

Reversible Jump MCMC for Model Selection

Bayesian and hierarchical modeling discussion group

February 5, 2003

1 Introduction

Methodologies for model selection using MCMC have been available since 1995. These techniques adopt a strategy akin to that of so-called switch-point analyses, such that all models in a set M under consideration are given a unique index $m \in M$ and then treating this index itself as a model parameter to be estimated. Model choice is then based upon the posterior weight of the model indices, rather than on likelihood ratios or information-theoretic criterion-based methods, such as AIC. The most obvious advantage of the MCMC approach is that, in general, it is capable of sampling from arbitrarily complex posterior distributions; specifically, this means that the model set may be very large without concern for computational limitations (*e.g.* King and Brooks, 2002)

The first approach was formulated by introduced by Carlin and Chib (1995), which estimates all possible parameters for a given set of models simultaneously, by sampling from a large joint posterior, or "supermodel":

$$p(\mathbf{y}, \theta_m, m) = p(\mathbf{y}|\theta_m, m)p(\theta_m|m)p_m$$

where y is observed data and θ_m is the parameter set corresponding to model m . Inferences about individual models of interest are gained by considering the conditional posterior distribution of parameter subsets. This approach is hampered by a number of practical and theoretical difficulties (Gamerman, 1997), and therefore will not be the focus of today's discussion.

In contrast, Green (1995) offers an alternate approach, reversible jump MCMC, that essentially generalizes the Metropolis-Hastings algorithm. Rather than simply exploring the posterior distribution of a single model, RJMCMC moves *between* models as well as within a given model. This approach is used effectively by King and Brooks (2002) to estimate the best model for a combined recovery/recapture data set, and will be explored in some detail here.

2 Review of Key Concepts

Before proceeding, I would like to reinforce/introduce some general concepts that are germane to this discussion:

2.1 Rejection Sampling

Rejection sampling techniques are those which generate samples of random variables in more than one step. It is used for obtaining random deviates from densities that are complex, or otherwise do not yield samples easily. The first step is the generation of a sample from a more elementary approximating distribution (proposal distribution). The second step is an evaluation of the sample based upon some criterion (acceptance function) related to the density of interest. Remarkably, this technique can be used to generate values from even partially-specified densities.

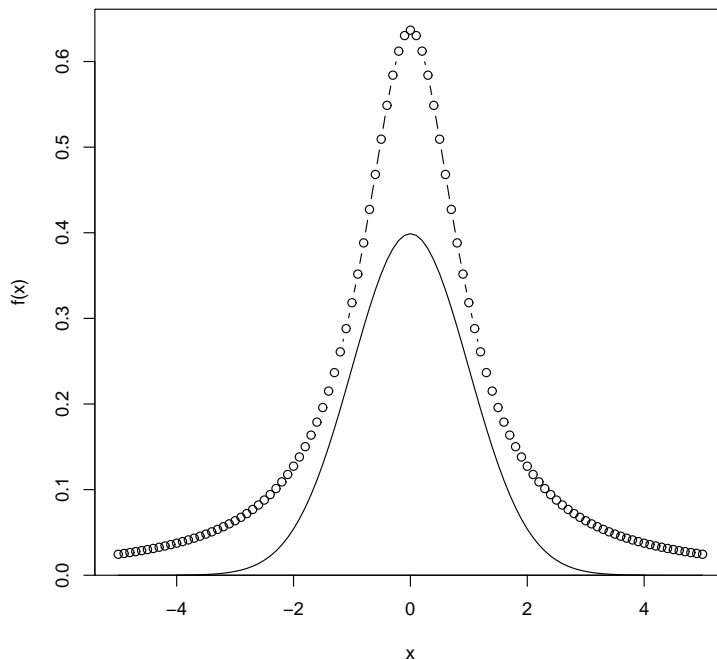


Figure 1: Enveloping a Normal distribution with a Cauchy distribution. The latter was scaled to ensure a complete envelope.

Consider a density of interest, π , that will be approximated by the proposal distribution q . Rejection sampling only requires that there exist some scalar, A , such that $\pi(x) \leq Aq(x)$ for every $x \in (-\infty, \infty)$; Aq is thus an envelope of π , and that π can be evaluated for every $x \in (-\infty, \infty)$. Provided this, one can draw N values from π as follows:

1. Set $n = 0$

2. While $N \leq n$:

- (a) Draw $x \sim q(\theta)$ and $u \sim U[0, 1]$
- (b) Accept x if $u \leq \pi(x)/Aq(x)$, reject otherwise
- (c) If x is accepted, increment n

It can be shown that this procedure effectively produces a sample of values from π (Gamerman, 1997). This rejection method is only one of many in a suite of resampling techniques, including weighted resampling and adaptive rejection, but it is the simplest and is easy to incorporate into reversible jump MCMC.

2.2 Reversible Markov Chains

Consider a homogeneous Markov chain $\{\theta^{(i)}\}_{i \geq 0}$ that is homogeneous (transition probabilities $p(x, y)$ are independent of n). One may be interested in studying this sequence of states in reverse order:

$$\theta^{(n)}, \theta^{(n-1)}, \dots, \theta^{(0)}$$

This sequence is also a Markov chain, since:

$$Pr(\theta^{(i)} = y | \theta^{(i+1)} = x_0, \theta^{(i+2)} = x_1, \dots) = Pr(\theta^{(i)} = y | \theta^{(i+1)} = x_0)$$

Therefore, the reverse transition probability may be re-expressed by applying Bayes' formula:

$$\begin{aligned} p_i(x, y) &= Pr(\theta^{(i)} = y | \theta^{(i+1)} = x) \\ &= \frac{Pr(\theta^{(i+1)} = x | \theta^{(i)} = y) Pr(\theta^{(i)} = y)}{Pr(\theta^{(i+1)} = x)} \\ &= \frac{p_i(y, x) \pi^{(i)}(y)}{\pi^{(i+1)}(x)} \end{aligned}$$

As $n \rightarrow \infty$ the chain becomes homogeneous, and is then considered to be *reversible*. A reversible chain satisfies the detailed balance equation:

$$\pi(x)p(x, y) = \pi(y)p(y, x), \forall x, y \in \mathcal{S}$$

Reversible chains are important because given *any* density π which obeys the detailed balance equation for an irreducible chain, then that chain is positive recurrent and has π as its stationary distribution. In other words, if you have a chain that is reversible, you get its stationary distribution for free!

2.3 Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) is the most general MCMC procedure, and therefore the most widely applicable. Recall that Gibbs sampling, the most popular sampling algorithm, relies on the conjugacy of prior and likelihood forms to facilitate posterior inference. When the prior and likelihood are not conjugate, a more general procedure such as Metropolis-Hastings is required.

If there exists some complex density $\pi(\theta)$ from which a sample is sought, a transition function is required to transit the current parameter value, θ (may be vector-valued, but I will consider a scalar for simplicity), to a new value. This may easily be done by constructing a reversible chain:

$$\pi(\theta)p(\theta, \phi) = \pi(\phi)p(\phi, \theta)$$

Though reversibility is not *necessary* to guarantee convergence of the posterior to π , it is certainly *sufficient*, as mentioned previously.

By implementing rejection sampling, the transition kernel $p(\theta, \phi)$ may be expressed as the product of an arbitrary proposal distribution, q , and an associated acceptance probability, α :

$$p(\theta, \phi) = q(\theta, \phi)\alpha(\theta, \phi), \forall \theta \neq \phi$$

As p is a density, this implies that the probability of no change in parameter value is:

$$p(\theta, \theta) = 1 - \int q(\theta, \phi)\alpha(\theta, \phi)d\phi$$

A logical choice for the acceptance probability is one which, in conjunction with the proposal density, satisfies the detailed balance equation. Hastings (1970) suggests the following:

$$\alpha(\theta, \phi) = \min\left[1, \frac{\pi(\phi)q(\phi, \theta)}{\pi(\theta)q(\theta, \phi)}\right]$$

It is this specification of transition kernel that characterizes the Metropolis-Hastings algorithmTM. Provided that q is irreducible and aperiodic, and α is a probability, then the chain is irreducible and aperiodic, with limiting distribution π as desired.

The algorithm straightforward:

1. Initialize $\theta^{(0)}$ arbitrarily
2. Repeat until convergence is satisfied:
 - (a) Propose a new value of θ : $\phi \sim q(\theta^{(i)}, \cdot)$
 - (b) Generate a uniform random deviate: $u \sim U[0, 1]$
 - (c) Accept the proposed move to ϕ if $u \leq \alpha(\theta^{(i)}, \phi)$, otherwise reject the transition

(d) If accepted, $\theta^{(i+1)} = \phi$, otherwise $\theta^{(i+1)} = \theta^{(i)}$

Notice that the acceptance of proposed parameter values is a function of the ratio of the proposed and current densities. If this ratio is low, the chain may stay in the current state for many iterations; thus, a high acceptance rate is guaranteed by a proposal distribution that generates values relatively "close to" the current value. On the other hand, very small moves inhibit a thorough exploration of the parameter space, which can severely slow convergence. The key to an efficient proposal distribution is a balance between the probability of acceptance and the rate of exploration.

3 Markov Chains with Jumps

Reversible jump MCMC implements the techniques outlined above to provide an algorithm for model selection from a potentially massive set of candidate models (Green, 1995). There are no special requirements for candidate models; they need not be nested, nor of similar functional form. In the reversible jump algorithm, the Markov chain "jumps" between parameter subspaces (models) of differing dimensionality, thereby generating samples from the joint distribution of parameters and model indices.

Consider a countable collection of candidate models, $\{m : m = 1, 2, \dots, M\}$, each having an associated vector of parameters θ_m of dimension d_m , which typically varies across models. We would like use MCMC to sample from the joint distribution:

$$\pi(\theta_m, m | \mathbf{y}) \propto p(\theta_m | m, \mathbf{y})p(m)$$

As was illustrated above, a convenient way to do this is to formulate a Markov chain that satisfies the detailed balance equation, here expressed as:

$$\pi(\theta, j)p((\theta, j), (\phi, k)) = \pi(\phi, k)p((\phi, k), (\theta, j))$$

where $p((\theta, j), (\phi, k))$ is the transition kernel for "jumping" from model j with associated parameters θ to model k and parameters ϕ . I have dropped the parameters subscripts for notational simplicity.

Also required are an arbitrary proposal distribution $q((\theta, j), (\phi, k))$ and acceptance probability $\alpha((\theta, j), (\phi, k))$. Note that according to the potential values of ϕ , there is a potentially infinite number of possible moves to model k from j . To accommodate this, we can specify:

$$B_{jk} = \int q((\theta, j), (\phi, k))d\phi$$

as the probability of jumping to model k from j , regardless of the parameters of the proposed model. In turn, this quantity can be used to specify a proposal transition function for θ , using the definition of conditional probability:

$$\beta_{\theta\phi} = q((\theta, j), (\phi, k))/B_{jk}.$$

However, one may consider the entire set of parameters (super-parameter) as partitioned into $\theta = (\theta_k, \theta_{-k})$; that is, those parameters which specify model k and those which do not. Given this, the actual probability of moving to ϕ_k is given by:

$$g(\phi_k) = q((\theta_k, \theta_{-k}, j), (\phi, \theta_{-k}, k)) / B_{jk}$$

A special concern arises when the dimensions of the current and proposed parameter spaces are unequal. Reversibility is only guaranteed when the parameter transition function is a bijection¹. If, for example, $d_k > d_j$, an additional random quantity \mathbf{u} of dimension $d_k - d_j$ is required to satisfy the bijection $\phi = g(\theta, \mathbf{u})$, allowing moves to be made in both directions with the same probability.

Of course, the chain may not jump from model j for every iteration of the algorithm; this may happen by either of two circumstances. First, a model jump may not be proposed for the current iteration, implying that $\sum_m B_{jm} = B_j \leq 1$. Second, the proposed move to model k may be rejected by the algorithm. So, the full transition kernel is a mixed distribution that admits a density for ϕ , a transition function for model k , and a probability that no moves are taken. The acceptance probability follows that of the Metropolis-Hastings algorithm:

$$\alpha((\theta, j), (\phi, k)) = \min \left[1, \frac{\pi(\phi, k)q((\phi, k), (\theta, j))}{\pi(\theta, j)q((\theta, j), (\phi, k))} \right]$$

The reversible jump MCMC algorithm is summarized by the following:

1. Initialize $(\theta^{(0)}, j)$ arbitrarily
2. Repeat until convergence is satisfied:
 - (a) Generate a uniform random deviate: $u_1 \sim U[0, 1]$
 - (b) Choose a proposed jump using probabilities $\{B_{jm}\}$; if $B_{jk} \leq u_1$ (jump is rejected), set $(\theta^{(i+1)}, j^{(i+1)}) = (\theta^{(i)}, j^{(i)})$ and begin a new iteration
 - (c) Generate ϕ_k from $g(\phi_k)$ and a second uniform random deviate u_2
 - (d) Calculate the acceptance probability $\alpha((\theta^{(i)}, j^{(i)}), (\phi, k))$; if $u_2 \leq \alpha$ then $(\theta^{(i+1)}, j^{(i+1)}) = (\phi, k)$, otherwise reject the transition and set $(\theta^{(i+1)}, j^{(i+1)}) = (\theta^{(i)}, j^{(i)})$

¹A function is bijective or a bijection or a one-to-one correspondence if it is both injective (no two values map to the same value) and surjective (for every element of the codomain there is some element of the domain which maps to it). That is, there is exactly one element of the domain which maps to each element of the codomain. Only bijective functions have inverses f' where $f(f'(x)) = f'(f(x)) = x$.

References

- Carlin, B. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, 57:473–484.
- Gamerman, D. (1997). *Markov Chain Monte Carlo: statistical simulation for Bayesian inference*. Chapman and Hall, London, first edition.
- Green, P. (1995). Reversible jump Markov chain Monte Carlo. *Biometrika*, 82:711–732.
- Hastings, W. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57:97–109.
- King, R. and Brooks, S. (2002). Model Selection for Integrated Recovery/Recapture Data. *Biometrics*.
- Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equations of state calculations by fast computing machine. *J. Chem. Phys.*, 21:1087–1091.