RESEARCH SUPPORT
UNIVERSITY OF MICHIGAN BUSINESS SCHOOL

APRIL 1997

GIBBS SAMPLING FOR BAYESIAN NONCONJUGATE MODELS USING AUXILLIARY VARIABLES

WORKING PAPER #9705-13

BY
PAUL DAMIEN
UNIVERSITY OF MICHIGAN BUSINESS SCHOOL,
STEPHEN WALKER
IMPERIAL COLLEGE, LONDON, UK
AND
JON WAKEFIELD
IMPERIAL COLLEGE, LONDON, UK

Gibbs Sampling For Bayesian Nonconjugate Models Using Auxilliary Variables

Paul Damien[1] Stephen Walker[2]
Jon Wakefield[3]

[1] Business School, University of Michigan, Ann Arbor, 48109-1234, USA.

[2] Department of Mathematics, Imperial College, 180 Queen's Gate, London SW7 2BZ

[3] Department of Epidemiology and Public Health, Imperial College School of Medicine at St Mary's, Norfolk Place, London, W2 1PG.

SUMMARY

We demonstrate the use of auxilliary (or latent) variables for sampling non-standard densities which arise in the context of a Gibbs sampler, involving Bayesian nonconjugate models. Their strategic use can result in a Gibbs sampler having standard full conditionals. We propose such a procedure to simplify or speed up an existing Markov chain Monte Carlo algorithm. The strength of this approach lies in its generality and its ease of implementation. The method is illustrated on posterior densities arising from Bayesian nonconjugate and hierarchical models. A feature of the paper, therefore, is to provide an alternative sampling algorithm to rejection based methods and other sampling approaches such as the Metropolis-Hastings algorithm.

Keywords: Gibbs sampler, Hierarchical model, Latent variable, Nonconjugate model.

1 Introduction

Markov chain Monte Carlo (MCMC) methods (Smith and Roberts, 1993; Tierney, 1994) allow Bayesian inference for highly complex models in which realistic distributional assumptions can be made. The Gibbs sampler, the most common of the MCMC algorithms, can often be difficult to implement due to the required conditional distributions assuming awkward forms. In this case the practitioner may turn to the Hastings-Metropolis algorithm and /or a rejection algorithm; see, for example, Metropolis et al. (1953); Hastings (1970); Tierney (1994); Devroye (1986). However, this may be difficult to set up and may require 'tuning' to achieve satisfactory performance (Bennett, Racine-Poon and Wakefield, 1995). In this paper we discuss a novel approach which, after the introduction of strategic auxilliary (or latent) variables, results in a Gibbs sampler having a set of standard full conditionals.

Suppose the required conditional distribution for a random variable X is denoted f. The basic idea is to introduce a latent variable U, construct the joint density of U and X, with marginal density for X given by f, and then extend the Gibbs sampler to include the extra full conditional for U. We demonstrate that in many cases it is possible to introduce a latent variable so that all full conditionals are standard and can be sampled directly. This is obviously appealing and, provided there is no dramatic loss in efficiency compared to the original chain(s), will be of particular interest to the MCMC practitioner.

In many cases the introduction of a single latent variable will 'split' a nonstandard full conditional disribution appearing in a Gibbs sampler into two standard full conditional distributions, effectively increasing the Gibbs sampler by one more full conditional. This then may lead to a more efficient method than the Metropolis-Hastings algorithm, the adaptive rejection method for log-concave densities (Gilks and Wild, 1992) and other rejection algorithms, in many contexts.

For a historical overview of Markov chain methods and the use of latent (auxilliary) variables the reader is referred to Besag and Green (1993). In particular our approach develops the original idea introduced by Edwards and Sokal (1988) and highlighted by Besag and Green in Section 5 of their paper. Recent progress with auxilliary variables is reported in Higdon (1996), and references therein.

The paper is organised as follows. In the next section, we develop the

theory underlying the new algorithm. We show that our method improves on a Metropolis independence chain as well as a rejection algorithm. In Section 3 we implement the approach for Bayesian hierarchical models; Section 4.1 for *generalised linear mixed models* (GLMMs) and Section 4.2 for *nonlinear mixed models* (NMMs). Section 5 contains numerical examples, followed by a concluding discussion in Section 6.

2 Preliminaries

The main result on which the algorithm developed in this paper is given in the following theorem; see, Damien and Walker (1996).

THEOREM 1. *Suppose we wish to generate random variates from a density f given by*

$$f(x) \propto \pi(x) \prod_{l=1}^{L} g_l(x),$$

where π is a density of known form and the g_l are nonnegative invertible functions (not necessarily densities), that is, if $g_l(x) > u$ then it is possible to obtain the set $A_l(u) = \{x : g_l(x) > u\}$. Then a Gibbs sampler for generating random variates from f exists in which all but one of the full conditionals are uniform densities and the remaining full conditional is a truncated version of π.

Proof. We introduce the latent variables $u = (u_1, ..., u_L)$ with each u_l defined on $(0, \infty)$ such that the joint density with x is given by

$$f(x, u_1, ..., u_L) \propto \pi(x) \prod_{l} I\{u_l < g_l(x)\}.$$

Clearly the marginal density for x is $f(x)$. A Gibbs sampler can now be implemented where obviously the full conditionals for each u_l is the uniform density on $(0, g_l(x))$. The full conditional for x is given by π restricted to the set $A(u) = \{x : g_l(x) > u_l, l = 1, ..., L\}$.

The decomposition appearing in Theorem 1 is very similar to an expression appearing in Besag and Green (1993, Section 5). However, they do not mention the significant advantages that having invertible g_l lead to. They say,

"When dealing with more complicated models, *direct simulation* from $f(x|u)$ is *unlikely to be available.*" [Italics ours.] As a consequence, they propose that sampling from π restricted to the set $A(u)$ be achieved using rejection algorithms; that is, sampling from π until the sample falls in $A(u)$. While this method would work in principle, our aim in this paper is to demonstrate that we *can* introduce latent variables in complex models which still permits *direct simulation* from $f(x|u)$; i.e., the use of rejection sampling can be obviated. If the decomposition indicated in Theorem 1 is possible then we will achieve substantial efficiency over the rejection sampling of Besag and Green. The class of densities having the appropriate decomposition seems to be large, and specifically, in the context of Bayesian models, the decomposition stated in Theorem 1 can be readily achieved.

Consider the posterior density given by $f(x) \propto l(x)\pi(x)$ and suppose it is not possible to sample directly from f. The general idea is to introduce latent variable Y, defined on the interval $(0, \infty)$ or more strictly the interval $(0, l(\hat{x}))$, where $\hat{x}$ maximises $l(.)$, and define the joint density with X by

$$f(x, y) \propto I\Big(y < l(x)\Big)\pi(x).$$

The full conditional for Y is $U(0, l(x))$ and the full conditional for X is π, restricted to the set $A_y = \{x : l(x) > y\}$.

We can show how this approach "improves" on a particular independent Hastings -Metropolis chain. The Metropolis algorithm is a Markovian scheme which may be used for obtaining samples from the posterior $f(x) \propto l(x)\pi(x)$. Given $x^{(t)}$, a proposal for $x^{(t+1)}$, $\tilde{x}$, is taken from, for example, $\pi(.)$, and a uniform random variable, u, is taken from the interval $(0, 1)$. Essentially, if $l(\tilde{x})/l(x^{(t)}) > u$ then $x^{(t+1)} = \tilde{x}$ else $x^{(t+1)} = x^{(t)}$. The chain either 'moves on' or 'stays where it is'. The convention is that $\tilde{x}$ is sampled first followed by u. Suppose we reverse this and sample u first. To 'move on' we need to sample $\tilde{x}$ from $\pi(.)$ such that $l(\tilde{x})/l(x^{(t)}) > u$. Suppose, therefore, we sample $\tilde{x}$ from $\pi(.)$ restricted to the set $A_u(t) = \{x : l(x) > ul(x^{(t)})\}$. The chain will always 'move on'. In fact we have just described a Gibbs sampler with standard full conditionals, detailed in Theorem 2:

THEOREM 2. *The Markovian scheme for generating* $\{x^{(t)}\}$ *given by* $x^{(t+1)} \sim \pi(.)$ *restricted to the set* $A_u(t) = \{x : l(x) > ul(x^{(t)})\}$, *where* u *is a uniform random variable from the interval* $(0, 1)$, *satisfies* $x^{(t)} \to_d x \sim f(x) \propto$

$l(x)\pi(x)$.

Proof. Define the joint density function of x and y by

$$f(x,y) \propto I\Big(y < l(x)\Big)\pi(x).$$

Clearly the marginal density for x is f. A Gibbs sampler can now be used to generate $\{x^{(t)}\}$ which satisfies the conclusion of the theorem. To implement the Gibbs sampler the full conditional densities need to be sampled in turn, updating the parameters as they are sampled. These full conditional densities are given by $f(y|x^{(t)})$, which is the uniform density on the interval $(0, l(x^{(t)}))$, and $f(x^{(t+1)}|y)$, which is $\pi(.)$ restricted to the set $A_y = \{x : l(x) > y\}$. Clearly such a scheme is the one described in Theorem 2, completing the proof.

Theorem 2 says that it is possible to remove the 'ties' from an independent Metropolis chain and remain with a simulated Markov chain from a pure Gibbs sampler. So why not just simulate from this Gibbs sampler in the first place? This is exactly what we are proposing, and intuitively we must improve the efficiency in exploration of the sample space since we have removed the 'stops and starts' of the Metropolis chain.

An additional burden with the Metropolis algorithm is that it may be difficult, in some instances, to obtain a good candidate distribution; see, for example, Chib and Greenberg (1995), who discuss the difficulties in this selection process.

The general case

The above algorithm works when $l(.)$ itself is invertible. It is more usual for it to decompose into a product of invertible functions. Therefore, we consider the case when

$$f(x) \propto \left\{\prod_{i=1}^{n} l_i(x)\right\}\pi(x).$$

Here we introduce the latent variable $y = (y_1, ..., y_n)$, where the y_i are mutually independent given x, and define the joint density of x and y by

$$f(x,y) \propto \left\{\prod_{i=1}^{n} I\left(y_i < l_i(x)\right)\right\}\pi(x).$$

The full conditional densities are given by $f(y_i|x)$, independent uniform densities on the intervals $(0, l_i(x))$, and $f(x|y)$, $\pi(.)$ restricted to the interval $A_y = \{x : l_i(x) > y_i, i = 1, ..., n\}$.

A simplification in the multivariate case

If x is multidimensional and it is not possible to obtain the multivariate set A_u then a simplification is to sample from $f(x|y)$ by sampling from $f(x_k|x_{-k}, y)$, for $k = 1, ..., p$, where p is the dimension of x. This would involve sampling from $\pi(x_k|x_{-k})$ restricted to the set $A_{ku} = \{x_k : l(x_k, x_{-k}) > y)\}$.

In the discussion above it is not required that l and π be a likelihood function and a prior dsitribution, respectively. More generally, suppose $f = h \times g$, and we wish to sample from f using g as a proposal distribution for a rejection algorithm within a Gibbs sampler. A standard rejection algorithm would proceed by first calculating the supremum of h. Using an argument similar to the one appearing in Theorem 2, it is easy to prove that the method proposed in Theorem 1 will be at least as efficient as a standard rejection algorithm; this, of course, means that the method developed in this section will be at least as efficient as *any* variant of the standard rejection algorithm, when used within a Gibbs sampler.

3 Bayesian nonconjugate models

Example 3.1 Poisson/log-normal model

Suppose we observe a random nonnegative integer, n, from a Poisson distribution with parameter $\exp(x)$. Without loss of generality we assume a $N(0, 1)$ prior for x. Then we obtain the posterior given by

$$f(x) \propto \exp\left(nx - \exp(x)\right) \exp(-0.5x^2),$$

where we assume without loss of generality that the prior is normal$(0, 1)$ and n is a nonnegative integer. We introduce the latent variable Y, defined on the interval $(0, \infty)$, such that the joint density with X is given by

$$f(x, y) \propto \exp(-y) I\left(y > \exp(x)\right) \exp\left(-0.5(x^2 - 2nx)\right),$$

which leads to the conditional densities given by

$$f(y|x) \propto \exp(-y)I(y > \exp(x))$$

and

$$f(x|y) \propto \exp\left(-0.5(x-n)^2\right) I\left(x < \log(y)\right),$$

a truncated $N(n,1)$ density; see Devroye (1986); Robert (1995); Cumbus et al. (1997) for information concerning the sampling of truncated normal densities.

Example 3.2 Bernoulli/logistic regression model

Consider the following Bernoulli regression model for which

$$w_i|[X = x], z_i \sim \text{Bernoulli}\left(\{1 + \exp(-\mu - xz_i)\}^{-1}\right),\ i = 1, ..., n,$$

with $X \sim N(0,1)$ as the prior (we assume μ is known). The posterior density for X is given, up to a constant of proportionality, by

$$f(x) \propto \exp(-0.5x^2) \prod_{i=1}^{n} \left(\{1 + \exp(-\mu - xz_i)\}^{-w_i} \{1 + \exp(\mu + xz_i)\}^{w_i - 1}\right).$$

We introduce the latent variables $u = (u_1, ..., u_n)$ and $v = (v_1, ..., v_n)$ such that their joint density with X is given, up to a constant of proportionality, by

$$f(x, u, v) \propto \exp(-0.5x^2)$$

$$\times \prod_{i=1}^{n} I\left(u_i < \{1 + \exp(-\mu - xz_i)\}^{-w_i}, v_i < \{1 + \exp(\mu + xz_i)\}^{w_i - 1}\right).$$

The full conditional densities $f(u_i|u_{-i}, v, x)$ and $f(v_i|v_{-i}, u, x)$ are all uniform:

$$f(u_i|u_{-i}, v, x) = \mathcal{U}\left(0, \{1 + \exp(-\mu - xz_i)\}^{-w_i}\right)$$

and

$$f(v_i|v_{-i}, u, x) = \mathcal{U}\left(0, \{1 + \exp(\mu + xz_i)\}^{w_i - 1}\right),$$

where $\mathcal{U}(\dashv, \lfloor)$ is the uniform distribution on (a, b). Let $\mathcal{S} = \{i : w_i = 1\} \cap \{i : z_i \neq 0\}$ and $\mathcal{R} = \{i : w_i = 0\} \cap \{i : z_i \neq 0\}$. Then

$$f(x|u, v) \propto \exp(-0.5x^2)I(x \in A_{uv}),$$

where $A_{uv} = (\max_{i\in\mathcal{S}}\{a_i\}, \min_{i\in\mathcal{R}}\{b_i\})$, $a_i = \{\log(1/u_i - 1) - \mu\}/z_i$ and $b_i = \{\log(1/v_i - 1) - \mu\}/z_i$. Note that if $\mathcal{S} = \emptyset$ then we replace $\max_{i\in\mathcal{S}}\{a_i\}$ by $-\infty$ and if $\mathcal{R} = \emptyset$ then we replace $\min_{i\in\mathcal{R}}\{b_i\}$ by $+\infty$.

Example 3.3 Probit model

Here we consider the posterior density given, up to a constant of proportionality, by

$$f(\beta) \propto \prod_{i=1}^{n} \{\Phi(\beta_0 + \beta_1 z_i)\}^{w_i} \prod_{i=1}^{n} \{1 - \Phi(\beta_0 + \beta_1 z_i)\}^{n_i - w_i} \pi(\beta),$$

where we assume a multivariate normal(μ, Σ) prior for β, and Φ is the standard normal distribution function. We introduce the latent variables $U = (U_1, ..., U_n)$ and $V = (V_1, ..., V_n)$ such that their joint density with β is given, up to a constant of proportionality, by

$$f(\beta, u, v) \propto \prod_{i=1}^{n} I\left(u_i < \{\Phi(\beta_0 + \beta_1 z_i)\}^{w_i}, v_i < \{1 - \Phi(\beta_0 + \beta_1 z_i)\}^{n_i - w_i}\right) \pi(\beta).$$

The full conditional densities $f(u_i|u_{-i}, v, \beta)$ and $f(v_i|v_{-i}, u, \beta)$ are uniform and so we focus attention on $f(\beta_k|\beta_{-k}, u, v)$. Let $a_i = \Phi^{-1}(\tau_i) - \beta_1 z_i$, $b_i = (\Phi^{-1}(\tau_i) - \beta_0)/z_i$, $c_i = \Phi^{-1}(\lambda_i) - \beta_1 z_i$ and $d_i = (\Phi^{-1}(\lambda_i) - \beta_0)/z_i$, where $\tau_i = u_i^{1/w_i}$ and $\lambda_i = 1 - v_i^{1/(n_i - w_i)}$. Note that b_i and d_i are only defined for those $z_i \neq 0$, a_i and b_i are only defined when $w_i > 0$ and c_i and d_i are only defined when $n_i > w_i$. Then

$$f(\beta_0|\beta_1, u, v) \propto \pi(\beta_0|\beta_1) I\left(\max_i\{a_i\} < \beta_0 < \min_i\{c_i\}\right)$$

and

$$f(\beta_1|\beta_0, u, v) \propto \pi(\beta_1|\beta_0) I\left(\max_i\{b_i\} < \beta_1 < \min_i\{d_i\}\right).$$

Example 3.4 Weibull proportional hazards model

The Weibull proportional hazards model is popular for modelling censored survival time data. The hazard function for the ith individual is given by

$$\lambda_i(t) = \lambda_0(t)\exp(X_i\beta),$$

where $\beta = (\beta_1, ..., \beta_p)$ is a vector of unknown parameters and $\lambda_0(t)$ is the baseline hazard. The Weibull model arises when $\lambda_0(t) = \alpha t^{\alpha-1}$ for some

unknown $\alpha > 0$. The conditional posterior distribution for β, given α and taking a normal multivariate normal prior for β, is given, up to a constant of proportionality, by

$$f(\beta) \propto \prod_{i=1}^{n} \exp\left(X_i\beta I(\delta_i = 0) - t_i^{\alpha}\exp(X_i\beta)\right)\exp\left(-0.5(\beta-\mu)'\Sigma^{-1}(\beta-\mu)\right),$$

where $\delta_i = 0$ indicates that t_i is an uncensored observation. Here we introduce the latent variable $U = (U_1, ..., U_n)$ such that the joint density with β is given, up to a constant of proportionality, by

$$f(\beta, u) \propto \prod_{i=1}^{n} e^{-u_i} I\left(u_i > t_i^{\alpha} e^{X_i\beta}\right)\exp\left(-0.5(\beta-\mu)'\Sigma^{-1}(\beta-\mu) + \nu\beta\right),$$

where $\nu = \sum_{i=1}^{n} X_i I(\delta_i = 0)$. The full conditional distributions for each of the u_i are independent exponential(1) distributions restricted to the sets $(t_i^{\alpha}\exp(X_i\beta), \infty)$. Sampling from $f(\beta_k|\beta_{-k}, u)$ requires

$$A_{ku} = \left\{\beta_k : \beta_k < \min_i\{\log(u_i/t_i^{\alpha})/X_{ki} - \sum_{l\neq k} X_{li}\beta_l/X_{ki}\}\right\}.$$

The full conditional for α with prior $\pi(\alpha)$ =constant (Dellaportas and Smith, 1993) is given by

$$\alpha^{\tilde{n}}\left(\prod_{\delta_i=0} t_i\right)^{\alpha} I\left(\max_{t_i<1}\left\{\frac{\log(u_i) - X_i\beta}{\log t_i}\right\} < \alpha < \min_{t_i>1}\left\{\frac{\log(u_i) - X_i\beta}{\log t_i}\right\}\right),$$

where $\tilde{n}$ is the number of uncensored observations. We can sample this density via the introduction of a latent variable V and define the joint density with α by

$$f(v, \alpha) \propto \alpha^{\tilde{n}} I\left(v < \left(\prod_{\delta_i=0} t_i\right)^{\alpha}\right) I(\lambda^- < \alpha < \lambda^+),$$

where λ^- and λ^+ are the bounds appearing in the full conditional for α. It is now seen that both $f(v|\alpha)$ and $f(\alpha|v)$ are of standard type and can be sampled using uniform random variables.

4 Bayesian hierarchical models

Hierarchical models are relevant when the observed variability in the data on a number of units can be conveniently partitioned into *within-* and *between-*unit components. At the first stage of the hierarchy observations from a particular unit are modelled, whilst at the second stage of the hierarchy between unit differences are modelled. In this paper we consider both (i) *generalised linear mixed models* and (ii) *nonlinear mixed models.* We concentrate on that situation in which the second stage distribution is specified parametrically, typically using normal or Student's t distributions. The use of such models is becoming increasingly common as both computational power increases and new computational techniques are developed. For example GLMMs are used in health services research (Gatsonis et al., 1995; Kahn and Raftery, 1996), disease mapping (Clayton and Kaldor, 1987), multicentre clinical trials (Skene and Wakefield, 1990), educational testing (Goldstein, 1995) and small area estimation (Ghosh and Rao, 1994).

NMMs are used, for example, in growth curve analysis (Berkey, 1982), and population pharmacokinetic/pharmacodynamic studies (Beal and Sheiner, 1982; Wakefield, 1996).

Computationally NMMs and GLMMs pose two problems; first, interest generally focuses on the second stage parameters and to obtain the likelihood for these parameters the unit-specific first-stage parameters (the random effects) need to be integrated out. For the models considered here these integrals are analytically intractable. Secondly problems arise when summarising the resultant marginal likelihood function or the posterior distribution if a Bayesian approach is taken. From a non-Bayesian perspective various approaches have been suggested including Maximum Likelihood Estimation, Generalized Estimating Equations and Quasi-Likelihood. There is now a large amount of literature on estimation in GLMMs; see, for example, Williams (1982), Breslow (1984), Stiratelli, Laird and Ware (1984), Gilmour, Anderson and Rae (1985), Liang and Zeger (1986), Goldstein (1987), Zeger, Liang and Albert (1988), Schall (1991) and Breslow and Clayton (1993).

Bayesian solutions for GLMMs have been suggested using the Laplace method (Kass and Steffey, 1989) and numerical integration (Skene and Wakefield, 1990). Neither of these techniques is completely satisfactory, however. Laplacian methods and numerical integration become infeasible as the dimensionality of the parameter space increases and for a given model and

dataset it is difficult to assess whether the posterior distribution is sufficiently well-behaved for the analytical and numerical approximations to be appropriate. MCMC techniques are far more appealling since they allow the complete posterior surface to be examined. Unfortunately for GLMMs the required conditional distributions do not assume standard forms and so specialist code is required. Zeger and Karim (1991) proposed the Gibbs sampler as a method for GLMMs but their rejection algorithm was not guaranteed to provide a bounding envelope and so strictly the Markov chain did not have the correct limiting distribution, though Tierney (1994) gave a Metropolis algorithm to correct for this. For a GLM with appropriate priors Dellaportas and Smith (1993) showed that the required conditional distributions are log-concave so the *adaptive rejection algorithm* (ARS) (Gilks and Wild, 1992) can be used. This approach could be adopted within a hierarchical framework and examples have been presented using the BUGS software (Spiegelhalter et al., 1995).

We turn now to NMMs. A good review of this area is provided by Davidian and Giltinan (1995). Non-Bayesian computational techniques have been suggested by, amongst others, Lindstrom and Bates (1990), Vonesh and Carter (1992) and Walker (1996).

Fearn (1975) provided an early Bayesian solution to growth curve analysis and Racine-Poon (1985) provided an EM-type algorithm. However it was not until MCMC techniques became available that the Bayesian approach was feasible generally. As with GLMMs the conditional distributions for the random effects do not assume standard forms. Wakefield et al. (1994) used the ratio-of-uniforms black-box random number generation method (Wakefield, Gelfand and Smith, 1991) but this technique is computationally expensive since numerical maximizations are required at each iteration and also requires specialist software. The Metropolis algorithm is an obvious candidate but this requires 'tuning' by the user. In this paper we provide an algorithm which, via the introduction of latent variables, provides an MCMC solution with all sampling being from standard forms.

4.1 Generalised linear mixed models

4.1.1 The model

Given $\{b_i\}$, a set of q-vector random effects, the observations $y_i : i = 1, ..., n$ are conditionally independent from the exponential family of distributions with mean $h(X_i\beta + Z_i b_i)$, where $h(.)$ is a non-negative invertible function, that is, $g^{-1} = h$ exists, X_i is a p-vector of explanatory variables, β a p-vector of unknown parameters, and Z_i a q-vector of explanatory variables, for the ith observation. The conditional variances are given by $\text{var}(y_i|b_i) = \phi v\{E(y_i|b_i)\}$ where v is a known variance function and ϕ an unknown dispersion parameter. The b_i are assumed to be independent and identically distributed (iid) from the multivariate normal distribution with mean 0 and covariance matrix Ω. Within a Bayesian framework conjugate prior distributions are assigned to the parameters ϕ, β and Ω. The prior for ϕ (if present) is typically an inverse gamma distribution, the prior for β a multivariate normal prior, say $N(\mu, \Sigma)$, and an inverse Wishart prior for Ω.

4.1.2 The algorithm

Here we present a general algorithm for sampling the conditional distributions of the GLMM. Suppose the full conditional distribution for β is given, up to a constant of proportionality, by

$$f(\beta) \propto \exp\Big(y_i X_i\beta - h(X_i\beta + Z_i b_i)\Big) N(\beta|\mu, \Sigma).$$

In this form the distribution is not of standard type and so cannot be sampled directly without recourse to specific software. However, with the introduction of latent variables standard forms can be recovered. We proceed by introducing the latent variables $u = (u_1, ..., u_n)$ and $v = (v_1, ..., v_n)$ such that the joint (full conditional) distribution with β is given, again up to a constant of proportionality, by

$$f(\beta, u, v) \propto \left\{\prod_{i=1}^{n} I\Big(u_i < \exp(y_i X_i\beta), v_i < \exp\{-h(X_i\beta + Z_i b_i)\}\Big)\right\} N(\beta|\mu, \Sigma).$$

Clearly the marginal distribution for β is as required. Some simple algebra gives the following full conditional distributions for each $\beta_k, k = 1, ..., p$,

$$f(\beta_k) \propto N(\beta_k|\mu_k^*, 1/e_{kk}) I(a_k < \beta_k < c_k),$$

where

$$\mu_k^* = \mu_k - \sum_{l \neq k} (\beta_l - \mu_l) e_{lk}/e_{kk},$$

e_{lk} is the lkth element of Σ^{-1},

$$a_k = \max_{X_{ki} \neq 0} \left\{ \left(y_i^{-1} \log u_i - \sum_{l \neq k} \beta_l X_{li} \right) / X_{ki} \right\}$$

and

$$c_k = \min_{X_{ki} \neq 0} \left\{ \left(g(-\log v_i) - \sum_{l \neq k} \beta_l X_{li} - Z_i b_i \right) / X_{ki} \right\}.$$

The introduction of the latent variables provides what can be described as a latent model and the 'new' Gibbs sampler includes the sampling of the full conditional distributions for u and v within each iteration. These are easily seen to be uniform distributions. The full conditional distribution for b_i is given by

$$f(b_i) \propto \left\{ \prod_{i=1}^{n} I\Big(u_i < \exp\{y_i Z_i b_i\}, v_i < \exp\{-h(X_i\beta + Z_i b_i)\}\Big) \right\} N(b_i|0, \Omega)$$

which, as with the full conditional for β, will lead to a truncated normal distribution.

4.1.3 Examples

Example 4.1.1 Random effects binomial model.
The model considered here is a random effects binomial model which allows for over-dispersion. If p_i is the probability of success for the ith observation then

$$y_i|p_i \sim \text{binomial}(p_i, n_i),$$
$$\text{logit} p_i = X_i\beta + b_i,$$
$$b_i \sim N(0, \lambda).$$

Independent priors are assigned to λ and β, typically a gamma and a multivariate normal, respectively. Of interest is the joint probability distribution of β, $b = (b_1, ..., b_n)$ and λ given, up to a constant of proportionality, by

$$f(\beta, b, \lambda) \propto \left\{ \prod_{i=1}^{n} \frac{\exp(y_i\theta_i)}{(1+\exp\theta_i)^{n_i}} \exp\Big(-0.5 b_i^2 \lambda\Big) \right\} \lambda^{n/2} \pi(\lambda, \beta),$$

where $\theta_i = X_i\beta + b_i$. Here we introduce the latent variables $u = (u_1, ..., u_n)$ and $v = (v_1, ..., v_n)$ such that the joint distribution with β, b and λ is given, again up to a constant of proportionality, by

$$f(\beta, b, \lambda, u, v) \propto \lambda^{n/2}\pi(\lambda, \beta)\times$$

$$\Big\{\prod_{i=1}^{n} e^{-u_i y_i - v_i(n_i - y_i)} I\Big(u_i > \log\{1 + e^{-\theta_i}\}, v_i > \log\{1 + e^{\theta_i}\}\Big)\exp\Big(-0.5b_i^2\lambda\Big)\Big\}.$$

The full conditional distribution for β_k is given by

$$f(\beta_k) \propto \pi(\beta_k|\beta_{-k})I\Big(\beta_k \in A_k\Big),$$

where A_k is the set

$$\Big(\max_{X_{ki}\neq 0}\Big\{\Big[-\log\{e^{u_i} - 1\} - \Sigma_{l\neq k}X_{li}\beta_l - b_i\Big]/X_{ki}\Big\},$$

$$\min_{X_{ki}\neq 0}\Big\{\Big[\log\{e^{v_i} - 1\} - \Sigma_{l\neq k}X_{li}\beta_l - b_i\Big]/X_{ki}\Big\}\Big).$$

The full conditional distribution for b_i is given by

$$f(b_i) \propto \exp\Big(-0.5b_i^2\lambda\Big)I\Big(b_i \in A_i\Big),$$

where A_i is the set

$$\Big(-\log\{e^{u_i} - 1\} - \Sigma_k X_{ki}\beta_k, \log\{e^{v_i} - 1\} - \Sigma_k X_{ki}\beta_k\Big).$$

The full conditional distributions for the latent variables are given by

$$f(u_i) \propto \exp(-u_i y_i)I\Big(u_i > \log\{1 + e^{-\theta_i}\}\Big),$$

a truncated exponential(y_i) distribution, and

$$f(v_i) \propto \exp\Big(-v_i(n_i - y_i)\Big)I\Big(v_i > \log\{1 + e^{\theta_i}\}\Big)$$

and the full conditional for λ is

$$f(\lambda) \propto \lambda^{n/2}\exp\Big(-\lambda/2\sum_i b_i^2\Big)\pi(\lambda).$$

All of these full conditionals are of known types. Only minor modifications are required if $y_i = 0$ or $y_i = n_i$.

Example 4.1.2 Random effects Poisson model.
Here we consider the random effects Poisson model given by

$$y_i|\theta_i \sim \text{Poisson}\big(\exp\theta_i\big),$$
$$\theta_i = X_i\beta + b_i,$$
$$b_i \sim N(0,\lambda).$$

Priors for β and λ are taken as in Example 1. The joint probability distribution of β, b and λ is given by

$$f(\beta,b,\lambda) \propto \exp\Big\{\sum_{i=1}^{n}\big(y_i\theta_i - \exp\theta_i - \lambda/2b_i^2\big)\Big\}\lambda^{n/2}\pi(\lambda,\beta).$$

Here we introduce the latent variables $u = (u_1,...,u_n)$ and $v = (v_1,...,v_n)$ such that the joint distribution with β, b and λ is given, up to a constant of proportionality, by

$$f(\beta,b,\lambda,u,v) \propto \lambda^{n/2}\pi(\beta,\lambda)\Big\{\prod_{i=1}^{n} e^{-v_i} I\big(u_i < e^{-y_i\theta_i}, v_i > e^{\theta_i}\big)\exp\big(-0.5b_i^2\lambda\big)\Big\}.$$

The full conditional distribution for β_k is given by

$$f(\beta_k) \propto \pi(\beta_k|\beta_{-k})I\big(\beta_k \in A_k\big),$$

where A_k is the set

$$\Big(0, \min_{X_{ki}\neq 0}\Big\{\Big[\log\{v_i\} - \Sigma_{l\neq k}X_{li}\beta_l - b_i\Big]/X_{ki}, \Big[-y_i^{-1}\log\{u_i\} - \Sigma_{l\neq k}X_{li}\beta_l - b_i\Big]/X_{ki}\Big\}\Big).$$

The full conditional distribution for b_i is

$$f(b_i) \propto \exp\big(-0.5b_i^2\lambda\big)I\big(b_i \in A_i\big),$$

where A_i is the set

$$\Big(0, \min\Big\{\log\{v_i\} - \Sigma_k X_{ki}\beta_k, -y_i^{-1}\log\{u_i\} - \Sigma_k X_{ki}\beta_k\Big\}\Big).$$

The full conditional distributions for the latent variables are given by

$$f(u_i) \propto I\big(u_i < \exp(-y_i\theta_i)\big),$$

$$f(v_i) \propto \exp(-v_i)I\big(v_i > \exp(\theta_i)\big)$$

and the full conditional for λ is

$$f(\lambda) \propto \lambda^{n/2}\exp\big(-\lambda/2\sum_i b_i^2\big)\pi(\lambda).$$

Again, only minor modifications are required for the case when $y_i = 0$.

4.2 Nonlinear mixed models

4.2.1 The model

In the following let i ($i = 1, ..., n$) index individuals and j ($j = 1, ..., n_i$), with $N = \sum_i n_i$, index observations within individuals. Let y_{ij} represent an observation or a transformation of the observation (for example the logarithm). The conditional probability model for the observations is given by

$$y_{ij}|\theta_i \sim N\Big(g(\theta_i, x_{ij}), \sigma^2\Big),$$

where θ_i is the random effect associated with the ith individual, x_{ij} an explanatory variable for the ijth observation and g a known nonlinear mean response function. From now on we will write $g(\theta_i, x_{ij})$ as $g_j(\theta_i)$. The θ_is are assumed to be normally distributed with mean μ and variance-covariance matrix Σ. Here σ, μ and Σ are the population parameters. Conjugate priors are assigned to these parameters in a manner described in Wakefield et al. (1994). Note that the full conditional density for θ_i is given, up to a constant of proportionality, by

$$f(\theta_i) \propto \Big\{ \prod_{j=1}^{n_i} \exp\Big(-0.5\, l_j(\theta_i)/\sigma^2 \Big) \Big\} \pi(\theta_i),$$

where $l_j(\theta_i) = (y_{ij} - g_j(\theta_i))^2$ and $\pi(\theta_i) = N(\theta_i|\mu, \Sigma)$. It is not possible to sample this distribution directly without specialist random number generation techniques. Note that here the ratio-of-uniforms method may be used but requires three numerical maximisations for each sample. The adaptive rejection sampling routine cannot be used since the conditional distributions are typically not log-concave. Gilks, Best and Tan (1995) proposed the Metropolis adaptive rejection sampling algorithm for such cases.

We can write this model in a different way by introducing a (latent) random effect u_{ij} for each observation. This latent model is given by

$$y_{ij}|u_{ij}, \theta_i \sim \mathcal{U}\Big(g_j(\theta_i) - \sqrt{u_{ij}}, g_j(\theta_i) + \sqrt{u_{ij}}\Big),$$

and

$$u_{ij} \sim G(3/2, \lambda/2),$$

where G denotes the gamma distribution and $\lambda = 1/\sigma^2$. It is easily seen that integrating over the u_{ij} returns the original model. Now the full conditional

distributions for the random effects are given by

$$f(\theta_i) \propto N(\theta_i|\mu, \Sigma) I(\theta_i \in A_i),$$

where

$$A_i = \left\{\theta_i : y_{ij} - \sqrt{u_{ij}} < g_j(\theta_i) < y_{ij} + \sqrt{u_{ij}} : j = 1, ..., n_i\right\}.$$

The full conditional distributions for the latent variables are given by

$$f(u_{ij}) \propto \exp(-\lambda u_{ij}/2) I(u_{ij} > [y_{ij} - g_j(\theta_i)]^2.$$

The full conditional distribution for λ, with prior λ^{-1}, is given by

$$G\left(3N/2, \sum_{i=1}^{n}\sum_{j=1}^{n_i} u_{ij}/2\right).$$

This latent model motivates the algorithm presented in the next section.

4.2.2 The algorithm

The target distribution is given, up to a constant of proportionality, by

$$f(\theta) \propto \left\{\prod_{j=1}^{m} \exp\left(-0.5\lambda l_j(\theta)\right)\right\}\pi(\theta),$$

where $l_j(\theta) > 0$ for all θ (here we have removed the subscripts i and we put $m = n_i$). The aim is to introduce a latent variable and use Gibbs sampling to generate random variates from the target distribution. Since inference is already enveloped within a Gibbs sampler the only extra computation required is the sampling of the full conditional distributions for each of the latent variables within each iteration. We define the joint density for θ and $u = (u_1, ..., u_m)$, given up to a constant of proportionality, by

$$f(\theta, u) \propto \prod_{j=1}^{m} \left\{\exp(-0.5\lambda u_j) I\left(u_j > l_j(\theta)\right)\right\}\pi(\theta).$$

Clearly the marginal density for θ has the required target density $f(.)$. The full conditional densities, required for the Gibbs sampling, are given by

$$f(u_j|\theta) \propto \exp(-0.5\lambda u_j) I\left(u_j > l_j(\theta)\right),$$

for $j = 1, ..., m$, and

$$f(\theta|u) \propto \pi(\theta) I(\theta \in A_u),$$

where $A_u = \{\theta : l_j(\theta) < u_j, j = 1, ..., m\}$.

Sampling from $f(u_j|\theta)$ is straightforward and we assume that sampling from $\pi(.)$ restricted to some set A is also easy (typically, in applications, $\pi(.)$ will be a known multivariate distribution, for example, the normal). Therefore the only remaining difficulty is the determination of the set A_u. It should first be pointed out that A_u is not empty. Within a Gibbs sampling algorithm it is easily seen that the current θ must be a member of A_u. If it is not possible to obtain the multivariate set A_u then an alternative approach is to sample from $f(\theta|u)$ by sampling from $f(\theta_k|\theta_{-k}, u)$ for $k = 1, ..., p$ where p is the dimension of θ. This would involve sampling from $\pi(\theta_k|\theta_{-k}) I(\theta_k \in A_{ku})$ where $A_{ku} = \{\theta_k : l_j(\theta_k, \theta_{-k}) < u_j, j = 1, ..., m\}$. Clearly each A_u or A_{ku} will depend on the likelihood $l_j(.)$. We now consider some examples.

Example 4.2.1 One compartment pharmacokinetic model.

Here the logged data is assumed to be normally distributed. This gives an approximate constant coefficient of variation, which in such applications mimics assay precision. A Bayesian analysis of a population pharmacokinetic data set (with a one compartment model) then involves the simulation of $\pi^*(.)$ with

$$l_j(\theta) = \Big(\log y_j - \log d - \theta_1 + x_j \exp(\theta_2)\Big)^2.$$

Here y_j represents the measured concentration of a drug at time x_j after administration of a dose of size d at time $x = 0$. Due to the difficult task of obtaining the two dimensional set A_u we concentrate on obtaining A_{1u} and A_{2u}. In this case let $z_j = \log y_j - \log d$. Then $l_j(\theta) < u_j$ implies

$$\Big(z_j - \theta_1 + x_j \exp(\theta_2)\Big)^2 < u_j.$$

This leads to

$$A_{1u} = \Big(\max_j\{a_j\}, \min_j\{b_j\}\Big),$$

where $a_j = z_j - \sqrt{u_j} + x_j \exp(\theta_2)$ and $b_j = z_j + \sqrt{u_j} + x_j \exp(\theta_2)$. If $\max_j\{\theta_1 - z_j - \sqrt{u_j}\} > 0$ then

$$A_{2u} = \Big(\max_{j \in \mathcal{S}}\{\alpha_j\}, \min_j\{\beta_j\}\Big),$$

where $S = \{j : \theta_1 - z_j - \sqrt{u_j} > 0\}$, $\alpha_j = \log\{(\theta_1 - z_j - \sqrt{u_j})/x_j\}$ and $\beta_j = \log\{(\theta_1 - z_j + \sqrt{u_j})/x_j\}$. If $\max_j\{\theta_1 - z_j - \sqrt{u_j}\} \leq 0$ then

$$A_{2u} = \big(-\infty, \min_j\{\beta_j\}\big).$$

Note that $\theta_1 - z_j + \sqrt{u_j} > 0$ for all j. Therefore if $\pi(.)$ is a bivariate normal distribution then sampling from $f(\theta_k|\theta_{-k}, u)$ is the sampling of a truncated univariate normal distribution.

Example 4.2.2 Logistic model.

For the logistic model we obtain

$$l_j(\theta) = \big(z_j - \theta_1 + \log\{1 + \exp(\theta_2 + \theta_3 x_j)\}\big)^2.$$

Therefore

$$A_{1u} = \big(\max_j\{a_j\}, \min_j\{b_j\}\big),$$

where $a_j = z_j - \sqrt{u_j} + \log\{1 + \exp(\theta_2 + \theta_3 x_j)\}$ and $b_j = z_j + \sqrt{u_j} + \log\{1 + \exp(\theta_2 + \theta_3 x_j)\}$. Let $\mathcal{S} = \{j : \exp(\theta_1 - \sqrt{u_j} - z_j) > 0\}$. If $\mathcal{S} \neq \emptyset$ then

$$A_{2u} = \big(\max_{j\in\mathcal{S}}\{\alpha_j\}, \min_j\{\beta_j\}\big),$$

where $\alpha_j = \log\{\exp(\theta_1 - \sqrt{u_j} - z_j) - 1\} - \theta_3 x_j$ and $\beta_j = \log\{\exp(\theta_1 + \sqrt{u_j} - z_j) - 1\} - \theta_3 x_j$ (note that $\theta_1 + \sqrt{u_j} - z_j) > 0$). If $\mathcal{S} = \emptyset$ then

$$A_{2u} = \big(-\infty, \min_j\{\beta_j\}\big).$$

Finally, if $\mathcal{S} \neq \emptyset$,

$$A_{3u} = \big(\max_{j\in\mathcal{S}}\{\gamma_j\}, \min_j\{\delta_j\}\big),$$

where $\gamma_j = [\log\{\exp(\theta_1 - \sqrt{u_j} - z_j) - 1\} - \theta_2]/x_j$ and $\delta_j = [\log\{\exp(\theta_1 + \sqrt{u_j} - z_j) - 1\} - \theta_2]/x_j$. If $\mathcal{S} = \emptyset$ then

$$A_{3u} = \big(-\infty, \min_j\{\beta_j\}\big).$$

Typically $\pi(\theta)$ will be the normal distribution $N(\theta|\mu, \Sigma)$. Then $f(\theta_k|\theta_{-k}, \mu, \Sigma)$ will be the univariate normal distribution $N(\theta_k|\mu_k^*/e_{kk}.1/e_{kk})$, where $\mu_k^* = \mu_k e_{kk} - \sum_{l\neq k} e_{lk}(\theta_l - \mu_l)$ and e_{lk} is the lkth element of Σ^{-1}. The algorithm reduces therefore to the sampling of the univariate normal distributions $N(\theta_k|\mu_k^*/e_{kk}, 1/e_{kk})$ restricted to the set A_{ku}.

5 Numerical examples

Example 5.1 Generalised linear model with logit link.

Our example is the binomial GLM with a logit link function and a quadratic logistic model given by

$$y_i|\pi_i \sim \text{binomial}(n_i, \pi_i)$$

and

$$\log(\pi_i/(1-\pi_i)) = \beta_1 + Z_i\beta_2 + Z_i^2\beta_3 = X_i\beta, \;\; i = 1, ..., n.$$

Further details are provided in Dellaportas and Smith (1993). With a multivariate normal prior for β, say $N(\mu, \Sigma)$, the posterior distribution is given by

$$f(\beta) \propto \left\{\prod_{i=1}^{n} e^{y_i X_i\beta}/(1+e^{X_i\beta})^{n_i}\right\} \exp\Big(-0.5(\beta-\mu)'\Sigma^{-1}(\beta-\mu)\Big).$$

We introduce the latent variable $u = (u_1, ..., u_n)$ such that the joint density with β is given, up to a constant of proportionality, by

$$f(\beta, u) \propto \prod_{i=1}^{n} I\Big(u_i < [1+\exp(X_i\beta)]^{-n_i}\Big) \exp\Big(-0.5(\beta-\mu)'\Sigma^{-1}(\beta-\mu) + \nu\beta\Big),$$

where $\nu = \sum_{i=1}^{n} y_i X_i$. The full conditional distributions for each of the U_i are uniform,

$$f(u_i|u_{-i}, \beta) = \mathcal{U}\Big(0, [1+\exp(X_i\beta)]^{-n_i}\Big).$$

The condition $u_i < (1+e^{X_i\beta})^{-n_i}$ implies $\exp(X_i\beta) < 1/u_i^{1/n_i} - 1$. Therefore, define the sets

$$A_{ku} = \Big\{\beta_k : \beta_k < \min_i\{\log(1/u_i^{1/n_i} - 1)/X_{ki} - X_{li}\beta_l/X_{ki} - X_{mi}\beta_m/X_{ki}\}\Big\},$$

where $\{k, l, m\}$ are, in some order, the elements $\{1, 2, 3\}$. Sampling from $f(\beta|u)$ can now be done by sampling successively from $f(\beta_k|\beta_{-k}, u)$ which involves sampling from a univariate normal distribution restricted to the set A_{ku}. This univariate normal distribution is given by $\pi(\beta_k|\beta_{-k})$ where $\pi(\beta)$ is the multivariate normal distribution with mean $\mu + \Sigma\nu$ and covariance matrix Σ.

We analyse a data set relevant to the above example. The data set and prior distribution used are given in Dellaportas and Smith (1993). We start the chain by taking β as the location of the prior distribution and then proceed to sample u and then back to β. We ran the chain for 5,000 iterations (taking under 15 seconds) and collected the last 2,000 for parameter estimation. We can report, as was to be expected, that our parameter estimates ($\hat{\beta}_1 = -2.36$, $\hat{\beta}_2 = 0.21$ and $\hat{\beta}_3 = -0.004$) coincide with those obtained by Dellaportas and Smith. These authors used the adaptive rejection sampling scheme (Gilks and Wild, 1992) which depends on the posterior density being log-concave. (We need no such condition.) This example demonstrates that not only is the algorithm simple to code, but it is also quick.

Example 5.2 Random effects logistic regression.

Here we analyse the data set presented in Table 3 of Crowder (1978) which involves binomial data, the proportion of seeds that germinated on each of 21 plates, in a 2×2 factorial layout by seed and type of root extract. The model for analysis is described in Example 1 (Section 2). Here we have

$$\theta_i = \beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2} + \beta_4 x_{i1} x_{i2} + b_i,$$

where x_{i1}, x_{i2} are the seed type and root extract of the ith plate. We encountered some problems with high autocorrelation associated with the Markov chain due to the introduction of the latent variables. To solve this problem we took every 100th sample for inference using a single Markov chain. This reduced the autocorrelation to satisfactory levels. Computing time was about half an hour.

Parameter	Estimate
β_1	-0.547
β_2	0.068
β_3	1.337
β_4	-0.812
σ	0.292

Table 1: Parameter estimates for Example 1

The parameter estimates for the Crowder data set are given in Table 1. The estimates compare well with those obtained using BUGS. Additionally we provide in Figure 1 the ergodic plots for the parameters obtained from the

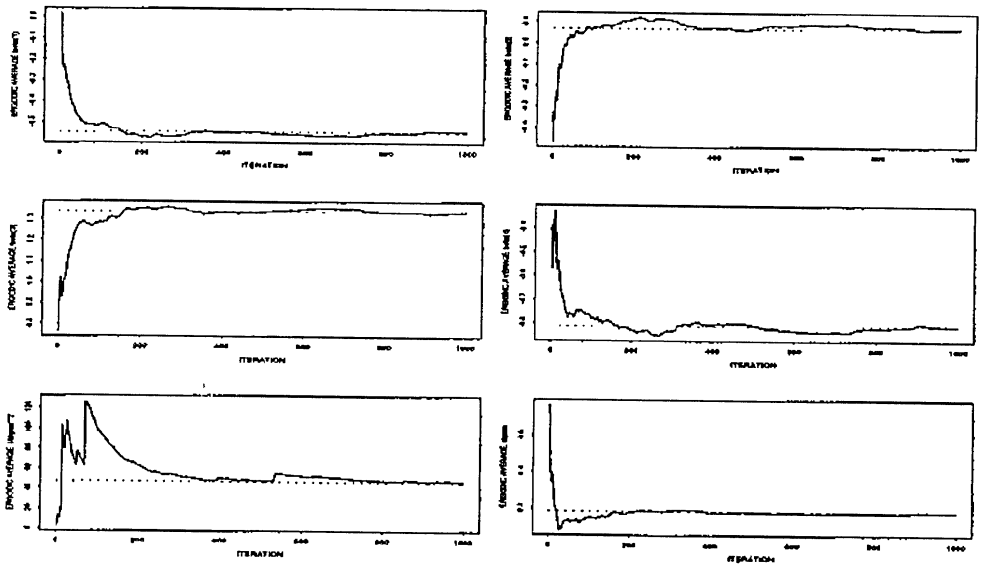

Figure 1: Ergodic averages for Example 1. Parameters are β_1, β_2 (row 1); β_3, β_4 (row 2) and σ^{-2}, σ (row 3).

Markov chain (thinned) output based on a sample of size 1000.

Example 5.3 Nonlinear random effects model.

The second example is taken from the paper of Lindstrom and Bates (1990). The data set consists of 5 (indexed by i) orange trees with 7 (indexed by j) trunk circumference measurements taken for each tree over an interval of time (denoted by x). The logistic model is used to fit the data giving

$$\log y_{ij} = \theta_{1i} - \log\Big\{1 + \exp\big(\theta_{2i} + x_{ij}\theta_{3i}\big)\Big\} + \varepsilon_{ij},$$

where y_{ij} are the observed trunk circumference measurements and ε_{ij} are independent $N(0, \sigma^2)$ error variables. As in Example 1 of this Section, the second stage assumes $\theta_i = (\theta_{1i}, \theta_{2i}, \theta_{3i})$ to be independent $N(\mu, \Sigma)$ variables with priors assigned to (σ^2, μ, Σ). The parameter estimates are presented in Table 2. (We took every 10th sample from the Markov chain output for parameter estimation). These are comparable with the maximum likelihood estimates of Lindstrom and Bates (1990). Histogram representations of the marginal posterior distributions are presented in Figure 2. In Figure 3 we

present the ergodic plots for the parameters based on a sample of size 1000 from the Markov chain (thinned) output. In this example we did not encounter the high autocorrelation which was met in Example 1.

Parameter	Estimate
μ_1	5.304
μ_2	2.041
μ_3	-0.00327
Σ_{11}	0.274
Σ_{22}	0.576
Σ_{33}	0.00203

Table 2: Parameter estimates for Example 2

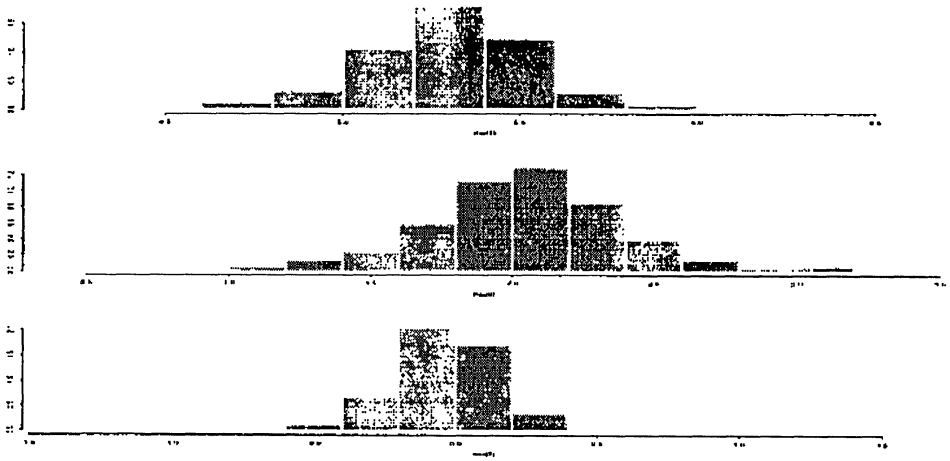

Figure 2: Estimated marginal densities for μ_1 (top), μ_2 (middle) and μ_3 (bottom) from Gibbs output for Example 2

6 Discussion, extensions and conclusions

In Section 5 we presented examples using the auxilliary variable method which resulted in quick and efficient MCMC algorithms. Additionally, the algorithm is easy to code, requiring only standard random variate generation

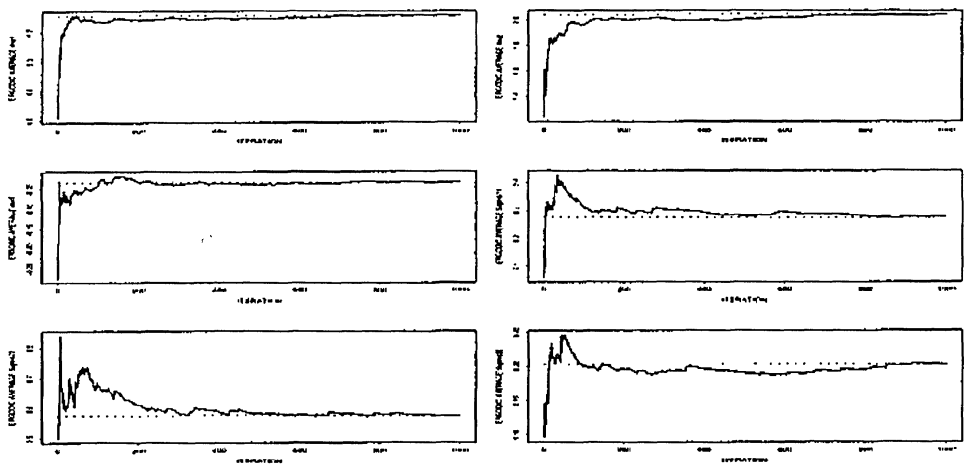

Figure 3: Ergodic averages for Example 2. Parameters are μ_1, μ_2 (row 1); μ_3, Σ_{11} (row 2) and Σ_{22}, Σ_{33} (row 3).

routines. However, we do not claim that superior efficiency will be the case in general. If there is an efficient Metropolis or rejection algorithm then, rather than introducing latent variables, these may be the preferred choice.

A broad question is, "Will a Gibbs sampler with more conditional distributions, all of which are uniform densities, be more efficient than a Gibbs sampler in which some or all of the full conditionals have to be sampled via rejection and/or Hastings type algorithms?" We are not aware of a definitive answer to this question. On the other hand, the question above points to the fact that ease in coding may very well outweigh gains in efficiency especially when the gains using an alternative approach may be neglible: in many contexts, particularly in examples such as the ones described in this paper, and which are very useful in applications, this appears to be a common trend.

In addition to the ease in coding, just like other algorithms, the method proposed in this paper is, in a sense, *general*. In this spirit the results of this paper can be compared to that of Meng and Van Dyk (1997), but from a Bayesian perspective. Generality combined with ease of computation of the latent variable approach, and gains in efficiency relative to other approaches, in a variety of contexts, are compelling reasons for popularising its use among statisticians. For example, improved efficiency in simulation from distributions arising from exponential power, Student t and stable laws, as well as

truncated versions of the latter laws, are also being studied. These results will be reported elsewhere.

From a practical perspective suppose that a "one-off" MCMC algorithm is required. Our method provides an "instant" solution and the only issue is that of convergence (which is the case for all chains). Algorithms based on Metropolis or rejection steps typically require "tuning" and even then there is no guarantee that a more efficient chain will emerge.

Results on rates of convergence are currently only available for narrow classes of models (see, for example, Polson, 1996). Polson, and the discussants of his paper point out that answers to questions such as the one stated in this section remain unresolved. But there is no specific reason why the introduction of latent variables should reduce efficiency. On the contrary, Polson, in his section titled, 'Using Latent Variables To Improve Convergence', reports that "Careful use of latent variables...can lead to vast improvements in efficiency." The examples in Section 4 of Polson's paper give support to the auxilliary variable approach for two types of distribution. Polson indicates there will be improved efficiency for these cases. That there should be a significant reduction in efficiency for *all* other types of distributions, with the introduction of auxilliary variables, may not, of course, follow. But Roberts (personal communication) has shown that our method (which Radford Neal at the University of Toronto first referred to as "slice-sampling") *always* converges geometrically, under *very* mild regularity conditions: this is untrue in general for other MCMC algorithms, which require far more stringent conditions to obtain geometric convergence.

We finally note that the method developed in this paper may be used for random variate generation in general. A comprehensive comparison of alternate methods to random variate generation for sampling well-known densities will be reported elsewhere.

Acknowledgements

The authors are grateful to a Joint Editor, several referees and Adrian Smith for critical comments on earlier drafts of the paper. Thanks, also, to Gareth Roberts for sharing his results on convergence of "slice-sampling" methods.

References

Beal,S.L. and Sheiner,L.B. (1982). Estimating population kinetics. *CRC Critical Reviews in Biomedical Engineering* **8**, 195-222.

Bennett,J.E.,Racine-Poon,A. and Wakefield,J.C. (1996). MCMC for nonlinear hierarchical models. In *Markov chain Monte Carlo methods in practice* (eds. W.R.Gilks,S.Richardson and D.J.Spiegelhalter), pp339-357. London:Chapman and Hall.

Berkey,C.S. (1982). Bayesian approach for a nonlinear growth model. *Biometrics* **38**, 953-961.

Besag, J. and Green, P.J. (1993). Spatial statistics and Bayesian computation. *J. Roy. Statist. Soc. B* **55**, 25-37.

Breslow,N.E. (1984). Extra-Poisson variation in loglinear models. *Applied Statistics* **33**, 38-44.

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

Chib,S. and Greenberg,E. (1995). Understanding the Metropolis-Hastings Algorithm. *Amer. Statist.* **49**, 327-335.

Clayton,D.G. and Kaldor,J. (1987). Empirical Bayes estimates of age standardised relative risks for use in disease mapping. *Biometrics* **43**, 671-682.

Crowder,M.J. (1978). Beta-binomial ANOVA for proportions. *Applied Statistics* **27**, 34-37.

Cumbus,C.,Damien,P. and Walker,S.G. (1997). Sampling truncated multivariate normal and other densities within the Gibbs sampler.

Damien, P. and Walker, S.G. (1996). Sampling probability densities via uniform random variables and a Gibbs sampler. *University of Michigan Business*

School Working Paper.

Davidian,M. and Giltinan,D. (1995). *Nonlinear Models for Repeated Measurement Data.* Chapman and Hall, London.

Dellaportas, P. and Smith,A.F.M. (1993). Bayesian inference for generalised linear and proportional hazards models via Gibbs sampling. *Appl. Statist.* **42**, 443-459.

Devroye, L. (1986). *Non-uniform random variate generation.* Springer, New York.

Edwards, R.G. and Sokal,A.D. (1988). Generalization of the Fortuin-Kasteleyn-Swendsen-Wang representation and Monte Carlo algorithms. *Phys. Rev.* D, **38**, 2009-2012.

Fearn,T. (1975). A Bayesian approach to growth curves. *Biometrika* **62**, 89-100.

Gatsonis,C.,Epstein,A.,Newhouse,J.,Normand,S. and McNeil,B. (1995). Variations in the utilization of coronary angiography for elderly patients with an acute myocardial infarction: an analysis using hierarchical logistic regression. *Medical Care* **33**, 625-642.

Ghosh,M. and Rao,J.N.K. (1994). Small area estimation: an appraisal. *Statistical Science* **9**, 55-93.

Gilks, W.R. and Wild,P. (1992). Adaptive rejection sampling for Gibbs sampling. *Appl. Statist.* **41**, 337-348.

Gilks,W.R.,Best,N.G. and Tan,K.K.C. (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *Applied Statistics* **44**, 455-472.

Gilmour,A.R.,Anderson,R.D. and Rae,A.L. (1985). The analysis of binomial data by a generalized linear mixed model. *Biometrika* **72**, 593-599.

Goldstein,H. (1995). Nonlinear multilevel models with an application to

discrete response data. *Biometrika* **78**, 45-51.

Goldstein,H. (1995). *Multilevel Statistical Models.* London: Edward Arnold.

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

Higdon, D. (1996). Auxilliary variable methods for Markov chain Monte Carlo with applications. *Technical Report*, Duke University.

Kahn,M.J. and Raftery,A.E. (1996). Discharge rates of Medicare stroke patients to skilled nursing facilities: Bayesian logistic regression with unobserved heterogeneity. *Journal of the American Statistical Association* **91**, 29-41.

Kass,R.E. and Steffey,D. (1989). Approximate Bayesian inference in conditionally independent hierarchial models (parametric empirical Bayes models). *Journal of the American Statistical Association* **84**, 717–726.

Liang,K. and Zeger,S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

Lindstrom,M.J. and Bates,D.M. (1990). Nonlinear mixed-effects models for repeated-measures data. *Biometrics* **46**, 673-687.

Meng,X.L. and Van Dyk,D. (1997). The EM algorithm — an old folk-song sung to fast new tune (with discussion). *J. Roy. Statist. Soc. B.* To appear.

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** 1087-1091.

Polson, N.G. (1996). Convergence of Markov chain Monte Carlo Algorithms. In *Bayesian Statistics 5*, pp. 297-321. J.M.Bernardo et al. (Eds.) Oxford University Press.

Racine-Poon, A. (1985). A Bayesian approach to nonlinear random effects

models. *Biometrics* **41**, 1015-1024.

Robert, C.P. (1995). Simulation of truncated normal variables. *Statist. Comput.* **5**, 121-125.

Schall,R. (1991). Estimation in generalized linear models with random effects. *Biometrika* **78**, 719-727.

Skene,A.M. and Wakefield,J.C. (1990). Hierarchical models for multivariate binary response studies. *Statistics in Medicine* **9**, 919-929.

Smith, A.F.M. and Roberts,G.O. (1993). Bayesian computations via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. Roy. Statist. Soc. B* **55**, 3-23.

Spiegelhalter,D.J.,Thomas,A..Best,N.G. and Gilks,W.R. (1995). *BUGS:Bayesian inference Using Gibbs Sampling, Version 0.50.* Cambridge: Medical Research Council Biostatistics Unit.

Stiratelli,R.,Laird,N. and Ware,J.H. (1984). Random effects models for serial observations with binary responses. *Biometrics* **40**, 961-971.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.* **22**, 1701-1762.

Vonesh,E. and Carter,R.L. (1992). Mixed effects nonlinear regression for unbalanced repeated measures. *Biometrics* **48**, 1-17.

Wakefield,J.C.,Smith,A.F.M.,Racine-Poon,A. and Gelfand,A.E. (1994). Bayesian analysis of linear and non-linear population models using the Gibbs sampler. *Applied Statistics* **43**, 201-221.

Wakefield,J.C.,Gelfand,A.E. and Smith,A.F.M. (1991). Efficient generation of random variates via the ratio-of-uniform method. *Statistics and Computing* **1**, 129-133.

Wakefield,J.C. (1996). The Bayesian analysis of population pharmacokinetic

models. *Journal of the American Statistical Association* **91**, 62-75.

Walker, S.G. (1996). An EM algorithm for nonlinear random effects models. *Biometrics* **52**, 934-944.

Williams,D.A. (1982). Extra-binomial variation in logistic linear models. *Applied Statistics* **31**, 144-148.

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

Zeger,S.L.,Liang,K.-Y. and Albert,P.S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics* **44**, 1049-1060.