M

Markov Chain Monte Carlo Methods

Siddhartha Chib

Abstract

MCMC methods, an important class of Monte Carlo methods, have played a major role in the growth of Bayesian statistics and econometrics. In an MCMC simulation, one samples a given distribution (say the posterior distribution in a Bayesian model) by simulating a suitably constructed Markov chain whose invariant distribution is the target distribution. The Metropolis–Hastings algorithm and its special case, the Gibbs sampler, are two common ways of devising an MCMC simulation. We discuss how these methods originate, discuss implementation issues and provide examples. The use of MCMC methods in Bayesian prediction and model choice problems is also discussed.

Keywords

Autocorrelation; Bayesian econometrics; Bayesian prior–posterior analysis; Bayesian statistics; Invariance; Latent variables; Marginal likelihood; Markov chain Monte Carlo methods; Model choice; Prediction; Proposal densities; Reversibility; Transition density

JEL Classifications

C10

Introduction

Markov chain Monte Carlo methods, popularly called MCMC methods, are a class of Monte Carlo methods for sampling a given univariate or multivariate probability distribution (the target distribution). These methods play a central role in the theory and practice of modern Bayesian methods where they are used for the numerical calculation of quantities (such as the moments and quantiles of posterior and predictive densities) that arise in the Bayesian prior–posterior analysis. They have transformed the fields of Bayesian statistics and econometrics.

Suppose that in a given Bayesian model the prior density is $\pi(\boldsymbol{\theta})$ and the sampling density or likelihood function is $f(\mathbf{y}|\boldsymbol{\theta})$, where $\mathbf{y}$ is a vector of observations and $\boldsymbol{\theta} \in \mathcal{R}^d$ is an unknown parameter. In the Bayesian context, inferences about $\boldsymbol{\theta}$ are based on the posterior density $\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \pi(\boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta})$. Now suppose that one is interested in finding the mean of the posterior density

This chapter was originally published in *The New Palgrave Dictionary of Economics*, 2nd edition, 2008. Edited by Steven N. Durlauf and Lawrence E. Blume.

Palgrave Macmillan (ed.), *The New Palgrave Dictionary of Economics*,
DOI 10.1057/978-1-349-95121-5_2042-1

$$E(\boldsymbol{\theta}|\,\mathbf{y}) = \int \boldsymbol{\theta}\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$$

but that the integral cannot be computed analytically. In that case one can compute the integral by Monte Carlo sampling methods. The general idea is to calculate the integral from a sample

$$\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)} \sim \pi(\boldsymbol{\theta}|\mathbf{y}),$$

that is drawn from the posterior density. This sample can be used to estimate the posterior mean and other features of the posterior density. For instance, the posterior mean can be estimated by the average of the sampled draws, and the quantiles of the posterior density by the quantiles of the sampled output.

The requisite sampling of the target density is made possible by MCMC methods. In a MCMC simulation, one samples the target density in an indirect way: by simulating a suitably constructed Markov chain whose invariant distribution is the target density. Then the draws beyond some chosen burn-in period are taken as a (correlated) sample from the target density. The defining feature of Markov chains is the property that the conditional density of $\boldsymbol{\theta}^{(j)}$ (the jth element of the sequence) conditioned on the entire preceding history of the chain depends only on the previous value $\boldsymbol{\theta}^{(j-1)}$. Denote this conditional density, the transition density of the Markov chain, by $p(\boldsymbol{\theta}^{(j-1)}, \cdot|\mathbf{y})$. Then, in the MCMC framework, a sample is produced by simulating the transition density as

$$\boldsymbol{\theta}^{(1)} \sim p\left(\boldsymbol{\theta}^{(0)}, \cdot\,|\mathbf{y}\right) \vdots\, \boldsymbol{\theta}^{(j)} \sim p\left(\boldsymbol{\theta}^{(j-1)}, \cdot\,|\mathbf{y}\right) \vdots$$

If we let the first n_0 cycles represent the burn-in phase, for some choice of n_0, the draws

$$\boldsymbol{\theta}^{(n_0+1)}, \boldsymbol{\theta}^{(n_0+2)}, \ldots, \boldsymbol{\theta}^{(n_0+M)}$$

are treated as those from $\pi(\boldsymbol{\theta}|\mathbf{y})$. Even though the sampled variates are correlated, laws of large numbers for Markov sequences can be used to show that, under regularity conditions, the sample average of any integrable function $g(\boldsymbol{\theta})$ converges to its posterior expectation:

$$M^{-1}\sum_{j=1}^{M} g\left(\boldsymbol{\theta}^{(j)}\right) \to \int g(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}, \qquad (1)$$

as M becomes large.

There are two common ways of constructing a transition density $p(\boldsymbol{\theta}^{(j-1)}, \cdot|\mathbf{y})$ whose limiting distribution is the required target density. One way is by a method called the Metropolis–Hastings (M–H) algorithm, which was introduced by Metropolis et al. (1953) and Hastings (1970). Key references about this method are Tierney (1994), and Chib and Greenberg (1995). A second approach is by the so-called Gibbs sampling algorithm. This method was introduced by Geman and Geman (1984), Tanner and Wong (1987) and Gelfand and Smith (1990), and was the impetus for the current interest in Markov chain sampling methods. A summary of many aspects of MCMC methods is contained in Chib (2001) while textbook accounts include Gilks, Richardson and Spiegelhalter (1996), Chen, Shao and Ibrahim (2000), Liu (2001) and Robert and Casella (2004).

Metropolis–Hastings Algorithm

Suppose that we are interested in sampling the target density $\pi(\boldsymbol{\theta}|\mathbf{y})$, where $\boldsymbol{\theta}$ is a vector-valued parameter and $\pi(\boldsymbol{\theta}|\mathbf{y})$ is a continuous density. The idea behind the M–H algorithm is to simulate a proposal value $\boldsymbol{\theta}'$ from a transition density $q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\,\mathbf{y})$ that is convenient to stimulate but does not necessarily have the correct limiting distribution and then to subject the proposal value to a specific randomization to ensure that the resulting Markov chain has the correct limiting distribution.

To define the M–H algorithm, let $\boldsymbol{\theta}^{(j-1)}$ be the current value. Then the next value $\boldsymbol{\theta}^{(j)}$ is produced by a two-step process consisting of a 'proposal step' and a 'move step'.

- *Proposal step*: Sample a proposal value $\boldsymbol{\theta}'$ from $q(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}|\mathbf{y})$ and calculate the quantity

$$\alpha\left(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}' \,|\, \mathbf{y}\right) = \min\left\{1, \frac{\pi(\boldsymbol{\theta}' \,|\, \mathbf{y})}{\pi\left(\boldsymbol{\theta}^{(j-1)} \,|\, \mathbf{y}\right)} \frac{q\left(\boldsymbol{\theta}', \boldsymbol{\theta}^{(j-1)} \,|\, \mathbf{y}\right)}{q\left(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}' \,|\, \mathbf{y}\right)}\right\}. \quad (2)$$

- *Move step*: Let $\boldsymbol{\theta}^{(j)} = \boldsymbol{\theta}'$ with probability $\alpha(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}'|\mathbf{y})$; remain at the current value $\boldsymbol{\theta}^{(j-1)}$ with probability $1 - \alpha(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}'|\mathbf{y})$.

In terms of nomenclature, the source density $q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})$ is called the candidate generating density or proposal density, and $\alpha(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}'|\mathbf{y})$ the *acceptance probability* or, more descriptively, the *probability of move*. Note also that the function $\alpha(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}'|\mathbf{y})$ in this algorithm can be computed without knowledge of the norming constant of the posterior density $\pi(\boldsymbol{\theta}|\mathbf{y})$. In addition, if the proposal density is symmetric, satisfying the condition $q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) = q(\boldsymbol{\theta}', \boldsymbol{\theta}|\mathbf{y})$, then the acceptance probability reduces to $\pi(\boldsymbol{\theta}'|\mathbf{y})/\pi(\boldsymbol{\theta}^{(j-1)}|\mathbf{y})$; hence, if $\pi(\boldsymbol{\theta}') \geq \pi(\boldsymbol{\theta}^{(j-1)}|\mathbf{y})$, the chain moves to $\boldsymbol{\theta}'$, otherwise it moves to $\boldsymbol{\theta}'$ with probability given by $\pi(\boldsymbol{\theta}'|\mathbf{y})/\pi(\boldsymbol{\theta}^{(j-1)}|\mathbf{y})$. The latter is the algorithm of Metropolis et al. (1953).

Remark 1: Derivation of the M–H algorithm A question of some interest is the justification of this two-step approach. This question was tackled by Chib and Greenberg (1995) who derived the method from the logic of reversibility. A Markov transition density $p(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})$ is said to be reversible for $\pi(\boldsymbol{\theta}|\mathbf{y})$ if the following condition holds for every $(\boldsymbol{\theta}, \boldsymbol{\theta}')$ in the support of the target distribution:

$$\pi(\boldsymbol{\theta}|\mathbf{y})p(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) = \pi(\boldsymbol{\theta}'|\mathbf{y})p(\boldsymbol{\theta}', \boldsymbol{\theta}|\mathbf{y}). \quad (3)$$

The reversibility condition is important because reversible chains are invariant. Invariance refers to the property that

$$\pi(\boldsymbol{\theta}'|\mathbf{y}) = \int p(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})\pi(\boldsymbol{\theta}|\mathbf{y})\, d\boldsymbol{\theta} \quad (4)$$

which means that, if the transition density is invariant for the target density, then, once convergence is achieved, a subsequent value $\boldsymbol{\theta}'$ drawn from the transition density is also from the target density. To see that reversibility implies invariance one simply integrates both sides of Eq. (3) over $\boldsymbol{\theta}$. This leads to the invariance condition since $\int p(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})\, d\boldsymbol{\theta}' = 1$ by virtue of being a transition density. Now consider the Markov chain induced by the proposal density $q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})$. Because this was formulated without the reversibility condition in mind it is unlikely to satisfy reversibility. Suppose that for a pair of points $(\boldsymbol{\theta}, \boldsymbol{\theta}')$ it is true that

$$\pi(\boldsymbol{\theta}|\mathbf{y})q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) > \pi(\boldsymbol{\theta}'|\mathbf{y})q(\boldsymbol{\theta}', \boldsymbol{\theta}|\mathbf{y}), \quad (5)$$

which means informally that the process moves from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$ too frequently and too rarely in the reverse direction. This situation can be corrected by reducing the flow from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$ by introducing probabilities $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})$ and $\alpha(\boldsymbol{\theta}', \boldsymbol{\theta}|\mathbf{y})$ of making the moves in either direction so that

$$\pi(\boldsymbol{\theta}|\mathbf{y})q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) = \pi(\boldsymbol{\theta}'|\mathbf{y})q(\boldsymbol{\theta}', \boldsymbol{\theta}|\mathbf{y})\alpha(\boldsymbol{\theta}', \boldsymbol{\theta}|\mathbf{y}).$$

One now sets $\alpha(\boldsymbol{\theta}', \boldsymbol{\theta}|\mathbf{y})$ to be as high as possible, namely, equal to 1. Solving for $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})$ one then gets

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) = \frac{\pi(\boldsymbol{\theta}'|\mathbf{y})}{\pi(\boldsymbol{\theta}|\mathbf{y})} \frac{q(\boldsymbol{\theta}', \boldsymbol{\theta}|\mathbf{y})}{q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})}.$$

Because one started from Eq. (5) this is clearly less than 1. On the other hand, if the inequality in Eq. (5) were reversed, the same argumentation leads to the conclusion that $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) = 1$. Thus, on combining these two cases we reproduce the expression of $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})$ given in Eq. (2).

Remark 2: Transition density of the M–H chain The transition density of the M–H chain has two components – one for the move away from $\boldsymbol{\theta}$ and given by $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})$ and one for the probability of staying at $\boldsymbol{\theta}$ given by $r(\boldsymbol{\theta}|\mathbf{y}) = 1 - \int \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})\, d\boldsymbol{\theta}'$. In particular,

$$p_{MH}(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) = \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) + \delta_\theta(\boldsymbol{\theta}') r(\boldsymbol{\theta}|\mathbf{y})$$

where $\delta_\theta(\boldsymbol{\theta}')$ is the Dirac-function at $\boldsymbol{\theta}$ defined as $\delta_\theta(\boldsymbol{\theta}')= 0$ for $\boldsymbol{\theta}' \neq \boldsymbol{\theta}$ and $\int \delta_\theta(\boldsymbol{\theta}')\, d\boldsymbol{\theta}' = 1$. It is easy to check that the integral of the transition density over all possible values of $\boldsymbol{\theta}$ is 1, as required.

Remark 3: Convergence properties The theoretical properties of the M–H algorithm (in particular the ergodic behaviour of the chain from an arbitrarily specified initial value) depend crucially on the nature of the proposal density. One requirement is that the proposal density be everywhere positive in the support of the posterior density, which means that the M–H chain can make a transition to any point in its support in one step. Further discussion of the conditions is given in Tierney (1994) and Robert and Casella (2004).

Remark 4: Mixing The sampled values from the M–H algorithm (as from any Markov chain) are correlated. The goal in any particular application is to ensure that the serial correlation is not excessive. One diagnostic to check for the degree of serial correlation in the sampled draws is the *autocorrelation time* or *inefficiency factor* of each component $\boldsymbol{\theta}_k$ of $\boldsymbol{\theta}$ defined as

$$a_k = \left\{ 1 + 2\sum_{s=1}^{M} (1 - \frac{s}{M}) \rho_{k_s} \right\},$$

where ρ_{ks} is the sample autocorrelation at lag s from the M sampled draws $\theta_k^{(n_0+1)}, \ldots, \theta_k^{(n_0+M)}$. One can interpret this quantity in terms of the *effective sample size*, or ESS, defined for the kth component of $\boldsymbol{\theta}$ as $ESS_k = \frac{M}{a_k}$. With independent sampling the autocorrelation times are theoretically equal to 1, and the effective sample size is M. When the inefficiency factors are high, the effective sample size is much smaller than M.

Choice of Proposal Density

One family of candidate-generating densities is given by $q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})= q(\boldsymbol{\theta}' - \boldsymbol{\theta})$. The candidate $\boldsymbol{\theta}'$ is thus drawn according to the process $\boldsymbol{\theta}' = \boldsymbol{\theta} + \mathbf{z}$, where $\mathbf{z}$ follows the distribution q, and is called the *random walk M–H* chain. The random walk M–H chain is quite popular in applications. One has to be careful in setting the variance of $\mathbf{z}$ because if it is too large the chain may remain stuck at a particular value for many iterations, while if it is too small the chain will tend to make small moves and move inefficiently through the support of the target distribution.

Another possibility is to let $q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})= q(\boldsymbol{\theta}'|\mathbf{y})$, an *independence M–H chain* in the terminology of Tierney (1994). One way to implement such chains is by tailoring the proposal density to the target at the mode by a multi-variate normal or multivariate-t distribution with location given by the mode of the target and the dispersion given by inverse of the Hessian evaluated at the mode (Chib and Greenberg 1994, 1995).

Yet another way to generate proposal values is through a Markov chain version of the accept–reject method (Tierney 1994; Chib and Greenberg 1995). To explain this method, suppose $c > 0$ is a known constant and $h(\boldsymbol{\theta})$ a source density. Let $C = \{\boldsymbol{\theta} : \pi(\boldsymbol{\theta}|\mathbf{y}) \leq ch(\boldsymbol{\theta})\}$ denote the set of value for which $ch(\boldsymbol{\theta})$ dominates the target density. Given $\boldsymbol{\theta}^{(j-1)} = \boldsymbol{\theta}$ the next value $\boldsymbol{\theta}^{(j)}$ is obtained as follows. First, a candidate value $\boldsymbol{\theta}'$ is obtained, independent of the current value $\boldsymbol{\theta}$, by applying the accept–reject algorithm with $ch(\cdot)$ as the 'pseudo-dominating' density. The candidates $\boldsymbol{\theta}'$ that are produced under this scheme have density $q(\boldsymbol{\theta}'|\mathbf{y}) \propto \min\{\pi(\boldsymbol{\theta}'|\mathbf{y}); ch(\boldsymbol{\theta}')\}$. Then, the M–H probability of move is given by

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) = \begin{cases} 1 & \text{if} \quad \boldsymbol{\theta} \in C \\ 1/w(\boldsymbol{\theta}) & \text{if} \quad \boldsymbol{\theta} \notin C, \boldsymbol{\theta}' \in C \\ \min\{w(\boldsymbol{\theta}')/w(\boldsymbol{\theta}), 1\} & \text{if} \quad \boldsymbol{\theta} \notin C, \boldsymbol{\theta}' \notin C \end{cases} \tag{6}$$

where $w(\boldsymbol{\theta})= c^{-1}\pi(\boldsymbol{\theta}|\mathbf{y})/h(\boldsymbol{\theta})$.

Example

To illustrate the M–H algorithm, consider the binary response data in Table 1, on the occurrence or non-occurrence of infection following birth by Caesarean section. The response variable y is 1 if the Caesarean birth resulted in an infection, and

Markov Chain Monte Carlo Methods, Table 1 Caesarean infection data

y(1/0)	$x1$	$x2$	$x3$
11/87	1	1	1
1/17	0	1	1
0/2	0	0	1
23/3	1	1	0
28/30	0	1	0
0/9	1	0	0
8/32	0	0	0

Source: Fahrmeir and Tutz (1994)

zero if not. There are three covariates: x_1, an indicator of whether the caesarean was non-planned; x_2, an indicator of whether risk factors were present at the time of birth; and x_3, an indicator of whether antibiotics were given as a prophylaxis. The data in the table contains information from 251 births. Under the column of the response, an entry such as 11/87 means that there were 98 deliveries with covariates (1,1,1) of whom 11 developed an infection and 87 did not. Suppose that the probability of infection for the ith birth ($i \leq 251$) is

$$\Pr(y_i = 1 | \mathbf{x}_i, \beta) = \Phi(\mathbf{x}_i'\beta), \tag{7}$$

$$\beta \sim N_4(0, 5\mathbf{I}_4) \tag{8}$$

where $\mathbf{x}_i = (1, x_{i1}, x_{i2}, x_{i3})^{\mathrm{T}}$ is the covariate vector, $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$ is the vector of unknown coefficients, Φ is the cdf of the standard normal random variable and $\mathbf{I}_4$ is the four-dimensional identity matrix. The target posterior density, under the assumption that the outcomes $\mathbf{y} = (y_1, y_2, ..., y_{251})$ are conditionally independent, is

$$\pi(\beta | \mathbf{y}) \propto \pi(\beta) \prod_{i=1}^{251} \Phi(\mathbf{x}_i'\beta)^{y_i} \{1 - \Phi(\mathbf{x}_i'\beta)\}^{(1-y_i)}$$

where $\pi(\beta)$ is the density of the $N(0,10\mathbf{I}_4)$ distribution.

To define the Chib and Greenberg (1994) tailored proposal density, let

$$\hat{\beta} = (-1.093022\ 0.607643\ 1.197543\ -1.904739)'$$

be the maximum likelihood estimate and let

$$\mathbf{v} = \begin{pmatrix} 0.040745 & -0.007038 & -0.039399 & 0.004829 \\ & 0.073101 & -0.006940 & -0.050162 \\ & & 0.062292 & -0.016803 \\ & & & 0.080788 \end{pmatrix}$$

be the symmetric matrix obtained by inverting the negative of the Hessian matrix (the matrix of second derivatives) of the log-likelihood function evaluated at $\hat{\beta}$. To generate proposal values, we use a multivariate-t density with 15 degrees of freedom, location given by $\hat{\beta}$ and dispersion given by $\mathbf{V}$. The M–H algorithm is run for 5000 iterations beyond a burn-in of 100 iterations. The prior–posterior summary is reported in Table 2. It contains the first two moments (the mean and the standard deviation) of the prior and posterior and the 2.5th (lower) and 97.5th (upper) percentiles of the marginal densities of β.

In addition, we plot in Fig. 1 the four marginal posterior densities. These are derived by smoothing the histogram of the simulated values with a Gaussian kernel. In the same plot we also report the autocorrelation functions (correlation against lag) for each of the sampled parameter values. The serial correlations decline quickly to zero indicating that the algorithm is mixing well.

Multiple-Block M–H Algorithm

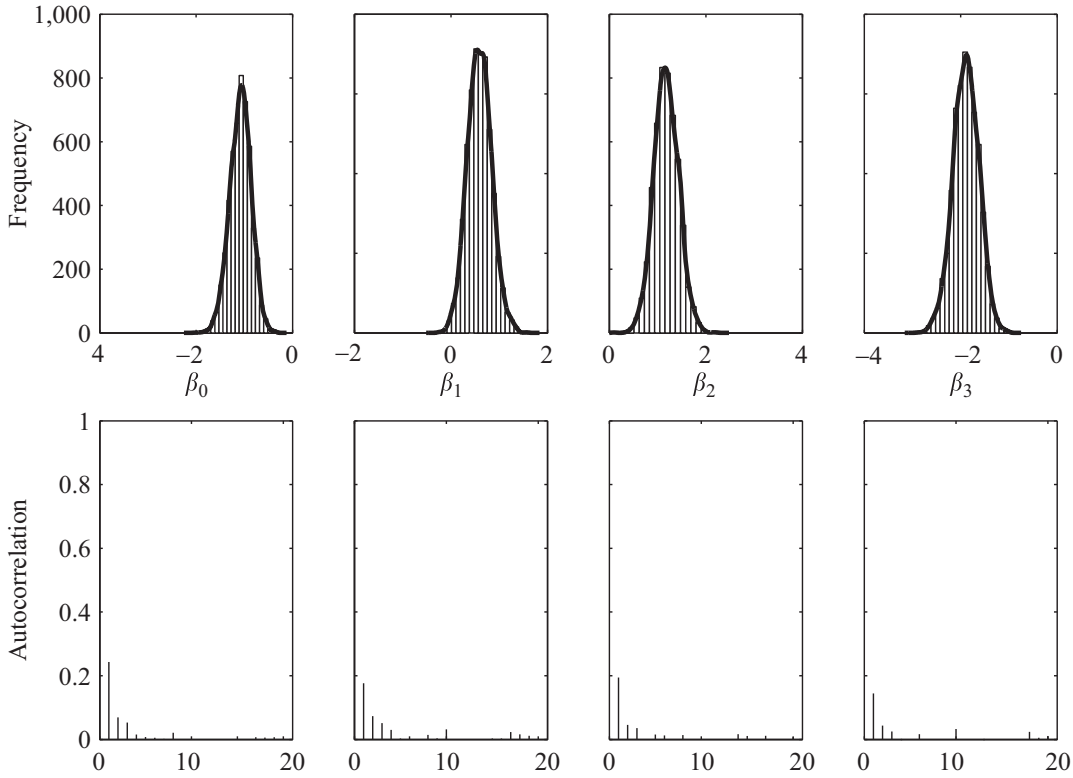
When the dimension of $\boldsymbol{\theta}$ is large it is often necessary to divide the parameters into smaller groups or blocks and then to sample the blocks in turn. For simplicity suppose that two blocks are adequate and that $\boldsymbol{\theta}$ is written as $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, with $\boldsymbol{\theta}_k \in \Omega_k \subseteq \mathcal{R}^{d_k}$. To sample these blocks let

$$q_1(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1' | \mathbf{y}, \boldsymbol{\theta}_2); q_2(\boldsymbol{\theta}_2, \boldsymbol{\theta}_2' | \mathbf{y}, \boldsymbol{\theta}_1),$$

denote the two proposal densities, one for each block $\boldsymbol{\theta}_k$, where the proposal density q_k may depend on the current value of the remaining block. Also, define

Markov Chain Monte Carlo Methods, Table 2 Caesarean data: prior–posterior summary based on 5000 draws (beyond a burn-in of 100 cycles) from the tailored M–H algorithm

	Prior		Posterior			
	Mean	Std dev	Mean	Std dev	Lower	Upper
β_0	0.000	2.236	− 1.080	0.220	− 1.526	− 0.670
β_1	0.000	2.236	0.593	0.249	0.116	1.095
β_2	0.000	2.236	1.181	0.254	0.680	1.694
β_3	0.000	2.236	− 1.889	0.266	− 2.421	− 1.385



Markov Chain Monte Carlo Methods, Fig. 1 Caesarean data with tailored M–H algorithm: marginal posterior densities (*top panel*) and autocorrelation plot (*bottom panel*)

$$\alpha(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1' | \mathbf{y}, \boldsymbol{\theta}_2) = \min\left\{\frac{\pi(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2 | \mathbf{y}) q_1(\boldsymbol{\theta}_1', \boldsymbol{\theta}_1 | \mathbf{y}, \boldsymbol{\theta}_2)}{\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{y}) q_1(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1' | \mathbf{y}, \boldsymbol{\theta}_2)}, 1\right\}$$

and

$$\alpha(\boldsymbol{\theta}_2, \boldsymbol{\theta}_2' | \mathbf{y}, \boldsymbol{\theta}_1) = \min\left\{\frac{\pi(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2 | \mathbf{y}) q_2(\boldsymbol{\theta}_2', \boldsymbol{\theta}_2 | \mathbf{y}, \boldsymbol{\theta}_1)}{\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{y}) q_2(\boldsymbol{\theta}_2, \boldsymbol{\theta}_2' | \mathbf{y}, \boldsymbol{\theta}_1)}, 1\right\}$$

as the probability of move for block $\boldsymbol{\theta}_k$ conditioned on the other block. Then, in what may be called the multiple-block M–H algorithm, one updates each block using an M-H step with the above probability of move, given the most current

value of the other block. The method can be extended to several blocks in the same way.

Remark 5 An important special case arises if each proposal density is the full conditional density of that block. Specifically, if we set

$$q_1(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1' | \mathbf{y}, \boldsymbol{\theta}_2) \propto \pi(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2 | \mathbf{y}),$$
$$q_1(\boldsymbol{\theta}_1', \boldsymbol{\theta}_1 | \mathbf{y}, \boldsymbol{\theta}_2) \propto \pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{y})$$

and

$$q_2(\boldsymbol{\theta}_2, \boldsymbol{\theta}_2' | \mathbf{y}, \boldsymbol{\theta}_1) \propto \pi(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2 | \mathbf{y}),$$
$$q_2(\boldsymbol{\theta}_2', \boldsymbol{\theta}_2 | \mathbf{y}, \boldsymbol{\theta}_1) \propto \pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{y})$$

then an interesting simplification occurs. The probability of move (for the first block) becomes

$$\alpha_1(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1' | \mathbf{y}, \boldsymbol{\theta}_2) = \min\left\{1, \frac{\pi(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2 | \mathbf{y})\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{y})}{\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{y})\pi(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2 | \mathbf{y})}\right\} = 1,$$

and similarly for the second block, implying that, if proposal values are drawn from their full conditional densities, then the proposal values are accepted with probability one. This special case is called the Gibbs sampling algorithm.

The Gibbs Sampling Algorithm

The Gibbs sampling was introduced by Geman and Geman (1984) in the context of image processing and then discussed in the context of missing data problems by Tanner and Wong (1987). It was brought into prominence by Gelfand and Smith (1990) who demonstrated its use in a range of Bayesian problems.

The Algorithm

Suppose that the parameters are grouped into two p blocks ($\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$,..., $\boldsymbol{\theta}_p$) with the associated set of full conditional distributions

$$\{\pi(\boldsymbol{\theta}_1 | \mathbf{y}, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_p); \pi(\boldsymbol{\theta}_2 | \mathbf{y}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_3, \ldots, \boldsymbol{\theta}_p); \ldots \pi(\boldsymbol{\theta}_p | \mathbf{y}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{d-1})\},$$

where each full conditional distribution is proportional to $\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2,\ldots, \boldsymbol{\theta}_p)$ |y). Then, one cycle of the Gibbs sampling algorithm is completed by simulating $\{\boldsymbol{\theta}_k\}_{k=1}^p$ from these distributions, recursively refreshing the conditioning variables.

Sufficient Conditions for Convergence

Under rather general conditions, the Markov chain generated by the Gibbs sampling algorithm converges to the target density as the number of iterations become large. Formally, if we let $p_G(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})$ represent the transition density of the Gibbs algorithm and let $p_G^{(M)}(\boldsymbol{\theta}', \boldsymbol{\theta}|\mathbf{y})$ be the density of the draw $\boldsymbol{\theta}'$ after M iterations given the starting value $\boldsymbol{\theta}_0$, then

$$\left\| p_G^{(M)}(\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}' | \mathbf{y} - \pi(\boldsymbol{\theta}' | \mathbf{y}) \right\| \to 0, \quad \text{as } M \to \infty. \tag{9}$$

Roberts and Smith (1994) (see also Chan 1993) have shown that this convergence occurs under the following weak conditions: (i) $\pi(\boldsymbol{\theta}|\mathbf{y}) > 0$ implies there exists an open neighbourhood N_θ containing $\boldsymbol{\theta}$ and $\varepsilon > 0$ such that, for all $\boldsymbol{\theta}' \in N_\theta$, $\pi(\boldsymbol{\theta}'|\mathbf{y}) \geq \varepsilon > 0$; (ii) $\int \pi(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}_k$ is locally bounded for all k, where $\boldsymbol{\theta}_k$ is the kth block of parameters; and (iii) the support of $\boldsymbol{\theta}$ is arc connected.

MCMC Sampling with Latent Variables

MCMC sampling can involve not just parameters but also latent variables. This idea was called data augmentation by Tanner and Wong (1987) in the context of missing data problems.

To fix notations, suppose that $\mathbf{z}$ denotes a vector of latent variables and let the modified target distribution be $\pi(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})$. If the latent variables are tactically introduced, the conditional distribution of $\boldsymbol{\theta}$ (or sub-components of $\boldsymbol{\theta}$) given $\mathbf{z}$ may be easy to derive. Then, a multiple-block M–H simulation is conducted with the blocks $\boldsymbol{\theta}$ and $\mathbf{z}$ leading to the sample

$$\left(\boldsymbol{\theta}^{(n_0+1)}, \mathbf{z}^{(n_0+1)}\right), \ldots, \left(\boldsymbol{\theta}^{(n_0+M)}, \mathbf{z}^{(n_0+M)}\right) \sim \pi(\boldsymbol{\theta}, \mathbf{z}|\,\mathbf{y}),$$

where the draws on $\boldsymbol{\theta}$, ignoring those on the latent data, are from $\pi(\boldsymbol{\theta}|\mathbf{y})$, as required.

To demonstrate this technique in action, consider the probit regression example discussed in section "Example". Albert and Chib (1993) introduced a technique for this and related models that capitalizes on the simplifications afforded by introducing latent data into the sampling. The Albert–Chib method has found wide use and has made possible the routine analysis of models for categorical responses. To begin, let

$$z_i|\beta \sim N\left(\mathbf{x}_i'\beta, 1\right),$$

$$y_i = I[z_i > 0], \quad i \le n,$$

$$\beta \sim N_k(\beta_0, \mathbf{B}_0). \tag{10}$$

This specification is equivalent to the probit model since $\Pr(y_i = 1|\,\mathbf{x}_i, \beta) = \Pr(z_i > 0|\,\mathbf{x}_i, \beta) = \Phi\left(\mathbf{x}_i'\beta\right)$. Now the MCMC sampling is based on the full conditional distributions

$$\beta|\mathbf{y}, \{z_i\}; \quad \{z_i\}|\mathbf{y}, \beta,$$

which are both tractable. In particular, the distribution of β conditioned on the latent data becomes independent of the observed data and has the same form as in the Gaussian linear regression model with the response data given by $\{z_i\}$ and is multivariate normal with mean $\hat{\beta} = \mathbf{B}\left(\mathbf{B}_0^{-1}\beta_0 + \sum_{i=1}^n \mathbf{x}_i z_i\right)$ and variance matrix $\mathbf{B} = \left(\mathbf{B}_0^{-1} + \sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'\right)^{-1}$. Next, the distribution of the latent data conditioned on the data and the parameters factor into a set of n independent distributions with each depending on the data through y_i:

$$\{z_i\}|\mathbf{y}, \beta \overset{d}{=} \prod_{i=1}^n z_i|y_i, \beta,$$

where the distribution $z_i|\mathbf{y}_i$, β is the normal distribution $z_i|\beta$ truncated by the knowledge of y_i; if $y_i = 0$, then $z_i \le 0$ and if $y_i = 1$, then $z_i > 0$. Thus, one samples z_i from $TN_{(-\infty,0)}\left(\mathbf{x}_i'\beta, 1\right)$ if $y_i = 0$ and from $TN_{(0,\infty)}\left(\mathbf{x}_i'\beta, 1\right)$ if $y_i = 1$, where $TN_{(a;b)}(\mu, \sigma^2)$ denotes the $N(\mu, \sigma^2)$ distribution truncated to the region (a,b).

We apply this method to the example considered in section "Example" above and report the results in Fig. 2. We see the close agreement between the two sets of results.

Strategies for Improving Mixing

In practice, while implementing MCMC methods it is important to construct samplers that mix well, where mixing is measured by the autocorrelation time, because such samplers can be expected to converge more quickly to the invariant distribution.

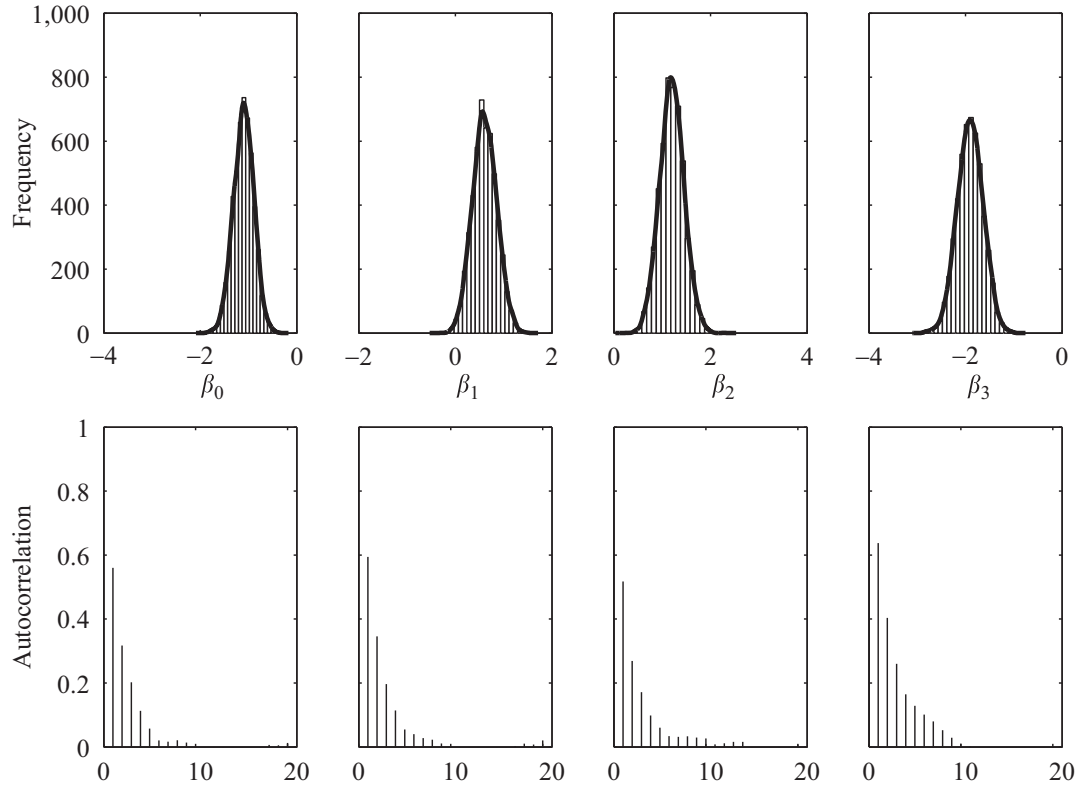
Choice of Blocking

As a general rule, sets of parameters that are highly correlated should be treated as one block when applying the multiple-block M–H algorithm. Otherwise, it would be difficult to develop proposal densities that lead to large moves through the support of the target distribution.

Blocks can be combined by the method of composition. For example, suppose that $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$ denote three blocks and that the distribution $\boldsymbol{\theta}_1|\mathbf{y}$, $\boldsymbol{\theta}_3$ is tractable (that is, can be sampled directly). Then, the blocks $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ can be collapsed by first sampling, $\boldsymbol{\theta}_1$ from $\boldsymbol{\theta}_1|\mathbf{y}, \boldsymbol{\theta}_3$ followed by $\boldsymbol{\theta}_2$ from $\boldsymbol{\theta}_2|\mathbf{y}$, $\boldsymbol{\theta}_1$; $\boldsymbol{\theta}_3$. This amounts to a two-block MCMC algorithm. In addition, if it is possible to sample $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ marginalized over $\boldsymbol{\theta}_3$ then the number of blocks is reduced to one. Liu (1994) discusses the value of these strategies in the context of a three-block Gibbs MCMC chain. Roberts and Sahu (1997) provide further discussion of the role of blocking in the context of Gibbs Markov chains used to sample multivariate normal target distributions.

Tuning the Proposal Density

The proposal density in an M–H algorithm has an important bearing on the mixing of the MCMC chain. Chib and Greenberg (1994, 1995), Tierney



Markov Chain Monte Carlo Methods, Fig. 2 Caesarean data with Albert–Chib algorithm: marginal posterior densities (*top panel*) and autocorrelation plot (*bottom panel*)

(1994), Tierney and Mira (1999) and Liu (2001) discuss various possibilities for formulating proposal density that can be helpful in a variety of problems.

Prediction and Model Choice

In some settings, for example in models for time series data, an important goal is prediction. In the Bayesian context, a future observation y_f is predicted through the (predictive) density defined as

$$f\left(y_f|\mathbf{y}\right) = \int f(y_f|\mathbf{y},\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\,\mathbf{y})\,d\boldsymbol{\theta},$$

where $f\left(y_f|\,\mathbf{y},\mathcal{M},\,\boldsymbol{\theta}\right)$ is the conditional density of y_f given (y, $\boldsymbol{\theta}$). In general, the predictive density is not available in closed form. It can be shown, however, that, if one simulates $y_f^{(j)} \sim f\left(y_f|\mathbf{y},\boldsymbol{\theta}^{(j)}\right)$ for each sampled draw $\boldsymbol{\theta}^{(j)}$ from the MCMC simulation, then the collection of simulated values $\left\{y_f^{(1)},\ldots,y_f^{(M)}\right\}$ is a sample from $f\,(y_f|\mathbf{y})$. This simulated sample can be summarized in the usual way.

MCMC methods have also been widely applied to the problem of the model choice. Suppose that there are K possible models $\mathcal{M}_1,\ldots,\mathcal{M}_K$ for the observed data defined by the sampling densities $\{f(\mathbf{y}|\,\boldsymbol{\theta}_k,\mathcal{M}_k)\}$ and proper prior densities $\{p(\boldsymbol{\theta}_k|\mathcal{M}_k\}$ and the objective is to find the evidence in the data for the different models. In the Bayesian approach this question is answered by placing prior probabilities $\Pr(\mathcal{M}_k)$ on each of the K models and using the Bayes calculus to find the posterior probabilities $\{\Pr(\mathcal{M}_1|\,\mathbf{y}),\ldots,\Pr(\mathcal{M}_K|\mathbf{y}\}$ conditioned on the data but marginalized over the unknowns $\boldsymbol{\theta}_k$

(Jeffreys 1961). Specifically, the posterior probability of $\mathcal{M}_k$ is given by the expression

$$\Pr(\mathcal{M}_k|\,\mathbf{y}) = \frac{\Pr(\mathcal{M}_k)m(\mathbf{y}|\,\mathcal{M}_k)}{\sum_{l=1}^{K}\Pr(\mathcal{M}_l)m(\mathbf{y}|\,\mathcal{M}_l)} \propto \Pr(\mathcal{M}_k)m(\mathbf{y}|\,\mathcal{M}_k), \quad (k \leq K)$$

where $m(\mathbf{y}|\,\mathcal{M}_k)$ is the marginal likelihood of $\mathcal{M}_k$.

A problem in estimating the marginal likelihood is that it is an integral of the sampling density over the prior distribution of $\boldsymbol{\theta}_k$. Thus, MCMC methods, which deliver sample values from the posterior density, cannot be used to directly average the sampling density. One method for dealing with this difficulty is due to Chib (1995). The starting point is the expression

$$m(\mathbf{y}|\,\mathcal{M}_k) = \frac{f(\mathbf{y}|\,\boldsymbol{\theta}_k, \mathcal{M}_k)p(\boldsymbol{\theta}_k|\,\mathcal{M}_k)}{\pi(\boldsymbol{\theta}_k|\,\mathbf{y}, \mathcal{M}_k)}$$

which is an identity in $\boldsymbol{\theta}_k$. From here an estimate of the marginal likelihood on the log-scale is given by

$$\log\hat{m}(\mathbf{y}|\,\mathcal{M}_k) = \log f\left(\mathbf{y}|\,\boldsymbol{\theta}_k^*, \mathcal{M}_k\right) + \log p\left(\boldsymbol{\theta}_k^*|\,\mathcal{M}_k\right) - \log\hat{\pi}\left(\boldsymbol{\theta}_k^*|\,\mathbf{y}, \mathcal{M}_k\right)$$

where $\boldsymbol{\theta}_k^*$ denotes an arbitrarily chosen point and $\hat{\pi}\left(\boldsymbol{\theta}_k^*|\,\mathbf{y}, \mathcal{M}_k\right)$ is the estimate of the posterior density at that single point. To estimate the posterior ordinate one utilizes the Gibbs output in conjunction with a decomposition of the ordinate into marginal and conditional components. Chib and Jeliazkov (2001) extend this approach for output produced by the M–H algorithm while Basu and Chib (2003) show how the method can be applied in semiparametric models.

In some cases one is interested in a large number of candidate models, each with parameters $\boldsymbol{\theta}_k \in B_k \subseteq R^{d_k}$. In such cases one can get information about the relative support for the contending models from a model space-parameter space MCMC algorithm. In these algorithms, the models are represented by a categorical variable $\mathcal{M}$ which is then sampled along with the parameters of each model. The posterior distribution of $\mathcal{M}$ is computed as the frequency of times each model is visited. Methods for doing this have been proposed by Carlin and Chib (1995) and Green (1995). Both methods are closely related as shown by Dellaportas et al. (2002) and Godsill (2001). Related methods for the problem of variable selection have also been developed starting with George and McCulloch (1993).

See Also

- Bayesian Econometrics
- Bayesian Statistics
- Econometrics
- Hierarchical Bayes Models
- Simulation-Based Estimation

Bibliography

Albert, J.H., and S. Chib. 1993. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* 88: 669–679.

Basu, S., and S. Chib. 2003. Marginal likelihood and Bayes factors for Dirichlet process mixture models. *Journal of the American Statistical Association* 98: 224–235.

Carlin, B.P., and S. Chib. 1995. Bayesian model choice via Markov chain Monte Carlo. *Journal of the Royal Statistical Society, Series B* 57: 473–484.

Chan, K.S. 1993. Asymptotic behavior of the Gibbs sampler. *Journal of the American Statistical Association* 88: 320–326.

Chen, M.H., Q.M. Shao, and J.G. Ibrahim. 2000. *Monte Carlo methods in Bayesian computation*. New York: Springer.

Chib, S. 1995. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* 90: 1313–1321.

Chib, S. 2001. Markov chain Monte Carlo methods: Computation and inference. In *Handbook of econometrics*, ed. J.J. Heckman and E. Leamer, Vol. 5. Amsterdam: North-Holland.

Chib, S., and E. Greenberg. 1994. Bayes inference in regression models with ARMA (p,q) errors. *Journal of Econometrics* 64: 183–206.

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

Chib, S., and I. Jeliazkov. 2001. Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association* 96: 270–281.

Dellaportas, P., J.J. Forster, and I. Ntzoufras. 2002. On Bayesian model and variable selection using MCMC. *Statistics and Computing* 12: 27–36.

Fahrmeir, L., and G. Tutz. 1994. *Multivariate statistical modelling based on generalized linear models*. Berlin: Springer.

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

Geman, S., and D. Geman. 1984. Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Transactions, PAMI* 6: 721–741.

George, E.I., and R.E. McCulloch. 1993. Variable selection via Gibbs sampling. *Journal of the American Statistical Association* 88: 881–889.

Gilks, W.K., S. Richardson, and D.J. Spiegelhalter. 1996. *Markov chain Monte Carlo in practice*. London: Chapman & Hall.

Godsill, S.J. 2001. On the relationship between Markov chain Monte Carlo methods for model uncertainty. *Journal of Computational and Graphical Statistics* 10: 230–248.

Green, P.J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82: 711–732.

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

Jeffreys, H. 1961. *Theory of Probability.* 3rd edn. Oxford: Oxford University Press.

Liu, J.S. 1994. The collapsed Gibbs sampler in Bayesian computations with applications to a gene-regulation problem. *Journal of the American Statistical Association* 89: 958–966.

Liu, J.S. 2001. *Monte Carlo strategies in scientific computing*. New York: Springer.

Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, et al. 1953. Equations of state calculations by fast computing machines. *Journal of Chemical Physics* 21: 1087–1092.

Robert, C.P., and G. Casella. 2004. *Monte Carlo statistical methods*. 2nd edn. New York: Springer.

Roberts, G.O., and S.K. Sahu. 1997. Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society, Series B* 59: 291–317.

Roberts, G.O., and A.F.M. Smith. 1994. Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stochastic Processes and their Applications* 49: 207–216.

Tanner, M.A., and W.H. Wong. 1987. The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association* 82: 528–550.

Tierney, L. 1994. Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics* 21: 1701–1762.

Tierney, L., and A. Mira. 1999. Some adaptive Monte Carlo methods for Bayesian inference. *Statistics in Medicine* 18: 2507–2515.