A Very Brief Introduction to MCMC

Malcolm Farrow
University of Newcastle upon Tyne

August 17, 2010

1 Monte Carlo Methods

1.1 Introduction

Suppose we have a (vector) unknown (parameter) θ, prior density $p(\theta)$ and likelihood $L(\theta; x)$. The main computational problem in Bayesian inference is usually the evaluation of an integral

$$\int_\theta q(\theta)p(\theta)L(\theta;x).d\theta \tag{1}$$

for some $q(\theta)$.

For example, if $q(\theta) = 1$ then (1) gives C^{-1} where C is the constant of proportionality for the posterior density. If $q(\theta)$ is some function of θ then the result of (1) multiplied by C gives the posterior expectation of that function.

We can evaluate the integral (1) approximately by choosing a set of values, $\theta_1, \ldots, \theta_n$ of θ, called *nodes*, and calculating

$$\sum_{i=1}^{n} w_i q(\theta_i)p(\theta_i)L(\theta_i;x) \tag{2}$$

where $w_1, \ldots, w_n$ are *weights.* For example, suppose θ is a scalar unknown such that $0 < \theta < 1$. We might choose $\theta_i = (i - 0.5)/n$ and $w_i = 1/n$. E.g., with $n = 1000$, we would have $\theta = 0.0005, 0.0015, 0.0025, \ldots, 0.9995$ and $w_i = 0.001$. Here w_i is the step size. If we generalise this to a vector of parameters then w_i is the product of the step sizes. More generally, the points need not be equally spaced so the w_i need not be equal.

We might want to place more nodes in the important parts of the parameter space, that is where the function we are integrating is greatest. One way to do this is to generate nodes as random samples from a distribution with pdf $g(\theta)$ and let $w_i = [ng(\theta_i)]^{-1}$. Then

$$\int_\theta q(\theta)p(\theta)L(\theta;x).d\theta = \int_\theta g(\theta)q(\theta)p(\theta)L(\theta;x)/g(\theta).d\theta$$

and we can think of this as the expected value of $q(\theta)p(\theta)L(\theta;x)/g(\theta)$ over the distribution with pdf $g(\theta)$. So we take random samples from the distribution with pdf $g(\theta)$ and average the values of $q(\theta)p(\theta)L(\theta;x)/g(\theta)$ obtained. Thus (2) will tend to approximate (1) for large n. The variance of (2) will be

$$n \operatorname{var}\left[\frac{q(\theta)p(\theta)L(\theta;x)}{ng(\theta)}\right] = \frac{1}{n}\operatorname{var}\left[\frac{q(\theta)p(\theta)L(\theta;x)}{g(\theta)}\right].$$

This will be small if $g(\theta)$ has a similar shape to $|q(\theta)p(\theta)L(\theta;x)|$ in which case sampling in this way is called *importance sampling.*

1.2 Example

Suppose we take $n+1$ observations $y_0, \ldots, y_n$ on a stationary first order autoregressive process

$$Y_i - \mu = \phi(Y_{i-1} - \mu) + \varepsilon_i$$

where $\varepsilon \sim N(0, \sigma^2)$.

The likelihood is

$$\begin{aligned} L(\theta;y) &= (2\pi)^{-(n+1)/2}\left(\frac{\sigma^2}{1-\phi^2}\right)^{-1/2}\exp\left[-\frac{1}{2\sigma^2/(1-\phi^2)}(y_0-\mu)^2\right] \\ &\quad \times(\sigma^2)^{-n/2}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\{y_i-\mu-\phi(y_{i-1}-\mu)\}^2\right] \end{aligned}$$

For large n and reasonably vague prior the posterior density is dominated by

$$(\sigma^2)^{-n/2}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\{y_i-\mu-\phi(y_{i-1}-\mu)\}^2\right].$$

We can approximate this shape by giving $p=(\sigma^2)^{-1}$ a gamma distribution, gamma$(1+n/2,(1/2)\sum[y_i-\mu^\star-\phi^\star y_{i-1}]^2)$ and, independently, ϕ and $\tilde{\mu}=\mu(1-\phi)$ a normal distribution with mean $\phi^\star,\mu^\star$, the least squares estimates of ϕ and $\tilde{\mu}$ based on regressing y_i on y_{i-1}, and variance matrix given by that of the least squares estimates.

Now suppose, for example, we want to find the marginal posterior distribution of ϕ. Then

$$q(\theta)=I_u=\begin{cases}1 & \phi<u\\ 0 & \phi\geq u\end{cases}$$

may be evaluated (for each of a range of values of u) along with $w_i, p(\theta_i)$ and $L(\theta_i;y)$ and the sum (2) formed by sampling $p,\phi,\tilde{\mu}$.

2 The Gibbs Sampler

2.1 Introduction

Fortunately it is no longer necessary to consider apparently complicated approximation methods as in the previous section every time we want to make an inference. Use of the Gibbs sampler (and other *Markov chain Monte Carlo* methods) has been developed so that the computation for models of many types can now be done relatively straightforwardly. In particular software is available to do this and, in this course, we will be using the software package "BUGS" (Bayes Using Gibbs Sampling).

Suppose we have two unknowns, X,Y. It may be difficult to write down their joint density function or integrate over it but suppose that for any value y we can sample from the conditional distribution of $X|Y=y$ and vice versa. Suppose we want to find the expected value of $q(X,Y)$. (This includes most, if not all, of the things we might want to find). Then the Gibbs sampling algorithm works as follows.

1. Choose starting values (x_0,y_0). (Let $i=1$).
2. Take a sample Y_i from the conditional distribution of $Y|X=x_{i-1}$.
3. Take a sample X_i from the conditional distribution of $X|Y=y_i$.
4. Go to 2. (Let i become $i+1$).

Actually the way a program like BUGS works involves a lot of details to do with how the model is specified and how samples are taken from the conditional distributions. For these you are referred to the BUGS manual, Gilks *et al* (1996) and the BUGS web site.

"Eventually" the sample (X_i,Y_i) are from the joint distribution of X,Y. Thus, having allowed for a suitable run-in period, we may accumulate values of $q(X,Y)$ and find an average (or even plot a histogram).

In fact X,Y may each be vectors (i.e. sets of unknowns) and we may have more than two such sets and cycle around them.

For further comments see Shaw (1993), Gilks *et al* (1996) etc. In particular there are important issues to do with convergence.

2.2 Examples

2.2.1 Example 1 (very simple)

This example is so simple that we would not really use Gibbs sampling in this case. However it illustrates how it works.

We have observations $y_1, \ldots, y_n$ from a normal distribution with mean μ and variance $\sigma^2 = p^{-1}$. The observations are independent given μ and p. We have independent priors for μ and p such that $\mu \sim N(\mu_0, \pi_0^{-1})$ and $p \sim$ gamma(α, β) (an "inverse gamma" prior for σ^2). Then the joint prior density is proportional to

$$p^{\alpha-1} \exp(-\beta p) \exp\{-\pi_0(\mu - \mu_0)^2/2\}.$$

the likelihood is proportional to

$$p^{n/2} \exp\left\{-p \sum (y_i - \mu)^2/2\right\}.$$

The posterior density is proportional to

$$p^{\alpha-1+n/2} \exp\left\{-\beta p - \left[\pi_0(\mu - \mu_0)^2 + p \sum (y_i - \mu)^2\right]/2\right\}.$$

Suppose we fix a value for μ, say μ_1. Then the conditional posterior density for p is proportional to

$$p^{\alpha-1+n/2} \exp\left\{-\left[\beta + \sum (y_i - \mu)^2/2\right] p\right\}.$$

This is a gamma$(\alpha + n/2, \beta + \sum [y_i - \mu]^2/2)$ distribution and we sample a value, say p_1, from this.

Next we fix $p = p_1$. Then the conditional posterior density for μ is proportional to

$$\exp\left\{-\left[\pi_0(\mu - \mu_0)^2 + p_1 \sum (y_i - \mu)^2\right]/2\right\}.$$

That is

$$\exp\left\{-\left[\mu^2(\pi_0 + np_1) - 2\mu\left(\pi_0\mu_0 + p_1 \sum y_i\right) + \pi_0\mu_0^2 + p_1 \sum y_i^2\right]/2\right\}.$$

This is proportional to

$$\exp\left\{-(\pi_0 + np_1)\left[\mu - \left(\pi_0\mu_0 + p_1 \sum y_i\right)(\pi_0 + np_1)^{-1}\right]^2/2\right\}.$$

This is a $N([\pi_0\mu_0 + p_1 \sum y_i][\pi_0 + np_1]^{-1}, [\pi_0 + np_1]^{-1})$ distribution and we sample a value, say μ_2, from this.

We can then sample a new value for p, given $\mu = \mu_2$ and so on.

2.2.2 Example 2 (Taken from Gelfand et al, 1990)

Suppose we have k samples of observations and observation j in sample i is

$$Y_{ij} = \theta_i + \varepsilon_{ij}$$

for $j = 1, \ldots, J$, where $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$ and $\theta_i \sim N(\mu, \sigma_\theta^2)$ (all independent).

Suppose we have independent priors for the three parameters:

$$\begin{aligned} \mu &\sim N(\mu_0, \sigma_0^2) \\ \sigma_\theta^2 &\sim \text{IG}(a_1, b_1) \\ \sigma_\varepsilon^2 &\sim \text{IG}(a_2, b_2) \end{aligned}$$

where IG stands for "inverse gamma" (i.e. $(\sigma^2)^{-1}$ has a gamma distribution).

Then

$$
\begin{aligned}
\sigma_\theta^2 | Y, \mu, \theta, \sigma_\varepsilon^2 &\sim \mathrm{IG}(a_1 + k/2, b_1 + (1/2) \sum [\theta_i - \mu]^2) \\
\sigma_\varepsilon^2 | Y, \mu, \theta, \sigma_\theta^2 &\sim \mathrm{IG}(a_2 + kJ/2, b_2 + (1/2) \sum \sum [Y_{ij} - \theta_i]^2) \\
\mu | Y, \theta, \sigma_\theta^2, \sigma_\varepsilon^2 &\sim N\left(\frac{\sigma_\theta^2 \mu_0 + \sigma_0^2 \sum \theta_i}{\sigma_\theta^2 + k\sigma_0^2}, \frac{\sigma_\theta^2 \sigma_0^2}{\sigma_\theta^2 + k\sigma_0^2} \right) \\
\theta | Y, \mu, \sigma_\theta^2, \sigma_\varepsilon^2 &\sim N\left(\frac{J\sigma_\theta^2}{J\sigma_\theta^2 + \sigma_\varepsilon^2} \bar{Y} + \frac{\sigma_\varepsilon^2}{J\sigma_\theta^2 + \sigma_\varepsilon^2} \mu \underline{1}, \frac{\sigma_\theta^2 \sigma_\varepsilon^2}{J\sigma_\theta^2 + \sigma_\varepsilon^2} I \right)
\end{aligned}
$$

where $\underline{1}$ is a $k \times 1$ vector of 1's and I is a $k \times k$ identity matrix.

2.3 Convergence

One of the main difficulties with the Gibbs sampler and other Markov chain Monte Carlo methods is convergence. We wish to obtain a sample of values from the posterior distribution. We know that the equilibrium distribution of the chain is the required posterior distribution but how do we know that the chain has reached equilibrium?

Of course we use a "burn-in" period when we take samples but discard them before we start collecting samples. In some straightforward models a burn-in of a few hundred samples may be reasonably assumed to be sufficient. In more complicated cases, involving posterior distributions with "diagonal" ridges or more than one local maximum, the chain can take a long time to move around the whole parameter space and a much longer burn-in may be required. We may also require a much larger sample to be sure that the chain has not just been temporarily "stuck" in one part of the parameter space.

Because of these difficulties much work has been done on *convergence diagnostics* and other output analysis. For further information, see, e.g., the CODA manual (Best, Cowles and Vines, 1995), section 4.

References

Best, N., Cowles, M.K. and Vines, K., 1995, CODA: Convergence Diagnostics and Output Analysis Software for Gibbs Sampling Output, Version 0.30. Available from BUGS Web Site.

BUGS Examples vols. 1 and 2: `http://www.mrc-bsu.cam.ac.uk/bugs/doc/doc.html`

BUGS Manual: `http://www.mrc-bsu.cam.ac.uk/bugs/doc/doc.html`

BUGS Web Site: `http://www.mrc-bsu.cam.ac.uk/bugs/`

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, *Journal of the American Statistical Association*, **85**, 972-985.

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

Shaw, J.E.H., 1993, Statistical computing for Bayesian applications. *Handout for R.S.S. One-day Workshop, 11th May 1993. (Attached).*