y}, z|\theta)}{f(\mathbf{y}, z|\theta^*)} f(z, \theta^*|\mathbf{y}) dz d\theta^* \tag{2}
 \end{aligned}$$

$$L(\theta) \propto \frac{1}{M} \sum_{i=1}^M \frac{f(\mathbf{y}, z_i|\theta)}{f(\mathbf{y}, z_i|\theta_i)} \tag{3}$$

where  $(z_i, \theta_i) \sim f(z, \theta|\mathbf{y})$ . Thus we sample the random effects  $z_i$  and parameters  $\theta_i$  from a joint posterior density and use them to evaluate the relative likelihood by a Monte Carlo approximation. This method uses the Gibbs sampling output in Bayesian analysis to evaluate the relative likelihood function in the classical analysis. Apart from solving the problem of choosing a proper reference point  $\theta_0$ , it requires no simulation of a new set of random effects each time the reference point is updated. For model selection, we have to rely on other methods to approximate the log-likelihood value.

In Bayesian inference, there are concerns that the specification of prior density  $h(\theta^*)$  may have an effect on the resultant parameter estimates. Note that the identity (2) holds for any choice of the prior density  $h(\theta^*)$  that is proper, hence the Monte Carlo approximation (3) will always be unbiased. From the point of view of importance sampling, the variance of the Monte Carlo approximation is expected to be small if the posterior distribution of  $\theta^*$  is concentrated around the ML estimate  $\hat{\theta}$ . Now if the sample size is large, the sample information is likely to outweigh the effect of any non-degenerated prior specification and the posterior simulation will be concentrated around the ML estimates  $\hat{\theta}$  automatically to give us a good approximation of  $L(\theta)$  around  $\hat{\theta}$ . If the sample size is not large enough, Kuk [13] suggested posterior sharpening and data duplication to improve the accuracy of the simulated likelihood function.

### 3. Example on the salamander mating data

We use the famous salamander mating example to illustrate our method. Salamander Mating experiment was conducted by Arnold and Verrell of the Department of Ecology and Evolution at the University of Chicago [17, p. 439–450]. The salamanders come from two populations: Rough Butt (RB) and Whiteside (WS). The objective of this experiment was

to investigate whether there were barriers to inter-breeding in the salamanders from these two geographically isolated populations.

There were three experiments, each involving 20 female and 20 male salamanders. Each female salamander in the experiment mated with three male salamanders from its population and another three from the other population under a crossed design. In total, there were 120 observations per experiment. The first experiment was done in summer in 1986 and the second in fall in the same year using the same animals. The third experiment was carried out at the same time but with a new set of salamanders. We will illustrate our method using data from the first experiment.

The responses, coded as 1 if the mating is successful and as 0 otherwise, are not independent because each female salamander is paired up with six male salamanders in the crossed design. However, the random effects, introduced to account for overdispersion and clustering, complicate the parameter estimation considerably. The crossed design hinders us from factorizing the likelihood function and hence the resulting likelihood function involves integrals of 20 dimensions and is beyond the capacity of numerical approximation. Many estimation methods have been proposed to overcome the difficulties in evaluating the likelihood function. For example, McCullagh and Nelder [17] used the estimating equation approach. Karim and Zeger [11] adopted the Gibbs sampling approach. Breslow and Clayton [1] and Lin and Breslow [16] used the uncorrected and corrected penalized quasi-likelihood approach, respectively. Shun [22] suggested the modified Laplace approximation while Kuk [12] introduced the Laplace importance sampling method. All of these methods use a logit link function. On the other hand, McCulloch [18] suggested the probit link through the Monte Carlo EM method (MCEM). Chan and Kuk [4] extended the MCEM to correlated binary data. In this paper, we adopt a logit link function and use the Monte Carlo approximation through Gibbs output to estimate the likelihood function. We compare the results with other researchers and study the goodness-of-fit of our method.

The generalized linear mixed model is defined as follows. Let  $Y_t$  be the binary response of mating (1 = success, 0 = failure) in the experiment, where  $t = 1, \dots, 120$  correspond to the matings. The fixed effects indicate which population the male and female salamander belongs to  $WSF_t = 1$  if the female salamander involved in the  $t$ th mating came from Whiteside, and 0 otherwise. Similarly,  $WSM_t$  equals to 1 if a male salamander involved in the  $t$ th mating came from Whiteside, and 0 otherwise.  $WSF_t \times WSM_t$  is the interaction between the two fixed effects. We define  $z_{1,j}$  and  $z_{2,j}$  as the random effects of the  $j$ th female and male salamanders with  $j = 1, \dots, 20$ . These random effects follow normal distributions with means equal to 0 and variance equal to  $\sigma_1^2$  and  $\sigma_2^2$ , respectively. We denote the vector of the random effects by  $\mathbf{z} = (z_{1,1}, \dots, z_{1,20}, z_{2,1}, \dots, z_{2,20})$  and the vector of parameters by  $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\sigma}) = (\beta_0, \beta_1, \beta_2, \beta_3, \sigma_1, \sigma_2)$ . The marginal density is given by

$$f(\mathbf{y}|\boldsymbol{\theta}) = \int f(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}) d\mathbf{z} = \int \dots \int f(\mathbf{y}|\mathbf{z}, \boldsymbol{\beta}) f(\mathbf{z}|\boldsymbol{\sigma}) dz_{1,1} \dots dz_{1,20} dz_{2,1} \dots dz_{2,20}$$

where

$$f(\mathbf{y}|\mathbf{z}, \boldsymbol{\beta}) = \prod_{t=1}^{120} \frac{\exp(\eta_t y_t)}{1 + \exp(\eta_t)} \quad (4)$$

$$f(\mathbf{z}|\boldsymbol{\sigma}) = \prod_{j=1}^{20} \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{z_{1,j}^2}{2\sigma_1^2}\right) \times \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{z_{2,j}^2}{2\sigma_2^2}\right) \quad (5)$$

$$\eta_t = \beta_0 + \beta_1 W S F_t + \beta_2 W S M_t + \beta_3 W S F_t \times W S M_t + z_{1,s'} + z_{2,s''} \quad (6)$$

$s'$  and  $s''$  correspond, respectively, to the female and male salamander involved in the  $t$ th mating.

To obtain the ML estimates, we evaluate the marginal likelihood by first sampling  $\mathbf{z}$  and  $\boldsymbol{\theta}$  from the joint posterior density  $f(\mathbf{z}, \boldsymbol{\theta}|\mathbf{y})$  adopting a vague prior  $h(\boldsymbol{\theta})$  for the parameters  $\boldsymbol{\theta}$  and using Gibbs sampler. Let  $(\mathbf{z}_i, \boldsymbol{\theta}_i)$  be the  $i$ th simulated set of random effects and parameters where  $\mathbf{z}_i = (z_{1,1,i}, \dots, z_{1,20,i}, z_{2,1,i}, \dots, z_{2,20,i})$  and  $\boldsymbol{\theta}_i = (\beta_{0i}, \beta_{1i}, \beta_{2i}, \beta_{3i}, \sigma_{1i}, \sigma_{2i})$ . We then use the Gibbs output to approximate the marginal likelihood function given by (3), where  $f(\mathbf{y}, \mathbf{z}_i|\boldsymbol{\theta})$  and  $f(\mathbf{y}, \mathbf{z}_i|\boldsymbol{\theta}_i)$  are calculated by

$$f(\mathbf{y}, \mathbf{z}_i|\boldsymbol{\theta}) = f(\mathbf{y}|\mathbf{z}_i, \boldsymbol{\beta}) f(\mathbf{z}_i|\boldsymbol{\sigma}) \quad (7)$$

$$f(\mathbf{y}, \mathbf{z}_i|\boldsymbol{\theta}_i) = f(\mathbf{y}|\mathbf{z}_i, \boldsymbol{\beta}_i) f(\mathbf{z}_i|\boldsymbol{\sigma}_i) \quad (8)$$

$$\eta_{ti} = \beta_0 + \beta_1 W S F_t + \beta_2 W S M_t + \beta_3 W S F_t \times W S M_t + z_{1,s',i} + z_{2,s'',i} \quad (9)$$

$$\eta_{ti}^* = \beta_{0i} + \beta_{1i} W S F_t + \beta_{2i} W S M_t + \beta_{3i} W S F_t \times W S M_t + z_{1,s',i} + z_{2,s'',i} \quad (10)$$

$f(\mathbf{y}|\mathbf{z}_i, \boldsymbol{\beta})$  and  $f(\mathbf{z}_i|\boldsymbol{\sigma})$  are given by (4) and (5), respectively, with  $\mathbf{z}_i$  replacing  $\mathbf{z}$  and  $\eta_{ti}$  replacing  $\eta_t$  in (6) and  $f(\mathbf{y}|\mathbf{z}_i, \boldsymbol{\beta}_i)$  and  $f(\mathbf{z}_i|\boldsymbol{\sigma}_i)$  are given by (4) and (5), respectively, with  $(\mathbf{z}_i, \boldsymbol{\beta}_i, \boldsymbol{\sigma}_i)$  replacing  $(\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\sigma})$  and  $\eta_{ti}^*$  replacing  $\eta_t$  in (6). Newton Raphson method is then used to obtain the ML estimates.

#### 4. Results

As the conditional density  $f(\mathbf{z}, \boldsymbol{\theta}|\mathbf{y})$  is not standard, sampling methods such as Metropolis–Hastings or adaptive rejection sampling are used. Our Gibbs output is obtained from the Bayesian software ‘WinBUGS’ [23], adopting a vague prior for  $\boldsymbol{\theta}$ . To approximate the likelihood function closely, the Gibbs output should be large in size, independent and stable to make sure that it follows the posterior distribution. We run a series of 220,000 iterations, discarding the first 20,000 observations in the burn-in period and taking every 10 observations resulting in a sample of  $M = 20,000$  sets of  $(\mathbf{z}_i, \boldsymbol{\theta}_i)$ . The convergence and auto-correlation of the Gibbs output are checked by history plots and auto-correlation functions and results show that the sample is satisfactory.

Calculation of the ML estimates using the Newton Raphson method requires the first and second derivatives of log-likelihood function with respect to  $\boldsymbol{\theta}$ , denoted by  $l'(\boldsymbol{\theta}; \mathbf{y})$  and  $l''(\boldsymbol{\theta}; \mathbf{y})$ , respectively. The relevant expressions are given in appendix. The ML estimates are updated iteratively until converge by

$$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} - [l''(\boldsymbol{\theta}^{(m)}; \mathbf{y})]^{-1} l'(\boldsymbol{\theta}^{(m)}; \mathbf{y})$$

where  $m$  is the number of iterations. See Kuk and Cheng [14] for the details of MCNR algorithm. In our example, the initial values are set to the moment estimates.

Table 1

Parameter estimates by our method, with standard errors in brackets, and other estimation methods

Estimates	$\beta_0$	$\beta_1$	$\beta_2$	$\beta_3$	$\sigma_1$	$\sigma_2$
Our method	1.30 (0.56)	-2.83 (0.82)	-0.50 (0.68)	3.22 (0.92)	1.30	0.38
Moment	0.97	-2.12	-0.30	2.26	1.17	0.84
Laplace importance sampling	1.39	-3.05	-0.45	3.29	1.31	0.50
Penalized quasi-likelihood (PQL)	0.79	-2.29	-0.54	2.82	1.19	0.30
Bias-corrected PQL (CPQL)	0.79	-2.29	-0.54	2.82	1.31	0.63
Gibbs sampling	1.48	-3.25	-0.50	3.62	1.53	0.37
Laplace approximation	1.39	-3.06	-0.45	3.31	1.34	0.50

Table 2

Observed and expected proportions of successful mating for different geographically located female and male salamanders involved in the mating and the average percentage errors by the different estimation methods

	Observed proportion	Expected proportion						
		Proposed	Moment	Lap. imp.	PQL	CPQL	Gibbs	Lap. ap.
$\pi_{WW}$	0.700	0.719	0.650	0.713	0.653	0.644	0.729	0.713
$\pi_{WR}$	0.233	0.232	0.294	0.217	0.229	0.242	0.214	0.218
$\pi_{RW}$	0.667	0.653	0.625	0.674	0.551	0.547	0.672	0.672
$\pi_{RR}$	0.730	0.735	0.677	0.745	0.655	0.646	0.748	0.743
Average error (in %)		0.015	0.117	0.029	0.091	0.104	0.039	0.028

Table 1 shows the model fits using our method for the experiment carried out in summer, 1986. We also include estimates obtained by other methods for comparison. Our estimates for  $\beta_0, \beta_1, \beta_2, \beta_3, \sigma_1$  are close to the estimates obtained by Kuk [12] using the method of Laplace importance sampling while the estimate for  $\sigma_2$  is close to the estimates obtained by Karim and Zeger [11] using Gibbs sampling. It is interesting that our estimates lie between the results of Bayesian and classical approach. While other methods do not provide standard error estimates, the standard errors estimates for  $\beta$  using our proposed method are reported in Table 1.

The goodness-of-fit can be assessed by the estimated proportions of successful matings for the different geographically located (WS or RB) female and male salamanders involved in the mating. We let  $\pi_{RW}$  to denote the proportion of successful matings between a female RB salamander and a male WS salamander. The observed and expected proportions are shown in Table 2 for the various methods listed in Table 1. The expected probabilities obtained by our method are close to the observed probabilities. The average percentage error is defined as

$$\frac{1}{4} \sum_{i,j} \frac{|\pi_{ij} - \hat{\pi}_{ij}|}{\pi_{ij}}, \quad i, j = W, R$$

where  $\pi_{ij}$  is the observed proportion and  $\hat{\pi}_{ij}$  is the expected proportion under a model. Our proposed model gives the smallest average error for the salamander data.

### 5. Example on exponential mixture of Poisson distributions

As suggested by Kuk [13], we will make use of a simple conjugate model to demonstrate the accuracy of our proposed method.

Let  $y_j, j = 1, \dots, n$  be independently distributed as *Poisson*( $\mu_j$ ) given the random effects  $u_j$  and the random effects  $u_j$  are independently distributed as *Exponential*( $\lambda$ ) so that  $S = \sum_{j=1}^n u_j$  follows *Gamma*( $n, \lambda$ ). Then it is well known that the marginal distribution is geometric with success probability  $\pi = \lambda/(\lambda + 1)$  and the marginal likelihood function is

$$L(\pi; \mathbf{y}) = \prod_{j=1}^n [(1 - \pi)^{y_j} \pi] = (1 - \pi)^T \pi^n \tag{11}$$

where  $T = \sum_{j=1}^n y_j$ . We specify a convenient prior *Beta*( $a, b$ ) to  $\pi$  and approximate the likelihood function using our proposed method as

$$L(\theta) = f(\mathbf{y}) \left\{ \frac{1}{M} \sum_{i=1}^M \left( \frac{\lambda}{\lambda_i} \right)^n \exp[-(\lambda - \lambda_i) S_i] \right\} \tag{12}$$

where the proportionality constant as in (3) is  $f(\mathbf{y}) = \frac{B(a + n, b + T)}{B(a, b)}$ .

The model is fitted to the bird hops data reported in Rice [20,p. 288] that records the number of flights for birds with  $n = 130$  and  $T = 233$ . According to the likelihood function (11), the ML estimate of  $\pi$  is

$$\hat{\pi}_{ML} = \frac{n}{n + T} = \frac{130}{363} = 0.358.$$

Using (12), we could have sampled  $S_i = \sum_{j=1}^n u_{ji}$  and  $\pi_i = \frac{\lambda_i}{\lambda_i + 1}$  directly from the posterior distributions, *Gamma*( $n+T, \lambda+1$ ) and *Beta*( $a+n, b+T$ ), respectively, which are standard distributions. Instead of drawing samples from exact posterior distribution, we demonstrate the use of Gibbs output which can be easily implemented by WinBUGS. From a Gibbs sampling chain of 12000 iterations, we discard the first 2000 iterations as in burn-in period and sample every 10th iteration thereafter resulting in a sample of  $M = 1000 \lambda_i$  and  $S_i$ . Two choices of parameters, (1,9) and (6,4) for ( $a, b$ ) are used so that the prior means for  $\pi$  are 0.1 and 0.6, respectively and they are about 0.25 on either side of  $\hat{\pi}_{ML}$ . The exact log-likelihood function and its approximations based on the Gibbs output for the two sets of ( $a, b$ ) are given in Fig. 1. The approximated ML estimates are  $\hat{\pi}_1 = 0.356$  when ( $a, b$ ) = (1, 9) and  $\hat{\pi}_2 = 0.358$  when ( $a, b$ ) = (6, 4). At  $\hat{\pi}_{ML} = 0.358$ , the exact log-likelihood value is  $-236.7968$  whereas the approximated log-likelihood values are  $-240.1715$  (1.4% error) and  $-234.8842$  (0.8% error) respectively for ( $a, b$ ) equals (1,9) and (6,4). This show that the Gibbs output mimic samples from the posterior distribution closely.

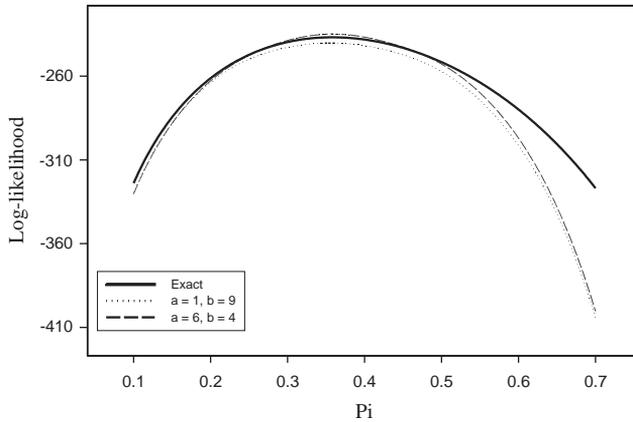


Fig. 1. Monte Carlo likelihood for the bird hops data based on posterior  $\theta$ .

**6. Conclusion**

We have shown that our method of assigning prior to the parameters and using the Gibbs output for Monte Carlo approximation is useful in obtaining the ML estimates for models that involve multivariate random effects. The use of prior information solves the problem of choosing a proper reference point and updating it through iterations in the Monte Carlo relative likelihood approach.

Our method is an innovation in making inference in the generalized linear mixed models through the Gibbs output for Monte Carlo approximation. It makes evaluating the marginal likelihood easier. We illustrate our method on the famous salamander mating data. The crossed random effects induce high-dimensional integrals in the likelihood function and the integrals cannot be factorized. The Monte Carlo approximation provides a practical solution. In a non-parametric approach, Chib [5] suggested ways to evaluate the marginal likelihood for a given model through Gibbs output. Using his idea, we can further extend our methodology when the prior density for the parameters is itself estimated. This is another interesting area to investigate.

Finally, to simulate the Gibbs output in the Bayesian step, a non-standard sampling method, such as Metropolis–Hastings, adaptive rejection sampling or ratio-of-uniform, is needed. The convergence and auto-correlation of the Gibbs output should be checked to make sure that it follows the desired distribution. A good starting solution is needed for the MCNR method because, like its classical counterpart, it is sensitive to starting values. In our analysis, moment estimates are used. The ML estimates are subject to both sampling and approximation errors in the Gibbs output and the Monte Carlo method, respectively. The efficiency of the estimates can be improved by increasing the size of the Gibbs output. Another remedy is to obtain several Gibbs outputs, approximate the likelihood function several times and average the resulting parameter estimates across the replicate runs,

$$\hat{\theta} = \frac{1}{R} \sum_{r=1}^R \hat{\theta}_r$$

and estimate the sampling variance–covariance matrix of the final estimates [21] by

$$var(\theta) = \frac{1}{R} \sum_{r=1}^R var(\hat{\theta}_r) + \left(1 + \frac{1}{R}\right) \frac{1}{R-1} \sum_{r=1}^R (\hat{\theta}_r - \hat{\theta})(\hat{\theta}_r - \hat{\theta})^T$$

where the first and second terms give the variability of estimates within and between replicates, respectively, and  $\hat{\theta}_r$  and  $var(\hat{\theta}_r)$  are the parameter estimates and the variance-covariance matrix for the  $r$ th replicate. In our analysis, we set the size of the Gibbs output to be large enough ( $M = 20,000$ ) to reduce both sampling and approximation errors in the Gibbs output and the Monte Carlo method and hence used no replicate ( $R = 1$ ).

Most of the current result in GLMM analysis of the salamander mating data assumes normal random effects. In practice, it may be more appropriate to use a wider class of random-effects distributions to widen the scope of applications. For example, Choy and Smith [7] suggested the use of scale mixtures of normal (SMN) distributions, that include the Student- $t$ , symmetric stable, exponential-power and Laplace distributions for robustness consideration because of the heavy-tailed behavior of these distributions. Choy et al. [8] analyzed the salamander data through a full Bayesian approach and the random effects are modeled with the Student- $t$  distribution expressed as SMN distributions which simplifies the Bayesian computation and enables the detection of potential outlying random effects. To allow for more flexibility, the degrees of freedom of the Student- $t$  distribution are assigned a suitable prior distribution. Results show that the adoption of Student- $t$  distribution for the random effects improves the model fit considerably. Using this idea, we can extend our proposed methodology to model with Student- $t$  or other heavy-tail distributed random effects for robustness consideration. Then Eq. (5) will be replaced by the density function of Student- $t$  distribution and Eqs. (A.1) and (A.2) modified accordingly.

Another practical application of the proposed methodology is to informative dropout modeling. See Chan and Chau [3]. The proposed methodology makes feasible the application of classical ML approach to diverse classes of models that involve complicated likelihood functions.

### Acknowledgements

The authors are grateful to the editor and referees for their valuable comments and helpful suggestions which led to improvements in the paper. The research of J.S.K. Chan was supported by the RGC competitive earmarked research grant HKU 7257/99H, the University of Hong Kong.

### Appendix A. A.

The likelihood function is:

$$L(\theta; \mathbf{y}) \propto \frac{1}{M} \sum_{i=1}^M \frac{f(\mathbf{y}, z_i | \theta)}{f(\mathbf{y}, z_i | \theta_i)}$$

and its logarithm is:

$$\ell(\boldsymbol{\theta}; \mathbf{y}) \propto \ln \left( \frac{1}{M} \sum_{i=1}^M \frac{f(\mathbf{y}, \mathbf{z}_i | \boldsymbol{\theta})}{f(\mathbf{y}, \mathbf{z}_i | \boldsymbol{\theta}_i)} \right)$$

where  $f(\mathbf{y}, \mathbf{z}_i | \boldsymbol{\theta})$  and  $f(\mathbf{y}, \mathbf{z}_i | \boldsymbol{\theta}_i)$  are given by (7) and (8), respectively.

1. Differentiate the log-likelihood function once with respect to  $\beta_k$  where  $k = 0, 1, 2, 3$ :

$$\frac{\partial \ell(\boldsymbol{\theta}; \mathbf{y})}{\partial \beta_k} \propto L(\boldsymbol{\theta}; \mathbf{y})^{-1} \frac{1}{M} \sum_{i=1}^M \frac{f(\mathbf{y}, \mathbf{z}_i | \boldsymbol{\beta}, \boldsymbol{\sigma}) Df_{\beta_k}(\mathbf{y} | \mathbf{z}_i, \boldsymbol{\beta})}{f(\mathbf{y}, \mathbf{z}_i | \boldsymbol{\beta}_i, \boldsymbol{\sigma}_i)}$$

where

$$Df_{\beta_k}(\mathbf{y} | \mathbf{z}_i, \boldsymbol{\beta}) = \frac{\partial \ln f(\mathbf{y}, \mathbf{z}_i | \boldsymbol{\beta}, \boldsymbol{\sigma})}{\partial \beta_k} = \sum_{t=1}^{120} \left( X_{tk} \left[ y_t - \frac{\exp(\eta_{ti})}{1 + \exp(\eta_{ti})} \right] \right)$$

2. To assure a positive value on the parameter  $\sigma^2$ , we differentiate the log-likelihood function with respect to  $\ln \sigma_l^2$  where  $l = 1, 2$ . We have:

$$\frac{\partial \ell(\boldsymbol{\theta}; \mathbf{y})}{\partial \ln \sigma_l^2} \propto L(\boldsymbol{\theta}; \mathbf{y})^{-1} \frac{1}{M} \sum_{i=1}^M \frac{f(\mathbf{y}, \mathbf{z}_i | \boldsymbol{\beta}, \boldsymbol{\sigma}) Df_{\sigma_l}(\mathbf{z}_i | \boldsymbol{\sigma})}{f(\mathbf{y}, \mathbf{z}_i | \boldsymbol{\beta}_i, \boldsymbol{\sigma}_i)}$$

where

$$\begin{aligned} Df_{\sigma_l}(\mathbf{z}_i | \boldsymbol{\sigma}) &= \frac{\partial \ln f(\mathbf{z}_i | \boldsymbol{\sigma})}{\partial \ln \sigma_l^2} = \frac{\partial}{\partial \ln \sigma_l^2} \sum_{j=1}^{20} \left[ -\frac{1}{2} \ln \sigma_l^2 - \frac{z_{l,j,i}^2}{2 \exp(\ln \sigma_l^2)} \right] \\ &= \sum_{j=1}^{20} \frac{z_{l,j,i}^2}{2 \sigma_l^2} - 10. \end{aligned} \tag{A.1}$$

3. Differentiate the log-likelihood function twice with respect to  $\beta_{k_1}$  and  $\beta_{k_2}$ :

$$\begin{aligned} &\frac{\partial \ell(\boldsymbol{\theta}; \mathbf{y})}{\partial \beta_{k_1} \partial \beta_{k_2}} \\ &\propto L(\boldsymbol{\theta}; \mathbf{y})^{-2} \left\{ L(\boldsymbol{\theta}; \mathbf{y}) \frac{1}{M} \right. \\ &\quad \times \sum_{i=1}^M \frac{[Df_{\beta_{k_1}}(\mathbf{y} | \mathbf{z}_i, \boldsymbol{\beta}) Df_{\beta_{k_2}}(\mathbf{y} | \mathbf{z}_i, \boldsymbol{\beta}) + D^2 f_{\beta_{k_1} \beta_{k_2}}(\mathbf{y} | \mathbf{z}_i, \boldsymbol{\beta})] f(\mathbf{y}, \mathbf{z}_i | \boldsymbol{\beta}, \boldsymbol{\sigma})}{f(\mathbf{y}, \mathbf{z}_i | \boldsymbol{\beta}_i, \boldsymbol{\sigma}_i)} \\ &\quad - \left[ \frac{1}{M} \sum_{i=1}^M \frac{f(\mathbf{y}, \mathbf{z}_i | \boldsymbol{\beta}, \boldsymbol{\sigma}) Df_{\beta_{k_1}}(\mathbf{y} | \mathbf{z}_i, \boldsymbol{\beta})}{f(\mathbf{y}, \mathbf{z}_i | \boldsymbol{\beta}_i, \boldsymbol{\sigma}_i)} \right] \\ &\quad \left. \times \left[ \frac{1}{M} \sum_{i=1}^M \frac{f(\mathbf{y}, \mathbf{z}_i | \boldsymbol{\beta}, \boldsymbol{\sigma}) Df_{\beta_{k_2}}(\mathbf{y} | \mathbf{z}_i, \boldsymbol{\beta})}{f(\mathbf{y}, \mathbf{z}_i | \boldsymbol{\beta}_i, \boldsymbol{\sigma}_i)} \right] \right\} \end{aligned}$$

where

$$D^2 f_{\beta_{k_1} \beta_{k_2}}(\mathbf{y}|\mathbf{z}_i, \boldsymbol{\beta}) = \frac{\partial Df_{\beta_{k_1}}(\mathbf{y}|\mathbf{z}_i, \boldsymbol{\beta})}{\partial \beta_{k_2}} = - \sum_{t=1}^{120} X_{tk_1} X_{tk_2} \frac{\exp(\eta_{ti})}{[1 + \exp(\eta_{ti})]^2}.$$

4. Differentiate the log-likelihood function twice with respect to  $\ln \sigma_1^2$  and  $\ln \sigma_2^2$  where for  $l, l' = 1, 2$ :

$$\begin{aligned} & \frac{\partial^2 \ell(\boldsymbol{\theta}; \mathbf{y})}{\partial \ln \sigma_l^2 \partial \ln \sigma_{l'}^2} \\ & \propto L(\boldsymbol{\theta}; \mathbf{y})^{-2} \left\{ L(\boldsymbol{\theta}; \mathbf{y}) \frac{1}{M} \right. \\ & \quad \times \sum_{i=1}^M \frac{[Df_{\sigma_l}(\mathbf{z}_i|\boldsymbol{\sigma}) Df_{\sigma_{l'}}(\mathbf{z}_i|\boldsymbol{\sigma}) + D^2 f_{\sigma_l \sigma_{l'}}(\mathbf{z}_i|\boldsymbol{\sigma})] f(\mathbf{y}, \mathbf{z}_i|\boldsymbol{\beta}, \boldsymbol{\sigma})}{f(\mathbf{y}, \mathbf{z}_i|\boldsymbol{\beta}_i, \boldsymbol{\sigma}_i)} \\ & \quad - \left[ \frac{1}{M} \sum_{i=1}^M \frac{f(\mathbf{y}, \mathbf{z}_i|\boldsymbol{\beta}, \boldsymbol{\sigma}) Df_{\sigma_l}(\mathbf{z}_i|\boldsymbol{\sigma})}{f(\mathbf{y}, \mathbf{z}_i|\boldsymbol{\beta}_i, \boldsymbol{\sigma}_i)} \right] \\ & \quad \left. \times \left[ \frac{1}{M} \sum_{i=1}^M \frac{f(\mathbf{y}, \mathbf{z}_i|\boldsymbol{\beta}, \boldsymbol{\sigma}) Df_{\sigma_{l'}}(\mathbf{z}_i|\boldsymbol{\sigma})}{f(\mathbf{y}, \mathbf{z}_i|\boldsymbol{\beta}_i, \boldsymbol{\sigma}_i)} \right] \right\} \end{aligned}$$

where

$$D^2 f_{\sigma_l \sigma_{l'}}(\mathbf{z}_i|\boldsymbol{\sigma}) = \frac{\partial Df_{\sigma_l}(\mathbf{z}_i|\boldsymbol{\sigma})}{\partial \ln \sigma_{l'}^2} = \begin{cases} -\frac{1}{2} \sum_{j=1}^{20} \frac{z_{l,j,i}^2}{\sigma_l^2} & \text{if } l = l' \\ 0 & \text{if } l \neq l' \end{cases}. \tag{A.2}$$

**References**

[1] N.E. Breslow, D.G. Clayton, Approximate inference in generalized linear mixed models, *J. Amer. Statist. Assoc.* 88 (1993) 9–25.  
 [2] N.E. Breslow, X. Lin, Bias correction in generalized linear mixed models with a single component of dispersion, *Biometrika* 82 (1995) 81–91.  
 [3] J.S.K. Chan, V.K.K. Chau, Informative drop-out models for longitudinal binary data with random effects (2003), submitted for publication.  
 [4] J.S.K. Chan, A.Y.C. Kuk, Maximum likelihood estimation for probit-linear mixed models with correlated random effects, *Biometrics* 53 (1997) 86–97.  
 [5] S. Chib, Marginal likelihood from the Gibbs output, *J. Amer. Statist. Assoc.* 90 (1995) 1313–1321.  
 [6] S. Chib, I. Jeliazkov, Marginal likelihood from the Metropolis–Hastings output, *J. Amer. Statist. Assoc.* 96 (2001) 270–281.  
 [7] S.T.B. Choy, A.F.M. Smith, Hierarchical models with scale mixtures of normal distribution, *TEST* 6 (1997) 205–211.  
 [8] B.S.T. Choy, J.S.K. Chan, H.K. Yam, Robust analysis of salamander data, generalized linear model with random effects, *Bayesian Statistics* 7 (2003) 477–484.  
 [9] C.J. Geyer, On the convergence of Monte Carlo maximum likelihood calculations, *J. Roy. Statist. Soc. B* 56 (1994) 261–274.

- [10] C.J. Geyer, E.A. Thompson, Constrained Monte Carlo maximum likelihood for dependent data, *J. Roy. Statist. Soc. B* 54 (1992) 657–699.
- [11] M.R. Karim, S.L. Zeger, Generalized linear models with random effects; salamander mating revisited. *Biometrics* 48 (1992) 631–644.
- [12] A.Y.C. Kuk, Laplace importance sampling for generalized linear mixed models, *J. Statist. Comput. Simulation* 63 (1999) 143–158.
- [13] A.Y.C. Kuk, Automatic choice of driving values in Monte Carlo likelihood approximation via posterior simulation, *Statist. Comput.* 13 (2003) 101–109.
- [14] A.Y.C. Kuk, Y.W. Cheng, The Monte Carlo Newton–Raphson Algorithm, *J. Statist. Comput. Simulation* 59 (1997) 233–250.
- [15] A.Y.C. Kuk, Y.W. Cheng, Pointwise and functional approximations in Monte Carlo maximum likelihood estimation, *Statist. Comput.* 9 (1999) 91–99.
- [16] X. Lin, N.E. Breslow, Bias correction in generalized linear mixed models with multiple components of dispersion, *J. Amer. Statist. Assoc.* 91 (1996) 1007–1016.
- [17] P. McCullagh, J.A. Nelder, *Generalized linear Models*, 2nd Edition, Chapman & Hall, London, 1989.
- [18] C.E. McCulloch, Maximum likelihood variance components estimation for binary data, *J. Amer. Statist. Assoc.* 89 (1994) 330–335.
- [19] C.E. McCulloch, Maximum likelihood algorithms for generalized linear mixed models, *J. Amer. Statist. Assoc.* 92 (1997) 162–170.
- [20] J.A. Rice, *Mathematical Statistics and Data Analysis*, 2nd Edition, Duxbury, Belmont, 1995.
- [21] D.B. Rubin, *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York, 1987.
- [22] Z. Shun, Another look at the salamander mating data: a modified Laplace approximation approach, *J. Amer. Statist. Assoc.* 92 (1997) 341–349.
- [23] D. Spiegelhalter, A. Thomas, N. Best, Bayesian Inference using Gibbs Sampling for Window Version WinBUGS, a computer software for Bayesian analysis using MCMC method and Gibbs sampler and it can be downloaded from <http://www.mrc-bsu.cam.ac/bugs>, the website of the BUGS project by the Biostatistics Unit of the Medical Research Council, in the University of Cambridge, 2000.
- [24] S.L. Zeger, M.R. Karim, Generalized linear models with random effects: A Gibbs sampling approach, *J. Amer. Statist. Assoc.* 86 (1991) 79–86.