

## Markov Chain Monte Carlo (MCMC)

MCMC is a method of approximating expectations of random variables. As the name implies, it uses Monte Carlo methods based on Markov chains.

Expectations of interest in the Bayesian statistics context:

- Posterior probabilities:

$$P(\boldsymbol{\theta} \in A | \mathbf{y}) = \int_{\Theta} I_A(\boldsymbol{\theta}) \pi(\boldsymbol{\theta} | \mathbf{y}) d\boldsymbol{\theta}$$

- Posterior mean of a component  $\theta_i$  of  $\boldsymbol{\theta}$ :

$$E_i = \int_{\Theta} \theta_i \pi(\boldsymbol{\theta} | \mathbf{y}) d\boldsymbol{\theta}$$

- Posterior variance of component  $\theta_i$ :

$$\int_{\Theta} (\theta_i - E_i)^2 \pi(\boldsymbol{\theta} | \mathbf{y}) d\boldsymbol{\theta}$$

## Markov chains

A *Markov chain* is a stochastic process  $\{X_1, X_2, \dots\}$  with the following property:

$$P(X_t \leq x_t | X_1 = x_1, \dots, X_{t-1} = x_{t-1}) = \\ P(X_t \leq x_t | X_{t-1} = x_{t-1})$$

for all  $t \geq 3$  and all choices of  $x_1, x_2, \dots, x_t$ .

In other words, the distribution of  $X_t$  given the whole history of the process is the same as the distribution of  $X_t$  given *just the most recent observation*.

For an arbitrary density  $g$ , MCMC provides a means of generating observations  $X_1, X_2, \dots$  that each have density  $g$ . In contrast to many familiar methods, though,  $X_1, X_2, \dots$  are *not* independent, but instead form a Markov chain.

Let  $h$  be a known function, and suppose we want to determine  $\mu = E[h(X)]$ , where  $X$  has density  $g$ .

Given  $X_1, X_2, \dots, X_n$  that have been generated from  $g$  using MCMC, we may approximate  $\mu$  using

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n h(X_i).$$

The fact that the  $X_i$ 's are not independent *does* have an effect on the accuracy of  $\hat{\mu}$ , but since we can generate as many  $X_i$ 's as we wish, this is usually not a big concern.

As with rejection sampling, an important aspect of MCMC is that it may be applied when  $g$  is known only up to a constant of proportionality.

This, of course, is the situation often faced in Bayesian analysis when we know

$$\pi^*(\boldsymbol{\theta}|\mathbf{y}) = f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}),$$

but not the constant  $m(\mathbf{y})$  such that

$$\pi^*(\boldsymbol{\theta}|\mathbf{y})/m(\mathbf{y})$$

is a density.

### *Metropolis-Hastings algorithm*

A very general MCMC algorithm is Metropolis-Hastings. Let  $g$  be the density from which observations are to be generated. The algorithm relies on a so-called *proposal distribution*, which we will denote  $q$ .

The proposal distribution is actually a *collection* of distributions. If  $x$  is any value in the support of  $g$ , then  $q(\cdot|x)$  is a density function. Specifically,  $q(\mathbf{y}|x) \geq 0$  for all  $x$  and  $\mathbf{y}$ , and

$$\int q(\mathbf{y}|x) d\mathbf{y} = 1 \quad \forall x.$$

Let  $\mathbf{X}_0, \mathbf{X}_1, \dots$  denote the random variables forming the Markov chain. Given a value of  $\mathbf{X}_0$ , the rest of the chain evolves as follows.

Suppose that  $\mathbf{X}_t$  has taken on value  $x_t$ . Then a value  $\mathbf{y}$  is generated from the proposal distribution  $q(\cdot|x_t)$ .

This candidate value  $\mathbf{y}$  is accepted with probability  $\alpha(x_t, \mathbf{y})$ , where

$$\alpha(\mathbf{u}, \mathbf{v}) = \min \left( 1, \frac{g(\mathbf{v})q(\mathbf{u}|\mathbf{v})}{g(\mathbf{u})q(\mathbf{v}|\mathbf{u})} \right).$$

If the candidate point is accepted, then  $\mathbf{X}_{t+1} = \mathbf{y}$ . Otherwise, the chain does not move, i.e.,  $\mathbf{X}_{t+1} = x_t$ .

We see that the previous algorithm only requires knowing  $g$  up to a constant of proportionality since it depends on  $g$  only through  $\alpha$ , and  $\alpha$  depends on  $g$  only through ratios of the form  $g(\mathbf{u})/g(\mathbf{v})$ .

The previous algorithm works in the following sense:

- For virtually any proposal distribution, the marginal distribution of  $X_t$  tends to  $g$  as  $t \rightarrow \infty$ .
- The chain has the property of being *ergodic*, which implies that a sample average as on p. 236N converges to  $E[h(X)]$ .

A partial justification for Metropolis-Hastings involves showing that if  $X_t$  has density  $g$ , then so does  $X_{t+1}$ . To prove this fact, note that

$$F(y|x) \equiv P(X_{t+1} \leq y | X_t = x) = \int_{-\infty}^y q(\theta|x)\alpha(x, \theta)d\theta + (1 - \rho(x))D_x(y),$$

where  $D_x$  is the cdf of a random variable that is degenerate at  $x$  and

$$\rho(x) = \int_{-\infty}^{\infty} q(\theta|x)\alpha(x, \theta) d\theta.$$

This says that the conditional distribution of  $X_{t+1}$  given  $X_t = x$  is a mixture of an absolutely continuous distribution and one that is degenerate at  $x$ , with the mixing proportions being  $\rho(x)$  and  $1 - \rho(x)$ .

We then have

$$\begin{aligned} & \int_{-\infty}^{\infty} g(x)F(y|x) dx = \\ & \int_{-\infty}^y \int_{-\infty}^{\infty} g(x)q(\theta|x)\alpha(x, \theta) dx d\theta \\ & + \int_{-\infty}^{\infty} g(x)(1 - \rho(x))D_x(y) dx. \end{aligned}$$

Now, the equation that defines  $\alpha$  implies that

$$g(x)q(\theta|x)\alpha(x, \theta) = g(\theta)q(x|\theta)\alpha(\theta, x).$$

Combining these results yields

$$\begin{aligned} & \int_{-\infty}^{\infty} g(x)F(y|x) dx = \\ & \int_{-\infty}^y \int_{-\infty}^{\infty} g(\theta)q(x|\theta)\alpha(\theta, x) dx d\theta \\ & + \int_{-\infty}^y g(x)(1 - \rho(x)) dx = \end{aligned}$$

$$\begin{aligned}
& \int_{-\infty}^y g(\theta) \int_{-\infty}^{\infty} q(x|\theta) \alpha(\theta, x) dx d\theta \\
& + \int_{-\infty}^y g(x)(1 - \rho(x)) dx = \\
& \int_{-\infty}^y g(\theta) \rho(\theta) d\theta + \int_{-\infty}^y g(x)(1 - \rho(x)) dx = \\
& \int_{-\infty}^y g(\theta) d\theta.
\end{aligned}$$

Since  $\int_{-\infty}^{\infty} g(x)F(y|x) dx$  is an expression for the cdf of  $X_{t+1}$  assuming that  $g$  is the density of  $X_t$ , the result is proven.

A complete justification for Metropolis-Hastings would involve showing that the marginal distribution,  $g_t$ , of  $X_t$  eventually gets sufficiently close to  $g$ .

Once  $g_t$  is close to  $g$ , then, for  $s > t$ ,  $g_s$  will be also according to the argument on pp. 240-242N.