J. R. Statist. Soc. B (1993)
55, *No.* 1, *pp.* 3–23

Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods

By A. F. M. SMITH† and G. O. ROBERTS

Imperial College of Science, Technology and Medicine, London, UK *University of Cambridge, UK*

[*Read before* The Royal Statistical Society *at a meeting on 'The Gibbs sampler and other Markov chain Monte Carlo methods' organized by the* Research Section *on Wednesday, May 6th, 1992,* Professor B. W. Silverman *in the Chair*]

SUMMARY

The use of the Gibbs sampler for Bayesian computation is reviewed and illustrated in the context of some canonical examples. Other Markov chain Monte Carlo simulation methods are also briefly described, and comments are made on the advantages of sample-based approaches for Bayesian inference summaries.

Keywords: BAYESIAN STATISTICS; CENSORED DATA; CONSTRAINED PARAMETER MODELS; GENERALIZED LINEAR MODELS; GIBBS SAMPLER; HASTINGS ALGORITHM; HIERARCHICAL MODELS; MARKOV CHAIN MONTE CARLO METHODS; MISSING DATA; TIME SERIES MODELS

1. INTRODUCTION

This paper reviews recent uses of Markov chain Monte Carlo (MCMC) methods for exploring and summarizing posterior distributions in Bayesian statistics. The emphasis and terminology will be Bayesian but, by considering prior specifications as constants, the discussion is equally applicable to (normalizable) likelihood surfaces, or even to non-statistical response surfaces.

In Section 2, we review briefly the computational problems which arise with Bayesian methods and comment on some previous proposals for carrying out such computations. In Section 3, we review basic ideas of MCMC methodology and, in particular, the Gibbs sampler, which will subsequently be the main focus of the paper. In Section 4, we comment on the advantages of sample-based methods for Bayesian computation. Sections 5 and 6 examine a very wide range of canonical problem types, illustrating the ways in which the Gibbs sampler leads to relatively straightforward implementation in many situations which seem highly intractable when viewed from the perspective of other approaches to Bayesian computation. Other forms of MCMC methods are illustrated briefly in Section 7 and a summary discussion is provided in Section 8. The style of presentation throughout is fairly informal, with a more technical summary presented in Appendix A. No attempt is made to provide a complete coverage of the rapidly growing literature of MCMC methods in statistics.

2. BAYESIAN COMPUTATION

In all but very stylized problems, the integrals required for Bayesian computation

†*Address for correspondence*: Department of Mathematics, Imperial College of Science, Technology and Medicine, Huxley Building, 180 Queen's Gate, London, SW7 2BZ, UK.

 0035-9246/93/55003 $2.00

require analytic or numerical approximation. Progress over the past decade towards developing suitable approximation techniques has focused on three main strategies: Laplace and related analytic techniques, adaptive quadrature based on classical numerical analysis and versions of standard Monte Carlo importance sampling. All have contributed to extending the Bayesian computational toolkit, but all suffer from limitations on their scope and ease of implementation. See Smith (1991) for a recent overview and detailed references.

The ideas that we shall be discussing in this paper are closest in spirit to the standard Monte Carlo methods, which estimate features of the posterior or predictive distribution of interest by using samples drawn from that distribution, or suitably reweighted samples drawn from some other appropriately chosen distribution. Often, particularly in high dimensional problems, this may be the only feasible approach. However, even when alternatives are available, there may be other considerations which make a Monte Carlo approach attractive. We shall expand more fully on what we see as the general advantages of sample-based approaches in Section 4. For now, we simply note the obvious ease with which a sample from a joint density can be used to form estimates of any distributions or functions of interest.

However, directly generating samples from an arbitrary, often high dimensional, joint distribution is in general not possible, thus seemingly making sample-based approaches of limited use. The methods described in this paper overcome this problem by an indirect approach to the required sampling based on Markov chains.

3. MARKOV CHAIN MONTE CARLO METHODS

3.1. *Markov Chain Monte Carlo*

The key idea is very simple. Suppose that we wish to generate a sample from a distribution $\pi(x)$ for $x \in \mathscr{X} \subseteq \mathbf{R}^n$ but cannot do this directly. However, suppose that we can construct a Markov chain with state space $\mathscr{X}$, which is straightforward to simulate from and whose equilibrium distribution is $\pi(x)$. If we then run the chain for a long time, simulated values of the chain can be used as a basis for summarizing features of $\pi(x)$ of interest. To implement this strategy, we simply need algorithms for constructing chains with specified equilibrium distributions.

For a rigorous mathematical discussion this idea has to be set in an appropriate theoretical framework. Our presentation will mainly use the elementary language and concepts corresponding to discrete state spaces (but with a brief account of some general theory in Appendix A). For more extensive theoretical accounts, the reader is referred to Roberts and Smith (1992), Besag and Green (1993) and the recent overview by Tierney (1991).

Under suitable regularity conditions, asymptotic results exist which clarify the sense in which the sample output from a chain with equilibrium distribution $\pi(x)$ can be used to mimic a random sample from $\pi(x)$ or to estimate the expected value, with respect to $\pi(x)$, of a function $f(x)$ of interest.

If $X^1, X^2, \ldots, X^t, \ldots$ is a realization from an appropriate chain, typically available asymptotic results include

$$X^t \xrightarrow[t\to\infty]{\mathrm{d}} X \sim \pi(x); \quad \frac{1}{t}\sum_{i=1}^{t} f(X^i) \xrightarrow[t\to\infty]{} E_\pi\{f(X)\} \quad \text{almost surely.}$$

Clearly, successive X^t will be correlated, so that, if the first of these asymptotic results is to be exploited to mimic a random sample from $\pi(x)$, suitable spacings will be required between realizations used to form the sample, or parallel independent runs of the chain might be considered. The second of the asymptotic results implies that ergodic averaging of a function of interest over realizations from a single run of the chain provides a consistent estimator of its expectation.

We now consider some particular forms of Markov chain schemes, beginning with one which has proved particularly convenient for a range of applications in Bayesian statistics.

3.2. *Gibbs Sampler*

Let $\pi(x)=\pi(x_1, \ldots, x_k)$, $x\in\mathbf{R}^n$, denote a joint density, and let $\pi(x_i|x_{-i})$ denote the induced full conditional densities for each of the components x_i, given values of the other components $x_{-i}=(x_j; j\neq i)$, $i=1, \ldots, k$, $1<k\leqslant n$.

A systematic form of the so-called Gibbs sampler algorithm (Geman and Geman, 1984) proceeds as follows. First, pick arbitrary starting values $x^0=(x_1^0, \ldots, x_k^0)$. Then successively make random drawings from the full conditional distributions $\pi(x_i|x_{-i})$, $i=1, \ldots, k$, as follows:

$$
\begin{aligned}
&x_1^1 \text{ from } \pi(x_1|x^0{}_{-1});\\
&x_2^1 \text{ from } \pi(x_2|x_1^1, x_3^0, \ldots, x_k^0);\\
&x_3^1 \text{ from } \pi(x_3|x_1^1, x_2^1, x_4^0, \ldots, x_k^0);\\
&\quad\vdots\\
&x_k^1 \text{ from } \pi(x_k|x^1{}_{-k}).
\end{aligned}
$$

This completes a transition from $x^0=(x_1^0, \ldots, x_k^0)$ to $x^1=(x_1^1, \ldots, x_k^1)$. Iteration of this cycle of random variate generation from each of the full conditional distributions in turn produces a sequence $x^0, x^1, \ldots, x^t, \ldots$ which is a realization of a Markov chain, with transition probability from x^t to x^{t+1} given by

$$K_G(x^t, x^{t+1})=\prod_{l=1}^{k}\pi(x_l^{t+1}|x_j^t, j>l, x_j^{t+1}, j<l).$$

The key feature of this algorithm is that we only sample from the full conditional distributions $\pi(x_i|x_{-i})$.

In many cases, it may be natural to work with a complete breakdown of x into all its scalar components ($k=n$). In other cases, the components could be subvectors or matrices. One important consideration in choosing the level at which the components for the conditionals are chosen is the correlation structure of $\pi(x)$. If highly correlated scalar components are treated individually, there could be painfully slow convergence of the chain to equilibrium as a result of very little movement at each conditional random variate generation step. If, however, correlated scalars are blocked together to form a subvector component this problem is avoided, but at the expense of having to perform a draw from a multivariate conditional distribution.

Another algorithm which involves successive drawing from various (not just full) conditionals is the substitution sampler, which is discussed in detail in Gelfand and Smith (1990), who examine its relation to the Gibbs sampler and to the work of Tanner

and Wong (1987). Developments and applications are described at length in Tanner (1991) and the reader is referred to these references for further details.

3.3. *Metropolis–Hastings Algorithm*

To construct a Markov chain $X^1, X^2, \ldots, X^t, \ldots$ with state space $\mathscr{X}$ and equilibrium distribution $\pi(x)$, the Metropolis–Hastings algorithm constructs the transition probability from $X^t = x$ to the next realized state X^{t+1} as follows. Let $q(x, x')$ denote a (for the moment arbitrary) transition probability function, such that, if $X^t = x$, x' drawn from $q(x, x')$ is considered as a proposed possible value for X^{t+1}. However, a further randomization now takes place. With some probability $\alpha(x, x')$, we actually accept $X^{t+1} = x'$; otherwise, we reject the value generated from $q(x, x')$ and set $X^{t+1} = x$. This construction defines a Markov chain with transition probabilities given by

$$p(x, x') = \begin{cases} q(x, x')\,\alpha(x, x') & \text{if } x' \neq x, \\ 1 - \sum_{x''} q(x, x'')\,\alpha(x, x'') & \text{if } x' = x. \end{cases}$$

If now we set

$$\alpha(x, x') = \begin{cases} \min\left\{\dfrac{\pi(x')\,q(x', x)}{\pi(x)\,q(x, x')}, 1\right\} & \text{if } \pi(x)\,q(x, x') > 0, \\ 1 & \text{if } \pi(x)\,q(x, x') = 0, \end{cases}$$

it is easy to check that $\pi(x)\,p(x, x') = \pi(x')\,p(x', x)$, which, provided that the thus far arbitrary $q(x, x')$ is chosen to be irreducible and aperiodic on a suitable state space, is a sufficient condition for $\pi(x)$ to be the equilibrium distribution of the constructed chain (see Appendix A).

This general algorithm is due to Hastings (1970); see, also, Peskun (1973). It is important to note that the (equilibrium) distribution of interest, $\pi(x)$, only enters $p(x, x')$ through the ratio $\pi(x')/\pi(x)$. In the context where $\pi(x)$ is a Bayesian posterior distribution this is quite crucial since it means that knowledge of the distribution up to proportionality (given by the likelihood multiplied by the prior) is sufficient for implementation.

Clearly, different specific choices of $q(x, x')$ will lead to different specific algorithms. Tierney (1991) provides a systematic taxonomy of the kinds of choice available. Among them, we note the following. If $q(x, x') = q(x', x)$, we have $\alpha(x, x') = \min\{\pi(x')/\pi(x), 1\}$, which is the well-known Metropolis algorithm (Metropolis *et al.*, 1953). If $q(x, x') = q(x' - x)$, the chain is driven by a random walk process, with multivariate normal and multivariate Student t- or split t-forms possible candidates for $q(x' - x)$; see Muller (1991). If $q(x, x') = q(x')$, we have $\alpha(x, x') = \min\{w(x')/w(x), 1\}$, with $\alpha(x, x')$ defined via the importance weights $w(x) = \pi(x)/q(x)$. Insight from general importance sampling methodology suggests that Student t- or split Student t-forms with small degrees of freedom are likely to be good candidates. Finally, we note that a range of hybrid strategies can be created by combining different chains in various ways, e.g. by cycling, mixing or using an MCMC method within the conditioning of the Gibbs sampler. See Tierney (1991) and Section 7 of this paper.

Our subsequent focus will be as follows. In Sections 5 and 6, we shall illustrate the ways

in which the Gibbs sampling algorithm provides a powerful general approach to Bayesian calculations for a range of problems. In Section 7, we shall briefly discuss applications of some other MCMC ideas in Bayesian statistics, noting that a more wide ranging account is given by Besag and Green (1992). First, we give a brief overview of some general implementation issues and then, in Section 4, discuss some of the general advantages of sample-based approaches to Bayesian calculation.

3.4. *General Implementation Issues*

Several different questions arise in implementing MCMC methods. Does a particular constructed Markov chain have a mathematical structure for which the desired asymptotic results hold? Are some forms of construction preferable to others in terms of their speed of convergence or the relative ease of the required random variate generation? Can one prespecify how long a chain should be run for, or how it should be monitored to decide whether it has been run for sufficiently long? Specific discussion of 'convergence' issues is given in the next section and in Appendix A. Here, we comment on some of the other pragmatic aspects of implementation.

Most publications relating to monitoring the realization of a chain, so-called 'output analysis' (see, for example, Ripley (1987)), reflect the fact that the objective in many applications of MCMC methods to physics or operational research problems is point estimation of a single unknown quantity, based on an ergodic average from a single long realization of the chain.

However, applications in Bayesian (or likelihood) statistics typically have the somewhat more amorphous objective of exploring complicated high dimensional uncertainty surfaces with a subsequent view to estimating or summarizing several different, perhaps as yet unarticulated, features. For such exploratory statistical analysis, a random sample from the distribution would clearly be preferable to a prespecified collection of point estimates.

As we remarked earlier, one way of mimicking random samples from the equilibrium distribution is to perform long runs of each of a number of independent chains in parallel, forming a sample by collecting the final states from each. An alternative is to perform one run of a single chain, ignoring an initial transient phase, and then forming a sample by collecting equally spaced outcomes, the gaps being chosen to render serial correlation negligible. In theory, the single-chain approach appears to be more efficient, in that only one transient phase is involved. However, particularly during the first tentative examination of a new problem, it can be argued that monitoring the evolutionary behaviour of several runs of the chain starting from a wide range of initial values is necessary (see Gelman and Rubin (1992)).

In either case, a key problem is to decide how long the chain should be run for, and whether this can be done in advance or needs to be determined by some kind of sequential stopping rule. Examining several parallel runs, or successive batches within a single run, can certainly provide (negative) evidence that a run is not sufficiently long. However, there can never be any (positive) empirical guarantee that a sufficiently long run has been taken. Rather, the issue is identifying implementation schedules which imply suitably high probabilities that the objectives of the simulation study (however defined) will be met.

Even then, it has to be frankly acknowledged that, however hard you try, someone can always invent a problem for which your intended (finite resource) schedule will

have a high probability of failure. Consider, for example, a two-dimensional surface which consists of a nearly infinitely high near zero volume spike sitting in the otherwise nearly flat plane. Clearly, the chances of discovering the form of this surface by Gibbs sampling could be rather small. Similarly, imagine a surface formed by a mixture of two bivariate normals with locations, along a bisector of the component axes, very widely separated relative to the spreads. Clearly, a naïvely designed chain could fall into long cycles of being trapped at one or other of the modes, with low, but non-zero, probability of moving to the other. Practical convergence is unlikely, even though theoretically the chain should converge. Worse still, the vagaries of computer arithmetic could result in rounding to 0, effectively disconnecting the two humps from each other and potentially violating a necessary condition for convergence.

For any given amount of sampling effort, variance reduction ideas from classical Monte Carlo simulation are still relevant. In particular, techniques of importance sampling, conditioning, and antithetic and control variates find a role within the Markov chain methodology. See, for example, Tierney (1991) and Green and Han (1992) for further discussion and illustration.

In addition, the cost of simulating is partly determined by the efficiency of random variate generation. Here, we can draw on a wealth of existing methodology (see, for example, Devroye (1986)), but we also note that the increased use of the MCMC method is itself stimulating further development and refinement of random variate generation techniques (see Carlin and Gelfand (1992), Gilks and Wild (1992) and Wakefield *et al.* (1991)). A related issue is that of reparameterization, either to break high correlations or to facilitate random variate generation (although these two aims may sometimes conflict); see Muller (1991) and Hills and Smith (1992).

3.5. *Specific Convergence Issues*

Although there is a reassuring theoretical literature concerning the convergence of MCMC methods (see, for example, Appendix A, Tierney (1991) and Besag and Green (1992)), results do not easily translate into clear guidelines for the practitioner. Theory does not, except in some important special cases, provide useful bounds on rates of convergence and some kind of pragmatic output analysis to estimate the length of the transient phase is inevitable. However, such output analysis necessarily reduces to monitoring output summaries for a relatively small set of features of the distribution, thus carrying with it the danger of overlooking some aspects of the multidimensional behaviour of the chain.

Before discussing approaches to output analysis, we first note various special cases where useful analytical bounds on rates of convergence can be found, so that estimates of the length of the transient phase can be made *a priori*. For Gaussian distributions, results are available on the relationship between the correlation structure of the target distribution and the convergence rate of the Gibbs sampler; see, for example, Amit and Grenander (1991), Barone and Frigessi (1989) and Amit (1991), who extends these results to bounded multiplicative perturbations of Gaussian densities. Here, theory adds weight to the heuristic discussion at the end of Section 3.2. Beyond the Gaussian case, Applegate *et al.* (1990) find, for grid-based Metropolis and Gibbs algorithms, workable bounds for rates of convergence in terms of a parameter measuring log-convexity, and a log-Lipschitz coefficient of the target distribution, and Rosenthal (1992) gives explicit bounds for a Bayesian variance component model.

We turn now to a summary discussion of various methods of output analysis ('convergence diagnostics') that have been proposed for statistical applications of MCMC methods. The most common approach seems to consist of monitoring ergodic averages of selected scalar quantities (e.g. first and second moments), for stationarity. Gelman and Rubin (1992) suggest augmenting this by carrying out multiple runs from dispersed starting points and checking both within- and between-output series variation. A useful adjunct is to monitor also a few arbitrary functions of the parameters of interest (e.g. linear combinations), but these methods necessarily suffer from the deficiency of possibly overlooking lack of convergence of some aspect of the distribution. Raftery and Lewis (1992), in the context of estimating a single marginal posterior distribution quantile to prespecified accuracy, propose a two-state Markov chain model fitting procedure based on pilot analysis of output from the original chain. However, it is not clear how to extend this approach if many posterior probabilities are to be simultaneously estimated. To measure convergence of the whole distribution, Roberts (1992) provides theory for estimating an L^2-distance between the target distribution and the chain distribution at any time. Details of practical implementation are given in Roberts and Hills (1992), who also provide a comparative study of various forms of output analysis.

4. BAYESIAN INFERENCE VIA SAMPLE-BASED METHODS

In this section, we discuss various ways in which the final output from an MCMC simulation might be used as the basis for inference reporting and diagnostics in Bayesian statistics. For concreteness, suppose that the equilibrium distribution corresponds to a posterior density $\pi(\theta) = \pi(\theta \mid y) \propto p(y \mid \theta)\, p(\theta)$, where $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$, with $1 < k \leqslant n$.

4.1. *Graphics and Exploratory Data Analysis*

Suppose that $\theta^1, \theta^2, \ldots, \theta^N$ is a random sample from $\pi(\theta)$. The agenda for exploring and summarizing features of $\pi(\theta)$ of interest might include

(a) examining the shapes of univariate marginal distributions for the individual θ_i, or for functions of θ_i,
(b) producing marginal moment or quantile summaries,
(c) examining bivariate marginals for pairs θ_i, θ_j, or pairs of functions of θ,
(d) examining trivariate distributions,
(e) checking for interesting special features and trying to obtain a global overview of the whole posterior and
(f) uses of the output for specific decision problems or predictive analysis.

It is striking that much of this agenda of exploration and summarization is very much the same as presents itself when one is interested in exploratory data analysis (EDA) of a point cloud of multivariate observations. The individual observation vectors are here replaced by the individual parameter vectors drawn from the posterior; the number of observations (cases) corresponds to the size of the sample drawn from the posterior. It follows that the exploratory, particularly graphical, toolkit developed by the multivariate EDA and visual EDA communities over the past two decades now finds an additional—and perhaps unexpected—role as an essential part of the Bayesian computational toolkit.

A further aspect of dynamic graphics which will undoubtedly attract increasing attention is the use of 'smooth animation' to explore summaries of posterior surfaces (see Tierney (1991)). The visual image is some functional form, e.g. a curve whose coefficients are model parameters, and successive parameter drawings from the posterior then generate a moving image. Perceived changes and stabilities in the image can then be related to different regions in the parameter space and these in turn related to high and low posterior probabilities. An interesting feature of this approach is that using successive correlated values from the chain is a positive advantage in producing smooth animation.

4.2. *Inference and Prediction*

As far as specific EDA and graphics-related tools are concerned, kernel density estimation finds a role in converting samples into posterior density curves; scatterplots and scatterplot matrices can provide a basic overview of pairwise aspects of the posterior, forming the basis of contour plots if required; point cloud rotation is available as a tool for exploring trivariate posterior marginals; grand tours may be undertaken, and projection pursuit used to search for interesting features.

That said, the ready computational availability of smoothing and graphics tools should not lead one to neglect to exploit any structural mathematical insights that might be available. As an example, consider the problem of producing the marginal density curve for a component parameter θ_i. This could be produced directly from the sample values $\theta_i^1, \ldots, \theta_i^N$, e.g. by using kernel smoothing. However, if we know the form of the conditional $\pi(\theta_i | \theta_{-i})$, the marginal could also be calculated pointwise by averaging the conditional density over the sample values of $\theta_{-i} = (\theta_j, j \neq i)$. The conditional procedure, akin to 'Rao–Blackwellization', might be more efficient (Gelfand and Smith, 1990, 1991).

For a predictive distribution—e.g. for some as yet unobserved data y' generated from the same model as $p(y|\theta)$—sample-based calculation is particularly straightforward since

$$\frac{1}{N}\sum_{i=1}^{N} p(y'|\theta^i) \approx p(y'|y) = \int p(y'|\theta)\,\pi(\theta|y)\,\mathrm{d}\theta$$

provides a direct pointwise estimate of the predictive density curve $p(y'|y)$.

4.3. *Diagnostics and Model Validation*

Because it invokes more modelling structure than other approaches do, Bayesian statistics should, if anything, have a more extensive toolkit for sensitivity analysis and diagnostics. However, viewed from the perspective of non-sample-based approaches to Bayesian computation, systematic reanalysis with varying assumptions has been seen as requiring enormous computational effort and has consequently not assumed its rightful place on the Bayesian agenda, in marked contrast with the development of a plethora of influence and other diagnostics in non-Bayesian approaches.

The problem of Bayesian sensitivity analysis is to investigate how inferences or predictions would change if we were to make perturbations to the assumed forms of the likelihood and/or prior specification. Among forms of perturbation of interest we note

(a) changes to the likelihood which take the form of omitting a subset of observations to examine their influence within the context of the assumed model forms
(b) changes to the functional form of the likelihood itself and
(c) changes to the functional form of the prior.

In all cases, the perturbation will result in a changed posterior form. For a general analysis, let us call $\pi(\theta)$ the posterior density resulting from an initial formulation $p(y|\theta)$, $p(\theta)$ and $\pi'(\theta)$ the posterior density resulting from a perturbed formulation $p'(y|\theta)$, $p'(\theta)$. To focus ideas, suppose that the inference of interest is an estimate of the posterior mean of the scalar function $g(\theta)$. With respect to the perturbed formulation, we wish to estimate the integral $\int g(\theta)\,\pi'(\theta)\,d\theta$. But, using standard importance sampling ideas, it is clear that

$$\frac{1}{N}\sum_{i=1}^{N}\frac{g(\theta^i)\,\pi'(\theta^i)}{\pi(\theta^i)} \approx \int \frac{g(\theta)\,\pi'(\theta)}{\pi(\theta)}\,\pi(\theta)\,d\theta$$

provides an estimate based on the sample from the posterior in the original formulation, a calculation which only requires the importance weights. If, more generally, we want to create a random sample from $\pi'(\theta)$, given a random sample $\theta^1, \ldots, \theta^N$ from $\pi(\theta)$, we might exploit standard random variate reject–accept techniques, or perhaps employ a form of weighted bootstrap, to resample from $\theta^1, \ldots, \theta^N$ to mimic a random sample from $\pi'(\theta)$ (see Smith and Gelfand (1992)).

The crucial point in all this is that, unlike with non-sample-based approaches to Bayesian calculation, we do not have to restart the whole computation from scratch for each variation in the likelihood–prior specification. Instead, reanalysis proceeds on the basis of reweighting or resampling output from the original analysis. In line with the theory of standard importance sampling, the viability and efficiency of this strategy will depend on sensitivity analysis being conducted in a local neighbourhood of the original model. See Gelfand, Dey and Chang (1992) for some detailed proposals.

For a more global assessment of the adequacy of a Bayesian model, a possible sample-based approach is the following. Given data y and the assumption that the currently entertained model is correct, our previous discussion of predictive calculations makes clear how to simulate new data which is 'as if' from the model, taking into account uncertainties about unknown model parameters in the light of $\pi(\theta|y)$. For a variety of data summaries of interest, the observed summaries from the actual data y can be calibrated against the corresponding predictive distribution. If too many of these lie in the tails of the respective distributions, we may be led to doubt the adequacy of the model. This approach to model criticism relates to the ideas of Box (1980) and Rubin (1984) but has rarely been seriously pursued becasue of computational difficulties. There are many issues to be resolved here, including those relating to multiple testing and the choice of and numbers of test summaries. What we wish to emphasize in this discussion is that sample-based methods free us to explore this and related ideas by removing previously perceived computational constraints.

5. GIBBS SAMPLER FOR BAYESIAN CALCULATIONS

We now shall assume that the equilibrium distribution of interest is a Bayesian posterior distribution for unknown parameters $\theta = (\theta_1, \ldots, \theta_k)$ having a density

$\pi(\theta) = \pi(\theta|y)$, defined up to proportionality by the product of a likelihood $p(y|\theta)$ and a prior density $p(\theta)$. The Gibbs sampler requires successive generation from the conditional forms $\pi(\theta_i|\theta_{-i})$, the latter immediately identifiable, up to proportionality, by simply regarding $\pi(\theta)$ as a function of θ_i only, considering θ_j, $j \neq i$, to be fixed.

We now present a range of problem types, chosen to illustrate various ways in which seemingly awkward or intractable Bayesian calculation problems become very much more straightforward when approached via the Gibbs sampler. For simplicity, our illustrations throughout are in terms of parametric model formulations. For illustration of uses of the Gibbs sampler in Bayesian nonparametric settings, the reader is referred to Escobar (1991), Gelfand and Kuo (1991) and Kuo and Smith (1992).

5.1. *Constrained Parameter Models*

Suppose that a posterior density $\pi(\theta)$ for $\theta = (\theta_1, \ldots, \theta_k)$ is constrained to have support $S^k \subset \mathbf{R}^n$. This can pose extremely awkward computational problems. As a result, parameter constraints which should be acknowledged in honest modelling are often ignored.

Consider, however, the full conditional forms required for implementing the Gibbs sampler. For any θ_i, the constraint derived from S^k given specified values of θ_{-i} is a cross-section of the form $S_i^k(\theta_{-i})$. In particular, if θ_i is a scalar this cross-section is typically an interval (at worst perhaps a union of intervals). The task of random variate generation required to implement the sampler thus reduces to generation from specified univariate shapes, truncated to intervals. This is clearly far more straightforward than the problem of directly approaching the evaluation of high dimensional integrals over complicated constraint volumes.

Ordered parameter problems are important special cases. If we have $\theta_i \in \mathbf{R}$, $i = 1, \ldots, n$, with $\theta_1 < \theta_2 < \ldots < \theta_n$, the cross-sectional constraint for θ_i given specified values of θ_{-i} is simply $\theta_{i-1} < \theta_i < \theta_{i+1}$. Thus, the full conditional for θ_i is typically just the posterior form for θ_i that would have been obtained from an unconstrained analysis, but now truncated to the interval $(\theta_{i-1}, \theta_{i+1})$.

There are obvious extensions. For example, parameters may be constrained to increase and then to decrease (or vice versa), so that $\theta_1 < \theta_2 < \ldots < \theta_r$, $\theta_r > \theta_{r+1} > \ldots > \theta_n$, where r may be known or unknown. Again, full conditionals for θ_i, $i = 1, \ldots, n$, will simply be truncated forms. As a somewhat more complicated extension, suppose that $\theta_1, \ldots, \theta_n$ represent 'responses' to a 'stimulus' at levels $s_1, \ldots, s_n$, and that the 'response curve' is known to be monotonic increasing and concave. This information implies that the parameters must obey the constraints

$$0 < \frac{\theta_{i+1} - \theta_i}{s_{i+1} - s_i} < \frac{\theta_i - \theta_{i-1}}{s_i - s_{i-1}}.$$

Once again, it is easy to see that the full conditionals will correspond to posterior forms that would have obtained in the unconstrained case, but now truncated to intervals whose end points are more complicated, but still easily identified, functions of the specified conditioning parameters.

The striking simplicity of the Gibbs sampler structure in these cases opens the way to straightforward routine Bayesian analysis in a range of important applications. Examples include models of reliability growth, categorical data and analysis-of-variance

models with ordinal factors, bioassay and survival modelling, and multiple changepoint models. See, for example, Gelfand, Smith and Lee (1992), Ramgopal *et al.* (1992) and Stephens (1991).

The case of multiple-changepoint models is particularly striking. A sequence of random variables $Y=(Y_1, \ldots, Y_m)$ is said to have k changepoints at $r_1, \ldots, r_k$ if, with $r_0=1$, $r_{k+1}=m$,

$$Y_{r_{i-1}+1}, \ldots, Y_{r_i} \sim p_i(y|\theta_i), \qquad i=1, k+1,$$

where, typically, the parametric model forms $p_i(y|\theta_i)$ are unchanged, but $\theta_1 \neq \theta_2 \neq \ldots \neq \theta_{k+1}$. With $\theta_1, \ldots, \theta_{k+1}$ and $r_1, \ldots, r_k$ unknown, the marginal (joint) posterior for the changepoints is often unavailable in closed form and, in any case, the awkward support of the distribution (all feasible k-tuples $r_1<r_2<\ldots<r_k$) makes direct calculation or simulation difficult. But, if we now consider all $\theta=(\theta_1, \ldots, \theta_{k+1})$ and $r=(r_1, \ldots, r_k)$ as unknowns and examine the full posterior conditionals of each of these individually, conditional on the data and on all the other unknowns, a dramatic simplification occurs. For example, with independent priors for $\theta_1, \ldots, \theta_{k+1}$, it is easy to see that

$$\pi(\theta_i|\theta_{-i}) = \pi(\theta_i|y_{r_{i-1}+1}, \ldots, y_{r_i}),$$
$$\pi(r_i|r_{-i}) = \pi(r_i|r_{i-1}, r_{i+1}, \theta, y).$$

The form for θ_i reduces, given θ_{-i} and r, to a straightforward 'non-changepoint' posterior. The form for r_i reduces to truncated inference in a single changepoint setting, with model sampling distribution parameters known. The required random variate generation is typically straightforward to implement (see Stephens (1991)).

5.2. *Hierarchical Models*

From a Bayesian modelling perspective, a very large class of problems treated by other statistical approaches under headings such as empirical Bayes models, random effects models or population models can be subsumed within a hierarchical framework, which specifies the model through successive conditioning of the form $p(y|w_1)$, $p(w_1|w_2), \ldots, p(w_{k-1}|w_k)$.

In many applications, $k=3$ and the three stages of the model are interpreted as follows: y denotes a collection of separate data vectors from a number of individual units (e.g. patients, geographical areas, . . .), the first-stage parameter w_1 is a collection of individual parameters, each defining an underlying response characteristic or profile of some kind for the individual units (e.g. drug concentration *versus* time after administration, mortality rate for a disease, . . .) and the second-stage parameter w_2 indexes a distribution which models the relatedness of the individual unit characteristics encapsulated in the first-stage parameters (e.g. similarity due to age and treatment regime, spatial correlation, . . .). Interest may focus on inference or prediction for individuals, via w_1, or on population features and variation via w_2. But in all applications the common feature is that the integrals required for a fully Bayesian analysis cannot be obtained in closed form.

Now consider what happens if we apply the Gibbs sampler to the general hierarchical form, taking the w_i as the components. Because of the hierarchical structure, we see immediately that the required full conditionals simplify considerably to give (omitting terms w_0 and w_{k+1})

$$\pi(w_i | w_{-i}, y) = \pi(w_i | w_{i-1}, w_{i+1}, y), \qquad i = 1, \ldots, k.$$

The specific distributional forms which arise vary from application to application, but for a very extensive range of commonly used models it has now been shown how to deal efficiently with the random variate generation problems posed by the forms of the full conditionals: see Gelfand *et al.* (1990), George *et al.* (1991) for exponential family random effects models and Wakefield *et al.* (1992) for linear and non-linear regression models, including non-normal and mean–variance relationship modelling.

A particular structure arising in Wakefield *et al.* (1992) which deserves separate mention concerns the modelling of heavy-tailed distributions as robust replacements for conventional normal assumptions. At first sight, it would seem that abandoning normal assumptions would lead to a considerable increase in computational difficulties. Consider, however, the general class of distributions obtained by forming scale mixtures of normals. Now just add the further scale parameter to the list of unknowns in the model and consider the conditional forms arising in the Gibbs sampler. All those which condition on the new scale parameter retain the same basic distributional forms as under the assumption of normality, so that the only further complication is the additional full conditional for the scale parameter. In many cases, this takes the form of a familiar distribution; even if not, simulation from one further univariate distribution is no great additional burden.

5.3. *Generalized Linear Models*

For most exponential family and link function combinations, the posterior density forms for the regression parameters β in generalized linear models are distinctly intractable and previously Bayesian analysis has depended on quite subtle uses of adaptive quadrature techniques, usually requiring sophisticated insight into suitable model reparameterization.

At first sight, it does not seem that the Gibbs sampler is particularly helpful. The full conditionals for the regression parameters are unfamiliar density forms, perhaps each requiring sophisticated tuning of a special purpose random variate generation algorithm. However, this is not so. Salvation lies in the well-known fact that, at least for canonical link parameters, the likelihood is log-concave, so that given a constant or log-concave prior the posterior is log-concave, resulting in log-concave full conditionals. Log-concavity of the likelihood can also be established for many specific non-canonical link functions for generalized linear models. Moreover, the log-concavity property can also be established for many commonly used parametric proportional hazards models; see Dellaportas and Smith (1993). The significance of this is that an extremely efficient random variate generation algorithm is then generally available for all these cases (see Gilks and Wild (1992)), so that Bayesian implementation becomes straightforward.

5.4. *Time Series Models*

The Bayesian literature on time series analysis is substantial. However, even for standard autoregressive moving average (ARMA) models proper computation of posteriors has not been the norm. Instead, conditional likelihoods ignoring stationarity and invertibility restrictions are typically employed, although with numerical integration and judicious reparameterization proper posteriors can be calculated in some low order cases (see, for example, Marriott and Smith (1992)).

However, using the Gibbs sampler Marriott *et al.* (1992) have succeeded in implementing a fully Bayesian analysis for general ARMA(p, q) models. Features of the Gibbs sampler structuring in this context include

(a) a convenient readily evaluated factorization of the full likelihood incorporating latent variables,
(b) a two-stage reparameterization in terms of partial autocorrelations which provides a one-to-one transformation of the constrained ARMA coefficients to R^{p+q} and
(c) efficient sampling of full conditional distributions.

In addition, the approach permits routine handling of typical, but awkward, time series problems such as missing data and outliers.

6. GIBBS SAMPLER AND INCOMPLETE DATA PROBLEMS

6.1. *Overview*

Among the incomplete data problems that we shall consider are those corresponding to the following: missing data, censoring and finite mixtures. What all these problems have in common from a Bayesian perspective is that the (often multiparameter) likelihoods which arise are extremely challenging computationally if approached via direct approximation of required integrals. Consider, for example, likelihoods derived from censoring. These will contain terms involving cumulative distribution functions, which are typically not available in closed form. Not only are direct evaluations of the (joint) posterior therefore expensive to compute, but also the awkwardness of the mathematical form usually denies the insight required for analytic approximation, or design of efficient importance sampling functions for direct Monte Carlo simulation.

But, at first sight, these complications also seem to preclude any effective simplification from using the Gibbs sampler. Since the required posterior full conditionals derive directly from the form of the joint posterior, they will also be expensive to compute and awkward to sample from. The solution, which we shall outline more fully in the following subsections, is to reintroduce the missing data as further unknowns, in addition to the unknown model parameters. It transpires, for a wide range of 'incomplete data' problems, that the resulting structure of the full conditionals for the 'augmented unknowns' problem is very simple, leading to a straightforward implementation of the Gibbs sampler. See Gelfand, Smith and Lee (1992) for a fuller account and Tanner (1991) for extensive use of related ideas.

6.2. *Missing Data*

We shall write $z=(y, y')$ to denote that z are the 'intended' data (e.g. from a structured designed experiment), but that only y is observed, so that y' is missing. The actual posterior for the unknown model parameters θ, given by $\pi(\theta|y)$, will typically be messy, whereas the intended posterior, $\pi(\theta|z)$, will typically be a tractable form.

Following the augmented unknowns prescription, we consider the unknowns to consist of (θ, y'). Treating, for a moment, θ and y' as the basic components for the Gibbs sampler, the required full conditionals have the forms

$$\pi(\theta|y', y)=\pi(\theta|z),$$
$$\pi(y'|\theta, y)=\pi(y'|\theta).$$

But now we notice something rather striking. The first of these is the 'typically tractable' form that would have arisen if we had not had any missing data. The second is simply the sampling distribution, under the model, of y' given θ. In most applications, neither of these will present problems for the random variate generation required in implementing the Gibbs sampler. As an aside, we note that generation from $\pi(\theta|z)$ might either proceed directly (using an appropriate importance sampling family), or by partitioning θ into subcomponents $\theta_1, \ldots, \theta_k$ and replacing $\pi(\theta|z)$ by $\pi(\theta_i|\theta_{-i}, z)$, $i=1, \ldots, k$, or by using an alternative form of MCMC method.

6.3. *Censored Data*

For concreteness, assume that $y=(y_1, \ldots, y_s)$ are exactly observed, but that the remaining data are subject to a censoring mechanism with outcomes $V_j \leqslant y_j \leqslant W_j$, $j=s+1, \ldots, m$, implying, under an assumed parametric model, a joint posterior of the form

$$\pi(\theta|V, W, y) \propto \prod_{i=1}^{s} p(y_i|\theta) \prod_{j=s+1}^{m} \int_{V_j}^{W_j} p(y_j|\theta)\, \mathrm{d}y_j\, p(\theta),$$

with $V=(V_{s+1}, \ldots, V_m)$ and $W=(W_{s+1}, \ldots, W_m)$.

Instead of regarding θ as the unknowns and deriving full conditionals from this joint posterior, we define as further unknowns $y'=(y_{s+1}, \ldots, y_m)$ the observations that we did not obtain exactly because of censoring. With θ and y' together constituting the augmented unknowns, the corresponding full conditionals have the forms

$$\pi(\theta|V, W, y', y) = \pi(\theta|y', y),$$

$$\pi(y'|V, W, \theta, y) = p(y'|V, W, \theta) = \prod_{j=s+1}^{m} \int_{V_j}^{W_j} p(y_j|\theta)\, \mathrm{d}y_j.$$

Again, we notice something striking. The first of these forms is just the joint posterior for θ that we would have obtained had there been no censoring. The second is just the joint distribution of the censored observations, given θ. Random variate generation from the former will typically be straightforward; generation from the latter is trivial and reduces to independent draws from the truncated model sampling distribution.

6.4. *Finite Mixtures*

If θ denotes the totality of distinct parameters in $\pi=(\pi_1, \ldots, \pi_k)$ and $(\phi_1, \ldots, \phi_k)$, the posterior for θ given a sample $y=(y_1, \ldots, y_m)$ from $p(y_i|\theta)=\pi_1 p_1(y_i|\phi_1)+\ldots+\pi_k p_k(y_i|\phi_k)$ has the form

$$\pi(\theta|y) \propto \prod_{i=1}^{m} \left\{ \sum_{j=1}^{k} \pi_j p_j(y_i|\phi_j) \right\} p(\theta),$$

which, as detailed, for example, by Titterington *et al.* (1986), typically has many unpleasant features.

Suppose, however, that we append to each observation y_i the missing label vector $z_i=(z_{i1}, \ldots, z_{ik})$, which would have identified the component of the mixture from which y_i were actually generated (so that, if y_i is from component j, $z_{ij}=1$ with $z_{ij'}=0$ for $j' \neq j$). If we now regard the $z_1, \ldots, z_m$ as further unknowns, in addition

to θ, it is easy to see that the conditional model components greatly simplify so that, for example,

$$p(y_i|z_i, \phi) = \prod_{j=1}^{k} p_j^{z_{ij}}(y_i|\phi_j),$$

$$p(z_i|\pi) = \prod_{j=1}^{k} \pi_j^{z_{ij}}.$$

Thus, with, say, a Dirichlet prior for $\pi = (\pi_1, \ldots, \pi_k)$ and independent priors for $\phi = (\phi_1, \ldots, \phi_k)$ and π, the resulting structural forms of full conditionals are easily seen to be

(a) for π given ϕ, z, a Dirichlet,
(b) for ϕ_j given ϕ_{-j}, π, z, the appropriate posterior corresponding to updating the prior based on just the y_i assigned (under the assumed values of z) to component j and
(c) for z given ϕ, π, a discrete distribution, whose form, like that of the posterior for ϕ_i, is well defined, and easily identified in terms of $p_j(y_i|\phi_j)$, $j = 1, \ldots, k$.

7. USES OF METROPOLIS–HASTINGS ALGORITHM

We briefly outline two settings which illustrate the potential of versions of the Metropolis–Hastings algorithm as a hybrid adjunct to a Gibbs sampling structure.

7.1. *Gene Mapping*

The problem of ordering and mapping genes which are known to belong to the same chromosome is of considerable current interest. One approach is to base inferences on data derived from observing possible recombinants, with recombinant fractions assumed to be monotone functions of the distances between the loci of genes on the chromosome. As a simplified version of the model and inference problems posed, consider the following. An interval contains k labelled locations, corresponding to the true locations of k specified genes. Thus, for example, with $k = 4$ and the genes labelled by A, B, C and D the actual order might be C A B D, with actual interloci distances given by δ_1, δ_2 and δ_3.

An observation process proceeds by labelling the genes 1, 2, . . ., k, leading to the specification of a likelihood (based on observed recombinants) in terms of distances between the i, j labelled pairs. However, the relation of these to the actual unknown $(\delta_1, \ldots, \delta_{k-1})$ depends on the unknown correspondence between $(1, \ldots, k)$ and $(A, B, \ldots)$. As a consequence, the form of the direct likelihood for the $(\delta_1, \ldots, \delta_{k-1})$ is extremely complicated since it implicitly derives from mixing over all the possible permutations linking $(1, \ldots, k)$ to $(A, B, \ldots)$. With, for example, a prior which assumes uniform order statistics for the locations, the posterior for $(\delta_1, \ldots, \delta_{k-1})$ inherits the nightmare qualities of the likelihood.

Suppose, however, that we introduce the unknown permutation explicitly into the model. Denoting the permutation by σ and the vector of unknown actual interloci distances by δ, the posterior of interest is now $\pi(\sigma, \delta|y)$, where y denotes the

totality of data. Intuitively, it is clear that inferences for δ would be straightforward if σ were known; similarly, knowledge of the δ leads directly to a well-defined discrete distribution over the set of all permutations. This suggests immediately the use of the Gibbs sampler, based on the conditionals $\pi(\delta|\sigma, y)$ and $\pi(\sigma|\delta, y)$, perhaps with $\pi(\delta|\sigma, y)$ replaced by the sequence of componentwise conditionals $\pi(\delta_i|\delta_{-i}, \sigma, y)$, $i=1, \ldots, k-1$.

With k small (say, 8 or smaller), the above scheme is easily implemented for typical likelihoods (Stephens and Smith, 1992). The conditionals for the δ_i are sampled by using rejection methods and the discrete distribution for σ is sampled directly using cumulative density function inversion. However, for larger values of k the latter becomes increasingly infeasible and direct sampling from $\pi(\sigma|\delta, y)$ is not possible. But now recall, from Section 3, that we are at liberty to form a hybrid simulation strategy, carrying out this step of the Gibbs structure by, within the step, carrying out a run of any other form of MCMC simulation that might be more suitably adapted to the problem.

For this, let us note that for this step the state space is the set of all $(k!)$ permutations on $(1, \ldots, k)$. At each drawing, direct simulation is involving all $k!$ possibilities; we therefore need a method which at each drawing involves a substantially smaller number of possibilities (while still being capable of moving eventually from any current state to any other).

The Metropolis–Hastings algorithm (Section 3.4), which uses an intermediate transition mechanism to generate candidate 'next states' in the evolution of a chain, offers considerable creative flexibility in designing such a method. An obvious possibility is to exploit the fact that the state space has a 'local neighbourhood' structure, two permutations being considered to be neighbours if they are only a small number of transpositions (for example, one or two) apart. In the notation of Section 3.4, we set $q(\sigma, \sigma')=0$ for all σ' not in the neighbourhood (however defined) of σ. In the absence of strong prior information, a possibility would be to set $q(\sigma, \sigma')=q(\sigma', \sigma)$ for all σ' in the neighbourhood of σ. This would lead to the Metropolis algorithm, within the Gibbs step for σ given δ, with

$$\alpha(\sigma, \sigma')=\min\left\{\frac{\pi(\sigma'|\delta, y)}{\pi(\sigma|\delta, y)}, 1\right\}.$$

However, if information were available about the relative ordering probabilities relating to locations involved in the transposition(s) differences between σ and σ' we would have $q(\sigma, \sigma') \neq q(\sigma', \sigma)$ and a general Hastings algorithm would be used within the Gibbs step for σ given δ.

7.2. *Image Modelling with Deformable Templates*

Markov random fields have been extensively used in image modelling as a stylized way of representing assumed local correlation structures, with the additional pleasant property that posterior calculations are often straightforwardly approached via the Gibbs sampler. However, such priors are not well suited to representing knowledge that is often available regarding the detailed global geometry of the underlying object in a noisy image. Here, we typically need to be able to describe a prior for an image in terms of (possibly interconnected) curves, collectively described as a template and

represented by some suitable parsimonious parameterization. Assigning prior distributions to the parameters induces a prior for the image in the form of probabilistic templates. The templates, in turn, induce priors over pixel values, which can be combined with a likelihood based on data in the form of noisy observations of pixels to produce a posterior for the template parameters. See Ripley and Sutherland (1990) and Amit *et al.* (1991) for applications of this methodology.

A Gaussian parametric application arises in Phillips and Smith (1992), in the context of a hierarchical construction of a prior distribution for the human face. Inner and outer (including hair) facial boundary templates are defined as parameterized curves. Conditional on the inner boundary, further parameters define locations and shape features for eyebrows, eyes, nose and mouth. The prior for the complete parameter vector is defined via carefully structured (conditional) Gaussian forms. Conditional on feature boundaries, further Gaussian distributed parameters define differing mean pixel levels.

Posterior simulations are based on an appropriately structured Gibbs sampler, with generation within each Gibbs step carried out with the random walk version of the Hastings algorithm (Section 3.4). The proposal distribution takes the form of a Gaussian distribution initially located at the realized value from the previous Gibbs cycle and with spread scaled to be about one-half of the magnitude of the prior spread (see Muller (1991) for detailed discussion of the choice of scale in the Gaussian random walk Hastings algorithm).

8. CONCLUDING REMARKS

At the time of writing, most reported applications of MCMC methods in Bayesian statistics have focused on the Gibbs sampler. In part, this is no doubt due to its relative ease of implementation and direct applicability to a wide range of commonly encountered problems. In part, it may also reflect the fact that, in a sense, it is a stochastic analogue of the EM algorithm, replacing expectations and maxima by random drawings, so that all the problems investigated over the years from a likelihood perspective using the EM algorithm become natural targets for a Bayesian Gibbs sampling approach.

However, along with increasing familiarity and experimentation with MCMC methods there will undoubtedly be new insights into which method is best suited to which problem, and how pure and hybrid algorithms should be tuned to best advantage.

ACKNOWLEDGEMENTS

Much of the work of the first author and co-workers referenced here has been supported by the UK Science and Engineering Research Council. Support has also been provided by the Mathematical Applications Section of CIBA-Geigy, Basle. We are grateful to referees for helpful comments on an earlier version of this material. Conversations with Alan Gelfand have been particularly helpful.

APPENDIX A: SOME CONVERGENCE THEORY

We outline some of the concepts required and results available for MCMC convergence

in a rather general mathematical setting: see Tierney (1991) and Roberts and Smith (1992) for some extensive discussion. Throughout, notation for distributions and densities will be used interchangeably.

Let $X=(X^0, \ldots, X^t, \ldots)$, $X^t \in E \subseteq \mathbf{R}^n$, be a Markov chain (MC) with transition kernel K: $E \times E \to \mathbf{R}$ such that, with respect to a σ-finite measure ν on $\mathbf{R}^n$,

$$P(X^t \in A \mid X^{t-1}=x)=\int_A K(x, x')\,\mathrm{d}\nu(x')+r(x)\,I[x \in A],$$

where

$$r(x)=P(X^t=x \mid X^{t-1}=x)=1-\int K(x, x')\,\mathrm{d}\nu(x').$$

If we define

$$K^{(t)}(x, x')=\int K^{(t-1)}(x, y)\,K(y, x')\,\mathrm{d}\nu(y)+K^{(t-1)}(x, x')\,r(x')+\{1-r(x)\}^{t-1}K(x, x'),$$

then $K^{(t)}_{x_0}=K^{(t)}(x_0, \ )$ is the density (with respect to ν) of X^t, given $X^0=x_0$, excluding realizations with $X^j=x_0$, $j=1, \ldots, t$. Let π denote the invariant distribution satisfying

$$\pi(A)=\int P(X^1 \in A \mid X^0=x)\,\pi(x)\,\mathrm{d}\nu(x)$$

for all ν-measurable A. To avoid distracting complications we assume that $E=\{x;\ \pi(x)>0\}$. We recall that X is called π irreducible if, for all $x_0 \in E$ and for some $t \geqslant 0$, $\pi(A)>0$ implies that $K^{(t)}_{x_0}(A)>0$ and is called aperiodic if there does not exist a measurable partition $E=(B_0, \ldots, B_{r-1})$, for some $r \geqslant 2$, such that $P(X^t \in B_{t \bmod(r)} \mid x_0 \in B_0)=1$ for all t. A key result implying the convergence in distribution and ergodic results discussed in Section 3.1 is the following.

Theorem 1. If K is π irreducible and aperiodic then, for all $x_0 \in E$,

(a) $|K^{(t)}_{x_0}-\pi|=\int_E |K^{(t)}_{x_0}(x)-\pi(x)|\,\mathrm{d}\nu(x) \to 0$, as $t \to \infty$, and

(b) for real-valued, π-integrable f,

$$t^{-1}\{f(X^1)+\ldots+f(X^t)\} \to \int f(x)\,\pi(x)\,\mathrm{d}\nu(x), \qquad \text{almost surely, as } t \to \infty.$$

Proof. See Tierney (1991), based on Nummelin (1984).

We are particularly interested in the Gibbs (with blocking $x=(x_1, \ldots, x_k)$, $1<k \leqslant n$) and Hastings algorithms, defined by

$$K_{\mathrm{G}}(x, x')=\prod_{i=1}^{K} \pi(x_i' \mid x_j, j>i, x_j', j<i)$$

and

$$K_{\mathrm{H}}(x, x')=q(x, x')\,\alpha(x, x'),$$

where q: $E \times E \to \mathbf{R}$ is an MC kernel (with respect to ν) and α: $E \times E \to [0, 1]$ is as defined in Section 3.3. For the Gibbs algorithm, $r_{\mathrm{G}}(x)=0$ for all $x \in E$; for the Hastings algorithm,

$$r_{\mathrm{H}}(x)=1-\int q(x, x')\,\alpha(x, x')\,\mathrm{d}\nu(x').$$

In both cases, π is an invariant distribution by construction. Sufficient conditions for the applicability of theorem 1 to many standard problems are given by the following lemmas.

Lemma 1 (Gibbs).

(a) If ν is a discrete measure, then K_G is well defined and aperiodic. Irreducibility of the discrete chain is then a sufficient condition.
(b) If ν is Lebesgue measure and π is lower semicontinuous, then K_G with $\pi(x_i|x_{-i}) = \pi(x)/\pi(x_{-i})$ is well defined and aperiodic. A sufficient (but by no means necessary) condition for π-irreducibility is that E be connected and each $(k-1)$-dimensional marginal, $\pi(x_{-i}) = \int \pi(x)\,dx_i$, be locally bounded.

Proof. See Roberts and Smith (1992).

Lemma 2 (Hastings). For general ν:

(a) if q is aperiodic, or $P(X^t = X^{t-1}) > 0$ for some t, then the Hastings chain is aperiodic;
(b) if q is π irreducible and $\alpha(x, x') > 0$ for all $(x, x') \in E \times E$, then the Hastings chain is π irreducible.

Proof. See Roberts and Smith (1992).

Regarding rates of convergence, two available results are the following.

Theorem 2. Suppose that there exists K^*: $E \to \mathbf{R}$ such that $K(x, x') \geq K^*(x')$ for all $(x, x') \in E \times E$ and such that

$$\int K^*(x')\,d\nu(x') = 1 - \rho, \qquad \text{for } 0 < \rho < 1.$$

Then, for all t and $x_0 \in E$,

$$|K_{x_0}^{(t)} - \pi| < 2\rho^t.$$

Proof. See Roberts and Polson (1992) for this and related results.

The above is an L^1-result, applicable under suitable conditions to both K_G and K_H. For a result applicable to K_G in an L^2-setting, define, for g, h: $E \to \mathbf{R}$,

$$\langle g, h \rangle = \int g(x)\, h(x)/\pi(x)\,d\nu(x)$$

and $\|g\|^2 = \langle g, g \rangle$, and define the Hilbert–Schmidt norm (in cases of $r(x) = 0$),

$$\|K\|_{HS} = \int \|K_x^{(1)}\|^2\, \pi(x)\,d\nu(x).$$

We then have the following theorem.

Theorem 3. If X is π irreducible and $\|K\|_{HS} < \infty$, then there exists ρ and M: $E \to \mathbf{R}^+$, with $0 < \rho < 1$, such that, for all t and $x_0 \in E$,

$$\|K_{x_0}^{(t)} - \pi\| < M(x_0)\rho^t.$$

Proof. See Schervish and Carlin (1990).

We note that the Hilbert–Schmidt condition is frequently easy to check for K_G, holding, for example, when $K(x\,x')/\pi(x')$ is bounded.

REFERENCES

Amit, Y. (1991) On rates of convergence of stochastic relaxation for Gaussian and non-Gaussian distributions. *J. Multiv. Anal.*, **38**, 82–99.

Amit, Y. and Grenander, U. (1991) Comparing sweep strategies for stochastic relaxation. *J. Multiv. Anal.*, **37**, 197–222.

Amit, Y., Grenander, U. and Piccioni, M. (1991) Structural image restoration through deformable templates. *J. Am. Statist. Ass.*, **86**, 376–387.

Applegate, D., Kannan, R. and Polson, N. G. (1990) Random polynomial time algorithms for sampling from joint distributions.

Barone, P. and Frigessi, A. (1989) Improving stochastic relaxation for Gaussian random fields. *Probab. Engng Inform. Sci.*, **43**, 369–389.

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

Box, G. E. P. (1980) Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *J. R. Statist. Soc.* A, **143**, 383–430.

Carlin, B. P. and Gelfand, A. E. (1990) An iterative Monte Carlo method for non-conjugate Bayesian analysis.

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

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

Escobar, M. (1991) Estimating normal means with a Dirichlet process prior.

Gelfand, A. E., Dey, D. K. and Chang, H. (1992) Model determination using predictive distributions with implementation via sampling-based methods. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. Berger, A. P. Dawid and A. F. M. Smith), pp. 147–167. Oxford: Oxford University Press.

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. *J. Am. Statist. Ass.*, **85**, 972–985.

Gelfand, A. E. and Kuo, L. (1991) Nonparametric Bayesian bioassay including ordered polytomous response. *Biometrika*, **78**, 657–666.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Ass.*, **85**, 398–409.

——(1991) Gibbs sampling for marginal posterior expectations. *Communs Statist. Theory Meth.*, **20**, 1747–1766.

Gelfand, A. E., Smith, A. F. M. and Lee, T.-M. (1992) Bayesian analysis of constrained parameter and truncated data problems. *J. Am. Statist. Ass.*, **87**, 523–532.

Gelman, A. and Rubin, D. B. (1992) A single series from the Gibbs sampler provides a false sense of security. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. Berger, A. P. Dawid and A. F. M. Smith), pp. 627–633. Oxford: Oxford University Press.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattn Anal. Mach. Intell.*, **6**, 721–741.

George, E. I., Makov, U. E. and Smith, A. F. M. (1991) Conjugate likelihood distributions. Submitted to *Scand. J. Statist.*

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

Green, P. J. and Han, X.-L. (1992) Metropolis methods, gaussian proposals, and antithetic variables. *Lect. Notes Statist.*, **74**, 142–164.

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

Hills, S. E. and Smith, A. F. M. (1992) Parameterization issues in Bayesian inference. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. Berger, A. P. Dawid and A. F. M. Smith), pp. 227–246. Oxford: Oxford University Press.

Kuo, L. and Smith, A. F. M. (1992) Bayesian computation for survival models via the Gibbs sampler. In *Survival Analysis and Related Topics* (eds J. P. Klien and P. K. Goel). New York: Dekker.

Marriott, J., Ravishanker, N. and Gelfand, A. E. (1992) Bayesian analysis of ARMA processes: complete sampling-based inferences under full likelihoods.

Marriott, J. and Smith, A. F. M. (1992) Reparameterization aspects of numerical Bayesian methodology of autoregressive moving average models. *J. Time Ser. Anal.*, **13**, 327–343.

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

Muller, P. (1991) A generic approach to posterior integration and Gibbs sampling.

Nummelin, E. (1984) *General Irreducible Markov Chains and Non-negative Operators*. Cambridge: Cambridge University Press.

Peskun, P. H. (1973) Optimum Monte Carlo sampling using Markov chains. *Biometrika*, **57**, 97–109.

Phillips, D. and Smith, A. F. M. (1992) Bayesian faces. *Technical Report*. Department of Mathematics, Imperial College of Science, Technology and Medicine, London.

Raftery, A. and Lewis, S. (1992) How many iterations in the Gibbs sampler? In *Bayesian Statistics 4* (eds J. M. Bernardo, J. Berger, A. P. Dawid and A. F. M. Smith), pp. 765–776. Oxford: Oxford University Press.

Ramgopal, P., Laud, P. K. and Smith, A. F. M. (1992) Nonparametric Bayesian bioassay with prior constraints on the shape of the potency curve. *Biometrika*, to be published.

Ripley, B. D. (1987) *Stochastic Simulation*. New York: Wiley.

Ripley, B. D. and Sutherland, A. I. (1990) Finding spiral structures in images of galaxies. *Phil. Trans. R. Soc. Lond.* A, **332**, 477–485.

Roberts, G. O. (1992) Convergence diagnostics of the Gibbs sampler. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. Berger, A. P. Dawid and A. F. M. Smith), pp. 777–784. Oxford: Oxford University Press.

Roberts, G. O. and Hills, S. E. (1992) Assessing distributional convergence of the Gibbs sampler. Submitted to *J. Am. Statist. Ass.*

Roberts, G. O. and Polson, N. G. (1992) A note on the geometric convergence of the Gibbs sampler. Submitted to *J. R. Statist. Soc.* B.

Roberts, G. O. and Smith, A. F. M. (1992) Some convergence theory for Markov chain Monte Carlo. Submitted to *Stoch. Processes Appl.*

Rosenthal, J. (1992) Rates of convergence for Gibbs sampling for variance component models.

Rubin D. B. (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.*, **12**, 1151–1172.

Schervish, M. J. and Carlin, B. P. (1990) On the convergence of successive substitution sampling.

Smith, A. F. M. (1991) Bayesian computational methods. *Phil. Trans. R. Soc. Lond.* A, **337**, 369–386.

Smith, A. F. M. and Gelfand, A. E. (1992) Bayesian Statistics without tears: a sampling–resampling perspective. *Am. Statistn.*, **46**, 84–88.

Stephens, D. A. (1991) Bayesian retrospective multiple changepoint detection. *Appl. Statist.*, to be published.

Stephens, D. A. and Smith, A. F. M. (1992) Bayesian multipoint gene mapping.

Tanner, M. A. (1991) Tools for statistical inference, observed data and data augmentation methods. *Lect. Notes Statist.*, **67**.

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

Tierney, L. (1991) Markov chains for exploring posterior distributions. Submitted to *Ann. Statist.*

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1986) *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley.

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

Wakefield, J. C., Smith, A. F. M., Racine-Poon, A. and Gelfand, A. E. (1992) Bayesian analysis of linear and non-linear population models using the Gibbs sampler. *Appl. Statist.*, to be published.