

## Lecture Note 7: Markov Chain Monte Carlo

This note is based in part on Geweke, Ch. 4, and Chib and Greenberg.

### Gibbs Sampler

As before, let  $\theta$  be the random vector we wish to simulate, and let  $p(\theta)$  be its density. The target density could be a posterior density  $p(\theta|z)$ , but we suppress any conditioning in the notation for simplicity.

In many cases we will not be able to generate random draws directly from  $p(\theta)$ . The Gibbs sampler can be used instead of pure random sampling, if various conditional distributions are tractable.

Let the vector  $\theta$  be partitioned into subvectors:

$$\theta = \begin{pmatrix} \theta_{(1)} \\ \vdots \\ \theta_{(B)} \end{pmatrix}.$$

Here, each  $\theta_{(b)}$  for  $b = 1, \dots, B$  is a subvector (or “block”) of  $\theta$ .

The key is to choose a blocking such that each conditional distribution

$$p(\theta_{(b)}|\theta_{(1)}, \dots, \theta_{(b-1)}, \theta_{(b+1)}, \dots, \theta_{(B)})$$

can be sampled from. Although this might seem like a strong requirement, we will see some examples below where this can be done.

The Gibbs sampler generates a sequence of draws  $\theta^{(j)} = (\theta_{(1)}^{(j)}, \dots, \theta_{(B)}^{(j)})'$ ,  $j = 0, \dots, J$ , as follows:

- Initialize the vector:  $\theta^{(0)} = (\theta_{(1)}^{(0)}, \dots, \theta_{(B)}^{(0)})'$  according to some distribution.
- Draw  $\theta_{(1)}^{(1)} \sim p(\theta_{(1)}|\theta_{(2)}^{(0)}, \dots, \theta_{(B)}^{(0)})$ .
- Draw  $\theta_{(2)}^{(1)} \sim p(\theta_{(2)}|\theta_{(1)}^{(1)}, \theta_{(3)}^{(0)}, \dots, \theta_{(B)}^{(0)})$ .
- $\vdots$
- Draw  $\theta_{(B)}^{(1)} \sim p(\theta_{(B)}|\theta_{(1)}^{(1)}, \dots, \theta_{(B-1)}^{(1)})$ .
- Draw  $\theta_{(1)}^{(2)} \sim p(\theta_{(1)}|\theta_{(2)}^{(1)}, \dots, \theta_{(B)}^{(1)})$ .

and so on. Under some conditions, it can be shown that:

1. For large  $J$ ,  $\theta^{(J)} = (\theta_{(1)}^{(J)}, \dots, \theta_{(B)}^{(J)})'$  is approximately a draw from  $p(\theta)$ .
2. Ergodic averages converge:

$$\frac{1}{J} \sum_{j=1}^J g(\theta^{(j)}) \xrightarrow{\text{as}} \int g(\theta) p(\theta) d\theta.$$

To get some intuition for why this algorithm works, suppose that at step  $j$ ,  $\theta^{(j)}$  is a draw from  $p(\theta)$ . Then, at step  $j + 1$ ,

$$\theta_{(1)}^{(j+1)} \sim p(\theta_{(1)} | \theta_{(2)}^{(j)}, \dots, \theta_{(B)}^{(j)})$$

is, marginally, a draw from the marginal density

$$p(\theta_{(1)}) = \int \dots \int p(\theta) d\theta_{(2)} \dots d\theta_{(B)}.$$

Similar,  $\theta_{(2)}^{(j+1)}$  is a draw from the correct marginal distribution. So the entire vector

$$\theta^{(j+1)} \sim p(\theta).$$

This shows that  $p(\theta)$  is a *stationary* distribution for this process.

It is usually better to use a small number of blocks (lower  $B$ ) if possible. If there is only one block, so that  $\theta = (\theta_{(1)})$ , then Gibbs sampling would be equivalent to direct sampling from  $p(\theta)$ .

## Examples of Gibbs Sampling

### 1. Data Augmentation

By redefining  $\theta$  in the Gibbs sampling algorithm to include both the model parameters and the latent data  $z_i$ , we see that data augmentation is a special case of the Gibbs sampler. Here are some other examples of models with latent variables, for which the data augmentation principle could be used:

- (a) Tobit Model:

$$y_i^* = x_i' \beta + \epsilon_i, \quad \epsilon_i | x \stackrel{\text{iid}}{\sim} N(0, \sigma^2),$$
$$y_i = \max\{0, y_i^*\}.$$

(b) Stochastic Volatility Model: this model is often used in finance to model autocorrelation in asset volatility.

$$y_t = u_t \sqrt{h_t}, \quad u_t \stackrel{\text{iid}}{\sim} N(0, 1),$$

$$\log h_t = \alpha + \delta \log h_{t-1} + \eta_t, \quad \eta_t \stackrel{\text{iid}}{\sim} N(0, \sigma_h^2).$$

Here, only  $\{y_t\}$  is observed.

2. Normal Model: let  $Y_i \stackrel{\text{iid}}{\sim} N(\mu, \tau^{-1})$  for  $i = 1, \dots, n$ . Assume the standard diffuse prior  $p(\mu, \tau) \propto \frac{1}{\tau}$ . Here,  $\theta = (\mu, \tau)$  and our target is  $p(\mu, \tau|z)$ , where  $z = (y_1, \dots, y_n)$ .

We could use the results of LN4 to simulate directly from the posterior distribution, but as an illustration we show how to do Gibbs sampling for this model.

Our blocking is  $\theta_{(1)} = \mu$ ,  $\theta_{(2)} = \tau$ . We need to figure out the conditional distributions

$$p(\mu|\tau, z)$$

and

$$p(\tau|\mu, z).$$

First consider  $p(\mu|\tau, z)$ . By the results in LN4,

$$\mu|\tau, z \sim N(\bar{y}, (\tau \cdot n)^{-1}).$$

Next consider  $p(\tau|\mu, z)$ . We start by calculating the joint density:

$$\begin{aligned} p(\tau, \mu|z) &\propto p(\mu, \tau) f(z|\mu, \tau) \\ &= \frac{1}{\tau} \prod_{i=1}^n \frac{\tau^{1/2}}{\sqrt{2\pi}} \exp\left(-\frac{\tau}{2}(y_i - \mu)^2\right) \\ &\propto \frac{1}{\tau} \tau^{n/2} \exp\left(-\frac{\tau}{2} \sum_i (y_i - \mu)^2\right) \\ &= \tau^{n/2-1} \exp\left(-\frac{\tau}{2} \sum_i (y_i - \mu)^2\right) \end{aligned}$$

Since

$$p(\tau|\mu, z) = \frac{p(\tau, \mu|z)}{p(\mu|z)},$$

we can write

$$p(\tau|\mu, z) \propto \tau^{n/2-1} \exp\left(-\frac{\tau}{2} \sum_i (y_i - \mu)^2\right).$$

So,

$$\tau|\mu, z \sim \frac{\chi_n^2}{\sum_i (y_i - \mu)^2}.$$

Notice that in direct sampling, we would draw from  $p(\tau|z)$ , then  $p(\mu|\tau, z)$ , generating a single joint draw from  $p(\mu, \tau|z)$ . In direct sampling, the draw from the (marginal) posterior for  $\tau$  is

$$\tau|z \sim \frac{\chi_{n-1}^2}{(n-1)s^2} = \frac{\chi_{n-1}^2}{\sum_i (y_i - \bar{y})^2}.$$

Compare this to the conditional posterior for  $\tau$ , which uses the current draw for  $\mu$ . This induces some autocorrelation in the Gibbs sampling (the current draw for  $\mu$  depends on the past draw for  $\tau$ , and vice versa), so the Gibbs sampler will typically require more iterations to give a good approximation to the posterior distribution.

## Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm is useful when the target density  $p(\theta)$  can be calculated up to a normalizing constant.

Let  $q(\theta^*|\theta)$  be a transition probability density function, which specifies a distribution for a new draw for  $\theta$  given a current draw for  $\theta$ . The transition density does not have to be related in a particular way to the target density, but we need to be able to generate random draws from it, and evaluate it.

The Metropolis-Hastings algorithm proceeds as follows:

First, initialize the process at some  $\theta^{(0)}$ . For  $j = 1, \dots, J$ :

- Generate a candidate draw  $\theta^* \sim q(\theta^*|\theta^{(j-1)})$ .
- “Accept” the draw by setting  $\theta^{(j)}$  with probability

$$\alpha(\theta^*|\theta^{(j-1)}) = \min \left\{ \frac{p(\theta^*)/q(\theta^*|\theta^{(j-1)})}{p(\theta^{(j-1)})/q(\theta^{(j-1)}|\theta^*)}, 1 \right\}.$$

Otherwise set  $\theta^{(j)} = \theta^{(j-1)}$ .

Thus, the MH algorithm will sometimes stay at the same value of  $\theta^{(j)}$  for two or more iterations.

One common choice for the transition density  $q(\theta^*|\theta)$  is a random walk: we could let

$$\theta^* = \theta^{(j-1)} + \epsilon,$$

where  $\epsilon \sim N(0, \sigma^2)$  (or some other distribution symmetric about 0). Then

$$\theta^* | \theta^{(j-1)} \sim N(\theta^{(j-1)}, \sigma^2),$$

and  $q(\cdot|\cdot)$  is a normal density. This is particularly convenient because then  $q(\theta^*|\theta) = q(\theta|\theta^*)$ , so that the formula for the acceptance probability  $\alpha$  does not require computation of  $q$ .

One application we have looked at where the MH algorithm may be useful is the logit model. The logit model can be written in latent variable form

$$\begin{aligned} y_i^* &= x_i' \beta + \epsilon_i, \\ y_i &= 1(y_i^* \geq 0). \end{aligned}$$

However, in the logit case, the disturbances  $\epsilon_i$  would have a logistic distribution, not a normal distribution. For a data-augmentation type algorithm to work, we would need to draw from the conditional distribution of  $\beta$  given the  $x_i$  and  $y_i^*$ , but there is not a simple way to do so. So data augmentation/Gibbs sampling is not convenient here. However, it is relatively easy to calculate the log-likelihood (and hence the unnormalized posterior density) for the logit model. So Metropolis-Hastings is relatively easy to implement.

### Mixed Gibbs-MH Algorithm

Sometimes there is a blocking  $\theta = (\theta_{(1)}, \dots, \theta_{(B)})$  such that most of the conditional distributions  $p(\theta_{(b)} | \theta_{-(b)})$  are easy to draw from, but some are not. If there is some  $b$  such that  $p(\theta_{(b)} | \theta_{-(b)})$  is not tractable, but can be calculated up to a normalizing constant, we could replace the exact draw for  $\theta_{(b)}$  with a Metropolis-Hastings draw. We will see an example of this strategy later on.

### General State Space Markov Chain Theory

Consider a general stochastic process  $\{\theta^{(j)}\}_{j=0}^{\infty}$ . The distribution of  $\theta^{(j+1)}$  given the past values can be characterized by

$$Pr(\theta^{(j+1)} \in A | \theta^{(0)}, \dots, \theta^{(j)})$$

where  $A$  is a measurable subset of  $\Theta$ . In the Gibbs Sampler, the Metropolis-Hastings algorithm, and other Markov Chain Monte Carlo algorithms, we can write

$$\begin{aligned} Pr(\theta^{(j+1)} \in A | \theta^{(1)}, \dots, \theta^{(j)}) &= Pr(\theta^{(j+1)} \in A | \theta^{(j)}) \\ &= \int_A k(\theta^{(j+1)} | \theta^{(j)}) d\theta^{(j+1)}, \end{aligned}$$

where  $k$  is a transition kernel or transition density, which does not depend on  $j$ .

Let  $f_j(\theta)$  denote the marginal density of  $\theta^{(j)}$ . If  $\theta^{(0)}$  is deterministic, then

$$f_1(\theta) = k(\theta|\theta^{(0)}),$$

$$f_2(\theta) = \int_{\Theta} k(\theta|\theta^{(1)})f_1(\theta^{(1)})d\theta^{(1)},$$

and in general,

$$f_j(\theta) = \int_{\Theta} k(\theta|\theta^{(j-1)})f_{j-1}(\theta^{(j-1)})d\theta^{(j-1)}.$$

As a shorthand, we can write this last equation (which determines the function  $f_j(\cdot)$  from  $f_{j-1}(\cdot)$ ) as

$$f_j = K f_{j-1}.$$

Then an invariant density is a density  $f^*$  such that

$$f^* = K f^*.$$

Thus we see that an invariant density is a fixed point of the linear operator  $K$ .

We say that  $K$  is  $f^*$ -irreducible if, for any measurable set  $A \subset \Theta$  such that

$$\int_A f^*(\theta)d\theta > 0,$$

we have, for some  $j \geq 1$  and all  $\theta^{(0)}$ ,

$$Pr(\theta^{(j)} \in A|\theta^{(0)}) > 0.$$

We also say that  $K$  is aperiodic if the chain does not cycle through a finite number of disjoint subsets of  $\Theta$ .

**Theorem:** if  $K$  is  $f^*$ -irreducible and has invariant density  $f^*$ , then  $f^*$  is the unique invariant density of  $K$ . If  $K$  is also aperiodic, then for  $f^*$ -almost every  $\theta^{(0)}$  and all measurable sets  $A \subset \Theta$ ,

$$\left| \int_A f_J(\theta)d\theta - \int_A f^*(\theta)d\theta \right| \rightarrow 0$$

as  $J \rightarrow \infty$ , which implies convergence in distribution of the marginal densities; and for all  $g$  integrable with respect to  $f^*$ ,

$$\frac{1}{J} \sum_{j=1}^J g(\theta^{(j)}) \xrightarrow{\text{as}} \int g(\theta)f^*(\theta)d\theta.$$