Markov Chain Monte Carlo
A Contribution to the Encyclopedia of Environmetrics

Galin L. Jones and James P. Hobert
Department of Statistics
University of Florida

May 2000

1 Introduction

Realistic statistical models often give rise to probability distributions that are computationally difficult to use for inference. Fortunately, we now have a collection of algorithms, known as Markov chain Monte Carlo (MCMC), that has brought many of these models within our computational reach. In turn, this has lead to a staggering amount of both theoretical and applied work on MCMC. Thus we do not propose a complete overview of MCMC in this article. Actually, we only hope to get the reader started in the right direction. To this end, we spend some time indicating when it is appropriate to consider using MCMC. Then we introduce the fundamental algorithms and address some general implementation issues. An obvious omission from this paper is an introduction to the Markov chain theory underpinning MCMC algorithms. Instead we refer the interested reader to Besag, Green, Higdon & Mengersen (1995), Robert & Casella (1999), and Roberts & Rosenthal (1998). However, throughout we will use the phrases "MCMC algorithm" and "Markov chain" interchangeably.

MCMC is most commonly used to evaluate the complicated integrals encountered in Bayesian hierarchical models. However, analytically intractable integrals are also encountered in maximum likelihood estimation (Geyer & Thompson 1992), spatial statistics (Besag & Green 1993), and image analysis (Besag 1993). They are even found within other algorithms, for example, within the "E" step of the EM algorithm (McCulloch 1997). Specific environmetrical applications where this

situation is encountered include remote sensing (Green & Strawderman 1994), capture-recapture studies (Feinberg, Johnson & Junker 1999), agricultural field experiments (Besag et al. 1995), animal breeding (Wang, Rutledge & Gianola 1993), and genetics (Wilson & Balding 1998). Of course, this list is not meant to be exhaustive. Indeed, as we attempt to accurately model natural phenomena this setting becomes the rule rather than the exception.

The following example is motivated by the salamander data set introduced by McCullagh & Nelder (1989). Specifically, this is an application where analytically intractable integrals are found even though we are using a fairly uncomplicated model.

EXAMPLE 1. Suppose that for a given salamander species we have I females and J males. Also, suppose that the experiment consists of observing all $N = IJ$ of the possible heterosexual crosses. At the conclusion of the experiment we will have N binary observations, y_{ij}, indicating the success or failure of the mating episode of the ith female with the jth male. Let u_i denote the (random) effect of the ith female and v_j the (random) effect of the jth male. Now assume that given u_i and v_j the y_{ij}'s are independent and that $y_{ij}|u_i, v_j \sim \text{Bernoulli}(\pi_{ij})$ where π_{ij} is the probability of a successful mating, i.e. $\pi_{ij} = \Pr(Y_{ij} = 1)$, and

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \mu + u_i + v_j$$

where μ is unknown. To finish the model specification we assume that the u_i are independently $\text{N}(0, \sigma_u^2)$ for $i = 1, 2, \ldots, I$ and that the v_j are independently $\text{N}(0, \sigma_v^2)$ for $j = 1, 2, \ldots, J$.

A standard frequentist analysis requires maximizing the marginal likelihood function which is given up to a multiplicative constant by

$$\begin{aligned} L(\mu, \sigma_u^2, \sigma_v^2; \boldsymbol{y}) \quad \propto \quad & \frac{1}{(\sigma_u^2)^{I/2}(\sigma_v^2)^{J/2}} \int\int \prod_{i,j} \frac{\exp\{y_{ij}(\mu + u_i + v_j)\}}{1 + \exp\{y_{ij}(\mu + u_i + v_j)\}} \\ & \exp\left\{-\frac{1}{2\sigma_u^2}\sum_{i=1}^{I} u_i^2 - \frac{1}{2\sigma_v^2}\sum_{j=1}^{J} v_j^2\right\} d\boldsymbol{u}d\boldsymbol{v}. \end{aligned} \tag{1}$$

Due to the crossed nature of the design (1) involves intractable integrals. It is natural to think of the random effects as missing data and consider using the EM algorithm (Dempster, Laird & Rubin 1977) to maximize the likelihood. But if we try to implement the standard EM algorithm

we will encounter an expectation within the "E" step that also involves intractable integrals. Of course, there are Monte Carlo versions of the EM algorithm (see e.g. Booth & Hobert 1999) that overcome this difficulty.

‖

What are we to do when faced with a situation, as in example 1, where there is a probability or expectation that is difficult to calculate? Well, a good approximation to the integral will probably be sufficient. In this case, we could use an analytical approximation, such as the Laplace approximation (Tierney, Kass & Kadane 1989), or a numerical approximation. If we choose a numerical method then there are several options available. Specifically, we could use standard numerical integration, e.g. quadrature methods (Abramowitz & Stegun 1964), quasi-Monte Carlo methods (Fang, Wang & Bentler 1994), classical Monte Carlo integration, or Markov chain Monte Carlo methods. Each of these methods have advantages and disadvantages when compared to the others making the appropriate choice highly problem specific. Generally, deterministic methods will converge more quickly than sampling based approaches when the dimension of the problem is small. On the other hand, as the dimension of the integrals increases the computational cost of deterministic methods becomes prohibitive (Traub & Wozniakowski 1994) making sampling based methods preferable. If it is available, Monte Carlo integration based on independent and identically distributed (iid) samples is preferred. But if the distribution of interest is too complicated to allow iid sampling then MCMC algorithms may provide the wherewithal to approximate the integrals.

In order to make our discussion concrete we require some notation. Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)'$ and suppose $\pi(\boldsymbol{x})$ is a density function defined on $\mathcal{X} \subseteq \mathbb{R}^p$ with $p \geq 1$. Also, let F_π be the probability distribution corresponding to π. Then the basic problem that we are concerned with throughout is that we want to calculate the analytically intractable expectation of some function h with respect to F_π, that is, $E_\pi h := \int_{\mathcal{X}} h(\boldsymbol{x})\pi(\boldsymbol{x})d\boldsymbol{x}$.

Classical Monte Carlo integration requires that we produce an iid sample $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(n)}$ from the target distribution F_π. This is often done with rejection sampling (Robert & Casella 1999). With the sample in hand the desired integral approximation may be obtained via an

ergodic average, that is,

$$\bar{h}_n := \frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{x}^{(i)}) \approx E_\pi h \tag{2}$$

when n is large. Of course, it would be helpful to know just how large n needs to be in order for the approximation to be good. Under iid sampling it is easy to obtain a measure of the accuracy of $\bar{h}_n$. Specifically, note that

$$\mathrm{var}(\bar{h}_n) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{var}(h(\boldsymbol{x}^{(i)}))$$

which is consistently estimated by

$$v_n := \frac{1}{n^2}\sum_{i=1}^{n} \left(h(\boldsymbol{x}^{(i)}) - \bar{h}_n\right)^2.$$

Because

$$\frac{\bar{h}_n - E_\pi h}{\sqrt{v_n}}$$

is approximately $\mathrm{N}(0,1)$ when n is large, we can construct a valid confidence interval for $E_\pi h$. Then we can say that $\bar{h}_n$ is sufficiently accurate (i.e. n is large enough) when the width of the confidence interval is small.

If we cannot produce independent samples from F_π then MCMC techniques may be useful. The Metropolis-Hastings algorithm and the Gibbs sampler are the basic MCMC algorithms. Additionally, there are algorithms, known as hybrid samplers, that use Metropolis updates and the Gibbs sampler simultaneously. Gilks, Richardson & Spiegelhalter (1996) provide a wonderful introduction to applied MCMC while Robert & Casella (1999) consider MCMC, as well as standard Monte Carlo integration, from a slightly more technical perspective.

2 MCMC Algorithms

MCMC algorithms provide a method that allows us to use ergodic averages (2) to approximate the desired integral, $E_\pi h$, with dependent draws, $\boldsymbol{x}^{(0)}, \boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$, from distributions that approximate F_π. Specifically, with MCMC we can obtain a sample $\boldsymbol{x}^{(0)}, \boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$ that is neither independent nor identically distributed and yet, remarkably, we can still use an ergodic average

(2) to obtain a strongly consistent estimate of $E_\pi h$. However, it is no longer easy to obtain a valid estimate of the standard error of $\bar{h}_n$ (Geyer 1992). We will revisit this issue in the next section.

For our purpose, the Metropolis-Hastings algorithm (Chib & Greenberg 1995) is the basic MCMC algorithm. To begin we must choose a conditional density, $q(\boldsymbol{y}|\boldsymbol{x})$ say, from which we may easily draw random variates. From this density we can then propose a move, $\boldsymbol{y}$, conditional on the current state, $\boldsymbol{x}$, of the algorithm. However, to ensure that we don't move to $\boldsymbol{y}$ too often we will only accept $\boldsymbol{y}$ with probability

$$\alpha(\boldsymbol{x}, \boldsymbol{y}) = \min\left(\frac{\pi(\boldsymbol{y})q(\boldsymbol{x}|\boldsymbol{y})}{\pi(\boldsymbol{x})q(\boldsymbol{y}|\boldsymbol{x})}\ ,\ 1\right) \tag{3}$$

otherwise the chain will remain at $\boldsymbol{x}$. Putting this together gives the basic Metropolis-Hastings algorithm.

1. Choose an arbitrary starting value $\boldsymbol{x}^{(0)}$.

2. Sample an observation $\boldsymbol{y}$ from $q(\boldsymbol{y}|\boldsymbol{x}^{(0)})$.

3. Set $\boldsymbol{x}^{(1)} = \boldsymbol{y}$ with probability

$$\alpha(\boldsymbol{x}^{(0)}, \boldsymbol{y}) = \min\left(\frac{\pi(\boldsymbol{y})q(\boldsymbol{x}^{(0)}|\boldsymbol{y})}{\pi(\boldsymbol{x}^{(0)})q(\boldsymbol{y}|\boldsymbol{x}^{(0)})}\ ,\ 1\right)$$

   otherwise reject $\boldsymbol{y}$ and set $\boldsymbol{x}^{(1)} = \boldsymbol{x}^{(0)}$.

4. Continue in this fashion until the sample $\{\boldsymbol{x}^{(0)}, \boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}\}$ has been obtained.

There are two especially noteworthy features of this algorithm. The first is that the normalizing constant for π is not needed to calculate $\alpha(\boldsymbol{x}, \boldsymbol{y})$. Secondly, the flexibility of this algorithm is evident since it works for almost any choice of q that has at least the same support as π.

In the next two examples we introduce common ways to choose the candidate density q. In both cases we illustrate the required calculations with "toy" examples that do not require MCMC in order to simulate from the target distribution. However, these examples should be useful for the reader encountering MCMC for the first time.

EXAMPLE 2. If we choose $q(\boldsymbol{y}|\boldsymbol{x}) = q(\boldsymbol{y})$ so that the proposed move is independent of the current state, then we have the *independence* sampler. As a specific application, suppose that the target distribution is an Exponential with mean 1, that is, the density is given by

$$\pi(x) = e^{-x} I(x \geq 0).$$

Then an independence sampler would result if we chose as the candidate an Exponential with mean β, or

$$q(y) = \beta e^{-\beta y} I(y \geq 0).$$

Given that the current value of the chain is x we will accept the proposal, y, with probability

$$\alpha(x, y) = \min\left(\frac{e^{-y}\beta e^{-\beta x}}{e^{-x}\beta e^{-\beta y}}\ ,\ 1\right) = \min\left(e^{x-y}e^{\beta(y-x)}\ ,\ 1\right)$$

and otherwise the chain will remain at x.

‖

EXAMPLE 3. The *random walk* Metropolis-Hastings algorithm results when $q(\boldsymbol{y}|\boldsymbol{x}) = q(\boldsymbol{y} - \boldsymbol{x}) = q(\boldsymbol{x} - \boldsymbol{y})$. For example, suppose that the target distribution is $\mathrm{N}(0, 1)$ and the proposal is drawn from a $\mathrm{N}(x, 1)$ distribution, i.e.

$$q(y|x) = \frac{1}{\sqrt{2\pi}} e^{-0.5(y-x)^2}.$$

If x is the current state of the Markov chain then accept the proposal y with probability

$$\alpha(x, y) = \min\left(\exp\left[0.5(x^2 - y^2)\right]\ ,\ 1\right).$$

‖

The Gibbs sampler (Gelfand & Smith, 1990, and Casella & George, 1992) is also a basic MCMC algorithm. Because there is an article by Geir Storvik devoted solely to the Gibbs sampler in this volume we will content ourselves with a brief description. This algorithm is useful when we can identify full conditional densities that are easy to sample from directly. As above, consider a

random vector $\boldsymbol{x} = (x_1, \ldots, x_p)'$. Let $\boldsymbol{x}_{-i} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i-1}, \boldsymbol{x}_{i+1}, \ldots, \boldsymbol{x}_p)'$. Then the full conditional densities $f_i(x_i|\boldsymbol{x}_{-i})$, for $i = 1, \ldots, p$ are

$$f_i(x_i|\boldsymbol{x}_{-i}) = \frac{\pi(\boldsymbol{x})}{\int \pi(\boldsymbol{x}) dx_i}.$$

Typically, these densities are straightforward to determine since they are proportional to the target density π. This allows us to use the standard techniques described by Gilks (1996) to identify them. The Gibbs sampling algorithm follows. Given $\boldsymbol{x}^{(n)} = (x_1^{(n)}, \ldots, x_p^{(n)})'$ generate

$$\begin{aligned} x_1^{(n+1)} &\sim f_1(x_1|x_2^{(n)}, \ldots, x_p^{(n)}) \\ x_2^{(n+1)} &\sim f_2(x_2|x_1^{(n+1)}, x_3^{(n)}, \ldots, x_p^{(n)}) \\ &\vdots \\ x_p^{(n+1)} &\sim f_p(x_p|x_1^{(n+1)}, \ldots, x_{p-1}^{(n+1)}). \end{aligned} \tag{4}$$

This constitutes one full update. If at least one of the x_i is multivariate this is called a *block* Gibbs sampler.

An important feature of both the Gibbs sampler and the Metropolis-Hastings algorithm is that they are easy to implement. However, it is not always clear which one should be used. In fact, it is unfortunate but true that it is impossible to make a general statement about which algorithm is best. The preferred algorithm is one that rapidly explores the support of π. That is, we want an algorithm that will quickly produce a sample that is representative of F_π. If the candidate q is chosen wisely, i.e. to closely mimic π, Metropolis-Hastings will usually accomplish this faster than Gibbs. On the other hand, if q is chosen to be very different than π there will probably be a prohibitive number of rejections and hence Metropolis-Hastings will investigate π slowly. One possible solution to this dilemma is to use Gibbs and intersperse an occasional Metropolis update. This is an example of a hybrid algorithm.

Hybrid MCMC algorithms may also be used when at least one of the full conditional densities is of a non-standard form. In this case, we could embed a Metropolis-Hastings step within the Gibbs sampler. This approach is illustrated in the next example where we consider a version of the normal means model with an improper prior specification. The prior that we use is the standard

uniform shrinkage prior and is a special case of the default priors recently developed by Natarajan & Kass (2000).

EXAMPLE 4. Suppose we observe a vector of data $\boldsymbol{y} = (y_1, y_2, \ldots, y_k)'$ conditional on the parameters $\boldsymbol{b} = (b_1, b_2, \ldots, b_k)'$ and β according to the following model. Specifically, we assume that conditional on b_i and β each y_i is independent. Also, conditional on θ each b_i is assumed independent. Now let ϕ be a known (or estimated from the data) positive constant. Let f denote a generic density. Then for $i = 1, 2, \ldots, k$

$$\begin{aligned} y_i|b_i, \beta &\sim \mathrm{N}(\beta + b_i, \phi) \\ b_i|\theta &\sim \mathrm{N}(0, \theta) \\ f(\theta) &\propto (\phi + \theta)^{-2} I(\theta > 0) \\ f(\beta) &\propto 1 \end{aligned}$$

where $I(\theta > 0) = 1$ if $\theta > 0$ and 0 otherwise.

Now all inference will proceed through the posterior π. Specifically, suppose for some function $h(\boldsymbol{b}, \beta, \theta)$ we want to calculate $E_\pi h$. This is a situation where we may want to use the Gibbs sampler so that we can approximate $E_\pi h$ with an ergodic average (2).

The first thing we need are the full conditional densities. Specifically, we want to do a block update of the b_i and univariate updates for β and θ. This will require the conditional densities for $\beta|\boldsymbol{b}, \theta, \boldsymbol{y}$, $\boldsymbol{b}|\beta, \theta, \boldsymbol{y}$, and $\theta|\beta, \boldsymbol{b}, \boldsymbol{y}$. Let $\bar{y} = \frac{1}{k}\sum y_i$ and $\bar{b} = \frac{1}{k}\sum b_i$. Then

$$\begin{aligned} \beta|\boldsymbol{b}, \theta, \boldsymbol{y} &\sim \mathrm{N}(\bar{y} - \bar{b}, \phi/k) \\ \boldsymbol{b}|\beta, \theta, \boldsymbol{y} &\sim \mathrm{N}_k(\boldsymbol{\mu}, V) \end{aligned}$$

where

$$\boldsymbol{\mu} = \left(\frac{(y_1 - \beta)\theta}{\theta + \phi}, \frac{(y_2 - \beta)\theta}{\theta + \phi}, \cdots, \frac{(y_k - \beta)\theta}{\theta + \phi} \right)' \quad \text{and} \quad V = \frac{\theta\phi}{\theta + \phi}\mathbf{I}_k.$$

Finally, the density for $\theta|\beta, \boldsymbol{b}, \boldsymbol{y}$ is of the following non-standard form

$$f(\theta|\beta, \boldsymbol{b}, \boldsymbol{y}) = \frac{1}{c}(\theta + \phi)^{-2}\theta^{-k/2} \exp\left\{ \frac{-1}{2\theta} \sum_{i=1}^{k} b_i^2 \right\}$$

where c is the normalizing constant. We propose (making no claims about optimality) to use an independence sampler with candidate density

$$q(\theta) = \frac{\left(\sum_{i=1}^{k} b_i^2/2\right)^{k/2-1}}{\Gamma(k/2-1)} \theta^{-k/2} \exp\left\{\frac{-1}{2\theta}\sum_{i=1}^{k} b_i^2\right\}.$$

That is, our candidate density is an Inverse Gamma $(k/2-1, \sum b_i^2/2)$. If we let $\theta^{(n)}$ denote the current value of the chain and θ be the proposed move then the acceptance probability for this step is

$$\alpha(\theta^{(n)}, \theta) = \min\left\{\left(\frac{\theta^{(n)}+\phi}{\theta+\phi}\right)^2, 1\right\}.$$

Now put all of this together and obtain the following hybrid sampler. Let $(\beta^{(0)}, \boldsymbol{b}^{(0)}, \theta^{(0)})$ be specified starting values. Then we can obtain $(\beta^{(1)}, \boldsymbol{b}^{(1)}, \theta^{(1)})$ in three steps.

1. Generate $\beta^{(1)}|\boldsymbol{b}^{(0)}, \theta^{(0)}, \boldsymbol{y} \sim \mathrm{N}(\bar{y} - \bar{b}^{(0)}, \phi/k)$ where $\bar{b}^{(0)} = \sum_{i=1}^{k} b_i^{(0)}/k$.

2. Generate $\boldsymbol{b}^{(1)}|\beta^{(1)}, \theta^{(0)}, \boldsymbol{y} \sim \mathrm{N}_k(\boldsymbol{\mu}^*, V^*)$ where

$$\boldsymbol{\mu}^* = \left(\frac{(y_1-\beta^{(1)})\theta^{(0)}}{\theta^{(0)}+\phi}, \frac{(y_2-\beta^{(1)})\theta^{(0)}}{\theta^{(0)}+\phi}, \cdots, \frac{(y_k-\beta^{(1)})\theta^{(0)}}{\theta^{(0)}+\phi}\right)' \quad \text{and} \quad V^* = \frac{\theta^{(0)}\phi}{\theta^{(0)}+\phi}\mathbf{I}_k.$$

3. Generate θ from an Inverse Gamma $\left(k/2-1, \sum\left(b_i^{(1)}\right)^2/2\right)$. Then set $\theta^{(1)} = \theta$ with probability

$$\alpha(\theta^{(0)}, \theta) = \min\left\{\left(\frac{\theta^{(0)}+\phi}{\theta+\phi}\right)^2, 1\right\}.$$

   and leave $\theta^{(1)} = \theta^{(0)}$ with probability $1 - \alpha(\theta^{(0)}, \theta)$.

4. Continue in this fashion until we have the sample

$$\{(\beta^{(0)}, \boldsymbol{b}^{(0)}, \theta^{(0)}), (\beta^{(1)}, \boldsymbol{b}^{(1)}, \theta^{(1)}), \ldots, (\beta^{(n)}, \boldsymbol{b}^{(n)}, \theta^{(n)})\}.$$

Observe that we could have chosen any permutation of this update order. That is, there is nothing special about updating β then $\boldsymbol{b}$ then θ. Also, it would have been possible in this case to do a separate univariate update for each b_i. Finally, note that a starting value for $\beta^{(0)}$ is not required.

‖

3 General Implementation Issues

Once we have chosen the algorithm there are still several operational issues of concern. Some standard questions are (1) How should we choose the starting values? (2) How many chains should we use? (3) When should the sampling begin? (4) When should we decide to stop sampling? (5) Is there software available?

(1) The choice of starting values is irrelevant if we run the chain long enough. That is, it is valid to use ergodic averages to estimate the desired expectation no matter what starting point is chosen. However, trying a few preliminary runs with dispersed starting values can help protect against starting in a local mode. When this occurs the chain can get "stuck" for long periods and thus require many iterations to explore the support of π.

(2) The number of chains to be employed remains a somewhat unsettled issue. Gelman & Rubin (1992a), Gelman & Rubin (1992b), Geyer (1992), and Raftery & Lewis (1992) provide a thorough discussion. Those that advocate multiple independent chains with dispersed starting points argue that this is the only way to ensure that the support of π has been thoroughly explored. Of course, this can require substantial computational resources. On the other hand, if the algorithm is fairly well behaved then a single chain should explore the support of the target density rapidly. Also, an ergodic average based on a single chain will yield a less variable estimate of $E_\pi h$ than an ergodic average based on several chains.

(3) This is the issue of *burn-in* or the idea that we should discard the first B iterations of the algorithm so that the Markov chain will "forget" the starting value. Actually, $\bar{h}_n$ is still a consistent estimator of $E_\pi h$ if there is no burn-in, i.e. $B = 0$. However, if burn-in is used then the preferred method is to determine B before any simulation starts (Rosenthal, 1995, Roberts & Tweedie, 1999). Unfortunately, this can be a difficult task for a realistic model (**?**). Thus we typically have to make use of so-called convergence diagnostics which try to determine when to stop burn-in and start sampling based on the output of the algorithm. Brooks & Roberts (1998), Cowles & Carlin (1996), and Robert & Casella (1999) contain thorough discussions on the use of convergence diagnostics.

(4) A reasonable way to decide when to stop sampling is to wait until the ergodic average $\bar{h}_n$ reaches some prespecified level of accuracy. Just as with iid sampling, if a central limit theorem holds for $\bar{h}_n$ then we can construct sensible confidence intervals for $\bar{h}_n$. That is, we can say that the estimate of $E_\pi h$ is sufficiently accurate if the width of the confidence interval is small.

In order to calculate this confidence interval it is crucial that we obtain good estimate of the variance, $\sigma^2(h)$ say, of the asymptotic normal distribution of $\bar{h}_n$. Of course, this is complicated by the fact that MCMC algorithms produce dependent sequences. The simplest way of estimating $\sigma^2(h)$ is by the method of batch means Geyer (1992). An alternative is the regenerative simulation method proposed by Mykland, Tierney & Yu (1995).

(5) BUGS (Bayesian inference Using Gibbs Sampling)(Spiegelhalter, Thomas & Best 1999) is the only, to our knowledge, software package that will allow one to analyze a wide range of models via MCMC. As its name suggests this package focuses on Gibbs sampling but it will include a Metropolis-within-Gibbs step if a non-standard full conditional density is encountered. BUGS may be freely obtained via the World Wide Web at `http://www.mrc-bsu.cam.uk/bugs/welcome.shtml`.

We should also note that SAS PROC MIXED will allow one to fit a limited range of Bayesian hierarchical linear models via a Metropolis-Hastings algorithm.

Free software is also available that allows implementation of convergence diagnostics and aids in the analysis of MCMC output. CODA (Convergence Diagnosis and Output Analysis) was developed specifically for handling BUGS output and may be obtained from the same web site as BUGS. Another option is BOA (Bayesian Output Analysis) which is more flexible than CODA and may be obtained from `http://www.public-health.uiowa.edu/BOA/`.

References

Abramowitz, M. & Stegun, I. A. 1964. *Handbook of Mathematical Functions*, Dover, New York.

Besag, J. 1993. Towards Bayesian image analysis, *Journal of Applied Statistics* **20**: 107–119.

Besag, J., Green, P., Higdon, D. & Mengersen, K. 1995. Bayesian computation and stochastic systems (with discussion), *Statistical Science* **10**: 3–66.

Besag, J. & Green, P. J. 1993. Spatial statistics and Bayesian computation (disc: P53-102), *Journal of the Royal Statistical Society, Series B, Methodological* **55**: 25–37.

Booth, J. G. & Hobert, J. P. 1999. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm, *Journal of the Royal Statistical Society,* Series B **61**: 265–285.

Brooks, S. P. & Roberts, G. O. 1998. Convergence assessment techniques for Markov chain Monte Carlo, *Statistics and Computing* **8**: 319–335.

Casella, G. & George, E. I. 1992. Explaining the Gibbs sampler, *The American Statistician* **46**: 167–174.

Chib, S. & Greenberg, E. 1995. Understanding the Metropolis-Hastings algorithm, *The American Statistician* **49**: 327–335.

Cowles, M. K. & Carlin, B. P. 1996. Markov chain Monte Carlo convergence diagnostics: A comparative review, *Journal of the American Statistical Association* **91**: 883–904.

Dempster, A. P., Laird, N. M. & Rubin, D. B. 1977. Maximum likelihood from incomplete data via the em algorithm, *Journal of the Royal Statistical Society,* Series B **39**: 1–22.

Fang, K.-T., Wang, Y. & Bentler, P. M. 1994. Some applications of number-theoretic methods in statistics, *Statistical Science* **9**: 416–428.

Feinberg, S. E., Johnson, M. S. & Junker, B. W. 1999. Classical multilevel and Bayesian approaches to population size estimation using multiple lists, *Journal of the Royal Statistical Society,* Series A **162**: 383–405.

Gelfand, A. E. & Smith, A. F. M. 1990. Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association* **85**: 398–409.

Gelman, A. & Rubin, D. B. 1992a. Inference from iterative simulation using multiple sequences (disc: P483-501, 503-511), *Statistical Science* **7**: 457–472.

Gelman, A. & Rubin, D. B. 1992b. A single series from the Gibbs sampler provides a false sense of security, *Bayesian Statistics 4. Proceedings of the Fourth Valencia International Meeting*, Clarendon Press (Oxford), pp. 625–631.

Geyer, C. J. 1992. Practical Markov chain Monte Carlo (with discussion), *Statistical Science* **7**: 473–511.

Geyer, C. J. & Thompson, E. A. 1992. Constrained Monte Carlo maximum likelihood for dependent data (disc: P683-699), *Journal of the Royal Statistical Society, Series B, Methodological* **54**: 657–683.

Gilks, W. R. 1996. Full conditionals, *in* W. R. Gilks, S. Richardson & D. J. E. Spiegelhalter (eds), *Markov chain Monte Carlo in practice*, Chapman and Hall/CRC(Boca raton), pp. 75–88.

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. E. 1996. *Markov Chain Monte Carlo in Practice*, Chapman & Hall, London.

Green, E. J. & Strawderman, W. E. 1994. Determining accuracy of thematic maps, *The Statistician* **43**: 77–85.

McCullagh, P. & Nelder, J. A. 1989. *Generalized Linear Models (Second Edition)*, Chapman & Hall.

McCulloch, C. E. 1997. Maximum likelihood algorithms for generalized linear mixed models, *Journal of the American Statistical Association* **92**: 162–170.

Mykland, P., Tierney, L. & Yu, B. 1995. Regeneration in Markov chain samplers, *Journal of the American Statistical Association* **90**: 233–241.

Natarajan, R. & Kass, R. 2000. Reference Bayesian methods for generalized linear mixed models, *Journal of the American Statistical Association* **95**: 227–237.

Raftery, A. E. & Lewis, S. M. 1992. Comment on "the Gibbs sampler and Markov chain Monte Carlo", *Statistical Science* **7**: 493–497.

Robert, C. P. & Casella, G. 1999. *Monte Carlo Statistical Methods*, Springer, New York.

Roberts, G. O. & Rosenthal, J. S. 1998. Markov chain Monte Carlo: Some practical implications of theoretical results (with discussion), *Canadian Journal of Statistics* **26**: 5–31.

Roberts, G. O. & Tweedie, R. L. 1999. Bounds on regeneration times and convergence rates for Markov chains, *Stochastic Processes and their Applications* **80**: 211–229.

Rosenthal, J. S. 1995. Minorization conditions and convergence rates for Markov chain Monte Carlo, *Journal of the American Statistical Association* **90**: 558–566.

Spiegelhalter, D. J., Thomas, A. & Best, N. G. 1999. WinBUGS Version 1.2, MRC Biostatistics Unit, Cambridge: UK.

Tierney, L., Kass, R. E. & Kadane, J. B. 1989. Fully exponential Laplace approximations to expectations and variances of nonpositive functions, *Journal of the American Statistical Association* **84**: 710–716.

Traub, J. F. & Wozniakowski, H. 1994. Breaking intractability, *Scientific American* **270**: 102–107.

Wang, C. S., Rutledge, J. J. & Gianola, D. 1993. Marginal inference about variance components in a mixed linear model using Gibbs sampling, *Genetics, Selection, and Evolution* **25**: 41–62.

Wilson, I. J. & Balding, D. J. 1998. Genealogical inference from microsatellite data, *Geneteics* **150**: 499–510.