Monte Carlo Markov Chains, simple examples in R

Danny Kaye

danny@danny-kaye.co.uk

Nottingham Trent University

Abstract

Bayesian statistics usually results in the production of either a posterior density or a sample from such a density. As the cases in which a realistic, analytic solution can be produced are few and, ideally the choice of model and prior should not be limited to those cases we are often faced with the production of a sample from the posterior using Monte Carlo Markov Chain (MCMC) methods.

Here we present simple examples of two MCMC systems, Metropolis Hastings and Gibbs, coded directly in R.

MCMC

Consider a sequence of random variables $X_0, X_1, \ldots$ representing the state of some system at times $0, 1, \ldots$. These define a stochastic process, said to be in state i at time n if $X_n = i$. Suppose the system is in state i at time n and the transition probability that it will be in state j at time $n+1$ is P_{ij}, then

$$P_{ij} \quad = \quad P\{X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ \ldots \ , X_1 = i_1, X_0 = i_0\}$$

.

If this depends only on the state i and on no previous state, i.e.

$$P_{ij} \quad = \quad P\{X_{n+1} = j | X_n = i\},$$

then this is called a Markov process. The values P_{ij} are called the transition probabilities of the Markov Chain and the process is said to have no memory in the sense that P_{ij} depends only on $X_n = i$.

MCMC

Within the state space of the process, if $(X_t)_{t\geq 0}$ is a Markov chain it is said that i leads to j (written as $i \rightarrow j$) if

$$P_i(X_t = j \text{ for some } t \geq 0) > 0.$$

Where $P_i(\cdot)$ is the probability of some event happening, given that the chain is in state i. It is held that i communicates with j (written $i \leftrightarrow j$) if both $i \leftarrow j$ and $i \rightarrow j$. This is an equivalence relation, having reflexivity, symmetry and transitivity. Sets of states that can communicate with each other can be considered to be equivalence classes. States belonging to different equivalence classes do not communicate, though one way transition is possible. If a Markov Chain has all its states belonging to one equivalence class it is said to be irreducible.

MCMC

In practical terms, if the process is not irreducible then it can become “stuck” in one section and cannot be said to produce a “fair” representation of the density. An irreducible process is one in which it is possible to reach any state, starting in any state.

It can be shown that the sample obtained from the MCMC tends to a sample from the distribution of interest after some (large) number of steps n.

Metropolis Hastings

Target density $\pi(x)$ known to proportionality
Construct a Markov process with probability of transition from state x to state y $q(x, y)$,
With $X_n = x_n$, generate a candidate value Y for X_{n+1}
This candidate is accepted with probability $\alpha(x, y)$ where

$$\alpha(x, y) = \begin{cases} \min\left\{\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1\right\} \text{ if } \pi(x)q(x, y) > 0, \\ 1 \text{ if } \pi(x)q(x, y) = 0. \end{cases}$$

$\alpha(x, y)$ is called the acceptance probability.

The acceptance probability only depends on π through the ratio $\pi(y)/\pi(x)$, so π is only needed to proportionality.

Metropolis Hastings - Special case

If the candidate generating density is symmetric $q(x, y) = q(y, x)$ and the probability of a move reduces to $\pi(y)/\pi(x)$, hence, if $\pi(y) \geq \pi(x)$, the chain moves to y, otherwise it moves with probability given by $\pi(y)/\pi(x)$.
If the jump is to a point with higher probability it is always accepted. If to a point with lower probability it is accepted with a non-zero probability defined by the target density.
If the Markov process is irreducible and the candidate generating density is symmetric, then $q(x, y) > 0$, $q(x, y) = q(y, x)$ for all x and y and

$$\alpha(x, y) = \min\left\{\frac{\pi(y)}{\pi(x)}, 1\right\}.$$

Problem

Consider some measurement, x, modelled as a value of a random variable X

$$X \sim N(\mu, \sigma^2)$$

μ represents the average value of the measurement,
σ^2 represents its variability.
prior knowledge about μ, $N(\mu_0, \sigma_0^2)$
$\mu_0 = 3.5, \ \sigma_0 = 1$
$\sigma^2 = 1$ is uninformative as a wide range $(0.5$ to 6.5 say$)$ of values are all quite likely.

Assume the parameter σ is known: $\sigma = 0.01$.

Sample Data

3.51	3.50	3.52	3.50	3.51
3.50	3.52	3.50	3.51	3.51

This can be summarised as

$$n = 10$$

$$\sum x_i = 35.08$$

Analytic approach

Inference about μ given $\sigma = 0.01$.
With prior $\mu \sim N(3.5, 1)$

$$p(\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\mu - 3.5)^2}$$

$\bar{x}$, the sample mean, is a sufficient statistic for μ.

$$\begin{aligned} l(\mu|\boldsymbol{x}) &\propto p(\bar{x}|\mu, \sigma^2) \\ &\propto \exp\left(\frac{-n}{2\sigma^2}(\mu - \bar{x})^2\right) \end{aligned}$$

Analytic approach

The posterior for μ is

$$\begin{aligned} p(\mu|\boldsymbol{x}) &\propto p(\mu)l(\mu|\boldsymbol{x}) \\ &\propto p(\mu)p(\bar{x}|\mu,\sigma^2). \end{aligned}$$

So, by inspection

$$p(\mu|\boldsymbol{x}) = \left(\frac{n}{2\pi\sigma^2}\right)^2 \exp\left(\frac{-n}{2\sigma^2}(\mu-\bar{x})^2\right) \qquad -\infty < \mu < \infty$$

Analytic approach

$\boldsymbol{x}$ is a random sample of size n from $N(\mu, \sigma^2)$ where σ is known and the prior for μ is $N(\mu_0, \sigma_0^2)$. The posterior is $\mu \sim N(\mu_n, \sigma_n^2)$ where

$$\mu_n = \frac{n\bar{x}/\sigma^2 + \mu_0/\sigma_0^2}{n/\sigma^2 + 1/\sigma_0^2}$$

and

$$\sigma_n^{-2} = n\sigma^{-2} + \sigma_0^{-2}.$$

The prior for μ is $N(3.5, 1)$, $\sigma = 0.01$ and $\bar{x} = 3.508$ so posterior belief about μ is as if μ had a $N(3.508, 0.00316^2)$ distribution.

Simulation

Construct a Metropolis-Hastings, MCMC where the target distribution is the posterior density given

$$
\begin{array}{rrcl}
\text{Prior} & p(\theta) & \sim & N(3.5, 1) \\
\text{Likelihood} & l(\theta|\boldsymbol{x}) & \propto & \exp\left(\frac{-n}{2\sigma^2}(\theta - \bar{x})^2\right) \\
 & & \propto & p(\bar{x}|\theta, \sigma)
\end{array}
$$

Simulation

So, using a symmetric generating density the probability of transition from θ_1 the current state to θ_2 the candidate state is

$$p(\theta_1, \theta_2) \quad = \quad \min\left(\frac{p(\theta_2)p(\bar{x}|\theta_2,\sigma)}{p(\theta_1)p(\bar{x}|\theta_1,\sigma)}, 1\right)$$

Simulation

```
sim_full <- function (theta=3.5, n = 10, w = 1000, l=f, p=dnorm)
{
  z <- rep(0, n)
  for(i in 1 : (w + n))
  {
    new <- theta + rnorm(1, 0, 0.01)
    prob <- (p(new) * l(new, n=n)) / (p(theta) * l(theta, n=n))
    alpha <- min(prob, 1)
    if(runif(1) < alpha) theta <- new
    if(i > w) z[i - w] <- theta
  }
  z
}
```

Simulation

Called with

```
sim_full()
```

we get

```
 [1] 3.506161 3.506161 3.506161 3.510729 3.510729 3.510729
 [9]  3.510729 3.510521 3.509618 3.509618
```

More interestingly

```
summary(sim_full(n=1000))
```

gives

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.507   3.508   3.508   3.508   3.508   3.509
```

Simulation

It is possible to draw a predictive sample from this density with the R commands

```
> samp<-sim_full(n=1000)
> y<-rnorm(1000, samp, 0.01)
> summary(y)
```

Giving

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.479   3.502   3.508   3.508   3.515   3.538
```

Gibbs samplers

The density is approximated in the following way:

1. Obtain all the full conditional densities in the form $f(\theta_i|\boldsymbol{\theta}_{-i})$ (where $\boldsymbol{\theta}_{-i}$ is the $n-1$ vector consisting of the n parameter vector of the distribution excluding θ_i).
2. Given values for the parameters $\theta_1^{(0)}, \theta_2^{(0)} \ldots \theta_n^{(0)}$ use the conditional distribution $g(\theta_1|theta_{-1})$ to obtain a new estimate for θ_1, $\theta_1^{(1)}$.
3. Use the estimated parameters $\theta_3^{(0)}, \theta_4^{(0)} \ldots \theta_n^{(0)}$ and the new value $\theta_1^{(1)}$ to obtain a new value for θ_2, $\theta_2^{(1)}$.
4. Repeat, to obtain a complete realisation of $\boldsymbol{\theta}^{(1)}$.

Gibbs samplers

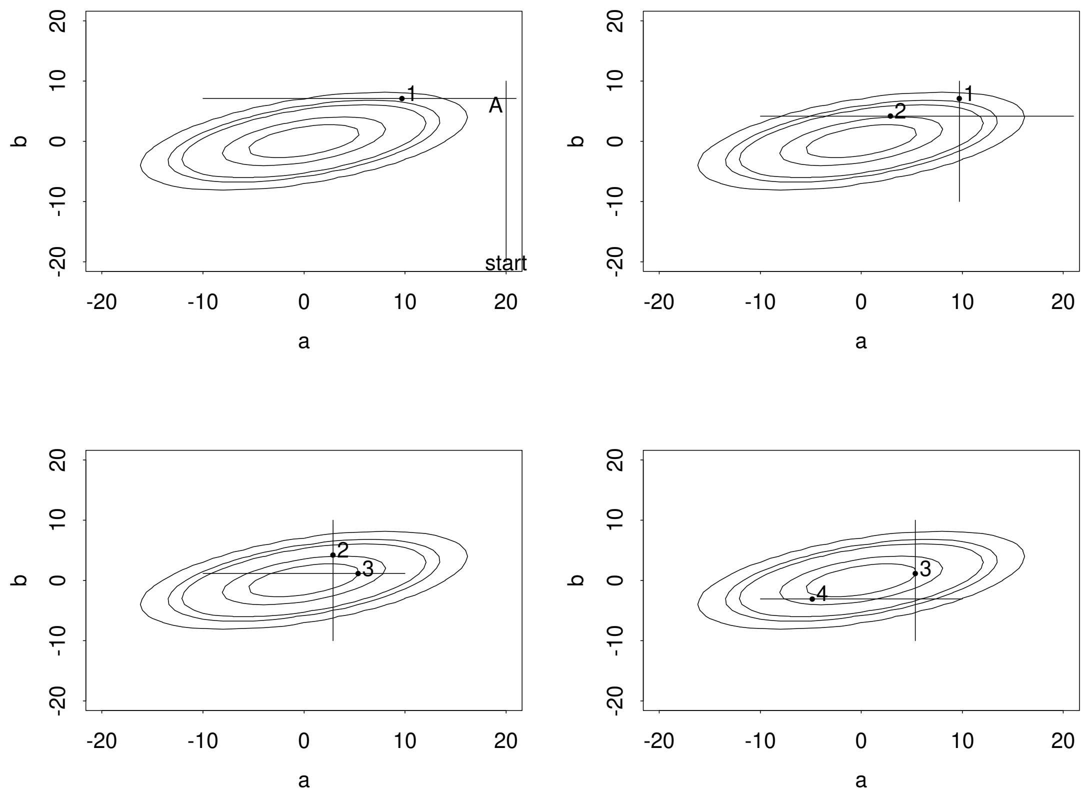

Figure 1: *The first 5 samples of a Gibbs sampler.*

Simulation

The density shown in the figure is a bivariate Normal with means zero, variances 2.5 and 5 and covariance $\rho = 0.5$. The well-known equations for the two conditional densities are:

$$p(X_2|X_1) = \frac{1}{\sqrt{2\pi(1-\rho^2)\sigma_2^2}} exp\left[-\frac{(x_2 - \rho(\sigma_2/\sigma_1)x_1)^2}{2(1-\rho^2)\sigma_2^2}\right],$$

the conditional expectations

$$E(X_2|X_1) = \rho(\sigma_2/\sigma_1)x_1,$$

$$E(X_1|X_2) = \rho(\sigma_1/\sigma_2)x_2,$$

and the variances

$$\text{var}\ (X_2|X_1) = (1-\rho^2)\sigma_2^2,$$

$$\text{var}\ (X_1|X_2) = (1-\rho^2)\sigma_1^2.$$

Simulation

```
gibbs <- function(x = c(10, 0, 5), y = c(-10, 0, 2.5),
                  rho = 0.5, n = 10, burn = 1000)
{
  FAC <- sqrt(1 - rho^2)
  out <- c(x[1], y[1])
  for(i in 1:n + burn)
  {
    ynew <- rnorm(1, (rho * y[3])/x[3] * x[1], FAC * y[3])
    xnew <- rnorm(1, (rho * x[3])/y[3] * ynew, FAC * x[3])
    if(i > burn)
    {
      x[1] <- xnew
      y[1] <- ynew
      out <- c(out, c(x[1], y[1]))
    }
  }
  out
}
```