Bayesian computational methods

By Adrian F. M. Smith

Department of Mathematics, Imperial College of Science, Technology and Medicine, London SW7 2BZ, U.K.

The bayesian (or integrated likelihood) approach to statistical modelling and analysis proceeds by representing all uncertainties in the form of probability distributions. Learning from new data is accomplished by application of Bayes's Theorem, the latter providing a joint probability description of uncertainty for all model unknowns. To pass from this joint probability distribution to a collection of marginal summary inferences for specified interesting individual (or subsets of) unknowns, requires appropriate integration of the joint distribution. In all but simple stylized problems, these (typically high-dimensional) integrations will have to be performed numerically. This need for efficient simultaneous calculation of potentially many numerical integrals poses novel computational problems. Developments over the past decade are reviewed, including adaptive quadrature, adaptive Monte Carlo, and a variant of a Markov chain simulation procedure known as the Gibbs sampler.

1. Introduction

The bayesian (or integrated likelihood) approach to statistical modelling proceeds by specifying a probability model, $p(x\,|\,\theta)$, for data realizations, x, together with an *a priori* (prior) probability weighting distribution, $p(\theta)$, for possible values of the unknown model parameters, θ. Statistical analysis (i.e. making inferences about unknown parameters, or predicting future data values) is then based on $p(\theta\,|\,x)$, the *a posteriori* (posterior) uncertainty distribution for unknown parameters (conditional on observed data), given by Bayes's Theorem,

$$p(\theta\,|\,x) = p(x\,|\,\theta)\,p(\theta)/\int p(x\,|\,\theta)\,p(\theta)\,\mathrm{d}\theta, \tag{1.1}$$

and the (predictive) uncertainty distribution for future data y generated by $p(y\,|\,\theta)$ is given by

$$p(y\,|\,x) = \int p(y\,|\,\theta)\,p(\theta\,|\,x)\,\mathrm{d}\theta.$$

In formal terms, this bayesian inference prescription is deceptively simple: we just specify $p(x\,|\,\theta)$ (usually referred to as the likelihood, $l(\theta\,;x)$, when considered as a function of θ) and $p(\theta)$, multiply the likelihood and prior probability functions together and normalize to form the posterior probability function.

From this joint posterior density over all unknown model parameters, by standard probability marginalization and transformation techniques we can straightforwardly obtain any particular univariate or bivariate, etc., summary in the form of densities, contours or moments as required.

However, calculation of the joint density for θ, together with the required marginalization and moment summaries, rests on an ability to perform a number of

Phil. Trans. R. Soc. Lond. A (1991) **337**, 369–386
Printed in Great Britain

(typically high-dimensional) integrations. In particular, it is necessary to find the normalizing constant of the full joint density and to eliminate complementary components of θ, or transformations of θ, to obtain marginal densities or marginal moment summaries. The technical problem of carrying out a bayesian analysis therefore reduces to that of performing or approximating a number of (typically high-dimensional) numerical integrals.

In practice, the mathematical forms of $l(\theta; x)$ and/or $p(\theta)$ make this process far from trivial, as we shall try to indicate by the following brief comments on a range of models and applications.

(*a*) *Image analysis*

Problems of image analysis, reconstruction, edge and feature detection, etc., arise throughout the sciences, medicine and technology. In these contexts, θ is the unknown true image, the θ-vector components being the true individual pixel values. The data vector x is the observed (noisy) image, with $p(x|\theta)$ describing the noise process and taking characteristically different forms (often the result of an intricate modelling process) in different application areas. The prior specification $p(\theta)$ characterizes (aspects of) the underlying true image and could take the form of a detailed probabilistic template, or a stylized statement about local correlation. Clearly, challenging computational problems arise, the dimension of θ often being of the order of 256×256 (and perhaps even 1000×1000).

(*b*) *Mixture analysis*

Titterington *et al.* (1986) detail numerous application areas where the data model, $p(x|\theta)$, is a weighed linear combination (a finite mixture) of component models. Thus, for example, x may be observed fish length, from a given species, and the component models correspond to length distributions at different fish ages. The unknown mixture weights are the proportions of different age groups in the underlying fish population; the unknown parameters of the component models characterize mean size and variability at different ages. Here, the prior specification, $p(\theta)$, needs to describe the relationships among these latter parameters, as well as information on age distribution. In the case of fisheries research, inference may focus on the mixture proportions; in other applications of mixtures, interest may focus on the component model parameters, or on deciding the number of components. Computational problems in mixture analysis arise both from potentially high dimensionality of θ (for example, a 12-component bivariate gaussian mixture has 71 unknown parameters) and the notoriously complex form of the likelihood function.

(*c*) *Change-point analysis*

Change-point models arise when a mechanism or structure (e.g. biological, economic or social) is characterized by periods of stability (although subject to statistical variability), but with sudden shifts from one stable form to another. Examples are provided by growth phases in biology, shifts in economic behaviour, archaeological phases corresponding to abandonment and resettlement of sites, and changes in clinical conditions, such as alternating periods of rejection and recovery in organ transplantation. The data model, $p(x|\theta)$, has components for each stable phase, with θ representing parameters for each component, as well as the unknown change-points (which, typically, refer to time). The prior specification needs to represent information about the change mechanism, as well as relationships among

parameters within and between phases. Again, dimensionality quickly becomes a problem (for example, a heteroscedastic segmented multiple regression with 10 regressors and up to five phases has at least 64 unknown parameters) and there are often awkward parametric constraints (for example, if segmented response curves are assumed continuous at change-points). In many applications, inference focuses on the change-points themselves. In some cases, for example economic forecasting, interest may focus on inference or prediction based just on the identified final data régime.

(*d*) *Hierarchical analysis*

Hierarchical models, often referred to under the label of Empirical Bayes models, arise as follows. The data model, $p(x\,|\,\theta)$, is a combination of data models, $p_i(x_i\,|\,\theta_i)$, from a number of separate contexts, and $p(\theta)$ has the structure

$$p(\theta) = \int p(\theta\,|\,\phi)\,p(\phi)\,\mathrm{d}\phi,$$

where $p(\theta\,|\,\phi)$ is a further layer of modelling describing an assumed relationship among the θ_i (with the 'hyperparameter', ϕ, functioning as an unknown parameter in $p(\theta\,|\,\phi)$). Examples include: population modelling in the pharmaceutical sciences, where the $p_i(x_i\,|\,\theta_i)$ model drug concentration profiles for individuals i, having kinetic and measurement noise parameters θ_i, with $p(\theta)$ modelling the variability of individual kinetic parameters in the population, as well as information about measurement noise; spatial modelling in epidemiology, where the $p_i(x_i\,|\,\theta_i)$ model disease incidence rates in different locations, with $p(\theta)$ modelling the spatial and other variability of underlying factors. Dimensionality again becomes a problem: for example, if drug concentration profiles for 50 individuals are modelled by a four-parameter function with additive gaussian noise, the combined vector of unknowns (θ, ϕ) may be 264-dimensional (even more if models are extended to deal with potential outliers or the need for data transformation). In many applications, interest focuses on the population characteristics, described by ϕ. In other cases, the individual θ_i (or predictions for individuals or locations) are of equal interest.

This brief summary of a selection of typical problem areas in which bayesian modelling and analysis is used serves, we hope, to indicate and illustrate the very real computational challenges posed by the dimensionality and complexity arising from realistic forms of $p(x\,|\,\theta)$ and $p(\theta)$. The remainder of this paper presents a review of some of the approaches developed over the past decade in response to these computational challenges.

In §2, we review various analytic approximation strategies that have been proposed. In §3, we outline the ways in which a bayesian response to the problem of simultaneously performing a number of high-dimensional numerical integrals suggests novel iterative, adaptive strategies based on mixes of cartesian product, spherical and importance sampling quadrature techniques. In §4, we outline recent ideas based on sampling and resampling strategies. In §5, we provide a brief comparative overview of the various techniques reviewed.

2. Analytic approximation

For completeness, we begin with a brief review of familiar analytic approaches to approximating posterior inference summaries.

(a) *The assumption of asymptotic normality*

If $\ln l(x;\theta)$ is denoted by $\mathrm{L}(\theta)$, suppressing the dependence on the given data x, it is well known that, for large sample sizes, the posterior distribution of θ is often approximately $N(\hat{\theta}, \hat{\Sigma})$, where $\hat{\theta}$ is the maximum likelihood estimate and $\hat{\Sigma}$ is the inverse of the hessian matrix evaluated at $\hat{\theta}$.

This approach has the powerful advantage that practically all forms of summary posterior inference are trivially calculated as by-products of the normality assumption. Moreover, computer programs are readily available for calculating $\hat{\theta}$ and $\hat{\Sigma}$: indeed, apart from differences in philosophical approach and interpretation, these latter quantities are precisely those widely used as the basis for summary inferences by proponents of likelihood inference.

However, there is a major problem with this strategy and one that receives far too little attention in practice, namely, how to check, in any specific application, that the assumption of approximate normality is really justified.

(b) *Lindley's approximations*

Lindley (1980) develops specific expansions to order n^{-1} for ratios of integrals of the form

$$\int w(\theta)\, l(x;\theta)\, \mathrm{d}\theta \Big/ \int v(\theta)\, l(x;\theta)\, \mathrm{d}\theta. \tag{2.1}$$

Thus, for example, with $w(\theta) = g(\theta)\, p(\theta), v(\theta) = p(\theta), \rho(\theta) = \ln p(\theta)$, Lindley shows that

$$E(g(\theta)\,|\,x) \approx g + \tfrac{1}{2}\sum_{i,j} (g_{ij} + 2g_i \rho_i)\, \sigma_{ij} + \tfrac{1}{2} \sum_{i,j,l,m} L_{ijm}\, g_l\, \sigma_{ij}\, \sigma_{ml}, \tag{2.2}$$

where subscripts on g, ρ and L denote differentiation with respect to specific components of θ, with all functions evaluated at $\theta = \hat{\theta}$, and the σ_{ij} are the elements of $\hat{\Sigma}$.

Lindley's expansions provide, in a sense, first-order corrections to the maximum likelihood approximations which incorporate some influence from the prior distribution. Unfortunately, however, the approximations require the evaluations of up to third-order derivatives, a task which quickly becomes irksome as the parameter dimension k increases.

(c) *The Laplace approximation*

More recently, Tierney & Kadane (1986) proposed a form of analytic approximation, which requires the evaluation of only first and second derivatives of slightly modified likelihood functions. The basic idea is to apply separately the Laplace method for integrals to both the numerator and denominator in (2.1). We can illustrate the method by considering again the special case $w(\theta) = g(\theta)\, p(\theta)$, $v(\theta) = p(\theta)$, for positive $g(\cdot)$. We take

$$M(\theta) = n^{-1}(L(\theta) + \ln p(\theta)) \approx M(\hat{\theta}_M) - (\theta - \hat{\theta}_M)^2/(2\sigma_M^2), \tag{2.3}$$

where $\hat{\theta}_M$ maximizes $M(\theta)$ and σ_M^2 is minus the inverse of the second derivative of $M(\cdot)$ evaluated at $\hat{\theta}_M$, and then approximate the denominator of (2.1) by

$$\int l(x;\theta)\, p(\theta)\, \mathrm{d}\theta = \int \mathrm{e}^{nM(\theta)}\, \mathrm{d}\theta \approx \sqrt{(2\pi)}\, \sigma_M\, n^{-\frac{1}{2}} \exp\{nM(\hat{\theta}_M)\}. \tag{2.4}$$

To approximate the numerator, we take

$$N(\theta) = n^{-1}\{\ln g(\theta) + L(\theta) + \ln p(\theta)\}$$

and evaluate the maximum $\hat{\theta}_N$ of $N(\theta)$, together with σ_N^2, which is equal to minus twice the second derivative of $N(\cdot)$ evaluated at $\hat{\theta}_N$, obtaining

$$\int g(\theta)\, l(x;\theta)\, p(\theta)\, \mathrm{d}\theta \approx \sqrt{(2\pi)}\, \sigma_N\, n^{-\frac{1}{2}} \exp\{nN(\hat{\theta}_N)\}. \tag{2.5}$$

From (2.4) and (2.5) we obtain the approximation

$$E[g(\theta)\,|\,x] \approx (\sigma_N/\sigma_M) \exp\{n[N(\hat{\theta}_N) - M(\hat{\theta}_M)]\}, \tag{2.6}$$

which has relative error of order n^{-2} and seems to be more accurate than conventional approximations for a range of problems.

The above developments have obvious direct application to the calculation of posterior moments and predictive densities. In addition, Tierney & Kadane extend these ideas to multiparameter situations and employ Laplace's method to perform the approximate integration of subsets of components of the parameter vector from the overall joint posterior density in order to obtain marginal joint posterior densities. Recent extensions and refinements of the methodology include Kass *et al.* (1989) and Tierney *et al.* (1989). An alternative analytic approximation technique is discussed by Leonard *et al.* (1989).

3. Numerical integration

(*a*) *Background*

As noted in Naylor & Smith (1982), the numerical calculation and marginalization of $p(\theta\,|\,x)$, as given in (1.1) above, involves the computation of integrals of the form

$$S_I[q(\theta)] = \int q(\theta)\, l(x;\theta)\, p(\theta)\, \mathrm{d}\theta_{I'}, \tag{3.1}$$

where, if $\theta_1, \ldots, \theta_k$ denote the components of θ, I is some index set $I \subseteq \{1, \ldots, k\}$ and $\theta_{I'}$, denotes the vector of components of θ whose subscripts are not elements of I. Thus for example, $I = \varnothing$, $q(\theta) = 1$ corresponds to the normalizing constant in (1.1), $S_I[1]/S_\varnothing[1]$ corresponds to the marginal posterior density ordinate $p(\theta_i\,|\,x)$, and $S_\varnothing[\theta_i\,\theta_j]/S_\varnothing[1]$ corresponds to the posterior expectation of $\theta_i\,\theta_j$. Other moments and predictive densities, etc., are similarly obtained from integrals having the form (3.1).

In the case of the k-dimensional integrals,

$$S[q(\theta)] = \int q(\theta)\, l(x;\theta)\, p(\theta)\, \mathrm{d}\theta, \tag{3.2}$$

it is well known that efficient quadrature formulae are available if the integrand in (3.2) can be well approximated by a function of the form

$$g(\theta) = h(\theta)\, n(\theta;\phi,\Sigma), \tag{3.3}$$

where $n(\theta;\phi,\Sigma)$ denotes a k-dimensional normal density, with known mean ϕ and known covariance matrix Σ, and $h(\theta)$ is a polynomial in the components of θ. In such cases, a constructive orthogonalizing and centring and scaling transformation of the form

$$\begin{aligned} \psi_1 &= \theta_1, \\ \psi_i &= \theta_i + \sum_{j=1}^{i-1} \beta_{ij}\, \psi_j, \quad i = 2, \ldots, k, \end{aligned} \tag{3.4}$$

with $$\beta_{ij} = -\text{cov}\,(\theta_i, \psi_j \,|\, x)/\text{var}\,(\psi_j \,|\, x),$$

followed by $$\xi_i = (\psi_i - \mu_i)/(\sqrt{2}\sigma_i), \tag{3.5}$$

where μ_i, σ_i^2 are the means and variances of the ψ_i, leads to the standardized form

$$\int g(\theta)\,\mathrm{d}\theta = \int h^*(\xi)\exp\{-(\xi_1^2 + \ldots + \xi_k^2)\}\,\mathrm{d}\xi, \tag{3.6}$$

where $h^*(\xi)$ is a polynomial in the components of ξ.

The class of functions which can be well approximated by polynomial × normal forms is, in fact, rather rich and covers many of the $S_I[q(\theta)]$ integrands we typically encounter, provided the individual parameter components have support $(-\infty, \infty)$. In many problems this is, of course, not the case since the likelihood involves variance components or proportions defined in the interval $(0, 1)$, or whatever. However, the range of application of the normal × polynomial approximation is greatly extended by, in such cases, redefining the likelihood and prior in terms of transformed individual components: for example, by working with the logarithms of variance components and the logits of proportions. Visual evidence of the effectiveness of such transformations is provided in Smith *et al.* (1985, 1987), and such individual parameter transformations are a key feature of the strategies to be described later.

(*b*) *Parametrization issues*

A diagnostic plot for assessing approximate marginal posterior normality and suggesting effective reparametrization can be developed as follows. Let $p(\theta \,|\, x)$ denote the joint posterior density, with a mode at $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)$, and let $p_j(\theta_j \,|\, x)$ denote the profile posterior for θ_j, defined by maximizing $p(\theta \,|\, x)$ over $\theta_i, i \neq j$, for each θ_j. A diagnostic proposed by Hills & Smith (1991*a*, *b*) is then to plot $T(\theta_j)$ against θ_j, where

$$T(\theta_j) = \text{sgn}\,(\theta_j - \hat{\theta}_j)\{-2[\ln p_j(\theta_j \,|\, x) - \ln p(\hat{\theta} \,|\, x)]\}^{\frac{1}{2}}.$$

To motivate this procedure, consider the case of $p = 1$. If $p(\theta \,|\, x)$ is actually normal, we have

$$p(\theta \,|\, x) = p(\hat{\theta} \,|\, x)\exp\{-(1/2\sigma^2)(\theta - \hat{\theta})^2\},$$

for some σ^2, from which it follows immediately that $T(\theta)$ is linear in θ. In general, a Taylor expansion gives

$$\ln p(\theta \,|\, x) = \ln p(\hat{\theta} \,|\, x) + \tfrac{1}{2}H(\theta - \hat{\theta})^2 + O[(\theta - \hat{\theta})^3],$$

where H is the second derivative of $\ln p(\theta \,|\, x)$ evaluated at $\hat{\theta}$. It follows that

$$T(\theta) = (\theta - \hat{\theta})[-H - 2O[(\theta - \hat{\theta})^3]/(\theta - \hat{\theta})^2]^{\frac{1}{2}}.$$

The first term in the square root expression is observed Fisher information (and does not depend on θ). If this term is large, it will dominate and $T(\theta)$ will be approximately linear in θ (the 'large sample' case). Departure from linearity will then indicate the importance of the cubic term.

This '$T(\theta)$ against θ' diagnostic plot (which we shall refer to as the Bayes t-plot) is discussed in detail in Hills & Smith (1991, 1992) where the behaviour of the diagnostic plot is studied for various stylized cases. As an example, suppose that in fact $\ln(\theta - c)$ were distributed as $N(\mu, \sigma^2)$. In this case, it can be shown that

$$T(\theta) = \sigma^{-1}[\ln(\theta - c) - (\mu - \sigma^2)],$$

which has an asymptote at $\theta = c$.

Based on this and other theoretical forms of $T(\theta)$ against θ, 'look-up' rules can be established (see Hills & Smith (1991) for detailed derivations and illustration). In practice, accuracy of choices of constants (such as c in the above) is not critical. In any case, a second diagnostic check can be carried out for the suggested reparametrization.

More specific forms of reparametrization strategy may be suggested by particular statistical model classes. Consider, for example the autoregressive moving average processes, an ARMA (p, q) being specified by $\phi(B)\, y_t = \theta(B)\, \epsilon_t$, where y_t denotes an observation at time t, ϵ_t denotes white noise, B is the backward shift operator, $B^k y_t = y_{t-k}$, $\phi(B) = 1-\phi_1 B-\phi_2 B^2-\ldots-\phi_p B^p$ and $\theta(B) = 1-\theta_1 B-\theta_2 B^2-\ldots-\theta_q B^q$. For stationarity, the roots of $\phi(B) = 0$ must lie outside the unit circle; for invertibility (to ensure a unique model corresponding to the likelihood) the roots of $\theta(B) = 0$ must lie outside the unit circle. These conditions induce a constraint region, $C_p \times C_q$, for the model parameters (ϕ, θ), where $\phi = (\phi_1, \ldots, \phi_p)$, $\theta = (\theta_1, \ldots, \theta_q)$.

To motivate a suitable reparametrization in this case, first recall that a polynomial in x, with real coefficients,

$$1-\tau_1 x-\tau_2 x^2-\ldots-\tau_m x^m,$$

can be factorized as

$$\prod_{j=1}^{\frac{1}{2}m} (1-a_{1j} x-a_{2j} x^2) \quad \text{for} \quad m \text{ even},$$

$$(1-a_0 x) \prod_{j=1}^{\frac{1}{2}(m-1)} (1-a_{1j} x-a_{2j} x^2) \quad \text{for} \quad m \text{ odd},$$

where the a are also real. When the polynomial corresponds to $\phi(B)$ or $\theta(B)$, the condition that (ϕ, θ) lies in $C_p \times C_q$ is satisfied by ensuring that each point (a_{1j}, a_{2j}) lies in the triangle $|a_{2j}| < 1$, $|a_{1j}| < 1-a_{2j}$, with $|a_0| < 1$ if p or q is odd. A suitable reparametrization is then achieved by setting

$$a_{1j}^* = \ln\left(\frac{1-a_{2j}+a_{1j}}{1-a_{2j}-a_{1j}}\right), \quad a_{2j}^* = \ln\left(\frac{1+a_{2j}}{1-a_{2j}}\right),$$

with $a_0^* = \ln[(1+a_0)/(1-a_0)]$ if p or q is odd. For further details (and discussion of the many-to-one problem in passing from the τ to the a) see Marriott & Smith (1991).

After performing individual component transformations of the kind discussed above, the remaining problem of carrying out the linear transformations described in §3*a* is that the β_{ij}, μ_i and σ_i are not known. The strategies now to be described approach this problem by a form of statistical sequential learning process.

(*c*) *An iterative product rule strategy*

The polynomial $h^*(\xi)$, assumed to be of order d, say, appearing in the standardized form (3.6) can always be written as a linear combination of monomials of the form

$$\xi_1^{\alpha_1} \ldots \xi_k^{\alpha_k},$$

where $\alpha_1+\ldots+\alpha_k \leqslant d$ and $\alpha_i \geqslant 0$ for $i = 1, \ldots, k$. The required integral is thus a linear combination of products of the form

$$\prod_{i=1}^{k} \int_{-\infty}^{\infty} \xi_i^{\alpha_i} \exp(-\xi_i^2)\, \mathrm{d}\xi_i.$$

The component integrals in this product are well-approximated by classical Gauss–Hermite quadrature rules, which have the general form

$$\int_{-\infty}^{\infty} f(t) \exp(-t^2)\, \mathrm{d}t \approx \sum_{i=1}^{n} w_i f(t_i),$$

where

$$w_i = 2^{n-1} n! \sqrt{\pi} / \{n^2 [H_{n-1}(t_i)]^2\},$$

and t_i is the ith zero of the Hermite polynomial $H_n(t)$ (see, for example, Davis & Rabinowitz 1984).

This motivates the approximation of the original integral (3.2) by a cartesian product rule, which takes a weighted average of the integrand values at the intersection points of the k-dimensional grid corresponding to the Hermite zeros in each direction, with weights equal to the product of the corresponding weights. The numbers of zeros used (i.e. the choices of n in the Gauss–Hermite rule) can of course be different in the different directions. As a result of the polynomial × normal underlying assumption, following individual parameter component transformation, and provided the grids are chosen large enough, the same zeros and weights will serve to calculate normalizing constants, marginal density values and moments. This is a key feature of the approach, leading to considerable gains in efficiency.

However, as we remarked earlier, the β_{ij}, μ_i and σ_i implicit in the transformations motivating this approximation are unknown. We therefore proceed as follows, by an iterative, adaptive technique. We begin with small grids based on, say, four points in each direction, and initial estimates (possibly based on maximum likelihood) of the mean and covariance matrix of the posterior distribution (which imply estimates of β_{ij}, μ_i and σ_i). Based on these estimates, we perform cartesian product rule integrals, using the implied grid and weights, to obtain new estimates of the first and second moments of the posterior distribution and hence improved estimates of β_{ij}, μ_i and σ_i. This process is then iterated. The successive μ_i and σ_i control the centring and scaling of the grids; the β_{ij} control the orientation.

The general iterative Gauss–Hermite quadrature strategy may therefore be described as follows.

(i) Reparametrize individual parameters so that the resulting working parameters all take values on the real line.

(ii) Using initial estimates of the joint posterior mean vector and covariance matrix for the working parameters, transform further to a centred, scaled, more 'orthogonal' set of parameters.

(iii) Using the derived initial location and scale estimates for these 'orthogonal' parameters, perform cartesian product integration of functions of interest using suitably dimensioned grids.

(iv) Iterate, successively updating the mean and covariance estimates, until stable results are obtained both within and between grids of specified dimension.

The 'convergence' criterion implicit in (iv) above is, frankly, pragmatic in nature, and complete mathematical treatment of this iterative quadrature strategy seems very difficult. Since the algorithm is driven by up-dating of the first and second moment parameters of the gaussian kernel, the process can be viewed as a nonlinear iterative map on this moment space. Study of convergence thus reduces to the study of the behaviour of this nonlinear map. Some progress is reported in Shaw (1988*a*), who exhibits cases in which unique fixed points exist, but also cases with period cycling and even chaotic behaviour.

Table 1. *Comparison of points required for product and spherical rules*

parameter dimension k	degree 5 spherical (2^k+2k)	three-grid product 3^k	degree 7 spherical $(2^{k+1}+4k^2)$	four-grid product 4^k
3	14	27	52	64
4	24	81	96	256
5	42	243	164	1024
6	76	729	272	4096
7	142	2187	452	16384
8	272	6561	768	65536
9	530	19683	1348	262144

(*d*) *An iterative spherical rule strategy*

The problem with the product rule strategy outlined above is that above five or six dimensions it quickly becomes prohibitively expensive in terms of numbers of integrand calculations required, even for small grids: for example, a 4^6 grid requires 4096 evaluations and a 4^7 grid requires 16384 evaluations. In practice, we may need seven- or eight-point rules in many directions to achieve convergence, and this clearly precludes the routine use of product rules in high dimensions.

However, considerable gains in efficiency can be obtained by transforming to a spherical polar coordinate system and constructing optimal integration formulae based on symmetric configurations over concentric spheres. Such rules, discussed in detail in Stroud (1971, §§2.6, 2.7), are based on the observation that, if we make the transformation from ξ to ψ, where

$$\left.\begin{aligned}
\xi_1 &= r\cos\psi_{k-1}\cos\psi_{k-2}\ldots\cos\psi_2\cos\psi_1,\\
\xi_2 &= r\cos\psi_{k-1}\cos\psi_{k-2}\ldots\cos\psi_2\sin\psi_1,\\
\xi_3 &= r\cos\psi_{k-1}\cos\psi_{k-2}\ldots\sin\psi_2,\\
&\vdots\\
\xi_k &= r\sin\psi_{k-1},
\end{aligned}\right\} \tag{3.7}$$

then the integral (3.6) becomes a product of integrals of the form

$$\int_{-\pi}^{\pi}(\cos\psi_j)^{a_j}(\sin\psi_j)^{b_j}\,\mathrm{d}\psi_j, \quad j=1,\ldots,k-1,$$

and

$$\int_0^{\infty}(r^2)^{c-1}\exp(-r^2)\,\mathrm{d}r,$$

for some a_j, b_j, c. The resulting optimal rules then take the form of a product of a Gauss–Laguerre rule for r^2, together with symmetric configurations on the spherical surfaces.

A rule which will integrate a (polynomial of degree d) × (multivariate normal) integrand exactly is called a 'degree d' rule. Table 1 compares the numbers of points required for optimal degree 5 and degree 7 product rules and spherical rules, respectively. The spherical rules are given in Stroud (1971); the product rules are the 3^k and 4^k Gauss–Hermite cartesian product rules.

It is clear from table 1 that spherical rules offer tremendous potential for efficient integration in the range of $4 \leqslant k \leqslant 9$, where the product grid strategy becomes too expensive.

(*e*) *An iterative importance sampling strategy*

The importance sampling approach to numerical integration is based on the observation that, if f and g are density functions,

$$\int f(x)\,\mathrm{d}x = \int [f(x)/g(x)]\,g(x)\,\mathrm{d}x = \int [f(x)/g(x)]\,\mathrm{d}G(x)$$
$$= E_G[f(X)/g(X)],$$

say, which suggests the 'statistical' approach of generating a sample from the distribution G and using the average of the values of the ratio f/g as an unbiased estimator of $\int f(x)\,\mathrm{d}x$. However, the variance of such an estimator clearly depends critically on the choice of G, it being desirable to choose g to be 'similar' to f.

In the univariate case, if we choose g to be heavier-tailed than f and if we work with $Y = g(X)$, the required integral is the expected value of $f[G^{-1}(X)]/g[G^{-1}(X)]$ with respect to a uniform distribution on the interval $(0, 1)$. The resulting periodic nature of the ratio function over this interval then suggests that we are likely to get a reasonable approximation to the integral by generating 'uniformly distributed' random numbers. If f is a function of more than one argument (k, say), an exactly parallel argument suggests that the choice of a suitable g followed by the use of a suitably 'uniform' configuration of points in the k-dimensional unit hypercube will prove an acceptable alternative to the 'costly' procedure of generating 'random' uniformly distributed points in k-dimensions.

However, the effectiveness of all this depends on choosing a suitable G, bearing in mind that we need to have available a flexible set of possible distributional shapes, for which G^{-1} is available explicitly.

A suitable class of univariate distributions for importance sampling (due to Shaw 1988*a*) can be developed as follows.

Let U denote a uniform $U(0, 1)$ random variable and let h denote a monotonic increasing function defined on the interval $(0, 1)$, such that $h(u) \to -\infty$ as $u \to 0$. Now define a family of random variables by

$$X_A = Ah(u) + (1-A)\,h(1-u),$$

where $0 \leqslant A \leqslant 1$. Clearly, $A = \frac{1}{2}$ defines a random variable with a symmetric distribution; as $A \to 0$ or $A \to 1$ we obtain increasingly skewed distributions (in opposite directions). The tail behaviour of the distribution is governed by the choice of the function h.

Among interesting choices of the latter, we note: $h(u) = -[-\ln(u)]^k, k > 0$, which gives the logistic distribution for $k = 1$, $A = \frac{1}{2}$ and the exponential distribution for $k = 1$, $A = 0$; $h(u) = -\tan[\frac{1}{2}\pi(1-u)]$, whose symmetric member ($A = \frac{1}{2}$) is the Cauchy distribution; $h(u) = 1-u^{-k}$, $k > 0$, which generalizes members of the Tukey-λ family of distributions.

If G_A, g_A denote, respectively, the distribution and density functions of X_A, we have:

$$G_A^{-1}(u) = x_A = Ah(u) + (1-A)\,h(1-u), \quad g_A(x) = G_A'(x).$$

If K_A, k_A denote the corresponding forms of G_A, g_A in terms of u, we have

$$K_A(u) = G_A(G_A^{-1}(u)) = [Ah'(u) + (1-A)\,h'(1-u)]^{-1}.$$

These forms guide the choices of h for which the corresponding importance

sampling density is easy to use. Moreover, the moments of these families of distributions are polynomials in A (of corresponding order), the median is linear in A, and so on, so that sample information about such quantities provides (for any given choice of h) operational guidance about the appropriate choice of A.

In addition, the strategy requires the specification of 'uniform' configurations of points in the k-dimensional unit hypercube, a problem which has been extensively studied by number theorists. Systematic experimentation with various suggested forms of 'quasi-random' sequences has identified effective forms of configuration for importance sampling purposes. Based on measures of 'good lattice structure' motivated by the need for efficient calculation of (lower-dimensional) joint densities, Shaw (1988*b*) presents detailed comparison of a number of rational, irrational and irregular quasi-random sequence approaches to generating a configuration in the hypercube. For mathematical details, see Shaw (1988*b*).

If we combine the importance sampling and quasi-random ideas, the resulting general strategy is the following.

(i) Reparametrize individual parameters so that the resulting working parameters all take values on the real line.

(ii) Using initial estimates of the joint posterior mean vector and covariance matrix for the working parameters, transform further to a centred, scaled, more 'orthogonal' set of parameters.

(iii) In terms of these transformed parameters, set

$$g(x) = \prod_{j=1}^{k} g_j(x_j),$$

for 'suitable' choices of $g, j = 1, \ldots, k$.

(iv) use the inverse cumulative distribution function transformation to reduce the problem to that of calculating an average over a 'suitable' uniform configuration (quasi-random sequence) in the k-dimensional hypercube.

(v) Use information from this 'sample' to learn about skewness, tailweight, etc., for each $j = 1, \ldots, k$ and hence choose 'better' $g_j, j = 1, \ldots, k$, as well as revising estimates of the mean vector and covariance matrix.

(vi) Iterate until the sample variance of replicate estimates of the integral value is sufficiently small.

A variant of this strategy, which can be more effective if the standardized posterior density is approximately spherically symmetric, is to transform the quasi-random random configuration in the hypercube to a configuration on a spherical surface. See Shaw (1988*a*) for details.

In implementing these various quadrature and Monte Carlo techniques, it is often convenient to work with hybrid schemes. For example, combining a product rule for parameters of interest with a Monte Carlo rule for nuisance parameters, in order to facilitate eventual inference summaries, graphics, etc., for the parameters of interest.

Of course, all the above ideas can be combined with standard techniques of variance reduction, such as the use of antithetic variates: see, for example, Geweke (1988), with further ideas on useful importance sampling families illustrated in Geweke (1989).

4. Iterated sampling and resampling approaches

In the sequel, densities will be denoted, generically, by square brackets so that joint, conditional and marginal forms appear as $[X, Y]$, $[X \mid Y]$ and $[Y]$. The usual marginalization by integration procedure will be denoted by forms such as $[X] = \int [X \mid Y]^*[Y]$.

(a) *Substitution sampling*

The substitution algorithm for finding fixed point solutions to certain classes of integral equations is a standard mathematical tool. Thus, for example, if X, Y, Z are random variables, so that, in the above notation

$$[X] = \int [X, Z \mid Y]^*[Y], \quad [Y] = \int [X, Y \mid Z]^*[Z], \quad [Z] = \int [Y, Z \mid X]^*[X],$$

the marginal density $[X]$ of X is the fixed point solution of the equation

$$[X] = \int h(X, X')^*[X'],$$

where the kernel is given by

$$h(X, X') = \int [X, Z \mid Y]^*[X'', Y \mid Z']^*[Y', Z' \mid X'],$$

a five-fold integral (with respect to X'', Y, Y', Z, Z').

The key idea of the stochastic substitution algorithm (see Tanner & Wong 1987; Gelfand & Smith 1990) is to estimate $[X]$ by successive stochastic simulation of random variates, drawn from the conditional distributions in the three above equations. The algorithm proceeds as follows: draw $X^{(0)}$ from an arbitrary $[X]_0$; draw $Y^{(0)'}, Z^{(0)'}$ from $[Y, Z \mid X^{(0)}]$; draw $X^{(0)'}, Y^{(1)}$ from $[X, Y \mid Z^{(0)'}]$; draw $X^{(1)}, Z^{(1)}$ from $[X, Z \mid Y^{(1)}]$ and then iterate. After t steps, with m replications of the process, we obtain $(X_j^{(t)}, Y_j^{(t)}, Z_j^{(t)})$, $j = 1, \ldots, m$. An estimate of the marginal density $[X]$ is then provided by

$$[\hat{X}]_t = \frac{1}{m} \sum_{j=1}^{m} [X \mid Y_j^{(t)}, Z_j^{(t)}].$$

In applications to bayesian inference, $[X, Y, Z]$ would denote the posterior distribution of unknown quantities of interest.

(b) *Gibbs sampling*

Throughout this section, we shall be dealing with collections of random variables $U_1, U_2, \ldots, U_k$, for which it is known that the joint density, $[U_1, U_2, \ldots, U_k]$, is uniquely determined by the full conditional densities $[U_s \mid U_r, r \neq s], s = 1, 2, \ldots, k$. Our interest is typically in the marginal distributions, $[U_s], s = 1, 2, \ldots, k$.

An algorithm for extracting marginal distributions from the full conditional distributions (in contrast to the form of conditionals used above in substitution sampling) was formally introduced as the Gibbs sampler in Geman & Geman (1984). The algorithm requires all the full conditional distributions to be 'available' for sampling, where 'available' is taken to mean that, for example, samples of U_s can be generated straightforwardly and efficiently from $[U_s \mid U_r, r \neq s]$, given specified values of the conditioning variables, $U_r, r \neq s$.

Gibbs sampling is a markovian updating scheme, which is a variant of the Metropolis algorithm. See, for example, Hastings (1970) and Peskun (1973) for seminal ideas on the use of Markov chain simulation algorithms for statistical problems. The Gibbs sampling algorithm proceeds as follows. Given an arbitrary starting set of values $U_1^{(0)}, \ldots, U_k^{(0)}$, we draw $U_1^{(1)}$ from $[U_1 \mid U_2^{(0)}, \ldots, U_k^{(0)}]$, then $U_2^{(1)}$ from $[U_2 \mid U_1^{(1)}, U_3^{(0)}, \ldots, U_k^{(0)}] \ldots$ and so on up to $U_k^{(1)}$ from $[U_k \mid U_1^{(1)}, \ldots, U_{k-1}^{(1)}]$ to complete one iteration of the scheme. After t such iterations we would arrive at $(U_1^{(t)}, \ldots, U_k^{(t)})$. Geman & Geman show under mild conditions that

$$U_s^{(t)} \xrightarrow{d} U_s \sim [U_s] \text{ as } t \to \infty.$$

Thus, for t large enough we can regard $U_s^{(t)}$ as a simulated observation from $[U_s]$.

Replicating this process m times produces m independent and identically distributed k-tuples $(U_{1j}^{(t)}, \ldots, U_{kj}^{(t)}), j = 1, \ldots, m$. For any s, the collection $U_{s1}^{(t)}, \ldots, U_{sm}^{(t)}$ can be viewed as a simulated sample from $[U_s]$. The marginal density could then be estimated by the finite mixture density.

$$[\hat{U}_s] = m^{-1} \sum_{j=1}^{m} [U_s \mid U_r = U_{rj}^{(t)}, r \neq s]. \tag{4.1}$$

(See Gelfand & Smith (1990) and Gelfand *et al.* (1990) for further discussion.)

Suppose interest centres on the marginal distribution for a variable V which is a function $g(U_1, \ldots, U_k)$ of $U_1, \ldots, U_k$. We note that evaluation of g at each of the $(U_{1j}^{(t)}, \ldots, U_{kj}^{(t)})$ provides samples of V, so that an ordinary kernel density estimate can readily be calculated.

In the bayesian framework, where U_s are unobservable, representing either parameters or missing data (and V can thus be a function of the parameters in which we are interested), all distributions will be viewed as conditional on the observed data, whence marginal distributions become the marginal posteriors needed for bayesian inference or prediction.

In many applications involving exponential families, some or all of the full conditionals may be familiar density forms from which sampling is straightforward. However, while 'conjugacy' simplifies the implementation of the Gibbs sampler it is not an essential element. In *any* Bayes model the full conditional distribution of any parameter is always identifiable from the joint density of the data and the parameters modulo normalizing constant. Using more sophisticated random variate generation approaches, such as the ratio of uniforms method (Devroye 1986; Wakefield *et al.* 1992), or methods which exploit features like log-concavity (Gilks & Wild 1991; Dellaportas & Smith 1991), we can sample the arbitrary non-normalized densities, although, of course, fine tuning of the sampling methodology, including 'clever' reparametrization, may be required to avoid highly inefficient random variate generation.

If the Gibbs sampler is run in order to generate an 'as if' independent sample from the joint posterior distribution (rather than simply to form estimates by ergodic averaging), this can be attempted either by replicate independent 'short' runs of the process or by extracting multiple sample values from a single 'long' run of the process. Output series need to be monitored by a stopping rule (formal or informal) which decides when 'convergence' has been achieved. In the case of replicate runs, the required 'independent' sample from the posterior is then formed by the final generated values from the stopped series.

The number of iterations to achieve 'convergence' is clearly a function of starting values and the correlation structure of the stochastic process generated by the Gibbs sampler. To try to get some insight into the importance and effect of such correlation (and hence into the possible importance of reparametrization to remove it), let us consider in detail the simple case of two parameters, where the joint posterior is actually zero-mean, unit-variance bivariate normal with correlation ρ. If we initialize the process at θ_2^0, say, the conditional distributions which drive the Gibbs sampler are given by

$$p(\theta_1^t, | \theta_2^{t-1}) \equiv N(\rho\theta_2^{t-1}, 1-\rho^2), \quad p(\theta_2^t | \theta_1^t) \equiv N(\rho\theta_1^t, 1-\rho^2),$$

from which it follows straightforwardly that the joint distribution of (θ_1^t, θ_2^t) has means $\rho^{2t-1}\theta_2^0, \rho^{2t}\theta_2^0$, variances $1-\rho^{4t-2}$, $1-\rho^{4t}$ and correlation $\rho[(1-\rho^{4t})/(1-\rho^{4t-2})]^{\frac{1}{2}}$, so that the effects of θ_2^0 and ρ on the rate of convergence can easily be examined.

The clear (and intuitive) message is that really high correlations disastrously slow down the convergence of the Gibbs sampler and that the higher the correlation the more serious are bad starting values. At the other end of the scale, small correlations (even of the order of $\rho = 0.8$) imply relatively trivial numbers of iterations to convergence in this simple case, with bad starting values quickly forgotten. Breaking high correlation can be achieved by means of an application after a few initial iterations (say, 10–15) of the orthogonalizing transformation defined earlier in §3*c*.

Complete implementation of the Gibbs sampler requires that a determination of t be made and that, across iterations, choice(s) of m specified. See Gelfand *et al.* (1990) for further discussion of convergence issues. See also, Gelfand & Smith (1990) and Carlin *et al.* (1991), as well as a discussion of other markovian updating procedures by Aykroyd & Green (1991).

(*c*) *Resampling*

As a first step towards motivating the resampling approach, we note the essential duality between a sample and the density (distribution) from which it is generated. Clearly, the density generates the sample; conversely, given a sample we can approximately recreate the density (as a histogram, an empirical cumulative distribution function or whatever).

Suppose we now shift the focus in (1.1) from densities to samples. In terms of densities, the inference process is encapsulated in the updating of the prior density, $p(\theta)$, to the posterior density, $p(\theta | x)$, through the medium of the likelihood function, $l(\theta; x)$. Shifting to samples, this corresponds to the updating of a sample from $p(\theta)$ to a sample from $p(\theta | x)$ through the likelihood function $l(\theta; x)$.

Generally, suppose that a sample of random variates is easily generated, or has already been generated, from a continuous density $g(\theta)$, but that what is really required is a sample from a density $h(\theta)$ absolutely continuous with respect to $g(\theta)$. Can we somehow utilize the sample from $g(\theta)$ to form a sample from $h(\theta)$? Slightly more generally, given a positive function $f(\theta)$ which is normalizable to such a density $h(\theta) = f(\theta)/\int f(\theta)\,\mathrm{d}\theta$, can we form a sample from the latter given only a sample from $g(\theta)$ and the functional form of $f(\theta)$?

We may approximately resample from $h(\theta) = f(\theta)/\int f(\theta)\,\mathrm{d}\theta$ as follows. Given θ_i, $i = 1, \ldots, n$, a sample from g, calculate $\omega_i = f(\theta_i)/g(\theta_i)$ and then

$$q_i = \omega_i / \sum_{j=1}^{n} \omega_j.$$

Draw θ^* from the discrete distribution over $\{\theta_1, \ldots, \theta_n\}$ placing mass q_i on θ_i. Then θ^*

is approximately distributed according to h with the approximation 'improving' as n increases (see Smith & Gelfand 1992). Note that this procedure is a variant of the by now familiar bootstrap resampling procedure. The usual bootstrap provides equally likely resampling of the θ_i, while here we have weighted resampling with weights determined by the ratio of f to g. See, also, Rubin (1988), who refers to this procedure as SIR (sampling/importance resampling).

Several obvious uses of this sampling–resampling perspective are immediate. In general, the translation from functions to samples provides a wealth of opportunities for creative exploration of bayesian ideas and calculations in the setting of computer graphical and exploratory data analysis tools. Also, we can easily approach problems of sensitivity of inferences to model specification, such as: How does the posterior change if we change the prior? How does the posterior change if we change the likelihood?

In the density function/numerical integration setting, such sensitivity studies are rather off-putting, in that each change of a functional input typically requires one to carry out new calculations from scratch. This is not the case with the sampling–resampling approach, as we now illustrate in relation to the questions posed above.

In comparing two models in relation to the second question, we note that change in likelihood may arise in terms of (i) change in distributional specification with θ retaining the same interpretation, e.g. a location; (ii) change in data to a larger data-set (prediction), a smaller data set (diagnostics), or a different data-set (validation).

To unify notation, we shall in either case denote two likelihoods by $l_1(\theta)$ and $l_2(\theta)$. We denote two different priors to be compared in relation to the first question by $p_1(\theta)$ and $p_2(\theta)$. For complete generality, we consider changes to both l and p, although in any particular application we would not typically change both. Denoting the corresponding posterior densities by $\tilde{p}_1(\theta), \tilde{p}_2(\theta)$, we easily see that

$$\tilde{p}_2(\theta) \propto [l_2(\theta)\, p_2(\theta)/l_1(\theta)\, p_1(\theta)]\, \tilde{p}_1(\theta). \tag{4.2}$$

Letting $v(\theta) = l_2(\theta)\, p_2(\theta)/l_1(\theta) p_1(\theta)$, we may directly apply the weighted bootstrap method to (4.2) taking $g = \tilde{p}_1(\theta), f = v(\theta)\, \tilde{p}_1(\theta)$ and $\omega_i = v(\theta_i)$. Resampled θ^* will then be approximately distributed according to f standardized, which is precisely $\tilde{p}_2(\theta)$.

5. Overview

(*a*) *Analytic against numerical integration procedures*

Of the analytic approximation techniques available, those based on the Laplace approximation (§2*c*) seem to be the most systematically studied and validated. The choice between these and numerical integration procedures rests largely on the dimensionality and complexity of the problem. As the latter becomes greater, implementation of the analytic techniques becomes very difficult indeed.

(*b*) *Quadrature against Monte Carlo integration*

Here again, dimensionality and complexity of the posterior functional forms are the main determining factors. Roughly speaking, for relatively well-behaved functions (typically following reparametrization) product rules can be effective in up to about six dimensions, with spherical rules extending the domain of quadrature up to nine dimensions. However, beyond that (or even for lower-dimensional problems with badly behaved functions or awkward parameter constraints) Monte Carlo methods are generally necessary.

(c) *Importance sampling against iterative sampling*

Use of importance sampling for high-dimensional problems is something of an art form, both in the choice of effective importance sampling functions and in the design of variance reduction sampling strategies. However, if a good procedure can be found, it is likely to be computationally more efficient than a Markov chain based iterative sampling procedure. On the other hand, the latter, and, in particular, the Gibbs sampler, has the merit of being (typically) very easy to implement, requiring very little numerical or stochastic simulation expertise. Moreover, in very complex problems it may simply prove too difficult to identify any suitable importance sampling strategy.

To give a concrete sense of the different 'flavour' of the two approaches, we return briefly to the mixture analysis problem of §1*b*.

Shaw (1988*c*) provides a detailed analysis for a $k(=5)$ component univariate normal mixture model for fish lengths. Lengths are binned into n class ranges, resulting in N_j fish assigned to class j, corresponding to the interval $[x_j, x_{j+1})$, with $N_1+\ldots+N_n = N$. If (μ_i, σ_i, p_i) denote the k component means, variances and proportions, it is easy to see that the log-likelihood is given by

$$\sum_{j=1}^{n} N_j \ln \hat{f_j} - N \ln (\hat{F}_{n+1} - \hat{F}_1),$$

where

$$\hat{F}_j = \sum_{i=1}^{k} p_i \Phi[(x_j - \mu_i)/\sigma_i],$$

and $\hat{f_j} = \hat{F}_{j+1} - \hat{F}_j$ (with $\Phi(\cdot)$ denoting the standard normal cumulative distribution function). This log-likelihood is then combined with a prior specification for the $3k-1$ $(=14)$ unknown parameters, the prior chosen to reflect basic information about fish growth, variability and abundance in five successive cohorts. In particular, such knowledge constrains all parameters to be positive, the means to be increasing and the proportions to lie in the simplex. This is a case where parameter transformation is vital and Shaw worked with

$$\begin{aligned}
\theta_1 &= \mu_1 - 11 && \text{(for numerical stability)},\\
\theta_{3i-2} &= \ln (\mu_i - \mu_{i-1}) && (i = 2, \ldots, k),\\
\theta_{3i-1} &= \ln \sigma_i && (i = 1, \ldots, k),\\
\theta_{3i} &= \ln \left[p_i / \left(1 - \sum_{j=1}^{i} p_j \right) \right] && (i = 1, \ldots, k-1).
\end{aligned}$$

A detailed exploration and summary of the resulting 14-dimensional posterior density for the θs was accomplished with a quasi-random spherical rule (as described at the end of the §3), using 10000 nodes on two spherical shells. Further details are given in Shaw (1988*c*). The point to note in the context of the present discussion is that the posterior full conditionals for each of the 14 parameters (i.e. for each parameter given the values of the other 13) seem to be extremely messy forms, as a consequence of the complicated form of the log-likelihood, so that Gibbs sampling looks no easier than direct numerical integration.

Now consider the same k component normal model, but with precisely observed (rather than grouped) observations, so that the likelihood takes the form

$$\prod_{l=1}^{N} \sum_{i=1}^{k} p_i \phi[(x_l - \mu_i)/\sigma_i],$$

where $\phi(\cdot)$ denotes the standard normal PDF. A surprisingly simple structure for the Gibbs sampler can be obtained by introducing, as further unknowns, indicator quantities $Z_1, \ldots, Z_N$, such that $Z_l = i$ corresponds to observation x_l actually coming from component i. In this case, if (μ_i, σ_i) are assigned conjugate normal-inverse-gamma priors, $(p_1, \ldots, p_k)$ a Dirichlet prior and the Z_l uniform priors, it is easy to show that successive generation from the full conditionals reduces to: draw the Zs from a specified discrete distribution; the ps from a Dirichlet distribution; the μs from a normal distribution and the σs from an inverse gamma distribution. Substantial iterative computation is then required, but the need for sophisticated numerical understanding on the part of the statistical analyst is obviated.

Much of the author's work reviewed here was supported by the SERC's Complex Stochastic Systems Initiative.

References

Aykroyd, R. G. & Green, P. J. 1991 Global and local priors and the location of lesions using gamma-camera imagery. *Phil. Trans. R. Soc. Lond.* A **337**, 323–342. (This volume.)

Carlin, B. P., Gelfand, A. E. & Smith, A. F. M. 1991 Hierarchical bayesian analysis of change-point problems. *Appl. Statist.* (In the press.)

Davis, P. J. & Rabinowitz, P. 1984 *Methods of numerical integration*, 2nd edn. Orlando, Florida: Academic Press.

Dellaportas, P. & Smith, A. F. M. 1991 Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling. *Appl. Statist.* (In the press.)

Devroye, L. 1986 *Non-uniform random variate generation*. Springer-Verlag: New York.

Gelfand, A. E. & Smith, A. F. M. 1990 Sampling based approaches to calculating marginal densities. *J. Am. statist. Ass.* **85**, 398–409.

Gelfand, A. E., Hills, S. E., Racine-Poon, A. & Smith, A. F. M. 1990 Illustration of bayesian inference in normal data models using Gibbs sampling. *J. Am. statist. Ass.* **85**, 972–985.

Geman, S. & Geman, D. 1984 Stochastic relaxation, Gibbs distributions and the bayesian restoration of images. *IEEE Trans. Patt. Analysis Mach. Int.* **6**, 721–741.

Geweke, J. 1988 Antithetic acceleration of Monte Carlo integration in Bayesian inference. *J. Econometrics* **38**, 73–90.

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

Gilks, W. R. & Wild, P. 1991 Adaptive rejection sampling for Gibbs sampling. *Appl. Statist.* (In the press.)

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

Hills, S. E. & Smith, A. F. M. 1991 Diagnostic plots for improved parameterization in Bayesian inference. (Submitted.)

Hills, S. E. & Smith, A. F. M. 1992 Parameterization issues in Bayesian inference. In *Bayesian statistics 4* (ed. J. Bernardo *et al.*). Oxford University Press.

Kass, R. E., Tierney, L. & Kadane, J. B. 1989 Asymptotics in Bayesian calculation. In *Bayesian statistics 3* (ed. J. M. Bernardo *et al.*), pp. 261–278. Oxford University Press.

Leonard, T., Hsu, J. S. J. & Tsu, K.-W. 1989 Bayesian marginal inference. *J. Am. statist. Ass.* **84**, 1051–1058.

Lindley, D. V. 1980 Approximate Bayesian methods. In *Bayesian statistics* (ed. J. M. Bernardo *et al.*), pp. 223–245. Valencia University Press.

Marriott, J. M. & Smith, A. F. M. 1991 Reparameterization aspects of Bayesian methodology for ARMA models. *J. Time Series Analysis.* (In the press.)

Naylor, J. C. & Smith, A. F. M. 1982 Applications of a method for the efficient computation of posterior distributions. *Appl. Statist.* **31**, 214–225.

Peskun, P. H. 1973 Optimum Monte Carlo sampling using Markov chains. *Biometrika* **60**, 607–612.

Rubin, D. B. 1988 Using the SIR algorithm to simulate posterior distributions. In *Bayesian statistics 3* (ed. J. M. Bernardo *et al.*), pp. 395–402. Oxford University Press.

Shaw, J. E. H. 1988*a* Numerical and graphical techniques for Bayesian inference. Ph.D. thesis, University of Nottingham.

Shaw, J. E. H. 1988*b* A quasi-random approach to integration in bayesian statistics. *Ann. Statist.* **16**, 895–914.

Shaw, J. E. H. 1988*c* Aspects of numerical integration and summarization. In *Bayesian statistics* 3 (ed. J. M. Bernardo *et al.*), pp. 411–428. Oxford University Press.

Smith, A. F M. & Gelfand, A. E. 1992 Bayesian statistics without tears: a sampling-resampling perspective. *Am. Statist.* (In the press.)

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

Smith, A. F. M., Skene, A. M., Shaw, J. E. H., Naylor, J. C. & Dransfield, M. 1985 The implementation of the Bayesian paradigm. *Commun. Statist.* A**14**, 1079–1102.

Stroud, A. H. 1971 *Approximate calculation of multiple integrals.* New Jersey: Prentice-Hall.

Tanner, M. & Wong, W. 1987 The calculation of posterior distributions by data augmentation. *J. Am. statist. Ass.* **82**, 528–550.

Tierney, L. & Kadane, J. B. 1986 Accurate approximation for posterior moments and marginal densities. *J. Am. statist. Ass.* **81**, 82–86.

Tierney, L., Kass, R. E. & Kadane, J. B. 1989 Approximate marginal densities of nonlinear functions. *Biometrika* **76**, 425–434.

Titterington, D. M., Smith, A. F. M. & Makov, U. E. 1986 *Statistical analysis of finite mixture distributions.* Chichester: Wiley.

Wakefield, J. C., Gelfand, A. E. & Smith, A. F. M. 1992 Efficient generation of random variates via the ratio-of-uniforms method. *Statist. Computing.* (In the press.)