Statistics and Computing (1991) **1**, 119–128

An iterative Monte Carlo method for nonconjugate Bayesian analysis

BRADLEY P. CARLIN[1] and ALAN E. GELFAND[2]

[1]*Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455, USA*
[2]*Department of Statistics, University of Connecticut, Storrs, CT 06268, USA*

Received February 1991 and accepted June 1991

The Gibbs sampler has been proposed as a general method for Bayesian calculation in Gelfand and Smith (1990). However, the predominance of experience to date resides in applications assuming conjugacy where implementation is reasonably straightforward. This paper describes a tailored approximate rejection method approach for implementation of the Gibbs sampler when nonconjugate structure is present. Several challenging applications are presented for illustration.

Keywords: Bayesian inference, Gibbs sampler, hierarchical models, logistic regression, nonlinear models, rejection method

1. Introduction

In earlier work (Gelfand and Smith, 1990; Gelfand *et al.*, 1990) a sampling-based approach using the Gibbs sampler (Geman and Geman, 1984) was offered as a means for implementing Bayesian data analysis. This approach is very broadly applicable but most straightforwardly effected when conjugacy is assumed. Very recent work (Dellaportas and Smith, 1991; Gilks and Wild, 1991; Tanner, 1991, p. 101; Wakefield *et al.*, 1991; Zeger and Karim, 1991) addresses the issue of sampling under nonconjugacy. Our effort here considers this issue as well in the case of continuous parameters but using a much different approach.

By way of clarification, in the context of a hierarchical Bayes model, conjugacy is taken to mean that for any parameter in the model specification (likelihood × prior), integration of this model with respect to this parameter may be achieved explicitly. This pragmatic definition differs a bit from formal versions in, for example, Diaconis and Ylvisaker (1979) or Morris (1983). Conjugacy allows us to simplify the implementation of the Gibbs sampler, enabling almost routine fully Bayesian analysis of many standard problems. However, more challenging modeling situations will not allow conjugacy, as the following examples suggest:

(1) Reference priors (Bernardo, 1979; Berger and Bernardo, 1989) and other versions of 'noninformative' priors (Berger, 1985) will not be conjugate with the likelihood.

(2) Nonlinear models, including generalized linear models, will have likelihoods as functions of the model parameters which typically do not admit a conjugate form.

(3) For hierarchical models, according to Berger (1985, p. 232) 'the choice of a form for the second or higher stage prior seems to have relatively little effect'. However, this is usually not the case at the first stage specification where the form of the prior—for example, body and tails—will substantially affect the inference. To assess model robustness requires Bayesian analysis when the first stage prior is nonconjugate.

To carry out the Gibbs sampler in the presence of nonconjugacy for at least some of the model parameters requires sampling from nonstandardized densities, as discussed below. A means of accomplishing such sampling is the rejection method (Devroye, 1986; Ripley, 1987), formally defined in Section 3. The purpose of this paper is to describe a tailored general rejection method approach for implementation of the Gibbs sampler when some nonconjugate structure is present.

To clarify how nonstandardized densities arise we note

that the Gibbs sampler requires independent draws from the complete conditional distributions of the model parameters (see Section 2). For any parameter in any hierarchical model, its complete conditional distribution is the conditional distribution of the parameter given the data and all other model parameters. But it is then clear that for each model parameter its complete conditional density is proportional to likelihood × prior. Often the hierarchical structure along with, for example, exchangeability assumptions greatly simplify these distributions.

In the next section we briefly review the Gibbs sampler. Since it is an iterative Markovian updating scheme which is usually replicated, the conditional levels for each complete conditional distribution which needs to be sampled will change with each iteration and replication. Standard use of the rejection method requires that a distinct envelope function be developed for each such sampling. Unfortunately, this envelope is used to generate but *one* observation. As an alternative, in Section 3 we first note that a good multidimensional envelope density will provide good complete conditional envelope densities. We then show how to create such a multidimensional envelope density which also possesses complete conditional distributions that are easy to sample.

In Section 4 we illustrate with three demanding modeling applications. Finally, in Section 5 we summarize, noting, in addition, when our proposed method is likely to work well and when not.

2. The Gibbs sampler

For convenience, in this section densities will be denoted generically by square brackets, so that joint, conditional and marginal forms for random variables U, V, appear as $[U, V]$, $[U \mid V]$ and $[V]$, respectively. The usual marginalization by integration is denoted by forms such as $[U] = \int [U \mid V] \times [V]$. For a collection of random variables $U_1, U_2, \ldots, U_k$ the complete conditional densities can thus be denoted by $[U_s \mid U_r,\ r \neq s]$, $s = 1, 2, \ldots, k$, and the marginal densities by $[U_s]$, $s = 1, 2, \ldots, k$.

Given the ability to draw random variate samples of U_s from $[U_s \mid U_r, r \neq s]$ for specified values of $\{U_r, r \neq s\}$, $s = 1, 2, \ldots, k$, the Gibbs sampler provides an iterative Markovian updating scheme which enables us to make sample-based estimates, $[\hat{U}_s]$, of the marginal densities, $[U_s]$, $s = 1, 2, \ldots, k$ (Gelfand and Smith, 1990). If the scheme is conducted in parallel m times each to t iterations, k-tuples $(U_{1j}^{(t)}, \ldots, U_{kj}^{(t)}), j = 1, 2, \ldots, m$, result. If t is large enough, each k-tuple is approximately distributed as $[U_1, \ldots, U_k]$. Assessment of convergence is a complex issue.

It is shown in Gelfand and Smith (1990) that a density estimate of the form

$$[\hat{U}_s] = \sum_{j=1}^{m} [U_s \mid U_{rj}^{(t)}, r \neq s]/m \qquad (1)$$

is better than a kernel density estimate for $[U_s]$. This is not surprising since Equation 1 takes advantage of the known structure in the model whereas the kernel density estimate does not. Equation 1 is a discrete mixture distribution and is essentially a Monte Carlo integration to accomplish the desired marginalization. Extension to expectations, $E[h(U_s)]$, and more generally to densities and expectations for functions $W(U_1, U_2, \ldots, U_k)$, is straightforward (for details, see Gelfand and Smith, 1991).

In the Bayesian context the U_i are the unknown parameters (or possibly unobserved data) in the model, and W would be any function of the parameters (or unobserved data) which is of interest. All distributions are viewed as conditional on the observed data, whence the marginal densities, $[U_s]$, become the desired marginal posterior distributions of the parameters (or unobserved data). Moreover, the joint density $[U_1, \ldots, U_k]$ becomes the joint density of all the model parameters/unknowns given the observed data. This density, only known modulo normalizing constant, will be denoted by $f(U_1, \ldots, U_k)$ where f is, in fact, likelihood × prior. Similarly, all complete conditional distributions will again be proportional to f and, in the absence of conjugacy, will not lead to familiar standard forms whence sampling will require random generation from nonstandardized densities. In Section 3 we suggest an approach to accomplish this using a tailored version of the rejection method.

As noted above, we prefer to use a density estimate of the form of Equation 1. In fact, using this form allows m to be much smaller (say, $m = 100$) than needed for kernel density estimates (say, $m = 5000$). However, calculation of Equation 1 will require, at the last iteration, m normalizations of f which in turn requires m one-dimensional numerical integrations. Simple trapezoidal or Simpson's rule integration to do this is quite fast, still yielding substantial overall savings in run time compared with kernel density estimation.

3. A tailored rejection method

In this section we develop a specialized version of the rejection method which is well suited to the sampling needs of the Gibbs sampler. First we review the basic rejection algorithm for sampling from a continuous density.

3.1. *The rejection algorithm*

The rejection algorithm for a nonstandardized integrable density $f(\boldsymbol{U})$, $\boldsymbol{U} = (U_1, \ldots, U_k)$ proceeds as follows:

(1) Identify a density $g(\boldsymbol{U})$ which may be readily sampled and such that there exists M for which $f(\boldsymbol{U})/g(\boldsymbol{U}) \leq M$ for all $\boldsymbol{U}$.

(2) Generate $\boldsymbol{U}^*$ from a $g(\boldsymbol{U})$.

(3) Generate X from a $U(0, 1)$ distribution.

(4) Accept $\boldsymbol{U}^*$ if $X \leq f(\boldsymbol{U}^*)/Mg(\boldsymbol{U}^*)$, otherwise return to (2).

It may be shown (Devroye, 1986; Ripley, 1987) that the distribution of $\boldsymbol{U}^*$ is $f(\boldsymbol{U})/\int f(\boldsymbol{U})$ and also that the acceptance probability associated with this algorithm is $M^{-1}\int f(\boldsymbol{U})$. Hence, the smaller we can make M, that is, the more g resembles f, the more efficient the sampling.

3.2. *Split-normal and split-t envelope functions*

Implementation of the Gibbs sampler requires sampling from f viewed as a function of, say, U_1 for fixed $\boldsymbol{U}_{-1} \equiv (U_2, \ldots, U_k)$. However, the value of $\boldsymbol{U}_{-1}$ changes with each iteration and each replication. Customary use of the rejection method then requires that a distinct envelope function $g_{\boldsymbol{U}_{-1}}(U_1)$ be developed for each $\boldsymbol{U}_{-1}$. Moreover, typically each such $g_{\boldsymbol{U}_{-1}}(U_1)$ is used to generate but *one* observation.

As an alternative, now viewing f as a k-dimensional function, we propose, before doing any sampling, to create a single k-dimensional density function $g(\boldsymbol{U})$ which is a good envelope for f and is such that for each U_i, g has complete conditional distributions which are easy to sample. Formalizing notation and still taking $i = 1$, we write $g(U_1, \ldots, U_k) = g_1(U_1 \mid \boldsymbol{U}_{-1})g_2(\boldsymbol{U}_{-1})$. Note that $g_1(U_1 \mid \boldsymbol{U}_{-1})$ serves as an envelope for the complete conditional distribution for U_1 arising from f. That is, if M is such that $f(\boldsymbol{U})/g(\boldsymbol{U}) \leq M$ for all $\boldsymbol{U}$, then, as a function of U_1 for fixed $\boldsymbol{U}_{-1}$, $f/g_1 \leq M' \equiv Mg_2(\boldsymbol{U}_{-1})$. In practice g_1, g_2 and M' are not calculated; acceptance of U_1^* is determined by the equivalent test, (4) above, evaluating f and g at $(U_1^*, \boldsymbol{U}_{-1})$.

How might a suitable $g(\boldsymbol{U})$ be developed? Writing $f(\boldsymbol{U}) = \text{likelihood}(\boldsymbol{U}) \times \text{prior}(\boldsymbol{U})$, if $\hat{\boldsymbol{U}}$ is the maximum likelihood estimate of $\boldsymbol{U}$ we may take $g(\boldsymbol{U}) = \text{prior}(\boldsymbol{U})$ with $M = \text{likelihood}(\hat{\boldsymbol{U}})$ to implement the rejection method. This choice of g has at least two drawbacks. First, since it only matches the prior, it need not be a good envelope for f so that very inefficient sampling may result. Second, it requires prior$(\boldsymbol{U})$ to be proper (since we must sample from g in the rejection method). Hence, while this choice of g may be viewed as a possible backup we seek a proper g which more closely resembles f.

In the context of noniterative Monte Carlo sampling with respect to a nonstandardized density, Geweke (1989) proposes the use of an importance sampling density which is a multivariate split-normal or split-t distribution. Such a density, g, is designed to make the variability of the ratio f/g over the space of $\boldsymbol{U}$ small under g which in turn makes the variance of the Monte Carlo integration small. Note that such a g is desirable for our purposes since the less variable f/g is, the smaller M will be, whence the greater the acceptance probability and the more efficient our sampling.

Recall that in the Bayesian framework, modulo normalization, f is viewed as the joint posterior density function of all the parameters (and perhaps any missing data) given the observed data. With an increasing amount of data, under usual regularity conditions f is approximately a multivariate normal density up to a proportionally constant (for example, Berger, 1985, p. 224). A convenient choice to approximate the mean of this normal distribution is $\hat{\boldsymbol{U}}$, the mode of f. With regard to an approximate covariance matrix, the preferred choice from an asymptotic viewpoint is the negative of the inverse Hessian evaluated at $\hat{\boldsymbol{U}}$. The Hessian matrix $\boldsymbol{H}$ is defined by $H_{ij} = \partial^2 \log f/\partial U_i \partial U_j$. In two-stage models we might use the log-likelihood rather than $\log f$, which amounts to replacing $\boldsymbol{H}$ by the information matrix $\boldsymbol{I}$.

Often both $\boldsymbol{H}$ and $\boldsymbol{I}$ are difficult to obtain since they require the existence evaluation of second derivatives. A differencing algorithm (such as in Kass, 1987) can be used to provide reliable derivative-free estimates for $\boldsymbol{H}$ or $\boldsymbol{I}$, thus avoiding formal differentiation. Since our objective is only to approximate the covariance matrix, $\boldsymbol{\Sigma}$, associated with f we need not use these asymptotic forms but may instead adopt alternative choices for $\hat{\boldsymbol{\Sigma}}$. One simple approach which avoids the differentiation problem is to approximate the surface $\log f(\boldsymbol{U})$ by a quadratic function such as $(\boldsymbol{U} - \hat{\boldsymbol{U}})^{\mathrm{T}}\boldsymbol{V}(\boldsymbol{U} - \hat{\boldsymbol{U}})$ whence $\hat{\boldsymbol{\Sigma}} = -\boldsymbol{V}^{-1}$. This approximation can be straightforwardly developed by usual least-squares methods, fitting the quadratic to a large set of $\log f$ values obtained by evaluating f at many points on a k-dimensional grid. It may in fact prove easiest to first transform $\boldsymbol{U}$ to the k-dimensional unit square, obtain the covariance matrix estimate and then transform this estimate back to the original scale by the delta method. When there are strong correlations among the Us or when $\log f$ is fairly flat $\boldsymbol{H}$, $\boldsymbol{I}$ and $\boldsymbol{V}$ may be nearly singular, making inversion awkward. This problem can be alleviated by appropriate reparametrization, that is, transformation of $\boldsymbol{U}$.

We note another approach which avoids both differentiation and inversion problems but at the expense of computational effort that will become infeasible with increasing dimensionality. We can obtain a piecewise uniform approximation to f and then obtain the covariance matrix associated with this approximation. For simplicity of illustration assume f is bivariate. Again, it may be easiest to transform (U_1, U_2) to the unit square with resulting density proportional to h. Partition the unit square into a grid of r^2 cells and evaluate h at, say, the midpoint of each cell in the grid. Denoting these values by h_{ij}, $i, j = 1, \ldots, r$, replace the density h by the constant h_{ij} for points in the (i, j)th cell to obtain the piecewise uniform approximation to h. Normalization of this approximate density and calculation of its moments is

straightforward. Thus we may approximate the covariance matrix associated with h and, again using the delta method, that associated with f.

The above discussion suggests taking $g(\boldsymbol{U})$ to be $N(\hat{\boldsymbol{U}}, \hat{\boldsymbol{\Sigma}})$ for some convenient $\hat{\boldsymbol{\Sigma}}$. However, for more interesting situations involving small to moderate amounts of data, although f will typically still be unimodal, it will likely be somewhat asymmetric and our choice of $\hat{\boldsymbol{\Sigma}}$ will likely be a weak covariance approximation. Geweke (1989) suggests that an appropriate split-normal or split-t distribution be used in place of $N(\hat{\boldsymbol{U}}, \hat{\boldsymbol{\Sigma}})$. We now develop the details of this approximation for our situation, including the required complete conditional distributions.

A standard univariate split-normal distribution denoted by $SN(0, q, r)$ is defined by the density

$$\begin{cases} \dfrac{1}{\sqrt{2\pi}q} \mathrm{e}^{-z^2/2q^2}, & z > 0 \\ \dfrac{1}{\sqrt{2\pi}r} \mathrm{e}^{-z^2/2r^2}, & z < 0 \end{cases}.$$

To generate $Z \sim SN(0, q, r)$ we draw $\epsilon \sim \eta(0, 1)$ and take $Z = q\epsilon$ if $\epsilon > 0$, $Z = r\epsilon$ if $\epsilon < 0$. Let $\boldsymbol{Z}^t = (Z_1, \ldots, Z_k)$ be a random vector such that the Z_i are independent with $Z_i \sim SN(0, q_i, r_i)$. A general multivariate split-normal arises by affine transformation of Z. In particular, it is proposed to take g to be the distribution of $\boldsymbol{U} = \hat{\boldsymbol{U}} + \hat{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{Z}$ for q_i, r_i given below.

Choices for q_i and r_i are intended to make g a better envelope than $N(\hat{\boldsymbol{U}}, \hat{\boldsymbol{\Sigma}})$. Geweke (1989) proposes that $q_i = \sup_{\Delta > 0} v_i(\Delta)$, $r_i = \sup_{\Delta < 0} v_i(\Delta)$ where

$$v_i(\Delta) = \frac{|\Delta|}{\sqrt{2(\log f(\hat{\boldsymbol{U}}) - \log f(\hat{\boldsymbol{U}} + \Delta\hat{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{\lambda}^{(i)}))}} \tag{2}$$

and $\boldsymbol{\lambda}^{(i)}$ is a unit vector in the ith coordinate direction. Geweke notes that the Δs yielding these maxima correspond to the positive and negative values, respectively, along the ith coordinate axis which maximize the ratio of the rate of decline of f, $f(\hat{\boldsymbol{U}} + \Delta\hat{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{\lambda}^{(i)})/f(\hat{\boldsymbol{U}})$, to the rate of decline of g, $g(\hat{\boldsymbol{U}} + \Delta\hat{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{\lambda}^{(i)})/g(\hat{\boldsymbol{U}})$. Choice of q_i, r_i in this manner gives f/g the same value at Δ such that $q_i = \sup_{\Delta > 0}(v_i(\Delta))$ $(r_i = \sup_{\Delta < 0}(v_i(\Delta)))$ as at $\Delta = 0^+$ (0^-). Such matching aids in making f/g 'more constant' in each coordinate direction. Exact calculation of q_i, r_i is an analytical problem generally without explicit solution. Practically, these values are obtained only approximately by evaluating $v_i(\Delta)$ over the set $\{\Delta = j/2, j = \pm 1, \pm 2, \ldots, \pm 12\}$.

A better choice for g to accommodate the tail behaviour of f might be a multivariate split-t distribution with suitable degrees of freedom. A standard univariate split-t with ν degrees of freedom, $ST(\nu; 0, q, r)$, arises as the distribution of $t = Z/\sqrt{Y/\nu}$, with $Z \sim SN(0, q, r)$ independent of Y, a χ^2 random variable with ν degrees of freedom. To generate $t \sim ST(\nu; 0, q, r)$ we draw $\xi \sim t_\nu$ and take $t = q\xi$ if $\xi > 0$, $t = r\xi$ if $\xi < 0$. More generally, let $\mathbf{t} = (t_1, \ldots, t_k)$ be a random vector where $t_i = Z_i/\sqrt{Y/\nu}$ with Z_i independent, $Z_i \sim SN(0, q_i, r_i)$ independent of $Y \sim \chi^2_\nu$. A general multivariate split-t arises by affine transformation of $\mathbf{t}$. In particular, it is proposed to take g to be the distribution of $\boldsymbol{U} = \hat{\boldsymbol{U}} + \hat{\boldsymbol{\Sigma}}^{1/2}\mathbf{t}$ with q_i and r_i calculated, replacing $v_i(\Delta)$ in Equation 2 with

$$v_i(\Delta) = \frac{|\Delta|}{\sqrt{\nu((f(\hat{\boldsymbol{U}})/f(\hat{\boldsymbol{U}} + \Delta\hat{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{\lambda}^{(i)}))^{2/(k+\nu)} - 1)}} \tag{3}$$

The remarks following Equation 2 are applicable here.

We comment that it seems preferable to transform (reparametrize) each U_i to have $\mathbb{R}^1$ as support before embarking on the creation of g to 'match' f.

Returning to the multivariate split-normal, it is perhaps easiest to think of the transformation from $\boldsymbol{Z}$ to $\boldsymbol{U}$ as arising from 2^k one-to-one transformations determined by the vector $\mathrm{sgn}\boldsymbol{Z} = (\mathrm{sgn}Z_1, \mathrm{sgn}Z_2, \ldots, \mathrm{sgn}Z_k)$. Index these transformations by $j = 1, 2, \ldots, 2^k$ with associated partitions of $\mathbb{R}^k$ denoted by $\boldsymbol{A}_j$. On $\boldsymbol{A}_j$ there will be an associated set of qs and rs. In fact, let $d_{ji} = q_i^2$ if $\mathrm{sgn}Z_i = 1$ on $\boldsymbol{A}_j$, r_i^2 if $\mathrm{sgn}Z_i = -1$ on $\boldsymbol{A}_j$, and $\boldsymbol{D}_j$ be a diagonal matrix with diagonal entries d_{ji}. Then on $\boldsymbol{A}_j$, $\boldsymbol{Z} \sim N(0, \boldsymbol{D}_j)$ and thus the density for $\boldsymbol{Z}$ is, in obvious notation,

$$h(\boldsymbol{Z}) = \sum_{j=i}^{2^k} N(0, \boldsymbol{D}_j)(\boldsymbol{Z}) \times 1_{\boldsymbol{A}_j}(\boldsymbol{Z}) \tag{4}$$

If $\boldsymbol{B}_j$ is the image of $\boldsymbol{A}_j$ under the transformation $\boldsymbol{U} = \hat{\boldsymbol{U}} + \hat{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{Z}$ then the density for $\boldsymbol{U}$ is

$$g(\boldsymbol{U}) = \sum_{j=1}^{2^k} N(\hat{\boldsymbol{U}}, \hat{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{D}_j(\hat{\boldsymbol{\Sigma}}^{1/2})^{\mathrm{T}})(\boldsymbol{U}) \times 1_{\boldsymbol{B}_j}(\boldsymbol{U}) \tag{5}$$

The Gibbs sampler requires sampling from the complete conditional distributions associated with f. By earlier remarks, this requires sampling from the complete conditional distributions associated with g. But, for example, what is $g(U_1 \mid \boldsymbol{U}_{-1})$ for the density given in Equation 5? We now show that this distribution is a univariate split-normal which can be easily sampled. Choose for $\hat{\boldsymbol{\Sigma}}^{1/2}$ the upper triangular (Cholesky) decomposition of $\hat{\boldsymbol{\Sigma}}$ which we denote by

$$\boldsymbol{T} = \begin{pmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{pmatrix}$$

Note that, by using $\boldsymbol{T}$, $\boldsymbol{U}_{-1}$ uniquely determines $\boldsymbol{Z}_{-1} = T_{22}^{-1}(\boldsymbol{U}_{-1} - \hat{\boldsymbol{U}}_{-1})$. Furthermore, $U_1 = W_1(\boldsymbol{U}_{-1}) + T_{11}Z_1$, where $W_1(\boldsymbol{U}_{-1}) = \hat{U}_1 + T_{12}T_{22}^{-1}(\boldsymbol{U}_{-1} - \hat{\boldsymbol{U}}_{-1})$. Hence $U_1 \mid \boldsymbol{U}_{-1} \sim U_1 \mid \boldsymbol{Z}_{-1} \sim (W_1(\boldsymbol{U}_{-1}) + T_{11}Z_1) \mid \boldsymbol{Z}_{-1} \sim W_1(\boldsymbol{U}_{-1}) + T_{11}Z_1$, that is, $U_1 \mid \boldsymbol{U}_{-1}$ has a univariate split-normal distribution. Moreover, $U_1 \mid \boldsymbol{U}_{-1}$ is easily sampled by drawing $Z_1 \sim SN(0, q_1, r_1)$ and making the linear transformation $W_1(\boldsymbol{U}_{-1}) + T_{11}Z_1$. We remark that $\boldsymbol{T}$ and $\hat{\boldsymbol{U}}$ do not change from iteration to iteration, from replication to replication. Given $\boldsymbol{U}_{-1}$, we need only calculate $W(\boldsymbol{U}_{-1})$ which just involves linear operations on $\boldsymbol{U}_{-1}$.

In the multivariate split-t case, $\mathbf{t} = (t_1, \ldots, t_k)$ arises from $t_i = Z_i\sqrt{Y/\nu}$ with Z_i independent, $Z_i \sim SN(0, q_i, r_i)$ independent of $Y \sim \chi^2_\nu$. Then, analogous to Equation 4, with obvious notation

$$h(\mathbf{t}) = \sum_{j=1}^{2^k} t_\nu(0, \boldsymbol{D}_j)(\mathbf{t}) \times 1_{A_j}(\mathbf{t}) \tag{6}$$

and for $\boldsymbol{U} = \hat{\boldsymbol{U}} + \hat{\boldsymbol{\Sigma}}^{1/2}\mathbf{t}$,

$$g(\boldsymbol{U}) = \sum_{j=1}^{2^k} t_\nu(\hat{\boldsymbol{U}}, \hat{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{D}_j(\hat{\boldsymbol{\Sigma}}^{1/2})^{\mathrm{T}})(\boldsymbol{U}) \times 1_{B_j}(\boldsymbol{U}) \tag{7}$$

Again we require associated complete conditional distributions. Careful evaluation shows that $U_1 \mid \boldsymbol{U}_{-1}$ now has a univariate split-t distribution. More precisely $U_1 \mid \boldsymbol{U}_{-1} \sim W_1(\boldsymbol{U}_{-1}) + T_{11}V_1(\boldsymbol{U}_{-1})t_1$ where $t_1 \sim ST(\nu + k - 1; 0, q_1, r_1)$ and $V_1(\boldsymbol{U}_{-1}) = ((\nu + k - 1)^{-1}(\nu + \Sigma_2^k Z_i^2/e_i))^{1/2}$ with the Z_i being components of $\boldsymbol{Z}_{-1}$ defined above and e_i equal to q_i^2 or r_i^2 according to whether $Z_i > 0$ or $Z_i < 0$. Given $\boldsymbol{U}_{-1}$, we need to calculate $V_1(\boldsymbol{U}_{-1})$ in addition to $W_1(\boldsymbol{U}_{-1})$. Note that each nonconjugate U_i, in being considered as a 'U_1', will require its own Cholesky decomposition. But then, recalling Equation 2 or 3, each U_1 will have its own set of qs and rs.

We commented earlier that in implementing the rejection method we would test a U_1^* generated from $U_1 \mid \boldsymbol{U}_{-1}$ by using (4) in Section 3.1. Computation is simplified by noting that $g(\boldsymbol{U}) = \Pi_{i=1}^k T_{ii}^{-1} \times h(\boldsymbol{T}^{-1}(\boldsymbol{U} - \hat{\boldsymbol{U}}))$ with h as in Equation 4 or 6 accordingly. However, it still remains to choose M. It seems natural to look at the ratio f/g at the mode $\hat{\boldsymbol{U}}$ but, as yet, g is undefined at $\hat{\boldsymbol{U}}$, that is, h is undefined at $\mathbf{0}$. Let $h(\mathbf{0}) = a\,\Pi_{i=1}^k \min(q_i^{-1}, r_i^{-1})$ where in the split-normal case $a = (2\pi)^{-k/2}$ while in the split-t case $a = \Gamma((\nu + k)/2)/(\Gamma(\nu/2)(\pi\nu)^{k/2})$. Then $g(\hat{\boldsymbol{U}}) = \Pi_{i=1}^k T_{ii}^{-1} \times h(\mathbf{0})$. Define $M(\hat{\boldsymbol{U}}) = f(\hat{\boldsymbol{U}})/g(\hat{\boldsymbol{U}})$. For both the split-normal and split-t cases as a result of the way h was chosen along with its definition at $\mathbf{0}$, $M(\hat{\boldsymbol{U}})$ will bound f/g in a neighborhood of $\hat{\boldsymbol{U}}$. In practice, choosing $M = bM(\hat{\boldsymbol{U}})$ with $1.2 \le b \le 5$ has provided an overall bound for f/g. In our experience the choice of ν to accommodate the tail behavior of f is more critical.

At this point a brief summary of the steps required to implement the proposed rejection method might be beneficial:

(1) Obtain $\hat{\boldsymbol{U}}$, the mode of f; obtain $\hat{\boldsymbol{\Sigma}}$ using one of the methods discussed earlier in this section.

(2a) For the lth nonconjugate variable rearrange rows and columns of $\hat{\boldsymbol{\Sigma}}$ so that this variable is 'U_1'.

(2b) For this rearranged $\hat{\boldsymbol{\Sigma}}$ obtain the unique $\boldsymbol{T} = \hat{\boldsymbol{\Sigma}}^{1/2}$ and $\boldsymbol{T}^{-1}$.

(2c) Using this $\boldsymbol{T}$ in Equation 2 or 3, obtain qs and rs; obtain M as well. Carry out all of (2) for each nonconjugate variable.

(3a) In the case of $g(\boldsymbol{U})$ being split-t (similarly for split-normal) at a given iteration and replication, for the lth variable treated as 'U_1' as in (2) draw $t_1 \sim ST(\nu + k - 1; 0, q_1, r_1)$.

(3b) Compute $U_1 = W_1(\boldsymbol{U}_{-1}) + T_{11}V_1(\boldsymbol{U}_{-1})t_1$ where $W_1 = \hat{U}_1 + T_{12}T_{22}^{-1}(\boldsymbol{U}_{-1} - \hat{\boldsymbol{U}}_{-1})$, $V_1 = \{(\nu + k - 1)^{-1} \times (\nu + \Sigma_2^k Z_i^2/e_i)\}^{1/2}$ and the Z_i are the components of $\boldsymbol{Z}_{-1} \equiv T_{22}^{-1}(\boldsymbol{U}_{-1} - \hat{\boldsymbol{U}}_{-1})$.

(3c) Calculate $f(U_1, \boldsymbol{U}_{-1})$; calculate $g(U_1, \boldsymbol{U}_{-1})$ using Equation 7 or more easily Equation 6 since t_1 is drawn and $\mathbf{t}_{-1}$ is uniquely determined from $\boldsymbol{U}_{-1}$. Check whether $f \le gM$. If not, consider the discussion at the end of this section and at the beginning of Section 4.

(3d) If $f \le gM$ generate $X \sim U(0, 1)$; accept U_1 if $XgM \le f$, otherwise return to (3a).

If, during the course of sampling, a $\boldsymbol{U}_0$ arises such that $f(\boldsymbol{U}_0)/g(\boldsymbol{U}_0)$ violates our bound, we do a local search in a neighborhood of $\boldsymbol{U}_0$ and revise M accordingly. Before M is revised the magnitude of $f(\boldsymbol{U}_0)/Mg(\boldsymbol{U}_0)$ provides a rough idea of the severity of the violation. Of course, if a violation occurs then some of the previously generated variates might not have been retained with this revised M and, more importantly, these variates were not sampled from the desired complete conditional distribution. Before exploring this point further, suppose that the change in M is small (as is typically the case in our experience) so that most of the previously generated variates would still be retained. Then we would expect the joint distribution of $\boldsymbol{U}$ at the current iteration of the sampler to be closer to the converged joint distribution than when we started. Thus we would expect no advantage to starting the sampler anew as opposed to continuing from the current iteration.

Continuing with these ideas, suppose for a given M we define $S_M = \{\boldsymbol{U}: f(\boldsymbol{U})/g(\boldsymbol{U}) > M\}$ with S_M^c denoting the complement. Following the argument which justifies the rejection method we may show that the distribution of $\boldsymbol{U}$ is actually

$$\begin{cases} \dfrac{f(\boldsymbol{U})}{\int_{S_M^c} f(\boldsymbol{U})\,\mathrm{d}\boldsymbol{U} + MP_g(S_M)}, & \boldsymbol{U} \in S_M^c \\[2ex] \dfrac{Mg(\boldsymbol{U})}{\int_{S_M^c} f(\boldsymbol{U})\,\mathrm{d}\boldsymbol{U} + MP_g(S_M)}, & \boldsymbol{U} \in S_M \end{cases} \tag{8}$$

where P_g denotes the probability under the density $g(\boldsymbol{U})$. Unfortunately, $\int_{S_M} f(\boldsymbol{U})\,\mathrm{d}(\boldsymbol{U}) > MP_g(S_M)$ so that even if $P_g(S_M)$ is very small we cannot be sure that Equation 8 is close to $f(\boldsymbol{U})/f(\boldsymbol{U})\,\mathrm{d}\boldsymbol{U}$. Hence, complete conditional distributions arising from Equation 8 need not be close to complete conditional distributions arising from $f(\boldsymbol{U})$. More optimistically, if for example, given $\boldsymbol{U}_{-1}$ the set of U_1 such that $f(\boldsymbol{U})/g(\boldsymbol{U}) > M$ is a null set then we are, in fact, sampling from the complete conditional distribution of $U_1 \mid \boldsymbol{U}_{-1}$ arising from f.

We conclude this section with an important remark. When k is large, development of $g(\boldsymbol{U})$ will be made difficult because of complications in obtaining $\hat{\boldsymbol{U}}$, $\hat{\boldsymbol{\Sigma}}$ and $\boldsymbol{T}$.

fact, sampling from the complete conditional distribution of $U_1 \mid \boldsymbol{U}_{-1}$ arising from f.

We conclude this section with an important remark. When k is large, development of $g(\boldsymbol{U})$ will be made difficult because of complications in obtaining $\hat{\boldsymbol{U}}$, $\hat{\boldsymbol{\Sigma}}$ and $\boldsymbol{T}$. However, in most applications $f(\boldsymbol{U})$ is a product of functions. Hence, if we need to sample from f viewed as a nonstandardized density for $U_1 \mid \boldsymbol{U}_{-1}$, we need only consider the terms in this product involving U_1 and only the variables, say, $U_2, \ldots, U_{k'}$, appearing in these terms. That is, we factor $f(\boldsymbol{U})$ as $f(U_1, \ldots, U_k) = f_1(U_1, U_2, \ldots, U_{k'}) \times f_2(\boldsymbol{U}_{-1})$ so that g need only be a k'-dimensional envelope function. Typically, k' is much smaller than k as, for instance, in exchangeable models.

4 Examples

In this section we apply our tailored rejection method to three nonconjugate modeling scenarios. Each has been chosen to illustrate one or more of points (1)–(3) of Section 1. In these examples the various ways of selecting $\hat{\boldsymbol{\Sigma}}$ were employed, as convenient. The choice of v, the degrees of freedom for the multivariate split-t, was made somewhat empirically to incur very infrequent (at most one in 10 000) violations of our bounds. There is a trade-off here. Smaller v produces fewer violations but longer run times. Another choice involves local search to increase M given a violation versus monitoring the incidence rate of violations and reducing v if the rate is too large. We adopted the latter strategy.

4.1. *Asymptotic regression model*

Consider a model having mean structure

$$E(Y_i) = \alpha - \beta\gamma^{X_i}, \qquad \alpha, \beta > 0, 0 < \gamma < 1 \tag{9}$$

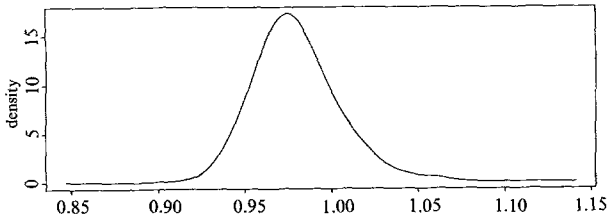
This equation describes a growth curve which has no inflection point and approaches an asymptote as X_i tends to infinity. Models of this type find agricultural, biological and engineering application. To complete the specification of the model, we assume independent $Y_i \sim N(E(Y_i), \sigma^2)$, $i = 1, \ldots, n$, and adopt the vague prior $\pi(\alpha, \beta, \gamma, \sigma) \propto (\alpha\sigma)^{-1}$ considered by Hills (1989). While the prior we have adopted is not a reference prior in the sense of Bernardo (1979), it is a vague prior in the spirit of point (1) in Section 1. In any event, the nonlinear structure in Equation 9 precludes conjugate priors as noted in point (2) in Section 1.

In order to implement the method of Section 3, we observe that f = likelihood × prior takes the form

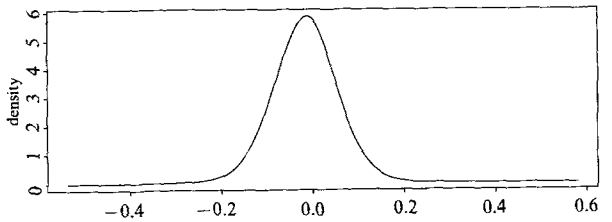
$$f(\alpha, \beta, \gamma, \sigma) = \alpha^{-1}\sigma^{-(n+1)} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha + \beta\gamma^{X_i})^2\right\}$$

so that none of the four required complete conditional distributions is available in closed form. Hence four split-t envelopes will be needed.

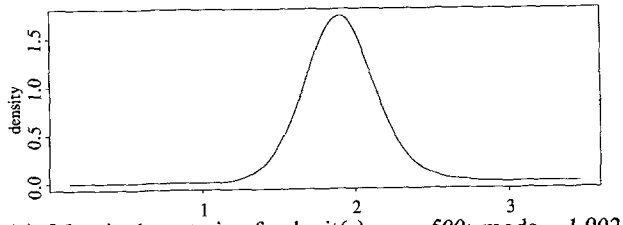
For numerical illustration we use a data set from Ratkowsky (1983), displayed in Table 1, which tallies length (Y) and age (X) for 27 captured samples of the



(a) Marginal posterior for logα, $m = 500$; mode = 0.975



(b) Marginal posterior for logβ, $m = 500$; mode = −0.014



(c) Marginal posterior for logit(γ), $m = 500$; mode = 1.902

Fig. 1. *Estimated posteriors, dugong data*

Table 1. *Length (Y) versus age (X) for the Sirenian species dugong*

X	1.0	1.5	1.5	1.5	2.5	4.0	5.0	5.0	7.0
Y	1.80	1.85	1.87	1.77	2.02	2.27	2.15	2.26	2.35
X	8.0	8.5	9.0	9.5	9.5	10.0	12.0	12.0	13.0
Y	2.47	2.19	2.26	2.40	2.39	2.41	2.50	2.32	2.43
X	13.0	14.5	15.5	15.5	16.5	17.0	22.5	29.0	31.5
Y	2.47	2.56	2.65	2.47	2.64	2.56	2.70	2.72	2.57

sirenian species dugong (sea cow). To implement our method we first transform each of the variables to R^1 by letting $U_1 = \log \alpha$, $U_2 = \log \beta$, $U_3 = \text{logit}(\gamma) = \log(\gamma/(1-\gamma))$, and $U_4 = \log \sigma$. We then approximate the covariance matrix, Σ of U, using the quadratic regression approach mentioned above. We obtain four Cholesky matrices T from $\hat{\Sigma}$ by permuting the elements of $\hat{\Sigma}$ appropriately to make each of the U_i in turn the first element of U. For this example we chose split-t distributions having $\nu = 5$ degrees of freedom.

Figure 1 shows the marginal posterior density estimates for U_1, U_2 and U_3 that result from the use of Equation 1 on $m = 500$ Gibbs iterates after completing $t = 50$ iterations of the algorithm. We remark that density estimates on the original scales could be obtained via routine transformation, as mentioned above (Gelfand and Smith, 1991). The posterior modes, 0.975, -0.014 and 1.902, are comparable to the least-squares estimates, 0.981, -0.028 and 1.932, obtained by Ratkowsky (1983, p. 96).

4.2. *Hierarchical event-rate model*

To model arrivals or events occurring over known lengths of time we may use an exchangeable hierarchical model. For example, if Y_i is the number of occurrences over an exposure time of length t_i, $i = 1, \ldots, k$, we might assume that each Y_i is a realization from an independent Poisson process having constant rate λ_i, that is, $Y_i^{\text{ind}} \sim P_0(\lambda_i t_i)$. We then assume that the λ_i are independent and identically distributed from some second stage distribution π. The conjugate choice for π would be a gamma distribution, so that the complete conditional distributions for the λ_i are updated gammas (see Gelfand and Smith, 1990). However, in order to allow for more dispersion and possible outliers in the rates, we might prefer a lognormal or logstudent-t prior for the λ_i, neither admitting closed-form complete conditionals for the λ_i.

To develop the competing models more explicitly in the gamma case, we have at the second stage $\lambda_i^{\text{iid}} \sim \text{gamma}(\alpha, \beta)$, $i = 1, \ldots, k$, where for convenience α is a known tuning constant. At the third stage of the hierarchy, we suppose $\beta \sim IG(c, d)$, where IG denotes the inverse (reciprocal) gamma distribution having mode $d/(c-1)$, and c and d are known constants. In the log t case, letting $\varepsilon_i = \log \lambda_i$, we have $\varepsilon_i^{\text{iid}} \sim t_\omega(\theta, \sigma)$, where θ and σ are unknown location and scale parameters, respectively, and ω is specified (note that this parameter has nothing to do with ν, the degrees of our envelope split-t distribution). At the third stage of this model, we suppose $\theta \sim N(\mu, \tau^2)$ and $\sigma^2 \sim IG(a, b)$, θ and σ^2 independent, μ, τ^2, a and b known. Taking ω sufficiently large leads to the lognormal model for λ_i mentioned above.

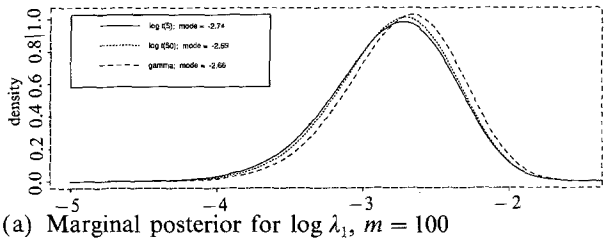
Implementation of the Gibbs sampler is routine in the gamma case (for details, see Gelfand and Smith, 1990). In the log t case none of the required $(k+2)$ complete conditionals is a standard distribution and hence, we apply the methods of Section 3. Since the likelihood factors into k pieces each involving only λ_i, θ and σ^2, the remark at the end of Section 3 may be used to reduce the dimensionality of each of the first k component problems. However, to streamline the computer code we chose to ignore these savings, simply using the same $(k+2)$-dimensional f function for each parameter under the parametrization $U_i = \epsilon_i = \log \lambda_i$, $i = 1, \ldots, k$, $U_{k+1} = \theta$, and $U_{k+2} = \log \sigma$. Here the covariance matrix is approximated using a derivative-free numerical Hessian.

Table 2. *Pump failures (t_j in thousands of hours)*

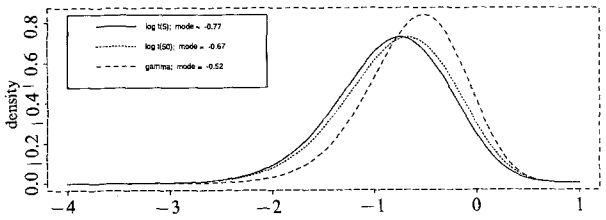
System j	X_j	t_j
1	5	94.320
2	1	15.720
3	5	62.880
4	14	125.760
5	3	5.240
6	19	31.440
7	1	1.048
8	1	1.048
9	4	2.096
10	22	10.480

The data in Table 2 are taken from Worledge *et al.* (1982), and record the number of failures of pumps over given lengths of time for several systems of a certain nuclear power plant. Gaver and O'Muircheartaigh (1987) also fit both the gamma and log t models described above to this data, but employ an empirical Bayes approach, using the data to estimate all the parameters at the second stage of the model instead of placing third stage prior distributions on them. We make our analysis somewhat comparable in the case of the gamma by choosing $\alpha = \hat{\alpha} = 1.802$, the value of the method of moments estimator of α based on the marginal distribution of the data $m(\boldsymbol{Y} \mid \alpha, \beta)$, and taking $c = 2.01$ and $d = 1.01$, so that β has prior mean 1 and prior standard deviation 10. In the log t case, we specify the priors on θ by letting $\mu = -1$, $\tau^2 = 1$, and the prior on σ^2 by letting $a = 2.01$ and $b = 1.01$ (again, a rather vague prior with mean 1 and standard deviation 10). We use split-t distributions taking $\nu = 10$.

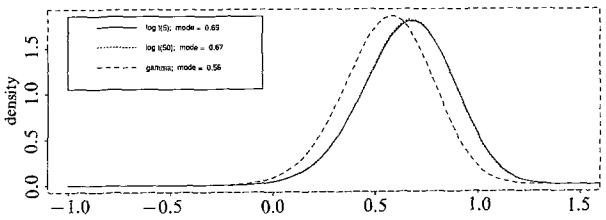
The estimated posterior distributions for ϵ_1, ϵ_5 and ϵ_{10} under the gamma model, the log t model with $\omega = 5$, and the log t model with $\omega = 50$ (essentially a lognormal model) are shown in Fig. 2 using Equation 1 after $t = 30$ iterations with $m = 100$ replications. The results are similar to those of Gaver and O'Muircheartaigh (1987, p. 11). We see that, as expected, the gamma model generally produces posterior distributions that are more highly peaked and less dispersed. Note also that the gamma model seems to encourage more shrinkage to the grand mean of the observed rates. This is especially true for ϵ_5, a rate corresponding to a system having a shorter history (smaller t_i).



(a) Marginal posterior for $\log \lambda_1$, $m = 100$



(b) Marginal posterior for $\log \lambda_5$, $m = 100$



(c) Marginal posterior for $\log \lambda_{10}$, $m = 100$

Fig. 2. *Estimated posteriors, pump failure data*

4.3. *Generalized logistic regression*

As a final illustration we consider a class of models that find broad application in the social and biological sciences, especially in the context of dose–response studies. Suppose we have a Bernoulli response variable Z and a single continuous predictor variable W. Typically one models the probability of a response at a given level of the predictor as

$$P(Z = 1 \mid w) = P(w) = \int_{-\infty}^{y} h(s)\, ds \qquad (10)$$

where $y = (w - \mu)/\sigma$, μ and σ unknown. The most common assumption is to let h be the logistic distribution, which enables the closed-form expression

$$P(w) = \exp y/(1 + \exp y) \qquad (11)$$

that is, the familiar logistic regression model. Prentice (1976) extended Equation 11, introducing the class of models generated by taking

$$h(s) = \exp(sm_1)(1 + \exp(s))^{-(m_1 + m_2)}/\beta(m_1, m_2), \quad m_1, m_2 > 0 \qquad (12)$$

where $\beta(\cdot, \cdot)$ represents the beta function. Prentice remarks that with appropriate choice of m_1, m_2 other familiar models for binary response data emerge. More importantly, he notes the potential improvement in fit afforded by the additional parameters. One convenient special case that enables an explicit form for $P(w)$ is to set $m_2 = 1$, obtaining

$$P(w) = [\exp y/(1 + \exp y)]^{m_1}. \qquad (13)$$

To effect a Bayesian analysis using Equation 13 we need to specify priors on μ, σ^2 and m_1. Broad modeling possibilities arise by letting $m_1 \sim \text{gamma}(a_0, b_0)$, $\mu \sim N(c_0, d_0^2)$, and $\sigma^2 \sim IG(e_0, f_0)$, m_1, μ and σ^2 a priori independent and a_0, b_0, c_0, d_0, e_0 and f_0 known. If we observe X_i responses out of n_i observations at predictor level w_i, $i = 1, \ldots, k$, f = likelihood × prior takes the form

$$\left\{\prod_{i=1}^{k} [P(w_j)]^{X_i}[1 - P(w_i)]^{n_i - X_i}\right\} \times \exp\left\{-\frac{1}{2}\frac{(\mu - c_0)^2}{d_0^2} - \frac{m_1}{b_0} - \frac{f_0}{\sigma^2}\right\}(m_1^{a_0 - 1}/\sigma^{2(e_0 + 1)})$$

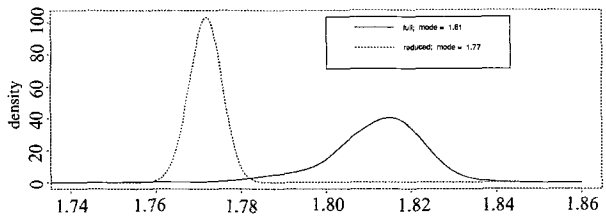
with $P(w)$ as given in Equation 13. Again, the three complete conditional distributions will be sampled using the rejection method.

Our illustrative data set for this model is taken from Bliss (1935), and gives the proportion of adult flour beetles killed after five hours of exposure to various levels of gaseous carbon disulphide. These data, displayed in Table 3, have been much analyzed in the literature since their variability cannot be adequately explained by the standard logit model (Equation 11). We will compare the posterior distribution of μ (= LD_{50} in the dose–response context) under the 'full' model (Equation 13) and the 'reduced' model (Equation 11).

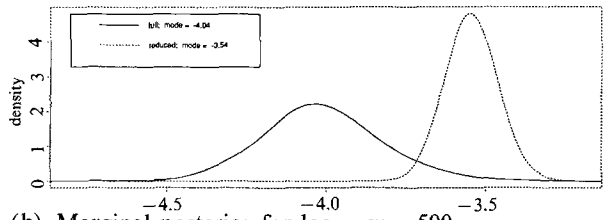
For prior specification, we let $a_0 = 0.25$ and $b_0 = 4$, so that m_1 has prior mean 1 (the 'reduced' value) and prior variance 4. We take rather vague priors on μ and σ^2 by letting $c_0 = 2$, $d_0 = 10$, $e_0 = 2.000004$, and $f_0 = 0.001$ (so that σ^2 has prior mean 0.001 and prior standard deviation 0.5). Using the obvious parametrization $U_1 = \mu$, $U_2 = \log \sigma$, and $U_3 = \log m_1$, empirical evidence suggests a multivariate split-t distribution with $\nu = 3$ to ensure adequate domina-

Table 3. *Observed flour beetle mortality data*

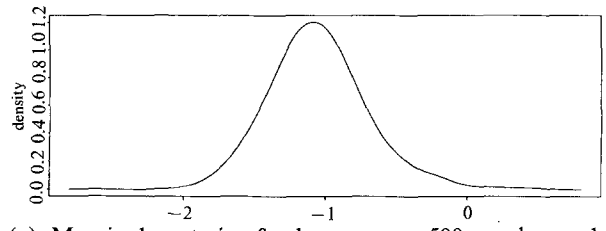
Dosage CS_2	*No. of beetles*	
	killed	*exposed*
1.6907	6	59
1.7242	13	60
1.7552	18	62
1.7842	28	56
1.8113	52	63
1.8369	53	59
1.8610	61	62
1.8839	60	60



(a) Marginal posterior for LD_{50}, $m = 500$



(b) Marginal posterior for $\log \sigma$, $m = 500$



(c) Marginal posterior for $\log m_1$, $m = 500$; mode $= -1.09$

Fig. 3. *Estimated posteriors, beetle data*

tion of f in the tails. The covariance matrix is approximated using the quadratic regression approach.

Figure 3(a) shows the posterior distribution of LD_{50} under the full model (solid line) and the reduced model (dashed line) arising from $t = 50$ iterations with $m = 500$ replications. The full model posterior mode of 1.81 is very close to the MLE value $\hat{\mu} = 1.82$ reported by Prentice (1976). The small overlap between the two posteriors in Fig. 3(a) is consistent with the alleged lack of fit of the reduced model. Figure 3(b) shows similar problems in estimating $U_2 = \log \sigma$; its posterior distribution under the reduced model is also inappropriately centered and too highly concentrated. Still further evidence of the inadequacy of the reduced model is provided by the marginal posterior of $U_3 = \log m_1$ in Fig. 3(c). Here we see that the value assumed under the reduced model, $U_3 = 0$, is located in the extreme right-hand tail of the estimated posterior.

5. Summary and comments

Use of the Gibbs sampler for Bayesian computation is attractive in that, by utilizing complete conditional distributions, multivariate concerns become univariate ones. Moreover, previous work (Gelfand and Smith, 1990; Gelfand *et al.*, 1990; Carlin *et al.*, 1991) shows this approach to be a reasonably straightforward means for implementing fully Bayesian inference under conjugacy. This paper demonstrates an approach which enables the Gibbs sampler to handle nonconjugate cases as well.

None the less, problems which plague other techniques for Bayesian calculation (such techniques are discussed in Naylor and Smith, 1982; 1988; Smith *et al.*, 1985; 1987; Tierney and Kadane, 1986; Geweke, 1989) will also cause difficulties for the Gibbs sampling approach. Such problems include disagreement between likelihood and prior, parametrization and flatness of the likelihood, strong posterior dependence among the parameters and, of course, high dimensionality. In such situations, successful use of the Gibbs sampler will require 'tweaking'. The recent paper of Hills and Smith (1991) expands upon these issues and provides an approach to assist with the required fine tuning.

We anticipate (and our examples support this) that the approach of Section 3 will be most effective in situations involving complicated likelihood but relatively low dimension or in higher-dimensional situations as described at the end of Section 3 where for any parameter the number of other model parameters entering into its complete conditional distribution will be somewhat small. This is because obtaining $\hat{U}$, $\hat{\Sigma}^{1/2}$, and so on, will become more difficult with increasing dimensionality. Our approach is intended for well-behaved (though, for example, not necessarily log-concave) likelihood $\times$ prior forms. Multiple modes, shoulders, very heavy tails, and so on, will be problematic. Finally, since our multivariate envelope is not guaranteed to cover f, our sampling is only approximately from the desired distributions.

Other promising tailored rejection methods for the Gibbs sampler are described in Gilks and Wild (1991), Tanner (1991, p. 101) and Wakefield *et al.* (1991). All of these methods create a distinct envelope function for each change of conditional information.

Acknowledgements

The authors acknowledge J. Berger for suggesting the use of Geweke's split distributions in the case of nonconjugate Gibbs sampling. A. E. Gelfand's research was supported in part by NSF Grant DMS-8918563.

References

Berger, J. O. (1985) *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, New York.

Berger, J. O. and Bernardo, J. (1989) Estimating a product of means: Bayesian analysis with reference priors. *Journal of American Statistical Association*, **84**, 200–207.

Bernardo, J. (1979) Reference posterior distributions for Bayesian inference (with discussion). *Journal Royal Statistics Society B*, **41**, 113–147.

Bliss, C. I. (1935) The calculation of the dosage–mortality curve. *Annals of Applied Biology*, **22**, 134–167.

Carlin, B., Gelfand, A. and Smith, A. F. M. (1991) Hierarchical Bayesian analysis of change point problems. *Applied Statistics* (forthcoming).

Dellaportas, P. and Smith, A. F. M. (1991) Bayesian inference for generalized linear models via Gibbs sampling. *Applied Statistics* (forthcoming).

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

Diaconis, P. and Ylvisaker, D. (1979) Conjugate priors for exponential families. *Annals of Statistics*, **7**, 269–281.

Gaver, D. P. and O'Muircheartaigh, I. G. (1987) Robust empirical Bayes analysis of event rates. *Technometrics*, **29**, 1–15.

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

Gelfand, A. E. and Smith, A. F. M. (1991) Gibbs sampling for marginal posterior expectations. *Communications in Statistics* (to appear).

Gelfand, A. E., Hills, S. E., Racine-Poon, A. and Smith, A. F. M. (1990) Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of American Statistical Association*, **85**, 972–985.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Geweke, J. (1989) Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, **57**, 1317–1339.

Gilks, W. and Wild, P. (1991) Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* (forthcoming).

Hills, S. E. (1989) The parametrization of statistical models. Unpublished PhD thesis, Dept of Mathematics, University of Nottingham.

Hills, S. E. and Smith, A. F. M. (1991) Parametrization issues in Bayesian inference, in *Bayesian Statistics*, Bernardo, J., Berger, J., Dawid, P. and Smith, A. F. M. (eds) (to appear).

Kass, R. E. (1987) Computing observed information by finite differences. *Communications in Statistics, B*, **16**, 587–599.

Morris, C. (1983) Natural exponential families with quadratic variance functions: statistical theory. *Annals of Statistics*, **11**, 515–529.

Naylor, J. and Smith, A. F. M. (1982) Applications of a method for the efficient computation of posterior distributions. *Applied Statistics*, **31**, 214–225.

Naylor, J. and Smith, A. F. M. (1988) Econometric illustrations of novel numerical integration strategies for Bayesian inference. *Journal of Econometrics*, **38**, 103–126.

Prentice, R. L. (1976) A generalization of the probit and logit model for dose response curves. *Biometrics*, **32**, 761–768.

Ratkowsky, D. (1983) *Nonlinear Regression Modeling*, Marcel Dekker, New York.

Ripley, B. (1987) *Stochastic Simulation*, J. Wiley and Sons, New York.

Smith, A. F. M., Skene, A. M., Shaw, J. E. H. *et al.* (1985) The implementation of the Bayesian paradigm. *Communications in Statistics, Theory and Methods*, **14**, 1079–1102.

Smith, A. F. M., Skene, A. M., Shaw, J. E. H. and Naylor, J. C. (1987) Progress with numerical and graphical methods for Bayesian statistics. *The Statistician*, **36**, 75–82.

Tanner, M. (1991) *Tools for Statistical Inference: Observed Data and Data Augmentation Methods*, Lecture Notes in Statistics 67, Springer-Verlag, New York.

Tierney, L. and Kadane, J. B. (1986) Accurate approximations for posterior moments and marginal. *Journal of American Statistical Association*, **81**, 82–86.

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

Worledge, D. H., Stringham, R. S. and McClymont, A. S. (1982) PWR power plant reliability data. Interim Report NP-2592, Electric Power Research Institute, Palo Alto, CA.

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