Quasi-Monte Carlo estimation in generalized linear mixed models

Pan, Jianxin and Thompson, Robin

2006

MIMS EPrint: **2007.169**

ISSN 1749-9097

Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 5765 – 5775
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Quasi-Monte Carlo estimation in generalized linear mixed models

Jianxin Pan[a,*], Robin Thompson[b]

[a]*School of Mathematics, University of Manchester, P.O. Box 88, Manchester M60 1QD, UK*
[b]*Biomathematics and Bioinformatics Division, Rothamsted Research, Harpenden, Hertfordshire AL5 2JQ, UK*

Received 10 January 2006; received in revised form 4 October 2006; accepted 5 October 2006
Available online 10 November 2006

Abstract

Generalized linear mixed models (GLMMs) are useful for modelling longitudinal and clustered data, but parameter estimation is very challenging because the likelihood may involve high-dimensional integrals that are analytically intractable. Gauss–Hermite quadrature (GHQ) approximation can be applied but is only suitable for low-dimensional random effects. Based on the Quasi-Monte Carlo (QMC) approximation, a heuristic approach is proposed to calculate the maximum likelihood estimates of parameters in the GLMM. The QMC points scattered uniformly on the high-dimensional integration domain are generated to replace the GHQ nodes. Compared to the GHQ approximation, the proposed method has many advantages such as its affordable computation, good approximation and fast convergence. Comparisons to the penalized quasi-likelihood estimation and Gibbs sampling are made using a real dataset and a simulation study. The real dataset is the salamander mating dataset whose modelling involves six 20-dimensional intractable integrals in the likelihood.

Keywords: Generalized linear mixed models; High-dimensional integrals; Maximum likelihood estimates; Quasi-Monte Carlo approximation; Salamander data analysis

1. Introduction

Generalized linear mixed models (GLMMs) have been widely used in the modelling of longitudinal and spatial data. The GLMM extends the generalized linear model (GLM) by incorporating random effects into the linear predictor to accommodate random variations and correlations from different sources. Nowadays it is well known that the GLMM is very useful in many statistical fields, for example, for clustered data and nonparametric penalized-spline smoothing.

Statistical inference of the GLMM, however, is highly hampered due to the incorporation of random effects. The likelihood function of the GLMM involves integrating out random effects from the joint density of the responses and random effects, which, except for a few cases such as a Gaussian density for the responses given random effects, is analytically intractable. In some circumstances the integration may be of very high dimension, making the analysis more considerably difficult. Existing estimation methods for the GLMM include: (a) analytically simplifying the problem, for example, by the use of Laplace approximation to the integrated likelihood, including the penalized quasi-likelihood (PQL) estimation (Breslow and Clayton, 1993) and the hierarchical generalized linear models (HGLM) procedure (Lee and Nelder, 2001); (b) using computation-intensive techniques such as the MCEM algorithm (Booth and Hobert, 1999),

* Corresponding author. Tel: +44 161 2755864; fax: +44 161 3063220.
E-mail addresses: jianxin.pan@manchester.ac.uk (J. Pan), robin.thompson@bbsrc.ac.uk (R. Thompson).

0167-9473/$ - see front matter
doi:10.1016/j.csda.2006.10.003

Markov Chain Monte Carlo (MCMC) (Zeger and Karim, 1991) and Gauss–Hermite quadrature (GHQ) approaches (Pan and Thompson, 2003).

However, the analytical simplification may not be always satisfactory. For example, when correlated data are binary and/or the designs are crossed, the use of PQL estimation may severely bias the estimates of the fixed effects and variance components of random effects. The GHQ approximation for the GLMM, on the other hand, is only useful for low-dimensional random effects since its computational loads increase exponentially with the dimension of random effects. The use of MCMC may also be very computationally intensive for the GLMM with multivariate random effects.

In computational mathematics, various heuristic optimization/approximation algorithms were developed in the past decade, such as Simulated Annealing, Threshold Accepting and the Quasi-Monte Carlo (QMC) approximation (Winker, 2000). Compared to the traditional Monte Carlo (MC) approach, heuristic algorithms have many advantages. For example, when applied to high-dimensional integral problems the QMC approach is superior to the traditional MC in terms of convergence rate and approximation accuracy. Application of heuristic algorithms to statistics, however, has lagged far behind, perhaps because not many statisticians are familiar with these algorithms.

In this paper, we propose to use the QMC approximation to solve the integration problem in the GLMM. In principle, the QMC approach generates integration nodes that are scattered uniformly on the integration domain. These nodes, called quasi-random numbers, tend to fill the space as fully as possible. The approach is also related to the computer space-filling technique (Bates et al., 1996). This paper is structured as follows. In Section 2, we briefly review the definition of the GLMM and highlight the problem of deriving the likelihood. In Section 3, we introduce the principle of the QMC approximation and discuss the generation of quasi-random numbers. In Section 4, the QMC approach is used to approximate the integrated likelihood and then calculate the maximum likelihood estimates (MLE) of fixed effects and variance components in the GLMM. In Section 5, the proposed approach is used to analyze the infamous salamander mating data, which involves six 20-dimensional integrals in the likelihood. In Section 6, simulation studies that have the same design protocol as the salamander data are conducted. Numerical comparisons are made to Gibbs sampling (Karim and Zeger, 1992) and the PQL estimation (Breslow and Clayton, 1993). In Section 7, discussions on some related issues are given. Technical details on the second-order derivatives of the log-likelihood are deferred to the Appendix.

2. Generalized linear mixed models

Suppose y_i $(i = 1, 2, \ldots, n)$ are the responses. Let x_i and z_i be $(p \times 1)$ and $(q \times 1)$ covariate vectors associated with fixed effects β $(p \times 1)$ and random effects b $(q \times 1)$, respectively. Given the random effects b, the responses y_i are assumed to be independent with expectation and variance:

$$E(y_i|b) = \mu_i \quad \text{and} \quad \text{var}(y_i|b) = \phi a_i^{-1} v(\mu_i), \tag{1}$$

where ϕ is scalar, a_i is the prior weight and $v(\cdot)$ is the variance function. The conditional expectation and variance have the form in (1) when the distribution of y_i given b comes from the quasi-likelihood family, in which a special case is the exponential family of distributions (McCullagh and Nelder, 1989). The definition of the GLMM is completed by introducing a monotone and differentiable function $g(\cdot)$ that links the conditional expectation μ_i to the linear predictor η_i through

$$g(\mu_i) = \eta_i \equiv x_i'\beta + z_i'b. \tag{2}$$

In the GLMM the random effects b are usually assumed to follow some distribution $F(\cdot)$ (e.g., Gaussian) with zero mean and variance–covariance matrix Σ, in other words, $b \sim F(0, \Sigma)$. The variance–covariance matrix Σ may depend on some unknown parameter vector θ $(m \times 1)$, i.e., $\Sigma \equiv \Sigma(\theta)$, known as variance components. In this paper we will focus on calculating the maximum likelihood estimates (MLE) of the fixed effects β and the variance components θ. Further statistical inferences including hypothesis test and model diagnostics can be made based on the estimates.

The integrated quasi-likelihood of (β, θ) must take the form

$$L(\beta, \theta) = \exp\{\ell(\beta, \theta)\} = \int \exp\left\{\sum_{i=1}^{n} \ell_i(\beta; b)\right\} \mathrm{d}F(b; \theta), \tag{3}$$

where

$$\ell_i(\beta; b) \propto \int_{y_i}^{\mu_i} \frac{a_i(y_i - u)}{\phi v(u)} \, du \tag{4}$$

defines the conditional log quasi-likelihood of the fixed effects β given the random effects b (Breslow and Clayton, 1993). The likelihood (3) is then maximized with respect to β and θ in order to obtain the MLE $\hat{\beta}$ and $\hat{\theta}$.

However, the likelihood $L(\beta, \theta)$ in (3) in general involves analytically intractable integrals. When the dimension q of the random effects b is low (e.g., $q < 5$), the GHQ approach is useful for approximating the likelihood (Pan and Thompson, 2003). As mentioned in Section 1, this technique is, however, no longer helpful for high-dimensional random effects. The Laplace approximation (Breslow and Clayton, 1993) and EM algorithm (Booth and Hobert, 1999) were proposed to calculate the MLEs of the parameters β and θ in the GLMM. MCMC was also used (Karim and Zeger, 1992).

3. Quasi-Monte Carlo integration

We propose to use the QMC approach to approximate the integrated quasi-likelihood $L(\beta, \theta)$ in (3). To gain an insight into the QMC approach, we first briefly review the traditional Monte Carlo (MC) approximation. Suppose $f(\cdot)$ is an integrable function on the q-dimensional unit cube $C^q = [0, 1)^q$. Consider the integral

$$I(f) = \int_{C^q} f(x) \, dx. \tag{5}$$

The MC integration draws a random sample $P_K = \{x_k : 1 \leqslant k \leqslant K\}$ from the uniform distribution on C^q and then approximates the integral in (5) through

$$\hat{I}_K(f, P_K) = \frac{1}{K} \sum_{k=1}^{K} f(x_k). \tag{6}$$

By the strong law of large numbers the estimate $\hat{I}_K(f, P_K)$ converges to $I(f)$ with probability one as $K \to \infty$. Moreover, the central limit theorem guarantees that $\hat{I}_K(f, P_K)$ is asymptotically normally distributed when the sample size K is large enough. The convergence rate for the MC integration is of the order $O(K^{-1/2})$, regardless of the integral dimensionality q. On the other hand, the previous statement regarding the convergence of the MC approximation is a probabilistic one, implying that the MC approximation in general may behave well but for a particular random sample may lead to a very poor approximation. Hence, it is necessary to make multiple draws of random samples and take the average of all the approximations, which may be computationally expensive.

The QMC approach aims to improve the MC approximation in terms of faster convergence rate and less computational load. The key idea is to use integration nodes that are scattered uniformly on C^q to replace the MC random samples. The reason behind this is due to the famous Koksma–Hlawka inequality:

$$|I(f) - \hat{I}_K(f, P_K)| \leqslant V(f) D(P_K), \tag{7}$$

where $V(f)$ is the bounded total variation of $f(\cdot)$ over C^q in the sense of Hardy and Krause (Fang and Wang, 1994, p. 64). The quantity $D(P_K)$ measures the evenness of spread of the point set P_K, defined by

$$D(P_K) = \sup_{x \in C^q} |U_K(x) - U(x)|, \tag{8}$$

where $U(x)$ is the uniform distribution on C^q and $U_K(x)$ is the empirical distribution of P_K. $D(P_K)$ in (8) is actually the Kolmogorov–Smirnov statistic but known as the discrepancy of P_K in analytical number theory. The inequality (7) implies that the absolute error of the integration approximation is bounded and dominated by $D(P_K)$ since $V(f)$ is a constant as far as $f(\cdot)$ is given. Therefore, the points P_K with the smallest discrepancy are the best integration nodes in this sense. It can be shown that the smallest discrepancy is of the order $O((\log K)^{q-1}/K)$ (Fang and Wang, 1994). Accordingly, the QMC integration has a faster convergence rate than the MC approximation. Unlike the MC approximation, the QMC integration nodes are deterministic so that multiple draws are not necessary.

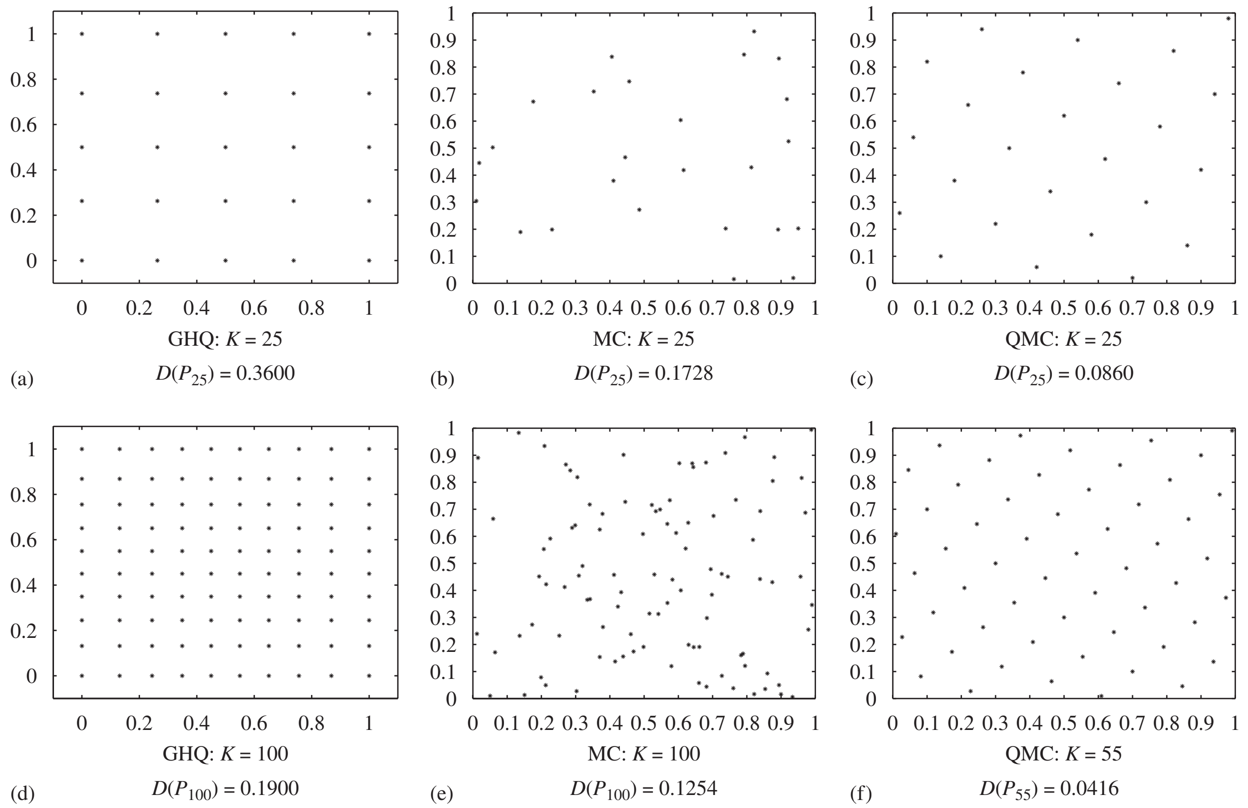

Fig. 1. Scatter plots of point sets. (a) and (d): confined GHQ points with $K = 25$ and $K = 100$, (b) and (e): MC points with $K = 25$ and $K = 100$, and (c) and (f): QMC points with generating vectors $(K; h_1, h_2) = (25;1,7)$ and $(K; h_1, h_2) = (55;1,34)$.

When $q = 1$, it is easy to generate the point set P_K that has the smallest discrepancy. In fact, it must be of the form $P_K = \{(2i - 1)/2K : i = 1, 2, \ldots, K\}$ and the smallest discrepancy is given by $D(P_K) = 1/2K$ (Fang and Wang, 1994). When $q \geqslant 2$, no analytical generating formula for P_K becomes available but there are a variety of methods available for generating the quasi-random numbers P_K with discrepancy close to the optimal order $\mathrm{O}((\log K)^{q-1}/K)$. For example, the so-called good lattice points (GLP) are one of those. The GLP quasi-random numbers can be generated as follows. Let $(K; h_1, h_2, \ldots, h_q)$ be a vector with integer components satisfying: $1 \leqslant h_j < K$ and $h_j \neq h_l$ $(j \neq l)$ where $q < K$. Suppose the greatest common divisor $(K, h_j) = 1$ $(j = 1, 2, \ldots, q)$. Denote $q_{kj} \equiv kh_j$ (modulo K) and $x_{kj} = (2q_{kj} - 1)/2K$ $(k = 1, 2, \ldots, K$ and $j = 1, 2, \ldots, q)$, where the operation modulo confines q_{kj} to the range between 1 and K. The point set $\{P_K = x_k = (x_{k1}, \ldots, x_{kq}) : k = 1, 2, \ldots, K\}$ is known as lattice points with the generating vector $(K; h_1, h_2, \ldots, h_q)$. If $(h_1, h_2, \ldots, h_q)$ is chosen such that the discrepancy of P_K reaches the minimum, then P_K is the GLP set. Fang and Wang (1994) gave some methods for finding the best generating vectors and tabulated their results for different numbers q and K.

For illustration, Fig. 1 gives two-dimensional scatter plots for the GHQ, MC, and GLP points with different K's. For comparison, their associated discrepancy values are also given underneath of the plots. Note the GHQ points have been scaled down here to the unit square $C^2 = [0, 1)^2$ in order to compare with the MC and QMC points on C^2. The QMC points in (c) and (f) are the GLP sets with generating vectors $(K; h_1, h_2) = (25; 1, 7)$ and $(K; h_1, h_2) = (55; 1, 34)$, respectively. Fig. 1 shows that the QMC points have the smallest discrepancy values, whereas the discrepancy of the GHQ points is the largest. At the first glance, however, it seems that the GHQ points in the Figs. 1(a) and (d) are as evenly spread as the QMC points in the Figs. 1(c) and (f) and so their discrepancy values should be quite close. This actually is not true because theoretical studies show that the discrepancy of the GHQ points is of the order $\mathrm{O}(K^{-1/q})$ (Fang and Wang, 1994, pp. 18–19). This is also the reason why the GHQ points are not recommended for the use in high-dimensional integration problems, in addition to having a computational intensive problem.

Table 1
Errors in $p - \hat{p}$ using the GHQ, MC and QMC approximations

K	GHQ	K	MC	K	QMC
64	−2.215e − 04	35	−2.121e−03	35	−1.162e −04
125	−1.415e − 04	101	1.937e −03	101	5.380e−05
729	−4.358e − 05	597	6.349e−05	597	−3.979e−06
1728	−2.451e − 05	1626	−6.560e −05	1626	6.688e−06
5832	−1.090e − 05	5037	−3.248e −05	5037	2.632e−07
39304	−3.072e − 06	39029	1.342e−05	39029	2.438e−09

To see how the QMC approximation improves the MC and GHQ approaches, we consider a simple example below. Let $X \sim N_3(0, I_3)$ and we want to calculate the three-dimensional integral

$$p = \int_0^1 \int_0^1 \int_0^1 (2\pi)^{-3/2} \exp\left\{-\frac{1}{2}(x_1^2 + x_2^2 + x_3^2)\right\} \mathrm{d}x_1 \, \mathrm{d}x_2 \, \mathrm{d}x_3.$$

Obviously, the integral is the probability $p = [\Phi(1) - \Phi(0)]^3 = 0.039772181953\ldots$, where $\Phi(\cdot)$ is the CDF of $N(0, 1)$. We generate the GHQ, MC and QMC/GLP points with different numbers of K and use (6) to approximate the integral. The resulting errors in $p - \hat{p}$ are listed in Table 1. Obviously, the QMC approximation outperforms the GHQ and MC approaches. This example was studied by Fang and Wang (1994, pp. 65–66).

4. Quasi-Monte Carlo estimation

Suppose $P_K = \{c_k : k = 1, \ldots, K\}$ is a QMC point set over the unit cube C^q. When P_K is used to approximate the integrated likelihood (3), the QMC approximated log-likelihood is of the form

$$\ell(\beta, \theta) = \log\left[\frac{1}{K}\sum_{k=1}^{K} \exp\left\{\sum_{i=1}^{n} \ell_i(\beta, \Sigma^{1/2} F^{-1}(c_k))\right\}\right], \tag{9}$$

where $F^{-1}(\cdot)$ is the inverse of the CDF $F(\cdot)$ and $\Sigma^{1/2}$ is the square root of Σ, for example, it can be taken as the Cholesky factor of Σ.

Let $b_k = F^{-1}(c_k)$, $\eta_{ik} = x_i'\beta + z_i'\Sigma^{1/2} b_k$ and $\mu_{ik} = h(\eta_{ik})$ where $h(\cdot)$ is the inverse of $g(\cdot)$, i.e., $h(\cdot) = g^{-1}(\cdot)$. Then the MLE $\hat{\beta}$ of β must be the solution of the score equations:

$$\dot{\ell}_\beta \equiv \frac{\partial}{\partial \beta}\{\ell(\beta, \theta)\} = \sum_{k=1}^{K} w_k \left\{\sum_{i=1}^{n} \frac{a_i(y_i - h(\eta_{ik}))}{\phi v(\mu_{ik}) \dot{g}(\mu_{ik})} x_i\right\} = 0, \tag{10}$$

where $\dot{g}(\cdot)$ is the derivative of $g(\cdot)$ and w_k is the weight given by

$$w_k = \frac{\exp\{\sum_{i=1}^{n} \ell_i(\beta, \Sigma^{1/2} c_k)\}}{\sum_{k=1}^{K} \exp\{\sum_{i=1}^{n} \ell_i(\beta, \Sigma^{1/2} c_k)\}}. \tag{11}$$

Note that the weight w_k above also depends on the unknown parameters β and θ, i.e., $w_k \equiv w_k(\beta, \theta)$. This has to be taken into account when calculating the second-order derivatives.

Similarly, the score equation for the jth variance component of θ must have the form

$$\dot{\ell}_{\theta_j} \equiv \frac{\partial}{\partial \theta_j}\{\ell(\beta, \theta)\} = \sum_{k=1}^{K} w_k \left\{\sum_{i=1}^{n} \frac{a_i(y_i - h(\eta_{ik}))}{\phi v(\mu_{ik}) \dot{g}(\mu_{ik})} z_i' \dot{\Sigma}_j^{1/2} b_k\right\} = 0, \tag{12}$$

where $\dot{\Sigma}_j^{1/2}=\partial\Sigma^{1/2}/\partial\theta_j$ $(j=1,2,\ldots,m)$. When the design of the GLMM is crossed, in other words, $\Sigma=\mathrm{diag}(\theta_1 I_{q_1},\ldots,\theta_m I_{q_m})$ or equivalently $\Sigma^{1/2}=\mathrm{diag}(\theta_1^{1/2}I_{q_1},\ldots,\theta_m^{1/2}I_{q_m})$ we then have

$$\dot{\Sigma}_j^{1/2}=\frac{\partial\Sigma^{1/2}}{\partial\theta_j}=\mathrm{diag}\left(0,\ldots,0,\frac{1}{2}\theta_j^{-1/2}I_{q_j},0,\ldots,0\right), \tag{13}$$

where $q=\sum_{j=1}^m q_j$.

The score equations (10) and (12) in general have no analytical solution. We then use the Newton–Raphson algorithm to solve these equations. The second-order derivatives of the QMC approximated log-likelihood (9) with respect to β and θ need to be calculated, which, though complicated, have analytically explicit forms. For ease of presentation, the technical expressions of these derivatives $\ddot{\ell}_{\beta\beta}$, $\ddot{\ell}_{\beta\theta}$ and $\ddot{\ell}_{\theta\theta}$ are deferred to the Appendix. Once an initial value of (β,θ) is given, the solution to the score equations (10) and (12) can be updated using

$$\begin{pmatrix}\beta\\ \theta\end{pmatrix}_{\text{new}}=\begin{pmatrix}\beta\\ \theta\end{pmatrix}_{\text{old}}+\begin{pmatrix}-\ddot{\ell}_{\beta\beta} & -\ddot{\ell}_{\beta\theta}\\ -\ddot{\ell}'_{\beta\theta} & -\ddot{\ell}_{\theta\theta}\end{pmatrix}^{-1}_{\text{old}}\begin{pmatrix}\dot{\ell}_{\beta}\\ \dot{\ell}_{\theta}\end{pmatrix}_{\text{old}}, \tag{14}$$

where $\dot{\ell}_\theta=(\dot{\ell}_{\theta_1},\ldots,\dot{\ell}_{\theta_m})'$. The above process is iterated until convergence of β and θ. At convergence, the MLEs $\hat{\beta}$ and $\hat{\theta}$ of the fixed effects and variance components are obtained. The asymptotic variance–covariance matrix of the MLEs can be approximated by calculating the inverse of the observed information matrix, evaluated at the MLEs $\hat{\beta}$ and $\hat{\theta}$.

It is noted that in some circumstances the prediction of the random effects b may be of interest. Once the MLEs $\hat{\beta}$ and $\hat{\Sigma}$ are obtained, the prediction of random effects can be updated using

$$\hat{b}=\hat{\Sigma}Z'\hat{V}^{-1}(\tilde{Y}-X\hat{\beta}), \tag{15}$$

where $\tilde{Y}=(\tilde{y}_1,\ldots,\tilde{y}_n)'$ with $\tilde{y}_i=\hat{\eta}_i+(y_i-\hat{\mu}_i)\dot{g}(\hat{\mu}_i)$ where $\hat{\eta}_i=x_i'\hat{\beta}+z_i'\hat{b}_i$ and $\hat{\mu}_i=g^{-1}(\hat{\eta}_i)$. The matrices Z, X and $\hat{V}$ are given by $Z=(z_1',\ldots,z_n')'$, $X=(x_1',\ldots,x_n')'$and $\hat{V}=\hat{W}^{-1}+Z\hat{\Sigma}Z'$ where $\hat{W}$ is the $(n\times n)$ diagonal matrix with the entries $\hat{\omega}_i=[\phi a_i v(\hat{\mu}_i)\{\dot{g}(\hat{\mu}_i)\}^2]^{-1}$. It is noted that the RHS of (15) also depends on $\hat{b}$, implying that an initial value of b substituted into the RHS of (15) yields a new prediction of b. This process is then iterated until the convergence of $\hat{b}$. It can be shown that the solution $\hat{b}$ actually provides the mode of the conditional distribution of the posterior density of the random effects, see Breslow and Clayton (1993) for more details.

5. Salamander mating data analysis

The salamander mating experiment (McCullagh and Nelder, 1989) involves two populations of salamanders: Rough Butt (RB) and Whiteside (WS). Ten males and 10 females from each population were mated in a crossed design, see Shun (1997) for the detailed design table. In a fixed period each animal had six matings with her/his opposite genders, three from her/his own population and three from the other. The experiment resulted in 120 correlated binary observations, depending on whether or not the mating was success. During the summer and autumn of 1986, the experiment was repeated three times. The primary objective of the experiment was to study whether or not the mating of salamanders between populations is as successful as within population. The secondary objective was to study if heterogeneity between individuals in the mating probability exists.

A GLMM is used to model the correlated binary responses resulting from each of the three experiments:

$$\mathrm{logit}\{E(y_{ij}|b_i^f,b_j^m)\}=x_{ij}'\beta+b_i^f+b_j^m, \tag{16}$$

where b_i^f and b_j^m are random effects from the female and male individuals in the pair and are assumed independent with $b_i^f\sim N(0,\sigma_f^2)$ and $b_j^m\sim N(0,\sigma_m^2)$ $(i,j=1,\ldots,20)$. The covariate vector x_{ij} is defined by $x_{ij}=(1,\mathrm{WS}_i^f,\mathrm{WS}_j^m,\mathrm{WS}_{ij}^{fm})'$ where WS_i^f is the indicator variable for the WS female (0 for RB and 1 for WS), WS_i^m is the variable for WS male (0 for RB and 1 for WS), and WS_{ij}^{fm} represents their interaction.

For each experiment the log-likelihood of the model (16) involves a 40-dimensional integral that can be further reduced to the sum of two 20-dimensional integrals because of the design. For example, the first five females of the first population were only mated with the first five males of each population (McCullagh and Nelder, 1989). In other words, $\ell = \ell_1 + \ell_2$ with ℓ_1 being equal to

$$\int \prod_{i=1}^{5} \left(\int \prod_{j=1}^{5} [y_{ij}|b_i^f, b_j^m][b_j^m] \prod_{j=11}^{15} [y_{ij}|b_i^f, b_j^m][b_j^m]\, \mathrm{d}b_1^m \ldots \mathrm{d}b_5^m\, \mathrm{d}b_{11}^m \ldots \mathrm{d}b_{15}^m \right) [b_i^f]\, \mathrm{d}b_1^f \ldots \mathrm{d}b_5^f$$
$$\prod_{i=16}^{20} \left(\int \prod_{j=1}^{5} [y_{ij}|b_i^f, b_j^m][b_j^m] \prod_{j=11}^{15} [y_{ij}|b_i^f, b_j^m][b_j^m]\, \mathrm{d}b_1^m \ldots \mathrm{d}b_5^m\, \mathrm{d}b_{11}^m \ldots \mathrm{d}b_{15}^m \right) [b_i^f]\, \mathrm{d}b_{16}^f \ldots \mathrm{d}b_{20}^f, \tag{17}$$

and ℓ_2 to

$$\int \prod_{j=6}^{15} \left(\int \left\{ \prod_{j=6}^{10} [y_{ij}|b_i^f, b_j^m][b_j^m] \right\} \left\{ \prod_{j=16}^{20} [y_{ij}|b_i^f, b_j^m][b_j^m] \right\} \mathrm{d}b_6^m \ldots \mathrm{d}b_{10}^m\, \mathrm{d}b_{16}^m \ldots \mathrm{d}b_{20}^m \right) [b_i^f]\, \mathrm{d}b_6^f \ldots \mathrm{d}b_{15}^f, \tag{18}$$

where $[y_{ij}|b_i^f, b_j^m]$ is the conditional probability mass function of y_{ij} given b_i^f and b_j^m, and $[b_i^f]$ and $[b_j^m]$ are the densities of b_i^f and b_j^m, respectively. That is,

$$[y_{ij}|b_i^f, b_j^m] = \frac{\exp\{y_{ij}\eta_{ij}\}}{1 + \exp\{y_{ij}\eta_{ij}\}},$$

where $\eta_{ij} = x_{ij}'\beta + b_i^f + b_j^m$, and

$$[b_i^f] = (2\pi\sigma_f^2)^{-1/2} \exp\{-b_i^{f2}/2\sigma_f^2\} \quad \text{and} \quad [b_j^m] = (2\pi\sigma_m^2)^{-1/2} \exp\{-b_j^{m2}/2\sigma_m^2\}.$$

Obviously, the integrals (17) and (18) are analytically intractable. When pooling the three experiment responses together, the log-likelihood becomes the sum of six 20-dimensional integrals and then the MLEs of the fixed effects β and the variance components σ_f^2 and σ_m^2 are extremely difficult to obtain. In the literature, the PQL and MCMC approaches were used to analyze the salamander data by Breslow and Clayton (1993) and Karim and Zeger (1992), respectively, but suffer from either being very biased or computationally intensive.

We now use the QMC approximation to calculate the MLEs of the fixed effects and variance components for modelling of the pooled data. We note that each of the six integrals is 20-dimensional but currently existing GLP sets are available only up to 18 dimensions (Fang and Wang, 1994). We then propose to generate quasi-random numbers on the unit cube $C^{20} = [0, 1)^{20}$ by replacing the generating vector $(h_1, \ldots, h_{20})'$ of the GLP method with the square roots of the first 20 prime numbers. The resulting quasi-random numbers have the discrepancy close to the optimal rate (Fang and Wang, 1994). Table 2 summarizes our numerical results for various sizes of the QMC quasi-random numbers. For comparison, in Table 2 we also reproduce Karim and Zeger's (1992) Gibbs sampling and Breslow and Clayton's (1993) PQL results.

From Table 2 we see that when increasing the number of the QMC integration nodes from 10,000 to 100,000 the maximized log-likelihood $\hat{\ell}_{\max}$ changes very little and the parameter estimates becomes stable quickly. It implies that any number of QMC points between 10,000 and 100,000 could produce reasonable estimates of the parameters. Note the integration approximation takes place on the 20-dimensional unit cube. The selected numbers of the QMC points, even up to $K = 100,000$, are still small in the 20-dimensional space but the approximation performs very well in terms of accuracy and stabilization. When using 100,000 QMC points and working on a Pentium(R) 4 PC (CPU 3.20 GHz) our Fortran code takes about 5 min to converge. In contrast, the use of the GLMMGibbs package in R for modelling of the pooling salamander data takes a few hours to obtain the similar results, where 1000 iterations of burn-in followed by 20,000 saved iterations are made.

Compared to the literature work, the proposed QMC approach produces numerical results close to those by Gibbs sampling. The latter involves conditional distribution specification and multiple draw of random samples, however. In contrast, the QMC method is easy to use for practitioners and is fast for obtaining the parameter estimates. The PQL approach, as criticized by many authors, results in very biased estimates for both fixed effects and variance components.

Table 2
MLEs of the parameters for modelling the pooling salamander data when varying K, the number of the QMC points (standard errors in parentheses)

K	β_0	β_1	β_2	β_3	σ_f	σ_m	$\hat{\ell}_{\max}$
10,000	0.92(.38)	−2.83(.51)	−0.58(.41)	3.57(.63)	1.11(.28)	0.98(.20)	−207.21
20,000	0.83(.37)	−2.80(.52)	−0.53(.44)	3.51(.61)	1.06(.23)	1.02(.23)	−207.70
30,000	1.28(.41)	−2.88(.54)	−0.99(.50)	3.64(.63)	1.25(.27)	1.16(.24)	−205.67
40,000	1.22(.41)	−2.83(.53)	−0.99(.49)	3.66(.63)	1.28(.28)	1.21(.26)	−206.19
50,000	1.21(.40)	−2.81(.53)	−1.03(.49)	3.70(.62)	1.25(.24)	1.24(.26)	−206.07
60,000	1.17(.39)	−2.80(.53)	−0.99(.49)	3.67(.63)	1.23(.24)	1.20(.26)	−206.41
70,000	1.21(.37)	−2.81(.53)	−0.96(.47)	3.68(.63)	1.30(.29)	1.22(.26)	−206.31
80,000	1.22(.38)	−2.86(.54)	−1.01(.49)	3.71(.64)	1.30(.29)	1.24(.26)	−206.35
90,000	1.21(.38)	−2.87(.54)	−0.99(.49)	3.69(.64)	1.28(.29)	1.22(.26)	−206.66
100,000	1.22(.39)	−2.91(.56)	−0.98(.49)	3.67(.64)	1.26(.29)	1.23(.27)	−206.83
Gibbs[a]	1.03(.43)	3.01(.60)	−0.69(.50)	3.74(.68)	1.22	1.17	—
PQL[b]	0.79(.32)	−2.29(.43)	−0.54(.39)	2.82(.50)	0.85	0.79	—

[a]by Karim and Zeger (1992).
[b]by Breslow and Clayton (1993).

Table 3
Parameter estimates in the simulation study, where the QMC approximation uses $K = 50,000$ integration nodes (simulated standard errors are given in parentheses)

	β_0	β_1	β_2	β_3	σ_f	σ_m
True	1.21	−2.81	−1.03	3.70	1.25	1.24
QMC	1.19(.39)	−2.92(.55)	−0.98(.44)	3.61(.63)	1.29(.27)	1.26(.26)
PQL	0.84(.35)	−2.40(.46)	−0.56(.38)	3.07(.45)	0.86(.25)	0.81(.23)

6. Simulation study

Based on the logistic model in (16), we simulate the salamander polling data that comprise 360 correlated binary observations. In the simulation study we run 1000 simulations. Similar to the real salamander data, the log-likelihood function for each simulated dataset involves six 20-dimensional integrals that are analytically intractable. We generate $K = 50,000$ QMC integration nodes on the cube C^{20} using the square roots of the first 20 prime numbers and then use (6) to approximate the integrated log-likelihood. The Newton–Raphson algorithm (14) is used to maximize the QMC approximated log-likelihood function.

The true values of the fixed effects $\beta = (\beta_1, \ldots, \beta_4)'$ and the variance components $\sigma^2 = (\sigma_f^2, \sigma_m^2)$ are chosen to be the MLEs obtained in the real data analysis using $K = 50,000$. For comparison, we also use the PQL approach to analyze the simulated binary data. Table 3 gives the average of the 1000 estimates for each parameter.

From Table 3 it is clear that the QMC approach produces satisfactory results. In contrast, the PQL procedure results in considerably biased estimates. In particular, it underestimates considerably the variance components of random effects. This phenomenon was noticed by other authors as well (e.g., Lin and Breslow, 1996).

We point out that, the proposed QMC-based Newton–Raphson algorithm may happen to converge to the boundary of variance components, that is, $\sigma_f^2 = 0$ or $\sigma_m^2 = 0$. In our 1000 simulations, we experienced 7 such cases, see Fig. 2. With those cases removed, the average of the remaining estimates of variance components may increase slightly, leading to slightly biased estimates of the variance components.

7. Discussion

In this paper we propose a heuristic approach, the QMC approximation to calculate the MLEs of the fixed effects and variance components in the GLMM. The salamander mating data analysis and the associated simulation study, involving six 20-dimensional analytically intractable integrals in the likelihood, show that the proposed QMC approach works reasonably well.

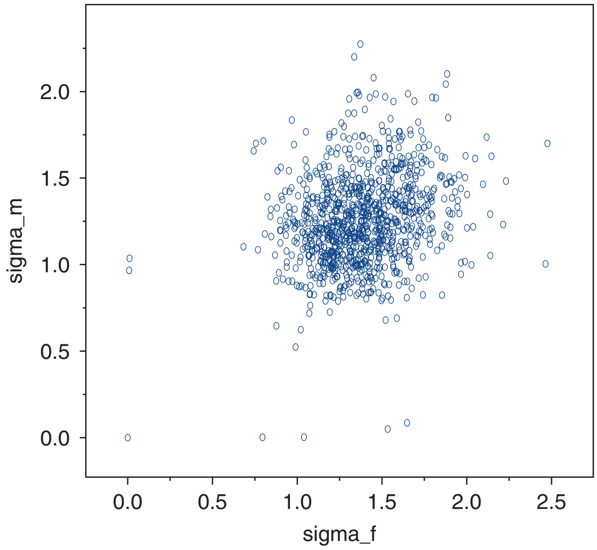

Fig. 2. Scatter plot of the estimates of variance components σ_m against σ_f in the 1000 simulation studies.

The key idea of the QMC approximation is to choose integration nodes that are scattered uniformly on the high-dimensional unit cube. In the literature, there were various methods proposed for generating the QMC points, including points generated by square roots of primes, good lattice points, Halton sequences, Faure sequences, and Sobol sequences among others. In this paper, we use the simplest one, square roots of primes, to generate 20-dimensional QMC points on the cube C^{20} and then approximate the log-likelihood for modelling of the salamander mating data. The numerical analysis shows that this simplest method works well. Recently, Al-Eid and Pan (2006) implemented more uniform QMC points including the Faure sequences and Sobol sequences into the QMC-based Newton–Raphson algorithm. They show that good QMC nodes do further improve the approximation.

Whatever the QMC points are chosen, a practical issue is on the selection of the number of the points. As a general rule, we would suggest to increase the number of the QMC points gradually until the resulting MLEs become stable, see Table 2 for example. A concern is that, the need to run the QMC algorithm repeatedly like this obviously increases the computational demands of the QMC approximation. Given that the QMC algorithm converges very rapidly, for example, it only takes about 5 min for the modelling of the salamander mating data when using $K = 100,000$ QMC points in 20-dimensional space, this issue is not problematic.

Another concern of the use of heuristic approach is that, no automatic random variation appears in the quasi-random numbers. As a result, there is no practical way of assessing the accuracy of parameter estimates in some circumstances. Owen (1998) proposed to randomize the QMC points and the resulting randomized points still keep the same discrepancy order in terms of uniformity. Randomized QMC (RQMC) approach, also called scrambled QMC approximation, combines the strength of the QMC and traditional MC methods. In such ways the variances of the parameter estimates can be estimated by taking a small number of independent replications in the RQMC approach. On the other hand, the RQMC method can also be useful for enhancing the performance of ordinary QMC applications. Typical RQMC points include the scrambled digital points and scrambled lattice points. One can refer to Owen (1998) for more details.

We also compare the QMC approximation with Gibbs sampling through modelling of the salamander mating data. We find that these two approaches yield very similar results but the QMC method is much faster than Gibbs sampling computationally. Also, the QMC approach is easy to use for practitioners, provided that the previous advice regarding the implementation of QMC points is acknowledged. In contrast, Gibbs sampling requires more experience such as in specifying conditional distributions and setting priors. Also, the speed of Gibbs sampling for models with high-dimensional random effects may be problematic. The PQL estimation approach, on the other hand, produces very biased estimates of the fixed effects and variance components when applied to modelling of correlated binary data.

It is noted that the HGLM approach proposed by Lee and Nelder (2001, 2006) also modeled the salamander mating data well, where the conjugate hierarchical generalized linear model, beta-binomial HGLM, was applied so that numerical integration is avoided. In contrast, the QMC approach proposed in this paper does not request a conjugate distribution assumption for responses and random effects, though the normality of random effects is assumed here. The issue of non-Normally distributed random effects will be addressed in our follow-up work.

Acknowledgements

The authors are grateful for many constructive suggestions from two referees, the associated editor and the co-editor. The authors also thank Dr Huajun Ye and Mr Eid Al-Eid for their help on the computational programming.

Appendix. The second-order derivatives of the log-likelihood

First, the second-order derivative of $\ell(\beta, \theta)$ with respect to β can be obtained, by taking derivative of $\dot{\ell}_\beta$ in (10) with respect to β, as follows:

$$
\begin{aligned}
\ddot{\ell}_{\beta\beta} &\equiv \frac{\partial^2}{\partial\beta\partial\beta'}\{\ell(\beta,\theta)\} \\
&= -\sum_{k=1}^{K} w_k \left\{\sum_{i=1}^{n} \frac{a_i(y_i - h(\eta_{ik}))}{\phi v(\mu_{ik})\dot{g}(\mu_{ik})} x_i\right\}\left\{\sum_{i=1}^{n} \frac{a_i(y_i - h(\eta_{ik}))}{\phi v(\mu_{ik})\dot{g}(\mu_{ik})} x_i'\right\} \\
&\quad + \left[\sum_{k=1}^{K} w_k \left\{\sum_{i=1}^{n} \frac{a_i(y_i - h(\eta_{ik}))}{\phi v(\mu_{ik})\dot{g}(\mu_{ik})} x_i\right\}\right]\left[\sum_{k=1}^{K} w_k \left\{\sum_{i=1}^{n} \frac{a_i(y_i - h(\eta_{ik}))}{\phi v(\mu_{ik})\dot{g}(\mu_{ik})} x_i'\right\}\right] \\
&\quad + \sum_{k=1}^{K} w_k \sum_{i=1}^{n} \left(\frac{a_i}{\phi}\right)[v(\mu_{ik})\{\ddot{g}(\mu_{ik})\}^2]^{-1} x_i x_i' \\
&\quad + \sum_{k=1}^{K} w_k \sum_{i=1}^{n} \left(\frac{a_i}{\phi}\right)\left[\frac{\dot{v}(\mu_{ik})\dot{g}(\mu_{ik}) + v(\mu_{ik})\ddot{g}(\mu_{ik})}{\{v(\mu_{ik})\dot{g}(\mu_{ik})\}^2\dot{g}(\mu_{ik})}\right](y_i - h(\eta_{ik}))x_i x_i',
\end{aligned}
$$

where $\ddot{g}(\cdot)$ is the second derivative of the link function $g(\cdot)$. When calculating the derivative of $\dot{\ell}_\beta$, the dependence of the weight w_k on β and θ needs to be taken into account. This is why the above expression becomes complicated.

Similarly, the (j, j')th element of the second-order derivative $\ddot{\ell}_{\theta\theta}$ can be obtained as

$$
\begin{aligned}
\ddot{\ell}_{\theta_j\theta_{j'}} &\equiv \frac{\partial}{\partial\theta_{j'}}\left[\frac{\partial}{\partial\theta_j}\{\ell(\beta,\theta)\}\right] \\
&= -\sum_{k=1}^{K} w_k \left\{\sum_{i=1}^{n} \frac{a_i(y_i - h(\eta_{ik}))}{\phi v(\mu_{ik})\dot{g}(\mu_{ik})} z_i'\dot{\Sigma}_j^{1/2} b_k\right\}\left\{\sum_{i=1}^{n} \frac{a_i(y_i - h(\eta_{ik}))}{\phi v(\mu_{ik})\dot{g}(\mu_{ik})} z_i'\dot{\Sigma}_{j'}^{1/2} b_k\right\} \\
&\quad + \left[\sum_{k=1}^{K} w_k \left\{\sum_{i=1}^{n} \frac{a_i(y_i - h(\eta_{ik}))}{\phi v(\mu_{ik})\dot{g}(\mu_{ik})} z_i'\dot{\Sigma}_j^{1/2} b_k\right\}\right]\left[\sum_{k=1}^{K} w_k \left\{\sum_{i=1}^{n} \frac{a_i(y_i - h(\eta_{ik}))}{\phi v(\mu_{ik})\dot{g}(\mu_{ik})} z_i'\dot{\Sigma}_{j'}^{1/2} b_k\right\}\right] \\
&\quad + \sum_{k=1}^{K} w_k \sum_{i=1}^{n} \left(\frac{a_i}{\phi}\right)[v(\mu_{ik})\{\ddot{g}(\mu_{ik})\}^2]^{-1} b_k'\dot{\Sigma}_j^{1/2} z_i z_i' \dot{\Sigma}_{j'}^{1/2} b_k \\
&\quad + \sum_{k=1}^{K} w_k \sum_{i=1}^{n} \left(\frac{a_i}{\phi}\right)\left[\frac{\dot{v}(\mu_{ik})\dot{g}(\mu_{ik}) + v(\mu_{ik})\ddot{g}(\mu_{ik})}{\{v(\mu_{ik})\dot{g}(\mu_{ik})\}^2\dot{g}(\mu_{ik})}\right](y_i - h(\eta_{ik}))b_k'\dot{\Sigma}_j^{1/2} z_i z_i' \dot{\Sigma}_{j'}^{1/2} b_k,
\end{aligned}
$$

where $j, j' = 1, 2, \ldots, m$. Finally, the jth column of the $(p \times m)$ second-order derivative matrix $\ddot{\ell}_{\beta\theta}$ must be of the form

$$\begin{aligned}
\ddot{\ell}_{\beta\theta_j} &\equiv \frac{\partial}{\partial\theta_j}\left[\frac{\partial}{\partial\beta}\{\ell(\beta,\theta)\}\right] \\
&= -\sum_{k=1}^{K} w_k \left\{\sum_{i=1}^{n} \frac{a_i(y_i - h(\eta_{ik}))}{\phi v(\mu_{ik})\dot{g}(\mu_{ik})} x_i\right\}\left\{\sum_{i=1}^{n} \frac{a_i(y_i - h(\eta_{ik}))}{\phi v(\mu_{ik})\dot{g}(\mu_{ik})} z_i' \dot{\Sigma}_{j'}^{1/2} b_k\right\} \\
&\quad + \left[\sum_{k=1}^{K} w_k \left\{\sum_{i=1}^{n} \frac{a_i(y_i - h(\eta_{ik}))}{\phi v(\mu_{ik})\dot{g}(\mu_{ik})} x_i\right\}\right]\left[\sum_{k=1}^{K} w_k \left\{\sum_{i=1}^{n} \frac{a_i(y_i - h(\eta_{ik}))}{\phi v(\mu_{ik})\dot{g}(\mu_{ik})} z_i' \dot{\Sigma}_{j'}^{1/2} b_k\right\}\right] \\
&\quad + \sum_{k=1}^{K} w_k \sum_{i=1}^{n} \left(\frac{a_i}{\phi}\right) [v(\mu_{ik})\{\dot{g}(\mu_{ik})\}^2]^{-1} x_i z_i' \dot{\Sigma}_{j'}^{1/2} b_k \\
&\quad + \sum_{k=1}^{K} w_k \sum_{i=1}^{n} \left(\frac{a_i}{\phi}\right) \left[\frac{\dot{v}(\mu_{ik})\dot{g}(\mu_{ik}) + v(\mu_{ik})\ddot{g}(\mu_{ik})}{\{v(\mu_{ik})\dot{g}(\mu_{ik})\}^2 \dot{g}(\mu_{ik})}\right] (y_i - h(\eta_{ik})) x_i z_i' \dot{\Sigma}_{j'}^{1/2} b_k,
\end{aligned}$$

where $j = 1, 2, \ldots, m$.

References

Al-Eid, E., Pan, J., 2006. Estimation in generalized linear mixed models using randomized Quasi-Monte Carlo approximation. Preprint.

Bates, R.A., Buck, R.J., Riccomagno, E., Wynn, H.P., 1996. Experimental design and observation for large systems. J. Roy. Statist. Soc. Ser. B 58, 77–94.

Booth, J.G., Hobert, J.P., 1999. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. J. Roy. Statist. Soc. Ser. B 61, 265–285.

Breslow, N.E., Clayton, D.G., 1993. Approximate inference in generalized linear mixed models. J. Amer. Statist. Assoc. 88, 9–25.

Fang, K.T., Wang, Y., 1994. Number-Theoretic Methods in Statistics. Chapman & Hall, London.

Karim, M.R., Zeger, S.L., 1992. Generalized linear models with random effects: salamander mating revisited. Biometrics 48, 681–694.

Lee, Y., Nelder, J.A., 2001. Hierarchical generalized linear models: a synthesis of generalized linear models, random effect models and structured dispersions. Biometrika 88, 987–1006.

Lee, Y., Nelder, J.A., 2006. Double hierarchical generalized linear models (with discussion). J. Roy. Statist. Soc. Ser. C 55, 139–185.

Lin, X., Breslow, N.E., 1996. Bias correction in generalized linear mixed models with multiple components of dispersion. J. Amer. Statist. Assoc. 91, 1007–1016.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. second ed. Chapman and Hall, London.

Owen, A.B., 1998. Latin supercubic sampling for very high dimensional simulations. ACM Trans. Model. Comput. Simulation 8, 71–102.

Pan, J., Thompson, R., 2003. Gauss-Hermite quadrature approximation for estimation in generalized linear mixed models. Comput. Statist. 18, 57–78.

Shun, Z., 1997. Another look at the salamander mating data: a modified Laplace approximation approach. J. Amer. Statist. Assoc. 92, 341–349.

Winker, P., 2000. Optimization Heuristics in Econometrics: Applications of Threshold Accepting. Wiley, New York.

Zeger, S.L., Karim, M.R., 1991. Generalized linear models with random effects: a Gibbs sampling approach. J. Amer. Statist. Assoc. 86, 79–86.