Statistics and Computing (1997) **7**, 57–68

Sampling from the posterior distribution in generalized linear mixed models

DANI GAMERMAN

Instituto de Matemática, Universidade Federal do Rio de Janeiro, Caixa Postal 68530, 21945–970 Rio de Janeiro, RJ, Brazil

Submitted February 1996 and accepted October 1996

Generalized linear mixed models provide a unified framework for treatment of exponential family regression models, overdispersed data and longitudinal studies. These problems typically involve the presence of random effects and this paper presents a new methodology for making Bayesian inference about them. The approach is simulation-based and involves the use of Markov chain Monte Carlo techniques. The usual iterative weighted least squares algorithm is extended to include a sampling step based on the Metropolis–Hastings algorithm thus providing a unified iterative scheme. Non-normal prior distributions for the regression coefficients and for the random effects distribution are considered. Random effect structures with nesting required by longitudinal studies are also considered. Particular interests concern the significance of regression coefficients and assessment of the form of the random effects. Extensions to unknown scale parameters, unknown link functions, survival and frailty models are outlined.

Keywords: Bayesian, blocking, longitudinal studies, Markov chain Monte Carlo, random effects, weighted least squares

1. Introduction

1.1. *Review of mixed generalized linear models*

The areas of generalized linear models (GLM) and random effects modelling have received a great deal of attention following the works of McCullagh and Nelder (1989) and Laird and Ware (1982) respectively. It did not take long for researchers to develop a framework to cope with both modelling approaches.

In the GLM setting, the data set consists of n observations with univariate response y_i and a p-dimensional vector of covariates $\mathbf{x_i}$, $i = 1, \ldots, n$. The observations are assumed to be independent with exponential family density

$$f(y_i|\theta_i) = \exp\{[y_i\theta_i - b(\theta_i)]/\phi_i\}c(y_i, \phi_i) \qquad (1)$$

The means $\mu_i = E(y_i|\theta_i)$ are related to the canonical parameters θ_i via $\mu_i = b'(\theta_i)$ and to the regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots \beta_p)'$ via the link relation

$$g(\mu_i) = \eta_i = \mathbf{x_i}'\boldsymbol{\beta}, \qquad i = 1, \ldots, n \qquad (2)$$

The link function g and the scale parameters ϕ_i are assumed to be known. The extension to the cases of unknown links and unknown ϕ_i are briefly considered in Section 4.

In the presence of q-dimensional random effects $\boldsymbol{\gamma}_i = (\gamma_{i1}, \ldots, \gamma_{iq})'$ explained by additional covariates $\mathbf{z_i}$, the link relation (2) is extended to

$$g(\mu_i) = \eta_i = \mathbf{x_i}'\boldsymbol{\beta} + \mathbf{z_i}'\boldsymbol{\gamma}_i, \qquad i = 1, \ldots, n \qquad (3)$$

The model is completed with a q-variate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$ with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$ for the random effects.

A typical situation here involves overdispersed data. The extra variation unaccounted for by the fixed effects model can be accommodated with the inclusion of a random effect (Breslow and Clayton, 1993). The first data set (Crowder, 1978, Table 3) comes from an experiment into the germination of seeds of different types and is reanalysed with fixed and mixed models. Data arriving in clusters or longitudinal studies provide another situation where random effects are called for. In these cases, however, the random effect structure is more elaborate where, depending on the sampling scheme, a multilevel hierarchical or nested structure is required (Goldstein, 1986). The second data set considered in this paper comes from a longitudinal medical study on the impact of a new drug in the treatment of epileptic patients (Thall and Vail, 1990). Evidence from previous analyses of this data suggests the presence of random effects at both unit and observation level. A reanalysis here shows new light over the possible distribution form for the random effects.

For the remainder of the paper, set $\mathbf{y} = (y_1, \ldots, y_n)'$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)'$, $\mathbf{X} = (\mathbf{x_1}, \ldots, \mathbf{x_n})'$, $\gamma = (\gamma_1, \ldots, \gamma_n)'$ and $\mathbf{Z} = (\mathbf{z_1}, \ldots, \mathbf{z_n})'$ so that $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ for the fixed effects GLM.

A Bayesian model is completed by the specification of a prior distributed for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. One possibility for $\boldsymbol{\beta}$ is $N(\mathbf{a}, \mathbf{R})$ used by West (1985) and Dellaportas and Smith (1993). It is of the same form as the random effects distribution, allows prior correlation and is likely to perform well in many applied problems. Its main weaknesses are thinness of its tails and unimodality. These problems are alleviated by respective use of scale and discrete mixtures of normals. Considering a mixing parameter λ, the first case corresponds to a prior specification $\boldsymbol{\beta}|\lambda \sim N(\mathbf{a}, \mathbf{R}/\lambda)$ and $\lambda \sim f$ for some f and the second one to $\boldsymbol{\beta}|\lambda = \lambda_l \sim N(\mathbf{a_l}\,\mathbf{R_l})$ and $Pr(\lambda = \lambda_l) = f_l$. The prior is completed with an inverse Wishart distribution for $\boldsymbol{\Sigma}$ with kernel $|\boldsymbol{\Sigma}|^{-\nu/2}\exp(-tr(\boldsymbol{\Sigma}^{-1}\mathbf{S})/2)$, denoted $IW(\nu, \mathbf{S})$. In the special case of a scalar random effect with $N(0, \sigma^2)$ distribution, an inverse Gamma prior is assumed for σ^2 with kernel $(\sigma^2)^{-(\nu/2)-1}\exp(-s/2\sigma^2)$, denoted $IG(\nu/2, s/2)$. Non-informative priors are obtained by letting $\mathbf{R}^{-1} \to \mathbf{0}$ and $\nu, s \to 0$.

When all these distributions are combined, the posterior distribution $\pi(\boldsymbol{\beta}, \gamma_1, \ldots \gamma_n, \boldsymbol{\Sigma})$ is formed where dependence on the observations $\mathbf{y}$ is suppressed throughout the paper for notational brevity. Particular interest may lie in assessing the form of the posterior distribution for the random effects or determining whether a specific regressor has any impact on the response. In all but very simple special cases, it is not possible to draw inferences analytically from this distribution and approximation methods must be used.

1.2. *Review of Markov chain Monte Carlo methodology*

In this paper, inference is performed through the Markov chain Monte Carlo (MCMC) methodology based on replacing the analytic expression of a density by a sample drawn from it. For the presentation below, consider a distribution (identified with its density) π on a random quantity $\mathbf{x}$ blocked into components $\mathbf{x_1}, \ldots \mathbf{x_m}$ that can themselves be vectors or matrices.

In its simplest form, the MCMC methodology is based on constructing a transition density $q(\mathbf{x}, \mathbf{x}^*)$ such that its Markov chain has equilibrium probability given by π. A draw $\mathbf{x}$ generated from π is obtained by:

1. starting with $\mathbf{x} = \mathbf{x^{(0)}}$ and setting $t = 1$;
2. sampling $\mathbf{x^{(t)}}$ from $q(\mathbf{x^{(t-1)}}, \mathbf{x})$;
3. increasing t by 1 and returning to Step 2.

For large enough t, $\mathbf{x^{(t)}}$ is a draw from π for any practical purpose. Some general convergence conditions are given in Tierney (1994) and Roberts and Smith (1994). A sample from π is formed by retaining enough values from the chain after convergence is established. Approximate independence between draws is achieved by retaining only every k generated value to break chain correlations. Alternatively, multiple chains may be run automatically ensuring independent draws after convergence. Important features for setting the convergence of a chain are the ergodic averages. These are formed by the consecutive averages of chain values and converge to the expected value of the limiting distribution π. A more thorough discussion of convergence checks and different schemes to form samples can be found in Gelman and Rubin (1992), Geyer (1992) and Gilks, Richardson and Spiegelhalter (1995).

One such scheme is Gibbs sampling (Geman and Geman, 1984; Gelfand and Smith, 1990). Its transition is formed by successively sampling from the full conditional densities $\pi_k(\mathbf{x_k}) = \pi(\mathbf{x_k}|\mathbf{x_{-k}})$ where $\mathbf{x_{-k}} = (\mathbf{x_1}, \ldots, \mathbf{x_{k-1}}, \mathbf{x_{k+1}}, \ldots, \mathbf{x_m})$, $k = 1, \ldots, m$. In many problems, this scheme works well but here the π_k's are hard to sample from.

Another scheme is provided by the Metropolis–Hastings algorithm (Metropolis *et al.*, 1953; Hastings, 1970). Consider a general transition density $q(\mathbf{x}, \mathbf{x}^*)$ and define

$$\alpha(\mathbf{x}, \mathbf{x}^*) = \min\left\{1, \frac{\pi(\mathbf{x}^*)q(\mathbf{x}^*, \mathbf{x})}{\pi(\mathbf{x})q(\mathbf{x}, \mathbf{x}^*)}\right\} \quad (4)$$

The move from state $\mathbf{x^{(t-1)}}$ to state $\mathbf{x^{(t)}}$ is made as follows: (a) sample $\mathbf{x}^*$ from $q(\mathbf{x^{(t-1)}}, \mathbf{x})$; (b) accept the move to $\mathbf{x}^*$ with probability $\alpha(\mathbf{x^{(t-1)}}, \mathbf{x}^*)$ and set $\mathbf{x^{(t)}} = \mathbf{x}^*$. Otherwise, stay at $\mathbf{x^{(t)}} = \mathbf{x^{(t-1)}}$.

The Markov chain formed by this procedure has equilibrium probability π and again, for large enough t, $\mathbf{x^{(t)}}$ is a draw from π.

Because of the acceptance stage (b), q is usually called proposal (transition) density. General guiding rules on the choice of proposal do not exist but empirical evidence suggests that the more it incorporates the structure of the problem the faster is the convergence. Tierney (1994) discusses and compares a few possible forms for the proposal.

In addition to these algorithms in pure form, a number of hybrid schemes are available. The most relevant ones here are those combining Metropolis steps within the Gibbs sampler (Muller, 1991). Consider a Gibbs sampling scheme where full conditionals π_k are formed but (some of them) are difficult to sample from. Then, Metropolis steps can be used to draw samples from $\mathbf{x_k}$ by forming proposal densities $q_k(\mathbf{x_k}, \mathbf{x_k^*})$ and acceptance probabilities $\alpha_k(\mathbf{x_k}, \mathbf{x_k^*})$ based on q_k and π_k. For the Gibbs sampler in pure forms blocking correlated quantities generally speeds up convergence but the same is not necessarily true for the Metropolis–Hastings algorithms. Once again, using the conditional independence structure of the model to dictate possible blocking turns out to be very important in practice.

1.3. *Outline of the paper*

Recently, Zeger and Karim (1991) proposed a Gibbs sampling approach and Breslow and Clayton (1993) proposed a penalized quasi-likelihood approach and these papers contain many of the relevant references to the subject. They are based on a normal distribution for the random effects. More recently, Lee and Nelder (1996) considered other possible distributions. Gelfand *et al.* (1996) also discussed reparametrization issues.

The scheme of Zeger and Karim (1991) involves use of two iterative procedures: one for the GLM fit and the other one for the Gibbs sampler. There is potential room for improvement in computations if these iterative procedures could be combined. Their approach requires setting of tuning constants that may change with every application, which make it unattractive as a general-purpose technique. Also, they only provide inference for the case of non-informative priors. The Gibbs sampling scheme of Dellaportas and Smith (1993) uses the adaptive rejection sampling technique (Gilks and Wild, 1992) to analyse GLM. The BUGS software considers the extension to include random effects in examples (Spiegelhalter *et al.*, 1993). Unfortunately, the adaptive rejection sampling is a univariate technique. It therefore can be very inefficient if parameters are highly correlated, which is frequently the case with regression coefficients.

The approach proposed here aims to overcome the difficulties presented above. Sampling from the posterior distribution is tackled by a MCMC approach using the Metropolis–Hastings algorithm. This is coupled with incorporation of the structure of the model, i.e. the form of likelihood *and* prior are taken into consideration, leading to an algorithm requiring a single iterative procedure. Prior distribution for regression coefficients and random effects distribution are not restricted to normality with the non-informative case providing a link with frequentist approaches. The resulting inference is based on samples from the posterior distribution of all model parameters. Standard assessments such as parameter significance and residual analysis can then be made without having to resort to asymptotic normality results.

The paper is organized as follows. Section 2 reviews the iterative weighted least squares (IWLS) procedure used in a GLM fit from both the frequentist and Bayesian viewpoint to motivate the method, introduced for GLM with normal prior. Section 3 extends the results to GLM with random effects. A comparison of the method with existing work is also returned to in more detail at the end of Section 3. Section 4 considers the extensions to non-normal priors and to models with additional hyper-parameters and discusses some computational issues. Section 5 draws some concluding remarks and indicates some possible research extensions.

2. Inference for the fixed effects model

The methods proposed in this paper will be introduced for the simpler case of a GLM without random effects and a normal prior distribution for the fixed effects β. They will be extended in Section 3 to mixed GLM and then, in Section 4, non-normal prior distributions for fixed and random effects will also be considered. For this section, the posterior distribution is

$$\pi(\beta) \propto \exp\left\{-\frac{1}{2}(\beta-\mathbf{a})'\mathbf{R}^{-1}(\beta-\mathbf{a}) + \sum_{i=1}^{n}\frac{y_i\theta_i - b(\theta_i)}{\phi_i}\right\} \tag{5}$$

where the likelihood term depends on β through θ_i, $i = 1, \ldots, n$.

2.1. *Iterative weighted least squares*

The maximum likelihood (ML) estimator in a GLM and its asymptotic variance are obtained by iterative use of weighted least squares (WLS) to transformed observations. Following McCullagh and Nelder (1989), define a vector of transformed observations $\tilde{\mathbf{y}}(\beta)$ and an associated diagonal matrix of weights $\mathbf{W}(\beta)$ with respective components

$$\tilde{y}_i(\beta) = \eta_i + (y_i - \mu_i)g'(\mu_i) \text{ and } W_i^{-1}(\beta) = b''(\theta_i)\{g'(\mu_i)\}^2, \quad i = 1, \ldots, n$$

The iterative weighted least squares algorithm is as follows:

1. start with $\beta = \mathbf{m}^{(0)}$ and set $t = 1$;
2. obtain $\mathbf{m}^{(t)}$, the WLS estimator of β as if $\tilde{\mathbf{y}}(\mathbf{m}^{(t-1)}) \sim \mathbf{N}(\mathbf{X}\beta, \mathbf{W}^{-1}(\mathbf{m}^{(t-1)}))$, and its associated covariance matrix $\mathbf{C}^{(t)}$;
3. increase t by 1 and return to Step 2.

At any stage of the iteration cycle, the values of $\mathbf{m}^{(t)}$ and $\mathbf{C}^{(t)}$ are respectively given by $(\mathbf{X}'\mathbf{W}(\mathbf{m}^{(t-1)})\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}(\mathbf{m}^{(t-1)})$ $\tilde{\mathbf{y}}(\mathbf{m}^{(t-1)})$ and $(\mathbf{X}'\mathbf{W}(\mathbf{m}^{(t-1)})\mathbf{X})^{-1}$. Note that they are both functions of the previous value $\mathbf{m}^{(t-1)}$.

The Bayesian version of the IWLS algorithm was developed by West (1985, Section 4) for the special case of canonical link $\theta_i = \eta_i, i = 1, \ldots, n$ but the extension to general link functions is straightforward. It provides the posterior mode and an approximate posterior covariance matrix for β. The idea is to combine Step 2 of the IWLS with a $N(\mathbf{a}, \mathbf{R})$ prior for β. This step is then replaced by:

2. obtain the $N(\mathbf{m}^{(t)}, \mathbf{C}^{(t)})$ 'posterior' distribution for β by combining the $N(\mathbf{a}, \mathbf{R})$ prior for β with 'observations' $\tilde{\mathbf{y}}(\mathbf{m}^{(t-1)}) \sim \mathbf{N}(\mathbf{X}\beta, \mathbf{W}^{-1}(\mathbf{m}^{(t-1)}))$;

The values of $\mathbf{m}^{(t)}$ and $\mathbf{C}^{(t)}$ now are respectively given by $(\mathbf{R}^{-1} + \mathbf{X}'\mathbf{W}(\mathbf{m}^{(t-1)})\mathbf{X})^{-1}\{\mathbf{R}^{-1}\mathbf{a} + \mathbf{X}'\mathbf{W}(\mathbf{m}^{(t-1)})\tilde{\mathbf{y}}(\mathbf{m}^{(t-1)})\}$ and $(\mathbf{R}^{-1} + \mathbf{X}'\mathbf{W}(\mathbf{m}^{(t-1)})\mathbf{X})^{-1}$. As before, they are both functions of the previous value $\mathbf{m^{(t-1)}}$. If the prior is non-informative ($\mathbf{R^{-1}} \rightarrow \mathbf{0}$), the original IWLS algorithm is recovered.

Both approaches use their IWLS algorithm combined with asymptotic results to make inferences based on approximate normality. We propose below a MCMC that preserves the structure of the IWLS algorithm without having to resort to possibly inadequate normality assumptions.

2.2. *A weighted least squares proposal*

The iterative schemes described in Sections 1.2 and 2.1 are very similar in nature and it seems natural to combine them in a single iteration cycle. The important steps in the cycles are Steps 2 which for the MCMC cycle requires a value sampled from a transition density and for the Bayesian IWLS cycle provides such a distribution but does not sample from it. A single iterative method then combines these two as follows:

1. start with $\boldsymbol{\beta} = \boldsymbol{\beta}^{(0)}$ and set $t = 1$;

2a. sample $\boldsymbol{\beta}^*$ from the $N(\mathbf{m^{(t)}}, \mathbf{C^{(t)}})$ proposal density and

2b. accept it with probability $\alpha(\boldsymbol{\beta}^{(t-1)}, \boldsymbol{\beta}^*)$ and set $\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}^*$. Otherwise, stay at $\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}^{(t-1)}$;

3. increase t by 1 and return to Step 2.

The moments of the proposal density are given by

$$\begin{aligned}\mathbf{m^{(t)}} &= (\mathbf{R^{-1}} + \mathbf{X'W}(\boldsymbol{\beta}^{\mathbf{(t-1)}})\mathbf{X})^{-1} \\ &\quad \times \{\mathbf{R^{-1}a} + \mathbf{X'W}(\boldsymbol{\beta}^{\mathbf{(t-1)}})\tilde{\mathbf{y}}(\boldsymbol{\beta}^{\mathbf{(t-1)}})\} \\ \mathbf{C^{(t)}} &= (\mathbf{R^{-1}} + \mathbf{X'W}(\boldsymbol{\beta}^{\mathbf{(t-1)}})\mathbf{X})^{-1} \qquad (6)\end{aligned}$$

so that the transition is made from the previous state $\boldsymbol{\beta}^{(t-1)}$. The acceptance probability, given in (4), is based on the ratio of the posterior densities (5) evaluated at $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}^{(t-1)}$ and also on the ratio $q(\boldsymbol{\beta}^{(t-1)}, \boldsymbol{\beta}^*)/q(\boldsymbol{\beta}^*, \boldsymbol{\beta}^{(t-1)})$ of proposal densities. The numerator is the density specified in Step 2a evaluated at $\boldsymbol{\beta}^*$ and the denominator $q(\boldsymbol{\beta}^*, \boldsymbol{\beta}^{(t-1)})$ is a $N(\mathbf{m}^*, \mathbf{C}^*)$ density evaluated at $\boldsymbol{\beta}^{(t-1)}$ where $\mathbf{m}^*$ and $\mathbf{C}^*$ have the same expression as $\mathbf{m^{(t)}}$ and $\mathbf{C^{(t)}}$ but depend on $\boldsymbol{\beta}^*$ instead of $\boldsymbol{\beta}^{(t-1)}$.

After convergence is reached, $\boldsymbol{\beta}^{(t)}$ corresponds to a draw from π. In the case of non-informative priors, draws are obtained from a normalized likelihood and comparisons with the ML approach can be made. The Markov chain formed by the above scheme leads to an algorithm with high acceptance rates due to the good approximation of the proposal to the true posterior distribution π and the chain moves are dictated by the structure of the model. Alternative chains are discussed in Section 4 but the WLS proposal combining prior and data information is retained throughout the paper.

Example 1: Logistic regression Consider the data in binomial form given by Crowder (1978, Table 3) consisting of the proportion of seeds that germinated in $n = 21$ plates. Relevant covariates are seed (2 types), root extract (2 types) and an interaction term. The success probabilities p_i are related to the covariates via $logit(p_i) = \mathbf{x}_\mathbf{i}'\boldsymbol{\beta}$ for $i = 1, \ldots, 21$. Table 1 presents the results for the fixed effects model obtained from retaining 500 samples every 10 iterations after an initial burn-in period of 1000 iterations with a non-informative prior for easy comparison with the ML results. Although only every 10th iterate was used to provide quasi-independent draws, there is no theoretical advantage in this due to ergodic theorems. Results obtained with successive samples were very similar providing further support for convergence of the chain.

The resulting posterior sample is very normal-like in form and its marginal quantiles coincide with those prescribed by the normal. This seems to hold in many applications of fixed effects GLM (Dellaportas and Smith, 1993; Spiegelhalter *et al.*, 1993) suggesting that in this case less elaborate schemes may be used. Also, the acceptance rate was 98% indicating similarity between the normal proposal

Table 1. *Estimation summary for Crowder's seed data. The table gives the values of the parameter estimates and a corresponding standard error (SE) for the fixed and mixed models using maximum likelihood (ML) and the MCMC methods proposed here. The values of standard errors are given by the asymptotic estimates for ML and the estimated posterior standard deviation from the MCMC sample*

	Fixed	GLM	Mixed	GLM
	ML	MCMC	ML	MCMC
Parameter	Estimate ± SE	Estimate ± SE	Estimate ± SE	Estimate ± SE
Intercept	−0.558 ± 0.126	−0.560 ± 0.132	−0.548 ± 0.167	−0.554 ± 0.192
Seed coef.	0.146 ± 0.223	0.144 ± 0.232	0.097 ± 0.278	0.099 ± 0.311
Extract coef.	1.318 ± 0.177	1.317 ± 0.184	1.337 ± 0.237	1.350 ± 0.284
Interaction coef.	−0.778 ± 0.306	−0.773 ± 0.312	−0.811 ± 0.385	−0.842 ± 0.448
σ	–	–	0.236 ± 0.110	0.284 ± 0.141

and the posterior distribution, and MCMC and ML results are very similar. The posterior correlation matrix is

$$\begin{pmatrix} 1.00 & & & \\ -0.60 & 1.00 & & \\ -0.73 & 0.41 & 1.00 & \\ 0.46 & -0.74 & -0.58 & 1.00 \end{pmatrix}$$

A point to notice here is the high correlation between some regression coefficients previously mentioned in Section 1 which frequently occurs in such models.

3. Inference for the mixed effects model

Consider now the model with random effects and link relation (3). From the Bayesian point of view, there is nothing different here and the model can be put in the form considered in the previous section by augmenting the vector of regression coefficients and the design matrix to

$$\beta^{(aug)} = \begin{pmatrix} \beta \\ \gamma_1 \\ \vdots \\ \gamma_n \end{pmatrix} \text{ and } \mathbf{X}^{(aug)} = \begin{pmatrix} \mathbf{x}_1' & \mathbf{z}_1' & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & & \vdots & & \\ \mathbf{x}_n' & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{z}_n' \end{pmatrix}$$

The model can be written as $\eta = \mathbf{X}^{(aug)}\beta^{(aug)}$ and completed with the prior $\beta^{(aug)}|\Sigma \sim N((\mathbf{a}',\mathbf{0}',\ldots,\mathbf{0}')'$, *diag* $(\mathbf{R},\Sigma,\ldots,\Sigma))$. This idea of extending regression coefficients to include random effects is used in some frequentist estimating procedures recently proposed. See, for example, the score equations of Breslow and Clayton (1993, Section 2.2) and Lee and Nelder (1996). The dimensionality of the parameter space increases dramatically however, which means that inversions of very large matrices may be required for large data sets. This may render the approach impractical for some applications based on surveys on moderate to large populations.

Here, a different route is pursued based on blocking correlated parameters. The parameter is divided into blocks $\beta, \gamma_1, \ldots, \gamma_n$ and Σ and the hybrid method described at the end of Section 1.2 is used to draw samples via Metropolis steps within a Gibbs updating scheme. The posterior distribution is

$$\pi(\beta,\gamma_1,\ldots,\gamma_n,\Sigma) \propto$$
$$\exp\left\{-\frac{1}{2}(\beta-\mathbf{a})'\mathbf{R}^{-1}(\beta-\mathbf{a}) + \sum_{i=1}^{n}\frac{y_i\theta_i - b(\theta_i)}{\phi_i}\right\}$$
$$\times|\Sigma|^{-n/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\gamma_i'\Sigma^{-1}\gamma_i\right\}$$

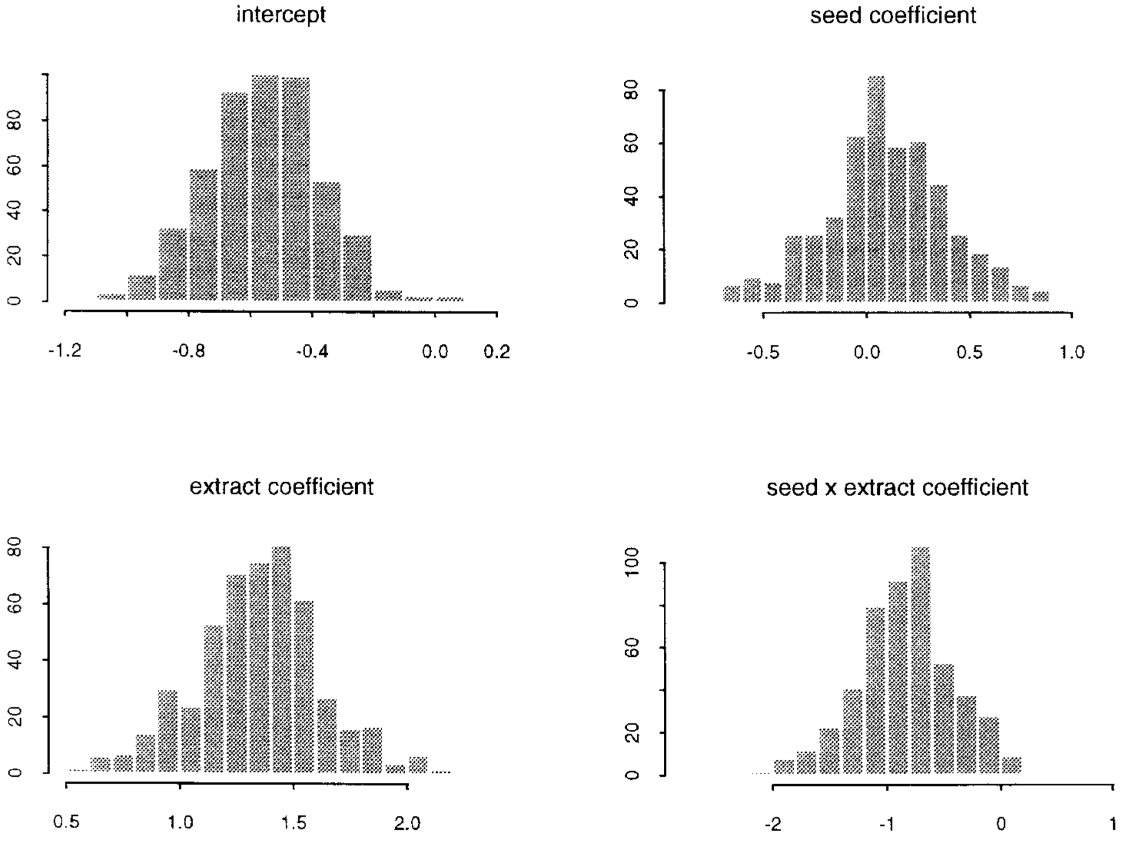

Fig. 1. *Histogram from the posterior distribution of the regression coefficients for the seeds data based on 500 samples*

$$\times|\boldsymbol{\Sigma}|^{-\nu/2}\exp\left\{-\frac{1}{2}tr(\boldsymbol{\Sigma}^{-1}\mathbf{S})\right\} \tag{7}$$

For the $\boldsymbol{\beta}$ block, the full conditional π_β still has the form (5) but the link now includes the known constants $\mathbf{z}_\mathbf{i}'\gamma_\mathbf{i}$, $i=1,\ldots,n$. These constants are known in the GLM methodology as offsets and the only changes required to the Metropolis step (2a–b) described in Section 2.2 is to replace the transformed variables by $\tilde{y}_i(\boldsymbol{\beta}^{(t-1)})-\mathbf{z}_\mathbf{i}'\gamma_i$, $i=1,\ldots,n$. So, the proposal is $q_\beta=N(\mathbf{m}^{(\mathbf{t})},\mathbf{C}^{(\mathbf{t})})$ with $\mathbf{m}^{(\mathbf{t})}$ and $\mathbf{C}^{(\mathbf{t})}$ given in (6) with the above change for $\tilde{\mathbf{y}}$.

For the γ_i block, the full conditional is obtained from (7) as

$$\pi_{\gamma i}(\gamma_i)\propto\exp\left\{-\frac{1}{2}\gamma_i'\boldsymbol{\Sigma}^{-1}\gamma_i+\frac{y_i\theta_i-b(\theta_i)}{\phi_i}\right\}$$

The posterior above is in the same form as (5) so the same approach with the WLS proposal can be used. The only changes are due to the presence of a single observation y_i and the offset now becomes $\mathbf{x}_\mathbf{i}'\boldsymbol{\beta}, i=1,\ldots,n$. The transformed observation and its weight are now written as $\tilde{y}_i(\gamma_i)$ and $W_i(\gamma_i)$. Following (6), the proposal q_{γ_i} is $N(\mathbf{m}_\mathbf{i}^{(\mathbf{t})},\mathbf{C}_\mathbf{i}^{(\mathbf{t})})$ with moments

$$\mathbf{m}_\mathbf{i}^{(\mathbf{t})}=(\boldsymbol{\Sigma}^{-1}+\mathbf{z}_\mathbf{i}W_i(\gamma_i^{(t-1)})\mathbf{z}_\mathbf{i}')^{-1}$$
$$\mathbf{z}_\mathbf{i}W_i(\gamma_i^{(t-1)})\{\tilde{y}_i(\gamma_i^{(t-1)})-\mathbf{x}_\mathbf{i}'\boldsymbol{\beta}\}$$
$$\mathbf{C}_\mathbf{i}^{(\mathbf{t})}=(\boldsymbol{\Sigma}^{-1}+\mathbf{z}_\mathbf{i}W_i(\gamma_i^{(t-1)})\mathbf{z}_\mathbf{i}')^{-1}$$

Finally, for the $\boldsymbol{\Sigma}$ block, the full conditional can be written as

$$\pi_\Sigma(\boldsymbol{\Sigma})\propto|\boldsymbol{\Sigma}|^{-(n+\nu)/2}\exp\left\{-\frac{1}{2}tr\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{S}+\sum_{i=1}^{n}\gamma_i'\gamma_i\right)\right]\right\}$$

which has an $IW(\nu+n,\mathbf{S}+\Sigma_i\gamma_i\gamma_i')$ form. In the case of a scalar random effect with variance σ^2, the full conditional for σ^2 is $IG((\nu+n)/2,(s+\Sigma_i\gamma_i^2)/2)$.

Example 1 (continued) The overdispersion noted by Crowder (1978) in the data was modelled by Breslow and Clayton (1993, Section 6.1) with the link relation changed to $logit(p_i)=\mathbf{x}_\mathbf{i}'\boldsymbol{\beta}+\gamma_i$ where the γ_i were assumed independent $N(0,\sigma^2), i=1,\ldots,n$.

The analysis was repeated with the MCMC method, a larger burn-in period of 2000 iterations and the non-informative prior $p(\boldsymbol{\beta},\sigma)\propto1/\sigma$. Figure 1 shows the marginal posterior histograms for the regression coefficients; the indications for normality are not so clear now. Table 1 presents numerical summaries from the posterior alongside those from the ML estimation. The point estimates are similar but the uncertainty measures are larger than those obtained from ML as in Dellaportas and Smith (1993, section 4.1). A possible cause is the asymmetry of the posterior which makes variances deviate more from those prescribed by asymptotic normality results. Similar comments are valid for point estimates for σ with the difference likely to be connected with the mode evaluation in ML as opposed to the posterior mean evaluated from a skew distribution. Figure 2 provides a summary of the inference for the random effects. It suggests that the prior assumption of normality for the random effects could be confirmed a posteriori. The posterior correlation matrix for the regression coefficients is very similar to that for the fixed GLM and the correlations between $\boldsymbol{\beta}$ and σ are all small in absolute value suggesting that separate blocking of random and fixed effects is reasonable.

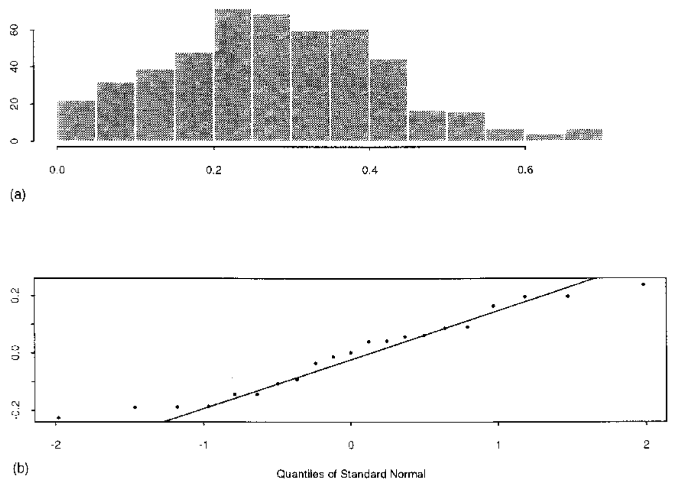

Fig. 2. *Summary of random effect inference: (a) histogram from the posterior distribution for σ; (b) Q–Q plot for normality of the random effects*

3.1. *Nested random effects*

In many applications, the data comes naturally divided in clusters. In longitudinal studies a collection of individuals is repeatedly observed through time. In these cases, the response may be influenced by random effects at both the subject and the unit levels. It is more appropriate here to change the notation to observations y_{ij} coming from an exponential family distribution with density (1), canonical parameter θ_{ij}, mean μ_{ij} and link

$$g(\mu_{ij})=\eta_{ij}=\mathbf{x}_\mathbf{ij}'\boldsymbol{\beta}+\mathbf{z}_\mathbf{ij}'\gamma_i+\mathbf{t}_\mathbf{ij}'\boldsymbol{\delta}_{ij},$$
$$j=1,\ldots,m_i,\quad i=1,\ldots,n$$

where the q-dimensional random effects γ_i are associated with the ith subject through the covariates $\mathbf{z}_\mathbf{ij}$ and the r-dimensional random effects $\boldsymbol{\delta}_{ij}$ are associated to the jth unit for the ith subject through covariates $\mathbf{t}_\mathbf{ij}$. Of course, more levels of nesting (Goldstein, 1986) can be added but the structure remains the same.

The model is completed with the independent random effects distributions $\gamma_i\sim N(\mathbf{0},\boldsymbol{\Sigma}_1)$ and $\boldsymbol{\delta}_{ij}\sim N(\mathbf{0},\boldsymbol{\Sigma}_2)$ and the independent prior distributions $\boldsymbol{\beta}\sim N(\mathbf{a},\mathbf{R}),\boldsymbol{\Sigma}_1\sim IW(\nu_1,\mathbf{S}_1)$ and $\boldsymbol{\Sigma}_2\sim IW(\nu_2,\mathbf{S}_2)$.

The analysis is pursued as before by appropriately blocking parameters according to their conditional independence structure. So, Metropolis-within-Gibbs sampling is again used for the blocks $\boldsymbol{\beta}, \gamma_1, \ldots, \gamma_n, \boldsymbol{\delta}_{11}, \ldots, \boldsymbol{\delta}_{nm_n}, \boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$.

For the $\boldsymbol{\beta}$ block, the Metropolis step is as in Section 3.1 but the link now includes the offsets $\mathbf{z}'_{\mathbf{ij}}\gamma_{\mathbf{i}} + \mathbf{t}'_{\mathbf{ij}}\boldsymbol{\delta}_{\mathbf{ij}}$ and the transformed observations are replaced by $\tilde{y}_{ij}(\boldsymbol{\beta}) - \mathbf{z}'_{\mathbf{ij}}\gamma_{\mathbf{i}} - \mathbf{t}'_{\mathbf{ij}}\boldsymbol{\delta}_{\mathbf{ij}}$.

For the γ_i block, the full conditional is

$$\pi_{\gamma_i}(\boldsymbol{\gamma}_i) \propto \exp\left\{ -\frac{1}{2}\boldsymbol{\gamma}'_i\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\gamma}_i + \sum_{j=1}^{m_i} \frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{\phi_{ij}} \right\}$$

The same Metropolis scheme with WLS proposal can be used for the GLM with the m_i observations y_{ij}, offsets $\mathbf{x}'_{\mathbf{ij}}\boldsymbol{\beta} + \mathbf{t}'_{\mathbf{ij}}\boldsymbol{\delta}_{\mathbf{ij}}$, transformed observations $\tilde{y}_{ij}(\boldsymbol{\gamma}_i)$ and weights $W_{ij}(\boldsymbol{\gamma}_i), j = 1, \ldots, m_i$. The proposal q_{γ_i} is $N(\mathbf{m}_{\mathbf{i}}^{(\mathbf{t})}, \mathbf{C}_{\mathbf{i}}^{(\mathbf{t})})$ with moments

$$\begin{aligned} \mathbf{m}_{\mathbf{i}}^{(\mathbf{t})} &= (\Sigma_1^{-1} + \mathbf{z}_{\mathbf{i}}\mathbf{W}_{\mathbf{i}}(\boldsymbol{\gamma}_{\mathbf{i}}^{(\mathbf{t-1})})\mathbf{z}'_{\mathbf{i}})^{-\mathbf{1}}\mathbf{z}_{\mathbf{i}}\mathbf{W}_{\mathbf{i}}(\boldsymbol{\gamma}_{\mathbf{i}}^{(\mathbf{t-1})}) \\ &\quad \times [\tilde{\mathbf{y}}_{\mathbf{i}}(\boldsymbol{\gamma}_{\mathbf{i}}^{(\mathbf{t-1})}) - \mathbf{x}'_{\mathbf{i}}\boldsymbol{\beta} - \mathbf{t}'_{\mathbf{i}}\boldsymbol{\delta}_{\mathbf{i}}] \\ \mathbf{C}_{\mathbf{i}}^{(\mathbf{t})} &= (\Sigma_1^{-1} + \mathbf{z}_{\mathbf{i}}\mathbf{W}_{\mathbf{i}}(\boldsymbol{\gamma}_{\mathbf{i}}^{(\mathbf{t-1})})\mathbf{z}'_{\mathbf{i}})^{-\mathbf{1}} \end{aligned}$$

where $\mathbf{W}_{\mathbf{i}} = diag(W_{i1}, \ldots, W_{im_i}), \mathbf{x}_{\mathbf{i}} = (\mathbf{x}_{\mathbf{i1}}, \ldots, \mathbf{x}_{\mathbf{im_i}})', \tilde{\mathbf{y}}_{\mathbf{i}} = (\tilde{y}_{i1}, \ldots, \tilde{y}_{im_i})', \mathbf{z}_{\mathbf{i}} = (\mathbf{z}_{\mathbf{i1}}, \ldots, \mathbf{z}_{\mathbf{im_i}})', \mathbf{t}_{\mathbf{i}} = diag(t_{i1}, \ldots, t_{im_i})$ and $\boldsymbol{\delta}_i = (\boldsymbol{\delta}'_{i1}, \ldots, \boldsymbol{\delta}'_{im_i})'$.

Similarly, for the $\boldsymbol{\delta}_{ij}$ block, the full conditional is

$$\pi_{\delta_{ij}}(\boldsymbol{\delta}_{ij}) \propto \exp\left\{ -\frac{1}{2}\boldsymbol{\delta}'_{ij}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\delta}_{ij} + \frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{\phi_{ij}} \right\}$$

The Metropolis scheme with WLS proposal is used for the GLM with a single observation y_{ij}, offset $\mathbf{x}'_{\mathbf{ij}}\boldsymbol{\beta} + \mathbf{z}'_{\mathbf{ij}}\gamma_{\mathbf{i}}$, transformed observation $\tilde{y}_{ij}(\boldsymbol{\delta}_{ij})$ and weight $W_{ij}(\boldsymbol{\delta}_{ij})$, $j = 1, \ldots, m_i$, $i = 1, \ldots, n$. The proposal $q_{\delta_{ij}}$ is $N(\mathbf{m}_{\mathbf{ij}}^{(\mathbf{t})}, \mathbf{C}_{\mathbf{ij}}^{(\mathbf{t})})$ with moments

$$\begin{aligned} \mathbf{m}_{\mathbf{ij}}^{(\mathbf{t})} &= (\boldsymbol{\Sigma}_2^{-1} + \mathbf{t}_{\mathbf{ij}}W_{ij}(\boldsymbol{\delta}_{ij}^{(t-1)})\mathbf{t}'_{\mathbf{ij}})^{-1}\mathbf{t}_{\mathbf{ij}}W_{ij}(\boldsymbol{\delta}_{ij}^{(t-1)}) \\ &\quad \times \{\tilde{y}_{ij}(\boldsymbol{\delta}_{ij}^{(t-1)}) - \mathbf{x}'_{\mathbf{ij}}\boldsymbol{\beta} - \mathbf{z}'_{\mathbf{ij}}\gamma_i\} \\ \mathbf{C}_{\mathbf{ij}}^{(\mathbf{t})} &= (\boldsymbol{\Sigma}_2^{-1} + \mathbf{t}_{\mathbf{ij}}W_{ij}(\boldsymbol{\delta}_{ij}^{(t-1)})\mathbf{t}'_{\mathbf{ij}})^{-1} \end{aligned}$$

The variance blocks have conjugate form with full conditionals $\boldsymbol{\Sigma}_1 \sim IW(\nu_1 + n, \mathbf{S_1} + \Sigma_i\gamma_i\gamma'_i)$ and $\boldsymbol{\Sigma}_2 \sim IW(\nu_2 + \Sigma_i m_i,\ \mathbf{S_2} + \Sigma_{i,j}\boldsymbol{\delta}_{ij}\boldsymbol{\delta}'_{ij})$. In the case of scalar random effects with variances σ_1^2 and σ_2^2, the full conditional for σ_1^2 is $IG((\nu_1 + n)/2, (s_1 + \Sigma_i\gamma_i^2)/2)$ and for σ_2^2 is $IG((\nu_2 + n)/2,\ (s_2 + \Sigma_{i,j}\delta_{ij}^2)/2)$.

A number of variations are possible here; unit level random effects may be dropped (Laird and Ware, 1982; Zeger and Karim, 1991) or may be allowed to have different variance matrices for each subject. These and other possibilities are straightforward to implement in the two-level scheme described above.

Example 2: Poisson log-linear model Consider the longitudinal data in Poisson form given by Thall and Vail (1990, Table 2) consisting of the counts of epileptic seizures in $n = 59$ patients. For each patient, the number of seizures in the 2 weeks preceding each of $m_i = 4$ clinic visits was recorded. The Poisson means μ_{ij} are related to the covariates via the log-linear model $\log \mu_{ij} = \mathbf{x}_{\mathbf{ij}}'\boldsymbol{\beta} + \gamma_i + \delta_{ij}$. The covariates included an average baseline count of seizures, treatment (new drug vs. placebo), their interaction, age and a fourth visit indicator V4 (see Thall and Vail (1990) for data and full details).

Analysis is based on 500 samples obtained from a single chain retaining draws every 20 iterations after a burn-in period of 2000 iterations. As before there is no theoretical advantage in using only every 20th iterate. The non-informative prior $p(\boldsymbol{\beta}, \sigma) \propto 1/\sigma$ is used to enable comparison

Table 2. *Estimation summary for Thall and Vail's epilepsy data. The table gives the values of the parameter estimates and a corresponding standard error (SE) for penalized quasi-likelihood (PQL), analysis with BUGS and the MCMC methods proposed here. The table also contains posterior estimates and standard deviation for the latter method, assuming a t_4 distribution for the unit random effects*

	Normal	random	effects	t_4
	PQL	BUGS	MCMC	MCMC
Parameter	Estimate ± SE	Estimate ± SE	Estimate ± SE	Estimate ± SE
Intercept	−1.27 ± 1.2	−1.43 ± 1.22	−1.41 ± 1.26	−1.33 ± 1.24
Baseline coef.	0.86 ± 0.13	0.89 ± 0.13	0.87 ± 0.14	0.89 ± 0.13
Treatment coef.	−0.93 ± 0.40	−0.94 ± 0.37	−0.98 ± 0.43	−0.93 ± 0.41
Interaction coef.	0.34 ± 0.21	0.34 ± 0.19	0.37 ± 0.21	0.34 ± 0.22
Age coef.	0.47 ± 0.35	0.49 ± 0.36	0.49 ± 0.37	0.46 ± 0.37
V4 coef.	−0.10 ± 0.09	−0.10 ± 0.09	−0.10 ± 0.09	−0.09 ± 0.08
σ_1	0.48 ± 0.06	0.50 ± 0.07	0.50 ± 0.07	0.49 ± 0.07
σ_2	0.36 ± 0.04	0.36 ± 0.05	0.36 ± 0.04	0.26 ± 0.04

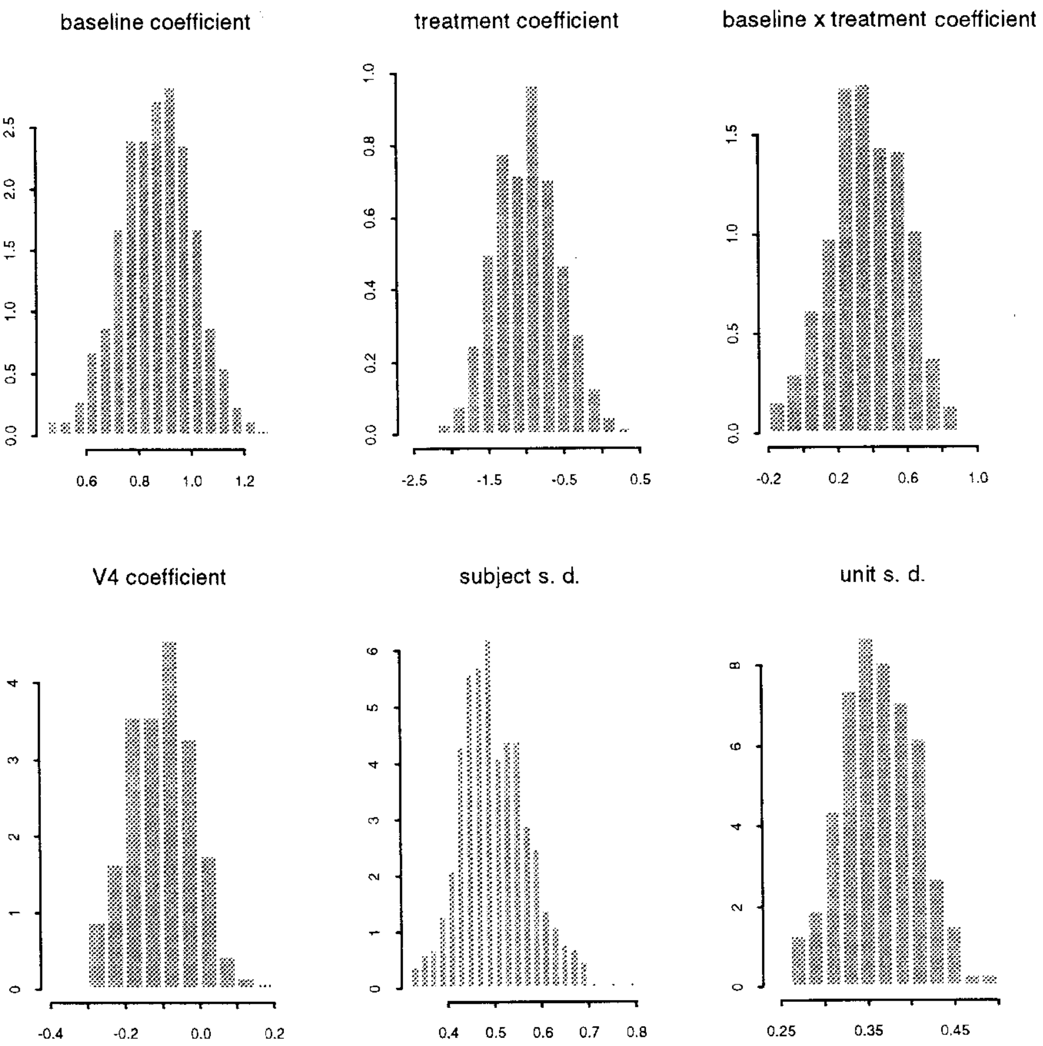

Fig. 3. *Histogram from the posterior distribution of the regression coefficients and standard deviation of random effects for the epilepsy data based on 500 samples*

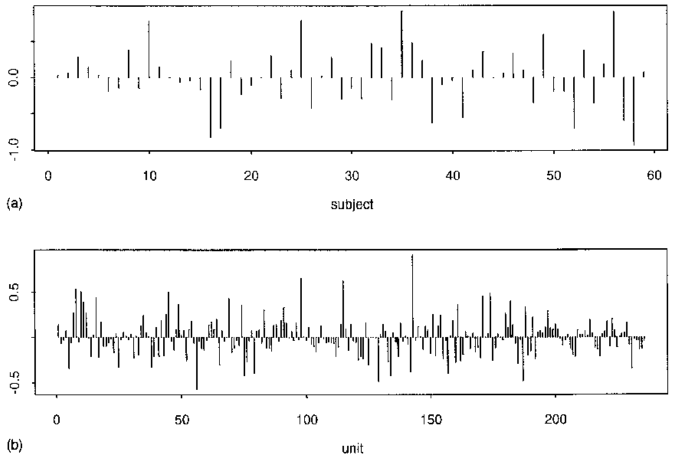

Fig. 4. *Sample averages over the 500 samples of the random effects: (a) subject level random effects; (b) unit level random effects. The units are ordered by running through first visit counts for all patients, then subsequently the second, then third and finally the fourth visit counts for all patients*

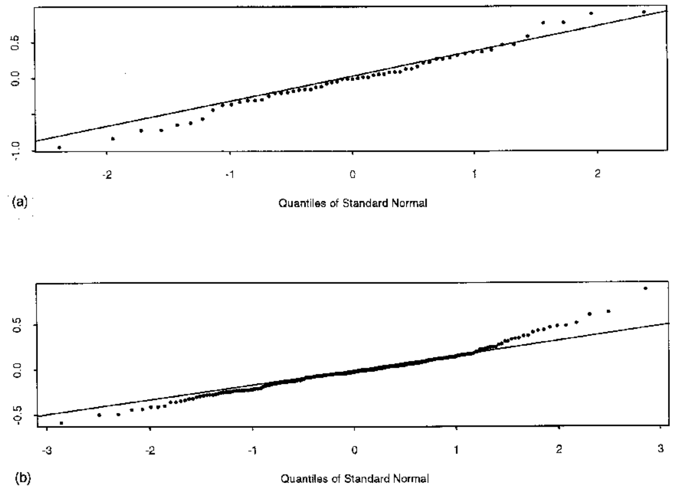

Fig. 5. *Q–Q plot for normality of the random effects: (a) subject level random effects; (b) unit level random effects*

with the penalized quasi-likelihood analysis (Breslow and Clayton 1993, Section 6.2) and the analysis with the BUGS software (Spiegelhalter *et al.*, 1993, Section 6). Table 2 presents a numerical summary of the estimation while Figure 3 displays marginal posterior histograms. Convergence is fast and the overall acceptance rate is above 85%. Non-normal forms for the regression coefficients are observed. Figure 4 presents point estimates for the random effects similar to those obtained by a likelihood approach (J. Nelder, personal communication, September 1994). Figure 5 provides a check on the posterior normality of the random effects. It indicates significant departures on both tails of the unit random effects, suggesting a heavier tail form. The correlation matrix for the regression coefficients is

$$\begin{pmatrix} 1.00 & & & & & \\ -0.12 & 1.00 & & & & \\ 0.09 & 0.57 & 1.00 & & & \\ -0.12 & -0.63 & -0.93 & 1.00 & & \\ -0.98 & -0.08 & -0.22 & 0.25 & 1.00 & \\ -0.02 & 0.04 & -0.07 & -0.06 & -0.01 & 1.00 \end{pmatrix}$$

Again there are several very high (negative) correlations but the correlations between β and the σ's are all again small.

3.2. *Comparison with previous work*

The work by Zeger and Karim (1991) was one of the first to use the Gibbs sampling methodology. They use rejection sampling from the posterior with a normal proposal distribution based on maximum likelihood estimation. Their calculation requires two iterative procedures: one for the IWLS algorithm and the other one for the Gibbs sampler. Also, they did not indicate how prior information about the regression coefficients may be incorporated. The approach proposed here combines the two iterations in a unified scheme while explicitly accounting for incorporation of prior opinions.

These difficulties are not present in the approach of Dellaportas and Smith (1993). The main drawback is the use of a univariate technique in such a structured problem. For the epilepsy data, Spiegelhalter *et al.* (1993, p.20) reported 'serious convergence problems' if the covariates are not centred. The high negative entries in the posterior correlation matrix of Section 3.2 help to explain why. Gelfand *et al.* (1996) showed how reparametrization can improve convergence in mixed GLM. Although this is a healthy exercise, simple use of a MCMC scheme incorporating blocking of correlated parameters suffices here. The calculations of Section 3.2 were done without centreing the covariates, and no convergence problems were encountered.

4. Extensions

4.1. *Non-normal priors*

The normal prior combines nicely with the WLS procedure to give the proposal density of Section 2.2. In some cases, however, it may not correspond to the prior information available. In Section 1.1, extensions covering a variety of opinions were discussed based on scale or discrete mixtures of normals. Irrespective of the choice of mixture, the method remains basically the same with an additional block λ. The only other parameter block affected by the mixture is β. It is affected in as much as its prior moments entering (6) are now given by $\mathbf{a}_l$ and $\mathbf{R}_l$ ($\mathbf{a}$ and $\mathbf{R}/\lambda$) for a discrete (scale) mixture of normals, following the notation of Section 1.1.

For the λ block, the full conditional is

$$\pi_\lambda(\lambda = \lambda_l) \propto f_l |\mathbf{R}_l|^{-1/2} \exp\left\{ -\frac{1}{2}(\beta - \mathbf{a}_l)'\mathbf{R}_l^{-1}(\beta - \mathbf{a}_l) \right\} \tag{8}$$

for the discrete case and

$$\pi_\lambda(\lambda) \propto f(\lambda)\lambda^{1/2} \exp\left\{ -\frac{\lambda}{2}(\beta - \mathbf{a})'\mathbf{R}^{-1}(\beta - \mathbf{a}) \right\} \tag{9}$$

for the scale mixture case. They are both univariate so Gibbs sampling can be used. In the common case of a t_ξ prior for β, $f = G(\xi/2, \xi/2)$ and $\pi_\lambda = G[(\xi + 1)/2, \{\xi + (\beta - \mathbf{a})'\mathbf{R}^{-1}(\beta - \mathbf{a})\}/2]$.

Similar comments are valid for the distribution of random effects with a hyperparameter λ_i associated with each random effect. Typically in the discrete mixture case for the random effects distribution, all the means $\mathbf{a}_l$ are set to ensure an overall zero mean, and the covariance

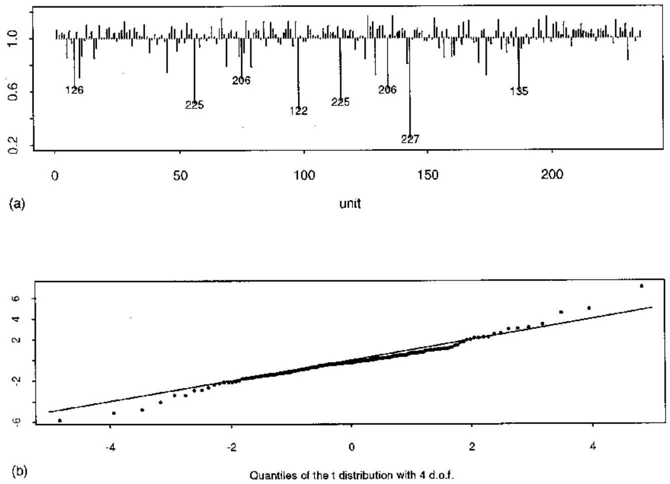

Fig. 6. *Summary of unit random effect inference: (a) point estimates of λ_{ij}s ordered as in Fig. 4, the most extreme ones are represented by the patient number; (b) Q–Q plot of standardized estimated random effects against quantiles of the t_4 distribution*

matrices $\mathbf{\Sigma}_l$ are restricted to $\mathbf{\Sigma}/\lambda_l$ to ensure enough information is available to estimate the variance. For the random effects model without nesting, changes for the γ_i blocks are as outlined above for $\boldsymbol{\beta}$ with corresponding changes in prior moments. For the $\mathbf{\Sigma}$ block, it is easy to see that the full conditional is $IW(\nu+n, \mathbf{S}+\Sigma_i\lambda_i\gamma_i\gamma_i')$. Finally, the full conditionals for λ_i will have forms similar to (8) in the discrete mixture case and (9) in the scale mixture case.

Example 2 (continued) The overdispersion of the random effects at the unit level shown in Fig. 5 was investigated in more detail by changing the distribution of the δ_{ij}'s from normal to a t_ξ with ξ small. In this case, mixing hyperparameters λ_{ij} are introduced with independent $G(\xi/2, \xi/2)$ priors. Their full conditionals are given by a $G\{(\xi+1)/2, (\xi+\delta_{ij}^2/\sigma_2^2)/2\}$ distribution. If the posterior distribution of λ_{ij} is concentrated around one, the normality assumption is reasonable. A more dispersed posterior indicates heavier tails needed to accommodate a more extreme random effect.

Estimation of the nested mixed GLM was repeated with a t_4 distribution for the unit random effects and is summarized in Table 2. Apart from σ_2 which is no longer the standard deviation of the random effects, all model parameters had very similar estimates. Figure 6a shows the estimates of the λ_{ij}s, and direct correspondence between the lowest estimates and the more extreme random effects in Fig. 4b can be easily made. Figure 6b shows that a more adequate fit results from the replacement of the normal assumption for the unit random effects by a t.

Another extension was considered by Lee and Nelder (1996) by allowing exponential family distributions for the random effects. In particular, they focused attention on conjugate double GLM in a nested data structure where the random effects γ_i have density with kernel $\exp\{h_1(\boldsymbol{\lambda})\gamma_i - h_2(\boldsymbol{\lambda})b(\gamma_i)\}$ where h_1 and h_2 are known functions of the hyperparameter $\boldsymbol{\lambda}$. In the case of the Poisson log-linear model, this would give a kernel $h_1\gamma_i - h_2\exp(\gamma_i)$. The link (3) is then given by

$$\mu_{ij} = e^{\mathbf{x}_{ij}'\boldsymbol{\beta}}w_i \tag{10}$$

where $w_i = \exp(\gamma_i)$ has Gamma distribution with variance λ and mean restricted to be one. Tsutakawa (1988) and Firth and Harris (1991) have also considered multiplicative mixed GLM similar to (10). Again, sampling can be done with the same blocking structure. The λ block is typically of low dimension but with an analytically untractable full conditional π_λ. Adaptive rejection in the case of log-concavity or adaptive rejection Metropolis (Gilks *et al.*, 1995) in the general case are particularly suitable for sampling in these situations.

4.2. *Additional parameters; survival and frailty models*

When the scale parameters ϕ_i are unknown, they can form another parameter block and be sampled from their full conditional. In the special case of normal observations, this is easily done if inverse Gamma priors are specified. In general, the full conditional will be very hard to sample from and each problem must be approached separately. If the scales are all equal, then adaptive rejection sampling can be an attractive procedure as the block becomes one-dimensional. West (1985, section 2) considered an inverse Gamma approximation to the full conditional distribution of the scale parameter. This can now be used as a proposal in a Metropolis step.

Similar comments are valid for the case of an unknown link function. Mallick and Gelfand (1994) considered a non-parametric approach to GLM with unknown link functions. Assuming instead a parametric family of link functions $g^{(\lambda)}(\mu) = \eta$, inference can proceed as before conditional on $\boldsymbol{\lambda}$. If $\boldsymbol{\lambda}$ is one- or two-dimensional, then their untractable full conditional can be efficiently sampled via adaptive rejection (Metropolis) sampling in a Gibbs step.

Another important example is the case of proportional hazards models (Cox, 1972) where the hazard function is given by $h(t) = h_0(t;\boldsymbol{\lambda})\exp(\mathbf{x}_i'\boldsymbol{\beta})$ with the baseline hazard h_0 parametrized in terms of $\boldsymbol{\lambda}$. Typical examples are the exponential hazard $h_0(t;\lambda) = \lambda t$ and the Weibull hazard $h_0(t;\boldsymbol{\lambda}) = \lambda_1 t^{\lambda_2}$. Aitkin and Clayton (1980) showed that, given $\boldsymbol{\lambda}$, the likelihood behaves as that from a Poisson sample with observations given by the survival indicators, canonical link, means $\mu_i = H_0(t_i)\exp(\mathbf{x}_i'\boldsymbol{\beta})$ and offsets $\log H_0(t)$ where H_0 is the integrated baseline hazard. Again, given the block $\boldsymbol{\lambda}$, analysis for the block $\boldsymbol{\beta}$ proceeds as in Section 3. For the $\boldsymbol{\lambda}$ block, Gibbs sampling can be used with adaptive rejection sampling.

Frailty models introduced by Clayton and Cuzick (1985)

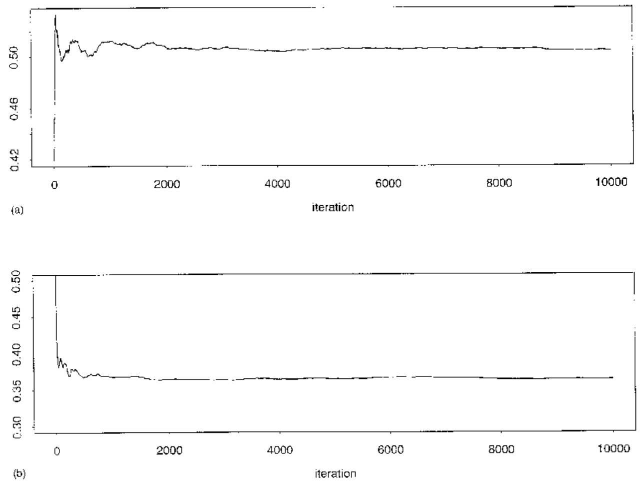

Fig. 7. *Ergodic averages of standard deviations of random effects for the epilepsy data in a single long run: (a) σ_1; (b) σ_2*

in a nested context extend survival models by including subject random effects w_i having Gamma distribution with mean one. By doing so, the means of the survival indicators are given by

$$\mu_{ij} = H_0(t_{ij})e^{\mathbf{x}'_{ij}\beta}w_i$$

The above specification is very similar to (10) and inference for conjugate double GLM outlined in Section 4.1 can be combined with the above discussion for survival models to produce Bayesian inference for frailty models. Full discussion of this topic is beyond the scope of this paper and results will be reported elsewhere.

4.3. *Computational issues*

The applications were performed using a single chain (Geyer, 1992). Parallel chains as advocated by Gelman and Rubin (1992) could also be used. For the applications of this paper, both lead to very similar results. The methods of the paper apply to both approaches with equal ease. Parallel chains may be computationally inefficient for situations where chains are slow to converge. It may be preferable to run a single long chain for models with non-nested random effects or measurement errors, that might lead to slow mixing chains. Convergence diagnostics help counteract some of the criticisms towards this approach. For the applications in this paper, the fast convergence reduces the importance of this point. However, the scientific debate about this point is still not settled (Gilks *et al.*, 1995, p. 13).

A single long chain can also be used to form ergodic averages, providing informal convergence checks. As an example, Fig. 7 suggests convergence through a single long chain around the estimates given in Table 2. Similar plots can easily be obtained for the parameters of this and other models used in this paper.

All chains were initialized with random effects and covariances set to zero and variances set to one. Updating was done for the β block first to position the chain in the right region of the parameter space. Then, the random effects were updated and finally the covariance matrices. Other schemes involving an initialization from the ML estimator or the use of a few steps of the WLS within each cycle of the Markov chain, with the full IWLS used by Zeger and Karim (1991) as a special case, are also possible.

The MCMC strategy adopted throughout this paper involved the blocking of the fixed and random effects separately as in Section 3 and use of the WLS proposal introduced in Section 2.2. Examples of other blockings are setting a single block $(\gamma_1, \ldots, \gamma_n)$ or even $(\beta, \gamma_1, \ldots, \gamma_n)$. Given the conditional independence of the random effects given β and Σ, there seems to be little to be gained from this strategy. Other examples of proposal densities include the prior distribution, which gives the simplest expression for the acceptance probability or random walk chains (Tierney, 1994), where the proposal is centred around the previous value possibly with increased variance. The structured blocking and the WLS proposal outperformed these combinations in the applications.

5. Concluding remarks

Markov chain Monte Carlo techniques are a powerful tool for solving long-standing Bayesian problems. This paper shows a simple procedure for the case of mixed GLM. The approach is based on general model-based rules that apply to any well specified mixed model and so can be incorporated in standard software available to perform inference on these models. In the special case of fixed GLM with normal priors considered in Section 2, all it requires is the trivial incorporation of the prior terms, a multivariate normal sampling and the evaluation of the probability (4).

The methodology can also be extended to consider linear models where the regression coefficients have an additional hierarchical structure (Gelfand *et al.*, 1990), additional dynamic structure (West *et al.*, 1985) or even an additional dynamic hierarchical structure (Gamerman and Migon, 1993). Extensions to non-linear models with random effects (Lindstrom and Bates, 1990; Wakefield *et al.*, 1994) however are not immediate given the important part played by linearity on the selection of the proposal density.

Acknowledgements

The research was performed during a visit to Imperial College London, whose hospitality is gratefully acknowledged, and was supported by a grant from CAPES-Brazil. The author has benefited from frequent discussions with Adrian Smith, John Nelder and Jon Wakefield; and also thanks David Phillips, Peter Muller, David Draper, the Editor and a referee for useful comments; Eduardo Gutierrez-Pena for the multivariate normal random generation routine and Penelope Vounatsou for computational help.

References

Aitkin, M. and Clayton, D. (1980) The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM. *Applied Statistics,* **29**, 156–63.

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

Clayton, D. and Cuzick, J. (1985) Multivariate generalisations of the proportional hazards model. *Journal of the Royal Statistical Society*, Series A, **148**, 82–117.

Cox, D. R. (1972) Regression models and life tables. *Journal of the Royal Statistical Society*, Series B, **34**, 187–220.

Crowder, M. J. (1978) Beta-binomial ANOVA for proportions. *Applied Statistics*, **27**, 34–7.

Dellaportas, P. and Smith, A. F. M. (1993) Bayesian inference for generalised linear and proportional hazards models via Gibbs sampling. *Applied Statistics*, **42**, 443–59.

Firth, D. and Harris, I. R. (1991) Quasi-likelihood for multiplicative random effects. *Biometrika*, **78**, 545–55.

Gamerman, D. and Migon, H. S. (1993) Dynamic hierarchical models. *Journal of the Royal Statistical Society* Series B, **55**, 629–42.

Gelfand, A. E., Hills, S. E., Racine-Poon, A. and Smith, A. F. M. (1990) Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association*, **85**, 972–85.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

Gelfand, A. E., Sahu, S. K. and Carlin, B. P. (1996) Efficient parametrizations for generalized linear mixed models. To appear in *Bayesian Statistics 5* (eds J. M. Bernardo *et al.*). Oxford: Oxford University Press.

Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, **7**, 457–72.

Geyer, C. J. (1992) Practical Markov chain Monte Carlo. *Statistical Science*, **7**, 473–511.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–41.

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

Gilks, W. R., Best, N. G. and Tan, K. K. C. (1995) Adaptive rejection Metropolis sampling within Gibbs sampling. *Applied Statistics*, **44**, 455–72.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996) Introducing Markov chain Monte Carlo. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks *et al.*), pp. 1–19. New York: Chapman and Hall.

Goldstein, H. (1986) Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, **73**, 43–56.

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Laird, N. M. and Ware, J. H. (1982) Random effects models for longitudinal data. *Biometrics*, **38**, 963–74.

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

Lindstrom, M. J. and Bates, D. M. (1990) Nonlinear mixed effects models for repeated measures data. *Biometrics*, **46**, 673–87.

Mallick, B. K. and Gelfand, A. E. (1994) Generalized linear models with unknown link functions. *Biometrika*, **81**, 237–45.

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

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953) Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–92.

Muller, P. (1991) Metropolis based posterior integration schemes. Technical Report 91-09 Department of Statistics, Purdue University.

Roberts, G. O. and Smith, A. F. M. (1994) Simple conditions for the convergence of the Gibbs sampler and Metropolis–Hastings algorithm. *Stochastic Processes and Their Applications*, **49**, 207–16.

Spiegelhalter, D. J., Thomas, A., Best, N. and Gilks, W. R. (1993) *BUGS Examples 0.30*, MRC Biostatistics Unit, Cambridge.

Thall, P. F. and Vail, S. C. (1990) Some covariance models for longitudinal count data with overdispersion. *Biometrics*, **46**, 657–71.

Tierney, L. (1994) Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics*, **22**, 1701–62.

Tsutakawa, R. K. (1988) Mixed model for analyzing geographic variability in mortality rates. *Journal of the American Statistical Association*, **83**, 37–42.

Wakefield, J. C., Smith, A. F. M., Racine-Poon, A. and Gelfand, A. E. (1994) Bayesian analysis of linear and non-linear population models by using the Gibbs sampler. *Applied Statistics*, **43**, 201–21.

West, M. (1985) Generalized linear models: outlier accommodation, scale parameters and prior distributions. In *Bayesian Statistics 2* (eds J. M. Bernardo *et al.*), 531–58. Amsterdam: North Holland.

West, M., Harrison, P. J. and Migon, H. S. (1985) Dynamic generalized linear models and Bayesian forecasting (with discussion). *Journal of the American Statistical Association*, **80**, 73–96.

Zeger, S. L. and Karim, M. R. (1991) Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association*, **86**, 79–86.