J. R. Statist. Soc. B (1999)
61, *Part* 2, *pp.* 331–344

Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables

Paul Damien,

University of Michigan, Ann Arbor, USA

Jon Wakefield

Imperial College School of Medicine at St Mary's, London, UK

and Stephen Walker

Imperial College of Science, Technology and Medicine, London, UK

[Received August 1996. Final revision May 1998]

Summary. We demonstrate the use of auxiliary (or latent) variables for sampling non-standard densities which arise in the context of the Bayesian analysis of non-conjugate and hierarchical models by using a Gibbs sampler. Their strategic use can result in a Gibbs sampler having easily sampled full conditionals. We propose such a procedure to simplify or speed up the Markov chain Monte Carlo algorithm. The strength of this approach lies in its generality and its ease of implementation. The aim 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; Non-conjugate 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, however, because the required conditional distributions assume awkward forms. In this case the practitioner may turn to the Metropolis–Hastings algorithm; see, for example, Metropolis *et al.* (1953), Hastings (1970) and Tierney (1994). Unfortunately, these algorithms may be difficult to set up and in particular may require 'tuning' to achieve satisfactory performance (Bennett *et al.*, 1996; Chib and Greenberg, 1995). Alternatively 'black box' random variate generation techniques such as the rejection algorithm (Devroye, 1986), adaptive rejection sampling for log-concave densities (Gilks and Wild, 1992) or the ratio-of-uniforms method (Wakefield *et al.*, 1991) may be used. The use of such techniques may be daunting to those who are unfamiliar with their use, however, since they also frequently require tuning to provide reliable and efficient algorithms. In this paper we discuss a novel approach which, after the introduction of strategic auxiliary (or latent) variables, results in a Gibbs sampler having a set of easily sampled 'standard' full conditionals.

Address for correspondence: Jon Wakefield, Department of Epidemiology and Public Health, Imperial College School of Medicine at St Mary's, Norfolk Place, London, W2 1PG, UK.
E-mail: j.wakefield@ic.ac.uk

 1369–7412/99/61331

Suppose that the required conditional distribution for a random variable X is denoted f. The basic idea is to introduce a latent variable U, to construct the joint density of U and X, with marginal density for X given by f, and then to 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, provided that there is no dramatic loss in efficiency compared with the original chain.

For a historical overview of Markov chain methods and the use of latent (auxiliary) 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 in section 5 of Besag and Green (1993). Recent progress with auxiliary variables is reported in Higdon (1998) and references therein.

The paper is organized as follows. In the next section, we develop the theory underlying the new algorithm. In particular, we show that our method improves on a Metropolis–Hastings independence chain. In Section 3 we discuss strategies for choosing latent variables and in Section 4 we implement the approach for Bayesian non-conjugate models. Section 5 considers hierarchical models, with Section 5.1 dealing with *generalized linear mixed models* and Section 5.2 with *non-linear mixed models*. Section 6 contains a numerical example, followed by a concluding discussion in Section 7.

2. Preliminaries

The main result on which the algorithm developed in this paper depends is given in the following theorem.

Theorem 1. Suppose that we wish to generate random variates from a density f given by

$$f(x) \propto \pi(x) \prod_{i=1}^{N} l_i(x),$$

where π is a density of known form and the l_i are non-negative invertible functions (not necessarily densities), i.e. if $l_i(x) > u$ then it is possible to obtain the set $A_u^i = \{x: l_i(x) > u\}$. Then a Gibbs sampler for generating random variates from f exists where all except 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 $U = (U_1, \ldots, U_N)$, with each U_i defined on $(0, \infty)$, such that the joint density with X is given by

$$f(x, u_1, \ldots, u_N) \propto \pi(x) \prod_{i=1}^{N} I\{u_i < l_i(x)\}.$$

Clearly the marginal density for X is $f(x)$. A Gibbs sampler can now be implemented with the full conditionals for each U_i being $\mathcal{U}\{0, l_i(x)\}$ where $\mathcal{U}(a, b)$ denotes the uniform density on the interval (a, b). The full conditional for X is given by π restricted to the set $A_u = \{x: l_i(x) > u_i, i = 1, \ldots, N\}$.

The decomposition appearing in the theorem is very similar to an expression appearing in Besag and Green (1993), section 5. However, they did not mention the significant advantages that an invertible l_i leads to. They stated that, 'When dealing with more complicated models,

direct simulation from $f(x|u)$ is *unlikely to be available*' (our italics). As a consequence, they proposed that sampling from π restricted to the set A_u may be achieved by sampling repeatedly from π until the sample falls in A_u. Although this method works in principle it will be inefficient in many situations. Our aim is to demonstrate that we *can* introduce latent variables in complex models in such a way that *direct simulation* from $f(x|u)$ is achieved. 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 density given by $f(x) \propto l(x)\,\pi(x)$ and suppose that it is not possible to sample directly from f. We assume that π is a density. The general idea is to introduce a latent variable U, defined on the interval $(0, \infty)$, or more strictly the interval $(0, l(\hat{x}))$ where $\hat{x}$ maximizes $l(\cdot)$, and define the joint density with X by

$$f(x, u) \propto I\{u < l(x)\}\,\pi(x).$$

The full conditional for U is $\mathcal{U}\{0, l(x)\}$, and the full conditional for X is π, restricted to the set $A_u = \{x: l(x) > u\}$. The decomposition $f(x) \propto l(x)\,\pi(x)$ is not unique, and we can exploit this fact when constructing the joint density containing the latent variable.

We now show that this approach is more efficient than a particular independence Metropolis–Hastings chain. The Metropolis–Hastings algorithm is a Markovian scheme which may be used for obtaining samples from the posterior $f(x) \propto l(x)\,\pi(x)$. Consider a specific version of this algorithm that uses $\pi(\cdot)$ as the proposal and let the sampled point be denoted $\tilde{x}$ and the current point be $x^{(t)}$. This point is accepted with probability $\min\{1, l(\tilde{x})/l(x^{(t)})\}$ and this condition is tested by sampling, independently of $\tilde{x}$, a uniform variate u. Essentially if $l(\tilde{x})/l(x^{(t)}) > u$ then $x^{(t+1)} = \tilde{x}$; otherwise $x^{(t+1)} = x^{(t)}$. The chain either 'moves on' or 'stays where it is'. The convention is that $\tilde{x}$ is sampled first, followed by u. Suppose that we reverse this and sample u first. To move on we need to sample $\tilde{x}$ from $\pi(\cdot)$ such that $l(\tilde{x})/l(x^{(t)}) > u$. Suppose, therefore, that we sample $\tilde{x}$ from $\pi(\cdot)$ restricted to the set $A_u(t) = \{x: l(x) > u\,l(x^{(t)})\}$. In this case, the chain will always move on. In fact, we have just described a Gibbs sampler with standard full conditionals, leading to the Markovian scheme for generating $\{x^{(t)}\}$ given by $x^{(t+1)} \sim \pi(\cdot)$ restricted to the set $A_u(t) = \{x: l(x) > u\,l(x^{(t)})\}$, where u is $\mathcal{U}(0, 1)$.

If X is multidimensional, and it is not possible to obtain the multivariate set A_u, then a simplification is to sample from $f(x|u)$ by sampling from $f(x_k|x_{-k}, u)$, for $k = 1, \ldots, p$, where p is the dimension of X. This would involve sampling from $\pi(x_k|x_{-k})$ restricted to the set $A_u^k = \{x_k: l(x_k, x_{-k}) > u\}$. In this case it is only required that $l_k(x_k) = l(x_k, x_{-k})$, given x_{-k}, be invertible for all k. The usefulness of this approach is demonstrated for non-linear mixed models in Section 5.2.

3. Choosing the latent variable(s)

In this section we discuss ways of introducing latent variables, other than the direct approach involving a uniform random variable outlined in theorem 1. Let us consider the non-conjugate case

$$f(x) \propto l(x)\,\pi(x),$$

where we assume that we can sample from truncated versions of π. In all the examples in this section we are not claiming that the 'best' way to sample from f is by using MCMC and latent variables; we are merely using these cases to illustrate the basic ideas of our approach.

3.1. Example 1: $l(x) = \exp\{-\exp(x)\}\ \mathcal{I}(-\infty < x < \infty)$

We wish to define a joint density in terms of X and a latent variable so that the marginal distribution for X corresponds to $f(\cdot)$. The obvious way to achieve this here is via the latent variable u defined through

$$f(x, u) \propto I[0 < u < \exp\{-\exp(x)\}]\, \pi(x).$$

Alternatively, we may introduce the variable V whose joint distribution with X is given by

$$f(x, v) \propto \exp(-v)\, I\{v > \exp(x)\}\, \pi(x).$$

The particular choice of latent variable depends on the context. The method works because $l(x) < 1$ for all x and hence $-\log\{l(x)\} > 0$. In general, if $l(x) < M$ then we can use $l^*(x) < 1$, where $l^*(x) = l(x)/M$. The conditional distributions for the second suggested choice are given by

$$f(v|x) \propto \exp(-v)\, I\{v > \exp(x)\}$$

and

$$f(x|v) \propto \pi(x)\, I\{x < \log(v)\}.$$

To perform an iteration of the Gibbs sampler we can take $v = \exp(\tilde{x}) + e$, where e is from the exponential distribution with mean 1 and $\tilde{x}$ is the current state of the chain. So the truncation set for X becomes $\{x\colon x < \log\{\exp(\tilde{x}) + e\}\}$.

3.2. Example 2: $l(x) = x^m(1+x)^{-n}\ \mathcal{I}(x > 0)$, $m < n$, and $\pi(x)$ is a gamma distribution with shape and scale parameters equal to 1

If we use $I\{u < (1+x)^{-n}\}$ and take x^m into the prior, then we have the joint density

$$f(x, u) \propto I\{u < (1+x)^{-n}\}\, \pi^*(x),$$

where π^* is the gamma distribution with mean $m+1$ and scale parameter 1. The conditional distributions are then given by

$$f(u|x) = \mathcal{U}\{0, (1+x)^{-n}\}$$

and

$$f(x|u) \propto \pi^*(x)\, I(x < 1/u^{1/n} - 1).$$

It is of interest to see how the truncation set for X depends on $\tilde{x}$. We can generate $\log(u) = -e - n\log(1+\tilde{x})$ and so the conditional distribution for X can be written as

$$f(x|\tilde{x}, e) \propto x^m \exp(-x)\, I\{x < (1+\tilde{x})\exp(e/n) - 1\}.$$

There are two considerations here. The size of m will determine the efficiency of sampling the truncated gamma distribution and n will control the size of the truncation set for X, but note that the 'minimum' set is $\{x\colon x < \tilde{x}\}$. If m is very large, and the sampling of the truncated gamma distribution becomes inefficient, then an alternative strategy is to introduce two latent variables based on $l_1(x) = x^m$ and $l_2(x) = (1+x)^{-n}$. The full conditionals are given by

$$f(u|x) = \mathcal{U}\{0, l_1(x)\},$$
$$f(v|x) = \mathcal{U}\{0, l_2(x)\}$$

and

$$f(x|u, v) \propto \exp(-x)\, I(u^{1/m} < x < v^{-1/n} - 1).$$

The full conditional for X can be written as

$$f(x|\tilde{x}, e_1, e_2) \propto \exp(-x)\, I\{\tilde{x} \exp(-e_1/m) < x < (1 + \tilde{x}) \exp(e_2/n) - 1\},$$

where e_1 and e_2 are independent exponential random variates with mean 1. This chain avoids the need to sample a truncated gamma distribution, but at the expense of an extra latent variable. The effect of this extra latent variable is evident from the two truncation sets — one is obviously smaller than the other. Problems of high autocorrelation will be encountered with the second chain if both m and n are large, which is clear from the truncation set for X.

If we have $l(x) = \exp(mx)\{1 + \exp(x)\}^{-n}$ and $\pi(\cdot)$ is normal, for example, then a similar approach can be taken.

3.3. Example 3: $l(x) = a^x\, I(x > 0)$ with $a > 0$

Here we introduce the latent variable u via $I(u < a^x)$. The truncation set depends on whether $a < 1$ or $a > 1$. If $a < 1$ the truncation set is given by $(0, \log(u)/\log(a))$ and, if $a > 1$, by $(\log(u)/\log(a), \infty)$.

These examples provide a brief summary of what is to follow. The selection of the appropriate latent variable(s) is usually self-evident but, in some cases, some thought may be required.

4. Bayesian non-conjugate models

In this section we implement the latent variable approach to sampling from posterior distributions arising from Bayesian non-conjugate models.

4.1. Example 4: Poisson–log-normal model

Suppose that we observe a random non-negative integer τ from a Poisson distribution with parameter $\exp(X)$. Without loss of generality we assume that the prior for X is $N(\cdot|0, 1)$. The posterior density is then given by

$$f(x) \propto \exp\{\tau x - \exp(x)\} \exp(-0.5x^2).$$

We notice that the $\exp(\tau x)$-term can be absorbed into the prior and, therefore, following example 1, we introduce the latent variable U, defined on the interval $(0, \infty)$, such that the joint density with X is given by

$$f(x, u) \propto \exp(-u)\, I\{u > \exp(x)\} \exp\{-0.5(x^2 - 2\tau x)\},$$

which leads to conditional densities given by

$$f(u|x) \propto \exp(-u)\, I\{u > \exp(x)\}$$

and

$$f(x|u) \propto \exp\{-0.5(x - \tau)^2\}\, I\{x < \log(u)\},$$

a truncated $N(\cdot|\tau, 1)$ density. See Devroye (1986) and Robert (1995) for methods for sampling from a truncated normal density.

4.2. Example 5: Bernoulli–logistic regression model

Here we consider a Bernoulli logistic regression model for which $w_i \sim \text{Bernoulli}(p_i)$ where $p_i^{-1} = 1 + \exp(-\mu - xz_i)$ and z_i is a known explanatory variable. We assume for simplicity that μ is known. We have

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

with, without loss of generality, $X \sim N(\cdot|0, 1)$ as the prior. The posterior density for X is given by

$$f(x) \propto \exp(-0.5x^2) \prod_{i=1}^{n} l_{1i}(x)\, l_{2i}(x),$$

where

$$l_{1i}(x) = \{1 + \exp(-\mu - xz_i)\}^{-w_i}$$

and

$$l_{2i} = \{1 + \exp(\mu + xz_i)\}^{w_i - 1}.$$

Using the standard approach, outlined in theorem 1, we introduce the latent variables $U = (U_1, \ldots, U_n)$ and $V = (V_1, \ldots, V_n)$, such that their joint density with X is given by

$$f(x, u, v) \propto \exp(-0.5x^2) \prod_{i=1}^{n} I\{u_i < l_{1i}(x),\ v_i < l_{2i}(x)\}.$$

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

$$f(u_i|u_{-i}, v, x) = \mathcal{U}\{0, l_{1i}(x)\}$$

and

$$f(v_i|v_{-i}, u, x) = \mathcal{U}\{0, l_{2i}(x)\}.$$

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 $\mathcal{S} = \varnothing$ then we replace $\max_{i \in \mathcal{S}}\{a_i\}$ by $-\infty$ and if $\mathcal{R} = \varnothing$ then we replace $\min_{i \in \mathcal{R}}\{b_i\}$ by ∞.

4.3. Example 6: 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

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

where $\beta = (\beta_1, \ldots, \beta_p)$ is a vector of unknown parameters and $\lambda_0(t)$ is the base-line 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 by

$$f(\beta|\alpha) \propto \prod_{i=1}^{n} \exp\{z_i\beta\delta_i - t_i^{\alpha} \exp(z_i\beta)\} \exp\{-0.5(\beta-\mu)'\Sigma^{-1}(\beta-\mu)\},$$

where $\delta_i = 1$ indicates that t_i is an uncensored observation, and $\delta_i = 0$ otherwise. Here, following example 1, we introduce the latent variable $U = (U_1, \ldots, U_n)$ such that the joint density with β is given by

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

where $\nu = \Sigma_{i=1}^{n} z_i\delta_i$ has been absorbed into the prior. The full conditional distributions for each of the u_i are independent exponential distributions with unit mean, restricted to the sets $(t_i^{\alpha} \exp(z_i\beta), \infty)$. Sampling from $f(\beta_k|\beta_{-k}, u, \alpha)$ requires

$$A_u^k = \left\{\beta_k: \beta_k < \min_i \left\{ \frac{\log(u_i/t_i^{\alpha})}{z_{ki}} - \sum_{l \neq k} \frac{z_{li}\beta_l}{z_{ki}} \right\}\right\}$$

and so involves sampling a normal distribution, truncated to A_u^k. The full conditional for α, with prior $\pi(\alpha) = \text{constant}$ (Dellaportas and Smith, 1993), is given by

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

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

$$f(v, \alpha) \propto \alpha^{\tilde{n}}\, I\left\{v < \left(\prod_{\delta_i=1} t_i\right)^{\alpha}\right\} I(\lambda^- < \alpha < \lambda^+),$$

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 form and can be sampled by using uniform random variables; see example 3.

5. Bayesian hierarchical models

Hierarchical models are relevant when the observed variability in the data on a number of units can be conveniently partitioned, in the simple two-stage model, into *within*- and *between*-unit components. At the first stage of the hierarchy observations from a particular unit are modelled, whereas at the second stage of the hierarchy between-unit differences are modelled. We consider both

(a) generalized linear mixed models and
(b) non-linear mixed models.

We concentrate on that situation in which the second-stage distribution is specified parametrically, typically using normal or Student's t-distributions.

5.1. Generalized linear mixed models

5.1.1. The model

Given $\{b_i\}$, a set of q-vector random effects, the observations y_i, $i = 1, \ldots, n$, are conditionally independent from the exponential family of distributions with mean $h(w_i\beta + z_ib_i)$,

where $h(\cdot)$ is a non-negative invertible function, i.e. $g = h^{-1}$ exists, w_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, if it is not equal to 1, is 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 λ^{-1}. Within a Bayesian framework conjugate prior distributions are assigned to the parameters ϕ, β and λ. The prior for ϕ is typically an inverse gamma distribution, the prior for β a multivariate normal prior $N(\cdot|\mu, \Sigma)$ and the prior for λ is a gamma or Wishart prior, depending on whether it is univariate or multivariate.

5.1.2. The algorithm

Here we present a general algorithm for sampling the conditional distributions of the generalized linear mixed model. The full conditional distribution for β is given by

$$f(\beta|b) \propto \exp\left[\sum_i \{y_i w_i \beta - h(w_i\beta + z_i b_i)\}\right] N(\beta|\mu, \Sigma).$$

In this form the distribution is not of standard type and so cannot easily be sampled directly. We could absorb the $\Sigma_i\, y_i w_i \beta$-term into the prior and then introduce a single latent variable. In general, however, this may not be the best strategy; see the discussion of example 2.

We proceed by introducing the latent variables $U = (U_1, \ldots, U_n)$ and $V = (V_1, \ldots, V_n)$ such that the joint (full conditional) distribution with β is given by

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

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

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

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}, the set (a_k, c_k) is obtained via the inequalities $y_i w_i \beta > \log(u_i)$ and $h(w_i\beta + z_i b_i) < -\log(v_i)$ for $i = 1, \ldots, n$.

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

$$f(b_i|u, v, \beta) \propto I[u_i < \exp(y_i z_i b_i),\, v_i < \exp\{-h(w_i\beta + z_i b_i)\}]\, N(b_i|0, \Omega)$$

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

5.1.3. Example 7: random effects Poisson model

Here we consider the random effects Poisson model given by

$$y_i|\theta_i \sim \text{Poisson}\{\exp(\theta_i)\}$$

$$\theta_i = w_i\beta + b_i,$$

$$b_i \sim N(0,\ \lambda^{-1}).$$

Priors for β and λ are taken as in Section 5.1.1. The joint probability distribution of β, b and λ is given by

$$f(\beta,\ b,\ \lambda) \propto \exp\left[\sum_{i=1}^{n} \{y_i\theta_i - \exp(\theta_i) - 0.5b_i^2\lambda\}\right] \lambda^{n/2}\ \pi(\lambda,\ \beta).$$

Here we introduce the latent variables $U = (U_1, \ldots, U_n)$ and $V = (V_1, \ldots, V_n)$ such that the joint distribution with β, b and λ is given by

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

The full conditional distribution for β_k is given by

$$f(\beta_k|\beta_{-k},\ b,\ \lambda,\ u,\ v) \propto \pi(\beta_k|\beta_{-k})\ I(\beta_k \in B_k),$$

where B_k is the set

$$\left(\max_{w_{ki}<0}\left\{\frac{a_{ki}}{w_{ki}},\ \frac{c_{ki}}{w_{ki}}\right\},\ \min_{w_{ki}>0}\left\{\frac{a_{ki}}{w_{ki}},\ \frac{c_{ki}}{w_{ki}}\right\}\right)$$

where

$$a_{ki} = \log(v_i) - \sum_{l\neq k} w_{li}\beta_l - b_i$$

and

$$c_{ki} = -y_i^{-1}\ \log(u_i) - \sum_{l\neq k} w_{li}\beta_l - b_i.$$

The full conditional distribution for b_i is

$$f(b_i|\beta,\ \lambda,\ u,\ v) \propto \exp(-0.5b_i^2\lambda)\ I(b_i \in A_i),$$

where A_i is the set

$$\left(-\infty,\ \min\left\{\log(v_i) - \sum_k w_{ki}\beta_k, -y_i^{-1}\ \log(u_i) - \sum_k w_{ki}\beta_k\right\}\right).$$

The full conditional distributions for the latent variables are given by

$$f(u_i|\beta,\ b,\ \lambda) \propto I\{u_i < \exp(-y_i\theta_i)\},$$

$$f(v_i|\beta,\ b,\ \lambda) \propto \exp(-v_i)\ I\{v_i > \exp(\theta_i)\}$$

and the full conditional for λ is

$$f(\lambda|\beta, b, u, v) \propto \lambda^{n/2} \exp\left(-0.5\lambda \sum_i b_i^2 \right) \pi(\lambda).$$

Only minor modifications are required for the case when $y_i = 0$.

5.2. Non-linear mixed models

5.2.1. The model

In the following let i index individuals and j index observations within individuals with $i = 1, \ldots, n$, $j = 1, \ldots, n_i$ and $N = \Sigma_i n_i$. Let y_{ij} represent the observation. The conditional probability model for the observations is given by

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

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 non-linear mean response function. We shall write $g(\theta_i, x_{ij})$ as $g_{ij}(\theta_i)$. The θ_i 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). As a consequence the conditional distributions for each of these parameters is of known form. The problem with implementing a Gibbs sampler is with the conditional for each of the θ_i. The conditional density for θ_i is given by

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

where $l_j(\theta_i) = \{y_{ij} - g_{ij}(\theta_i)\}^2$ and $\pi(\theta_i)$ is $N(\theta_i|\mu, \Sigma)$. It is not possible to sample this distribution directly without specialist random number generation techniques. The ratio-of-uniforms method may be used but requires, in its usual implementation, three numerical maximizations for each sample (Wakefield *et al.*, 1991). The adaptive rejection sampling routine cannot be used since the conditional distributions are typically not log-concave. Gilks *et al.* (1995) proposed the Metropolis adaptive rejection sampling algorithm for such cases. Care must be taken when such chains are constructed, however; see Gilks *et al.* (1997).

5.2.2. The algorithm

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

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

and

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

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

The full conditional distributions for the θ_i random effects are given by

$$f(\theta_i|u_i) \propto \pi(\theta_i)\, I(\theta_i \in A_i),$$

where

$$A_i = \{\theta_i\colon\ l_j(\theta_i) < u_{ij},\ j = 1,\ \ldots,\ n_i\},$$

which is

$$A_i = \{\theta_i\colon\ y_{ij} - \sqrt{u_{ij}} < g_{ij}(\theta_i) < y_{ij} + \sqrt{u_{ij}},\ j = 1,\ \ldots,\ n_i\}.$$

Therefore we can sample θ_i from π restricted to this set. The full conditional distributions for the latent variables are given by

$$f(u_{ij}|\theta_i) \propto \exp(-\lambda u_{ij}/2)\, I\{u_{ij} > l_j(\theta_i)\}.$$

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

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

In the following, for notational convenience, we have removed the subscripts i and put $m = n_i$. Recall that

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

Generally we will not be able to find the set A_i analytically and so instead we sample each element of θ separately. We sample from $f(\theta|u)$ by sampling from $f(\theta_k|\theta_{-k},\ u)$, for $k = 1,\ \ldots,\ p$, where p is the dimension of θ. This involves sampling from $\pi(\theta_k|\theta_{-k})\, I(\theta_k \in A_u^k)$ where

$$A_u^k = \{\theta_k\colon\ l_j(\theta_k,\ \theta_{-k}) < u_j,\ j = 1,\ \ldots,\ m\}.$$

Clearly the specific form of A_u and A_u^k will depend on the likelihood $l_j(\cdot)$.

5.2.3. *Example 8: logistic model*

For the logistic model we obtain

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

We shall concentrate on finding the sets A_u^k, $k = 1,\ \ldots,\ 3$, since once we have done this the algorithm is straightforward. Now

$$A_u^1 = \left(\max_j\{a_j\},\ \min_j\{b_j\}\right),$$

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

$$A_u^2 = \left(\max_{j\in\mathcal{S}}\{\alpha_j\},\ \min_j\{\beta_j\}\right),$$

where

$$\alpha_j = \log[\exp\{\theta_1 - \sqrt{u_j} - \log(y_j)\} - 1] - \theta_3 x_j$$

and

$$\beta_j = \log[\exp\{\theta_1 + \sqrt{u_j} - \log(y_j)\} - 1] - \theta_3 x_j$$

(note that $\theta_1 + \sqrt{u_j} - \log(y_j) > 0$). If $\mathcal{S} = \varnothing$ then

$$A_u^2 = \Big(-\infty,\ \min_j\{\beta_j\}\Big).$$

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

$$A_u^3 = \Big(\max_{j\in\mathcal{S}}\{\gamma_j\},\ \min_j\{\delta_j\}\Big),$$

where

$$\gamma_j = (\log[\exp\{\theta_1 - \sqrt{u_j} - \log(y_j)\} - 1] - \theta_2)/x_j$$

and

$$\delta_j = (\log[\exp\{\theta_1 + \sqrt{u_j} - \log(y_j)\} - 1] - \theta_2)/x_j.$$

If $\mathcal{S} = \varnothing$ then

$$A_u^3 = \Big(-\infty,\ \min_j\{\beta_j\}\Big).$$

6. Numerical example

In this section we consider a non-linear mixed model example and compare our auxiliary variable Gibbs sampler with a Metropolis–Hastings algorithm.

6.1. Non-linear random effects model

The example is taken from Lindstrom and Bates (1990). Let y_{ij} denote the observed trunk circumference measured on the ith orange-tree, $i = 1, \ldots, 7$, at time x_{ij}, $j = 1, \ldots, 5$. The logistic model (Section 5.2.3) models the relationship between trunk circumference and time:

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

where y_{ij} are the observed trunk circumference measurements and ϵ_{ij} are IID normal with mean 0 and variance σ^2. The second stage assumes that $\theta_i \sim N(\theta_i|\mu, \Sigma)$ where $\theta_i = (\theta_{1i}, \theta_{2i}, \theta_{3i})$. Conjugate priors are assumed for σ^2, μ and Σ.

We shall compare our algorithm with a Metropolis–Hastings algorithm which is used for sampling from the full conditional distribution for θ_i. A typical MCMC implementation for this model (see, for example, Bennett *et al.* (1996)) would be to use a Metropolis–Hastings chain with a random walk algorithm for θ_i, $i = 1, \ldots, n$. The proposal prior may be taken as a multivariate normal distribution, centred at the current point, and with covariance matrix given by a scalar multiple of the asymptotic covariance matrix evaluated at a point close to the posterior mean (calculated from an initial run for example) or the maximum likelihood estimate. The aim is to select the scalar to control the size of the steps in the random walk. If too large a value is chosen then few moves will be made; if too small a value is taken the walk will only take small steps.

The lengths of the Fortran code that implemented each algorithm were approximately equal. Similar run times were obtained for 10000 iterations of each algorithm but the Metropolis–Hastings algorithm required preliminary runs to obtain a desirable acceptance probability (54% for the final algorithm; for a discussion of optimal rates see Roberts *et al.* (1997)). Finally, we compare the 'worst' case of autocorrelation for each of the algorithms. In

the random walk Metropolis–Hastings algorithm this was with the μ_3-parameter and with the auxiliary variable Gibbs sampler this was with the Σ_{33}-parameter. The lag 1 autocorrelations of each of these parameters were 0.79 and 0.75 respectively, with the autocorrelations dying away very slowly for the Metropolis–Hastings algorithm (0.39 at lag 40) but falling to 0 by lag 9 for the auxiliary variable sampler.

7. Discussion, extensions and conclusions

In Section 6 we presented an example, using the auxiliary variable method, which resulted in a quick and efficient MCMC algorithm. Additionally, the algorithm was easy to code, requiring only standard random variate generation 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, this 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 an MCMC sampler in which some or all of the full conditionals have to be sampled via rejection and/or Metropolis–Hastings-type algorithms?'. We are not aware of a definitive answer to this question. However, 'efficiency' may be measured in several different ways and for many practitioners ease of coding will be the dominating criterion, particularly in 'one-off' applications.

The assessment of convergence remains a major problem with the use of MCMC algorithms. Results on rates of convergence are currently only available for narrow classes of models (Polson, 1996). Latent variables have a long history within the MCMC literature. In addition to the statistical physics work referred to in Besag and Green (1993) their use has also been proposed in a variety of models, e.g. with applications involving binary and polychotomous data (Albert and Chib, 1993), discrete regression models (Carlin and Polson, 1992), Student t-distributions (Wakefield *et al.*, 1994) and for constructing log-concave densities (Polson, 1996). In the data augmentation algorithm (Tanner and Wong, 1987) the latent variables represent 'missing' data which combine with the observed data to provide a 'standard' posterior for the parameters.

As far as the resultant Markov chain is concerned, Polson (1996) stated, 'Careful use of latent variables ... can lead to vast improvements in efficiency' and the examples in section 4 of Polson (1996) give support to the auxiliary variable approach for two types of distribution. Polson indicated that 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 auxiliary variables, does not, of course, follow.

Acknowledgements

The authors are grateful to a Joint Editor, several referees and Adrian Smith for critical comments on earlier drafts of the paper.

References

Albert, J. H. and Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *J. Am. Statist. Ass.*, **88**, 669–679.

Bennett, J. E., Racine-Poon, A. and Wakefield, J. C. (1996) MCMC for nonlinear hierarchical models. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), pp. 339–357. London: Chapman and Hall.

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

Carlin, B. P. and Polson, N. G. (1992) Monte Carlo Bayesian methods for discrete regression models and categorical time series. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 577–586. Oxford: Oxford University Press.

Chib, S. and Greenberg, E. (1995) Understanding the Metropolis–Hastings algorithm. *Am. Statistn*, **49**, 327–335.

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

Devroye, L. (1986) *Non-uniform Random Variate Generation*. New York: Springer.

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

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

Gilks, W. R., Neal, R. M., Best, N. G. and Tan, K. K. C. (1997) Corrigendum: Adaptive rejection Metropolis sampling. *Appl. Statist.*, **46**, 541–542.

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

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

Higdon, D. (1998) Auxiliary variable methods for Markov chain Monte Carlo with applications. *J. Am. Statist. Ass.*, **93**, 585–595.

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

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953) Equations of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1091.

Polson, N. G. (1996) Convergence of Markov chain Monte Carlo algorithms. In *Bayesian Statistics 5* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 297–321. Oxford: Oxford University Press.

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

Roberts, G. O., Gelman, A. and Gilks, W. R. (1997) Weak convergence and optimal scalings of random walk Metropolis algorithms. *Ann. Appl. Probab.*, **7**, 110–120.

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

Tanner, M. A. and Wong, W. H. (1987) The calculation of posterior distributions by data augmentation (with discussion). *J. Am. Statist. Ass.*, **82**, 528–550.

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

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

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. *Appl. Statist.*, **43**, 201–221.